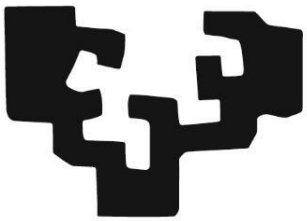


Inteligencia Computacional en Sistemas de Tele-asistencia en Domicilios

Memoria de Tesis Doctoral presentada por

Aitor Moreno Fdez. De Leceta

En el Departamento de Ciencias de la Computación e Inteligencia Artificial

Directores de tesis:

Prof. Manuel Graña at the University of the Basque Country and

Dr. Jose Manuel Lopez-Guede the University of the Basque Country

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Donostia-San Sebastián

2018

Agradecimientos.

Tras la gran aventura que ha supuesto para mí la realización de esta tesis doctoral, me gustaría mostrar mi agradecimiento a todas aquellas personas que, de una forma u otra, me han ayudado a alcanzar este objetivo.

Quisiera agradecer, en primer lugar, la paciente y valiosa dedicación de mis directores de tesis, Manuel Graña y Jose Manuel López, por sus consejos, por su paciencia, ayuda y por su buen hacer.

Gracias a mis padres y a mis hermanos, porque siempre me han apoyado en todas mis decisiones, y me han animado a seguir trabajando a pesar de las dificultades.

Agradecer especialmente a Maite y a Juan, por su paciencia y ayuda a la hora de poner en práctica las diferentes fases del proyecto. Gracias por estar ahí siempre que os necesito.

A todos mis compañeros de I3b, que siempre me han dado muestras de apoyo, especialmente a Nora y a Iñaki, que desde el primer día, han creído en mí a ciegas. Si no hubiera sido por ellos, esto no hubiera sido posible. Estamos viviendo unos tiempos apasionantes gracias a la Inteligencia Artificial.

Y por último, quiero agradecer y dedicar este trabajo a las personas que han decidido pasar la vida conmigo. Ellos saben más que nadie el trabajo que esta tesis lleva detrás, y sin su ayuda, apoyo incondicional, y soporte de todo tipo, con todo el cariño del mundo, no hubiera podido siquiera comenzar este largo camino. Gracias por compartir los enfados, berrinches, cansancios, ausencias, problemas, soluciones, alegría y días de buen y el mal humor vividos. Esta tesis es vuestra, porque sois el motor de mi vida, y sin vosotros nada de esto tendría sentido. Gracias Leire, por ser mi guía y compañera. Gracias Eneko, Jone y Maitane, por formar parte de este viaje.

Dedicado a Izarbe: lo hadi, maitia.

Inteligencia Computacional en Sistemas de Tele-asistencia en Domicilios

por

Aitor Moreno Fdez. De Leceta

Resumen

Este trabajo presenta un sistema inteligente de predicción del comportamiento, enfocado a prevenir situaciones de riesgo en el hogar, y orientado principalmente a personas mayores. El sistema presenta un modelo híbrido de detección de posibles alertas basado en Reglas Heurísticas introducidas por expertos en el dominio, complementadas por un Módulo de Detección Automática de Patrones, construido en base a un sistema combinado de algoritmos basados en técnicas de Inteligencia Artificial. El sistema ha sido probado en entornos reales y productivos en diversos domicilios. La detección temprana de accidentes y la prevención de comportamientos extraños en personas mayores, que viven solas en el hogar, tienen una demanda creciente. Esa demanda aún no ha sido resuelta por medio de sistemas de monitoreo manual de una manera efectiva. El sistema descrito en este trabajo solventa automáticamente este problema, evitando riesgos en el hogar mediante un método analítico avanzado. Se basa en el principio de no intrusión. Por ello, utiliza sensores plug-and-play y algoritmos de aprendizaje automático que modelizan la actividad habitual dentro del domicilio. Si el sistema detecta que algo inusual sucede (en un sentido amplio), se envía una alarma los agentes asistenciales. Una vez recibida la alarma, se comprueba su veracidad por parte del receptor de la misma, dado que la solución se configura como un sistema de apoyo a la decisión en tiempo real para los operadores de Teleasistencia. Para lograrlo, el sistema utiliza por un lado la información de sensores simples en el hogar, por otro, el conocimiento de sus actividades físicas recopiladas por aplicaciones móviles y por último, la información de salud personalizada de cada usuario basada en informes clínicos escritos en textos libres y codificados por el sistema. Actualmente, se está implantando en condiciones reales, con una precisión superior al 81%.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

Capítulo 1

Introducción

En este capítulo, primeramente repasamos el contexto socio-económico que justifica los trabajos de desarrollo tecnológico y científicos que constituyen la aportación de esta tesis doctoral. A continuación damos las guías filosóficas que hemos seguido en los desarrollos técnicos y validaciones empíricas, así como una descripción del contexto concreto en el que se han realizado los trabajos de esta tesis. Las siguientes secciones se dedican a detallar las contribuciones relevantes, los resultados conseguidos a nivel académico, en forma de publicaciones, y la estructura de la memoria.

1.1 Contexto socio-económico

El envejecimiento ha aumentado de una manera espectacular en nuestras sociedades a lo largo del siglo XX. Según la ONU[1], la esperanza de vida en España ha pasado de 34,8 años en 1900 a los 80,2 años en el año 2000. Durante el siglo XX, la población mayor se ha multiplicado por ocho en términos generales. Entre 1991 y 2001 los mayores de 80 años aumentaron en un 42%. En el año 2025, se estima que una de cada cuatro personas tendrá más de 65 años y la mitad serán mayores de 50 años. Un hecho importante en la evolución de la estructura de la población se plasma en el incremento de personas de edad avanzada: aquellas que han superado los 80 años. En el Padrón Municipal de Habitantes (2009) había contabilizadas en Euskadi 117.297 personas de más de 80 años[2] , lo que supone un 5,4% de la población total y un 21,6% de la población mayor de 65 años. Las estimaciones a nivel del Estado español avanzan que en 2060 el porcentaje de población octogenaria alcanzará el 13,1% de la población total y el 44,0% de la población mayor de 65 años. Sin duda alguna, el envejecimiento demográfico representa un éxito de las mejoras sanitarias y sociales sobre la enfermedad y la muerte. Pero también trae consigo importantes desafíos que afectan a la vida de las personas, a las familias, a la economía, a las finanzas públicas, a las prioridades de investigación y a la reorganización de los sistemas sanitario y social. El aumento de la población de edad avanzada se traduce en un incremento en las situaciones de dependencia. En el futuro más inmediato, la vivienda, la salud, y la asistencia estarán progresivamente interrelacionadas, por lo que vivienda y envejecimiento, constituirán una prioridad de modo conjunto.

[1] United Nations; Department of Economic and Social Affairs, Population Division. World Population Prospects, The 2008 Revision, Volume 1: Comprehensive Tables. New York: Author; 2009: http://www.un.org/esa/population/publications/wpp2008/wpp2008_highlights.pdf

[2] http://www.eustat.eus/productosservicios/catalogo_prod_c.html#axzz4YlAJi2NJ

En las conclusiones de su reunión de 4 de febrero de 2011[3], el Consejo Europeo respalda el lanzamiento de una Asociación Europea de Innovación para el Envejecimiento Activo y Saludable (EIP-AHA) especificando que "la innovación contribuye a afrontar los desafíos sociales más críticos a los que nos enfrentamos: la experiencia y los recursos europeos deben movilizarse de manera coherente y deben fomentarse las sinergias entre la UE y los Estados miembros para garantizar que las innovaciones con un beneficio social lleguen al Mercado más rápido".

La EIP-AHA coordina las estrategias de innovación para mejorar la calidad de vida a medida que la gente envejece. Su plan ejecutivo estratégico prevé un primer conjunto de acciones específicas:

- Cooperación para ayudar a prevenir el declive funcional y la fragilidad, con especial énfasis en la desnutrición;
- Difundir y promover modelos innovadores de cuidados integrados para las enfermedades crónicas entre los pacientes ancianos utilizando, por ejemplo, medios de monitorización remota.
- Mejorar la adopción de soluciones de vida independientes interoperables de TIC a través de estándares globales para ayudar a las personas mayores a mantenerse independientes, móviles y activos durante más tiempo.

Es importante destacar que la preferencia de los europeos a envejecer en su propio hogar se ha incrementado en los últimos años, incluso entre los casos que necesitan atención sanitaria. Según los estudios del Centro de Investigaciones Sociológica (CIS, 2009)[4], las personas mayores, en España, prefieren vivir en sus hogares antes que en residencias de ancianos o con los familiares y las personas adultas mayores de 16 años afirman que proyectan la estancia habitual en su vejez en sus domicilios. Según el Barómetro del CIS de mayo 2009, dedicado a las personas mayores y al envejecimiento, se estima que el 64,4% de la población española quiere vivir en su hogar cuando sea mayor de 65 años y, si ya lo hace, quiere seguir en la misma situación. En segundo lugar, un 15,6% afirma que preferiría vivir en casa de un hijo o hija u otros familiares. La opción de una residencia/urbanización o ciudad residencial para personas mayores se ubica en tercer lugar con un 12,5%. El porcentaje de preferencia sobre su independencia domiciliaria ha aumentado 18 puntos en ocho años, mientras que el número de los que prefieren mudarse a la casa de un hijo se ha reducido a la mitad.

Siguiendo estas preferencias, la tendencia europea respecto a la provisión de servicios dirigidos a las personas mayores se centra en el fomento de los servicios domiciliarios para facilitar que las personas se mantengan en su propio hogar. La evolución del ratio de cobertura de los Servicios de Asistencia Domiciliaria así lo demuestra. En España se ha pasado de una cobertura del 1,1% a mediados de los años 90, al 4,2% actual, siendo uno de los países que más ha avanzado en la cobertura de estos servicios en este periodo de tiempo, Sin embargo, todavía está muy lejos del 25,1% de Dinamarca o del 21,1% de los Países Bajos [Hub09]. Adicionalmente, la aplicación de los principios de la Ley de Promoción de la Autonomía Personal y Atención a las Personas Dependientes (LAAD) asegurará el ejercicio del poder de decisión y elección a las personas

[3] European Innovation Partnership on Active and Healthy Ageing (EIP AHA). http://ec.europa.eu/research/innovation-union/index_en.cfm?section=active-healthy-ageing

[4] http://www.cis.es/cis/export/sites/default/-Archivos/Marginales/2800_2819/2801/Cru2801_enlace.html

dependientes. La autonomía de la persona adquiere aquí una importancia particular, exigiendo prioritariamente el desarrollo de servicios de atención domiciliaria que maximicen dicha autonomía.

De acuerdo con estas preferencias, se está proponiendo un nuevo modelo de atención teniendo en cuenta las características y necesidades de cada persona, que se conceptualiza bajo el término anglosajón ***"Housing with Care***" o "***Extra Care Housing***" aludiendo a situaciones de dependencia en los que las personas viven en sus hogares (generalmente como propietarios, pero no únicamente) y cuentan generalmente con las siguientes facilidades [Per10]:

1. Tienen una oferta de servicios de cuidado de salud.
2. El personal profesional de cuidado está disponible las 24 horas del día.
3. Existe acceso sencillo a servicios sanitarios, tecnológicos y/o sociales, así como facilidades de interacción social comunes – por ejemplo un restaurante común o un área de reunión –.
4. Están diseñadas para un grupo específico –ya sea personas mayores sin discapacidad, personas con discapacidad pero sin deterioro cognitivo, etc. –.
5. Están diseñadas para promover la independencia y autonomía personal –es decir, no hay restricción de movimiento ni de decisión de uso de las viviendas– a lo largo de los últimos años de vida.

Las personas mayores tienen unas necesidades específicas para poder mantener su calidad de vida. La necesidad de seguridad es la principal preocupación de las personas con edad avanzada, especialmente de aquellas que viven solas. Entre las personas mayores los accidentes suponen la quinta causa de morbilidad y la séptima de mortalidad. Aproximadamente, el 80% de los accidentes se producen en la esfera privada. De éste, el mayor porcentaje se produce en el hogar y la causa es una caída: según la sociedad española de geriatría y gerontología, aproximadamente el 30% de las personas mayores de 65 años sufren una caída una vez al año. Para los mayores de 80 años, ese porcentaje se eleva hasta el 50%[5]. Este aspecto es de especial gravedad en el caso de las personas que viven solas pues puede pasar mucho tiempo hasta que se detecta la caída. Por otra parte, los sistemas de detección de caídas actuales se basan en dispositivos que la persona ha de llevar consigo permanentemente, que son usualmente invasivos o incómodos. Además suelen ser dispositivos grandes y, a veces, activan alertas cuando no existe riesgo (falsos positivos), lo que hace que en muchas ocasiones sean rechazados por los usuarios.

Las personas mayores buscan el desarrollo de su vida cotidiana bajo unas condiciones de seguridad adecuadas y demandan soluciones que sean capaces de cubrir su día a día, infundiendo confianza y tranquilidad, con el convencimiento de que ante cualquier incidencia (por ejemplo, caídas), tendrán a su disposición una asistencia adecuada y oportuna. Además, esta preocupación es compartida por sus familiares y personas más cercanas, cuidadoras de las mismas o no, quienes necesitan tener la seguridad de que no se produzca ninguna incidencia grave en los momentos en los que las personas mayores se encuentran solas en su domicilio, y que, en el caso de que está se produzca, la reacción será rápida y eficaz. Cuando dicha

[5] Documento de consenso sobre prevención de fragilidad y caídas en la persona mayor Estrategia de Promoción de la Salud y Prevención en el SNS Documento aprobado por el Consejo Interterritorial del Sistema Nacional de Salud el 11 de junio de 2014.https://www.msssi.gob.es/profesionales/saludPublica/prevPromocion/Estrategia/docs/FragilidadyCaidas_personamayor.pdf

necesidad no se cubre, entre otras consecuencias, se produce un incremento en la inversión del número de horas aplicadas a la atención por parte de los familiares (con las repercusiones negativas que tiene en la vida de estas personas, el impacto negativo en su privacidad, en su independencia y su economía) y los trastornos que pueden implicar los cambios de lugar residencia cuando se ven en la necesidad, más o menos voluntaria, de salir de su domicilio para ir a vivir con familiares y/o en instituciones sociales. Los resultados de los trabajos de esta tesis buscan posibilitar la puesta en ***producción real*** de un sistema que va a mejorar la percepción de seguridad en el hogar de las personas mayores y de sus familias, y con ello, mejorar la calidad de vida de ambos colectivos, favoreciendo un sentimiento de tranquilidad sobre dos situaciones posibles: que no se produzcan accidentes o situaciones de riesgo en el hogar, y que si éstas se producen, la respuesta sea inmediata.

Hoy en día, las personas que viven solas en su hogar, en general, son autosuficientes y no necesariamente enfermas o con necesidades de cuidados intensivos y, sin embargo, en situaciones de caída, pérdida de orientación, indisposición, malestar general o incomodidad, es difícil que puedan solicitar asistencia directamente. Un sistema de soporte automático debe tener conocimiento del contexto de los eventos que están sucediendo, asociados a un patrón de comportamiento del usuario concreto. Este objetivo es complicado de conseguir si no hay un complejo sistema de sensores a lo largo de toda la casa conectados a una red digital de datos. Por ejemplo, si una persona está mirando la televisión en la sala de estar, y esta persona está acostada en el sofá, el sistema podría detectar si hay o no un problema con la persona, comprobando si el que esté tumbado es algo habitual, o no, dado que puede estar viendo una película, pero también puede que esté indispuesta. Pero para ello se necesitan sensores de presión en el sofá o máquinas de visión artificial. Otro ejemplo: si una persona está en la cocina, y el fuego está encendido, se podría inferir que él / ella está cocinando y esto también es normal. La situación anormal se detecta cuando en la cocina se está cocinando algo, (la cocina está encendida) y la persona está en el dormitorio durante un tiempo más largo que de costumbre. Pero para ello, se necesitaría un análisis de la potencia por electrodoméstico, etc… Estos sistemas complejos que trabajan sobre diferentes tipos de sensores o entradas de dispositivos para inferir las situaciones de contexto (dormir, comer, ver la televisión, etc.), se ha demostrado que funcionan correctamente con una gran confianza [Cha14], pero es complejo desplegarlos (indicadores de potencia, cámaras, etc…) en múltiples domicilios, principalmente con los sistemas en los que hay muchos usuarios, y con características particulares y específicas diferentes en cada hogar.

Los sistemas AAL (Ambient Assisted Living) actuales prometen muchas oportunidades para la posibilidad de la vida de forma independiente de nuestros ancianos y personas mayores, así como para mejorar sus condiciones de salud. Diversas tecnologías emergentes están haciendo posible los sistemas AAL: gestión de aplicaciones móviles, sensores portátiles, robots asistenciales, casas inteligentes y tejidos inteligentes. Consecuentemente, las técnicas computacionales avanzadas están ayudando a dar valor a los datos suministrados por estas tecnologías. Pero todavía hay muchos desafíos que necesitan ser abordados por los investigadores en el futuro, como se desprende de la descripción del Estado del Arte [Ras13]. Estos retos son:

- ***Tecnología de Sensores***: La nueva generación de sensores deben ser más cómodos de usar y menos intrusivos. Para lograrlo, tales dispositivos deben incorporar las ventajas de futuras tecnologías

que mejoren la gestión de energía y la potencia en la transmisión vía inalámbrica. También, los investigadores deben abordar las preocupaciones relativas a la absorción de energía electromagnética por el cuerpo humano empleando dispositivos con baja potencia de transmisión y ciclos de trabajo bajos.

- ***Tecnología de Robótica Asistida:*** Los robots actuales de asistencia no soportan una variedad de tareas diarias, sino que cada robot es construido para prestar asistencia con un conjunto muy limitado de tareas [Sma11]. En el futuro, deberían realizarse más estudios de usuarios para la aceptación de los robots por los adultos mayores, así como para medir las expectativas de los adultos mayores ante tales robots asistenciales. Los robots no sólo deben ser capaces de ayudar a los adultos mayores en su día a día, sino también deberían ser capaces de adaptarse a su deterioro físico y cognitivo gradual, así como a sus cambios repentinos.

- ***Seguridad y Privacidad:*** La implementación de tecnologías AAL está provocando nuevas preocupaciones sobre la seguridad, debido al almacenamiento y transmisión de una gran multitud de datos personales. Los futuros sistemas AAL deberían emplear una variedad de métodos no invasivos de autenticación de usuarios basadas en características biométricas y fisiológicas para salvaguardar la privacidad del usuario. Deberían concederse distintos niveles de seguridad a diferentes usuarios en sistemas tan complejos, y la comunicación de los enlaces deben ser seguras y confiables.

- ***Factores humanos***: En general, la usabilidad y la experiencia del usuario son cuestiones de suma importancia en el diseño de sistemas AAL. Además de los ancianos, los desarrolladores de sistemas y los investigadores deben prestar atención a las otras partes interesadas, como los cuidadores, médicos y equipos hospitalarios. Además, es importante proporcionar a los usuarios la formación y la información adecuada, ya que. Sin estas acciones, muchos usuarios ancianos podrían rechazar usar tales sistemas debido a su complejidad de uso.

.

- ***Algoritmos***: La mayoría de las técnicas actuales, como el reconocimiento de actividad y la detección de la ubicación en interiores aún deben mejorarse para ser más confiables y más precisos para su uso en entornos reales. Además, algunas asunciones deben reformuladas, dado que en entornos reales no se producen, como que hay un solo residente y tenemos la actividad etiquetada en general para todos los hogares. Adicionalmente, hay una necesidad de construir bases de datos y normalizar conceptos relativos a sistemas AAL en una referencia estándar internacional.

- ***Legal y ético***: Actualmente no hay regulaciones estructuradas al respecto de los beneficios de las herramientas de AAL, o respecto malas prácticos o negligencias en sistemas complejos de teleasistencia. Además, para proteger sus derechos como consumidores, los residentes deben estar bien informados sobre las posibles consecuencias de la instalación en sus domicilios de soluciones AAL.

Los sistemas AAL se definen como el uso de la tecnología de la información y la comunicación para conformar ambientes dinámicos e inteligentes que reaccionan a las necesidades de los usuarios, brindando asistencia relevante y ayudándoles a mantener una vida totalmente independiente. Los usuarios finales son las partes interesadas en el ecosistema AAL: ciudadanos, proveedores de servicios formales e informales, proveedores de servicios, proveedores de tecnología y responsables de formular políticas. Los beneficiarios serán aquellas personas que deseen evitar la dependencia en sus hábitos diarios, es decir, ancianos que prefieren seguir viviendo independientemente en sus propios hogares. En estos casos, la asistencia puede ser necesaria en cualquier aspecto de la vida cotidiana, desde la seguridad y la salud hasta la integración social, el apoyo, ocio y la movilidad. El consejo directivo de la Asociación Europea de Innovación para el Envejecimiento Activo y Saludable (EIP-AHA) afirma[6]: "Las soluciones TIC pueden prolongar la vida independiente de las personas mayores y extender el tiempo que permanezcan activas y seguras en su entorno preferido. También tienen un enorme potencial para mejorar la inclusión social y la participación de las personas mayores, reducir las tasas de depresión, mejorar la calidad del trabajo para los cuidadores y hacer económicamente sostenible la provisión de atención (por ejemplo, evitando y reduciendo las estancias hospitalarias)".

1.2 Contexto de trabajo del doctorando

Es de extremada importancia precisar que el trabajo de la presente tesis se ha desarrollo en el contexto de una empresa dedicada a la investigación aplicada. Por tanto no puede desligarse de los trabajos realizados en proyectos financiados en los que el doctorando ha participado y cuyos resultados son convergentes con los resultados de la tesis. A continuación reseñamos los proyectos más relevantes:

- ***Chiron*** (Cyclic and person-centric Health management Integrated appRoach for hOme, mobile and clinical eNvironments)[7] es un proyecto de la convocatoria ARTEMIS liderado por Barco, que está orientado a la creación de un sistema holístico basado en el conocimiento para pacientes con cardiopatías severas. Ibermática colaboró en la creación del sistema experto para la toma de decisiones, y el doctorando participó en el mismo como investigador principal. La monitorización y predicción de anomalías en episodios sobre alteraciones cardiovasculares y posibilidad de infartos de miocardio en teleasistencia fue el trabajo principal de este proyecto, y el "germen" de esta tesis.
- ***HOI***[8] es un proyecto de la convocatoria Etorgai de 2011 liderado por Onkologikoa en el que Ibermática está creando productos para el acceso a fuentes de información muy diversas, con el objetivo de perfilar al paciente de la manera lo más precisa posible. El doctorando participó en la generación y arquitectura del sistema de estructuración de los Historiales Clínicos y su transformación en Resúmenes Médicos.

[6] https://ec.europa.eu/research/innovation-union/pdf/active-healthy-ageing/steering-group/implementation_plan.pdf
[7] https://artemis-ia.eu/project/17-chiron.html
[8] http://www.onkologikoa.org/content/participaci%C3%B3n-de-onkologikoa-en-proyectos-de-idi

- ***OSI+***[9] es un proyecto de la convocatoria Etorgai de 2009 liderado por Ibermática, en el que se desarrollaron productos relacionados con la provisión de servicios de salud para ámbitos distribuidos, obteniendo soluciones usables que proporcionan acceso remoto a la información, y cuentan con novedosas tecnologías de interacción hombre-máquina. En este proyecto se desarrollaron y los sistemas expertos de control de eventos en sistemas de teleasistencia. El doctorando construyó el sistema experto de modelado de reglas heurísticas para dicho control de eventos.
- ***HousGai***[10] es un proyecto de la convocatoria Etorgai de 2010 liderado por Matia,[11] en el que se ha colaborado en la creación de un sistema experto que modela el comportamiento de un usuario mayor en el hogar. En este proyecto se comenzaron a incorporar los primeros modelos analíticos de modelado automático de comportamiento de los ancianos en casa, en un entorno de laboratorio.
- ***Ebizi***[12] es un proyecto de la convocatoria Etorgai de 2013 liderado por Ibermática en el que se ha desarrollado una plataforma de coordinación de servicios socio-sanitarios enfocados a las personas mayores que viven solas de forma independiente en el hogar. En este proyecto, además de la continuación sobre los análisis de modelos, se trabajó en los sistemas de recomendación de pautas en la búsqueda de mejoras en la calidad de vida de cada usuario.
- **Guarantee**[13] es un proyecto de la convocatoria ITEA 2009 liderado por Philips en que se ha colaborando con otros socios para la creación de sistemas TIC para la monitorización de hogares con niños y personas mayores. Este proyecto fue la base de la primera recogida de datos masiva en domicilios reales, los cuales se analizan en esta tesis
- ***REAAL***[14] es un proyecto europeo del FP7 cuyo objetivo es investigar aplicaciones y servicios *Ambient Assisted Living* (AAL) para permitir que personas en riesgo de perder su independencia puedan mantenerla por más tiempo. Este proyecto es la segunda fuente de datos masiva en domicilios reales. Específicamente, estos datos constituyen la base de datos sobre la que se demuestran las técnicas de predicción y detección de anomalías que analiza este trabajo.

Por tanto, este trabajo de tesis se ha desarrollado dentro de un marco de colaboración entre la Universidad y la Empresa, en dónde tecnología estudiada en entornos académicos, se ha implantado en un entorno real, no controlado, a través de un centro tecnológico empresarial, y en dónde se ha podido validar su correcto funcionamiento en un condiciones operativas reales, con diversos perfiles de domicilios, personas, e infraestructuras.

---

[9] http://ibermatica.com/nueva-solucion-informatica-pionera-la-gestion-centros-asistenciales/

[10] http://ceit.es/en/industrial-sectors/health-a-food/ambient-assisted-living-and-e-health-solutions/31-electronics-a-communications/digital-signal-processing-and-digital-implementation/1055-housgai-project

[11] http://www.matiafundazioa.net/

[12] http://www.euskaditecnologia.com/atencion-a-personas-mayores-gracias-a-e-bizi-de-ibermatica/

[13] http://ibermatica.com/ibermatica-lidera-consorcio-espanol-del-proyecto-europeo-guardian-casa-aumentar-la-seguridad-hogar/

[14] http://www.cip-reaal.eu/home/

1.3 Guías filosóficas de diseño

La demanda de sistemas para la predicción de accidentes y comportamientos anómalos de personas ancianas que viven solas en el hogar está creciendo. Sin embargo, el problema hoy en día es abordado por sistemas de monitoreo manual de eficacia limitada, que no detectan adecuadamente accidentes domésticos muy frecuentes, como las caídas de personas mayores mientras están solas en su domicilio, accidentes u otros síntomas de alerta [Ogr10]. El sistema automático descrito en este trabajo evita tales riesgos mediante un método analítico avanzado apoyado por un sistema de conocimiento experto. La filosofía de diseño es la siguiente: implantar un sistema mínimamente intrusivo, utilizando sensores de instalación inmediata y algoritmos de aprendizaje automáticos que permitan modelar la actividad diaria del anciano teniendo en cuenta también la información extraída de sus registros de salud. Si el sistema detecta que sucede algo inusual (en un sentido amplio), o si algo está mal en relación con los hábitos de salud del usuario o respecto a sus recomendaciones médicas, el sistema es capaz de enviar alarmas en tiempo real a la familia, centro de atención o agentes médicos y de intervención. El sistema se alimenta de la información de los sensores desplegados en el hogar, del conocimiento del estado actividades físicas del sujeto, recolectado por aplicaciones móviles ubicadas en sus terminales móviles, además de información personalizada acerca de la salud del usuario obtenida a partir de los informes clínicos codificados en el sistema.

El desafío abordado en este trabajo es la aplicación en un entorno real de algoritmos matemáticos probados previamente en “test” de laboratorio, y aplicado posteriormente a decenas de hogares. Esta implementación requiere resolver problemas como el análisis de la calidad de los datos, el ruido, los prevención y corrección de los efectos de fallos en los sensores, la resolución de la ambigüedad en la información, y, principalmente, el aprendizaje y la previsión de acciones ante situaciones inesperadas, como la llegada de visitantes, reuniones familiares en la casa o temporadas vacacionales, situaciones no consideradas en entornos de prueba o de ensayos en laboratorio.

En este caso, hay dos condicionantes de procedimiento sobre los que se ha construido el sistema:

- la primera, sólo hay información acerca de personas que se mueven en sus hogares, sin identificación personalizada de las identidades de las personas en cada domicilio, y
- la segunda, el sistema automático no tiene un histórico al principio de la instalación: este histórico se va recopilando según transcurren los meses gracias al registro y grabación de la información recogida diariamente sobre las costumbres de cada usuario.

Estas restricciones son inevitables, ya que es requerimiento necesario construir un entorno automático de ayuda al soporte de los operadores que monitorizan los domicilios, medir su validez y estudiar la profundidad histórica mínima para determinar con una buena precisión en los modelos predictivos en un entorno real. El objetivo de este trabajo de tesis ha sido desarrollar un sistema que sirva para detectar situaciones anormales en el hogar, superando la necesidad de tener anotadas las actividades usuales de los usuarios (sería imposible anotarlas a mano en un producto comercial), como se plantea en otros trabajos [Men17], que codifican de forma manual los eventos en base a una medida combinada obtenida de los valores de los sensores en el tiempo en forma de clases semánticas “desayuno, dormido, etc.”.

El modelado de la información de base abarca el proceso de extracción, conceptualización, y validación de la información que servirá de soporte al sistema en el que se ha desarrollado en este trabajo. Las fuentes de información se convierten en datos esenciales para un apropiado razonamiento del servicio de telecuidado siendo fundamental el procedimiento a seguir en el modelado de dicha información. La decisión sobre qué metodología usar para adquirir el conocimiento relativo al telecuidado no es obvia, debido al reducido número de métodos que existen en teleasistencia. El conocimiento del estado de salud del usuario es el punto de partida para poder comenzar a identificar los patrones de comportamiento y poder evaluar su estado, puesto que este indicador puede indicar qué, cómo y cuándo debe tomar ciertos medicamentos, cuáles son las recomendaciones de actividades saludables y nutricionales, que restricciones, si las hay, pueden afectar a las interacciones sociales, si existe posibilidad de pérdida de memoria temprana, desorientación, caídas en el hogar, síntomas de debilidad, cansancio o fatiga. Se puede afirmar que estamos diseñando un sistema de detección de intenciones. Diversos experimentos sobre ambientes AAL han demostrado el potencial y las complejidades de la detección de intenciones [Gir08]. La medición de los indicadores (KPIs) sobre las acciones que realizan los usuarios y la supervisión en tiempo real de la experiencia del usuario permiten una nueva gama de sistemas innovadores y mejoras clave para los productos existentes [Med09].

En base a estos requerimientos de conocimiento, los principales objetivos de este trabajo son:
• Desarrollar un sistema de gestión del conocimiento capaz de almacenar y comprender el estado clínico del usuario, la actividad, el contexto y la situación consciente, permitiendo integrar esta información semántica en el sistema inteligente, para detectar eventos anormales de salud.
• Crear servicios de monitorización inteligente de un usuario y sus problemas médicos, para que el sistema se adapte a él, creando automáticamente reglas que determinen los valores habituales de cada individuo y evolucionen con el sujeto bajo vigilancia, para que estén siempre actualizados. Estas reglas permiten lanzar alertas completamente personalizadas sin intervención humana.
• Crear un sistema de teleasistencia de terceros basado en un sistema experto y un motor de inferencia que puede detectar automáticamente situaciones peligrosas disminuyendo falsos positivos, disparando alarmas sólo en circunstancias anormales.
Las acciones humanas están fuertemente influenciadas por el contexto, el conocimiento o la experiencia de las dependencias entre las acciones ejecutadas y las expectativas de cómo va a desarrollarse la situación en función del estado actual. El comportamiento y los hábitos humanos se caracterizan por tres atributos de las actividades diarias: el tiempo, la duración y la frecuencia. Las desviaciones en el comportamiento se pueden identificar analizando los cambios en cualquiera de estos tres atributos. Por ejemplo, si analizamos las conductas de sueño y de siesta de las personas durante un periodo determinado, comparándolo con periodos anteriores significativos, cualquier cambio sutil en la duración del sueño o de la siesta puede ser un signo de una enfermedad grave, especialmente para los ancianos, o un indicador en el progreso de una enfermedad mental, como la enfermedad de Alzheimer, a largo plazo [Sur14], mientras que en ancianos con otro tipo de patologías este patrón no es necesario tenerlo en cuenta. Mientras que la mayoría de los trabajos revisados en el Estado del Arte se han enfocado en mejorar las precisiones en la calidad en el reconocimiento de actividad física con aplicaciones de interior o al aire libre utilizando diferentes conjuntos de sensores, las tendencias de

investigación se están moviendo hacia la comprensión del comportamiento humano, de tal manera que los hábitos y rutinas diarias de las personas puedan ser descubiertas por sistemas automáticos y permitan analizar las causas de dichos patrones [Sal10].

1.4 Contribuciones

En el desarrollo de esta tesis Doctoral hemos implementado soluciones tecnológicas basadas en técnicas de Inteligencia Artificial orientadas a satisfacer las necesidades de atención en personas dependientes de edad avanzada, buscando ofrecer servicios con rapidez, eficacia y economía de recursos. El desarrollo de este trabajo contribuye a mejorar la calidad de vida de las personas mayores, favoreciendo su permanencia en el entorno domiciliario. Las contribuciones concretas que aporta este trabajo son las siguientes:

- **Mejoras en la conectividad:**

Mejoramos las capacidades de conectividad que se ofrecen los sistemas de teleasistencia, con sistemas locales y remotos sincronizados y controlados de forma automática. Estas mejoras no sólo afectan de cara a los sistemas instalados en el interior del hogar, donde se tiene que facilitar la conectividad entre los distintos dispositivos (sensores, mobiliario, electrodomésticos) que se vayan a desplegar para dar cobertura a las aplicaciones de teleasistencia, sino también de cara a la conectividad externa. Se busca la universalidad del servicio, de forma que se pueda llegar a los hogares remotos o rurales que no disponen de un acceso de banda ancha (cable/fibra) hasta el hogar.

- **Integración de tecnologías:**

El middleware utilizado está basado en OSGi, que se ha convertido en estándar de facto y muestra de ello es el acuerdo de colaboración firmado a finales de 2009 por HGI y OSGi[15]. OSGi es la plataforma idónea que permite el despliegue remoto de nuevas aplicaciones, dinámicamente y sin requerir de reinicios ni configuraciones manuales. Además, se integran otros middlewares y tecnologías con OSGi, para facilitar el desarrollo de nuevas aplicaciones que se van a desplegar el futuro. Dichas tecnologías pueden ser UPnP[16], que facilita la conectividad “Plug&Play” de dispositivos ya existentes en el mercado, o la tecnología MHP, que facilitará la interactividad con el usuario mediante la TV.

- **Mejoras en la interfaz de gestión del sistema:**

Los usuarios finales son diferentes, con distintas necesidades, y con aplicaciones que deben contemplar la personalización de cada uno de ellos. Todas las pasarelas (de distintos hogares con distintas necesidades) se gestionan desde el mismo servidor de gestión, por lo que se proveen facilidades de gestión de grupos de pasarelas, y se facilitan soluciones para los proveedores de servicio para que la gestión de dispositivos y aplicaciones heterogéneas se pueda realizar de forma sencilla, intuitiva y amigable, en una mejora importante en cuanto a la usabilidad y experiencia del usuario.

[15] OSGI Alliance Web Page: https://www.osgi.org/

[16] UPnP open development tools, Reference: Available from: http://pupnp.sourceforge.net/

- **Sistemas extracción de conocimiento y soporte a la decisión**

Se utilizan técnicas de Inteligencia Artificial (aprendizaje automático) para identificar, a partir de la información y señales recogidas (de dispositivos ambientales como de la actividad del propio usuario), los patrones de riesgo de los usuarios y gestionar las correspondientes alertas (a los propios usuarios, servicios emergencias, cuidadores, etc.) según reglas establecidas “a priori” por expertos, pero con adaptaciones automáticas a las pautas de comportamiento personales de cada usuario. Esto permite al sistema ir infiriendo y personalizando sus decisiones y sugerencias de acción en base a la personalización, contexto y situación de cada usuario en un entorno dinámico de aprendizaje automático, desasistido, transparente y continuo, así como, a su vez, permitirá inducir reglas de comportamiento generales y particulares a aplicar sobre nuevos usuarios futuros. En definitiva, permite mejorar la confiabilidad y precisión adaptándose a lo que su entorno de trabajo, un entorno real, variante y dinámico.

- **Legal y Ético**

En todas las transmisiones y comunicaciones se han establecido todos los controles, encriptados y consentimientos informados que regula la normativa actual, con lo que la plataforma es completamente segura en estos aspectos. Adicionalmente, se ha impartido formación a los residentes y a todos los actores alrededor de la plataforma propuesta.

Se puede afirmar que este trabajo es altamente innovador y tiene una clara ventaja competitiva sobre la actual estrategia de provisión de este tipo de servicios ya que:

- Puede dar cobertura a todo el colectivo de personas mayores.
- Ofrece un servicio integral ante las situaciones de riesgo actuando tanto en la prevención, como en la detección y asistencia de las mismas.
- Ofrece un servicio personalizado y personalizable, que cuenta con la actuación de diversos profesionales y que se apoya en la integración de distintas soluciones tecnológicas.
- Las soluciones tecnológicas proporcionan seguridad pasiva sin que las personas mayores tengan que intervenir en la activación de la alarma, ni portar las 24 horas ningún dispositivo.
- Está siendo probado en entornos reales, no de laboratorio, obteniéndose datos reales y actualmente sigue en proceso de expansión, incrementándose los domicilios de acción y recogida de información.
- No existe actualmente en el mercado, ningún producto ni servicio global de prevención automática de riesgos para personas mayores similar funcionando en entornos reales.

1.4.1 Diferencias computacionales con otros sistemas

La mayoría de sistemas propuestos en la literatura y en la industria se basan en el modelado del contexto más que en el modelado del usuario. En este trabajo hemos desarrollado técnicas para poder inferir comportamientos particulares en cada domicilio para cada usuario, analizando los movimientos a lo largo del espacio y del tiempo, utilizando sensores no invasivos y muy baratos. La principal hipótesis que este trabajo ha demostrado, es que, con el análisis de la secuencia de las posiciones de los habitantes en su deambular cotidiano por la casa, es posible entrenar a un sistema automático de los patrones particulares de esos

usuarios y, por tanto, determinar si hay algo extraño en sus propios comportamientos. Hemos sido capaces de extraer esta secuencia de posiciones a partir de información “bruta” recogida con sensores baratos y fáciles de instalar, y que, además, son no invasivos. Adicionalmente, somos capaces de agregar a la información de los sensores una interpretación semántica de los eventos, usando Reglas de Proceso predefinidas, (simplemente con la información de posición y tiempo, podemos generar inferencias sobre en qué estados está el usuario) y podemos complementarlos con otros datos externos al domicilio (meteorológicos, demográficos, etc.). Con estos dos conjuntos de datos (básicos y codificados semánticamente), hemos realizado una comparativa estadística para determinar cuál de los dos sistemas se ajusta mejor a los patrones de los usuarios, y por lo tanto, cuál detecta de manera más efectiva situaciones comprometidas. Con estas alertas, finalmente, generamos una serie de notificaciones que enviamos a un operador especializado para que las valide manualmente, y decida si activa los protocolos de acción asistencial. La diferencia en nuestro trabajo se basa en dos aspectos: por un lado, la profundidad y variabilidad de la información, y por otro, el entorno de captación de la misma. Usualmente, los trabajos anteriores están analizando conjuntos de datos bien conocidos, con bases de datos de muestra, por ejemplo, los obtenidos del proyecto CASAS de WSU (Aruba CASAS dataset), con sólo un usuario anotado. Nosotros estamos trabajando sobre datos reales obtenidos en casas reales con usuarios reales, viviendo de forma totalmente desatendida y obteniendo datos de sensores de forma constante y diaria. En este trabajo, en un primer estadio, se intenta determinar los patrones o eventos de los usuarios en función en una aproximación alineada con el Estado del Arte, en la codificación de los eventos recogidos por los sensores en “n” actividades generales, sin perder el objetivo final de predecir dichas actividades, y en un segundo estadio, se intenta determinar los patrones o eventos de los usuarios en función del análisis de la ubicación de los mismos en base a un análisis de las frecuencias de los usuarios por estancia más habitual en distintas secuencias temporales, en un enfoque más innovador.

Como se verá más adelante, se demuestra que el horizonte mínimo requerido para que los modelos que se han desarrollado en este trabajo sean mínimamente efectivos en un entorno real, es de un mes de profundidad mínimo, ante otros trabajos que estiman 8-9 semanas [Sur13]. Así, en un entorno productivo, un mes es el tiempo mínimo que necesitaría un producto en el mercado para poder funcionar de forma fiable, con modelos aprendidos. Para paliar este “desfase” entre la instalación y la producción de alertas automáticas personalizadas, el sistema se complementa con un conjunto de Reglas Heurísticas predefinidas por expertos en Teleasistencia, introducidas a mano en el sistema, que permite suplir este intervalo de aprendizaje con un sistema de control heurístico previo.

1.4.2 Publicaciones conseguidas en el desarrollo de la tesis

- ***Título: Lynx: Automatic Elderly Behavior Prediction in Home Telecare***
  Autores: Jose Manuel Lopez-Guede, Aitor Moreno-Fernandez-de-Leceta, Alexeiw Martinez-Garcia, and Manuel Graña, “Lynx: Automatic Elderly Behavior Prediction in Home Telecare,” BioMed Research International, vol. 2015, Article ID 201939, 18 pages, 2015. doi:10.1155/2015/201939

- ***Título: An Automatic Telemonitoring System for Elderly People at Home***
  Autores: Aitor Moreno-Fernandez- de-Leceta, Pedro de la Peña, David M. Barrios, Beñat G. Granciaenteparaluceta, Jose M. Lopez-Guede and Manuel Graña
  Revista: International Journal of Sensors Wireless Communications and Control (SWCC) ISSN: 2210-3279 (Print) / 2210-3287 (Online) Volumen: 4 Número: 2 Páginas: 57-66

- ***Título: Behavior prediction in home telecare systems.***
  Tipo participación: Ponencia invitada
  Autores: Jose Manuel Lopez-Guede, Aitor Moreno-Fernandez- de-Leceta, Manuel Graña. Congreso: 6th International Conference on Applied Informatics and Computing Theory (AICT 2015)
  Lugar celebración: Salerno, Italy

- ***Título: Real implantation of an expert system for elderly home care***
  Tipo participación: Ponencia
  Autores: Aitor Moreno-Fernandez- de-Leceta, Unai Arenal Gómez, Jose Manuel Lopez-Guede, Manuel Graña
  Lugar celebración: Bilbao (Spain)
  Congreso: 10th International Conference on Hybrid Artificial Intelligence Systems (HAIS 2015)
  Fecha: 22th-24th June, 2015

- ***Título: Real prediction of elder people abnormal situations at home***
  Tipo participación: Ponencia
  Autores: Aitor Moreno-Fernandez-de-Leceta , Jose Manuel Lopez-Guede , Manuel Graña , Juan Carlos Cantera
  Lugar celebración: San Sebastián (Spain)
  Congreso: 11th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2016)
  Fecha: 19th – 21st October, 2016

- ***Título: A novel methodology for clinical semantic annotations assessment.***
  Autores: Aitor Moreno-Fernandez-De-Leceta, Jose Manuel Lopez-Guede, Leire Ezquerro Insagurbe, Nora Ruiz de Arbulo, Manuel Graña (2018).
  Journal of Applied Logic. Elsevier. "In Press"

1.5 Estructura de la tesis

La estructura de la tesis doctoral es la siguiente:

- En el capítulo 2 se presenta el estado del arte, incluyendo una revisión bibliográfica al respecto de los sistemas de teleasistencia y gestión automática de alertas en domicilio.
- En el capítulo 3 se detallan los fundamentos de los algoritmos utilizados en la caracterización de las actividades y patrones de comportamiento modelados a lo largo del desarrollo del trabajo, presentando especial énfasis en aquellos conceptos estadísticos que engloban la validación de los resultados presentados.
- En el capítulo 4 se describe la arquitectura del sistema, explicando en detalle cómo se integran los distintos módulos que la componen y el software utilizado.
- En el capítulo 5 se detallan los componentes más relacionados con el "hardware" y la sensórica empleada.
- En el capítulo 6 explica más en detalle el "Módulo de Detección Automática de Patrones", que compone el sistema autónomo de modelado y predicción de patrones.
- En el capítulo 7 se explica el proceso implementación del Sistema Experto, a nivel de procedimiento, consideraciones éticas y resolución de problemas.
- En el capítulo 8 de desgranan los resultados obtenidos en los diseños experimentales, el proceso de tratamiento de los datos, la metodología de validación y los resultados finales de los modelos.
- Finalmente, en el capítulo 9 se exponen las conclusiones del estudio realizado y se discuten las posibles líneas futuras de trabajo.

1.6 Esquema Funcional del Sistema Presentado

A continuación se detalla el esquema funcional del Sistema desarrollado en este trabajo. El sistema se constituye en tres funcionalidades:

1. ***Sistema de Agregación de Datos***

   Está funcionalidad aúna datos de tres contextos (figura 1.1):

   - Datos provenientes de los sensores referentes a los cambios de estado de los usuarios en un domicilio.
   - Datos obtenidos a partir de los datos clínicos de los usuarios.
   - Datos meteorológicos externos al domicilio.

   En todos los casos, existe una transformación de los datos bruto, en datos codificados y normalizados, de la siguiente forma:

   - Los datos referentes a los sensores se codifican en función del contexto, por ejemplo, estar en el dormitorio a la noche se codifica como un estado de "durmiendo".
   - Los datos referentes a los historiales clínicos, originariamente en lenguaje natural, pasan por un proceso de transformación, codificación y filtrado, obteniéndose al final de este proceso un

resumen clínico que ayuda a los expertos clínicos a poder marcar pautas de alertas o de seguimiento a los usuarios.

- Los datos referentes a datos meteorológicos, se codifican en base a una nomenclatura interna definida en este trabajo.

Para dichas codificaciones y filtrados, se utilizan una serie de reglas incorporadas al sistema que denominamos “Reglas de Proceso”.

2. ***Sistema Experto: Generación de Reglas.***

   Esta funcionalidad, ver figura 1.2, es la que construye los sistemas de control y detección de alertas ante situaciones de riesgo en el domicilio. En este trabajo se han implantado dos aproximaciones de abordar este problema:

   1. Una aproximación determinista, que denominamos “Módulo de Reglas Heurísticas”, que está compuesto por aquellas reglas introducidas manualmente por dos grupos de expertos diferentes:
      a. Los Expertos en Teleasistencia: Personas cuyo trabajo habitual es el de monitorizar a personas mayores que viven solas, y conocen bien los patrones generales en los que hay que alertar a los familiares o los agentes asistenciales.
      b. Los Expertos Clínicos: que, en base a los datos clínicos resumidos que suministra el sistema, también son capaces, de forma manual, de crear reglas de propósito general sobre posibles alertas relacionadas con ciertas patologías concretas, sobre la toma de ciertos fármacos específicos, o sobre la detección del no seguimiento de recomendaciones clínicas o terapéuticas.
   2. Una aproximación algorítmica, en base al estudio de los históricos de los datos que se van almacenando con la evolución de lo que ocurre en cada domicilio. Esta aproximación también se basa en dos aproximaciones diferentes a la hora de detectar alertas o situaciones de riesgo:
      a. La primera se basa en automatizar el descubrimiento de patrones por medio de técnicas aprendizaje automático, para después, poder cotejar lo que realmente está ocurriendo con lo que el modelo predice que debiera estar sucediendo. Si el estado real y teórico no cuadran con un determinado umbral de confianza y con una cierta frecuencia, se puede pensar que existe una posible alerta. En este trabajo se ha trabajado con diversas aproximaciones en la extracción de dichos modelos, tanto algorítmicas como en la estrategia de la normalización previa de los datos de entrada. En el mismo proceso de generación de los modelos, también se ha realizado el análisis de la validez de los mismos, extrayendo informes sobre la confianza de la exactitud de los modelos.
      b. La segunda aproximación se basa en la detección automática de situaciones anormales, en base a análisis matemáticos que comparan la similitud o diferencia de los distintos eventos con los de sus grupos cercanos.

3. ***Sistema Experto: Aplicación y sistema de Notificaciones.***

Esta tercera funcionalidad es la responsable de aplicar los dos tipos de modelos del punto anterior (heurísticos y automáticos), obtenidos a partir de reglas manuales o del análisis de históricos, en función de cada caso, a los eventos en tiempo real que están ocurriendo en cada uno de los domicilios. De esta forma, esta funcionalidad aplica la lógica que determina si existe una probabilidad de que el usuario entre en un estado de riesgo o no. En caso de que esta posibilidad existiera, se notifica de forma automática a un operador, que, a su vez, manualmente, revisa los datos de origen que han generado la posible alerta, los verifica y comprueba, registra las notificaciones como correctas o incorrectas según el caso, y si es pertinente, activa los servicios asistenciales que corresponda. Este control manual, permite ir ajustando el sistema automático de forma más precisa, dado que la fase de generación de modelos automáticos se procesa diariamente, incluyendo los nuevos eventos que se van generando junto con el "veredicto" dado por el operador, en caso de que haya existido una probabilidad de alerta, (marcando los falsos positivos y los verdaderos positivos como tales). Finalmente, este control permite medir la validez global del sistema. Toda esta gestión de notificación se realiza a través de un sistema de notificación digital.

