Programa de Doctorado en Ingeniería de la Información y del Conocimiento

APRENDIZAJE AUTOMÁTICO E INTELIGENCIA ARTIFICIAL APLICADO A MODELOS DE CLASIFICACIÓN Y REGRESIÓN

Tesis Doctoral presentada por JESÚS MANUEL PUENTES GUTIÉRREZ

Directores:
DR. MIGUEL ÁNGEL SICILIA URBÁN

DR. SALVADOR SÁNCHEZ ALONSO

Alcalá de Henares, Julio 2022

Agradecimientos

Quisiera agradecer, con total sinceridad, la paciencia, el esfuerzo realizado, el detalle prestado y la profesionalidad reflejada por mis tutores y directores de este doctorado. A pesar de mi edad, me alegra mucho sorprenderme todavía con la calidad de las personas y su buen hacer, algo que me ha animado a seguir adelante hasta conseguir mi objetivo. También, he de admitir que ha sido complicado compaginar mi trabajo, en ocasiones muy exigente, con la realización del doctorado, aunque también me ayudó que las áreas de negocio e investigación estuvieran relacionadas.

Quisiera agradecer al Dr. Salvador Sánchez Alonso por ayudarme a iniciar este viaje, por su apoyo incondicional en todo momento, por su buen hacer profesional y por todos sus consejos y experiencia aportados que lograron ayudarme mucho durante este recorrido. Igualmente agradecer al Dr. Miguel Ángel Sicilia Urbán por el esfuerzo y tiempo dedicados, su profesionalidad, sus grandes propuestas de colaboración y por su visión e ideas que en muchas ocasiones me ayudaron a ampliar mi percepción de la investigación. También a la Dra. Elena García Barriocanal por su esfuerzo y tiempo dedicado, por su buen hacer y por su atención al detalle consiguiendo ver aquello que en ocasiones se me ha escapado a la vista. Sinceramente, gracias, sin vosotros no habría llegado hasta este momento.

También quisiera agradecer la paciencia y la comprensión de mi familia que también me han acompañado en todo este recorrido, en especial a mi hijo Jorge que en ocasiones ha compartido sus horas de estudio con las mías y a mi mujer Mercedes, con su gran apoyo moral, ambos realizando un gran esfuerzo a la hora de escuchar mis deducciones mostrando interés.

Resumen

Las técnicas de aprendizaje automático e inteligencia artificial son herramientas basadas en el análisis de datos para poder calcular la probabilidad de que sucedan determinados hechos o resultados, o para identificar la pertenencia a un determinado grupo basándose en sus propiedades. Mediante el uso del aprendizaje supervisado, en el cual se conocen previamente los resultados, se han realizado predicciones gracias a los datos obtenidos de los departamentos de administración y de atención primaria de un hospital, aunque el uso de estas mismas herramientas se puede extrapolar a otras áreas de conocimiento. Concretamente se ha estudiado los días que permanecen ingresados los pacientes debido a la causa que originó su ingreso a nivel hospitalario, donde se innova al no tratar de forma independiente los departamentos del hospital, y también se estudia las tasas de readmisión hospitalaria producidas por los pacientes al volver a ingresar en el hospital por motivos relacionados con la admisión previa, donde se mejoran las tasas predictivas gracias al uso de las técnicas más recientes y al empleo de redes neuronales combinadas con series temporales. Gracias al presente trabajo y a las técnicas utilizadas se conoce el comportamiento actual y futuro de los casos de uso sobre salud analizados, permitiendo incluso aprender con los datos analizados para adaptarse a los nuevos datos que puedan llegar en un futuro, potenciando así su uso.

Palabras clave

Aprendizaje supervisado, aprendizaje automático, inteligencia artificial, aprendizaje profundo, análisis predictivo, redes neuronales, salud.

Abstract

Machine learning and artificial intelligence techniques are tools based on the analysis of data in order to calculate the probability of certain events or outcomes occurring, or to identify membership of a certain group based on its properties. Using supervised learning, in which the results are known in advance, predictions have been made thanks to data obtained from the administration and primary care departments of a hospital, although the use of these same tools can be extrapolated to other areas of knowledge. Specifically, we have studied the number of days that patients remain hospitalised due to the cause that led to their admission in the hospital, where we have innovated by not examining the hospital departments independently, and we have also studied the hospital readmission rates produced by patients on re-admission for reasons related to the previous admission, where predictive rates are improved thanks to the use of the most recent techniques and the use of neural networks combined with time series. Thanks to this work and the techniques used, the current and future behaviour of the health use cases analysed is known, even allowing learning with the data analysed to adapt to new data that may arrive in the future, thus enhancing their use.

Keywords

Supervised learning, machine learning, artificial intelligence, deep learning, predictive analysis, neural networks, healthcare.

Índice General

Índice de Ilustraciones

Índice de Tablas

1 INTRODUCCIÓN

A lo largo de los últimos años, se han desarrollado multitud de investigaciones y estudios sobre tecnologías computacionales basadas en el análisis de datos que permiten calcular la probabilidad de que sucedan determinados hechos o resultados. Mediante el uso de estas tecnologías se construyen los modelos predictivos con los que se buscan soluciones a situaciones de la vida real, para así poder adelantarse y prepararse sobre los recursos que se van a necesitar a la hora de abordar una tarea o para poder evitar que ocurran sucesos no deseados y de esta forma poder prevenirlos con antelación, siempre que no se tenga influencia de factores externos que puedan alterar el comportamiento del modelo. Estos hechos son deseables en la mayoría de las actividades empresariales y cotidianas, como pueden ser en banca, donde predecir qué va a ocurrir con los diferentes activos puede suponer la diferencia entre obtener beneficios o sufrir pérdidas, como se analiza en el estudio publicado por Mamani (2020), en el que se predice el riesgo de morosidad en créditos bancarios. Otro ejemplo ilustrativo en el que se aplican técnicas predictivas puede ser la conducción autónoma de vehículos, aunque en este caso se trata de realizar predicciones en tiempo real que no se aplican en esta investigación, sí que muestra los posibles usos y utilidad de las tecnologías utilizadas en la predicción. En este ejemplo, adelantarse a la reacción de los peatones puede suponer salvar vidas y evitar accidentes, como se muestra en el artículo publicado por Mozaffari, Al-Jarrah, Dianati, Jennings y Mouzakitis (2020), en el cual se busca evitar peligros inminentes al predecir situaciones que pueden ocurrir en un horizonte cercano mientras se realiza la conducción de un vehículo. Este tipo de tecnologías también pueden aplicarse en las ciencias de la salud, ámbito de esta investigación, donde conocer con antelación que existe la posibilidad de sufrir una grave enfermedad puede ayudar a prevenirla, seleccionar el mejor tratamiento o incluso evitarla. Dentro del área de la salud, en el estudio que realizan Kalafi et al. (2019), el modelo propuesto busca predecir las posibilidades de supervivencia al cáncer de mama para seleccionar el mejor tratamiento que se debe aplicar. Como ejemplo de otros estudios realizados, puede verse la recopilación de artículos que investigan con este tipo de tecnologías y que se realiza en (Artetxe, Beristain y Graña, 2018) para predecir la readmisión hospitalaria. Hoy en día, dadas las actuales circunstancias donde las enfermedades y las pandemias cobran una especial relevancia, hace que los estudios sobre la salud sean especialmente demandados y relevantes.

Los sistemas que asisten a los humanos, como son los sistemas de salud, son uno de los sistemas de más utilidad y serán uno de los entornos de aplicación en los que más se evolucione en esta área de inteligencia artificial y aprendizaje automático, dadas las previsiones realizadas por entidades como el McKinsey Global Institute, el cual estima que en 2030, el 70% de las empresas a nivel mundial habrá adoptado algún tipo de inteligencia artificial, o por afirmaciones realizadas por Iglesias, García, Puig y Benzaqué (2020), comentando que los procesos relacionados con la inteligencia son tecnologías dirigidas a realizar tareas de forma más ágil, más eficiente y ayudando a las personas a tomar mejores decisiones incluso sin intervención humana. La evolución en los sistemas

de salud es previsible no solo por la mejora en la calidad de vida de las personas que puede suponer, sino por el impacto económico que puede suponer su ayuda al aliviar caídas económicas mediante la previsión, tal y como se ha comprobado con la pandemia del COVID-19, que ha provocado notables caídas en las previsiones económicas de crecimiento en las principales economías mundiales. Las tecnologías más recientes relacionadas con el Deep Learning o aprendizaje profundo están logrando grandes avances en este tipo de tecnologías, ya que son capaces de aprender sin la intervención humana previa, extrayendo del propio algoritmo las conclusiones acerca de la semántica embebida dentro de los datos. En (Shinde y Shah, 2019) podemos observar cómo se analizan las diferentes aplicaciones de aprendizaje automático y aprendizaje profundo, comprobando las múltiples áreas de uso. El poder detectar enfermedades en imágenes como radiografías o escáneres en las que el ojo humano no es capaz de detectar, o el poder entender instrucciones habladas para que un brazo robotizado pueda realizar operaciones de alta precisión, son algunas de las capacidades que se consiguen actualmente con estas últimas tecnologías relacionadas con el aprendizaje profundo.

La observación de las investigaciones analizadas en el estado del arte llevó al planteamiento de tratar de mejorar las actuales metodologías predictivas con las últimas tecnologías y algoritmos conocidos para aplicarlos a los sistemas de salud, partiendo de la gran base de conocimiento que suponen los datos disponibles en las bases de datos actuales, siempre respetando las leyes de protección de datos de todos los usuarios de los sistemas sanitarios. Aunque estas técnicas pueden aplicarse a infinidad de campos y plantear solución a varias consultas, se utilizaron los departamentos de administración y de atención primaria de los sistemas de salud como área de investigación por su especial relevancia y su amplio campo de aplicación. La información contenida en estos departamentos, en principio, puede parecer genérica, pero con el tratamiento y estudio adecuado se pueden extraer conclusiones relevantes como las obtenidas en esta investigación. Un ejemplo de este hecho es la investigación realizada sobre la readmisión hospitalaria (ver caso de uso del punto 5.4), donde se estudian las tasas de readmisión producidas por los pacientes de un hospital al volver a ingresar por motivos relacionados con la admisión previa. Estas tasas se estudian en función del número de días que transcurren entre la primera admisión y la siguiente readmisión relacionada, siendo las readmisiones a 30 días las más comunes.

La importancia de las tecnologías de aprendizaje automático se pone de manifiesto al manejar grandes cantidades de datos y al aprender de los resultados ya obtenidos con anterioridad. La capacidad humana no puede retener y procesar cantidades tan elevadas de datos como las que se manejan actualmente. Como afirman Dash, Shakyawar, Sharma y Kaushik (2019), según el International Data Corporation (IDC), estiman que en el año 2005 toda la información digital ocupaba 130 exabytes (EB) y en el año 2020 ya ocupa unos 40.000 EB, y continúa creciendo cada vez más según aumentan también nuestras necesidades. Con estas cantidades de datos es complejo ser capaces de distinguir diferentes casuísticas, teniendo en cuenta todos los casos observados, sin errores, y aprendiendo de todo lo ocurrido con anterioridad. Intentar predecir lo que va a ocurrir o tratar de conocer las posibles enfermedades que una persona va a tener, basándose en toda la experiencia anterior de otros pacientes, es la verdadera potencia del aprendizaje automático. Además, toda esta tecnología es

neutral, sin influencia de factores humanos como el cansancio, comportamientos éticos erróneos o decisiones subjetivas que afecten a las decisiones o detecciones de enfermedades.

También debe tenerse en cuenta que todos estos sistemas no serían posibles sin el componente principal de todos los algoritmos: el dato. Tener los datos apropiados, así como tenerlos de la forma adecuada y libre de errores, es fundamental para poder llegar a la fase final que supone el tratamiento con las propias tecnologías de aprendizaje automático o aprendizaje profundo. Por ello, es preciso también conocer las distintas formas en que se puede presentar la información, disponer de datos con calidad suficiente y pasar por una serie de procesos que permitan obtener el dato de las fuentes de datos de diversas procedencias. Será necesario prepararlo y limpiarlo para que se encuentre libre de errores y así poder procesarlo adecuadamente con los algoritmos. Se necesitará pasar por diferentes fases que permitan llegar a obtener conocimiento suficiente del dato, de forma que se conozca su distribución, que permita saber si la información se encuentra desbalanceada, si contiene errores que puedan obtener conclusiones erróneas o si son insuficientes para poder llegar a realizar un estudio de este tipo o no. Un estudio previo de la información que se dispone y pasar por los procesos mencionados es básico y necesario antes de poder acometer este tipo de estudios.

Un último proceso que también debe llevarse a cabo es el análisis de los resultados finales y el empleo de métricas para evaluar los resultados. Estas métricas deben permitir conocer la calidad de los resultados obtenidos y saber si los resultados y conclusiones finales son fruto de los buenos índices obtenidos o debido a la ausencia suficiente de datos, de forma que se estén obteniendo conclusiones erróneas.

De forma general, en investigaciones anteriores, al utilizar tecnologías de reciente creación y que se encuentran en evolución continua, tienden a quedar obsoletas por investigaciones más recientes con nuevas tecnologías que mejoran las precisiones en la predicción, o con nuevas herramientas que mejoran la forma de tratar o seleccionar los datos para obtener mejores resultados. Esta evolución en las metodologías se ha detectado en estudios como el realizado por Whitlock et al. (2010), donde realizan un análisis predictivo sobre la readmisión hospitalaria a 30 días con pacientes que padecen pancreatitis aguda. En este estudio obtienen una precisión del 83% utilizando el algoritmo de Logistic Regression, el cual es un algoritmo muy utilizado y, aunque es perfectamente válido, existen nuevas tecnologías como una de las utilizadas en esta investigación que pueden mejorar los resultados. En concreto, las redes neuronales convolucionales LSTM basadas en secuencias temporales mejora los resultados de Whitlock et al. (2010) en un 9,8% basándose en pacientes de patologías similares y en readmisiones con el mismo número de días. Esta tendencia en la mejora de los resultados y en la adopción de nuevas técnicas puede observarse en la revisión realizada por Artetxe, Beristain y Graña (2018), donde analizan las metodologías utilizadas en la readmisión hospitalaria con 76 artículos desde el año 1992 hasta el año 2018. Tal y como muestran de forma gráfica y por escrito, se evoluciona con el tiempo desde técnicas de regresión hasta técnicas de aprendizaje automático como Random Forest o redes

neuronales en los últimos años. También confirman la mejora realizada en la precisión, además de la adopción de técnicas más recientes.

Centrándose en el ámbito de esta investigación dentro del ramo de salud, además se comprueba que las investigaciones se encuentran centradas sobre patologías concretas y pertenecientes a especialidades específicas, lo que provoca que se reduzca la visión global de las investigaciones para análisis a una mayor escala. Por ejemplo, la investigación realizada por Chang y Lu (2016) se centra sobre una patología concreta, estudiando mujeres que han sido diagnosticadas con fibroma uterino, limitando su utilidad. En el caso de los estudios realizados por Rouzbahman, Jovicic y Chinell (2017) y Lowell et al. (1997), se centran en una especialidad específica, analizando los pacientes que son ingresados en el departamento de la unidad de cuidados intensivos y los pacientes en hospitales psiquiátricos, respectivamente. El hecho de disponer de una visión global tiene ventajas, como el poder tomar decisiones económicas para elegir más convenientemente las inversiones a realizar sobre especialidades concretas en las que es más necesario, aprovechando mejor los recursos disponibles.

En la investigación que se ha realizado, se han aplicado las últimas técnicas que se han ido descubriendo e implementando, consiguiendo mejorar los resultados y las precisiones, aplicando incluso técnicas basadas en la temporalidad de los hechos sobre las últimas tecnologías de redes neuronales, buscando descubrir nuevas vías de mejora de los resultados. Desde los inicios de la investigación, se eligieron técnicas asentadas con algoritmos como K Nearest Neighbours (KNN), Logistic Regression o la gama de algoritmos pertenecientes a Support Vector Machine (SVM), tal y como realizan investigaciones como (Tanuja, Acharya y Shailesh, 2011) y (Aghajani y Kargari, 2016). A todas estas técnicas se les aplicaron mejoras mediante los algoritmos optimizadores denominados *Hyperopt* y *GridSearchCV*, los cuales permiten elegir la parametrización que mejores resultados ofrece para el tipo de datos utilizado, aplicando de esta forma las últimas técnicas disponibles. No solo se aplicaron estas técnicas evolucionadas, sino que también se implementaron los últimos avances en redes neuronales combinando tipos de redes diferentes logrando mejorar ligeramente los resultados de cómo se obtenían de forma independiente, combinando redes neuronales convolucionales LSTM con redes en series temporales, confirmando la existencia de un patrón temporal al mejorar los resultados. Siguiendo la misma línea, se utilizó una tecnología denominada *Autosklearn,* la cual logra obtener los mejores resultados permitiendo al propio método elegir la técnica que considere como la más apropiada para el conjunto de datos facilitados. Aunque no se logró mejorar los resultados obtenidos por los algoritmos que buscaban la mejor parametrización de las técnicas comentadas anteriormente, sí que permitió corroborar que se estaban eligiendo los parámetros más adecuados para las técnicas empleadas. También se ha escogido una necesidad ya investigada, pero aplicándola a varios departamentos con distintas especialidades, de forma que permite compararlos entre sí y tener una visión más amplia en la mejora de la problemática tratada. En esta investigación, dentro de los casos de uso mostrados en los que se aplican las metodologías de aprendizaje automático, se analizó los días de estancia que permanecen ingresados los pacientes de un hospital. De esta forma se estudió el comportamiento de las predicciones en cualquiera de los departamentos, logrando extraer una visión global de los departamentos que mejores tasas predictivas obtienen,

en lugar de centrarse en un único departamento o especialidad, tal y como hacen en el estudio realizado por Yang, Wei, Yuan y Schoung (2010), donde centran su análisis en la unidad de quemados. Analizando un único departamento no se puede evaluar si este departamento obtiene resultados acordes a otros departamentos o si podría necesitarse algún tipo de corrección por tener tasas predictivas altas o bajas, tal y como si podría deducirse a partir de la investigación aquí realizada.

2 OBJETIVO

El objetivo principal de esta investigación es aplicar las últimas tecnologías en aprendizaje automático e inteligencia artificial sobre modelos clasificatorios, para así mejorar los resultados realizados por investigaciones previas existentes en cuanto a tasas predictivas y en cuanto a cubrir posibles carencias o evolucionar dichas investigaciones. Otros objetivos secundarios derivados del principal serían:

- Extraer conocimiento a partir de los datos de salud genéricos obtenidos de un sistema hospitalario mediante la ampliación del ámbito de investigación

- Mejorar la visión global del entorno de aplicación o del área de investigación de trabajos previos mediante análisis de ámbito más amplio, de forma que permita aumentar su utilidad y poder aplicarlo a otras unidades de aplicación como la financiera.

- Aplicar las últimas tecnologías de Inteligencia Artificial y aprendizaje automático a las bases de datos de salud buscando una mejor precisión y eficiencia en las predicciones

- La creación de un algoritmo específico sobre la detección de diagnósticos relacionados con un tipo de patología específica

- La creación de un algoritmo sobre la detección de un reingreso hospitalario en función del rango de días que se defina

- Obtener los mejores resultados predictivos posibles aplicando redes neuronales convolucionales LSTM (Long Short-Term Memory) y técnicas sobre selección de características correlacionadas, optimizadores bayesianos de los parámetros de las tecnologías y selector automático de tecnología de aprendizaje automático sobre los mismos modelos, de modo que se pueda elegir el mejor método para cada caso particular

3 METODOLOGÍA

En esta investigación, para alcanzar los objetivos propuestos, se ha diseñado una serie de casos de uso. Estos casos de uso fueron diseñados buscando ir evolucionando hasta lograr el objetivo principal, para así aplicar y aprender las últimas tecnologías existentes sobre la inteligencia artificial y el aprendizaje automático, siempre buscando su utilidad al perseguir mejorar los resultados obtenidos por investigaciones precedentes. El primer caso de uso, consistente en un análisis en profundidad del conjunto de datos disponible para abordar la investigación, fue seleccionado en primer lugar para lograr ir evolucionando y plantear la base introductoria hacia el objetivo principal. Con este caso de uso se tiene una visión clara de las líneas de investigación que se pueden seguir para obtener resultados aceptables y también permite conocer las líneas que no llegarían a materializarse en investigaciones viables. No realizar este caso de uso suele derivar en largas investigaciones que pueden llegar a obtener resultados no coherentes o incluso resultados demasiado evidentes en los que no hubiera sido necesario una investigación de este tipo para obtener la misma conclusión. El análisis se comenzó realizando reuniones con el personal del hospital para consultas y buscando información sobre los conceptos de salud desconocidos, para así conocer exactamente cada característica o campo de los que se compone la base de datos facilitada por el hospital. El conocer esta información permite saber, por ejemplo, si un registro nulo es correcto o no, o si los distintos valores que presenta una característica son todos los que existen o es que faltan casos en los que no se ha registrado ninguna casuística de ese tipo durante el periodo de tiempo en el que se ha realizado la extracción de datos. Posteriormente se realizaron representaciones gráficas mediante histogramas, gráficas de distribución, tablas, estudios de correlación y análisis de resultados, para de esta forma conocer en profundidad el conjunto de datos de salud del sistema hospitalario y estudiar las líneas de investigación a seguir. Este primer caso de uso, además de acercarnos al objetivo principal, ha permitido alcanzar uno de los objetivos secundarios de la investigación. Ha permitido extraer conocimiento sobre los datos de salud, ofreciendo una visión global sobre los pacientes del hospital y sobre la situación y distribución del propio hospital. Finalmente, también ha permitido la orientación de las posibles líneas de investigación a seguir.

El segundo caso de uso se eligió como consecuencia del resultado obtenido a partir del primero, gracias a la combinación de dos características existentes en la base de datos para obtener una nueva. Analizando las fechas disponibles, se calcula una nueva característica mediante la resta de las fechas de salida y de ingreso al hospital. De esta forma se obtiene una nueva característica que se llama días de permanencia y que facilita el número de días que un paciente permanece ingresado en el hospital, siendo este valor un factor hospitalario muy estudiado para optimizar recursos y mejorar la calidad de atención al paciente. Esta nueva característica es la que se utiliza para realizar el segundo caso de uso, prediciendo cual va a ser el número de días de ingreso de un paciente. Se realiza una revisión del estado del arte sobre las técnicas empleadas y sobre qué contexto se predicen los días de ingreso para estudiar si existe la posibilidad de innovar o de mejorar los resultados existentes. Tras el análisis, se aplican todas las

distintas técnicas que han sido utilizadas en los distintos estudios aplicando las últimas innovaciones a los algoritmos utilizados, aunque los resultados no mejoran los resultados ya publicados por otros estudios. Pero se descubre que todos los estudios encontrados se centran en especialidades o departamentos concretos, no existiendo ningún estudio que realice un estudio comparativo a nivel hospitalario. De esta forma, se estudian las especialidades más aconsejables para analizar los departamentos en los que mejor pueden predecirse los días de permanencia, de modo que sean más aconsejables ahorrando costes, aunque reservando menos recursos. Por consiguiente, el segundo caso realiza este estudio sobre todos los departamentos gracias a la disponibilidad de estos datos en el conjunto de datos utilizado. Se realiza un extenso análisis utilizando todos los algoritmos empleados por otros estudios en cada departamento, aplicando mejoras a cada algoritmo con las últimas técnicas disponibles, como son bibliotecas para emplear el algoritmo más óptimo, selector de parámetros más óptimos para cada técnica utilizada, un selector de las características más relevantes del conjunto de datos disponible y herramientas para el particionado de datos que permiten obtener representaciones equilibradas para los datos de entrenamiento y pruebas. Gracias a este caso de uso conseguimos alcanzar dos de los objetivos secundarios de esta investigación. Se realiza un análisis de mayor alcance al empleado por otros estudios y se consigue tener una visión global sobre una característica estudiada por los hospitales, pero que no ha sido planteada anteriormente y que ofrece nuevos enfoques de uso. También se logra aplicar nuevas tecnologías de aprendizaje automático para obtener los mejores resultados posibles.

Finalmente, en el tercer caso de uso se decidió investigar sobre la readmisión hospitalaria, otra de las grandes áreas de estudio utilizadas en salud hospitalaria y que, a su vez, existe la posibilidad de aplicarlo con el limitado conjunto de datos disponible. Aunque en un principio se pensó en tratar exclusivamente los modelos clasificatorios, finalmente se utilizaron también los modelos de regresión para demostrar las bondades del modelo elegido, como ocurre en el tercer caso de uso de esta investigación. La técnica aplicada en el tercer caso combina series temporales de datos con una red neuronal LSTM convolucional, habitualmente empleada en modelos de regresión, por lo que también se prepararon los datos para utilizarlos con tecnologías de regresión y así comparar los resultados con el modelo elegido. Igualmente, el modelo diseñado en este caso de uso se comparó preparando los datos y aplicando otras técnicas existentes basadas en árboles de decisión, ya que son técnicas muy utilizadas dados los buenos resultados obtenidos con este tipo de datos. De esta forma se consigue tener una referencia del comportamiento del modelo seleccionado al compararlo con las técnicas más extendidas y que mejores resultados obtienen. En este caso no se utilizó la técnica más utilizada en las predicciones sobre readmisión hospitalaria, Logistic Regression. Se trata de una técnica bastante asentada y no se buscaba únicamente mejorar los resultados, sino también innovar aplicando las últimas técnicas aparecidas en inteligencia artificial dentro del aprendizaje profundo. Gracias al modelo diseñado se consiguió mejorar los resultados y además detectar que presentan un patrón temporal. También se consigue, gracias a la técnica utilizada, que el modelo vaya aprendiendo y mejorando según se le alimente con nuevos datos según estén disponibles. Por medio de este caso de uso se alcanza el objetivo principal y tres de los objetivos secundarios, dos de ellos al utilizar las últimas técnicas disponibles usando métodos de aprendizaje

profundo, como es el tipo de red neuronal utilizado en este caso de uso. El tercer objetivo secundario se consigue debido a que para este caso de uso se necesita elaborar algoritmos propios de forma que permitan, en primer lugar, detectar diagnósticos sobre la patología de estudio elegida (patologías de digestivo) y, en segundo lugar, detectar cuando se trata de una readmisión concreta según el número de días que se defina. De esta forma se preparan los conjuntos de datos especializados que tratará posteriormente la red neuronal.

Para realizar los casos de uso, se han seguido unas pautas generales comunes en las investigaciones con este tipo de tecnologías, aplicándolas, adaptándolas y enfocándolas al área de salud. Igualmente ha sido necesario conocer el conjunto de datos, extraer conocimiento y adaptarlo a las necesidades particulares de cada caso de uso que, en este caso, al tratarse del mismo conjunto de datos, los procesos han sido comunes. A continuación, se detallan las fases que se han seguido en los casos de uso aplicados.

3.1 DIRECTRICES METODOLÓGICAS GENERALES DE MACHINE LEARNING

Tal y como se analiza en (Studer et al., 2021), todo proyecto de Machine Learning debe ser fiel y pasar por una serie de fases, basándose en el modelo CRISP–DM (Cross Industry Standard Process for Data Mining, ver definición en (Saltz y Hotz, 2022)), pero adaptado a las tecnologías de Inteligencia Artificial y aprendizaje automático, al cual denominan CRISP–ML. Tal y como también exponen en (Studer et al., 2021), donde proponen un modelo de procesos que se adapta al desarrollo de aplicaciones de aprendizaje automático, tanto para las organizaciones como para los investigadores y, en general, para usuarios de este tipo de tecnologías, es necesario seguir una guía a través del ciclo de vida de los proyectos de aplicaciones sobre aprendizaje automático.

A continuación, se exponen las fases o procesos generales de la metodología ML (del inglés Machine Learning o aprendizaje automático), relacionándola con alguna de las fases seguidas en esta investigación.

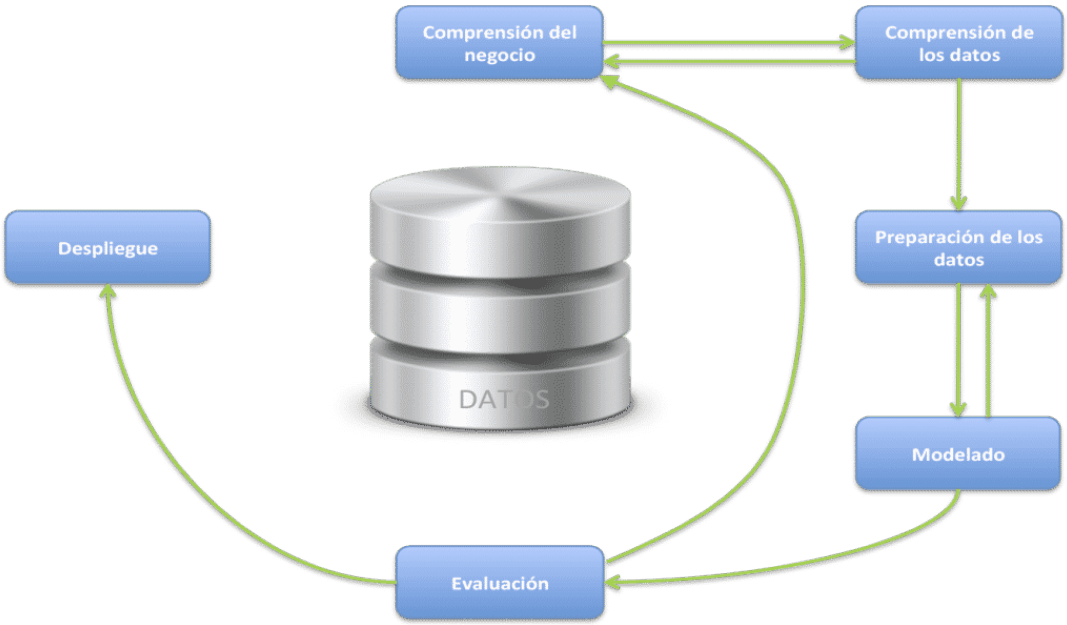


Ilustración 1. Modelo CRISP (Cross Industry Standard Process)

- *Análisis*. En primer lugar, se fijan unos objetivos claros iniciales que permitan alcanzar la meta que se quiere conseguir y analizar si es factible conseguirlo, ya que puede no ser posible llegar al objetivo con los datos que se disponen. En ocasiones se pueden llegar a obtener soluciones perfectas para problemas diferentes a los que se habían fijado como objetivo. Para llegar a superar esta fase inicial es necesario entender los datos que se disponen y comprender la lógica de funcionamiento del área de estudio que se está analizando. Como se está investigando sobre datos de salud, es necesario recabar información y entender términos médicos para comprender el significado de las características que se están analizando. Además, en el primer caso de estudio de esta investigación se necesitó relacionar dos características para obtener la característica-objetivo buscada en ese caso: el número de días de estancia de los pacientes. En el segundo caso de estudio, fue necesario analizar la relación existente entre los diagnósticos realizados por los médicos especialistas para detectar si estaban relacionadas con la enfermedad específica que se buscaba analizar. Este proceso junto al de preparación de los datos es uno de los que más esfuerzo requieren y en los que más tiempo es necesario invertir. De hecho, es aconsejable no avanzar a las siguientes fases sin haber conseguido los objetivos fijados y sin estar seguros de que las metas a conseguir son las que se necesitan.

