UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Análisis avanzado de biomarcadores de imagen de volumetría cerebral extraídos de resonancia magnética

Trabajo Fin de Máster

Máster Universitario en Ingeniería y Tecnología de Sistemas Software

Autora: Mª José Martínez Suárez

Tutora: Mª José Ramírez Quintana

Tutor externo: Fabio García Castro

2020-2021

Análisis avanzado de biomarcadores de imagen de volumetría cerebral extraídos de resonancia magnética

Resumen

La medicina actual, y todas las ciencias biomédicas y de la salud en general, se hallan inmersas en un cambio de paradigma. La cantidad de información relativa a pacientes que se maneja actualmente necesita de un tratamiento regido por los procedimientos de la ciencia de datos. Con ello, los métodos de diagnóstico están abocados a un cambio radical. La imagen radiológica presenta un cúmulo de información que puede ser convenientemente analizada, medida e interpretada. A partir de una imagen radiológica podemos extraer biomarcadores de imagen. Esta nueva información cuantitativa permite, una vez tratada por la ciencia de datos, su explotación posterior para poder determinar y adelantarse a enfermedades futuras. Este TFM se propone dilucidar el procedimiento para, a partir de una base de datos con información de biomarcadores de imagen de volumetría cerebral, arbitrar un sistema de predicción o determinación de la edad cerebral de un paciente, de manera que, contrastada esta con su edad real, pueda revelar la existencia potencial de futuras patologías.

Palabras clave: biomarcadores de imagen, volumetría cerebral, minería de datos, predicción de enfermedades.

Abstract

Modern medicine, as well as health and biomedical sciences in a whole, are currently undergoing deep changes in their models. These sciences have to manage a big quantity of information on their patients. This volume of information can be treated according to the rules of big data proceedings, and this is why diagnostics methods are about to change dramatically. Radiology images offer so much information that it can be analysed, measured and interpreted. From a radiology image we can obtain imaging biomarkers that, when properly treated according to the big data analysis, can be useful in order to predict and prevent future illnesses. This TFM is grounded in a data base of imaging biomarkers on brain volumetry and is trying to elucidate a proceeding to predict or establish how old a patient's brain is, in order to compare it with the real patient's age and to predict, through this comparaison, the hypothetical manifestation of coming diseases.

Keywords : imaging biomarkers, brain volumetry, data mining, diseases prediction.

Tabla de contenidos

CAPÍTULO 1: INTRODUCCIÓN

1.1. Motivación

La medicina actual, y todas las ciencias biomédicas y de la salud en general, se hallan inmersas en un cambio de paradigma. Desde hace años asistimos al surgimiento de un nuevo modelo de ciencia médica que tiende, cada vez más, a una práctica proactiva, que busca la preservación de la salud de la población, así como un tratamiento cada vez más preciso y hasta cierto punto personalizado de diversas enfermedades. Cada vez parece más evidente que el combate contra muchas de las patologías que nos acechan a los humanos resulta tanto más exitoso cuanto más ajustados a los casos individuales de los afectados sean los tratamientos. Y no solo los tratamientos. También los métodos de diagnóstico están abonados a un cambio radical en la manera de proceder que incide en la precisión de sus constataciones, llegándose a alcanzar resultados que hasta ahora resultaban inimaginables.

En este campo particular de las técnicas de diagnóstico, existe actualmente una gran cantidad de ellas que nos permiten medir aspectos extremadamente sutiles y relevantes de las enfermedades, tanto para su diagnóstico como para su seguimiento; aspectos que hasta ahora no eran apreciables al ojo humano del profesional médico, por muy experimentado que este fuese. Y ello, aunque la experiencia y capacidad de diagnóstico de este no ha dejado de tener, desde luego, un papel esencial en la práctica médica. Pero precisamente esa experiencia y conocimientos se ven actualmente reforzados y potenciados por el uso de las variadas técnicas a las que nos referimos.

Analicemos, por ejemplo, el ámbito de la radiología. La imagen radiológica ha sido tradicionalmente considerada como una técnica básica de diagnóstico en la que la experiencia y conocimientos del radiólogo desempeñaban un papel crucial. Este cambio de enfoque en la medicina del que venimos hablando, en cambio, se enfrenta a la imagen radiológica como lo que en realidad es: un ingente cúmulo de información que puede ser conveniente analizada, medida, interpretada. La cantidad de información implícita en la imagen radiológica, y solo interpretable cuando es tratada con procedimientos estadísticos, es tal que sobrepasa lo que el experto proceder humano es capaz de procesar. A partir de una imagen radiológica, por ejemplo, no es posible extraer biomarcadores de imagen, como el volumen de sustancia gris, si no utilizamos técnicas de análisis computacionales. El estudio de los biomarcadores y su análisis a través de las imágenes halla su definición más autorizada en el estudio coordinado por Martí-Bonmatí y Alberich-Bayarri[1]. Se considera un biomarcador de imagen a cualquier medida que se pueda obtener de manera objetiva de la imagen, siempre que sea reproducible y que sirva como indicador de un proceso patológico, de la evolución del mismo o de una respuesta al tratamiento.

Esta nueva información cuantitativa que es posible extraer hoy en día en la práctica clínica permite su explotación posterior mediante infinidad de técnicas, desde redes neuronales convolucionales hasta modelos de regresión lineal o logística, según las características de la cuestión que se esté estudiando. Nos hallamos, en definitiva, ante una nueva realidad en la manera de proceder de la práctica diagnóstica, muy distinta al enfoque

[1] Luis Martí-Bonmatí y Ángel Alberich-Bayarri, eds. *Imaging Biomarkers. Development and Clinical Integration*. Springer-Verlag, 2017

tradicional: el análisis computacional de los datos de la imagen radiológica, y solo accesibles y manejables mediante dicho análisis avanzado, nos puede permitir, entre otros:

- La segmentación automática del hígado mediante técnicas de inteligencia artificial (IA) a partir de imagen por resonancia magnética (RM), para posteriormente calcular la fracción grasa y el hierro presentes en el tejido [Jim et al. 2021].
- La detección automática de anomalías en placas de rayos X (RX) de tórax[2].
- La estimación del riesgo de fractura en hueso a partir de imágenes obtenidas con tomografía computarizada (TC) [3].
- La segmentación automática de los distintos tejidos cerebrales a partir de imagen de RM, incluyendo la parcelación de las áreas pertenecientes a la sustancia gris, obteniendo los volúmenes de áreas tan específicas como, por ejemplo, el hipocampo. Basándose en los resultados obtenidos con la segmentación de áreas de sustancia gris, y disponiendo de valores de normalidad para sujetos sanos en los rangos de edad bajo estudio, sería posible la generación de modelos para la predicción de la edad del paciente en función de sus biomarcadores de imagen de volumetría cerebral. Esto permitiría detectar anomalías potenciales como consecuencia de la posible disparidad entre la edad real del cerebro estudiado (entendido como la desviación de los volúmenes cerebrales respecto a la normalidad para su rango de edad) y de la persona. Ante ese escenario, sería posible, como decíamos anteriormente, actuar de manera preventiva, no reactiva, dependiendo, por supuesto, de los tratamientos disponibles para patologías concretas. Lo que, no cabe duda, supone un desiderátum de la medicina actual.

El uso de las nuevas tecnologías en la toma de decisiones en el ámbito de la medicina es un factor clave para poder determinar y adelantarse a enfermedades futuras[4]. La prevención, el diagnóstico precoz, la anticipación en la evolución de las patologías son, todas ellas, cuestiones a las que la nueva medicina debe atender, obligada por las condiciones sociales y económicas en que la práctica médica debe proceder. En tal sentido, la pandemia mundial que en 2020 sumió al mundo en el desconcierto más absoluto es un recordatorio más de la importancia vital que, en un mundo de creciente complejidad, tiene la capacidad de las sociedades, con la ciencia al frente de ellas, para anticipar escenarios futuros. Baste

---

[2] Liang, C. H., Liu, Y. C., Wu, M. T., Garcia-Castro, F., Alberich-Bayarri, A., & Wu, F. Z. (2020). "Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice". *Clinical radiology*, *75*(1), 38-45.

[3] Arnold, E. L., Clement, J., Rogers, K. D., Garcia-Castro, F., & Greenwood, C. (2020). "The use of μCT and fractal dimension for fracture prediction in osteoporotic individuals". *Journal of the mechanical behavior of biomedical materials*, *103*, 103585.

[4] De hecho, y a pesar de su carácter relativamente novedoso, el uso de la ciencia de datos en la detección de potenciales patologías cuenta ya con una importante bibliografía. Véanse al respecto los siguientes títulos: Milovic, B., & Milovic, M. (2012). "Prediction and decision making in health care using data mining". *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *1*(12), 126. 2012. Y esta otra referencia de título bien elocuente: Murdoch, T. B., & Detsky, A. S. The inevitable application of big data to health care. *Jama*, *309*(13), 1351-1352. 2013.

mencionar las numerosas iniciativas para estudiar la evolución de esta enfermedad a partir de los datos y el gran número de estudios y artículos científicos publicados desde que se inició la pandemia[5]. Todos estos estudios han puesto de manifiesto que, en el nivel más cercano a las personas, la cantidad de información relativa a pacientes que se maneja actualmente necesita de un tratamiento y una organización que permita obtener información acerca de la salud de dichos pacientes y nos ayude a ver el futuro para poder llevar a cabo decisiones con crecientes posibilidades de éxito.

1.2 Objetivos

El presente TFM se enmarca en un proyecto que ha sido posible gracias a la colaboración de la empresa Quibim. Quibim S.L. es una empresa líder en el sector del análisis de imagen médica mediante la aplicación de técnicas que están a la vanguardia del procesado de imagen y la inteligencia artificial (IA). Las técnicas desarrolladas por esta empresa permiten la extracción de biomarcadores de imagen de las imágenes adquiridas mediante diversas modalidades de imagen radiológica, como RM, TC o RX, y de medicina nuclear, como tomografía por emisión de positrones (PET). Por lo tanto, y en la línea con lo expresado en el apartado anterior, la empresa Quibim S.L. desarrolla su labor en un contexto de medicina de vanguardia, extrayendo, clasificando y analizando información que escaparía a la observación y medios de diagnóstico empíricos de la práctica médica clásica, sobrepasando lo que "el ojo humano puede ver" y mejorando, por tanto, las técnicas de diagnóstico radiológico gracias a los ya mencionados biomarcadores de imagen.

El objetivo general de este TFM sería dilucidar el procedimiento para, a partir de una base de datos con información de biomarcadores de imagen de volumetría cerebral, poder arbitrar un sistema de predicción o determinación fehaciente de la edad de un paciente con respecto a sus volúmenes de tejido cerebral, es decir, estableciendo si existe atrofia fuera de rangos de normalidad en su rango de edad, con la finalidad de poder revelar la potencial existencia de enfermedades. Esta posibilidad sería revelada si el sistema llegase a advertir una falta de correlación entre la edad real y la edad cerebral, de acuerdo con los datos aportados por los biomarcadores. Hemos expresado este objetivo en condicional de manera intencionada, pues de entrada hemos de decir que, por la complejidad del mismo y la amplitud de los fenómenos implicados, el presente TFM solo es un primer paso que pretende abrir la puerta a ulteriores investigaciones. Nuestro trabajo, por lo tanto, quiere arrojar luz en una línea de investigación sobre el apoyo que las tecnologías de la información pueden aportar a la biomedicina, en un campo, además, el de las enfermedades

[5] Ofrecemos las referencias siguientes, tan recientes como reciente es el problema que abordan:

-Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), 337-339. 2020.

-Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., ... & Islam, M. T. "Can AI help in screening viral and COVID-19 pneumonia?". *IEEE Access*, *8*, 132665-132676. 2020.

neurodegenerativas, que consideramos de crucial importancia por sus crecientes implicaciones sociales. Considérese, si no, el caso, por ejemplo, de la atrofia cerebral, un proceso que se produce por causas patológicas de distinta índole, con síntomas que varían de manera significativa dependiendo de las áreas afectadas: la atrofia frontotemporal conlleva un deterioro cognitivo específico, mientras que la atrofia del córtex motor provocará distintos grados de alteración en la motricidad del paciente. Las consecuencias de estas patologías para los pacientes afectados y para su entorno social son de una considerable gravedad. Sin embargo, es habitual encontrar atrofia cerebral en personas de edad avanzada sin que esto conlleve ningún proceso patológico, con lo que la atrofia puede considerarse una consecuencia normal del envejecimiento. Dilucidar gracias a los biomarcadores, si ello fuera posible, cuándo los datos detectados sugieren un envejecimiento normal y cuándo advierten de una patología, supondría, sin duda, un gran avance para el control y detección temprana de atrofia cerebral.

En su campo de trabajo Quibim ha desarrollado algoritmos que permiten la segmentación automática de sustancia gris, sustancia blanca y líquido cefalorraquídeo, y que permiten llegar a explorar incluso áreas de tamaño muy reducido, como el hipocampo, pero que tienen una importancia extraordinaria para diagnosticar diversos tipos de demencia (figura 1).

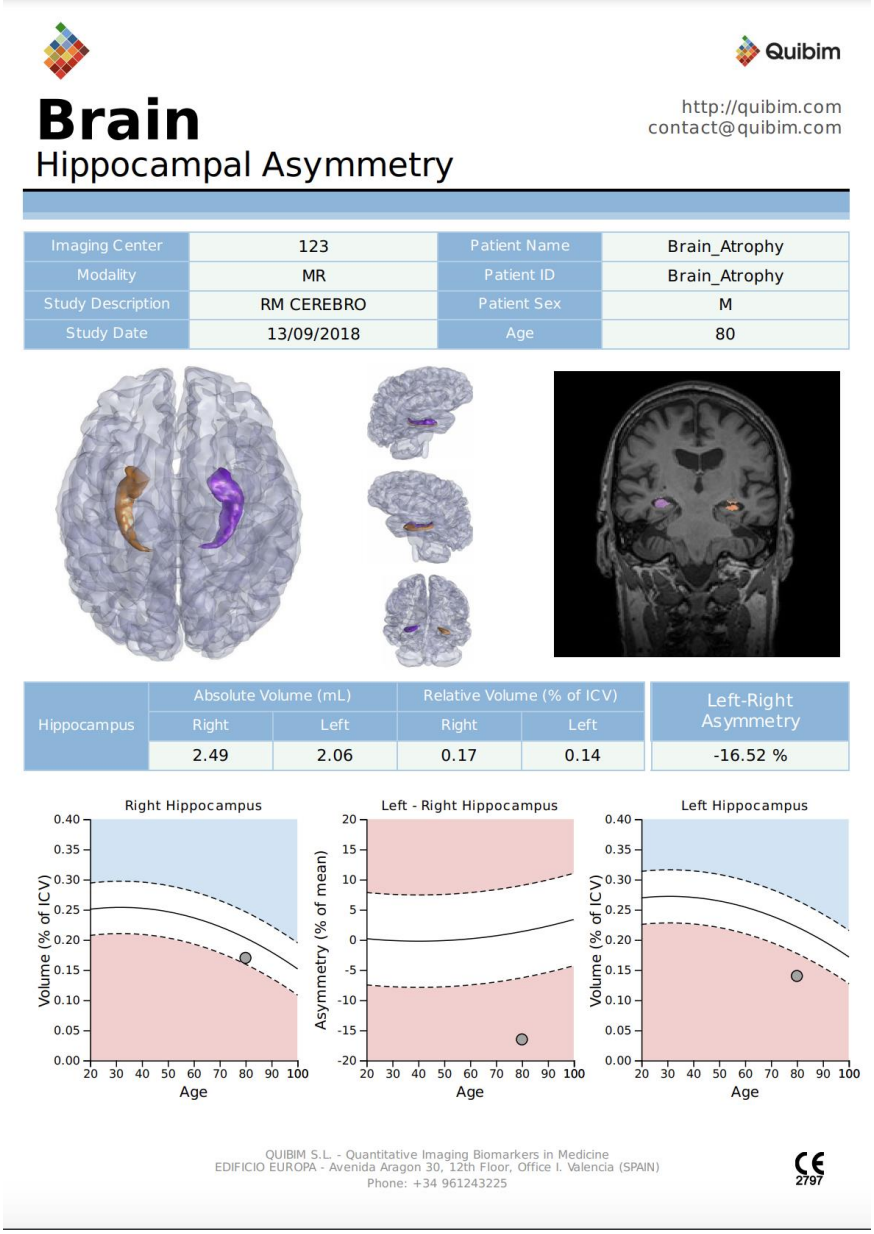


Figura 1: Informe ejemplo de atrofia y asimetría hipocampal de Quibim S.L. (imagen cortesía de Quibim S.L.)

En otras palabras, los métodos y algoritmos desarrollados por Quibim permiten conocer de manera objetiva los volúmenes de estas áreas y evaluar si se encuentran dentro de valores considerados normales, con lo que esto supone en cuanto a mejora del diagnóstico por imagen y, por consiguiente, el seguimiento del paciente de forma muy estrecha y precisa. Al tratarse de técnicas automatizadas, es posible obtener de manera sistemática estas medidas para todos los sujetos que se someten a una RM, generándose así una base de datos cuantitativa en cualquier centro que disponga del sistema Quibim Precision® de Quibim S.L. Sin embargo, hasta el momento Quibim S.L. no ha realizado estudios para predecir la edad cerebral en función de los volúmenes obtenidos y de las desviaciones de estos datos respecto a la normalidad. Este tipo de aproximaciones aporta

un punto de vista muy novedoso a la hora de, no sólo detectar diversas patologías que provocan atrofia, sino también para detectar el envejecimiento prematuro.

1.3 Metodología

El método de trabajo de este TFM consiste esencialmente en un análisis exhaustivo de datos, análisis que vamos a llevar a cabo basándonos en la metodología CRISP-DM[6].

CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada de las etapas o fases, también denominado ciclo de vida, de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas.

La metodología CRISP-DM contempla el proceso de análisis de datos de forma científica y generalmente adaptado a un entorno empresarial, pues el análisis y tratamiento de los datos es llevado a cabo por empresas especializadas, y sus objetivos suelen adecuarse a las necesidades y demandas de un cliente. En un contexto de creciente especialización de las ciencias biomédicas, la metodología CRISP-DM y su capacidad de tratar ingentes volúmenes de datos puede tener un papel determinante en la investigación y tratamiento de enfermedades de indudable trascendencia social y humana.

Todas las técnicas de extracción y tratamiento de datos, que conforman el campo más amplio de lo que usualmente se conoce como minería de datos, constan de una serie de fases bien definidas y protocolizadas. La que vamos a utilizar en nuestra investigación, la ya mencionada CRISP-DM, consta de seis fases que aparecen gráficamente explicadas en la figura siguiente:

[6] Para la descripción de la tecnología CRISP-DM nos servimos de la referencia clásica al respecto: Wirth, R., & Hipp, J. "CRISP-DM: Towards a standard process model for data mining". *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). Londres, abril de 2000. Springer-Verlag.