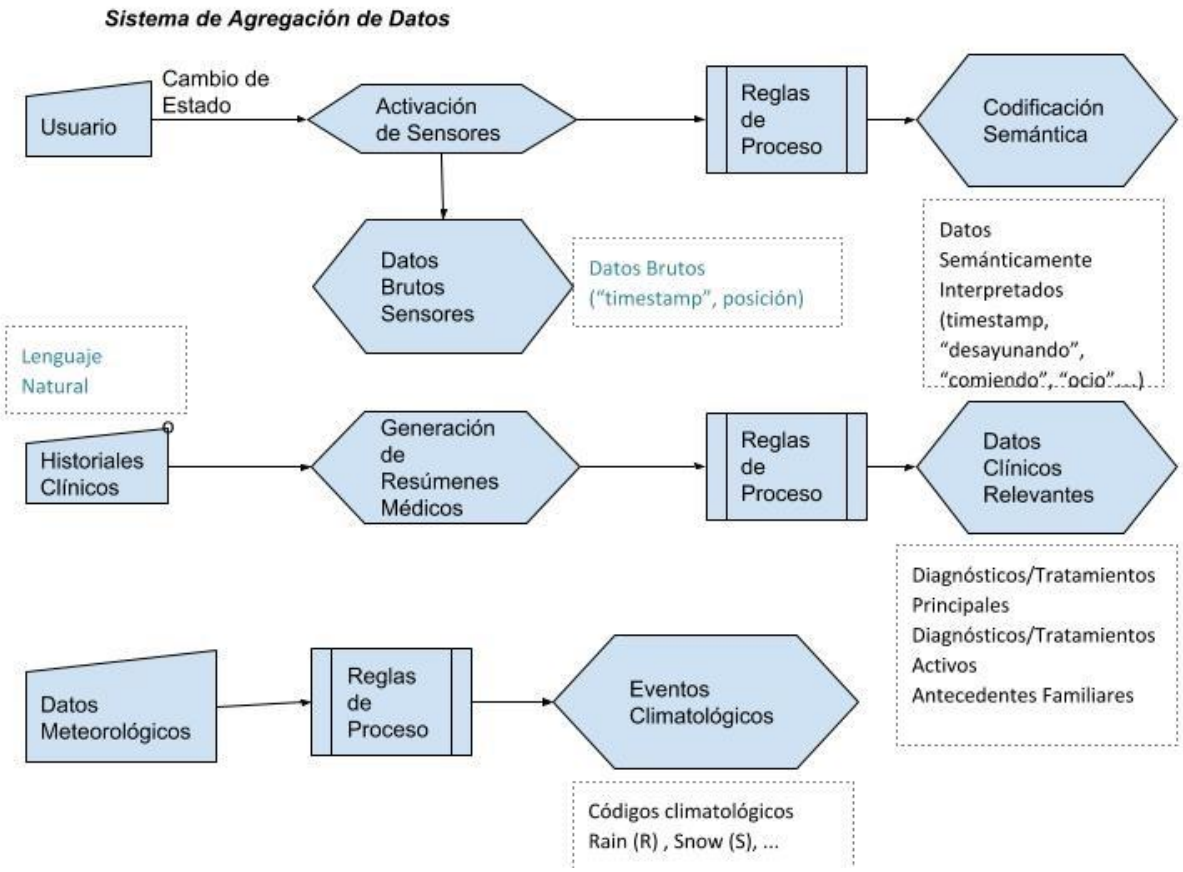


Figura 1.1 Esquema Funcional del Sistema de Agregación de Datos

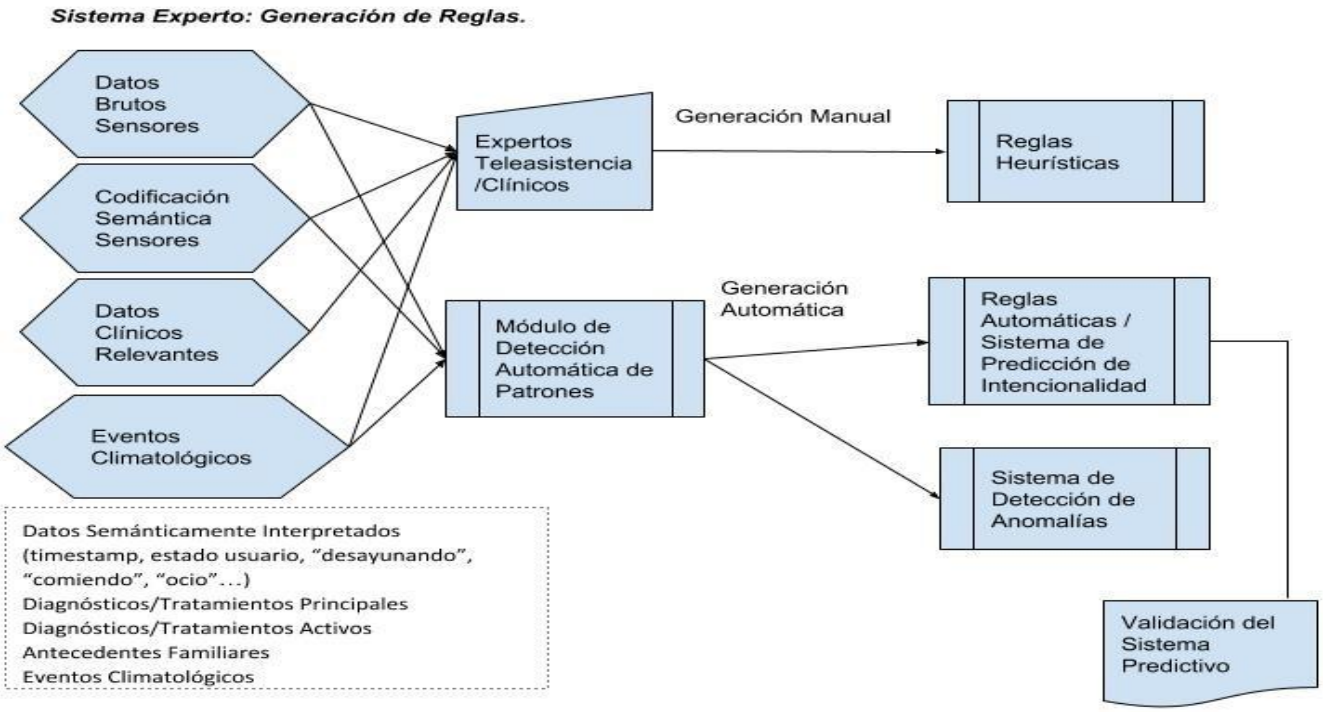


Figura 1.2 Esquema Funcional del Sistema Experto

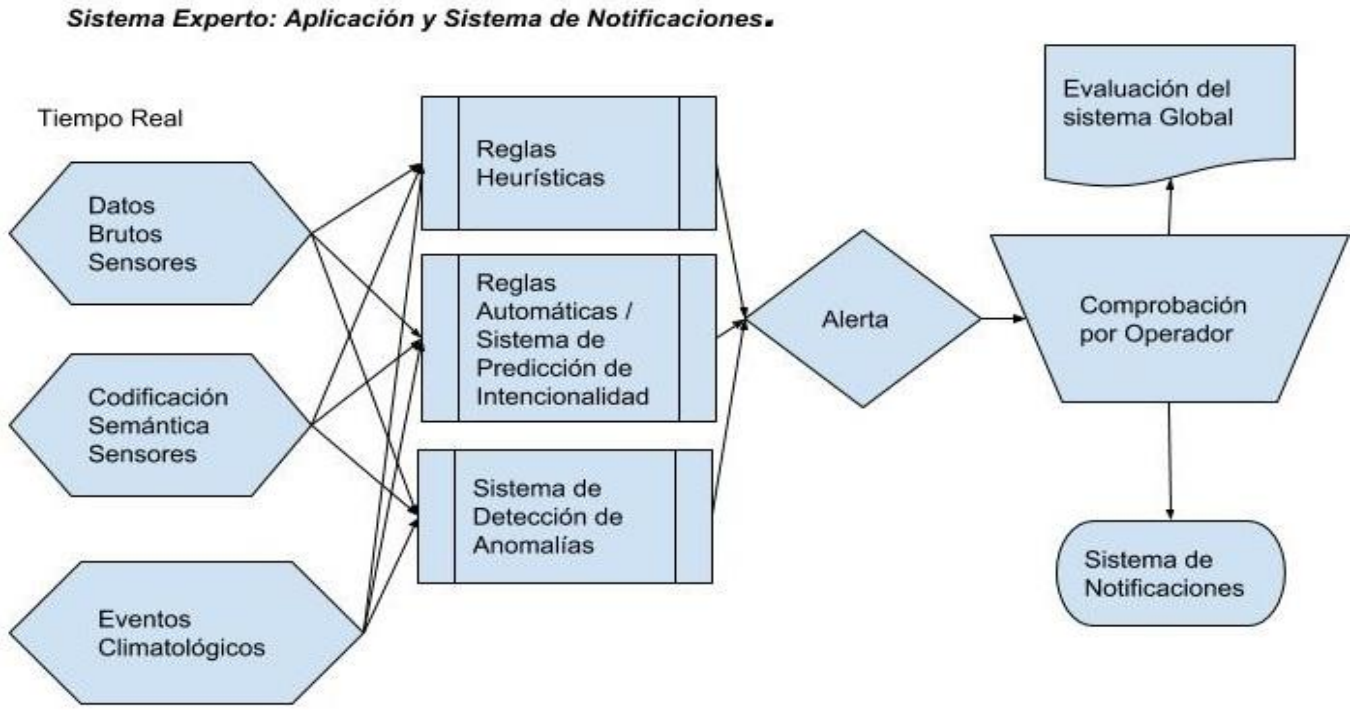


Figura 1.3 Esquema Funcional del Sistema de Notificaciones

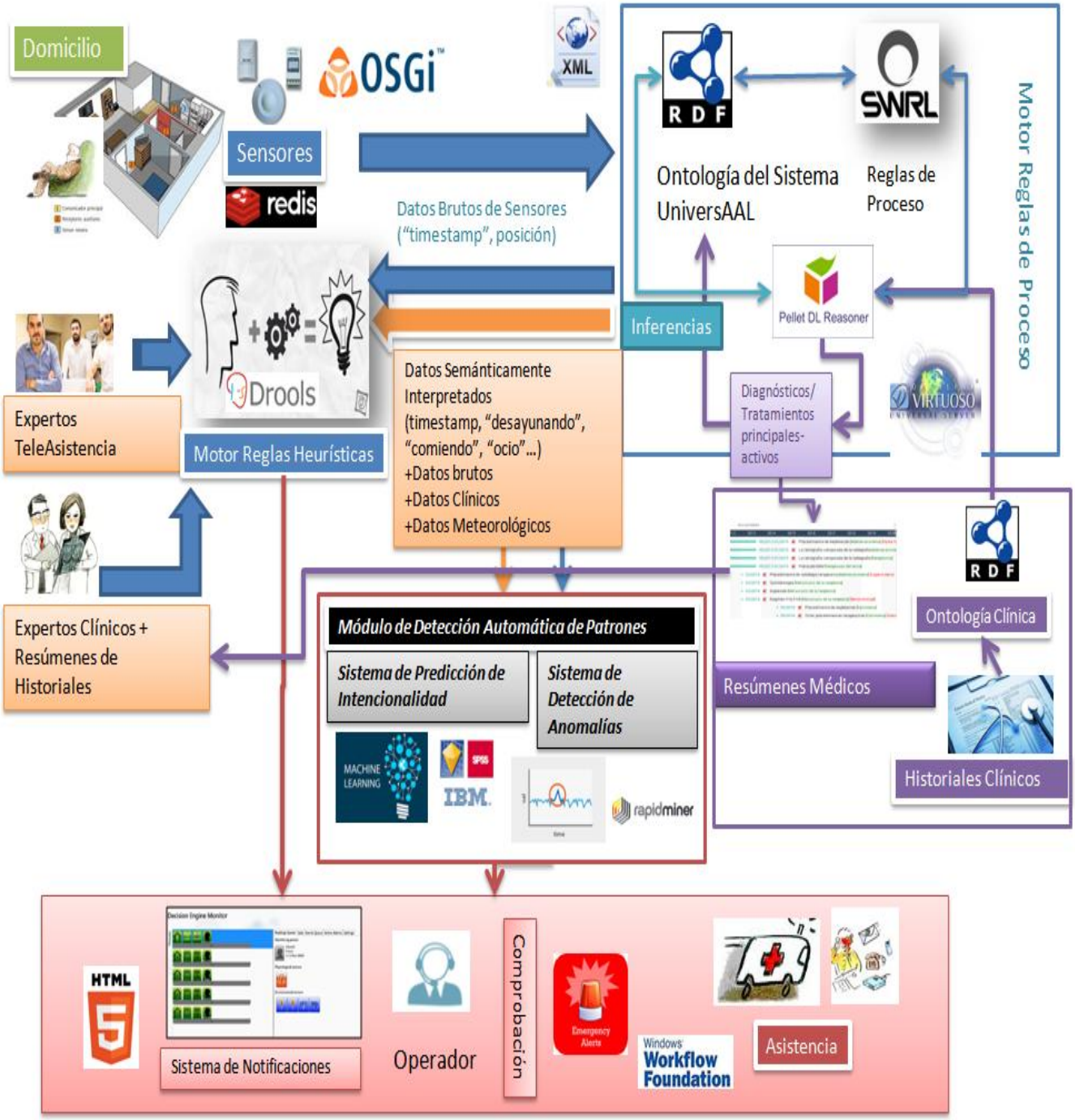


Figura 1.4 Esquema Funcional del Sistema y tecnologías aplicadas.

Capítulo 2

Antecedentes y Estado del Arte

En este capítulo recogemos el estado del arte relativo a los trabajos de esta tesis doctoral y sus contribuciones. Primeramente, relatamos el paisaje de proyectos de investigación financiados por instituciones públicas que son más relevantes para describir los esfuerzos de desarrollo e innovación realizados en los últimos años. A continuación, desgranamos los servicios, tanto públicos como privados que se dan alrededor del contexto de esta tesis, principalmente en el marco del País Vasco, que es en dónde se ha realizado. A continuación se muestra un resumen de los proyectos empresariales que están "naciendo" con iniciativas similares, para finalmente, realizar un repaso el estado del arte actual en cuanto a plataformas tecnológicas actuales relacionadas.

2.1 Proyectos de investigación relevantes

A nivel Internacional [Ras13] la tabla 2.1 resume varios proyectos de viviendas inteligentes:

- El proyecto CASAS [Ras09]: en este proyecto, realizado por la Universidad Estatal de Washington, se detalla un sistema que proporciona una ayuda no invasiva en un entorno para pacientes con demencia en el hogar.
- El "Aging in Place " de la Universidad de Missouri tiene como objetivo un modelo de atención a largo plazo para personas de la tercera edad en términos de salud de apoyo [Ran11]. "Elite Care" es un proyecto en un centro de vida asistida con sensores que permiten monitorear indicadores tales como tiempo en la cama, peso corporal, y la inquietud del sueño utilizando sensores heterogéneos [Ada10].
- El "Aware Home project" de Georgia Tech [Che11] emplea una variedad de sensores, tales como sensores de suelo inteligentes, así como robots de apoyo para monitorear y ayudar a los ancianos.
- Otro notable proyecto de casa inteligente es el proyecto DOMUS [Bou07] de la Universidad de Sherbrooke, y el proyecto "House_n" del MIT [Tap04].

En el estado español, existe mucha actividad en la investigación de la mejora en las plataformas de Teleasistencia. En primer lugar, avanzando en la investigación de plataformas tecnológicas, como es el caso de eMobility, Networked and Electronic Media, Artemis y ENIAC. También es de interés mencionar las plataformas tecnológicas nacionales como eMOV, INES, Prometeo y la Plataforma Tecnológica del Hogar Digital. A nivel Europeo, el desarrollo de proyectos de Teleasistencia en el FP7 ha sido claro durante los últimos años, muy relacionado con los planes de trabajo de ICT (Sociedad de la Información y las Comunicaciones) y Health (Salud).

Tabla 2.1 Relación de proyectos recientes relacionados con la tesis.

Project	Institution
Aging In Place	U. Of Missouri
Aware Home	Georgia Tech
CareLab	Germany
CareNet(MIDAS)	U. of Wales, UK
CASAS	Washington State U.
DOMUS	U. de Sherbrooke
Elite Care	OHSU
ENABLE	Netherlands
Gator Tech	UF
HIS	Grenoble U., France
MavHome	U. of Texas at Arlington
Millennium Home	Brunel U.
ProSAFE	LAAS, France
SELF	ETL, Japan
Smart Medical Home	Rochester U.
Ubiquitous Home	UCG, Japan
WTH	JMITI, Japan
-	UNSW, Australia

Más en concreto hay que resaltar los “Challenges” 5 y 7 que también están estrechamente relacionadas con este proyecto:

- Challenge 5: Towards sustainable and personalized healthcare
- Challenge 7: ICT for Independent Living, Inclusion and Governance. Es de resaltar en concreto el punto 7.1, donde se hace mención a:
    - Incremento en la eficiencia del cuidado y prolongación de la independencia y calidad de vida de las personas mayores y sus cuidadores
    - Uso de plataformas y herramientas abiertas para la creación y gestión de productos y servicios ICT integrados para el envejecer mejor, cuidado personalizado y gestión de la energía en el hogar.

Dentro de estos proyectos europeos, se pueden mencionar los siguientes proyectos relacionados con la temática de Teleasistencia:

- CONFIDENCE (FP7 ICT 7.1): (http://www.confidence-eu.org/: este proyecto tiene por objetivo el desarrollo de un sistema de cuidados y atención para las personas mayores capaz de detectar situaciones anormales.
- PERFORM (FP7 ICT 5.1): http://www.perform-project.com/ este proyecto aspira a la investigación de un sistema novedoso e inteligente de monitorización de la evolución de enfermedades neurodegenerativas mediante el empleo de una redes de sensores en una red de comunicación personal (Body Area Network) y la utilización de algoritmos avanzados de fusión y procesamiento de información.
- OASIS: Open architecture for accessible services integration and standardisation (http://www.oasis-project.eu) plataforma abierta que garantice la interoperabilidad y la integración de los servicios y la información enfocada a mejorar las actividades de la vida diaria de las personas mayores.

- HERMES: Cognitive care and guidance for active aging (http://www.fp7-hermes.eu) se orienta hacia la llamada Asistencia Cognitiva (Cognitive Care). Esta asistencia cognitiva se materializa mediante el uso de tecnología que combina las habilidades funcionales de la persona mayor para reducir el impacto de su declive y asistirle cuando sea necesario. Para ello, el proyecto emplea procesamiento y razonamiento inteligente de señales de audio y video para contribuir a la mejora de las capacidades cognitivas del usuario, y en última instancia, para mejorar su calidad de vida.
- AALIANCE: European ambient assisted living innovation alliance (http://www.aaliance.eu) Entre otros objetivos este proyecto pretende establecer una red formada por todos agentes implicados (empresas, agentes tecnológicos, centros de investigación, proveedores de servicios, etc.) y además coordinar las actividades europeas industriales y de investigación en el área de Ambient Assisted Living
- MonAMI. Mainstreaming on Ambient Intelligence: El objetivo general del proyecto MonAMI es la universalización de la accesibilidad a los servicios y bienes de consumo, incluyendo servicios públicos, a través de desarrollos e investigaciones aplicadas y usando tecnologías avanzadas que ayudan a asegurar una igualdad en el acceso, una vida independiente y la participación de todos en la Sociedad de la Información. Servicios y aplicaciones serán seleccionados a través de estándares tales como aplicaciones de confort (control del hogar, interfaz de comunicaciones personalizado, planeamiento de actividades), salud (monitorización, medicación), seguridad (seguridad en el hogar, validación de visitas, detección de actividades). La plataforma tecnológica será derivada de tecnología estándar. Integrarán elementos emergentes tales como redes fiables auto-organizadas, dispositivos portátiles, tecnología “e-inclusión” de interacción con el usuario, capacidad de monitorización, infraestructura de servicios que aseguran la calidad de los propios servicios, confianza y privacidad.
- SOPRANO. Desarrollo de entornos inteligentes dirigidos a las personas mayores: El proyecto SOPRANO propone desarrollar servicios de asistencia basados en Tecnología de la Información. El proyecto persigue dos objetivos: (1) desarrollar nuevos procedimientos que permitan integrar en los hogares de los usuarios sistemas basados en la Tecnología de la Información (tecnología de asistencia, tele-atención y tele-salud) que proporcionen asistencia, y poder investigar las dificultades motrices, sensoriales y cognitivas que padecen las personas mayores; (2) mejorar los canales de comunicación con los usuarios basados en la visión, la voz o los demás sentidos. El objetivo es desarrollar un entorno doméstico completamente comunicado. Los dispositivos e instrumentos integrados para ayudar a los usuarios a realizar sus actividades diarias y los sistemas avanzados de tele-asistencia y tele- salud realizarán el seguimiento de la salud y el bienestar, de forma que harán posible el envío de ayuda cuando sea necesaria.
- PERSONA. Espacios perceptivos que promuevan el envejecimiento independiente. El objetivo es desarrollar una plataforma para el acceso instantáneo y transparente, desde cualquier lugar, para dar soporte a servicios dirigidos a personas mayores, con el fin de que pudieran estar el mayor tiempo posible en su entorno habitual. El objetivo del proyecto Persona es avanzar en el paradigma de Inteligencia Ambiental, a través de la armonización de tecnologías y conceptos AAL. El sistema puede detectar situaciones concretas de emergencia, avisar a los médicos, bomberos y personal de seguridad, así como encender o apagar de manera automática los electrodomésticos, las luces, la

cocina o el televisor. En el ámbito del proyecto se desarrolla una plataforma tecnológica abierta y escalable, para construir y analizar un amplio catálogo de Servicios AAL.

- NETCARITY. Sistema multi-sensor en red para personas mayores: atención sanitaria y seguridad en el domicilio (INT): NETCARITY propone un nuevo paradigma para apoyar la independencia de las personas mayores que viven solas en su domicilio. El proyecto promueve el desarrollo de una infraestructura tecnológica "light" que se instalará en los domicilios de las personas mayores a un bajo coste, de manera que permita asegurar tanto el apoyo en las actividades básicas de la vida diaria y la detección de situaciones críticas de salud, como el apoyo social y psicológico necesario para que las personas mayores mantengan el bienestar emocional, su dignidad y calidad de vida. En NETCARITY, se abordan los factores sociales y psicológicos que explican los problemas que supone "envejecer en casa". Se reforzará la comunicación de las personas mayores con sus amigos y cuidadores, reduciendo el aislamiento y la sensación de soledad. Se estimulará la realización de las actividades diarias para mantener niveles altos de motivación y una correcta percepción de sus propias habilidades.
- CONFIDENCE. Sistema ubicuo de atención para apoyar la vida independiente: El objetivo que persigue el proyecto CONFIDENCE es crear un sistema que facilite a la persona mayor un control de sus conductas cotidianas así como de su estado de salud, de manera que se prolonguen los años de autonomía en el hogar. El proyecto propone nuevas soluciones para facilitar la vida independiente de las personas mayores, permitiendo monitorizar su situación y estado actual de salud. Los datos se obtendrían tanto en interiores como en el exterior con un solo sistema, de manera permanente, permitiendo a las personas realizar una vida normal sin miedo a los sobresaltos, permitiendo una atención rápida y bien informada.

2.2 Servicios Públicos y Privados de Teleasistencia

2.2.1 El Servicio Público de Teleasistencia de Euskadi

Actualmente, en la CAPV, a fin de prevenir caídas en las personas mayores, éstas y/o sus familiares pueden beneficiarse de este servicio. Se trata de un servicio público, en formato de copago, que da respuesta a situaciones de emergencia, servicios de urgencias, atención psicosocial, compañía, seguimiento e información y orientación sobre recursos de atención social. Pretende fomentar la autonomía personal de las personas mayores, ya que les permite permanecer en el entorno habitual con una mayor sensación de seguridad, sustentada en una respuesta rápida y profesional ante cualquier incidencia. Requiere que la persona lleve las 24 horas consigo un pulsador, a modo de medallón, para alertar de forma inmediata de una posible indisposición, mediante su pulsado, y acceder a un contacto especializado con el que poder hablar en modo "manos libres" para explicar la situación, dando a su vez conocimiento de la misma a los servicios de emergencia. Señalar que este tipo de Servicio se ofrece también de forma privada (véase como ejemplo http://www.tunstall.es/es/pdfs/detector_de_caidas.pdf.) El hecho de que la persona sea la que tenga que

activar el pulsador, hace que quedan excluidas de este servicio aquellas personas mayores con deterioro cognitivo que no son capaces de activar el pulsador. Además, se sabe que muchas personas mayores, usuarias del Servicio de Teleasistencia, mantienen el pulsador guardado en todo momento y/o se lo quitan por la noche, aun cuando suelen levantarse varias veces a lo largo de la misma para ir al baño.

2.2.2 Otros Servicios Privados de Teleasistencia y alarmas personales de detección de caídas

Existen también en el mercado, por parte de ciertos distribuidores como Attendo Systems S.L.U. Neat Group, Tunstall Televida, servicios de Teleasistencia, Asispa, Asist, servicios de teleasistencia, como el descrito anteriormente, pero también otros basados en alarmas personales de detección de caídas.
Estos últimos, son dispositivos que detectan si una persona se ha caído y puede necesitar ayuda. El detector puede implementarse de diversas formas:

- Algunos dispositivos detectan la pérdida de verticalidad de la persona, es decir, cuando la persona sufre una inclinación superior a 60º en cualquier dirección (por ejemplo, el detector de caídas de PC Compatible).
- En cambio, otros detectores miden si ha habido un cambio de posición brusco (oscilación) con impacto, tanto si la posición inicial es vertical (de pie) u horizontal (tumbado) (por ejemplo, el detector de caídas de Attendo).

En estos casos, los dispositivos debe portarlos la persona en la cintura, a modo de cinturón. En caso de detectar una caída, generan una señal de aviso (una señal acústica o una vibración) que la persona puede anular en un tiempo determinado (en general inferior al minuto). Si no se anula, se dispara una alarma alertando de que la persona se ha caído.

2.4 Plataformas tecnológicas

Existen una serie de plataformas y redes tanto a nivel internacional como estatal especialmente dedicadas, o que tienen grupos de trabajo específicos, relacionados con la mejora de la calidad de vida y el fomento de la vida independiente e incluyen algunas de las líneas de investigación relacionadas con este trabajo.
A nivel internacional cabe destacar, además de ARTEMIS o ITEA anteriormente mencionadas, a NESSI, que es la Plataforma Europea de Software y Servicios. Dentro de la plataforma existe un grupo de trabajo vertical relacionado con la eSalud (eHealth WG). Dicho grupo de trabajo ha identificado una serie de áreas estratégicas de trabajo que se incluyen en el “Nessi eHealth Group Manifesto” y el correspondiente “Position Paper”. De las áreas identificadas en el “Manifesto”, las directamente relacionadas con esta tesis son:

1. Proporcionar información más efectiva y personalizada relacionada con el paciente.
2. Proporcional a los médicos sistemas de soporte a las decisiones clínicas basados en la fusión de diferentes fuentes de datos heterogéneas.
3. Integración de información y extracción de conocimiento.

A nivel estatal cabe destacar la plataforma “eVia10 - Plataforma Tecnológica Española de tecnologías para la Salud, El Bienestar y la Cohesión Social”. La plataforma cuenta con la participación de un gran número de empresas (43%), centros tecnológicos y universidades, entre los miembros de dicha plataforma se encuentra una importante representación de empresas, centros tecnológicos y universidades de la CAPV.

2.5 Estado del Arte

La Inteligencia Computacional (I.C.) supone una herramienta de gran valor para mejorar la eficiencia de los servicios asistenciales que se prestan a las personas mayores, además de ser un apoyo para las y los profesionales, posibilitar una mayor cobertura de los servicios, y ahorrar costes en la prestación de los mismos. Además, también suponen un apoyo en el cuidado para las familias. La tarea crítica del sistema I.C., es el perfilado automático de los comportamientos de los usuarios.

En los últimos años se han introducido distintas tecnologías para abordar soluciones referentes a las cuestiones de cuidado de personas de la tercera edad, principalmente basadas en sistemas de detección de caídas [Hua14, Igu13, Pan14]. Sin embargo, la mayoría de estas plataformas son demasiado caras para el uso en masa o son de baja calidad: la mayoría de las soluciones comerciales se basan solamente en la detección de caídas, y en el soporte “a posteriori” [Kal10]. Otras soluciones están enfocadas a botones de pánico, pero siempre desde un punto de vista “reactivo”, no “predictivo”. El hecho de que la persona sea la que tenga que activar el pulsador, hace que quedan excluidas de este servicio aquellas personas mayores con deterioro cognitivo que no son capaces de activar el pulsador. Además, se sabe que muchas personas mayores, usuarias del Servicio de Teleasistencia, mantienen el pulsador guardado en todo momento y/o se lo quitan por la noche, aun cuando suelen levantarse varias veces a lo largo de la misma para ir al baño, lo cual, resta efectividad a este tipo de soluciones. Los componentes más estudiados en los sistemas AAL en el actual estado del arte son los siguientes:

- ***Reconocimiento de Actividades Humanas basadas en sensores de bajo nivel.***

Uno de los componentes más importantes de los sistemas AAL es el componente de "reconocimiento de actividad humana" o HAR en sus siglas en inglés (Human Activity Recognition). “HAR” es el responsable de reconocer los patrones de actividad humana con información obtenida de diferentes sensores de bajo nivel. En la literatura se pueden encontrar diferentes enfoques referentes en cuanto a la detección de actividades: la detección de actividades como objetivo principal, como complemento a la modelización de patrones o como base en la detección de anomalías en las rutinas de comportamiento y signos de indicios de posibles enfermedades. Algunos ejemplos de cada uno de estos enfoques son los siguientes: [Han12] propuso un marco de salud de cuatro capas para predecir el riesgo de depresión y mediante la vigilancia de la actividad relacionada con la enfermedad a largo plazo y la generación de patrones de actividad a largo plazo. Cuando los síntomas de estas enfermedades aparecen en personas con patrones de actividad irregulares, la información se envía a los médicos y cuidadores para la detección temprana y la prevención de la depresión y la diabetes. Según [Ras13] las herramientas AAL deben estar soportadas por varios algoritmos y técnicas, en

concreto, las siguientes: *el reconocimiento de la actividad, el modelado del contexto, la identificación de ubicación, y la planificación y detección de anomalías.*

- ***Reconocimiento de Actividad basados en dispositivos móviles.***

Estas técnicas están basadas en los sensores que contienen los "Smartphone" actuales, como el acelerómetro y el giroscopio, y en su transformación en series temporales. Las acciones más simples tales como caminar, trotar y correr pueden ser representados en forma de patrones de series de tiempo periódicos y tratadas de diferentes maneras para su modelado, como por ejemplo, la extracción de Fourier y FFT [Keo04]. Este tipo de modelado requiere modelos supervisados de aprendizaje para descubrir patrones de actividad [Tan05]. Lara et al. [Lar12] propuso Centinela, un sistema que monitorea continuamente cinco actividades (caminar, correr, sentarse, ascender y descender) utilizando un único dispositivo de detección y un teléfono móvil. Chernbumroong et al. [Che14a] propuso un práctico sistema de reconocimiento de actividad multi-sensor para analizar las actividades diarias de los residentes, incluyendo ejercicio, alimentación, planchado, lectura, fregado, paseos, lavar, ver y limpiar usando siete tipos de sensores conectados al cuerpo.

- ***Reconocimiento de la Actividad analizando la interacción con el entorno.***

Las técnicas englobadas en este grupo sirven para reconocer actividades más complejas, y está basada en una red de sensores interconectados que se usan para modelar las actividades del residente, en concreto con su interacción con el medio (por dónde se mueve, cuándo, etc…). En [Sur13] desarrollaron un sistema inteligente de monitoreo domiciliario para detectar cambios de comportamiento y pronosticar el comportamiento de las personas mayores, con sensores inalámbricos ubicados en aparatos (tostadoras, cama, televisión, radiadores), y con una aproximación estadística al etiquetado de actividad. Sin embargo, la mayoría de los algoritmos de reconocimiento de actividad ambiental son supervisados, y se basan en datos etiquetados previamente para su entrenamiento [Wad08]. Estos métodos incluyen árboles de decisión [Mau06], redes neuronales [Moz68], razonamientos basados en casos [Mat05] y métodos de modelado de Marcov [Lia06]. A pesar de su prevalencia, los métodos supervisados no escalan bien en el mundo real. En primer lugar, la suposición de actividades predefinidas no se cumple en la realidad. Debido a factores físicos, mentales, culturales y de estilo de vida, no todas las personas realizan el mismo conjunto de tareas. Por otro lado, recoger los datos y transformarlos en diversas actividades etiquetadas es una tarea muy lenta y laboriosa. Para abordar estos problemas, se utilizan métodos de minería de datos, como "activity streaming mining" [Ras10] o "minería de actividad secuencial" [Ras11], o técnicas como el aprendizaje semi-supervisado [Tom18], y otras como "transfer learning" [Zhe10]. Una de las propuestas más interesantes es el análisis basado en el paradigma del "Context Aware" o de la "Situación Contextual". Los sistemas AAL deben representar muchos tipos diferentes de información que varía según el contexto temporal y espacial, así, la misma información del mismo sensor interior (ubicación en la cocina), puede no indicar la misma actividad en función de la hora del día (desayuno, comida o cena), la época (Navidad = Reunión Familiar), o en función de los perfiles de usuario y preferencias (distintas informaciones temporales, clínicas o espaciales, por ejemplo, la disposición de la residencia y sus alrededores). Algunas aproximaciones, como [Cic16] se basan en diagramas UML sobre un sistema de agentes para modelar la actividad en función del contexto entre varios sensores. Otros trabajos se

basan en el razonamiento basado en el contexto, por ejemplo en [Kwo12]. El objetivo de este trabajo es mejorar la precisión del diagnóstico del estado de salud de una persona con un único sensor de actividad. La deficiencia de esta técnica es que cuando los datos de la muestra son incompletos y algo inconsistentes (principalmente debido a las condiciones del sensor), el razonamiento del proceso no es muy exacto, es decir, el sistema es muy sensible a la calidad de los datos. Según los estudios, los modelos basados en ontologías son interesantes en el modelado del conocimiento [Qin15], ya que proporcionan un acuerdo explícito común en la representación de los conceptos de forma jerárquica, utilizando propiedades, subclases y superclases, y permiten etiquetar de una forma rápida y efectiva los datos en bruto de los sensores, en base a su "situación contextual". Este tipo de modelado es particularmente interesante a la hora de representar la información contextual de la actividad, ya que es capaz de esquematizar los conceptos de una manera jerárquica que ha sido explícitamente acordada. Además, las ontologías y sus estándares permiten compartir el conocimiento y mejoran la interoperabilidad semántica, así como, adicionalmente, proporcionan mecanismos de razonamiento sobre los hechos de la propia ontología. Un ejemplo del dinamismo de estas tecnologías es, como muestra, una ontología estándar para la programación de aplicaciones basadas en la información de contexto (SOUPA) [Che04]. Como ejemplos de investigaciones previas, [Yej15a] propone un modelo de ontología formal reutilizable para describir el contexto del entorno inteligente. El modelo consta de cuatro componentes: objeto, ubicación, sensor y actividad. Además, enfoca dos tipos de variables cuantitativas respecto a las actividades: tiempo y duración. Por ejemplo, la actividad "Desayuno" debe ocurrir en la mañana (6 am a 12 am) y la actividad "tomar la ducha" no debe durar más de 5 min. Chen et al. [Che12] propuso un modelo formal de ontología ADL para establecer vínculos entre las actividades e información contextual a través de las propiedades basadas en la actividad. Utilizaron el estándar OWL (Ontología Web Lenguaje) para el modelado ontológico y su representación y manipulación. La característica principal de su modelo de ontología es que puede modelar el conocimiento del dominio en dos niveles de abstracción: el nivel conceptual, en el cual una clase de actividad es descrita por una serie de propiedades según el conocimiento de la actividad genérica; y el nivel específico, en el cual la manera especial que un usuario realiza una actividad puede ser modelada como una instancia. Un trabajo similar se puede encontrar en [Oke14]. Las representaciones del conocimiento basadas en ontologías muestran claras ventajas para el conocimiento del contexto compartido entre diferentes entidades mediante el uso de formalismos OWL. También funcionan bien como una solución para capturar las informaciones de los sensores en términos de heterogeneidad, interoperabilidad, y usabilidad gracias a las herramientas de gestión de grafos semántico, (por ejemplo, Protegé), aunque requieren un conocimiento fuerte del dominio, y por si solas, no tienen capacidad de razonamiento sobre contextos con incertidumbre. Con respecto al reconocimiento de actividad, en la literatura se han descrito enfoques de segmentación de actividad (de datos concurrentes), basados en mediciones de similitud entre los eventos que devuelven los sensores. Por ejemplo, [Yej15a] desarrolló un enfoque basado en un método de segmentación de datos brutos basado en la semántica temporal, espacial y de objeto del sensor y de sus eventos. Ésta medida semántica se utiliza para evaluar la similitud semántica entre dos eventos de sensor adyacentes y por consiguiente, se traduce en un método para dividir, de manera dinámica, una secuencia de eventos en segmentos, y a cada segmento, se le asigna una actividad. [Yej15b] propuso un mecanismo multiusuario en tiempo real basado en la semántica de la actividad para dividir los valores continuos de un sensor en una secuencia de fragmentos. En este caso, la

segmentación se realiza utilizando patrones de eventos obtenidos analizando la disimilitud semántica entre eventos sensoriales. En comparación con otros enfoques, la segmentación semántica puede detectar el límite de las actividades concurrentes con mayor precisión, así como producir un número de particiones mejor ajustadas a las actividades reales. Otros enfoques hibridan razonamiento semántico, con razonamiento probabilístico, como en [Rib16] para refinar las hipótesis de actividad inicial semánticas.

- ***Reconocimiento de la actividad basada en la visión y otros métodos***

Otras técnicas se basan en el reconocimiento de la actividad basada en técnicas de visión artificial, que proporcionan información contextual muy detallada. Sin embargo, también se enfrentan a dificultades importantes respecto a variaciones en los entornos naturales, complejidad algorítmica y, principalmente, reticencias de los residentes y preocupación sobre la privacidad. Gjoreski et al. [Gjo14] propuso monitorear la actividad diaria de los usuarios gracias a la combinación de dos acelerómetros y un sensor de electrocardiograma (ECG). Zhuang et al. [Zhu09] describió un sistema de detección de caídas para distinguir el ruido proveniente de las caídas de otros ruidos en el entorno doméstico inteligente. En su sistema ellos sólo utilizan un micrófono de campo lejano para identificar varios sonidos. Usan un sistema de vectores GMM (Gaussian Mixture Models) para modelar cada segmento de caída o ruido y una máquina de soporte vectorial construida sobre un supervisor de GMM.

- ***Predicción del comportamiento y detección de anomalías***

La detección de anomalías se refiere al problema de encontrar patrones en los datos que no se ajustan a la conducta prevista. Así, las técnicas de Inteligencia Ambiental Predictiva (PAI) utilizadas en el entorno de un hogar “inteligente” se diseñan con el fin de analizar el comportamiento del habitante bajo un entorno de monitorización. El sistema recopila la información de las redes de los sensores. Los datos recopilados se usan para modelar el comportamiento del habitante en diferentes momentos utilizando métodos de predicción. La predicción implica la extracción de patrones relacionados con las activaciones del sensor, es decir, el reconocimiento de actividades visto previamente. Finalmente, estos modelos predictivos se utilizan para clasificar la secuencia de actividades y predecir la actividad siguiente, y chequear con la realidad inmediata. En algunos estudios se ha utilizado un banco de pruebas usando el “Sistema de Inferencia Fuzzy Online Adaptativo” (AOFIS), consistente en un mecanismo de predicción en varias fases para el aprendizaje, control y adaptación, en basado en reglas difusas [Doc04]. Otros métodos incluyen redes neuronales y utilizan técnicas de aprendizaje automático para extraer los patrones ADL (“Activities of Daily Living”) de las actividades diarias observadas. Estos patrones se utilizan posteriormente como modelos predictivos [Brd09]. Sin embargo, estas técnicas, en general, necesitan de una solución alternativa (necesitan actualizaciones periódicas) si se cambia el entorno de ejecución, con lo que pueden existir problemas de inadecuación de datos para adaptarse a un nuevo entorno. Existen otros métodos para derivar modelos de actividad anormal, desde un modelo normal general a través de un Kernel de Regresión no lineal (KNLR) y Máquinas de Vector de Soporte (SVM) y Modelos Ocultos de Marcov, incluyendo modelos para reducir la tasa de falsos positivos en un sistema no supervisado [Cha12]. En [Bed12], los autores propusieron una metodología de clasificación para reconocer el

movimiento humano usando datos de aceleración para diferentes clases de movimientos, como conducir un coche, estar en un tren, y caminar, y lo modelaron comparando diferentes técnicas de aprendizaje (Random Forests, SVM y Naive Bayes). Los autores mostraron que el “Random Forest” proporciona una mayor precisión promedio superando a los SVMs y Naive Bayes. Existen otros enfoques para detectar y estudiar automáticamente patrones de comportamiento en el hogar. Uno de ellos [Bam10] utiliza, como nuestro trabajo, una red de sensores domésticos para rastrear el movimiento del usuario y diferentes etapas en su hogar. Usando la zona de inicio y el tiempo de ocupación, crean un conjunto de códigos que definen la zona, hora del día, duración de la presencia, con el fin de descubrir frecuentes secuencias de códigos en un conjunto de datos de 30 días. Diane Cook ha estado estudiando la minería de datos y el enfoque de aprendizaje de máquina para reconocer las actividades de la vida cotidiana, con el fin de encontrar el patrón más común en los datos de sensores de movimiento [Spa14]. Otros trabajos recientes, como Jakkula [Jak11] [Got15] se centraron en el uso de máquinas vectoriales de apoyo (SVM) para clasificar el comportamiento anómalo utilizando un conjunto de datos basado en sensores de puerta en casa, pero anotados manualmente. En [Vuo11] se modelan automáticamente los patrones de los pacientes para detectar patrones anormales en pacientes con demencia. [Kim09] tratan de distinguir patrones erróneos de patrones normales en un domicilio utilizando la información de tiempo y ubicación. En [Cam10] han desarrollado también métodos para detectar clases de secuencias normales, y generar alertas cuando dichas secuencias son inusuales utilizando una red neuronal. Respecto a la predicción enfocada a la detección de anomalías, se han propuesto algunos métodos basados en la agrupación, métodos estadísticos, y métodos de información teórica, entre otros [Cha09]. En [Phu09], Phua et al. propuso un sistema de reconocimiento de “planes erróneos· para detectar anomalías en las actividades diarias de personas ancianas con demencia, mediante el uso de sensores desplegados en el domicilio para monitorear diariamente las ocupaciones de sus residentes. Cuando se detecta un error, se envían oportunamente avisos de audio o visuales a los pacientes con demencia para reemplazar parte de su memoria disminuida y mejorar sus habilidades de resolución de problemas. Trabajos similares sobre demencia se pueden encontrar en [Rib15]. Ordóñez et al. [Ord15] desarrollaron un sistema automatizado de análisis del comportamiento para las personas mayores que viven solos en casa gracias a la captura de las mediciones de varios sensores, que detectan las actividades de cada usuario y son capaces de detectar comportamientos anómalos que reflejen cambios en el estado de salud, mediante el aprendizaje estándar de patrones de comportamiento. [Das05] utilizaron técnicas de predicción basada en Markov, con una confianza de hasta un 86% de precisión para los datos que contienen variaciones. En 1998, Michael C. Mozer De la Universidad de Colorado, implementó un sistema llamado ACHE “Adaptive Control of Home Environments” [Moz95] que utilizó redes neuronales para modelar patrones del usuario y realizar acciones sin la ayuda del usuario. El sistema monitoriza el medio ambiente, observaba las acciones tomadas por el usuario y trataba de aprender los patrones. ACHE está equipado con sensores para Informar sobre el estado del medio ambiente. Otros modelos, se han basado en técnicas de extracción de reglas temporales [Ras10]. [Nou12] han "demostrado la viabilidad para producir datos simulados que imitan los datos recogidos por presencia de sensores en condiciones de campo "; y que puedan generar una alarma siempre que los datos recogidos reales sean significativamente diferentes de los datos simulados”. Otros estudios han enfocado los análisis con otras técnicas, como en [Sur13], que analizan series temporales sobre entornos de laboratorio, en concreto, con la

base de datos CASAS [Ras09]. Por lo general, el análisis orientado por los intereses tiende a pasar por alto los patrones inesperados en los datos. Para evitar esta inconveniencia, algunos autores proponen la utilización de algoritmos no supervisados (modelos de agrupación y asociación), pero enfocados al análisis de anomalías [Niu07]. La herramienta conocida como “Anomaly Detection (AD)”, es una herramienta extremadamente útil, que funciona junto con las técnicas de Clustering, y que permite el reconocimiento de conjuntos de datos cuyo comportamiento es muy diferente del resto de los datos, o con un patrón desconocido o que no puede etiquetar los datos de forma fiable. A menudo, estos elementos se conocen como “Outliers” o atípicos [Mil89]. La DA también se conoce como detección de desviaciones, porque los objetos anómalos tienen valores de atributos con una desviación significativa de los valores esperados típicos. Aunque estas anormalidades a menudo se tratan como ruido o error en muchas operaciones, son una valiosa herramienta en la búsqueda de comportamientos atípicos [Liu14]. Otro autores utilizan para el mismo objetivo el algoritmo denominado LOF (Factor Local Outlier)[Bre00]. Este algoritmo compara la densidad de instancias de datos alrededor de una instancia dada, y en base a dichos valores, se determina qué instancias son anómalas. Las distancias en la búsqueda de estos atípicos se pueden medir mediante técnicas basadas en modelos estadísticos o distancias basadas en la “densidad de cada región” [Sin95], en dónde los objetos situados en regiones de baja densidad y relativamente distantes de sus vecinos se consideran anómalos. La característica principal de este algoritmo es que se considera un aprendizaje no supervisado y es capaz de asignar una puntuación a cada instancia que refleja el grado en que la instancia es anómala.

- ***Verificación de Identidad***

La verificación de la identidad es otro componente recurrente en aplicaciones AAL. Se trata de identificar quién está realizando alguna actividad en el hogar (pasando por el pasillo o tomando el medicamento). Para distinguir a los residentes se han adoptado enfoques generales, un enfoque de identificación activa, que utiliza herramientas tales como tarjetas RFID para identificar a los residentes [Wan09], mientras que otro método, como el denominado “enfoque anónimo”, utiliza métodos de aprendizaje automático para construir modelos únicos de movimiento de cada residente [Chi10]. Los métodos de identificación activa son más precisos, pero por otro lado, los métodos anónimos proporcionan una solución menos invasiva y más viable cara a su comercialización, dado que el usuario no está obligado a llevar un sensor encima.

- ***Arquitectura Orientada a Servicios***

La arquitectura necesaria dentro del trabajo realizado para la coordinación de los distintos elementos está orientada a dar soporte a esta granularidad es una Arquitectura Orientada a Servicios (Service Oriented Architecture) mediante el uso de servicios web. Se basa en la utilización de servicios para dar soporte a una o varias aplicaciones. Esto significa que a la hora de implementar aplicaciones, en lugar de desarrollar una gran aplicación que haga cientos de cosas, se desarrollan pequeños servicios independientes. Estos servicios son utilizados por la aplicación, que es mucho más compacta y flexible. En nuestro caso, esta información es capturada desde los dispositivos ubicados en las casas de los distintos clientes, para una vez almacenada y procesada, ser expuesta como servicios para su consumo desde los dispositivos móviles o desde el sistema

central. Las ventajas principales de la utilización de servicios son: a) la facilidad de mantenimiento, ya que cada servicio es independiente del resto, b) la escalabilidad del sistema, y es que para añadir nuevas funcionalidades basta con añadir nuevos servicios y c) la poca acoplación de las aplicaciones, ya que cada una de los servicios se comporta de manera independiente al resto. Los Servicios Web definidos por el W3C como sistemas software diseñados para soportar una interacción interoperable máquina a máquina sobre una red. Los Servicios Web suelen ser APIs Web que pueden ser accedidas dentro de una red (principalmente Internet) y son ejecutados en el sistema que los aloja.

Los dos tipos de arquitecturas de servicios más extendidas en la actualidad son:

- Servicios web basados en el protocolo (SOAP). Los Servicios Web pueden ser implementados siguiendo los conceptos de la arquitectura SOA, donde la unidad básica de comunicación es el mensaje, más que la operación. Esto es típicamente referenciado como servicios orientados a mensajes. Los Servicios Web basados en SOA son soportados por la mayor parte de desarrolladores de software y analistas. Al contrario que los Servicios Web basados en RPC, este estilo es débilmente acoplado, lo cual es preferible ya que se centra en el "contrato" proporcionado por el documento WSDL, más que en los detalles de implementación subyacentes.
- REST (REpresentation State Transfer). Los Servicios Web basados en REST intentan emular al protocolo HTTP o protocolos similares mediante la restricción de establecer la interfaz a un conjunto conocido de operaciones.

- ***UniversAAL: la plataforma europea de referencia en sistemas de teleasistencia***

Es importante resaltar que el sistema realizado en este trabajo ha sido admitido en el estándar universAAL [Kor18], que es la plataforma europea de referencia en sistemas de teleasistencia. Para formar parte de la arquitectura de referencia UniversAAL, el sistema debe cumplir una serie de requisitos para hacerlo. Una particularidad del software es que se debe ejecutar sobre una plataforma OSGi [Sta18]. UniversAAL utiliza OSGi y RDF como su pilar de conocimiento. Conjuntamente con una API llena de funciones útiles para usar en el dominio AAL, UniversAAL es una base poderosa para desarrollar aplicaciones de Teleasistencia. La tecnología OSGi facilita la integridad de los distintos módulos y aplicaciones de software y asegura la gestión remota y la interoperabilidad de aplicaciones y servicios a través de una amplia variedad de dispositivos. Construir sistemas desde módulos OSGi aumenta la productividad del desarrollo y los hace mucho más fáciles de modificar y evolucionar. La anotación semántica de la información en formatos RDF (Resource Description Framework) es la otra base de la potencia de la plataforma desarrollada, y que encaja con la tecnología de universAAL. El modelo de datos RDF es similar a los enfoques clásicos de modelado conceptual, tales como los diagramas entidad-relación o clase, ya que se basa en la idea de hacer declaraciones sobre recursos (en particular recursos web) en forma de expresiones sujeto-predicado-objeto, es decir, en relaciones basados en la lógica formal de primer orden. Estas expresiones se conocen como triplas o tripletas en la terminología RDF. El sujeto denota el recurso, y el predicado denota rasgos o aspectos del recurso y expresa una relación entre el sujeto y el objeto.

2.6 Sensórica y Redes Inalámbricas

Las redes inalámbricas de sensores, conocidas en inglés como Wireless Sensor (WSN), ofrecen la posibilidad de realizar las instalaciones en los hogares sin necesidad de realizar instalaciones profesionales. Estas redes inalámbricas de sensores, están constituidas por un conjunto de dispositivos autónomos llamados nodos, distribuidos por la zona a monitorizar, que son capaces de obtener información del entorno y enviarla de forma inalámbrica a un punto central o coordinador, como los expuestos en el punto anterior. Todas las redes WSN presentan unas características comunes, asociadas a la utilización de un entorno inalámbrico, entre ellas, se encuentran:

- Generalmente son redes Ad-hoc, sin infraestructura específica y sin necesidad de realizar un despliegue detallado de sensores, ya que no hará falta una previsión de cableado de la zona.
- Escasa capacidad de recursos, condicionado por el consumo energético, por lo que apenas podrán tener software de gestión.
- Fallos en la transmisión mayores que en las soluciones cableadas, lo que obliga a un tratamiento de la información recogida

Las tecnologías más habituales para la implementación de redes inalámbricas de sensores son:

- Wifi: El estándar de comunicaciones IEEE 802.11, muy utilizado en redes de sensores con comunicación con PC
- Bluetooth: muy orientado a la transmisión de pequeños volúmenes de datos, en entornos
- Zigbee estándar 802.15.4: que ofrece una comunicación entre sensores a un bajo costo, así como un reducido consumo de energía
- Redes de hardware abierto, muy relacionado con proyectos de investigación y “startups” relacionadas con IOT (Internet Of the Things) En este aspecto destacan los proyectos que utilizan hardware como Arduino o similar y sensores que inalámbricos que operan en los 433Mhz que son tecnologías específicamente diseñadas para las redes de sensores/actuadores, con un muy bajo consumo energético, y menos saturada que aquellas situadas en los 2.4 Ghz, como las anteriores.

En la tabla 2.2 se ha realizado una comparativa preliminar sobre las principales características de las 4 tecnologías. En este trabajo se ha incorporado, principalmente, tecnología ZigBee. El término ZigBee es el nombre de la especificación de un conjunto de protocolos de alto nivel de comunicación inalámbrica para su utilización con radios digitales de bajo consumo, basada en el estándar IEEE 802.15.4 de redes inalámbricas de área personal (wireless personal area network, WPAN). Su objetivo son las aplicaciones para redes Wireless que requieren comunicaciones seguras y fiables con baja tasa de envío de datos , coste contenido y maximización de la vida útil de sus baterías. El protocolo es el trabajo de más de 70 compañías., entre las que se encuentran Motorola, Mitshubishi, HoneyWeb, que desde la primera versión del consorcio en 1998, se han asociado formando formado la Alianza ZigBee. El estándar ZigBee enfoca a un segmento del mercado no atendido por los estándares anteriores, que se caracteriza por tener una baja necesidad en cuanto a la transmisión de datos, bajo ciclo de servicio de conectividad, que cubre nichos de mercado como la monitorización de instalaciones, lo que le hace un firme candidato a su utilización en soluciones de

monitorización de hogares como en el sistema. ZigBee es un estándar de hardware y software basado en el estándar IEEE 802.15.4. que permite la interoperabilidad entre dispositivos fabricados por compañías diferentes.

Tabla 2.2 Comparativa en Tecnologías Inalámbricas

	Wifi	Bluetooth	ZigBee	IoT
Velocidad	<50 Mpbs	1 Mpbs	<250 kpbs	<64 kpbs
Núm. nodos	32	8	255/65535	255/65535
Duración batería	Horas	Días	Años	Años
Consumo transm.	400 ma	40 ma	30ma	6ma
Consumo reposo	20 ma	0.2 ma	3ua	0.2 ua
Precio	Caro	Medio	Medio	barato
Configuración	Compleja	Compleja	Simple	Muy simple

El estándar IEEE 802.15.4. define el hardware y el software, en los términos de conexión de redes, como la capa físicas (PHY), y la capa de control de acceso al medio (MAC). Mientras que la alianza ZigBee ha añadido las especificaciones de las capas red (NWK), y aplicación (APL) para completar lo que se llama la pila o stack ZigBee. En las instalaciones con el hardware ZigBee existen tres tipos distintos de dispositivo según su papel en la red:

- Coordinador ZigBee (ZigBee Coordinator, ZC). El tipo de dispositivo más completo. Debe existir al menos uno por red. Sus funciones son las de encargarse de controlar la red y los caminos que deben seguir los dispositivos para conectarse entre ellos.
- Router ZigBee (ZigBee Router, ZR). Interconecta dispositivos separados en la topología de la red, además de ofrecer un nivel de aplicación para la ejecución de código de usuario.
- Dispositivo final (ZigBee End Device, ZED). Posee la funcionalidad necesaria para comunicarse con su nodo padre (el coordinador o un router), pero no puede transmitir información destinada a otros dispositivos. De esta forma, este tipo de nodo puede estar dormido la mayor parte del tiempo, aumentando la vida media de sus baterías. Un ZED tiene requerimientos mínimos de memoria y es por tanto significativamente más barato.

También se pueden clasificar los elementos hardware de la instalación basándose en su funcionalidad, planteándose una segunda clasificación:

- Dispositivo de funcionalidad completa (FFD): También conocidos como nodo activo. Es capaz de recibir mensajes en formato 802.15.4. Gracias a la memoria adicional y a la capacidad de computar, puede funcionar como Coordinador o Router ZigBee, o puede ser usado en dispositivos de red que actúen de interfaz con los usuarios.
- Dispositivo de funcionalidad reducida (RFD): También conocido como nodo pasivo. Tiene capacidad y funcionalidad limitadas (especificada en el estándar) con el objetivo de conseguir un bajo coste y una gran simplicidad. Básicamente, son los sensores/actuadores de la red.

Entre las principales ventajas que aporta para soluciones en el hogar como en el sistema esta que permite fácilmente crear conexiones punto a punto y punto a multipunto, así como múltiples topologías de red: estática, dinámica, estrella y malla, lo que permitiría una mayor flexibilidad a la hora de crear redes de sensores en el hogar. Además cuenta con tecnologías como la detección de energía (ED) y un bajo ciclo de trabajo , lo que proporciona una mayor duración de la batería . Además cuenta con un cifrado 128-bit AES de cifrado y mecanismos de identificación, autenticación y autorización., lo que permitiría despliegues en los hogares cifrados de manera segura. Entre sus desventajas se encuentra esta que proporcionalmente más cara que otras alternativas open Source, y al estar implementa toda la pila OSI, implica mayor complejidad en los desarrollos acometidos.

Capítulo 3

Razonamiento y Aprendizaje Automático

En este capítulo revisamos los conceptos de razonamiento y aprendizaje automático utilizados en los trabajos de la tesis. Puesto que son técnicas bien conocidas en la comunidad científica/tecnológica, con una amplia literatura que los ha popularizado, no nos extendemos en los detalles. También proporcionamos una revisión de los aspectos metodológicos, específicamente la validación cruzada. La sección 3.1 introduce los sistemas basados en reglas. La sección 3.2 discute los conceptos de aprendizaje automático y extracción de conocimiento a partir de los datos. La sección 3.3 discute el concepto de Detección de Anomalías, y la sección 3.4 presenta los algoritmos que se han usado con respecto al análisis de series temporales.

3.1 Sistemas Expertos basados en reglas

Los sistemas expertos representan el conocimiento formal para resolver problemas humanos. Este tipo de sistemas son aplicables a cualquier dominio y están presentes hoy en casi cualquier aplicación que requiera un alto costo computacional para automatizar procesos con algún razonamiento. En general, son adecuados para tareas específicas que requieren mucho conocimiento, derivado de una experiencia de dominio particular como diagnósticos, instrucciones, predicciones o consejos a situaciones reales que surgen y también pueden servir como herramientas de entrenamiento, imitando el comportamiento humano. La representación explícita y formal del conocimiento sobre un problema requiere el uso de técnicas particulares. En el campo de la representación simbólica del conocimiento, dentro de la inteligencia artificial, se han propuesto diversas formas de representación. En general, una forma de representación del conocimiento debe satisfacer los siguientes requisitos:

- Formal. La representación no debe presentar ambigüedades. Por ejemplo, el lenguaje natural no se considera representación del conocimiento debido a las ambigüedades que presenta.
- Expresiva. La representación debe ser suficientemente rica como para capturar los diferentes aspectos que sea necesario distinguir. Por ejemplo, las fórmulas lógicas de cálculo de predicados constituyen una representación más expresiva que la que se maneja cálculo proposicional.
- Natural. La representación debe ser suficientemente análoga a formas naturales de expresar conocimiento. En este sentido, las representaciones matemáticas tradicionales y cuantitativas (por ejemplo, las matrices) pueden resultar muy artificiales para emular procesos de razonamiento.
- Tratable. La representación se debe poder tratar computacionalmente, es decir, deben existir procedimientos suficientemente eficientes para generar respuestas a través de la manipulación de los elementos de las bases de conocimiento.

Algunas de las más conocidas técnicas que satisfacen los anteriores requisitos son técnicas básicas como las **reglas**, los **marcos**, las **restricciones**, las **cláusulas lógicas**, y otras más específicas como los **árboles de decisión**, usados en este trabajo. Estas formas de representación han sido ampliamente utilizadas en la construcción de sistemas inteligentes con las que, progresivamente, se ha abordado la construcción de sistemas más complejos[17]. Para considerar de una forma más estructurada el desarrollo de sistemas en problemas con mayor volumen o complejidad de conocimiento se han propuesto técnicas adicionales que complementan las técnicas básicas. Por ejemplo, los **contextos** son una forma de modularización que se basa en realizar una partición en bases separadas de forma que cada contexto puede corresponder, por ejemplo, a un área de conocimiento relativamente independiente o a una fase del razonamiento. Los contextos mejoran la eficiencia dado que las búsquedas se hacen de forma local y mejoran el mantenimiento de la base pero son limitados en problemas complejos dado que aportan una estructura plana con un único motor de inferencia. Otra solución es ***reunir varias formas de representación en una representación múltiple con un único motor de inferencia***. Aunque esta última solución potencia la representación mediante la suma de varias técnicas, es una opción que puede dar lugar a bases de conocimiento heterogéneas de difícil mantenimiento. Este tipo de soluciones parciales pueden considerarse útiles en la construcción particular de ciertos sistemas cuyas características hagan adecuadas el empleo de alguna de dichas soluciones. Sin embargo, en problemas complejos, es necesario ir a enfoques más avanzados, que permitan una adecuada modularización y estructuración. En la descripción de sistemas inteligentes es útil plantear un nivel superior, al que se puede denominar nivel de representación simbólica, en donde el sistema se contempla formado por bases de conocimiento con representaciones como reglas, marcos, etc. además de procedimientos de inferencia. Este es el nivel en el que se realiza el diseño del sistema inteligente haciendo uso de las técnicas tradicionales de representación del conocimiento del campo de inteligencia artificial

Principalmente existen tres tipos de sistemas expertos [Fde00]:

- Basados en reglas previamente establecidas.
- Basados en casos o CBR (Case Based Reasoning).

Los componentes de un sistema experto, básicamente son los siguientes:

- Base de conocimiento. Es la parte del sistema experto que contiene el conocimiento sobre el dominio. hay que obtener el conocimiento del experto y codificarlo en la base de conocimientos. Una forma clásica de representar el conocimiento en un sistema experto son lar reglas. Una regla es una estructura condicional que relaciona lógicamente la información contenida en la parte del antecedente con otra información contenida en la parte del consecuente.
- Base de hechos (Memoria de trabajo). Contiene los hechos sobre un problema que se han descubierto durante una consulta. Durante una consulta con el sistema experto, el usuario introduce la información del problema actual en la base de hechos. El sistema empareja esta información con el conocimiento disponible en la base de conocimientos para deducir nuevos hechos.

