TESIS DOCTORAL

‘’Aplicación de Técnicas de Machine Learning en la Predicción de Hospitalizaciones y Reingresos de Pacientes con Esquizofrenia en Castilla y León”

AUTORA

Susel Góngora Alonso

DIRECTORES

Isabel de la Torre Díez

Manuel Franco Martín

Universidad de Valladolid

Escuela Técnica Superior de Ingenieros de Telecomunicación

Departamento de Teoría de la Señal y Comunicación e Ingeniería Telemática

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS DE LA INFORMACIÓN Y LAS TELECOMUNICACIONES

TESIS DOCTORAL

Aplicación de Técnicas de *Machine Learning* en la Predicción de Hospitalizaciones y Reingresos de Pacientes con Esquizofrenia en Castilla y León

Tesis presentada por:

Susel Góngora Alonso

para optar al grado de Doctora por la Universidad de Valladolid

Dirigida por:

Dr. Isabel de la Torre Díez y Dr. Manuel Franco Martín

2023

Valladolid, España

A mi familia y amigos
por su apoyo incondicional,
especialmente a Roberto y Pilar

Defensa

TÍTULO	Aplicación de Técnicas de Machine Learning en la Predicción de Hospitalizaciones y Reingresos de Pacientes con Esquizofrenia en Castilla y León
AUTOR	Susel Góngora Alonso
DIRECTORES	Dr. Isabel de la Torre Díez y Dr. Manuel Franco Martín
DEPARTAMENTO	Teoría de la Señal y Comunicaciones e Ingeniería Telemática

---

TRIBUNAL

PRESIDENTE	Dr. D.
VOCAL	Dr. D.
SECRETARIO	Dr. D.

acuerda otorgarle la calificación de

En Valladolid, a de del

Universidad de Valladolid

Escuela Técnica Superior de Ingenieros de Telecomunicación

Dpto. de Teoría de la Señal y Comunicaciones e Ingeniería Telemática

Research Stay for the International Mention

City:	Covilhã (Portugal)
Institution:	Instituto de Telecomunicações, Delegação da Covilhã
Research group:	Network Architectures and Protocols
Dates:	01/09/2021 - 30/11/2021
Duration:	91 Days (3 Months)
Tutor of the stay	Prof. Joel José Puga Coelho Rodrigues

Agradecimientos

En primer lugar, agradecer a mis directores de Tesis, la Dra. Isabel de la Torre Díez y el Dr. Manuel Franco Martín.

Han pasado 7 años ya desde que comencé este camino con la profesora Isabel de la Torre. En un inicio, con el Trabajo de Fin de Máster y posteriormente con la Tesis Doctoral. Han sido años de mucho trabajo, esfuerzo y dedicación, pero han valido la pena y me han llevado a superarme a mí misma en lo personal y lo profesional. Gracias por guiarme durante mi trayectoria investigadora.

Agradecer a todos los miembros del Grupo de Telemedicina y eSalud, por el apoyo prestado. A los profesores Miguel López Coronado y Beatriz Sainz de Abajo, por su colaboración, los consejos y el cariño brindado. A Gonçalo Marques y Deevyankar Agarwal por su ayuda incondicional en diferentes etapas de la Tesis.

A mis colegas del laboratorio Gema Castillo, Rosmeri Martínez y Maidelis Trujillo, que han sido parte de esta última etapa. Gracias por animarme, escucharme y ser incondicionales. Gracias por los cafés, las charlas, las risas y los buenos momentos.

Con un especial cariño a Isabel Herrera, por tanta entrega, confianza, generosidad y apoyo constante. Sabes que eres un pilar fundamental en mi vida, y espero poder acompañarte y retribuirte de la misma manera en el camino que estas transitando como investigadora.

Y, por último, mi más sincero agradecimiento a mi familia, mi madre y mi esposo Roberto que han sido un soporte durante todos estos años, brindándome todo el amor y la fuerza necesaria para llegar a la meta. De igual modo, a los amigos cercanos que siempre han estado presentes, brindándome su cariño y respaldo.

A todos, muchas gracias.

Resumen

La esquizofrenia es un trastorno mental grave que se caracteriza por síntomas como las alucinaciones, delirios, trastornos del pensamiento, los sentimientos y la conducta. El 1% de la población mundial padece este trastorno. Los síntomas tienden a iniciar durante la adolescencia y la edad adulta temprana, causando déficits cognitivos, deterioro social, ansiedad, depresión y alteraciones del comportamiento. Es un síndrome clínico heterogéneo, de ahí que, los síntomas se presenten de forma variable dependiendo de las características de la persona.

La esquizofrenia es un grave problema a nivel mundial, con altas tasas de discapacidad y una importante carga sanitaria, social, laboral y económica. En 2016, se estimó una prevalencia puntual global estandarizada por edad de 0.28%, un total de 21 millones de casos en todo el mundo; con un 70.8% de estos casos, en el grupo de edad de 25 a 54 años. Las personas con esquizofrenia se asocian con un mayor riesgo de abuso de sustancias, suicidio y mortalidad en comparación con la población general. Presentan tasas de hospitalización de un 20-40% en un año, lo que deriva en altos costes en el sistema sanitario y afecta la calidad de vida de los pacientes y los familiares.

En la actualidad, las altas tasas de reingreso hospitalario se han convertido en un problema para los servicios de salud mental, ya que se asocia directamente con la calidad de la atención al paciente. Factores de riesgo como la edad, el número de hospitalizaciones previas, comorbilidades físicas, abuso de sustancias, falta de apoyo familiar y social, e incumplimiento del tratamiento, han sido identificados en personas con esquizofrenia que reingresan.

Según los datos del Instituto Nacional de Estadística (INE) de España, en 2020 se registraron un total de 32 628 hospitalizaciones de pacientes con esquizofrenia. De estas hospitalizaciones, 1 480 corresponden a los hospitales

públicos de Castilla y León (CyL). La estancia hospitalaria representa el coste directo más relevante del trastorno de esquizofrenia, alcanzado en España, el 37.6% de los costes sanitarios totales. En consecuencia, esto motiva la necesidad de identificar el riesgo de reingreso de estos pacientes y los factores asociados al mismo, para tomar medidas preventivas y mitigar los altos costos asociados a estas hospitalizaciones.

Existen métodos computacionales que ayudan en el diagnóstico, tratamiento y la toma de decisiones de trastornos como la esquizofrenia. El uso de técnicas de *Machine Learning* (ML), permite analizar patrones de los datos mediante métodos estadísticos, y crear modelos que aprenden y generalizan el comportamiento de los datos. El desarrollo de modelos predictivos para evaluar el riesgo de reingreso en los centros hospitalarios ayuda a desarrollar medidas preventivas en el tratamiento de estos pacientes.

En CyL, reducir el número de hospitalizaciones y de reingresos es de suma importancia para los servicios de psiquiatría. Por tanto, en esta Tesis Doctoral se plantea la hipótesis que *la aplicación de algoritmos de ML ayuda a identificar los factores de riesgo de hospitalización y predecir el reingreso de pacientes con esquizofrenia.* En consecuencia, el objetivo principal de esta investigación es desarrollar y evaluar nuevos modelos predictivos utilizando algoritmos de ML, con el fin de ayudar en la predicción de hospitalizaciones y reingresos de pacientes con esquizofrenia en CyL.

Para alcanzar este objetivo, se utilizaron 11 126 registros administrativos que corresponden a 5 412 pacientes hospitalizados con esquizofrenia, de dos bases de datos diferentes. Estas bases de datos corresponden a los 11 hospitales públicos de CyL en dos períodos de tiempo diferentes 2005-2015 y 2015-2020. Los registros son datos globales, no están basados en la psicopatología clínica del paciente; incluyen información demográfica, características de episodios de hospitalización, diagnósticos y procedimientos referentes al paciente hospitalizado. Estos registros se analizaron automáticamente utilizando técnicas de clasificación de ML, y se crearon

modelos predictivos para predecir el riesgo de reingreso de estos pacientes en CyL. En este sentido, se propuso una metodología que consta de 4 fases. Una primera fase de preprocesamiento de los datos a través de un análisis exploratorio general. Posteriormente, se realizó una fase de selección de características donde se determinaron las variables predictivas de la investigación. En la tercera fase se aplicaron diferentes algoritmos de ML supervisado para detectar automáticamente el riesgo de reingreso de las personas con esquizofrenia. Los modelos han sido validados con el método de validación cruzada y se han utilizado las curvas de características operativas del receptor (ROC: *Receiver Operating Characteristics*) para la interpretación de los modelos creados. Por último, se ha desarrollado una aplicación web que permite trasladar la principal contribución de esta Tesis Doctoral a la práctica clínica.

Se obtuvo un alto rendimiento con el enfoque de ML propuesto. Se compararon los diferentes modelos creados a partir de sus métricas de rendimiento, y se obtuvo que el algoritmo *Random Forest* (RF) es el que mejor predice el riesgo de reingreso de los pacientes con esquizofrenia en CyL. Este modelo RF alcanzó una exactitud (Acc: accuracy) de un 81.7% y un área bajo la curva ROC (AUC) del 87.9%. Estos valores sugieren que el modelo tiene una capacidad de discriminación razonable para predecir el reingreso de estos pacientes. Variables como la edad, la duración de la estancia, diagnósticos con códigos V, de abuso de sustancias y trastornos mentales, se identificaron como las variables más predictivas del modelo. Estas variables indican los posibles factores de riesgo asociados al reingreso de estos pacientes con esquizofrenia. Por último, se desarrolló una aplicación web que tiene la capacidad de calcular el riesgo de reingreso de un paciente cuando le van a dar el alta hospitalaria.