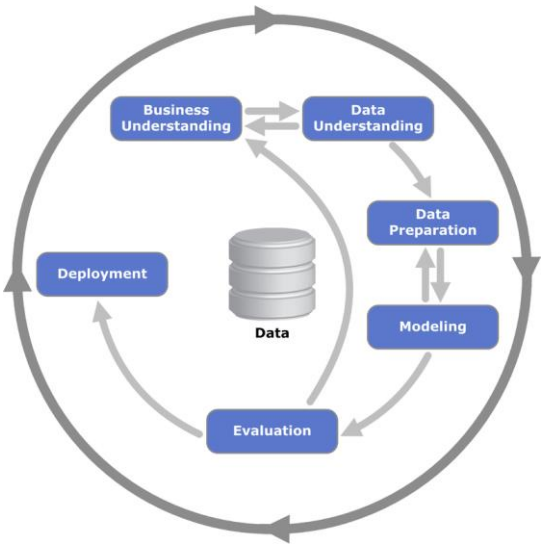


Figura 2: Ciclo de vida de la metodología CRISP-DM [Fuente: Wikipedia]

La metodología CRISP-DM, por lo tanto, establece un proyecto de minería de datos como una secuencia de fases, cada una de ellas con diferentes objetivos, fácilmente deducibles del nombre de cada una de las mismas:

1. Análisis del problema: en este punto se pretende conocer el objetivo que pretendemos alcanzar. En el caso de nuestro TFM, el problema objeto de análisis es la necesidad de predecir futuras patologías cerebrales. Esta "necesidad" viene dictada por la indudable trascendencia médica y social de este tipo de patologías.
2. Descripción de los datos. Consiste, en efecto, en una descripción del contenido de tales datos y de todas las variables que los conforman. Esta es la labor que llevamos a cabo en el capítulo 3 del presente trabajo, donde procedemos a una descripción de las variables documentadas de cada uno de los datos incluidos en la base.
3. Preparación de los datos. Antes de comenzar a realizar un modelado, es decir, de someter los datos al tratamiento de un modelo concreto, debemos prepararlos para cada modelo, para, de esta forma, poder seleccionar la mejor técnica de modelado.
4. Modelado. En nuestro trabajo vamos a someter los datos a distintas técnicas de modelado, que son las que se describen en el capítulo 2 y posteriormente son implementadas en el capítulo 4: regresión lineal, regresión logística, árbol de decisión, *random forest* y red neuronal.
5. Evaluación. Una vez los datos han sido sometidos a distintas técnicas de modelado, y a la vista de los resultados que cada una de ellas arroja, son evaluados con arreglo a diferentes métricas (sistemas de evaluación que arrojan resultados numéricos sobre la exactitud y validez del modelo).
6. Explotación. Con este término (de indudables resonancias económico-empresariales), nos referimos al fin ideal de toda investigación, que sería su aplicación práctica en el ámbito económico, científico, empresarial o cualquier otro en cuyo contexto había surgido la necesidad de la investigación.

En nuestro caso, la culminación ideal de la línea de investigación emprendida sería determinar un sistema capaz de predecir, a partir de los datos contenidos en las imágenes

radiológicas, la futura manifestación de patologías cerebrales. Tal y como ya hemos anunciado (vid. supra, 1.2. "Objetivos"), el presente TFM aspira a erigirse en una señal que indique los pasos a seguir en el futuro para alcanzar tal objetivo, la consecución del cual solo sería posible explotando al máximo el trabajo de equipos multidisciplinares y complementarios, contexto en que actualmente se desarrolla la investigación científica.

Hasta aquí, la explicación de cada una de las fases del método de trabajo CRISP-DM. Cabe añadir, además, que la secuencia de estas fases no es rígida ni unívoca, ni tampoco está estrictamente predeterminada. Al contrario, la metodología permite movimientos hacia adelante y hacia atrás entre diferentes fases. Es el resultado de cada fase, una vez concluida, el que determina cuál de ellas, o incluso qué tarea particular de una fase, hay que desarrollar a continuación. Las flechas que aparecen en el esquema anterior solo indican las dependencias más importantes y frecuentes.

El círculo externo en el que se inscribe todo el esquema simboliza la naturaleza cíclica que tienen los proyectos de análisis de datos. El proyecto no se termina una vez que la solución se despliega. Por el contrario, la información descubierta durante todo el proceso y la solución desplegada al final del mismo pueden producir nuevas iteraciones del modelo. Los procesos de análisis subsecuentes se beneficiarán de las experiencias previas, creándose así una serie de interacciones a lo largo del proceso de gran complejidad y riqueza.

Este sería, en esencia, la manera en que trabaja la metodología CRISP-DM que adoptamos como columna vertebral de nuestro TFM. Aunando lo dicho sobre este método de trabajo con lo expuesto en el apartado referente a la motivación y con los objetivos declarados más arriba, pensamos que conviene ajustar la estructura de la memoria de este trabajo a los apartados que especificamos a continuación.

1.4. Estructura de la memoria

A tenor de lo ya explicado en la sección de motivación, de los objetivos igualmente especificados en la sección correspondiente, y la metodología descrita, todo lo cual compone el grueso del capítulo primero del presente trabajo, este se organiza de la siguiente manera:

En el capítulo dos describimos los conceptos previos técnicos necesarios para la comprensión del trabajo realizado, así como las herramientas de que nos hemos provisto para internarnos en el análisis de los datos. El capítulo tres lleva a cabo una descripción de los datos de partida, así como del proceso de "comprensión" y "limpieza" de los mismos, a fin de poder trabajar directamente con ellos. En los capítulos cuatro y cinco desarrollamos las dos aproximaciones definidas para dar solución al problema planteado: aproximación directa, que aborda la estimación de la edad de los pacientes directamente como una tarea de regresión o bien de clasificación (capítulo cuatro), mientras que en el capítulo cinco presentamos la aproximación en la que primero se segmenta el conjunto de datos en grupos para pasar a estimar la edad en cada uno de los grupos. Finalmente, en el capítulo seis se indican las conclusiones y las futuras líneas de investigación y trabajo que, a nuestro juicio, quedan abiertas tras las iniciativas apuntadas en el presente TFM.

CAPÍTULO 2: CONCEPTOS PREVIOS

2.1.- Minería de datos

Una vez introducida la metodología que vamos a utilizar, CRISP-DM, y antes de abordar el problema que queremos resolver, debemos hacer una referencia a la minería de datos, puesto que éste es el novedoso procedimiento de investigación en cuyo ámbito se desarrolla nuestro trabajo

La **minería de datos** o **exploración de datos** es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. La minería de datos utiliza métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Recientemente, se ha popularizado el término *data science* para referirse a la ciencia centrada en el estudio de los datos, la cual por lo tanto aglutina a la minería de datos.

La tarea de minería de datos real es el análisis automático o semiautomático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos. Estos patrones (también llamados modelos) pueden ser de dos tipos:

- **descriptivos:** explican las interrelaciones y afinidades entre los atributos o los datos.
- **predictivos**: predicen valores futuros de atributos de interés, que se denominan variables objetivo o dependientes, en función de los valores del resto de atributos, que reciben el nombre de variables predictivas o independientes.

Entre los problemas que se pueden resolver mediante modelos descriptivos distinguimos el agrupamiento y las reglas de asociación, mientras que entre los problemas descriptivos se encuentran la clasificación y la regresión.

2.2.- Técnicas de modelización predictiva

El análisis predictivo es, tal vez, una de las tareas potencialmente más interesantes de la minería de datos y es, además, una de las más útiles para las necesidades de la mayoría de las empresas que precisan del análisis de datos para llevar a cabo sus cometidos. Son muchos los sectores que utilizan el análisis predictivo, mediante el cual pueden llegar a anticipar el comportamiento de sus clientes, para, según el sector de que se trate, ajustar su oferta, o los precios, o prever el sentimiento futuro que se experimentará hacia su marca, u optimizar su productividad, o prevenir enfermedades... Las potenciales aplicaciones del análisis predictivo en entornos complejos son extensísimas.

Existen dos tipos de modelos predictivos: de regresión y de clasificación.

Los modelos de regresión son los que nos permiten predecir valores numéricos. Por ejemplo, en el ámbito comercial, cuál es el beneficio estimado que obtendremos de un determinado cliente, o cuáles van a ser las ventas de un producto.

Los modelos de clasificación permiten predecir variables categóricas, es decir variables que pueden tomar un número limitado de valores. A estos valores habitualmente se les denomina clases. Por ejemplo, y siguiendo con el caso del ámbito comercial, podríamos clasificar a los clientes de una compañía según su grado de inclinación a abandonar los servicios de la misma, o bien podemos clasificar los clientes en buenos o no dependiendo de su perfil de compra. Cuando solo hay dos posibles valores o clases, por ejemplo, sí y no (o alternativamente, 0 y 1) se dice que el problema (y el modelo) son binarios; mientras que si el número de clases es mayor que dos entonces el problema es multiclase. También es posible que los modelos predigan una probabilidad de pertenencia a una clase en lugar de la propia etiqueta de clase. Es decir, nos pueden decir, por ejemplo, que un cliente tiene una probabilidad de abandono de la compañía del 89%.

Aunque hay algunas técnicas que son específicas de clasificación y otras de regresión, la mayoría de las técnicas pueden usarse para resolver ambos tipos de problemas[1]. A continuación, se describe someramente el modo en que proceden algunas de las principales técnicas predictivas que se van a utilizar en este trabajo.

1.Regresión lineal

La regresión lineal es una técnica de modelado estadístico que estima el valor de la variable a predecir como una combinación lineal del resto de las variables predictoras. Cuando el problema es univariante (es decir, una única variable predictora) estamos ante un modelo de regresión lineal simple, mientras que cuando hay varias variables predictoras nos encontramos ante un modelo de regresión lineal múltiple. Los modelos de regresión lineal simple tienen la ventaja de que pueden representarse gráficamente mediante una recta tal y como se muestra en la figura 3.

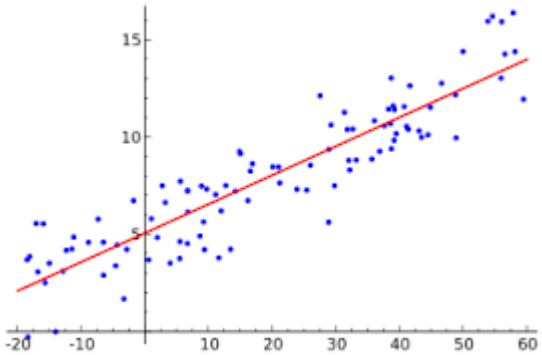


Figura 3: Representación gráfica de una regresión lineal para el caso univariante [Fuente: Wikipedia].

La regresión requiere que todas las variables predictoras sean cuantitativas. No obstante, existen extensiones para tratar variables dependientes multicategóricas y/o ordinales, tales como la regresión politómica. Hoy en día, casi todas las implementaciones de esta técnica permiten tratar variables predictoras categóricas, que son automáticamente

transformadas en cuantitativas (aplicando alguna de las técnicas que se han propuesto en la literatura).

El entrenamiento de una regresión consiste en estimar el valor de los coeficientes de la combinación lineal (es decir, el término independiente o *intercept* y los coeficientes que acompañan a cada variable predictora). El método más utilizado es el de los mínimos cuadrados que consiste en minimizar la suma de las diferencias al cuadrado entre los valores reales y los valores predichos.

Esta técnica se adapta a una amplia variedad de situaciones, el análisis de regresión se utiliza para predecir un amplio rango de fenómenos, desde medidas económicas hasta diferentes aspectos del comportamiento humano. En física se utiliza para caracterizar la relación entre variables o para calibrar medidas.

2.Regresión logística

La regresión logística es un método de clasificación basado en regresión que se usa para predecir el resultado de una variable categórica binaria cuyos valores se representan como 1 y 0. Básicamente, consiste en convertir la variable respuesta, que es cualitativa, en una probabilidad, que es cuantitativa. Esto se consigue aplicando la función logística. La extensión de esta técnica para tratar el caso multiclase es conocida como *logit multinomial*.

Una de las principales aplicaciones de un modelo de regresión logística es clasificar la variable cualitativa en función de valor que tome el predictor. Para conseguir esta clasificación, es necesario establecer un umbral de probabilidad a partir del cual se considera que la variable pertenece a uno de los niveles. En el caso binario, como la probabilidad de pertenecer a una clase es un valor comprendido entre 0 y 1, se suele usar el umbral 0.5, tal que si la probabilidad estimada es mayor que 0.5 se predice la clase positiva (clase 1) y, se predice clase negativa (alternativamente, clase 0) en caso contrario.

Un ejemplo de este modelo sería calcular la probabilidad de obtener una matrícula de honor al final del bachillerato en función de la nota que se ha obtenido en matemáticas.

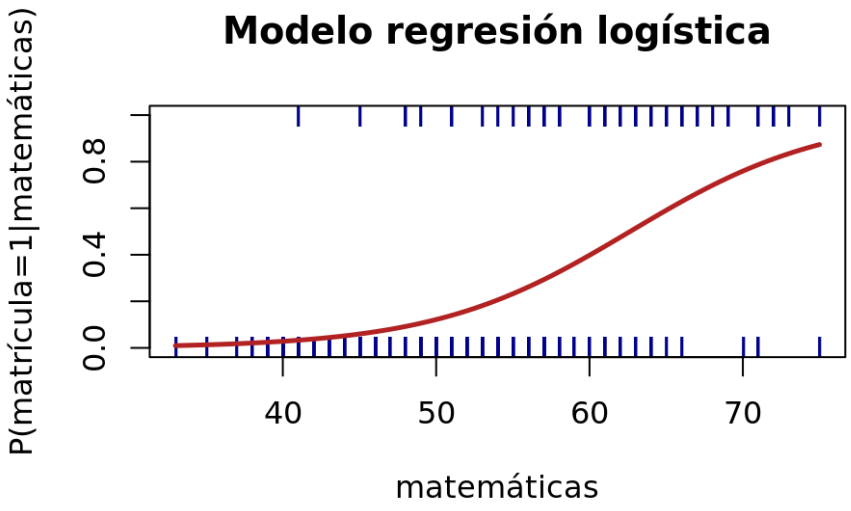


Figura 4: Representación gráfica de una regresión logística [Fuente: ciencia de datos.net]

3. Árboles de decisión

Son modelos de clasificación muy utilizados que construyen una estructura jerárquica en la que el conjunto de datos inicial se va partiendo hasta conseguir conjuntos puros, es decir conjuntos en los que todos los datos tienen la misma etiqueta de clase. Para ello, en cada nodo interno del árbol, se selecciona un atributo en base a cuyos valores se parte el conjunto de datos. El proceso se repite en cada nodo hijo hasta que el conjunto es puro, en cuyo caso el nodo se convierte en una hoja del árbol. Para clasificar una nueva instancia se recorre el árbol de arriba hacia debajo de acuerdo a los valores que la instancia tiene para cada uno de los atributos usados como test en los nodos del árbol hasta alcanzar una hoja. Entonces se asigna a la instancia la clase de dicha hoja. El criterio de partición es el que establece cuál es el atributo que se debe seleccionar en un nodo dado para continuar el crecimiento del árbol. Las distintas implementaciones de esta técnica difieren en el criterio usado (por ejemplo, *information gain*, o *gini impurity*).

Los árboles de decisión son de gran ayuda a la hora de determinar las decisiones que deben adoptarse para la consecución de los objetivos de que se trate, ya que pueden representarse gráficamente, pero también pueden expresarse como un conjunto de reglas de implicación lógica, del tipo *si...entonces; no… sino*, donde las condiciones son tests sobre los atributos. Por este motivo, se considera que los árboles de decisión son un claro ejemplo de modelos comprensibles.

Existen parámetros adicionales a la hora de configurar las características de un árbol de decisión, estos parámetros son los siguientes:

- La profundidad de un árbol, que viene determinada por el número máximo de nodos de una rama.
- Tamaño mínimo de las hojas del árbol (número mínimo de ejemplos en ese nodo para ser considerado una hoja).

Los árboles de decisión tienen una marcada tendencia al sobreajuste (*overfit*). Esto quiere decir que tienden a aprender muy bien los datos de entrenamiento, pero su posterior generalización no es tan buena. Para mejorar mucho más la capacidad de generalización de los árboles de decisión, deberemos combinar varios de tales árboles.

4. Random forest

Es un modelo de análisis de datos basado en el anterior, pero más complejo, puesto que lleva a cabo una combinación aleatoria de diversos árboles. Es una técnica de las que se denominan multiclasificadores, ya que se basan en la idea de combinar las decisiones de varios modelos. Pero para que esta combinación haga predicciones más precisas que cualquiera de sus modelos constituyentes, es necesario que éstos sean diferentes entre sí. Esto se puede conseguir haciendo que cada árbol se entrene con distintas muestras de los datos iniciales. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que arroja un mejor comportamiento en la posterior fase de generalización.

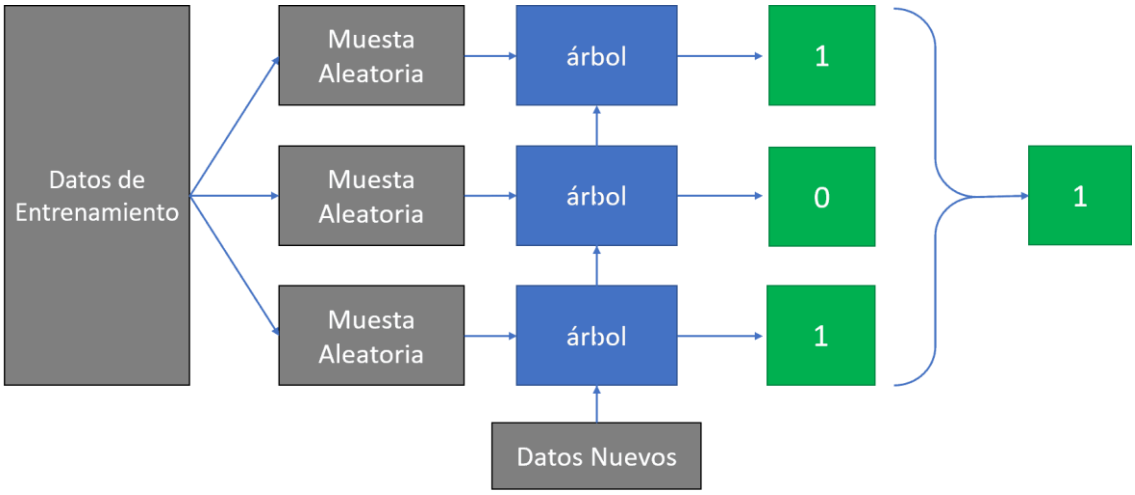


Figura 5: Random forest [Fuente: iartificial.net]

Para clasificar una nueva instancia usando un modelo aprendido mediante esta técnica, primero se aplican los árboles constituyentes a la nueva instancia y una vez recopiladas todas las clases predichas se combinan para tomar la decisión final sobre qué clase se asignará a la instancia. La forma más simple de combinación es el voto mayoritario (es decir, se asigna la clase que más veces ha sido predicha), aunque existen otras formas de tomar la decisión final.