[17] Molina, Martin. Métodos de resolución de problemas: Aplicación al diseño de sistemas inteligentes. Martin Molina, 2006.http://oa.upm.es/14207/

- Motor de inferencia. El sistema experto modela el proceso de razonamiento humano con un módulo conocido como el motor de inferencia. Dicho motor de inferencia trabaja con la información contenida en la base de conocimientos y la base de hechos para deducir nuevos hechos. Contrasta los hechos particulares de la base de hechos con el conocimiento contenido en la base de conocimientos para obtener conclusiones acerca del problema.
- Subsistema de explicación. Una característica de los sistemas expertos es su habilidad para explicar su razonamiento. Usando el módulo del subsistema de explicación, un sistema experto puede proporcionar una explicación al usuario de por qué está haciendo una pregunta y cómo ha llegado a una conclusión. Este módulo proporciona beneficios tanto al diseñador del sistema como al usuario. El diseñador puede usarlo para detectar errores y el usuario se beneficia de la transparencia del sistema.
- Interfaz de usuario. La interacción entre un sistema experto y un usuario se realiza en lenguaje natural. También es altamente interactiva y sigue el patrón de la conversación entre seres humanos. Para conducir este proceso de manera aceptable para el usuario es especialmente importante el diseño del interfaz de usuario. Un requerimiento básico del interfaz es la habilidad de hacer preguntas. Para obtener información fiable del usuario hay que poner especial cuidado en el diseño de las cuestiones. Esto puede requerir diseñar el interfaz usando menús o gráficos.

Las formas de razonamiento diagnóstico tienen similitud con los razonamientos de los sistemas expertos:

- Probabilísticas. Se basan en la frecuencia de ocurrencia de los patrones de comportamientos y consideran variables como sexo, edad, peso, frecuencia y la probabilidad asociada entre indicadores-acción.
- Causales. Encuentran relaciones entre los eventos y las relacionan con los efectos que causan, que pueden ser datos antecedentes, como por ejemplo, el tiempo atmosférico y la tendencia a salir de casa.
- Deterministas. Son mucho más directos, ya que identificando cada estado, se asocia con una regla que lleva directamente hacia la conclusión

En las ciencias de la computación, hay dos métodos básicos para buscar una solución de un problema de razonamiento, ambos basados en la regla de inferencia Modus Ponens. El primer método es el impulsado por datos y es conocido como encadenamiento hacia delante o “forward chaining”, y el segundo es dirigido por las consultas, y se llama encadenamiento hacia atrás o “backward chaining”. El hecho V (el lado izquierdo de la regla) suele ser un conjunto de datos o información v1, v2,…, vn. La conclusión o inferencia obtenida, W, La derivación de encadenamiento hacia delante de W (la derecha de la regla), necesita verificar si el conjunto de informaciones que componen V, (v1, v2,…, vn), pertenecen al conjunto de hechos aceptados en el contexto del problema. El reto es de estos algoritmos es chequear si para todas las reglas

$$v1v2 \dots vn \rightarrow W$$

donde

$$\{v1, v2, \dots, vn\} \subseteq Deduced\ or\ not,$$

es decir, si todas las informaciones necesarias para inducir el hecho W, pueden ser generadas en base a otras reglas previas. Un algoritmo sencillo para esta comprobación consiste en enumerar todas las posibles reglas e ir generando el conjunto de deducciones de forma recurrente, y desarrollar un patrón secuencial de

coincidencia entre ellos. En el marco de los sistemas Expertos Charles L. Forgy en su Ph.D. Disertación en la Universidad Carnegie-Mellon en 1979 propuso una solución que llamó algoritmo Rete [For79] [For88]. Rete es hoy en día la base de muchos sistemas expertos muy famosos, incluyendo CLIPS, Jess, y Drools [Thi07]. Un sistema experto basado en Rete construye una red de nodos, donde cada uno de ellos (excepto el nodo raíz) representa un patrón que aparece en la parte izquierda (el condicional) de una regla. Por lo tanto, el camino desde el nodo raíz a una hoja define la parte condicional entera de una regla. Cada nodo tiene una memoria de hechos que satisfacen su patrón. A medida que se añaden o modifican hechos, se propagan los cambios por la red, haciendo que los nodos que se activan con el patrón se activen. Cuando un hecho o un conjunto de ellos hacen que todos los patrones de una regla se satisfagan, se llega a un nodo hoja y la regla es activada. Básicamente, el algoritmo Rete sacrifica memoria para incrementar velocidad de procesamiento. En la mayoría de los casos el incremento de velocidad comparado con la implementación simple es de varios órdenes de magnitud (porque teóricamente el rendimiento de Rete es independiente del número de reglas del sistema). En sistemas expertos muy grandes, sin embargo, Rete suele presentar problemas por su gran cantidad de consumo memoria.

Otra vertiente de la aplicación de los sistemas expertos es su aplicación sobre ontologías. Las ontologías son representaciones del conocimiento humano sobre un dominio (área de interés) determinado. Se trata de describir los conceptos y relaciones existentes entre los componentes de ese dominio de tal manera que puedan ser interpretados y manipulados por software. En el dominio de tecnologías de la información, una ontología es un listado de términos, propiedades, relaciones, etc. que pretenden definir de una forma genérica un área de conocimiento, de tal forma que pueda ser reutilizado por diferentes grupos y aplicaciones. Una de las razones fundamentales de que las ontologías sean tan utilizadas es que permiten un entendimiento común y compartido de algún dominio y que puede ser comunicado a través de las personas y sistemas, es decir, se crea un “idioma” específico para un área concreta. Las funcionalidades principales que tiene una ontología son las siguientes:

- Clarificación de la estructura de conocimiento: creando conceptualizaciones que subyacen al conocimiento lograremos un vocabulario válido para representar el conocimiento.
- Compartición del conocimiento: las ontologías facilitan este proceso.

El Razonamiento automatizado se puede aplicar sobre un conjunto de datos que se materializa en estructuras semánticas (tripletas) inferidas que explotan los axiomas de ontología [Jup11]. Las ontologías aportan la tecnología base para hacer posible el acceso eficiente a los datos, la creación de nuevo conocimiento, así como la reutilización de la información. Estas tecnologías son: RDF, ontologías OWL y Notation 3, que de modo muy breve se explican a continuación.

- RDF (*Resource Description Framework*) proporciona el estándar de representación de conocimiento y modelado de la información dentro de la Web Semántica, siendo la base sobre la que se apoyan el resto de especificaciones.
- OWL (*Ontology Web Language*) proporciona una familia de estándares para la serialización de ontologías, que gozan de una gran aceptación dentro de la comunidad médica ya que la gran mayoría de las Ontologías OBO (*Open Biomedical Ontologies*) están siendo portadas a OWL debido a las grandes ventajas que ofrece.

- Notation 3 es una forma abreviada de serialización no-XML de modelos en RDF, diseñado pensando en la legibilidad por parte de humanos: N3 es mucho más compacto y fácil de leer que la notación RDF/XML [Ber08].

El razonamiento también se puede utilizar para comprobar la conformidad de los datos dentro de la ontología, especialmente con herramientas como Pellet[18]. Por último, algunas propuestas ofrecen la posibilidad de explotar la semántica de OWL ("web ontology language"), por ejemplo, en Virtuoso, utilizando la propiedad transitiva entre conceptos, se puede utilizar su extensión en las consultas SPARQL utilizando la palabra clave "transitivo". Por otra parte, OWLim[19] ofrece la posibilidad de explotar los fragmentos de la semántica de OWL mediante la aproximación de las semánticas con conjuntos de reglas. Por su naturaleza, los sistemas semánticos trabajan sobre la representación de datos estructurados (esquemas), basados en el conocimiento ya almacenado "a priori" en la propia estructura semántica, y completado por el conocimiento de expertos en forma de reglas heurísticas manuales, basadas en la evidencia, con capacidad de razonamiento utilizando un motor de inferencia. Esto significa que las reglas son bien conocidas, y siempre son verdaderas (no es posible modelar la incertidumbre). El uso de RDF ("Resource Description Framework"), y por lo tanto, representaciones asociadas, tales como RDF Schema y OWL, ofrece la posibilidad de inferencia al recuperar y consulta de información, de una manera muy similar al lenguaje natural humano, y esta es la ventaja en los sistemas de consulta de respuesta. Aunque este razonamiento automatizado no es fiable con grandes volúmenes de datos, los investigadores han comenzado recientemente a estudiar los problemas y las soluciones técnicas que se deben abordar con el fin de construir un sistema distribuido [Liu12]. El modelo apropiado debe abordar desde la definición de los conocimientos a adquirir, hasta la conceptualización y formalización de la información recopilada, tanto de las fuentes humanas y del conocimiento del usuario, como de los entradas "automáticas" de los sensores, para modelar el funcionamiento del sistema inteligente. Existen estudios de las metodologías principales existentes para extraer conocimiento, como GROVER, CommonKADS y Brulé [Ama17], y en varios trabajos se indica que las más adecuadas puedan ser las características de GROVER y CommonKADS para modelar el sistema del servicio de teleasistencia[20]. Estas metodologías indican que, una vez especificadas las fases de recogida de conocimiento, basadas en 6 etapas (Técnica para la adquisición de conocimiento, Dominio del problema, Identificación de los problemas, Conceptualización y Formalización, Implementación y Validación), todas ellas tareas claves, en la parte tecnológica, hay que plasmar de una manera formal la implementación del modelo del conocimiento. A nivel formal, la descripción de todas estas relaciones tiene su representación en las ontologías. Por su naturaleza, los sistemas semánticos (basados en el conocimiento) de apoyo a la toma de decisiones de gestión (MDSS) trabajan sobre la representación de datos estructurados o

[18] Clark, Kendall, et al. Pellet: Owl 2 reasoner for java.*https://www.w3.org/2001/sw/wiki/Pellet*

[19] Stoilov, D., and Bishop B. (2012), OWLIM-SE Reasoner [online]. Available at: http://bulgariana.eu/display/OWLIMv54/OWLIM-SE+Reasoner

[20] Metodología de adquisición de conocimiento para telecuidado inteligente en el hogar digital Ana Peñalver Blanco, Miguel Ángel Valero Duboy e Iván Pau Departamento de Ingeniería y Arquitecturas Telemáticas, Universidad Politécnica de Madrid, Madrid, España http://www.imaginar.org/taller/ecollecter/fullpapers/p62-artMetodologiaDeAdquisicionDeConocimiento.pdf

"schema". El conocimiento es persistente en las bases de datos, y el conocimiento experto (reglas del sistema) son instrucciones basadas en la evidencia heurística, con capacidad de razonamiento inferencial usando un motor de inferencia. A estas reglas semánticas se les denomina SWRL (Semantic Web Rule Language)[21]. En este trabajo se han implementado una ontología, la denominada "Ontología Clínica", con la herramienta Protegé, desarrollada por la Universidad de Stanford [Wan18]. Protegé es un programa informático "open source". Su función consiste en la creación, edición y mantenimiento de ontologías. Además de ser un programa exquisitamente diseñado en cuanto a interfaz y facilidad de funcionamiento, posee la capacidad de generar de manera automática el código fuente de cualquier ontología creada con Protegé en formato RDF/OWL. Por otro lado, la gestión de los datos que integran las dos ontologías del sistema (la Ontología de Resúmenes Clínicos, y la Ontología de la plataforma de Teleasistencia UniversAAL), se gestionan mediante el sistema de almacenamiento no-sql comercial (Virtuoso) [Sak18]. Finalmente, el motor que gestiona las reglas semánticas (SWRL) es el razonador Pellet[22].

3.2 Aprendizaje Automático.

Como complemento al punto anterior, existen MDSS no basados en el conocimiento, que aprenden de los datos brutos (semi / no estructurados), y se basan en técnicas probabilísticas: los patrones se toman como ejemplos o casos en el pasado y el sistema tiene capacidad de aprendizaje probabilístico. Históricamente, la identificación de patrones útiles en grandes conjuntos de datos se ha denominado minería de datos, donde el análisis estadístico ha sido predominante. El descubrimiento de conocimiento en bases de datos (Knowledge Discovery and Data Mining, KDD) se refiere a la utilización de técnicas de inteligencia artificial para extraer conocimiento útil a partir de datos. Algunos de los estándares como el CRISP-DM (Cross Industry Standard Process for Data Mining) es utilizado en el ámbito industrial por más de 160 empresas e instituciones de todo el mundo y su génesis se debe a la necesidad de concretar una estandarización. CRISP-DM propone un modelo neutral para la industria y herramientas, como también un modelo general de procesos para proyectos de minería de datos. KKD consta de una secuencia de cinco fases [Her04], sin embargo, existe una fase previa a la que [Kur06] hace referencia y que incluye el modelo presentado por Two Crows Corporation, también llamado por ellos "Minería de Datos para el Descubrimiento de Conocimiento". El desarrollo de descubrimiento del conocimiento es iterativo e interactivo, por lo que las fases del proceso pueden ser en cualquier momento interrumpidas para volver a comenzar en alguno de los pasos anteriores. Para que estas tecnologías sean capaces de extraer conocimiento a partir de grandes volúmenes de información, especialmente en áreas médicas y asistenciales, y sean herramientas útiles, se deben realizar de forma automática las siguientes acciones:

[21] https://www.w3.org/Submission/SWRL/

[22] https://github.com/stardog-union/pellet

- Selección automática de casos de éxito similares en el pasado.
- Unificación de pautas de actuación.
- Estudio de correlación entre indicadores y resultados.
- Control de anomalías (alertas).

Para ello se usan diferentes técnicas de Inteligencia Artificial, entre las que destacan los ***Sistemas automáticos de extracción de reglas***, o árboles de decisión. Los ***sistemas de soporte a la decisión*** o DSS son sistemas informáticos interactivos que tienen como objetivo el ayudar a los decisores en la utilización de datos y modelos para resolver problemas no estructurados. La extracción de conocimiento en bases de datos se basa en técnicas inductivas y de aprendizaje automático. Mediante los modelos extraídos se aborda la solución a problemas de predicción, clasificación y segmentación. Dependiendo del objetivo, se diferencian dos grandes grupos de problemas:

• ***Predictivos***: Predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos. En los problemas de predicción se intenta obtener un modelo que sea capaz de pronosticar la solución en casos futuros.

• ***Descubrimiento del conocimiento***: En el caso de descubrimiento de conocimiento, por el contrario, se trata de obtener información nueva a partir de los datos ya existentes. Se descubren patrones y tendencias en los datos.

Cabe hacer una distinción entre los siguientes aspectos, a nivel de funcionalidad algorítmica en la resolución de los dos tipos de problemas anteriores:

- Técnicas de verificación, en las que el sistema se limita a comprobar las hipótesis suministradas por el usuario.
- Métodos de descubrimiento, en los que se han de encontrar patrones potencialmente interesantes de forma automática.

El ***Aprendizaje Automático*** es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender en base a unos históricos, de cara a poder aplicar dicho aprendizaje a situaciones, datos o contextos no conocidos, con éxito. Una situación en la que se requiere aprender es cuando un problema no tiene una solución fácilmente abordable por técnicas de programación clásicas, entornos con multitud de variables explicativas, cuando no existe experiencia humana transferible en sistemas expertos, o esta debe ser complementada por sistemas automáticos, o cuando el problema no es fácilmente modelable. Otra situación es cuando el problema a resolver cambia en el tiempo o depende del entorno particular. El Aprendizaje Automático transforma los datos en conocimiento y proporciona sistemas de propósito general que se adaptan a las circunstancias. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. En muchas ocasiones el campo de actuación del Aprendizaje Automático se solapa con el de la Estadística, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el Aprendizaje Automático se centra más en el estudio de la Complejidad Computacional de los problemas. Muchos problemas son de clase *NP-complejo*, por lo que gran parte de la investigación realizada en Aprendizaje Automático está enfocada al diseño de soluciones factibles a esos

problemas. Entre las técnicas usadas en la Minería de Datos y el Aprendizaje Automático se encuentran, entre otras:

• ***Análisis multivariante:*** El análisis multivariante se dedica al estudio de varias variables de modo simultáneo y a determinar relaciones simultáneas entre ellas.

• ***Árboles de decisión:*** Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la resolución de un problema.

• ***Agrupamiento o Clustering:*** Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se trata de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes, y permiten detectar aquellas situaciones anómalas en función de si distancia al centro del segmento. Aplicaciones tales como detección de anomalías para la prevención de patrones anormales sobre lo que es el perfil “normal” de cada segmento extraído por el sistema son de gran ayuda a la hora de detectar patrones inusuales de una forma preventiva, con lo que los costes de correctivos posteriores se reducen.

En este trabajo, el objetivo principal es encontrar soluciones que nos ayuden a extraer conocimiento a partir de los datos, de manera que podamos, de una manera descriptiva, modelar los comportamientos o patrones de vida en cada domicilio en particular, y de un modo predictivo, poder antecederse a situaciones anómalas o de alerta, de esta forma, de un modo prescriptivo, podremos generar acciones preventivas. Este conocimiento puede obtenerse a partir de la búsqueda de conceptos, ideas o patrones estadísticamente confiables, que no son evidentes a primera vista, desconocidos anteriormente y que pueden derivarse de los datos originales.

Hay dos tipos de modelos de generador de datos, generalmente llamados "algoritmos de aprendizaje automático":

- ***El aprendizaje supervisado:*** se basa en una tarea de aprendizaje previa en función de los datos de entrenamiento etiquetados con el fin de predecir el valor de una entrada válida. Ejemplos comunes de aprendizaje supervisado incluyen la clasificación de mensajes de correo electrónico como correo no deseado, etiquetado páginas Web de acuerdo con su género, y el reconocimiento de escritura a mano. Muchos algoritmos se utilizan para crear estudiantes supervisados, las redes neuronales son más comunes, Máquinas de Vectores Soporte (SVM), Árboles de Decisión y los clasificadores de Bayes.
- ***El aprendizaje no supervisado***, se encarga de dar sentido a los datos sin ningún tipo de ejemplos de lo que es correcto o incorrecto. Se utiliza más comúnmente para agrupar entrada similar en grupos lógicos. También se puede utilizar para reducir el número de dimensiones en un conjunto de datos con el fin de centrarse en sólo los atributos más útiles, o para detectar tendencias. Enfoques comunes para el aprendizaje sin supervisión incluyen K-means, “Clustering Jerárquico”, y los mapas auto-organizados.

En este trabajo se han utilizado distintos métodos analíticos para el análisis de los datos brutos obtenidos, tanto de los sensores como de los patrones clínicos de los pacientes, en concreto, se han implementado clasificadores supervisados basados en Redes Neuronales, SVM, Bayes, y Árboles de Decisión, en dos plataformas diferentes, una comercial (SPSS Modeler, perteneciente a IBM), y en RapidMiner (una plataforma de Analítica Avanzada “OpenSource”). La dirección de las últimas investigaciones es la adición de ambas

capacidades (Sistemas Expertos basados en reglas + Sistema Automático de extracción de Reglas) en una plataforma de motor híbrido [Zho07]. Este enfoque híbrido es el que se utiliza en este trabajo, como se detalla en el Capítulo 4 de Arquitectura y en el Capítulo 6 en la Aplicación de Reglas para la Detección de Automática de Patrones.

3.2.1 Métodos de discretización de variables continuas.

Habitualmente, en las técnicas de aprendizaje automático de sistemas expertos basados en reglas, es necesario transformar valores de señales continuas (datos de sensores numéricos enviados a través de una señal variable), en datos discretos, que permitan un proceso de generalización o inducción más amplio, evitando posibles situaciones de sobreajuste o sobre aprendizaje en los sistemas de extracción de patrones y comportamientos. Esta discretización no es necesaria en otros métodos de aprendizaje, como redes neuronales o SVM, pero sí en aquellas cuya base se encuentra en la extracción de reglas más comprensibles para los técnicos humanos. Existen métodos estadísticos que permiten esta manipulación de los datos continuos de una forma matemática, creando automáticamente nuevos campos nominales en función de los valores de uno o varios campos continuos (rango numérico) existentes. Por ejemplo, puede transformar un campo de medidas en cierto sensor continuo en un campo categórico nuevo que contenga grupos de valores con el mismo tamaño muestral, o como desviaciones desde la media. Los intervalos pueden resultar útiles por varias razones, entre ellas:

- **Requisitos de algoritmos.** Algunos algoritmos como las Redes Bayesianas o las Regresión logística requieren entradas categóricas.
- **Rendimiento.** El rendimiento de algoritmos como los de regresión logística multivariables es mayor si se reduce el número de valores distintos de los campos de entrada. Por ejemplo, utilice el valor de la media o la mediana para cada intervalo en lugar de los valores originales.
- **Privacidad de los datos.** La información personal confidencial, como los salarios, se puede registrar en rangos en lugar de cifras salariales reales para proteger la privacidad.

Existen hay varias opciones de discretización: intervalos de igual amplitud, basadas en el cálculo de cuantiles, por minimización de la entropía, o por métodos de discretización proporcional y de frecuencia fija [Yan09]. En este trabajo, ha resultado de vital importancia discretizar la cantidad de tiempo que una persona permanece en una determinada ubicación en el domicilio, a una hora y día determinados, puesto que, si modelamos con los valores de frecuencia “brutos”, puede que sobreentrenemos el sistema (un residente ha pasado exactamente 43,5 minutos en una habitación a las 15:00 horas), o puede que los modelos no convergen hacía ninguna conclusión general. Así, el método utilizado en este trabajo es el denominado “***método de intervalos de cuantil***”, en el que se crean campos nominales basados en la división del indicador continuo en grupos de percentiles, o cuartiles, deciles, etc. para que, de este modo, cada grupo contenga el mismo número de registros, o bien, la suma de los valores de cada uno de ellos sea la misma. Los distintos intervalos se recodifican según un orden, así, los registros se clasifican en orden ascendente en función del valor del campo de intervalo especificado: los registros con los valores más bajos de la variable de intervalo seleccionada se les asigna un rango de 1, al siguiente conjunto de registros un rango de 2, y así sucesivamente. Los valores de

umbral de cada intervalo se generan automáticamente según los datos y el método de generación empleado. Los p-tiles disponibles son:

- **Cuartil.** Genera cuatro intervalos, cada uno con el 25% de los casos.
- **Quintil.** Genera cinco intervalos, cada uno con el 20% de los casos.
- **Decil.** Genera 10 intervalos, cada uno con el 10% de los casos.
- **Veintil.** Genera 20 intervalos, cada uno con el 5% de los casos.
- **Percentil.** Genera 100 intervalos, cada uno con el 1% de los casos.
- **N personalizado.** Seleccione esta opción para especificar el número de intervalos. Por ejemplo, un valor de 3 produciría 3 categorías agrupadas (2 puntos de corte), cada una de las cuales contendría el 33,3% los casos.

Debido a que los cuantiles son parámetros del tipo de la mediana, su cálculo se realiza de forma análoga:

$$P_k = L_i + \frac{\frac{kN}{100} - F_{i-1}}{f_i} a_i \quad k=1,2,..,99$$

Donde

- L_i es el límite inferior de la casa dónde se encuentra el percentil.
- N es la suma de las frecuencias absolutas.
- F_{i-1} es la feecuencia acumulada anterior a laclase del percentil.
- a_i es la amplitud de la clase.

Existen dos métodos diferentes utilizados para asignar registros a los intervalos.

- **Recuento de registros.** Trata de asignar el mismo número de registros a cada intervalo.
- **Suma de los valores.** Trata de asignar registros a intervalos de forma que la suma de los valores de cada intervalo sea la misma.

3.2.2 Árboles de Decisión

Dentro de la disciplina del aprendizaje automático, los árboles de decisión son unos de los métodos de aprendizaje inductivo más populares. Quinlan en [Qui90] desarrolla ID3 con la heurística de ganancia de información para desarrollar sistemas expertos desde ejemplos almacenados en bases de datos en 1970. En un árbol de decisión, cada nodo del árbol es un atributo y de él parten o nacen tantas ramas como valores puede tener ese atributo. Las hojas o nodos terminales de estos árboles representan conjuntos ya clasificados y '*etiquetados*' con el nombre de una clase. Los árboles de decisión se utilizan para modelar funciones discretas con el objetivo de identificar el valor combinado de una serie de variables y a partir de esa combinación de valores, determinar la acción que ha de ser tomada. Por tanto, el punto de partida de un árbol de decisión es un problema descrito mediante un conjunto de atributos a partir del cual se pretende encontrar una decisión en base al valor combinado de dichos atributos. Para clasificar una instancia del juego de datos,

el árbol es recorrido de arriba abajo, hasta llegar a un nodo terminal. En un árbol de decisión podemos encontrar los siguientes tipos de elementos:

- Nodo interno, representa la evaluación del valor de alguno de los atributos.
- Nodo de probabilidad, representa un evento aleatorio en relación con la naturaleza del problema. Normalmente se representa como un nodo circular.
- Nodo hoja, representa el valor que el árbol de decisión determina como resultado de la evaluación.
- Rama, representa un posible camino en el árbol que se pueden tomar en base a la decisión (valor del atributo) tomada.

La utilidad de los árboles de decisión se adapta especialmente bien a determinados tipos de problemas, en concreto los problemas para los que son más apropiados son aquellos en los que:

- Los ejemplos del juego de datos pueden ser representados como pares de '*valor-atributo*'.
- La función objetivo toma valores discretos.
- Las hipótesis pueden ser expresadas mediante disyunciones.
- Existen atributos para los que sus valores son desconocidos.

Debido a la estructura de los árboles de datos y la capacidad de generar fácilmente reglas, los árboles de decisión son seguramente la técnica más utilizada para representar modelos. Estos árboles pueden ser usados tanto en tareas de exploración como de predicción. Un árbol de clasificación es la representación gráfica de una serie de reglas de decisión. A partir de un nodo raíz, que incluye todos los casos, el árbol se va ramificando en diferentes nodos "hijo" que contienen un subgrupo de casos. El criterio de ramificación (o partición) es seleccionado de manera óptima después de examinar todos los posibles valores de todas las variables predictivas disponibles. En los nodos terminales ("hojas" del árbol) se obtiene una agrupación de los casos de la manera más homogénea posible en cuanto al valor de la variable dependiente. Dependiendo de cómo se lleve a cabo la partición de los nodos, se distinguen diferentes tipos de árbol de clasificación: CART, C5.1, CHAID, etc. En este trabajo, las conclusiones finales vienen dadas por la aplicación de varios algoritmos de árbol de decisión, en concreto el algoritmo C5.1 y el algoritmo CHAID. El algoritmo C5.1 ha sido desarrollado por IBM en su plataforma SPSS Modeler. Los modelos C5.1 dividen la muestra en función del campo que ofrece la máxima ***ganancia de información*** (al contrario que el algoritmo CHAID, que se basa en detección automática de interacciones mediante la metodología Chi-cuadrado, en la que se comprueba que las frecuencias (número de valores) en cada categoría o grupo generados son estadísticamente diferentes entre sí [23]. Un aspecto esencial de los algoritmos basado en árboles de decisión es la elección del mejor criterio para la división de los datos. Una de las mejores opciones es realizar esta selección basándose en el concepto de entropía:

$$Entropia(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

23 https://www.ibm.com/support/knowledgecenter/en/SS4QC9/com.ibm.solutions.wa_an_overview.2.0.0.doc/chaid_classification_tree.html

donde p_i es la probabilidad de la clase i. Valores más bajos de la entropía hacen más predecible el caso. Dicho de otra forma: un valor más alto de la entropía produce más impredecibilidad, mientras que un valor más bajo de la entropía se interpreta como una menor impredecibilidad. La entropía caracteriza la impureza de una colección de datos, que se define como la entropía de Shannon [Bul18]. La ganancia de información es la reducción de la entropía causada por el particionado de los datos por un determinado atributo, y se define de la siguiente manera:

$$Ganacia(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Dónde:

- *S* es una colección de objetos
- *A* son los atributos de los objetos
- *V(A) es el* conjunto de valores que A puede tomar
- Sv es el subconjunto de S formado por aquellas instancias que en el atributo A toman el valor v.

Aunque la ganancia de información es una buena medida para la elección de atributos relevantes para el sistema, no es perfecta, ya que en algunos casos puede que atributos con gran ganancia de información, que identifiquen unívocamente la solución, pueden tomar infinidad de distintos valores y generan árboles de decisión muy exactos (y extensos) pero ineficaces en la generalización de la solución. Por tanto, la propuesta que se hace para sobrellevar esta debilidad de la ganancia de información es la utilización de la medida de la *proporción* o *ratio de ganancia de información*. La ***proporción de ganancia de información*** influye en la creación del árbol de decisión de manera que evita utilizar atributos que pueden tomar un gran número de valores distintos. El ratio o proporción de ganancia de información se representa mediante la siguiente fórmula matemática:

$$ratio(S, A) = \frac{ganancia(S, A)}{info_{part}(S|A)}$$

Donde

$$info_{part}(S|A) = -\sum_{(i=1)}^{n} \frac{|S_i|}{S} * \log_2 \frac{|S_i|}{S}$$

Una vez seleccionado el atributo que mayor proporción de ganancia de información suministra, las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las submuestras de nuevo, siempre bajo el mismo criterio. Por último, se vuelven a examinar las divisiones del nivel inferior, y se eliminan o podan las que no contribuyen significativamente con el valor del modelo. ***De esta forma, a partir de un conjunto de datos históricos, esta algoritmia es capaz de generar reglas sobre dichos datos, en función de un objetivo determinado, de forma automática.***

Ambos algoritmos, (C5.1, CHAID), pueden generar dos tipos de modelos: árboles de decisión y conjuntos de reglas. Un **árbol de decisión** es una descripción sencilla de las divisiones que se han encontrado en el

algoritmo. Los distintos nodos terminales (o "de hoja") describen un subconjunto de datos de entrenamiento, y cada uno de los casos incluidos en los datos de entrenamiento pertenece exactamente a un nodo terminal del árbol. ***En otras palabras, es posible realizar exactamente una predicción para cada registro de datos específico presente en un árbol de decisión***. En cambio, un **conjunto de reglas** es, como su propio nombre indica, un grupo de reglas que intenta realizar predicciones de registros individuales. Los conjuntos de reglas derivan de los árboles de decisión y, en cierto modo, representan una versión simplificada de la información que se incluye en estos árboles. Por lo general, los conjuntos de reglas pueden retener la mayor parte de la información significativa de un árbol de decisión completo, aunque utilizan un modelo menos complejo. Debido a las diferencias de funcionamiento de los conjuntos de reglas, sus propiedades son distintas de las de los árboles de decisión. La diferencia más importante consiste en que con un conjunto de reglas, puede aplicarse más de una regla a cualquier registro específico o no aplicar ninguna regla. Al aplicar varias reglas, cada una de ellas obtiene un "voto" ponderado basado en la confianza que se asocia a dicha regla. La predicción final se alcanza mediante la combinación de los votos ponderados de todas las reglas que se aplican al registro en cuestión. Si no se aplica ninguna regla, se asignará al registro una predicción predeterminada. Los árboles de decisión C5.1 son bastante más robustos cuando aparecen problemas como datos perdidos y un número elevado de campos de entrada. Por lo general no precisan de largos tiempos de entrenamiento para calcular las estimaciones. Además, los modelos C5.1 suelen ser más fáciles de comprender que algunos tipos de modelos, ya que la interpretación de las reglas derivadas del modelo es muy directa. C5.1 también ofrece el eficaz método del **aumento** para obtener una mayor precisión en tareas de clasificación. Las ventajas e inconvenientes de utilizar árboles de decisión son las siguientes:

Ventajas:

- Los árboles de decisión son fáciles de utilizar y eficientes.
- Las reglas que generan son fáciles de interpretar.
- Escalan mejor que otros tipos de técnicas.
- Tratan bien los datos con ruido.

Inconvenientes:

- No manejan de forma sencilla los atributos continuos.
- Tratan de dividir el dominio de los atributos en regiones rectangulares y no todos los problemas son de ese tipo.
- Tienen dificultad para trabajar con valores perdidos.
- Pueden tener problemas de sobreaprendizaje.
- No detectan correlaciones entre atributos.

3.2.3 Redes Neuronales

Considerando un escenario de aprendizaje supervisado, se dispone de un conjunto de datos etiquetados $\{(x^{(i)}, y^{(i)})\}$ donde $x^{(i)}$ representa las características e $y^{(i)}$ las etiquetas del ejemplo de entrenamiento i-ésimo. Las redes neuronales proporcionan una forma de representar una función compleja no lineal $h_w(x)$ de la variable de entrada x. La función $h_w(x)$ está parametrizada por una matriz de pesos W, la cual podemos amoldar a los datos. La figura 3.1 muestra una red neuronal simple que consiste en tres unidades o neuronas de entrada, x_{11} , x_{12} y x_{13} una neurona de salida tal que $y = h_w(x_{11}, x_{12}, x_{13})$

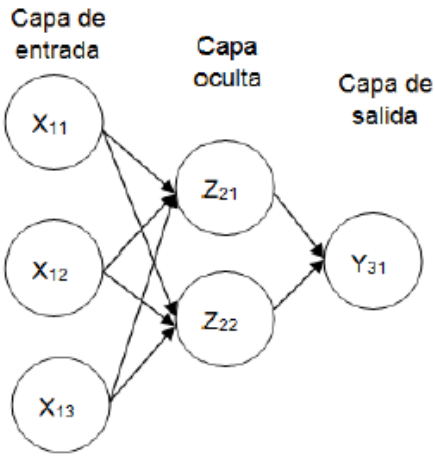


Figura 3.1 Estructura del Perceptrón multicapa

Una red neuronal se organiza en varias capas. En el ejemplo de la figura 3.1 se compone de tres capas: la capa de entrada, la capa oculta, y la capa de salida. Como se observa en el diagrama, las neuronas se enlazan entre capas adyacentes mediante un conjunto de conexiones. Aunque en la figura se muestra una red completamente conectada, donde cada neurona está conectada a todas las neuronas de la capa anterior, esta no es una condición necesaria en la estructura de una red neuronal. El patrón de conectividad de una red se denomina arquitectura de la red. Sin contar con las neuronas de la capa de entrada, cada neurona x_i de la red toma como entrada los valores de las neuronas de la capa precedente que alimentan a x_i . Como ejemplo concreto, las entradas a la neurona z_{21} en la red neuronal de la figura 3.1 son x_{11} , x_{12} y x_{13} y para y_{31} las entradas son z_{21} y z_{22} . Una neurona calcula primero una combinación lineal ponderada de sus entradas, tal que:

$$a_j = \sum_{i=1}^{n} w_{ji} x_i + b_j$$

Donde w_{ji} es un parámetro que describe la interacción entre z_j y la neurona de entrada x_i. El término b_j es un sesgo asociado a la neurona z_j. Después a a_j se le aplica una función de activación no lineal. Algunas funciones de activación comunes son la sigmoidea y las funciones de tangente hiperbólicas. En particular, la activación o el valor de la neurona z_j se define como:

$$z_j = h(a_j) = h\left(\sum_{i=1}^{n} w_{ji}x_i + b_j\right)$$

Donde h en este caso es la función de activación sigmoidea o lineal. Como la activación de cada neurona sólo depende de los valores de las neuronas de las capas anteriores, se calculan las activaciones a partir de la primera capa oculta (que sólo depende de los valores de entrada) y aprendiendo así a través de la red. Este proceso en el que la información se propaga a través de la red se denomina etapa forward-propagation. Al final de la etapa de forward-propagation, se obtiene un conjunto de salidas $y = h_w(x)$. Cuando se está realizando una clasificación binaria, la salida y puede verse como resultado de la clasificación de la entrada x. Dado el conjunto de entrenamiento etiquetado $\{(x^{(i)}, y^{(i)})\}$, el objetivo es aprender los parámetros W a fin de minimizar una función objetivo o pérdida. Esta minimización se puede emplear Stochastic Gradient Descent (SGD), back-propagation u otros métodos. En este trabajo se ha realizado un aprendizaje denominado “Backpropagation”, consistente en minimizar una función del error entre la salida proporcionada por la red y el valor deseado de la variable objetivo, por regla general, proporcional al error cuadrático medio.

$$ep^2 = \frac{1}{2}\sum_{k=1}^{l}(\delta_k^2)$$

donde

ep^2: Error cuadrático medio para cada patrón de entrada p.

δ_k: Error en la neurona k de la capa de salida con l neuronas.

Para ello, se busca un extremo relativo de la función de los pesos, esto es, un punto donde todas las derivadas parciales de la función del error respecto a los pesos se anulan:

$$\frac{\partial E}{\partial w_{ijl}^{ahl}} = 0; \quad con\ l \leq i \leq N$$

Las redes neuronales artificiales (RNA) tienen muchas ventajas dado están basadas en la estructura del sistema nervioso, principalmente el cerebro. Las ventajas son las siguientes:

- Aprendizaje: Las RNA tienen la habilidad de aprender mediante una etapa que se llama etapa de aprendizaje. Esta consiste en proporcionar a la RNA datos como entrada a su vez que se le indica cuál es la salida (respuesta) esperada.

- Auto organización: Una RNA crea su propia representación de la información en su interior, descargando al usuario de esto.
- Tolerancia a fallos: Debido a que una RNA almacena la información de forma redundante, ésta puede seguir respondiendo de manera aceptable aun si se daña parcialmente.
- Flexibilidad: Una RNA puede manejar cambios no importantes en la información de entrada, como señales con ruido u otros cambios en la entrada (ej. si la información de entrada es la imagen de un objeto, la respuesta correspondiente no sufre cambios si la imagen cambia un poco su brillo o el objeto cambia ligeramente)
- Tiempo real: La estructura de una RNA es paralela, por lo cual si esto es implementado con computadoras o en dispositivos electrónicos especiales, se pueden obtener respuestas en tiempo real.

Y las desventajas son las siguientes:

- Complejidad de aprendizaje para grandes tareas, cuanto más cosas se necesita que aprenda la red, más complicado será enseñarle
- No permite interpretar lo que se ha aprendido, la red por si sola proporciona una salida, un número, que no puede ser interpretado por ella misma sino que se requiere de la intervención del programador y de la aplicación en si para encontrarle un significado la salida proporcionada.
- Elevada cantidad de datos para el entrenamiento, cuanto más flexible se requiere que sea la red neuronal, más información tendrá que enseñarle para que realice de forma adecuada la identificación
- Los sobreentrenamientos deben ser controlados en la fase de aprendizaje.

3.2.4 Máquinas de Vectores de soporte (SVM)

La máquina de vectores soporte es un método de aprendizaje basado en muestras para la realización de clasificadores y regresores. Este algoritmo generaliza el método “generalized portrait”, propuesto por Vapnik y Lerner (Vapnik y Lerner, 1963) para la resolución de problemas de clasificación linealmente separables mediante lo que se denomina hiperplano óptimo de separación (optimal hyperplane decision rule, OHDR). La formulación de la SVM parte del concepto clásico de hiperplano óptimo de separación, cuyo vector director queda expresado en función de las muestras de entrenamiento. Así mismo, incorpora una serie de aspectos derivados de la teoría del aprendizaje estadístico que confieren a la máquina de vectores soporte una capacidad de generalización superior a la de otros métodos de aprendizaje.

Consideremos el problema de clasificación de un punto cuyas características están dadas por el vector x tal que $x = \left(x_1, \dots, x_p\right)^T$ y este pertenece a una de dos clases posibles. Supongamos que tenemos las funciones $f_1(x)$ y $f_2(x)$ que definen las clases 1 y 2 y nosotros clasificamos al punto x dentro de la clase 1 si

$$f_1(x) < 0, f_2(x) < 0,$$

o clasificamos al punto x dentro de la clase 2 si

$$f_1(x) < 0, f_2(x) > 0,$$

A estas funciones las llamamos funciones de decisión. Al proceso de encontrar las funciones de decisión a partir de pares de entrada-salida es llamado entrenamiento. Los métodos convencionales de entrenamiento determinan las funciones de decisión de tal forma que cada par entrada-salida sea correctamente clasificado dentro de la clase a la que pertenece. La Figura 3.2 muestra un ejemplo. Asumiendo que los cuadros pertenecen a la clase 1 y los círculos pertenecen a la clase 2, resulta claro que los datos de entrenamiento no se intersectan en ningún momento y es posible trazar una línea separando los datos de manera perfecta.

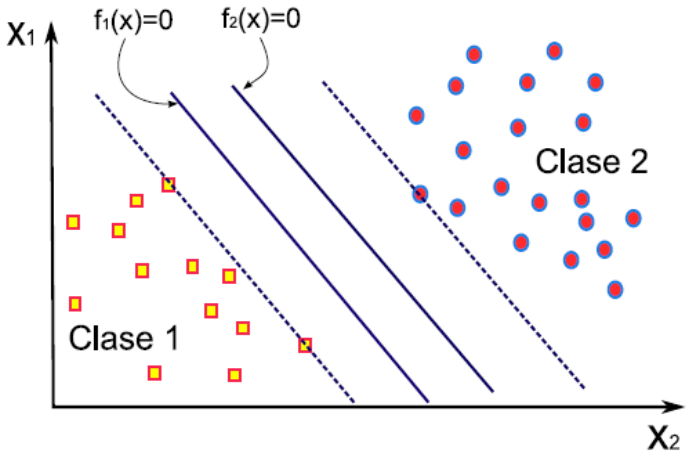


Figura 3.2 Estructura del SVM

Sin embargo, ya sea que la función de decisión $f_1(x)$ o la función $f_2(x)$ se muevan hacia la línea punteada de su propio lado, el conjunto de datos de entrenamiento aún sigue siendo correctamente clasificado, dándonos la certeza de que es posible encontrar un conjunto infinito de hiperplanos que correctamente clasifiquen los datos de entrenamiento. Sin embargo, es claro que la precisión de clasificación al generalizar será directamente afectada por la posición de las funciones de decisión. Las SVM a diferencia de otros métodos de clasificación consideran esta desventaja y encuentra la función de decisión de tal forma que la distancia entre los datos de entrenamiento es maximizada. Esta función de decisión es llamada función de decisión óptima o hiperplano de decisión óptima [Cri00]. El objetivo de SVM es encontrar un función lineal $f(x) = (w, x) + b$, entre todos los hiperplanos canónicos que clasifican correctamente los datos, aquel con menor norma, o, equivalentemente, con mínimo $\|w\|^2$. Es interesante notar que la minimización de $\|w\|^2$es equivalente a encontrar el hiperplano separador para el cual la distancia entre dos envolturas convexas (las dos clases del conjunto de datos de entrenamiento, asumiendo que son linealmente separables), medida a lo largo de una línea perpendicular al hiperplano, es maximizada. Esta distancia se conoce como margen. El problema de maximización del margen se formula de la siguiente manera:

$$\underset{w,b}{Min} \quad \frac{1}{2}\|w\|^2$$

sujeto a

$$y_i \cdot (w^T \cdot x_i + b) \geq 1 \qquad i = 1, \ldots, m,$$

A partir de esta formulación se construye el dual mediante la técnica de los multiplicadores de Lagrange. La formulación dual se corresponde con la siguiente fórmula:

$$Max_\alpha \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,s=1}^{m} \alpha_i \alpha_s y_i \, y_s x_i x_s$$

sujeto a

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \;\; i = 1, \ldots, m$$

donde α_i representan los multiplicadores de Lagrange asociados a las restricciones de $\underset{w,b}{Min} \quad \frac{1}{2}\|w\|^2$.

Los multiplicadores que cumplen con $\alpha_i > 0$ son llamados “Support Vectors”, ya que son los únicos que participan en la construcción del hiperplano de clasificación.

Las grandes ventajas que tiene SVM son:

- Una excelente capacidad de generalización, debido a la minimización del riesgo estructurado.
- Existen pocos parámetros a ajustar; el modelo solo depende de los datos con mayor información.
- La estimación de los parámetros se realiza a través de la optimización de una función de costo convexa, lo cual evita la existencia de un mínimo local.
- El modelo final puede ser escrito como una combinación de un número muy pequeño de vectores de entrada, llamados vectores de soporte.

Si bien las SVM han demostrado un gran potencial en tareas de clasificación principalmente por su buena capacidad de generalización, debido a que están fundamentadas en la teoría de aprendizaje estadístico, posee varias desventajas desde el punto de vista práctico
En particular, las desventajas de las SVM son las siguientes:

- La predicción del clasificador no tiene significado probabilístico.
- SVM sufre de otros problemas como la selección de la mejor función kernel y SVM presenta problemas computacionales al aplicarse sobre conjuntos grandes de datos de entrenamientos.

3.2.5 Redes Bayesianas

El teorema de Bayes fue enunciado por Thomas Bayes en 1763. Este teorema expresa la probabilidad condicional de un evento aleatorio. Se define como probabilidad condicional, la probabilidad de que ocurra un suceso A sabiendo que también sucede otro evento B. La manera formal de escribirlo es $P(A|B)$ y la definición es:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Por el teorema de la multiplicación sabemos que si los sucesos A y B son independientes, entonces $P(A|B) = P(A)P(B)$, por tanto $P(A|B) = P(A)$ e igualmente ocurre con $P(B|A) = P(B)$. Un error muy común es asumir que $P(A|B)$ y $P(B|A)$ son casi iguales pero la verdadera relación entre estos dos términos se expresa en el teorema de Bayes con la siguiente ecuación:

$$P(A|B) = P(B|A) \cdot \frac{P(A)}{P(B)}$$

Este enunciado del teorema de Bayes es muy sencillo y sólo para dos sucesos, el teorema tal y como lo enunció Thomas Bayes es:

Sea $\{A_1, A_2, \dots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionadas $P(A|B_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = P(B|A) \cdot \frac{P(A_i)}{P(B)}$$

donde:

$P(A_i)$ son las probabilidades a priori.
$P(A|B_i)$ es la probabilidad de B en la hipótesis Ai.
$P(A_i|B)$ son las probabilidades a posteriori.

Según el teorema de las probabilidades totales, la probabilidad de B se puede expresar como:

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

y por tanto otra forma de enunciar el teorema de Bayes es:

$$P(A_i\,|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)}$$

Las redes bayesianas son una alternativa a la hora de implementar un sistema experto probabilístico ya que poseen ciertas cualidades que otras técnicas no permiten, como por ejemplo, admiten el aprendizaje sobre relaciones de dependencia y causalidad, permiten la combinación de conocimiento con datos, evitan el sobre-ajuste continuo de los datos y pueden manejar bases de datos incompletas. Formalmente, una red bayesiana es un grafo acíclico dirigido (DAG) en el cual cada nodo representa una variable y cada arco una dependencia probabilística que especifica la probabilidad condicional de cada variable dados sus padres. La red bayesiana se puede ver como un conjunto formado por tres partes:

- un conjunto de variables del dominio que se quiere representar.
- un grafo acíclico dirigido (DAG) cuyos nodos están etiquetados con los elementos del anterior conjunto.
- una distribución conjunto sobre las variables.

Las ventajas e inconvenientes de utilizar Redes Bayesianas son las siguientes:

Ventajas:

• Es fácil de implementar

• Obtiene buenos resultados en gran parte de los casos

Desventajas:

• Asumir que las variables tienen independencia condicional respecto a la clase lleva a una falta de precisión.

3.2.6 El problema de la validación: La Validación Cruzada

Existen diversas técnicas para validar los métodos de clasificación, como son:

- La comparación de los resultados obtenidos en un modelado con los obtenidos a su vez mediante modelos físicos teóricos o con simulaciones,
- La utilización de nuevos conjuntos de datos conocidos para comparar con los obtenidos
- El uso de técnicas de validación cruzada.

Las técnicas que hemos utilizado en este trabajo pertenecen a este último grupo. Básicamente, existen dos métodos de validación cruzada, ***hold-out*** y ***k-fold***. El método hold-out es el más sencillo de los distintos métodos de validación cruzada. Este separa el conjunto de datos disponibles en dos subconjuntos, uno utilizado para entrenar el modelo y otro para realizar el test de validación [Arl10]. De esta manera, se crea un modelo únicamente con los datos de entrenamiento. Con el modelo creado se generan datos de salida que se comparan con el conjunto de datos reservados para realizar la validación (que no han sido utilizados en el entrenamiento, por lo que no han sido utilizados para generar el modelo [Haw03]. Los estadísticos obtenidos con los datos del subconjunto de validación son los que nos dan la validez del método empleado en términos

de error. Una aplicación alternativa de este método consiste en repetir el proceso hold-out, tomando distintos conjuntos de datos de entrenamiento (aleatorios) un determinado número de veces, de manera que se calculan los estadísticos de la regresión a partir de la media de los valores en cada una de las repeticiones. El otro método utilizado, k-fold, está basado en el método anterior, pero con mayor utilidad cuando el conjunto de datos es pequeño [Yan14]. En este caso, el total de los datos se dividen en k subconjuntos, de manera que aplicamos el método hold-out k veces, utilizando cada vez un subconjunto distinto para validar el modelo entrenado con los otros k-1 subconjuntos [Jun15]. El error medio obtenido de los k análisis realizados nos proporciona una estimación del error de generalización cometido por el método.

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i$$

Si comparamos los dos métodos, el método k-fold tiene la ventaja de que todos los datos son utilizados para entrenar y validar, por lo que se obtienen resultados más representativos a priori. Mientras que, para el método hold-out, se realiza el proceso n veces de manera aleatoria, lo que no garantiza que los casos de entrenamiento y validación no se repitan. Concretamente, en los trabajos de esta Tesis hemos utilizado el método k-fold de validación cruzada, coincidiendo con las recomendaciones en la literatura [Per15]. El resultado de los métodos de validación cruzada usualmente se muestra en las llamadas "matrices de confusión". Una matriz de confusión nos permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador, que representa el resultado de la prueba de un modelo de predicción. Cada columna de la matriz representa las instancias de una clase predicha, mientras que cada fila representa las instancias de una clase real. Para evaluar correctamente los resultados y los distintos métodos usaremos los conceptos de *exactitud, sensibilidad ("recall") y precisión*. La ***precisión*** es una medida de la exactitud de los elementos que se sugieren como entidades o relaciones, y se mide típicamente como la proporción de verdaderos positivos (elementos sugeridos correctamente) sobre todos los elementos sugeridos. "***Recall***" o sensibilidad designa la proporción a la que se reconocen las entidades o relaciones y se mide generalmente como la relación de verdaderos positivos sobre todos los elementos que deben ser reconocidos. La ***exactitud*** mide la proporción de aquellos elementos bien clasificados (verdaderos positivos y verdaderos negativos), con respecto a todos los elementos seleccionados para la validación.

- tp: Verdadero Positivo (éxito).
- tn: Verdadero Negativo (rechazos correctos).
- fp: Falso Positivo (Falsa alarma, error tipo I).
- fn: Falso Negativo (Ausencia de alarma, error tipo II).
- Precisión = $\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}$
- Recall (Sensibilidad) = $\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$
- Exactitud (Accuracy) = $\frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{tn}+\mathrm{fp}+\mathrm{fn}}$

3.3 Detección de Anomalías

3.3.1 Local Outlier Factor

El algoritmo LOF (Local Outlier Factor) es un algoritmo no supervisado que trata de buscar datos atípicos de forma local y que proporciona un valor de cuánto de atípico es un punto basado en la diferencia de densidad que tiene con respecto a sus vecinos más cercanos. Debemos entender que intuitivamente un punto presenta una densidad alta si sus vecinos se encuentran muy próximos a él, mientras que diremos que tiene una densidad baja cuando sus vecinos estén muy próximos entre sí, pero no se encuentren cerca de él. En la figura 3.3 podemos observar como el punto A tiene una densidad mucho menor que los demás puntos, por lo que el algoritmo debería devolver un valor elevado de anomalía.

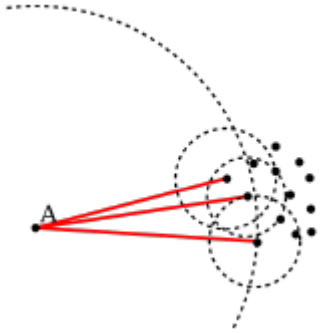


Figura 3.3 Estructura LOF

Para ejecutar dicho algoritmo, debemos definir antes algunos términos que se emplearán. Primeramente, definiremos la distancia *k* de un punto $p \in D$, denotada por k-distancia(p). Para cualquier entero positivo *k*, la ***k*-distancia** de un punto p, se define como la distancia $d(p,o)$ entre p y un punto $o \in D$ tal que:

(i) para al menos *k* puntos $o' \in D \setminus \{p\}$ se tiene que $d(p, o') \leq d(p,o)$ y

(ii) para al menos *k-1* puntos $o' \in D \setminus \{p\}$ se tiene que $d(p, o') < d(p,o)$

Otro concepto que es necesario definir es el de vecindad de distancia *k* de un punto p . Dada la **k-distancia** de p, la vecindad de distancia *k* de p contiene a todo punto cuya distancia a *p* sea menor o igual que la ***k-distancia***, es decir, $N_{\text{k-distancia}}(p) = \{ q \in D \setminus \{p\} \mid d(p,q) \leq \text{k-distancia}\}$. Por simplicidad y cuando no haya lugar a confusión emplearemos $N_k(p)$ en lugar de $N_{\text{k-distancia}(p)}$.

Es necesario también definir la "reach-distance" o "***distancia de accesibilidad***" de un punto p con respecto a otro *o*, de la siguiente forma:

$\text{reach-dist}_k(p, o) = \max\{\text{k-distancia}(o), d(p, o)\}$, siendo *k* un número natural.

Ahora tenemos definidos estos conceptos para cualquier *k*, sin embargo a la hora de ejecutar el algoritmo debemos fijar este valor, denotado por MinPts (mínimo de puntos), para mantener fija la condición de densidad. Definimos entonces la *local reach density* de un punto p, denotada por *lrd(* p*)*, como:

$$lrd_{MinPts}(p) = \left(\frac{\sum_{o \in N_{MinPts}(p)} reach - dist_{MinPts}(p,o)}{|N_{MinPts}(p)|}\right)^{-1}$$

Finalmente, definimos el local outlier factor de un punto p como:

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

Atendiendo a la expresión de la fórmula, observamos que los valores altos (los atípicos) se obtendrán cuando la lrd(p) sea pequeña, mientras que las de sus vecinos más próximos (MinPts-vecinos) sean grandes.

3.3.2. Análisis individual de atípicos

Con el objetivo de reducir los falsos positivos sin que el número de anomalías encontradas se vea afectado, presentaremos en esta sección una técnica que tratará de analizar cada punto individualmente para determinar si es un verdadero atípico o un falso positivo. Comenzaremos ejecutando el LOF, para posteriormente realizar un estudio de cada dato marcado como atípico. Una vez obtenido este conjunto formado por los posibles datos atípicos utilizaremos la siguiente regla: si un dato se encuentra a una distancia superior a 3 desviaciones típicas de la media de los q vecinos más cercanos etiquetados como datos normales en un porcentaje alto de dimensiones, comparado con el número de dimensiones para las cuales sus valores se encuentran a una distancia de 2 desviaciones típicas, entonces esta es una verdadera anomalía (tiene muchas dimensiones con valores extremos), mientras que en el caso contrario nos encontraríamos con un falso positivo (un valor alejado de la media, pero no extremo en un elevado número de dimensiones). El valor crítico que tomaremos para diferenciar los verdaderos atípicos de los falsos positivos será 4, es decir, si el porcentaje de dimensiones para las cuales el dato se encuentra a más de 3 desviaciones típicas multiplicado por 4 es menor que el porcentaje de dimensiones que se alejan 2 desviaciones típicas, es un falso positivo, y en caso contrario es un verdadero atípico. Es importante destacar que para la aplicación de esta técnica es necesario contar con un conjunto de datos etiquetados como no atípicos, ya que éstos formarán la matriz de la cual se extraerán las medias y desviaciones típicas por columnas imprescindibles para el cálculo del valor crítico.

