UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIERÍA INDUSTRIAL VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DISEÑO DE UN SISTEMA DE APRENDIZAJE BASADO EN PROYECTOS EN CIENCIA DE DATOS BIOMÉDICOS UTILIZANDO LA BASE DE DATOS PÚBLICA MIMIC III

AUTOR: Luis Alcalá Pérez

TUTOR: Juan Miguel García Gómez

COTUTOR: Carlos Sáez Silvestre

Curso Académico: 2019-20

AGRADECIMIENTOS

A mi tutor, Carlos, por ser tan buen guía desde el primer día, por el apoyo que me ha dado y su amabilidad, haciendo que todo sea más fácil.

A mi madre y a mi padre, por estar ahí siempre que los necesito y fijarse siempre en los pequeños detalles.

Gracias a mis amigos, por el apoyo incondicional, siempre disponibles y aconsejándome.

Por último, gracias a mi compañera de viajes, porque aunque haya sido un verano difícil, lo hemos sacado adelante juntos, como siempre hacemos.

RESUMEN

Las asignaturas de Sistemas de Información y Telemedicina y Data Quality and Interoperability del Grado y Máster de Ingeniería Biomédica de la Universidad Politécnica de Valencia, España, abordan los objetivos de aprendizaje relacionados con el manejo y procesado de bases de datos biomédicas, utilizando estándares de información en salud para la obtención y el intercambio de datos, el análisis de la calidad de estos, y el desarrollo de modelos de *machine learning.* Estos objetivos de aprendizaje cubren un amplio abanico de actividades distintas en el ciclo de vida de los datos biomédicos, lo que puede dificultar el proceso de aprendizaje en el tiempo limitado asignado para cada asignatura. Proponemos un aprendizaje basado en proyectos abordando el ciclo de vida completo de los datos biomédicos utilizando el *dataset* público MIMIC-III (Medical Information Mart for Intensive Care III), una base de datos accesible gratuitamente que contiene información relacionada con la admisión de pacientes en la unidad de cuidados intensivos. Por medio de este enfoque de aprendizaje activo, los estudiantes pueden lograr todos los objetivos de aprendizaje de las asignaturas de forma integral: entendiendo el modelo de datos MIMIC-III, utilizando los estándares de información como *International Classification of Diseases 9th Edition* (ICD-9), mapeando a estándares de interoperabilidad, consultando datos, creando tablas de datos y abordando la calidad de datos con el objetivo de aplicar estadísticas fiables y análisis *machine learning* y*,* desarrollando múltiples modelos como los de predicción de mortalidad hospitalaria. Para ello, se han desarrollado tres nuevos *datasets* extrayendo información de importancia de la base de datos MIMIC III mediante diferentes consultas SQL, se han preprocesado estos datos para un posterior modelado predictivo de mortalidad hospitalaria, con el objetivo de crear un aprendizaje activo para los alumnos y de poder, en un futuro, incluir estos modelos en un sistema de ayuda a la decisión médica.

RESUM

Les assignatures de Sistemes d'Informació i Telemedicina i Data Quality and Interoperability del Grau i Màster d'Enginyeria Biomèdica de la Universitat Politècnica de València, Espanya, aborden els objectius d'aprenentatge relacionats amb el maneig i processament de bases de dades biomèdiques, utilitzant estàndards d'informació en salut per a l'obtenció i l'intercanvi de dades, l'anàlisi de la qualitat d'aquests, i el desenvolupament de models de *machine learning*. Aquests objectius d'aprenentatge cobreixen un ampli ventall d'activitats diferents en el cicle de vida de les dades biomèdiques, la qual cosa pot dificultar el procés d'aprenentatge en el temps limitat assignat per a cada assignatura. Proposem un aprenentatge basat en projectes abordant el cicle de vida complet de les dades biomèdiques utilitzant el *dataset* públic MIMIC-III (Medical Information Mart for Intensive Care III), una base de dades accessible gratuïtament comprenent informació relacionada amb l'admissió de pacients en la unitat de vigilància intensiva. Per mitjà d'aquest enfocament d'aprenentatge actiu, els estudiants poden aconseguir tots els objectius d'aprenentatge de les assignatures de manera integral: entenent el model de dades MIMIC-III, utilitzant els estàndards d'informació com International Classification of Diseases 9th Edition (ICD-9), convertint a estàndards d'interoperabilitat, consultant dades, creant taules de dades i abordant la qualitat de dades amb l'objectiu d'aplicar estadístiques fiables i anàlisis *machine learning* i, desenvolupant múltiples models com els de predicció de mortalitat hospitalària. Per a això, s'han desenvolupat tres nous datasets extraient informació d'importància de la base de dades MIMIC III mitjançant diferents consultes SQL, s'han preprocessat aquestes dades per a un posterior modelatge predictiu de mortalitat hospitalària, amb l'objectiu de crear un aprenentatge actiu per als alumnes i de poder, en un futur, incloure aquests models en un sistema d'ajuda a la decisió mèdica.

ABSTRACT

The subjects Health Information Systems and Telemedicine and Data Quality and Interoperability of the Degree and Master in Biomedical Engineering of the Universitat Politècnica de València, Spain, address learning outcomes related to managing and processing biomedical databases, using health information standards for data capture and exchange, data quality assessment, and developing machine-learning models from these data. These learning outcomes cover a large range of distinct activities in the biomedical data life-cycle, what may hinder the learning process in the limited time assigned for the subject. We propose a project based learning approach addressing the full life-cycle of biomedical data on the MIMIC-III (Medical Information Mart for Intensive Care III) Open Dataset, a freely accessible database comprising information relating to patients admitted to critical care units. By means of this active learning approach, students can achieve all the learning outcomes of the subject in an integrated manner: understanding the MIMIC-III data model, using health information standards such as International Classification of Diseases 9th Edition (ICD-9), mapping to interoperability standards, querying data, creating data tables and addressing data quality towards applying reliable statistical and machine learning analysis and, developing predictive models for several tasks such as predicting in-hospital mortality. Therefore, three new datasets have been created extracting key information from the MIMIC III database through SQL querys, this data has then been preprocessed to create in-hospital mortality prediction models with the aim of creating an active learning methodology for students and, in a future, implement those models into a clinical decision support system.

ÍNDICE

ÍNDICE MEMORIA

ÍNDICE PRESUPUESTO

ÍNDICE TABLAS

ÍNDICE TABLAS MEMORIA

ÍNDICE TABLAS PRESUPUESTO

ÍNDICE FIGURAS

MEMORIA

1. Introducción

1.1 Motivación

La digitalización de los sistemas de salud ha evolucionado a saltos agigantados durante las últimas décadas y está transformando la medicina tal y como la conocemos, tanto desde el punto de vista del médico como del paciente. Esto conlleva innumerables retos en logística, interoperabilidad, análisis de datos y seguridad. La inmensa cantidad de datos, que crece exponencialmente ("Big Hopes for Big Data," 2020), supone una de las mayores revoluciones que se han conocido en el campo de la medicina.

Uno de los campos donde mayor volumen de datos se generan son las Unidades de Cuidados Intensivos (UCI) (Thomas et al., 2017). La monitorización de los pacientes en estos entornos es mayor que en cualquier otro lugar de un hospital. Es crítico, debido a la situación del paciente (Bricon-Souf & Newman, 2007), la obtención de importantes parámetros derivados de la correcta interpretación de estos datos, donde el uso de modelos de aprendizaje automático sería de gran ayuda para los profesionales de la salud. Sin embargo, la complejidad en la interpretación de esta inmensa cantidad de datos generados y las consecuencias que ello conlleva debido al entorno donde se está trabajando resalta la importancia en la formación de ingenieros cualificados.

Por otro lado, es en asignaturas como Sistemas de Información y Telemedicina (SIT) y Data Quality and Interoperability (DQI), impartidas en el grado y master de Ingeniería Biomédica de la Universitat Politècnica de València respectivamente, donde se prepara a los alumnos para el correcto análisis y utilización de cualquier tipo de dato derivado de las actividades médicas. Desde la primera toma de contacto con la información, su análisis estadístico previo y la calidad de la información, hasta la generación de modelos de predicción o de ayuda al diagnóstico. Es de gran importancia para el alumnado poder aprender y trabajar con unos datos lo más parecidos a la realidad posibles, encontrándose así con las dificultades reales que se muestran ante los profesionales día a día y aprendiendo a sortearlas.

Por lo tanto, la formación de profesionales capaces de desarrollar modelos de aprendizaje automático para pacientes en estado crítico es clave para la mejora en el bienestar del paciente, el apoyo al profesional sanitario y una mejora en la eficacia del sistema.

1.2 Objetivos

El objetivo principal del trabajo es el desarrollo de un modelo de aprendizaje activo para las asignaturas SIT y DQI basado en el conjunto de datos (o en inglés *dataset*) en abierto Medical Information Mart for Intensive Care (MIMIC) III, el cual cubra las distintas fases del ciclo de vida de datos biomédicos, desde su captura, post-proceso, análisis y generación de conocimiento.

Los objetivos secundarios derivados del desarrollo de dicho modelo son los siguientes:

- O1: Realizar una descripción detallada de los datos presentes en el *dataset* MIMIC-III
- O2: Extracción y preparación de los datos necesarios para la creación de una base de datos con la que el alumnado pueda trabajar.
- O3: Desarrollo de diferentes modelos predictivos de mortalidad a estudiar en la asignatura SIT en base a los nuevos parámetros.
- O4: Desarrollo de nuevos modelos predictivos de mortalidad de mayor potencia.
- O5: Análisis de la calidad de datos y su posible influencia en estos modelos en base a lo estudiado en la asignatura DQI
- O6: Alinear el trabajo realizado con los objetivos de aprendizaje de las asignaturas SIT y DQI y plantearlo como método de aprendizaje de las mismas

En el Trabajo Final de Grado se deben de poner en manifiesto ciertas competencias adquiridas a lo largo del Grado de Ingeniería Biomédica. En particular, este proyecto se centra en el área del tratamiento de datos clínicos y extracción de conocimiento de los mismos.

1.3 Contribuciones

En primer lugar, se ha generado una propuesta de aprendizaje activo para las asignaturas SIT y DQI del Grado y Máster en Ingeniería Biomédica, respectivamente, la cual queda recogida en la presente memoria de Trabajo Final de Grado (TFG). La utilización de esta metodología será evaluada en los próximos cursos académicos.

En segundo lugar, los resultados de este TFG han sido recopilados en forma de artículo el cual, previa presentación de este TFG, ha sido aceptado en el congreso INNODOCT/20 – Born Virtual.

- Luis Alcalá, Juan M García-Gómez. Carlos Sáez. (2020). *Project based learning in Biomedical Data Science using the MIMIC III Open dataset*. Aceptado en la Conferencia Internacional sobre Innovación, Documentación y Tecnologías Docentes (INNODOCT/20). València, Noviembre 2020.

2. Antecedentes

2.1 Aprendizaje basado en proyectos (ABP)

En los últimos años, las metodologías docentes centradas en el aprendizaje del estudiante y que ofrecen una mayor implicación en la enseñanza de este se han instaurado en las aulas universitarias, llamadas “metodologías activas” (Vega et al., 2014). Una de las más importantes es el ABP, por la que el alumno adquiere los conocimientos y aptitudes requeridas trabajando durante un periodo de tiempo extendido sobre un proyecto de aprendizaje real, implicándose en la toma de decisiones y la actividad investigadora, mientras que el profesor es un guía de dicho proceso (Blumenfeld et al., 1991).

2.2 Asignaturas SIT y DQI

SIT y DQI son dos asignaturas impartidas en el Grado en Ingeniería Biomédica y Master en Ingeniería Biomédica por la Universitat Politècnica de València, España. Los objetivos principales de aprendizaje de SIT incluyen el procesado de repositorios electrónicos de salud, el desarrollo de modelos de *machine learning* y el desarrollo de sistemas de ayuda a la decisión utilizando datos biomédicos. A su vez, los objetivos de aprendizaje principales de DQI se centran en la descripción de la calidad de los datos biomédicos, observando así las consecuencias que puede tener el análisis de estos en el desarrollo de sistemas de soporte a la decisión.