En un *random forest* existen parámetros que cabe tener en cuenta a la hora del modelado como son, el número de árboles de que se compone, así como la profundidad máxima de cada uno de ellos, parámetro que cabe recordar, también se ha definido dentro de los parámetros a tener en cuenta en un árbol de decisión.

4. Redes neuronales

La inteligencia artificial y el d*eep learning* han puesto muy de moda esta técnica tan sofisticada de reconocimiento de patrones que imita las neuronas del cerebro humano, y que es capaz de modelar relaciones extremadamente complejas.

Una red neuronal está constituida por una colección de elementos de procesamiento (nodos o neuronas) altamente interconectados que transforma un conjunto de datos de entrada en un conjunto de datos de salida deseado. Es una técnica de estimación y clasificación perteneciente a la inteligencia artificial, que tiene como principio el proceso de aprendizaje que intenta simular la conducta cognitiva del cerebro humano. Está formado por un conjunto de nodos conocidos como neuronas artificiales que están conectadas y transmiten señales entre sí. Estas señales se transmiten desde la entrada hasta generar una salida. Para ello, la red neuronal está formada por las capas de entrada y salida y varias capas internas. Una de las dificultades en el entrenamiento de las redes neuronales es determinar cuál es el número de capas ocultas y cuántas neuronas deben contener.

El método de redes neuronales suele utilizarse cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y los de salida.

En las redes neuronales existen parámetros que se deben tener en cuenta. En este TFM nos vamos a centrar en poder determinar cuándo se detendrá el proceso de aprendizaje; es necesario establecer una condición de detención, que como se verá, puede ser el número de ciclos y/o pasos de entrenamiento.

2.2.- Técnicas de modelización descriptiva

El análisis descriptivo es, tal vez, una de las técnicas más interesantes de la minería de datos y es, además, una de las más útiles para las necesidades de la mayoría de las empresas que precisan del análisis de datos para llevar a cabo sus cometidos. Son muchos los sectores que utilizan el análisis descriptivo, mediante el cual pueden llegar a clasificar grandes volúmenes de información.

Aunque existen diversas técnicas de clasificación descriptiva, en este TFM nos hemos centrado en la técnica de clasificación descriptiva utilizando k-medias.

1. K-medias

K-medias es un método que tiene como objetivo generar una partición de un conjunto de n observaciones en k grupos. Cada grupo está representado por el promedio de los puntos que lo componen. El representante de cada grupo se denomina centroide. La cantidad de grupos a descubrir, k, es un parámetro que se debe fijar a priori. El método de clustering o agrupamiento comienza con k centroides ubicados de forma aleatoria, y asigna cada observación al centroide más cercano. Después de asignarlos, los centroides se mueven a la ubicación promedio de todos los datos asignados a él, y se vuelven a reasignar los puntos de acuerdo a las nuevas posiciones de los centroides. Este proceso continúa hasta que la posición de los centroides no se modifica en dos iteraciones consecutivas.

El objetivo de K-medias es agrupar las observaciones de forma tal que todas las que se encuentren en el mismo grupo sean lo más semejantes entre sí y que las pertenecientes a grupos distintos sean lo más desemejantes entre sí. Las medidas de distancia, como la euclídea, son utilizadas para medir la semejanza y desemejanza. Una medida para indicar lo bien que los centroides representan a los miembros de su grupo es la suma de los errores al cuadrado.

El algoritmo siempre termina, ya que no necesariamente encuentra la configuración óptima, la que se corresponde con el mínimo de la función objetivo. Hallar un mínimo de la función, a pesar de que no se trate del mínimo absoluto, garantiza un agrupamiento en el que los grupos son poco dispersos y se encuentran separados entre sí. El algoritmo es significativamente sensible a los centroides que se seleccionan inicialmente de manera aleatoria. Este efecto se puede reducir realizando varias iteraciones del método.

2.3.- Evaluación y métricas de los modelos predictivos y descriptivos

Una vez realizado el modelado de los datos y obtenido los distintos modelos predictivos, debemos evaluarlos y analizar aquellos que más nos interesen. Antes que nada, debemos destacar que las medidas de evaluación de regresión y de clasificación son diferentes, como vamos a tratar de mostrar en este apartado.

La finalidad de un modelo es predecir el valor de la variable respuesta en observaciones futuras o en observaciones que el modelo no ha "visto" antes. El error mostrado por defecto tras entrenar un modelo suele ser el error de entrenamiento, el error que comete el modelo al predecir las observaciones que ya ha "visto". Si bien estos errores son útiles para entender cómo está aprendiendo el modelo (estudio de residuos), no es una estimación realista de cómo se comporta el modelo ante nuevas observaciones. Para conseguir una estimación más certera, se tiene que recurrir a un conjunto de test o emplear otras estrategias de validación.

Los métodos de validación, también conocidos como *resampling*, son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento. La idea en la que se basan todos ellos es la siguiente: el modelo se ajusta empleando un subconjunto de observaciones del conjunto de entrenamiento y se evalúa con las observaciones restantes. Este proceso se repite múltiples veces y los resultados se agregan y promedian. Gracias a las repeticiones, se compensan las posibles desviaciones que puedan surgir por el reparto aleatorio de las observaciones. La diferencia entre métodos suele ser la forma en la que se generan los subconjuntos de entrenamiento/validación.

Aunque la técnica de validación más sencilla es la partición del conjunto de datos en dos subconjuntos (uno para el entrenamiento y otro para el test), la técnica de validación que se utiliza más frecuentemente es la validación cruzada o *cross validation*.

Con la validación cruzada o *cross validation* se garantiza la independencia de la partición entre los datos de entrenamiento y los de prueba. Consiste en partir el conjunto de datos en n subconjuntos (denominados pliegues) de aproximadamente el mismo tamaño. En cada iteración se reserva un pliegue para test y se entrena un modelo con los (n-1) pliegues restantes. Este proceso se repite n veces, de forma que al terminar cada pliegue ha sido usado 1 vez como conjunto de test. Finalmente se calcular la media aritmética de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica. Como veremos va a ser la técnica que vamos a utilizar en este TFM.

A continuación, resumimos las medidas de evaluación más utilizadas para la evaluación de los distintos modelos predictivos, incidiendo en el hecho de si son de regresión o de clasificación. Cada una de las métricas que se presentan se puede utilizar para cuantificar el comportamiento del modelo. La elección de la métrica depende del tipo de modelo, el tipo de datos y el campo específico de trabajo o de investigación de que se trate.

A. Evaluación de los modelos de regresión

Hay diferentes métricas que se pueden usar para la evaluación del modelo de regresión, pero la mayoría de ellas se basan en la similitud (o diferencia) de los valores estimados y reales. Una de las métricas más simples para calcular la precisión de nuestro modelo de regresión es el error, calculado como la diferencia promedio entre los valores predichos y los reales para todas las filas.

- Error cuadrático medio (MSE): es la media de los errores al cuadrado. Está más extendida que el error medio absoluto porque esta medida penaliza más cuanto mayor es la discrepancia entre el valor estimado y el real. Esto se debe a que el término al cuadrado aumenta exponencialmente los errores más grandes en comparación con los más pequeños, haciendo que pesen más en la evaluación.
- Raíz cuadrada del error cuadrático medio (RMSE). Esta es una de las métricas más extendidas de entre las de evaluación, porque es interpretable en las mismas unidades que el vector de respuesta, haciendo así fácil la correlación con la información. En este TFM vamos a aplicar esta medida de evaluación para poder razonar sobre las predicciones en términos de años (ya que ésta es la magnitud de los valores de la variable objetivo).

B. Evaluación de los modelos de clasificación:

Las métricas de evaluación del modelo de clasificación explican el rendimiento de un modelo. Básicamente, se comparan los valores reales del conjunto de pruebas con los valores pronosticados por el modelo a fin de calcular la precisión de este. Las métricas de evaluación del modelo de clasificación proporcionan un papel clave en el desarrollo de un modelo, ya que ofrecen pistas de las áreas que pueden mejorarse. Hay diferentes métricas de evaluación de modelos, que pueden definirse a partir de la matriz de confusión. Presentamos el caso de dos clases, ya que las definiciones se pueden fácilmente extender para el caso multiclase.

La matriz de confusión es una representación tabular del comportamiento de un clasificador, al representar los aciertos y fallos que el clasificador comete en el conjunto de instancias de cada clase. Para ello, las filas representan los valores predichos por el clasificador mientras que las columnas representan los valores reales. En el caso binario, la matriz es de 2x2 ya que la variable objetivo (la salida del modelo) puede tener dos valores: que se suelen denominar positivo y negativo, respectivamente. La figura 6 muestra esquemáticamente la estructura de una matriz de confusión.

Figura 6: Matriz de confusión [Fuente: Wikipedia].

Las medidas de evaluación son:

Accuracy o exactitud se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. También se conoce como Verdadero Positivo (o "True positive"). Se representa como la proporción de resultados verdaderos (tanto verdaderos positivos (VP) como verdaderos negativos (VN)) dividido entre el número total de casos examinados (verdaderos positivos, falsos positivos, verdaderos negativos, falsos negativos).

En forma práctica, la exactitud es la cantidad de predicciones positivas que fueron correctas.

$$\text{Exactitud} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}$$

La precisión se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor es la precisión. Se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos). En forma práctica es el porcentaje de casos positivos detectados:

$$\text{Precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

La sensibilidad o *recall*, también se conoce como Tasa de Verdaderos Positivos (*True Positive Rate*) o TP. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo. Se calcula como:

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

C. Evaluación de los modelos de agrupamiento

En el caso de los algoritmos de agrupamiento o *clustering*, es difícil definir cuando el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado.

Como el objetivo del clustering es agrupar objetos similares en el mismo clúster y objetos diferentes ubicarlos en diferentes clústeres, las métricas de validación interna están basadas usualmente en los criterios de **cohesión** dentro del grupo y **separación** entre grupos.

Una de las medidas más utilizadas es el índice de **Davis Bouldin** (DB). Este índice combina los dos criterios de evaluación que se acaban de mencionar, por lo que se basa en la idea de que una buena partición se caracteriza por poseer una alta separación entre los clústeres resultantes, así como un alto grado de compactación y de homogeneidad en el interior de cada uno. Dado de que el objetivo es alcanzar una mínima dispersión de cada clúster y una máxima separación entre los clústeres, un índice DB menor se considera el valor óptimo.

D. Significancia de los resultados

Habitualmente, se suele hacer uso de los test estadísticos para determinar la significancia de los resultados obtenidos. En particular, algunas técnicas de aprendizaje usan este tipo de test para determinar la importancia de las variables predictoras (o su significancia) a la hora de estimar la variable objetivo.

Por ejemplo, la técnica de regresión lineal se sirve del valor del test de probabilidad p-value para determinar la significancia de las variables. Así, se puede observar las variables más correlacionadas en nuestro modelo de regresión lineal. Se define el p-*value c*omo la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. En términos simples, el valor p ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos. Este valor es un valor de probabilidad, oscila entre 0 y 1. Si el valor *p-value* es menor que un cierto umbral, ello significa que el resultado es estadísticamente significativo. Lo más habitual es hacer este test con un nivel de significancia del 95%, lo que significa que dado un valor v, si p-value(v) < 0,05 entonces v es estadísticamente significativo.

2.4. Rapidminer

Para poder realizar tanto el análisis, como el modelado y la evaluación de nuestro proyecto, necesitamos una herramienta. Existen numerosas plataformas para el desarrollo de sofisticados modelos estadísticos y el análisis predictivo de grandes paquetes de datos: SAS, Oracle Advanced Analytics, KNIME, WEKA, IBM, Neural Designer, etc.

La herramienta que se ha elegido para desarrollar este TFM es Rapidminer. Esta herramienta dispone de un potente entorno visual para el encadenamiento de operadores, extensiones para importar y manejar diferentes tipos de datos, compatibilidad con R y

Python, y una abundantísima documentación mediante vídeos tutoriales y aplicaciones prediseñadas para diversos escenarios. Además, dispone de una versión de escritorio (RapidMiner Studio) que se puede descargar gratuitamente e instalar en diversas plataformas.

RapidMiner participa en todos y cada uno de los pasos del proceso de *data mining*, interviniendo también en la visualización de los resultados. La herramienta está formada por tres grandes módulos: RapidMiner Studio, RapidMiner Server y RapidMiner Radoop, cada uno encargado de una técnica diferente de minería de datos. Asimismo, RapidMiner prepara los datos antes del análisis y los optimiza para su rápido procesamiento. Para cada uno de estos tres módulos hay una versión gratuita y diferentes opciones de pago.

El punto fuerte de RapidMiner, si se compara con el resto de software de data mining, reside en los análisis predictivos, es decir, en la previsión de desarrollos futuros basándose en los datos recopilados.

En este TFM además de la herramienta Rapidminer vamos a utilizar una extensión de Weka para Rapidminer. Weka es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Es software libre distribuido bajo la licencia GNU-GPL y además como hemos adelantado es posible integrarlo en otras plataformas.

CAPÍTULO 3. COMPRENSIÓN Y PREPARACIÓN DE LOS DATOS

3.1. El conjunto de datos

Para el objetivo que nos hemos planteado partimos de una base de datos con información de 1400 pacientes de diferentes países y con edades comprendidas entre 20 y 90 años. La base de datos es propiedad de Quibim S.L. y fue obtenida a partir de secuencias de resonancia magnética T1[7] 3D. Estas secuencias se consideran "standard of care", es decir, son secuencias que pueden obtenerse mediante los equipos de resonancia magnética (RM) que se encuentran en cualquier hospital del mundo, tanto con equipos de 1.5T (Teslas) como 3T. Estas secuencias también son estándar para los fabricantes más habituales de dispositivos de RM: Philips Healthcare, Siemens, General Electric y Canon. Las secuencias T1 3D proporcionan una alta resolución espacial, además de un muy buen contraste entre sustancia gris, sustancia blanca y líquido cefalorraquídeo (figura 7).

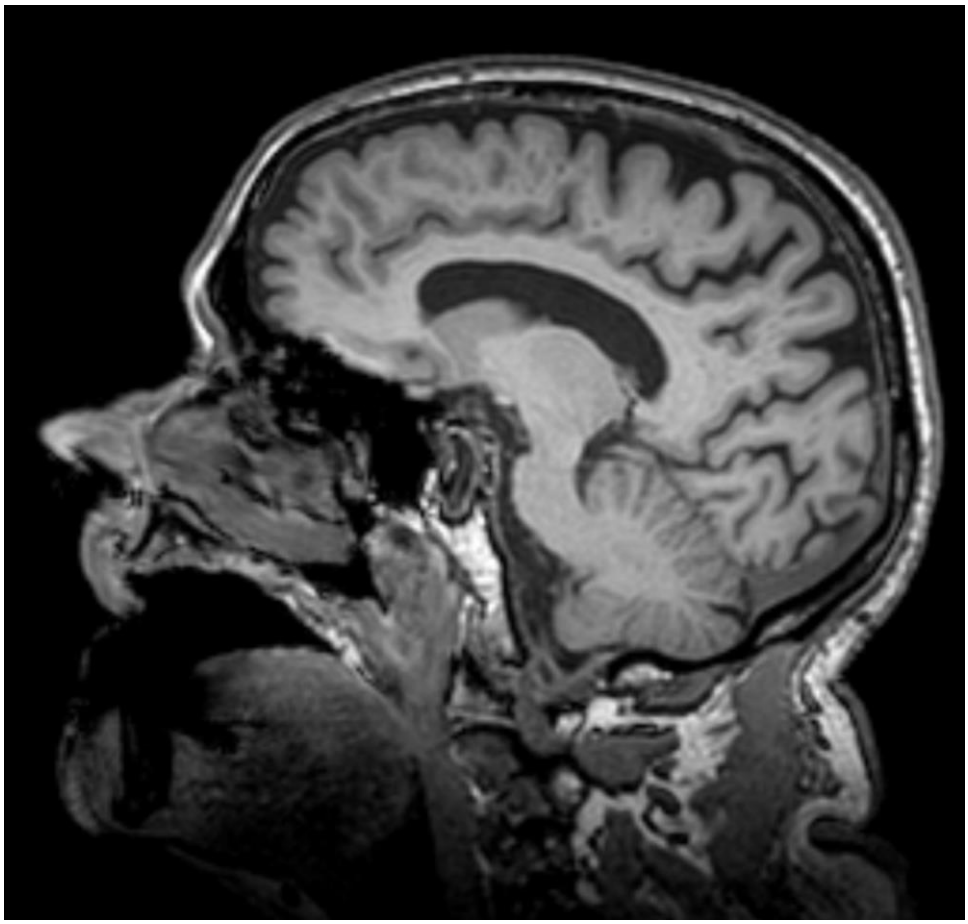


Figura 7: Ejemplo de corte sagital de una secuencia T1 3D cerebral de RM (imagen cortesía de Quibim S.L.)

La metodología seguida por Quibim S.L. para la extracción de los biomarcadores de imagen de volumetría cerebral se muestra en la figura 7. Sin entrar en detalle en las

[7] Por consenso, el **T1** de un tejido es el tiempo que tarda en recuperar el 63% de la magnetización longitudinal.

particularidades de los algoritmos implicados, se describe a continuación cada uno de los pasos:

1. Adquisición de secuencia de RM T1 3D.
2. Corrección de inhomogeneidades, intensidad y reducción de ruido mediante distintos algoritmos de filtrado.
3. Normalización de la imagen del espacio nativo de la adquisición al espacio MNI[8].
4. Utilización de plantillas de tejido en el espacio MNI para la segmentación de sustancia gris, blanca y líquido cefalorraquídeo.
5. Parcelación de áreas de sustancia gris mediante la utilización de un atlas.