3.4 Series Temporales. Modelo ARIMA.

La detección de patrones de repetición en secuencias temporales diferentes, la detección de dichas secuencias, su ubicación temporal y la predicción del objetivo en función de los patrones en secuencia encontrados se denomina análisis de Series Temporales. Una serie temporal se define como una colección de

observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equi-espaciados. El estudio descriptivo de series temporales se basa en la idea de descomponer la variación de una serie en varias componentes básicas. Este enfoque no siempre resulta ser el más adecuado, pero es interesante cuando en la serie se observa cierta tendencia o cierta periodicidad. Las componentes o fuentes de variación que se consideran habitualmente son las siguientes:

1. ***Tendencia***: Se puede definir como un cambio a largo plazo que se produce en relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.
2. ***Efecto Estacional***: Muchas series temporales presentan cierta periodicidad o dicho de otro modo, variación de cierto periodo (anual, mensual ...). Por ejemplo, el paro laboral aumenta en general en invierno y disminuye en verano. Estos tipos de efectos son fáciles de entender y se pueden medir explícitamente o incluso se pueden eliminar del conjunto de los datos (desestacionalizar la serie original).
3. ***Componente Aleatoria***: Una vez identificados los componentes anteriores y después de haberlos eliminado, persisten unos valores que son aleatorios. Se pretende estudiar qué tipo de comportamiento aleatorio presentan estos residuos, utilizando algún tipo de modelo probabilístico que los describa.

De las tres componentes reseñadas, las dos primeras son componentes determinísticas, mientras que la última es aleatoria. Así, se puede denotar que Xt = Tt + Et + It donde Tt es la tendencia, Et es la componente estacional, que constituyen la señal o parte determinística, e It es el ruido o parte aleatoria. En 1970, Box y Jenkins [Box70] desarrollaron un cuerpo metodológico destinado a identificar, estimar y diagnosticar modelos dinámicos de series temporales en los que la variable tiempo juega un papel fundamental. Una parte importante de esta metodología está pensada para liberar al investigador de la tarea de especificación de los modelos dejando que los propios datos temporales de la variable a estudiar nos indiquen las características de la estructura probabilística subyacente. El método desarrollado se denomina ARIMA (Modelos Autorregresivos Integrados de Medias Móviles).

- ***Modelos Autorregresivos***

Definimos un modelo como autorregresivo si la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores añadiendo, como en los modelos estructurales, un término de error. En el caso de procesos estacionarios con distribución normal, la teoría estadística de los procesos estocásticos dice que, bajo determinadas condiciones previas, toda Yt puede expresarse como una combinación lineal de sus valores pasados (parte sistemática) más un término de error (innovación). Los modelos autorregresivos se abrevian con la palabra AR tras la que se indica el orden del modelo: AR(1), AR(2),....etc. El orden del modelo expresa el número de observaciones retrasadas de la serie temporal analizada que intervienen en la ecuación. Así, por ejemplo, un modelo AR (1) tendría la siguiente expresión:

$$Y_t = \emptyset_0 + \emptyset_1 Y_{t-1} + a_t$$

El término de error de los modelos de este tipo se denomina generalmente ruido blanco cuando cumple las tres hipótesis siguientes: media nula, varianza constante, covarianza nula entre errores correspondientes a

observaciones diferentes. La expresión genérica de un modelo autorregresivo, no ya de un AR (1) sino de un AR (p) sería la siguiente:

$$Y_t = \emptyset_0 + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \ldots + \emptyset_p Y_{t-p} + a_t$$

- ***Modelo de Medias Móviles***

Un modelo de los denominados de medias móviles es aquel que explica el valor de una determinada variable en un período t en función de un término independiente y una sucesión de errores correspondientes a períodos precedentes, ponderados convenientemente. Estos modelos se denotan normalmente con las siglas MA, seguidos, como en el caso de los modelos autorregresivos, de orden entre paréntesis. Así, un modelo con q términos de error MA(q) respondería a la siguiente expresión:

$$Y_t = \mu + a_t + \emptyset_1 a_{t-1} + \emptyset_2 a_{t-2} + \ldots + \emptyset_q Y_{t-q}$$

- ***Modelos ARIMA (p,q)***

La extensión de los modelos AR(p) y Ma (q) es un tipo de modelos que incluyen tanto términos autorregresivos como de medias móviles y se definen como ARIMA(p, 0, q). Se representan por la ecuación:

$$Y_t = \emptyset_0 + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \ldots + \emptyset_p Y_{t-p} + a_t - \emptyset_1 a_{t-1} - \emptyset_2 a_{t-2} - \ldots - \emptyset_q a_{t-q}$$

El proceso ARIMA(p, q) es estacionario si lo es su componente autorregresiva, y es invertible si lo es su componente de medias móviles.

- ***Proceso Autorregresivo Integrado y de Media Móvil ARIMA(p, d, q)***

Los modelos de series de tiempo anteriores se basan en el supuesto de estacionalidad, esto es, la media y la varianza para una serie de tiempo son constantes en el tiempo y la covarianza es invariante en el tiempo. Pero se sabe que muchas series de tiempo no son estacionarias, porque pueden ir cambiando de nivel en el tiempo o sencillamente la varianza no es constante en el tiempo. A este tipo de proceso se les considera procesos integrados. Por consiguiente, se debe diferenciar una serie de tiempo “d” veces para hacerla estacionaria y luego aplicarla a esta serie diferenciada un modelo ARIMA(p,q), se dice que la serie original es ARIMA (p,d,q), es decir, una serie de tiempo autorregresiva integrada de media móvil, donde “p” denota el número de términos autorregresivos, “d” el número de veces que la serie debe ser diferenciada para hacerla estacionaria y “q” el número de términos de la media móvil invertible. Si además, la serie es estacional de periodo “s”, se deberán determinar los parámetros (p,d,q) para definir la parte no estacional, y (P,D,Q), para la parte estacional, y el proceso se describe como un ARIMA(p,d,q)$(P,D,Q)_s$.

La construcción de los modelos ARIMA (p,d,q)(P,D,Q) se lleva de manera iterativa mediante un proceso en el que se puede distinguir cuatro etapas:

- ***Identificación***. Utilizando los datos ordenados cronológicamente se intenta sugerir un modelo ARIMA(p,d,q) eficiente. El objetivo es determinar los valores p, d, q que sean apropiados para reproducir la serie temporal. En esta etapa es posible identificar más de un modelo candidato que pueda describir la serie.
- ***Estimación***. Considerando el modelo apropiado para la serie de tiempo se realiza inferencia sobre los parámetros.
- ***Validación***. Se realizan contraste de diagnóstico para validar si el modelo seleccionado se ajusta a los datos, y si no es así, escoger el próximo modelo candidato y repetir los pasos anteriores.
- ***Predicción***. Una vez seleccionado el mejor modelo candidato ARIMA(p, d, q) se pueden hacer pronósticos en términos probabilísticos de los valores futuros.

Capítulo 4

Arquitectura del Software

En este capítulo se describe la arquitectura implantada en el desarrollo de la tesis. En la sección 4.1 se introducen los requerimientos que han llevado a la definición de la misma. En el 4.2 se explicitan los distintos componentes que conforman la arquitectura. En la sección 4.3 se describe el sistema local de adquisición y tratamiento de la información sensórica. En la sección 4.4. se detallan los módulos de agregación de datos y procesos de tratamiento más avanzados, centrándonos particularmente en el tratamiento de historiales clínicos sobre lenguaje natural. Además, se expone la arquitectura del Sistema Experto y del Sistema de Envío de Notificaciones.

4.1 Introducción y motivación

El sistema propuesto en este trabajo está diseñado para cumplir con cuatro requisitos principales (ver figura 1.1, en el Capítulo 1):

- En primer lugar, la extracción, transformación y carga de la información de los sensores se llevará a cabo de manera sencilla: el sensor se conecta a la red y sus datos sin procesar se integran automáticamente en la plataforma, en un formato “bruto”, que posteriormente se agregan, transforman y cargan en la Ontología del Sistema de una forma automática, en base a una serie de Reglas de Proceso.
- En segundo lugar, la plataforma debe poder registrar conocimiento de expertos en el contexto de la Teleasistencia (técnicos en Asistencia y personal médico), de una forma sencilla y manual (en forma de Reglas Heurísticas), y poder aplicar dicho conocimiento al conjunto de datos del sistema, en tiempo real.
- En tercer lugar, la plataforma debe ser capaz de modelar de forma automática los hábitos de los usuarios, y sus variaciones, a partir de los datos brutos, agregados y transformados, con el fin de monitorizar su comportamiento para encontrar desviaciones de sus tareas diarias (por ejemplo, cuando se despiertan, hábitos de sueño, paseos diarios, etc.).
- Finalmente, con toda esta información generada en históricos, y recogida en tiempo real, junto con el conocimiento extractado (Reglas Heurísticas y Modelos Automáticos), el sistema de poder proveer un resumen detallado a los agentes sanitarios o a la familia sobre su estado, evolución y detectar situaciones de riesgo para los usuarios.

Además, el diseño del sistema busca satisfacer los siguientes requisitos:

- Interoperabilidad: el sistema debe tener la capacidad de intercambiar procesos y/o datos con el resto de componentes/entidades del sistema y utilizar la información intercambiada.

- Escalabilidad: el sistema debe tener la capacidad de reaccionar y adaptarse en todo momento sin perder calidad, manejar el crecimiento continuo de los datos de manera fluida, por tanto estar capacitado para hacerse más grande sin que se degrade la calidad en los servicios ofrecidos.
- Tolerancia de Fallos: el sistema debe tener la capacidad de poder acceder a información aun en caso de producirse algún fallo. Como el problema puede ser debido a diversos motivos, (fallo de hardware, fallo de comunicación, …), la tolerancia de fallos requiere que el sistema guarde la información de forma redundante en más de un componente de hardware.
- Seguridad: el sistema trata de ser seguro y confiable en todo momento. Para ello empleamos una serie de estándares, protocolos, procedimientos, métodos, reglas, y/o herramientas para minimizar posibles riesgos en la infraestructura y la información.
- Privacidad: el sistema debe controlar en todo momento quién tiene acceso a la información que posee cada uno de los usuarios de la plataforma.
- Mantenibilidad, auditabilidad, flexibilidad e interacción con otros sistemas de información.

Para cumplir con estos requisitos, los componentes del sistema se instalan tanto en el hogar de la persona usuaria, como en un conjunto de servicios alojados en un servidor cloud. La figura 4.1 presenta esta estructura conceptual.

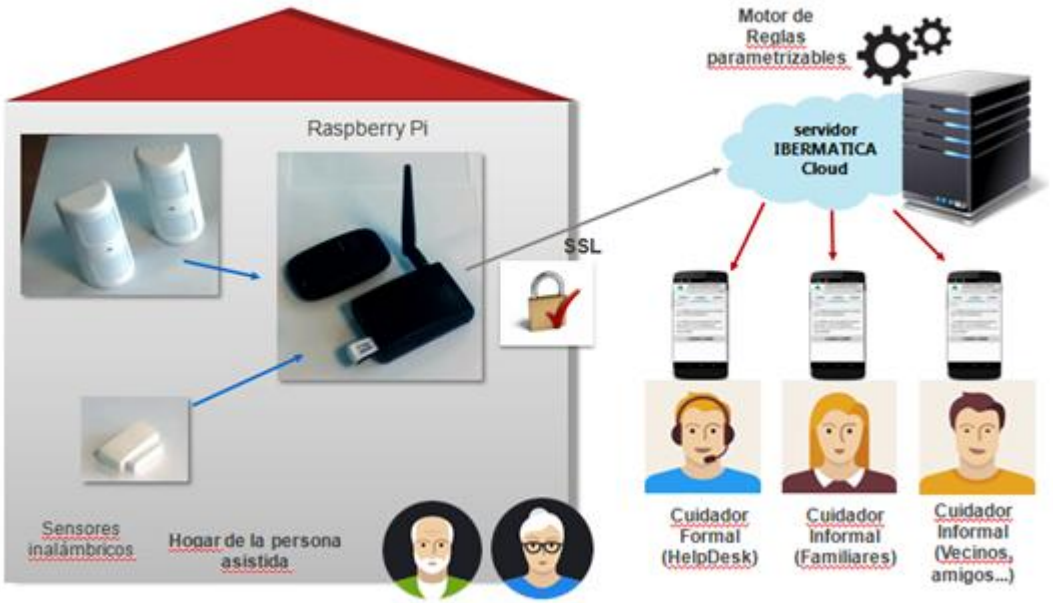


Figura 4.1 Estructura de comunicaciones del sistema propuesto e instalado

4.2 Componentes de la arquitectura

La arquitectura del sistema está configurada por dos grandes bloques de sistemas: un primer sistema, al que denominamos ***Sistema Local***, que gestiona toda la información local recogida en el domicilio, y un segundo

sistema, al que denominamos, ***Sistema en "Cloud"***, que contiene toda la lógica "inteligente" a la hora de analizar las posibles situaciones de riesgo que puedan ocurrir en el domicilio.

1. El ***Sistema Local***, instalado en el domicilio, a su vez, se compone de los siguientes componentes:
    - Un ***Componente de Captación de Información*** que se conforma como un conjunto de diversos sensores de bajo coste y poca necesidad de mantenimiento, que serán la base de la identificación de patrones y de situaciones de peligro.
    - Un ***Componente de Gestión Local*** de información, que, utilizando un micro computador de bajo coste, Raspberry, con una base de datos local, ágil y no relacional, permite que la información se almacene antes de su envío al "cloud", y se mantenga en un "buffer" lógico ante posibles caídas de comunicación o incidencias del sistema.
2. El ***Sistema en "Cloud"***, a su vez, está compuesto por los siguientes componentes::
    - ***Sistemas de soporte*** que permiten ofrecer, dentro del entorno distribuido y basado en servicios, herramientas de almacenamiento, seguridad, coordinación y almacenamiento de información, siendo el módulo central del sistema que interconecta los hogares de los usuarios con el resto de servicios ofrecidos en cloud.
    - ***Sistema de Agregación de Datos***: Este componente permite integrar, transformar y almacenar la información obtenida, tanto a partir del Componente de Gestión Local, como de fuentes externas al domicilio, pero relevantes para su perfilado, como lo son la información clínica y la meteorológica. La primera se obtiene a partir de historiales clínicos, y permiten realizar recomendaciones terapéuticas, seguimientos y chequeos de comportamientos relacionados con ciertas patologías o tratamientos. Además, es el sistema que integra también datos meteorológicos a través de servicios web públicos.
    - ***Sistema Experto***. Este módulo es el "cerebro" del sistema. A su vez, se compone de dos módulos::
        - ***Módulo de Reglas Heurísticas***: Este módulo permite integrar conocimiento de los técnicos, tanto de expertos en procedimientos de Teleasistencia (en concreto, personal de Matia), como de expertos clínicos en forma de reglas introducidas de forma manual.
        - ***Módulo de Detección Automática de Patrones***: Este módulo permite analizar y modelizar los patrones de comportamiento personalizados en cada domicilio, y de esta forma, generar reglas de forma automática que complementan el conocimiento de los expertos.
    - ***Sistema de Envío de Notificaciones***: Este módulo implementa un flujo de trabajo basado en el uso de "Windows Workflow Foundation" que permite asegurar que desde que se lanza una alerta desde un hogar, ésta es atendida por las personas indicadas y dentro de los parámetros de tiempo prefijados. Este sistema asegurará que la persona objeto del evento es atendida en el menor tiempo posible. Para ello, se encargará de seguir el flujo de trabajo con los eventos a generar, así como las notificaciones y las contestaciones de las personas indicadas, para lo que utilizará un servidor de notificación. Para poder interactuar con las personas usuarias se han creado tanto una APP Android, en las que recibir estas notificaciones, como una aplicación

web para la persona profesional y/o responsable en el que poder y trabajar con las distintas notificaciones.

La figura 4.2 muestra cómo los componentes anteriores están organizados de una manera esquemática.

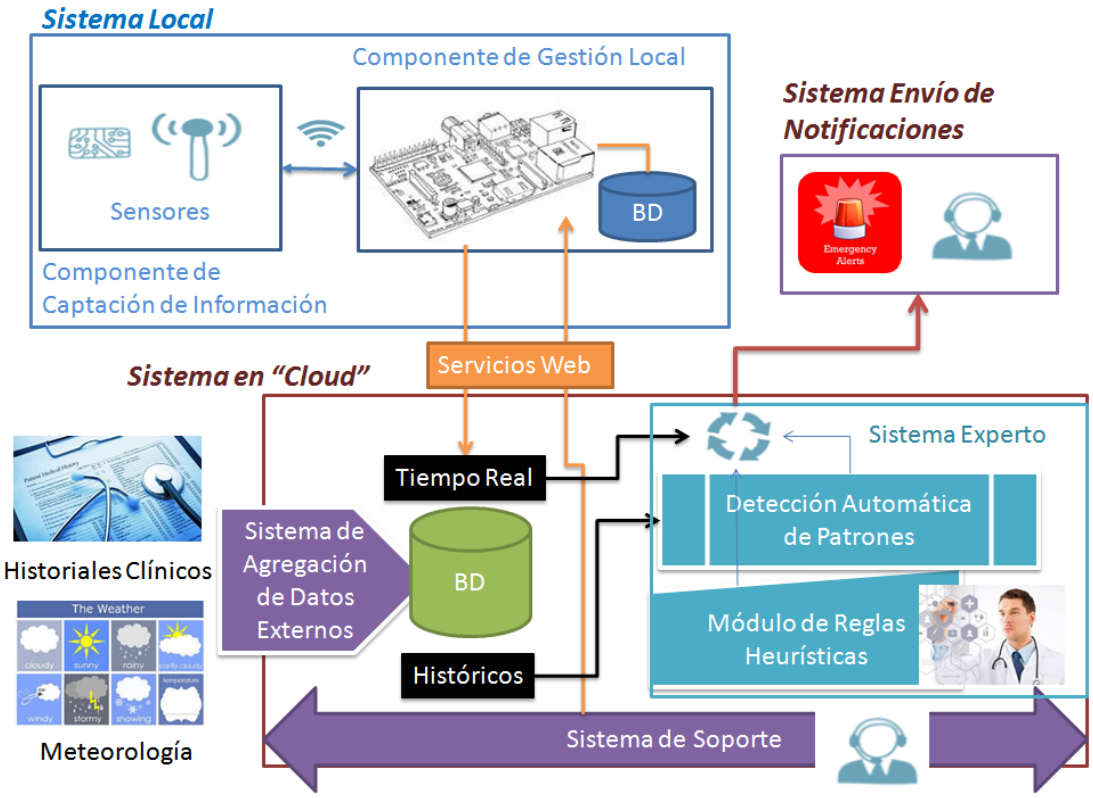


Figura 4.2 Componentes de la Arquitectura del sistema

4.3 Sistema Local

4.3.1 Componente de Captación de Información

La versatilidad del sistema es, básicamente, la inclusión del denominado "Universal Sensor Connection", que recibe automáticamente datos de los sensores, los almacena temporalmente en una base de datos temporal, y finalmente, los envía a un sistema central externo (Sistema Cloud), independientemente del contenido que contengan. Posteriormente, el Sistema Cloud infiere cómo estos valores afectan el comportamiento de gestión del objeto de análisis. Por lo tanto, el componente de captación de información puede incorporar una multitud de sensores[24] para detectar información básica sobre el estado de los usuarios y el estado de su hogar, de forma abierta, plástica, flexible y sin afectar al “kernel” del sistema.

[24] https://github.com/universAAL/ontology/wiki/ActivityHub

Los sensores en general son los dispositivos responsables de convertir eventos del mundo real en señales electromagnéticas. Hoy en día, existe una gran variedad de sensores que pueden clasificarse aproximadamente en tres grandes categorías, como se describe en [Kon12]. La primera categoría se refiere a sensores inalámbricos que pueden medir y monitorizar métricas ambientales tales como temperatura, humedad, ubicación, aceleración, presión, etc. Esta categoría incluye sensores que se despliegan sobre un área para detectar y supervisar sus métricas en el ámbito de una aplicación específica. La segunda categoría se refiere a teléfonos inteligentes, PDAs y otros dispositivos manuales o incorporados con interfaces de comunicación como Bluetooth, GPS y Wi-Fi. Estos dispositivos pueden ser considerados como sensores, ya que son capaces de comunicar datos sensibles a la ubicación a los agregadores de datos. La tercera categoría se refiere a los sensores audiovisuales. Se pueden colocar cámaras y / o micrófonos para monitorear las áreas de interés y enviar los datos recopilados a una capa arquitectónica superior. El Componente de Captación de Información admite actualmente todos ellos. Los sensores envían la información del entorno a través de la red doméstica hacia el Componente de Almacenamiento Local, que además, revisa cada cierto tiempo si los sensores están activos. Para la comunicación con el Componente de Almacenamiento Local, utilizamos el protocolo UPnP, que no sólo cubre protocolos de Internet como TCP / IP, HTTP, SOAP, UDP o XML, sino que también integra Zigbee, USB, IEEE802.11, BT, BLE Wi -Fi y consideraciones de seguridad utilizando técnicas de seguridad como certificados X.509. Además, este protocolo está abierto y se puede extender (por ejemplo, para definir un tipo específico de mensaje que añade atributos adicionales).

4.3.2 Componente de Gestión Local

El agregador de datos es la pieza necesaria para realizar las siguientes funciones:

- Recopilar y agregar los datos de todos los sensores en una base de datos local.
- Chequear que los sensores está activos.
- Intercambiar datos entre el Sistema Local y el Sistema en “Cloud”.

En el presente trabajo se ha elegido un Raspberry Pi como la Unidad Computacional básica para ejecutar UniversAAL y también para conectarse a los sensores inalámbricos de la casa. Debido al hecho de que los sensores están típicamente desplegados alrededor de la casa con una simple estructura cableada de dos pares, o con sensores inalámbricos, se necesita un registrador de datos doméstico para proporcionar conectividad y control con componentes de nivel superior. Este componente puede trabajar con dos tipos de entradas; por una parte, con sensores binarios que disparan una señal al sistema cuando detecta un evento, como alarmas de incendio, detectores de presencia o contactos de las puertas, que recibe a través de una LAN o una red Wi-Fi. Por otro lado, se pueden conectar sensores de valores continuos al agregador de datos, tales como sondas de temperatura, que proporcionan información constante en lugar de disparar un evento al sistema. Este tipo de sensores se usan típicamente para monitorizar las condiciones ambientales en una habitación. Para la comunicación con el Sistema en “Cloud”, utiliza una gran varios protocolos de transporte como HTTP o XML, a través de Servicios Web y la comunicación es bidireccional, el Componente de Gestión Locl puede ser gestionado remotamente desde el Sistema “Cloud”. El sistema, que está desatendido, debe ser calibrado por un operador, y su mantenimiento y control es totalmente transparente para los residentes en el

domicilio. Para ello, se ha implementado una arquitectura de software basada en una estructura en capas con un sistema operativo (Linux, Android o Windows), una capa de lenguaje Java, una capa de servicio OSGi y diferentes "paquetes" de aplicaciones de software. OSGi es un sistema de módulos y una plataforma de servicio para el lenguaje de programación Java pueda implementar modelos de componentes completos y dinámicos. Las aplicaciones o los componentes (que vienen en forma de paquetes para el despliegue) se pueden instalar, iniciar, detener, actualizar y desinstalar remotamente sin necesidad de reiniciar el sistema. Por otro lado, el sistema presenta la necesidad de almacenar y tratar los datos obtenidos. Al ser un dispositivo local, en el domicilio, limitado en su computación, no se pueden utilizar sistemas de gestión de base de datos convencionales, por lo que la agregación de datos de ha realizado sobre una base de datos no-SQL, denominada ***Redis***. Redis es un motor de base de datos que opera principalmente en memoria RAM, basado en el almacenamiento en tablas de hashes (clave/valor), con capacidad de ejecutarse en entornos con escasa capacidad de cálculo. Está liberado bajo licencia BSD por lo que es considerado software de código abierto. Redis basa la estructura de datos en el uso de un diccionario o tabla de hashes que relaciona una llave a un contenido almacenado en un índice. La principal diferencia entre Redis y otros sistemas similares es que los valores no están limitados a ser de tipo “string”, ya que da soporte además de la habitual estructura clave-valor, otras estructuras diferentes de datos más complejas.

4.4. Sistema en “Cloud”.

El módulo Cloud es el servidor central que da soporte y almacena todos los datos recogidos de las diferentes fuentes de datos mencionadas en el proyecto. Para ello dispone de una base de datos en la que se almacenarán los diferentes datos recogidos. La forma de recibir dichos datos es mediante llamadas a servicios web WCF (Windows Communication Foundation), tanto desde el módulo local como desde Internet. Estos servicios se despliegan en el Cloud, más concretamente, en el servidor web de Windows IIS (Internet Information Services). El módulo “Cloud” también dispone de las herramientas necesarias para hacer el análisis de los datos almacenados y determinar los diferentes perfiles, principalmente utilizando técnicas de Datamining (Minería de Datos) con las plataformas SPSS Modeler y RapidMiner instaladas en su sistema central de procesamiento. Otro módulo disponible en el servidor es el Workflow encargado de gestionar la forma de actuar ante una alerta y que se encarga de mandar las notificaciones de las posibles alertas a la persona indicada en cada momento (cuidadores de la persona asistida). En resumen, se incorporan los siguientes elementos “cloud” a la plataforma:

- Un Servidor Web: servidor para gestionar las peticiones recibidas por el Servidor Central.
- Un servidor FTP: repositorio que contiene los ficheros de configuración necesarios para llevar a cabo las actualizaciones del Gateway (HSB) del hogar.
- Servidor de aplicaciones web: servicio de aplicaciones web que el usuario utilizará de forma remota.
- Servidor de base de datos: servidor que proporciona la persistencia de la información y/o datos de los servicios de monitorización de parámetros fisiológicos y dietas saludables.

- Servidor de Interoperabilidad: servidor encargado de integrar los dispositivos móviles del hogar con la plataforma remota a través de Web Services.
- Servidor de workflow: servidor disponible para gestionar el flujo de las alarmas de seguridad que se desencadenan.
- Servidor de sistema experto, tratamiento de datos clínicos, y detección automático de patrones: servidor encargado de extraer, y generar conocimiento de la información “base” recogida por los sensores del hogar, agregada con información clínica y externa de otros sistemas, y procesada para la generación de alertas y recomendaciones.
- Subsistema de Persistencia: el sistema debe preservar la información recibida de forma permanente permitiendo que sea reutilizable en cualquier momento. Así, será posible que el receptor de la información no tenga que estar operativo al mismo tiempo que se realiza la comunicación
- Subsistema de Comunicaciones: el sistema debe permitir que existan comunicaciones entre las distintas entidades que forman el sistema para poder transmitir información en todo momento. En este caso, deberá existir comunicación entre todos los hogares y el Sistema Central de Servicios.

4.4.1 Sistema de Agregación de Datos

4.4.1.1 Información extraída del Sistema Local.

El Sistema de Agregación de Datos pide de forma asíncrona, vía servicio Web, al Componente de Gestión Local, que le suministre la información que tienen en memoria cada cierto tiempo. Una vez recibida esta información, el Sistema de Agregación de Datos verifica si esta información, en primer lugar, es correcta, y en segundo lugar, si se ha genera un cambio en sus valores sobre lecturas anteriores o no. Si la nueva información que llega es diferente sobre ciertos umbrales límites de variación con la última información recibida, el evento se registra en la Ontología del Sistema. El motor que gestiona la calidad de los datos, y la lógica de creación de eventos son las denominadas “Reglas de Proceso”, sitas en la Ontología del Sistema. Posteriormente, los eventos ya generados se transforman de datos en bruto a datos de contexto, “codificados semánticamente”, en base otro subconjunto de reglas (SWRL), introducidas manualmente en la Ontología del Sistema, dentro del mismo grupo de “Reglas de Proceso”. Por ejemplo, “estar en la cocina por la mañana después de una ducha” se traduce en <desayunar>. El procesar las reglas a partir del razonador de la Ontología ayuda a generar este tipo de instancias de forma automática. Pueden verse un ejemplo de este tipo de reglas en la figura 7.6. A continuación se muestran algunos ejemplos de esta funcionalidad:

- El sistema deduce si el anciano “está comiendo a una hora específica del día”: el sistema de localización en interiores es consciente de la hora del día y la cantidad de tiempo que las personas mayores permanecen en la cocina. El uso de electrodomésticos también se podría tiene en cuenta para inferir si el anciano ha estado preparando su comida. Las puertas de contacto en muebles de cocina también se utilizan para seguir esta situación.

- Durante el día, el sistema deduce si los usuarios están realizando “tareas domésticas”. El sistema de localización en interiores es consciente de la hora del día y la cantidad de tiempo que el anciano pasa por las distintas dependencias de su residencia habitual, y puede deducir este escenario en base a un seguimiento constante (cada cierto periodo) de su posición. También se puede detectar ausencia de movimiento, como factor importante en los modelos predictivos.
- El ejercicio físico, como caminatas o paseos, se puede detectar a través del teléfono inteligente, habilitando una aplicación de seguimiento del sistema GPS, basado en el movimiento exterior de los ancianos. La posición, la velocidad y la trayectoria serán analizadas para medir la cantidad de kilómetros realizados cada día.

4.4.1.2 Información adicional extraída del registro electrónico de salud.

El cuidado de los adultos mayores requiere un enfoque multidisciplinario y puede incluir desde la monitorización del estado de salud de adultos mayores o personas, pasando por las recomendaciones de hábitos de vida, control de situaciones en condiciones diagnósticas crónicas y la inclusión de la necesidad del control de la realización de elementos terapéuticos o la adherencia a uno o más medicamentos recetados para su uso regular. Este contexto complejo debe tenerse en cuenta en un sistema de teleasistencia [Aca13]. El conocimiento de la ***información médica personal*** de las personas mayores en un sistema de teleasistencia permite mejorar la calidad de sus hábitos de vida, previendo situaciones peligrosas derivadas de sus patologías. En este trabajo, la información clínica de los usuarios viene dada en forma de historiales clínicos, en texto libre. Para facilitar la introducción de reglas dentro del Módulo de Reglas Heurísticas a los expertos clínicos y de atención asistencial, se ha desarrollado, dentro del Sistema de Agregación de datos externos, un componente cuyo objetivo es resumir en un diagrama temporal (ver figura 4.4), en forma de resumen, el historial clínico de los usuarios, resaltando cuáles son los diagnósticos y tratamientos (fármacos) principales, secundarios y activos. Estos resúmenes se integran dentro de una ontología, que denominamos “***Ontología Clínica***”.

4.4.1.3 Creación Automática de Resúmenes Médicos basados en Evolutivos escritos en Lenguaje Natural

Hoy en día, los registros de salud electrónicos (EHR) almacenan la mayor parte de la información referente a la evolución de los pacientes, en informes almacenados en lenguaje natural (80% de la información pertinente) [Gil16], escrita por el personal médico, de manera que el porcentaje de información estructurada que contiene datos cuantitativos es mínimo. La información en texto libre es fundamental para los investigadores biomédicos, que necesitan información detallada sensible los eventos que ocurren a lo largo del tiempo. La extracción de la información de los EHR no es una tarea fácil y requiere el desarrollo de algoritmos de correlación, reducción de ruido e inferencia. La codificación de los datos dentro de los registros de salud

electrónicos y bases de datos clínicos es esencial para la difusión y el intercambio entre sistemas heterogéneos. Los desafíos actuales para la adopción de normas de terminología incluyen el uso de un vocabulario de codificación local y los modismos particulares de un centro, más el esfuerzo necesario para mapear de sistemas dispares a un nivel dado, así como la identificación de las normas apropiadas para cada contexto [Kar11]. En general, este objetivo implica el uso de una combinación de diferentes enfoques (automatizados, semiautomáticos y manuales) y la combinación de distintos sistemas de terminología (Master Drug Base de Datos [MDDB], RxNorm, Nomenclatura Sistemática de Medicina-Clínica [SNOMED CT] Medical Language System [UMLS] Metathesaurus) para identificar conceptos de una forma no ambigua [Mat12].

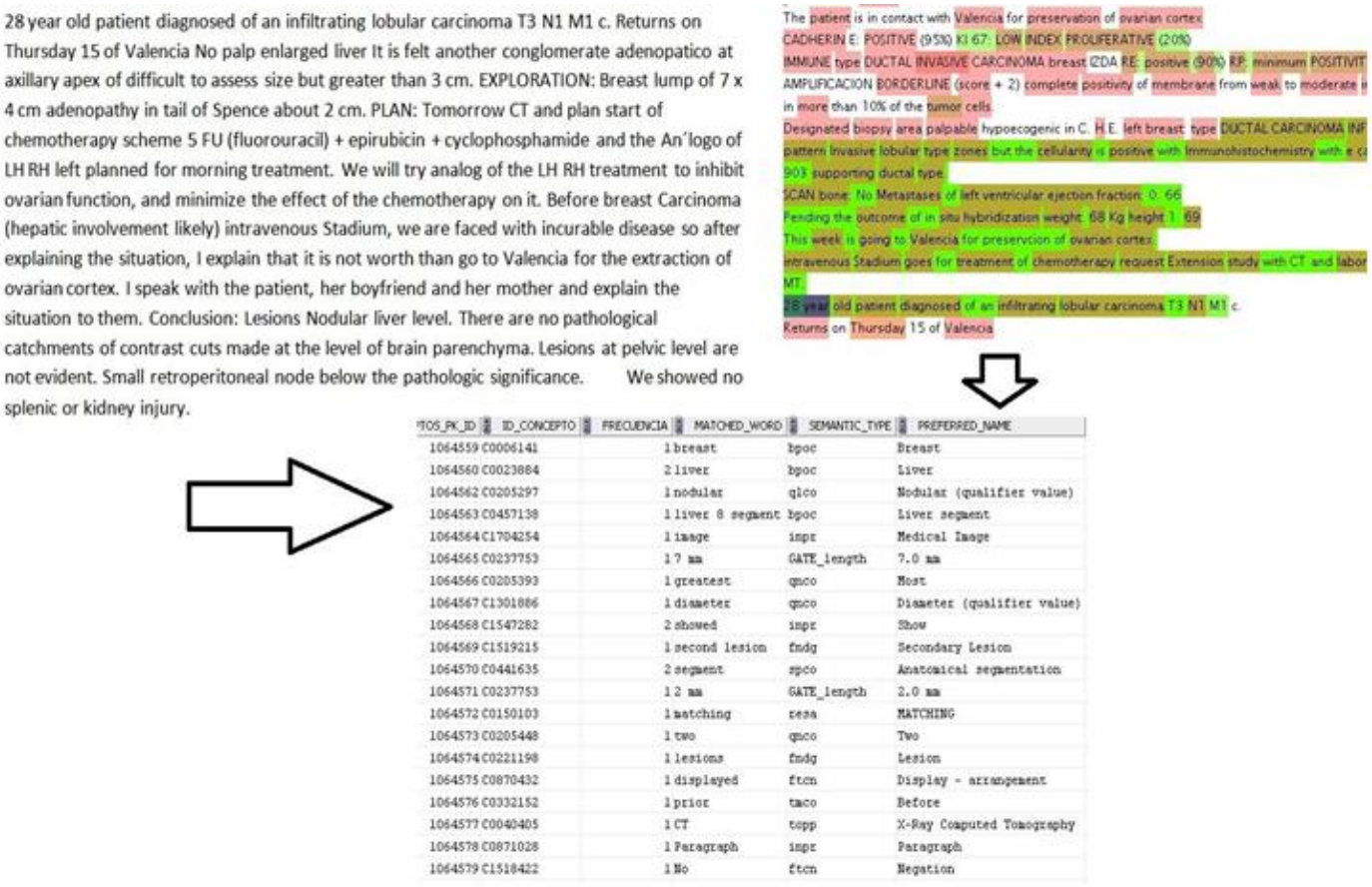


Figura 4.3 Anotación en base al tesauro UMLS.

La meta en este punto del trabajo es extraer un conjunto de los eventos clínicos más relevantes relacionados con la evolución del paciente a partir de sus historiales, con dos objetivos:

1. Alimentar a la plataforma de teleasistencia con los datos clínicos más relevantes obtenidos de una manera no asistida de los historiales médicos, a partir de las denominadas “Reglas de Proceso” y mover los procedimientos terapéuticos, tratamientos o recomendaciones médicas resumidos a la Ontología Central (UniversAAL).
2. Generar unos informes digitales en los que, en formato de secuencia temporal (resumen médico), se muestren dichos eventos de una forma gráfica a los expertos médicos de cara a ayudar la generación de recomendaciones de control y seguimiento de los pacientes en sus domicilios (Reglas Heurísticas).

Los EHR normalmente contienen descripciones de diferentes episodios escritos en lenguaje natural por el personal médico [Lin15], y a menudo son multilingües (español, euskera, etc.), con sus impresiones diagnósticas, tratamientos, procedimientos, etc. Para conformar dicho resumen, es necesario, en un primer lugar, extractar e identificar cuáles son o han sido los diagnósticos principales, cuáles los diagnósticos secundarios, y cuáles son los tratamientos o secuencia de acciones asociadas a los mismos, para después, asignarles una importancia y un orden. Por ejemplo, en un contexto oncológico, probablemente, un esguince no sea un diagnóstico principal, sino secundario. No es sencillo identificar cuáles son los diagnósticos principales o secundarios, y los tratamientos asociados, o cuál es la enfermedad actual a partir de la información presente en los historiales clínicos. Existen descripciones sobre los antecedentes familiares, sobre otros tratamientos secundarios y una serie de referencias a enfermedades relacionadas en el pasado que quizás no sean relevantes para el estado actual del usuario, o sí, en función del paciente. Para solventar este problema, en este trabajo, se ha desarrollado un sistema que permite, en base a los textos extraídos de los evolutivos e historiales clínicos de los usuarios, detectar la enfermedad principal y sus diagnósticos (principales y secundarios), los principales procedimientos médicos asociados, y los tratamientos y medicamentos que impactan directamente en la vida del usuario, incorporándose a la plataforma de teleasistencia como un nuevo indicador. Los cambios normales relacionados con la edad pueden ir acompañados de problemas de salud crónicos como la diabetes o las enfermedades cardíacas. El tratamiento de muchas de estas enfermedades crónicas puede incluir uno o más medicamentos recetados diariamente para su uso regular. Combinados, estos factores aumentan la complejidad del diseño de los sistemas de teleasistencia. El sistema desarrollado es capaz de traducir textos médicos literales en una estructura semántica (la ontología central), con el fin de encontrar las correlaciones entre la situación clínica del usuario (enfermedades, tratamientos, fármacos), y relacionarla con los requisitos personales de estilo de vida de cada usuario que tenemos que el domicilio con asistencia remota. Existen técnicas matemáticas que se utilizan para capturar la estructura semántica de documentos basadas en correlaciones entre elementos textuales dentro de ellos [Kri15], sin embargo, dentro del trabajo propuesto, se ha desarrollado un nuevo método para crear resúmenes médicos, apoyado por un sistema híbrido estadístico / semántico (ver ejemplo en la figura 4.4).

Los algoritmos de anotación propuestos trabajan sobre diferentes fuentes, lenguajes y tipos de datos, basados en vastas fuentes de información de texto multidisciplinares (historias clínicas e informes médicos que comprenden un amplio conjunto de disciplinas médicas). Conseguir este objetivo implica las siguientes acciones:

• Extraer información de informes médicos de lenguaje natural en EHR, que comprenden un amplio contexto médico y una gran variabilidad entre los diferentes campos médicos, países e incluso hospitales.

• Definir un nuevo modelo capaz de comprender, normalizar y estructurar la información contenida en los diferentes informes clínicos generando un único descriptor estructurado en un formato estándar.

• Enriquecer la información de otras fuentes, tanto externas como internas, de datos heterogéneos.

• Comprimir los modelos generados y los descriptores obtenidos en estructuras sencillas que soporten la compacidad, la respuesta ante preguntas en tiempo real, y la inferencia en los resúmenes semánticos.

• Desarrollar un modelo de ontología semántica capaz de estructurar, normalizar, enriquecer y compactar las historias clínicas escritas, informes de ensayos y registros clínicos, independientemente de la disciplina

médica, país, idioma, hospital y profesional. Esta tarea se centrará en el modelado de información relevante para aplicaciones de práctica clínica, por ejemplo, la integración con otros datos clínicos recopilados y el uso de esta información para la búsqueda de nuevas investigaciones médicas.

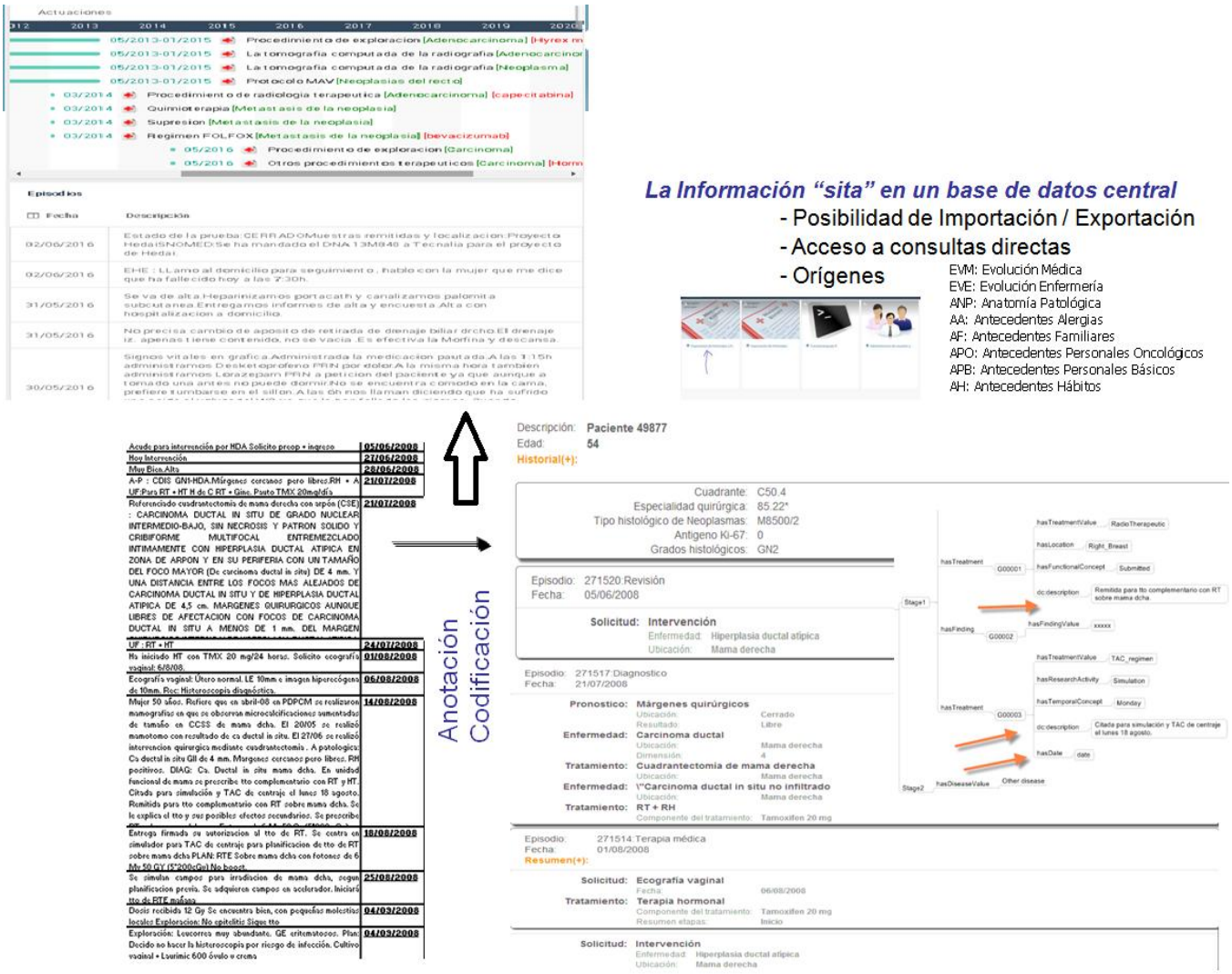


Figura 4.4. Ejemplo de Resúmenes Médicos para un paciente en base a su Historial Clínico en Lenguaje Natural

El sistema tiene en cuenta la correlación de todos los conceptos en los registros clínicos sobre el diagnóstico primario del paciente, calculando diferentes vectores de información relacionada con dicho diagnóstico en base al algoritmo de frecuencia "tf/idf", que permite la ponderación de los términos necesaria para representar un documento y permitir su posterior recuperación. Esto implica que se debe determinar el poder de resolución de los términos de la colección, o lo que es lo mismo, la capacidad de los términos para representar el contenido de los documentos en la colección, que permitan identificar cuáles son relevantes o no ante la consulta de un usuario. Al valor e índice que es capaz de determinar este extremo se le denomina "peso del término" o

"ponderación del término" y su cálculo implica determinar la "Frecuencia de aparición del término TF" y la "Frecuencia inversa del documento para un término IDF".

Posteriormente, ejecutamos un proceso de “mapeo” entre conceptos que analiza la correlación directa de variables secundarias con los diagnósticos principales, extrayendo sus relaciones y recodificados bajo el tesauro del Sistema de Lenguaje Médico Unificado (UMLS). El tesauro de UMLS es una gran base de datos de vocabularios médicos, multipropósito y multilingüe organizada por conceptos semánticos. La versión actual incluye más de 1,5 millones de términos biomédicos de más de 100 fuentes diferentes. Los términos sinónimos se agrupan para formar un concepto o grupo único. Los conceptos están vinculados a otros conceptos por medio de diversos tipos de relaciones, almacenado en un grafo complejo. Además, se proporciona una categorización consistente de todos los conceptos representados en el tesauro de UMLS (ver Figura 4.3), así como información sobre el conjunto de Tipos Semánticos básicos o categorías que pueden asignarse a esos conceptos. En este trabajo hemos anotado, en base a los historiales clínicos, 133 tipos semánticos en 54 posibles relaciones entre ellos. Para la consecución de los objetivos iniciales, en este trabajo se ha construido una “Ontología Clínica”, basada en arquetipos OpenEHR, que es utilizada como base principal para construir las clases semánticas, subclases y propiedades de todo el modelado clínico. (ilustrada en la figura 4.5)

Figura 4.5. Ontología Médica General del Trabajo

La ontología desarrollada sobre tecnología OWL es capaz de almacenar información en un formato semántico, episodio por episodio, sobre cuáles son los conceptos más importantes extraídos en cada uno de ellos, si se ha producido un diagnóstico secundario, o si el episodio habla de un diagnóstico general previo. También recoge si ha habido algún tipo de tratamiento, procedimiento médico, cirugía, si se ha recomendado al usuario algún fármaco en particular, si se ha modificado el tratamiento, si se ha detectado alguna patología específica, o simplemente, si el paciente ha sido informado de algo. La ontología propuesta cumple la norma CEN/ISO EN13606, que es una norma europea del Comité Europeo de Normalización (CEN), diseñada para lograr la interoperabilidad semántica en la comunicación electrónica de registro sanitario. Está demostrado que la tecnología OWL logra una alta eficiencia, precisión, escalabilidad y efectividad [Che14b]. También es necesario almacenar la evolución de los pacientes (actuaciones clínicas) denominada "Observaciones" en OpenEHR. Para unir estos conceptos a nuestra ontología, enriquecemos nuestra ontología base con una ontología específica que define de forma estándar las “Observaciones Clínicas”, incluyendo la identificación de

elementos del modelo de información, conceptos de vocabulario y tipos de datos de estándares clave como HL7 / RIM, Modelos Clínicos Detallados (DCM), la Arquitectura de Documentos Clínicos (CDA) y el Modelo de Tabulación de Datos de Estudio[25].

El algoritmo general para extractar la información relevante de los EHR, contiene las siguientes fases:

1. ***Fase1: Extracción, Anotación y Codificación de los Conceptos Médicos.***

   En esta Fase se extraen los "tokens" o conceptos relevantes de oraciones a lo largo de diferentes episodios en todo el registro de salud.

   - Todo el texto del registro clínico se divide en frases usando el "tokenizer", parte de los módulos de tagger y splitter de la arquitectura de la Arquitectura General para la Ingeniería de Texto (GATE) para la ingeniería de texto y transformándose en la codificación UMLS. (Ver figura 4.6).
   - Para ello se utilizan algoritmos de "stopwords", lematización, diccionarios, traducciones, corrección ortográfica, y generación de n-gramas.
   - Para la anotación de los conceptos o n-gramas, se utilizan los algoritmos de anotación de MetaMap[26] (una herramienta que lleva a cabo un análisis de textos biomédicos y que presenta un elevado grado de configurabilidad, obtenidos tras procesar una selección amplia de documentos médicos extraídos de la base de datos biomédica MedLine2010), con una configuración muy restrictiva para evitar resultados muy ambiguos al asignar texto a la ontología UMLS, resolviendo la expansión de los acrónimos conocidos.
   - Se corrigen o eliminan las cadenas de información (conceptos o n-gramas) que no son reconocidos por el anotador UMLS.

2. ***Fase 2: Jerarquización de Conceptos en grupos principales y secundarios:***

   Cada uno de estos códigos UMLS se asigna a un grupo principal en la ontología del trabajo, indicando a qué parte del proceso dentro del episodio pertenece dicho concepto. Un Diagnóstico no tiene el mismo peso que un Tratamiento o que una ubicación corporal. Con nuestro algoritmo de ponderación, el sistema selecciona todas las entradas, elige los episodios más relevantes y, dentro de ellos, selecciona las enfermedades principales y secundarias para cada episodio:

   Cada uno de estos códigos UMLS se asigna a un grupo principal en la ontología del trabajo, indicando a qué parte del proceso dentro del episodio pertenece dicho concepto. Un Diagnóstico no tiene el mismo peso que un Tratamiento o que una ubicación corporal.

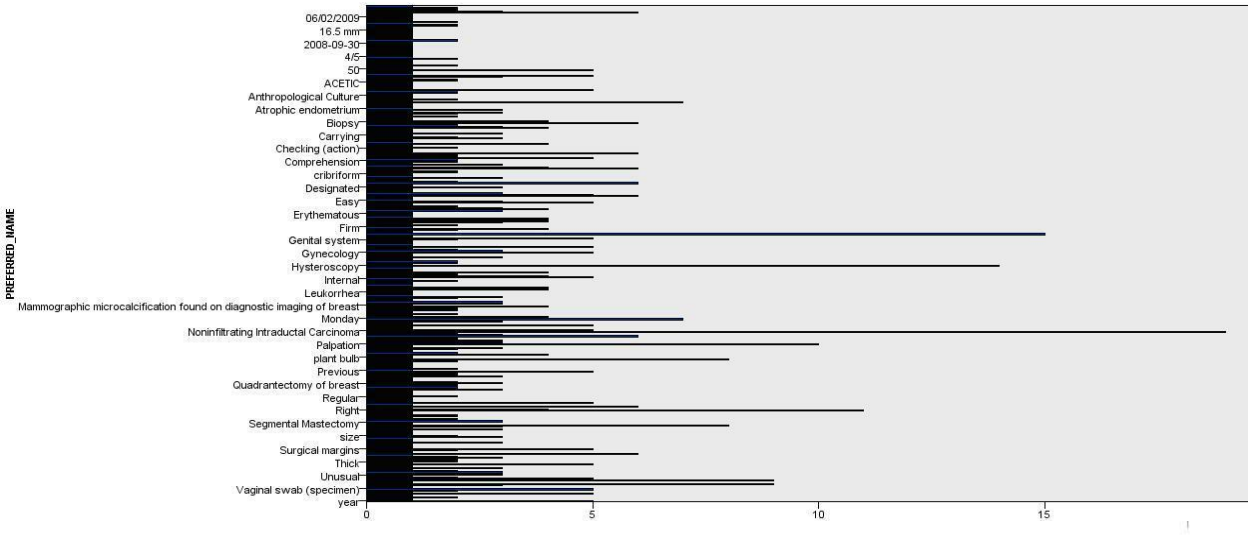
   Con nuestro algoritmo de ponderación, el sistema selecciona todas las entradas, elige los episodios más relevantes y, dentro de ellos, selecciona las enfermedades principales y secundarias para cada episodio:

   - Esta asignación requiere un mecanismo adecuado de selección, basado en la adición de reglas semánticas mediante el lenguaje de reglas de red semántica (SWRL) y en un motor de inferencia que

[25] Russler, Dan and Moores, Matt and Chen, Helen and Mirhaji, Parsa and Richesson, Rachel and Pathak, Jyoti and Kashyap, Vipul, "RDF/OWL Representation of HL7/RIM v3.0" (2008). https://www.w3.org/wiki/HCLS/ClinicalObservationsInteroperability/RIMRDFOWL

[26] National Library of Medicine. MetaMap. http://mmtx.nlm.nih.gov/.

mantiene una lógica de asignación basado en la cercanía de los conceptos a sus diagnósticos o tratamientos principales o secundarios (Ver Figura 4.8).



Woman year 50 - ccss's mammography microcalcifications - right breast - quadrantectomy surgical intervention 4 mm ductal pathological - close free margins - RH
TM: 3/irreg rule itching (Pruritus) Scar - puncture - endovaginal Normal size - LE utero secretor

Disease:

C0007124	1	dcis	neop	Noninfiltrating Intraductal Carcinoma
C0337354	1	quadrantectomy	topp	Quadrantectomy of breast
C0222600	1	right breast	bpoc	Right breast
C1176475	1	ductal carcinoma	neop	Ductal Carcinoma
C1318216	1	4	qnco	4
C0237753	1	4 mm	GATE length	4.0 mm

Treatment:

scan	Radionuclide Imaging
secretion	Bodily secretions
tamoxifen	Tamoxifen

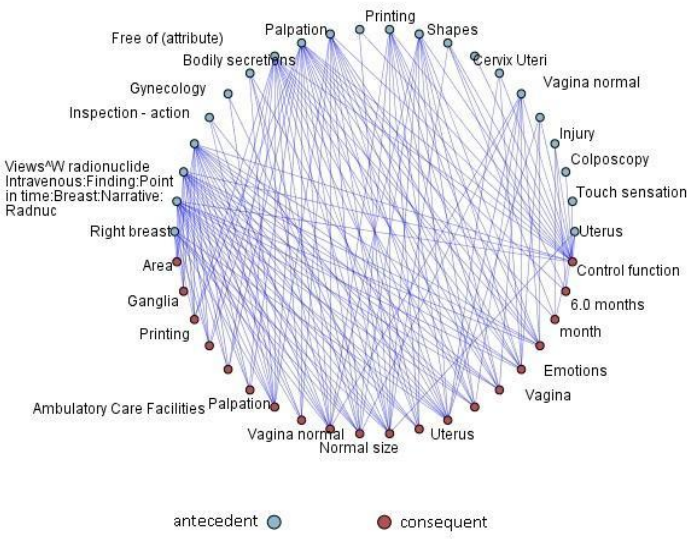
Figura 4.6. Extracción, Anotación y Codificación de los Conceptos Médicos.

- En realidad, es una lógica para determinar los puntos de ruptura en cada episodio con respecto a una enfermedad principal, a un tratamiento principal, y que permite agrupar las recomendaciones médicas y la transformación de los conceptos encontrados en cada grupo.

- Para incorporar la parte semántica en el proceso [Lin15], capaz de desambiguar términos con varias interpretaciones, se utilizan dos fuentes externas de anotación: las jerarquías de codificación UMLS y un diccionario médico de sinónimos y acrónimos.