La adquisición de las competencias necesarias para la creación de modelos de predicción y *machine learning*, así como comprender los problemas que conlleva el pre-procesado y calidad de los datos biomédicos puede resultar abstracta al alumnado si se enfoca desde un punto de vista exclusivamente teórico. Es por ello que optar por un enfoque en ABP, provocando así que el alumno pueda recorrer todo el ciclo de vida de los datos biomédicos, desde el pre-procesado y análisis de la calidad de datos hasta el desarrollo de modelos resulta ser el mejor planteamiento para que el aprendizaje del alumno sea lo más eficiente posible.

En la actualidad, DQI (Sáez, Mañas, et al., 2017) tiene un ABP propio en funcionamiento. Sin embargo, la complementariedad de esta asignatura con SIT ofrece la posibilidad de utilizar un ABP común y continuo que puede mejorar el aprendizaje utilizando datos biomédicos con casuísticas reales.

2.3 Sistemas de ayuda a la decisión médica (CDSS)

2.3.1 Introducción a los CDSS

Los CDSS son una herramienta derivada de la aplicación de Inteligencia Artificial en el ámbito de la salud (Wasylewicz & Scheepers-Hoeks, 2018). Estos sistemas computacionales tienen como objetivo principal aportar un conocimiento específico al profesional sanitario y servirle así de apoyo en la toma de decisiones tanto en planes de tratamiento, como en diagnóstico del paciente, pretendiendo así mejorar la atención sanitaria individualizada. Por otro lado, el objetivo secundario de la implantación de los CDSS es provocar un aumento en la eficiencia de los sistemas sanitarios, utilizando la enorme cantidad de datos generados por estos, para mejorar el aprovechamiento de los recursos sanitarios disponibles y reducir lo máximo posible los errores médicos. En resumen, los CDSS tratan de aportar un valor añadido a las tareas de los profesionales de la sanidad permitiendo que estos manejen un mayor número de variables de las que serían capaces por ellos mismos.

En particular, los modelos de predicción de mortalidad hospitalaria son un tipo de CDSS, utilizados comúnmente en situaciones con pacientes críticos (Johnson et al., 2017), como pacientes ingresados en las Unidades de Cuidados Intensivos (UCI), y son en los que se va a centrar este trabajo.

Hoy en día, existen dos grandes escalas dirigidas a la predicción de mortalidad en la UCI, el sistema Acute Physiology and Chronic Health Evaluation (APACHE) (Niewiński et al., 2014) y el sistema Simplified Acute Physionolgy Score (SAPS) (Alvear-Vega & Canteros-Gatica, 2018), ambos teniendo diferentes versiones. Asignan una puntuación para cada paciente, basándose en diferentes parámetros fisiológicos. Sin embargo, como se ha comentado anteriormente, se trata de dos escalas no basadas en sistemas de aprendizaje automático.

La mortalidad es el indicador más importante de los resultados de la calidad y la eficacia en la atención de las UCI. Estas son las áreas hospitalarias con menor cantidad de camas en el hospital, entre el 5 y el 10% y sin embargo, son las que más recursos consumen proporcionalmente (30% de los recursos para pacientes agudos y más del 10% de todos los costes hospitalarios) (*ICU Outcomes | Philip R. Lee Institute for Health Policy Studies*, n.d.).

Han sido 5 los algoritmos elegidos para el desarrollo de los CDSS, a continuación se va realizar una pequeña introducción a cada uno de ellos para comprender su funcionamiento, su utilidad y cuándo es posible su aplicación.

2.3.2 Análisis Discriminante Lineal y Cuadrático

El Análisis Discriminante Lineal (cuyas siglas en inglés son LDA) es un conocido método de clasificación supervisado, es decir, el objetivo es clasificar correctamente las nuevas instancias (James, G., Witten, D., Hastie, T., Tibshirani, 2013). Las variables en cuestión deben ser variables cuantitativas, además, los grupos de salida, en este caso paciente fallecido o paciente no fallecido, deben de ser conocidos a priori. Cuando llega una nueva observación, esta se clasifica en uno de estos grupos en función de sus características. LDA se basa en el teorema de Bayes (Ecuación 1) para así estimar la probabilidad de que dicha observación entrante pertenezca a una u otra clase. La clase asignada a la observación será la que mayor probabilidad predicha haya obtenido.

$$P(pertenecer\ a\ la\ clase\ k|valor\ x\ observado) = \frac{P(pertenecer\ a\ la\ clase\ k\ y\ observar\ x)}{P(observar\ x)} \quad (Ec.1)$$

Por otro lado, el Análisis Discriminante Cuadrático (cuyas siglas en inglés son QDA) se asemeja en gran medida al LDA (James, G., Witten, D., Hastie, T., Tibshirani, 2013). El método es el mismo pero QDA utiliza una matriz de covarianza propia para cada clase a predecir, por lo tanto, la función discriminante deja de tener forma lineal (Ecuación 2) y pasa a tener forma cuadrática (Ecuación 3).

$$\log\big(\mathrm{P}(\mathrm{Y}=\mathrm{k}|\mathrm{X}=\mathrm{x})\big) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad (Ec.2)$$

$$\log\big(\mathrm{P}(\mathrm{Y}=\mathrm{k}|\mathrm{X}=\mathrm{x})\big) = -\frac{1}{2}\log\left|\sum k\right| - \frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k) + \log(\pi_k) \qquad (Ec.3)$$

2.3.3 Algoritmo K Vecinos Más Próximos

El algoritmo K Vecinos Más Próximos (cuyas siglas en inglés son KNN) es el algoritmo más conocido de la familia *lazy learning methods* (James, G., Witten, D., Hastie, T., Tibshirani, 2013). El KNN es un método de aprendizaje supervisado, como LDA y QDA, y de una gran sencillez. Trata de clasificar cada nueva observación en su grupo correspondiente calculando la distancia Euclídea entre esta nueva observación y las ya existentes (Figura 2-1). La variable k determina el número de observaciones vecinas tenidas en cuenta a la hora de clasificar. El método necesita de variables cuantitativas y una clase salida conocida.

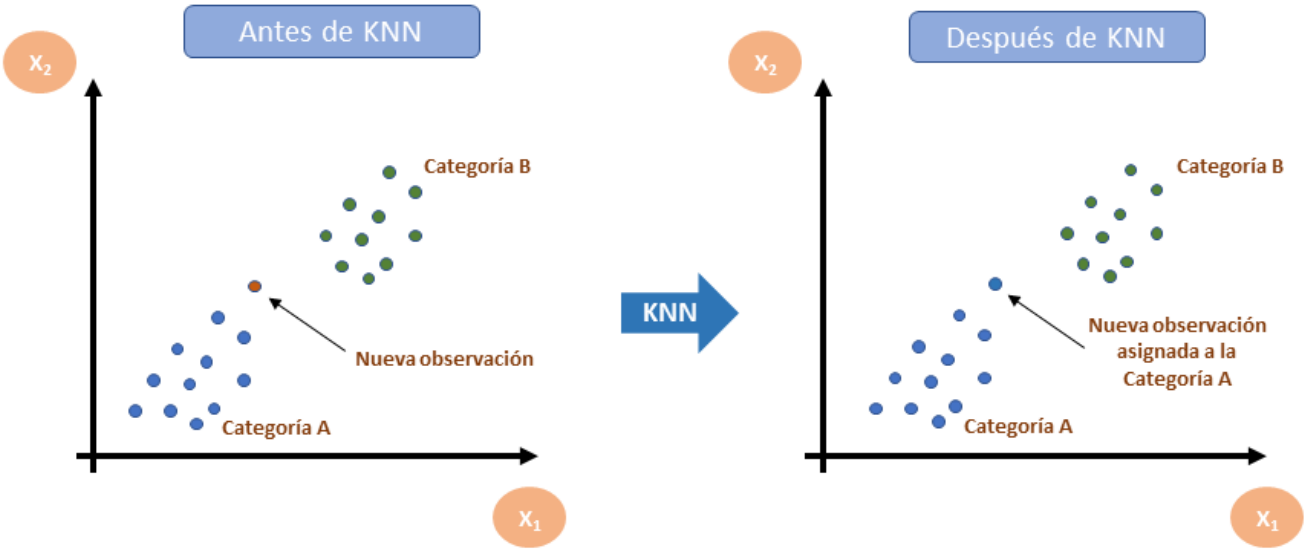

*Figura 2-1. Clasificación de una nueva observación mediante el algoritmo KNN (*K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*, n.d.))*

2.3.4 Random Forest

El algoritmo *Random Forest* es una técnica de clasificación que se basa en el promediado de los resultados de conjuntos de clasificadores (árboles de decisión) (James, G., Witten, D., Hastie, T., Tibshirani, 2013). Cada uno de ellos es entrenado con un subconjunto de características distinto, derivado del *dataset* principal. Tanto el número de clasificadores como el número de características son parámetros a determinar en el algoritmo. Estos clasificadores del subconjunto son denominados *weak classifiers* debido a que tienen al menos un 51% de éxito, por lo que no son clasificadores muy buenos. Sin embargo, al combinar un gran número de estos *weak classifiers*, se construye así un *strong classifier* que tiene un porcentaje de éxito mucho mayor. Es un método muy eficiente y una de las técnicas de clasificación más precisas actualmente.

2.3.5 Gradient Boosting

Como *Random Forest, Gradient Boosting* también es un método que convierte *weak classifiers* en *strong classifiers* (James, G., Witten, D., Hastie, T., Tibshirani, 2013). Sin embargo, cada uno de los clasificadores en este algoritmo se entrena de manera gradual, aditiva y secuencial. El método asigna gradientes después de cada clasificador a las variables en función de si han resultado fáciles o difíciles de clasificar.

2.4 MIMIC

El proyecto MIMIC surgió en 1992 debido a la creciente necesidad de obtener datos de prueba reproducibles, fiables y bien caracterizados, para así tener la capacidad de desarrollar sistemas de soporte a la decisión automatizados centrados en pacientes ingresados en la UCI (PhysioNet, 2015).

Es una base de datos relacional desarrollada por el Laboratorio de Fisiología Computacional del Massachusetts Institute of Technology (MIT). En ella se encuentra información desidentificada relacionada con la estancia de pacientes en la unidad de cuidados intensivos (UCI) del Beth Israel Deaconess Medical Center, Boston.

La primera versión de este *dataset*, MIMIC - I, fue desarrollada entre 1992 y 1999, a partir de 90 pacientes de UCI (Moody & Mark, 1996). Los datos obtenidos se centraban en señales y medidas registradas en la monitorización de cada paciente en cama, así como datos clínicos obtenidos de la historia clínica de este. Esta versión fue el primer intento en la creación de una base de datos de parámetros múltiples basada en pacientes ingresados en la UCI. Sin embargo, la calidad y la cantidad de los datos, así como su almacenamiento, estaban restringidos a la tecnología y los conocimientos de la época.

A partir del conocimiento desarrollado con esta versión inicial, se llevó a cabo un nuevo proyecto, llamado MIMIC – II (PhysioNet, 2009). Este tuvo lugar entre 2001 y 2008. La idea principal era la misma: obtener datos fisiológicos, clínicos y señales vitales durante la estancia de los pacientes en la UCI, así como información médica adicional obtenida de los archivos del hospital. Esta segunda versión, al contrario que la primera, contiene formas de onda fisiológicas de alta resolución y series numéricas temporales de medidas fisiológicas con datos tomados minuto a minuto. El proyecto está separado en dos bases de datos diferentes:

- Waveform Database:

Base de datos con las formas de onda de las señales fisiológicas (tales como electrocardiogramas continuos o gráficos de presión sanguínea) y de las series temporales de medidas periódicas de señales vitales (tales como presión sanguínea media, sistólica, diastólica, frecuencia cardiaca o frecuencia respiratoria). Esta base de datos está formada con un total de alrededor de 4000 pacientes.