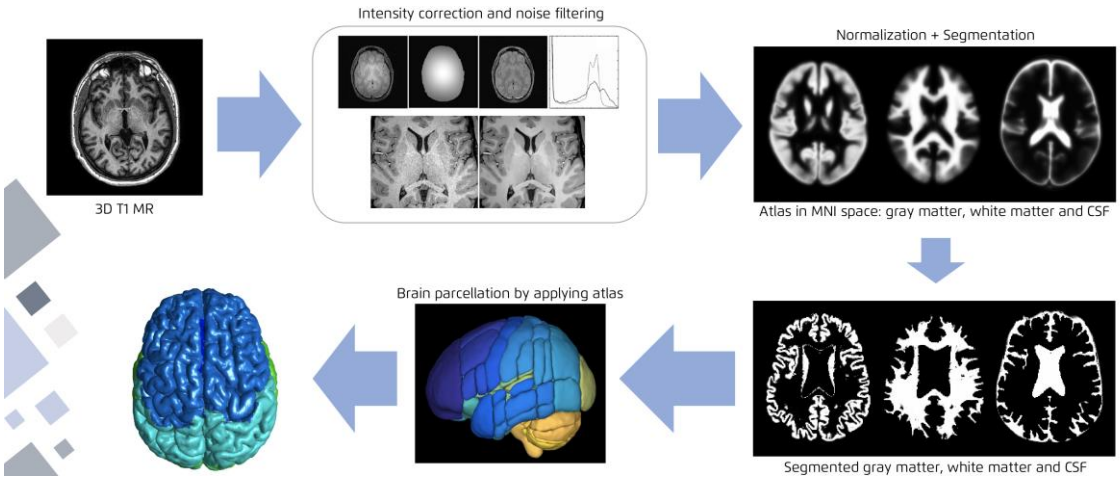


Figura 8: Diagrama de flujo con los algoritmos necesarios para la extracción de biomarcadores de volumetría cerebral con la metodología de Quibim S.L. (imagen cortesía de Quibim S.L).

Una de las principales características de la información de la base de datos es que no es posible identificar a los pacientes mediante nombre, identificador de paciente o fecha de nacimiento (la anonimia es completa). Por otro lado, los datos más relevantes para nuestro estudio son la edad de los pacientes (variable a predecir) y el sexo de los mismos. Ambas variables deben tenerse en cuenta y así ha sido a la hora de realizar este TFM.

Además de esas dos variables, edad y sexo, existen otras nueve que corresponden a los biomarcadores extraídos a partir de las imágenes cerebrales y que ofrecen información de los volúmenes que aparecen reflejados en una radiografía. Son las siguientes, según la denominación común en inglés y su interpretación en castellano:

- Graymattervolume-Value: volumen en ml de la sustancia gris del paciente.
- Whitemattervolume-Value: volumen en ml de la sustancia blanca del paciente.
- Cerebrospinalfluidvolume-Value: volumen en ml del líquido cefalorraquídeo del paciente.
- Righthippocampusvolume-Value: volumen de hipocampo derecho.
- Lefthippocampusvolume-Value: volumen de hipocampo izquierdo.
- Righttemporalvolume-Value: volumen de temporal derecho.
- Lefttemporalvolume-Value: volumen de temporal izquierdo.

[8] MNI: espacio estandarizado desarrollado por el Instituto Neurológico de Montreal, cuyo objetivo era definir un mapa del cerebro lo más representativo posible de la población.

- Rightenthorynalcortexvolume-Value: volumen del córtex entorrinal derecho.
- Leftenthorynalcortexvolume-Value: volumen del córtex entorrinal izquierdo.

Junto a ellas, existe aún otra variable más que debe ser tenida en cuenta durante el trabajo, que es la que nos da información sobre la fracción de parénquima cerebral (BPF en inglés). La fracción de parénquima viene expresada en la relación:

$$\mathrm{BPF} = \frac{(\mathrm{SG} + \mathrm{SB})}{(\mathrm{SG} + \mathrm{SB} + \mathrm{LCF})}$$

donde SG es el volumen de sustancia gris, SB es el volumen de sustancia blanca y LCF es el volumen de líquido cefalorraquídeo.

Es decir, la fracción de parénquima cerebral se nos presenta como una ratio de cuanto representa la suma de sustancia gris y sustancia blanca respecto al total intracraneal. Un paciente con una atrofia cerebral pronunciada perdería especialmente sustancia gris y sustancia blanca. El espacio correspondiente va siendo ocupado por un volumen creciente de líquido cefalorraquídeo, lo que comporta que en pacientes con distintos tipos de demencia la BPF sea más baja que en sujetos sanos.

El volumen total intracraneal es el resultado de sumar los volúmenes de sustancia gris, blanca y líquido cefalorraquídeo. Se considera que, a grandes rasgos, estos tres tejidos componen la totalidad del interior del cráneo.

Hasta aquí la descripción del significado de las variables que se hallan presentes en la base de datos utilizada para nuestro estudio. El paso siguiente consiste en, según lo expuesto en nuestro apartado metodológico (vid. supra 1.3), preparar los datos con los que vamos a trabajar (lo que incluye, entre otras acciones, explorar en profundidad los datos y la limpieza de los mismos).

3.2. Análisis exploratorio

El análisis exploratorio de datos (*Exploratory Data Analysis*, EDA) es un paso previo e imprescindible a la hora de comprender los datos con los que se va a trabajar. Su objetivo es explorar, describir, resumir y visualizar la naturaleza de los datos recogidos en las variables del conjunto de datos con el que se va a trabajar, mediante la aplicación de técnicas simples de resumen de datos y métodos gráficos sin asumir asunciones para su interpretación.

A continuación, se describen las características cualitativas/cuantitativas o atributos de nuestra base de datos.

- PatientSex variable cualitativa ordinal cuyos valores pueden ser (M, F).
- PatientAge variable cuantitativa discreta cuyos valores oscilan entre [20-90]
- Graymattervolume-Value: variable cuantitativa continua [270-1000]
- Whitemattervolume-Value: variable cuantitativa continua [215-1031]
- Cerebrospinalfluidvolume-Value: variable cuantitativa continua [40-1490]

- Righthippocampusvolume-Value: variable cuantitativa continua [1.3-6.7]
- Lefthippocampusvolume-Value: variable cuantitativa continua [1.1-8.7]
- Righttemporalvolume-Value: variable cuantitativa continua [15-82]
- Lefttemporalvolume-Value: variable cuantitativa continua [16-72]
- Rightenthorynalcortexvolume-Value: [0.047-2.235]
- Leftenthorynalcortexvolume-Value: [0.047-2.335]

Aunque esta primera vista de los datos proporciona información sobre el rango de valores que cada variable toma dentro del conjunto de datos, los resúmenes gráficos permiten también visualizar la distribución de los mismos, siendo una herramienta especialmente útil para la detección de valores anómalos.

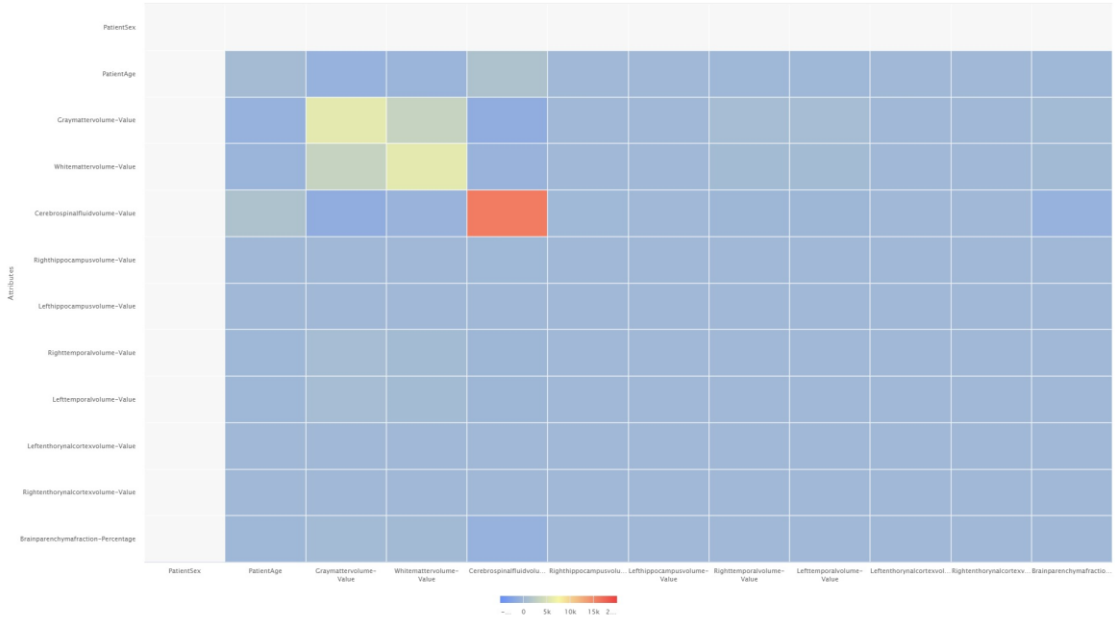


Figura 9: Matriz de covarianza de todas las variables que vamos a utilizar en este TFM. Los identificadores en los ejes representan los nombres de los atributos en el mismo orden de la lista incluida anteriormente.

Las variables que se muestran en la figura 9 y que se detallan a continuación son: sexo, edad, y los volúmenes de sustancia blanca, sustancia gris, líquido cefalorraquídeo, hipocampo izquierdo, hipocampo derecho, temporal izquierdo, temporal derecho, córtex entorrinal izquierdo y derecho.

Si analizamos la figura anterior, cabe destacar la alta covarianza que presenta el líquido cefalorraquídeo, seguido del volumen de sustancia gris y blanca. Se observa que el volumen de sustancia gris y blanca también están altamente correlacionados. Esto puede significar que para nuestro estudio no necesitemos todas las variables, y que de las altamente correlacionadas solamente necesitemos una de ellas.

A continuación, se muestra en la figura 10 los volúmenes de las tres variables más correlacionadas que hemos obtenido de la figura anterior.

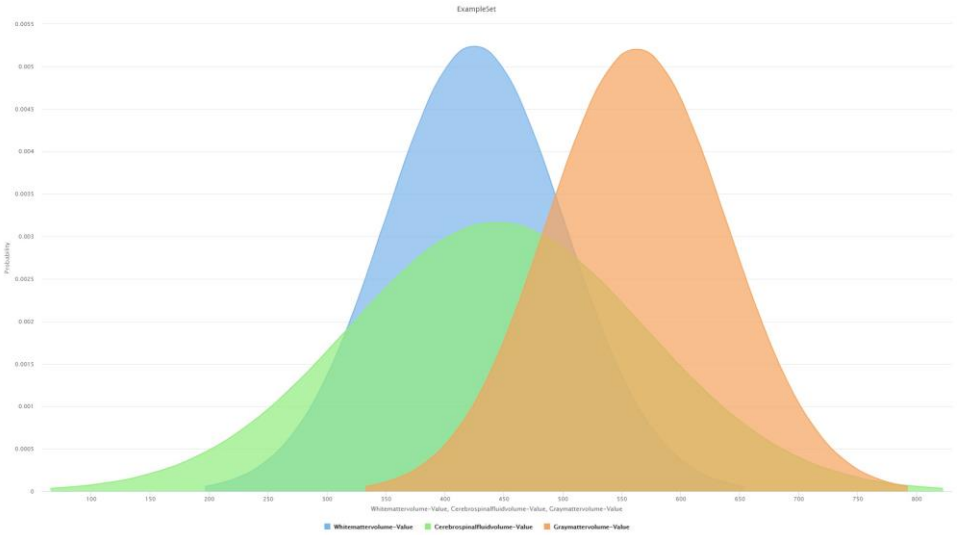


Figura 10: Gráfico curva donde se muestran las frecuencias de los valores de los volúmenes de sustancia gris (representada en color azul), blanca (representada en color verde) y líquido cefalorraquídeo (representado en color verde).

Como se puede observar, los volúmenes de sustancia gris y blanca dibujan la misma forma, una normal, si lo analizamos los datos del volumen de sustancia gris como vimos en la figura 10 están más desplazados hacia la derecha de la figura, en cambio los de sustancia blanca representados en azul están ubicados más hacia la izquierda y de esta forma se definen en las características cuantitativas de las variables al principio.

La figura 11 muestra el histograma con la frecuencia de valores de la variable dependiente **edad**. Para ello, se han agrupado los datos en intervalos (o bins) de 2 años. El gráfico muestra una distribución bastante próxima a la uniforme, cuya media es de 54 años y una desviación estándar de 18,48 años.

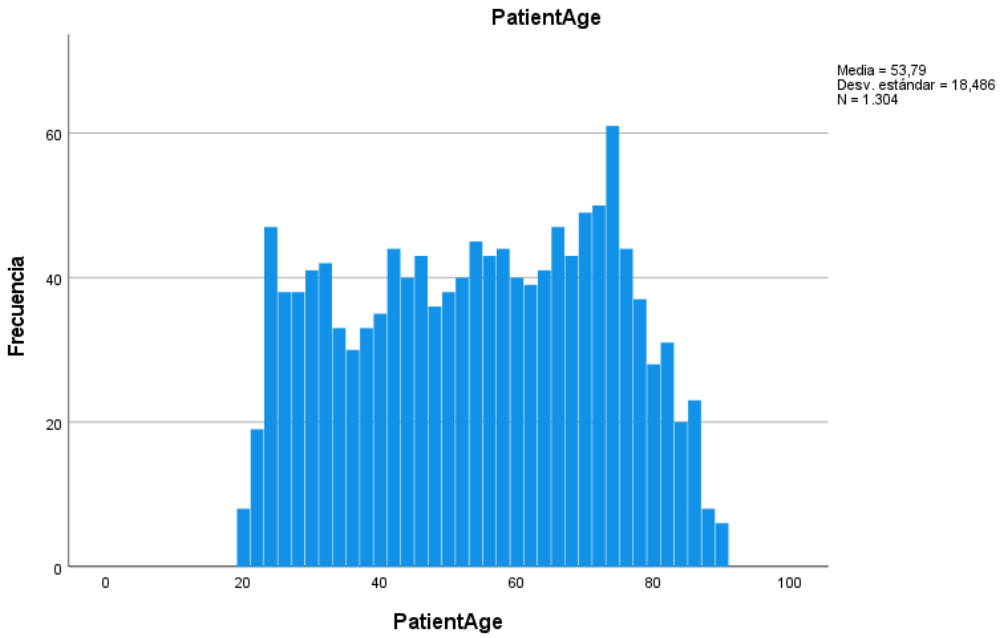


Figura 11: Histograma de frecuencias de la variable edad.

Se observa que la muestra tiene un pico entre los 24 y 26 años de edad y otro en los 76 años que es realmente la frecuencia máxima en el gráfico. Como era de suponer, las frecuencias en los extremos del gráfico son ascendentes y descendentes respectivamente, indicando que hay menos datos muestrales de personas muy jóvenes (edad<25) y muy mayores (edad>80).

Otra de las variables incluida en el conjunto de datos es el **sexo** de los pacientes; los valores de esta variable son M: masculino y F: femenino.

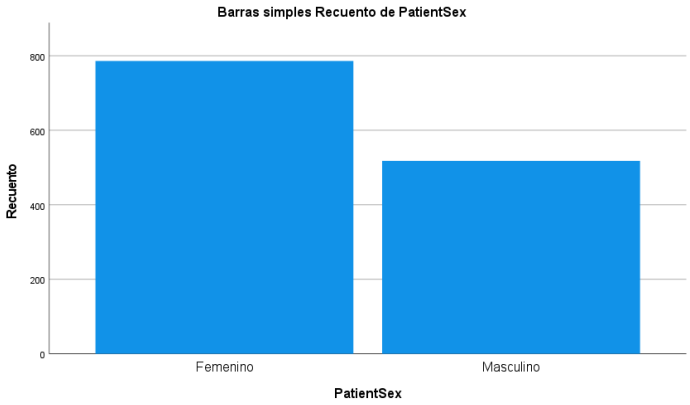


Figura 12: Recuento por sexo (femenino y masculino).

Se observa en la figura 12 que el porcentaje de mujeres es superior al de hombres en la muestra disponible. Puede ello deberse o bien a que las mujeres están más preocupadas por su salud, mayor preocupación que ha quedado reflejada en la muestra mediante un mayor número de casos de femeninos que masculinos; o bien la razón puede hallarse en la mera estructura de la población, en la que el número de mujeres excede el de hombres.

A continuación, se analizan las frecuencias de valores del volumen temporal izquierdo y derecho, respectivamente.

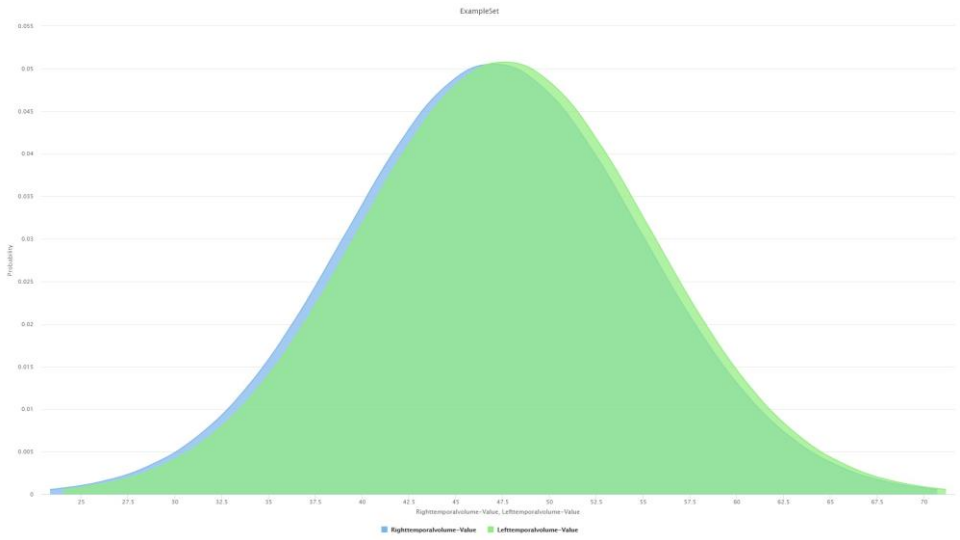


Figura 13: Representación gráfica del volumen temporal izquierdo (representado en color verde) y derecho (representado en color azul).

En la figura 13 se puede observar que ambas variables presentan una mínima variación. Como ya se expuso, el volumen temporal izquierdo varía de 16-72 y el derecho de 15 a 82. Los valores que representan son muy similares. Estos representan una distribución normal y simétrica.