3. Fase 3: Filtrado estadístico de los tipos semánticos.

En este proceso, se seleccionan, por métodos estadísticos (distribución Zipf [Pia14]) , cuáles son los conceptos o conjuntos de conceptos que mejor representan al conjunto de episodios, y en base a esta información, con un algoritmo de ponderación y asociación estadística, se seleccionan las clases o pares de tipos semánticos o jerarquías relacionadas que mantienen una ganancia de información mínima que soporta los resúmenes generales, como se muestra en la figura 4.7. Estas relaciones, además, nos indican qué tipo de propiedades o tripletas pueden coexistir en el sistema. Para cada clase o par de tipos semánticos únicos se asignan manualmente a una regla de anotación, que se utilizará en la generación de instancias de la Fase 6. (Ver Figura 4.9),



Support	Item 1	Item 2
0.055	Free of attrib	
0.048	month	
0.048	menp	
0.048	medd	
0.048	inpo	
0.048	Shapes	
0.048	Mammograp	
0.117	spco	bpoc
0.110	spco	qnco
0.090	spco	diap
0.062	spco	topp
0.076	spco	fndg
0.048	spco	neop
0.048	spco	dsyn
0.048	spco	GATE_length
0.048	spco	Uterus
0.048	spco	Shapes
0.090	bpoc	qnco
0.055	bpoc	diap
0.062	bpoc	topp
0.069	bpoc	fndg
0.062	bpoc	Vagina
0.055	bpoc	Uterus
0.055	bpoc	Right breast
0.048	bpoc	Shapes
0.055	qnco	diap
0.069	qnco	topp
0.055	qnco	tmco

Figura 4.7. Tipos semánticos y su importancia en el Corpus.

4. Fase 4: Ponderación de la importancia estadística

Posteriormente, se calcula el peso (la importancia) de cada episodio en el conjunto del Historial y el peso de cada concepto en el episodio. Con esta información simple, somos capaces de ordenar los conceptos en una lista jerárquica de importancia con el fin de:

- Presentar al personal médico únicamente las relaciones pertinentes entre conceptos, medida estadísticamente. No todos los episodios de un expediente médico tienen el mismo peso en el

historial del paciente, y también dentro de cada episodio, no todas las oraciones son relevantes: puede haber referencias a antecedentes, observaciones médicamente no significativas o no suficientemente relevantes para aparecer en un resumen médico ni para un posterior proceso analítico. Por lo tanto, el sistema registra todos los conceptos de la ontología, pero más adelante, en la presentación del resumen final, solo se muestra la información de aquellos episodios que se consideran más relevantes, es decir, aquellos que tienen una correlación objetiva más fuerte con el diagnóstico principal de cada paciente.

- Analizar las relaciones desconocidas entre las enfermedades, los procedimientos, los tratamientos y los datos personales del paciente como temas familiares, sexo, edad, situación demográfica o económica, contexto personal, fármacos [Ahl15], etc.

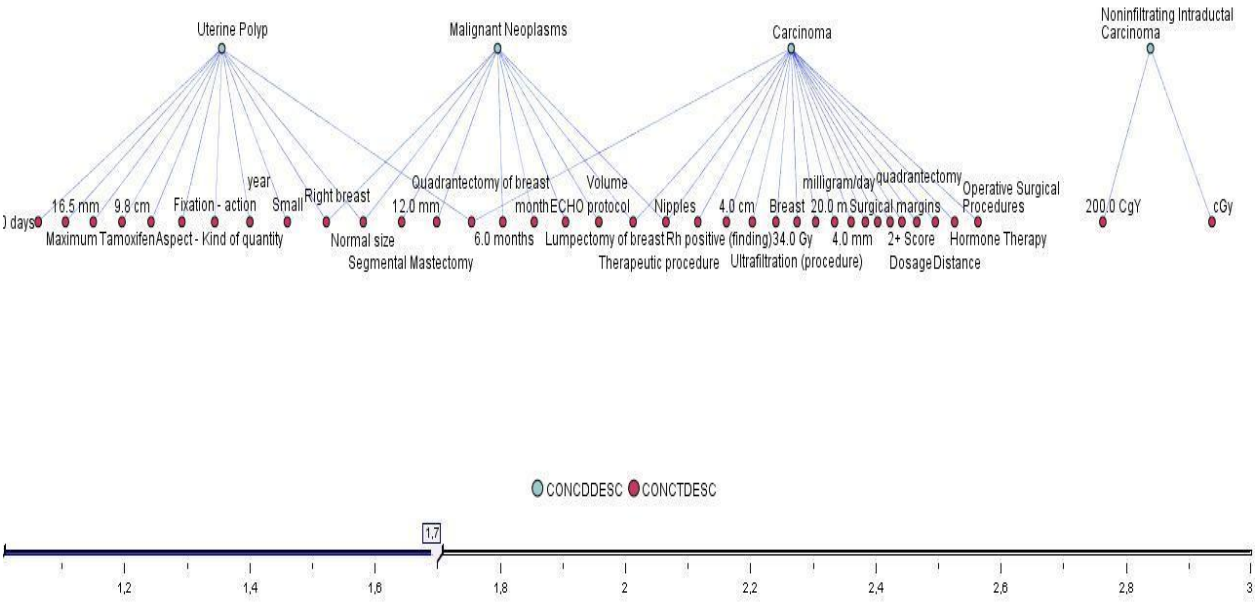


Figura 4.8. Relación estadística entre conceptos y jerarquización. .

5. Fase 5: Filtrado semántico de la información anotada.

En esta fase, se selecciona sólo la información relevante, medida semánticamente. Para ello, se utilizan las relaciones jerárquicas generadas en la Fase 3, con el objetivo de crear las tripletas semánticas en un formato o sintaxis de lógica predicativa o lógica de primer orden, agregando solamente aquellos tipos de relaciones que permiten una mejora en la calidad de los resúmenes. En este caso, se realiza un filtrado previo de las clases más relevantes que se deben anotar (previamente definidas por expertos médicos, en base a los intereses de cada especialidad o perfil del médico), gracias a la aplicación de las Reglas de Procesos (SWRL) en un proceso de inferencia sobre la ontología, gracias al motor de razonamiento Pellet. De esta forma, sólo se seleccionan ciertos tipos

semánticos de información (por ejemplo, diagnósticos y tratamientos, pero no “revisiones periódicas”), o conjuntos de tipos semánticos previamente definidos por los expertos médicos.

6. ***Fase 6: Generación de instancias en la ontología.***

   Finalmente, se anotan los conceptos en la ontología de los resúmenes con un conjunto de reglas que permiten realizar y explicitar la equivalencia en los conceptos, o la relación entre conceptos a partir de las relaciones o propiedades que los unen. Cada par de conceptos se transforman en una unión de relación (un grafo), y cada grafo se corresponde a una categoría diferente de información (por ejemplo, enfermedades, síntomas y signos o medicamentos). Para crear estos grafos, necesitamos anotar esta información bajo los conceptos de la Ontología Clínica en forma de "tripletas" de acuerdo con un conjunto de reglas de alto nivel (Reglas de Proceso) definidas en el sistema de inferencia Pellet. De esta manera, el sistema es capaz de transformar los conceptos del texto en una estructura semántica jerárquica. Así mismo, el sistema utiliza el mismo motor de inferencia para asignar las tripletas o grafos creados al grupo de diagnóstico que le corresponde (principal o secundario). La ventaja de esta aproximación es que las Reglas de Proceso (filtrado semántico, creación de grafos y asignación a diagnósticos principales o secundarios) son completamente configurables por el personal médico, sin conocimiento informático, estadístico o semántico, y por lo tanto, de una manera muy dinámica y elástica, permitiendo una variación de diferentes contextos médicos. La figura 4.10 muestra un registro clínico en la Ontología Clínica del trabajo.

7. ***Fase 7: Instanciación de los Datos Clínicos en la Ontología del Sistema (UniversAAL)***

   Es en esta Fase en dónde los resúmenes automáticos generados en la Fase 6 alimentan a la Ontología de Teleasistencia (UniversAAL) con los datos clínicos más relevantes obtenidos de una manera no asistida de los resúmenes médicos y en dónde los procedimientos terapéuticos, tratamientos o recomendaciones médicas de los resúmenes médicos se complementan con el resto de datos del Sistema (tanto de eventos en el domicilio, como de datos meteorológicos). Para construir esta integración, gracias al proceso de anotación semántica, sólo necesitamos unir los conceptos de los resúmenes clínicos en la Ontología Clínica (ver figura 4.11) con la conocida propiedad "Same_as" enlazándolos con la Ontología Central (ver figura 4.12), tal y como recomiendan las buenas prácticas en la gestión en la vinculación de datos semánticos [Dam12]. Una vez almacenada toda la información en la Ontología Central, es accesible mediante consultas SPARQL con acceso por medio de “endpoints” o puntos de consulta semánticos, vinculando esta información de forma directa a otras plataformas clínicas como "LinkedLifeData" [Kam14].

REGLA...	REGLASRESUMENE...	REGLAS...	REGLASRESUMENES_S...	REGLASRESUMENES_PROPIEDADONTO	REGLASRESUMENES_CLASEO...	REGLA...	REGLASRESUMENE...	REGLASRESUMENES_CLA...
10	T	P	acty$topp	hasTreatment	Treatment	instance	therapeutic-act	Tratamiento
47	Anormalidad an...	S	anab	hasAnatomicalAbnormality	Anatomical_Abnormality	instance	(null)	Anormalidad anatomica
161	secundario	S	blor	hasBodyPart	Body_Part	instance	(null)	Parte del cuerpo
187	Localizacion d...	S	blor	hasBodyLocation	Location	instance	(null)	Ubicacion en el cuerpo
44	Disciplina Bio...	P	bmod	hasBiomedicalOccupation	Biomedical_Ocupation	instance	medical-histor...	Ocupacion biomedica
48	secundario	S	bpoc	hasBodyPart	Body_Part	instance	(null)	Parte del cuerpo
141	Parte del cuer...	S	bpoc$spco	hasBodyPart	Body_Part	instance	(null)	Parte del cuerpo
102	Espacio del cu...	S	bsoj	hasLocation	Location	instance	(null)	Ubicacion
143	Para el breast...	P	clna	hasResearchActivity	Research_Activity	instance	therapeutic-act	Actividad de invest...
41	Procedimientos...	P	diap	hasDiagnosticProcedure	Diagnostic_Procedure	instance	therapeutic-act	Procedimiento diagn...
61	Atrofia	S	dsyn	hasDisease	Disease	instance	(null)	Enfermedad
45	Hallazgos	S	fndg	hasFinding	Finding	instance	clinical-exami...	Hallazgo
49	secundario	S	ftcn	hasFunctionalConcept	Functional_Concept	instance	(null)	Concepto funcional
46	Parte del cuer...	S	ftcn$bpoc	hasBodyPart	Body_Part	instance	(null)	Parte del cuerpo
81	T	S	GATE_Date	hasDate	(null)	date	(null)	(null)
182	T	S	GATE_DimDesconocida	hasValue	(null)	string	(null)	(null)
25	T	S	GATE_Gray	hasGray	(null)	string	(null)	(null)
21	T	S	GATE_length	hasLength	(null)	string	(null)	(null)
23	T	S	GATE_mass					
26	T	S	GATE_time					
24	T	S	GATE_Voltage					
22	T	S	GATE_volume					
42	Citacion, inte...	P	hlca					
64	herida, o algo...	S	inpo					
62	Lista de espera	S	inpr					
101	Procedimiento ...	S	lbpr					

Annotations | Usage | Rules

Rules:

PrincipalTemp(?x), hasPrincipal(?x, Treated_with) -> hasImportancia(?x, "Secundario")

CancerTypes(?y), PrincipalTemp(?x), hasFunctionalConcept(?x, Free_of_attribute), hasPrincipal(?x, ?y) -> hasImportancia(?x, "Secundario")

CancerTypes(?y), PrincipalTemp(?x), hasFinding(?x, Medical_History), hasPrincipal(?x, ?y) -> hasImportancia(?x, "Secundario")

CancerTypes(?y), PrincipalTemp(?x), hasHealthCareActivity(?x, Patient_History), hasPrincipal(?x, ?y) -> hasImportancia(?x, "Secundario")

CancerTypes(?y), Functional_Concept(?z), PrincipalTemp(?x), hasFunctionalConcept(?x, ?z), hasFunctionalConcept(?x, Negation), hasPrincipal(?x, ?y) -> hasImportancia(?x, "Secundario")

Figura 4.9. Reglas de Proceso referentes a la Anotación Semántica en la Ontología Clínica.

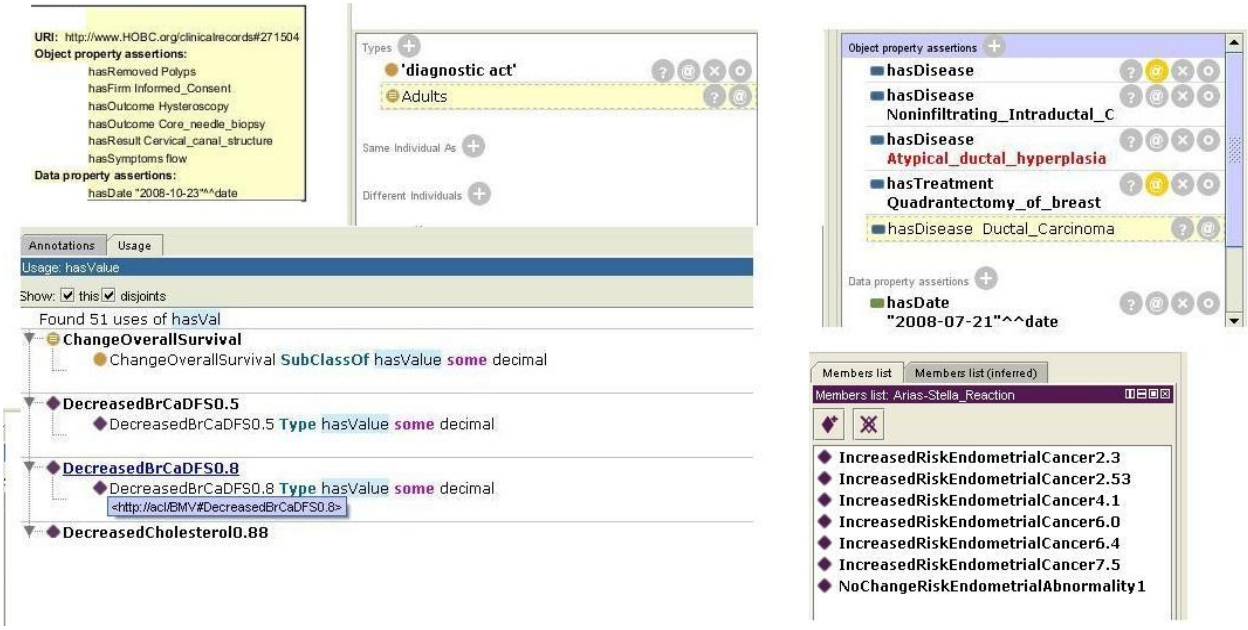


Figura 4.10. Ejemplo de Registro Clínico en la Ontología Médica del Trabajo.

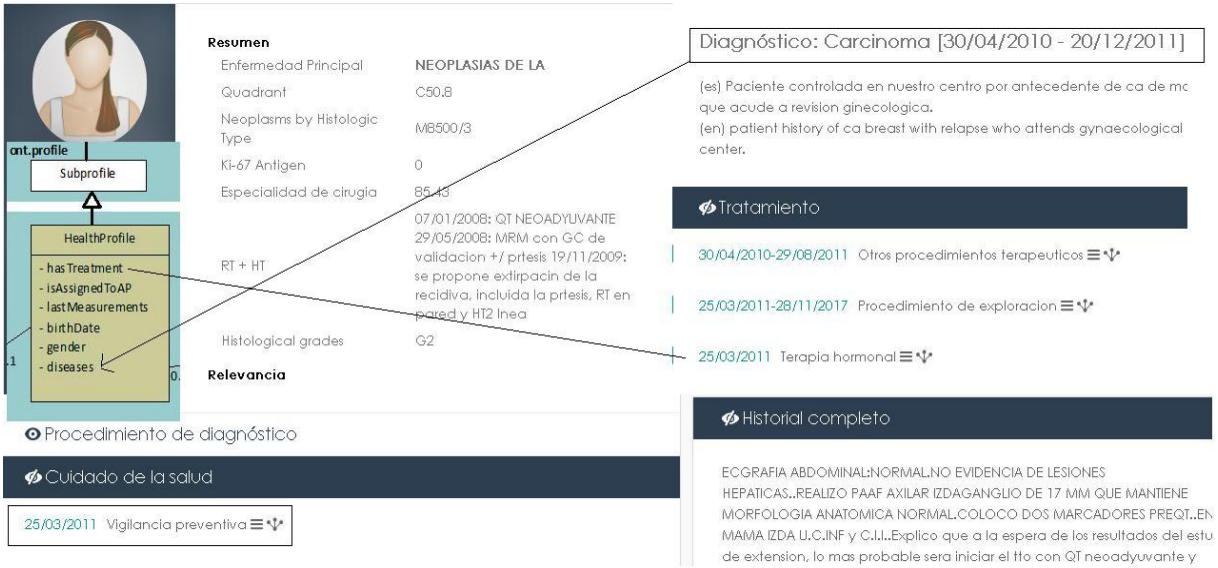


Figura 4.11. Integración ente la Ontología Médica y la Ontología de Teleasistencia.

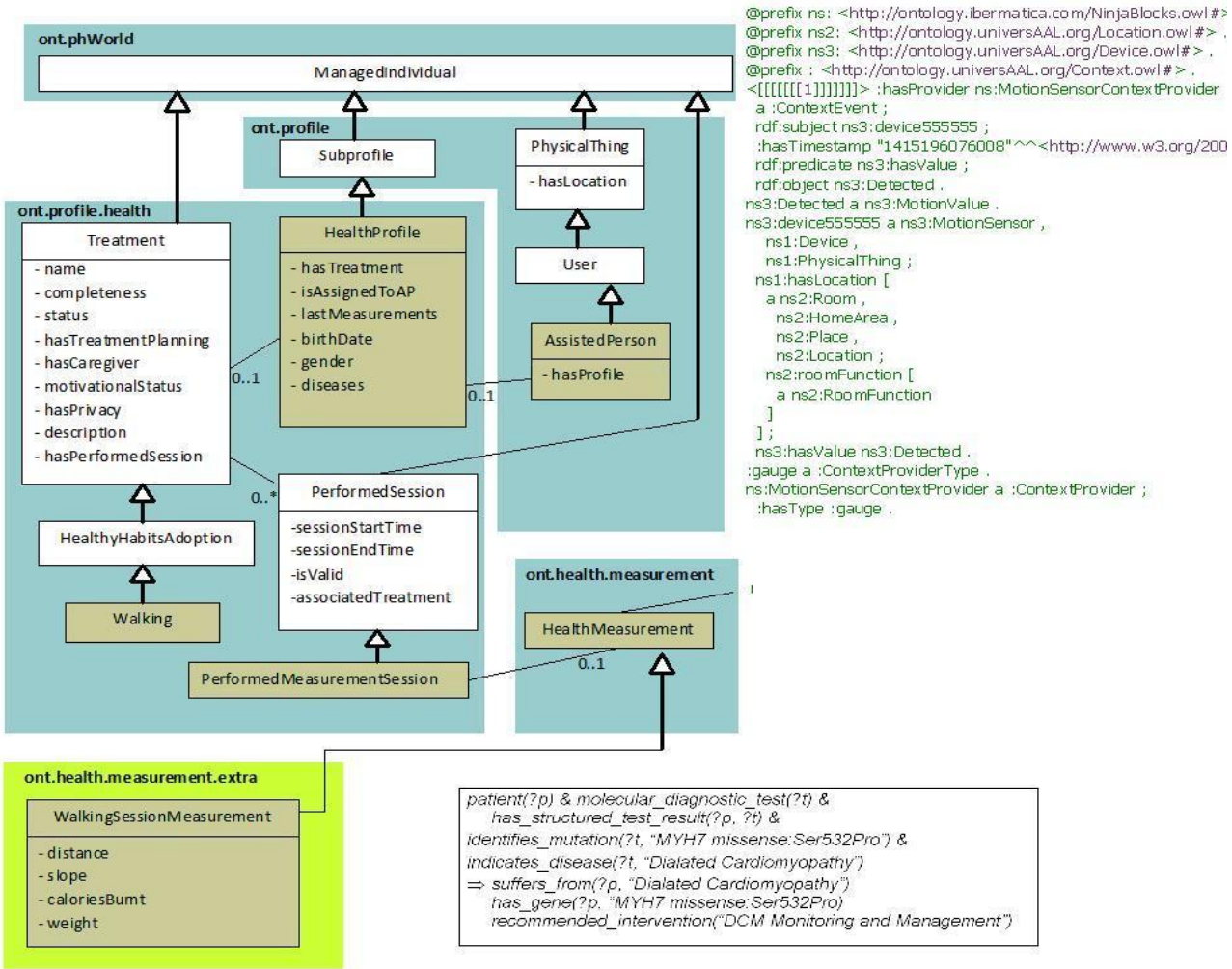


Figura 4.12. Fusión, Arquitectura, reglas y consultas en la Ontología Unificada del Trabajo.

4.4.2 Sistema Experto: Sistema basado en reglas y Detección Automática de Patrones.

El Sistema Experto asume dos funciones diferenciales:

a) ***Módulo de Reglas Heurísticas:*** Este primer módulo permite incorporar el conocimiento de expertos en materia de Teleasistencia y recomendaciones clínicas en forma de reglas deterministas introducidas manualmente. Los enfoques basados en reglas son adecuados para los servicios de contexto altamente dinámicos [Dan07]. EL objetivo es integrar y ejecutar una serie de “Reglas de Expertos” que gestionen situaciones comprometidas sobre situaciones anómalas, incidencias o alertas preventivas en el periodo de tiempo en el que no existe un histórico previo que permita automatizar este tipo de modelos. Dado que estas reglas son fáciles de adaptar, alterar y mantener, esta característica las convierte en una solución atractiva para los expertos. El cuidador es capaz de definir y modificar directamente sobre la plataforma las reglas que especifican el comportamiento de un sistema en una situación dada de una forma adaptable y usable [Ald11].

b) ***Módulo de Detección Automática de Patrones***: Además de integrar las “Reglas Heurísticas” anteriormente descritas, el Sistema Experto también es el encargado de extractar, analizar y “interiorizar” patrones sobre los hábitos de vida personales de los usuarios, a partir de un almacén ya consolidado de eventos. Esta segunda funcionalidad, trata de generar, complementar e incluso sustituir las reglas heurísticas por reglas autogeneradas a partir de los históricos personalizados de cada paciente.

La conjunción de las dos funcionalidades unidas en el mismo sistema soporta el objetivo principal del Sistema Experto*:* ejecutar una serie de reglas de negocio con el objetivo de sugerir recomendaciones de salud o enviar alertas sobre situaciones anormales sobre los comportamientos del anciano, basadas en conjunto de "reglas heurísticas", complementadas por un conjunto de modelos o reglas generados automáticamente por el sistema de Detección Automática de Patrones.

4.4.3 Sistema de Envío de Notificaciones

Una notificación “Push” es un tipo de comunicación entre un dispositivo cliente y un servidor en el que es este último es el que inicia la petición, es decir, el servidor notifica al dispositivo cliente sobre algún evento sin que el usuario final tenga que realizar acción alguna. En este trabajo el Sistema de Envío de Notificaciones cubre la necesidad de enviar a los dispositivos móviles de los familiares la información sobre una incidencia ocurrida en el hogar. Las notificaciones Push tienen como ventaja, frente a la técnica del “polling” (peticiones periódicas al servidor para averiguar si hay nuevos eventos pendientes de notificar), que consume menos recursos y las notificaciones llegan al instante y no al cabo de un determinado periodo de tiempo. Para implementar esta funcionalidad, se ha utilizado el estándar Advanced Message Queuing Protocol (AMQP) sobre la plataforma

RabbitMQ[27]. RabbitMQ está liberado bajo la licencia Mozilla Public License, lo cual permite una mayor flexibilidad y control sobre el producto. RabbitMQ utiliza dos componentes principales:

- El servidor RabbitMQ: que se comportara como el bróker del servicio de mensajería creando una cola de mensajes a la que deberán suscribirse los distintos dispositivos
- Aplicación Android: que será la aplicación cliente que se conectará con el bróker para recibir las notificaciones, lo hará mediante la implementación de un servicio Android que mantendrá activa la conexión al servicio de notificaciones en segundo plano.

[27] https://www.rabbitmq.com/

Capítulo 5

Sensorización, Redes Inalámbricas y hardware en el domicilio

El hardware que desplegado en el sistema, sito en el hogar de los usuarios, debe permitir la captura de los hábitos y actividades con cierta probabilidad de peligrosidad del usuario en el hogar, para lo que se utilizarán sensores de movimiento (ver ejemplo de instalación en la Figura 5.1). Dentro de este apartado, y en relación al hardware necesario para la captura de la información, relacionada con la captura de los hábitos de los residentes que debe instalarse en el hogar del anciano, existen múltiples alternativas en el mercado. No obstante como el objetivo del proyecto en el sistema está enfocado a la creación de un sistema de bajo coste, se ha adquirido únicamente aquellas soluciones de bajo coste que cumplen los siguientes requisitos funcionales:

- Poca necesidad de mantenimiento
- Bajo coste
- Bajo consumo
- En el caso de los sensores, se han seleccionado aquellos que no requieren instalación compleja, inalámbricos, y que funcionan a pilas y/o baterías.

Figura 5.1 Configuración de sensores en la entrada de la vivienda

Se hizo un análisis preliminar del estado del arte en Sensórica y Redes Inalámbricas, detallado en el Capítulo 2, y en base a los requerimientos necesarios, (precio muy ajustado, bajos requisitos en cuanto a las necesidades de configuración y de mantenimiento), se descartaron tanto las tecnologías wifi como bluetooth, ya que, aunque dan unos requerimientos muy buenos en cuanto a la capacidad de transmitir la información, su complejidad en la instalación y mantenimiento, unido al coste, las hacen inviables para su uso en el contexto del sistema. Es por ello que se ha decidido implantar sensores con redes inalámbricas basadas en soluciones tanto de ZigBee, como las relacionadas con las soluciones open source de IOT, que son las que mejor se ajustan a las necesidades de en el sistema (RaspBerry).

5.1 Hardware del Componente de Captación de Información.

El sistema desarrollado contiene tres tipos de sensores: sensores ambientales, fisiológicos, integrados y audiovisuales, conectados a uno o más “hubs” de recepción de datos. El servicio de monitoreo de seguridad incluye diferentes conjuntos de dispositivos:

- ***Sensores corporales***: el sistema se ha probado con un cinturón que mide una serie de signos vitales como respiraciones por minuto, conductividad eléctrica de la piel o posición en los ejes x, aceleración relativa, frecuencia cardiaca y temperatura. Para conectar estos dispositivos con el sistema central se utiliza una conexión directa ECG Bluetooth con la pasarela (FTTH OSGi Gateway). En casos de uso real, se ha utilizado un sensor “Zephyr Bio Harness” (ver figura 5.2) que permite medir la frecuencia cardíaca, la frecuencia respiratoria, la temperatura de la piel, la posición / postura, la actividad medida en VMU, la aceleración en los tres ejes a 16G, el nivel de conductividad de la piel. La información que suministra es almacenada por el sensor y transmitida a la puerta de enlace cuando hay una conexión Bluetooth. Se utilizó en las primeras fases en las pruebas de laboratorio, pero después, en el despliegue final en domicilios, se desechó por ser incómodo para los usuarios, y por ser un elemento que encarece el producto final.

Figura 5.2 Utilización del sensor Zephyr Bio Harness

- ***Sensores ambientales simples***: Estos sensores están conectados a un Hub conectado a la pasarela vía RF433. Son los siguientes:
    - ***Sensores de Presencia PIR***: envían una señal por radiofrecuencia 433 MHz cuando detecta movimiento o presencia en la casa. Esta señal es recogida por la RaspBerry.
    - ***Sensores de Puertas*:** envían una señal por radiofrecuencia 433 MHz cuando la puerta se abre. Esta señal es recogida por la RaspBerry.
    - ***Sensores de Humo/Gas:*** envían una señal por radiofrecuencia 433 MHz
- ***Sensores ambientales avanzados***:
    - ***Sensor de Video-Análisis***: Adicionalmente, el sistema incluye sensores avanzados de vídeo para el análisis del reconocimiento facial [Mor09], situaciones de caídas por posición, análisis de las expresiones faciales. La cámara está conectada a la RaspBerry por el puerto de la cámara (en nuestro caso un puerto USB) y se utiliza para capturar imágenes y vídeos del entorno y analizarlas en local, es decir, será la propia RaspBerry la que gestione la posibilidad de lanzar una alerta analizando las imágenes y de mandar sólo la información de la incidencia pero sin llegar a enviar, en ningún caso, imágenes/capturas ni vídeos al exterior. Por lo tanto, una vez conectada la cámara a la RaspBerry y desplegado el desarrollo realizado mediante Software Libre OpenCV, tendremos un sistema autónomo de detección de situaciones peligrosas o anómalas en tiempo real. La cámara funciona básicamente como un sistema de detección de presencia en zonas, y de tracking de personas. Si la cámara detecta presencia continuada de una persona en ciertas zonas parametrizadas como de alerta, se lanza un aviso al Sistema Experto, que gestiona dicho evento.
    - ***Micrófono de Detección de ruido ambiental***: La RaspBerry lleva acoplado un micrófono que es capaz de analizar el ruido ambiental, y detectar peticiones de auxilio, o gritos, en base a una caracterización del ruido ambiental.
    - **Sensor de ocupación de cama.** Este sensor se coloca sobre el colchón de la cama, por debajo de la sábana bajera y es el encargado de enviar una señal indicando si la persona está o no en la cama. Al igual que en los casos anteriores, esta señal también es recogida por la RaspBerry.

Estos sensores avanzados, al igual que en el caso de los sensores corporales, se han utilizado solamente en las pruebas de laboratorio, pero no se han incorporado en el despliegue en los domicilios reales, debido principalmente al coste de su implementación.

Adicionalmente, el sistema incluye efectores de ayuda a los ancianos, que son activados a través del dispositivo central (Raspberry), bajo comandos enviados por el Sistema Experto. Estos efectores son los siguientes:

- **Luces LED**

Estas luces van colocadas en la habitación, bien a la altura del rodapié o a lo largo de la parte baja de la propia cama. Se manejan desde la RaspBerry mediante el puerto GPIO.

- **Luces X10**

Estas luces también se manejan desde la RaspBerry pero esta vez por medio del protocolo X10. Para ello la RaspBerry cuenta con un emisor de señales X10 para mandar las órdenes requeridas a las luces.

Este emisor, está conectado a la RaspBerry mediante un puerto USB al cual se le da las órdenes de encender o apagar las luces y éste enviará dichas órdenes a través de la electricidad a los actuadores X10.Las órdenes enviadas por el emisor X10 son recibidas por unos aparatos llamados actuadores X10, que serán los que ejecuten las instrucciones emitidas. Para instalar los actuadores X10 hay que hacer una pequeña obra y manipular ligeramente el sistema eléctrico del hogar.

- **Botones bluetooth**

Los botones están emparejados a la RaspBerry mediante BlueTooth y se encarga de enviarle ciertas acciones configuradas (haciendo clic o doble clic).

Estos botones debido a su ligereza y largo alcance (aproximadamente 10 metros), permiten a los usuarios llevarlos consigo cuando lo requieran, y tenerlos al alcance sin molestar con el resto del mobiliario. Por ejemplo pueden dejar el que controla las luces en la mesilla, y los de las alertas uno en cada estancia implicada en el proceso (mesilla del dormitorio, pasillo y baño).

Dependiendo del botón que se pulse se llevará a cabo una tarea u otra:

- Botón amarillo: pulsando una vez (1 clic) se envía la orden a la RaspBerry de encender las luces, tanto LED como de tipo X10, implicadas en el proceso, es decir los LED del dormitorio y las luces tanto del pasillo como del baño. Pulsando dos veces seguidas (doble clic) se envía la orden de apagarlas.
- Resto de botones: pulsando una vez (1 clic) se envía la orden de lanzar una alerta instantánea para que el sistema avise a los cuidadores correspondientes. Pulsando dos veces seguidas (doble clic) el usuario puede desactivar las alertas que estén activas en su sistema.

5.2 Hardware del Componente de Gestión Local.

El Componente de Gestión Local (ver Capítulo 4), está íntegramente embebido en un Microcontrolador RaspBerry. Tiene varias funciones, entre las que se encuentra la de ser el encargado de recoger los valores de los diferentes sensores (movimiento, puerta y ocupación cama) desplegados por el hogar. Los sensores de movimiento y puerta se recogen mediante un receptor de radiofrecuencia 433 MHz. Todas las lecturas recibidas se van guardando en una base de datos NoSQL (Redis), para su posterior envío al Servidor Cloud mediante una serie de servicios web WCF (Windows Communication Foundation). Otra de sus funciones es la de manejar ciertas luces de casa (LED y/o x10). El manejo de las luces de tipo LED se realiza mediante una interfaz específica, y mediante el protocolo X10 se controlan las luces asociadas a los conectores X10. También es el encargado de analizar las imágenes obtenidas a través de la cámara conectada a la RaspBerry. La conexión de la cámara se realiza a través del puerto USB. En cuanto a las imágenes, se procesan directamente en local, en la propia RaspBerry, y a partir de ese análisis particular, se decide si existe o no una

situación de alerta. Por último, el microcontrolador es también el encargado de recoger la actividad de una serie de botones (una pulsación simple, pulsación doble o mantener el botón pulsado), configurables para realizar acciones (activar los efectores o realizar una notificación al sistema Cloud). Las características del Componente de Gestión Local instalado en todos los domicilios son las siguientes:

- SoC Broadcom BCM2835 (CPU + GPU + DSP + SDRAM + puerto USB) CPU ARM1176JZF-S a 700 MHz (familia ARM11)
- GPU Broadcom VideoCore IV, OpenGL ES 2.0, -2 y VC-1 (con licencia), 1080p30 H.264/MPEG-4 AVC
- Memoria (SDRAM) 512 MB (compartidos con la GPU) Puertos USB 2.0 2 (vía hub USB integrado)
- Entradas de vídeo Conector [MIPI] CSI que permite instalar un módulo de cámara desarrollado por la RPF
- Salidas de vídeo Conector RCA (PAL y NTSC), HDMI (rev1.3 y 1.4), Interfaz DSI para panel LCD
- Salidas de audio Conector de 3.5 mm, HDMI
- Almacenamiento integrado SD / MMC / ranura para SDI
- Conectividad de red 10/100 Ethernet (RJ-45) vía hub USB
- Periféricos de bajo nivel: 8 x GPIO, SPI, IC, UART Consumo energético 700 mA, (3.5 W)
- Fuente de alimentación 5 V vía Micro USB o GPIO header Dimensiones: 85.60mm x 53.98mm

Capítulo 6

Módulo de Detección Automática de Patrones

Este capítulo introduce el concepto de “Detección Automática de Patrones”, y su implantación, en sus dos vertientes: la predicción de intencionalidad, y la detección de anomalías. En ambos casos, se justifica su implantación como módulos del Sistema Experto.

6.1 Justificación

Los sistemas de detección de alertas en el hogar actuales, tal y como se ha detallado en el Capítulo 1, se basan, en general, en dispositivos que la persona ha de llevar consigo permanentemente, usualmente invasivos o incómodos. Además, son ineficaces en el sentido de que, o son reactivos (se debe pulsar un botón cuando algo ocurre), o se debe llevar encima constantemente, lo que hace que en muchas ocasiones sean rechazados. En este escenario, este trabajo tiene como objetivo resolver los problemas anteriores por medio del denominado “***Módulo de Detección Automática de Patrones***”, que permite el seguimiento de la actividad de los ancianos en el hogar con el único objetivo de evitar posibles accidentes, incidentes en base, principalmente, a la detección de cambios de conductas en sus patrones de uso habituales. Por ello, la funcionalidad deseada a resolver mediante algoritmia computacional se resume como la implantación, en entornos reales, de un conjunto de servicios que permitan una vigilancia inteligente de los usuarios en sus domicilios, de forma que el sistema se adapte a cada usuario, creando automáticamente modelos que determinen su pauta general, evolucionándolos diariamente. La generación de estos modelos en base a un aprendizaje automático, por un lado, y la detección de anomalías automáticas por otro, en un sistema de entrenamiento diario, se aplican sobre los nuevos datos de entrada, con una frecuencia cuarto horaria, buscando detectar cambios en los patrones que indiquen una disminución de atención, agilidad, rendimiento, o alguna situación de riesgo. A futuro, estos cambios de comportamientos pueden denotar, además, un indicio sobre el incremento de la gravedad en ciertos trastornos, que sin ser evidencias, pueden ser “indicaciones” para los clínicos de suma importancia. La Detección Automática de Patrones se configura como un módulo dentro del Sistema Experto, explicado en el Capítulo 4, y que funciona como complemento al Módulo de Reglas Heurísticas del Sistema. A su vez, el Módulo de Detección Automática de Patrones se compone de los dos submódulos siguientes:

1. ***Sistema de Predicción de Intencionalidad:*** Este sistema predice cuál es la situación teórica del anciano en función de una determinada situación, de una forma supervisada.
2. ***Sistema de Detección de Anomalías:*** Este sistema detecta anomalías en los comportamientos del anciano en casa, de una forma no supervisada.

6.2 Sistema de Predicción de Intencionalidad

Definir y "comprender" los modelos de comportamiento e interacción de los usuarios de forma personalizada, y dinámica, en función del contexto externo, es un desafío clave cuando se considera el problema de predecir la intencionalidad. La capacidad de predecir las actividades del residente basándonos en su actividad histórica habitual es el objetivo del Sistema de Predicción de Intencionalidad. El Sistema se modula como un sistema automático de clasificación supervisada, que con la suficiente profundidad de históricos, irá complementando la generalidad de las Reglas Heurísticas con predicciones más personalizadas. Esta arquitectura permite cumplir las expectativas demandas en los sistemas de teleasistencia domiciliaria en dos puntos:

- Detectar situaciones no previstas (no registradas en la heurística), en un modelo *predictivo*.
- Detectar comportamientos similares de usuarios afines por actividad o perfiles, como punto de mejora futura de cara a los procesos asistenciales, no sólo en la calidad sino también en la optimización de de sus procesos de servicio.

En general, después de la instalación de la sensórica en cada domicilio, el Sistema Experto ejecuta las siguientes tareas:

- Paso 1: Anotación de los comportamientos observables del usuario.
- Paso 2: Ejecución las Reglas Heurísticas para detectar riesgos deterministas, y almacenar los resultados en el histórico de eventos.
- Paso 3: Iniciar el proceso de aprendizaje automático, diario, en base a los históricos.
- Paso 4: Aplicar el modelo de comportamiento aprendido en el paso previo, sobre los datos recogidos en el Paso 1, siempre que las reglas generadas por los modelos supervisados superen una confianza y un soporte parametrizado previamente por domicilio.
- Paso 5: Comparar las predicciones con el estado real, y determinar si existe una probabilidad de alerta.
- Paso 6: En el caso de probabilidad de alerta, verificarla manualmente, y almacenar los resultados en el histórico de eventos, tanto si la alerta era positiva como si era negativa.
- Paso 7, se vuelve a comenzar desde el Paso 1.

Una de las maneras más sencillas de detectar desviaciones en el comportamiento es comparando la el estado del usuario en un momento determinado contra una predicción teórica de este estado en ese mismo momento. Pero para hacer predicciones correctamente, no sólo se tiene que considerar los datos propios del usuario y su interacción con el ambiente más cercano (con los sensores ambientales, por ejemplo), sino también es recomendable, como se demuestra en el capítulo 7 de Implementación, la necesidad de agregar información

externa, como el clima exterior, o el estado meteorológico. Parece plausible intuir que, por ejemplo, si está lloviendo, aunque sea verano, existe mayor probabilidad de que un anciano no salga al aire libre a menos que le guste caminar bajo la lluvia, o tenga una necesidad u obligación ineludible. Por lo tanto, parece importante que éstas observaciones externas deben ser contempladas como parte de los modelos y que el Sistema de Predicción de Intencionalidad sea capaz de incorporar este tipo de entradas. Al ser un sistema de clasificación supervisado, los modelos que los componen necesitan un "campo objetivo" a modelizar, y un histórico de "hechos conocidos" que desembocan en dicho "campo objetivo". En nuestro caso, el campo objetivo es el estado personalizado de cada anciano, y los "hechos conocidos" son indicadores tales como el espacio en dónde se encuentra el anciano, su ubicación temporal, estado meteorológico, etc… En este trabajo se han realizado distintas aproximaciones a este problema, comenzando con unas primeras de laboratorio, con un sólo domicilio, y evolucionando las soluciones hasta su despliegue en los domicilios reales. Tal y cómo se explica en el capítulo 7 de Implementación, finalmente se ha optado por el desarrollo de un sistema basado en una ventana deslizante de eventos segmentados por horas, a los que se aplica un árbol de decisión que predice la próxima acción del usuario en el segmento horario a analizar. Para cada ventana de tiempo, el árbol de decisión toma en cuenta un conjunto multivariado de valores que generan un modelo predictivo y extrae el siguiente estado del usuario, con ciertos niveles de confianza y probabilidad.

6.3 Sistema de Detección de Anomalías

Como se ha descrito en el Estado del Arte, en este trabajo, a la hora de detectar patrones anómalos, y como complemento al Sistema de Predicción de Intenciones, se integra también la herramienta "AD" (Anomaly Detection). Es una técnica no supervisada, que permite descubrir anomalías en un subconjunto de datos marcando aquellos eventos cuyo comportamiento es muy diferente al resto de datos, o que contienen un patrón desconocido que no ha sido previamente anotado. De esta forma, además de prever comportamientos formalizados o codificados y analizar la "adherencia" del anciano a dichos comportamientos, también chequeamos si existen otros patrones desconocidos que no "cuadran" con el conjunto de datos general, como, por ejemplo, presencia simultánea de varias personas en el domicilio, ausencia de información o señales, tiempos "extraños" en estancias "a priori" determinadas como correctas por el Sistema de Predicción de Intenciones, o cualquier otra circunstancia que se salga de la normalidad, sin necesidad de tener codificada la razón. Para su implantación (ver capítulo 7), se ha decidido utilizar, de los distintos tipos de algoritmos existentes en el estado del arte, el algoritmo LOF. Se aplica el algoritmo a un conjunto de eventos y se chequea el nivel de anomalía de cada evento con respecto a sus vecinos, de cara a determinar de forma automática si realmente dicho evento tiene una cierta probabilidad a ser una anomalía. En el caso de que sea una verdadera anomalía, esta se comprueba de forma manual, y en caso afirmativo, se lanza un nivel de advertencia desde el Sistema Experto a los servicios asistenciales a través del Sistema de Notificaciones. Es importante destacar que el análisis de anomalías se realiza tanto para los datos históricos como para las predicciones generadas por el Sistema de Predicción de Intencionalidad, sustituyendo los valores teóricos en el futuro como si fueran los reales. De esta forma, podemos analizar si va a existir anomalías en base a las predicciones de estados futuros de una forma desasistida.

Capítulo 7
Implementación del Sistema Experto

En este capítulo tratamos brevemente la implementación del sistema y las distintas fases por las que ha evolucionado hasta el despliegue final del mismo. La sección 7.1 describe una implementación piloto que sirvió para configurar el sistema final. La sección 7.2 describe las pruebas iniciales. La sección 7.3 describe pruebas de laboratorio. La sección 7.4 discute la prueba en ambientes reales. La sección 7.5 proporciona consideraciones éticas sobre el sistema. La sección 7.6 presenta la gestión de incidencias planificada. La sección 7.7 describe las actividades de evaluación que dieron paso a la recogida sistemática de datos que sirven de base para los resultados expuestos en el Capítulo 8.

7.1 Implementación piloto en domicilios controlados.

Al inicio del trabajo, incluso antes de comenzar a realizarse los primeros modelos analíticos, se instaló un sistema piloto centrado en la sensórica y en las primeras pruebas respecto a la capacidad del Componente de Captación de Información, en ciertos domicilios controlados, con el objetivo de conocer las opiniones de los usuarios al respecto de usabilidad y utilidad de su uso, de primera mano. Además, sirvió también de chequeo para analizar la viabilidad del uso del Sistema Local de cara a su futura implantación en un entorno de producción real. Se implantaron, en un primer momento, los siguientes tipos de sensores:

- ***Sensores de Presencia:*** Son sensores que se activan cuando detectan presencia en sus inmediaciones. Son sensores muy sencillos, comúnmente utilizados en comercios, accesos para abrir puertas de forma automática, etc…
- ***Sensores de apertura de puertas:*** Al igual que los anteriores, son sensores muy sencillos que se colocan en las puertas, principalmente en las de entrada y salida de los domicilios, y que simplemente indican si la puerta se ha abierto o cerrado.
- ***Sensores de Temperatura:*** Indican la temperatura de las habitaciones, pensados para su instalación a la hora de conocer si hay ambientes fríos o calientes, el estado de la calefacción, etc…
- ***Sensores de humo:*** Instalados en la cocina, principalmente para atender alertas de incendios.
- ***Sensores de humedad:*** Instalados en la cocina, pensados para detectar inundaciones.
- ***Sensores biomédicos:*** Utilizando bluetooth, en concreto era un cinturón pensado para medir los valores físicos como frecuencia respiratoria, frecuencia cardiaca, temperatura corporal, posición.
- ***Cámaras de visión artificial:*** Instalados en las habitaciones, servirían para identificar a las personas que están en el domicilio, y analizar estados, si están tumbados (posibles caídas), o en movimiento, etc…

Después de unas semanas de seguimiento, y en base a diferentes cuestionarios suministrados a los usuarios y cuidadores, se obtuvieron las siguientes conclusiones que sirvieron de base para los requerimientos iniciales del sistema:

- Los sensores biomédicos, como el cinturón, es un impedimento para los hábitos de vida independiente usual de los usuarios, molesto incómodo, y que pocas veces se "acordaban" de ponérselo. Actualmente existen diversas pulseras que pueden suplir el cinturón, no en todos los aspectos, y que se pueden incluir en el sistema. Sin embargo, existe cierto rechazo a este tipo de sensórica intrusiva.
- La instalación de cámaras incomoda a los usuarios, con dudas sobre la invasión de intimidad, a pesar de que las imágenes son procesadas siempre en local, y nunca se transmiten fuera del domicilio. Además, encarecen la instalación, por lo que, después de los primeras pruebas, se desecharon como parte de la sensórica del sistema.

De esta forma, finalmente, y gracias a este estudio preliminar, en las instalaciones reales realizadas, sólo se han ubicado en los domicilios sensores no invasivos (un sensor de presencia por cada habitación y cuartos de baños, un sensor de apertura de puertas en la entrada principal, un sensor de humo y otro de temperatura). En total, se han instalado entre 7 y 12 sensores por domicilio:

- Los sensores de Presencia (PIR)
- Los sensores de contacto de puertas
- Los sensores de humo
- Sensores de gas

Los costes por casa, en total, suman 500 € de coste, incluyendo los sensores, y el agregador de datos. En cuanto a la seguridad, la instalación usa protocolos SSL para cifrar las credenciales de autenticación y evitar la suplantación de identidad.

7.2 Pruebas preliminares en laboratorio.

Antes de la implantación en entornos reales del sistema completo, fueron obligatorias la verificación técnica y las pruebas preliminares íntegras en un entorno de laboratorio. Las pruebas de laboratorio tuvieron lugar durante los meses de noviembre y diciembre del 2014 en un "Living Lab"[28] o entorno controlado, y fueron supervisadas en H-ENEA, que es la iniciativa H-ENEA Living Lab ACEDE (Cluster Home), un espacio participativo de experimentación y validación conjunta con los usuarios. El equipo H-ENEA está formado por un equipo multidisciplinario (Antropología, Sociología y Diseño e Innovación Aplicadas), que utilizan metodologías ágiles para la innovación más cercanas a las necesidades de los clientes y los resultados de los usuarios. A partir de estas pruebas, tras obtener resultados satisfactorios, se definió un Plan de Implantación en domicilios reales. El perfil de los usuarios seleccionado fue el de personas jubiladas que no eran dependientes, pero que necesitan chequeos médicos regularmente, y con una necesidad adicional de estar en contacto con alguien obligatoriamente todos los días, porque en su franja de edad (entre 65 y 78), hay riesgo de que aparezcan los

[28] http://h-enea.org/es/living-lab/

primeros signos de empeoramiento cognitivo. Los usuarios seleccionados estaban especialmente motivados para hacer los ensayos y reportar sus impresiones y objeciones sobre la plataforma, que fueron de enorme utilidad para corregir errores y problemas de interacción del usuario. En la fase de pruebas en laboratorio, se seleccionaron dos personas que respondían a estos perfiles, con su pleno consentimiento informado, y que fueron contratados para probar los servicios en el Living Lab. Siguiendo el procedimiento del Living Lab de H-enea, las personas seleccionadas experimentaron una secuencia de entrevistas, entrenamientos y una etapa de formación para resolver una serie de desafíos en el ambiente controlado. Toda esta experiencia se observó presencialmente, y también, mediante grabaciones de video, siempre con su consentimiento formal, con el fin de extraer los factores clave en su experiencia, que sirvieron de base para una serie de reajustes en el sistema final.

7.3 Implementación y prueba en ambientes reales.

La implementación y prueba en ambientes reales comenzaron en enero del 2015, y duran hasta la actualidad, en más de 50 domicilios. La instalación implicó la instalación del hardware en cada casa. Hubo un paso previo de selección de domicilios en función de su ubicación geográfica, y se dividió el trabajo entre 3 técnicos para hacer esta tarea tan concurrente cómo fue posible. Existen domicilios monitorizados en Bilbao, San Sebastián y Vitoria. La configuración del sistema se realiza el mismo día de la instalación. El sistema está preconfigurado como un sistema plug and play. Si el domicilio tiene conexión a Internet, el Componente de Gestión Local se conecta directamente a la wifi del domicilio; en caso de que no exista conexión a Internet, el propio Sistema Local se configura con una conexión internet de datos con una línea de datos muy sencilla. Después de la configuración del Sistema Local, se instalan los sensores de presencia, puertas, humo/gases y temperaturas a lo largo del domicilio, y se personaliza el registro de cada uno de los sensores en las distintas ubicaciones del domicilio a través de una aplicación móvil creada a tal efecto, de modo que el técnico “in situ” parametriza la instalación, y desde el primer momento, comienza a enviar información a Sistema “Cloud” remoto de forma inmediata. Adicionalmente, las pruebas básicas realizadas para comprobar si el sistema funciona correctamente pueden realizarse de forma remota, o en local, a través de la misma aplicación móvil o a partir de una página web también dedicada a tal efecto. Todas las familias que participan en el proyecto han recibido un curso de capacitación del cuidador formal sobre los beneficios de la plataforma, qué tipo de datos serán recibidos por el sistema y a qué tipo de información puede ser accedida por el usuario. Además, se ha proporcionado al usuario toda la información necesaria para ponerse en contacto con los proveedores en cualquier momento, junto con un documento de Consentimiento Informado, que han sido firmados por todos los usuarios y sus familiares, en donde se detalla todo el procedimiento y los datos de contacto con los proveedores de la atención telefónica o de atención de urgencias.

En cuanto a la plataforma tecnológica, se ha utilizado dos bases de datos no-sql, una, para el almacenamiento de los datos de los sensores, en el “Sistema Local”, soportada en una base de datos “Redis”, y otra, en el Sistema “Cloud”, conformada en una base de datos en formato de grafo (Virtuoso), que sirve de base para almacenamiento de la información en el sistema central. Para el sistema de reglas de inferencia, se ha utilizado la herramienta “Drools” como contenedor de las “Reglas Heurísticas” de los expertos en teleasistencia y

clínicos, y “Pellet” para las “Reglas de Proceso”, en base a la información obtenida directamente de las tripletas semánticas de Virtuoso. Algunos ejemplos de “Reglas Heurísticas” se muestran en la Figura 7.1.

```
// Esta regla es para el caso de no haberse detectado la puerta en las ultimas 24 horas pero esta en casa y hay inactividad
rule 'Caida_No_Activo'
    when
        ed : EventDetection (DiaSem : sDayOfWeek, numDate : EventDate, numTime : EventTime, numTimeTicks : EventDateTimeTicks)
        con : ConfigTiempos ( ConfType == "NoActivity", MinConf : ConfTime )
        not Evento ( SensorTypeName == "EnvDoorContact")
        o : Evento ( SensorTypeName == "EnvPresence", MinutesLastEvent > MinConf )
        not AlarmExceptions ( StartDate <= numDate, EndDate >= numDate)
        not AlarmExceptions ( sDayOfWeek == DiaSem )
        not AlarmExceptions ( StartTimeTicks <= numTimeTicks, EndTimeTicks > numTimeTicks )
   then
        Controller.NotificarAlarma(6, o.endEventDateTime, o.MonitoringPersonId, 6);
        Console.WriteLine("Caida_No_Activo");
end
rule 'Regla_Caida_Activo'
    when
        con : ConfigTiempos (ConfType == "LongActivity", MinConf : ConfTime, Habitac : RoomTipe)
        con2 : ConfigTiempos (ConfType == "NoActivity", MinConf2 : ConfTime)
        o : Evento (SensorTypeName == "EnvPresence", eventRoomType == Habitac, MinutesFirstEvent > MinConf, MinutesLastEvent < MinConf2, ultimaAct : MinutesLastEvent, primerAct
        not Evento (SensorTypeName == "EnvDoorContact", MinutesLastEvent < (ultimaAct + 5))
        not Evento (SensorTypeName == "EnvDoorContact", MinutesLastEvent < (primerAct - 5), MinutesLastEvent > (ultimaAct+5))
    then
        Controller.NotificarAlarma(9, o.startEventDateTime, o.MonitoringPersonId, 7);
        Console.WriteLine("Regla_Caida_Activo");
end
rule 'Caida_Night_Levanta_Cama'
    when
        ed : EventDetection ( DiaSem : sDayOfWeek, numDate : EventDate, numTime : EventTime, numTimeTicks : EventDateTimeTicks )
        con : ConfigTiempos ( ConfType == "NoActivityNight", MinConf : ConfTime )
        o3 : Evento ( SensorTypeName == "EnvDoorContact", HoraPuerta : MinutesFirstEvent )
        o2 : Evento ( SensorTypeName == "EnvPresion", SensorValue == 0 )
        o : Evento ( SensorTypeName == "EnvPresence", eventRoomType == "Dormitorio", EventRoomPrincipal == "1",  MinutesLastEvent < (HoraPuerta-5), MinutesLastEvent > MinConf )
        not AlarmExceptions ( StartDate <= numDate, EndDate >= numDate)
        not AlarmExceptions ( sDayOfWeek == DiaSem)
        a: AlarmExceptions ( StartTimeTicks <= numTimeTicks, EndTimeTicks >= numTimeTicks )

    then
        Controller.NotificarAlarma(24, o.endEventDateTime, o.MonitoringPersonId, 8);
        Console.WriteLine("Caida_Night_Levanta_Cama");
end
```

Figura 7.1 Ejemplos de Reglas Heurísticas

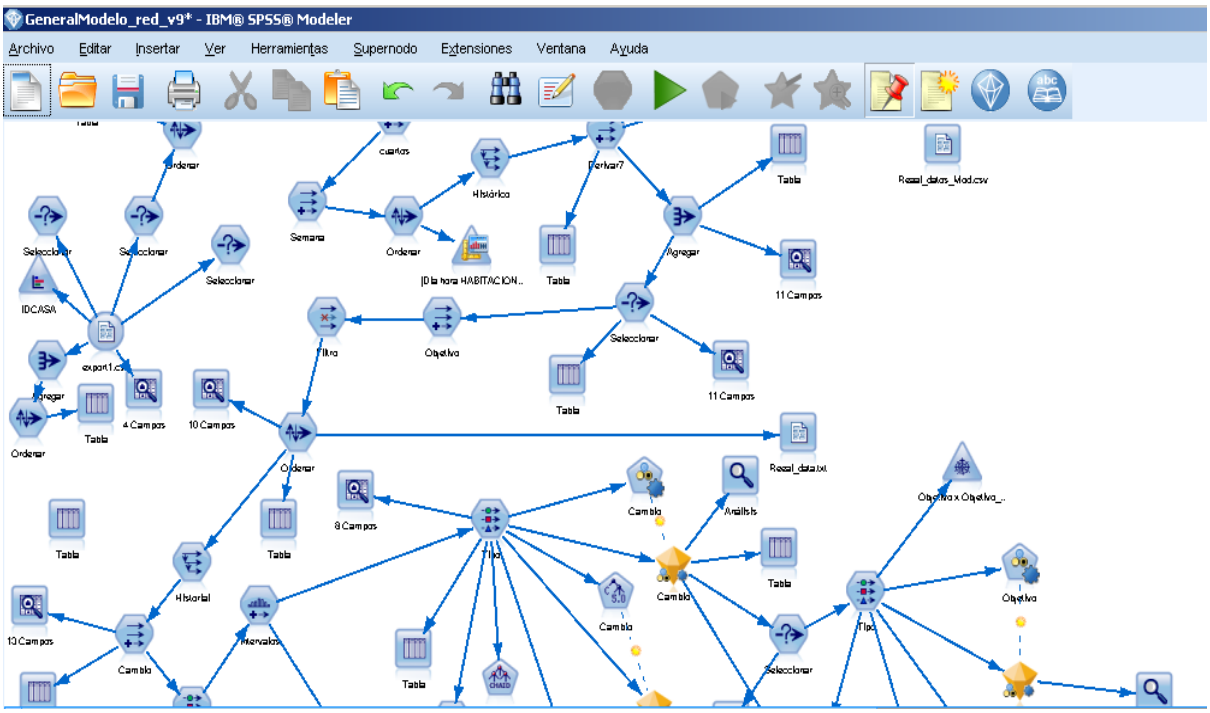


Figura 7.2 Modelización con SPSS

Tanto el sistema de análisis de la calidad de los datos (completitud y consistencia), como los distintos modelados del sistema se han realizado con dos herramientas de análisis predictivo, en concreto, la plataforma de RapidMiner y la plataforma de SPSS Modeler de IBM, con el objetivo de validar los resultados en ambos sistemas. En la figura 7.2 se muestra un ejemplo de una de las rutas utilizadas en la modelización con SPSS.