- Clinical Database

Base de datos con la información clínica. Esta incluye información general (demografía de los pacientes, fechas de admisión y alta etc…), información fisiológica, información sobre medicamentos, datos relacionados con test de laboratorio, etc. Todos estos datos han sido tomados sobre 26000 pacientes durante los 7 años que duró el proyecto.

Finalmente, en el año 2012, finalizó la toma de datos para la última versión disponible de esta base de datos MIMIC – III (Laboratory For Computational Physiology, 2015). Esta está compuesta por los datos presentes en MIMIC-II e incrementada con datos obtenidos entre los años 2008 y 2012.

El desarrollo de este trabajo se basa en esta última versión, utilizando sus datos como herramienta principal para el desarrollo de modelos de predicción de mortalidad y de la metodología ABP.

Cabe destacar que el 13 de agosto de 2020, durante la realización de este trabajo, se publicó MIMIC IV (Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, 2020), con el objetivo de mejorar la usabilidad de los datos y permitir una mayor creación de aplicaciones a nivel investigatorio.

3. Materiales

A continuación, se va a realizar una introducción al *dataset* público MIMIC-III, para así entender su procedencia, su importancia y lo interesante que resulta para el desarrollo de modelos de predicción tanto para investigación como para docencia.

3.1 MIMIC III

MIMIC III fue publicada en Physionet[1] el 25 de agosto de 2015 en su primera versión (MIMIC III v1.0). Se utilizó principalmente para el testeo interno de los datos y no fue hasta la versión 1.1, publicada el 24 de septiembre de ese mismo año, que se empezó a utilizar por terceros. A partir de entonces, surgieron tres versiones más, todas ellas enfocadas en la corrección de errores, la mejora de la calidad y el aumento del número de datos. Finalmente, la versión MIMIC III v1.4[2] es la última disponible, publicada el 2 de septiembre de 2016, y con la que se va a desarrollar este trabajo.

Como se puede ver en la Figura 3-1, la organización de la base de datos es bastante compleja, debido a la proveniencia y la categoría de los datos presentes.

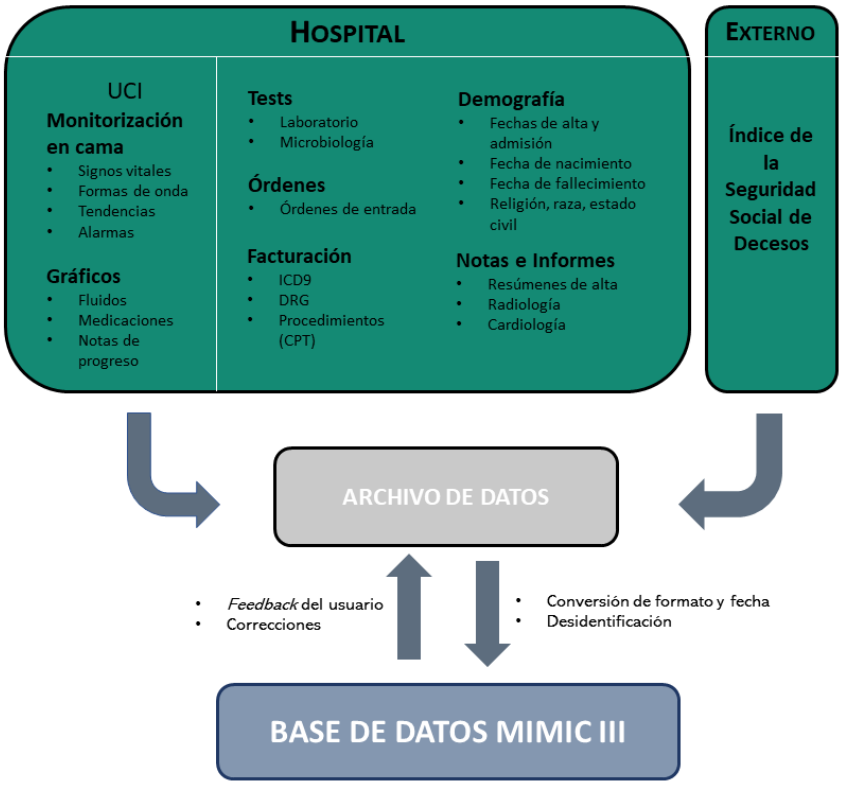

Figura 3-1. Clasificación de los datos provenientes tanto del hospital como externamente utilizados para la creación de la base de datos MIMIC III

[1] https://physionet.org/
[2] https://physionet.org/content/mimiciii/1.4/

Toda la información presente en la base de datos ha sido obtenida de tres fuentes principales, existiendo tipos de dato muy diferentes (Tabla 3-2):

- Los sistemas de archivos de la UCI

Se han utilizado dos sistemas de información de la UCI en la recogida de datos. Por un lado, el sistema *Philips CareVue Clinical Information System* y, por otro lado, el *iMDsoft MetaVision ICU*. Ambos sistemas tratan los datos y los almacenan de manera muy similar, con la excepción de los datos relacionados con la ingesta de fluidos por parte del paciente. Esto ha permitido una relativamente sencilla fusión de los dos sistemas al construir las múltiples tablas, sin embargo, los datos que no han podido juntarse han sido derivados a diferentes tablas, una por cada sistema de información, y con su sufijo correspondiente: _CV para el sistema *CareVue* y _MV para el sistema *MetaVision*.

De ambos sistemas de información se han obtenido datos como la documentación de progreso del paciente con las anotaciones de los cuidadores, medidas fisiológicas con una marca de tiempo particular verificada por una enfermera (documentación por hora de frecuencia respiratoria, frecuencia cardiaca o presión arterial) o los balances fluídicos y medicaciones intravenosas por goteo continuo.

Cabe destacar que todos estos datos provienen de la UCI del hospital, que está dividida en 5 subunidades principales (Tabla 3-1).

Tabla 3-1. Diferentes subunidades de UCI del hospital y sus correspondientes porcentajes de pacientes y admisiones únicos

Subunidad de la UCI	% Pacientes únicos	% Admisiones únicas
Cardiac/Coronary Care Unit (CCU)	14,7%	14,6%
Cardiac Surgery Intensive Care Unit (CSRU)	20,9%	18,4%
Medical Intensive Care Unit (MICU)	35,4%	39,7%
Surgical Intensive Care Unit (SICU)	16,5%	16,3%
Trauma and Surgical Intensive Care Unit (TSICU)	12,5%	11%

- Las bases de datos de historia clínica electrónica

De la historia clínica electrónica se ha obtenido la información relacionada con la demografía de los pacientes y la mortalidad interna del hospital, los resultados de los test de laboratorio, los estudios sobre las imágenes y los informes de los electrocardiogramas, así como los informes de las altas y la información obtenida de la codificación *International Classification of Disease, 9th Edition* (ICD-9) (Slee, 1978), de los códigos *Diagnoses Related Groups* (DRG) y *Current Procedural Terminology* (CPT) (Thorwarth, 2008).

- *Death Master File* de la administración de la seguridad social

El *Death Master File* se ha utilizado para la obtención de la fechas de fallecimiento fuera del hospital de los pacientes que han estado ingresados en la UCI durante cierto periodo de tiempo.

Tabla 3-2. Tipos de dato presentes en MIMIC III y su descripción.

Tipo de dato	Descripción
Descriptivo	Detalles demográficos, tiempos de admisión y alta, fechas de fallecimiento
Diccionario	Tablas de búsqueda para referenciación cruzada de identificadores con etiquetas asociadas
Facturación	Datos codificados principalmente con propósitos administrativos y de facturación
Fisiológicos	Signos vitales medidas aproximadamente cada hora y verificados por las enfermeras
Informes	Informes de texto libre de estudios electrocardiógrafos e imágenes médicas
Intervenciones	Procedimientos tales como la diálisis, estudios de imágenes, etc
Laboratorio	Resultados de los test de microbiología, análisis urinarios, hematología, etc
Medicaciones	Registros de la administración de medicaciones intravenosas y medicamentos recetados.
Notas	Notas de texto libre y resúmenes de altas hospitalarias

La base de datos MIMIC III v1.4 está compuesta por 26 tablas que recopilan toda la información disponible (ver 10.1), ocupando un total de 43,3 GB. Los datos provienen de un total de 53 423 admisiones distintas en el hospital, todas ellas relacionadas con la UCI, es decir, en todas las admisiones existentes el paciente ha pasado al menos una vez por la UCI. Un total de 38 597 pacientes adultos diferentes y 7 870 neonatos han sido partícipes en este estudio. Un 55,9% de los pacientes son varones, mientras que la edad media de todos los pacientes se sitúa en 65,8 años y la mortalidad hospitalaria en 11,5%.

Los datos recopilados en las 26 tablas de esta base de datos tienen ciertas características específicas debido a los procesos de desidentificación aplicados por los creadores del *dataset*. Por un lado, todas las fechas de cada una de las admisiones en el hospital han sido modificadas aleatoriamente a una fecha futura situada entre los años 2100 y 2200. Es decir, un mismo paciente, en una admisión concreta, tiene todas sus fechas desplazadas en el tiempo futuro. Esto se ha llevado a cabo de manera coherente con cada admisión, si para un paciente se le han añadido, por ejemplo, 89 años a sus fechas, este valor permanece constante para todas las fechas de su estancia, tanto en los datos administrativos como las fechas relacionadas con la toma de medicamentos, medición de variables fisiológicas, etc. Asimismo, se ha mantenido de la mejor manera posible el momento del día, el día de la semana y la estación de cada una de las fechas. Por otro lado, y posterior a este cambio, a las personas cuya edad supera los 89 años, se les ha aleatorizado la fecha de nacimiento, pudiendo resultar en pacientes que aparecen con más de 300 años. Por último, se ha procedido a modificar el peso de igual manera que con la edad de los pacientes.

En este trabajo nos vamos a centrar en cinco tablas en concreto para la creación de los diferentes modelos de predicción de mortalidad (Tabla 3-3):

Tabla 3-3. Principales tablas utilizadas en el trabajo, su tamaño en MB, el número de variables presentes y su descripción.

Nombre de la tabla	Tamaño (MB)	Número de variables	Descripción
ADMISSIONS	12,2	19	Admisiones en el hospital asociadas con una estancia en la UCI
CHARTEVENTS	34 480	15	Eventos que ocurren en la historia clínica de un paciente

DIAGNOSES_ICD	18,7	5	Diagnósticos relacionados con la admisión en el hospital y codificados con el sistema ICD-9
ICUSTAYS	6,2	12	Listado de las admisiones en la UCI
PATIENTS	2,6	8	Pacientes asociados con una estancia en la UCI

En estas cinco tablas, existen tres tipos de identificadores únicos relacionados con los pacientes:

- subject_id: identificador único para cada paciente
- hadm_id: identificador único para cada admisión en el hospital de cada paciente
- icustay_id: identificador único para cada estancia en la UCI de cada paciente

Teniendo en cuenta que un único paciente puede ser admitido en el hospital múltiples veces, y que en una misma admisión un paciente puede ingresar en la UCI varias veces, un único subject_id puede tener asociado múltiples hadm_id, a su vez, un único hadm_id puede tener asociado múltiples icustay_id.

A lo largo de la última década, se han desarrollado numerosos modelos predictivos de mortalidad hospitalaria basados en el *dataset* MIMIC III (Johnson et al., 2017). Un gran porcentaje de estos modelos utilizan las escalas SAPS y APACHE en su predicción (Silva et al., 2012) (Sadeghi et al., 2018), mientras que algunos desarrollan modelos propios (Caballero & Akella, 2015). En este trabajo no se van a utilizar las escalas mencionadas, sino que se van a desarrollar modelos de aprendizaje automático basados en diferentes algoritmos de clasificación.

3.2 Curso *CITI-Training* de acceso a MIMIC III

La base de datos MIMIC III es pública. Sin embargo, para poder obtener el acceso a toda su información y hacer uso de esta, se debe completar indispensablemente el curso titulado “Human Research Data or Specimens Only Research 1 – Basic Course” ofrecido por el programa Collaborative Institutional Training Initiative (CITI) bajo la supervisión de Massachusetts Institute of Technology Affiliates, cuyo certificado de obtención se puede ver en el punto 10.2.