Por tanto, los resultados obtenidos en esta Tesis Doctoral sugieren que algoritmos de ML como el RF, tienen la capacidad de aprender características complejas de los datos y predecir el riesgo de reingreso de pacientes hospitalizados

con esquizofrenia, en CyL. Se considera que los modelos desarrollados en esta investigación pueden ayudar a la toma de decisiones, mejorando la calidad de la atención al paciente y desarrollando tratamientos preventivos en función de reducir el número de hospitalizaciones. Además, la implementación de la aplicación web desarrollada en esta investigación, en los hospitales públicos de CyL, puede ser de gran utilidad al personal sanitario en función de reducir los altos costos asociados a estas hospitalizaciones.

Abstract

Schizophrenia is a severe mental disorder characterized by symptoms such as hallucinations, delusions, disturbances of thought, feelings and behavior. The 1% of world's population suffers from this disorder. Symptoms tend to initiate during adolescence and early adulthood, causing cognitive deficits, social impairment, anxiety, depression and behavioral disturbances. It is a heterogeneous clinical syndrome, hence, symptoms present variably depending on the characteristics of the person.

Schizophrenia is a serious problem worldwide, with high rates of disability and a significant health, social, occupational and economic burden. In 2016, a global age-standardized point prevalence of 0.28% was estimated, a total of 21 million cases worldwide; with 70.8% of these cases, in the 25-54 age group. People with schizophrenia are associated with an increased risk of substance abuse, suicide and mortality compared to the general population. They have hospitalization rates of 20-40% in a year, which results in high costs to the healthcare system and affects the life quality of patients and family members.

Currently, high rates of hospital readmission have become a problem for mental health services, as it is directly associated with the quality of patient care. Risk factors such as age, number of previous hospitalizations, physical comorbidities, substance abuse, lack of family and social support, and noncompliance with treatment have been identified in people with schizophrenia who are readmitted to hospital.

According to data from the National Institute of Statistics, Spain, a total of 32 628 hospitalizations of patients with schizophrenia were recorded in 2020. Of these hospitalizations, 1 480 correspond to public hospitals in Castilla y León (CyL). Hospital stay represents the most relevant direct cost of schizophrenia disorder,

reaching 37.6% of total health care costs in Spain. Consequently, this motivates the need to identify the readmission risk of these patients and the factors associated with it, in order to take preventive measures and mitigate the high costs associated with these hospitalizations.

There are computational methods that help in the diagnosis, treatment and decision making of disorders such as schizophrenia. The use of Machine Learning (ML) techniques makes it possible to analyze data patterns using statistical methods and create models that learn and generalize the behavior of the data. The development of predictive models to assess the readmission risk to hospitals helps to develop preventive measures in the treatment of these patients.

In the CyL, reducing the number of hospitalizations and readmissions is of great importance for psychiatric services. Therefore, in this Doctoral Thesis it is hypothesized that *the application of ML algorithms helps to identify risk factors for hospitalization and predict readmission of patients with schizophrenia*. Consequently, the main objective of this research is to develop and evaluate new predictive models using ML algorithms, in order to help in the prediction of hospitalizations and readmissions of patients with schizophrenia in CyL.

To achieve this objective, 11 126 administrative records corresponding to 5 412 hospitalized patients with schizophrenia from two different databases were used. These databases correspond to the 11 public hospitals in CyL in two different time periods 2005-2015 and 2015-2020. The records are global data, not based on the clinical psychopathology of the patient; they include demographic information, characteristics of hospitalization episodes, diagnoses and procedures concerning the hospitalized patient. These records were automatically analyzed using ML classification techniques and predictive models were created to predict the readmission risk of these patients in CyL. In this regard, a methodology consisting of 4 phases was proposed. A first phase of data preprocessing through a general exploratory analysis. Subsequently, a phase of characteristics selection was carried

out to determine the predictive variables of the research. In the third phase, different supervised ML algorithms were applied to automatically detect the readmission risk of people with schizophrenia. The models have been validated with the cross-validation method and receiver operating characteristics (ROC) curves have been used for the interpretation of the created models. Finally, a web application has been developed to transfer the main contribution of this Doctoral Thesis to clinical practice.

High performance was obtained with the proposed ML approach. The different models created from their performance metrics were compared, and the RF algorithm was found to best predictor the readmission risk of patients with schizophrenia in CyL. The RF model achieved an accuracy (Acc) of 81.7% and an area under the ROC curve (AUC) of 87.9%. These values suggest that the model has a reasonable discrimination capacity to predict the readmission of these patients. Variables such as age, length of stay, diagnoses with codes V, substance abuse and mental disorders were identified as the most predictive variables of the model. These variables indicate possible risk factors associated with the readmission of these patients with schizophrenia. Finally, a web application was developed that has the ability to calculate the readmission risk of a patient at the time of discharge from the hospital.

Therefore, the results obtained in this Doctoral Thesis suggest that ML algorithms such as RF have the ability to learn complex features from the data, and predict the readmission risk of hospitalized patients with schizophrenia in CyL. It is considered that the models developed in this research can help in decision making, improving the quality of patient care and developing preventive treatments in order to reduce the number of hospitalizations. In addition, the implementation of the web application developed in this research, in public hospitals of CyL, can be very useful to health personnel in order to reduce the high costs associated with these hospitalizations.

Acrónimos

AB	*Adaptative boosting*
Acc	*Accuracy*
AUC	*Area under ROC curve*
BD1	Base de datos de los 11 hospitales públicos de CyL de 2005-2015
BD2	Base de datos de los 11 hospitales públicos de CyL de 2015-2020
CEIm	Comité de Ética de la Investigación con Medicamentos área de salud de Valladolid
CIE-10	Clasificación internacional de enfermedades décima revisión
CIE-9	Clasificación internacional de enfermedades novena revisión
CMBD	Conjunto Mínimo Básico de Datos
CyL	Castilla y León
DT	*Decision tree*
FN	Falsos negativos
FP	Falsos positivos
GBM	*Gradient boosting machine*
IG	Índice de Gini
INE	Instituto Nacional de Estadística
kNN	*k-Nest Neighbor*
LR	*Logistic regression*
ML	*Machine Learning*
MLP	*Multi-layer perceptron*
NB	Naïve bayes
NN	*Neural network*s
OMS	Organización Mundial de la Salud
RF	*Random forest*

ROC	*Receiver operating characteristics*
SD	*Standard deviation*
SVM	*Support vector machine*
SVMradial	*Support Vector Machine with Radial Basis Function kernel*
VN	Verdaderos negativos
VP	Verdaderos positivos
XGboost	*Extreme gradient boosting*
ZBS	Zona básica de salud

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

1.1 Contexto y Motivación

La esquizofrenia es un trastorno mental grave que afecta alrededor del 0.75% de la población mundial, se caracteriza por alteraciones cognitivas, de percepción y conductuales. Este trastorno puede presentarse de forma heterogénea, lo que ha llevado a investigadores y clínicos a determinar que puede ser un síndrome clínico, compuesto por varios patrones de expresión de síntomas en lugar de una sola enfermedad. La esquizofrenia puede provocar síntomas como deterioro cognitivo social, déficits en la memoria de trabajo, disfunción neurocognitiva, aislamiento social, desmotivación, déficit en la función ejecutiva y la velocidad de procesamiento (Kirkpatrick et al., 2001; Rantala et al., 2022).

Este trastorno mental presenta un inicio de los síntomas en la adolescencia y la edad adulta temprana. La prevalencia está alrededor de los 40 años con un descenso en los grupos de mayor edad (Charlson et al., 2018). Los adultos con esquizofrenia tienen un riesgo notablemente mayor de muerte prematura. Aproximadamente el 60% de las muertes por este trastorno pueden atribuirse a enfermedades físicas prevenibles, del tipo cardiovascular, respiratoria, accidentes cerebrovasculares y complicaciones de diabetes (Harris et al., 2013; Lurie et al., 2021; Olfson et al., 2015). El suicidio es también un problema que afecta a esta población, es más frecuente entre las personas que padecen este trastorno, se estima que el 4.9% mueren por suicidio en comparación con el 0.013% de la población general (Huberts et al., 2022).

El impacto del trastorno de esquizofrenia influye a nivel personal, familiar y social, lo que supone una carga excesiva de cuidados para el paciente y un elevado coste médico (Chi et al., 2016). En una revisión sistemática de 2020 (Christensen et al., 2020), realizada para caracterizar el costo de los trastornos mentales, determinan que la esquizofrenia se asoció con el costo social medio más alto por paciente a nivel mundial (Kotzeva et al., 2023). La tasa de hospitalización de estas personas está entre el 20-40% por año, dependiendo de factores relacionados con el tratamiento, las patologías asociadas, la población y factores sociales (Ruetsch et al., 2018). El reingreso en los servicios de salud mental es una métrica fundamental de la calidad de la atención al paciente. Este indicador mide la capacidad de los sistemas de salud para apoyar al paciente y proporcionar una mejor atención hospitalaria (Wolff et al., 2019). Los reingresos en este servicio se consideran un mal resultado y tienen un impacto negativo en el bienestar del paciente, el estado emocional de las familias y los costes en el sistema de salud (Baeza et al., 2018). Esto motiva la necesidad de identificar el riesgo de reingreso de estos pacientes y los factores asociados al mismo, para tomar medidas preventivas y mitigar los altos costos asociados a estas hospitalizaciones.

En la actualidad han surgido nuevos métodos computacionales para apoyar el diagnóstico, la toma de decisiones y el tratamiento de trastornos de salud mental como la esquizofrenia. El uso de técnicas de ML en esta área se ha ido incrementando en las dos últimas décadas (Dipnall et al., 2016). A través de estas técnicas se pueden obtener correlaciones y patrones de grandes conjuntos de datos para crear nuevos conocimientos (Awad et al., 2017); (Tai et al., 2019).

La extracción de datos predictivos de personas con esquizofrenia permite construir modelos de ayuda a la decisión para determinar la gravedad de los síntomas comunes, identificar y medir la progresión del trastorno, así como la efectividad del tratamiento. Por tanto, esta Tesis Doctoral pretende ayudar a predecir el riesgo de reingreso de pacientes con esquizofrenia en CyL, utilizando algoritmos de ML.