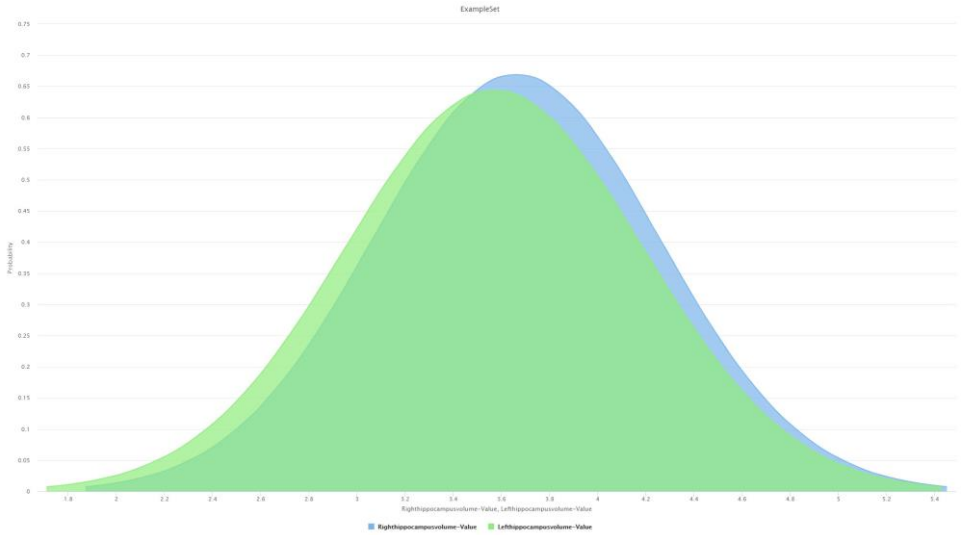


Figura 14: Representación gráfica de frecuencias del volumen del hipocampo izquierdo (representado en color verde) y derecho (representado en color azul).

Si analizamos el gráfico representado en la figura 14, observamos que dibuja una normal simétrica, existe una pequeña diferencia entre ambos que ya quedó reflejada en la

descripción de los mismos. En el caso del hipocampo izquierdo los valores varían de 1.1 a 8.7 y en el caso del hipocampo derecho de 1.3 a 6.7. Tal y como se muestra en la figura los valores máximos del hipocampo izquierdo son mayores que los del hipocampo derecho.

3.4. Limpieza de datos

Una vez realizado el análisis exploratorio, la información obtenida es usada para efectuar la limpieza de los datos. El proceso de limpieza de datos debe ajustarse a los siguientes tres criterios básicos: completitud, validez y unicidad.

- Completitud: lo que significa que todos los valores de las diferentes variables están presentes (de no cumplirse esta condición se deberán usar técnicas de modelado que sean capaces de trabajar con valores faltantes, lo que limita las posibilidades del estudio).
- Validez/Corrección: representa la ausencia de valores anómalos (valores que no se ajustan a la distribución de los datos). Estos valores anómalos pueden ser ruido o bien tratarse de datos erróneos. Nuevamente, la presencia de valores anómalos no erróneos limita el tipo de técnicas de modelado a usar ya que no todas ellas son robustas a este tipo situaciones.
- Unicidad: indica que el conjunto de datos no tiene registros repetidos.

Siguiendo las directrices que se acaban de indicar, se han descartado aquellos registros de información que no contenían toda su información completa. Se ha optado por esta opción ya que el porcentaje de datos incompletos era relativamente pequeño.

En un segundo paso, se han revisado los volúmenes de los biomarcadores facilitados y se han descartado aquellos registros que presentaban valores negativos (al tratarse de valores erróneos).

Asimismo, se han descartado los datos duplicados de pacientes y también, en un primer análisis, los datos anómalos. Consideramos datos anómalos aquellos cuyos valores están alejados del resto de datos, es decir, que se desvían en alguna dirección respecto al comportamiento general del resto del conjunto de datos. Hay que ser muy escrupuloso en esta fase, ya que, si dejamos datos anómalos, estos pueden afectar negativamente a los resultados de los métodos que apliquemos.

Como resultado de este proceso de limpieza, de los 1400 registros iniciales de pacientes de distintos países, se ha pasado a un conjunto de datos con 1303 registros los cuales se usarán para efectuar la etapa de modelado. Si analizamos este proceso de limpieza de forma estadística, de los 1400 registros iniciales que suponen un 100% de la muestra, se han descartado 97 registros, que supone un 6,93%.

Finalmente, dado que algunas de las técnicas de aprendizaje que se van a aplicar son sensibles a la magnitud de los valores de las variables (como, por ejemplo, la regresión lineal o las técnicas basadas en distancias), se ha procedido a normalizar todas las variables cuantitativas. La normalización consiste en la transformación de escala de la distribución de una variable con el objetivo de poder hacer comparaciones respecto a conjuntos de

elementos y a la media mediante la eliminación de los efectos de influencias. Se han normalizado todas las variables exceptuando la edad, que recordemos es nuestra variable dependiente.

CAPÍTULO 4. MODELADO DE DATOS: APROXIMACIONES DIRECTAS

Una vez preparados los datos con los que se va a trabajar, y habiendo realizado en el capítulo 3 una descripción de los mismos, se presentan las diferentes aproximaciones que se han seguido para crear los modelos (usando las técnicas descritas en el capítulo 2) que den respuesta al objetivo principal del TFM: la predicción de la edad (física) del paciente. Para cada una de las aproximaciones que se van a estudiar en este TFM, se mostrarán los resultados obtenidos por cada una de las técnicas empleadas usando las medidas de evaluación presentadas en el capítulo 2.

Se proponen dos aproximaciones generales:

- **Directas**: dado que la variable a predecir es numérica, el problema puede verse como un problema de regresión (estimando directamente la variable de salida partir de las predictoras), pero también puede resolverse como un problema de clasificación, previa discretización de la variable de salida para convertirla en cualitativa.
- **Basadas en segmentación**: previo a resolver el problema como una regresión o una clasificación, en esta aproximación se segmentarán los datos y se resolverá el problema de forma local para cada segmento de registros.

Las aproximaciones directas se presentan en este capítulo, mientras que las basadas en segmentación se presentan en el capítulo siguiente. Para todos los experimentos se ha aplicado una validación cruzada, pero también se ha comparado con la técnica de validación con segmentación de datos para poder compararlas.

4.1. Aproximaciones directas: regresión

La forma más sencilla de resolver el problema que nos ocupa es predecir directamente la variable edad. Para ello se aplica la técnica de regresión lineal, ya que la variable a predecir es una variable numérica. Esto hace pensar que la regresión lineal, tal y como se define en el capítulo 2, es una técnica que parece ser acertada para este problema. Antes de continuar, se debe hacer especial hincapié en que predecir la edad del paciente a partir de los biomarcadores puede ser un problema complejo porque los datos cubren pacientes en un amplio rango de edad (entre los 20 y 90 años).

En el primer experimento realizado consideramos todas las variables predictoras. RapidMiner al mostrar el resultado de la regresión (el modelo) también informa sobre la importancia estadística de cada variable predictora con respecto a la variable a predecir. La figura 15 muestra una captura de pantalla de RapidMiner con los resultados de la regresión lineal. La columna “Code” indica la significancia estadística de cada variable predictora, lo que se representa mediante una secuencia de “*” (desde la secuencia vacía, indicando que la variable no es estadísticamente importante, hasta la secuencia “****” que representa la máxima importancia). Esta importancia de las variables se calcula a partir del *p-value* aplicando un test estadístico con un nivel de significancia del 95%.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
PatientSex = F	-1.248	0.725	-0.033	0.999	-1.723	0.085	*
PatientSex = M	1.248	0.725	0.033	0.999	1.721	0.085	*
Graymattervolume-Value	-10.656	0.546	-0.577	0.421	-19.506	0	****
Whitemattervolume-Value	-0.558	0.395	-0.030	0.806	-1.413	0.158	
Cerebrospinalfluidvolume-Value	9.336	0.473	0.506	0.563	19.753	0	****
Lefthippocampusvolume-Value	-1.474	0.413	-0.080	0.737	-3.570	0.000	****
Righttemporalvolume-Value	4.817	0.569	0.261	0.389	8.469	0.000	****
Lefttemporalvolume-Value	-3.473	0.546	-0.188	0.422	-6.362	0.000	****
Leftenthorynalcortexvolume-Value	-0.842	0.373	-0.046	0.905	-2.257	0.024	**
Rightenthorynalcortexvolume-Value	-0.730	0.364	-0.040	0.950	-2.006	0.045	**
Brainparenchymafraction-Percentage	1.947	0.707	0.105	0.252	2.755	0.006	***
(Intercept)	54.018	∞	?	?	0	1	

Figura 15: Regresión lineal realizada con Rapidminer teniendo en cuenta todas las variables.

Se observa que para este modelo las variables más importantes a la hora de predecir la edad son los siguientes volúmenes: sustancia gris, líquido cefalorraquídeo, temporal izquierdo y derecho y el del hipocampo izquierdo. Otras variables importantes a tener en cuenta son la fracción de parénquima cerebral y el volumen de los córtex izquierdo y derecho. Mientras que las variables menos importantes son la variable sexo y la sustancia blanca.

La evaluación de este modelo de regresión lineal da como resultado un valor de RMSE de 12.391, lo que significa que de media el modelo se equivoca en 12 años al estimar la edad de un paciente. Dicho de otra manera y con un ejemplo concreto: si la edad de un paciente es de 31 años, el modelo nos podría predecir una edad de 19 años o de 43 años. La tasa de error es muy elevada, por lo que no consideramos que sea un buen modelo para nuestro objetivo.

Este primer experimento ha mostrado que las variables sexo y sustancia blanca son poco importantes para el problema. Por ello se decide repetir el experimento, pero eliminando estas variables predictoras.

El modelo aprendido, excluyendo la variable sexo, tiene un valor de RMSE de 12.399, ligeramente peor que el resultado obtenido en el primer experimento, por lo que podemos concluir que, aunque el sexo no es muy importante, sí que aporta información al modelo y por lo tanto es una variable que se debe tener en cuenta.

A continuación, se plantea el mismo reto que con el sexo, pero con la variable volumen de sustancia blanca. El modelo de regresión lineal que se obtiene sin usar esta última variable tiene un valor de RMSE de 12.380 que es ligeramente mejor que el RMSE de los modelos generados hasta ahora.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
PatientSex = F	-1.208	0.724	-0.032	0.999	-1.668	0.096	*
PatientSex = M	1.207	0.724	0.032	0.999	1.666	0.096	*
Graymattervolume-Value	-10.373	0.544	-0.562	0.425	-19.077	0	****
Cerebrospinalfluidvolume-Value	8.043	0.464	0.436	0.582	17.317	0	****
Lefthippocampusvolume-Value	-1.541	0.411	-0.083	0.742	-3.746	0.000	****
Righttemporalvolume-Value	4.812	0.569	0.261	0.389	8.463	0.000	****
Lefttemporalvolume-Value	-3.460	0.546	-0.187	0.421	-6.337	0.000	****
Leftenthorynalcortexvolume-Value	-0.838	0.373	-0.045	0.904	-2.248	0.025	**
Rightenthorynalcortexvolume-Value	-0.739	0.364	-0.040	0.950	-2.033	0.042	**
Brainparenchymafraction-Percent...	0.422	0.703	0.023	0.254	0.601	0.548	
(Intercept)	54.010	∞	?	?	0	1	

Figura 16: Resultado del modelado utilizando regresión lineal excluyendo el volumen de sustancia blanca como variable predictora.

La figura 16 muestra los detalles de este último modelo de regresión lineal (sin tener en cuenta la variable de sustancia blanca). De acuerdo a la columna p-value, la variable fracción parénquima cerebral no tiene significancia estadística (se trata de la variable con el p-value más alto y mayor que 0,05). Esto sugiere que se puede volver a aprender una regresión lineal, pero eliminando también esta variable. Este último modelo tiene un valor de RMSE de 12.345, que es, por lo tanto, el más bajo que hemos obtenido hasta ahora.

Regresión lineal	RMSE
Todas las variables predictoras	12.391
Excluyendo la variable sexo	12.399
Excluyendo sustancia blanca	12.380
Excluyendo sustancia blanca y BPM	**12.345**

Figura 17: Resultados modelado utilizando la técnica de regresión lineal variando el conjunto de variables predictoras. La mejor configuración se resalta en negrita.

Como conclusión de este estudio en el que el problema se resuelve mediante una regresión, tal y como muestra la figura 17, los mejores resultados se obtienen cuando no se consideran las variables volumen de sustancia blanca y la fracción parénquima cerebral. Por lo tanto, los volúmenes que resultan ser informativos para resolver el problema por regresión son:

- Sustancia gris
- Líquido cefalorraquídeo
- Hipocampo derecho
- Hipocampo izquierdo
- Córtex entorrinal derecho
- Córtex entorrinal izquierdo

4.2. Aproximaciones directas: clasificación

Los resultados obtenidos resolviendo el problema como una regresión ponen de manifiesto que intentar predecir la edad concreta de un paciente es un problema complejo, teniendo en cuenta que para dicha predicción únicamente disponemos de los volúmenes biométricos extraídos de una imagen. Una forma alternativa de aproximarnos al problema es predecir no una edad concreta, sino el intervalo de edad al que pertenece y concluir si, por ejemplo, tal paciente tiene una edad entre los 30 y los 40 años. Aunque pueda parecer que esta solución es menos concreta, todavía puede ser de utilidad desde el punto vista médico, puesto que puede servir para identificar el segmento de la población al que pertenece el paciente. Para ello debemos discretizar la variable de salida "edad", convirtiéndola en una variable cualitativa, con tantos valores diferentes como intervalos de edad se definan.

En primer lugar, vamos a discretizar la variable edad, agrupando los valores de 10 años en 10 años. Como el rango de valores de esta variable es de 20 a 90 años, esta discretización da lugar a 7 bins o intervalos. En la figura 18 se muestra el histograma por edades discretizadas en estos 7 bins.

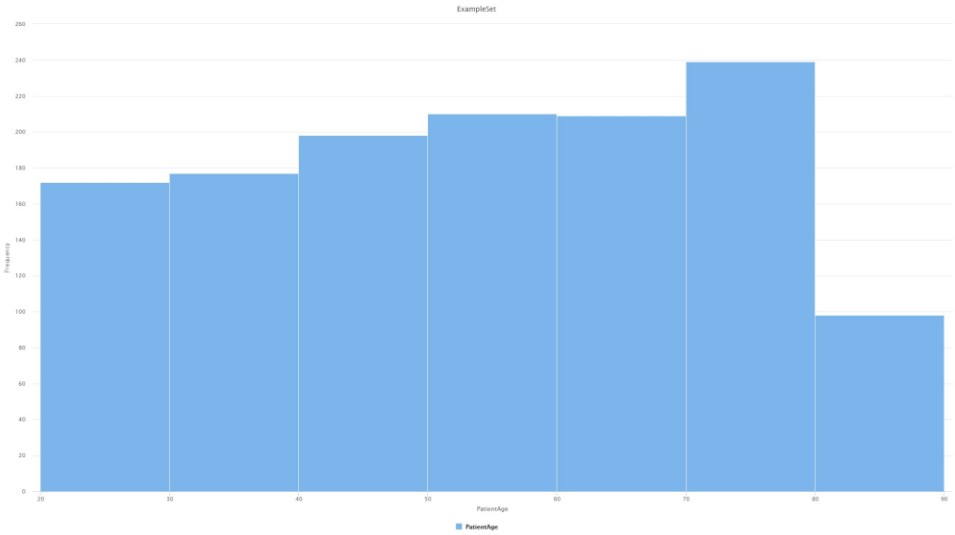


Figura 18: Histograma de la variable edad discretizada en 7 bins de igual talla (10 años).

A partir de la figura 18, se resalta que de la muestra de 80 a 90 años hay menos registros que del resto. De 70 a 80 años es donde más muestra tenemos, podríamos pensar que es en este tramo de edad dónde más enfermedades neuronales se detectan y donde más preocupados estamos por nuestra salud cerebral.

A partir de este punto, se plantean dos discretizaciones más: una reduciendo el tamaño de los bins (por lo tanto, incrementando su número) y la otra haciendo que los bins sean de diferente tamaño reduciendo al mismo tiempo su número. En el primero de los casos se plantea la utilización de 14 bins, es decir, el doble de bins que, en la primera opción, que se obtiene agrupando las edades de 5 años en 5 años. La última discretización consiste

en dividir las edades por categorías. Se plantean 4 categorías que se detallan a continuación, así como el número de ítems que hay en cada una de ellas:

- Categoría 1: 20-35 años 280 ítems jóvenes
- Categoría 2: 35-50 años 285 ítems adultos
- Categoría 3: 50-70 años 569 ítems adultos mayores
- Categoría 4: 70-90 años 169 ítems ancianos

Cada una de estas tres discretizaciones constituye un escenario de clasificación diferente, sobre el que podemos aplicar las técnicas descritas en el capítulo 2: regresión logística, árboles de decisión, *random forest* y redes neuronales.

4.2.1. Regresión logística (W-logistic)

La primera técnica de clasificación que se aplica es la regresión logística (ver capítulo 2). Aunque inicialmente esta técnica se utiliza para predecir el resultado de una variable categórica binaria, existen versiones que permiten una clasificación multiclase. Para ello se va a utilizar la técnica W-logistic de Weka la cual está disponible en RapidMiner. Inicialmente se van a utilizar todas las variables predictoras. La métrica de evaluación que se utiliza es el accuracy o acierto en la predicción.

Los resultados de evaluación de los modelos aprendidos para los tres escenarios de discretización se muestran en la figura 19.

Discretización	**Accuracy (%)**
7 bins	37.61
14 bins	20.11
4 bins	**59.79**

Figura 19: Resultados de *accuracy* de los modelos inferidos usando la técnica *W-logistic* con diferentes discretizaciones de la edad.

Como se puede observar, el porcentaje de *accuracy* o acierto no es muy elevado. Aunque aumenta en el último escenario con menos bins (discretizando la edad en cuatro categorías), en general las tasas de acierto de los modelos son bastante bajas. Analizando la matriz de confusión obtenida en el último escenario (4bins) que se muestra en la figura 20, se observa que, aunque la exactitud del modelo es del 59.79%, que sería una tasa de acierto aceptable, el *recall d*e la clase 2 es muy bajo (del 25.26%) por lo que el modelo falla la mayoría de las instancias de esta clase (o categoría).

accuracy: 59.79% +/- 4.57% (micro average: 59.79%)

	true 1	true 2	true 3	true 4	class precision
pred. 1	171	84	27	1	60.42%
pred. 2	35	72	58	0	43.64%
pred. 3	73	126	447	79	61.66%
pred. 4	1	3	37	89	68.46%
class recall	61.07%	25.26%	78.56%	52.66%	

Figura 20: Matriz de confusión obtenida discretizando la edad en 4 categorías y utilizando la técnica de modelado W-logistic.

Aunque el concepto de *accuracy* y *recall* se explicó en el capítulo 2, cabe recordar que la accuracy nos da la calidad de la predicción, es decir el porcentaje de los que hemos dicho que son de una clase y realmente lo son. En cambio, el *recall* nos da el porcentaje de instancias de una clase que el modelo ha sido capaz de identificar.