El programa del curso contiene 8 apartados principales sobre los que se realiza una posterior evaluación:

- *Belmont Report and Its Principles*
- *History and Ethics of Human Subjects Research*
- *Basic Institutional Review Board (IRB) Regulations and Review Process*
- *Records-Based Research*
- *Genetic Research in Human Populations*
- *Populations in Research Requiring Additional Considerations and/or Protections*
- *Research and HIPAA Privacy Protections*
- *Conflicts of Interest in Human Subjects Research*

3.3 Software empleado

Como se puede ver en la Tabla 3-3, la cantidad de datos presente en las tablas, sobre todo la tabla CHARTEVENTS (34,48 GB), es muy grande, por lo que se ha optado por un software de tratamiento de base de datos, PostgreSQL 10, para la extracción de toda la información necesaria para este trabajo

Para el tratamiento estadístico y desarrollo de modelos se ha utilizado el lenguaje de programación *R* (versión 3.6.2) y el entorno de programación *RStudio* (versión 1.3.1073).

4. Metodología

4.1 Análisis estadístico general

Se ha parametrizado la base de datos mediante un análisis estadístico genérico. Para ello, se han considerado a los 38 597 pacientes adultos.

El primer paso ha sido dividirlos en tres categorías para así poder entender con mayor profundidad las diferentes características de cada grupo. Estas son: pacientes fallecidos en el hospital, pacientes fallecidos 90 días después del alta y pacientes que no han fallecido a los 90 días del alta. Para cada uno de los grupos, se ha analizado la edad, el género, y la estancia en el hospital, tanto ingresados en la UCI como no.

Por otro lado, se ha observado la distribución de los pacientes en función de la UCI en la que han sido ingresados, tanto el número de pacientes admitidos en cada UCI específica como la mortalidad en estas.

Todo esto ha permitido tener una visión global de la distribución y mortalidad de los pacientes presentes en esta base de datos.

4.2 Extracción de los datos en el servidor local PostgreSQL

Una vez terminado con el análisis estadístico preliminar, se ha procedido a la extracción de todos los datos necesarios para poder posteriormente crear los diferentes *datasets* necesarios para la modelización. Toda la extracción de datos se ha llevado a cabo mediante lenguaje SQL en un servidor local de PostgreSQL. Cabe destacar que, en un principio, se utilizó el servicio en la nube de Google, *BigQuery*. La base de datos MIMIC III se puede enlazar de forma gratuita al servidor online, pero, debido a la cantidad de datos manejados y los objetivos del TFG, se ha optado por realizarlo todo en el servidor local PostgreSQL.

4.2.1 Selección de la cohorte

El primer paso ha consistido en seleccionar el grupo de pacientes con el que vamos a trabajar. Para ello, se han optado por diferentes criterios de exclusión (Tabla 4-1) basados en la finalidad del proyecto.

Por un lado, se ha optado por excluir a todos los pacientes menores de 16 años en el momento de la admisión en la UCI, esto incluye a todos los pacientes neonatos, que no forman parte del objetivo del estudio. Por otro lado, todos los pacientes cuyas estancias en la UCI han sido menores a 4 horas también han sido eliminados debido a

que estas estancias corresponden, en su gran mayoría, a situaciones donde los modelos de predicción de mortalidad tendrían muy poco valor, como puede ser, por ejemplo, situaciones de preparación quirúrgica. A su vez, las cuentas relacionadas con pacientes donantes de órganos no se han tomado en cuenta. Por último, las admisiones sin medidas de ritmo cardiaco del paciente, así como registro tanto de admisión como de alta incompleto y sin medidas en la tabla CHARTEVENTS se han considerado admisiones inválidas, en su mayor parte debidas a errores administrativos, y han sido eliminadas.

Tabla 4-1. Criterios y métodos de selección de la cohorte

Criterio de exclusión	Método de exclusión
Pacientes menores de 16 años	Filtrado basado en el cálculo de la edad en el momento de la admisión
Estancia menor de 4 horas	Filtrado basado en el tiempo de estancia en la UCI
Donantes de órganos	Filtrado por código asociado a la cuenta de donante de órganos
Admisiones inválidas	Comprobación del cumplimiento de los criterios y eliminación

Resulta importante mencionar que una gran cantidad de admisiones en la UCI son en realidad readmisiones de un mismo paciente. Cada una de estas readmisiones se ha tratado como si fuese un paciente nuevo, sin embargo, en los modelos de predicción desarrollados posteriormente se han tomado medidas para que esto no distorsione los resultados.

Para proceder con el filtrado de los pacientes de interés según los criterios enumerados anteriormente, se ha creado una nueva tabla llamada pm_cohorte en el servidor local PostgreSQL. En esta tabla, a cada una de las admisiones se le ha asignado un “flag” indicando si es válida o no, siendo el 1 indicador de admisión inválida, y 0 indicador de admisión válida.

De una cantidad inicial de más de 50 000 admisiones, el grupo final, después de la aplicación de los diferentes criterios de exclusión, contiene un total de 37 765 admisiones.

4.2.2 Selección de las variables de interés

Una vez que el grupo de pacientes de interés ha sido elegido, se ha procedido a determinar cuáles son las variables, tanto fisiológicas como administrativas, a tomar en cuenta para el desarrollo posterior de los modelos.

En primer lugar, se ha definido una ventana de tiempo para extraer las series temporales de las variables fisiológicas en cuestión. Este intervalo temporal da comienzo en el momento de la admisión del paciente en la UCI y finaliza a sus 24 horas. Por lo tanto, todos los datos fisiológicos de los pacientes van a extraerse durante las primeras 24 horas de ingreso en la UCI.

En segundo lugar, se ha procedido con la selección de las variables fisiológicas clave para el estudio. Esta selección se ha basado en la literatura ya existente (Johnson et al., 2017). Estas variables son: frecuencia cardiaca (fc), presión arterial sistólica (pas), diastólica (pad) y media (pam), frecuencia respiratoria (fr), temperatura corporal (T), peso (w), saturación de oxígeno en sangre(O2) y glucosa en sangre (g), que, como se ha mencionado anteriormente, se encuentran en la tabla CHARTEVENTS.

Se ha desarrollado un query SQL en el servidor local PostgreSQL, utilizando tanto la tabla pm_cohorte creada anteriormente, como la tabla CHARTEVENTS, para extraer toda la información necesaria mencionada para cada uno de los pacientes. Además, se ha extraído información no vital como puede ser la edad, el género y los identificadores, tanto de la admisión como del paciente.

4.3 Creación de los nuevos *dataset*

Una vez se ha terminado de trabajar con la tabla más pesada, CHARTEVENTS, y por lo tanto en el servidor local, se ha pasado al software RStudio donde se han generado tres *datasets* de interés con el fin de desarrollar diferentes modelos de aprendizaje automático.

En primer lugar, se ha obtenido la última columna, común para todos los *datasets*. Esta contiene la variable (flag) sobre la cual se van a basar los modelos posteriores para clasificar a los pacientes. Representa la mortalidad hospitalaria en cada una de las estancias presentes en el *dataset*. El flag tiene un valor de 1 si el paciente ha fallecido durante esa estancia, mientras que el valor es de 0 si ha sobrevivido hasta el alta.

Asimismo y válido para los tres *datasets*, cada una de las filas corresponde a una estancia en la UCI distinta. Estas estancias están identificadas por el identificador único icustay_id, y corresponden al paciente identificado como subject_id. Cabe destacar, y

siendo válido también para todos los *datasets*, que se ha añadido una columna relacionada con la edad del paciente en el momento de la admisión en la UCI.

4.3.1 *Dataset* A: valores fisiológicos

En esta primera tabla (Figura 4-1.A) se ha incluido toda la información relacionada con las variables fisiológicas del paciente. Al estar las variables fisiológicas en forma de series temporales, y para poder posteriormente utilizar el *dataset* para los diferentes modelos, se ha optado por parametrizar cada una de las señales. Es por ello que se ha tomado los siguientes valores clave que permiten definir cada una de las series: mínimo (min), máximo (max), media (mean) y desviación estándar (std).

El primer objetivo ha sido buscar datos anómalos, en inglés *outliers,* en las series temporales de las variables fisiológicas de cada estancia de los pacientes. En estos casos, se ha tomado la decisión de corregir estos valores, si se encuentran aislados, sustituyéndolos por la media de la estancia del paciente (*mean imputation*).

A continuación, se ha parametrizado las señales mediante los valores comentados en 4.2.2 y se ha continuado con la búsqueda de valores *outliers* en estos parámetros. Estos valores *outliers* pueden ser indicativos de errores de registro y pueden afectar en los resultados del modelo. Para ello se ha realizado análisis exploratorio mediante un número de gráficos explicativos utilizado el paquete *ggplot2* de R.

Por otro lado, se ha comprobado la cantidad de datos fisiológicos faltantes, presentes como *NA* en la base de datos. Al ser series temporales, se ha optado por el siguiente criterio para combatir los datos NA: si en una serie temporal (excluyendo el peso) existen más de 6 horas sin registro de datos, se eliminaba la estancia de ese paciente. Se ha optado por este criterio debido a que, en un periodo de 6 horas, un paciente ingresado en la UCI puede generar información muy relevante sobre su estado de salud y se ha considerado esta como una ventana de tiempo adecuada.

El *dataset* A final tiene, por lo tanto, 41 columnas (siendo cada una una estancia en particular): un identificador de la fila, un identificador de la estancia en la UCI, un identificador del paciente, la edad del paciente, 36 columnas correspondientes a los parámetros definidos de las variables fisiológicas, y una columna de “flag”.

4.3.2 *Dataset* B: codificación ICD-9

En este *dataset* se ha utilizado la información relacionada con los códigos ICD-9 presentes en la tabla DIAGNOSES_ICD.

Los códigos ICD-9 son un sistema creado por la Organización Mundial de la Salud que ha sido diseñado fundamentalmente para mejorar la comparación de las estadísticas de mortalidad internacionalmente. Cada uno de los diagnósticos emitidos para el paciente son traducidos a códigos médicos basándose en la estructura de la ICD-9. Estos códigos están divididos en 19 capítulos distintos, como se puede ver en la Tabla 4-2. Cada uno de estos capítulos están a su vez subdivididos en subcapítulos. Por ejemplo, el código 011 corresponde al diagnóstico de tuberculosis pulmonar, que se encuentra dentro del subcapítulo tuberculosis (010 – 018), que a su vez se encuentra dentro del capítulo de enfermedades infecciosas y parasitarias (001 – 139).

En este caso en particular, cada una de las admisiones tiene relacionados varios códigos ICD-9. En primer lugar, se ha asociado todos los códigos con su capítulo correspondiente. Posteriormente, se han pasado estas variables a variables *dummy* es decir, se ha creado una columna por cada capítulo existente, tomando el valor 1 si el código se encuentra en el capítulo en cuestión y 0 si no. El *dataset* B (Figura 4-1.B) se ha creado teniendo 22 columnas (siendo cada una una estancia en particular): un identificador de la fila, un identificador de la estancia en la UCI, un identificador del paciente, la edad del paciente, 17 columnas correspondientes a los capítulos de las códigos ICD-9 y una columna de “flag”.