- *Definición de criterios de evaluación*. En esta fase de la metodología se busca la forma de evaluar el modelo que se ha elegido para validar su calidad. La forma habitual de evaluar los modelos de aprendizaje automático es obtener el valor del error cometido en las medidas de predicción realizadas. Típicamente para problemas de regresión se suele utilizar el error cuadrático medio y para problemas de clasificación la entropía cruzada (véase (Bischel y Salmerón, 2013;

Foody, 1995)). En esta investigación, al utilizarse principalmente problemas de clasificación, los métodos de evaluación se basan en las medidas mediante la entropía cruzada y se utiliza la desviación media con respecto al valor real aplicado como una de las medidas a los problemas de clasificación que se han utilizado como casos de estudio. También se utilizan otras medidas como la precisión y la exhaustividad.

- *Análisis del modelo a implementar*. Se debe estudiar si existen otros modelos que solucionen el mismo problema y analizar si se pueden mejorar los resultados existentes, al igual que se ha realizado en esta investigación (ver sección 4) mediante un estudio previo de las materias a investigar por medio de una revisión del estado del arte. También, posteriormente a la elección del modelo para cada caso de uso estudiado, debe realizarse un nuevo análisis para comprobar que, en caso de existir un estudio similar, se mejora la propuesta, o para verificarse que el modelo creado es un modelo no existente. También debe estudiarse el rendimiento de la solución elegida para comprobar si es factible su aplicación. Como norma general, en la práctica, un modelo simple con mayor número de datos suele funcionar mejor que un modelo complejo con menor número de datos. Este punto se ha comprobado particularmente en el segundo caso de uso sobre readmisión hospitalaria (ver sección 5.3), donde el modelo elegido funcionaba con más fiabilidad y mayor precisión en los casos en los que tenían más registros para permitir mejorar el aprendizaje y el proceso de entrenamiento. En este punto puede ser necesario volver al punto de partida para elegir otro modelo que encuentre la solución buscada, si no cumple con los requisitos mencionados en esta fase.

- *Preparación de los datos*. Es bastante común tener que adaptar los datos según los algoritmos y técnicas que se hayan elegido para abordar la solución elegida. La base de datos con la que se trabaje habitualmente no tendrá el formato correcto que acepten las tecnologías que se van a utilizar. Por lo tanto, puede ser necesario aplicar varios procesos hasta adecuar la información al formato correcto, tal y como fue necesario en el segundo caso de uso de esta investigación, donde los requerimientos del algoritmo utilizado precisaban la utilización de valores numéricos para poder realizar los cálculos internos a base de distancias entre elementos. Por ello, se necesitó preparar los datos para realizar conversiones de variables categóricas a variables numéricas, de forma que se asignase un valor numérico a cada variable categórica, como puede ser el diagnóstico de cada paciente. Este caso particular, aunque varios diagnósticos se identifican mediante cifras numéricas, otros se identifican mediante valores que contienen letras, lo que los convierte en valores no numéricos, siendo necesario también realizar esta transformación. También es normal que la información disponible pueda contener datos incompletos, datos erróneos, errores tipográficos, datos insuficientes, etc. Para corregir estas formas incorrectas de la información y poder trabajar con un conjunto de datos con calidad suficiente, puede ser necesario eliminar aquellos datos incompletos que no son parte de

una característica concreta y siempre que sean una minoría. También puede ser posible sustituir aquellos valores que no existan con un valor razonable y que tengan sentido. Por ejemplo, puede sustituirse un valor que no exista con el valor medio del resto de muestras de esa característica concreta. También es viable no realizar ninguna acción dependiendo de la técnica utilizada, porque pueda manejar datos incompletos, al igual que ocurre al utilizar el algoritmo del segundo caso de uso de esta investigación. En cambio, las particularidades de algunas de las tecnologías utilizadas en el primer caso de uso sobre la estancia de los pacientes en el hospital, no permitía la utilización de valores nulos o no existentes entre los elementos del conjunto de datos utilizado. En este caso se utilizó la sustitución de los valores nulos por el valor 0. También es posible, tal y como fue necesario en esta investigación, obtener una nueva característica a partir de las características ya existentes, de forma que se obtienen nuevas funcionalidades. Esta fase es una de las que más tiempo necesita para dejar completamente los datos preparados.

- *Normalizar datos*. En muchos casos es útil normalizar los datos para hacerle más fácil el aprendizaje a la técnica utilizada o evitar efectos no deseados, como es el sobre ajuste de los datos. En esta fase se ajustan todos los datos a una escala similar, aplicando técnicas ya preparadas mediante bibliotecas existentes o de forma manual. Esta técnica fue aplicada sobre los datos numéricos de la red neuronal utilizada en el caso de uso sobre la readmisión hospitalaria, consiguiendo una menor dispersión de los datos y un mejor comportamiento.

- *Construir el modelo*. Esta fase, en ocasiones, no requiere un esfuerzo excesivo gracias a que se puede elegir un modelo preexistente que se adapte a la problemática disponible, tal y como se realiza en el segundo caso de uso de esta investigación, donde se utiliza una red neuronal básica aplicada en otras investigaciones y se adapta al caso particular de las readmisiones hospitalarias en los sistemas de salud. En concreto, se aprovechan las primeras capas de nodos o neuronas de la red neuronal, común a estudios de aprendizaje supervisado, optimizando las configuraciones y adaptándose a los datos disponibles. Las últimas capas de neuronas también son utilizadas como base. Existen varias bibliotecas de código abierto de aprendizaje automático con modelos funcionales para poder ser reutilizadas. De esta forma se transfiere y reutiliza el conocimiento, siendo solo necesario la adaptación de los procedimientos a las circunstancias particulares de cada problemática. Aunque se debe tener en cuenta que, al reutilizar la configuración genérica de otros modelos, es posible que se genere un error que puede ser elevado para el modelo seleccionado, por lo que será necesario estudiar y adaptar los parámetros a cada modelo concreto.

- *Evaluación, análisis de errores*. En esta fase se ha obtenido un primer modelo y se necesita saber si este modelo es bueno o no con respecto al objetivo que se fijó en la fase de construcción. Por ello, se necesita una forma de analizar la calidad que se obtiene de un modelo estimando el comportamiento futuro del

modelo con respecto a objetos del dominio al que pertenece, como, por ejemplo, conocer el número de días que permanecerá ingresado en el hospital un paciente, conociendo datos personales, como sexo o edad y otros datos como el diagnóstico por el que ingresó o el procedimiento principal que se le aplicó. También, se deben analizar los errores para comprender qué es lo que se debe hacer para mejorar los resultados obtenidos. En esta parte también se debe intentar asegurar que el modelo es capaz de generalizar, es decir, debe evaluarse que el modelo es capaz de producir buenos resultados cuando se utilicen nuevos datos que no han sido utilizados previamente. Por ello se separa el conjunto de datos en dos partes, una parte para entrenar el algoritmo de forma que aprenda a detectar, por ejemplo, posibles reingresos hospitalarios, y otra parte, para comprobar la efectividad del algoritmo detectando datos que no conoce. En función de los resultados puede ser necesario iterar sobre fases anteriores varias veces, para conseguir un modelo más ajustado al objetivo inicial.

- *Integrar el modelo en un sistema*. Aunque esta fase no ha sido llevada a cabo en esta investigación, sí que sería factible realizarla, dejando los modelos desarrollados preparados para ser integrados con pocas variaciones. Además de utilizarse los modelos para demostrar hipótesis o para responder cuestiones planteadas en las investigaciones, la evolución natural también es ser integrado en un sistema de trabajo. Para poder integrar el modelo en un sistema, primero será necesario crear interfaces de datos para que se puedan obtener los datos de forma automática, recogiendo los datos del CMBD del hospital para volcarlo a ficheros con la estructura que se precise. Después será necesario comunicar el modelo con otras partes del sistema, de forma que las predicciones y los resultados puedan ser utilizados. Por ejemplo, al introducir los datos de un paciente, se realizará una estimación de forma directa de los días que permanecerá ingresado o se calculará el riesgo de padecer una enfermedad con un intervalo de confianza dado. Finalmente, también es útil poder monitorizar los errores del modelo para que avise en caso de que se incrementen con el tiempo, pudiendo reconstruir y realimentar el modelo con nuevos datos.

A modo de conclusión, todo modelo debe ser un sistema en continua revisión y mejora para que no quede obsoleto y para que sus resultados no lleven a tomar decisiones erróneas por un desvío en el comportamiento, debido a la aparición de nuevos patrones de comportamiento en los datos. Por ello se debe planificar dar un seguimiento al modelo, de manera que se reevalúe y realimente de forma continua. Se trata de un proceso continuo y reiterativo.

3.2 ALGORITMOS DE CLASIFICACIÓN

Aunque existe un mayor número de algoritmos de los que se comentan en esta sección, se van a tratar aquellos que se han utilizado en esta investigación. Se han seleccionado aquellos algoritmos más comunes y que mejores resultados han dado en otras investigaciones (ver sección 4.4).

Los algoritmos utilizados, por orden de utilización en las investigaciones, se exponen a continuación.

- *K Nearest Neighbours (KNN)*. Este algoritmo es un clasificador simple y eficiente que permite obtener buenos resultados en aquellos conjuntos de datos que no tienen un comportamiento lineal y que no presentan valores atípicos, como ocurre en el conjunto de datos que disponemos. Se debe tener en cuenta que es preferible utilizar otro método en caso de detectar problemas de eficiencia por el tamaño del conjunto de datos. El método que utiliza este algoritmo calcula la distancia euclídea entre las diferentes muestras para clasificarlas en función de las distancias obtenidas a las K muestras más cercanas definidas previamente. De esta forma clasifica cada muestra por clases y predice a qué clase pertenece.

- *Support Vector Machines (SVM)*. Se trata de un algoritmo que se compone de varios métodos dependiendo del núcleo o función principal que se utilice para su aplicación. Es útil para aquellos conjuntos de datos con un número elevado de dimensiones y que permiten separar los datos mediante una función matemática que sigan este patrón. Este algoritmo se comporta bien con valores atípicos, ya que es capaz de encontrar la separación lineal que exista entre estos valores mediante una función matemática. El algoritmo busca un conjunto de muestras para construir un plano que limite la separación entre dos clases distintas de muestras y así poder clasificarlas. Dependiendo de la función que se utilice para separar las muestras, tendremos un tipo diferente de algoritmo *SVM*. Las funciones utilizadas con este algoritmo en esta investigación fueron una función polinómica (*Polynomial kernel*), una función con base radial (*RBF*) y el grupo de *Support Vector Classifiers (SVC).* A su vez, dentro de este grupo *SVC*, se utilizaron funciones lineales (*SVC – Linear kernel*), mediante la técnica *one-vs-the-rest* que compara cada muestra con el resto, y no lineales (*SVC – Non Linear kernel*), con funciones basadas en el kernel *RBF* mediante la técnica *one-against-one* que compara una clase contra otra.

- *Naïve Bayes*. Este algoritmo parte de la base de que todas las características son independientes unas de otras, por lo que calcula la variación que se produce entre cada clase. Tiene un buen comportamiento con características poco o nada correlacionadas y con datos balanceados, con lo que sería apropiado para el conjunto de datos disponible en esta investigación, con características poco correlacionadas, pero balanceando las muestras. También es apropiado para muestras que sigan una distribución Normal o Gaussiana de los datos, al igual que suelen seguir todos los fenómenos naturales, tal y como se afirma en (Steven, 2009).

- *Decision Tree Classifier*. Este es uno de los algoritmos que mejores resultados han dado en el primer caso de uso de esta investigación, al igual que el algoritmo Random Forest, ya que está basado en este mismo algoritmo. Es un clasificador

que tiene un buen comportamiento con conjuntos de datos que presentan valores atípicos y que pueden tener datos desbalanceados, ya que clasifican todos los valores mediante reglas que van tomando decisiones según se avanza por una estructura en forma de árbol. Se debe prestar atención al utilizar este algoritmo, ya que puede producir un efecto de sobre ajuste. Este efecto puede ofrecer buenos resultados en los valores de entrenamiento al clasificar todo, pero puede dar resultados con baja precisión en los datos de prueba que son conocidos, pudiendo no tener reglas definidas en los nuevos valores.

- *Random Forest*. Al igual que el algoritmo anterior, ha obtenido los mejores resultados, ya que se basa en una estructura con múltiples árboles de decisión de bajo nivel de profundidad. Tiene los mismos usos y peculiaridades que el algoritmo *Decision Tree Classifier*, pero selecciona la mejor clase en función de los resultados obtenidos por los múltiples árboles de decisión generados.

- *Redes Neuronales*. Aunque se han utilizado varios modelos de redes neuronales, principalmente se comenta la que se ha estudiado con mayor profundidad en esta investigación. En general, las redes neuronales son inmunes al ruido en los datos (ver definición de ruido en (Gupta y Gupta, 2019)), tienen un aprendizaje adaptativo según se entrenan con nuevos datos y son capaces de auto gestionarse. La red neuronal que se ha utilizado en el último caso de uso es una red basada en secuencias temporales, ya que la idea de predecir una readmisión hospitalaria conlleva una temporalidad en la secuencia de detección de los reingresos. De hecho, a modo de comparación, se utilizó una red neuronal que no tenía en cuenta la temporalidad y, posteriormente, se utilizó una red neuronal convolucional LSTM (Long Short-Term Memory), que si la tiene en cuenta. Se comprueba que los resultados en la precisión mejoran ligeramente, lo que nos indica que la temporalidad es un factor que afecta al aprendizaje y al algoritmo utilizado en las predicciones.

3.3 MÉTRICAS

Se necesita disponer de una forma de analizar la calidad que se obtiene de los modelos utilizados, estimando el comportamiento futuro de esos modelos con respecto a objetos del dominio al que pertenecen dichos modelos. Para ello se utilizan diferentes métricas de análisis, algunas de ellas como las analizadas por Mishra (2018), que nos indicarán si los modelos y los algoritmos utilizados tienen un comportamiento correcto o no.

En esta investigación se han tratado los modelos clasificatorios principalmente. Por ello, este tipo de modelos se han evaluado midiendo la capacidad que tienen para predecir correctamente la clase a la cual pertenecen objetos que no se han utilizado en la construcción del modelo. En el segundo caso de uso (sección 5.3), se utilizaron métricas que consisten en medir la calidad del modelo mediante el cálculo del porcentaje de clasificaciones incorrectas que efectúa el modelo a partir de un conjunto de observaciones, siendo cada una de las operaciones clasificadas incorrectamente un

error de clasificación. El parámetro de calidad que se utiliza en este caso es la tasa de error:

$$Tasa\ de\ error = \frac{Errores}{Casos}$$

También se puede considerar como parámetro de calidad la tasa de aciertos del modelo, que se denomina precisión del modelo:

$$Precisión = \frac{Éxitos}{Casos}$$

Estas métricas utilizadas se calcularon bien mediante una función propia que calculaba las métricas o mediante funciones ya creadas pertenecientes a las bibliotecas propias del lenguaje Python. En concreto las bibliotecas utilizadas para estos cálculos corresponden a sklearn.metrics, y los procedimientos usados en este caso son accuracy_score y std() para la desviación estándar media producida en las medidas. En el tercer caso de uso (sección 5.4), se utilizaron otras métricas adicionales para tener un mayor control sobre los resultados, ya que, en este caso, se necesitaba ver con más detalle el comportamiento del modelo para ver si aprendía correctamente y para comprobar si no presentaba efectos adversos como el sobreajuste. Las métricas usadas en este caso de uso fueron las siguientes:

- *Sensibilidad*. También llamado *recall*, da la probabilidad de que, dada una observación realmente positiva, el modelo la clasifique realmente así.
- *Exactitud*. Calcula la probabilidad de que, dada una predicción positiva, el valor real también sea positivo. También es denominado *precision* en inglés. Mide lo confiable que es una clase.
- *Precisión*. También llamado *accuracy*, calcula el total de aciertos del modelo.
- *F1 Score*, donde mide la media armonía entre *precision* y *recall*. Combina *precision* y *recall* en una sola métrica.
- *ROC AUC*, donde mide la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio, permitiendo ver la calidad de las predicciones del modelo.
- *Matriz de confusión.* Se trata de una representación gráfica de la precisión que ofrece el modelo con dos clases: la clase predicha y la clase real. Muestra el número de muestras que cumplen los criterios que se muestran en la Tabla 1.

		Clase predicha	
		+	**-**
Clase real	**+**	Positivos verdaderos	Falsos negativos
	-	Falsos positivos	Negativos verdaderos

Tabla 1. Matriz de Confusión

La forma de definir los valores de la tabla sería la siguiente:

- *Positivos verdaderos* o número de predicciones correctas de la clase positiva (la clase que es cierta)
- *Falsos positivos* o número de predicciones incorrectas de la clase positiva
- *Falsos negativos* o número de predicciones incorrectas de la clase negativa (la que no es cierta)
- *Negativos verdaderos* o número de predicciones correctas de la clase negativa

En el tercer caso de estudio (sección 5.4) se utilizaron algoritmos de regresión para poder tener un punto de referencia con el que comparar los resultados obtenidos por los modelos de clasificación, dado que no existía un modelo similar publicado hasta la fecha. Por este motivo, se utilizaron las siguientes métricas propias de los modelos de regresión para poder evaluar el modelo:

- *Error Absoluto Medio*. Se calcula como un promedio de diferencias absolutas entre los valores objetivo y las predicciones realizadas. Se trata de una métrica que penaliza los errores grandes, por lo que no es tan sensible a los valores atípicos como si lo es el error cuadrático medio. Matemáticamente se calcula utilizando la fórmula:

$$Error\ Absoluto\ Medio = \frac{1}{N}\sum_{i=1}^{N}|y_{i-}\hat{y}_i|$$

- *Error Cuadrático Medio*. Mide el error cuadrado promedio de las predicciones realizadas. Cuanto mayor es el valor, peor es el modelo, siendo una métrica útil si se tienen valores inesperados. Si se tienen datos con ruido o no confiables, la cuadratura empeorará aún más el error. La bondad del modelo medida con esta métrica es relativa y debe realizarse con el apoyo de otras métricas y/o definiendo lo que se considera un buen resultado. Matemáticamente se calcula mediante la fórmula:

$$Error\ Cuadrático\ Medio = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Todas estas métricas también cobran especial importancia teniendo en cuenta, además, la relevancia del ámbito de aplicación. Se debe tener en cuenta el coste de predecir un falso positivo o un falso negativo según la aplicación a la que se destina el modelo. Por ejemplo, si queremos crear un clasificador que prediga correctamente si un paciente puede sufrir una enfermedad determinada, un falso positivo pondrá en tratamiento a una persona sana y un falso negativo la dejará sin él. En este caso las métricas se hacen especialmente necesarias y se debe crear un modelo más estricto.

3.4 CRITERIOS DE OPTIMIZACIÓN

Dentro de la metodología utilizada en esta tesis, se buscaron otros criterios que permitieran corroborar los resultados y/u optimizarlos en caso de que fuese posible. Uno de ellos es una técnica que permite seleccionar las condiciones óptimas para cada algoritmo utilizado. La herramienta utilizada es un optimizador bayesiano que aplica diferentes rangos de parámetros o funciones al algoritmo específico seleccionado. Se crea un listado en forma de elementos individuales o rangos numéricos, se le envía al optimizador y éste se encarga de testear todos los elementos mostrando el que mejores resultados obtiene. Se utilizó un optimizador bayesiano a partir de las bibliotecas *Hyperopt* del lenguaje *Python*, tal y como se explica en (Bergstra, Yamins y Cox, 2013).

Otro criterio utilizado ha sido un método que permitiese validar los resultados obtenidos mediante los algoritmos clasificatorios. Para ello se utilizó la tecnología llamada *AutoSklearn*, analizada en los artículos (Feurer et al., 2015; Feurer, Eggensperger, Falkner, Lindauer y Hutter, 2020), la cual consiste en un algoritmo que selecciona automáticamente el algoritmo predictivo de aprendizaje automático más óptimo, así como los parámetros más adecuados mediante un optimizador, incluyendo éste las mejoras que suponen los optimizadores bayesianos. Aunque se seleccionan los mejores parámetros, se deben informar los límites en cuanto a memoria y tiempo de ejecución. En concreto se utilizó el procedimiento *AutoSklearnClassifier()* de la biblioteca *autosklearn.classification* dentro del grupo de herramientas *auto-sklearn* versión 0.12.6.

3.5 ANÁLISIS E INTERPRETACIÓN DEL CONJUNTO DE DATOS

3.5.1 INTRODUCCIÓN

El dato es la materia prima de todos los recursos y algoritmos del aprendizaje automático y de la inteligencia artificial y, por lo tanto, es el punto de partida hacia el entendimiento y el conocimiento de las diferentes técnicas que se utilizan en la inteligencia artificial. Para poder saber cómo aplicar los algoritmos o cómo utilizar las herramientas para extraer conocimiento, se debe conocer el formato que tiene la información, cómo se encuentra estructurada, qué elementos la componen, cómo presentarla y cómo se va a poder tratar. De igual forma, para ser capaces de analizar y conocer el comportamiento de las áreas de conocimiento a las que se refiere cada característica o campo del conjunto de datos sobre el que se va a trabajar, se necesita realizar un estudio previo sobre los datos extrayendo su estructura y composición.

En el manejo de las técnicas de aprendizaje automático se utiliza una clasificación simple -numérico, texto o fecha- para que se utilicen de forma sencilla, eliminando la complejidad de las diferentes categorizaciones que realizan algunos lenguajes de programación o la que puedan emplear internamente los sistemas de inteligencia artificial. La misma sencillez se aplica al clasificar los tipos de variables, categóricas y cuantitativas o numéricas, tal y como se puede observar en la Ilustración 2, donde se

observa cómo se estructuran estos tipos, permitiendo centrarse en el contenido en sí mismo y en extraer conocimiento de forma más eficiente.

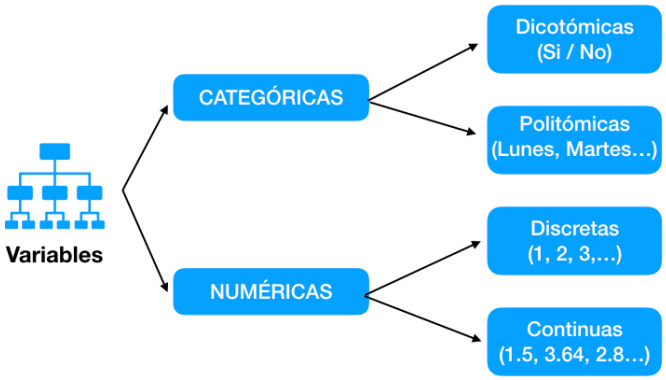


Ilustración 2. Clasificación de variables categóricas y cuantitativas

También se debe tener en cuenta que en España existen actualmente leyes como la LOPD (Ley Orgánica 3/2018 de Protección de Datos Personales y garantía de los derechos digitales), las cuales protegen el tratamiento de los datos personales como un derecho fundamental de forma que se limita por ley el uso de la informática para garantizar la intimidad personal y familiar de los ciudadanos. Por esta razón, se debe tener especial cuidado cuando se utilicen los datos públicos para no estar contraviniendo las leyes vigentes. Por este motivo el conjunto de datos que se ha utilizado en esta investigación no contiene ningún tipo de dato personal que permita identificar al individuo originario de la información, a pesar de que se traten de datos reales.

3.5.2 MATERIALES Y MÉTODOS

3.5.2.1 INTERPRETACIÓN DE LOS DATOS

El conjunto de datos con el que se trabaja en esta investigación se obtiene de una base de datos de un hospital español, a partir de un modelo extraído de un sistema de información estandarizado, en el que se anonimizaron los datos eliminando todos los detalles personales. La base de datos se utiliza en el ámbito de la Comunidad de Madrid y a nivel Nacional, y se denomina CMBD (Conjunto Mínimo Básico de Datos), creado por Consejería de Sanidad y Consumo (2004). Se trata de un conjunto de datos clínicos y administrativos que trabaja con el registro de las historias clínicas hospitalarias

codificadas. También utiliza la Clasificación Internacional de Enfermedades (9ª revisión) de Diagnósticos y Procedimientos (CIE-9-MC), elaborado por Ministerio de Sanidad, Servicios Sociales e Igualdad (2014). El CMBD contiene los datos de referencia del Sistema Nacional de Salud para el análisis comparativo de la casuística y del funcionamiento de los hospitales. Esta base de datos tiene una estructura fija estipulada por el SERMAS (Servicio Madrileño de Salud) y es de obligado cumplimiento para todos los hospitales de la Comunidad de Madrid.

Los datos obtenidos corresponden concretamente a los datos básicos recogidos con el alta de cada episodio de hospitalización, donde se extrae tanto la información genérica de los pacientes como la codificación de procedimientos y diagnósticos llevados a cabo. Este conjunto de datos contiene 45 características diferentes (o campos) con un total de 63.932 registros, obtenidos durante un periodo aproximado de 5 años, desde junio de 2010 hasta septiembre de 2015. En la Tabla 2 podemos observar la composición de este modelo estándar de datos (CMBD), así como el porcentaje de valores sin rellenar que nos encontramos en el conjunto de datos sobre el que se ha trabajado.

Nombre de la característica	Tipo	Descripción y valores	% valores nulos
HISTORIA_COD	Numérico	Código de historia clínica	0%
FECNAC	Fecha	Fecha de nacimiento	0%
SEXO	Numérico	Código numérico que identifica el género. Valores: 1 – hombre, 2 – mujer	0%
FECING	Fecha	Fecha de ingreso al hospital	0%
SERVING	Carácter	Código de caracteres que indica el departamento de ingreso	0%
SECCING	Carácter	Código de caracteres que indica la sección de ingreso	16%
FECALT	Fecha	Fecha de alta del hospital	0%
SERVALT	Carácter	Código de caracteres que indica el departamento de alta	0%
SECCALT	Carácter	Código de caracteres que indica la sección de alta	0%
TIPALT	Numérico	Código numérico que indica el motivo del alta del hospital (1 – Casa, 2 – Desvío a otro hospital, 3 – Alta voluntaria, 4 – Fallecimiento, 5 – Desvío a centro social de salud, 6 – Huida, 7 – Hospitalización en domicilio del paciente).	0%
FECINTERV	Fecha	Fecha de intervención quirúrgica en caso de que exista	71%
D1	Carácter	Código de caracteres que identifican el diagnóstico principal en el alta hospitalaria.	0%
D2	Carácter	Código de caracteres que identifican otros diagnósticos secundarios (D2 a D13).	5%
D3	Carácter		11%
D4	Carácter		18%
D5	Carácter		26%
D6	Carácter		34%
D7	Carácter		42%
D8	Carácter		50%
D9	Carácter		57%
D10	Carácter		63%
D11	Carácter		69%
D12	Carácter		76%
D13	Carácter		86%
P1	Carácter	Código de caracteres que identifican el procedimiento médico principal, quirúrgico y/o obstétrico.	0%
P2	Carácter	Código de caracteres que identifican otros procedimientos médicos secundarios, quirúrgicos y/o obstétricos (P2 a P18).	10%
P3	Carácter		23%
P4	Carácter		34%
P5	Carácter		46%
P6	Carácter		58%
P7	Carácter		70%
P8	Carácter		78%
P9	Carácter		85%

Nombre de la característica	Tipo	Descripción y valores	% valores nulos
P10	Carácter		90%
P11	Carácter		93%
P12	Carácter		95%
P13	Carácter		97%
P14	Carácter		98%
P15	Carácter		98%
P16	Carácter		99%
P17	Carácter		99%
P18	Carácter		99%
ED	Numérico	Edad del paciente	0%
Peso	Numérico	Peso relativo al coste relacionado con el coste medio que suponen las altas, cuyo valor medio es 1 (Valores tipo: X.XXXX)	0%
GRDS	Numérico	Valor decimal que indica el grupo relacionado por el diagnóstico	0%

Tabla 2. Listado de características del conjunto de datos

Como puede observarse, existen características como los diagnósticos y procedimientos secundarios que tienen un alto porcentaje de valores nulos o no utilizados. Esto se debe a que no es necesario rellenarlos, ya que no todos los pacientes tienen porqué tener muchos síntomas o aplicárseles varios procedimientos médicos. En general, el motivo principal por el que se acudió al centro es el que debe estar relleno.

3.5.2.2 EXTRACCIÓN DEL CONJUNTO DE DATOS INICIAL

En ocasiones, dependiendo del tipo de estudio que se realice, puede ser necesario extraer una parte del conjunto de datos disponible, debido a que algunas de las características no tengan correlación con el objeto de la investigación a realizar y, por lo tanto, no influyan en los resultados. De esta forma, utilizando solo una parte del conjunto de datos obtiene los mismos resultados y agiliza los procesos de cálculo. Para el análisis de la segunda parte de esta investigación (ver sección 5.3), se utilizó un subconjunto de las características disponibles del total. Es cierto que cuanta más información se tenga disponible, más posibilidades se tienen de descubrir conocimiento y que la propia ausencia de información es, en sí mismo, nueva información, pero para este caso se decidió extraer una parte. Los motivos por los cuales se llegó a esta conclusión fueron:

- Los algoritmos y técnicas utilizadas no necesitaban tanta información para extraer las predicciones y el uso de toda la información disponible suponía utilizar gran cantidad de recursos y ejecuciones de varios días con el riesgo de errores o cortes que obligaban a repetir los procesos. Para este tipo de modelos no es necesario disponer de grandes cantidades de datos para su análisis. Además, en un uso real del estudio con las técnicas utilizadas no harían viable su utilización. Al reducir el conjunto de datos la eficiencia aumentó notablemente, quedándose en unas horas por cada prueba.
- Se realizó un análisis del conjunto de datos utilizando el coeficiente de correlación de Pearson para descartar características que no tenían relación alguna con la característica que se iba a utilizar en esta primera parte (el número de días de estancia). Se realizó una prueba contrastando los resultados del conjunto de características seleccionado mediante el coeficiente de correlación

de Pearson contra el conjunto global de características y los resultados obtenidos al predecir los días de estancia apenas sufrieron variación (por debajo del 1%). Para extraer el conjunto de características definitivo y así corroborar los resultados del coeficiente de correlación de Pearson, se utilizó un algoritmo llamado *SelectKBest* que selecciona las mejores características entre un conjunto de datos siguiendo funciones estadísticas, extrayendo el mismo número de características que se habían analizado inicialmente. Para extraer las mejores características se utilizó la función *chi cuadrado* ($\mathcal{X}^2$), ya que es la función más adecuada para tareas de clasificación con características no negativas.

Para la primera parte de la investigación del segundo caso de uso (sección 5.3), se analizó más en profundidad el conjunto de datos, ya que se realizó un análisis de los días de estancia por cada uno de los departamentos que los pacientes habían visitado en el periodo de tiempo antes comentado (punto 3.5.2.1). Los departamentos que fueron visitados durante este periodo, así como la frecuencia de dichas visitas, pueden verse en la Tabla 3.