1.2 Hipótesis

Las altas tasas de reingreso de los pacientes con esquizofrenia afectan la salud y la calidad de vida de estos pacientes. Son indicadores negativos de la calidad de la atención médica de los servicios psiquiátricos con respecto a los pacientes hospitalizados (Baeza et al., 2018). La estancia hospitalaria representa el coste directo más relevante del trastorno de esquizofrenia. La proporción del coste de hospitalización entre los costes médicos directos varía del 27-92% en la Unión Europea (Kovács et al., 2018). En España, el coste estimado de trastornos de esquizofrenia alcanza un 37.6% de los costes sanitarios totales (Jin & Mosweu, 2017; *Situación de La Salud Mental En España. Psiquiatría Clínica*, 2017).

El abuso de sustancias se ha identificado como uno de los factores influyentes en las hospitalizaciones de pacientes con esquizofrenia (Rozin et al., 2019). Factores de riesgo demográficos como el estado civil, la edad, el género, las hospitalizaciones previas, la falta de apoyo familiar, apoyo social y el incumplimiento de la medicación, se han determinado como causas de reingreso de los pacientes con trastorno de esquizofrenia (Grudnikoff et al., 2019; Lorine et al., 2015; Sugisawa et al., 2022). En consecuencia, reducir el número de hospitalizaciones y de reingresos es de suma importancia para los servicios de psiquiatría en CyL.

El uso de algoritmos de ML permite extraer patrones de comportamiento, desarrollar modelos predictivos, reducir riesgos e identificar información útil de grandes bases de datos (Alonso et al., 2017; Pirooznia et al., 2012). El desarrollo de herramientas predictivas para evaluar los factores de riesgo asociados con el reingreso ofrece oportunidades para la selección de tratamientos e implementar medidas preventivas asociadas con este trastorno (Holderness et al., 2019).

Sobre la base de esta problemática, en esta Tesis Doctoral se plantea la siguiente hipótesis: *"La aplicación de algoritmos de ML ayuda a identificar los factores de riesgo de hospitalización y predecir el reingreso de pacientes con esquizofrenia en CyL".*

1.3 Objetivos

El objetivo principal de esta Tesis Doctoral es desarrollar y evaluar nuevos modelos predictivos utilizando algoritmos de ML, con el fin de ayudar en la predicción de hospitalizaciones y reingresos de pacientes con esquizofrenia, en CyL. Para alcanzar el objetivo principal se plantean los siguientes objetivos específicos:

- Comparar y seleccionar algoritmos de ML que mejor se ajusten a los datos de hospitalización de pacientes con esquizofrenia en CyL, utilizando modelos predictivos.
- Identificar los factores de riesgo asociados a los pacientes hospitalizados con trastornos de esquizofrenia, en CyL.
- Desarrollar y evaluar un modelo predictivo de aplicación clínica, para ayudar a predecir el reingreso de los pacientes con esquizofrenia en esta población.
- Desarrollar e implementar un prototipo de aplicación web para visualizar los principales resultados obtenidos en la Tesis Doctoral.

1.4 Metodología

A lo largo de esta Tesis Doctoral se utilizaron un total de 11 126 registros administrativos que corresponden a 5 412 pacientes con trastornos de esquizofrenia, de dos bases de datos diferentes. Estas bases de datos contienen los registros de hospitalizaciones de los 11 hospitales públicos de CyL, en el período de 2005-2015 y de 2015-2020. Con el análisis preliminar a la base de datos en el período de 2005-2015, se identificó el trastorno de esquizofrenia, como el diagnóstico principal de salud mental prevalente en los pacientes hospitalizados en esta región.

Las principales técnicas de ML aplicadas a salud mental, específicamente al trastorno de esquizofrenia se identificaron con una revisión del estado del arte. Para ello se definieron las bases de datos científicas y los criterios de inclusión y exclusión de los estudios a revisar. La revisión se ha centrado en los estudios

relacionados con la predicción de hospitalizaciones y reingresos de personas con esquizofrenia y trastornos mentales en general.

En este sentido, para alcanzar los objetivos propuestos en esta investigación se estableció la siguiente metodología:

1. Preprocesamiento de las bases de datos utilizadas en la investigación. Se realizó un análisis exploratorio general y se transformaron las variables del conjunto de datos para aplicar diferentes algoritmos de ML.

2. Selección de características del conjunto de datos. Para la selección de características se han empleado dos criterios. Un primer criterio se basa en la experiencia clínica del psiquiatra experto, seleccionando las variables que están relacionadas con los factores de riesgo de este tipo de pacientes, mientras que el segundo criterio se basa en determinar la importancia de las variables aplicando el algoritmo RF. Con la selección de características se identifican las variables predictivas del conjunto de datos, y se asocian estas variables a los factores de riesgo o patrones comunes en pacientes hospitalizados con trastorno de esquizofrenia, en CyL.

3. Desarrollo de modelos predictivos con pacientes hospitalizados con esquizofrenia utilizando algoritmos de ML. En la primera parte de esta Tesis Doctoral, se ha comparado el rendimiento de varios algoritmos de ML enfocados en la predicción de pacientes hospitalizados con esquizofrenia. Este primer estudio (Góngora Alonso et al., 2022), tiene como objetivo identificar el algoritmo que mejor se ajusta a los datos existentes. Posteriormente, teniendo en cuenta los resultados ya obtenidos y la revisión del estado del arte, se desarrollan varios modelos que predicen el riesgo de reingreso de estos pacientes. Los

modelos desarrollados en esta investigación se han validado aplicando validación cruzada k=10.

4. Desarrollo de una aplicación web que muestra los principales resultados de esta Tesis Doctoral. En la fase final de esta investigación se desarrolla y evalúa un modelo predictivo de aplicación clínica que determina el reingreso de pacientes con esquizofrenia en los diferentes complejos asistenciales públicos de CyL. Este modelo se desarrolla con el objetivo de trasladar nuestra contribución a la práctica clínica a través de la aplicación web. Para construir el modelo final se usan las dos bases de datos y el clasificador RF, que muestra los valores más altos de Acc y AUC en comparación con el resto de los algoritmos. La aplicación usa el modelo final entrenado y calcula el riesgo de reingreso de los pacientes hospitalizados con esquizofrenia en CyL.

1.5 Estructura de Tesis

Esta Tesis Doctoral se estructura en 8 capítulos. En la presente sección se ha descrito la motivación de la investigación, la hipótesis, objetivos, metodología y la estructura del documento. A continuación, se resume el contenido de cada capítulo.

- En el Capítulo 2 se describe el trastorno de esquizofrenia en cuanto a prevalencia, factores de riesgo, hospitalizaciones y epidemiología. A continuación, se presenta una descripción detallada de los algoritmos de ML más utilizados en la predicción de hospitalizaciones, y, en consecuencia, del reingreso de personas con esquizofrenia.

- En el Capítulo 3 se describe la población objeto de estudio y las bases de datos utilizadas para el desarrollo de la investigación.

- En el Capítulo 4 se plantea el proceso metodológico de esta investigación. En la fase inicial se describe el preprocesamiento de los datos. Posteriormente, se explica el proceso de selección de características del conjunto de datos. A continuación, se describen los algoritmos de ML y las pruebas estadísticas utilizadas en esta Tesis Doctoral. Por último, se describe el diseño de la aplicación web, que permite trasladar la principal contribución de esta Tesis Doctoral a la práctica clínica.

- En el Capítulo 5 se plantea un análisis de las características de la población de ambas bases de datos. Se muestran los resultados obtenidos a partir de las métricas de rendimiento de cada uno de los modelos. Se identifica el algoritmo que mejor se ajusta a los datos de hospitalización de pacientes con esquizofrenia en CyL, y a partir de este algoritmo, se crea un modelo final que predice el riesgo de reingreso de estos pacientes. Por último, se muestran los resultados de la aplicación web desarrollada sobre la base del modelo final construido.

- En el Capítulo 6 se analiza cada uno de los resultados presentados en el Capítulo 5. Se realiza una comparación con estudios similares y se plantean las limitaciones de la investigación.

- En el Capítulo 7 se plantean las conclusiones obtenidas de la presente Tesis Doctoral, las principales contribuciones y las líneas futuras derivadas de esta investigación. En el Capítulo 8 se presenta la traducción en inglés del Capítulo 7.

Capítulo 2

Estado del Arte

2.1 Esquizofrenia

2.1.1 Definición y prevalencia

La esquizofrenia es un trastorno mental grave, que se caracteriza por diferentes síntomas como delirios, trastornos del pensamiento, alucinaciones y déficits cognitivos (El-Missiry et al., 2011; Kendler, 2016). Sin embargo, la esquizofrenia es un síndrome clínico heterogéneo, la principal característica fenomenológica es la variedad de la sintomatología y la ausencia de un síntoma o signo patognomónico (American Psychiatric Association, 2013). A nivel neuropatológico, la esquizofrenia no tiene una única característica diagnóstica. En general, parece estar caracterizado por déficits sinápticos, alteraciones en la neurotransmisión de glutamato y dopamina e hipofrontalidad (Harris et al., 2013).

Según la Organización Mundial de la Salud (OMS), la esquizofrenia afecta a una de cada 300 personas, alrededor de 24 millones a nivel mundial (*World Health Organization-Schizophrenia*, 2022). El inicio de los síntomas y el diagnóstico suele ser durante la segunda y tercera década de la vida. Las personas con este trastorno pueden mostrar déficits cognitivos y de cognición social, variación del patrón del sueño, fobia, ansiedad, alteración del comportamiento y depresión. Los déficits cognitivos están asociados al deterioro del lenguaje, la memoria de trabajo, la memoria declarativa, la velocidad de procesamiento y la función ejecutiva; mientras que los déficits de cognición social están asociados a la capacidad de atención de la

persona, para determinar la importancia de sucesos irrelevantes (American Psychiatric Association, 2013).