4.2.2. Árbol de decisión

Tal y como se hizo con la técnica de regresión logística, se va a realizar el mismo proceso con las tres discretizaciones planteadas, pero ahora utilizando la técnica de aprendizaje de árboles de decisión. La métrica que se va a utilizar es la misma que en el caso anterior, el accuracy o precisión.

Pero antes de comenzar, debemos definir algunos parámetros, antes de proceder a la ejecución. Tal y como se define en el capítulo 2, los parámetros a tener en cuenta son la profundidad máxima del árbol, que se ha estimado en 10 niveles, así como el tamaño mínimo de las hojas, que se ha definido en 2.

Una vez estipulados los parámetros del árbol de decisión, se procede a la ejecución del mismo. Se obtienen los siguientes resultados:

Discretización de la edad	Accuracy (%)
7 bins	26.55
14 bins	14.05
4 categorías	**46.58**

Figura 21: Tasa de acierto (*accuracy*) obtenida por los modelos de árbol de decisión con diferentes discretizaciones de la edad.

La figura 21 muestra los resultados de *accuracy* obtenidos por los árboles de decisión en los diferentes escenarios. Se observa que los valores de accuracy son mejores cuánto menor es el número de bins. Lógicamente, al ser los intervalos mayores es más fácil para el modelo predecir el intervalo correcto. Aun así, una tasa de acierto del 46.58%, obtenida en el mejor de los casos, está todavía lejos del acierto que se pretende conseguir. Se adjunta en el Anexo II la descripción del árbol de decisión discretizando la edad en 4 categorías: joven, adulto, adulto-mayor y anciano.

A continuación, se muestra una representación gráfica del árbol de decisión obtenido, realizando una discretización de la edad en cuatro categorías que ya se definieron: joven, adulto, adulto-mayor y anciano.

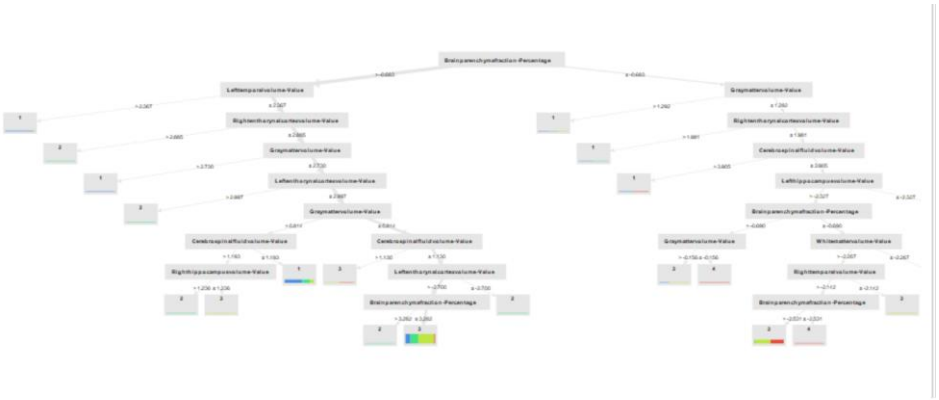


Figura 22. Representación gráfica de un árbol de decisión con la discretización de la variable edad en cuatro categorías (joven, adulto, adulto-mayor, anciano).

A partir de la figura 22, se observa que el árbol está balanceado. La profundidad máxima del mismo es de 10 niveles, cabe recordar que en la configuración inicial del árbol de decisión se estableció el parámetro máximo de profundidad máxima del árbol de 10 niveles, por lo que se puede concluir que es un árbol bastante profundo. Las variables que se muestran en el gráfico representado en la figura 22 son las siguientes (volúmenes): BPM, temporal izquierdo, córtex derecho, sustancia gris, córtex izquierdo, sustancia gris, sustancia blanca, hipocampo izquierdo, líquido cefalorraquídeo, hipocampo derecho y volumen temporal derecho. Se adjunta como Anexo I la descripción de la regresión logística realizada con todas las variables. Por lo que para el árbol de decisión se han tenido en cuenta todas las variables de partida a excepción del sexo.

Si ejecutamos el árbol de decisión, pero ahora sin tener en cuenta la variable sexo, observamos que la accuracy es la misma de un 46.58%. Por todo ello podemos concluir que la variable sexo no aporta información a este modelo.

Si se comparan los datos de la técnica de modelado con árbol de decisión con los datos de la técnica utilizando W-logistic, se observa que el árbol de decisión arroja peores resultados. Una forma de mejorar la generalización de los árboles de decisión es combinar varios árboles, esta técnica de modelado se conoce como random forest. Por ello, se plantea continuar con esta técnica.

4.2.3. Random Forest

La técnica de modelización *Random Forest* es una evolución de la técnica de árboles de decisión, tal y como se detalló en el capítulo 2. Se ha generado un modelo de clasificación usando esta técnica para cada uno de los escenarios de discretización.

Pero antes de continuar se establecen los parámetros iniciales, a partir de los cuales se ejecutará el modelo. Tal y como se define en el capítulo 2, los parámetros iniciales son el número máximo de árboles que podrá tener nuestro *Random Forest*, y que se han estimado en 100. Otro parámetro a establecer un valor inicial es la profundidad de los mismos que se han estimado en 10. Este parámetro, cabe recordar, que también se configuró como parámetro inicial en el punto anterior, y la medida de profundidad que se estimó fue de 10 niveles, los mismos que en el caso que vamos a abordar.

Discretización	Accuracy (%)
7 bins	33.00
14 bins	15.04
4 categorías	**51.11**

Figura 23: Tasa de acierto (*accuracy*) obtenida por los modelos *random forest* con diferentes discretizaciones de la edad.

La figura 23 resume los resultados de accuracy obtenidos. Se adjunta el Anexo III con la descripción del random Forest obtenido. Por motivo de simplicidad se muestra la información de los dos primeros árboles obtenidos, aunque el número de árboles que se han formado al realizar la modelización es de 100, que recordemos era el máximo que nos habíamos definido. Si comparamos estos resultados con los obtenidos anteriormente, aunque los resultados del Random Forest son mejores que los del árbol de decisión, vemos que se sigue sin alcanzar una tasa de acierto buena, por lo que debemos seguir planteando distintas alternativas. Si bien estos valores de accuracy son los mejores de los obtenidos hasta el momento, estos resultados son un poco sorprendentes (en el sentido de que se esperaba que fueran más altos) ya que, generalmente, *Random Forest* es una técnica cuyos modelos suelen tener una alta tasa de acierto en problemas de clasificación de todo tipo de dominios. Todo ello pone nuevamente de manifiesto la dificultad del problema que se está abordando.

4.2.4.- Red Neuronal

El concepto de red neuronal se definió en el capítulo 2. El objetivo que dio origen a las redes neuronales artificiales era el de construir un modelo que fuera capaz de reproducir el método de aprendizaje del cerebro humano. Las células encargadas de este aprendizaje son las neuronas interconectadas entre sí a través de complejas redes.

Uno de los parámetros que se debe estimar, antes de comenzar a realizar la modelización, es el número de ciclos de entrenamiento máximo que se van a utilizar para nuestro modelo, y que se ha estimado en 200.

A continuación, se muestra, la representación gráfica del modelado utilizando la técnica de red neuronal con la edad discretizada en cuatro categorías: joven, adulto, adulto-mayor y anciano.

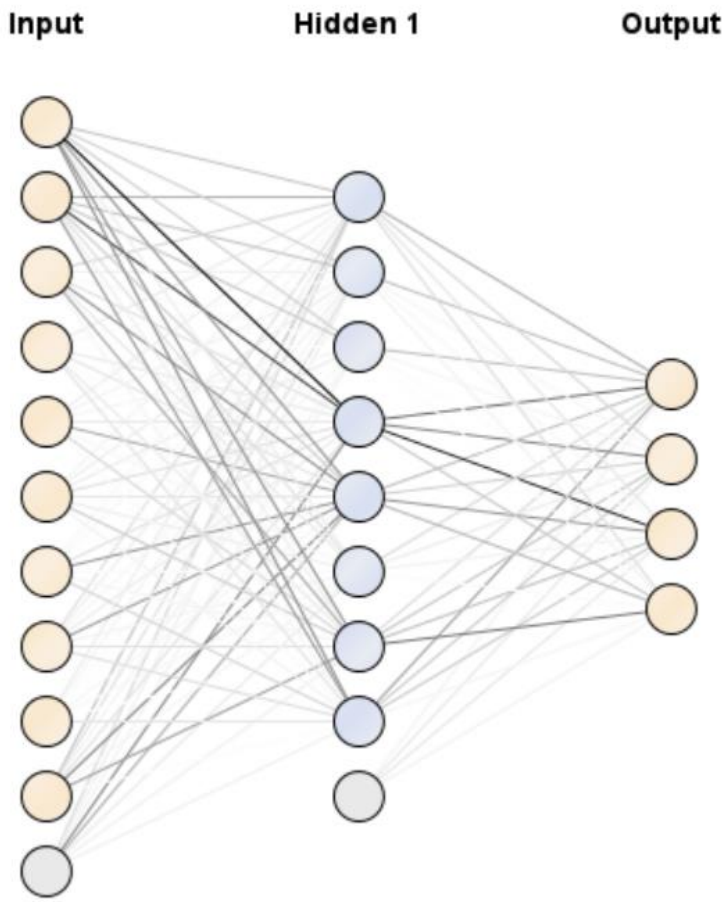


Figura 24: Topología de Red Neuronal con la discretización de la edad en cuatro categorías.

En la figura 24 se puede ver la topología que se ha obtenido utilizando la técnica de modelado de red neuronal con la edad discretizada en cuatro categorías. Tal y como se puede observar presenta una topología de 3 capas: con 11 entradas que son las variables que nos hemos definido, una capa intermedia con 9 nodos ocultos y 4 salidas. La descripción de esta topología que realiza rapidminer se muestra en el Anexo IV.

La figura 25 muestra los resultados de accuracy de las redes neuronales entrenadas para cada uno de los tres escenarios de discretización.

Discretización	Accuracy (%)
7 bins	37.68
14 bins	19.42
4 categorías	**58.48**

Figura 25: Tasa de acierto (*accuracy*) obtenida por los modelos de redes neuronales con diferentes discretizaciones de la edad

Se puede observar que nuevamente el escenario con el menor número de bins obtiene el mejor valor de accuracy, mientras que la tasa de acierto de los modelos los escenarios de 7 y 14 bins es bastante baja.

En el caso de la discretización de la edad en 14 bins, que recordemos se agrupaban las edades de 5 años en 5 años, se podría achacar el mal resultado de accuracy al tamaño de la muestra inicial, que, cabe recordar es de 1303 registros, por lo que al discretizar en 14 bins, los bins son muy pequeños y tienen poca muestra para poder modelar.

4.2.5. Discusión

La figura 26 muestra las tasas de acierto de todos los modelos de clasificación generados para los diferentes escenarios de discretización de la variable edad:

Discretización/ Accuracy	W-logistic	Árbol de decisión	Random Forest	Red Neuronal
7 bins	37.61	26.55	33.00	37.68
14 bins	20.11	14.05	15.04	19.42
4 categorías	**59.79**	46.58	51.11	58.48

Figura 26: Resumen de precisiones de las distintas técnicas de modelización utilizadas discretizando la edad según los criterios especificados.

De acuerdo a los resultados, se observa que la tasa de acierto mejora cuántas menos categorías existen, es decir, conforme aumenta el número de bins las tasas de acierto de los modelos disminuyen notablemente, por lo que debemos descartar este tipo de discretización.

Si analizamos la discretización de 7 bins, que ha sido la primera que nos habíamos planteado y en la que las edades se agrupan de 10 en 10 años, se observa que la tasa de acierto de todos los modelos no es muy buena, oscila entre el 26% y el 37%. Aun así, estos modelos superan el accuracy de un modelo aleatorio cuya tasa de acierto es del 14% (1/7). Estos resultados tan pobres nos indican que con este escenario la resolución del problema no alcanza ese umbral mínimo de acierto para poder considerarse útil, según los expertos. En el caso de la discretización en 4 categorías, las cuales se definieron como joven, adulto, adulto-mayor y anciano, observamos que la tasa de acierto varía dependiendo de la técnica de modelización utilizada. El mejor resultado se obtiene con la técnica de modelado por regresión logística que alcanza un valor del 59.79%, la cual está muy próxima al umbral mínimo por lo que sería un modelo aceptable para resolver automáticamente el problema. Por ello, nos planteamos realizar más experimentos usando este escenario de discretización para profundizar un poco más en este modelo.

Una posible mejora consiste en seleccionar solo aquellas variables predictoras que sean realmente importantes para resolver el problema teniendo en cuenta la técnica de aprendizaje que es la regresión logística. Si recordamos, en la regresión lineal vimos que al no tener en cuenta las dos variables menos significativas para el modelo (la sustancia blanca y la fracción parénquima cerebral) hacían que el comportamiento del modelo en términos

de la tasa de error mejorara ligeramente. La figura 27 muestra la matriz de confusión del modelo aprendido.

accuracy: 59.79% +/- 4.57% (micro average: 59.79%)

	true 1	true 2	true 3	true 4	class precision
pred. 1	171	84	27	1	60.42%
pred. 2	35	72	58	0	43.64%
pred. 3	73	126	447	79	61.66%
pred. 4	1	3	37	89	68.46%
class recall	61.07%	25.26%	78.56%	52.66%	

Figura 27: Matriz de confusión utilizando la técnica de modelado de Regresión logística con la variable edad discretizada en cuatro categorías.

Si analizamos detenidamente esta matriz, llama la atención que la clase 2, que representaría las edades de 35 a 50 y que definimos como adulta, tiene un *recall* muy bajo, de un 25%. Si recordamos los datos del histograma, de los 35 a los 50 años hay 285 ítems. Estos 285 ítems representan un 21% de los datos. En el caso de la clase 4, la precisión es de un 68% pero el *recall* es de un 52.66%. La precisión del modelo podría considerarse acertada.

Una mejora que se plantea para ver si podemos mejorar los resultados de la clase 2, es dividirla en 2 subclases. Una vez dividida vamos a volver a modelar para poder estudiar su comportamiento. Subdividimos el grupo 2 que va de 35 a 50 años en dos grupos, el primero de 35 a 42 años y el segundo de 42 a 50 años.

accuracy: 59.48% +/- 3.38% (micro average: 59.48%)

	true 1	true 2	true 3	true 4	true 5	class precision
pred. 1	194	59	49	41	1	56.40%
pred. 2	1	7	5	2	0	46.67%
pred. 3	5	3	5	9	0	22.73%
pred. 4	79	58	95	480	79	60.68%
pred. 5	1	1	3	37	89	67.94%
class recall	69.29%	5.47%	3.18%	84.36%	52.66%	

Figura 28: Regresión logística discretización 5 categorías.

Si analizamos los resultados de estos últimos experimentos que se muestran en la figura 28, vemos que la tasa de acierto disminuye, y el *recall* de las clases 2 y 3 que hemos creado nuevas es muy bajo, de un 5,47% y un 3.18%. Por lo que descartamos esta modificación del modelado.

Se puede concluir en este punto que las técnicas de modelado que hemos ejecutado hasta el momento no obtienen una tasa de acierto que se pueda considerar apta para el problema que nos planteamos. En el siguiente capítulo se presenta la segunda aproximación en la que se estudiara si resolver el problema localmente a cada grupo de pacientes mejora los resultados obtenidos con la primera aproximación.

CAPÍTULO 5. MODELADO DE DATOS: APROXIMACIONES BASADAS EN SEGMENTACIÓN

En el capítulo anterior se resolvió el problema directamente utilizando técnicas de predicción tanto de regresión como de clasificación. Las técnicas que se utilizaron fueron la regresión lineal, regresión logística, árbol de decisión, *random forest* y red neuronal. Los resultados obtenidos no se consideran buenos para el objetivo que nos planteamos, que debemos recordar, es predecir la edad de un paciente a partir de la información de biomarcadores cerebrales. Por ello, se presenta en este capítulo una segunda aproximación. Básicamente, consiste en un procedimiento en dos fases: primero se crean agrupaciones de pacientes, y después se entrenan modelos para la estimación de la edad en cada grupo.

5.1.- Agrupamiento o clustering

El agrupamiento o *clustering,* tal y como se explicó en el capítulo 2, *e*s una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como clústeres. El centroide de un clúster se define como el punto equidistante de los objetos pertenecientes a dicho clúster.

La métrica que vamos a utilizar es el índice de Davis-Bouldin. Valores pequeños para el índice Davis-Bouldin indican clústeres compactos y cuyos centros están bien separados los unos de los otros. Consecuentemente, el número de clústeres que minimiza el índice Davis-Bouldin se toma como el óptimo.

Uno de los algoritmos que vamos a utilizar es K-medias. Se trata de un algoritmo de aprendizaje no supervisado que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster.

El proceso que vamos a seguir es, en primer lugar, determinar cuál es el mejor valor de k para realizar el agrupamiento. Una vez se tiene el agrupamiento realizado, se pasa a entrenar un modelo que prediga la edad. La técnica elegida es la regresión lineal.

Una vez explicados los conceptos, y el proceso a seguir, se comienza con el agrupamiento tal y como se ha indicado. Para efectuar este agrupamiento y dado que podemos representar cada registro o instancia de nuestro conjunto de datos como un vector de valores numéricos (al ser todos los atributos cuantitativos), para calcular la distancia entre instancias se usará la distancia euclídea.

Como inicialmente no se puede establecer un valor de k óptimo, se plantea comenzar realizando pruebas del algoritmo k-medias con los siguientes valores de k: 3, 4, 5 y 6, para después elegir el valor con el que se obtenga una segmentación de los datos con el mejor índice de DB. La figura 29 muestra los valores de DB obtenidos.

K-medias	DB
3	-1.100
4	-1.230
5	-1.000
6	-1.031

Figura 29: Resultado del índice de DB para el agrupamiento usando k-medias para distintos valores de k.

A la vista de la figura anterior, y tal y como se explicó, el mejor índice de DB es el menor, ya que nos indica clústeres compactos. Por lo tanto, el mejor valor de k para nuestro conjunto de datos es 4. Esto significa que se va a agrupar la información en 4 clústeres.

Tras ejecutar el algoritmo k-medias con k=4, se obtienen los siguientes clústeres:

Clúster 0: 336 items
Clúster 1: 612 items
Clúster 2: 345 items
Clúster 3: 10 items
Número total de ítems: 1303

RapidMiner permite mostrar gráficamente el resultado del proceso de agrupamiento (figura 30).