Nombre del Departamento	Código del Departamento	Número de casos	% de casos	Descripción
ANGIOLOGÍA Y CIRUGÍA VASCULAR	ACV	1.328	2,08%	Especialidad médica que estudia las enfermedades del Sistema circulatorio y departamento de cirugía vascular.
ANESTESIOLOGÍA Y REANIMACIÓN	ANR	249	0,39%	Especialidad médica que se centra en la administración de anestesia y la resucitación.
CARDIOLOGÍA	CAR	9.142	14,30%	Rama de la medicina que trata los trastornos y enfermedades cardiacas, así como partes del sistema circulatorio.
CIRUGÍA CARDIACA	CCA	1.063	1,66%	Cirugía cardiaca o de grandes vasos.
CIRUGÍA GENERAL Y DIGESTIVA	CGD	5.721	8,95%	Cirugía general y digestiva.
CIRUGÍA MAXILOFACIAL	CMF	356	0,56%	Cirugía que se especializa en el tratamiento de muchas enfermedades, heridas y defectos en la cabeza, cuello, cara, mandíbula y en los tejidos duros y blandos orales (boca) y en las regiones maxilofaciales (mandíbula y cara).
CIRUGÍA PEDIÁTRICA	CPE	661	1,03%	Subespecialidad de cirugía enfocada a la cirugía de fetos, bebes, niños y adolescentes hasta los 16 años de edad.
CIRUGÍA PLÁSTICA Y REPARADORA	CPL	584	0,91%	Cirugía especializada en corregir trastornos funcionales causados por quemaduras, heridas traumáticas, cáncer, tumores, anormalidades congénitas y anormalidades en el desarrollo, por infecciones o por enfermedades.
CIRUGÍA TORÁCICA	CTO	582	0,91%	Cirugía especializada en el tratamiento quirúrgico de órganos internos del tórax (pecho)—tratamiento genérico de condiciones del corazón (enfermedades cardiacas) y del pulmón (enfermedades pulmonares).
DERMATOLOGÍA	DER	235	0,37%	Rama de la medicina que trata la piel, uñas, pelo y sus enfermedades.
APARATO DIGESTIVO	DIG	2.343	3,66%	Se centra en el tracto gastrointestinal y los órganos accesorios digestivos.
ENDOCRINOLOGÍA	END	347	0,54%	Rama de la medicina y la biología que trata con el sistema endocrino, sus enfermedades y sus secreciones específicas llamadas hormonas.
GINECOLOGÍA	GIN	1.716	2,68%	Especialidad médica que trata la salud de los órganos reproductores femeninos.
HEMATOLOGÍA CLÍNICA	HEM	514	0,80%	Estudio y gestión de todos los desórdenes hematológicos generales.
MEDICINA INTERNA	MIR	12.346	1,31%	Especialidad médica que trata la prevención, diagnóstico y el tratamiento de enfermedades en los adultos.
MEDICINA INTENSIVA	MIV	2.250	3,52%	Rama de la medicina que trata el diagnóstico y la gestión de condiciones peligrosas para la vida que

Nombre del Departamento	Código del Departamento	Número de casos	% de casos	Descripción
				puedan requerir el apoyo de órganos sofisticados y la monitorización invasiva.
NEFROLOGÍA	NEF	1.875	2,93%	Especialidad médica que estudia la anatomía de los riñones y sus funciones.
NEONATOLOGÍA	NEO	77	0,12%	Subespecialidad de la Pediatría que consiste en el cuidado de recién nacidos, especialmente con aquellos que tienen alguna enfermedad o han nacido de forma prematura.
NEUMOLOGÍA	NML	1.834	2,87%	Especialidad médica que trata enfermedades relacionadas con el tracto respiratorio.
NEUROCIRUGÍA	NRC	681	1,07%	Especialidad médica que trata la prevención, diagnóstico, tratamientos quirúrgicos y rehabilitación de aquellos desórdenes que pueden afectar en cualquier medida al sistema nervioso, incluyendo el cerebro, la médula espinal, los nervios periféricos y el sistema cerebrovascular.
NEUROLOGÍA	NRL	1.521	2,38%	Rama de la medicina relacionada con el estudio y tratamiento de desórdenes del sistema nervioso.
OBSTETRICIA	OBS	1.760	2,75%	Especialidad que se centra en el estudio del embarazo, el nacimiento y el periodo post-parto.
OFTALMOLOGÍA	OFT	380	0,59%	Especialidad médica que trata la anatomía, psicología y las enfermedades de la órbita y el globo oculares.
ONCOLOGÍA MÉDICA	ONC	776	1,21%	Rama de la medicina que estudia la prevención, diagnóstico y el tratamiento del cáncer.
ONCOLOGÍA RADIOTERÁPICA	ONR	378	0,59%	Especialidad que utiliza la energía de la radiación para estudiar, tratar y gestionar el cáncer y otras enfermedades.
OTORRINOLARINGOLOGÍA	ORL	761	1,19%	Especialidad médica enfocada en desórdenes del oído, nariz y garganta.
PEDIATRÍA	PED	993	1,55%	Especialidad médica enfocada a los niños.
PSIQUIATRÍA	PSQ	2.841	4,44%	Especialidad médica dedicada al diagnóstico, prevención, estudio y tratamiento de desórdenes mentales.
REHABILITACIÓN	REH	2	0,00%	Rehabilitación física utilizando la fuerza mecánica y el movimiento.
REUMATOLOGÍA	REU	626	0,98%	Subespecialidad médica de la Medicina Interna dedicada al diagnóstico y la terapia de enfermedades reumáticas.
TRAUMATOLOGÍA	TRA	3.370	5,27%	Especialidad que estudia las lesiones y heridas causadas por accidentes o por violencia hacia una persona. También trata la terapia quirúrgica y la reparación de los daños producidos.
UNIDAD DEL DOLOR	UDO	3	0,00%	Trata la terapia de la gestión del dolor de una persona.
URGENCIAS	URG	4.170	6,52%	Especialidad médica que trata el cuidado de pacientes imprevistos de forma indiscriminada con enfermedades o lesiones que requieren una atención médica inmediata.
UROLOGÍA	URO	2.445	3,82%	Especialidad médica que trata enfermedades médicas o quirúrgicas del sistema urinario masculino y femenino, y los órganos reproductores masculinos.
OTROS	-	2	0,00%	Especialidades no conocidas o erróneas.

Tabla 3. Departamentos visitados por los pacientes de 2010 a 2015.

Para la tercera parte de la investigación (sección 5.4) se utilizó el conjunto de datos completo, ya que el algoritmo de redes neuronales utilizado no vio mermada su eficiencia y además el algoritmo no necesita otras adaptaciones, salvo las del propio algoritmo. En este tipo de modelo predictivo, cuantos más datos se disponga más óptimo será el modelo y mejores serán los resultados en las predicciones, ya que se adaptan y aprenden con los datos nuevos.

3.5.2.3 ANÁLISIS Y CONJUNTO DE DATOS FINAL

Algunas de las características de la base de datos CMBD, como la sección de ingreso o de alta del hospital, tienen menos influencia en los resultados de los análisis predictivos sobre los objetivos planteados en esta investigación, debido a los resultados mostrados por el coeficiente de correlación de Pearson o por el algoritmo *SelectKBest* definido en (Pedregosa et al., 2011). Se ha analizado la posible correlación existente entre las principales características del conjunto de datos para comprobar el grado de influencia. El resto de las características no mostradas casi no presentaban relación entre ellas por lo que se desecharon para la primera parte de la investigación, ya que apenas influyen en los resultados. En la Ilustración 3 se puede observar la relación entre estas características mediante una matriz de correlación. Para obtener esta matriz se ha utilizado el método llamado *corr()* de la biblioteca *Pandas DataFrame* (Pandas, 2022, Febrero) del lenguaje de programación *Python*. En concreto el coeficiente calculado y que se muestra en el gráfico es el coeficiente de correlación de Pearson.

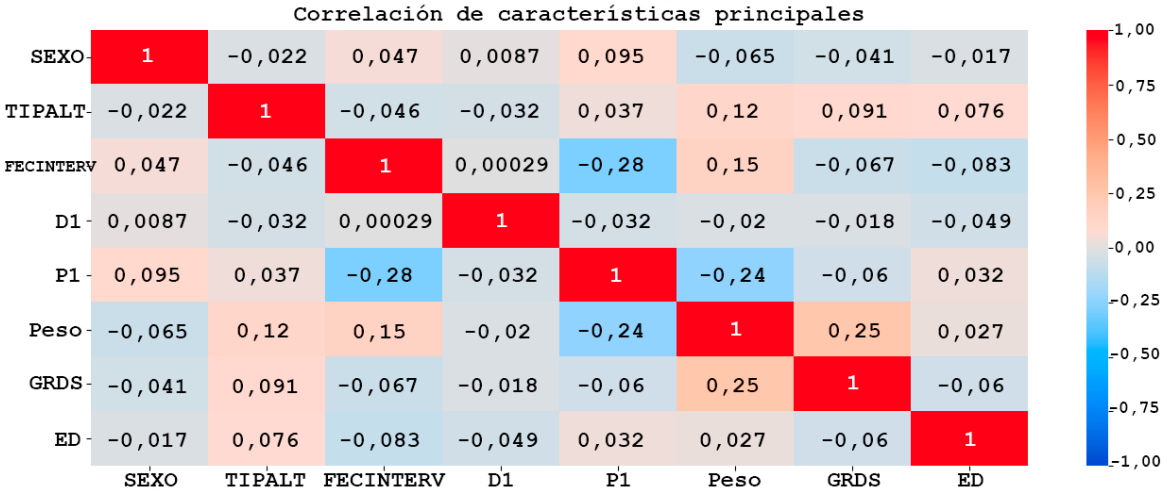


Ilustración 3. Matriz de correlaciones de características principales

Se puede destacar la mayor relación existente entre el grupo relacionado por el diagnóstico (*GRDS*), característica definida por el Ministerio de Sanidad y Consumo (1999), con el peso relativo al coste (*Peso*) y también el *Peso* con la fecha de intervención quirúrgica (*FECINTERV*). También se puede observar que existe relación (correlación negativa) entre el procedimiento principal (*P1*) y la fecha de intervención, y entre el *Peso* y el procedimiento principal.

Para una mejor comprensión del conjunto de características principal, se puede observar su distribución en la Tabla 4.

Variable	Composición	Número de casos	% de población
Género			
	Femenino	32.920	51,64%
	Masculino	30.832	48,36%

Variable	Composición	Número de casos	% de población
Fecha de ingreso			
	Mín.:	1 junio 2010	
	Máx.:	29 septiembre 2015	
Especialidad de ingreso			
	(ver listado de Tabla 3)		
Motivo del alta			
	Alta a casa	61.059	95,78%
	Alta a otro hospital	1.018	1,60%
	Alta voluntaria	194	0,30%
	Alta por fallecimiento	1.481	2,32%
	Alta por transferencia a otro centro social-sanitario	-	0,00%
	Alta por huida	-	0,00%
	Alta por hospitalización en casa	-	0,00%
Cirugía			
	si:	18.487	29,00%
	no:	45.265	71,00%
Diagnóstico principal			
	Insuficiencia cardiaca congestiva	2.349	3,68%
	Aterosclerosis coronaria en arteria coronaria nativa	1.403	2,20%
	Exacerbación aguda	1.312	2,06%
	Neumonía	1.124	1,76%
	Fibrilación auricular	833	1,31%
	Desorden respiratorio	795	1,25%
	Infarto subendocárdico	674	1,06%
	Insuficiencia cardiaca izquierda	648	1,02%
	Dolor precordial	609	0,96%
	Insuficiencia respiratoria aguda	552	0,87%
	Otros (por debajo del 0,85%)	53.453	83,85%
Procedimiento médico principal			
	Cariotipo	5.694	8,93%
	Electrocardiograma	3.534	5,54%
	Radiografía de tórax	2.928	4,59%
	Angioplastia coronaria	2.426	3,81%
	Tomografía axial computarizada de la cabeza	1.653	2,59%
	Ecocardiograma Doppler	1.398	2,19%
	Evaluación de estado mental	1.298	2,04%
	Tomografía cardiaca	1.134	1,78%
	Ultrasonido abdominal	1.045	1,64%
	Inyección de antibióticos	1.026	1,61%
	Escisión de tejido cardiaco	949	1,49%
	Otros (por debajo del 1,36%)	40.667	63,79%
Edad			
	30 años o menos	4.576	7,18%
	31 – 60 años	18.800	29,49%
	Mayor de 60 años	40.376	63,33%
Peso relativo al coste			
	Por debajo del valor medio	15.417	24,18%
	Por encima del valor medio	48.335	75,82%
Grupos relacionados por el diagnóstico			
	Nacimiento con esterilización	2.756	4,32%
	Procedimientos renales	2.413	3,78%
	Dilatación y legrado	2.401	3,77%
	Sistema respiratorio	2.019	3,17%
	Vías aéreas superiores	1.331	2,09%
	Otros procedimientos respiratorios	953	1,49%
	Procedimientos de rehabilitación	907	1,42%

Variable	Composición	Número de casos	% de población
	Otros procedimientos de rehabilitación	873	1,37%
	Otros (por debajo del 1,35%)	50.099	78,58%
Días transcurridos entre la admisión y el alta			
	0 – 7 días	37.425	58,70%
	8 – 14 días	15.342	24,07%
	15 – 21 días	5.555	8,71%
	22 – 28 días	2.355	3,69%
	Más de 28 días	3.075	4,82%

Tabla 4. Distribución de las características principales

Debido al elevado número de diagnósticos y procedimientos utilizados, se muestra la distribución de los principales hasta un porcentaje de uso por debajo del 1,36%. Por el mismo motivo, en el caso de la edad se muestra la distribución en forma de rangos, al igual que en el caso de los días de estancia.

3.5.3 PRE-PROCESADO DE LOS DATOS

El hecho de disponer de los datos no quiere decir que puedan ser utilizados directamente. Se debe proceder a prepararlos para poder aplicar las técnicas seleccionadas en la creación del modelo que se va a aplicar. Esta metodología que se aplica a los datos suele llamarse "*limpieza de datos*" y consiste en asegurar los siguientes objetivos:

- *Los datos deben disponer de la calidad suficiente*. Es decir, que no contengan errores, redundancias o que tengan otros problemas. La calidad del dato se entiende como la propiedad que asegure la calidad del modelo resultante. En el primer caso de estudio de la investigación fue necesario no utilizar aquellos registros en los que el departamento del hospital era desconocido o inexistente, ya que se trataba de ver los resultados en cada departamento del hospital.
- *Los datos disponibles tienen que ser los necesarios para la creación del modelo*. Es posible que tengamos datos que no vayan a ser necesarios por no aportar una mejora en los resultados (correlación nula) o que sea necesario añadir otros. En el primer caso pueden desecharse esos datos y mejorar de esta forma la eficiencia de los algoritmos. En el segundo caso normalmente supone añadir nuevas características a los conjuntos de datos que pueden ser calculados en función de otros o que pueden obtenerse a partir de nuevas relaciones con otros conjuntos de datos o con otras bases de datos. En la primera parte de la investigación el objetivo que se buscaba era predecir el número de días que un paciente permanecería ingresado. Como no se disponía de esta característica, pero se podía calcular a partir de otras, se obtuvo creando una nueva característica (ver Tabla 4 – *días transcurridos entre la admisión y el alta*) que se obtiene a partir de una función propia que calcula el número de días transcurridos entre la fecha de ingreso y el alta hospitalaria del paciente.

- *Los datos se deben encontrar con el formato adecuado*. Muchos de los métodos que existen exigen que los datos se encuentren en un formato determinado, el cual no tiene por qué ser el que se dispone en el conjunto de datos. Si existen muchas diferencias entre los formatos disponibles, lo más habitual será tener que efectuar varias transformaciones hasta obtener el formato adecuado. Lo más habitual es disponer de datos continuos y que el método utilizado sólo admita valores discretos. En esta investigación ha sido necesario realizar varias transformaciones para poder utilizar varios métodos de aprendizaje automático:
    - Fue necesario realizar la transformación de datos en formato texto a datos en formato numérico por exigencia del algoritmo. Aquellos algoritmos, como las redes neuronales, que utilizan pesos relativos o cálculo de distancias entre los distintos registros, precisan de un valor numérico que les permita calcular la diferencia entre ellos. La técnica utilizada para tratar los datos categóricos (formato texto) se denomina *one-hot encoding* y convierte todos los valores a 0 o 1, de forma que se crean tantas columnas como valores categóricos distintos existan. Se puede observar un ejemplo de uso de esta técnica en el artículo de Zhibin et al. (2020), donde los autores convierten las variables categóricas en numéricas para utilizarlas en una red neuronal. A modo de ejemplo, tal y como puede observarse en la Tabla 5, si tuviésemos la característica *sexo* definida con las letras H (Hombre) y M (Mujer), al aplicar la técnica se crearían 2 nuevas columnas, *sexo_H* y *sexo_M*, identificando a un hombre con el valor (1, 0) para las columnas *sexo_H* y *sexo_M*, respectivamente, y el valor (0, 1) para identificar a una mujer. Para realizar este proceso se utilizó el procedimiento *get_dummies* de la biblioteca libre *Pandas* del lenguaje Python
    - También fue necesaria la transformación de los formatos numéricos mediante un proceso de normalización que ajustase los valores a un rango predeterminado, para así evitar efectos de sobreajuste en el modelo y mejorar los resultados. El modelo utilizado en el segundo caso de la investigación daba un efecto de sobreajuste debido a que la capacidad del modelo era demasiado elevada con relación a la complejidad del conjunto de datos. Esto provocaba un error elevado en el entrenamiento generando un desajuste entre las fases de entrenamiento y test, haciendo el modelo inestable. Este proceso se utilizó en la segunda parte de la investigación en el uso de las redes neuronales. Para realizar la normalización de los datos se utilizó la clase *MinMaxScaler* dentro del procedimiento *sklearn.preprocessing* de la biblioteca *scikit-learn* versión 0.24.2.

Tabla original

Nº Registro	sexo
1	H
2	M
3	M
…	…

Tabla tras aplicar técnica one-hot encoding

Nº Registro	sexo_H	sexo_M
1	1	0
2	0	1
3	0	1
…	…	…

Tabla 5. Ejemplo de uso de técnica one-hot encoding

3.5.4 DATOS DE ENTRENAMIENTO Y DATOS DE PRUEBA

Cuando se aplican técnicas de aprendizaje automático es habitual separar el conjunto de datos disponible en 2 subconjuntos de datos con unas determinadas proporciones. De esta forma se utiliza el primer subconjunto de datos (el mayor de ellos) para entrenar el modelo y, el segundo subconjunto, para probar la eficiencia del modelo. Las proporciones más habituales suelen ser de 80% para entrenamiento y 20% para pruebas, como utilizan en (Wu, Pang y Kwong, 2015), donde utilizan las mismas proporciones para predecir la presión sistólica sanguínea, o también 70% / 30%, como realizan Aghajani y Kargari (2016), quienes utilizan estas proporciones como las más adecuadas para realizar predicciones sobre el número de días de estancia en un hospital dentro del departamento de cirugía. Esta separación es importante para que la prueba del algoritmo utilice una muestra de datos totalmente desconocida y que de esta forma no obtenga resultados demasiado optimistas por ya conocer previamente los resultados.

A lo largo de esta investigación se utilizaron varias técnicas de selección de datos para ambas fases de prueba y entrenamiento. Además de algoritmos propios de selección de los datos, se utilizaron procedimientos de bibliotecas de Python ya existentes para estas funciones.

En una de las principales técnicas de selección se seleccionaba el conjunto de datos de entrenamiento y pruebas con una proporción de 70% / 30%, mediante una función en la que se le pasaba estos datos en forma de parámetros, dando a la salida dos conjuntos de datos con estas proporciones. Otro importante parámetro que se le pasaba era la petición de que la selección de las muestras individuales se hiciera de forma aleatoria, así como un parámetro en forma de valor entero para controlar la variabilidad de la aleatoriedad. La biblioteca concreta utilizada para esta técnica fue la clase *train_test_split* dentro del procedimiento *sklearn.model_selection* de la biblioteca *scikit-learn*.

La segunda técnica principal utilizada fue mediante la validación cruzada, llamada *5-fold cross-validation*. Esta técnica divide el conjunto de datos en 5 partes, una de las cuales se reserva para la realización de las pruebas. Esto provoca que el porcentaje sea de un 80% / 20%. El uso de esta técnica y, por lo tanto, el mismo porcentaje para entrenamiento y pruebas, es utilizado también por otros autores, tal y como hacen Howell, Coory, Martin y Duckett (2009) en su investigación para identificar pacientes en riesgo de necesitar readmisión hospitalaria. Debido a que las muestras se encontraban

desbalanceadas, esta técnica se utilizó con estratificación para mantener equilibradas las clases, ya que en caso contrario el modelo no podría aprender a generalizar, produciéndose un efecto de infrajuste (*underfitting* en inglés) en el que solo puede reconocer una clase. Para realizar la validación cruzada de 5 partes con estratificación mediante la selección aleatoria de muestras se utilizó la clase *StratifiedKFold* dentro del procedimiento *sklearn.model_selection* de la biblioteca *scikit-learn*.

4 ESTADO DEL ARTE

4.1 INTRODUCCIÓN

Aunque el aprendizaje automático es una disciplina reciente, sus bases y sus algoritmos matemáticos de los que parte no lo son tanto. Algunos teoremas y teorías que sentaron las bases de esta tecnología fue, por ejemplo, el teorema de Bayes (1812), que definió la probabilidad de que un evento ocurriese basándose en el conocimiento de las condiciones previas que pudieran estar relacionadas con dicho evento, o el trabajo de científicos como el matemático Alan Turing (1950), que planteó por primera vez la pregunta de si era posible que las máquinas pudieran pensar, dando lugar al inicio de la creación de computadoras de "inteligencia artificial", capaces de replicar de forma autónoma tareas típicamente humanas tales como la escritura o el reconocimiento de imágenes.

En general, el aprendizaje automático evolucionó desde estudios de reconocimiento de patrones y de teorías de aprendizaje computacional en inteligencia artificial, como el estudio realizado por Samuel (1959) sobre el juego de las damas aplicando procedimientos de aprendizaje automático. Además, está fuertemente relacionado con la estadística computacional y se encuentra enfocado en realizar predicciones mediante ordenadores. En ocasiones, se combina con la Minería de Datos para la realización de análisis exploratorio de los datos. Cuando se combina con la analítica de datos, permite crear modelos y algoritmos que obtienen predicciones por ellos mismos, lo que se conoce como analítica predictiva.

Analizando el ámbito de estudio de esta investigación, existen multitud de avances en el campo de la salud como la robotización, el láser, la impresión 3D o los sistemas de realidad aumentada que facilitan la labor de los médicos y del personal sanitario. Pero uno de los mayores avances se encuentra en la inteligencia artificial y el aprendizaje automático, donde se está consiguiendo mejorar el diagnóstico precoz de enfermedades, la asistencia primaria y la administración eficaz de los recursos sanitarios. Como se afirma en (Arwinder y Ashima, 2019), se trata de una de las áreas de investigación más prometedoras, donde se consiguen predicciones más precisas. Los investigadores en esta área se centran en utilizar los datos para realizar predicciones, tal y como realizan en (Ajay, Rama y Arvind, 2019) donde se centran en datos clínicos y datos sobre el genoma humano para analizarlos con algoritmos de aprendizaje automático.

Dependiendo del tipo de aprendizaje que se quiera emplear, se utilizarán diferentes tipos de metodologías. Esta investigación se ha basado en el tipo de aprendizaje supervisado, donde se parte de unos datos previos en los que se tienen disponibles los resultados anteriores y se conoce la respuesta a la cuestión que se quiere resolver. Con esta información en forma de registros de datos, se trata de aprender a partir de los datos conocidos previamente para saber cuál va a ser su comportamiento futuro, prediciendo su respuesta ante datos de entrada de los que no se sabe la respuesta. Por ejemplo, tal y como se ha estudiado en esta investigación, se ha buscado predecir

cuántos días permanecerá ingresado un paciente en un hospital en función de los datos genéricos del paciente y de los datos de los departamentos en los que ingresa o de lo que se le diagnostica. Previamente se disponía de estos mismos datos conociendo los días que permaneció el paciente en el hospital, tal y como realizan en (Hachesu, Ahmadi, Alizadeh y Sadoughi, 2013), donde se predicen los días de estancia que permanecerán ingresados los pacientes con problemas cardiacos. Otro ejemplo también estudiado ha sido predecir si un paciente va a reingresar o no al hospital debido a alguna dolencia relacionada con la que ingresó en primera instancia, partiendo también de los mismos datos genéricos y de los diagnósticos realizados a dicho paciente, tal y como también se realiza en (Peiró, Librero y Martínez, 1996), donde se estudia el reingreso hospitalario según varios márgenes de días para considerarlo reingreso en pacientes con enfermedades digestivas y hepatobiliares.

Los algoritmos de aprendizaje automático que se utilizan también dependen del tipo de aprendizaje utilizado y en función de las características que dispongan los datos. Es habitual utilizar varios tipos para analizar los que mejores resultados se obtienen para cada caso analizado, como realizan Morton, Marzban, Giannoulis, Aparasu y Kakadiaris (2014), donde se realiza una comparación de diferentes algoritmos de aprendizaje supervisado para predecir los días de estancia en un hospital con aquellos pacientes que son diabéticos. En este caso se utilizan algoritmos como *Random Forest*, *Support Vector Machines* o *Linear Regression*. Otros estudios, como el realizado por Aghajani y Kargari (2016), utilizan *Support Vector Classifiers*, *K Nearest Neighbors* y *Decision Trees* para predecir los días de estancia en pacientes del departamento de cirugía general, o Bergese et al. (2019), en el que utilizan Redes Neuronales y Árboles de decisión para predecir la readmisión hospitalaria en el departamento de emergencias de Pediatría. De la misma forma que realizan estos y otros autores, todas estas técnicas también han sido utilizadas en las investigaciones aquí realizadas. Al igual que se realiza en la última parte llevada a cabo en esta investigación (sección 5.4), durante éstos últimos años se aplican las últimas tecnologías relacionadas con el machine learning, más concretamente el *Deep learning*, tal y como utilizan en el artículo (Zhibin, Ding, Wang y Zou, 2020), donde se aplica una red neuronal convolucional basadas en ciclos temporales para identificar las cadenas de ADN en el genoma del arroz.

Por norma general, los algoritmos que mejores resultados suelen obtener son aquellos basados en estructuras en forma de árbol en los que se van tomando diferentes caminos en función de resultados o preguntas planteadas, como son los algoritmos *Random Forest*, que está basado en múltiples árboles de decisión y el propio algoritmo *Decision Tree*. Esta tendencia, que también dependerá de los datos, ha sido confirmada por los resultados obtenidos en esta investigación y por otros estudios como los realizados por Chuang, Hu, Tsai, Lo y Lin (2015) o LaFaro et al. (2015). Las redes neuronales también obtienen buenos resultados cuando los datos lo permiten y realizando una configuración adecuada. Además, permiten reutilizar el modelo de red generado de forma que aprenden con las nuevas muestras que les van llegando, mejorando así su comportamiento. Como se puede ver en el análisis realizado por Futoma, Morris y Lucas (2015), los mejores resultados obtenidos por la comparativa entre diferentes estudios para predecir la readmisión hospitalaria fue con los algoritmos Random Forest y las redes neuronales. En un trabajo reciente (Zhibin et al., 2020) también se utiliza un

modelo de red neuronal como único algoritmo en el que se obtienen muy buenos resultados, donde también sugieren reutilizar su modelo de forma que se optimice en futuras investigaciones.

El estado del arte nos ha permitido observar los resultados y los campos estudiados para comprobar la posibilidad de mejorar o innovar dentro de las áreas de estudio seleccionadas. Para realizar los casos de estudio de esta investigación, se ha realizado una revisión sistemática de la literatura cuya metodología y alcance se describen con detalle en las siguientes secciones.

4.2 MÉTODO DE REVISIÓN

Como se comenta en (Manterola et al., 2013), una revisión sistemática se sigue con el objetivo de resumir la información existente respecto de la temática analizada, revisando tanto sus aspectos cuantitativos como cualitativos.

Como parte de la metodología seguida en la revisión sistemática se han utilizado una serie de criterios de inclusión o exclusión de los artículos analizados. Se han excluido todos aquellos artículos que no utilizasen las técnicas de aprendizaje automático o no predijesen la temática elegida según el área de estudio seleccionada. Los criterios de exclusión concatenados utilizados se han realizado por fases:

- Exclusión basada en el título, donde la especificación de un tema totalmente ajeno al estudiado era descartada, como por ejemplo ocurría al buscar sobre las temáticas "*Días de estancia*" y "a*prendizaje automático*". Aunque dentro del propio artículo se pueda comentar algo sobre estas temáticas, el propio título ya indica que el artículo trata sobre un tema totalmente distinto, como es el caso del artículo cuyo título es "*A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis*". En este artículo se puede observar que, aunque utiliza técnicas de aprendizaje automático, predice el índice de mortalidad y no los días de estancia hospitalaria.
- Exclusión basada en el resumen, donde al explicar los objetivos cumplidos o la metodología seguida permitía descartar dichos artículos. En este caso, la lectura del tema es insuficiente para verificar el descarte del artículo, como ocurrió al buscar sobre las temáticas "*Readmisión hospitalaria*", "*Digestivo*" y "*aprendizaje automático*". Se encontró el artículo "*Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records*", para el cual se necesitó una lectura más en profundidad sobre el resumen de forma que se puede comprobar que, aunque se utilizan múltiples bases de datos que pueden incluir patologías de digestivo, también se utilizan técnicas de aprendizaje automático, pero para ver la relación existente entre características demográficas.
- Exclusión realizada por la revisión del contenido, donde la lectura en profundidad del artículo permite corroborar el descarte del mismo. En ocasiones, la lectura del resumen no permite descartar totalmente el artículo y es necesaria una lectura más en profundidad del artículo completo.

El criterio de inclusión de los artículos analizados se basa en el tratamiento de la temática elegida tanto en metodologías a estudiar como en objetivos elegidos, y son comentados en el siguiente punto sobre el criterio de búsqueda.

4.3 CRITERIO DE BÚSQUEDA

Como criterio de búsqueda se ha utilizado la cadena de texto asociada a cada área de estudio, desarrollada para poder encontrar los artículos relacionados con la temática elegida y con las técnicas de aprendizaje automático seleccionadas en cada caso particular. Además, los artículos encontrados se ordenan por orden de relevancia, de mayor a menor.

En el primer caso de estudio realizado en esta investigación sobre los días de permanencia en un hospital, como criterio de búsqueda se utiliza la cadena de texto "*Hospital length of stay machine learning*", de forma que permita encontrar todos los artículos relacionados con la temática del caso de estudio y que aplique las tecnologías de aprendizaje automático que se quieren abordar. Se obtienen 148.000 referencias con los criterios aplicados en este caso particular. Al estar ordenados por relevancia, se analizan los primeros 510 artículos más citados y se obtienen 35 artículos científicos, conferencias, reportes médicos e incluso borradores que tratan la misma temática. Se decide finalizar el análisis tras comprobar que en las últimas 130 referencias de los 510 artículos analizados no se encuentra nada relacionado con la temática del primer caso de uso y, al encontrarse ordenados por relevancia, muchas de las referencias restantes son análisis o estudios pendientes de publicación. En la búsqueda realizada para este caso, aunque se trata de una investigación sobre algoritmos de clasificación, se encuentran artículos sobre clasificación y regresión, pero no son descartados los de regresión ya que son usados como referencias de uso de la propia temática sobre los días de estancia. Finalmente, se considera relevante referenciar 25 de ellos al elaborar la exposición final.

En el segundo caso de estudio sobre readmisión hospitalaria, como criterio de búsqueda se utiliza la cadena de texto "*Hospital readmission digestive machine learning*". En este caso, al tratar una temática mucho más específica, se encuentra un número bastante menor de artículos. Se localizan 993 referencias con los criterios de búsqueda aplicados, de los cuales 16 artículos tratan sobre la temática elegida aplicando técnicas de aprendizaje supervisado.

Con los criterios elegidos se encuentran y estudian todos los análisis realizados para comprobar que no se inicia una investigación previamente realizada. En caso de existir el mismo estudio, se puede sopesar la posibilidad de ampliar el estudio previo, mejorarlo o rebatirlo con nuevos resultados.