La prevalencia global estandarizada por edad de la esquizofrenia se estimó en un 0.28% en 2016, un total de 21 millones de casos en todo el mundo; con un 70.8% de estos casos en el grupo de edad de 25 a 54 años (Charlson et al., 2018). La prevalencia en cuanto a la proporción de sexos es controvertida, es mayor en hombres que en mujeres en países desarrollados, pero no en países en desarrollo (Rantala et al., 2022). Esta proporción está equilibrada cuando se incluyen más síntomas del estado de ánimo y cuadros breves, pero en los grupos clínicos tiene mayor prevalencia en los hombres, como se encontró en el actual estudio español (Orrico-Sánchez et al., 2020).

Las personas con esquizofrenia presentan tasas de morbilidad y mortalidad más elevadas que la población general. Las tasas de comorbilidad están asociadas con trastornos como el abuso de sustancias, trastornos del pánico, de ansiedad y obsesivos-compulsivos (American Psychiatric Association, 2013; Lurie et al., 2021). Las personas con esquizofrenia por lo general tienen una esperanza de vida de 15 años menos que la población general, una tasa reproductiva más baja, tienden a ingresar con frecuencia en los centros hospitalarios y presentan tasas elevadas de discapacidad y mortalidad (Rantala et al., 2022). Presentan una probabilidad de muerte prematura entre el 40-60% mayor que la población en general (Jin & Mosweu, 2017; *World Health Organization-Schizophrenia*, 2022). Una gran proporción de las muertes prematuras en personas con este trastorno se debe especialmente a comorbilidades médicas crónicas, como las enfermedades cardiovasculares, respiratorias, infecciosas, accidente cerebrovascular y complicaciones por diabetes (de Pedro Cuesta et al., 2016; Jørgensen et al., 2017; McGrath et al., 2008). Sin embargo, el suicidio es también una causa relevante de muerte en estas personas, alcanzando el 5% de ellos. De hecho, se considera un evento subestimado, ya que alrededor del 25-50% de pacientes con este trastorno

intentan suicidarse a lo largo de su vida (Hor & Taylor, 2010). Según el estudio (Berardelli et al., 2021), los autores indican que los comportamientos suicidas en personas con esquizofrenia pueden estar asociado con el inicio temprano del trastorno, y determinan que los intentos de suicido están fuertemente relacionados con una mayor frecuencia de hospitalización.

2.1.2 Factores de riesgo

Los factores de riesgo asociados a personas con diagnóstico de esquizofrenia se determinan como factores biológicos, psicológicos y sociales. Partiendo de que las causas de la esquizofrenia siguen siendo desconocidas, la predisposición genética y el abuso de sustancias tienen un impacto significativo en este trastorno, específicamente en personas con vulnerabilidad biológica para desarrollarlo (Tsoi et al., 2008). La Tabla 2.1 muestra una clasificación de los factores relacionados con el trastorno de esquizofrenia.

Tabla 2.1 Factores relacionados con el trastorno de esquizofrenia. Fuente: History, aetiology, and symptomatology of schizophrenia. Psychiatry, 2008 (Tsoi et al., 2008).

	Biológicos	**Psicológicos**	**Sociales**
Factores de predisposición	Antecedentes familiares de esquizofrenia Complicaciones obstétricas y perinatales	Deterioro cognitivo Personalidad esquizotípica	Nacimiento en entornos urbanos Emigración
Factores desencadenantes	Abuso de sustancias	Alta emoción expresada	Estrés
Factores perpetuantes	Incumplimiento con la medicación	Alteración de la percepción	Falta de vivienda Desempleo

El riesgo de una persona de desarrollar esquizofrenia es de aproximadamente el 1% en la población general. Los factores genéticos y fisiológicos contribuyen a

determinar el riesgo de padecer la enfermedad (American Psychiatric Association, 2013). Se ha demostrado que los parientes en primer grado de las personas con esquizofrenia tienen un riesgo de 2 a 9% mayor de desarrollar este trastorno en comparación con los parientes de personas sanas (Tsoi et al., 2008). Si bien los factores genéticos son claramente importantes en la etiología de la esquizofrenia, en la literatura relacionan determinados factores de riesgo ambientales para la enfermedad. Estos factores ambientales incluyen complicaciones obstétricas, infecciones prenatales, postnatales y otros factores que pueden actuar durante este período crucial de desarrollo del cerebro (Khan et al., 2022).

El abuso de sustancias, en particular el cannabis, es considerado también como un factor de riesgo, especialmente en personas con alta vulnerabilidad biológica para desarrollar esquizofrenia (Tsoi et al., 2008). Según el estudio (Arranz et al., 2018), las sustancias más consumidas por los pacientes con este trastorno son: el cannabis (12-42%), la cocaína (15-50%), el alcohol (20-60%) y la nicotina (80-90%). El abuso de estas sustancias en personas con este trastorno provoca una mayor tendencia al comportamiento agresivo e impulsivo, y un mayor riesgo de suicidio, derivando en recaídas y reingresos hospitalarios (Teixeira et al., 2022).

Los factores psicológicos están asociados al deterioro cognitivo de la persona y trastornos de la personalidad. El deterioro cognitivo en pacientes con esquizofrenia muestra una asociación consistente con los índices de capacidad funcional cotidiana. En los estudios previos, el hallazgo más consistente ha sido, un deterioro generalizado en las medidas neuropsicológicas que persisten en cada estado clínico y a lo largo de la vida de estos pacientes (Schaefer et al., 2013). Los factores de personalidad en niños que desarrollan esquizofrenia se traducen en problemas de desarrollo, comportamiento social, habilidades motoras y rendimiento académico, mientras que en los adultos tienden a tener rasgos de personalidad esquizoide y esquizotípico (Tsoi et al., 2008).

El estrés psicosocial y los síntomas depresivos constituyen un factor importante en el curso clínico de la esquizofrenia, pueden desencadenar episodios de psicosis o agravar los síntomas psicóticos (Mizrahi, 2016; Rantala et al., 2022). Existen evidencias de que los pacientes con este trastorno a menudo viven una vida social estresante, con un apoyo social limitado y un entorno familiar crítico. La falta de control es uno de los principales factores para provocar una respuesta de estrés fisiológico (Lange et al., 2017).

Los factores sociales también influyen en el desarrollo de este trastorno. Factores como la migración, la crianza en zonas urbanas, el bajo nivel socioeconómico, el aislamiento social, la falta de vivienda y el desempleo, están asociados con diversos trastornos mentales, específicamente con el riesgo de psicosis y esquizofrenia (Akdeniz et al., 2014). Estos factores por sí solos no aportan evidencia de riesgo de padecer esquizofrenia, la propensión a este trastorno viene derivada del conjunto de factores descritos en esta sección.

2.1.3 Hospitalizaciones

El reingreso hospitalario ha sido uno de los problemas más importantes en el ámbito de la salud mental en los últimos años (Cronin et al., 2019; J. Edgcomb et al., 2019; Innes et al., 2015; Wolff et al., 2019). Es un indicador de calidad de la atención hospitalaria, y refleja la capacidad de los sistemas de salud mental, para proporcionar una atención coordinada a este tipo de pacientes (Baeza et al., 2018; Castillo-Sánchez et al., 2022; Neto et al., 2021). Una alta tasa de reingresos se asocia a resultados negativos, afecta el bienestar del paciente, desencadena un deterioro emocional de las familias y un aumento de los costes sanitarios (Rumshisky et al., 2016; Shadmi et al., 2018; P. Zhao & Yoo, 2021).

A pesar de la aplicación de diversos tratamientos personalizados, se estima que la tasa de recaída entre las personas con esquizofrenia se sitúa entre el 50-92% (Adebiyi et al., 2018). Esto implica una alta morbilidad y una elevada tasa de

reingreso. En España, la tendencia de los reingresos psiquiátricos es del 10%, con 1.6 hospitalizaciones por cada 1 000 habitantes (*National Health System Annual Report 2020-2021*, 2022). Como consecuencia, estas tasas tienen un alto coste para el sistema sanitario y los servicios comunitarios.

La mayoría de los pacientes psiquiátricos hospitalizados pueden ser dados de alta sin necesidad de un seguimiento exhaustivo. Sin embargo, los pacientes con una enfermedad mental como la esquizofrenia necesitan cuidados posteriores a largo plazo (Higgins et al., 2018). Los largos períodos de hospitalización se asocian con un mayor riesgo de reingreso en estos pacientes. En un estudio de 2017 (Hung et al., 2017) los autores obtienen una tasa de reingreso de pacientes con esquizofrenia del 15.2% dentro de los 3 meses posteriores al alta, y el 33.3% dentro del año. Las mejores prácticas recomiendan hospitalizaciones breves y seguimientos posteriores al alta para mejorar la integración social y la recuperación (Fleury et al., 2019). La atención psiquiátrica sigue siendo necesaria para un pequeño subgrupo de pacientes que no pueden ser tratados de forma segura o eficaz en casa (Steeg et al., 2018).

Aunque es difícil prever las causas de reingreso de estos pacientes, debido a que existen muchas razones por las cuales una persona puede ser hospitalizada, se han encontrado varios estudios que han identificado factores de riesgo asociados al reingreso de pacientes con esquizofrenia (Artetxe et al., 2018; Morel et al., 2020). El abuso de sustancias es uno de los factores más influyentes en estos pacientes, se estima que un tercio de ellos consumen psicofármacos como el cannabis (Rozin et al., 2019). El riesgo de consumir alcohol, tabaco y drogas es dos veces más común entre las personas con trastornos esquizofrénicos que en la población general (Thomsen et al., 2018). Se han identificado otros factores de riesgo relacionados con las hospitalizaciones previas (Grudnikoff et al., 2019), la raza, el estado civil, la edad, el género (Sugisawa et al., 2022), determinantes sociales como la falta de apoyo familiar, de relaciones sociales (Lorine et al., 2015), el acceso a la atención ambulatoria y el incumplimiento de la medicación (Portela et al., 2022).