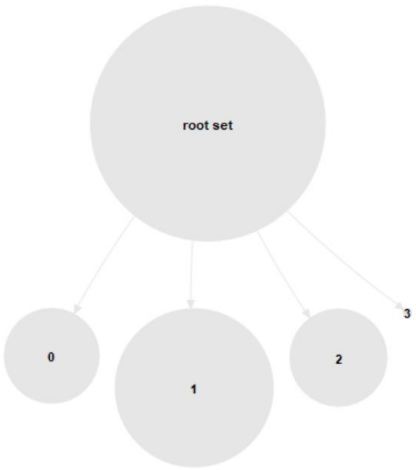


Figura 30: Gráfico de nodos utilizando clustering técnica 4-medias.

Lo primero que se observa es que hay 3 grupos grandes y otro con muy pocas instancias. De entre los grupos más numerosos, destaca el clúster 1 con casi el doble más de instancias que los otros dos grupos. Los clústeres 0 y 2 están más balanceados, aunque el 2 es ligeramente mayor.

Respecto al clúster 3, al tener tan poca muestra no se puede aplicar ninguna técnica predictiva para estimar la edad. Aunque no podemos confirmarlo, podríamos pensar que las instancias del clúster 3, podían presentar alguna anomalía o enfermedad.

A continuación, la figura 31 muestra una tabla con los centroides de cada clúster y la figura 32 la representación gráfica de los mismos.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Graymattervolume-Value	0.520	0.394	0.298	0.100
Whitemattervolume-Value	0.334	0.249	0.180	0.716
Cerebrospinalfluidvolume-Value	0.376	0.317	0.478	0.277
Righthippocampusvolume-Value	0.521	0.430	0.318	0.368
Lefthippocampusvolume-Value	0.390	0.322	0.244	0.323
Righttemporalvolume-Value	0.611	0.467	0.358	0.097
Lefttemporalvolume-Value	0.726	0.553	0.426	0.122
Leftenthorynalcortexvolume-Value	0.468	0.372	0.322	0.192
Rightenthorynalcortexvolume-Value	0.445	0.370	0.329	0.186
Brainparenchymafraction-Percentage	0.557	0.558	0.349	0.656

Figura 31: Tabla con los centroides de los grupos generados con el algoritmo k.medias para k=4.

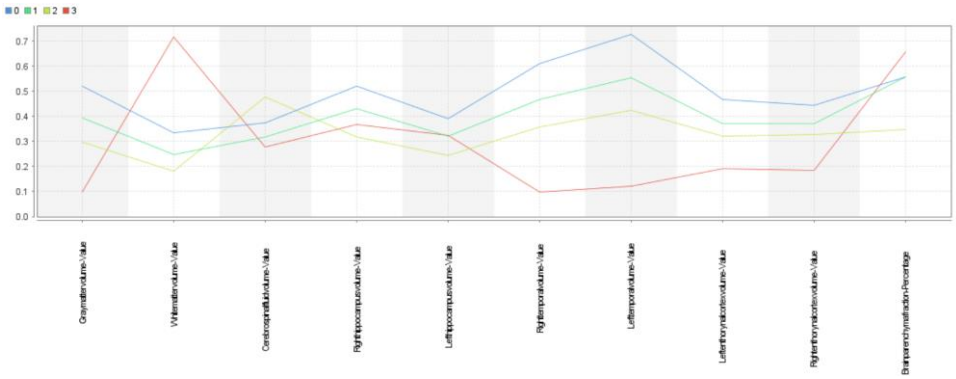


Figura 32: Representación gráfica de los centroides de los grupos generados con el algoritmo k.medias para k=4

El clúster 3 se caracteriza por volúmenes altos de sustancia blanca y fracción parénquima cerebral, y valores bajos del volumen temporal izquierdo y derecho, así como el volumen del córtex entorrinal izquierdo y derecho.

En los clústeres 0, 1 y 2, observamos que los centroides siguen el mismo patrón de forma paralela con una diferencia pequeña en los valores de las variables (0.1). Esto significa que los centroides de estos tres grupos están bastante cerca. Además, cabe añadir que en el clúster 2 el volumen de líquido cefalorraquídeo es el dato con mayor valor que para el resto de los clústeres.

Una vez se han analizado los 4 clústeres de forma genérica, vamos a pasar a realizar la estimación de la edad en cada uno de ellos. Tal y como se explicó al principio de este capítulo, la técnica que vamos a utilizar dado el carácter numérico de los datos es la regresión lineal.

5.2.- Regresión lineal en los clústeres.

Como ya se ha comentado, se descarta para este estudio el clúster 3 por no tener suficientes muestras. Para estos dos clústeres la predicción podría ser directamente obtenida como la media de las edades de las instancias en el clúster.

Por lo tanto, en los experimentos se ha entrenado una regresión lineal usando como datos de entrenamiento las instancias en cada uno de los clústeres 0, 1 y 2. Como en el

capítulo anterior, al tratarse de modelos de regresión, la métrica que se va a tener en cuenta es el RMSE.

La figura 33 muestra los resultados de la evaluación de los modelos de regresión lineal generados.

Regresión lineal	RMSE
Clúster 0	12.360
Clúster 1	12.689
Clúster 2	11.761

Figura 33: RMSE de los modelos de regresión lineal entrenados con las instancias de los clústeres 0, 1 y 2.

Se puede observar que todos los clústeres tienen un RMSE elevado, el clúster 0 y el 1 muestran un valor ligeramente peor que los resultados obtenidos por la regresión lineal entrenada con todos los datos y El clúster 2 en cambio presenta un ligero mejor resultado (tal y como se muestra en la sección 4.1, figura 17).

Para obtener más información de este modelado, vamos a analizar los p-values de todas las variables en el clúster 0.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Graymattervolume-Value	-0.221	0.030	-0.812	0.656	-7.426	0.000	****
Whitemattervolume-Value	0.068	0.019	0.218	0.992	3.621	0.000	****
Cerebrospinalfluidvolume-Value	0.131	0.034	0.923	0.903	3.867	0.000	****
Righthippocampusvolume-Value	0.982	1.577	0.029	0.999	0.623	0.534	
Righttemporalvolume-Value	0.607	0.283	0.217	0.636	2.147	0.032	**
Lefttemporalvolume-Value	-0.501	0.276	-0.169	0.668	-1.816	0.070	*
Leftenthorynalcortexvolume-Value	-5.695	3.235	-0.085	0.998	-1.761	0.079	*
Rightenthorynalcortexvolume-Value	-3.243	3.532	-0.045	0.990	-0.918	0.359	
Brainparenchymafraction-Percentage	1.845	0.610	0.692	0.754	3.025	0.003	***
(Intercept)	-40.703	44.286	?	?	-0.919	0.359	

Figura 34: Regresión lineal entrenada con los datos del clúster 0.

Tal y como muestra la figura 34, el sexo y el volumen del hipocampo izquierdo son variables no significativas para estimar la edad. De hecho, el mismo programa, las ha omitido a la hora de realizar la regresión lineal. Si repetimos los experimentos excluyendo estas variables (tal y como hemos hecho en el capítulo anterior), se obtiene un modelo cuyo RMSE es ligeramente mejor pasando a ser de 12.332.

Volviendo de nuevo a la figura 34, se observa que la variable córtex entorrinal derecho, está en la misma situación que las variables sexo y volumen del hipocampo izquierdo, por lo que también se decide descartarlas y volver a entrenar un nuevo modelo de regresión. En este caso, la mejoría en el valor de RMSE en relación al obtenido por el modelo anterior (sin considerar la variable sexo) es ínfima de apenas dos centésimas (valor de RMSE =12.219, frente a 12.332). De todas formas, éste es el mejor resultado en el clúster 0.

El análisis que se ha realizado en el clúster 0, también se va a realizar con el clúster 1. Se procede de la misma forma.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Whitemattervolume-Value	0.093	0.033	0.613	0.892	2.798	0.006	***
Righttemporalvolume-Value	-0.320	0.350	-0.133	0.976	-0.916	0.361	
Rightenthorynalcortexvolume-Value	-8.502	4.240	-0.143	0.998	-2.005	0.046	**
Lefttemporalvolume-Value	0.566	0.288	0.238	0.948	1.966	0.051	*
Lefthippocampusvolume-Value	-3.377	1.662	-0.152	0.930	-2.031	0.044	**
Graymattervolume-Value	0.067	0.049	0.234	0.930	1.372	0.172	
Cerebrospinalfluidvolume-Value	-0.112	0.056	-0.951	0.291	-1.999	0.047	**
Brainparenchymafraction-Percent...	-3.770	1.358	-1.531	0.296	-2.776	0.006	***
(Intercept)	301.418	84.583	?	?	3.564	0.000	****

Figura 35: Regresión lineal entrenada con los datos del clúster 1.

La figura 35 muestra que no se han utilizado la información de todas las variables, en concreto no se han utilizado las variables: sexo, volumen hipocampo derecho, volumen córtex entorrinal izquierdo. Por todo ello, se vuelve a ejecutar el modelo sin tener en cuenta estas tres variables. El RMSE que se obtiene es de 12.605 que es ligeramente mejor del obtenido con todas las variables, que cabe recordar 12.689.

A continuación, se realiza el mismo estudio con el clúster 2.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Graymattervolume-Value	-0.142	0.038	-0.547	0.461	-3.710	0.000	****
Whitemattervolume-Value	0.061	0.030	0.210	0.987	2.071	0.039	**
Cerebrospinalfluidvolume-Value	0.052	0.066	0.316	0.761	0.779	0.437	
Lefthippocampusvolume-Value	-1.961	1.231	-0.073	0.991	-1.594	0.112	
Righttemporalvolume-Value	0.291	0.303	0.111	0.536	0.961	0.338	
Lefttemporalvolume-Value	-0.509	0.282	-0.194	0.517	-1.805	0.072	*
Leftenthorynalcortexvolume-Value	-4.786	3.136	-0.080	0.999	-1.526	0.128	
Rightenthorynalcortexvolume-Value	-1.234	3.796	-0.018	0.991	-0.325	0.745	
Brainparenchymafraction-Percentage	-0.076	1.450	-0.021	0.528	-0.052	0.958	
(Intercept)	113.860	105.081	?	?	1.084	0.279	

Figura 36: Regresión lineal entrenada con los datos del clúster 2.

La figura anterior muestra que no se han utilizado la información de todas las variables, en concreto no se han utilizado las variables sexo y volumen hipocampo derecho. Por todo ello, se vuelve a ejecutar el modelo sin tener en cuenta estas tres variables. El RMSE que se obtiene es de 11.739 que es ligeramente mejor del obtenido con todas las variables, que cabe recordar 11.761.

Tras analizar los resultados obtenidos en los clústeres 0,1 y 2, podemos concluir que con la selección de variables no se consiguen modelos con la calidad suficiente. Esto se debe a que los clústeres 0, 1 y 2 parecen ser bastante heterogéneos. Con objeto de intentar que los grupos sean más homogéneos, en la siguiente sección planteamos volver a aplicar el clustering en estos tres grupos.

5.3. Agrupamiento en dos niveles

La idea es volver a aplicar el algoritmo k-medias, pero únicamente sobre los clústeres 0, 1 y 2, que son para los que las regresiones lineales tienen el RMSE mayor. Una vez generados los nuevos clústeres procederemos como en la sección 5.2, se entrenará una regresión lineal en cada subgrupo y se evaluará usando RMSE. Para poder distinguir los nuevos grupos de los generados anteriormente, los clústeres 0, 1 y 2 se denominarán clústeres de primer nivel, mientras que los que surjan de la segunda aplicación del algoritmo de agrupamiento los denominaremos clústeres de segundo nivel.

Pero para ello, lo primero es determinar el mejor valor de k en cada caso, utilizando el índice de DB como métrica. La figura 35 muestra el valor del índice DB del resultado de aplicar k-medias para valores de k desde 3 hasta 6.

k	Clúster 0	Clúster 1	Clúster2
3	**-1.758**	**-1.232**	-1.543
4	-1.586	-1.145	**-1.676**
5	-1.591	-1.085	-1.675
6	-1.596	-1.161	-1.476

Figura 37: Resultados del índice de DB para los distintos agrupamientos aplicando k-medias a los clústeres 0,1 y 2 para valores de k desde 3 hasta 6.

Como se puede observar, para los clústeres 0 y 1 el mejor k es k=3. En cambio, para el clúster 2 el mejor k es un k=4 de segundo nivel. A continuación, se analiza este segundo nivel de clustering para cada uno de los clústeres de primer nivel.

5.3.1 Regresión lineal sobre el clúster 0 de primer nivel

Una vez se ha analizado el clúster 0, se procede de la misma forma con el clúster 1. Cabe recordar que el mejor k-medias que se había obtenido para este nivel es para k=3.

Cluster 0: 153 ítems

Cluster 1: 183 ítems

Cluster 2: 1 ítem

Número total de ítems: 337

En una primera observación se puede descartar el clúster 2, ya que solamente presenta 1 ítem. Se tendría que estudiar detenidamente los clústeres que se van descartando ya que, seguramente, son individuos que presentan anomalías en su imagen cerebral en algún parámetro que sería conveniente estudiar. El resto de los clústeres presentan una muestra muy equilibrada en cuanto al número de instancias que los componen.

A continuación, pasamos a realizar una regresión lineal en el segundo nivel de clustering aplicada a los clústeres 0 y 1 de segundo nivel. Pero, antes de continuar, debemos anticipar que la cantidad de ítems de cada uno de los nuevos clústeres es muy pequeña, en concreto de 153 y 183 ítems, lo que supone un 12% y un 14% respectivamente. Los modelos que se obtengan no van a ser representativos.

Segundo nivel – clúster 0	RMSE
Clúster 0	10.218
Clúster 1	13.145

Figura 38: Resultados de la regresión lineal realizada en un segundo nivel de clustering a partir del clúster 1 de primer nivel.

Si se analizan los resultados mostrados en la figura 38, se observa que el RMSE solamente se mejora en el clúster 0, en el resto de los clústeres los resultados son ligeramente peores.

En este segundo nivel de clustering, se observa que los resultados no son los que cabría esperar para la predicción de la edad, presentan una tasa de error elevada.

5.3.2 Regresión lineal sobre el clúster 1 de primer nivel

Tras aplicar k-medias con k=3 sobre el clúster 1, se obtienen los siguientes clústeres de segundo nivel:

Cluster 0: 278 ítems

Cluster 1: 194 ítems

Cluster 2: 139 ítems

Número total de ítems: 611

Aunque este clúster era el que más ítems tenía, se debe recalcar que a cantidad de ítems de cada uno de los nuevos clústeres es muy pequeña. Ello nos lleva a pensar que, aunque apliquemos regresión lineal no se va a mejorar los resultados que se han obtenido hasta ahora al tener tan pocos datos de entrenamiento.

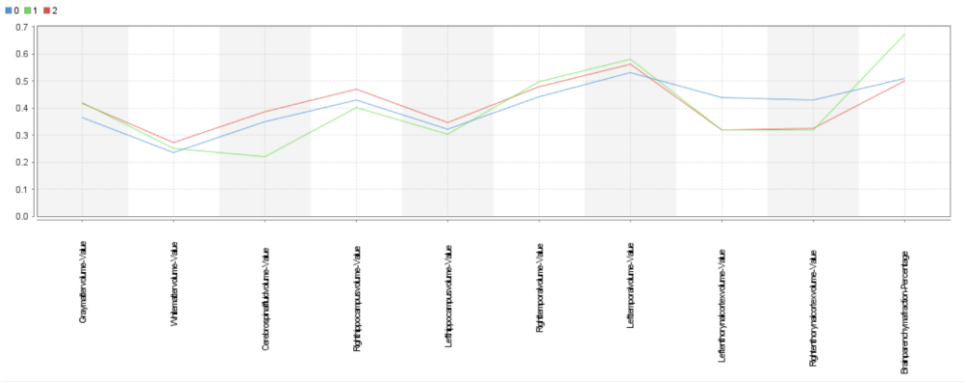


Figura 39: Gráfico del clúster 0 de primer nivel y de 3-medias de 2 nivel.

La figura 39 muestra los centroides de los clústeres de segundo nivel. Se observa que los centroides presentan valores similares. Se destaca que el clúster1 presenta un valor menor en el volumen de líquido cefalorraquídeo y mayor en el BPM.

A partir de aquí, y dados los datos de segundo nivel, aplicamos regresión lineal para los clústeres de segundo nivel 0,1 y 2.

Segundo nivel – clúster 1	RMSE
Clúster 0	12.570
Clúster 1	11.100
Clúster 2	13.218

Figura 40: Valores de RMSE de los modelos de regresión lineal para los clústeres de segundo nivel (0, 1 y 2) obtenidos a partir del clúster 0 de primer nivel.

Si se analizan los datos de la figura 40, se observa que los RMSE obtenidos en los clústeres 0 y 2 son peores. En el caso del clúster 1 mejora ligeramente. Tal y como se pone de manifiesto al principio del segundo nivel de clúster, al realizar el agrupamiento disminuye notablemente la cantidad de muestra, por lo que no es posible generar un modelo adecuado.

5.3.3 Regresión lineal sobre el clúster 2 de primer nivel

Una vez se han analizado los clústeres 0 y 1, se procede de la misma forma con el clúster 2. Cabe recordar que el mejor k-medias que se había obtenido para este nivel es para k=4.

Cluster 0: 131 ítems

Cluster 1: 53 ítems

Cluster 2: 94 ítems

Cluster 3: 67 ítems

Número total de ítems: 345

En una primera observación vemos que los clústeres de segundo nivel presentan muy poca muestra. Por lo que los resultados que se obtengan de este análisis utilizando regresión lineal de segundo nivel no van a ser significativos. Por este motivo no se va a realizar la regresión lineal de segundo nivel en el clúster 2.

Como conclusión al clustering de segundo nivel, cabe reseñar que al realizar el segundo nivel de clustering se reduce mucho la muestra, así, los resultados que se han obtenido no van a ser significativos para nuestro modelo y, además, cabe reseñar que tampoco aumentan la tase de acierto. Por lo que concluimos que no es un modelo válido.

Capítulo 6. Conclusiones

En las páginas precedentes hemos expuesto en repetidas ocasiones que el proyecto de nuestro TFM se dirigía a un objetivo difícil de alcanzar por las dificultades que entraña y por la complejidad del objetivo en sí. No obstante, se considera que la investigación llevada a cabo seguía siendo igualmente válida por razones de muy diversa índole. En primer lugar, porque podía servir de indicador de futuras líneas de investigación. Además, resulta obvio que resaltar las dificultades encontradas es el requisito necesario para poder llegar a superarlas en posteriores proyectos de investigación. Y una tercera razón que nos lleva a defender la validez de este objeto de investigación es la ineludible tendencia de la práctica médica (y en general de todas las ciencias con una aplicación práctica inmediata) a trabajar sobre grandes volúmenes de datos que no simplemente describen la realidad, sino que la anticipan.