4.4 RESULTADOS DEL ANÁLISIS

Cada caso de estudio realizado en esta investigación ha supuesto el análisis de diferentes tecnologías, las cuales en ocasiones han coincidido en varios casos de estudio. Cada técnica aplicada a cada caso de estudio particular se ha analizado realizando un estudio de investigaciones similares que han arrojado diferentes resultados para cada tecnología. Estos resultados se han recogido para realizar una valoración global de los resultados por cada algoritmo aplicado. En general, la mayor parte de las referencias encontradas se localizan durante los últimos años al tratarse de tecnologías recientes y en continua evolución, sobre todo en el segundo caso de estudio que, además, se ha realizado más recientemente. La Ilustración 4 permite ver la distribución por año de los artículos referenciados en los análisis realizados.

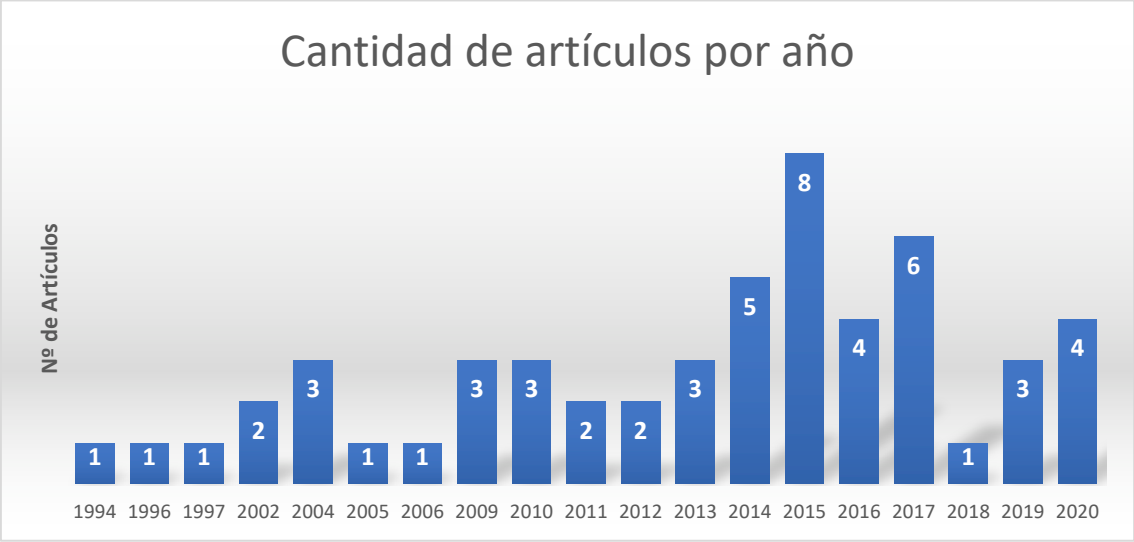


Ilustración 4. Cronología de artículos que han sido referenciados

En la Ilustración 5 se puede observar la precisión media obtenida por los algoritmos de aprendizaje supervisado con métodos de Clasificación. Estos resultados se han obtenido de todos los artículos referenciados que así lo han reflejado en sus análisis. Al tratar esta investigación sobre métodos clasificatorios, se muestra la precisión de estos métodos concretos.

Esta gráfica (Ilustración 5), permite conocer la media de precisión que obtiene cada algoritmo concreto y muestra que algoritmos obtienen mejores resultados. Estos resultados tan solo deben considerarse orientativos, ya que depende también en gran medida de los conjuntos de datos que se dispongan para cada caso concreto.

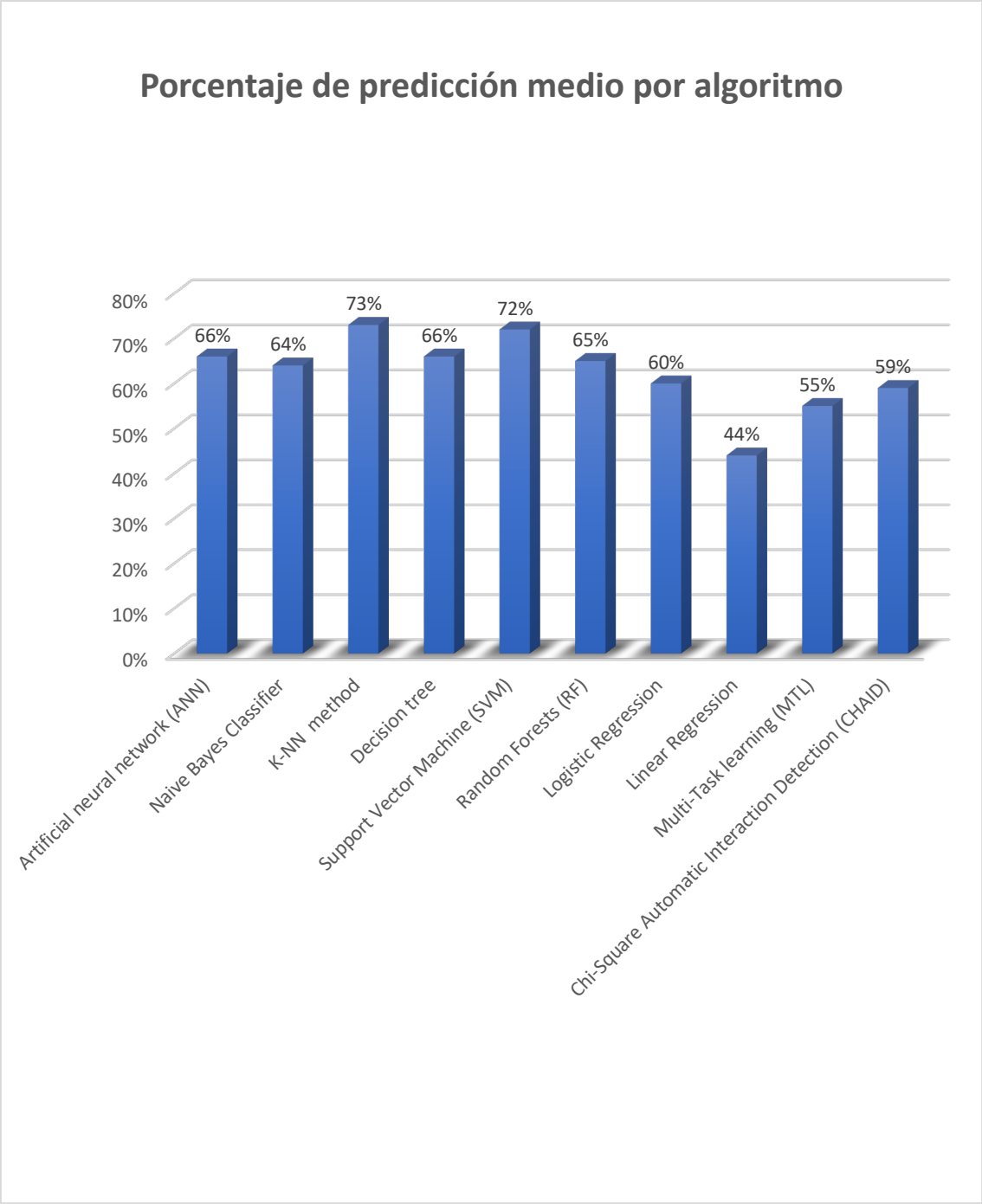


Ilustración 5. Precisión media obtenida por los distintos métodos de aprendizaje automático en los artículos referenciados

Los resultados reflejados en la Ilustración 5 de precisiones, los algoritmos que mejores resultados obtienen son *KNN* y *SVM*. En el caso del primero tan solo se han encontrado resultados en 2 artículos, por lo que el resultado es relativo y en el caso del segundo algoritmo, se trata de un conjunto de métodos diferentes en los que se usan diferentes núcleos (*kernels*) con distintas funciones matemáticas, por lo que se encuentra muy orientado a casos concretos y por eso refleja mejores resultados.

En general se puede observar que, en los algoritmos restantes, los que mejores resultados obtienen son las redes neuronales, los árboles de decisión y *Random Forest*, al igual que ha ocurrido en esta investigación tal y como veremos más adelante (sección 5.3). También se debe tener en cuenta que se debe optimizar de manera adecuada los algoritmos y que existen casos en los que los datos disponibles no permiten obtener buenos resultados o no se perseguía obtener buenos resultados sino analizar el algoritmo desde otro punto de vista, como ocurre en (Jamei, Nisnevich, Wetchler, Sudat y Liu, 2017), donde buscan demostrar que las redes neuronales son mejores para conjuntos de datos complejos con respecto a otros algoritmos independientemente de la precisión, obteniendo tan solo una precisión del 24%. Aunque en este artículo confirman una mejora del 4% con respecto a la tecnología LACE con el mismo tipo de datos, supone una reducción aproximada del 330% con respecto a otros resultados obtenidos para este tipo de algoritmo sobre readmisiones.

4.5 FUENTES DE INFORMACIÓN

En este tipo de estudios se necesita realizar una búsqueda amplia sobre todas las fuentes de información más relevantes. Como recomiendan Rattan, Bhatia y Singh (2013), es aconsejable buscar información en las fuentes más conocidas y extendidas. Aunque en esta investigación se ha realizado una búsqueda exhaustiva en todas las fuentes que se han encontrado, podemos confirmar que también se han utilizado dichas fuentes tras encontrar nuestras fuentes de información dentro de parte del listado referenciado en dicho artículo. Se dispone del siguiente listado de las principales fuentes de información utilizadas en esta investigación, ordenadas de mayor a menor número de referencias:

- IEEE eXplore – 15% (https://ieeexplore.ieee.org/Xplore/home.jsp)
- Elsevier – 13% (www.sciencedirect.com)
- Springer – 11% (www.springerlink.com)
- Oxford Academic – 6% (https://global.oup.com/academic/?cc=es&lang=en&)
- PLOS ONE – 4% (https://journals.plos.org/plosone/)
- Journal of the American Medical Association – 4% (https://jamanetwork.com/)

El resto de fuentes de información puede encontrarse en la gráfica de medios de publicación que se muestra a continuación en la Ilustración 6.

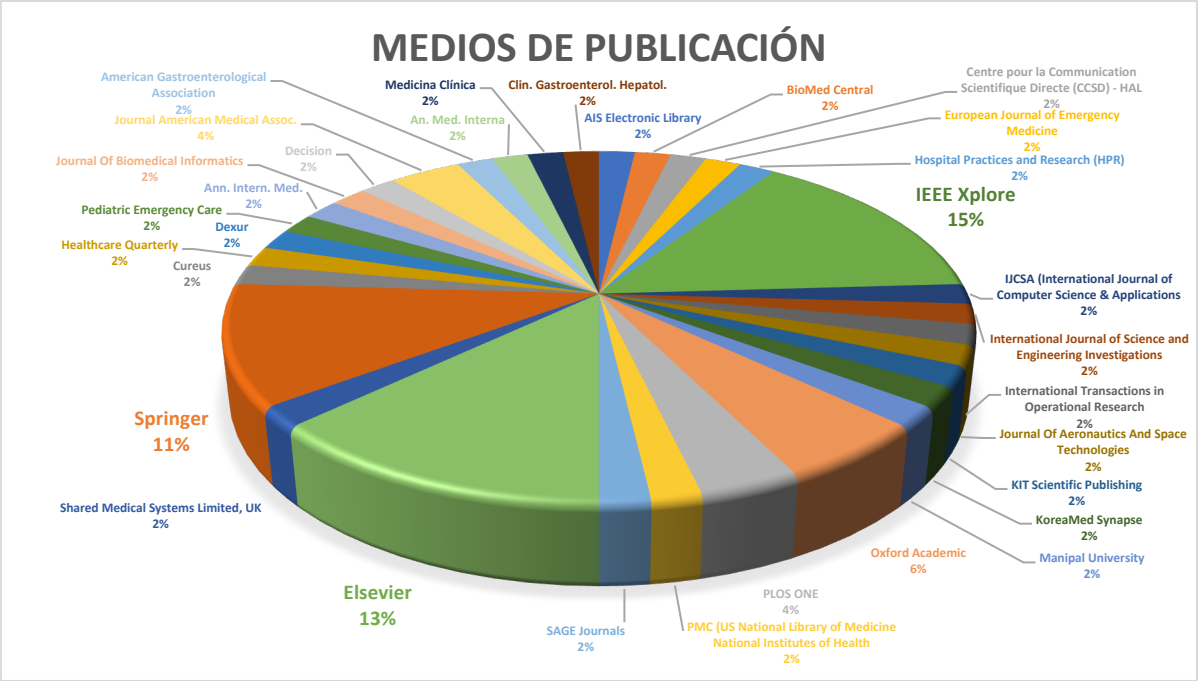


Ilustración 6. Análisis del uso de los medios de publicación utilizados

Los artículos localizados en las búsquedas como consecuencia de las investigaciones sobre los casos de uso que han sido utilizados para el análisis de esta sección y de la sección 4.4, se pueden localizar en la última parte del ANEXO.

4.6 DISCUSIÓN SOBRE EL ESTADO DEL ARTE

El estado del arte es un proceso que debe realizarse para estudiar la situación del área de estudio que quiere implementarse y supone un consumo de tiempo que conviene planificar. Este proceso es aconsejable realizarlo previamente al inicio de la propia investigación, ya que puede dar no solo una idea de si ya ha sido analizado anteriormente, sino de las diferentes técnicas que han utilizado otros investigadores y en qué forma. Se trataría de un conocimiento ya existente al que ya no sería necesario dedicar esfuerzo, porque ya se encuentra implementado. Al igual que ha ocurrido en el segundo caso de estudio aquí realizado (sección 5.3), también puede aprovecharse la investigación realizada por otros investigadores para iniciar una nueva área de estudio a partir del trabajo ya realizado. En el estudio de la readmisión hospitalaria se utilizaron las distintas patologías de digestivo con sus correspondientes codificaciones elaboradas por expertos médicos que realizaron en el artículo (Peiró et al., 1996). Gracias a este listado fue posible conocer si se trataba de una readmisión derivada de este tipo de patologías.

Este proceso también puede ayudar a conocer si se está realizando la investigación de manera adecuada o no. Se puede ver, por el trabajo realizado por otros autores, si los resultados son acordes a otros trabajos realizados anteriormente, ya que puede ocurrir que se estén obteniendo resultados excelentes por algún error en algún cálculo o en la forma de utilizar alguna técnica, en vez de por haber realizado un gran descubrimiento.

Gracias al análisis realizado se puede corroborar que los resultados obtenidos son correctos.

En general se trata de una fase que conviene llevarse a cabo en todo proceso de investigación y que aporta numerosas ventajas. Permite evitar, entre otras ventajas, pérdidas de tiempo por iniciar estudios ya realizados, por llegar a conclusiones erróneas o por dedicar esfuerzos innecesarios a investigar o desarrollar técnicas que ya han sido implementadas y probadas con anterioridad. También permite corroborar y afianzar los resultados que se obtienen contrastándolos con otros estudios.

5 CASOS DE ESTUDIO

5.1 INTRODUCCIÓN

En esta sección se van a analizar los tres casos de estudio realizados en esta investigación como consecuencia de las carencias detectadas en la revisión del estado del arte realizado (ver sección 4), partiendo del conjunto de datos de salud disponible para la investigación. Estos casos de estudio comprenden las temáticas sobre el análisis del conjunto de datos de salud, un estudio sobre los días que deben permanecer ingresados los pacientes de un hospital a nivel hospitalario global y una investigación sobre las readmisiones de los pacientes que se producen tras haber sido dados de alta del centro hospitalario.

El primer caso de estudio abordado es necesario no solo en esta investigación, sino en todas aquellas en las que se trabaje con datos de cualquier área de conocimiento. Se necesita conocer qué representan los datos para saber cómo trabajar con ellos, cómo combinarlos con otros datos para extraer nuevo conocimiento o para tener la capacidad de discernir si un dato se trata de un error tipográfico o se trata de una anomalía aun siendo un dato correcto. El no abordar este tipo de estudio puede suponer obtener conclusiones erróneas o sin ningún sentido. En este caso de estudio se analiza qué temática concreta de salud se está tratando dentro del conjunto de datos disponible. Se trata de un conjunto de datos que contiene la actividad hospitalaria en el que se analiza cómo se representan y su significado. Se estudia cómo se distribuye la información según los diagnósticos realizados, según la edad y el sexo de los pacientes, los motivos del alta hospitalaria y los días de permanencia según la edad y el sexo.

El siguiente caso de estudio trata sobre conocer de antemano el número de días que van a permanecer ingresados los pacientes en un departamento concreto del hospital. Se predicen los días según varios algoritmos predictivos para obtener los mejores resultados, se ofrece un listado de los departamentos más aconsejables para realizar dichas predicciones, ordenados de más a menos aconsejable, y se ofrece los mejores resultados descubiertos para patologías y departamentos concretos del hospital estudiado. Este caso de estudio, además de por su importancia en la optimización de los recursos y en la mejora de la calidad de atención al paciente, cubre la falta de un análisis de los días de estancia de los pacientes a nivel hospitalario, aportando una visión que facilita la gestión global de los recursos y que mejora aún más la calidad del paciente. Las investigaciones realizadas previamente estudiaban los casos para patologías particulares o departamentos concretos, existiendo incluso departamentos sin disponer de este tipo de estudio.

El último caso de uso estudia la predicción de las tasas de readmisión de un hospital, pasando por la detección de la propia readmisión, la cual debe estar relacionada con el motivo que causó el ingreso y dentro de un rango concreto de días, llegando hasta la predicción de las readmisiones que se van a producir en un futuro. Este caso es elegido por su importancia como índice para medir la calidad de atención al paciente y por ser un estudio muy extendido, ya que permite aumentar la calidad al paciente y reducir las

posibles penalizaciones gubernamentales. También es elegido por haberse detectado que no se había implementado una de las últimas tecnologías de aprendizaje profundo y por no haber sido tratado como una secuencia de sucesos temporal, a pesar de que para detectar una readmisión se debe tener en cuenta la línea temporal.

5.2 ANÁLISIS GLOBAL DEL CONJUNTO DE DATOS

Como se ha comentado previamente (sección 3.5.2.1), se ha utilizado una parte de la base de datos que se utiliza a nivel de la Comunidad de Madrid y a nivel Nacional, llamada CMBD. En concreto es la parte que fue facilitada por un hospital de la Comunidad de Madrid, cuyas características, tipos y descripciones pueden verse en la tabla 2. Se trata de una fuente de datos homologada fundamental para el análisis de la actividad hospitalaria y para el conocimiento del proceso asistencial. El pilar fundamental es la codificación de la información clínica relativa al paciente y su proceso de atención. Dicha codificación es realizada por documentalistas clínicos y técnicos especializados.

La parte codificada que corresponde a los datos clínicos recogidos en la parte del CMBD que se dispone (diagnósticos, procedimientos, causas externas), se puede encontrar en el Ministerio de Sanidad, Servicios Sociales e Igualdad (2014), donde es definida como la clasificación de referencia utilizada para la codificación de los diferentes datos clínicos, también llamado CIE9MC. Este manual de referencia contiene un índice alfabético de enfermedades y otro de procedimientos, con sus correspondientes codificaciones, así como listas tabulares. Este manual habitualmente se utiliza para codificar enfermedades y procedimientos aplicados a pacientes por especialistas médicos, los cuales deben conocer la terminología médica y entender las características, terminología y convenciones. En cambio, en esta investigación se ha utilizado en sentido inverso, es decir, conociendo su codificación se acude al manual (CIE9MC) para identificar a qué procedimiento o diagnóstico concreto corresponde. De esta forma se utiliza para consulta y no se hace necesario el conocimiento de un especialista médico. En la Tabla 6 podemos observar los 25 diagnósticos más frecuentes observados en el conjunto de datos utilizado, dentro del rango de fechas disponible. En la tabla se ve la codificación aplicada en el CMBD con la descripción correspondiente según el CIE9MC.

Código de Diagnóstico	Nº de Casos	Porcentaje	Definición
414.01	1.085	11,87%	Otras formas de cardiopatía isquémica crónica (414) – Aterosclerosis coronaria (414.0) de arteria coronaria nativa (414.01)
427.31	545	5,96%	Disritmias cardiacas (427) – Fabricación y flutter auricular (427.3) – Fibrilación auricular (427.31)
410.71	450	4,92%	Infarto agudo de miocardio (410) – Infarto subendocárdico (410.7) – Infarto no transmural
427.32	418	4,57%	Disritmias cardiacas (427) – Fibrilación y flutter auricular (427.3) – Flutter auricular (427.32)
427.89	399	4,36%	Disritmias cardiacas (427) – Otras disritmias cardiacas especificadas (427.8) – Otras (427.89)

Código de Diagnóstico	Nº de Casos	Porcentaje	Definición
786.51	394	4,31%	Síntomas que implican al aparato respiratorio y otros síntomas torácicos (786) – Dolor torácico (786.5) – Dolor Precordial (786.51)
428.00	370	4,05%	Insuficiencia cardiaca (428) – Insuficiencia cardiaca congestiva, no especificada (428.0)
413.90	341	3,73%	Angina de pecho (413) – Otra angina de pecho y angina de pecho no especificada (413.9)
V53.31	332	3,63%	Colocación y ajuste de otro dispositivo (V53) – Dispositivo cardiaco (V53.3) – Marcapasos cardiaco (V53.31)
410.11	309	3,38%	Infarto agudo de miocardio (410) – De otra pared anterior (410.1) – anteroapical con porción contigua del tabique interventricular (410.11)
411.10	274	3,00%	Otras formas agudas y subagudas de cardiopatía isquémica (411) – Síndrome coronario intermedio (411.1)
410.41	234	2,56%	Infarto agudo de miocardio (410) – De otra pared inferior (410.4) – pared diafragmática (con porción contigua del tabique interventricular) (410.41)
426.00	206	2,25%	Trastornos de conducción (426) – Bloqueo auriculoventricular completo (426.0)
780.20	205	2,24%	Síntomas generales (780) – Síncope y colapso (780.2)
410.31	182	1,99%	Infarto agudo de miocardio (410) – De la pared inferoposterior (410.3) – (410.31)
425.40	164	1,79%	Miocardiopatía (425) – Otras miocardiopatías primarias (425.4)
416.00	162	1,77%	Enfermedad cardiopulmonar crónica (416) – Hipertensión pulmonar primaria (416.0)
996.72	146	1,60%	Complicaciones propias de ciertos procedimientos especificados (996) – Otras complicaciones de dispositivo profético, implante e injerto internos (biológico) (sintético) (996.7) – Por otro dispositivo, implante e injerto cardiaco (996.72)
428.10	120	1,31%	Insuficiencia cardiaca (428) – Insuficiencia cardiaca izquierda (428.1)
424.10	115	1,26%	Otras enfermedades de endocardio (424) – Trastornos de la válvula aórtica (424.1)
416.80	110	1,20%	Enfermedad cardiopulmonar crónica (416) – Otras enfermedades cardiopulmonares crónicas (416.8)
410.81	101	1,10%	Infarto agudo de miocardio (410) – De otros sitios especificados (410.8) – De músculo papilar (410.81)
V53.32	98	1,07%	Colocación y ajuste de otro dispositivo (V53) – Dispositivo cardiaco (V53.3) – Desfibrilador cardíaco automático implantado (V53.32)
410.01	89	0,97%	Infarto agudo de miocardio (410) – De la pared anterolateral (410.0) – (410.01)
428.90	84	0,92%	Insuficiencia cardiaca (428) – Fallo cardiaco no especificado (428.9)

Tabla 6. Diagnósticos más comunes del conjunto de datos de estudio según codificación CIE9MC.

La parte del CMBD utilizada es una parte que contiene datos genéricos sobre los pacientes, de forma que se puede acceder a ella con facilidad en caso de querer replicar los resultados o utilizarlos para aplicar los algoritmos que se tratan en esta investigación. También se puede utilizar para abrir otras líneas de investigación analizando la distribución de los datos disponibles. En la Ilustración 7 se puede observar cómo se distribuyen los datos en función de alguna de las variables más comunes de una forma gráfica. Se puede apreciar cómo los hombres tienen más ingresos a edades más tempranas que las mujeres, teniendo éstas una curva de ingresos más lineal, o cómo en la gráfica de la parte de abajo, en *Evolución del número de pacientes*, cómo existió un aumento de los ingresos durante los años 2013 y 2014.

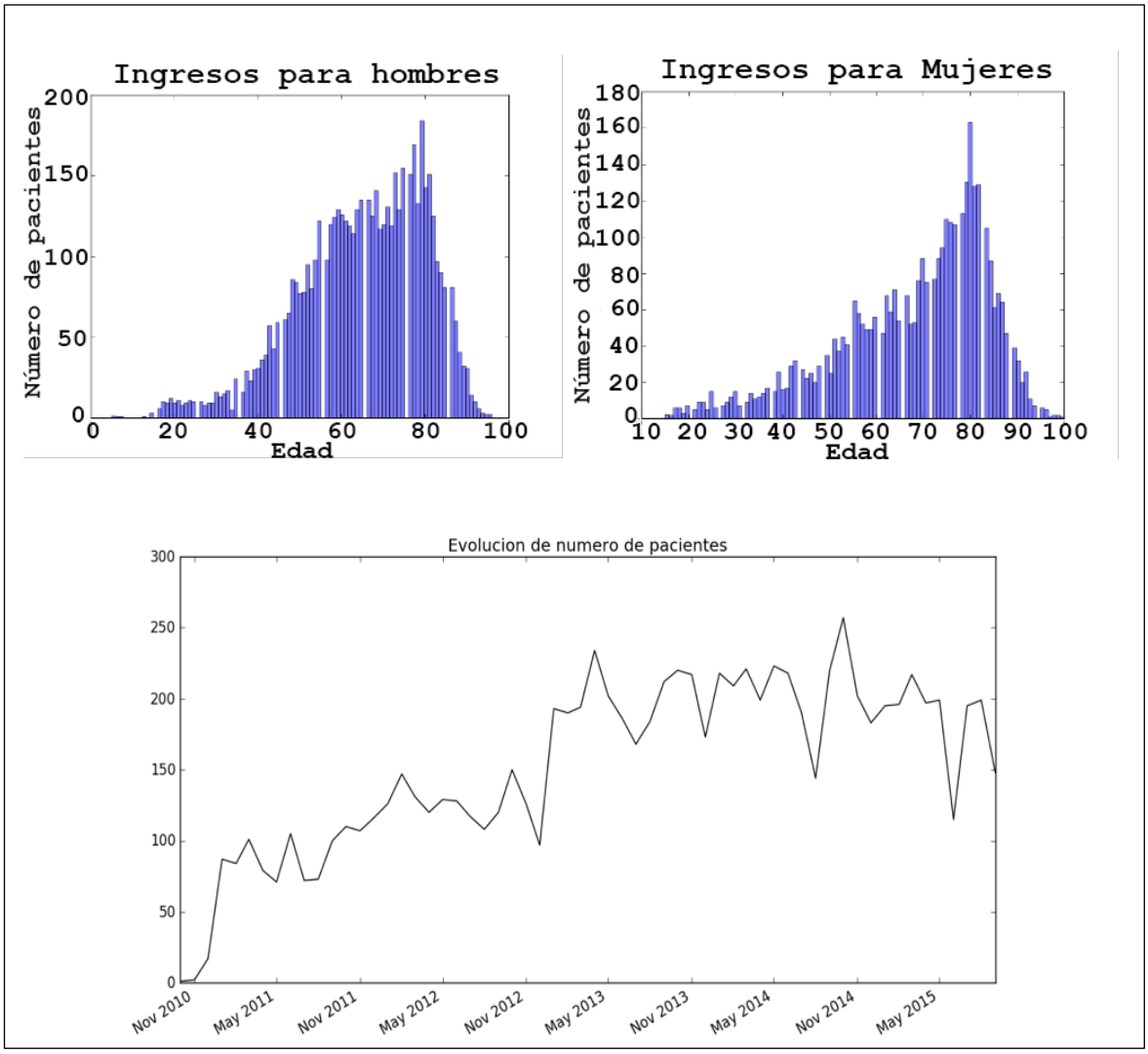


Ilustración 7. Gráficas de distribución en función del sexo y la edad, y evolución temporal de los ingresos en el rango de estudio.

También se puede analizar el conjunto de datos relacionando la distribución de dos o más variables entre sí de forma gráfica y, de esta forma, descubrir posibles nuevos patrones. Por ejemplo, como se puede ver en la Ilustración 8, podemos analizar cómo se distribuye el motivo que causa el alta hospitalaria en función del sexo del paciente, la edad y los días de estancia, mostrando la línea de tendencia en cada caso particular. Cabe destacar que se producen más altas voluntarias en los hombres, que los traslados a otro hospital se producen en mujeres de más avanzada edad que los hombres y que los fallecimientos en mujeres se producen después de permanecer más días ingresadas. En la Ilustración 9, se puede observar el número de días que permanece ingresado un paciente en función de su edad y del sexo, donde se puede comprobar que hay más mujeres de avanzada edad ingresadas que hombres, lo que parece ser indicativo de que las mujeres son más longevas en lugar de indicar que permanecen más tiempo ingresadas a esas edades. Al combinar esta información con la Ilustración 8 (*TIPALT* = 4

- Exitus), en la que vemos que hay más fallecimientos de hombres que de mujeres, el hecho de que las mujeres son más longevas parece la explicación más plausible.

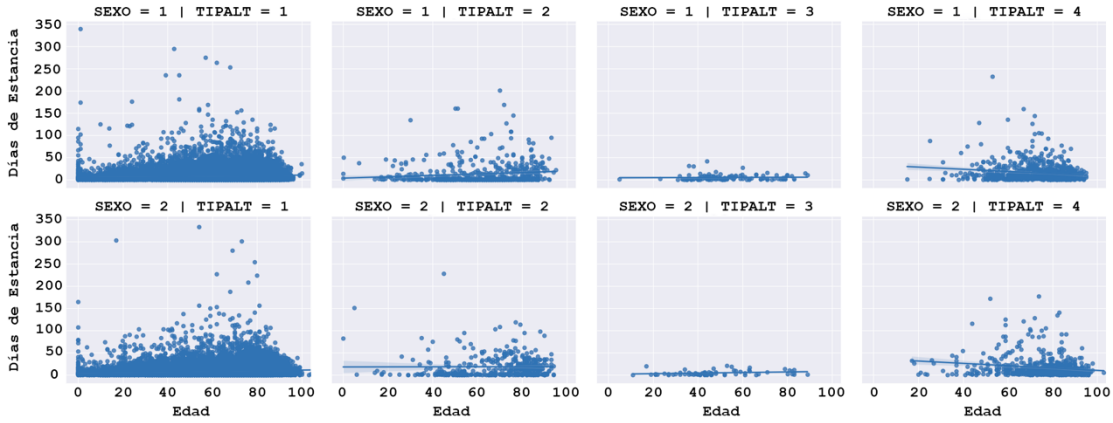


Ilustración 8. Gráficas de distribución de varias características. El sexo puede ser 1 (Hombre) o 2 (Mujer). El tipo de alta puede ser 1 (Domicilio particular), 2 (Traslado a otro hospital), 3 (Alta voluntaria) o 4 (Éxitus).