7.4 Diseño experimental

En el trabajo desarrollado, partimos de una serie de datos bruto, de los sensores, recogidos durante un periodo de varios meses, en varios domicilios, que soportan la información sobre los movimientos de los usuarios a lo largo del domicilio, junto con información de los movimientos de la puerta de entrada/salida del domicilio. Esa es la base de datos del sistema. Con estos datos, el objetivo es determinar cuándo existe una actividad o situación anormal en el sistema con respecto al patrón usual de comportamiento en cada domicilio. Para lograr este objetivo, se han realizados dos tipos de diseños experimentales:

A. ***Análisis con datos semánticamente interpretados:*** Intentar predecir la intencionalidad de los usuarios tomando como base los datos de los sensores, pero interpretados semánticamente y etiquetados como actividades, que serán la base, junto con otros datos agregados, como la climatología del momento, del estudio de intencionalidad.
B. ***Análisis con datos brutos***: Intentar predecir la intencionalidad de los usuarios tomando como base los datos en bruto de los sensores, esto es, solamente las diferentes variaciones de la posición de los usuarios a lo largo del tiempo en cada domicilio.

En ambos casos, la infraestructura utilizada es la misma, sólo que los algoritmos de tratamiento de información en el Sistema “Cloud” son diferentes en un caso que en otro, y los resultados, también varían considerablemente. En este trabajo se han ido evolucionando las distintas estrategias funcionales y técnicas de resolución del problema inicial, la predicción del estado futuro del usuario, en una evolución de procesos, cuyos pasos secuenciales han sido los siguientes:

1. Integración de datos y análisis de la Calidad de los mismos.
2. Transformación y Análisis con datos semánticamente Interpretados
    a. Modelado Supervisado y Selección de la mejor algoritmia a implementar
    b. Incorporación de series de estados al modelado anterior.
3. Análisis con datos no codificados semánticamente
    a. Discretización de eventos en tiempo, duración y frecuencia.
    b. Aproximación estática a la predicción del siguiente estado del usuario.
    c. Aproximación en base a análisis de series temporales.
    d. Aproximación en base a una predicción binaria sobre la probabilidad de que exista o no cambio en la ubicación del usuario.
    e. Clasificador Jerárquico Final.
4. Sistema de Detección de Anomalías.

7.4.1 Integración de datos y análisis de la Calidad de los mismos.

7.4.1.1 Análisis de la Calidad de los Datos en la Anotación de Historiales Clínicos

De cara a incorporar reglas personalizadas en función del estado clínico de cada usuario, se ha generado un complejo sistema de codificación médica semántica, que se está utilizando en este proyecto, y en otros contextos, en fase de validación. Estos contextos son los siguientes:

- La generación automática de Resúmenes Médicos,
- La Triangulación y Segmentación de pacientes, de cara a mejorar el Análisis Diagnóstico
- Análisis y recomendación de tratamientos en la mejora de la Eficiencia Terapéutica
- Análisis y mejora de Procesos Asistenciales.

A partir de este trabajo, se ha generado una plataforma de extracción de conocimiento a partir de historiales clínicos escritos en lenguaje natural, en castellano, en la que se ha recogido información de 85 historiales clínicos, en dos contextos muy concretos, pacientes con cáncer de mama y pacientes con cáncer de colon. Los resultados de dichas anotaciones han sido los siguientes:

- Se han procesado una media de 20.000 términos en dos contextos diferentes, en total, aproximadamente 40.000 términos.
- Dichos términos se codifican en un total de media de 4.500 conceptos gracias a al proceso de desambiguación desarrollado en esta investigación. (Ver en el capítulo 4, el apartado de Creación Automática de Resúmenes Médicos basados en Evolutivos escritos en Lenguaje Natural).

Para la base de la representación de la Ontología Clínica se ha utilizado OWL-DL *(OWL Description Logic)*, un subset de OWL que nos permite personalizar el servicio en función de cada usuario. Entre muchas otras, OWL-DL es compatible con capacidades de razonamiento[29] y además, permite la desambiguación de conceptos [Sow14]. Se ha diseñado con la herramienta Protegé. Para el filtrado semántico de la información anotada (cuáles son los diagnósticos principales, activos, y los tratamientos principales y activos de la Ontología Clínica, ver Capítulo 4), se utilizan reglas semánticas denominadas “Reglas de Proceso” sobre tecnología SWRL [Gar14]. Dada la gran cantidad de información a ser gestionada, necesitamos utilizar técnicas de almacenamiento capaz de gestionar eficientemente grandes volúmenes de datos de una forma plástica y facilitar el acceso y la gestión eficiente de la información (Big Data paradigma) [Mer14]. Por lo tanto, en este proyecto estamos utilizando una base de datos no-SQL, almacenando la información en forma de tripletas semánticas (RDF/OWL) sobre el sistema Virtuoso. Para la medición de la prevalencia estadística del anotador, se han creado una serie de indicadores relacionados con la calidad de los datos (completitud, consistencia, análisis de extremos y atípicos), así como la precisión y la sensibilidad encontradas en la clasificación de términos. Los resultados son los siguientes:

[29] OWL Web Ontology Language Semantics and Abstract Syntax W3C Recommendation 10 February 2004 New Version Available: OWL 2 (Document Status Update, 12 November 2009) https://www.w3.org/TR/owl-semantics/

- Las anotaciones tienen un buen indicador en la calidad global de los datos (un 90% , ver Figura 7.3), excepto en el apartado de atípicos, en dónde tenemos un 69% de datos sin valores atípicos o extremos.

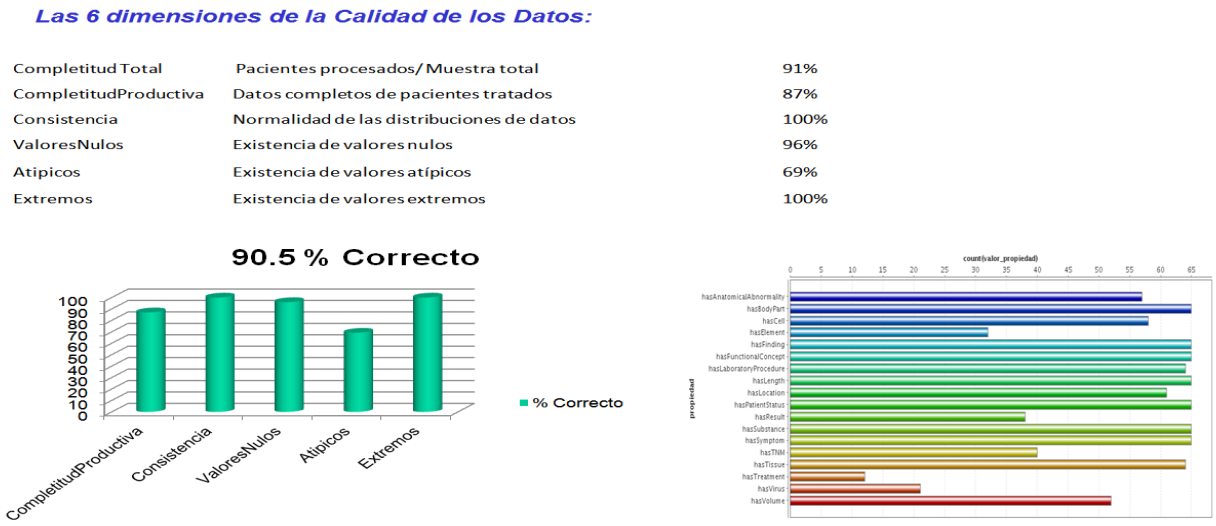


Figura 7.3 Calidad de Datos en las Anotaciones de Historiales Clínicos

En cuanto a la calidad de anotación, comparando los términos anotados semánticamente con los términos extractados estadísticamente, y comprobando ambos resultados, se extrae una precisión global del 90%, con una sensibilidad global del 69%, confianzas similares a los de otros trabajos [Sav10].

7.4.1.2 Análisis de la Calidad de los Datos en la Información de la Sensórica.

Los datos que utilizamos contienen un conjunto de registros con series de eventos obtenidos a partir de la información recogida en los sensores instalados en los domicilios. De 60 domicilios analizados, en total, se tienen datos con una calidad de datos suficientemente válida (sin nulos, datos sin conexiones, errores en la transmisión, etc…), en 29 domicilios. En estos domicilios, los datos de origen tienen la siguiente naturaleza y restricciones:

- En cada hogar habrá una o varias personas viviendo habitualmente, esta información es conocida.
- Los usuarios pueden tener varias visitas, ya sea cuidadores o miembros de la familia.
- Los eventos son todas las detecciones de medidas de sensores y sólo los cambios de estado de los sensores o la medición se guardan en la base de datos, en un período de cinco minutos.
- Si no hay cambios, se registra el mismo estado o medición en la base de datos cada 5 minutos
- En total, se han almacenado 556.972 eventos, guardados entre 2014-10-03 y 2015-12-17 (hoy en día los sistemas están trabajando y recolectando más datos), con una distribución muy diferente entre los hogares analizados. (Ver figura 7.4).

- Es evidente que no todas las señales son relevantes, por ejemplo, en los puntos intermedios, yendo de una habitación a otra, el sensor del pasillo se activará muchas veces, pero con una frecuencia muy pequeña, y por lo tanto, estas señales deben de ser consideradas como no relevantes.
- El archivo de datos de entrada es muy sencillo, consta de tres valores, el identificador de la casa, la fecha, hora y minutos del sensor activado, y el sensor que se ha activado, que corresponde con la presencia de alguna persona en alguna habitación, como se muestra en la figura 7.5.

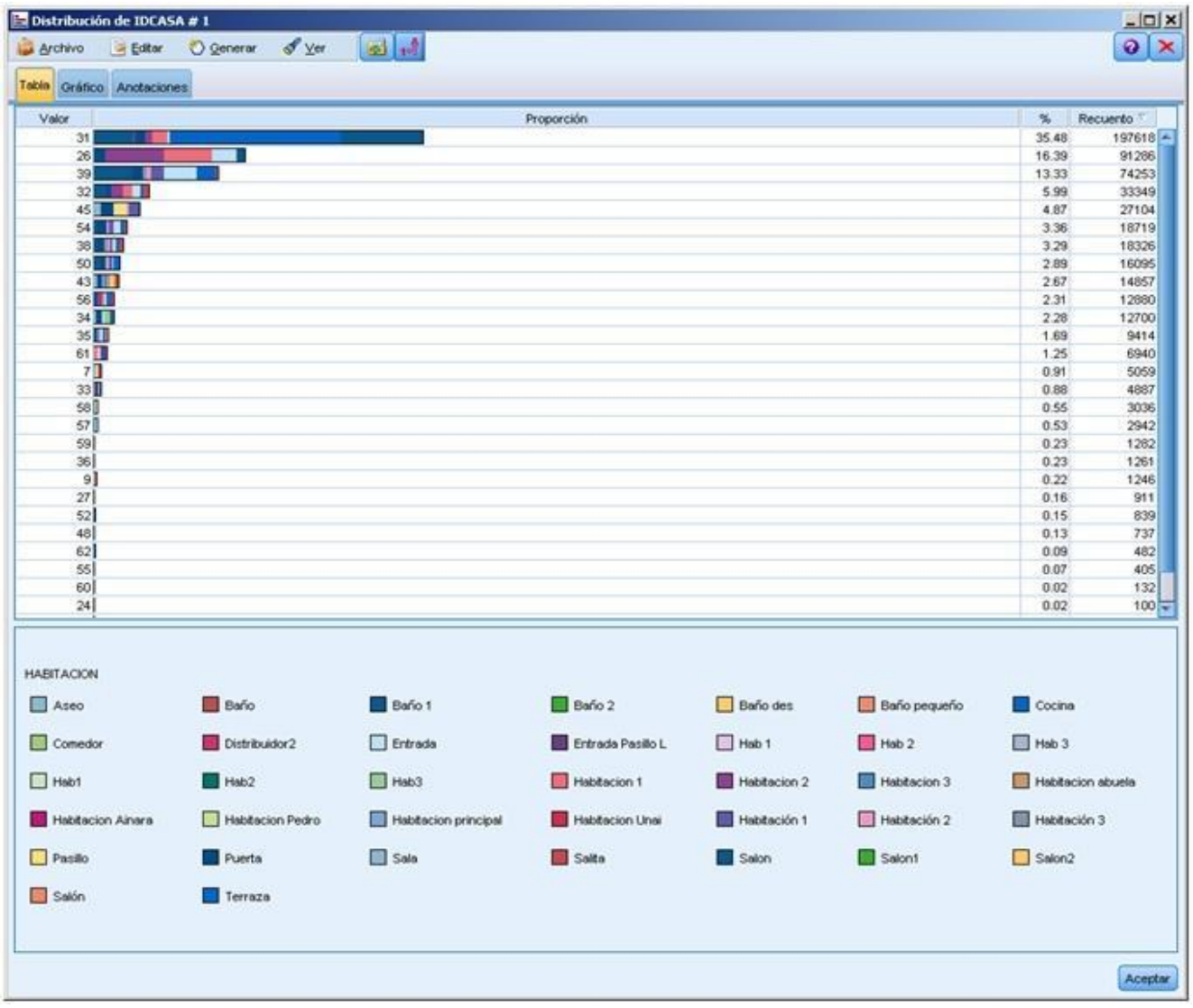


Figura 7.4 Distribución de Eventos por Domicilio

A partir de esta información simple, debemos ser capaces de generar conocimiento de los patrones de usuario.

IDCASA	CP	ESR_DATETIME	HABITACION
31	48003	2015-06-25 13:20:00	Habitacion 2
31	48003	2015-06-25 13:20:43	Puerta
31	48003	2015-06-25 13:22:40	Salita
31	48003	2015-06-25 13:22:44	Entrada
31	48003	2015-06-25 13:23:38	Salon
31	48003	2015-06-25 13:25:44	Cocina
31	48003	2015-06-25 13:29:15	Habitacion 1
31	48003	2015-06-25 13:30:37	Cocina
31	48003	2015-06-25 13:37:33	Salon
31	48003	2015-06-25 13:38:08	Entrada
31	48003	2015-06-25 13:39:42	Puerta
31	48003	2015-06-25 13:40:07	Puerta
31	48003	2015-06-25 13:40:30	Habitacion 1
31	48003	2015-06-25 13:40:36	Entrada
31	48003	2015-06-25 13:41:19	Puerta
31	48003	2015-06-25 13:41:24	Habitacion 1
31	48003	2015-06-25 13:41:31	Salita
31	48003	2015-06-25 13:42:08	Habitacion 2
31	48003	2015-06-25 13:46:54	Entrada
31	48003	2015-06-25 13:47:04	Habitacion 1
31	48003	2015-06-25 13:48:07	Entrada
31	48003	2015-06-25 13:49:14	Habitacion 1
31	48003	2015-06-25 13:49:19	Puerta
31	48003	2015-06-25 13:49:19	Entrada
31	48003	2015-06-25 13:50:10	Habitacion 2
31	48003	2015-06-25 13:53:00	Salon
31	48003	2015-06-25 13:53:06	Puerta

Figura 7.5: Ejemplo de datos de entrada al sistema

7.4.2 Transformación y Análisis con datos semánticamente Interpretados

La aproximación realizada parte de una primera transformación de los datos de los "sensores" en una serie de actividades determinadas. Este proceso tiene la ventaja de que esta codificación se abstrae de la morfología del domicilio, y es la misma para todos los casos de estudio, por lo que el "vector" de actividades resultantes es el mismo en todos los estudios, y permite una normalización del comportamiento de los residentes, lo que facilita estudios posteriores. Así, este modelado se basa, según la literatura analizada, en "segmentar" o "traducir" los datos de los sensores en etiquetas que semánticamente nos suministran más información sobre los eventos o acciones que está realizando el usuario en el domicilio. En esta primera aproximación del problema, se formalizan los datos, en base a una serie de "Reglas de Proceso" sitas en la Ontología del Sistema, de la siguiente manera:

- En una primera etapa, los datos en bruto de los sensores, tanto del domicilio como de los sensores fisiológicos, así como de la aplicación móvil, son procesados por el sistema experto para determinar cuál es el contexto del anciano en cada momento, formateando los datos en una tabla estructurada con la información sobre la persona, la fecha, la hora y la etapa en ese momento.
- Los datos de los sensores, en función de las Reglas de Proceso semánticas introducidas en el sistema (SWRL), se transforman en las siguientes clases o acciones: Dormir (S), Cocinar (C), Comer (E), Hacer el trabajo doméstico (D), salir al aire libre (O), Deporte al aire libre (U), Usar la Tablet (T), Tiempo de Ocio (P), Hablar por el móvil (X). Se muestra un ejemplo de Reglas de Procesos de este tipo en la Figura 7.6.
- El sistema también verifica los sensores ambientales, como humo, temperatura y humedad, lanzando alertas cuando se activa.
- En cada momento, el sistema toma datos externos climatológicos para integrarlos en los datos de los usuarios, con una serie de valores: Haze (C), Fog (N), Low Fog (N), Fog (I), Precipitation (P), Drizzle (L), Rain (U), Torn Rain (V), Tornado Sight (R), Rain Shower (H), Rain (E), Snow (E), Shower Hail (T), Freezing Rain (T).

- Las fechas de los datos históricos recogidos por el sistema se formalizan en nuevos campos tales como la semana, el día del mes, la hora, el cuarto horario de la acción, el trimestre, y el día de la semana.

Adicionalmente, como se ha comentado en el capítulo 4 de arquitectura, existe otra serie de “Reglas Heurísticas”, soportadas en la plataforma “Drools”, e introducidas manualmente por usuarios expertos, en una primera instancia, en relación a alertas sobre estados del entorno. De esta forma, los datos se gestionan de forma más compacta y sólo se administra la información pertinente a las alertas. Como ejemplo, la regla general: "Si el usuario está en estado de {salir al aire libre} cuando hora > {hora_salida_límite} de la noche → Alerta", la variable {hora_salida_limite}, es la hora máxima en la que un usuario puede salir de casa sin que salten las alarmas. Esta regla se pauta por norma general las 12 de la noche para todos los usuarios, y sin embargo, en base a los históricos, este umbral se va modificando en relación a los datos de salidas recogidos del usuario a lo largo de su historial, es decir, se personalizan las reglas que generan las alarmas de una forma dinámica.

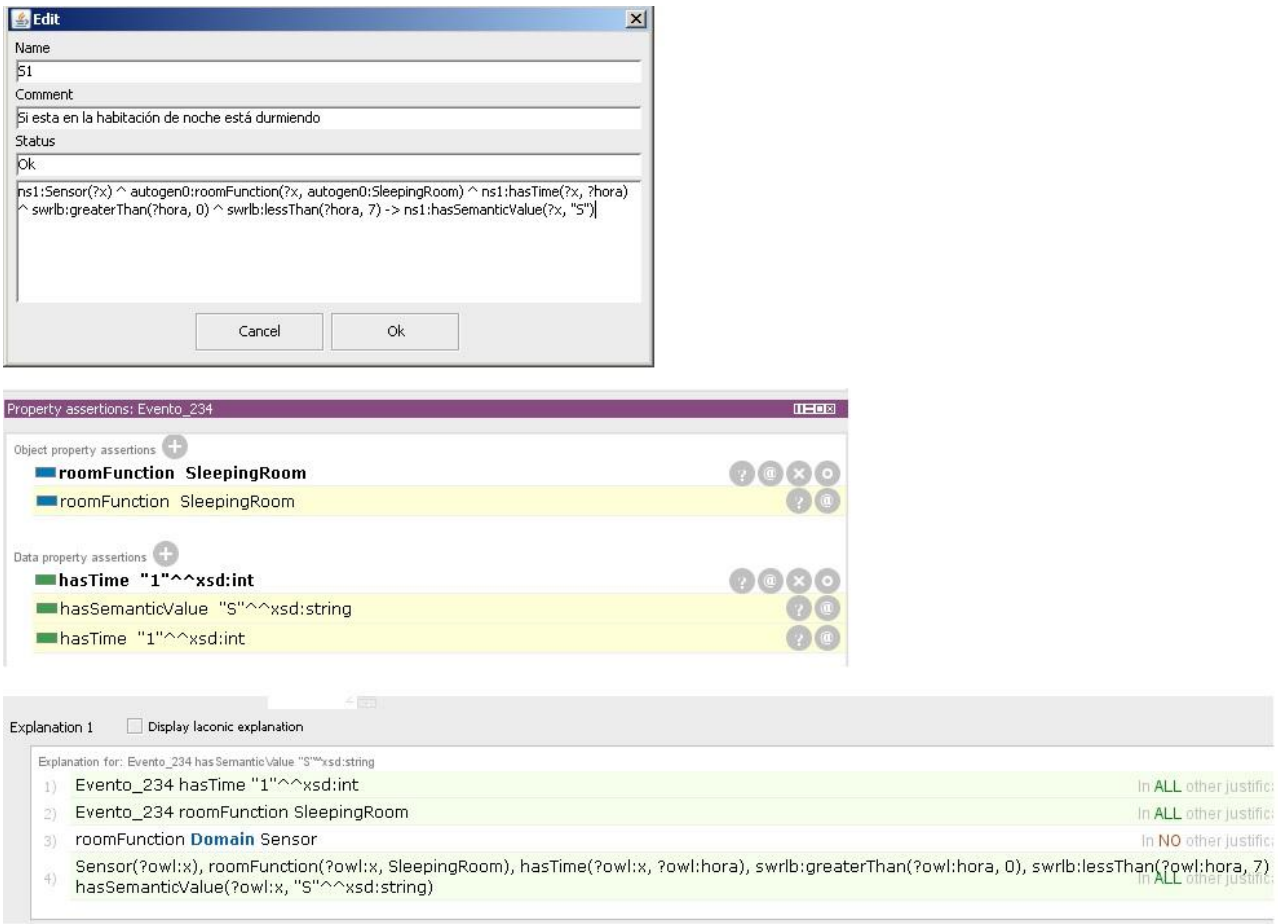


Figura 7.6. Ejemplo de Regla de Proceso en la Codificación Semántica

7.4.2.1 Modelado Supervisado y Selección de la mejor algoritmia a implementar

Una vez que el conjunto de datos ha sido correctamente etiquetado semánticamente, se “presentan” a una serie de algoritmos supervisados para modelar e inferir el comportamiento de los usuarios. El objetivo es modelar la acción concreta que el usuario está realizando en con respecto a los datos de entrada que tenemos, que son el tiempo de la acción a modelar, y el estado meteorológico externo. Los datos fisiológicos y del entorno, se excluyeron del análisis supervisado por las razones ya comentadas anteriormente, en un entorno real de producción. Sólo se mantienen para alertas en base a las Reglas Heurísticas (alerta de humo, por ejemplo). De esta forma, el modelo es una función de la siguiente forma:

Estado_Usuario= f(mes, semana_mes, dia_semana, hora, cuarto_horario, trimestre, estado_meteorológico)

en dónde “f” se ha analizado con cada uno de las varios métodos analíticos propuestos en el apartado de “Aprendizaje Automático” descritos en el capítulo 3. El siguiente paso es analizar cuáles de los distintos métodos algorítmicos es el que mejor confianza de exactitud nos ofrece. De esta manera, se aplican los siguientes algoritmos, a los mismos datos de entrada, y se analiza la precisión general de cada modelo:

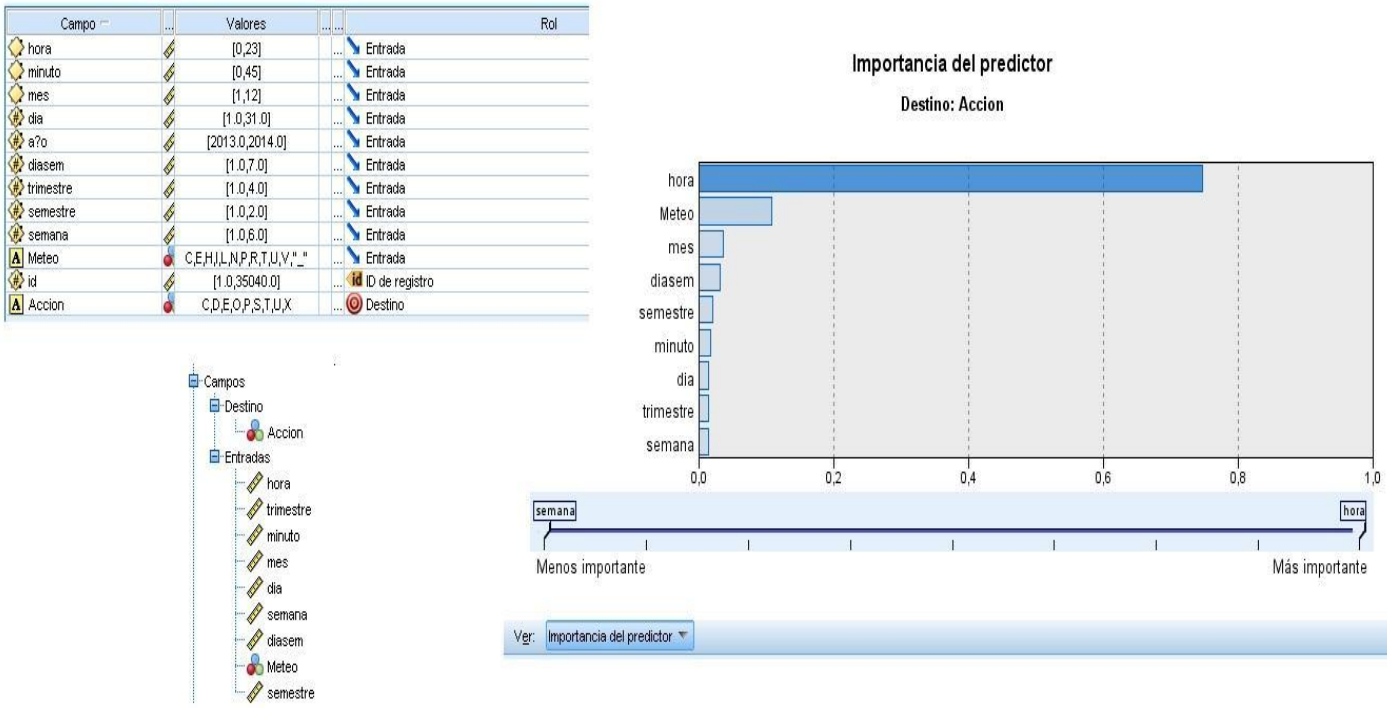
- Árboles de Decisión (C5.1, Quest, C&R, CHAID)
- Red Neuronal Perceptrón
- Red Bayesiana
- Regresión Logística
- Análisis Discriminante

Como se puede observar en la figura 7.6, el mejor modelo (con una exactitud de un 77,69% de confianza), es el algoritmo de árbol de decisión (C5.1), seguido por la familia de algoritmos de árboles de decisión, después las redes neuronales, y finalmente, las redes bayesianas, y regresiones logísticas. La Figura 7.7 muestra ordenados por precisión los distintos modelos. En primer lugar, la figura muestra en forma gráfica, por cada uno de los estados posibles, los datos reales y predichos para el conjunto de test, después, el modelo utilizado, los tiempos de generación en segundos, la precisión general el sistema, y por último, el número de campos de entrada utilizados como predictores. Por fortuna, los árboles de decisión, adicionalmente, cuentan con la ventaja de que su salida es comprensible por los expertos del dominio, y además, es posible incorporar su salida dentro de la ontología del sistema, bien como reglas semánticas (SWRL), o como relaciones de equivalencia semánticas (OWL). Si realizamos un análisis de la importancia de las variables, con respecto a la ganancia de información respecto al objetivo del análisis, es decir, la Acción del usuario como campo a predecir, y el resto de indicadores (mes, día, hora, climatología, etc..) como entrada, se deduce que lo que más “impacta” en la confianza de los resultados es la hora de la acción, en primer lugar, la climatología externa, en segundo lugar, y el mes y el día de la semana en último lugar. (Ver Figura 7.8.).

Ascendente Descendente | Eliminar modelos no usados | Ver: Conjunto de entrenamiento

Gráfico	Modelo	Tiempo de generación (min)	Precisión general (%)	Nº de campos utilizados
	C5 1	5	77,694	9
	Quest 1	5	64,926	10
	Árbol C&R 1	5	64,212	8
	Red neuronal 1	5	61,949	10
	CHAID 1	5	60,945	5
	Red bayesiana 1	5	55,996	10
	Regresión logística 1	5	44,241	10
	Discriminante 1	5	43,787	9

Figura 7.7. Precisión de los distintos modelos.



Campo	Valores	Rol
hora	[0,23]	Entrada
minuto	[0,45]	Entrada
mes	[1,12]	Entrada
dia	[1.0,31.0]	Entrada
a?o	[2013.0,2014.0]	Entrada
diasem	[1.0,7.0]	Entrada
trimestre	[1.0,4.0]	Entrada
semestre	[1.0,2.0]	Entrada
semana	[1.0,6.0]	Entrada
Meteo	C,E,H,I,L,N,P,R,T,U,V,"_"	Entrada
id	[1.0,35040.0]	ID de registro
Accion	C,D,E,O,P,S,T,U,X	Destino

Campos
- Destino
  - Accion
- Entradas
  - hora
  - trimestre
  - minuto
  - mes
  - dia
  - semana
  - diasem
  - Meteo
  - semestre

Importancia del predictor

Destino: Accion

hora
Meteo
mes
diasem
semestre
minuto
dia
trimestre
semana

0,0 0,2 0,4 0,6 0,8 1,0

semana hora

Menos importante Más importante

Ver: Importancia del predictor

Figura 7.8. Importancia de los indicadores con respecto al Objetivo

Sin embargo, utilizando sólo esa información, la exactitud alcanzada por el mejor de los modelos, en un análisis de validación cruzada, (k-fold con k=10), es bastante pobre (sólo 67,95%), como puede verse en la matriz de confusión de la figura 7.9.

accuracy: 68.12% +/- 0.16%										
	true S	true P	true E	true T	true O	true C	true D	true U	true X	class precision
pred. S	14450	1065	13	0	68	50	1	0	1	92.34%
pred. P	940	3800	695	405	326	1016	52	29	10	52.25%
pred. E	10	320	592	105	41	358	185	0	0	36.75%
pred. T	0	123	177	1566	293	91	201	39	0	62.89%
pred. O	102	296	70	298	1203	242	9	342	28	46.45%
pred. C	44	703	888	40	224	1854	484	85	0	42.90%
pred. D	0	31	175	38	0	138	183	0	0	32.39%
pred. U	0	13	0	73	215	2	0	218	14	40.75%
pred. X	0	0	0	0	1	0	0	2	3	50.00%
class recall	92.95%	59.83%	22.68%	62.02%	50.74%	49.43%	16.41%	30.49%	5.36%	

Figura 7.9. Matriz de Confusión Caso 1

7.4.2.2. Incorporación de series de estados al modelado anterior.

De cara a mejorar la exactitud del modelo, se pensó en introducir dentro del modelo, un atributo más: la acción anterior realizada por el usuario. De esta forma, la hipótesis a analizar era conocer si acciones previas realizada por un usuario eran relevantes a la hora de inferir nuevas acciones. Para comprobar dicha hipótesis, se incluyó en la función del modelado la "acción anterior" a la acción a modelizar:

$Estado\_Usuario_n = f(mes, semana\_mes, dia\_semana, hora, cuarto\_horario, trimestre, estado\_meteorológico, Estado\_Usuario_{n-1})$

Con esta configuración, se realiza un primer análisis de correlación basado en la ganancia de información, con los siguientes resultados (ver figura 7.10):

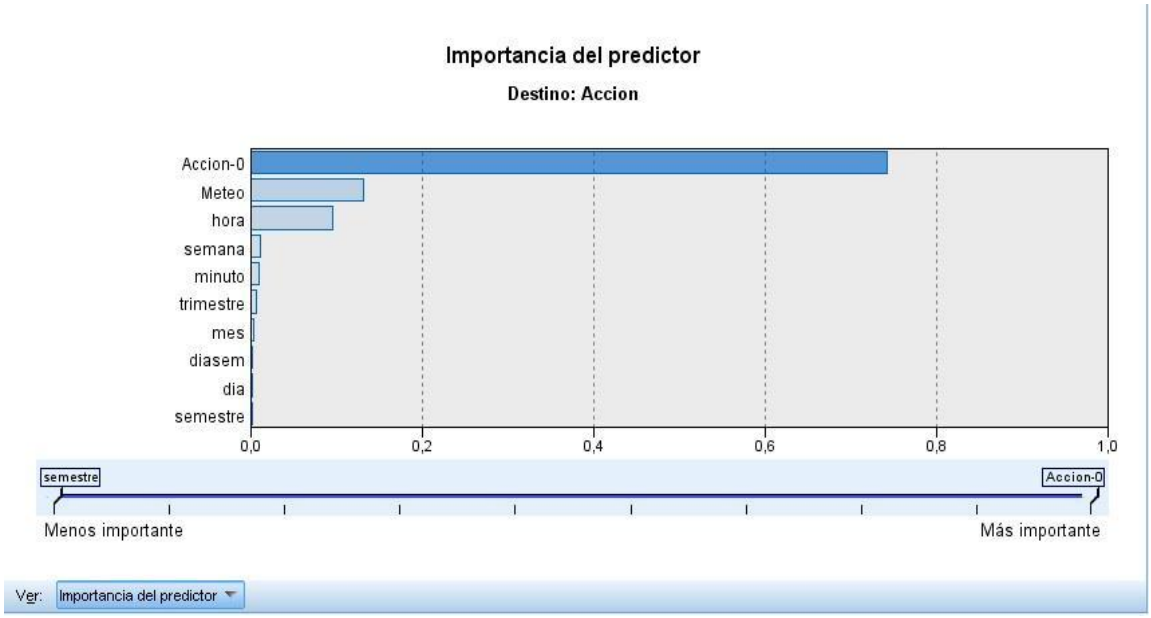


Figura 7.10. Importancia de los indicadores con respecto al Objetivo, Caso 2.

Se observa que la variable que tiene la mayor importancia con respecto al objetivo analizado es, justamente, la acción anterior realizada (“Acción-0”). La segunda variable en importancia es la que detalla las condiciones climáticas exteriores, y la tercera es la hora del día en la que ocurre el evento.
Si entrenamos de nuevo un árbol de decisión C5.1 con el nuevo conjunto de datos, y lo evaluamos, la exactitud del sistema aumenta de forma significativa (véase la figura 7.11), alcanzando el 81,80% de confianza, como se muestra en la matriz de correlación obtenida por el método de validación cruzada.

accuracy: 81.80%										class
	true S	true P	true E	true T	true O	true C	true D	true U	true X	precision
pred. S	7404	270	9	2	74	23	0	0	3	95.11%
pred. P	313	2318	136	51	66	207	23	2	0	74.39%
pred. E	16	175	851	84	6	68	124	0	0	64.27%
pred. T	0	73	17	994	126	27	6	38	1	77.54%
pred. O	0	87	44	6	843	119	1	72	14	71.08%
pred. C	36	197	244	42	3	1321	56	0	0	69.56%
pred. D	2	25	0	83	8	75	345	0	0	64.13%
pred. U	0	28	2	0	59	33	2	242	5	65.23%
pred. X	0	0	0	0	1	0	0	3	2	33.33%
class recall	95.28%	73.05%	65.31%	78.76%	71.08%	70.53%	61.94%	67.79%	8.00%	

Figura 7.11. Matriz de Confusión Caso 2

La aplicación de este modelo de forma práctica se realiza chequeando el estado real del usuario (inferido constantemente en base a los datos de entrada por el sistema de reglas), con el estado teórico en el que debiera estar en función a este clasificador, generado en base a el árbol de decisiones. Si el estado predicho no coincide con el estado real, y ***esta situación tiene una alta tasa de significación (****en este modelo, la confianza de la predicción debe ser mayor que 0.753 para una precisión correcta por encima del 95% de las predicciones****)***, el sistema envía una alerta a Servicio de Notificación para su comprobación manual. Este sistema es un nuevo desarrollo y una reformulación del modelo actual de cuidado teniendo en cuenta las características y preferencias de cada persona, obteniendo un modelo de comportamiento personal para cada uno de los usuarios, como se puede observar en la figura 7.12. En dicha figura, se muestra el comportamiento de un usuario particular (un ejemplo de extracción de reglas con el árbol C5.1 utilizado) , demostrándose que el nodo central del árbol es el campo “Acción-0” (Acción previa del usuario), y dependiendo de dicha acción, viajamos hasta otro subconjunto del árbol (cada uno de los círculos que rodean el centro), en donde, a su vez, el siguiente campo que pauta el comportamiento es la hora en la que se quiere analizar la acción que se va a realizar, y en función de esta hora, en algunos casos podemos ya determinar cuál será la siguiente acción, y en otros, necesitaremos viajar a otro subárbol cuyo centro sea el cuarto horario, o el tiempo meteorológico.

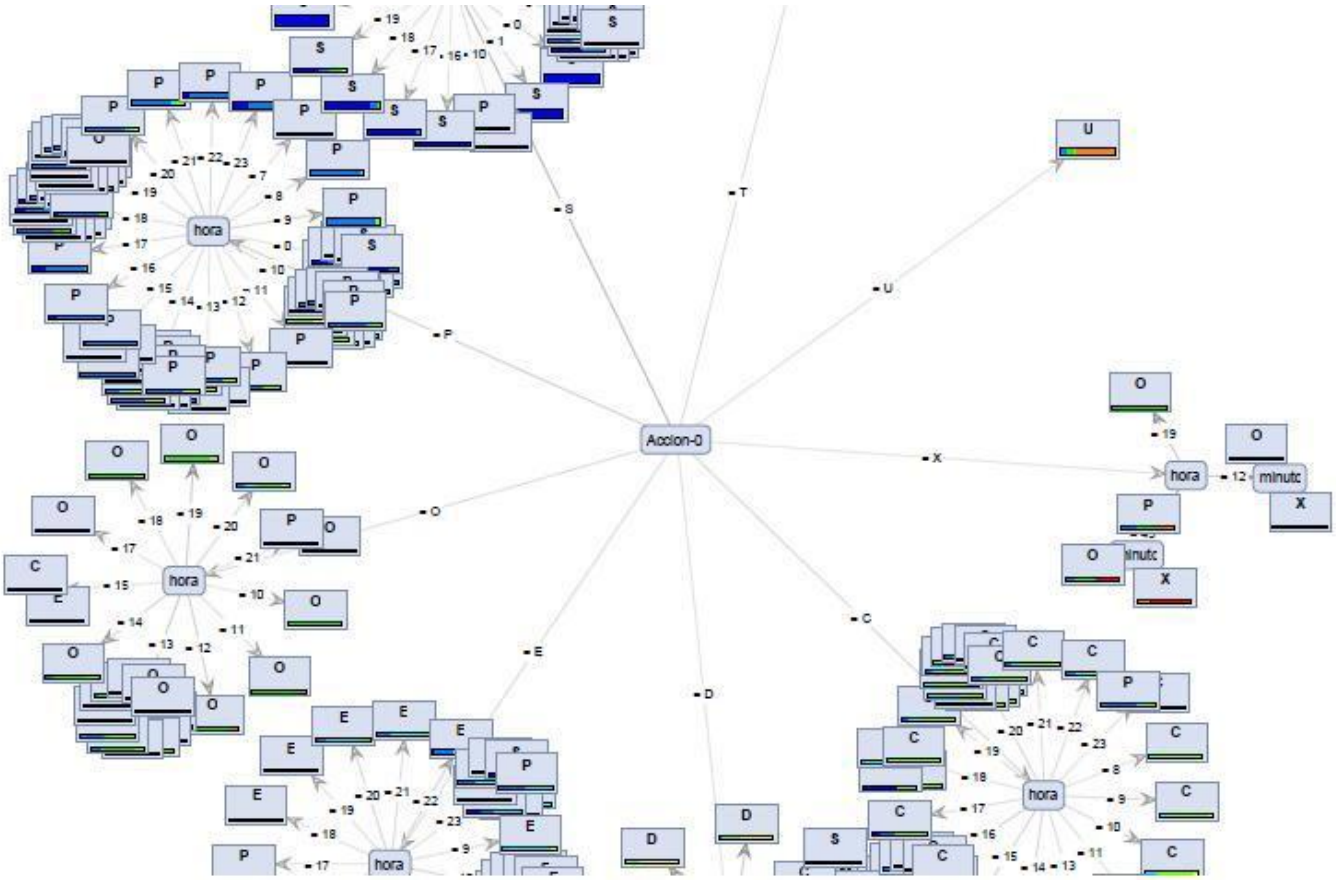


Figura 7.12: Patrones particulares de un usuario concreto

7.4.3. Análisis con datos no codificados semánticamente: discretización de eventos en tiempo, duración y frecuencia.

En el caso anterior, analizamos las actividades de los usuarios, codificadas de forma semántica a través de una "inferencia" del estado de su actividad, basada en los datos de los sensores y su contexto. Pero la codificación semántica tienen un coste: la necesidad de conocimiento previo de expertos para la "traducción" de combinaciones de eventos, tiempos y contexto a una codificación formalizada. De cara a evitar este "coste" en la implantación, y como parte del trabajo de investigación, se propuso investigar y demostrar la hipótesis de si, simplemente con la información de los sensores, y conociendo la ubicación de los usuarios en un momento y lugar determinado, es posible modelizar la ubicación teórica de un individuo en un momento determinado. Una vez conocida la ubicación teórica, podríamos comparar la ubicación real obtenida de los sensores, contra la teórica, y decidir si dicha ubicación es usual o anómala. Finalmente, si podemos determinar si esta anomalía ocurre con una cierta cadencia o sostenibilidad en el tiempo, podríamos generar una notificación. Partimos de la base de que *el comportamiento y los hábitos humanos se caracterizan por tres atributos de las actividades diarias, a saber: tiempo, duración y frecuencia [Noy95]. Los desvíos en el comportamiento pueden ser identificados mirando los cambios en esos atributo*s.

En este sentido, hay que tener en cuenta varios factores ambientales que no ocurren en estudios de laboratorio, como lo son:

- El ruido ambiente o las medidas de error en sensores
- La activación de sensores en el mismo instante, en ubicaciones diferentes (pasar por el pasillo puede activar la cocina o el comedor por presencia, aunque sea tan sólo un instante).
- Activación de múltiples sensores cuando los habitantes de la casa son visitados por amigos, durante celebraciones, asistentes a domicilio trabajando a ciertas horas, largas ausencias (vacaciones, viajes), etc.

De cara a la comprobación de si la hipótesis inicial es suficiente a la hora de generar un proceso predictivo, se configuró un experimento en un escenario con una profundidad de datos de al menos 25 meses, de forma que nos permita encontrar estacionalidad y periodicidad sobre la serie temporal de transición entre los valores de los sensores, aún sabiendo que en un producto real en el mercado, no podemos esperar que el nivel de recolección de datos sea tan profundo para comenzar a analizar los procesos. Es por ello, que en este análisis, era también muy importante poder determinar la profundidad y el horizonte mínimo de datos necesarios en el que podamos comenzar a inferir intencionalidad, de cara a la instalación del resultado de esta investigación como producto final en el mercado.

7.4.3.1 Discretización de eventos en tiempo, duración y frecuencia.

Al igual que en el primer caso de análisis, se lleva a cabo un proceso de modelado y transformación sobre la entrada de datos:

- La fecha se divide en atributos de día, día de semana, semana, hora y cuarto-horario
- Con los datos de entrada, se calcula la frecuencia de la acción (campo Derivar7 en la figura 7.13), hallando la diferencia en tiempos entre la última acción registrada y la acción presente, es decir, la diferencia en minutos en el cambio de un estado a otro.
- Con esta información, eliminamos de los eventos de memoria del sistema que tienen una frecuencia menor que 15 minutos (Derivar7 < 15). Hay que tener en cuenta que anulando esta información, perdemos la trazabilidad entre un evento y otro (eventos muy cortos de frecuencia desaparecen), pero ganamos suministrando mayor importancia al evento.
- Una vez seleccionados estos eventos, generamos una ventana deslizante que genera un nuevo dato: cuál es el evento anterior (estancia en determinada ubicación) que causa el nuevo evento (Habitación_1), ambos eventos con duración igual o mayor a 15 minutos. (Ver Figura 7.13).

Otros trabajos [Kri14] discretizan eventos en series discretas de 15 minutos, aunque el evento se repite en el tiempo, o se selecciona el evento más representativo en cada cuarto de hora si hay más de un evento por hora. En nuestro caso, creemos que la clave es determinar si el evento es relevante o no (> 15 minutos), y posteriormente, si es relevante, asignaremos a este evento su duración y posición como atributos y, finalmente, asignaremos a este evento su marca de tiempo como el atributo final. Por ejemplo, si el usuario está realizando varias tareas en casa, moviéndose a través de varias salas en una actividad de limpieza, por ejemplo, y permanece en cualquier habitación menos de 15 minutos, ignoramos los registros intermedios.

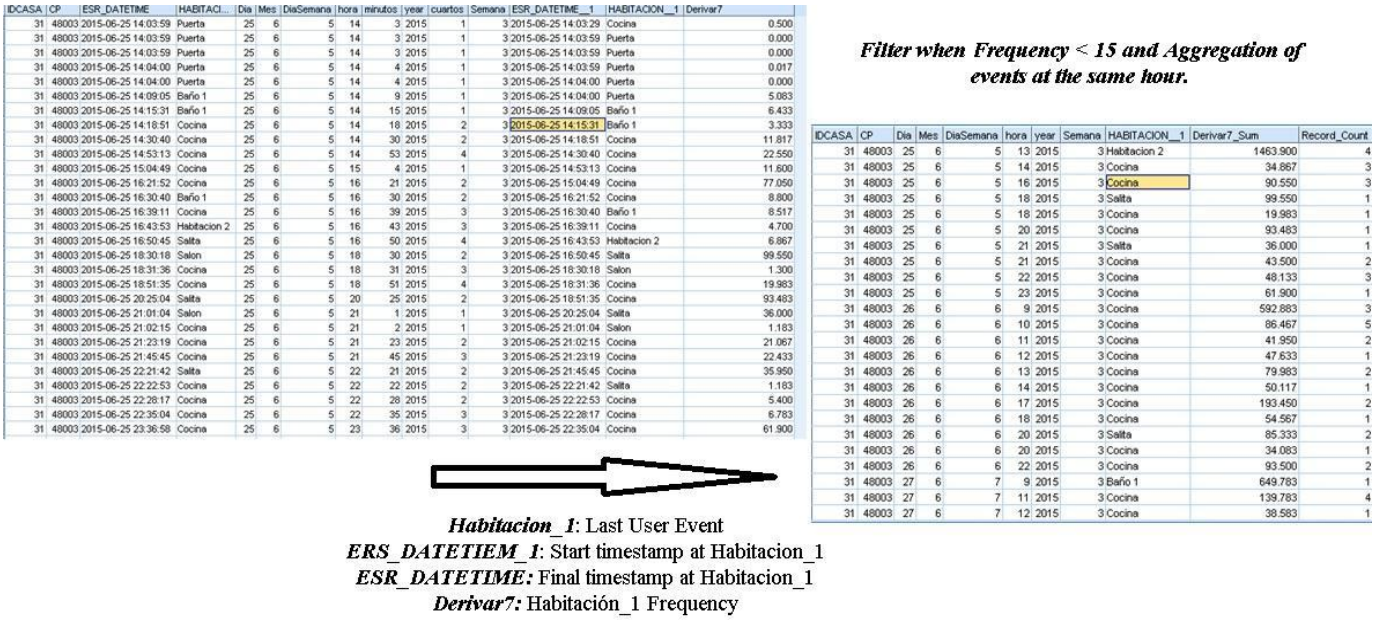


Figura 7.13. Filtrado y generación de la ventana deslizante de eventos.

Esta decisión significa que sólo consideramos en el conjunto de entrenamiento eventos que duran igual o más de 15 minutos y, por lo tanto, sólo podríamos comparar en un entorno productivo eventos con una frecuencia igual o mayor de 15 minutos, el resto se obvia. Pero, para nuestros propósitos de control, es una buena medida, ya que sólo eventos largos tienen la suficiente información y evidencia para evitar falsos positivos en el sistema de alarmas y notificaciones.

Por lo tanto, en proceso final, tenemos un vector de atributos con la siguiente estructura:

- ***Evento (Habitación, Frecuencia, TimeStamp, Habitación_Anterior)***
- donde:
- ***Habitación***: Es el evento relacionado con la presencia de un usuario en una ubicación determinada de igual o más de 15 minutos.
- ***Frecuencia***: Tiempo de Estancia en la Habitación de referencia (minutos de estancia).
- ***Timestamp***: Fecha, hora y minutos de inicio del Evento.
- ***Habitación_Anterior***: Estancia anterior del usuario, cuya duración fue igual o más de 15 minutos.

Hay que tener en cuenta que los nombres de las habitaciones en cada domicilio se parametrizan en el momento de realizar la instalación, y por un lado, no todos los domicilios tienen el mismo número de habitaciones, baños o cocinas, y por otro lado, tampoco reciben las mismas denominaciones. Por lo tanto, cada serie de eventos en cada domicilio difiere en el número de eventos, habitaciones y modelos registrados. Esta característica es importante, ya que nos obliga a modelar a cada casa por separado como una tarea individual. Esto se debe a que nuestro objetivo es predecir en qué habitación ocurrirá el siguiente evento y comparar esta estimación con la sala real registrada por el sistema (evento, obviamente, con una frecuencia mayor o igual de 15 minutos). Si hay alguna diferencia, es posible que algo esté mal. Intuitivamente, la longitud de la historia de los eventos registrados puede tener impacto en los resultados (incluso si no está claro cuál puede ser este impacto), es decir, la secuencia de los eventos en el pasado inmediato, (cómo se mueve el usuario entre eventos y su secuencia), puede incidir en el próximo evento a realizar. Por ejemplo, de la

Habitación al Baño, y del Baño a la Cocina, a las mañanas, puede ser un patrón repetitivo. Así, que debemos validar que las secuencias anteriores afectan a los eventos posteriores, por lo que en el modelo de clasificación desarrollado, generamos una ventana deslizante con la información referente a los 5 eventos anteriores, que se conforman como “entradas” en el modelo predictivo. De todos los atributos disponibles en el conjunto de datos original, seleccionamos la siguiente lista para crear nuestro conjunto de datos y crear un conjunto de datos "históricos":

- Habitación (Evento a predecir)
- Día del mes
- Día de la semana
- Mes
- Hora
- Frecuencia de la habitación
- Los 5 Eventos anteriores dentro de una ventana temporal.

Por último, decidimos no incluir en el conjunto de datos predictivos los siguientes atributos, por ser demasiado específicos, de cara a no provocar sobreentrenamientos o ruido en el sistema:

- Año y Semana del Mes (por lo general, sólo hay un año de datos por domicilio).
- El cuarto horario en el que se produce el Evento (es demasiado preciso). Esto implica que en la misma hora pueden coexistir eventos diferentes, con duraciones diferentes. Lo que la diferencia es la secuencia de eventos, en concreto, el evento anterior.

7.4.3.2. Aproximación estática a la predicción del siguiente estado del usuario.

El primer enfoque es recopilar los eventos relevantes, en una secuencia de estado puro, y tratar de correlacionar cada resultado de eventos (habitación) con el tiempo (día de semana, hora), en un modelo estacional, ya que no tenemos otra información más relevante. En este caso, estamos usando un sistema híbrido de clasificadores, construido por un conjunto de algoritmos: Naïve Bayes (NB), Máquinas de Vector de Soporte (con Optimización Mínima Secuencial, SMO), Redes Neuronales Artificiales (específicamente Multi Layer Perceptron, MLP) Y Random Trees (RandTree). El objetivo es modelar los datos de entrada, es decir, el día, el día de la semana, la hora, los estados anteriores, y con esta información, predecir, para el día y la marca de tiempo a analizar, en qué estado o habitación debería estar el usuario. para finalmente, cotejarla con la situación real. En una primera aproximación estática, usando los clasificadores anteriores, los resultados fueron muy pobres (véase la figura 7.14), incluso en el modelo con inclusión de los 5 estados anteriores, cuando, en teoría, debería mejorar la predicción si existiera un efecto secuencial en el orden de las estancias.

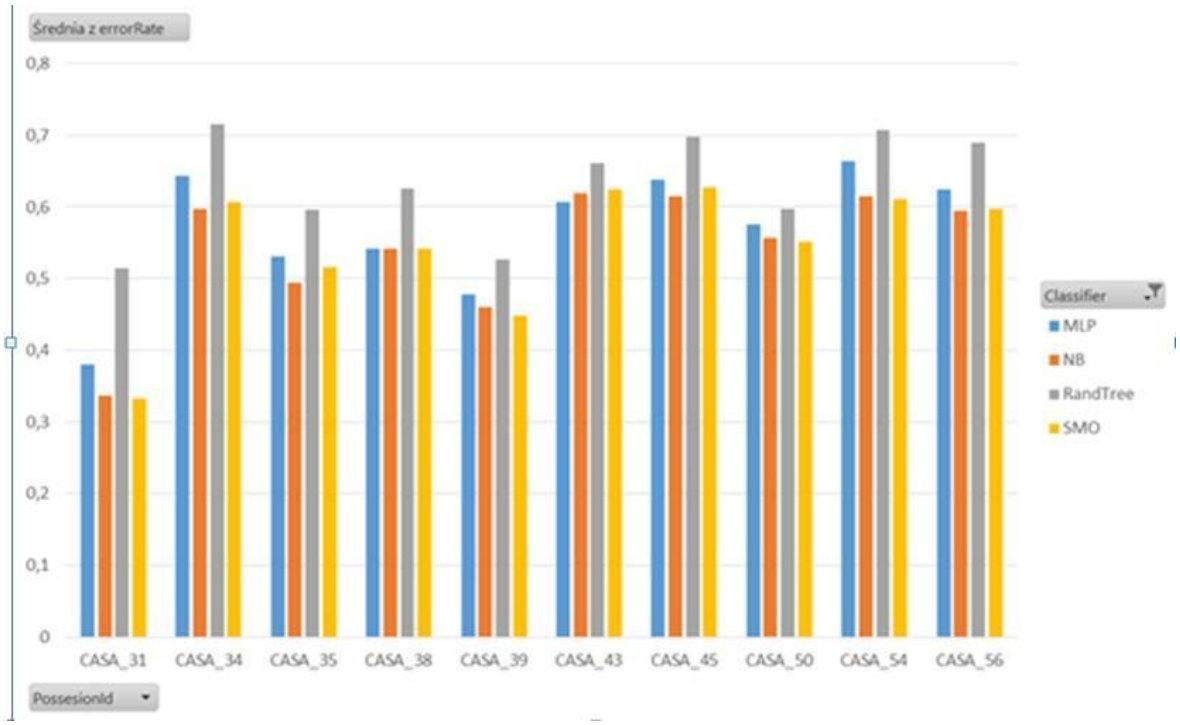


Figura 7.14. Errores de los clasificadores Caso 3.