Así, en el presente TFM, a partir de una base de datos con biomarcadores de imagen de volumetría cerebral de RM de distintos pacientes/sujetos facilitada por la empresa Quibim,SL. se ha realizado un proceso de minería de datos para poder dilucidar si los biomarcadores de imagen almacenados en dicha base de datos permiten obtener modelos que predigan la edad del sujeto/paciente. Se podría evaluar así cualquier discrepancia entre la edad predicha y la edad real y, en consecuencia, detectar envejecimiento prematuro basado en una atrofia inesperada. Hipotéticamente sería posible, además, poder realizar un diagnóstico temprano de patologías relacionadas con la atrofia cerebral. Insistimos en la ambición de tal objetivo a la vista de los datos de los que actualmente disponemos, que claramente deben ser enriquecidos con datos adicionales.

La base de datos disponible incluye pacientes y sujetos, pero no se dispone de información individualizada por patología. Esto significa que se es completamente ciego a la hora de generar los modelos y evaluar su bondad. El “ground truth" utilizado es la edad del paciente/sujeto, pero si estamos ante un paciente de 40 años con una atrofia avanzada, posiblemente su edad predicha se aproxime más a la de una persona de 80 años. Esta puede ser una de las causas por la que los modelos de regresión tienen un error de entre 10 y 12 años. Sería necesario que los expertos médicos revisaran los casos en los que hay mayor discrepancia para comprobar si efectivamente hay alguna causa médica que explique el que la edad inferida por el modelo a partir de la imagen cerebral sea mayor o menor que la edad física.

El objetivo previsto, que cabe recordar era predecir la edad de un paciente a partir de once biomarcadores de imagen cerebral obtenidos a partir DM, se ha ido abordando utilizando la minería de datos y en concreto la metodología CRISP-DM.

Las técnicas que se han utilizado han sido muy variadas, intentando en todo momento buscar una solución al problema que se había planteado. Se han explorado diversas aproximaciones en las que se han aplicado diversas técnicas, y los modelos generados se han evaluado mediante las métricas más usadas en la literatura. A pesar de todos los esfuerzos realizados, los modelos que se han obtenido no son todavía aplicables en práctica clínica ya que el error que se ha obtenido en todos ellos es un error elevado y por ello poco asumible desde un punto de vista médico. No obstante, cabe mencionar que el estudio y sus resultados sí que son prometedores y bastante satisfactorios desde el punto de vista del análisis de datos. Baste indicar que, las alternativas más sencillas para estimar la edad de los pacientes serían o bien usar un estadístico como la media aritmética (si se

afronta el problema como una regresión), que claramente cometería un error mayor que los modelos de regresión que se han presentado, o un clasificador aleatorio (si se afronta el problema como una clasificación) cuya tasa de acierto sería del 25% usando una discretización en cuatro bins (la configuración con la que se han obtenido los mejores resultados con una tasa de acierto próxima al 60%). Por lo tanto, este desarrollo inicial abre un gran abanico de posibilidades para trabajos posteriores de Quibim SL.

Es por ello que nos planteamos un futuro trabajo que profundice en la línea de investigación aquí apuntada. En este TFM se ha trabajado con un primer nivel de datos obtenidos a partir de biomarcadores de imagen cerebrales. Actualmente, y gracias a las técnicas que utiliza Quibim, SL., es posible obtener mucha más información a nivel de volumetría cerebral a partir de RM. Los últimos desarrollos de Quibim SL permiten obtener volúmenes para más de 100 regiones distintas en sustancia gris. Utilizando una base de datos de un número elevado de pacientes y sujetos, con rangos de normalidad definidos para cada edad y donde, idealmente, se conozca si el sujeto sufre algún tipo de patología o si no se habían encontrado alteraciones de significación en su RM, dotaría de mucha mayor potencia estadística a una segunda iteración de este trabajo. Además, al disponer de volúmenes para muchas más áreas de la sustancia gris, se espera una menor correlación entre las variables obtenidas, puesto que la atrofia del hipocampo no tiene por qué estar íntimamente correlacionada con la atrofia del córtex motor en distintas patologías. Pero esto último excede del trabajo que se había planteado para este TFM. Por todo ello, dejamos la puerta abierta a un futuro trabajo de investigación para ahondar más en el objetivo que queríamos alcanzar, un reto que podríamos considerar de una envergadura superior al trabajo que hemos realizado. Sin embargo, esta primera iteración será utilizada por Quibim SL como punto de partida para seguir desarrollando la línea de envejecimiento prematuro, permitiéndole desarrollar modelos más complejos una vez se recojan suficientes datos.

Bibliografía

Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., ... & Islam, M. T. "Can AI help in screening viral and COVID-19 pneumonia?". *IEEE Access*, *8*, 132665-132676. 2020.

Han, J., Kamber, M., & Pei, J. *Data mining: concepts and techniques*. Morgan Kaufmann. 2012.

Hernández Orallo, J., Ferri Ramírez, C., & Ramírez Quintana, M. J. *Introducción a la Minería de Datos*. Pearson Prentice Hall. 2004.

Jiménez-Pastor et al. "Precise whole liver automatic segmentation and quantification of PDFF and R2* on MR images". *European Radiology, 1-12.*

Milovic, B., & Milovic, M. "Prediction and decision making in health care using data mining". *Kuwait Chapter of the Arabian Journal of Business and Management Review*, *1*(12), 126. 2012

Murdoch, T. B., & Detsky, A. S. *The inevitable application of big data to health care*. *Jama*, *309*(13), 1351-1352. 2013

Provost, F., & Fawcett, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc. 2013.

Schutt, R., & O'Neil, C. *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reilly. 2013.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. "Artificial Intelligence (AI) applications for COVID-19 pandemic". *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), 337-339. 2020

Wirth, R., & Hipp, J. "CRISP-DM: "Towards a standard process model for data mining". En *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, abril de 2000. Springer-Verlag.

ANEXO I

Resultado de la regresión logística realizada con W-logistic y la edad discretizada en cuatro categorías (joven, adulto, adulto-mayor y anciano)

W-Logistic

Logistic Regression with ridge parameter of 1.0E-8Coefficients...
ClassVariable 1 2
3==========Graymattervolume-Value 3.2733 1.1183 -0.0685

Whitemattervolume-Value -1.2515 -0.1931 -0.4014

Cerebrospinalfluidvolume-Value -1.8984 -0.916 0.4845

Righthippocampusvolume-Value 0.1921 0.2508 0.1928

Lefthippocampusvolume-Value 0.5437 0.502 0.4273

Righttemporalvolume-Value -1.0977 -0.5046 -0.1562

Lefttemporalvolume-Value 1.124 0.851 0.706

Leftenthorynalcortexvolume-Value -0.1599 -0.184 -0.4114

Rightenthorynalcortexvolume-Value 0.4323 0.4799 0.3509

Brainparenchymafraction-Percentage 0.7016 1.7984 2.1359

PatientSex -0.4744 -0.3292 -0.2192

Intercept 1.6532 2.1756 3.0998Odds Ratios... ClassVariable
1 2
3==========Graymattervolume-Value 26.3993 3.0597 0.9338

Whitemattervolume-Value 0.2861 0.8244 0.6694

Cerebrospinalfluidvolume-Value 0.1498 0.4001 1.6234

Righthippocampusvolume-Value 1.2118 1.2851 1.2127

Lefthippocampusvolume-Value 1.7224 1.652 1.5331

Righttemporalvolume-Value 0.3337 0.6037 0.8554

Lefttemporalvolume-Value 3.077 2.342 2.0258

Leftenthorynalcortexvolume-Value 0.8522 0.832 0.6628

Rightenthorynalcortexvolume-Value 1.5408 1.616 1.4203

Brainparenchymafraction-Percentage 2.017 6.0398 8.4644

PatientSex 0.6222 0.7195 0.8032

ANEXO II

Resultado descriptivo utilizando la técnica de modelado de Árbol de decisión con la variable edad discretizada en cuatro categorías: joven, adulto, adulto-mayor, anciano mostrando una parte del árbol.

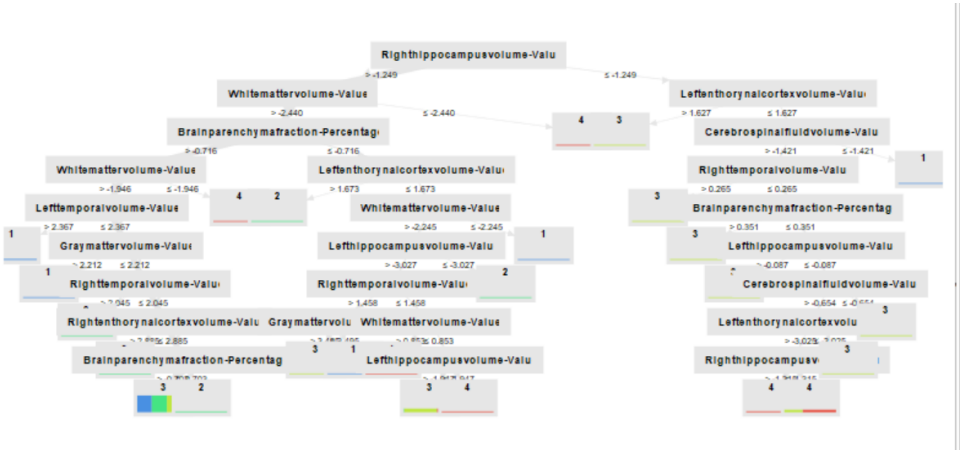
Tree

```
Brainparenchymafraction-Percentage > -0.683
|   Lefttemporalvolume-Value > 2.367: 1 {1=18, 2=1, 3=0, 4=0}
|   Lefttemporalvolume-Value ≤ 2.367
|   |   Rightenthorynalcortexvolume-Value > 2.885: 2 {1=0, 2=4, 3=0, 4=0}
|   |   Rightenthorynalcortexvolume-Value ≤ 2.885
|   |   |   Graymattervolume-Value > 2.730: 1 {1=2, 2=0, 3=0, 4=0}
|   |   |   Graymattervolume-Value ≤ 2.730
|   |   |   |   Leftenthorynalcortexvolume-Value > 2.997: 2 {1=0, 2=2, 3=0, 4=0}
|   |   |   |   Leftenthorynalcortexvolume-Value ≤ 2.997
|   |   |   |   |   Graymattervolume-Value > 0.814
|   |   |   |   |   |   Cerebrospinalfluidvolume-Value > 1.193
|   |   |   |   |   |   |   Righthippocampusvolume-Value > 1.236: 2 {1=0, 2=2, 3=0, 4=0}
|   |   |   |   |   |   |   Righthippocampusvolume-Value ≤ 1.236: 3 {1=0, 2=0, 3=4, 4=0}
|   |   |   |   |   |   Cerebrospinalfluidvolume-Value ≤ 1.193: 1 {1=134, 2=62, 3=29, 4=0}
|   |   |   |   |   Graymattervolume-Value ≤ 0.814
|   |   |   |   |   |   Cerebrospinalfluidvolume-Value > 1.130: 3 {1=0, 2=0, 3=1, 4=1}
|   |   |   |   |   |   Cerebrospinalfluidvolume-Value ≤ 1.130
|   |   |   |   |   |   |   Leftenthorynalcortexvolume-Value > -2.700
|   |   |   |   |   |   |   |   Brainparenchymafraction-Percentage > 3.282: 2 {1=0, 2=2, 3=0, 4=0}
|   |   |   |   |   |   |   |   Brainparenchymafraction-Percentage ≤ 3.282: 3 {1=122, 2=200, 3=383, 4=21}
|   |   |   |   |   |   |   Leftenthorynalcortexvolume-Value ≤ -2.700: 2 {1=0, 2=2, 3=0, 4=0}
Brainparenchymafraction-Percentage ≤ -0.683
```

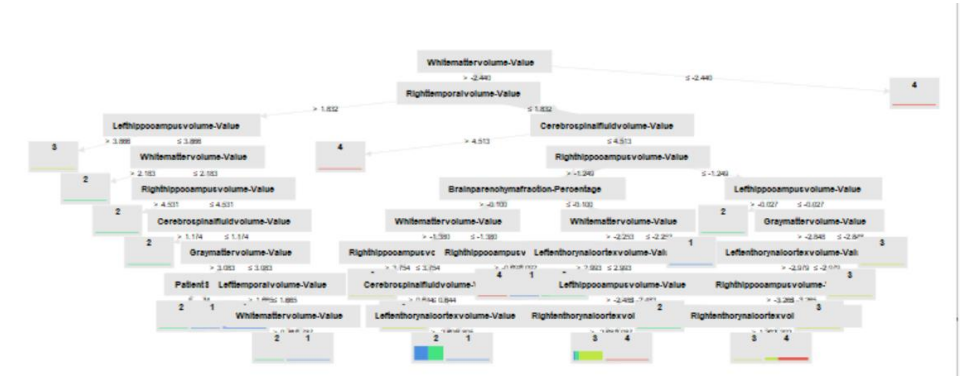
Anexo III

Representación gráfica de los dos primeros árboles de decisión obtenidos utilizando la técnica de modelización mediante *random forest.* A continuación, se muestran los dos primeros árboles obtenidos de los 100 totales.

Árbol 1



Árbol 2



Anexo IV

Descripción utilizando la técnica de modelización con Red Neuronal con la edad discretizada en cuatro categorías: joven, adulto, adulto-mayor y anciano.

ImprovedNeuralNet

Hidden 1

========

Node 1 (Sigmoid)

Graymattervolume-Value: -2.123

Whitemattervolume-Value: 2.641

Cerebrospinalfluidvolume-Value: -1.461

Righthippocampusvolume-Value: -0.645

Lefthippocampusvolume-Value: 0.564

Righttemporalvolume-Value: 0.395

Lefttemporalvolume-Value: 0.216

Leftenthorynalcortexvolume-Value: 1.158

Rightenthorynalcortexvolume-Value: 0.655

Brainparenchymafraction-Percentage: 2.286

Bias: -1.403

Node 2 (Sigmoid)

Graymattervolume-Value: -1.645

Whitemattervolume-Value: 2.133

Cerebrospinalfluidvolume-Value: -0.447

Righthippocampusvolume-Value: -0.151

Lefthippocampusvolume-Value: 0.441

Righttemporalvolume-Value: 0.616

Lefttemporalvolume-Value: 0.215

Leftenthorynalcortexvolume-Value: 0.541

Rightenthorynalcortexvolume-Value: 0.348

Brainparenchymafraction-Percentage: 1.129

Bias: -0.529

Node 3 (Sigmoid)

Graymattervolume-Value: -1.701

Whitemattervolume-Value: 1.996

Cerebrospinalfluidvolume-Value: 0.135

Righthippocampusvolume-Value: -0.049

Lefthippocampusvolume-Value: 0.307

Righttemporalvolume-Value: 0.657

Lefttemporalvolume-Value: -0.039

Leftenthorynalcortexvolume-Value: 0.264

Rightenthorynalcortexvolume-Value: 0.118

Brainparenchymafraction-Percentage: 0.385

Bias: -0.419

Node 4 (Sigmoid)

Graymattervolume-Value: 9.196

Whitemattervolume-Value: -6.818

Cerebrospinalfluidvolume-Value: -1.041

Righthippocampusvolume-Value: -0.716

Lefthippocampusvolume-Value: 0.731

Righttemporalvolume-Value: -1.111

Lefttemporalvolume-Value: 1.093

Leftenthorynalcortexvolume-Value: 0.662

Rightenthorynalcortexvolume-Value: -0.030

Brainparenchymafraction-Percentage: 1.009

Bias: -4.415

Node 5 (Sigmoid)

Graymattervolume-Value: -3.970

Whitemattervolume-Value: 1.774

Cerebrospinalfluidvolume-Value: 4.094

Righthippocampusvolume-Value: -0.896

Lefthippocampusvolume-Value: -3.042

Righttemporalvolume-Value: 0.784

Lefttemporalvolume-Value: -3.793

Leftenthorynalcortexvolume-Value: -3.643

Rightenthorynalcortexvolume-Value: 0.278

Brainparenchymafraction-Percentage: -4.462

Bias: -2.948

Node 6 (Sigmoid)

Graymattervolume-Value: -0.618

Whitemattervolume-Value: 1.132

Cerebrospinalfluidvolume-Value: 0.774

Righthippocampusvolume-Value: 0.059

Lefthippocampusvolume-Value: 0.261

Righttemporalvolume-Value: 0.571

Lefttemporalvolume-Value: -0.097

Leftenthorynalcortexvolume-Value: 0.404

Rightenthorynalcortexvolume-Value: 0.240

Brainparenchymafraction-Percentage: -0.312

Bias: -0.723

Node 7 (Sigmoid)

Graymattervolume-Value: 4.124

Whitemattervolume-Value: 0.542

Cerebrospinalfluidvolume-Value: -3.192

Righthippocampusvolume-Value: 0.709

Lefthippocampusvolume-Value: 1.419

Righttemporalvolume-Value: -1.856

Lefttemporalvolume-Value: 0.364

Leftenthorynalcortexvolume-Value: 0.988

Rightenthorynalcortexvolume-Value: 0.008

Brainparenchymafraction-Percentage: 3.497

Bias: 0.513

Node 8 (Sigmoid)

Graymattervolume-Value: -4.238

Whitemattervolume-Value: 3.781

Cerebrospinalfluidvolume-Value: 0.334

Righthippocampusvolume-Value: 0.959

Lefthippocampusvolume-Value: -0.150

Righttemporalvolume-Value: 0.406

Lefttemporalvolume-Value: -1.522

Leftenthorynalcortexvolume-Value: -1.186

Rightenthorynalcortexvolume-Value: -0.881

Brainparenchymafraction-Percentage: -0.076

Bias: 0.506

Output

======

Class '1' (Sigmoid)

Node 1: -2.887

Node 2: -2.308

Node 3: -2.296

Node 4: 5.048

Node 5: -2.339

Node 6: -1.313

Node 7: 1.257

Node 8: -3.379

Threshold: 0.412

Class '2' (Sigmoid)

Node 1: 1.390

Node 2: 0.271

Node 3: -0.315

Node 4: -4.230

Node 5: -1.693

Node 6: -0.680

Node 7: 2.094

Node 8: -1.894

Threshold: -0.784

Class '3' (Sigmoid)

Node 1: 0.873

Node 2: 0.262

Node 3: 0.075

Node 4: -7.634

Node 5: -3.721

Node 6: -0.418

Node 7: -2.307

Node 8: 1.920

Threshold: 0.843

Class '4' (Sigmoid)

Node 1: -1.180

Node 2: -0.436

Node 3: -0.069

Node 4: -2.018

Node 5: 2.639

Node 6: 0.152

Node 7: -4.912

Node 8: -0.339

Threshold: -0.450