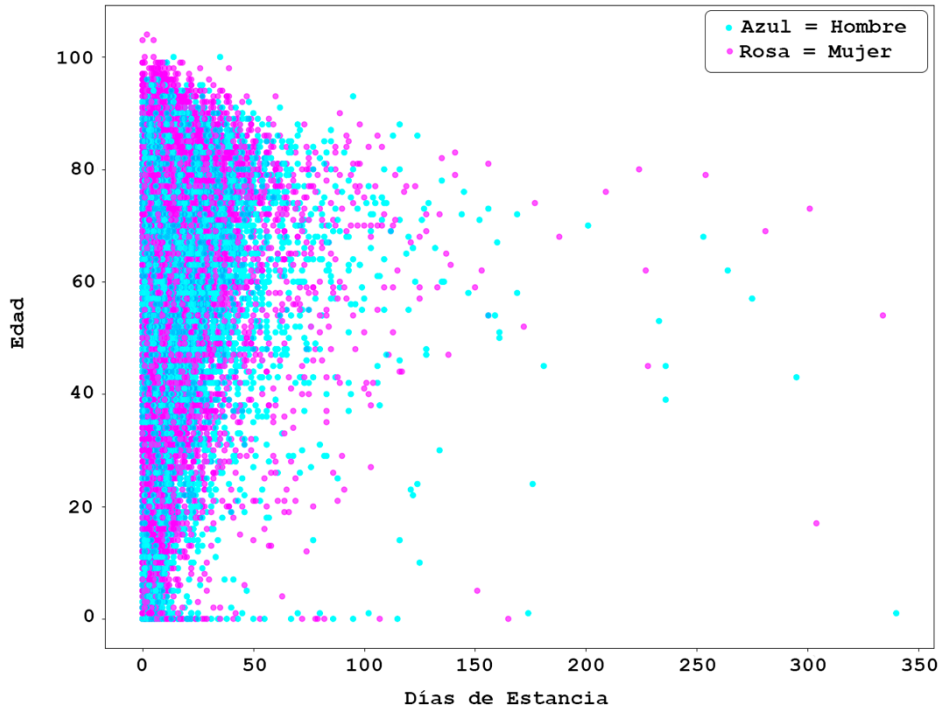


Ilustración 9. Gráfica de distribución que muestra cómo se distribuyen los pacientes en función del sexo, edad y días de estancia.

5.3 DÍAS DE PERMANENCIA

En el caso de estudio aquí tratado, se analiza el número de días que un paciente permanece ingresado en cada uno de los departamentos encontrados del hospital analizado. Este tipo de análisis también se conoce como LOS, del inglés Length Of Stay

(días de estancia). De esta forma, se puede comparar en que especialidad o departamento concreto es más alta la precisión a la hora de predecir el número de días de estancia, para así poder decidir, por ejemplo, en qué departamentos es más aconsejable dedicar más recursos de forma que se gestionen más eficientemente. En la Ilustración 10 podemos observar los departamentos encontrados en el hospital tratado, así como el número de días medio que los pacientes permanecen ingresados en cada departamento y su desviación media.

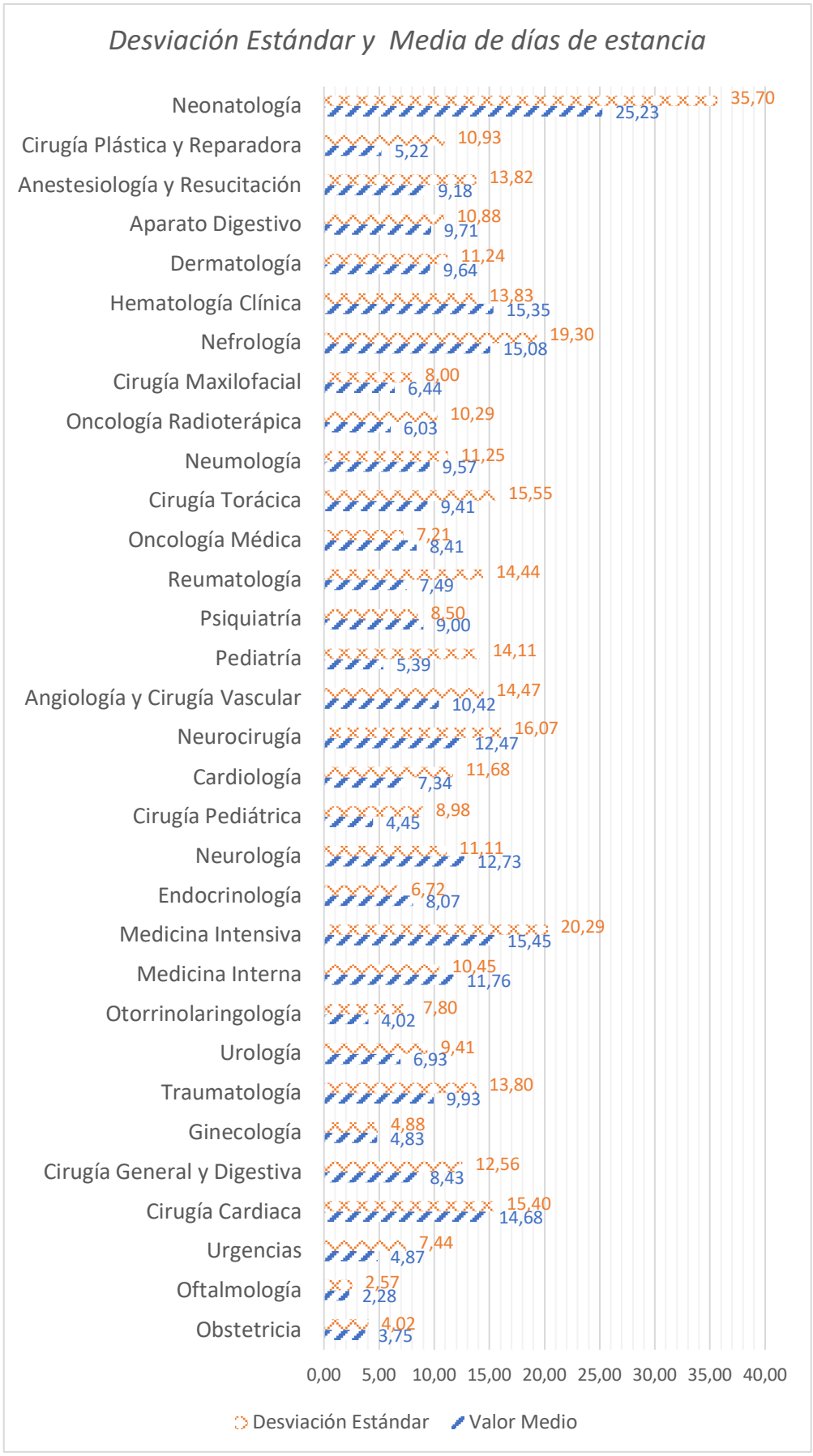


Ilustración 10. Distribución de los Departamentos del Hospital analizados

Las razones por las que se debe permanecer ingresado pueden deberse a múltiples motivos: enfermedades crónicas, estancias prolongadas debidas a operaciones quirúrgicas, recuperación por enfermedades, pacientes mayores con problemas de dependencia, etc. (Panchami y Radhika, 2014). De igual forma, como también se afirma en (Parag y Hitesh, 2014), el número de días que se tiene que permanecer ingresado dependerá de cada tipo de enfermedad o del motivo del ingreso, por lo que contar con un conjunto de datos que contenga esta información será aconsejable para poder predecir los días de estancia de forma adecuada y de manera que los algoritmos puedan detectar algún patrón por el que puedan deducirlo. Tal y como podemos observar en la Tabla 4, dentro del punto 3.5.2 sobre los materiales disponibles, disponemos de información sobre el diagnóstico principal y el procedimiento principal que se realizó a los pacientes ingresados en el hospital, los cuales nos indican la enfermedad o el motivo de ingreso y ayudarán en la predicción de los días que podrá permanecer ingresado un paciente. También influyen otros factores que a priori pueden parecer no relacionados, tal y como podemos comprobar en la gráfica realizada sobre correlaciones existentes entre las características del conjunto de datos utilizado. En ella se puede observar que el peso (variable relacionada con los costes del servicio) y el grupo relacionado por el diagnóstico (GRDS – variable que relaciona los tipos de pacientes con el coste que representa su asistencia) también influyen en la variable que contiene el número de días de estancia.

Una estancia prolongada de los pacientes en un hospital supone un coste de recursos elevado, además de incomodidad para los pacientes. Por ello, es importante la necesidad de planificar y gestionar correctamente los recursos disponibles ante la posible demanda que pueda necesitarse, de forma que se consiga un equilibrio entre los costes y la calidad del servicio. Un exceso de recursos reservados que posteriormente no se utilicen, puede suponer malgastar parte del presupuesto del hospital. De la misma forma, un defecto de recursos puede derivar en retrasos en la admisión de pacientes y en una pobre calidad de los servicios sanitarios. Por estos motivos, al igual que afirma Morton et al. (2014) en su estudio sobre los días de estancia con pacientes diabéticos, el estudio del número de días de estancia en un hospital es un tema ampliamente estudiado desde finales de los años 60 que, mediante una correcta planificación, puede suponer un ahorro de costes y una mejora en la calidad de la atención sanitaria.

Para poder realizar este estudio es necesario disponer de una característica de salida que nos indique el número de días que un paciente ha permanecido ingresado, de forma que se puedan predecir este número de días. Aunque esta característica no era una de las disponibles en el conjunto de datos, sí que se pudo obtener mediante el cálculo de la diferencia de días entre la fecha de ingreso y la fecha de alta del paciente. Durante la revisión de la característica de salida obtenida, se observó que los días que los pacientes permanecían ingresados estaban englobados en un pequeño y concreto grupo de valores enteros, por lo que se decidió tratar el estudio como un problema de clasificación en lugar de regresión, dado también el hecho de que la investigación trata de una comparativa entre departamentos y no es imprescindible conocer el número de días exacto.

De igual manera, como características de salida se seleccionaron las más influyentes, elegidas entre las de mayor índice de correlación mediante el coeficiente de Pearson. Para ello se utilizó el procedimiento *SelectKBest* comentado en el punto 3.5.2.3, utilizando el test estadístico chi-cuadrado, ya que es el óptimo para cálculos de clasificación. Como datos de entrada, se eligieron las 12 características con mayor correlación por motivos de eficiencia y porque el añadir más características no mejoraba los resultados.

En este caso de estudio se utilizaron varios algoritmos de aprendizaje automático, tanto para poder comparar los resultados como para analizar cuál es el que mejores resultados predictivos obtiene para este caso de estudio particular. Los algoritmos empleados fueron:

- K Nearest Neighbours (KNN)
- Support Vector Machines (SVM)
    - Polynomial Kernel
    - RBF Kernel
    - Support Vector Classifiers (SVC)
        - Non Linear Kernels
        - Linear Kernel
- Naïve Bayes
- Decision Tree Classifier
- Random Forest
- Neural Network

Los resultados obtenidos se evaluaron según la precisión obtenida y el error cometido, tal y como puede observarse en la Ilustración 11 y en la Ilustración 12.

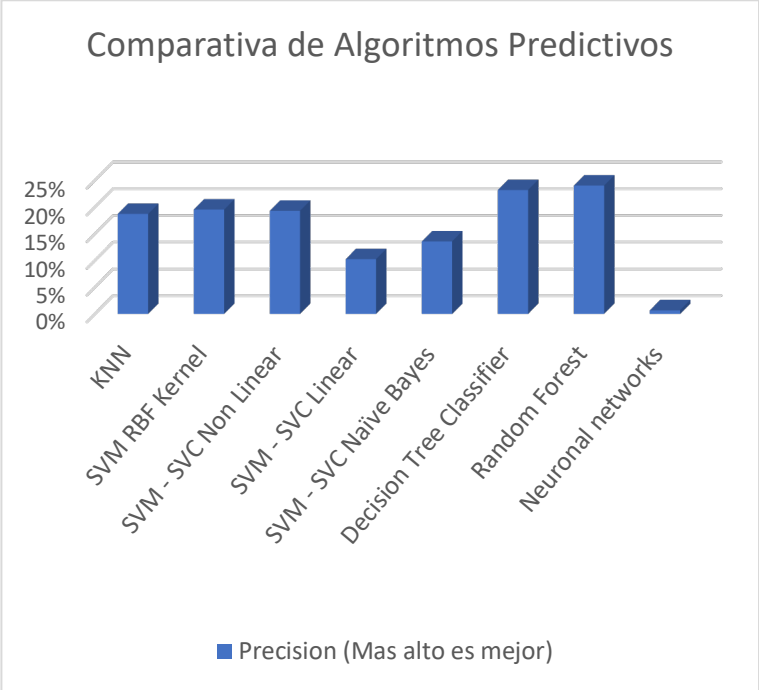


Ilustración 11. Comparativa de algoritmos según la precisión obtenida

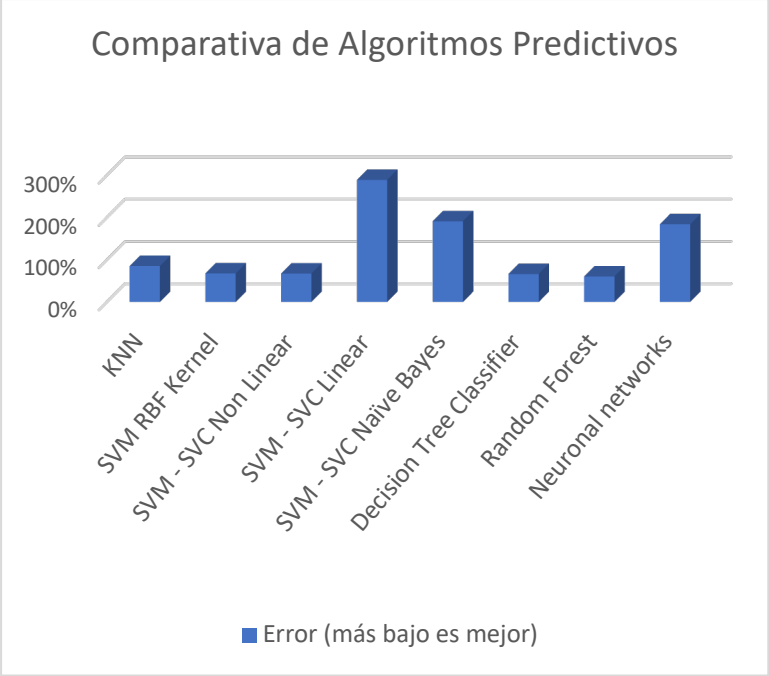


Ilustración 12. Comparativa de algoritmos según el error cometido

Los algoritmos basados en árboles de decisión sobresalieron sobre los demás a la hora de ofrecer los mejores resultados, como son *Decision Tree Classifier* y *Random Forest*. También obtuvieron buenos resultados aquellos basados en el kernel RBF (núcleo basado en *Radial Basis Function*), como son los algoritmos *Support Vector Machines* (SVM) y *Support Vector Classifier* (SVC) *Non Linear Kernel*, ambos basados en el kernel RBF. Los departamentos con los mejores resultados para predecir LOS fueron Cirugía General y Digestiva, y otros departamentos de cirugía (Angiología y Cirugía Vascular, Cirugía Maxilofacial, Cirugía Cardiaca y Neurocirugía), todos ellos con el algoritmo *Random Forest* como mejor método. Los departamentos con el error más bajo fueron Obstetricia y Oftalmología, estando también estos departamentos en los que mayor precisión obtuvieron en las predicciones. Ambas métricas, error cometido y precisión ayudan a la hora de elegir el departamento más adecuado para la predicción de LOS.

En la Tabla 7 podemos observar un listado de departamentos, ordenado de más aconsejable a menos aconsejable, con los resultados obtenidos en la medición del error cometido por las predicciones, medido en días. Como ejemplo, cabe destacar el departamento de Obstetricia con un error de 1,21 días y el departamento de Oftalmología con un error de 0,70 días en la realización de las predicciones, datos obtenidos con las muestras de test elegidas de forma independiente a las de entrenamiento.

DEPARTAMENTO	Valor med. De días de están.	Valor máx. de días de están.	KNN	SVM Kernel Polinomial	SVM Kernel RBF	SVM – SVC No Lineal	SVM – SVC Lineal	SVM – SVC Naïve Bayes	Árboles de Decisión	Random Forest	Redes Neur.
Oftalmología	2,28	29	0,84	32,45	0,84	0,90	1,74	3,00	0,70	0,78	2,26
Obstetricia	3,75	69	1,55	76,33	1,48	1,42	4,03	3,23	1,21	1,27	3,15
Endocrinología	8,07	46	4,63	70,29	4,56	4,49	12,25	10,00	4,73	3,41	7,10
Oncología Radioterápica	6,03	102	2,99	308,13	3,73	4,19	15,25	12,08	3,22	2,68	30,05
Ginecología	4,83	57	3,35	84,21	3,80	4,34	5,32	5,53	2,33	2,20	4,33
Urología	6,93	159	3,83	295,21	3,62	3,62	13,22	18,53	3,36	3,42	7,53
Cirugía Cardiaca	14,68	153	11,89	315,75	10,99	11,57	30,74	24,20	7,23	7,11	9,00
Otorrinolaringología	4,02	114	2,07	166,07	1,95	1,98	2,76	4,59	2,18	1,99	7,07
Urgencias	4,87	103	2,69	288,88	2,49	2,48	25,72	12,11	2,49	2,37	10,97
Cirugía Pediátrica	4,45	151	2,55	358,21	2,66	2,95	58,95	3,84	2,32	2,29	2,98
Reumatología	7,49	281	5,90	660,54	6,22	6,31	5,92	6,31	4,49	3,90	25,96
Cirugía Torácica	9,41	150	7,97	480,48	6,64	6,53	10,01	14,48	5,62	4,93	14,75
Cirugía General y Digestiva	8,43	295	5,69	608,02	5,06	5,03	52,25	17,83	4,98	4,43	11,52
Traumatología	9,93	301	7,72	854,78	7,01	6,99	8,29	20,27	5,76	5,37	51,61
Cirugía Plástica y Reconstructiva	5,22	156	3,30	377,12	3,06	2,93	12,13	11,91	2,88	2,94	15,00
Medicina Intensiva	15,45	304	13,41	1069,73	13,99	14,20	41,78	16,83	8,64	9,07	10,54
Cirugía Maxilofacial	6,44	78	3,83	138,14	3,64	3,77	5,38	8,50	4,18	3,76	15,43
Pediatría	5,39	340	3,38	934,51	3,05	3,07	13,73	11,17	3,15	3,17	3,48
Medicina Interna	11,76	188	7,70	316,28	7,03	7,02	17,80	10,34	7,03	6,94	7,16
Oncología Médica	8,41	56	5,62	92,82	5,39	5,46	13,80	9,47	5,35	5,00	14,08
Neurología	12,73	107	7,78	175,44	8,01	7,69	18,60	28,10	8,07	8,14	13,01
Neurocirugía	12,47	172	9,02	397,47	9,23	8,62	32,56	15,58	8,02	7,53	14,97
Angiología y Cirugía Vascular	10,42	169	7,86	579,71	7,25	7,26	24,17	42,87	7,48	6,37	12,53
Cardiología	7,34	275	5,73	846,73	4,67	4,77	30,22	20,19	4,54	4,57	7,21
Psiquiatría	9,00	78	9,40	202,86	6,24	6,57	33,31	72,20	5,60	5,81	25,94
Neonatología	25,23	174	133,06	1398,25	18,64	18,70	381,5	15,95	62,30	29,16	131,76
Aparato Digestivo	9,71	120	8,11	320,57	6,80	6,66	12,43	23,98	6,37	6,42	17,34
Hematología Clínica	15,35	88	13,79	195,57	10,14	10,92	16,02	20,26	10,43	10,64	23,05
Anestesiología y Resucitación	9,18	83	7,11	261,80	7,15	7,09	24,07	11,77	6,07	7,85	29,17
Nefrología	15,08	334	15,60	955,71	14,30	14,03	45,77	44,29	13,44	11,50	20,33
Neumología	9,57	236	8,02	435,34	8,72	8,65	14,50	20,20	8,06	7,33	12,76
Dermatología	9,64	128	9,79	116,21	11,38	7,75	29,03	12,79	9,33	11,33	17,52

Tabla 7. Listado de departamentos con el error cometido en días ordenado de menor a mayor error

En la Tabla 8 se muestra un listado con los departamentos del hospital, mostrando los resultados obtenidos en referencia a la precisión de cada algoritmo, ordenado de más adecuado a menos adecuado. Si observamos los departamentos más adecuados según el listado de la Tabla 7, Obstetricia y Oftalmología obtienen una precisión de 37,85% y

64,05%, respectivamente. Cabe destacar que en listado de la Tabla 8, los mejores departamentos según la precisión fueron Oftalmología y Oncología Radioterápica.

DEPARTAMENTO	KNN (K = 15)	SVM Kernel Polinomial	SVM Kernel RBF	SVM – SVC No Lineal	SVM – SVC Lineal	SVM – SVC Naïve Bayes	Árboles de Decisión	Random Forest	Redes Neur.
Oftalmología	58,69%	0,00%	59,83%	57,62%	57,07%	9,17%	64,05%	61,34%	1,29%
Oncología Radioterápica	49,28%	0,00%	38,57%	30,53%	19,99%	25,94%	49,62%	51,15%	0,26%
Reumatología	22,15%	0,00%	19,39%	17,88%	8,60%	22,93%	36,20%	38,24%	0,18%
Obstetricia	30,41%	0,00%	32,99%	33,94%	2,46%	31,55%	37,85%	37,46%	0,51%
Otorrinolaringología	32,62%	0,12%	33,80%	32,20%	20,66%	25,11%	34,72%	35,82%	0,91%
Urgencias	33,33%	0,05%	35,11%	35,60%	20,78%	23,24%	35,11%	35,54%	0,53%
Pediatría	23,01%	0,00%	25,96%	25,64%	4,85%	25,19%	30,17%	33,61%	2,07%
Cirugía Pediátrica	28,49%	0,00%	26,31%	27,05%	26,45%	21,76%	33,35%	32,40%	1,08%
Anestesiología y Resucitación	20,95%	0,78%	22,12%	22,77%	4,22%	24,33%	33,34%	32,18%	0,00%
Cirugía Plástica Reconstructiva	24,98%	0,00%	30,57%	32,34%	15,73%	11,56%	30,86%	30,36%	0,51%
Cardiología	20,58%	0,00%	21,69%	21,38%	11,19%	14,47%	27,41%	28,52%	2,66%
Endocrinología	22,14%	0,00%	22,70%	24,74%	17,70%	8,12%	24,59%	26,07%	0,00%
Ginecología	16,01%	0,00%	14,83%	15,08%	10,09%	13,41%	25,88%	25,98%	0,42%
Urología	17,82%	0,00%	19,32%	18,68%	8,27%	15,47%	24,67%	23,59%	0,70%
Cirugía General y Digestiva	17,07%	0,00%	17,77%	17,82%	14,62%	12,21%	21,45%	22,83%	1,57%
Angiología y Cirugía Vascular	16,45%	0,00%	15,97%	15,71%	6,49%	10,20%	19,93%	21,74%	0,82%
Cirugía Torácica	9,13%	0,00%	12,97%	15,93%	5,53%	11,33%	18,20%	20,86%	0,16%
Psiquiatría	13,37%	0,04%	16,75%	15,40%	5,38%	1,59%	20,68%	20,71%	0,11%
Traumatología	13,61%	0,00%	14,08%	14,50%	8,38%	9,69%	19,93%	20,25%	0,12%
Neonatología	6,00%	0,00%	18,43%	18,43%	1,39%	19,51%	8,22%	14,37%	0,00%
Cirugía Maxilofacial	16,54%	0,00%	17,80%	18,26%	3,71%	16,32%	15,39%	19,26%	0,60%
Oncología Médica	14,06%	0,12%	12,87%	11,73%	3,86%	10,62%	16,12%	17,60%	0,25%
Aparato Digestivo	12,44%	0,00%	12,83%	14,30%	7,20%	9,14%	14,82%	16,96%	0,70%
Cirugía Cardiaca	12,68%	0,00%	13,20%	11,19%	5,62%	5,93%	14,90%	16,13%	0,87%
Medicina Intensiva	11,85%	0,00%	9,42%	9,18%	4,21%	11,72%	14,64%	14,13%	1,17%
Neurocirugía	8,89%	0,00%	9,91%	9,49%	3,72%	7,32%	12,70%	12,94%	0,72%
Dermatología	9,49%	0,00%	6,65%	6,49%	3,07%	6,39%	11,84%	9,26%	0,00%
Neumología	7,50%	0,00%	8,67%	8,77%	5,58%	7,44%	9,55%	11,24%	1,04%
Nefrología	8,14%	0,00%	7,98%	7,19%	2,60%	5,23%	10,50%	10,88%	0,66%
Neurología	8,48%	0,00%	8,87%	9,85%	8,88%	2,67%	8,48%	8,94%	0,40%
Medicina Interna	7,65%	0,00%	9,06%	8,99%	4,59%	8,55%	9,09%	9,09%	1,10%
Hematología Clínica	4,99%	0,00%	8,31%	8,84%	5,43%	6,01%	7,88%	8,89%	0,37%

Tabla 8. Listado de departamentos con la precisión obtenida por los algoritmos ordenados de mayor a menor precisión

Utilizando herramientas de inspección visual y estadística descriptiva, como pueden ser los histogramas o los diagramas de barras, se detectaron patrones de comportamiento

que mejoraban los resultados obtenidos, tanto en precisión como en error cometido. Por ello se prepararon conjuntos de datos especiales sobre varios departamentos para ser procesado por los algoritmos utilizados. Estos conjuntos de datos fueron preparados en la investigación para comprobar la mejora en los resultados. Los departamentos y patrones obtenidos, con sus correspondientes mejoras, fueron:

- *Oftalmología*. Aquellos pacientes con edades menores o iguales a 75 años mejoran el error cometido en un 12,01%, pasando de 0,91 días de error a 0,63 días de error. En cuanto a la precisión, la mejora supone un 8,53%
- *Obstetricia*. Las pacientes con embarazos prolongados (post-término) de alto riesgo, excediendo las 42 semanas de gestación con riesgo debido a una mayor mortalidad perinatal y morbilidad, en el cálculo del error cometido se reduce (mejora) en un 13,65%, pasando de 1,21 días de error a 0,83 días de error. La precisión mejora en un 9,56%.
- *Urgencias*. En este departamento, los resultados mejoran cuando los pacientes son mujeres cuyo motivo de ingreso se debe a enteritis (inflamación intestinal). El error cometido en este caso se reduce en un 13,65%, pasando de 2,37 días a 1,71 días de error, y la precisión mejora en un 16,11%.
- *Cardiología*. Si el motivo de ingreso se debe a una fibrilación o a una palpitación del corazón, el error cometido se reduce en un 20,57%, pasando de 4,54 días a 3,03 días de error. Pero si además los pacientes son hombres, la reducción del error alcanzada es de un 21,61%, llegando a ser de 2,95 días. La mejora alcanzada en este caso con respecto a la precisión es de un 23,83%.

Esta investigación logra obtener un estudio de los días de estancia de los pacientes en todos los departamentos de un hospital, en lugar de centrarse en uno concreto o en una única patología, de forma que se obtiene una visión global a nivel hospitalario de esta magnitud, algo que hasta ahora no se había realizado. Se debe tener en cuenta que los departamentos de Rehabilitación y de la Unidad de Control del Dolor en Urgencias, también pertenecientes al hospital, se excluyeron del estudio debido a la carencia de muestras suficientes para ser tratados con los algoritmos. Con los resultados de este análisis se puede abordar una reducción de costes y una mejora en la calidad de atención al paciente en un hospital. Con ello también se redistribuyen los recursos de forma más eficiente, como puede ser la reserva de un número de camas acorde a la previsión de pacientes que se estima van a ingresar próximamente o como puede ser la previsión de personal que va a ser necesario para atender a los futuros pacientes en los departamentos correctos, todo esto consultando las listas ordenadas sobre la previsión de los departamentos. También resultan de interés los patrones encontrados sobre casos específicos en determinados departamentos, como puede ser la mejora en las predicciones y la reducción del error en, por ejemplo, el caso de pacientes masculinos con problemas debidos a fibrilaciones o palpitaciones del corazón en el departamento de Cardiología, donde se logran mejoras en la precisión de hasta un 23,83%.

5.4 READMISIÓN HOSPITALARIA

Este caso de estudio trata sobre la predicción de las readmisiones de los pacientes que se producen en un hospital tras haber sido dados de alta. La diferencia con los reingresos es que las readmisiones se producen debido al mismo motivo que provocó el ingreso o a otro relacionado con la misma patología o enfermedad. En cambio, los reingresos pueden ser debidos a cualquier motivo. La readmisión hospitalaria es un indicador importante y muy utilizado, ya que se utiliza para comparar la calidad entre hospitales, para mejorar la calidad en el cuidado al paciente y para mejorar la administración de los recursos hospitalarios. Este indicador cobra especial importancia, ya que puede penalizar a los hospitales con altas tasas de readmisión, tal y como se afirma en (Greysen, Stijacic, Auerbach y Covinsky, 2015), donde comentan cómo el seguro médico privado Medicare penaliza a los hospitales en la actualidad. También se manifiesta en este artículo que las readmisiones afectan a entre un 15% y un 30% de los pacientes de Medicare, con costes que superan los 17 billones de dólares anuales. También es un indicador importante por utilizarse para analizar las tasas de reentrada, de forma que sirva para validar la efectividad de la atención al paciente entre los hospitales y centros de salud.

Este caso de estudio analiza diferentes tipos de readmisiones para compararlas entre ellas. Los tipos de readmisión analizados son las readmisiones a 7 días, 30 días, 60 días y 90 días, ya que son los casos más apropiados y utilizados, tal y como se afirma en (Kiran et al., 2004) (Kansagara et al., 2011), donde realizan una revisión sistemática de artículos sobre readmisiones en diferentes años.

Para poder elaborar un conjunto de datos sobre readmisiones, se necesita conocer las patologías que están relacionadas con la enfermedad concreta que provocó un ingreso. Este conocimiento debe aportarlo un doctor en medicina. En este caso de estudio la readmisión se estudia sobre patologías de digestivo, ya que se utiliza el conocimiento aportado por las tablas que contiene el artículo científico (Peiró et al., 1996). Esta investigación trata sobre patologías de digestivo y hepatobiliares, y aporta un listado de todas las patologías relacionadas, elaboradas por expertos en medicina. Gracias a este listado se elaboró, en primer lugar, un algoritmo que permitiera identificar todos los pacientes que ingresaron al hospital por alguna patología de digestivo en alguno de los 13 posibles diagnósticos que se pueden tener. La base de datos utilizada para identificar las patologías de digestivo es la misma que la utilizada en el caso de estudio anterior y la misma que se describe en el punto 3.5.2. Esta base de datos permite introducir hasta 13 posibles diagnósticos para un mismo paciente en su ficha de ingreso hospitalaria. Posteriormente, con el conjunto de datos de pacientes con patologías de digestivo elaborado por el primer algoritmo, se utiliza como entrada a un segundo algoritmo para identificar los distintos tipos de readmisión. Como parámetro de entrada se introducirá el número de días según el tipo de readmisión que se desea obtener. De esta forma, en nuestro caso, se ejecuta 4 veces este algoritmo, una por cada tipo de readmisión a estudiar. En la Ilustración 13 podemos ver un esquema de funcionamiento de estos dos algoritmos creado para esta investigación.