Tabla 4-2. División en capítulos de la codificación International Classification of Disease, 9th Edition

Capítulo	Rango de los códigos	Descripción
1	001 – 139	Enfermedades Infecciosas y Parasitarias
2	140 – 239	Neoplasias
3	240 – 279	Enfermedades endocrinas, nutricionales y metabólicas, trastornos inmunitarios
4	280 – 289	Enfermedades de la sangre y de los orgánulos formadores de sangre
5	290 – 319	Trastornos mentales
6	320 – 389	Enfermedades del sistema nerviosos y órganos del sentido
7	390 – 459	Enfermedades del sistema circulatorio
8	460 – 519	Enfermedades del sistema respiratorio
9	520 – 579	Enfermedades del sistema digestivo
10	580 – 629	Enfermedades del sistema genital-urinario
11	630 – 679	Complicaciones del embarazo, nacimiento y puerperio
12	680 – 709	Enfermedades de la piel y tejido subcutáneo
13	710 – 739	Enfermedades del sistema musculo-esquelético y del tejido conectivo
14	740 - 759	Anomalías congénitas
15	760 – 779	Ciertas condiciones originadas en el periodo perinatal
16	780 – 799	Síntomas, signos y condiciones definidas
17	800 - 999	Lesiones y envenenamientos

4.3.3 *Dataset* C: variables fisiológicas y codificación ICD-9

Este último *dataset* (Figura 4-1.C) se obtiene de la combinación de los dos anteriores, mezclando tanto variables fisiológicas como las variables *dummy* derivadas de los códigos ICD-9. Al tener el *dataset* B menores restricciones en cuanto al filtrado de estancias, y por lo tanto, un número mayor de filas, se ha optado por utilizar las mismas estancias presentes en el *dataset* A, eliminando así estancias del *dataset* B.

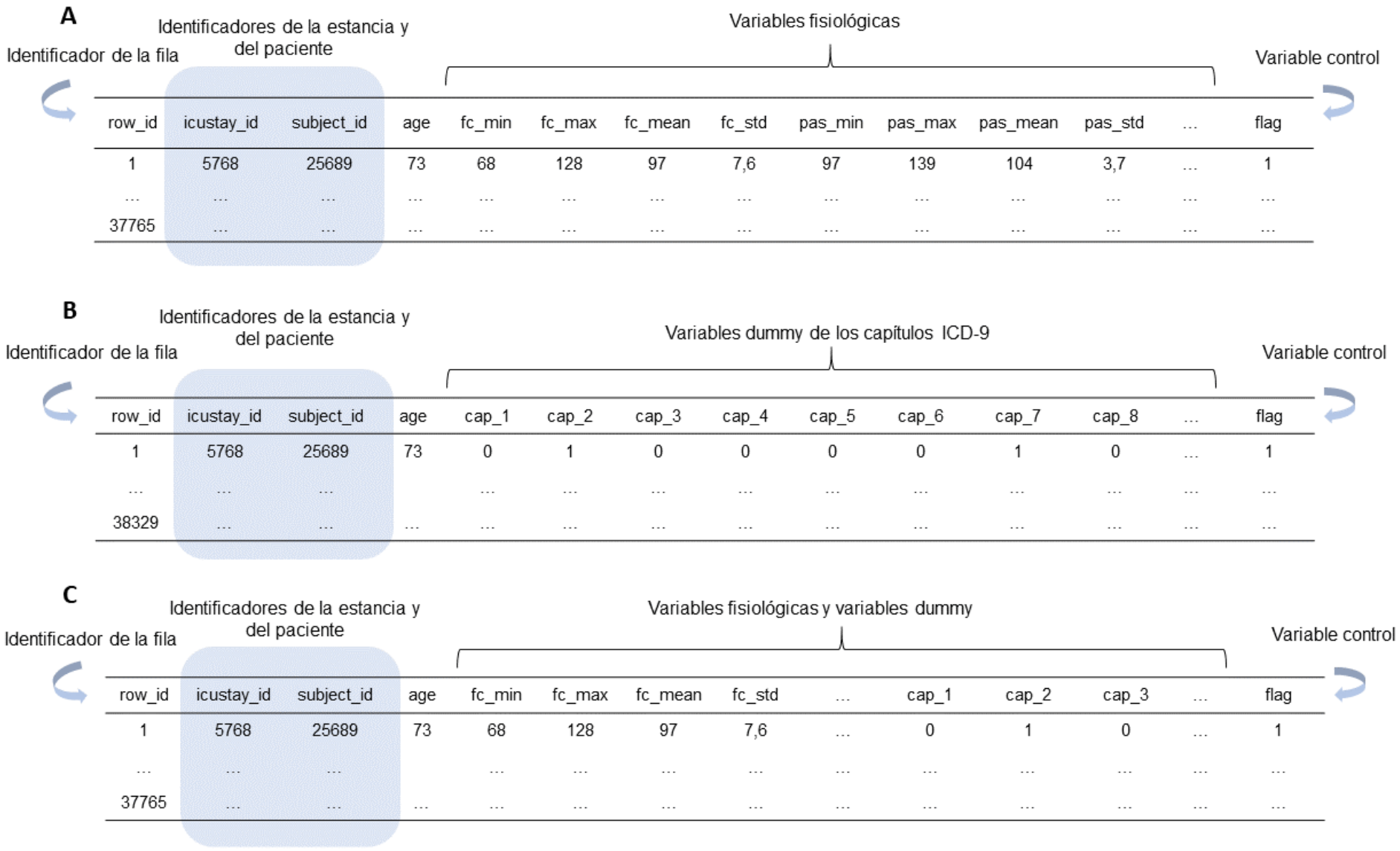

Figura 4-1. Estructura de los 3 datasets generados, siendo las columnas 1, 2 y 3 de cada dataset identificadores, la última columna la clase control (paciente fallecido o no fallecido), y las columnas centrales las variables a tener en cuenta por los modelos. A) Dataset A, B) Dataset B, C) Dataset C.

4.4 Modelos de predicción de mortalidad

Para cada uno de los tres *datasets* diferentes se han desarrollado diferentes modelos. Estos *datasets* han sido tratados de la misma forma a la hora de proceder con la separación de los datos de entrenamiento y de los datos de validación. Dado el gran tamaño muestral, se ha optado por un particionamiento *hold-out* frente a una validación cruzada. De este modo, se ha separado un 70% de los datos para el set de entrenamiento y un 30% de los datos para el set de validación. Esta separación ha sido aleatoria, sin embargo, se ha garantizado que los pacientes que presentan varios ingresos distintos en la UCI se encuentran dentro del mismo set para no influir en los resultados.

El cálculo de las tres métricas determinadas para comprobar el rendimiento de los modelos se basa en la matriz de confusión mostrada en la Figura 4-2:

		Predicho	
		Positivo	**Negativo**
Real	**Positivo**	Verdadero Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadero Negativo (VN)

Figura 4-2. Matriz de confusión

Las métricas utilizadas han sido las siguientes:

- *Accuracy*: cantidad de aciertos del modelo respecto a la cantidad de observaciones totales

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \qquad (Ec. 4)$$

- *Precision*: cantidad de positivos verdaderos respecto a los positivos predichos totales

$$Precision = \frac{VP}{VP + FP} \qquad (Ec. 5)$$

- *Recall*: cantidad de positivos respecto a los positivos actuales totales

$$Recall = \frac{VP}{VP + FN} \qquad (Ec. 6)$$

4.4.1 Modelos utilizando el *dataset* A

Las variables fisiológicas presentes en este *dataset* son todas numéricas. Por lo tanto, se ha procedido con un modelado en primera estancia mediante *LDA*, *QDA* y *KNN.*

En primer lugar, se ha reducido la dimensionalidad de la matriz, debido al gran número de columnas. Para ello se ha realizado sobre el *dataset* un Análisis de Componentes Principales (cuyas siglas en inglés son PCA) (Pearson K., 1901). De esta reducción de dimensionalidad se ha derivado un nuevo *dataset* formado por las ocho primeras componentes que ha sido utilizado para el modelado.

Posteriormente se ha entrenado y validado el modelo *LDA*. A su vez, se ha realizado lo mismo con el modelo *QDA*. Por último, se ha procedido al modelado mediante *KNN*. Para este último, se ha utilizado un parámetro k = 10 debido a que es el valor por defecto. Sin embargo, los alumnos en un futuro trabajo, podrán variar este parámetro y comprobar si existen así mejoras en los resultados.

4.4.2 Modelos utilizando el *dataset* B

Las variables predictoras del *dataset* B están formadas por variables *dummy*, con dos valores posible. Se ha utilizado los modelos *Random Forest* y *Gradient Boosting.*

El modelo *Random Forest* se ha ejecutado con una cantidad de 700 árboles de decisión. Además, se ha definido en 4 el número de variables utilizadas en cada iteración. Asimismo, para el modelo *Gradient Boosting* se han utilizado 700 árboles y una distribución de Bernoulli debido a que la salida de la regresión es binaria (0 o 1).

4.4.3 Modelos utilizando el *dataset* C

Por último, el *dataset* C, al tener información tanto numérica como categórica, ha sido empleado en la modelización mediante *Random Forest* y *Gradient Boosting.* Los parámetros utilizados en *Random Forest* han sido 1000 árboles y 6 variables por iteración, mientras que para el algoritmo *Gradient Boosting* se han utilizado 1000 árboles y una distribución de Bernoulli.

4.5 Calidad de los datos

Se ha llevado a cabo un análisis parcial de la calidad de los datos sobre las variables fisiológicas determinadas previamente de interés. Para ello, se ha medido mediante diferentes herramientas presentes en el lenguaje R la completitud y la exactitud (Sáez

et al., 2012) (Sáez, Robles, et al., 2017). Estas medidas se han realizado en un *dataset* A paralelo al creado anteriormente. En la creación de este *dataset* no se han utilizado criterios de eliminación. Las medidas de consistencia y la estabilidad temporal y multifuente se realizará en un trabajo futuro.

5. Modelado Predictivo

5.1 Resultados

5.1.1 Análisis estadístico general

En la Tabla 5-1 se puede observar la distribución básica de los pacientes en la base de datos MIMIC III separados en los tres grupos definidos.

Tabla 5-1. Análisis estadístico general basado en la distribución en tres grupos de la mortalidad de los pacientes

Grupo	Edad media (años) (Q1 – Q3)	Mujeres (%)	Media de días en el hospital (Q1 – Q3)	Media de días en la UCI (Q1 – Q3)	Número de pacientes (%)
Pacientes fallecidos en el hospital	74 (60–83)	45,5%	6 (2-13)	3,1 (1,3–7,5)	5842 (15%)
Pacientes fallecidos 90 días después del alta	76 (64-84)	46,2%	9 (5-16)	2,7(1,5-5,3)	3286 (8,2%)
Pacientes no fallecidos 90 días después del alta	64 (51-76)	42,5%	7 (4-11)	2 (1,2-3,5)	29469 (76.8%)

Cabe destacar que los pacientes fallecidos tanto en el hospital como a los 90 días de darles el alta son notablemente más mayores (74 (60-83) y 76 (64-84) años respectivamente) que los pacientes no que han fallecido a los 90 días, que son, de media, 10 años más jóvenes. La mortalidad tanto hospitalaria como a los 90 días se sitúa en un 23,2%.

Ahora bien, si observamos la mortalidad en cada una de las UCI existentes (Figura 5-1) se puede observar cómo la de mayor mortalidad es la, MICU que también es la que presenta un mayor número de ingresos (ver Figura 5-2).