El mejor resultado es en la casa 31, con un 63% de exactitud, un resultado peor que el registrado por el método de clasificación semántica anterior (81,80%, ver figura 7.10). Sin embargo, en este modelo, aunque la validez no es la deseada, se extraen varias conclusiones importantes:

- Como era de esperar, el indicador “día de la semana” influye mejor que el “día del mes” en la confianza del modelo final.
- El mes produce “overfitting” en los clasificadores. (Es claro, no hay suficientes casas con más de dos años de historia).
- Los resultados esperados no dependen en gran medida del volumen de datos. Es muy importante con respecto al producto final.
- Con la ventana deslizante, es decir, tomando “n” situaciones anteriores para modelar la situación actual, (los 5 estados objetivo mencionados anteriormente), las predicciones empeoran, a diferencia de lo esperado (el sistema se sobreentrena).

7.4.3.3. Aproximación en base a análisis de series temporales.

Los resultados anteriores nos obligaron a cambiar la estrategia de modelado en la búsqueda de un sistema que al menos, fuera similar al comentado en la primera aproximación. Como segunda opción, secuenciamos el tiempo en bloques de 15 minutos y generamos una serie temporal cada 15 minutos en la que situamos los eventos de usuario más comunes en cada intervalo. En la literatura hay ejemplos similares aplicados al mismo contexto [Sur14]. En este caso, estamos extrapolando un día en el futuro, para ver cómo será la predicción cada cuarto de hora en ese día a futuro. A partir de esta configuración, se realizó un análisis de tendencias, con un sistema ARIMA combinado con otros clasificadores, (red neuronal, máquina vectorial de apoyo, ...),

basado en ventanas de tiempo con un retraso de 24 eventos (es decir, 6 horas) en series cronológicas, como posible ensayo. En los resultados, seleccionando el mejor modelo ARIMA (0.1.3)(1.0.1), obtenemos un coeficiente de determinación R^2 = 0.18787, lo cual indica que tampoco es una buena aproximación para la resolución del problema, al menos, en nuestro caso en concreto.

7.4.3.4. Aproximación en base a una predicción binaria sobre la probabilidad de que exista o no cambio en la ubicación del usuario.

La resolución del problema por medio de series temporales tampoco dio los resultados esperados, por lo que tuvimos que idear un último enfoque diferente. Uno de los resultados interesantes en las pruebas anteriores era que el nivel de éxito de los eventos predichos se corresponde, en cada hogar, con el número de casos en los que no hay cambios de habitación. Es decir, teniendo el evento anterior más representativo al momento que queremos predecir, cuándo el siguiente evento no cambia (es decir, el usuario sigue en la misma habitación en el siguiente ciclo de 15 minutos), el predictor tiene éxito, pero en los casos de cambio, el sistema no es capaz de predecir con suficiente confianza cuál va a ser el siguiente evento. Como puede verse en la figura 7.15, si no hay cambio de habitación, el número de errores es muy pequeño y el número de predicciones correctas es bastante alto.

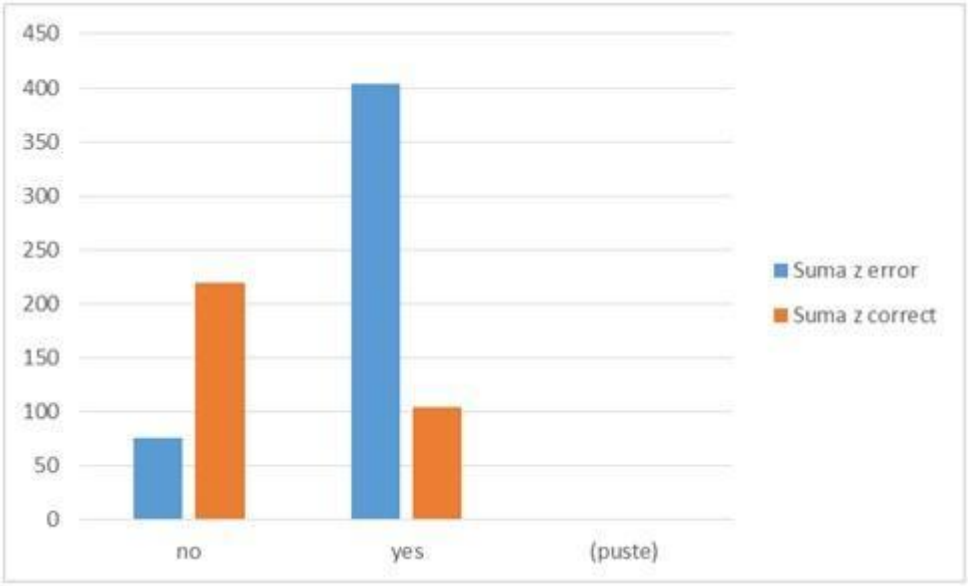


Figura 7.15. Errores de los clasificadores Caso 4.

En el caso de cambio de ubicación, el sistema comete muchos errores y clasifica correctamente sólo un pequeño número de eventos. Entendemos que esta situación ocurre porque en el conjunto de datos hay muchos eventos repetitivos en momentos diferentes, pero, para el objetivo en el que estamos trabajando, es importante analizar cada hora, o cada cuarto de hora, y chequear si la situación en la que está el residente es usual o no, puesto que un retardo de más de 20 minutos en una respuesta adecuada puede ser vital. Es por ello que, como segunda aproximación y bajo este axioma, primero debemos analizar si existe un patrón válido

(como parece lógico), indicando en qué situaciones habrá un cambio de estado del evento (por ejemplo, en ciertas horas de la noche, el usuario se levanta al baño, o en los meses de verano, el usuario sale para dar un paseo normalmente a las tardes). Para ello, en primer lugar, en el conjunto de datos anteriores, generamos un nuevo campo indicando si hay un cambio de escenario o no, como un indicador adicional en la entrada del modelo definitivo. Así, no sólo es importante saber cuál fue la última ubicación del usuario, (en donde el usuario debe permanecer o el usuario no debería estar), sino también cuánto tiempo ha estado en esa ubicación y cuál fue la ubicación anterior, y adicionalmente, cuánto tiempo estuvo en dicha ubicación. Si construyéramos un modelo predictivo que incluye como variable de entrada los valores brutos del tiempo en cada instancia (exactamente 16 minutos, o 32 minutos, por ejemplo), podemos sobreentrenar el sistema si dejamos muchos ciclos de aprendizaje, puesto que los datos de predicción de entrada serán tan específicos que predecirán exactamente en función del día, la hora y el tiempo exacto de estancia en una habitación, pero el conocimiento no se generaliza, por lo que decidimos discretizar los valores habituales de permanencia en cada ubicación. Esta discretización es particular para cada casa. Para cada domicilio, se genera una tabla de discretización en base a un algoritmo de agrupación basado en el método "***intervalos de cuantil***", en concreto, en base a una partición de "deciles", y con un método basado en el recuento de registros. Por ejemplo, tenemos dos ejemplos de discretización para dos casas diferentes (con diferentes grados de profundidad de datos y número de eventos), como puede verse en la figura 7.16.

Quantile Grouping Method

Home:	31		*Home:*	43	
Interval	*Below*	*Upper*	*Interval*	*Below*	*Upper*
1	>= 15,01666667	< 18,83333333	1	>= 15,03333333	< 17,76666667
2	>= 18,83333333	< 23,9	2	>= 17,76666667	< 20,8
3	>= 23,9	< 31,41666667	3	>= 20,8	< 24,76666667
4	>= 31,41666667	< 39,33333333	4	>= 24,76666667	< 29,81666667
5	>= 39,33333333	< 49,83333333	5	>= 29,81666667	< 36,21666667
6	>= 49,83333333	< 62,1	6	>= 36,21666667	< 44,63333333
7	>= 62,1	< 79,6	7	>= 44,63333333	< 59,05
8	>= 79,6	< 106,81666667	8	>= 59,05	< 90,75
9	>= 106,81666667	< 175,58333333	9	>= 90,75	< 177,7
10	>= 175,58333333	< 1463,9	10	>= 177,7	<= 95739,05

Figura 7.16. Discretización personalizada para las frecuencias de permanencia en estancias en distintos domicilios.

Una vez discretizados estos valores, se predice, en primer lugar, si para la combinación del vector de entrada (estado previo, frecuencia de tiempo actual discretizada, frecuencia anterior discretizada, día, hora y semana del mes), hay un patrón que nos puede mostrar si el usuario debería continuar en la misma ubicación anterior o en una nueva, es decir, si podemos predecir si en el estado actual debiera existir un cambio de evento. Por lo tanto, el objetivo del clasificador es el indicador de cambio de estado:

Cambio_Estado_Usuario= f(estado_previo, frecuencia_estado_anterior, frecuencia_estado_actual, día, hora, semana_mes)

Aplicando ***el mismo clasificador híbrido*** que en el punto anterior, conseguimos una exactitud con una confianza entre 75.82% y 80.49% (dependiendo del domicilio) con respecto a la predicción sobre si hay o no cambio de ubicación. (Ver la matriz de confusión de ejemplo en dos domicilios en la figura 7.17):

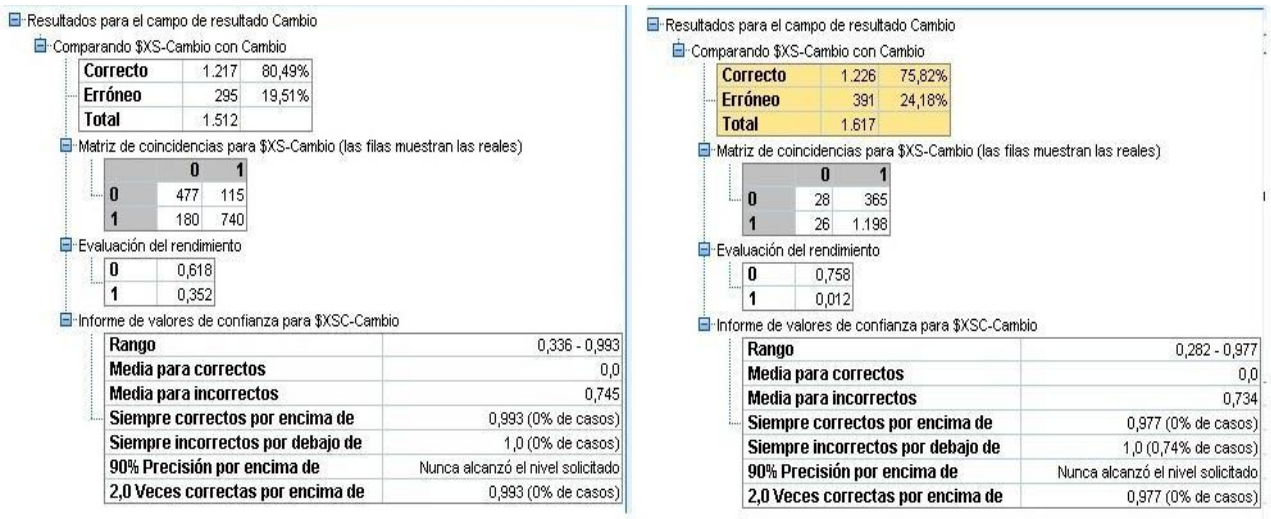


Figura 7.17. Matriz de Confusión para el Objetivo de Cambio de Estado

Con estos valores, podemos generar alertas cuando se den estas dos situaciones:

- Cuando el usuario no está en la misma ubicación esperada:
  Si es habitual continuar en la misma ubicación, (por ejemplo, en el dormitorio, entre ciertas horas), y en realidad el evento indica movimiento o cambio de posición, con una cadencia de más de más de 15 minutos, puede que esté ocurriendo un problema.
- Si no hay un cambio donde se predice:
  Al contrario del punto anterior, si lo habitual, a ciertas horas, es un cambio de estado, (a la mañana, del baño a la cocina, para desayunar), y no hay dicho cambio, también puede suponer una posible alerta.

En este punto, simplemente con este clasificador, podemos ya desarrollar un producto capaz de determinar anormalidades, al menos en una probabilidad de cambios de estado, basada en patrones de comportamiento del usuario. Sin embargo, nuestro objetivo es conocer si podemos discriminar, en base a esta nueva transformación de los datos, y con esta nueva entrada que indica si hay posibilidad de cambio de estado o no, el evento más probable hacía dónde se dirige el nuevo estado.

7.4.3.4 Clasificador Jerárquico Final.

Una vez que el clasificador anterior ha determinado si existe cambio en el estado del usuario, podemos predecir el siguiente estado, sólo para aquellos puntos donde el sistema predice que hay un cambio de estado,

porque cuando no hay cambio, es obvio dónde debe estar el usuario. Este simple filtro minimiza la probabilidad de Falsos Positivos, como se ha demostrado en el punto anterior. Por lo tanto, tenemos el primer clasificador que indica la probabilidad de s existe o no un cambio de evento en el estado del usuario, basado en la frecuencia del estado actual, la frecuencia del estado anterior y el componente estacionario día de la semana y hora. Si no hay predicción de cambio, podemos inferir la probabilidad de que el usuario permanece en el mismo estado anterior, con la confianza suministrada por el primer clasificador, y si hay una probabilidad de cambio, vamos a aplicar un segundo clasificador que previamente ha sido entrenado para aquellos eventos en los que hay cambio en el histórico, obviando del conjunto de datos los registros donde no hay un cambio de evento, es decir, quitando las secuencias continuas del modelo. Se aplican todos los métodos supervisados descritos en el capítulo 3, y en este caso, se selecciona como el mejor algoritmo de aplicación un árbol de clasificación el método CHAID, siendo más efectivo que el método C5.1 (ver figura 7.18).
Los atributos de entrada de este segundo clasificador son las mismas que el anterior, sólo cambia el objetivo del mismo:

- Día de la Semana
- Hora del Evento
- Semana del mes
- Estado anterior
- Frecuencia de tiempo (discretizada) en el estado anterior
- Frecuencia de tiempo /discretizada) en el estado actual.

Estado_Usuario= f(estado_previo, frecuencia_estado_anterior, frecuencia_estado_actual, día, hora, semana_mes)

Figura 7.18. Clasificador para la modelización del Estado del Usuario

La confianza final de que un usuario permanezca en la misma ubicación, o cambie de ubicación viene dada por la confianza conjunta de los dos clasificadores, como se puede observar en un ejemplo en la figura 7.19.
Además, los modelos proporcionan un “Umbral de Confianza” personalizado para cada domicilio, que va a

permitir, en el método de validación expuesto más adelante, certificar y anotar con seguridad los eventos que se vayan produciendo a lo largo de la vida de la instalación en cada domicilio.

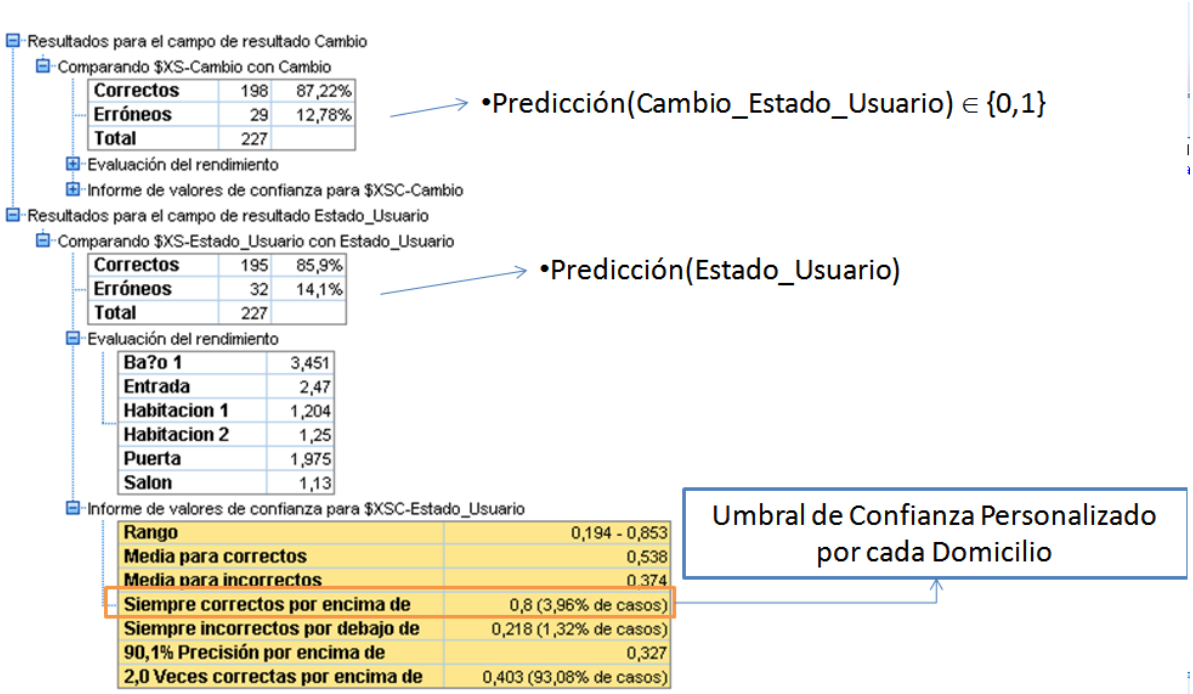
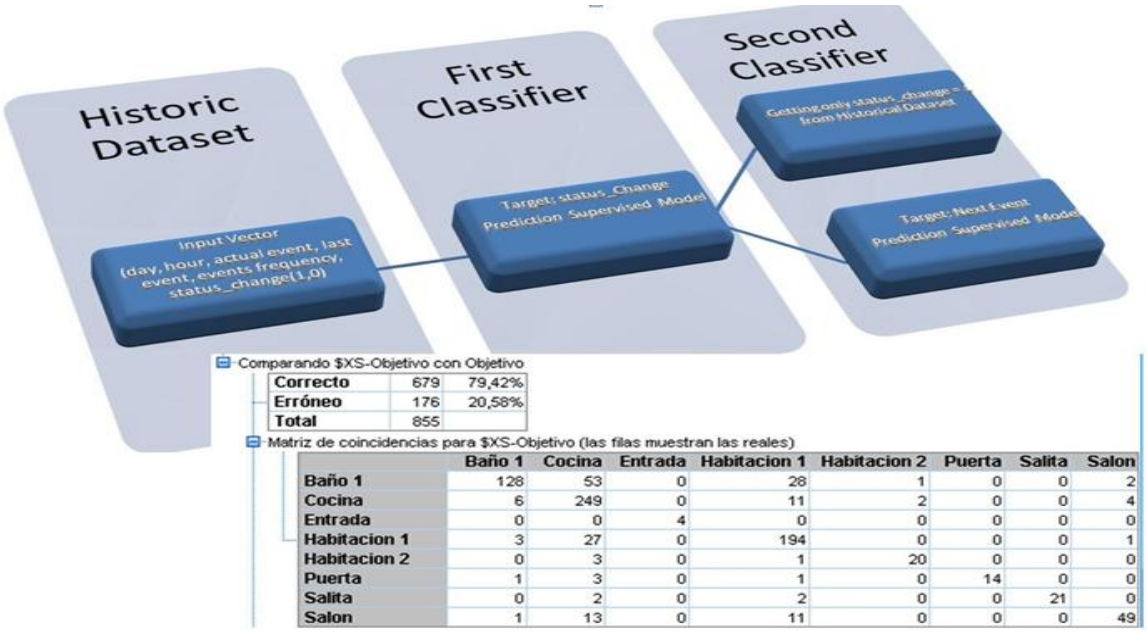


Figura 7.19. Confianza conjunta en el Clasificador Jerárquico.

Analizando con el método de validación cruzada este nuevo modelo jerárquico, vemos que la confianza mejora significativamente que en los casos anteriores (ver figura 7.20).



	Baño 1	Cocina	Entrada	Habitacion 1	Habitacion 2	Puerta	Salita	Salon
Baño 1	128	53	0	28	1	0	0	2
Cocina	6	249	0	11	2	0	0	4
Entrada	0	0	4	0	0	0	0	0
Habitacion 1	3	27	0	194	0	0	0	1
Habitacion 2	0	3	0	1	20	0	0	0
Puerta	1	3	0	1	0	14	0	0
Salita	0	2	0	2	0	0	21	0
Salon	1	13	0	11	0	0	0	49

Figura 7.20. Matriz de Validación para el MultiClasificador para la predicción del Estado del Usuario

Si tomamos las dos muestras más representativas, como en el inicio, el domicilio con identificador 31, se observa que, con esta aproximación, mejoramos los resultados de una probabilidad global del 63% de éxito a

una confianza del 79% para predecir el nuevo lugar de estado. En la figura 7.21 se detallan ejemplos de dos conjuntos de reglas extraídas automáticamente del sistema, en dos domicilios diferentes. Se observa que cada usuario genera un patrón diferente, personalizando sus patrones de cambio de estado sólo cuando existe probabilidad de cambio.

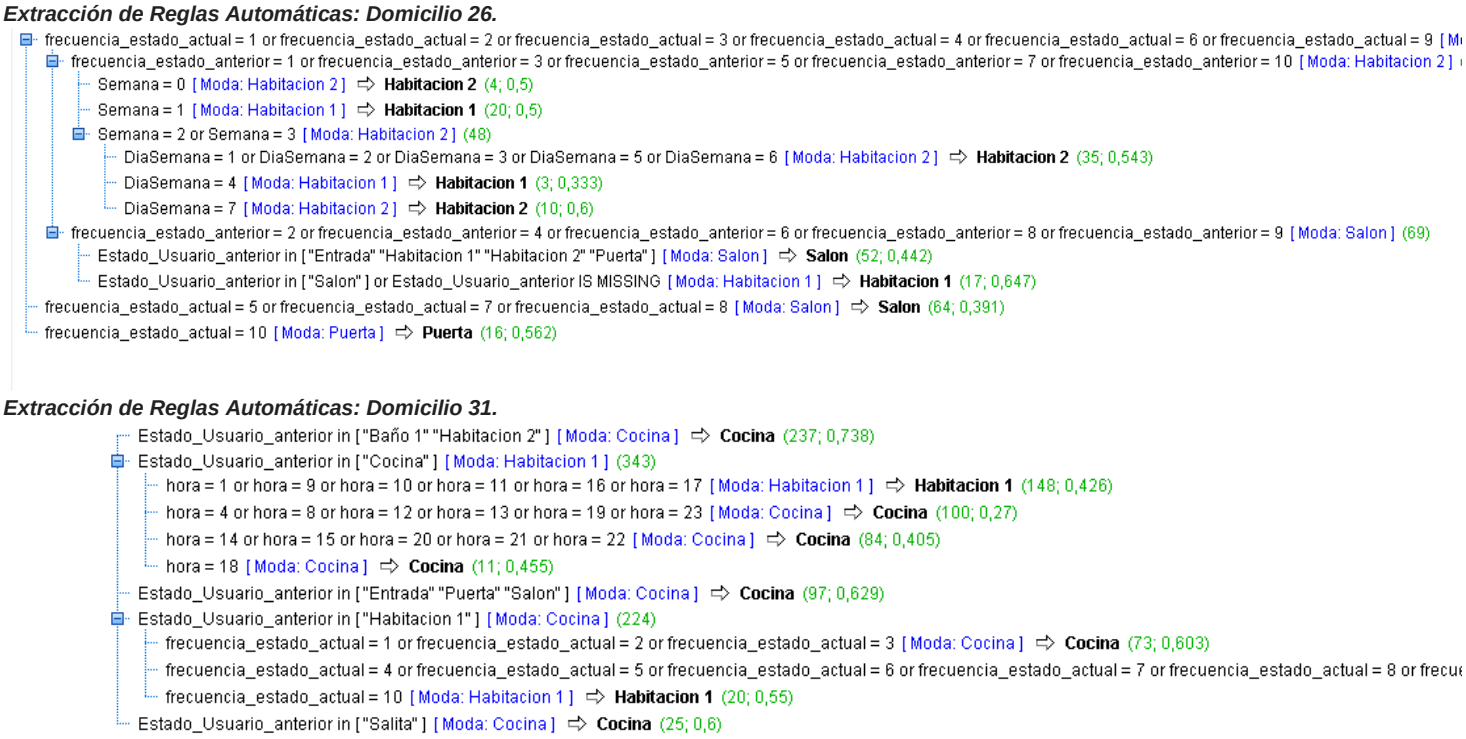


Figura 7.21. Ejemplos de Reglas obtenidas por el Clasificador Jerárquico en Distintos Domicilios

Obviamente, hay algunos casos (domicilios), donde el conjunto de datos es tan caótico, (mucha gente en casa, o muy pocos históricos), que los resultados son poco representativos. En este sentido, la calidad de los datos de entrada, los fallos en las comunicaciones, pérdidas de datos o sensores apagados, afectan a este modelo de forma directa.

7.4.3.4 Profundidad necesaria en los históricos para garantizar el aprendizaje del Módulo de Detección Automática de Patrones

La última pregunta a analizar es conocer cuál es la ganancia de información mínima, es decir, el número mínimo de registros necesarios para obtener una precisión mínima que permita automatizar el Módulo de Detección Automática de Patrones del Sistema Experto. Para ello, partimos de la base que se están analizando todos los hogares con más de 1.000 eventos diferentes (el resto, los rechazamos para el análisis por no tener un número mínimo de registros que podamos valorar como representativos). En la tabla 7.1 se recogen los datos de los domicilios con las siguientes etiquetas:

- Porcentaje (%): Número de registros recogidos en el domicilio con respecto al total analizados.

- Records: Número de registros recogidos en el domicilio
- Accuracy 1: Confianza en la exactitud del primer clasificador (Cambio de estado, 1/0)
- Accuracy 2: Confianza en la exactitud del segundo clasificador: (Predicción de la siguiente Ubicación cuando hay cambio de Ubicación).

Home	%	Records	accurancy 1	accurancy 2	Persons at Home
36	0,23	1.261	90	96	5
59	0,23	1.282	90	95,71	1
33	0,88	4.887	85,6	87,42	1
26	16,39	91.286	88,53	86,62	1
58	0,55	3.036	85,6	84,88	1
32	5,99	33.349	83,85	79,67	1
31	35,48	197.618	86,55	79,42	1
57	0,53	2.942	82,45	76,16	1
56	2,31	12.880	78,28	67,15	2
39	13,33	74.253	88,61	65	2
35	1,69	9.414	81,75	64,34	1
61	1,25	6.940	76,85	63,93	1
50	2,89	16.095	83,14	57	1
38	3,29	18.326	83,23	54,5	2
7	0,91	5.059	78,07	46,35	2
45	4,87	27.104	97,6	41,54	1
34	2,28	12.700	78,63	36,99	1
27	0,16	911	83,02	35,85	1
9	0,22	1.246	79,02	31,71	2
43	2,67	14.857	76,65	28,34	1
54	3,36	18.719	83,12	17,34	2

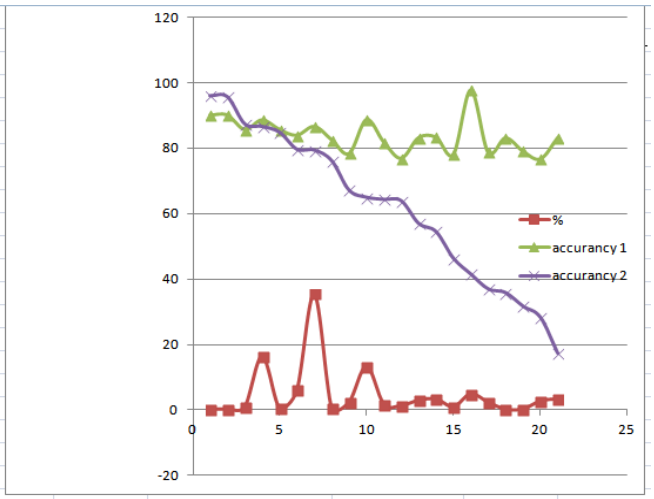


Tabla 7.1. Resultados de la confianza en la exactitud del sistema MultiClasificador en los distintos domicilios.

De los resultados de la tabla anterior, podemos extraer algunas conclusiones interesantes:

- No hay una relación directa entre filas de números de eventos por hogar, y la exactitud global del clasificador.
- El primer clasificador (predictor de cambio de evento) es un buen clasificador para detectar si habrá un cambio de evento (sin saber qué evento será). (Por encima del 75% de exactitud, para domicilios con más de 1.000 eventos recogidos).
- El segundo clasificador es muy dependiente de la naturaleza de los datos de cada domicilio. La exactitud media de los domicilios con más de 1000 eventos es de un 62% de confianza. La mitad de estos domicilios tienen una exactitud con una confianza mayor del 65%, con una media de una exactitud del 81,80% si sólo tomamos este subconjunto de domicilios.
- Aquellas viviendas con una exactitud cuya confianza está por encima del 75%, tienen una horquilla media de de 14 a 41 días de profundidad de históricos (considerando 96 eventos por día de media).
- En ambos casos, el sistema es independiente del número de personas que viven en cada hogar.

7.4.4. Sistema de Detección de Anomalías.

Adicionalmente a los modelos supervisados vistos anteriormente, en el sistema desarrollado también se ha incorporado un sistema de detección de anomalías de patrones no supervisado, basado en el análisis de las diferencias de los distintos estados en los que está cada usuario con referencia a un perfil de comportamiento

“tipo” (ver algoritmia LOF el Capítulo 3). Con este modelo, podemos comprobar automáticamente cuáles son los eventos anómalos en la vida normal de una persona y detectar cuál es el evento atípico y el por qué. Esta información se envía al Sistema de Notificación para decidir o no lanzar una alerta a la familia o a los centros médicos. Este modelo se entrena con todo el histórico por domicilio, para modelar las situaciones normales y anómalas, de forma general, y después se aplica para cada franja de las 3 horas anteriores, para detectar si dentro de esas 3 horas, incluyendo los últimos 15 minutos, hay eventos “extraños” que puedan denotar comportamientos erráticos, peligrosos o indicios de alertas. Sin embargo, en el propio análisis completo de todo el histórico, también podemos analizar si existen anomalías de una forma más global, y descriptiva, como se puede ver un caso como ejemplo en la figura 7.22.

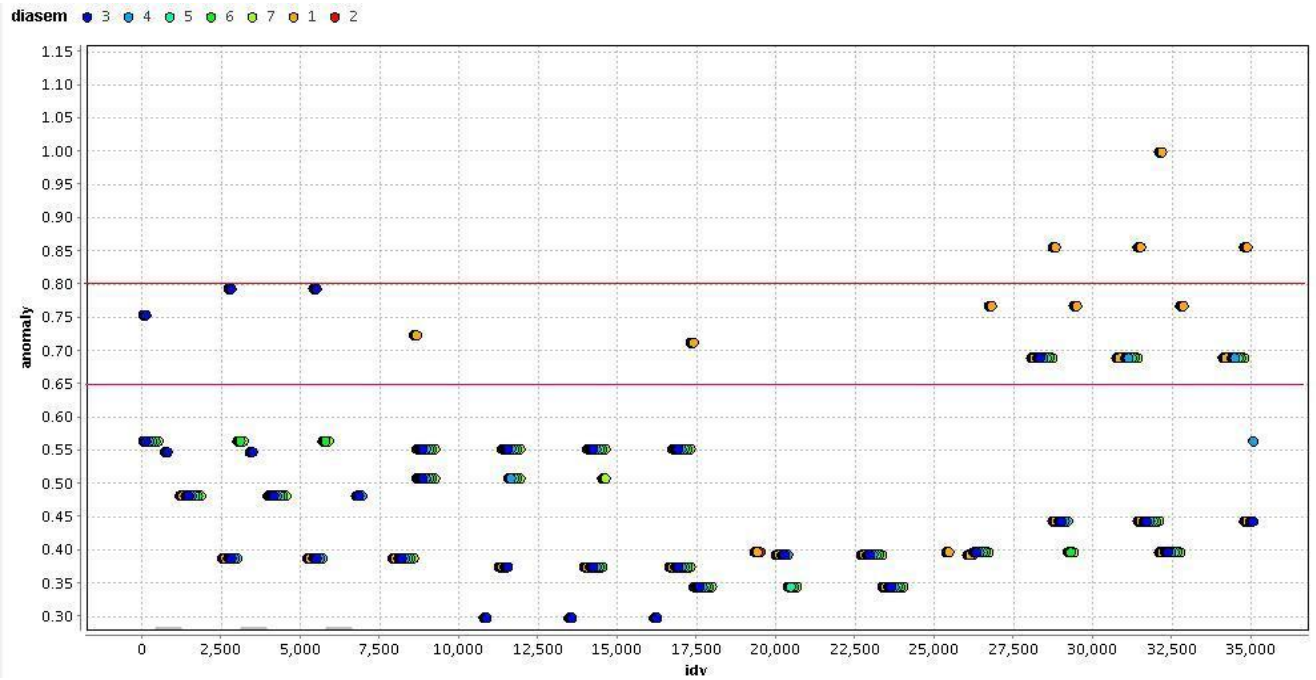


Figura 7.22. Ejemplo de Detección de Anomalías sobre el Histórico de un domicilio.

.

En dicho ejemplo, se muestran los eventos para todo un año y un domicilio, en el análisis de sus acciones cada 15 minutos, y dentro de este espacio temporal, qué eventos son anormales. Se puede observar fácilmente que las conductas de anomalía ocurren generalmente en domingo, con acciones que no son usuales durante el resto de la semana, principalmente por una acumulación importante de eventos de ocio y descanso, y de horarios “caóticos” sobre acciones, salidas y entradas en el domicilio, etc… que no se dan entre semana.

Esta herramienta se ha demostrado como una potente utilidad de uso en el análisis de patrones de comportamiento en este contexto, con una tasa de acierto sobre indicios extraños muy alta (hasta el momento, el 100% de los casos detectados como anomalías lo han sido, pero en un 75% de los casos, han sido anomalías justificadas por el usuario como situaciones excepcionales, pero no críticas en cuanto criterios de peligrosidad o de salud). De cara a minimizar esta realidad *(anomalías verdaderas, pero no peligrosas con respecto a la salud*), durante el tiempo que los históricos generan la suficiente profundidad como para solventar estas situaciones ambiguas, es preciso incorporar una capa semántica de interpretación de las anomalías en

función de la acción que las provoca, que evite falsos positivos con respecto al riesgo, aunque la anomalía estadísticamente sea válida. Por ejemplo, si un domingo puntual hay mucho movimiento de entradas y salidas de la casa, puede que haya comida familiar, y es una anomalía, pero no es claro que tenga que generar una alerta. Esta capa se opera, actualmente, en el Sistema de Notificaciones, de forma manual, dentro de la Metodología de Control Global del Sistema, que se explica a continuación.

7.5 Metodología de Control y Chequeo Global del Sistema

La validación de los modelos supervisados implementados se basan en la comparativa entre los resultados obtenidos en los procesos de entrenamiento, con los valores ya conocidos reales esperados, en un proceso de análisis cruzado denominado "k-fold de validación cruzada," y se demuestran examinando la "Matriz de Confusión" (ver el apartado de Validación en el Capítulo 3). Así, para cada predicción propuesta por el Módulo de Detección Automática de Patrones, en concreto, en el Clasificador Jerárquico, tenemos una probabilidad determinada de que el usuario esté en cierta ubicación, y dicha probabilidad viene dada por la confianza proporcionada por el método de validación del clasificador. Pero al ser un trabajo implantado en un ambiente real, y productivo, añadimos a dicha probabilidad una comprobación adicional, gestionada por expertos en teleasistencia, puesto que en un estado de producción, las predicciones realizadas no podemos compararlas con ningún dato existente, dado que aún no sabemos si la conclusión del modelo (por ejemplo, la siguiente actividad de un residente), es correcta o no hasta que no se haya transformado en un histórico. De este modo, se ha implementado una comprobación adicional en la que cada vez que se genera una alerta, esta, se revisa de forma manual, por un experto, y se valida si es un verdadero positivo, o un falso positivo.

Para poder realizar este control adicional, el sistema realiza una serie de acciones previas (ver figura 7.23):

- Para el instante actual a analizar, se chequea si la predicción del evento teórico en donde el usuario debiera estar supera un cierto ***umbral de confianza*** (determinado por la validación del entrenamiento previa, y particular para cada domicilio).
- Si el evento a estudiar supera dicho ***umbral***, se analiza si existe una ***predicción de cambio de ubicación*** (0/1), es decir, si la ubicación actual debiera haber cambiado respecto a la actividad anterior. Si la hay:
    - Si la predicción del cambio de situación ***no coincide*** con la real (es decir, si el residente debiera haber cambiado de posición, y no lo ha hecho), se activa una "propuesta de alerta", que se envía al controlador interno del sistema, junto con las posiciones anteriores y actuales (tanto la real como la teórica), para que el operador valide si es una alerta real (verdadero positivo), o un falso positivo.
    - Si la predicción del cambio de situación ***coincide*** con lo real (es decir, hay cambio de situación, o no, pero que se valida con la realidad), se chequea si la ubicación predicha coincide con la ubicación real.
        - Si la ***ubicación no coincidiera,*** se envía, igual que en el caso anterior, una "propuesta de alerta" al operador.

        - Si la ***ubicación coincide***, no se realiza ninguna propuesta de alerta, pero se marca este evento como un "***Verdadero Negativo***".
    - El operador, una vez que ha recibido la información de contexto referente a la "propuesta de alerta", la valida, en función de su "expertis", y si decide que es una alerta real, notifica dicha alerta a los cuidadores, o a los familiares, y si no recibe respuesta, hace una llamada directa al residente. Además, notifica dicha alerta como un "***Verdadero Positivo***". La información de contexto está compuesta por los siguientes "servicios de información:
        - Servicio de información sobre la Actividad Actual: informa sobre cuál es la actividad o estado que los usuarios están realizando en el momento de la consulta, en tiempo real.
        - Servicio de información sobre las Actividades de las últimas 14 horas: informa sobre las actividades realizadas durante las últimas 24 horas divididas en franjas horarias de 15 minutos.
        - Servicio Predictivo: informa sobre la estimación de cuál será la próxima actividad pronosticada en los próximos 15 minutos.
    - Si para el operador experto, la "propuesta de alerta" no es suficientemente sostenible, no realiza ninguna notificación, pero marca la incidencia como un "***Falso Positivo***".

En este sentido, obtenemos unos indicadores de exactitud y precisión globales. La ***exactitud global*** del sistema son aquellas alertas positivas (Verdaderos Positivos y Verdaderos Negativos) que se recogen del total de registros analizados (aquellos que superan el umbral de confianza determinado para cada domicilio), la ***precisión global*** la conforman sólo aquellos Verdaderos Positivos detectados como alertas sobre el total de "propuestas de alertas". La ***sensibilidad*** o "recall" no podemos calcularla, dado que nos sabemos cuántos Falsos Negativos se están produciendo, hasta que no se dé un caso del que nos informen los cuidadores o familiares. Esta Metodología de Validación se aplica tanto para el Módulo de Reglas Heurísticas como para el Módulo de Detección Automática de Patrones.

7.6 Consideraciones Éticas.

Como desarrolladores del sistema, hemos tenido que certificar que los servicios que proveemos no incluyen sistemas de vigilancia en el sentido de capturar video e imágenes intrusivas, ni que dicha información sale del domicilio. Además, todas las consideraciones éticas deben garantizar la privacidad, la seguridad, el acceso a la información involucrada, la autonomía, la dignidad, la integridad y la confidencialidad, por lo tanto, se han tenido en cuenta las siguientes disposiciones:

- Seguridad y seguridad de las aplicaciones.
- Privacidad de toda la información de todos los actores involucrados.
  Transparencia en todos los procedimientos con cada parte interesada

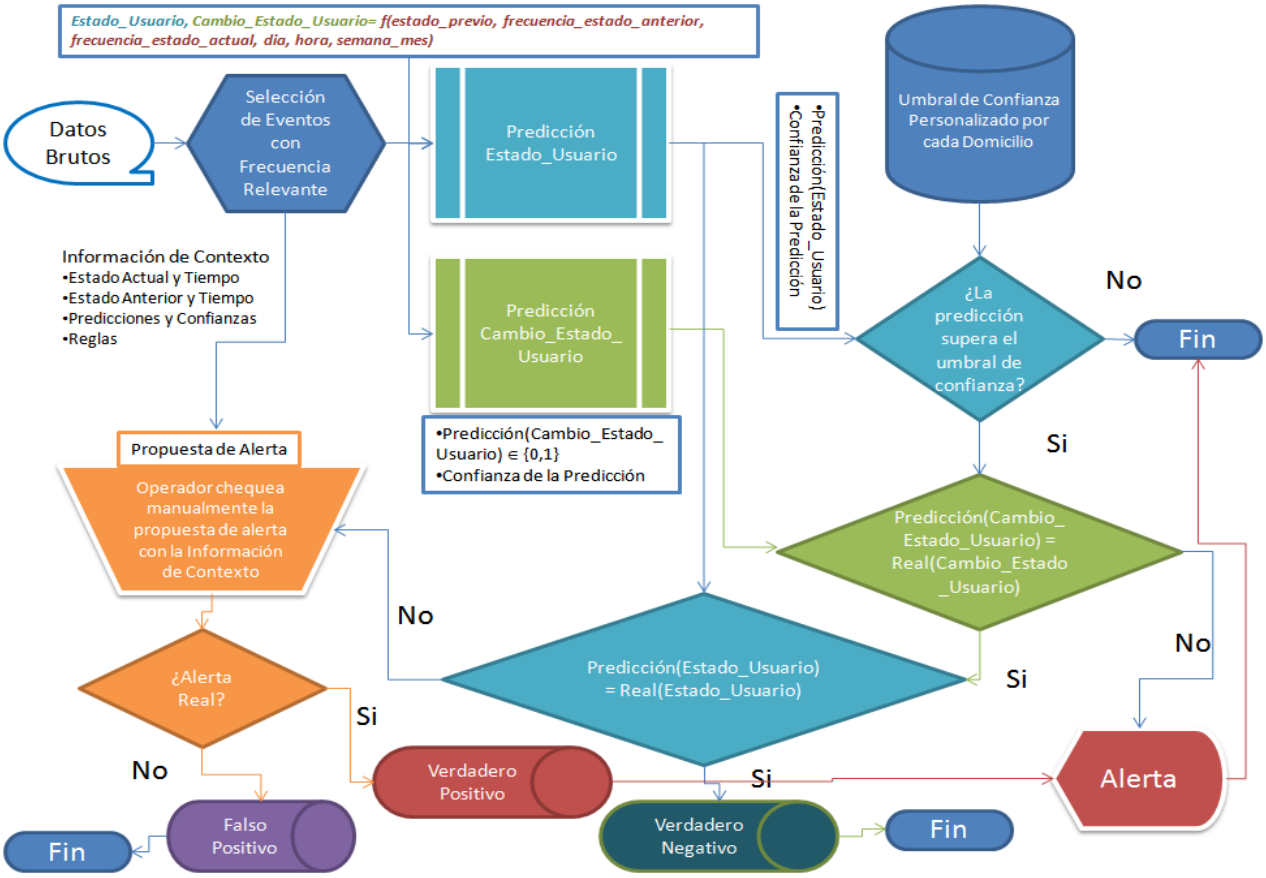


Figura 7.23. Flujo de Proceso en la Metodología de Control y Chequeo.

Al mismo tiempo se garantizarán los siguientes derechos:

- Artículo 8 del Convenio Europeo de Derechos Humanos "Derecho al respeto de la vida privada y familiar".
    - Toda persona tiene derecho al respeto de su vida privada y familiar, de su domicilio y de su correspondencia.
    - No habrá injerencia de una autoridad pública en el ejercicio de este derecho, excepto en los casos en que esté de acuerdo con la ley y sea necesaria en una sociedad democrática en interés de la seguridad nacional, la seguridad pública o el bienestar económico de la sociedad. Para la prevención del desorden o del crimen, para la protección de la salud o la moral, o para la protección de los derechos y libertades de los demás".
- Artículo 7 de la Carta de los Derechos Fundamentales de la Unión Europea "Respeto por la vida privada y familiar". Toda persona tiene derecho al respeto de su vida privada y familiar, de su hogar y de sus comunicaciones".
- Artículo 8 de la Carta de los Derechos Fundamentales de la Unión Europea "Protección de datos personales".
    - Toda persona tiene derecho a la protección de los datos personales que le conciernan.

- Dichos datos deberán ser tratados de forma justa para fines específicos y sobre la base del consentimiento del interesado o de algún otro fundamento legítimo establecido por la ley.
- Toda persona tiene derecho de acceso a los datos que se han recopilado sobre él y el derecho a rectificarlos.
- El cumplimiento de estas normas estará sujeto al control de una autoridad independiente.

- Artículo 16 del Tratado de Lisboa
  - Toda persona tiene derecho a la protección de los datos personales que le concierne.
  - El Parlamento Europeo y el Consejo, de conformidad con el procedimiento legislativo ordinario, establecerán las normas relativas a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por las instituciones, órganos y organismos de la Unión y por el Los Estados miembros cuando desarrollen actividades que entren en el ámbito de aplicación del Derecho de la Unión, así como las normas relativas a la libre circulación de estos datos. El cumplimiento de estas normas estará sujeto al control de autoridades independientes.
  - Las normas adoptadas con arreglo al presente artículo se entenderán sin perjuicio de las normas específicas establecidas en el artículo 39 del Tratado de la Unión Europea.

7.7 Gestión de incidentes y problemas

Integrado junto con la plataforma se suministra un sistema de soporte de tres niveles de acceso. Consiste en un nivel de alcance más alto a un nivel más detallado de resolución de problemas dependiendo del evento que los genere. Los niveles son los siguientes:

- ***Nivel 1.*** Todos los usuarios tienen el contacto de información para cualquier tipo de solución de problemas. Se proporciona un Help Desk que abordará cualquier pregunta entrante en diferentes departamentos. El 70% de las preguntas son consultas sobre operaciones básicas..
- ***Nivel 2***. El 30% restante de las consultas están relacionadas con la plataforma subyacente y el sistema de hardware. Aunque el sistema de sensores inalámbricos está diseñado para funcionar durante un año sin supervisión (debido a la duración de la batería), se da soporte en caso de solución de problemas o cualquier problema relacionado con la instalación / despliegue / operación. Las situaciones anormales que pueden aparecer pueden ser:
  - Robos: lo que implica la destrucción de los sensores, robo, mal uso de la tecnología, manipulación inadecuada y así sucesivamente.
  - Sobretensión debido a una mala instalación eléctrica.
  - Apagado General: que detiene el servicio trabajando (y enviando lecturas de la casa) por un largo tiempo (más de 1-2 días).
- ***Nivel 3.*** En todos los casos, el reemplazo del hardware / software se realiza por el personal técnico. Sin embargo, cuando el problema de hardware no pueda ser resuelto por el personal técnico, un reemplazo se pide al proveedor. En caso de problemas de software, como la búsqueda de fallos en la plataforma UniversAAL, no se supondrá ninguna interacción física con las viviendas ya que se sustituirá el software de forma remota y el usuario no sentirá ningún cambio.

En todos los casos, corresponde al usuario final la decisión de continuar o detener la participación en el proyecto. El usuario siempre puede solicitar la salida y la recuperación de sus datos tal y como se describe en la normativa publicada por la Agencia Española de Protección http://www.agpd.es/. No se ha producido ningún caso de este tipo a lo largo de los tres años que el sistema está funcionando ininterrumpidamente. La instalación técnica del sistema a nivel operativo no requiere más de una hora, chequeando que todo funcionase correctamente, pero además, el técnico recogía en el momento de la instalación información del estado de la casa, del usuario, de su agilidad en la misma, conversaba con el usuario, resolvía dudas, etc. de cara a tener un nivel mayor de información ante posibles incidencias futuras, por lo que la instalación en un domicilio podía durar el doble o incluso el triple de lo previsto inicialmente. Finalmente, como apunte adicional, fue necesaria la contratación de un seguro ante incidencias en las instalaciones, protegiendo posibles caídas de los sensores y rotura de algún mueble, o "desconchamiento" de las pinturas de las paredes, al "despegar" los sensores de presencia o los de puerta de las mismas, obligando a repintar dichas ubicaciones.

7.8 Actividades de Evaluación

A finales del 2015 y principios del 2016, se realizaron varias evaluaciones del proceso de adaptación desde la perspectiva de los propios usuarios y de los agentes responsables de la seguridad de los residentes al respecto de la plataforma. Los objetivos fueron evaluar los logros del piloto en su conjunto, desde la perspectiva del proveedor de servicios y desde el punto de vista de los residentes (ancianos, familiares), en base a diferentes reuniones con los diferentes grupos de actores participantes. Se evaluó la experiencia, usabilidad, utilidad, y el beneficio de la plataforma para el usuario, en todas las fases de implantación, recogiendo sus experiencias con las aplicaciones (la teleasistencia y la aplicación móvil de seguimiento de actividades), sus opiniones con respecto a las instalaciones técnicas, el impacto en los usuarios y los cuidadores formales e informales. Por otro lado, se realizó la evaluación de los distintos módulos del sistema, controlando los siguientes aspectos:

- El sistema de comunicación y la calidad de los datos, principalmente en sus aspectos de completitud y consistencia.
- Se comprobaron las alertas generadas por las "Reglas Heurísticas" formalizadas por los especialistas en teleasistencia, para verificar la veracidad de las alertas que se producían.
- Por otro lado, y en paralelo, se cotejaron las alertas generadas por los módulos automáticos (Sistema de Predicción de Intencionalidad y Sistema de Anomalías), recogiéndose las posibles alertas que pudieran provocar una acción de seguimiento.
- En ambos casos se seguía el siguiente procedimiento:
    - El sistema detecta una posible alerta. Dicha alerta se coteja con los datos que la provocan (tanto datos reales como previstos), de forma manual, por personal interno, con el objetivo de minimizar los falsos positivos. La alerta se anota, y si se decide que es pertinente, se manda una notificación automática a los cuidadores o a los familiares. Si no se obtiene respuesta de estos en un intervalo de tiempo, se llamaba directamente al residente. Si la alerta no era validada, se anotaba como falso positivo, y en caso contrario, como verdadera positiva.

Capítulo 8

Resultados del Proyecto

A lo largo del 2015 se ha desplegado una plataforma automatizada de Teleasistencia, en la que se basa este trabajo, en 60 domicilios, que, hasta el momento actual, siguen funcionando, analizando constantemente la vida y costumbres diarias de los distintos usuarios, en tres segmentos de clientes diferentes: ancianos dependientes, ancianos cuyos hábitos están empeorando debido al envejecimiento y personas mayores que sufren los primeros síntomas de demencia.

8.1 Resultados en el Tratamiento de Historiales Clínicos.

En primer lugar, se ha generado una aplicación automática de tratamiento de Historiales Clínicos, con el objetivo de extractar toda aquella información relevante referente a los pacientes residentes. Usualmente, la información médica sita en los historiales está descrita en lenguaje natural, y con este sistema, de dichos historiales, se extrae la siguiente información:

- Información general relativa a los residentes (edad, sexo, últimos diagnósticos principales, estado de salud general)
- Información clínica particular: Consiste en información relacionada con su calidad de vida, y que puede afectar a los modelos de intencionalidad de comportamientos, así como a las recomendaciones de mejora de hábitos de vida, o al seguimiento de la toma de medicamentos, guías de actividad diaria en base a ejercicios, o en base a una aplicación, generada a medida de este proyecto, para registrar hábitos de vida saludables, como lo son recomendaciones de paseos, ejercicio, y otros..

Como se ha detallado en el Capítulo 7 de Implementación, la confianza del anotador semántico se plasma en una precisión global del 90%, con una sensibilidad global del 69%, pero aún existe un amplio recorrido de investigación en esta área, debido a ciertos problemas detectados y no resueltos aún, como lo son:

- Existen términos recogidos por medios estadísticos y no reconocidos por la ontología de referencia (UMLS) (1.9 %)
- Existen conceptos ambiguos sin resolver semánticamente, por falta de información de contexto (1.6%). Por ejemplo, “alta”, no es lo mismo en fiebre que en el estado de un paciente, o no se codifica igual un cáncer de mama masculino que femenino.
- Conceptos no anotados por falta de contexto o complementariedad (6.1%).

Es decir, se anotan y enlazan correctamente un 90% de los conceptos presentados en los historiales médicos. Sin embargo, el 10% restante, clínicamente es representativo, dado que en este 10% se encuentran problemas de:

- Anotaciones incorrectas de Estadio de pacientes.
- Anotaciones incorrectas de Diagnósticos, por problemas de unificación en la codificación.
- Anotaciones ambiguas que producen incertidumbre en el correcto diagnóstico del cuadro clínico del paciente.

8.2 Efectividad del Sistema Experto

El sistema se comenzó a implantar en enero del 2015, y en un proceso iterativo, se han ido añadiendo domicilios y usuarios de una forma progresiva (ver figura 8.1). A medida que se comenzaban a recibir datos reales, y a procesarlos en el Sistema Experto, se comenzaban a recibir alertas, y en un proceso constante, se procedía a su comprobación manual. Esto ha permitido ir refinando el sistema dinámicamente. Este refinamiento se ha logrado gracias a la plasticidad suministrada por el sistema a la hora de la gestión del "aprendizaje" en el Módulo de Detección Automática de Patrones, concretamente en el Sistema de Predicción de Intencionalidad (ver Capítulo 6) poder aplicar reglas de control sobre el sistema experto de la plataforma.

Figura 8.1. Interacciones de los Usuarios en la Plataforma

En un principio, las alertas producidas se basaban en la generación de ciertas reglas manuales generalistas para todos los domicilios, "Reglas Heurísticas", como una manera de prever falsos positivos en base a un sistema automático del cual no se tenían evidencias de su fiabilidad. Estas reglas se describen en base a la experiencia de los expertos técnico en teleasistencia, y gestionaban las alertas en un primer nivel de control. En paralelo, se iban agregando los datos de los sensores al sistema de forma diaria, y se iba validando el Sistema de Predicción de Intencionalidad automático, así como el Sistema de Detección de Anomalías.
En esta sección, se desgranan los resultados obtenidos por el Módulo de Reglas Heurísticas, y a continuación, se comparan con los resultados obtenidos por el Módulo de Detección Automática de Patrones.