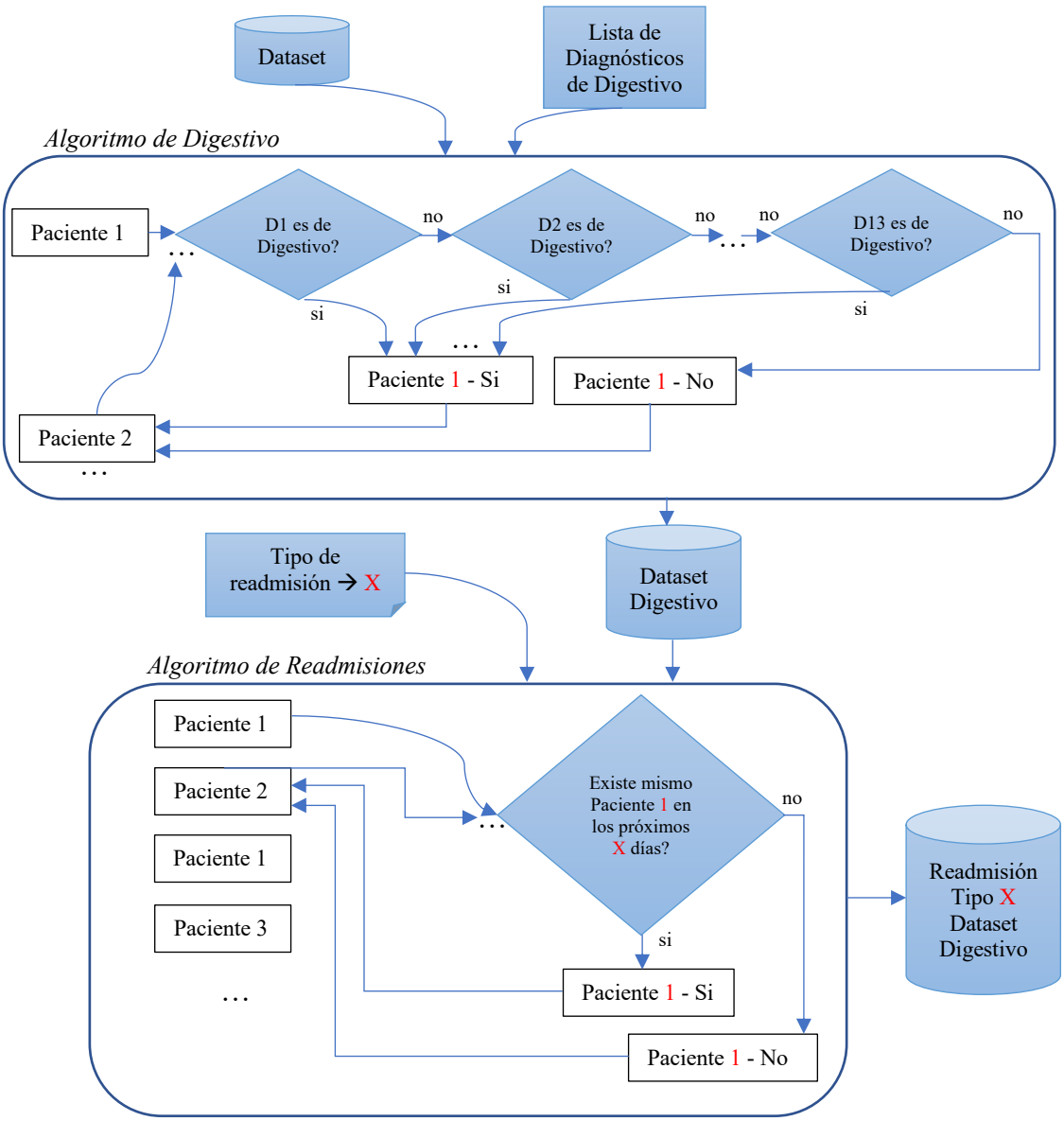


Ilustración 13. Esquema de los algoritmos de identificación de patologías de Digestivo y de cálculo de Readmisiones

El segundo algoritmo de readmisión marcará los pacientes como readmisiones de cada tipo concreto, buscando a futuro si va a producirse un próximo reingreso con la misma patología de digestivo. De esta forma, las predicciones que se realizarán con el modelo que se creará serán para predecir si se prevé que vaya a tener una readmisión en los próximos X días (según sean 7, 30, 60 o 90 días de reingreso). El resultado que se obtiene a la salida de este segundo algoritmo es un conjunto de datos para cada tipo de reingreso que se haya seleccionado. Este conjunto de datos es el que se utilizará como entrada para realizar las predicciones.

Las tasas de readmisión detectadas en cada uno de los tipos fueron de 1,10% para las readmisiones de 7 días, un 8,68% para las de 30 días, un 15,57% para las de 60 y un 19,72% para las de 90 días de readmisión. Las tasas obtenidas en las readmisiones de 7 días concuerdan con los resultados publicados por Peiró et al. (1996), donde obtienen un 1,77% para este tipo de readmisión en pacientes con la misma patología. El resto de readmisiones que obtienen es mayor que la obtenida en nuestra investigación: 4,67% para readmisiones de 30 y 60 días, y un 8,97% para las de 90. Estas desviaciones con respecto a nuestra investigación son razonables, ya que los autores de (Peiró et al., 1996) sólo consideran el diagnóstico principal y en nuestra investigación se consideran todos los diagnósticos posibles. En cambio, en el estudio desarrollado por Futoma et al. (2015) se obtienen porcentajes de 19% y 27% para readmisiones a 30 y 90 días, los cuales son más cercanos a los nuestros, teniendo en cuenta que estudian enfermedades sobre inflamaciones intestinales en vez de enfermedades de digestivo y hepatobiliares.

Para realizar las predicciones se utiliza una técnica de aprendizaje profundo basada en redes neuronales. En concreto se utiliza una red neuronal convolucional Long Short-Term Memory (LSTM) utilizando secuencias temporales y aplicada a modelos de Clasificación. Se utiliza esta técnica por obtener buenos resultados, lograr aprender con los datos ya introducidos previamente y por ser una técnica reciente que no se ha utilizado en este tipo de contexto. En general otros autores como George et al. (2019) y Whitlock et al. (2010), utilizan la técnica de Regresión Logística que es la más utilizada para estos tipos de estudios sobre readmisiones. Solo unos pocos autores como Bergese et al. (2019) y Jamei et al. (2017), utilizan modelos comunes de redes neuronales de forma conjunta con otras metodologías.

Como las readmisiones dependen de la fecha de ingreso al hospital, es razonable pensar que pueden existir patrones con ciclos temporales. Por este motivo se utilizan estas redes neuronales aplicando bucles con las distintas readmisiones detectadas en cada tipo de readmisión. Consecuentemente, el conjunto de datos de entrada se adapta para aplicar el modelo de la red neuronal sobre estos bucles temporales, de forma que se eligen diferentes ciclos, desde secuencias de 2 en 2 hasta de 30 en 30 elementos ordenados cronológicamente, estudiando el comportamiento en todos ellos. El número de muestras del conjunto de datos disponible no permitía estudiar las secuencias con grupos de elementos mayores. La estructura utilizada como entrada para las distintas readmisiones del modelo de red neuronal consiste en ternas del tipo (número de muestras, secuencia temporal, número de características). De este modo, para un conjunto de datos con 310 muestras en grupos de 3 elementos como secuencia temporal y 21.392 características correspondería a una estructura (310, 3, 21392). Se ha de tener en cuenta que se obtienen 21.392 características debido a que se aplica la técnica *one-hot encoding* comentada en el apartado 5.3. Esta técnica se usa para convertir todas las características categóricas en numéricas, ya que el algoritmo de red neuronal, para poder trabajar, debe disponer de este tipo de datos.

Adicionalmente, se probaron varias configuraciones sobre el modelo de red neuronal para obtener el modelo más eficiente. En concreto se optimizaron los siguientes parámetros:

- *Epochs*. Este parámetro contiene el número de veces que se ejecutan los algoritmos *backpropagation* y *forwardpropagation* en la red. Se usó un rango de entre 15 y 90 veces y se eligió el 15 como el número de veces más óptimo.
- *Batch_size*. Este parámetro contiene el número de muestras que se prueban en cada iteración, para los cuales se ejecutan los algoritmos comentados en el anterior parámetro. Se probó el rendimiento para valores entre 64 y 5.000 muestras y se eligieron 64 muestras como el valor más adecuado.
- *Algoritmo de optimización*. Se evaluaron los algoritmos *SGD, RMSprop, Adadelta, Adagrad, Adam, Adamax y Nadam*. El algoritmo de optimización que mejores resultados obtuvo y que se utilizó para el modelo fue *Adam*.

También se utilizaron los siguientes parámetros como mejor compromiso entre eficacia y eficiencia para la generación del modelo de red neuronal:

- *Función loss*. Este parámetro indica el comportamiento de la red neuronal mediante la evaluación del desvío que se produce durante el entrenamiento entre los valores predichos y los valores reales. La función utilizada fue *categorical_accuracy*, ya que es la función más adecuada para trabajar con tipos de datos categóricos para el conjunto de datos de readmisiones disponible. Esta función también es la más apropiada para obtener la mejor precisión (*accuracy*) con el tipo de datos creado mediante la técnica *one-hot encoding*.
- *Función de activación*. Se utilizó la función de activación lineal rectificada (*ReLU*).
- *Número de nodos*. Se utilizaron 100 nodos intermedios.
- *Función de salida*. Se utilizó la función *softmax* con los nodos de salida.

El motivo por el que se utilizó el rango comentado para los parámetros *epochs* y *batch_size* fue debido a que con los últimos valores del rango no se lograba mejorar los resultados y la eficiencia se reducía significativamente.

Como no existen estudios similares en estudios sobre readmisiones de patologías de digestivo, se realizan pruebas con tecnologías diferentes sobre el mismo conjunto de datos utilizado en el algoritmo principal de este caso de estudio. Se probaron las mismas readmisiones de 7, 30, 60 y 90 días con otra red neuronal convolucional LSTM bidireccional, pero aplicando regresión, que es el uso más común de este tipo de redes neuronales. Igualmente se estudiaron para las mismas secuencias temporales: desde 2 hasta 30 elementos. Es decir, se predice el número de días transcurridos hasta la próxima readmisión del paciente. En esta técnica se obtiene el valor concreto de días transcurridos hasta la próxima readmisión, en lugar de identificar el tipo de readmisión al que pertenece (clasificación). Se obtiene un error absoluto medio de 112,05 y un error cuadrático medio de 46.455,25, como valores medios entre todas las mediciones obtenidas de todos los grupos de elementos analizados. Como con la regresión no se pueden aplicar las mismas métricas que para los casos de clasificación, no se puede comparar directamente los resultados en ambos algoritmos. Por ello, sobre los valores obtenidos en las predicciones realizadas con el método de regresión, se clasifican para comprobar a cuál de los 4 tipos de readmisiones estudiados pertenece cada predicción, y de esta forma tener una medida relativa que permita poder comparar ambos métodos. De la misma forma, se utiliza otro algoritmo de aprendizaje automático genérico que

obtiene buenos resultados con otros conjuntos de datos similares, para poder comparar así su eficiencia con el algoritmo principal de la investigación. Este último algoritmo utilizado se llama *Árbol de Decisión*. De esta forma, en la Ilustración 14 podemos comparar los resultados obtenidos comparando los 3 algoritmos utilizados.

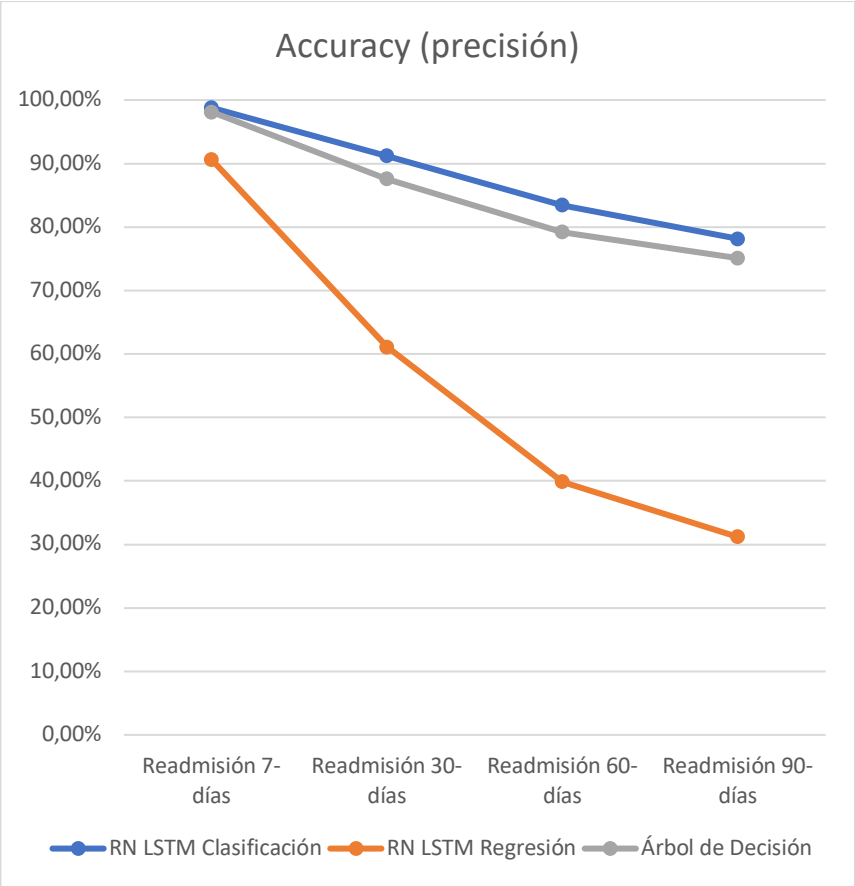


Ilustración 14. Comparación de diferentes algoritmos con el conjunto de datos de Readmisión

Como se puede comprobar en la Ilustración 14, los mejores resultados son los obtenidos por el algoritmo principal (RN LSTM Clasificación) de este caso de estudio que, aunque obtiene resultados parejos a los árboles de decisión, en el caso del algoritmo principal, puede seguir aprendiendo y mejorando sobre el conjunto de datos que se procese.

Para evaluar el algoritmo principal se utilizaron las métricas *accuracy* (precisión), *precision* (exactitud), *recall* (sensibilidad), *F1 Score* y *ROC AUC*, comentadas en el punto 3.3. En la Tabla 9 y la Tabla 10 podemos observar los resultados obtenidos para las pruebas realizadas a los 4 tipos de readmisiones analizados.

TABLA IX

	Readmisión 7-días					Readmisión 30-días				
Secuencia temporal	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
2	98,86%	0%	0	0	0,29052	89,57%	12,82%	0,04237	0,06369	0,36921
3	98,83%	0%	0	0	0,41399	89,56%	15,79%	0,03529	0,05769	0,42870
4	98,86%	0%	0	0	0,24048	91,34%	20,00%	0,01724	0,03175	0,46089
5	98,93%	0%	0	0	0,42101	90,94%	14,29%	0,02174	0,03774	0,44305

TABLA IX

	Readmisión 7-días					Readmisión 30-días				
Secuencia temporal	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
6	98,72%	0%	0	0	0,52070	89,98%	0,00%	0,00	0,00	0,43905
7	99,01%	0%	0	0	0,07349	92,29%	0,00%	0,00	0,00	0,48853
8	98,58%	0%	0	0	0,35331	91,76%	0,00%	0,00	0,00	0,45078
9	99,04%	0%	0	0	0,41720	90,73%	0,00%	0,00	0,00	0,48177
10	98,93%	0%	0	0	0,51799	91,46%	0,00%	0,00	0,00	0,53624
11	98,83%	0%	0	0	0,43478	91,80%	0,00%	0,00	0,00	0,41578
12	98,72%	0%	0	0	0,31241	90,60%	0,00%	0,00	0,00	0,38079
13	98,61%	0%	0	0	0,17527	91,20%	0,00%	0,00	0,00	0,38418
14	99,01%	0%	0	0	0,04020	91,04%	0,00%	0,00	0,00	0,35311
15	98,93%	0%	0	0	0,1	91,44%	0,00%	0,00	0,00	0,35330
16	98,30%	0%	0	0	0,47881	92,61%	0,00%	0,00	0,00	0,34592
17	98,79%	0%	0	0	0,37730	91,52%	0,00%	0,00	0,00	0,36471
18	98,72%	0%	0	0	0,34740	91,03%	0,00%	0,00	0,00	0,42153
19	98,65%	0%	0	0	0,14726	89,19%	0,00%	0,00	0,00	0,33582
20	98,57%	0%	0	0	0,24457	92,86%	0,00%	0,00	0,00	0,42885
21	98,51%	0%	0	0	0,28599	92,54%	0,00%	0,00	0,00	0,43992
22	98,44%	0%	0	0	0,43651	91,41%	0,00%	0,00	0,00	0,54157
23	99,18%	0%	0	0	0,33884	90,16%	0,00%	0,00	0,00	0,55303
24	98,29%	0%	0	0	0,13478	90,60%	0,00%	0,00	0,00	0,46741
25	99,11%	0%	0	0	0,01802	90,18%	0,00%	0,00	0,00	0,69579
26	98,15%	0%	0	0	0,34906	91,67%	0,00%	0,00	0,00	0,55948
27	99,04%	0%	0	0	0,51456	91,35%	0,00%	0,00	0,00	0,42047
28	99,00%	0%	0	0	0,03030	93,00%	0,00%	0,00	0,00	0,37327
29	98,97%	0%	0	0	0,94792	89,69%	0,00%	0,00	0,00	0,44944
30	98,92%	0%	0	0	0,43478	92,47%	0,00%	0,00	0,00	0,47674

Tabla 9. Resultados de métricas para readmisiones de 7 y 30 días

TABLA X

	Readmisión 60-días					Readmisión 90-días				
Secuencia temporal	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
2	81,55%	2,88%	0,13551	0,18239	0,40229	74,38%	27,04%	0,19557	0,22698	0,40395
3	79,98%	29,21%	0,17219	0,21667	0,41845	74,87%	24,44%	0,11579	0,15714	0,44702
4	80,97%	8,82%	0,02830	0,04286	0,46233	74,43%	18,06%	0,09702	0,12621	0,48794
5	80,82%	17,95%	0,08434	0,11475	0,44866	75,13%	14,58%	0,06604	0,09091	0,47470
6	80,38%	9,52%	0,02667	0,04167	0,53724	75,27%	17,14%	0,06452	0,09375	0,47808
7	84,33%	30,77%	0,06897	0,11267	0,50301	79,85%	20,00%	0,04167	0,06897	0,46006
8	82,67%	0,00%	0,00	0,00	0,48116	75,28%	8,00%	0,03030	0,04396	0,48329
9	80,51%	7,69%	0,02	0,03175	0,52103	73,48%	11,54%	0,04762	0,06742	0,54038
10	81,14%	0,00%	0,00	0,00	0,50111	78,65%	9,09%	0,01961	0,03226	0,45546
11	84,77%	0,00%	0,00	0,00	0,42756	78,91%	30,77%	0,08163	0,12903	0,42315
12	82,91%	0,00%	0,00	0,00	0,43763	72,65%	17,39%	0,08163	0,11111	0,45918
13	79,63%	8,33%	0,02941	0,04348	0,51034	78,24%	20,00%	0,02273	0,04082	0,50357
14	86,57%	0,00%	0,00	0,00	0,42615	78,11%	28,00%	0,21212	0,24138	0,40350
15	86,63%	0,00%	0,00	0,00	0,58988	78,61%	20,00%	0,05882	0,09091	0,52797
16	85,23%	100,00%	0,03704	0,07143	0,52821	78,41%	16,67%	0,02941	0,05	0,50508
17	83,03%	0,00%	0,00	0,00	0,47075	77,58%	20,00%	0,02941	0,05128	0,47968
18	83,97%	0,00%	0,00	0,00	0,56718	81,41%	0,00%	0,00	0,00	0,43606
19	83,78%	0,00%	0,00	0,00	0,38348	76,35%	12,50%	0,03448	0,05405	0,41698
20	80,71%	0,00%	0,00	0,00	0,41033	82,14%	0,00%	0,00	0,00	0,45510
21	83,58%	0,00%	0,00	0,00	0,47890	79,10%	50,00%	0,03571	0,06667	0,49326
22	85,94%	0,00%	0,00	0,00	0,47601	80,47%	0,00%	0,00	0,00	0,41166
23	82,79%	0,00%	0,00	0,00	0,48798	76,23%	16,67%	0,04	0,06452	0,50804
24	83,76%	0,00%	0,00	0,00	0,59533	76,92%	28,57%	0,08333	0,12903	0,48253
25	85,71%	0,00%	0,00	0,00	0,48633	81,25%	0,00%	0,00	0,00	0,44584
26	84,26%	0,00%	0,00	0,00	0,37621	80,56%	100,00%	0,04546	0,08696	0,37579
27	84,62%	0,00%	0,00	0,00	0,58427	82,69%	100,00%	0,05263	0,1	0,49474
28	87,00%	0,00%	0,00	0,00	0,41998	83,00%	0,00%	0,00	0,00	0,50602
29	84,54%	0,00%	0,00	0,00	0,57886	79,38%	0,00%	0,00	0,00	0,56104
30	87,10%	0,00%	0,00	0,00	0,40741	82,80%	0,00%	0,00	0,00	0,57305

Tabla 10. Resultados de métricas para readmisiones de 60 y 90 días

Estas tablas muestran los resultados obtenidos por las métricas comentadas en cada secuencia temporal y con cada tipo de readmisión. Se puede observar que la clase o tipo de readmisión se puede detectar razonablemente bien en los casos con bajo índice de *recall* y con una alta confiabilidad, en los casos con niveles altos de *precision*. Estos casos pueden observarse en readmisiones de 60 y 90 días, como en la secuencia con 16 elementos en readmisiones a 60 días, y en las secuencias de 21, 26 y 27 elementos para las readmisiones a 90 días (Tabla 10). Según la métrica *ROC AUC*, el valor medio de las mediciones para 7, 30, 60 y 90 días son 32,40%, 44,34%, 47,99% y 47,22%, respectivamente. Por lo tanto, la readmisión de 60 días es la que obtiene la mejor calidad en sus predicciones, seguida de cerca de las de 90 y 30 días. En lo que se refiere a las secuencias temporales, el mejor valor del número de elementos para clasificar predicciones son aquellas secuencias de 29 elementos para readmisiones de 7 días con un 94,79%, según la métrica *ROC AUC*, las secuencias de 25 elementos para readmisiones de 30 días con un 69,58%, las secuencias de 24 elementos para readmisiones de 60 días con un 59,53% y las secuencias de 30 elementos en readmisiones de 90 días con un 57,31%.

A pesar de los buenos resultados en precisión (*accuracy*) durante los entrenamientos, el modelo obtiene muchos valores con un 0 en exactitud (*precision*), *recall* y *F1 Score*. Este comportamiento se debe a que el modelo no es capaz de detectar valores reales positivos durante las pruebas de test debido a un número bajo de valores positivos. Las métricas *precision* y *recall* miden los valores positivos y la métrica *F1 Score* mide la relación existente entre *recall* y *precision*.

Los mejores resultados en términos de precisión (*accuracy*) se obtienen para las readmisiones a 7 días, pero debe tenerse en cuenta que estos resultados se deben al bajo número de casos positivos durante la fase de prueba, en la que se dispone de pocas muestras, lo que hace que se tenga un conjunto de datos desbalanceado. De forma global, teniendo en cuenta todas las métricas, los mejores resultados los obtiene la readmisión a 60 días, con un rango medio de secuencias temporales. Como se observa en la Ilustración 15, el mejor resultado se obtiene usando secuencias temporales con grupos de 16 elementos, que corresponde con una precisión de 85,23%, un 100% de exactitud y un nivel de calidad medio. Con respecto a los resultados obtenidos por las readmisiones de 30 y 90 días, los mejores resultados se obtienen con secuencias temporales de 4 y 27 elementos con precisiones de 91,34% y 82,69%, respectivamente.

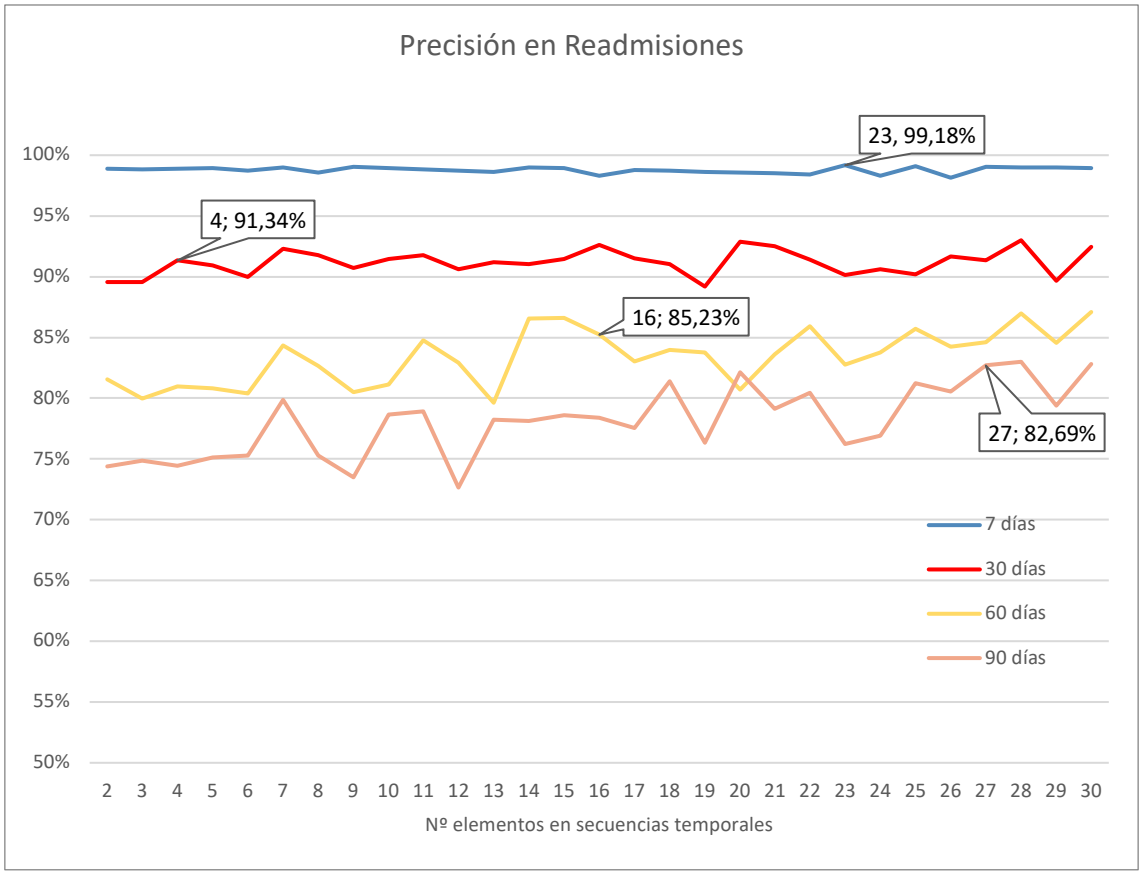
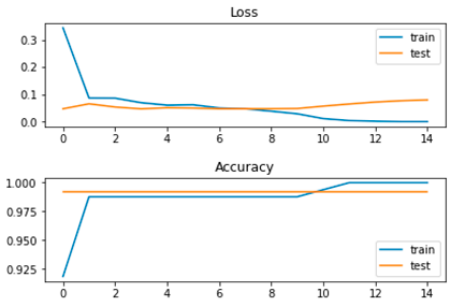


Ilustración 15. Comparativa de precisiones según los diferentes tipos de readmisión analizados

Se puede observar en la Ilustración 16 un análisis con más detalle para los resultados obtenidos en estos puntos concretos. Se puede ver el comportamiento de la función *loss*, la desviación entre los valores predichos y los valores reales, y la precisión obtenida, tanto en los procesos de entrenamiento como en los de pruebas. También se muestra la matriz de confusión para mostrar el comportamiento del modelo en cada uno de los casos analizados (ver punto 3.3 – Matriz de Confusión).

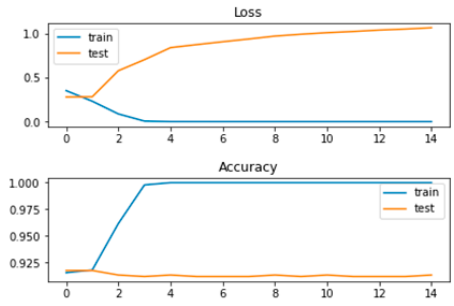
Readmisión 7 días - Ciclo Temporal 23




```
Matriz de Confusión:

[[121   0]
 [  1   0]]
```

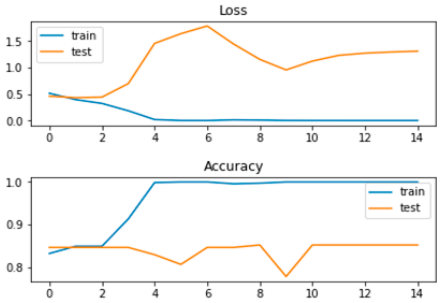
Readmisión 30 días - Ciclo Temporal 4




```
Matriz de Confusión:

[[642   4]
 [ 57   1]]
```

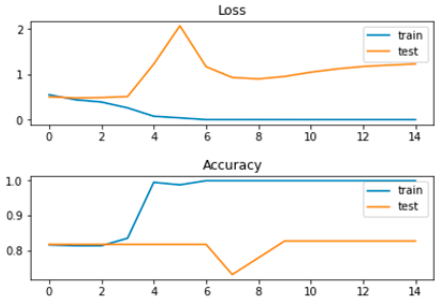
Readmisión 60 días - Ciclo Temporal 16




```
Matriz de Confusión:

[[149   0]
 [ 26   1]]
```

Readmisión 90 días - Ciclo Temporal 27




```
Matriz de Confusión:

[[85   0]
 [18   1]]
```

Ilustración 16. Comparativa del detalle de los resultados más óptimos para cada tipo de readmisión

Según el modelo utilizado en esta investigación, basado en estructuras de datos con secuencias temporales, el tipo de readmisión más equilibrado es el de 60 días, pero también se obtiene buenos resultados para los otros tipos de readmisión. En el caso de la readmisión a 7 días, sería aconsejable disponer de un número mayor de muestras, observando el comportamiento de todas sus métricas para confirmar los resultados. Otra ventaja de este modelo es el poder seleccionar el tipo concreto de readmisión que se desea analizar, gracias al algoritmo de selección de readmisión que genera el conjunto de datos para la readmisión elegida, sin necesidad de tener que localizar uno específico por otros medios. También cabe destacar que este modelo mejora los resultados obtenidos por otros algoritmos más comúnmente utilizados, como es el algoritmo *Logistic Regression*. Tal y como afirman Whitlock et al. (2010), utilizan este algoritmo para estudiar readmisiones a 30 días y en el cual obtienen una precisión de 83%, frente al 90,17% obtenido en este caso de estudio. Los buenos resultados obtenidos también son confirmados analizando otros posibles algoritmos que permitan comprobar su validez, tal y como se puede ver en la Ilustración 14.

6 RESULTADOS

Con la planificación planteada y el recorrido seguido en esta investigación se consigue alcanzar el objetivo principal propuesto al inicio de este documento. Se logra estudiar y aplicar las últimas tecnologías en aprendizaje automático e inteligencia artificial sobre modelos clasificatorios. Adicionalmente, también se aplican modelos de regresión como una derivada de este objetivo principal, para de esta forma poder demostrar y comparar el funcionamiento de los modelos clasificatorios, con lo que se amplía el alcance conseguido por esta investigación. Utilizando los modelos de regresión, también se deben utilizar nuevas técnicas de evaluación del modelo que permitan realizar una comparativa global, ya que no se puede realizar una comparación directa entre ambos modelos. Al utilizar las tecnologías más recientes y tras comprobar los resultados, se consiguen mejorar los resultados de forma adicional, lo que permite validar que se ha alcanzado el objetivo principal de la investigación.

Tras culminar todos los casos de estudio, se consigue llegar al objetivo principal. En concreto se alcanza con los casos de estudio 2 y 3 (secciones 5.3 y 5.4), ya que abordan las tecnologías más recientes sobre aprendizaje automático y se aplican técnicas de inteligencia artificial, utilizando las técnicas más modernas sobre selección de características, sobre optimizadores de parámetros para aplicar a cada algoritmo, sobre selección de la mejor tecnología para el conjunto de datos disponible, sobre conversión de variables para adaptarlas a cada algoritmo, sobre técnicas de medición y normalización, y sobre redes neuronales que permiten aprender a partir de nuevos datos y ejecuciones.