2.1.4 Epidemiología

Según la OMS (*World Health Organization-Schizophrenia*, 2022), la tasa de esquizofrenia en adultos a escala mundial es de 1 por cada 222 personas, que representa un 0.45%. Las personas con este trastorno presentan una probabilidad de muerte prematura de 2-3 veces mayor que la población general. En España, la prevalencia registrada de trastornos mentales según el (*National Health System Annual Report 2020-2021*, 2022), es de 286.7 casos por cada 1 000 habitantes. Respecto al trastorno de esquizofrenia, en la Figura 2.1 se muestran los casos registrados desde 2011 hasta 2019. En 2018 se registraron alrededor de 183 735 casos, representando la cifra más alta de estos años.

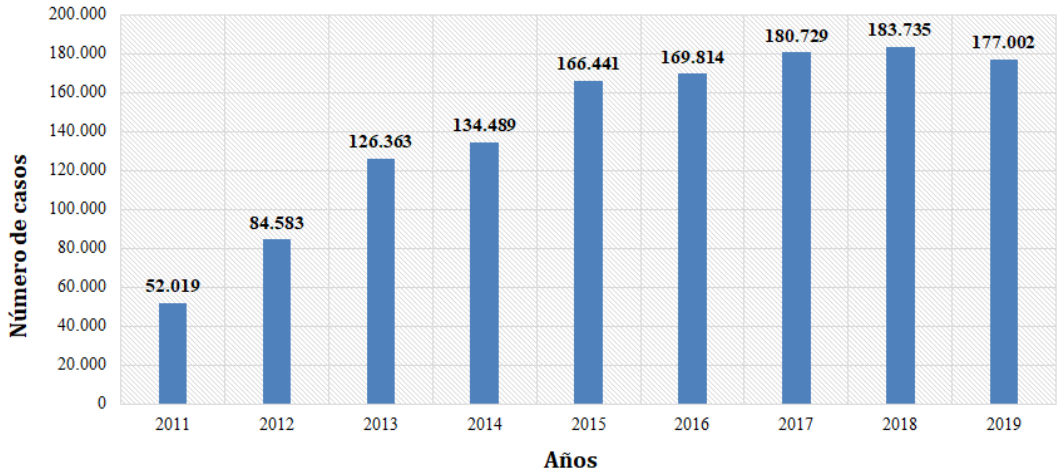

Figura 2.1 Número de casos de esquizofrenia registrados en España desde 2011 hasta 2019. Fuente: Ministerio de Sanidad, Servicios Sociales e Igualdad.

Según los últimos datos del INE (*National Institute of Statistics*, 2020), que corresponden al 2020, se registraron un total de 32 628 altas hospitalarias, (19 809 hombres y 12 819 mujeres) que representan un 60.71% y un 39.29% respectivamente. Las hospitalizaciones por grupos de edad son mayores en el intervalo de 25-64 años, siendo mayor en el intervalo de 35-44 años en los hombres y de 45-54 años en las mujeres (Ver Figura 2.2). El comportamiento del género en este

grupo de intervalos es más prevalente en los hombres de 15-74 años, mientras que a partir de los 75 años las mujeres tienden a hospitalizar más, aunque el número de hospitalizaciones sigue siendo menor que el registrado en el intervalo de 25-64 años.

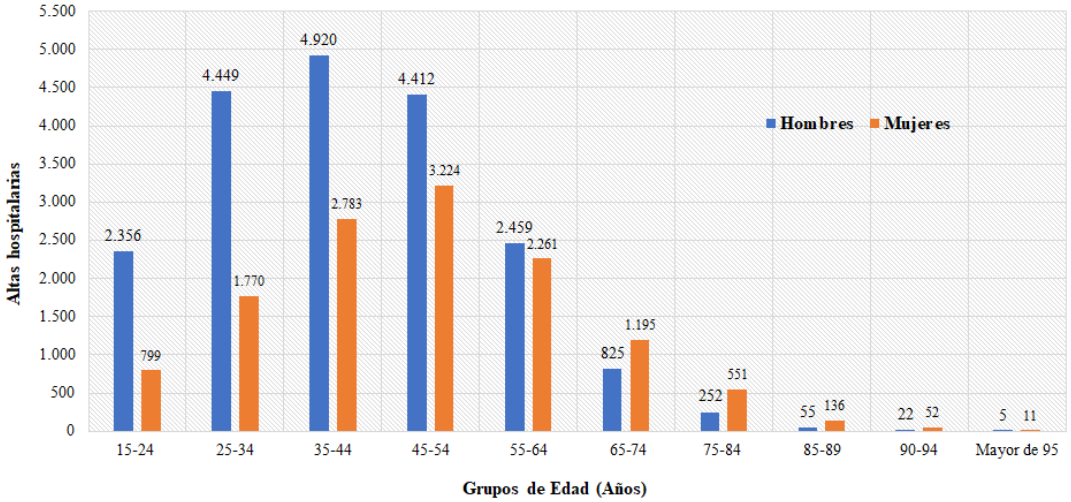

Figura 2.2 Altas hospitalarias de pacientes con esquizofrenia según el género y el grupo de edad, registradas en España en 2020. Fuente: INE.

En la comunidad de CyL, la tasa de morbilidad hospitalaria por 100 000 habitantes, de pacientes con trastornos de esquizofrenia, fue de 62 en el 2020. En la Figura 2.3 se muestra esa tasa por cada una de las provincias de CyL. Estos datos corresponden sólo a los 11 hospitales públicos que tiene la comunidad. Las mayores tasas se registran en la provincia de Soria y Palencia (93 y 90 respectivamente).

La prevalencia registrada de altas hospitalarias en CyL es de 1 480, siendo más elevada en los hombres (873) que en mujeres (607), lo que representa una diferencia del 17.98%. En la Figura 2.4 se muestra el comportamiento del número de altas hospitalarias por género y provincia. Los valores más altos de hospitalización se registraron en Valladolid (278) y León (282), aunque es importante destacar que en estas provincias se registran datos de dos hospitales públicos, mientras que el resto solo incluyen un hospital. La prevalencia de hospitalizaciones sigue siendo mayor en

hombres, excepto en Segovia que registró en ese año una diferencia de 9.68% más de mujeres que hombres.

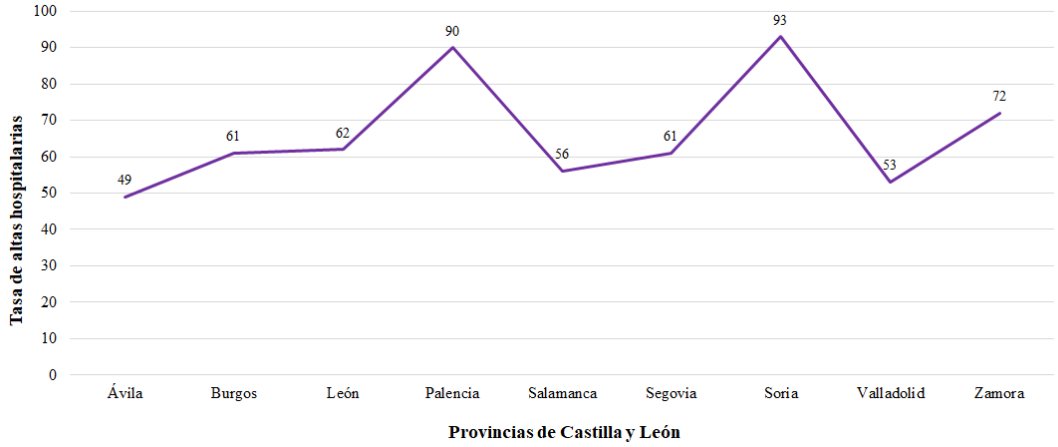

Figura 2.3 Tasas de morbilidad hospitalaria por 100 000 habitantes (pacientes con esquizofrenia), registradas en las provincias de CyL en 2020. Fuente: INE.

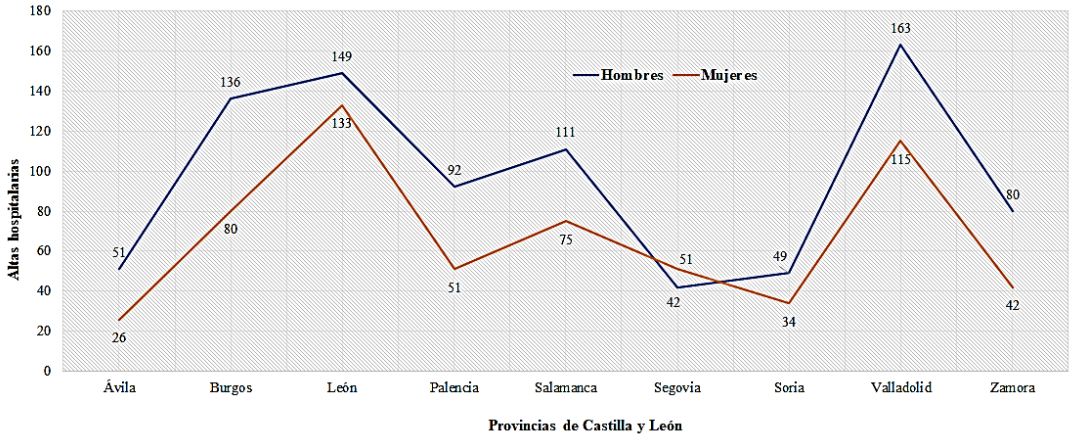

Figura 2.4 Altas hospitalarias de pacientes con esquizofrenia según el género, registradas en las provincias de CyL en 2020. Fuente: INE.