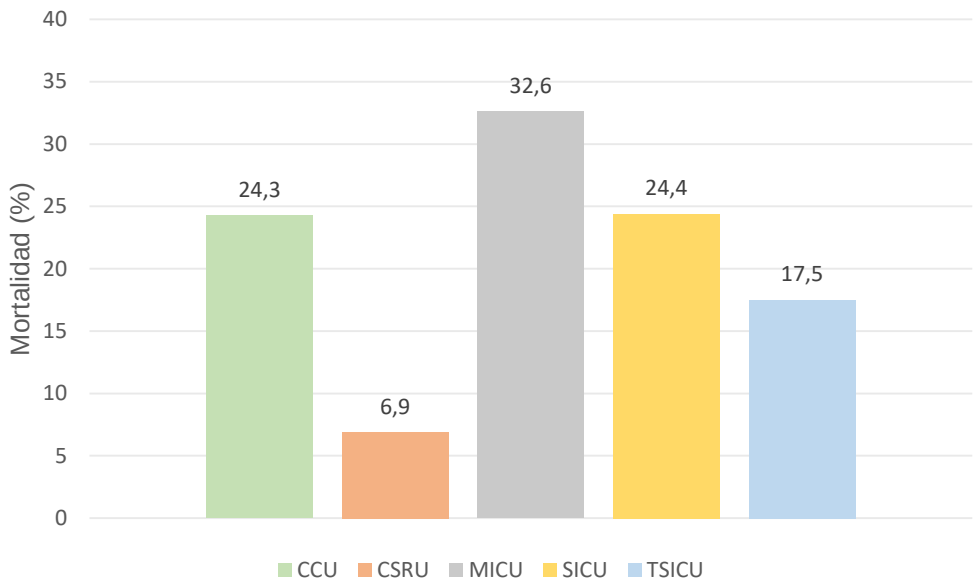

Figura 5-1. Mortalidad, tanto hospitalaria como a los 90 días, en las diferentes UCI. (CCU: Cardiac/Coronary Care Unit, CSRU: Cardiac Surgery Intensive Care Unit, MICU: Medical Intensive Care Unit, SICU: Surgical Intensive Care Unit, TSICU: Trauma and Surgical Intensive Care Unit)

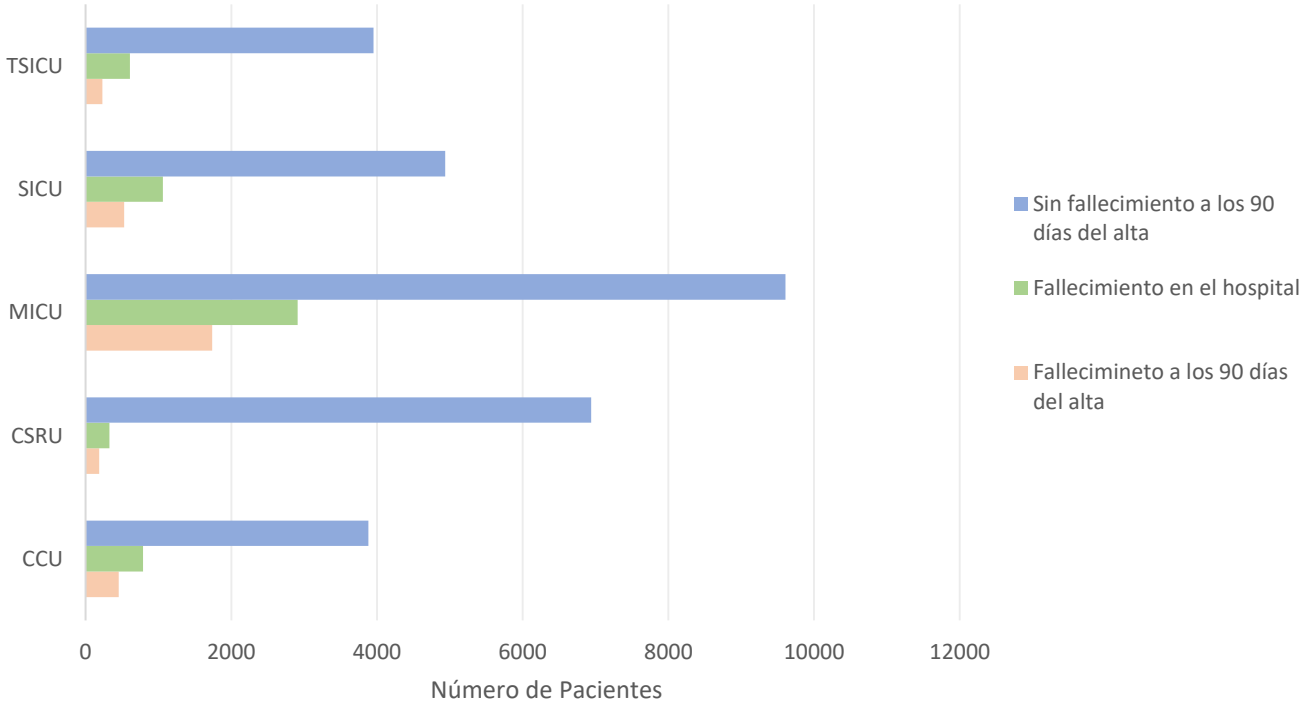

Figura 5-2. Distribución de los pacientes ingresados en la UCI

Cabe destacar que, la menor tasa de ingresos la presenta la unidad CCU teniendo sin embargo, una mortalidad que supera el 24%. Por otro lado, la unidad CSRU, teniendo el segundo mayor número de ingresados, tiene una mortalidad mucho menor que el resto (6,9%).

5.1.2 Dataset A: valores fisiológicos

En primer lugar, se han encontrado 107 valores *outliers* en las medidas obtenidas de las series de variables fisiológicas, con los que se ha procedido a realizar la *mean imputation*. Además, se han encontrado 23 *outliers* en los parámetros extraídos de las series temporales de medidas fisiológicas en el *dataset*. Al ser un número tan pequeño en comparación con la cantidad de estancias disponibles, se ha procedido a su eliminación. En segundo lugar, la cantidad de estancias con valores NA durante más de seis horas ha sido de 541. Se ha procedido con la eliminación de esas estancias a su vez, resultando en una cantidad de 564 estancias eliminadas.

A continuación, podemos observar en la Figura 5-3 como la primera variable resultante de la PCA permite explicar un 62,1% de la varianza, mientras que con 8 variables se consigue explicar el 95% de la varianza.

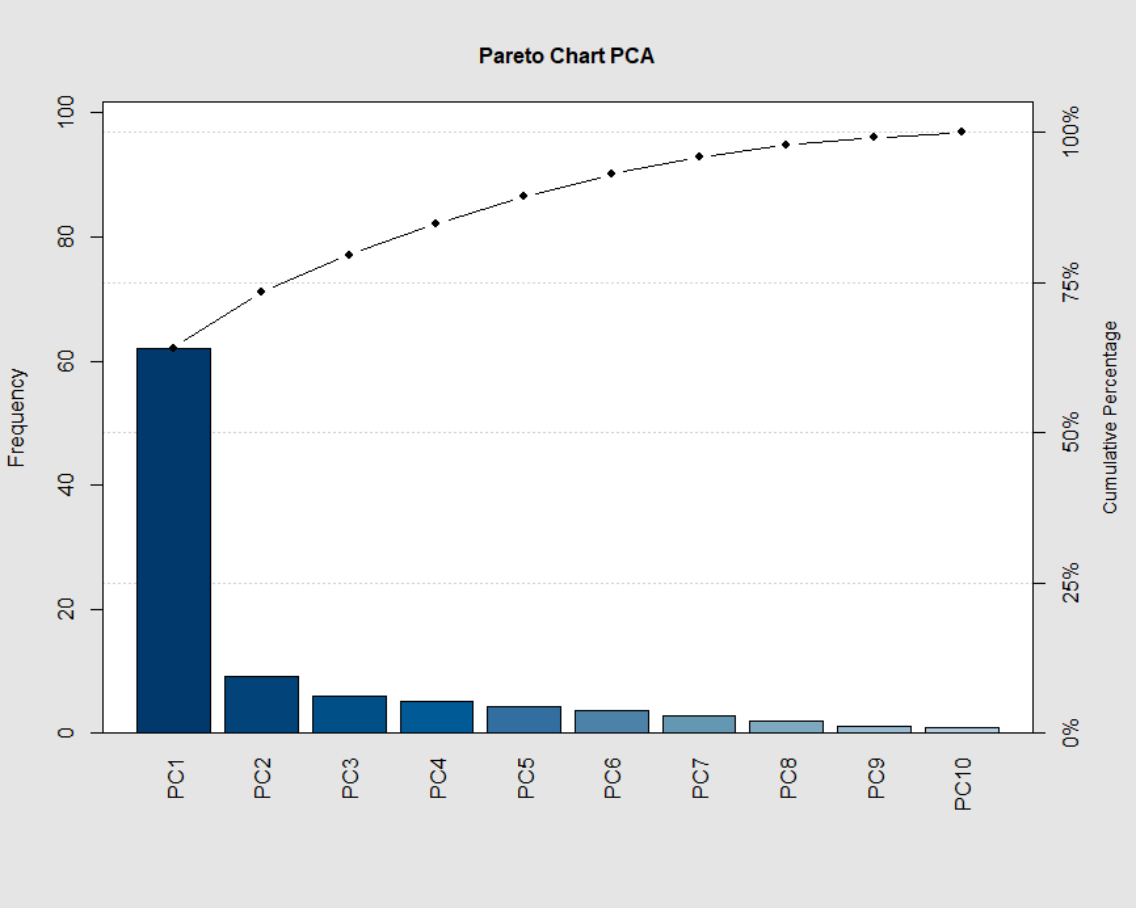

Figura 5-3. Gráfica Pareto mostrando la varianza explicada por cada variable generada por la PCA

Con estas 8 variables se ha procedido a la realización de los tres modelos introducidos anteriormente y cuyos resultados se pueden observar en la Tabla 5-2.

Tabla 5-2. Resultados de los modelos desarrollados para el dataset A

	ACCURACY	PRECISION	RECALL
LDA	0,68	0,61	0,63
QDA	0,63	0,59	0,57
KNN	0,76	0.74	0,75

5.1.3 *Dataset* B: codificación ICD-9

Los resultados de los dos modelos desarrollados con este *dataset*, donde no se ha realizado ninguna reducción de dimensionalidad, se pueden ver en la Tabla 5-3.

Tabla 5-3. Resultados de los modelos desarrollados para el dataset B

	ACCURACY	PRECISION	RECALL
RANDOM FOREST	0,79	0,73	0,75
GRADIENT BOOSTING	0,78	0,73	0,74

Los resultados mejoran sensiblemente con respecto a los modelos desarrollados con el primer *dataset* y resultan ser muy parecidos entre ellos.

5.1.4 *Dataset* C: valores fisiológicos y codificación ICD-9

Como se puede ver en la Tabla 5-4, los resultados derivados de la conjunción de los *datasets* A y B mejoran notablemente.

Tabla 5-4. Resultados de los modelos desarrollados para el dataset C

	ACCURACY	PRECISION	RECALL
RANDOM FOREST	0,89	0,85	0.86
GRADIENT BOOSTING	0,87	0,82	0,84

5.1.5 Calidad de datos

En un primer lugar se ha comprobado con éxito que los identificadores de paciente, de admisión tanto en el hospital como en la UCI son únicos.

En segundo lugar, se ha obtenido el porcentaje de pacientes donde no existía ninguna medida de las variables fisiológicas determinadas (Tabla 5-5). Las variables de edad y género no presentan ningún valor faltante.

Tabla 5-5. Porcentaje de pacientes sin ninguna medida en las variables fisiológicas de interés

Variables fisiológicas	Porcentaje de pacientes sin datos
Frecuencia Cardiaca	1,6%
Presión Arterial Sistólica	1.6%
Presión Arterial Diastólica	1.6%
Presión Arterial Media	1,6%
Frecuencia Respiratoria	1,7%
Temperatura	1,6%
Peso	1,9%
Saturación de Oxígeno	2,1%
Glucosa en sangre	2,5%

A su vez, se ha comprobado la existencia de valores anómalos en las variables fisiológicas, resultando en unos porcentajes inferiores al 1% en todas las variables.

5.2 Discusión

5.2.1 Modelización

En el trabajo desarrollado en este estudio, se ha intentado desarrollar una serie de modelos con el objetivo principal de que los alumnos puedan, en un futuro, reproducir sus fases de desarrollo y aprender de la experiencia propuesta. Estos modelos han tomado como partida y se han inspirado en los desarrollados anteriormente en esta misma base de datos (Johnson et al., 2017).

Los resultados obtenidos con los *dataset* A y B no resultan ser los esperados para un modelo CDSS robusto en comparación con los resultados del *dataset* C. La

utilización de las variables fisiológicas escogidas en solitario en el *dataset* A no ofrece resultados de gran precisión a la hora de desarrollar los modelos. A pesar de que los resultados obtenidos con el modelo KNN no son malos si se contrastan con los obtenidos ante el mismo *dataset* A con los modelos LDA y QDA, parametrizar las diferentes series temporales de manera diferente y añadir nuevas variables fisiológicas obtenidas de medidas de laboratorio podría mejorar sustancialmente los resultados del modelo.