El siguiente cuadro muestra un resumen del objetivo principal planteado y el momento en el que se consigue alcanzar. También muestra los objetivos secundarios, bajo que acciones se consigue su cumplimiento y en qué secciones de la investigación son logrados.

Objetivo Principal
Aplicar últimas tecnologías en aprendizaje automático e inteligencia artificial sobre modelos clasificatorios y de esta forma mejorar los resultados previos
Alcance
El objetivo principal se consigue tras realizar los casos de uso 2 y 3 (Días de estancia y Readmisión hospitalaria), ya que se utilizan las últimas tecnologías sobre aprendizaje automático y se aplican técnicas de inteligencia artificial, utilizando no solo modelos clasificatorios sino también de regresión, consiguiendo mejorar los resultados de otros estudios

Objetivo secundario	Acciones para su cumplimiento	Sección donde se obtiene
1. Extraer conocimiento a partir de los datos de salud genéricos, obtenidos de un sistema hospitalario mediante la ampliación del ámbito de investigación	Encontrar patrones de comportamiento a partir de resultados y datos globales. Generar una nueva característica partiendo de las ya existentes, la cual se utiliza para estudiar un ámbito nuevo (días de estancia)	Caso de uso 2 (sección 5.3)
2. Mejorar la visión global del entorno mediante análisis de ámbito más amplio, de forma que permita aumentar su utilidad y poder aplicarlo a otras unidades de aplicación como la financiera	Realizar un estudio global sobre un hospital completo, al contrario que los estudios analizados, que estudian departamentos concretos. De esta forma se descubren nuevos ámbitos de aplicación	Caso de uso 2 (sección 5.3)
3. Aplicar las tecnologías más recientes de Inteligencia Artificial y aprendizaje automático a las bases de datos de salud, buscando una mejor precisión y eficiencia en las predicciones	Se aplican tecnologías sobre algoritmos existentes (SelectKBest, etc.). También se aplican las últimas tecnologías aparecidas sobre redes neuronales, incluso combinando varias de ellas	Casos de uso 2 (sección 5.3) y 3 (sección 5.4)
4. Creación de algoritmos específicos sobre la detección de diagnósticos relacionados con un tipo de patología específica y sobre la detección de un reingreso hospitalario en función del rango de días que se defina	Se crean nuevos algoritmos adaptados a las necesidades particulares de la investigación. En concreto se realiza para poder abordar y preparar el entorno, para así aplicar el modelo predictivo estudiado	Caso de uso 3 (sección 5.4)

Objetivo secundario	Acciones para su cumplimiento	Sección donde se obtiene
5. Obtener los mejores resultados predictivos posibles aplicando redes neuronales convolucionales LSTM y técnicas sobre selección de características correlacionadas, optimizadores bayesianos de los parámetros de las tecnologías y selector automático de tecnología de aprendizaje automático sobre los mismos modelos, de modo que se pueda elegir el mejor método para cada caso particular	Se utiliza una red neuronal convolucional, la cual se optimiza para utilizar los parámetros que obtengan los resultados más óptimos, utilizando otras tecnologías existentes. Estos métodos de apoyo a la optimización también se aplican sobre otras técnicas de aprendizaje automático.	Caso de uso 3 (sección 5.4 - mejora de resultados y optimización). Caso de uso 2 (sección 5.3 - uso de técnicas de optimización)

Tabla 11. Cuadro de objetivos de la investigación

Las distintas fases que se han seguido a lo largo de la investigación son necesarias realizarlas en ese orden para poder cumplir los objetivos propuestos. Como podemos comprobar en el cuadro de la Tabla 11, en el primer caso de estudio no se alcanzan ninguno de los objetivos secundarios que se habían fijado, pero sin este caso de estudio no se podrían haber realizado los siguientes casos de estudio ni alcanzar el objetivo principal o los objetivos secundarios.

El objetivo secundario 1 es alcanzado al obtener nuevo conocimiento sobre el comportamiento de determinados sectores de los pacientes al comprobar que, para determinadas patologías en los departamentos de Oftalmología, Obstetricia, Cardiología y Urgencias, se puede mejorar la eficacia al predecir el número de días que van a permanecer ingresados en el hospital. Por consiguiente, también permite conocer los departamentos que son más adecuados para optimizar los recursos disponibles, consiguiendo mejorar la calidad de atención al paciente y consiguiendo utilizar estos recursos de forma más eficiente. Este objetivo secundario también se logra ampliando el ámbito de investigación mediante la creación de una nueva característica que calcula el tiempo de permanencia de los pacientes, lo que da pie al estudio del segundo caso de estudio sobre los días de estancia.

En el caso del objetivo secundario 2, se alcanza al innovar realizando una investigación que no se había realizado con anterioridad, estudiando la característica de los días de estancia, ya estudiada por otros investigadores, pero a nivel hospitalario en lugar de sobre una patología y/o un departamento particular. De esta forma se consigue tener una visión global de la característica y también se logra ampliar su utilidad.

El objetivo secundario 3 logra alcanzarse, además de mediante la utilización de tecnologías ya existentes con las últimas actualizaciones y mejoras, se alcanza con el uso de otras nuevas que permiten seleccionar las mejores características, los mejores parámetros o la mejor selección de datos de entrenamiento y de prueba, sin que se produzca estratificación de los datos y por medio de validación cruzada para garantizar que son independientes. También se utilizan técnicas recientes, ya existentes, de aprendizaje profundo, combinándolas entre sí para mejorar los resultados, tal y como se aplica en el tercer caso de uso (sección 5.4).

Al objetivo secundario 4 se llega como consecuencia de las propias necesidades del tercer caso de uso, el cual necesita preparar los datos para poder obtener un conjunto de datos que permita aplicar el modelo que se va a utilizar en este caso de uso. Se alcanza creando dos algoritmos específicos adaptados a las necesidades del modelo elegido. El primer algoritmo detecta y marca las patologías del conjunto de datos que están relacionadas con enfermedades de digestivo y hepatobiliares, de forma que se obtiene un nuevo conjunto de datos específico sobre este tipo de patologías. Posteriormente, se crea otro algoritmo que trabaja sobre este último conjunto de datos generado en el algoritmo anterior, para así poder identificar y marcar los distintos tipos de readmisión que se pueden dar. Se elige el tipo de readmisión que se quiere identificar en función del número de días necesario para considerarse como readmisión y que se introduce como parámetro de entrada. De esta forma se elige el tipo de readmisión en función del número de días elegido, cuyo producto de salida será un nuevo conjunto de datos específico, que será utilizado como el conjunto de datos definitivo para el modelo del tercer caso de uso.

El último objetivo secundario se logra al mejorar los resultados obtenidos por otras investigaciones previas. Aunque siempre se busca obtener los mejores resultados y superar a los ya existentes, no siempre se puede conseguir. Por ejemplo, no siempre se dan las circunstancias óptimas, como puede ser tener un conjunto de datos cuyo comportamiento se adapta a una función matemática concreta, o un conjunto de datos suficientemente grande o con la suficiente calidad como para detectar un patrón de comportamiento. En el segundo caso de uso no se logra mejorar los resultados, pero si se consigue en el tercer caso de uso al utilizar redes neuronales convolucionales LSTM, tratándolas como secuencias temporales. Éstas obtendrán mejores resultados aprendiendo sobre los datos existentes y con los datos nuevos que se introduzcan a la red. Esta mejora no solo se consigue aplicando esta técnica, sino que también se logra buscando la parametrización más óptima de los parámetros involucrados en la configuración de la red, aplicando técnicas de selección de características correlacionadas, utilizando optimizadores bayesianos de los parámetros de las tecnologías y con el uso de un selector automático de tecnología de aprendizaje automático, técnicas que también son utilizadas en el segundo caso de uso.

Los resultados que se obtienen en esta investigación están relacionados y estructurados de acuerdo con el seguimiento y desarrollo de los distintos casos de estudios tratados. El primer caso de estudio sobre el análisis en profundidad de los datos (sección 5.2), es necesario para toda investigación realizada sobre aprendizaje automático y es la base necesaria para abordar los siguientes casos de estudio que se han planteado. Este

análisis, además de aportar el conocimiento necesario sobre el área de estudio que se analiza, permite conocer y elaborar las líneas de investigación que se pueden seguir con el conjunto de datos disponible. Esta parte analítica utiliza diferentes técnicas y gráficas para poder conocer la distribución de la información. También es necesario conocer el significado de la unidad de negocio que implican las características que se tienen en el conjunto de datos. Todo esto implica dedicar el tiempo y esfuerzo que sean necesarios hasta comprender el significado del conjunto de datos disponible, ya que se pueden obtener buenos resultados predictivos sobre una determinada característica, pero carecer de relevancia. Como consecuencia de este análisis de los datos, se consigue extraer información para realizar un estudio de los dos casos de estudio siguientes (secciones 5.3 y 5.4). Se estudian los días de estancia de un paciente (sección 5.3) y los reingresos hospitalarios (sección 5.4), no por ser las únicas posibilidades, sino también por su relevancia en los sistemas de salud. También podría haberse realizado un seguimiento de alguna patología concreta y haber realizado un nuevo caso de estudio, pero es probable que no se hubiesen tenido suficientes datos o podría no haberse encontrado ningún patrón que permitiese obtener buenos resultados predictivos.

Hay que tener presente, que dependiendo de la calidad los datos de los que se disponga para realizar las predicciones, los resultados pueden variar significativamente. También, dependiendo de la precisión que se quiera tener en las predicciones, se puede mejorar la precisión (*accuracy*) obtenida en los algoritmos. Por ejemplo, si se necesita predecir el número de días que van a transcurrir hasta la próxima visita de un paciente a su médico habitual y no es necesario saber este número de días de forma exacta, puede ser suficiente con conocer el rango de días que van a transcurrir hasta la próxima visita, como pueden ser rangos de 5 en 5 días. De esta forma, si la calidad del dato es insuficiente como para predecir el número de días exacto, puede ser suficiente si se predice un rango determinado de días. En el segundo caso de estudio de esta investigación (sección 5.3), aunque los resultados no hayan obtenido tasas altas de predicción, el objetivo buscado era obtener un listado ordenado de los mejores departamentos de un hospital para poder predecir los días de estancia y obtener así los mejores departamentos a los que dedicar los mayores esfuerzos, para así obtener el mejor rendimiento con los recursos disponibles. Como se puede observar en este caso, aunque la calidad del dato no es suficiente, para obtener altas tasas de predicción con las mediciones exactas, si se pudo aplicar las técnicas predictivas para conseguir el objetivo buscado.

El segundo y tercer caso de estudio (secciones 5.3 y 5.4) tratan sobre dos áreas de conocimiento de salud diferentes, aunque en ambos se aplican tecnologías de aprendizaje automático. En el segundo caso de estudio se utiliza el conjunto de datos preparado en el primer caso de estudio, aunque extrayendo el conjunto de datos específico para cada departamento del hospital en las predicciones sobre el número de días de estancia. En el tercer caso de estudio, aunque también se parte del conjunto de datos preparado en el primer caso de estudio, se hace necesario una mayor adaptación y preparación de los datos para poder tratarlo con las redes neuronales. En este caso, además de elaborar dos algoritmos que extraigan la información necesaria para adaptarla a la patología y al tipo específico de readmisión, también se hace necesario utilizar técnicas como la de normalización. Esta técnica se hace necesaria para reducir el

error de generalización producido por el propio algoritmo, el cual es detectado en la fase de entrenamiento, ya que reducía su efectividad. También se hace necesario preparar los datos para trabajar sobre matrices de tres dimensiones, cuya tercera dimensión refleja el número de pasos temporales que contiene cada secuencia, y las otras dos dimensiones reflejan el número de características y el número total de muestras. La elaboración del conjunto de datos es más compleja en este tercer caso de uso. Aunque en el segundo caso no se logra mejorar la tasa predictiva del modelo, como si realiza en el tercer caso, en ambos se logran innovar aportando hallazgos y elementos no investigados hasta la fecha.

En el segundo caso de estudio (sección 5.3) se logra aportar nuevo conocimiento global gracias al estudio realizado sobre los días de estancia de los pacientes, no solo en un departamento completo o de una patología particular, sino sobre todos los departamentos de un hospital, donde, además, no se había realizado en investigaciones previas. También se aporta conocimiento nuevo gracias a los patrones de comportamiento encontrados en el área de estudio sobre los días de estancia, en los departamentos de Oftalmología, Obstetricia, Urgencias y Cardiología. Aunque estos casos son particulares del hospital estudiado, pueden extrapolarse a otros hospitales, teniendo en cuenta los hallazgos encontrados sobre rasgos específicos de los pacientes y sobre las patologías concretas. En el tercer caso de estudio también se logra aportar nuevo conocimiento al combinar las técnicas más recientes de aprendizaje profundo sobre redes neuronales con las secuencias temporales, encontrando que existe una relación y logrando mejorar los resultados predictivos realizados por investigaciones previas. De esta forma, también se logra una aportación que no había sido realizada previamente, utilizando el modelo de datos con la técnica de redes neuronales como una secuencia temporal.

Dentro del aprendizaje supervisado utilizado en esta investigación, los resultados obtenidos son diferentes dependiendo del tipo de tarea que se aplique. Si se aplica una tarea de clasificación se clasifica un conjunto finito de etiquetas, en el que se busca obtener la pertenencia a una etiqueta concreta. Si se aplica una tarea de regresión, se busca predecir una variable de salida continua, donde la relevancia consiste en poder caracterizar la variación que se produce en la variable dependiente que se predice con respecto al resto de características independientes. En estos casos, las métricas a utilizar son diferentes y la forma de evaluar los resultados también son distintas. En concreto, en el tercer caso de estudio de esta investigación (sección 5.4) se realiza, como línea principal de investigación, una clasificación mediante una red neuronal convolucional LSTM, utilizando diferentes secuencias temporales. Como no existe una investigación similar y en este tipo de redes neuronales es más común utilizarlas en regresión, se aplica el modelo de regresión al mismo conjunto de datos para poder compararlas. También se utiliza una red neuronal similar con las mismas secuencias temporales. Del mismo modo, se utiliza un tercer algoritmo (Árbol de Decisión) para comprobar si existe mejora en los resultados del algoritmo principal de la investigación. La comparación del algoritmo principal mediante clasificación (red neuronal) y el algoritmo con Árbol de Decisión fue posible porque ambos utilizan clasificación y las métricas son las mismas, pero en el caso de la comparativa con el algoritmo de regresión, al utilizar métricas diferentes, la comparación es relativa, ya que en la clasificación se pueden estudiar

métricas sobre la predicción, como la exactitud, precisión, etc. En la regresión se mide la desviación producida en las medidas predichas y el error cometido. Por ello, tras realizar las predicciones sobre los resultados de la regresión, se hizo una conversión de estos resultados, clasificando la pertenencia a los diferentes grupos utilizados en la clasificación, para así tener una medida relativa de los resultados que permita realizar una comparativa aproximada. Como los resultados fueron bastante inferiores (ver sección 5.4), se puede concluir que el modelo clasificatorio mediante redes neuronales obtiene mejores resultados que el modelo de regresión, y también obtiene mejores resultados que el tercer algoritmo utilizado con Árbol de Decisión. Finalmente, en este caso de estudio, se logra implementar un modelo que vaya aprendiendo con el tiempo y con los nuevos datos que se vayan utilizando, lo que conseguirá mejorar y optimizar los resultados, adaptándose a los posibles nuevos cambios de comportamiento que se puedan producir, con lo que el modelo no se queda obsoleto, sino que se adapta.

7 CONCLUSIONES

El desarrollo de esta investigación y la elaboración de esta tesis permiten obtener las conclusiones mostradas a continuación a partir de las distintas fases que se han ido desarrollando.

El análisis del conjunto de datos inicial ha permitido tener un profundo conocimiento dentro del área de salud y del comportamiento de las entidades participantes en las áreas hospitalarias analizadas, gracias al estudio de la distribución de los datos y a la búsqueda de información a través del estado del arte y a través de la consulta de profesionales del sector. Todo este estudio permite la comprensión de la información disponible, algo que es necesario para avanzar hacia otras áreas de estudio, mostrando la necesidad de aplicar esta metodología utilizada para abordar estudios sobre análisis de datos, como el que se realiza en esta investigación sobre aprendizaje automático. Esta primera fase ha permitido conocer la posibilidad de estudiar y avanzar en el estudio de dos temáticas muy utilizadas, como son la readmisión hospitalaria y la duración de las estancias de los pacientes. También permite evidenciar la necesidad de disponer de apoyo de profesionales del sector para poder estudiar áreas más especializadas, como son las patologías digestivas y sus relaciones con otras patologías, pudiendo así demostrar la vinculación con la patología analizada.

Tras las fases previas, se consigue el objetivo principal de la investigación aplicando las últimas técnicas disponibles en aprendizaje automático e inteligencia artificial sobre aprendizaje supervisado, creando dos casos de estudio que las aplican (secciones 5.3 y 5.4). En el caso de estudio sobre días de estancia, se aplican estas técnicas con éxito mediante la optimización automática de los parámetros y por medio de la elección de la mejor tecnología, consiguiendo además extraer nuevo conocimiento sobre patologías y departamentos hospitalarios. En el caso de estudio sobre reingreso hospitalario se aplica una reciente tecnología mediante redes neuronales convolucionales, en la que se utiliza una comparativa propia con otras tecnologías similares, consiguiendo mostrar la mejora en precisión y eficiencia. También se crean dos algoritmos propios, uno para la detección de diagnósticos relacionados con el diagnóstico elegido y otro para la detección de un reingreso hospitalario según el rango de días que se quiera utilizar. Estos algoritmos son creados para poder preparar los conjuntos de datos que se van a utilizar con esta tecnología. Con el desarrollo y elaboración de los dos casos de estudio comentados anteriormente, se logran también los objetivos secundarios derivados del principal que se habían definido, los cuales se definieron como la aplicación de las tecnologías más recientes, la extracción de nuevo conocimiento, la creación de los algoritmos específicos necesarios para la elaboración de los modelos y la mejora de los resultados predictivos en la mayor medida posible mediante las tecnologías necesarias.

Adicionalmente, tras los resultados obtenidos a lo largo de esta investigación, en base a los algoritmos utilizados, se obtienen varias conclusiones sobre los métodos y algoritmos utilizados en los distintos casos de uso.

- *K Nearest Neighbours*. Se trata de un clasificador simple y eficiente en el que se obtienen buenos resultados gracias a que los datos tratados tienen poco ruido y a que las características no son independientes unas de otras.
- *Logistic Regression*. Este algoritmo calcula la probabilidad de que la combinación de las características de entrada pertenezca a una de las clases de salida, es decir, a uno de los días en este caso (sección 5.3). Obtiene buenos resultados gracias a que la salida de datos se puede expresar como una función matemática que es una combinación lineal de las características de entrada. El hecho de obtener buenos resultados también indica que se dispone de la cantidad suficiente de datos y que las interrelaciones entre las características no son muy complejas.
- *Support Vector Machines*. Se trata de un conjunto de algoritmos que tratan de construir planos para maximizar el límite entre las distintas clases existentes. En general obtiene buenos resultados salvo para el caso en el que se usa un núcleo polinomial como función del algoritmo, ya que el comportamiento de los datos no se asemeja a funciones de grado elevado. Aunque el tiempo de procesamiento es elevado, se logran buenos resultados porque los valores atípicos tienen poco impacto en los resultados y porque las clases de la variable de salida se pueden separar fácilmente.
- *Naïve Bayes*. Este algoritmo es similar al conjunto de algoritmos anterior y se suele agrupar conjuntamente. La diferencia principal es que considera que las características de entrada son independientes y asume que los datos se comportan siguiendo una distribución gaussiana. Obtiene buenos resultados porque las características utilizadas no tienen un alto grado de dependencia y porque funciona bien con tareas de clasificación múltiple, como el conjunto de datos disponible.
- *Decision Trees*. Es un sistema efectivo y eficiente que se basa en reglas de decisión simples deducidas a partir de las características de entrada. Con este algoritmo se han obtenido buenos resultados gracias a que se adaptan bien a cualquier tipo de datos y gracias a la optimización y ajuste de sus parámetros, lo que ha hecho que se evite la tendencia que tienen a sobre-ajustar o generalizar.
- *Random Forest*. Este algoritmo también obtiene buenos resultados ya que utiliza múltiples algoritmos tipo *Decision Tree* de forma aleatoria y con bajo nivel de profundidad. Se obtienen buenas tasas predictivas ya que son capaces de manejar gran cantidad de características de entrada sin excluir ninguna, son capaces de descubrir las características más relevantes y, al igual que el modelo anterior, con un buen ajuste de sus parámetros, evita la tendencia que tienen a sobre-ajustar.
- *Redes Neuronales*. Estas técnicas son funciones que aprenden la salida de datos esperada para las características de entrada a partir de datos de entrenamiento. Se obtienen buenos resultados gracias a la correcta configuración de sus parámetros y a su capacidad para aprender a partir de los datos disponibles. En general son buena opción para la mayoría de los conjuntos de datos, aunque sean datos complejos y siempre que se tengan datos claros e informativos.

Como conclusión del estudio realizado, la aplicación de análisis de aprendizaje automático y de inteligencia artificial permite conocer el comportamiento actual de las áreas de conocimiento analizadas y exponer una proyección del comportamiento futuro con diferentes grados de confianza, en función de la bondad que ofrecen los modelos utilizados. Estas técnicas permiten incluso simular el comportamiento humano aprendiendo de sí mismos y de la nueva información que les llegue en forma de datos, lo que las hace tener amplias posibilidades de aplicación para multitud de áreas de conocimiento.

8 TRABAJO FUTURO

El objeto de estudio de esta investigación trataba sobre el análisis y la aplicación de las tecnologías de aprendizaje automático e inteligencia artificial, enfocado en un sistema de datos real dentro del ramo de salud. Los modelos aplicados y las líneas de investigación seguidas dan lugar a enumerar las siguientes líneas de trabajo futuro consideradas importantes y apropiadas para evolucionar las aportaciones realizadas:

- Un trabajo futuro a realizar dentro de las investigaciones realizadas es el estudio en mayor profundidad de tecnologías de regresión aplicado a las mismas áreas utilizadas en esta investigación. El trabajo realizado se ha enfocado principalmente sobre tecnologías de clasificación que permiten responder a cuestiones más concretas, donde se han utilizado técnicas de regresión para comparar los resultados obtenidos mediante clasificación. El mismo trabajo puede aplicarse mediante regresión modificando el modelo, permitiendo un estudio más amplio de forma que se analice el comportamiento desde una perspectiva orientada al resultado numérico en lugar de comprobar a qué grupo pertenece o de responder a una pregunta concreta.

- Dado que el conjunto de datos utilizado en esta investigación son datos comunes que se recogen en los hospitales de manera habitual, una línea de trabajo futuro a seguir sería la obtención de más datos de diferentes años y/o de diferentes hospitales para corroborar o mejorar los resultados. De esta forma, se pueden repetir los análisis realizados en los siguientes casos:

    - En aquellos métodos de aprendizaje supervisado del caso de uso sobre el número de días de estancia en los que no se obtuvieron resultados concluyentes por falta de muestras.
    - También, con el caso de uso sobre el reingreso hospitalario, gracias a los algoritmos de detección de patologías relacionadas y de detección de tipo de reingreso, se pueden preparar los datos para su estudio de forma que pueda mejorar los resultados obtenidos y seguir aprendiendo con los nuevos datos, adaptándose a posibles cambios de tendencia que puedan aportar las nuevas muestras.

- Otra línea de investigación a seguir sería complementar la investigación realizada mediante la aplicación de aprendizaje no supervisado, en lugar del aprendizaje supervisado aquí realizado. Al utilizar aprendizaje no supervisado, no se necesitan conocer las respuestas o los resultados con anterioridad para así predecir un nuevo resultado. Esta línea a seguir permite ver la distribución de los datos y definirlos, explicando el comportamiento de las características de entrada del modelo de datos, intentado descubrir patrones de comportamiento.

REFERENCIAS

Mamani Choque, G. (2020). *Modelo predictivo de riesgo de morosidad para crédito bancario a partir de datos simulados de la Caja Rural de Ahorro y Crédito Los Andes – Puno* (Tesis doctoral). Universidad Nacional del Altiplano.

Mozaffari S., Al-Jarrah O. Y., Dianati M., Jennings P. y Mouzakitis A. (2020). *Deep Learning-based Vehicle Behaviour Prediction for Autonomous Driving Applications: a Review*. arXiv:1912.11676v2 [cs.CV], pp. 1-15

Kalafi E. Y., Nor N. A. M., Taib N. A., Ganggayah M. D., Town C. y Dhillon S. K. (2019). *Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data*. Folia Biol (Praha), PMID: 32362304, vol. 65 (5-6), pp. 212-220

Artetxe A., Beristain A. y Graña M. (2018). *Predictive models for hospital readmission risk: A systematic review of methods*, Computer Methods and Programs in Biomedicine, Science Direct, Elsevier, vol. 164, pp. 49-64

Iglesias E., García A., Puig P. y Benzaqué I. (2020). *Inteligencia Artificial: la gran oportunidad del siglo XXI. Documento de reflexión y propuesta de actuación*. Biblioteca Felipe Herrera, Banco Interamericano de Desarrollo, 1300 New York, N.W. (Washington D.C.). IDB-MG-904

Shinde P. P. y Shah S. (2019). *A Review of Machine Learning and Deep Learning Applications*, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), doi: 10.1109/ICCUBEA.2018.8697857, pp. 1-6

Dash S., Shakyawar S. K., Sharma M. y Kaushik S. (2019). *Big Data in healthcare: management, analysis and future prospects*. J Big Data 6, 54. https://doi.org/10.1186/s40537-019-0217-0

Whitlock T. L., Tignor A., Webster E. M., Repas K., Conwell D., Banks P. A. y Wu B. U. (2010). *A scoring system to predict readmission of patients with acute pancreatitis to the hospital within thirty days of discharge*. Clin. Gastroenterol. Hepatol., vol. 9 (2), pp. 175-180

Chang C.-L. y Lu P.-Y. (2016). *The Study on Evaluating Length of Hospital Stay for Myomectomy*. Vol. 5(59), ISSN: 2251-8843.

Rouzbahman M., Jovicic A. y Chignell M. (2017). *Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU?*. IEEE Journal of Biomedical and Health Informatics, vol. 21(3), pp. 851-858.

Lowell W., Davis G., Lajousky W., Stieffel S., Davis G., Breau M. y Shirazi H. (1997). *A field trial using Artificial Neural Networks to predict psychiatric inpatient Length- of-stay*. Neural Computing & Applications, vol. 5(3), pp. 184-193.

Aghajani S. y Kargari M. (2016). *Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department*. Hospital Practices Res., vol. 1, no. 2, pp. 51–56
Tanuja S., Acharya U. D. y Shailesh K. R. (2011). *Comparison of different data mining techniques to predict hospital length of stay*. Journal of Pharmaceutical and Biomedical Sciences, ISSN NO- 2230 – 7885, vol. 7 (15)

Yang C.-S., Wei C.-P., Yuan C.-C. y Schoung J.-Y. (2010). *Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages*. Ed. Elsevier, Decision Support Systems, vol. 50(1), pp. 325–335.