8.2.1 Evaluación del Módulo de Reglas Heurísticas

Como se detalla en el Capítulo 7 de Implementación, las "Reglas Heurísticas" ubicadas en el Sistema Experto se disparan cuando se cumplen los antecedentes que las modelan, en tiempo real, sobre la información de los sensores de actividad, presencia, humo y temperatura, con una lógica generalista para todos los domicilios por igual. Además, según el método de validación propuesto, se analizaron una serie de eventos susceptibles de interés (ver el apartado de Metodología de Validación), que se revisan por un operador, de cara a determinar su validez real. En este sentido, a lo largo del año 2015, y respecto a las "reglas heurísticas", se revisaron 223 posibles eventos en 29 domicilios seleccionados con una buena calidad de datos, (199 eventos sólo relacionados con posibles incidencias en cuanto a la actividad, los 24 restantes, fueron alertas relacionadas con gas, humo o temperatura). De los 199 eventos relacionados con alertas en actividad, se analizaron como posibles un total de 58 "propuestas de alertas", es decir, eventos que generaron posibles alerta a revisar por el operador (Positivos). En resumen, una media de dos alertas por domicilio, de las cuales:

- Las alertas por Actividad Prolongada, fueron Verdaderos Positivos (VP) en un 11% (2 de 18). El resto son "Falsos Positivos", ya que el operador no las consideró alertas reales.
- Las alertas por Inactividad, fueron "VP" en un 13% (5 de 40).
- 141 eventos de los analizados, no eran alertas, y tampoco generaron alertas. ("VN").

El resto de alertas no relacionadas con la actividad fueron las siguientes:

- Las alertas por Gas, fueron Verdaderos Positivos (VP) el 100%, (1 de 1).
- Las alertas por Humo, fueron "VP" un 20% (2 de 10).
- Las alertas por Anomalía en Temperaturas, fueron "VP" un 50% (1 de 2).
- Las alertas por Puertas Abiertas, fueron "VP" un 30% (3 de 10).

En definitiva, y centrándonos sólo en las alertas por Actividad, se obtuvo una "***precisión global***" (tasa de verdaderos positivos entre las "propuestas de alertas") de un 11% en la detección de actividades prolongadas (2 casos de 18), y de un 13% (5 casos de 40) en la detección de alertas por inactividad, es decir, en la detección de cambio de actividad no realizada. En general, la exactitud del modelo "heurístico" (los verdaderos positivos y negativos, con respecto al total de eventos analizados), y sólo con respecto a las alertas de Actividad, es de un 74%. (Ver la matriz de confusión en la Tabla 8.1)

Las razones de estos resultados generales son las siguientes:

- Existieron problemas relacionados con la calidad de la señal y la cobertura en las comunicaciones en ciertos domicilios, que se fueron resolviendo con el tiempo.
- Muchos falsos positivos se generaban con el sensor de humo, en la cocina, al cocinar, debido principalmente a la ubicación de los sensores y su cercanía a los fuegos de cocina. Las instalaciones deben ser cuidadosas en este sentido.
- Otros muchos falsos positivos se refieren a momentos de inactividad por periodos vacacionales u otros eventos, con el sistema en activo, y sin incorporar un control de ausencia prolongada en el domicilio, o simplemente, no "apagar" la plataforma en periodos estivales.
- Otros problemas que afectan directamente a las "Reglas Heurísticas" se refieren a la multi-actividad en casa, cuando varias personas en casa realizan actividades simultáneas no previstas, como celebraciones en fechas señaladas, que disparan el sistema de reglas.

Tabla 8.1. Matriz de confusión de Resultados por Actividad según el Modelo Heurístico

Total de Eventos revisados por Actividad	199	*VP*	*FN*
Total propuesta de Alertas por Actividad	58	**7**	**0**
Total de Verdaderos Positivos por Actividad	7	**51**	**141**
Total de Falsos Positivos por Actividad	51	*FP*	*VN*
Total de Verdaderos Negativos	141		
Exactitud 74%			
Precisión 14%			

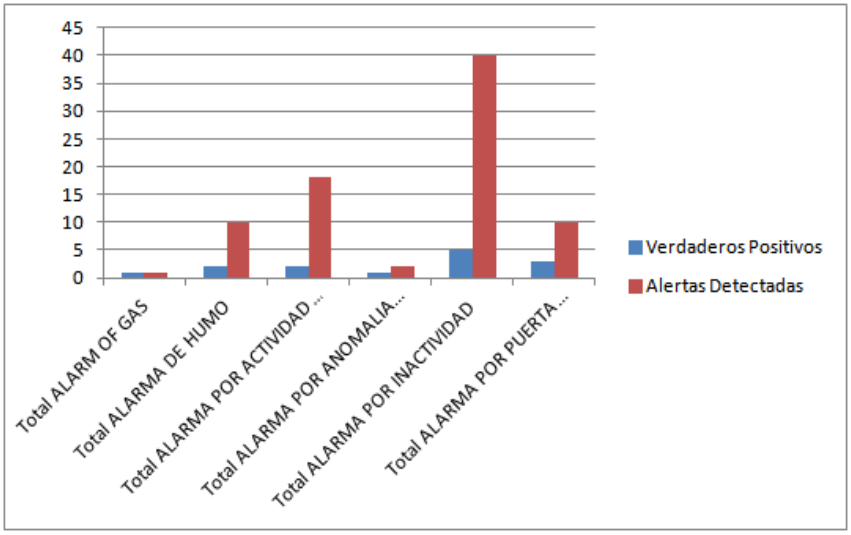


Figura 8.2. Propuestas de Alertas revisadas y validadas

Como punto positivo a reseñar, durante los 12 meses de operación, se detectaron 2 casos de alertas críticas, en donde las alertas por actividad prolongada se dispararon correctamente, se activaron los protocolos, ante dos casos de caídas graves en el domicilio, y el aviso a los familiares que permitieron una gestión proactiva exitosa. En otros 5 casos de alerta por inactividad, también se avisó a los familiares, principalmente en ausencias no justificadas en casa y 1 caso de desorientación.

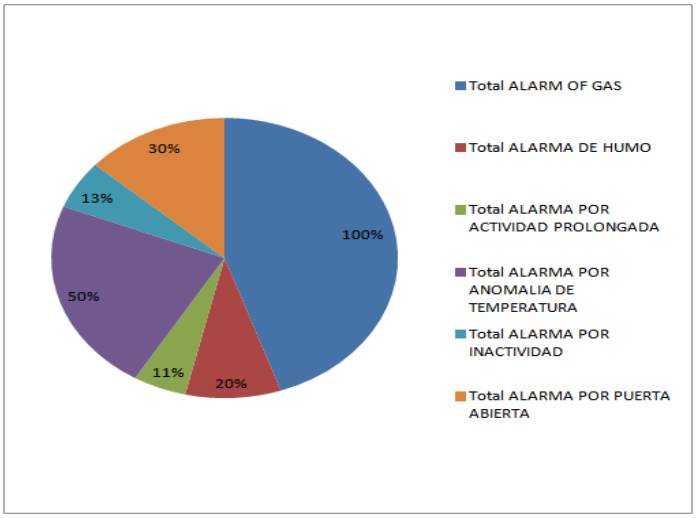


Figura 8.3. Proporción de Propuestas de Alertas

8.2.2 Evaluación del Módulo de Detección Automática de Patrones.

En paralelo, a pesar de que las reglas heurísticas eran las que más confianza daban a los expertos en teleasistencia, con los datos recogidos se iba alimentando los históricos del sistema, y se iba evaluando, en paralelo, el sistema automático de predicción de intencionalidad, basado en la segunda aproximación vista en el capítulo de Resultados del Proyecto, basado en el sistema MultiClasificador (predicción de cambio o no de situación, y predicción de la ubicación estimada en dicho cambio de estado). Se analizaron 532 eventos y se predijeron 27 posibles alertas de anormalidad sobre los patrones de comportamientos en 20 domicilios con más de 1.000 eventos representativos. De dichas alertas, 16 fueron Verdaderos Positivos (se confirmó la alerta como válida), y 11 fueron Falsos Positivos, es decir, se obtuvo una precisión del 59% de confianza. (Tabla 8.2). Los Verdaderos Positivos fueron casos en los que:

- Se detectó actividad prolongada en el Baño (más de una hora, y en un horario no habitual), y se confirmó una caída.
- Se detectó actividad no habitual en otro domicilio en la cocina, a una hora no usual para dicho evento, y se confirmó una caída.
- En otro caso, en horas no habituales, la estancia en una habitación no usual, se detectó malestar en el residente.
- En dos casos diferentes, se detectó la permanencia más de lo habitual a la noche en el salón, (horas): el residente se quedó dormido en el salón viendo la tele. Se notificó.
- En otro caso, la estancia en el baño un tiempo prolongado (más de una hora), cuando normalmente el residente estaba fuera de casa. Se detectó malestar e indisposición en el residente.
- En otro caso, el residente estuvo en el salón a la tarde, de menos de una hora, pero que el sistema indicaba como anómalas, dado que esperaba, curiosamente, que el residente saliera de casa (habitualmente, daba un paseo a esas horas). Se notificó, y el residente comentó que se encontraba cansado.

- Otros casos de conductas no habituales, pero confirmadas como buenas, lo fueron al detectarse principalmente actividad en la cocina a horas no habituales y ocupación de habitaciones no usuales, en dos domicilios, en diciembre y en enero, debido a que en esas fechas, los hijos conviven unos días con los padres en el domicilio, y el sistema detectaba cierta actividad inusual, de forma correcta, debido a comportamientos de los hijos no “aprendidos” previamente.

En cuanto a los Falsos Positivos, principalmente se dieron por las siguientes causas:

- Movimientos a la habitación, a la noche, a descansar, a horas usuales, pero desde ubicaciones inusuales (del baño a la habitación, cuando lo usual era ir desde el salón).
- Salidas y entradas de casa, detectados como no usuales, pero mal clasificados. Este es uno de los casos que más Falsos Positivos ha generado. De hecho, incluso en las reglas heurísticas, se tuvo que quitar el lanzamiento de esta alerta, dado que las entradas y salidas de los residentes eran muy caóticas.

Como conclusión, el sistema automático de detección de intencionalidad detectó, de más de 556.972 registros de actividad extraídos por los sensores, sólo 559 eventos susceptibles de ser analizados (predicciones por encima del umbral de confianza personalizado por cada domicilio), de los cuáles, el sistema clasificó como posibles alertas 27, y lo fueron, realmente 16. Es decir, el sistema tiene una confianza de precisión del 59%, frente al 14% del sistema heurístico, con lo que se demuestra que el sistema automático es más efectivo que el sistema propuesto de reglas generales, tanto en precisión, como en número de alertas posibles a validar (de 58 alertas por actividad a chequear por el sistema heurístico, a sólo 27 del sistema automático).

Tabla 8.2. Matriz de confusión de Resultados por Actividad según el Modelo Automático

Total de Eventos revisados por Actividad	559	*VP*	*FN*
Total propuesta de Alertas por Actividad	27	**16**	**0**
Total de Verdaderos Positivos por Actividad	16	**11**	**532**
Total de Falsos Positivos por Actividad	11	*FP*	*VN*
Total de Verdaderos Negativos	532		
Exactitud 98%			
Precisión 59%			

8.3 Evaluación por parte de los usuarios, cuidadores y otros agentes.

8.3.1 Impresiones Generales de los usuarios y agentes implicados.

Después de la utilización de la plataforma basada en la investigación de este trabajo, se mantuvieron varias reuniones con todos los agentes implicados (usuarios, cuidadores, agentes sanitarios, familiares), y se les remitió una encuesta de satisfacción para conocer de primera mano sus opiniones al respecto de la plataforma. A continuación, se hace un resumen general de las conclusiones obtenidas:

- En general, hubo respuestas heterogéneas entre los asistentes. Mientras que algunas de las personas asistidas no han percibido mejoras en su estilo de vida, otro grupo de personas asistidas tuvo un sentimiento muy positivo sobre las aportaciones realizadas por el sistema, con comentarios como: "Me siento más seguro ahora y quisiera continuar utilizando la aplicación de la misma manera".
- Por otro lado, ciertos residentes acudieron a la entrevista acompañados por sus familiares, quienes a su vez, destacaron los beneficios que para la familia de la persona asistida y la tranquilidad que suponía el haber podido disfrutar de la plataforma durante un cierto tiempo. Se daban situaciones en la que los familiares vivían a 100 kilómetros de su madre y que la plataforma había sido beneficiosa al no tener que desplazarse al tener monitorizada la actividad del residente en remoto.
- Se recabaron algunas reacciones negativas entre algunos usuarios en cuanto a las cuestiones de problemas técnicos. El problema más común era la conectividad de red 3G. En las zonas rurales había muchos problemas para mantener el sistema conectado, especialmente al comienzo de la puesta a prueba de los servicios.
- Se detectó cierto rechazo a la aplicación de seguimiento fuera de casa, principalmente debido a que las personas residentes no se sienten cómodas con un seguimiento continuo, y prefieren que este seguimiento sea selectivo, o a demanda, cuando sean los propios residentes los que quieran activarlo, dado que las persona asistida, son perfectamente capaces de darse cuenta de que está tomando el camino equivocado, sin necesidad de activar alertas.

En el capítulo de propuestas, en general, se comentó como mejora importante el interés de poder “empaquetar” la plataforma en un producto fácilmente instalable por familiares o usuarios, sin necesidad de tener un perfil TI. Esta mejora impacta directamente sobre la fortaleza de implementar el segundo método de gestión de intencionalidad, basado en sensores simples y de instalación básica, como lo son los sensores de presencia, puertas, humo y gas.

8.3.2. Resultados de las Evaluaciones.

A fecha del 6 de junio del 2016, había 300 usuarios de la plataforma, según la distribución de la tabla 8.3.

Tabla 8.3. Distribución de Usuarios por Roles.

ROLES	Plataforma de Teleasistencia	
	N	%
Residentes	194	64,7%
Cuidadores Profesionales	96	32,0%
Cuidadores no Profesionales	10	3,3%
Other	0	0,0%
TOTAL	300	100,0%

Los perfiles de los residentes en sus domicilios se muestran en la tabla 8.4:

Tabla 8.4. Perfiles de los Residentes en sus Domicilios

Perfil Residente	Immediate Aid Provider	
	N	%
TOTAL	194	
Hombres	24	12,4%
Mujeres	170	87,6%
Edad Media	74,8	
Edad Mínima	21	
Edad Máxima	111	

Según muestran las interacciones de los usuarios con la aplicación de teleasistencia, instalada en sus teléfonos móviles, para la configuración y el chequeo del sistema y la infraestructura local de cada hogar, para cada usuario hay una interacción por día en el 95% de casos (keep-alive + sistema de inicio de verificación por el sistema PUSH). Hay pocos casos donde hay menos de una interacción, y se ha comprobado que coincide con los periodos vacacionales de los residentes y han apagado el sistema durante ese período. Del total de usuarios activos (300), 39 de ellos (13%) cumplieron con la adhesión en el uso de la aplicación sobre las expectativas, 245 usuarios (82%) usaron los servicios de forma regular, mientras que sólo el 5% (16 usuarios) utilizaron la aplicación menos de lo esperado (menos de 1 conexión por día). (Ver figura 8.4).

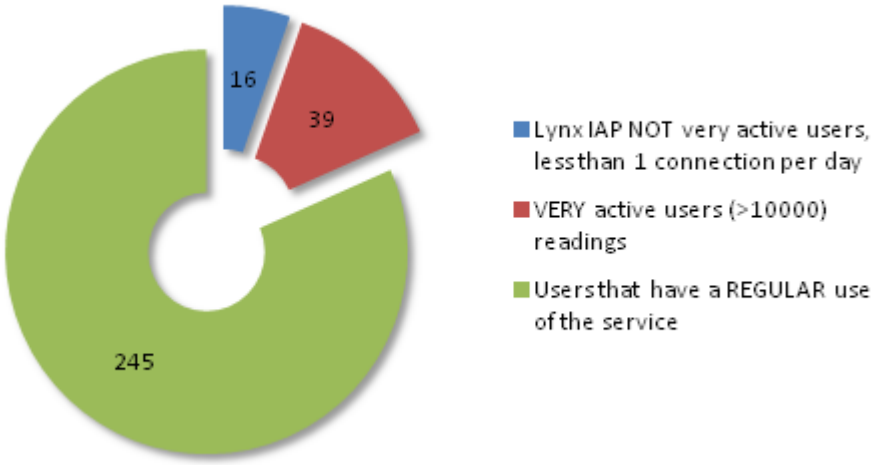


Figura 8.4. Adherencia de los Usuarios al Servicio.

Las tabla 8.5 indica el valor que los residentes dan a la utilidad de la plataforma. Las valoraciones están normalizadas en una escala del 1 al 5, donde 5 indica la mayor opinión positiva sobre la utilidad de la aplicación.

Tabla 8.5. Valor que los residentes dan a la utilidad de la plataforma

Indicador de Utilidad	N	Utilidad por el usuario	Sensación de seguridad	Utilidad por el Cuidador	Sensación de control	Usuario Activo	Fiabilidad en el Uso
Plataforma de Teleasistencia	67	3,4	-	3	3,2	2	3,2

En cuanto al volumen de satisfacción, como se puede observar en la Figura 8.5, el 44% de los usuarios no contesta ni afirmativamente ni en contra de la herramienta, y de los que sí o hacen, el 38% está satisfecho, contra un 16% de usuarios insatisfechos.

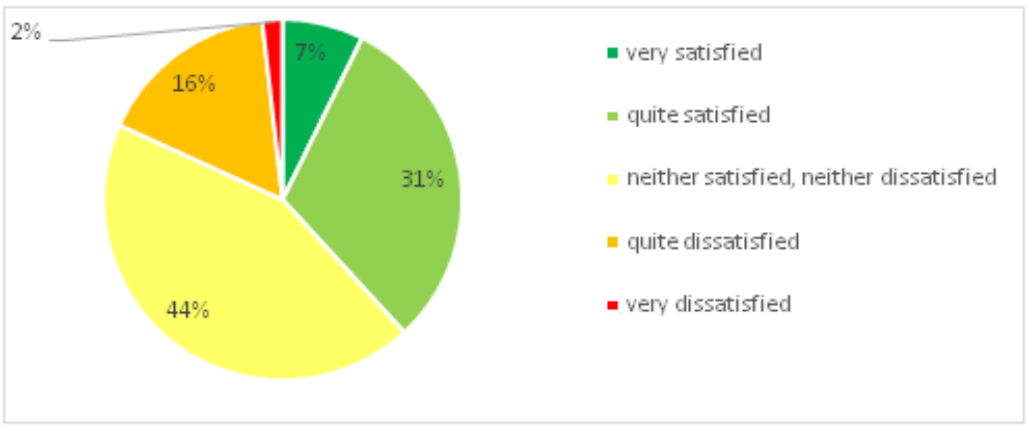


Figura 8.5. Volumen de Satisfacción de los Usuario

Capítulo 9
Conclusiones

En este trabajo hemos presentado un sistema real de atención asistida a personas que viven solas en sus domicilios, con el objetivo final de proporcionar una solución robusta, fácilmente desplegable y de bajo costo que permita garantizar la seguridad de los residentes, tanto desde el punto de la seguridad fisiológica como de la ambiental. El sistema propuesto se basa en un Sistema de Predicción de Intencionalidad, pero que no es capaz de personalizar correctamente los patrones de cada usuario (en el Módulo de Detección Automática de Patrones) hasta el primer mes de funcionamiento (de 15 a 40 días), que es el tiempo que requiere para aprender a interiorizar los hábitos de cada domicilio en particular. Entre tanto, el sistema se alimenta de un sistema experto soportado por reglas generalistas, (Módulo de Reglas Heurísticas), introducidas manualmente por expertos en el contexto de la Teleasistencia, para poder prevenir situaciones anormales generales, como, por ejemplo, la estancia en el baño más de una hora a la noche (que pueden inducir, claramente, a una posible caída o desmayo), o incumplimiento de ciertas pautas clínicas recomendadas por los sanitarios (por ejemplo, pasear cada dos días).

Se ha demostrado que una vez que el Sistema de Intencionalidad Automático a alcanzado un nivel de aprendizaje estable, su confianza en la precisión y exactitud es mayor que en la del sistema de Reglas Heurísticas generales (ver tabla 9.1), con lo que se demuestra que un sistema basado en técnicas de aprendizaje personalizado es más efectivo que un sistema heurístico generalista.

Tabla 9.1. Comparativa de los Módulos en el Sistema de Predicción de Intencionalidad

Sistema	***Precisión***	***Exactitud***
Módulo de Reglas Heurísticas	74%	14%
Módulo de Detección Automática de Patrones	98%	59%

Para el aprendizaje automático de patrones, se han modelado dos procesos de intencionalidad de acción siguiendo principios totalmente diferentes:

- Un sistema complejo, en el que los datos en bruto se “interpretaban” por el sistema y se traducen a conceptos de actividades complejas, en base a una serie de reglas semánticas de fácil implementación en el sistema. Además, en este sistema se incorporan datos externos sobre el tiempo local o el uso de electrodomésticos, e incluso sensores de mediciones fisiológicas y cámaras de visión artificial.
- Otro sistema, totalmente diferente, simplemente basado en sensores sencillos, que únicamente se activan con la presencia de una persona a su paso, y en el que el reto era demostrar si era posible detectar situaciones anormales registrando solamente la posición de los usuarios en casa a lo largo del tiempo, sin incorporar datos externos

Después de evaluar ambos sistemas, se concluye que el primer sistema complejo es más robusto, menos sensible a la variabilidad y a la calidad de los datos obtenidos de los sensores, (principalmente debido a que su capa de interpretación “oculta” deficiencias en los datos a nivel más “fino”), y su confianza de exactitud es algo mayor, debido precisamente a la incorporación de datos externos. Por otro lado, este tipo de instalaciones es más costosa, conlleva un mayor nivel de mantenimiento, dado que los sensores son más complejos, una mayor complejidad de instalación en el hogar, (requiere en algún caso cableado), y además, genera dos problemas prácticamente insalvables de cara a la introducción de dicha plataforma en el mercado:

- La utilización de sensores fisiológicos constantemente en el domicilio es una opción no real, dada la reticencia de las personas mayores a llevar una pulsera, cinturón o sensor encima, principalmente debido a la incomodidad que ello les supone.
- La utilización de cámaras genera grandes dudas en los residentes en cuanto a cuestiones de privacidad.
- La instalación de cableado supone un sobrecosto a la venta del sistema, así como la posibilidad de ciertas obras menores a realizar en cada domicilio (conectores en las paredes, interconexión con electrodomésticos, etc…), que encarecen el sistema.
- La codificación semántica de los eventos no es evidente, y necesita cierto conocimiento implícito de las costumbres de los usuarios, lo que encarece y complica la implantación.

Por otro lado, el segundo sistema propuesto, la modelización en base a un sistema básico soportado solamente en sensores de presencia, asume un tratamiento más costoso a nivel de software y modelado, debido a la necesidad de desarrollar un sistema MultiClasificador, que detecte, en un primer estadio, una probabilidad alta de que exista un cambio en la ubicación del usuario, y por otro lado, intentar determinar cuál va a ser la siguiente posición del residente si el cambio de ubicación es probable. Este segundo modelo se demuestra menos eficaz que el anterior, (un ratio de una exactitud media de un 63% de confianza en las medidas de todos los domicilios con más de 1000 eventos representativos, contra un 82% en el caso complejo), pero lo suficientemente válido como para poder implantarlo en entornos reales (la mitad de los domicilios ofrecen de resultado una exactitud mayor que el 65%, con una confianza media de un 81%, muy similar a la de la primera aproximación) y poder utilizarlo como producto de detección final. Hay que tener en cuenta que estas medidas son sólo en aquellos casos en los que el sistema prevé que va a existir un cambio de ubicación, en el resto de situaciones, la predicción de persistencia en la misma ubicación previa a la detección de cambio de ubicación tiene una confianza en la exactitud de un 80%. En definitiva, con este soporte, si el estado de predicción no coincide con el estado real, y esta situación tiene una tasa significativa, dependiente del modelo particular de cada domicilio, en base al umbral de confianza por encima del cual las predicciones son siempre correctas en el set de entrenamiento en un 95%, el sistema envía una alerta al Servicio de Notificación de atención remota, que se revisa y permite iniciar de inmediato los protocolos de asistencia. Por otra parte, el método de detección de anomalías utilizado, es un complemento al sistema supervisado, con el fin de detectar comportamientos inusuales. La capacidad de analizar estos patrones para mejorar el sistema, e incluso como punto de partida de ampliar la información de los comportamientos domésticos de cara a mejorar los sistemas de atención de los gobiernos y agentes de salud, es un hecho real, gracias a esta investigación.

9.1 Crítica, Despliegue y Lecciones Aprendidas

La implantación en un entorno real de la plataforma, ha suministrado una experiencia enormemente valiosa a la hora de solventar algunas deficiencias, tanto técnicas, como procedimentales y operativas. Hay una amplia gama de experiencias diferentes en muchos ámbitos del trabajo. Por un lado, el reclutamiento de usuarios ha implicado un gran volumen de actividades, involucrando a un gran número de profesionales multidisciplinares, dado que fue necesario cubrir una serie de tareas relacionadas con la comunicación, actividades de marketing, la búsqueda y selección de usuarios interesados, dentro de los criterios de inclusión y exclusión predefinidos, la gestión legal y ética, y sobre todo, la estrategia de segmentación de los usuarios en diferentes grupos para facilitar el despliegue. Los usuarios finales se dividieron en 3 grupos: grupo A: personas realmente convencidas de la necesidad de un sistema autónomo de teleasistencia, grupo B: personas con necesidades parciales en soluciones asistenciales/clínicas, y grupo C: personas que potencialmente puedan usar esta aplicación en función de sus estados físicos y clínicos. Esta segmentación es una de las lecciones más útiles que hemos aprendido, ya que nos ha permitido hacer más fáciles los despliegues dando prioridad a los diferentes usuarios en función de sus necesidades. Como punto de partida en esta segmentación, es vital un conocimiento previo de su estado de salud y clínico, obtenido, si es posible, de sus historiales sanitarios o validado por un especialista sanitario con el objetivo de perfilar muy bien los usuarios en función de sus necesidades de hábitos de vida, restricciones y limitaciones, si las hubiera, y recordatorios de pautas necesarias, directamente relacionadas con su estado clínico, y sus costumbres en el domicilio.

La experiencia con los usuarios finales fue bastante heterogénea, pero hubo un problema inicial y generalizado, que fue la comprensión del propósito final del servicio en sí. Al igual que muchos nuevos avances tecnológicos, esta es una barrera común que algunos residentes tienen problemas o ciertas reticencias para adoptar estas novedades. La solución para estos problemas es mantener el mayor contacto posible con las personas mayores y realizar una comunicación sosegada y constante, en formato de pregunta-respuesta con el fin de que no exista ninguna duda sobre el servicio.

La correcta ubicación de los sensores a la hora de su instalación, y su implantación física en el domicilio, aunque parece algo evidente, no es trivial. Se debe adquirir experiencia en la búsqueda en la mejor ubicación para los distintos sensores, dado que, aunque sean sensores sencillos, de presencia, por ejemplo, se debe chequear que un movimiento en cierta ubicación no “dispara” varios sensores, por ejemplo, o que ciertos elementos en las habitaciones no impiden la activación de los mismos, la sincronización con el “Home Box” se debe verificar, etc… Se necesita personal entrenado y con sensibilidad para estas tareas, dado que son vitales y tienen correlación directa con la calidad de datos recogidos, y por lo tanto, con la calidad de las respuestas ante posibles alertas.

Por otro lado, hay que poner en valor la utilización del estándar UniversAAL. Hemos aprendido que uno de los puntos más fuertes del proyecto ha sido utilizar la capacidad semántica que proporciona, herramienta imprescindible a la hora de desarrollar las aplicaciones alrededor del sistema propuesto. El valor actual de UniversAAL se deriva de la capacidad de reutilizar el modelo estándar de ontologías de la plataforma y adaptarlo a las necesidades del trabajo realizado. También fue muy valioso ver, en el caso de aplicaciones

importadas, como los datos pueden ser compartidos de una manera inteligente y fácil de integrar usando semántica en la interoperabilidad. Con estas utilidades embebidas, ha sido más fácil desarrollar ciertas partes de las aplicaciones y asegurar su calidad y mantenibilidad. Gracias a este trabajo, se ha demostrado que el filtrado de posibles alertas y la detección de las mismas a partir de un sistema automático es más eficiente que los sistemas de “Reglas Heurísticas” o de expertos, que no recogen la especificidad de los patrones de actividad de cada domicilio, y por lo tanto, tienen una precisión menor a la hora de valorar posibles alertas. Otra de las lecciones aprendidas es el hecho de que es vital y crítico un chequeo constante, por cada instalación, de la calidad de los datos, sobre todo en sus dimensiones de completitud y consistencia. Estas dimensiones indican cuántos valores de deben recoger por domicilio y sensor aproximadamente, y si se está recogiendo dicha tasa de valores, y con qué sesgo. Si la completitud no es buena, puede indicar que hay problemas de conectividad, y este hecho es crítico para el sistema, puesto que si no recogemos datos, y por ejemplo, hay una posible alerta de caída, no la detectaremos, y se produciría un verdadero negativo provocado por un fallo en las comunicaciones, y no por el sistema de clasificación. Ha ocurrido algún caso de este tipo, y se ha detectado a posteriori. Esto implica un mal funcionamiento del sistema en su integración, y lo que es peor, puede provocar desconfianza en los cuidadores, que deben estar chequeando cada poco tiempo si hay o no conectividad con el domicilio.

En resumen, con la plataforma desplegada, y el sistema MultiClasificador propuesto, es posible implementar un sistema de teleasistencia sencillo, que, con sensores simples y accesibles, de instalación no cableada y sencilla, asegura un ratio de éxito suficientemente bueno como para poder atender situaciones de riesgo de forma autónoma, y alertar a cuidadores y familiares cuando dichos eventos anómalos se produzcan.

9.2 Posibles Mejoras y Trabajos Futuros

Después de la elaboración de este trabajo, aún se sigue recogiendo a día de hoy datos de residentes en distintos domicilios, lo que está incrementando nuestra base de datos de conocimiento. Nuestros próximos objetivos son expandir el sistema a más domicilios, reforzar los sistemas inteligentes para mejorar la exactitud y precisión del sistema y tratar de generalizar el conocimiento de algunos ancianos a otros (mediante el paradigma de “transfer learning”), todo de una manera no asistida en tiempo real, para evitar el “vacío” de los primeras semanas de aprendizaje basado únicamente en reglas heurísticas generales. Por otro lado, la acumulación de la información que tenemos, (millones de filas referentes a datos brutos de sensores sobre el comportamiento de residentes en domicilios), y la capacidad de analizar estos patrones, gracias a este trabajo, lo vemos como una oportunidad e incluso como un punto de partida en la generación de nuevas innovaciones, en aras de mejorar los sistemas de atención sanitaria, principalmente para las Administraciones Públicas, sistemas sanitarios y de teleasistencia, así como punto de investigación para los Agentes de Salud. Existen ciertas mejoras al sistema, como incluir ciertos calendario estivales, por ejemplo, o indicar a través de la aplicación cuando los residentes abandonan el domicilio por vacaciones, o cuando tienen visitas, o familiares va a convivir con ellos ciertas temporadas, de manera que el sistema automático pueda minimizar el impacto de falsos positivos por comportamientos habituales provocados por este tipo de eventos. Por otro lado, la

amplitud del horizonte de históricos a 25 meses mejoraría la precisión de los modelos, pudiendo contener patrones con mayor estacionalidad dada la repetitividad de los meses, indicador que ahora mismo no se está teniendo en cuenta en los modelos. En ese momento, podremos tener en cuenta indicadores adicionales como el día del mes, o el mes, que ahora no se incluyen como predictores válidos. Otra mejora sustancial es, en base al conocimiento que se está adquiriendo, sustituir, al menos en parte, el control manual de los operadores sobre las alertas en el Sistema de Notificaciones, en base a nuevas “Reglas Heurísticas” incluidas en dicho módulo. Nuestros próximos pasos son, en primer lugar, transformar esta propuesta en una referencia en las plataformas de teleasistencia en casa, reales y comercializables, y por otro lado, permitir a la comunidad investigadora y social de una herramienta que permita mejorar los servicios actuales de teleasistencia sanitaria en domicilios para personas mayores independientes. En siguientes trabajos, queremos reforzar el Módulo de Detección Automática de Patrones aprovechando la información que tenemos actualmente, incrementándola con nueva información adicional obtenida de los registros clínicos, diagnósticos médicos y tratamientos, y cotejarla a un nivel de estudio más epidemiológico sobre cómo puede estar afectando los datos clínicos al comportamiento habitual de diferentes perfiles de pacientes en su vida cotidiana. Como complemento a la extracción de la información clínica, nuestra intención es investigar en el desafío de crear un nuevo método para el análisis automático de datos no estructurados sobre los resúmenes médicos (obtenidos a partir de registros clínicos, usualmente escritos en lenguaje natural) para descubrir nuevas relaciones entre diagnósticos, procedimientos médicos, tratamientos y su correlación con la medicina personalizada en el hogar, y los patrones de comportamiento particulares de cada usuario. En conclusión, la plataforma desarrollada cumple con el objetivo final, que no ha sido otro que proporcionar una solución robusta, fácilmente desplegable y de contenido compartido para garantizar la seguridad de las personas mayores y su seguridad a través de una infraestructura multisensor, conectándose tanto con el cuidador como con la familia, en situaciones teóricamente comprometidas.

Bibliografía

[Aca13] Acampora, G., Cook, D. J., Rashidi, P., & Vasilakos, A. V. (2013). A survey on ambient intelligence in healthcare. Proceedings of the IEEE, 101(12), 2470-2494.

[Ada10] Adami, A. M., Pavel, M., Hayes, T. L., & Singer, C. M. (2010). Detection of movement in bed using unobtrusive load cell sensors. IEEE Transactions on Information Technology in Biomedicine, 14(2), 481-490.

[Ahl15] Ahlin, Č., Stupica, D., Strle, F., & Lusa, L. (2015). medplot: a web application for dynamic summary and analysis of longitudinal medical data based on R. PloS one, 10(4), e0121760.

[Ald11] Aldin, Laden and de Cesare, Sergio, "A literature review on business process modelling: new frontiers of reusability", Enterprise Information Systems 5, 3 (2011), pp. 359-383.

[Ama17] Amatriain, H. G., Merlino, H., Martins, S., & Bianco, S. (2017). Modelo de proceso de gestión para proyectos de ingeniería del conocimiento. In XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017)..

[Arl10] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

[Bam10] Bamis, A., Lymberopoulos, D., Teixeira, T., & Savvides, A. (2010). The BehaviorScope framework for enabling ambient assisted living. Personal and Ubiquitous Computing, 14(6), 473-487.

[Bed12] Bedogni, L., Di Felice, M., & Bononi, L. (2012, November). By train or by car? Detecting the user's motion type through smartphone sensors data. In Wireless Days (WD), 2012 IFIP (pp. 1-6). IEEE.

[Ber08] Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., & Hendler, J. (2008). N3logic: A logical framework for the world wide web. Theory and Practice of Logic Programming, 8(3), 249-269.

[Bou07] Bouchard, B., Giroux, S., & Bouzouane, A. (2007). A keyhole plan recognition model for Alzheimer's patients: First results. Applied Artificial Intelligence, 21(7), 623-658.

[Box70] Box, G. E. P., & Jenkins, G. M. (1970). Times series Analysis Forecasting and Control. Holden-Day San Francisco.

[Brd09] Brdiczka, O., Crowley, J. L., & Reignier, P. (2009). Learning situation models in a smart home. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1), 56-63.

[Bre00] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.

[Bul18] Bulinski, A., & Dimitrov, D. (2018). Statistical estimation of the Shannon entropy. arXiv preprint arXiv:1801.02050.

[Cam10] Campo, E., Chan, M., Bourennane, W., & Estève, D. (2010, August). Behaviour monitoring of the elderly by trajectories analysis. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE (pp. 2230-2233). IEEE.

[Cha09] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[Cha12] Chan, M., Estève, D., Fourniols, J. Y., Escriba, C., & Campo, E. (2012). Smart wearable systems: Current status and future challenges. Artificial intelligence in medicine, 56(3), 137-156.

[Cha14] Chaaraoui, A. A., Padilla-López, J. R., Ferrández-Pastor, F. J., Nieto-Hidalgo, M., & Flórez-Revuelta, F. (2014). A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context. Sensors, 14(5), 8895-8925

[Che04] Chen, H., Perich, F., Finin, T., & Joshi, A. (2004, August). Soupa: Standard ontology for ubiquitous and pervasive applications. In Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004. The First Annual International Conference on (pp. 258-267). IEEE.

[Che11] Chen, T. L., King, C. H., Thomaz, A. L., & Kemp, C. C. (2011, March). Touched by a robot: An investigation of subjective responses to robot-initiated touch. In Proceedings of the 6th international conference on Human-robot interaction (pp. 457-464). ACM.

[Che12] Chen, L., Nugent, C. D., & Wang, H. (2012). A knowledge-driven approach to activity recognition in smart homes. IEEE Transactions on Knowledge and Data Engineering, 24(6), 961-974.

[Che14a] Chen, H., Chen, X., Gu, P., Wu, Z., & Yu, T. (2014). OWL reasoning framework over big biological knowledge network. BioMed research international, 2014.

[Che14b] Chernbumroong, S., Cang, S., & Yu, H. (2014). A practical multi-sensor activity recognition system for home-based care. decision support systems, 66, 61-70.

[Chi10] Chiang, Y. T., Hsu, K. C., Lu, C. H., Fu, L. C., & Hsu, J. Y. J. (2010, October). Interaction models for multiple-resident activity recognition in a smart home. In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on (pp. 3753-3758). IEEE.

[Cic16] Cicirelli, F., Fortino, G., Giordano, A., Guerrieri, A., Spezzano, G., & Vinci, A. (2016). On the design of smart homes: A framework for activity recognition in home environment. Journal of medical systems, 40(9), 200.

[Cri00] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

[Dam12] Damljanovic, D., Stankovic, M., & Laublet, P. (2012, May). Linked data-based concept recommendation: Comparison of different methods in open innovation scenario. In Extended Semantic Web Conference (pp. 24-38). Springer, Berlin, Heidelberg.

[Dan07] Daniele, L., Costa, P. D., & Pires, L. F. (2007, July). Towards a rule-based approach for context-aware applications. In Meeting of the European Network of Universities and Companies in Information and Communication Engineering (pp. 33-43). Springer, Berlin, Heidelberg.

[Das05] Das, S. K., & Cook, D. J. (2005, December). Designing smart environments: A paradigm based on learning and prediction. In International Conference on Pattern Recognition and Machine Intelligence (pp. 80-90). Springer, Berlin, Heidelberg.

[Doc04] Doctor, F., Hagras, H., Callaghan, V., & Lopez, A. (2004, June). An adaptive fuzzy learning mechanism for intelligent agents in ubiquitous computing environments. In Automation Congress, 2004. Proceedings. World (Vol. 16, pp. 101-106). IEEE.

[Fde00] Fdez Riverola, F., & Corchado, J. M. (2000). Sistemas híbridos neuro-simbólicos: una revisión. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 4(11).

[For79] Forgy, C. L. (1979). On the efficient implementation of production systems (Doctoral dissertation, Carnegie-Mellon University).

[For88] Forgy, C. L. (1988). Rete: A fast algorithm for the many pattern/many object pattern match problem. In Readings in Artificial Intelligence and Databases (pp. 547-559).

[Gar14] Garcia-Valverde, T., Muñoz, A., Arcas, F., Bueno-Crespo, A., & Caballero, A. (2014). Heart health risk assessment system: a nonintrusive proposal using ontologies and expert rules. BioMed research international, 2014.

[Gil16] Gil, E. (2016). Big data, privacidad y protección de datos.XIX Edición del Premio Protección de Datos Personales de Investigación de la Agencia Española de Protección de Datos.

[Gir08] Giroux, S., Bauchet, J., Pigot, H., Lussier-Desrochers, D., & Lachappelle, Y. (2008, July). Pervasive behavior tracking for cognitive assistance. In Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments (p. 86). ACM.

[Gjo14] Gjoreski, H., Rashkovska, A., Kozina, S., Lustrek, M., & Gams, M. (2014, May). Telehealth using ECG sensor and accelerometer. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on (pp. 270-274). IEEE.

[Got15] Gottfried, B., Aghajan, H., Wong, K. B. Y., Augusto, J. C., Guesgen, H. W., Kirste, T., & Lawo, M. (2015). Spatial health systems. In Smart Health (pp. 41-69). Springer, Cham.

[Han12] Han, Y., Han, M., Lee, S., Sarkar, A. M., & Lee, Y. K. (2012). A framework for supervising lifestyle diseases using long-term activity monitoring. Sensors, 12(5), 5363-5379..

[Haw03] Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. Journal of chemical information and computer sciences, 43(2), 579-586.

[Her04] Hernández Orallo, J., Ferri Ramirez, C, & Ramirez Quintana, M. J. (2004). Introducción a la Minería de Datos. Pearson Prentice Hall,.

[Hua14] Huang, C. N., & Chan, C. T. (2014). A zigbee-based location-aware fall detection system for

improving elderly telecare. International journal of environmental research and public health, 11(4), 4233-4248

[Hub09] Huber, M., Rodrigues, R., Hoffmann, F., Gasior, K., & Marin, B. (2009). Facts and figures on long-term care. Europe and North America, Wien: European Centre for Social Welfare Policy and Research.

[Igu13] Igual, R., Medrano, C., & Plaza, I. (2013). Challenges, issues and trends in fall detection systems. Biomedical engineering online, 12(1), 66.

[Jak11] Jakkula, V. R., & Cook, D. J. (2011). Detecting Anomalous Sensor Events in Smart Home Data for Enhancing the Living Experience. Artificial intelligence and smarter living, 11(201), 1.

[Jun15] Jung, Y., & Hu, J. (2015). AK-fold averaging cross-validation procedure. Journal of nonparametric statistics, 27(2), 167-179.

[Jup11] Jupp, S., Klein, J., Schanstra, J., & Stevens, R. (2011, December). Developing a kidney and urinary pathway knowledge base. In Journal of biomedical semantics (Vol. 2, No. 2, p. S7). BioMed Central.

[Kal10] Kaluža, B., Mirchevska, V., Dovgan, E., Luštrek, M., & Gams, M. (2010, November). An agent-based approach to care in independent living. In International joint conference on ambient intelligence (pp. 177-186). Springer, Berlin, Heidelberg.

[Kam14] Kamdar, M. R., Zeginis, D., Hasnain, A., Decker, S., & Deus, H. F. (2014). ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of biomedical informatics, 47, 112-130.

[Kar11] Karim, N. A., Ahmad, M., & Mohamed, N. (2011, December). A framework for electronic health record (EHR) implementation impact on system service quality and individual performance among healthcare practitioners. In Proceedings of the 10th WSEAS International Conference on E-Activities (E-ACTIVITIES'11) (pp. 197-201).

[Keo04] Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: A survey and novel approach. In Data mining in time series databases (pp. 1-21).

[Kim09] Kim, K. J., Hassan, M. M., Na, S. H., & Huh, E. N. (2009, December). Dementia wandering detection and activity recognition algorithm using tri-axial accelerometer sensors. In Ubiquitous Information Technologies & Applications, 2009. ICUT'09. Proceedings of the 4th International Conference on (pp. 1-5). IEEE.

[Kon12] Konstantinou, N. (2012). Converting raw sensor data to semantic web triples: a survey of implementation options. International Journal of Sensors Wireless Communications and Control, 2(1), 44-52.

[Kor18] Kor, A. L., Pattinson, C., Yanovsky, M., & Kharchenko, V. (2018). IoT-Enabled Smart Living. In Technology for Smart Futures (pp. 3-28). Springer, Cham.

[Kri14] Krishnan, N. C., & Cook, D. J. (2014). Activity recognition on streaming sensor data. Pervasive and mobile computing, 10, 138-154.

[Kri15] Krishnamurthi, K. A. R. T. H. I. K., Griet, V. R. P., & Jntuh, V. V. B. (2015). Capturing the semantic structure of documents using summaries in supplemented latent semantic analysis. WSEAS Transactions on Computers, 14, 314-323.

[Kur06] Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, 21(1), 1-24.

[Kwo12] Kwon, O., Shim, J. M., & Lim, G. (2012). Single activity sensor-based ensemble analysis for health monitoring of solitary elderly people. Expert Systems with Applications, 39(5), 5774-5783.

[Lar12] Lara, O. D., Pérez, A. J., Labrador, M. A., & Posada, J. D. (2012). Centinela: A human activity recognition system based on acceleration and vital sign data. Pervasive and mobile computing, 8(5), 717-729.

[Lia06] Liao, L., Fox, D., & Kautz, H. (2006). Location-based activity recognition. In Advances in Neural Information Processing Systems (pp. 787-794).

[Lin15] Lin, K. C., Yeh, C. L., & Tsai, H. C. (2015). Developing knowledge-based emr services using semantic web technology. Internet Technology Journal, 16(3), 403-414.

[Liu12] Liu, C., Qi, G., Wang, H., & Yu, Y. (2012). Reasoning with large scale ontologies in fuzzy pD* using MapReduce. IEEE Computational Intelligence Magazine, 7(2), 54-66.

[Liu14] Liu, X., Chen, F., & Lu, C. T. (2014). On detecting spatial categorical outliers. GeoInformatica, 18(3), 501-536.

[Mat05] Ma, T., Kim, Y. D., Ma, Q., Tang, M., & Zhou, W. (2005, August). Context-aware implementation based on CBR for smart home. In Wireless And Mobile Computing, Networking And Communications, 2005.(WiMob'2005), IEEE International Conference on (Vol. 4, pp. 112-115). IEEE.

[Mat12] Matías, L. L., & Sicilia, M. Á. (2012). Combining ontologies and rules with clinical archetypes (Doctoral dissertation, Universidad de Alcalá).

[Mau06] Maurer, U., Smailagic, A., Siewiorek, D. P., & Deisher, M. (2006, April). Activity recognition and monitoring using multiple sensors on different body positions. In Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on (pp. 4-pp). IEEE.

[Med09] Medjahed, H., Istrate, D., Boudy, J., & Dorizzi, B. (2009). A fuzzy logic system for home elderly people monitoring (EMUTEM). Fuzzy Systems.

[Men17] Meng, L., Miao, C., & Leung, C. (2017). Towards online and personalized daily activity recognition, habit modeling, and anomaly detection for the solitary elderly through unobtrusive sensing. Multimedia Tools and Applications, 76(8), 10779-10799.

[Mer14] Merelli, I., Pérez-Sánchez, H., Gesing, S., & D'Agostino, D. (2014). Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. BioMed research international, 2014.

[Mil89] Milligan, G. W. (1989). A validation study of a variable weighting algorithm for cluster analysis. Journal of Classification, 6(1), 53-71.

[Mor09] Moreno Fdz de Leceta, A. & Rincón, M. (2009, June). Access Control to Security Areas Based on Facial Classification. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 254-263). Springer, Berlin, Heidelberg.

[Moz95] Mozer, M. C., Dodier, R. H., Anderson, M., Vidmar, L., Cruickshank, R. F., & Miller, D. (1995). The neural network house: an overview. Current trends in connectionism, 371-380.

[Niu07] Niu, K., Huang, C., Zhang, S., & Chen, J. (2007, May). ODDC: outlier detection using distance distribution clustering. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 332-343). Springer, Berlin, Heidelberg.

[Nou12] Noury, N., & Hadidi, T. (2012). Computer simulation of the activity of the elderly person living independently in a Health Smart Home. Computer methods and programs in biomedicine, 108(3), 1216-1228.

[Noy95] Noyes, J. (1995). A review of:"Human Reliability Analysis: Context and Controli", by ERIK HOLLNAGEL, Academic, London (1993), pp. xxvi+ 336. Ergonomics, 38(12), 2614-2615.

[Ogr10] O'Grady, M. J., Muldoon, C., Dragone, M., Tynan, R., & O'Hare, G. M. (2010). Towards evolutionary ambient assisted living systems. Journal of Ambient Intelligence and Humanized Computing, 1(1), 15-29.

[Oke14] Okeyo, G., Chen, L., & Wang, H. (2014). Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes. Future Generation Computer Systems, 39, 29-43.

[Ord15] Ordóñez, F. J., de Toledo, P., & Sanchis, A. (2015). Sensor-based Bayesian detection of anomalous living patterns in a home setting. Personal and Ubiquitous Computing, 19(2), 259-270.

[Pan14] Pannurat, N., Thiemjarus, S., & Nantajeewarawat, E. (2014). Automatic fall monitoring: a review. Sensors, 14(7), 12900-12936.

[Per10] Percival, J. (2010). Simon Evans, Community and Ageing: Maintaining Quality of Life in Housing with Care Settings, Policy Press, Bristol, UK, 2009, 168 pp. Ageing and Society, 30(7), 1280-1282.

[Per15] Pérez Planells, L., Delegido, J., Rivera-Caicedo, J. P., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. Revista Española de Teledetección, 2015, vol. 44, p. 55-65.

[Phu09] Phua, C., Foo, V. S. F., Biswas, J., Tolstikov, A., Maniyeri, J., Huang, W., ... & Chu, A. K. W. (2009, December). 2-layer erroneous-plan recognition for dementia patients in smart homes. In e-Health Networking, Applications and Services, 2009. Healthcom 2009. 11th International Conference on (pp. 21-28). IEEE.

[Pia14] Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review, 21(5), 1112–1130.

[Qin15] Ni, Q., García Hernando, A. B., & de la Cruz, I. P. (2015). The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development. Sensors, 15(5), 11312-11362.

[Qui90] Quinlan, J. R. (1990). Learning logical definitions from relations. Machine learning, 5(3), 239-266.

[Ran11] Rantz, M. J., Skubic, M., Koopman, R. J., Phillips, L., Alexander, G. L., Miller, S. J., & Guevara, R. D. (2011, June). Using sensor networks to detect urinary tract infections in older adults. In e-Health Networking Applications and Services (Healthcom), 2011 13th IEEE International Conference on (pp. 142-149). IEEE.

[Ras09] Rashidi, P., & Cook, D. J. (2009). Keeping the resident in the loop: Adapting the smart home to the user. IEEE Transactions on systems, man, and cybernetics-part A: systems and humans, 39(5), 949-959.

[Ras10] Rashidi, P., & Cook, D. J. (2010, December). Mining sensor streams for discovering human activity patterns over time. In Data Mining (ICDM), 2010 IEEE 10th International Conference on (pp. 431-440). IEEE.

[Ras11] Rashidi, P., Cook, D. J., Holder, L. B., & Schmitter-Edgecombe, M. (2011). Discovering activities to recognize and track in a smart environment. IEEE transactions on knowledge and data engineering, 23(4), 527-539.

[Ras13] Rashidi, P., & Mihailidis, A. (2013). A survey on ambient-assisted living tools for older adults. IEEE journal of biomedical and health informatics, 17(3), 579-590.

[Rib15] Riboni, D., Bettini, C., Civitarese, G., Janjua, Z. H., & Helaoui, R. (2015, March). Fine-grained recognition of abnormal behaviors for early detection of mild cognitive impairment. In Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on (pp. 149-154). IEEE.

[Rib16] Riboni, D., Sztyler, T., Civitarese, G., & Stuckenschmidt, H. (2016, September). Unsupervised recognition of interleaved activities of daily living through ontological and probabilistic reasoning. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 1-12). ACM.

[Sak18] Sakr, S., Wylot, M., Mutharaju, R., Le Phuoc, D., & Fundulaki, I. (2018). Fundamentals. In Linked Data (pp. 9-32). Springer, Cham.

[Sal10] Salah, A. A., Gevers, T., Sebe, N., & Vinciarelli, A. (2010, August). Challenges of human behavior understanding. In International Workshop on Human Behavior Understanding (pp. 1-12). Springer, Berlin, Heidelberg.

[Sav10] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), 507-513.

[Sin95] Singh, A., & Nocerino, J. M. (1995). Robust procedures for the identification of multiple outliers. In Chemometrics in Environmental Chemistry-Statistical Methods (pp. 229-277). Springer, Berlin, Heidelberg.

[Sma11] Smarr, C. A., Fausset, C. B., & Rogers, W. A. (2011). Understanding the potential for robot assistance for older adults in the home environment. Georgia Institute of Technology.

[Sow14] Sowa, J. F. (Ed.). (2014). Principles of semantic networks: Explorations in the representation of knowledge. Morgan Kaufmann.

[Spa14] Spagnolo, P., Mazzeo, P., & Distante, C. (2014). Human behavior understanding in networked sensing. Springer.

[Sta18] Stavrotheodoros, S., Kaklanis, N., & Tzovaras, D. (2018). A Personalized Cloud-Based Platform for AAL Support to Cognitively Impaired Elderly People. In Precision Medicine Powered by pHealth and Connected Health (pp. 87-91). Springer, Singapore.

[Sur13] Suryadevara, N. K., Mukhopadhyay, S. C., Wang, R., & Rayudu, R. K. (2013). Forecasting the behavior of an elderly using wireless sensors data in a smart home. Engineering Applications of Artificial Intelligence, 26(10), 2641-2652.

[Sur14] Suryadevara, N. K., & Mukhopadhyay, S. C. (2014). Determining wellness through an ambient assisted living environment. IEEE Intelligent Systems, 29(3), 30-37.

[Tan05] Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time-series motif from multi-dimensional data based on MDL principle. Machine Learning, 58(2-3), 269-300.

[Tap04] Tapia, E. M., Intille, S. S., & Larson, K. (2004, April). Activity recognition in the home using simple and ubiquitous sensors. In International conference on pervasive computing (pp. 158-175). Springer, Berlin, Heidelberg.

[Thi07] Thirumalainambi, R. (2007). Pitfalls of JESS for Dynamic Systems. In Artificial Intelligence and Pattern Recognition (pp. 491-494).

[Tom18] Tomforde, S., Dehling, T., Haux, R., Huseljic, D., Kottke, D., Scheerbaum, J., ... & Wolf, K. H. (2018). Towards Proactive Health-enabling Living Environments: Simulation-based Study and Research Challenges.

[Vuo11] Vuong, N. K., Chan, S., Lau, C. T., & Lau, K. M. (2011, May). Feasibility study of a real-time wandering detection algorithm for dementia patients. In Proceedings of the First ACM MobiHoc Workshop on Pervasive Wireless Healthcare (p. 11). ACM.

[Wad08] Wadley, V. G., Okonkwo, O., Crowe, M., & Ross-Meadows, L. A. (2008). Mild cognitive impairment and everyday function: evidence of reduced speed in performing instrumental activities of daily living. The American Journal of Geriatric Psychiatry, 16(5), 416-424.

[Wan09] Wang, L., Gu, T., Tao, X., & Lu, J. (2009, November). Sensor-based human activity recognition in a multi-user scenario. In European Conference on Ambient Intelligence (pp. 78-87). Springer, Berlin, Heidelberg.

[Wan18] Wang, M. (2018). A Web-Based System for KPI-Based Workplace Learning. In E-Learning in the Workplace (pp. 129-138). Springer, Cham.

[Yan09] Yang, Y., & Webb, G. I. (2009). Discretization for naive-Bayes learning: managing discretization bias and variance. Machine learning, 74(1), 39-74.

[Yan14] Yang, Y., & Huang, S. (2014). Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models–a case study. Forestry: An International Journal of Forest Research, 87(5), 654-662.

[Yej15a] Ye, J., Stevenson, G., & Dobson, S. (2015). USMART: An unsupervised semantic mining activity recognition technique. ACM Transactions on Interactive Intelligent Systems (TiiS), 4(4), 16.

[Yej15b] Ye, J., Stevenson, G., & Dobson, S. (2015). KCAR: A knowledge-driven approach for concurrent activity recognition. Pervasive and Mobile Computing, 19, 47-70.

[Zhe10] Zheng, V. W., Zheng, Y., Xie, X., & Yang, Q. (2010, April). Collaborative location and activity recommendations with GPS history data. In Proceedings of the 19th international conference on World wide web (pp. 1029-1038). ACM.

[Zho07] Zhou, X., Zhang, X., & Hu, X. (2007, January). Semantic Smoothing of Document Models for Agglomerative Clustering. In IJCAI (pp. 2928-2933).

[Zhu09] Zhuang, X., Huang, J., Potamianos, G., & Hasegawa-Johnson, M. (2009, April). Acoustic fall detection using Gaussian mixture models and GMM supervectors. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on (pp. 69-72). IEEE.