Por otro lado, el *dataset* B ofrece una visión de la morbilidad y comorbilidad de los pacientes seguramente demasiado amplia, utilizando únicamente los capítulos principales de la codificación ICD-9. Resultaría interesante realizar un modelo de mayor complejidad, exclusivamente centrado en ICD-9, pero aumentando la precisión de esta morbilidad/comorbilidad aumentando el rango de variables a los subcapítulos de los códigos ICD-9.

Sin embargo, el *dataset* C ofrece unos resultados acordes a los esperado, situándose prácticamente a la par con otros estudios realizados hasta la fecha con esta base de datos (Johnson et al., 2017). La *accuracy* de ambos modelos supera el 85%, llegando prácticamente al 90% en el caso de *Random Forest.* Además, los valores tanto de precisión como de *recall* se mantienen por encima del 80%.

Así pues, resulta importante destacar que la conjunción de las variables fisiológicas escogidas y la terminología de diagnóstico ofrece buenos resultados. Tanto *Random Forest* como *Gradient Boosting* resultan ser dos algoritmos utilizados frecuentemente en estudios similares (Lin et al., 2019), también con resultados prometedores. Además, el dataset C está formado por los datos presentes en el *dataset* A y B, por lo que las sugerencias mencionadas para la mejora de los resultados de estos dos datasets podría afectar de manera positiva a los resultados del *dataset* C.

Sin embargo, es complicado comparar los resultados obtenidos por estos modelos y que esta comparación sea fidedigna debido a las diferencias en la realización tanto en la extracción de los datos como en el desarrollo de los modelos. En general, la mayoría de *papers* explican sus metodologías con un detalle bastante limitado y con una heterogeneidad muy amplia. Se omiten detalles importantes para comprender la cohorte seleccionada, como la estancia mínima requerida para un paciente, como se tratan las readmisiones, etc.

En este trabajo en particular, para aumentar la calidad de los resultados obtenidos, una opción interesante sería endurecer los criterios a la hora de seleccionar la cohorte, realizando así modelos de mayor especificidad. Desarrollando un modelo, por ejemplo,

para cada tipo de UCI existente, pudiendo obtener, por ende, diferentes variables fisiológicas centradas en las enfermedades más frecuentes en esas UCI.

5.2.2 Calidad de datos

La calidad de los datos está inherentemente ligada al contexto a la finalidad y al uso de estos (Kahn et al., 2015). Resulta complicado, sin embargo, un análisis exhaustivo de la calidad en una base de datos con tal cantidad de tablas y tanta información como en MIMIC III. Es por ello que se ha optado por un análisis de mayor simplicidad centrándose en el *dataset* A sin filtrar. Los resultados obtenidos destacan, como se menciona en (Johnson et al., 2016), el trabajo que se ha llevado a cabo en esta base de datos para mejorar la calidad de los mismos. Será de gran interés para los alumnos, en un futuro, aplicar las medidas de calidad de datos y analizar las diferentes dimensionalidades sobre la base de datos completa.

Cabe destacar que, debido al estricto proceso de desidentificación al que se ha sometido esta base de datos, resultando, entre otras, en una aleatoriedad de todas las fechas presentes, no se ha podido llevar a cabo un análisis de la variabilidad temporal de los datos.

.

6. Propuesta de aprendizaje

6.1 Resultados

La propuesta de aprendizaje planteada para las asignaturas de SIT y DQI parte de replicar la experiencia desarrollada en este TFG con la base de datos MIMIC III, cubriendo así los objetivos de aprendizaje de ambas asignaturas tal y como se puede observar en la Tabla 6-1.

La tercera y más extendida unidad didáctica (UD) de la asignatura SIT está enfocada en los CDSS. Además, las UD1 y UD2 se centran en el manejo de bases de datos biomédicos y el procedimiento a la hora de trabajar con diferentes terminologías, como puede ser ICD-9 o códigos DRGs, respectivamente. Además, las UDs 1 y 2 de DQI cubren la descripción y comprensión de las dimensiones de calidad de datos como son completitud, consistencia, exactitud, contextualización y estabilidad temporal y multifuente. Como se puede ver en la Tabla 6-1, la utilización de los datos biomédicos ofrecidos por el proyecto MIMIC engloba las UDs de mayor peso tanto de SIT como de DQI.

Tabla 6-1. Unidades Didácticas de SIT y DQI cubiertas por el proyecto

Unidades Didacticas	**Tarea desarrollada por el proyecto MIMIC III**	**Objetivo de aprendizaje específico cubierto**
SIT – UD 1: Organización de los Sistemas de Información en Salud	Comprensión y obtención de los datos en MIMIC III, creando bases de datos para modelos de predicción	Manejo y utilización de bases de datos hospitalarias
SIT – UD 2: Sistemas de Historia Clínica Electrónica	Utilización del estándar ICD-9	Descripción y utilización de los estándares de la historia clínica electrónica
SIT – UD 3: CDSS	Desarrollo de modelos de predicción hospitalaria utilizado diferentes algoritmos de “machine learning”	Desarrollo de modelos de predicción y CDSS utilizando algoritmos de “machine learning”

DQI – UD 1: Introducción a la Calidad de Datos y las Dimensiones de Calidad de Datos	Descripción de las dimensiones de calidad de datos	Descripción de los diferentes aspectos de calidad de datos y clasificación en las diferentes dimensiones
DQI – UD 2: Dimensiones de Calidad de Datos	Medición y ajuste de completitud, consistencia, exactitud, contextualización y estabilidad temporal y multifuente	Medición de las dimensiones de calidad de datos y saneamiento de los datos

En relación a las competencias en Ingeniería Biomédica por la UPV, este proyecto ha cubierto principalmente :

- 40(ES) Capacidad para el auto-aprendizaje, la consolidación y la actualización de nuevos conocimientos en el área de la ingeniería biomédica, y para emprender estudios posteriores con alto grado de autonomía.
- 43(GE) Capacidad para el aprendizaje de nuevas técnicas y herramientas de análisis, modelización, diseño y optimización.
- 5(ES) Poseer conocimientos de herramientas informáticas para analizar, calcular, visualizar, representar y obtener la información necesaria para apoyar las tareas de análisis, cálculo, diseño, desarrollo y gestión relacionadas con la ingeniería biomédica.
- 8(ES) Capacidad de integrar conocimientos multidisciplinares asociados a la ingeniería, biología y medicina.
- 11(ES) Ser capaz de entender las características técnicas y funcionales de los sistemas, métodos y procedimientos que se utilizan en prevención, diagnóstico, terapia y rehabilitación.

6.2 Discusión

En relación al desarrollo del nuevo ABP, resulta haber sido un éxito en relación a la propuesta inicial, demostrando que los datos presentes en la base de datos MIMIC III permiten ajustarse de gran manera a los objetivos de aprendizaje tanto de SIT como de DQI. La relación tan complementaria que existe entre SIT y DQI permitirá en un futuro mejorar la adquisición de conocimientos por parte de los estudiantes durante la puesta en práctica de la metodología ABP diseñada durante el Grado y el Máster. Además,

resultaría interesante realizar el mismo trabajo con versiones anteriores de MIMIC III, debido a que la versión utilizada en este trabajo, la MIMIC III v.1.4, resulta estar íntegramente enfocada al análisis de la calidad de los datos y su saneamiento, por lo que versiones anteriores podrían ofrecer mejores resultados a la hora de la enseñanza. A su vez, también puede ser de interés reproducir este proyecto en la versión posterior aparecida en Agosto de 2020, MIMIC IV. Esta versión contiene una mayor cantidad de datos tomados sobre pacientes hasta el año 2019.

El trabajo desarrollado en este estudio ofrece la capacidad no solo de mostrar ejemplos teóricos basados en estos datos, si no crear trabajos finales de asignatura semi guiados que prosigan los pasos seguidos en el estudio, mostrando a los estudiantes las dificultades reales y los métodos adecuados a la hora de tratar datos biomédicos reales y tangibles.

En la actualidad, no se tiene constancia de ninguna metodología de enseñanza como la desarrollada en este trabajo que utilice la información disponible en la base de datos MIMIC III. Se ha encontrado únicamente un curso en la plataforma Coursera, “Clinical Data Models and Data Quality Assessments” (*Clinical Data Models and Data Quality Assessments* | *Coursera*, 2019), basado en MIMIC III, centrado en modelización y ciertos aspectos de la calidad de datos.

7. Líneas Futuras

Basado en los resultados llevados a cabo en este trabajo, sería de interés el desarrollo de modelos de predicción de mortalidad más específicos. Centrándose, por ejemplo, en la unidad de cuidados intensivos con mayor mortalidad, como es la MICU o en alguna enfermedad en particular. A su vez, se podría estudiar la mortalidad en diferentes momentos de la estancia en la UCI (a las 24h o 48h) o en el alta.

A su vez, resultaría interesante estudiar las señales fisiológicas temporales en función de lo impartido en diferentes asignaturas de la carrera, enlazando así este trabajo con asignaturas donde se analicen tipos de datos más específicamente.

Por otro lado, y a pesar de las limitaciones que crea la desidentificación en relación a las fechas, se podría estudiar si la estacionalidad tiene alguna influencia en los resultados de los modelos de predicción.

Por último, el siguiente paso sería testear el ABP desarrollado con los estudiantes, estudiar y comparar su curva de aprendizaje con el ABP utilizado actualmente, y recoger sus opiniones para mejorar el nuevo método.

8. Conclusiones

El objetivo principal de este trabajo ha sido el desarrollo de un nuevo ABP que abarque tanto a la asignatura SIT como a la asignatura DQI ofrecidas en el Grado y Master en Ingeniería Biomédica en la Universitat Politècnica de València. Este objetivo se ha cumplido, mostrando cómo los objetivos de aprendizaje y las unidades didácticas de ambas asignaturas concuerdan en gran medida con el trabajo desarrollado. Además, se ha conseguido extraer tres tablas de datos basadas desde la base de datos pública MIMIC III diseñados para la creación de CDSS y con los que los alumnos serán capaces de trabajar, ya que puede resultar de gran complejidad utilizar datos en crudo tal y como se proporcionan. Utilizar estos *dataset* de manera continua, tanto como ejemplos prácticos como teóricos, en SIT y DQI, permitirá mejorar la curva de aprendizaje de los alumnos en estas áreas de conocimiento biomédico.

Por otro lado, se han desarrollado diferentes modelos de predicción de mortalidad con el *dataset* MIMIC III con una capacidad de predicción alta equivalente al estado del arte. Resulta prometedor la obtención de estos buenos resultados debido al margen de mejora todavía existente, el cual puede obtenerse con consecuentes trabajos de los alumno siguiendo el ABP propuesto. Esto abre las puertas al desarrollo de CDSS más robustos y precisos por profesionales de la Ingeniería Biomédica que puedan mejorar la eficacia en la atención al paciente y la utilización de los recursos médicos disponibles.