Studer S., Bui T. B., Drescher C., Hanuschkin A., Winkler L., Peters S. y Müller K.-R. (2021). *Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology*. Mach. Learn. Knowl. Extr., https://doi.org/10.3390/make3020020, vol. 3, pp. 392-413

Saltz J. y Hotz N. (2022). *What is CRISP DM?*. Data Science Process Alliance, CRISP DM. https://www.datascience-pm.com/crisp-dm-2/

Bischel S. H. y Salmerón A. (2013). *El método de la entropía cruzada. Algunas aplicaciones*. TFM Escuela Politécnica Superior y Facultad de Ciencias Experimentales, Universidad de Almería, http://repositorio.ual.es/bitstream/handle/10835/3322/Trabajo.pdf?sequence=1&isAllowed=y

Foody G. M. (1995). *Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data*. ISPRS Journal of Photogrammetry and Remote Sensing, Elsevier, https://doi.org/10.1016/0924-2716(95)90116-V, vol. 50 (5), pp. 2-12

Steven A. F. (2009). *The common patterns of Nature*. J Evol Biol., doi: 10.1111/j.1420-9101.2009.01775.x, vol. 22 (8), pp. 1563-1585

Gupta S. y Gupta A. (2019). *Dealing with noise problem in Machine Learning Data-sets: A systematic review*. Procedia Computer Science, Elsevier B. V., vol 161, pp. 466-474

Mishra A. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science, Canada. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Bergstra J., Yamins D. y Cox D. D. (2013). *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures*. To appear in Proc. of the 30th International Conference on Machine Learning (ICML 2013), vol. 28. http://hyperopt.github.io/hyperopt/#hyperopt-distributed-asynchronous-hyper-parameter-optimization

Feurer M., Klein A., Eggensperger K., Springenberg J., Blum M. y Hutter F. (2015). *Efficient and Robust Automated Machine Learning*. Advances in Neural Information Processing Systems 28, pp. 2962-2970

Feurer M., Eggensperger K., Falkner S., Lindauer M. y Hutter F. (2020). *Auto-Sklearn 2.0: The Next Generation*. arXiv:2007.04074 [cs.LG], vol. 1, pp. 1-18

Agencia Estatal Boletín Oficial del Estado (2018). *Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales*. https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673

Consejería de Sanidad y Consumo (2004). *Manual del CMBD (Conjunto Mínimo Básico de Datos)*, Comunidad de Madrid. http://www.madrid.org/cs/BlobServer?blobcol=urldata&blobtable=MungoBlobs&blobheadervalue1=filename%3DCMBD.pdf&blobkey=id&blobheadername1=Content-Disposition&blobwhere=1119153506764&blobheader=application%2Fpdf

Ministerio de Sanidad, Consumo y Bienestar Social (2016). *ICMBD: Indicadores y Ejes de Análisis del CMBD*. Proyecto de Implantación y Soporte del Modelo de Indicadores y Ejes de Análisis de los Datos del Conjunto Mínimo Básico de Datos de Hospitalización del SNS (2001-2016). https://icmbd.sanidad.gob.es/icmbd/login-success.do

Ministerio de Sanidad, Servicios Sociales e Igualdad (2014). *CIE-9-MC Clasificación Internacional de Enfermedades 9ª Edición*. Informes y estadísticas sanitarias 2013. https://www.mscbs.gob.es/estadEstudios/estadisticas/docs/CIE9MC_2014_def_accesible.pdf

Comunidad de Madrid (2022). *SERMAS - Servicio Madrileño de Salud de la Comunidad de Madrid*. https://www.comunidad.madrid/servicios/salud

Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830. https://scikit-learn.org/stable/index.html

Pandas (2022, Febrero). Pandas documentation. Versión 1.4.1. https://pandas.pydata.org/docs/index.html

Ministerio de Sanidad y Consumo (1999). *Análisis y desarrollo de los Grupos de Diagnósticos Relacionados (GDR) en el sistema Nacional de Salud*. Centro de Publicaciones del Ministerio de Sanidad y Consumo. https://www.mscbs.gob.es/estadEstudios/estadisticas/docs/analisis.pdf

Zhibin L. V., Ding H., Wang L. y Zou Q. (2020). *A convolutional neural network using dinucleotid one_hot encoder for identifying DA N6-Methyladenine sites in the rice genome*. Neurocomputing, Elsevier B. V., vol. 422 (2021), pp. 214-221

Wu T. H., Pang G. K. y Kwong E. W. (2015). *Predicting Systolic Blood Pressure Using Machine Learning*. 7th International Conference on Information and Automation for Sustainability, doi: 10.1109/ICIAFS.2014.7069529, pp. 1-6

Howell S., Coory M., Martin J. y Duckett S. (2009). *Using routine inpatient data to identify patients at risk of hospital readmission*. BioMed Central Health Services Research, Springer Nature, vol. 9 (96), pp. 1-9

Samuel A. L. (1959). *Some Studies in Machine Learning Using the Game of Checkers*. IBM Journal of Research and Development, doi: 10.1147/rd.33.0210, vol. 3 (3), pp. 210-229

Arwinder D. y Ashima S. (2019). *Machine Learning in Healthcare Data Analysis: A Survey*. Journal of Biology and Today's World, doi: 10.15412/J.JBTW.01070206, vol. 8 (2), pp. 1-10

Ajay K., Rama S. y Arvind K. T. (2019). *Cancer survival analysis using machine learning*. International Conference on Sustainable Computing in Science, Technology & Management, Elsevier SSRN, pp. 657-662

Hachesu P. R., Ahmadi M., Alizadeh S. y Sadoughi F. (2013). *Use of data mining techniques to determine and predict length of stay of cardiac patients*. Healthcare Informat. Res., vol. 19, no. 2, pp. 121–129

Peiró S., Librero J. y Martínez A. B. (1996). *Factors associated to emergency hospital readmittance in digestive and hepatobiliary diseases*. Medicina Clínica, Valencia, Spain, vol. 107, num. 1, pp. 4-13

Morton A., Marzban E., Giannoulis G., Aparasu A. P. R. y Kakadiaris I. A. (2014). *A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients*. 13th International Conference on Machine Learning and Applications, pp. 428-431

Bergese I., Figerio S., Clari M., Castagno E., De Clemente A., Ponticelli E., Scavino E. y Berchialla P. (2019). *An Innovative Model to Predict Pediatric Emergency Department Return Visits*. Pediatric Emergency Care, vol. 35 (3), pp. 231-236

Chuang M.-T., Hu Y.-H., Tsai C.-F., Lo C.-L. y Lin W.-C. (2015). *The identification of prolonged length of stay for surgery patients*. Proc. IEEE Int. Conf. Syst., Man, Cybern., pp. 3000–3003

LaFaro R. J., Pothula S., Kubal K. P., Inchiosa M. E., Pothula V. M., Yuan S. C. y Inchiosa M. A. (2015). *Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables*. PLoS ONE, Public Library Sci., vol. 10, pp. 1–19

Futoma J., Morris J. y Lucas J. (2015). *A comparison of models for predicting early hospital readmissions*. Journal of Biomedical Informatics, Elsevier, Durham, USA, vol. 56, pp. 229-238

Manterola C., Astudillo P., Arias E., Claros N. y Grupo MINCIR (Metodología e Investigación en Cirugía) (2013). *Revisiones Sistemáticas de la literatura. Qué se debe saber acerca de ellas*. Elsevier, Cirugía Española, vol. 91 (3), pp. 149-155

Jamei M., Nisnevich A., Wetchler E., Sudat S. y Liu E. (2017). *Predicting all-cause risk of 30-day hospital readmission using artificial neural networks*. PLoS ONE, Canada, vol. 12 (7), pp. 1-14

Rattan D., Bhatia R. y Singh M. (2013). *Software Clone Detection: A Systematic Review*. Elsevier, Information and Software Technology, vol. 55, pp. 1165-1199

Panchami V. U. y Radhika N. (2014). *A Novel Approach for Predicting the Length of Hospital Stay With DBSCAN and Supervised Classification Algorithms*. In Proc. 5th Int. Conf. Appl. Digit. Inf Web Technol. (ICADIWT), pp. 207-212

Parag C. P. y Hitesh K. (2014). *Machine Learning Techniques for Predicting Hospital Length of Stay in Pennsylvania Federal and Specialty Hospitals*. Int. J. Comput. Sci. Appl., vol. 11(3), pp. 45-56

AlShehery M. Z., Duraisamy B., Zaidi A. R. Z., AlShehry N. F., Zaidi F. Z., Rababah A. A., Assiri A. A., AlGhamdi M. S., Al Mutair A. and Al-Omari A. (2020). *COVID-19 and palliative care services: Comparative patterns of inpatient, outpatient, and consultation services in a tertiary care center in riyadh*. Cureus, vol. 12, no. 12, pp. 1–11

Greysen S. R., Stijacic C. I., Auerbach A. D y Covinsky K. E. (2015). *Functional impairment and Hospital readmission in Medicare seniors*. JAMA Intern. Med., vol. 175 (4), pp. 559-565

Kiran R. P., Delaney C. P., Senagore A. J., Steel M., Garafalo T. y Fazio V. W. (2004). *Outcomes and prediction of hospital readmission after intestinal surgery*. American College of Surgeons, Elsevier Inc., vol. 198 (6), pp. 877-883

Kansagara D., Englander H., Salanitro A., Kagen D., Theobald C., Freeman M. y Kripalani S. (2011). *Risk prediction models for hospital readmission. A systematic review*. Journal American Medical Association, vol. 306 (15), pp. 1688-1698

George L. A., Martin B., Gupta N., Shastri N., Venu M. y Naik A. S. (2019). *Predicting 30-day readmission rate in inflammatory bowel disease patients: Performance of LACE index*. Crohn's & Colitis 360, Ed. Decision, Chicago, IL, vol 1 (1), pp. 1-7

ANEXO

Repositorios GitHub código Python

En este anexo se incluye el código Python de las principales herramientas y utilidades que son especialmente relevantes para su uso en la aplicación del aprendizaje automático. Todas ellas han sido utilizadas a lo largo de las investigaciones realizadas.

Se ha creado expresamente un contenedor de repositorios online en GitHub, de forma que se tenga acceso como referencia a las principales utilidades, a los algoritmos más representativos y a algunos notebooks relevantes de las investigaciones. El repositorio creado se encuentra en la ubicación:

https://github.com/JesusManuelPuentesG

Los repositorios que han sido creados son los que se enumeran a continuación:

- *ML_Python_Metricas*. Contiene fragmentos de código sobre las métricas utilizadas en aprendizaje Automático para Clasificación y Regresión. Las métricas contempladas son mostradas a continuación.
    - Clasificación
        - Matriz de confusión
        - Precisión (en inglés *Accuracy*)
        - Exactitud (en inglés *Precision*)
        - Sensibilidad (en inglés *Recall*)
        - F1 Score
        - ROC AUC
    - Regresión
        - Error cuadrático medio
        - Error absoluto medio
- *ML_Python_Optimizadores*. Código utilizado para seleccionar los parámetros más óptimos en un conjunto de datos. Los optimizadores utilizados son:
    - Optimizador Bayesiano (*Hyperopt*)
    - Optimizador exhaustivo por hiperparámetros (*GridSearchCV*)
- *ML_Python_Selector_Algoritmo_Optimo*. Código necesario para seleccionar el algoritmo de aprendizaje Automático más óptimo para realizar predicciones. La biblioteca utilizada se denomina *AutoSklearn*.
- *ML_Python_Normalizacion*. Código utilizado para Normalizar las características de un conjunto de datos. La biblioteca utilizada en la Normalización se denomina *MinMaxScaler*.
- *ML_Python_Conversor_Categorico_Numerico*. Código necesario para convertir variables categóricas a numéricas mediante la técnica *One-hot encoding*, realizada mediante la función *get_dummies* perteneciente a la biblioteca *pandas*.

- *ML_Python_Selector_Caracteristicas_Relevantes*. Código necesario para seleccionar las características más relevantes de un conjunto de datos según una función estadística. Esta función es realizada mediante la biblioteca *SelectKBest*.
- *ML_Python_Particionado_Datos*. Código Python con varias formas de particionar los datos con y sin estratificación. Se han recogido dos técnicas principalmente:
    - Particionamiento aleatorio sin estratificar (*Train_test_split*)
    - Particionamiento estratificado con validación cruzada (*StratifiedKFold*)
- *Salud_Algoritmos_Python*. Algoritmos relacionados con Salud desarrollados para investigar reingresos hospitalarios.
    - Algoritmo para la detección de patologías médicas
    - Algoritmo para el marcado de readmisiones hospitalarias
- *ML_Notebooks_Python*. Principales notebooks creados sobre desarrollos de Salud para Machine Learning, creados con Jupyter Notebook.
    - Extracción de precisiones al predecir los días de estancia de un conjunto de datos de un hospital para algoritmos Naïve Bayes y SVC con Kernels No Lineal y Lineal.
    - Extracción de precisiones y errores al predecir los días de estancia en un departamento de un hospital para algoritmos KNN, SVM con Kernel RBF, Árbol de Decisión y Random Forest, con optimizador Bayesiano.
    - Obtención de resultados de una Red Neuronal Convolucional LSTM para predecir el tipo de Readmisión seleccionada en un conjunto diferente de ventanas temporales

A continuación, se muestran los principales algoritmos en la investigación sobre readmisión hospitalaria.

Algoritmo para detectar patologías médicas

Algoritmo para detectar y marcar aquellos pacientes (uno por registro) que contienen alguna patología de digestivo entre sus posibles diagnósticos. Tras finalizar el conjunto de datos queda marcado con una nueva característica que indica si el paciente pertenece o no pertenece al grupo considerado como un reingreso de digestivo.

```
…
# Biblioteca necesaria
import pandas as pd

# Leemos el fichero CSV que contiene las Enfermedades digestivas
# relacionadas realizado por doctores en medicina. Este archivo
# contiene un listado de las enfermedades según la codificación
# internacional CIE9MC2014
enf = pd.read_csv('EnfermedadesDig_CIE9MC2014v2.csv')
…
# Funciones necesarias

# Función para comprobar si tiene “-“
def tiene_guion(entrada):
    texto = str(entrada)
    result = texto.find("-")
    if result < 0:
```

```
        return False
    else:
        return True

# Función para comprobar si tiene V
def tiene_V(entrada):
    texto = str(entrada)
    result = texto.find("V")
    if result < 0:
        return False
    else:
        return True

# Función para extraer los números Si tiene guión y SI tiene V (Ejem.:
'V44.1-V44.4')
def extraeNums_g_V(entrada):
    texto = str(entrada)
    num1 = float(texto[1:texto.find("-")])
    num2 = float(texto[texto.find("-")+2:len(texto)])
    return num1,num2

# Función para extraer los números Si tiene guion y NO tiene V (Ejem.:
'571.0-571.3')
def extraeNums_g(entrada):
    texto = str(entrada)
    num1 = float(texto[0:texto.find("-")])
    num2 = float(texto[texto.find("-")+1:len(texto)])
    return num1,num2

# Función para extraer el número Si NO tiene guión y SI tiene V
(Ejem.: 'V12.54')
def extraeNum_V(entrada):
    texto = str(entrada)
    num = float(texto[1:len(texto)])
    return num

# Función para saber si es un número
def es_numero(n):
    try:
        float(n)
    except ValueError:
        return False
    return True

# Función para saber si es nan (NaN)
def es_nan(n):
    var = str(n)
    if var == 'nan':
        return True
    else:
        return False

# Función para saber si NO es nan (NaN)
def noes_nan(n):
    var = str(n)
    if var == 'nan':
        return False
    else:
        return True

# Función para buscar si es de Digestivo un diagnóstico
def busca_digestivo(diagnostico, enfermedades_digestivo, dataset):
    for c in range(len(dataset)):
```

```
        for e in range(len(enfermedades_digestivo)):
            if noes_nan(dataset[diagnostico][c]):
                if dataset['DIGESTIVO'][c] == 0:
                    if tiene_guion(enfermedades_digestivo['id'][e]):
                        if tiene_V(enfermedades_digestivo['id'][e]):
                            if tiene_V(dataset[diagnostico][c]):
                                numc =
extraeNum_V(dataset[diagnostico][c])
                                inf, sup =
extraeNums_g_V(enfermedades_digestivo['id'][e])
                                if numc >= inf and numc <= sup:
                                    dataset['DIGESTIVO'][c] = 1
                                else:
                                    dataset['DIGESTIVO'][c] = 0
                            else:
                                dataset['DIGESTIVO'][c] = 0
                        else:
                            if tiene_V(dataset[diagnostico][c]):
                                dataset['DIGESTIVO'][c] = 0
                            else:
                                if es_numero(dataset[diagnostico][c]):
                                    inf, sup =
extraeNums_g(enfermedades_digestivo['id'][e])
                                    if float(dataset[diagnostico][c])
>= inf and float(dataset[diagnostico][c]) <= sup:
                                        dataset['DIGESTIVO'][c] = 1
                                    else:
                                        dataset['DIGESTIVO'][c] = 0
                                else:
                                    dataset['DIGESTIVO'][c] = 0
                    else:
                        if tiene_V(dataset[diagnostico][c]):
                            numc =
extraeNum_V(dataset[diagnostico][c])
                            if
tiene_V(enfermedades_digestivo['id'][e]):
                                nume =
extraeNum_V(enfermedades_digestivo['id'][e])
                                if numc == nume:
                                    dataset['DIGESTIVO'][c] = 1
                                else:
                                    dataset['DIGESTIVO'][c] = 0
                            else:
                                dataset['DIGESTIVO'][c] = 0
                        else:
                            if
tiene_V(enfermedades_digestivo['id'][e]):
                                dataset['DIGESTIVO'][c] = 0
                            else:
                                if es_numero(dataset[diagnostico][c]):
                                    if float(dataset[diagnostico][c])
== float(enfermedades_digestivo['id'][e]):
                                        dataset['DIGESTIVO'][c] = 1
                                    else:
                                        dataset['DIGESTIVO'][c] = 0
                                else:
                                    dataset['DIGESTIVO'][c] = 0

…
# Creamos una columna nueva (DIGESTIVO) con todos los valores a 0
DIGESTIVO = cmbd['GRDS'].astype('int')
DIGESTIVO.name = 'DIGESTIVO'
cmbd2 = cmbd.join(DIGESTIVO)
```

```
cmbd2['DIGESTIVO'] = 0
…
# Buscamos en todos los diagnósticos si existe alguna enfermedad
# relacionada con Digestivo en el conjunto de datos (cmbd2). En
# este caso tenemos 13 posibles diagnósticos (D1 a D13)
for i in range(13):
    busca_digestivo('D' + str(i+1), enf, cmbd2)
…
```

Algoritmo para marcar readmisiones hospitalarias

Algoritmo para marcar y crear un conjunto de datos con el tipo de readmisiones hospitalarias que se desee, según el parámetro que se utilice, para componer el conjunto de datos de un tipo determinado de Readmisión.

```
…
# Biblioteca necesaria
from datetime import date

# Funciones necesarias

# Función para calcular los días transcurrido entre 2 fechas

def dias_entre(d1, d2):
    """
    Función que calcula los días transcurrido entre 2 fechas

    Parámetros de entrada:
    d1: Fecha inicial
    d2: Fecha final

    Salida:
    Devuelve el número de días transcurrido entre las 2 fechas
    """

    return abs(d2 - d1).days

# Función que devuelve la fecha de ingreso de una historia clínica
# concreta a partir de un índice determinado y de un dataset que
# se le pasa como entrada

def busca_historia(historia_clinica, desde, dataset):
    """
    Función que busca una Historia Clínica a partir de un índice
determinado para buscar su fecha de ingreso en el hospital

    Parámetros de entrada:
        historia_clinica: Código de Historia Clínica
        desde: índice desde el que empezar a buscar la Historia
Clínica
        dataset: Conjunto de datos donde busca los datos

    Salida:
        Devuelve la fecha de ingreso en formato fecha
    """

    if desde <= len(dataset):
        for c in range(desde, len(dataset)):
            if dataset['HISTORIA_COD'][c] == historia_clinica:
```

```
                fecha = date(int(dataset['FECING'][c][0:4]),
int(dataset['FECING'][c][5:7]), int(dataset['FECING'][c][8:]))
                return fecha

# Función que modifica el dataset de entrada con el número de días
# transcurridos en las readmisiones

def calcula_dias_Readmision(desde, dataset):
    """
    Función que busca Readmisiones para calcular el número de días
    que ha pasado desde que ingresa hasta su próxima readmisión

    Parámetros de entrada:
        desde: índice desde el que empezar a buscar la historia
clínica
        dataset: Conjunto de datos donde se encuentra los casos
(registros) a estudiar

    Salida:
        El dataset de entrada será modificado (dataset['READMISION'])
indicando si el número de días pasado desde el ingreso hasta la
readmisión
    """

    if desde <= len(dataset):
        # Recorremos el dataset
        for c in range(desde, len(dataset)):
            # Guardamos la fecha de ingreso de la posición actual
            fecha1 = date(int(dataset['FECING'][c][0:4]),
int(dataset['FECING'][c][5:7]), int(dataset['FECING'][c][8:]))
            # Buscamos un reingreso a futuro y recuperamos su fecha de
ingreso
            fecha2 = busca_historia(dataset['HISTORIA_COD'][c], c+1,
dataset)
            if fecha2 is None:
                num_dias = 0
            else:
                num_dias = dias_entre(fecha1, fecha2)

            # Solo modificamos el dataset si encuentra una readmisión
            if num_dias > 0:
                dataset['READMISION'][c] = num_dias
…
# Ejecutamos algoritmo para rellenar la característica READMISION con
# los valores a futuro del número días readmisión
calcula_dias_Readmision(0, conjunto_de_datos)

# Ejemplo para Readmisiones a 7 días
# Creamos dataset con readmisiones de 7 o menos días
conjunto_de_datos_r7_pre = conjunto_de_datos[conjunto_de_datos
['READMISION'] < 8]

# Creamos dataset con readmisones mayores que 0, es decir, que
# sean readmisiones (0 = no hay readmisión)
conjunto_de_datos_r7 =
conjunto_de_datos_r7_pre[conjunto_de_datos_r7_pre['READMISION'] > 0]
…
```

Artículos utilizados para el análisis del estado del arte (sección 4.4)

Aghajani S. y Kargari M. (2016). *Determining Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department*. Hospital Practices and Res., vol. 1, pp. 51-56

Artetxe A., Beristain A. y Graña M. (2018). *Predictive models for hospital readmission risk: A systematic review of methods*. Computer Methods and Programs in Biomedicine, Science Direct, Elsevier, vol. 164, pp. 49-64

Awad A., Bader-El-Den M. y McNicholas J. (2017). *Patient length of stay and mortality prediction: A survey*. Health Serv Manage Res, vol. 30(2), pp. 105-120. doi: 10.1177/0951484817696212. PMID: 28539083.

Azari A., Janeja V. P. y Mohseni A. (2012). *Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach*. 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 17-24, doi: 10.1109/ICDMW.2012.69

Barnes S., Hamrock E., Toerper M., Siddiqui S. y Levin S. (2016). *Real-time prediction of inpatient length of stay for discharge prioritisation*. J. Amer. Med. Inform. Assoc., vol. 23, no. e1, pp. e2-e10

Baylis P. (2009). *Better Healthcare with Data Mining*. SPSS Inc., Shared Medical Systems Limited, UK

Bergese I., Frigerio S., Clari M., Castagno E., De Clemente A., Ponticelli E., Scavino E. y Berchialla P. (2019). *An Innovative Model to Predict Pediatric Emergency Department Return Visits*. Pediatric Emergency Care, vol. 35(3), pp. 231-236

Cai X., Perez-Concha O., Coiera E., Martin-Sanchez F., Day R., Roffe D. y Gallego B. (2015). *Real-time prediction of mortality, readmission, and length of stay using electronic health record data*. J Am Med Inform Assoc., vol. 23(3), pp. 553-561. doi: 10.1093/jamia/ocv110. PMCID: PMC7839923.

Chang C.-L. y Lu P.-Y. (2016). *The Study on Evaluating Length of Hospital Stay for Myomectomy*. Int. J. Sci. Eng. Invest., vol. 5(59), pp. 157-162

Chuang M., Hu Y., Tsai C., Lo C. y Lin W. (2015). *The Identification of Prolonged Length of Stay for Surgery Patients*. In Proc. IEEE Int. Conf. Syst., Man, Cybern, pp 3000-3003.

Chuang M., Hu Y. y Lo, C. L. (2016). *Predicting the prolonged length of stay of general surgery patients: A supervised learning approach*. International Transactions in Operational Research, vol. 25(1). DOI 10.1111/itor.12298.

Combes C., Kadri F. y Chaabane S. (2014). *Predicting Hospital Length Of Stay Using Regression Models: Application To Emergency Department*. In Proc. 10th Conférence Francophone de Modélisation, Optimisation et Simulation (MOSIM). Available: https://hal.archives-ouvertes.fr/hal-01081557

El-Darzi E., Abbi R., Vasilakis C., Gorunescu F., Gorunescu M. y Millard P. (2009). *Length of Stay-Based Clustering Methods for Patient Grouping*. In Intelligent Patient Management (Series Studies), Berlin, Germany: Springer, vol. 189, pp. 39-56.

Evain S., Bourigault C., Juvin M.-E., Corvec S. y Lepelletier D. (2019). *Carbapenemase-producing Enterobacteriaceae (CPE) digestive carriage: How many carriers are already positive at hospital readmission and what is the role of antibiotic exposure for the remaining positive patients during hospital stay?*. In ScienceDirect, Elsevier, vol. 102(1), pp. 25-30

Ferrão J. C., Duarte M., Janela F. y Martins H. (2015). *Predicting Length of Stay and Assignment of Diagnosis Codes during Hospital Inpatient Episodes*. In Proc. 1st Karlsruhe Service Summit Workshop Adv. Service Res., Karlsruhe, Germany: KIT Scientific, pp. 65-72.

Freitas A., Silva-Costa T., Lopes F., Garcia-Lema I., Teixeira-Pinto A., Brazdil P. y Costa-Pereira A. (2012). *Factors influencing hospital high length of stay outliers*. BMC Health Services Res., vol. 12(1), p. 265

Futoma J., Morris J. y Lucas J. (2015). *A comparison of models for predicting early hospital readmissions*. Journal of Biomedical Informatics, Elsevier, Durham, USA, vol. 56, pp. 229-238

George L. A., Martin B., Gupta N., Shastri N., Venu M. y Naik A. S. (2019). *Predicting 30-day readmission rate in inflammatory bowel disease patients: Performance of LACE index*. Crohn's & Colitis 360, Ed. Decision, Chicago, IL, vol 1(1), pp. 1-7

Goodman P. H., Kaburlasos V. G., Egbert D. D., Carpenter G. A., Grossberg S., Reynolds J. H., Rosen D. B. y Hartz A. J. (2002). *Fuzzy ARTMAP neural network compared to linear discriminant analysis prediction of the length of hospital stay in patients with pneumonia*. [Proceedings] 1992 IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pp. 748-753, doi: 10.1109/ICSMC.1992.271536.

Greysen S. R., Stijacic C. I., Auerbach A. D. y Covinsky K. E. (2015). *Functional impairment and Hospital readmission in Medicare seniors*. JAMA Intern. Med., vol. 175(4), pp. 559-565

Gul, M. y Guneri, A.F. (2014). *Forecasting patient length of stay in an emergency department by artificial neural networks*. Journal of Aeronautics and Space Technologies (Havacilik ve Uzay Teknolojileri Dergisi), vol. 8, pp. 1-6.

Hachesu P. R., Ahmadi M., Alizadeh S. y Sadoughi F. (2013). *Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients*. Healthcare Informat. Res., vol. 19(2), pp. 121-129

Houthooft R., Ruyssinck J., Van Der Herten J., Stijven S., Couckuyt I., Gadeyne B. y De Turck F. (2015). *Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores*. Artif. Intell. Med., vol. 63(3), pp. 191–207

Howell S., Coory M., Martin J. y Duckett S. (2010). *Using routine inpatient data to identify patients at risk of hospital readmission*. BioMed Central Health Services Research, Springer Nature, vol. 9(96), pp. 1-9

Jamei M., Nisnevich A., Wetchler E., Sudat S. y Liu E. (2017). *Predicting all-cause risk of 30-day hospital readmission using artificial neural networks*. PLoS ONE, Canada, vol. 12(7), pp. 1-14

Kansagara D., Englander H., Salanitro A., Kagen D., Theobald C., Freeman M. y Kripalani S. (2011). *Risk prediction models for hospital readmission. A systematic review*. Journal American Medical Association, vol. 306(15), pp. 1688-1698

Kiran R. P., Delaney C. P., Senagore A. J., Steel M., Garafalo T. y Fazio V. W. (2004). *Outcomes and prediction of hospital readmission after intestinal surgery*. J Am Coll Surg. Vol. 198(6), pp. 877-883, doi: 10.1016/j.jamcollsurg.2004.01.036. PMID: 15194068.

LaFaro R. J., Pothula S., Kubal K. P., Inchiosa M. E., Pothula V. M., Yuan S. C. y Inchiosa M. A. (2015). *Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre-Incision Variables*. PLoS ONE, Public Library Sci., vol. 10, pp. 1-19

Lee J., Govindan S., Celi L. A., Khabbaz K. R., Subramaniam B. (2013). *Customized Prediction of Short Length of Stay Following Elective Cardiac Surgery in Elderly Patients Using a Genetic Algorithm*. World J Cardiovasc Surg, vol. 3(5), pp. 163-170. doi: 10.4236/wjcs.2013.35034. PMID: 24482754; PMCID: PMC3904130.

Li J.-S., Tian Y., Liu Y.-F., Shu T. y Liang M.-H. (2013). *Applying a BP Neural Network Model to Predict the Length of Hospital Stay*. Springer-Verlag Berlin Heidelberg 2013, pp. 18-29, 10.1007/978-3-642-37899-7_2.

Li Q., Yan M. y Xu J. (2020). *Optimizing Convolutional Neural Network Performance by Mitigating Underfitting and Overfitting*. 2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS), pp. 126-131

Liu P., El-Darzi E., Vasilakis C., Chountas P., Huang W. y Lei L. (2004). *Comparative Analysis of Data Mining Algorithms for Predicting Inpatient Length of Stay*. In Proc. PACIS, vol. 86, pp. 1087-1097.

Liu P., Lei L., Yin J., Zhang W., Naijun W. y El-Darzi E. (2006). *Healthcare Data Mining: Prediction Inpatient Length of Stay*. In Proc. 3rd Int. IEEE Conf. Intell. Syst., pp. 832-837.

Lowell W. y Davis G. (1994). *Predicting Length of Stay for Psychiatric Diagnosis-related Groups Using Neural Networks*. Journal of the American Medical Informatics Association. Vol. 1 (6), pp. 459-466

Lowell W., Davis G., Lajousky W., Stieffel S., Breau M. y Shirazi H. (1997). *A field trial using Artificial Neural Networks to predict psychiatric inpatient Length-of-stay*. Neural Comput. Appl., vol. 5, no. 3, pp. 184-193.

Maied Z. A., Balaji D., Zaidi A. R. Z., Nawal F. A., Zaidi F. Z., Rababah A. A., Assiri A. A., AlGhamdi M. S., Al Mutair A. y Al-Omari A. (2020). *COVID-19 and Palliative Care Services: Comparative Patterns of Inpatient, Outpatient, and Consultation Services in a Tertiary Care Center in Riyadh*. Cureus, vol. 12(12), pp. 1-11

Mekhaldi R. M., Caulier P., Chaabane S., Chraibi A. y Piechowiak S. (2020). *Using Machine Learning to Predict the Length of Stay in a Hospital Setting Trends and Innovations in Information Systems and Technologies*. Cham, Switzerland: Springer, 2020, vol. 1159, pp. 202-211

Morton A., Marzban E., Giannoulis G., Patel A., Aparasu R. y Kakadiaris I. A. (2014). *A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients*. 2014 13th International Conference on Machine Learning and Applications, pp. 428-431. DOI 10.1109/ICMLA.2014.76

Palomares M. T. D.-L., De la Iglesia F., Nicolás R., Pellicer C., Ramos V. y Martínez F. D.-L. (2002). *Factors that predict unplanned hospital readmission of patients discharged from a Short Stay Medical Unit*. An. Med. Interna, Madrid, vol. 19(5), pp. 9-13

Panchami V. U. y Radhika N. (2014). *A Novel Approach for Predicting the Length of Hospital Stay With DBSCAN and Supervised Classification Algorithms*. In Proc. 5th Int. Conf. Appl. Digit. Inf Web Technol. (ICADIWT), pp. 207-212

Parag C. P. y Hitesh K. (2014). *Machine Learning Techniques for Predicting Hospital Length of Stay in Pennsylvania Federal and Specialty Hospitals*. Int. J. Comput. Sci. Appl., vol. 11(3), pp. 45-56

Peiró S., Librero J. y Martínez A. B. (1996). *Factors associated to emergency hospital readmittance in digestive and hepatobiliary diseases*. Medicina Clínica, Valencia, Spain, vol. 107(1), pp. 4-13

Riascos A. y Serna N. (2017). *Predicting Annual Length-Of-Stay and its Impact on Health*. Proceedings of The First Workshop Medical Informatics and Healthcare held with the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining, PMLR, vol. 69, pp. 27-34

Rouzbahman M., Jovicic A. y Chignell M. (2017). *Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU?*. IEEE J. Biomed. Health Informat., vol. 21(3), pp. 851-858

Shailesh K. R., Tanuja S. y Acharya D. (2011). *Comparison of different data mining techniques to predict hospital length of stay*. Journal of Pharmaceutical and Biomedical Sciences, vol. 7(15). ISSN NO- 2230 - 7885

Stoean R., Stoean C., Sandita A., Ciobanu D. y Mesina C. (2015). *Ensemble of Classifiers for Length of Stay Prediction in Colorectal Cancer*. In Advances in Computational Intelligence, Cham, Switzerland: Springer, vol. 9094, pp. 444-457

Suter E., Oelke N. D., Adair C. E. y Armitage G. D. (2009). *Ten Key Principles for Successful Health Systems Integration*. Healthcare Quart., Toronto, Ont., vol. 13, pp. 16–23

Turgeman L., May J. H. y Sciulli R. (2017). *Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission*. Expert Syst. Appl., vol. 78, pp. 376–385

Walsh P., Cunningham P., Rothenberg S. J., O'Doherty S., Hoey H. y Healy R. (2004). *An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis*. Eur J Emerg Med. 2004 Oct, vol. 11(5), pp. 259-264. doi: 10.1097/00063110-200410000-00004. PMID: 15359198.

Whitlock T. L., Tignor A., Webster E. M., Repas K., Conwell D., Banks P. A. y Wu B. U. (2010). *A scoring system to predict readmission of patients with acute pancreatitis to the hospital within thirty days of discharge*. Clin. Gastroenterol. Hepatol., vol. 9(2), pp. 175-180

Yao Z., Liu P., Lei L. y Yin J. (2005). *R-C4.5 decision tree model and its applications to health care dataset.* Proceedings of ICSSSM '05. 2005 International Conference on Services Systems and Services Management, 2005, vol. 2, pp. 1099-1103, doi: 10.1109/ICSSSM.2005.1500165.

Yang C.-S., Wei C.-P., Yuan C.-C. y Schoung J.-Y. (2010). *Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages*. Decis. Support Syst., vol. 50(1), pp. 325–335

Yasinski E., Reilly D., Duggal N., Walker B. S., Carpintero E., Nag S. y Hentz C. (2017). *Understanding & Predicting Length of Stay (LOS) using Machine Learning*. In Proc. Dexur. 15th Floor, 575 Fifth Avenue, New York, NY, USA, vol. 10017. Available: https://dexur.com/a/ml-research-los/6/

Zhibin L. V., Ding H., Wang L. y Zou Q. (2020). *A convolutional neural network using dinucleotid one_hot encoder for identifying DA N6-Methyladenine sites in the rice genome*. Neurocomputing, Elsevier B. V., vol. 422(2021), pp. 214-221