En la Figura 2.5 se muestra el porcentaje de reingresos urgentes en CyL en el período de 2015-2020. La variación de reingresos en mujeres fue de 11.19-17.15% mientras que en los hombres fue del 10.3-11.7%. Tal como muestra la Figura 2.5,

cuando se analizan los casos de hospitalizaciones psiquiátricas en general las mujeres tienden a tener un mayor número de reingresos, mientras que en personas con esquizofrenia ocurre el efecto contrario (Ver Figura 2.2)

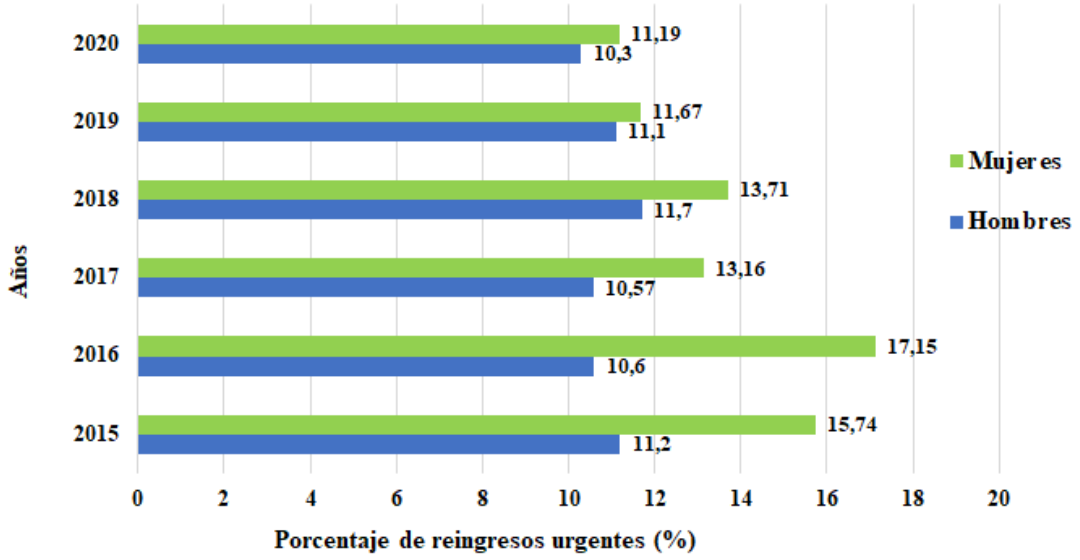

Figura 2.5 Porcentaje de reingresos urgentes psiquiátricos según el género, registrados en CyL de 2015 a 2020. Fuente: INE.

En las hospitalizaciones de este tipo de pacientes el intervalo de estancia más prevalente es de 1-30 días. En el 2020 en España, se registraron un total de 25 317 altas hospitalarias en ese intervalo de estancia, de los cuales CyL registró 1 119 (Ver Figura 2.6). Para esta región esos datos nos dan una estancia media clasificada para el intervalo de 1-30 días, de 12.25 días (*National Institute of Statistics*, 2020).

2.2. Técnicas de Machine Learning

ML generalmente se conoce como un área de la inteligencia artificial que se encarga del reconocimiento de patrones a partir de un conjunto de datos (Veronese et al., 2013; Vieira et al., 2020). Según la naturaleza del etiquetado de datos, se puede dividir en supervisado, no supervisado y de refuerzo. En el desarrollo de esta Tesis

Doctoral utilizamos algoritmos de clasificación de aprendizaje supervisado. Este tipo de aprendizaje tiene como objetivo predecir o clasificar un resultado específico de interés. Dependiendo de si la variable objetivo es una variable categórica o continua, es un problema de clasificación o de regresión, respectivamente (El Naqa et al., 2015).

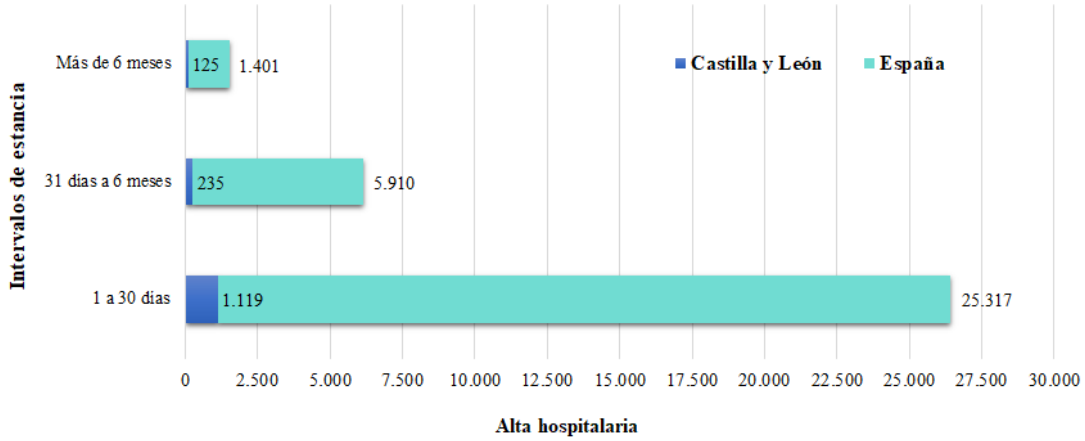

Figura 2.6 Altas hospitalarias por intervalos de estancia de pacientes con esquizofrenia, registradas en España y CyL en 2020. Fuente: INE.

En la actualidad, las técnicas de ML se usan para explorar y analizar patrones de datos mediante métodos estadísticos y de inteligencia artificial (Hou et al., 2020; Pachange et al., 2015; Z. Zhao et al., 2019). Permiten crear modelos predictivos precisos para aprender, analizar y generalizar el comportamiento de los datos (Rojas et al., 2018; van Mens et al., 2020; Xue et al., 2018). El desarrollo de estos modelos para evaluar el riesgo de reingreso en los centros hospitalarios ayuda a desarrollar medidas preventivas en el tratamiento de los pacientes (J. B. Edgcomb et al., 2021; Holderness et al., 2019).

Para la construcción de un modelo predictivo el desarrollo consta de diferentes fases. Una de las fases más importantes de este proceso es la selección de características, debido a que permite identificar las variables más significativas del conjunto de datos. En el área de salud, permite la identificación de los factores claves

asociados a una enfermedad o a una condición de riesgo específica. Tiene la capacidad de reducir el sobreajuste y la complejidad de los modelos.

Los métodos de selección de características se clasifican en tres tipos, métodos de filtrado, de envoltura y métodos integrados que combinan los dos métodos anteriores. Los estudios relacionados con el reingreso hospitalario que se han encontrado en la literatura utilizan principalmente los métodos de envoltura. En el estudio (Morel et al., 2020), los autores emplean el algoritmo Boruta de RF para obtener la importancia de las características, mientras que en el estudio (Ying et al., 2021) utilizan el algoritmo *extreme gradient boosting* (XGboost). El algoritmo CART ha sido utilizado en los estudios (J. Edgcomb et al., 2019; Fond et al., 2019) para identificar las variables predictoras del reingreso en pacientes con trastornos mentales y la recaída psicótica en personas con esquizofrenia respectivamente. En estudios como (Rojas et al., 2018; Y. Zhao et al., 2020), utilizan el índice de Gini (IG) para la selección de características. En (Rojas et al., 2018) tienen en cuenta la opinión de los expertos clínicos para la selección, y comprueban la validez clínica del modelo desarrollado a través del IG. Esta medida de importancia de las variables permite visualizar la contribución de las variables predictoras en el modelo. En (Y. Zhao et al., 2020), los autores identifican los factores de riesgo para el reingreso de pacientes con trastornos mentales a través del algoritmo RF. Aplican el IG para obtener la importancia de las variables, y el algoritmo *logistic regression* (LR) para evaluar la asociación entre los factores de riesgo identificados y el reingreso. Otros estudios como (Loreto et al., 2020) utilizan la ganancia de información y el método de envoltura con *naïve bayes* (NB) para la selección de características, mientras que en (Baeza et al., 2018) utilizan una LR multivariable para determinar los predictores del reingreso de pacientes hospitalizados con trastornos mentales.

En la revisión sistemática (Huang et al., 2021), sobre el uso de algoritmos de ML en la predicción de reingresos hospitalarios, los autores encontraron 43 estudios relacionados con el tema. Los algoritmos más comunes encontrados en el estudio son

los basados en árboles (23 estudios), estos algoritmos son: *decision tree* (DT), RF, XGboost y *adaptative boosting* denominado *adaboost* (AB). En esta revisión también plantean los estudios existentes que utilizan algoritmos de *neural network*s (NN) (14 estudios), LR (12 estudios) y *support vector machine* (SVM) (10 estudios). La mayoría de los estudios de esta revisión usan la métrica de rendimiento *area under ROC curve* (AUC).

En la clasificación de los pacientes con trastornos de esquizofrenia se utilizan diferentes algoritmos de ML, según indican los estudios encontrados (Jahmunah et al., 2019; Johannesen et al., 2016; Santos-Mayo et al., 2017; Shim et al., 2016). Algoritmos como el SVM, RF, *k-Nest Neighbor* (kNN) y *multi-layer perceptron* (MLP), se utilizan para clasificar a estos pacientes utilizando características combinadas de electroencefalograma.

En el estudio (Jo et al., 2020), los autores utilizaron el SVM, NB, RF, y *gradient boosting machine* (GBM) para predecir pacientes con esquizofrenia y controles sanos, obteniendo como resultado el mejor valor de *Acc* con RF. Algoritmos como RF y SVM se han aplicado en los estudios (Deng et al., 2019; Lee et al., 2018) para discriminar entre pacientes con esquizofrenia y pacientes sanos, utilizando imágenes de resonancia magnética. En (Lee et al., 2018), mostraron una alta tasa de rendimiento utilizando 504 características. RF tuvo una sensibilidad = 0.876 y una especificidad = 0.959, y SVM tuvo una sensibilidad = 0.895% y una especificidad = 0.945%. En (Zhu et al., 2021), los autores clasificaron a los pacientes con esquizofrenia utilizando el nivel de expresión del ARN mensajero en sangre periférica. Compararon algoritmos como NN, XGBoost, SVM, DT y RF. Los resultados mostraron que SVM era el mejor modelo con un AUC de 0.993, sensibilidad = 1.000 y especificidad = 0.895.