9. Bibliografía

Alvear-Vega, S., & Canteros-Gatica, J. (2018). Performance evaluation of APACHE II and SAPS III in an intensive care unit. *Revista de Salud Publica*, *20*(3), 373–377. https://doi.org/10.15446/rsap.v20n3.59952

Big hopes for big data. (2020). *Nature Medicine*, *26*(1), 1–1. https://doi.org/10.1038/s41591-019-0740-8

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating Project-Based Learning: Sustaining the Doing, Supporting the Learning. *Educational Psychologist*, *26*(3–4), 369–398. https://doi.org/10.1080/00461520.1991.9653139

Bricon-Souf, N., & Newman, C. R. (2007). Context awareness in health care: A review. In *International Journal of Medical Informatics* (Vol. 76, Issue 1, pp. 2–12). Elsevier. https://doi.org/10.1016/j.ijmedinf.2006.01.003

Caballero, K., & Akella, R. (2015). Dynamically modeling Patient's health state from electronic medical records: A time series approach. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 69–78. https://doi.org/10.1145/2783258.2783289

Clinical Data Models and Data Quality Assessments | Coursera. (2019). https://www.coursera.org/learn/clinical-data-models-and-data-quality-assessments

ICU Outcomes | Philip R. Lee Institute for Health Policy Studies. (n.d.). Retrieved August 22, 2020, from https://healthpolicy.ucsf.edu/icu-outcomes

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer*. https://www.springer.com/gp/book/9781461471370%0Ahttp://www.springer.com/us/book/9781461471370

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2020). *MIMIC-IV v0.4*. https://physionet.org/content/mimiciv/0.4/

Johnson, A. E. W., Pollard, T. J., & Mark, R. G. (2017). Reproducibility in critical care: a mortality prediction case study. *Proc. Mach. Learn. Res.*, *68*, 361–376.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 1–9.

https://doi.org/10.1038/sdata.2016.35

K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. (n.d.). Retrieved August 26, 2020, from https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., Schilling, L. M., Weiskopf, N. G., Williams, A. E., & Zozus, M. N. (2015). Transparent Reporting of Data Quality in Distributed Data Networks. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *3*(1), 7. https://doi.org/10.13063/2327-9214.1052

Laboratory For Computational Physiology, M. (2015). *The MIMIC III Clinical Database*. 1–14. https://doi.org/10.13026/C2XW26

Lin, K., Hu, Y., & Kong, G. (2019). Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *International Journal of Medical Informatics*, *125*, 55–61. https://doi.org/10.1016/j.ijmedinf.2019.02.002

Moody, G. B., & Mark, R. G. (1996). A database to support development and evaluation of intelligent intensive care monitoring. *Computers in Cardiology 1996*, 657–660. https://doi.org/10.1109/CIC.1996.542622

Niewiński, G., Starczewska, M., & Kański, A. (2014). Prognostic scoring systems for mortality in intensive care units - The APACHE model. In *Anaesthesiology Intensive Therapy* (Vol. 46, Issue 1, pp. 46–49). Via Medica. https://doi.org/10.5603/AIT.2014.0010

Pearson K. (1901). Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.

PhysioNet. (2009). *MIMIC II Databases*. https://archive.physionet.org/mimic2/

PhysioNet. (2015). *The MIMIC Database*. https://doi.org/10.13026/C2JS34

Sadeghi, R., Banerjee, T., & Romine, W. (2018). *Early hospital mortality prediction using vital signals*. https://doi.org/10.1016/j.smhl.2018.07.001

Sáez, C., Mañas, A., Muñoz-soler, V., & García-, J. M. (2017). *Project-based learning based on a national pilot project for the data quality control and standardization of maternal and child information applied to Biomedical Engineering University teaching*. *October*, 75–85.

Sáez, C., Martínez-Miranda, J., Robles, M., & García-GóMez, J. M. (2012). Organizing

data quality assessment of shifting biomedical data. *Studies in Health Technology and Informatics*, *180*, 721–725. https://doi.org/10.3233/978-1-61499-101-4-721

Sáez, C., Robles, M., & García-Gómez, J. M. (2017). Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Statistical Methods in Medical Research*, *26*(1), 312–336. https://doi.org/10.1177/0962280214545122

Silva, I., Moody, G., Scott, D. J., Celi, L. A., & Mark, R. G. (2012). Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in cardiology challenge 2012. *Computing in Cardiology*, *39*, 245–248.

Slee, V. N. (1978). The International Classification of Diseases: ninth revision (ICD-9). In *Annals of internal medicine* (Vol. 88, Issue 3, pp. 424–426). https://doi.org/10.7326/0003-4819-88-3-424

Thomas, M. M., Kannampallil, T., Abraham, J., & Marai, G. E. (2017). Echo: A large display interactive visualization of ICU data for effective care handoffs. *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*, 47–54. https://doi.org/10.1109/VAHC.2017.8387500

Thorwarth, W. T. (2008). CPT®: An Open System That Describes All That You Do. In *Journal of the American College of Radiology* (Vol. 5, Issue 4, pp. 555–560). Elsevier. https://doi.org/10.1016/j.jacr.2007.10.004

Vega, F., Portillo, E., Cano, M., & Navarrete, B. (2014). Experiencias de aprendizaje en ingeniería química: Diseño, montaje y puesta en marcha de una unidad de destilación a escala laboratorio mediante el aprendizaje basado en problemas. *Formacion Universitaria*, *7*(1), 13–22. https://doi.org/10.4067/S0718-50062014000100003

Wasylewicz, A. T. M., & Scheepers-Hoeks, A. M. J. W. (2018). Clinical decision support systems. In *Fundamentals of Clinical Data Science* (pp. 153–169). Springer International Publishing. https://doi.org/10.1007/978-3-319-99713-1_11

10. Anexo

10.1 Resumen de las tablas presentes en la base de datos MIMIC III

Nombre de la tabla	Tamaño (MB)	Número de variables	Descripción
ADMISSIONS	12,3	19	Admisiones en el hospital asociadas con una estancia en la UCI
CALLOUT	6,2	24	Registro del momento en el que los pacientes estaban preparados para el alta y tiempo real del alta
CAREGIVERS	0,2	4	Listado de cuidadores asociados con una estancia en la UCI
CHARTEVENTS	34480	15	Eventos que ocurren en la historia clínica de un paciente
CPTEVENTS	56,8	12	Eventos registrados con la CPT
D_CPT	0,014	9	Diccionario de alto nivel de la CPT
D_ICD_DIAGNOSES	1,4	4	Diccionario de la ICD-9 (diagnósticos)
D_ICD_PROCEDURES	0,3	4	Diccionario de la ICD-9 (procedimientos)
D_ITEMS	0,9	10	Diccionario de los ítems no relacionados con el laboratorio
D_LABITEMS	0,043	6	Diccionario de los ítems relacionados con el laboratorio
DATETIMEEVENTS	513,5	14	Eventos relacionados con una fecha
DIAGNOSES_ICD	18,7	5	Diagnósticos relacionados con la admisión en el hospital y codificados con el sistema ICD-9

DRGCODES	10,2	8	Estancias en el hospital clasificadas utilizando el sistema DRG
ICUSTAYS	6,2	12	Listado de las admisiones en la UCI
INPUTEVENTS_CV	2407	22	Eventos relacionados con administración de fluidos a pacientes cuyos datos fueron guardados originalmente en la base de datos CareVue
INPUTEVENTS_MV	952,4	31	Eventos relacionados con administración de fluidos a pacientes cuyos datos fueron guardados originalmente en la base de datos MetaVision
LABEVENTS	1810,8	9	Eventos relacionados con test de laboratorio
MICROBIOLOGYEVENTS	70,8	16	Eventos relacionados con test de microbiología
NOTEVENTS	3913,8	11	Anotaciones asociadas con estancias en el hospital
OUTPUTEVENTS	387,1	13	Outputs registrados durante la estancia en la UCI
PATIENTS	2,6	8	Pacientes asociados con una estancia en la UCI
PRESCRIPTIONS	752,3	19	Medicamentos prescritos
PROCEDUREEVENTS_MV	47,6	25	Tiempos de inicio y final de los procedimientos registrados para los pacientes de la base de datos MetaVision
PROCEDURES_ICD	6,6	25	Procedimientos relacionados con la admisión en el hospital codificados usando el sistema ICD-9
SERVICES	3,4	6	Servicios bajo los que los pacientes estuvieron en su estancia en el hospital
TRANSFERS	24,5	13	Localización de los pacientes durante su estancia en el hospital

10.2 Certificado de “Human Research Data or Specimens Only Research 1 – Basic Course”, necesario para el acceso a la base de datos

CITI PROGRAM

Completion Date 12-Oct-2019
Expiration Date 11-Oct-2022
Record ID 33692534

This is to certify that:

Luis Alcalá

Has completed the following CITI Program course:

Human Research (Curriculum Group)
Data or Specimens Only Research (Course Learner Group)
1 - Basic Course (Stage)

Under requirements set by:

Massachusetts Institute of Technology Affiliates

CITI
Collaborative Institutional Training Initiative

Verify at www.citiprogram.org/verify/?w33fee887-8ae0-4927-83a7-75d5b5411121-33692534

PRESUPUESTO

1. Introducción

Todos los costes relacionados con la elaboración del presente trabajo se recogen en este documento. La estimación económica del Trabajo Final de Grado es uno de los objetivos de este. En función de la naturaleza del trabajo desarrollado, los tipos de costes pueden variar, en este en concreto, se han tomado en consideración tres tipos de costes bien diferenciados:

- Coste de personal
- Coste de *hardware*
- Coste de *software*

Por último, se ha desarrollado un presupuesto final donde se recopilan todos los costes mencionados anteriormente.

2. Presupuesto detallado

2.1 Coste de personal

Se van a detallar los costes correspondientes a las personas que han tomado parte en la elaboración del proyecto (tabla 1), siendo estas:

- D. Carlos Sáez Silvestre: Investigador en Ciencias Biomédicas, realizando las tareas de planteamiento, guiado, supervisión y corrección del trabajo.
- D. Luis Alcalá Pérez: Estudiante del Grado en Ingeniería Biomédica y responsable del desarrollo y realización del proyecto.

Tabla 2-1. Desglose del presupuesto relacionado con la mano de obra requerida para la realización del proyecto.

Perfil	**Número de horas**	**Coste Unitario (€)**	**Coste Total (€)**
Estudiante Ingeniería Biomédica	300	15,00	4500
Doctor e Investigador	35	25,00	875
Subtotal			5375

El coste total del personal es por lo tanto de **cinco mil trescientos setenta y cinco euros** (5375 €).

2.2 Coste de *Hardware*

En relación a los costes relacionados con el *hardware* se han calculado teniendo en cuenta los periodos de amortización de los equipos empleados. Como se puede ver en la Tabla 2-2, para la realización del trabajo se ha utilizado un ordenador portátil, un disco duro externo de 2 TB.

Tabla 2-2. Desglose del presupuesto relacionado con el hardware utilizado para la realización del proyecto.

Hardware	**Coste sin IVA (€)**	**Periodo de amortización (meses)**	**Periodo de uso (meses)**	**Coste (€)**
Ordenador portátil LG Gram	1264	60	10	210,70
Disco Duro Seagate 2 TB	63,2	96	10	6,60
Subtotal				217,30

El coste total del hardware es de **doscientos diez y siete con treinta euros** (217,30 €).

2.3 Coste de *Software*

Los costes derivados de la adquisición y utilización de los *software* empleados en este trabajo se pueden observar en la Tabla 2-3.

Tabla 2-3. Desglose del presupuesto relacionado con el software utilizado para la realización del proyecto.

Software	**Coste de la licencia sin IVA (€)**	**Periodo de amortización (meses)**	**Periodo de uso (meses)**	**Coste (€)**
R Studio®	0	De por vida	10	0
Microsoft Office 2019	54,50	12	10	45,42
PostgreSQL	0	De por vida	10	0
Subtotal				45,42

El coste total del *software* es de **cuarenta y cinco con cuarenta y dos euros** (45,42 €).

3. Presupuesto final

La obtención del presupuesto final pasa por la suma de todos los componentes mencionados anteriormente, obteniendo así el coste de ejecución material. Los gastos generales se calculan como un 13% del coste de ejecución material. Se debe a su vez calcular el beneficio industrial como el 6% del coste de ejecución material. Se obtiene así el presupuesto bruto, al que se le debe añadir el 21% de IVA. Todo ello se puede ver detallado en la Tabla 3-1.

Tabla 3-1. Desglose del presupuesto final del proyecto.

Descripción	Coste
Coste de personal	5.375 €
Coste de *Hardware*	217,30 €
Coste de *Software*	45,42 €
Presupuesto de ejecución material	**5637,72 €**
Gastos generales (13%)	732,90 €
Beneficio industrial (6%)	338,26 €
Presupuesto final bruto	**6708,88 €**
IVA (21%)	1408,86 €
Presupuesto final neto	**8117,74 €**

Por lo tanto, el presupuesto final de este trabajo final de grado es de **ocho mil ciento diez y siete con setenta y cuatro euros** (8117,74 €).