Otros estudios utilizaron un enfoque de ML para predecir la función cognitiva en la esquizofrenia (Lin et al., 2022), identificar a estas personas a través de redes sociales (Bae et al., 2021; Birnbaum et al., 2017) e identificar la violencia en

pacientes con este trastorno (Wang et al., 2020). En estudios como (Berardelli et al., 2021), los autores trataron de encontrar nuevos recursos clínicos disponibles que identificaran con exactitud el riesgo de suicidio en la esquizofrenia. En (Hettige et al., 2017), los autores desarrollaron un modelo predictivo clínicamente útil para identificar a los pacientes con esquizofrenia que intentan suicidarse, en una muestra de 345 participantes. Los resultados mostraron las mejores métricas en el modelo SVM y LR regularizada, con una *Acc* del 0.670 y un AUC de 0.700 y 0.710, respectivamente.

Existen estudios recientes de predicción del reingreso donde aplican algoritmos de ML como XGBoost (Morel et al., 2020), NB (Artetxe et al., 2018), RF (Deschepper et al., 2019), LR (Y. Zhao et al., 2020), DT, NN (Huang et al., 2021), AB (Tong et al., 2016) y SVM (Salem et al., 2019). Además, identifican factores de riesgo asociados al reingreso (Morel et al., 2020). Estos estudios se centran en trastornos de salud mental, en general. Por tanto, tal y como hemos descrito en esta sección, aunque existe un número considerable de publicaciones relacionadas con la predicción de reingresos en centros hospitalarios, se ha encontrado un número reducido de estudios enfocados específicamente en pacientes con esquizofrenia (Fond et al., 2019; Huberts et al., 2022; Thongkam & Sukmak, 2014; Ying et al., 2021). Esta investigación pretende mejorar los valores de rendimiento de las métricas teniendo en cuenta algoritmos iguales y diferentes a los que se utilizan en estos estudios similares. Además, pretende trasladar los resultados de esta investigación a la práctica clínica a través de una aplicación web, para ayudar al manejo hospitalario, respaldar la toma de decisiones clínicas y reducir el número de hospitalizaciones en CyL.

En la revisión (Kansagara et al., 2011), los autores investigaron modelos validados en la predicción del riesgo de reingreso hospitalario. Encontraron 14 modelos basados en datos administrativos. En su mayoría, incluyeron variables de comorbilidad y el uso de servicios médicos previos, consideraron variables de salud

mental y determinantes sociales. Existe una revisión más actualizada que tiene en cuenta los resultados del estudio anterior (Artetxe et al., 2018). Esta nueva revisión plantea una descripción general de los modelos de predicción para el reingreso hospitalario, teniendo en cuenta los métodos de análisis de datos y los algoritmos utilizados para construir los modelos y sintetizar sus resultados. En un estudio de 2019 (Deschepper et al., 2019), los autores aplican ML en un conjunto de datos administrativos para la predicción de reingresos hospitalarios no planificados. Estos datos administrativos incluyen la edad, datos de facturación, datos de logística y de patologías del paciente, los datos han sido estructurados según la clasificación internacional de enfermedades décima edición (CIE-10). En este estudio, los autores obtienen la predicción más precisa para el reingreso no planificado con el algoritmo RF (AUC de 0.77), y obtienen como variables más predictivas, los ingresos previos, la edad, el tipo de ingreso y los factores patológicos relacionados con el alcohol y las drogas.

Capítulo 3

Materiales

En este capítulo se describen las bases de datos utilizadas en el desarrollo de esta Tesis Doctoral. Estas bases de datos corresponden a pacientes hospitalizados con trastornos mentales de 11 hospitales públicos de CyL, en dos períodos de tiempos diferentes. En la primera etapa, se ha identificado en la base de datos de 2005-2015 el trastorno de esquizofrenia como uno de los diagnósticos principales prevalentes en los pacientes hospitalizados de CyL. Por tanto, se han incluido sólo los registros de pacientes hospitalizados con esquizofrenia para el desarrollo de los objetivos específicos de esta investigación. Los modelos predictivos se han construido teniendo en cuenta estos datos. En la fase final de esta investigación, se ha incluido la base de datos de 2015-2020. Las dos bases de datos se han unido para crear un modelo final de aplicación clínica, que predice el riesgo de reingreso de los pacientes hospitalizados con esquizofrenia.

3.1 Base de datos de hospitales públicos de CyL (2005-2015)

En esta Tesis Doctoral se analiza una base de datos de psiquiatría, que contiene 53 461 registros administrativos de pacientes hospitalizados con trastornos de salud mental. Los registros son anonimizados y corresponden a 11 complejos asistenciales públicos de CyL, en el período de 2005-2015. Esta base de datos denominada (DB1) ha sido proporcionada por el Hospital de Zamora y la Junta de CyL, y aprobada por el Comité de Ética de la Investigación con medicamentos (CEIm) área de salud Valladolid este (PI 20-1780).

En el desarrollo de esta investigación se incluyen sólo las unidades de agudo de cada hospital, con un total de 48 337 registros de ingresos. Este criterio de selección se define teniendo en cuenta que las unidades de agudos permiten la hospitalización de pacientes con episodios agudos en una corta estancia de tiempo. Los pacientes que ingresan en este tipo de unidad necesitan un tratamiento intensivo para controlar los síntomas y el riesgo que trae consigo el trastorno mental del que padecen. Se eliminaron los registros que no contenían información, obteniendo un total de 47 805 registros de hospitalización (Ver Figura 3.1). Se seleccionó un conjunto de datos mínimo que sólo incluía a los pacientes ingresados con un diagnóstico principal de esquizofrenia, en el período de 2005-2015 (6 822 registros). Se excluyeron los registros con un 80% de datos nulos, y se analizaron 6 089 registros administrativos de hospitalización, correspondientes a 3 065 pacientes con esquizofrenia con un rango de edad entre 15 y 96 años. Los datos siguen la clasificación internacional de enfermedades novena edición (CIE-9) (Commission on Professional and Hospital Activities, 2014).

Para dar cumplimiento a uno de los objetivos específicos de esta Tesis Doctoral, se seleccionaron además de los registros de esquizofrenia, un total de 5 916 registros de hospitalización que corresponden a 3 931 pacientes con otros trastornos mentales. Los trastornos mentales incluidos como diagnóstico principal fueron: el trastorno bipolar, trastornos degenerativos, depresión, trastornos por drogas y abuso de sustancias, trastornos afectivos y otras psicosis. Estos registros fueron seleccionados por el psiquiatra experto teniendo en cuenta la mayor prevalencia del diagnóstico en la base de datos. En la Figura 3.1 se muestra el flujo de registros utilizados en la investigación.

Los datos incluidos en la BD1 son los establecidos en el Conjunto Mínimo Básico de Datos al alta hospitalaria (CMBD) (*Registro de Altas-CMBD Estatal. Manual de Definiciones y Glosario de Términos*, 2017), e incluyen información

demográfica, características de episodios de hospitalización, diagnósticos y procedimientos referentes al paciente hospitalizado.

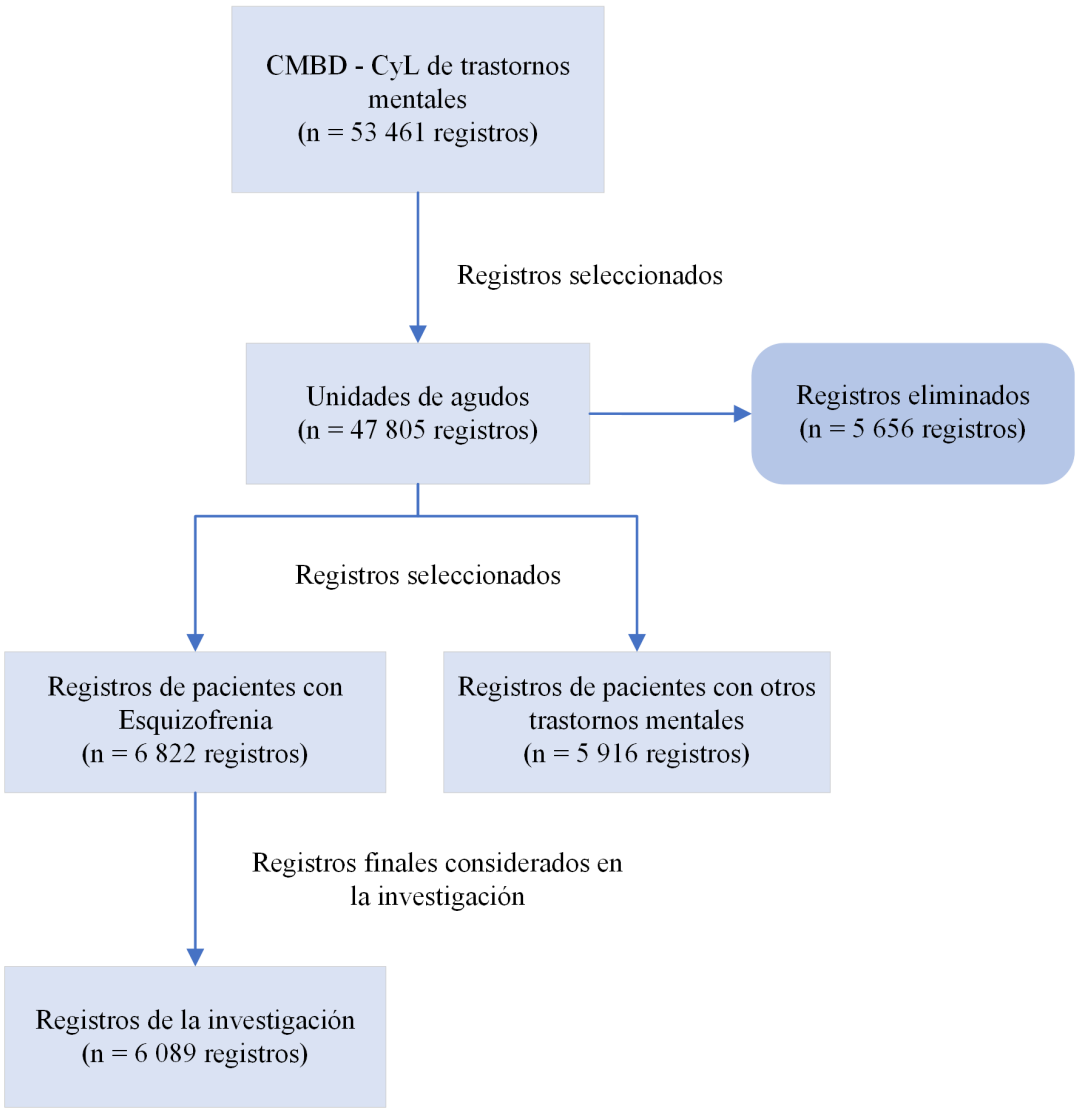

Figura 3.1 Diagrama de selección de registros de la investigación.

Los datos que contiene la BD1 son: a) nombre del centro sanitario donde hospitaliza, b) edad, c) género, d) zona básica de salud (ZBS), e) fecha de hospitalización, f) fecha de alta hospitalaria, g) identificación de la unidad de psiquiatría en la que se ingresa al paciente, h) tipo de ingreso, i) tipo de alta

hospitalaria, j) D1: diagnóstico principal, que incluye los diferentes subtipos de esquizofrenia descritos en el CIE-9 con los códigos 295.0-295.9, k) D2-D10: diagnósticos secundarios, l) PROC1-PROC3: procedimientos realizados al paciente durante la hospitalización. Estos datos son de carácter administrativo, son datos globales del paciente y no están basados en su psicopatología clínica.

3.2 Base de datos de hospitales públicos de CyL (2015-2020)

En la fase final de esta Tesis Doctoral, se analizó la base de datos de los hospitales públicos de CyL en el período de 2015-2020 (BD2), que contiene un total de 24 046 registros administrativos de pacientes hospitalizados. Esta base de datos ha sido proporcionada por el Hospital Universitario Río Hortega de Valladolid, contiene datos anonimizados y ha sido aprobada por el CEIm área de salud Valladolid oeste Ref. 22-PI009.

Teniendo en cuenta los criterios de inclusión de la BD1, se han seleccionado los registros de pacientes hospitalizados con esquizofrenia como diagnóstico principal, obteniendo un total de 5 037 registros que corresponden a 2 347 pacientes. Esta base de datos es más completa que la anterior, sigue los criterios establecidos en el CMBD CyL (*Manual de Procedimientos Del Conjunto Mínimo Básico de Datos. Castilla y León*, 2019). Los datos siguen la clasificación internacional de enfermedades décima edición (CIE-10) e incluyen la información de episodios de hospitalización, las características del paciente, los diagnósticos y procedimientos aplicados durante la estancia.

La información de los registros de la BD2 incluyen las siguientes variables: 1) género, 2) edad, 3) régimen de financiación, 4) fecha de ingreso, 5) fecha de alta hospitalaria, 6) ZBS, 7) identificación del servicio en que se hospitaliza al paciente, 8) tipo de alta hospitalaria, 9) nombre del hospital, 10) centro específico al que pertenece el paciente, 11) médico responsable del alta o atención del paciente, 12) D1: diagnóstico principal, 13) D2-D20: diagnósticos secundarios que coexisten con

el diagnóstico principal en el momento del ingreso o se desarrollan a lo largo de la asistencia sanitaria del paciente, 14) PROC1-PROC20: procedimientos realizados en el centro sanitario relacionados con el diagnóstico, 15) PROCEXT1-PROCEXT6: procedimientos realizados en otros centros distinto al que hospitaliza, 16) ingreso en unidad de cuidados intensivos (UCI), 17) días en UCI, 18) entrada quirúrgica del paciente, 19) salida quirúrgica del paciente, 20) tipo de anestesia utilizada en los procedimientos quirúrgicos, 21) morfología de las neoplasias reflejadas tanto en el diagnóstico principal como en los diagnósticos secundarios. Además, la BD2 incluye unas variables de desagregación, que se encargan de agrupar diferentes grupos de diagnósticos que sean clínicamente homogéneos. Estas variables de desagregación se asocian a los niveles de severidad y riesgo de mortalidad del paciente.

Capítulo 4

Métodos

En este capítulo se describe la metodología llevada a cabo durante el desarrollo de esta Tesis Doctoral. En la Figura 4.1 se muestra el proceso de desarrollo de esta investigación. Se realizó un preprocesamiento de los datos a través de un análisis exploratorio general. Posteriormente, se seleccionan las características del conjunto de datos que proporcionan información relevante, y se obtienen los factores de riesgo asociados a esta población. Con las características seleccionadas y utilizando diferentes algoritmos de ML, se desarrollan modelos para la predicción de hospitalización y reingreso. En esta fase, se plantea un primer estudio relacionado con los algoritmos de ML que mejor se ajustan a este tipo de datos. En este sentido, se han comparado las métricas de rendimiento de varios algoritmos, enfocados en la hospitalización de estos pacientes. Posteriormente, teniendo en cuenta los resultados de la investigación inicial (Góngora Alonso et al., 2022) y la literatura previa, se han desarrollado varios modelos que predicen el reingreso de los pacientes con esquizofrenia. A continuación, se unen ambas bases de datos para crear el modelo final con el algoritmo que obtiene el mejor rendimiento, RF. Por último, se desarrolla e implementa con el modelo final una aplicación web, que calcula el riesgo de reingreso de los pacientes hospitalizados con esquizofrenia en CyL. En el análisis, desarrollo y validación de esta investigación se utilizó el entorno de desarrollo integrado RStudio del lenguaje de programación R.

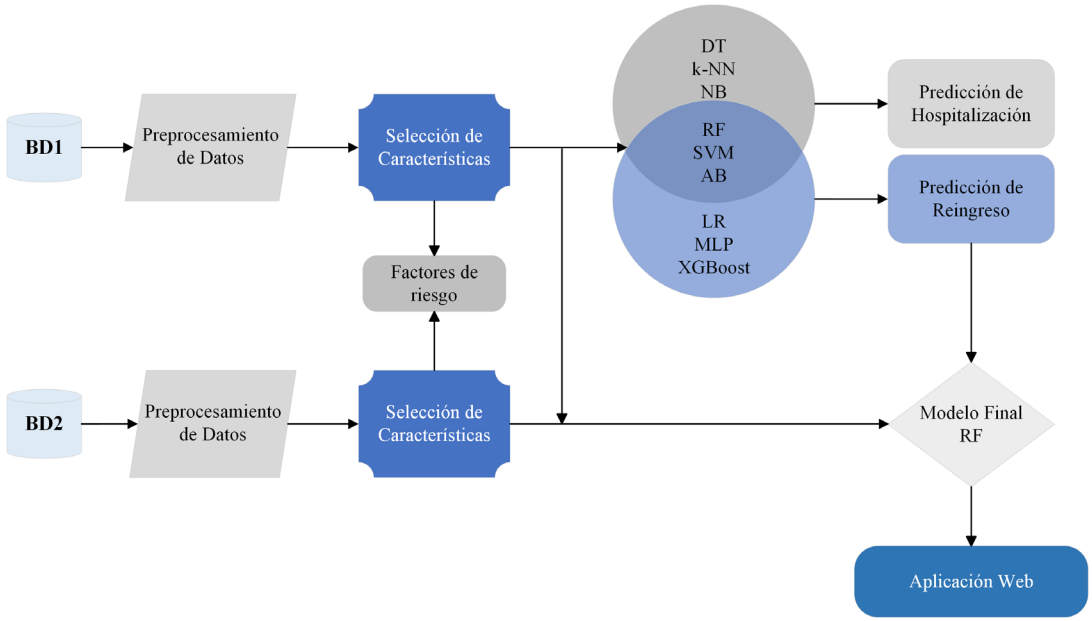

Figura 4.1 Esquema de la metodología general de esta Tesis Doctoral.

4.1 Preprocesamiento de los datos

En esta fase de preprocesamiento se realizó un análisis exploratorio general del conjunto de datos. Se realizaron transformaciones en las variables que se reflejan en la Tabla A.1.1 y B.1.1 del Apéndice A y B de esta Tesis Doctoral respectivamente.

En la BD1 se creó la variable duración de la estancia calculando la diferencia entre la fecha de hospitalización y el alta del paciente. Las bases de datos administrativas presentan los diagnósticos secundarios del paciente sin clasificarlos por tipo. En este sentido, se transformaron las variables de diagnósticos secundarios (D2-D10) del conjunto de datos original agrupándolas por tipo de comorbilidad: D2) Trastornos mentales, D3) Abuso de sustancias, D4) Enfermedades infecciosas, D5) Neoplasias, D6) Enfermedades endocrinas, D7) Enfermedades de la sangre, D8) Enfermedades del sistema circulatorio, D9) Enfermedades del aparato respiratorio, D10) Enfermedades del aparato digestivo, D11) Enfermedades del aparato

genitourinario, D12) Enfermedades de la piel y del tejido subcutáneo, D13) Enfermedades del sistema osteomioarticular, D14) Síntomas, signos y afecciones mal definidas, D15) Lesiones y envenenamientos y D16) Códigos de diagnóstico V (V02.52-V88.01). Los códigos de diagnóstico V se registran como "diagnósticos" o "problemas" cuando el paciente presenta una circunstancia o problema que influye en su estado de salud, pero que no constituye una enfermedad o lesión clasificada en las categorías 001-999 de la CIE-9. Las variables Proc1-Proc3 contienen los procedimientos o terapias aplicadas a las comorbilidades del paciente descritas en la CIE-9 (Códigos 01-99). Estas variables identifican los procedimientos que se realizan en el hospital, que implican un riesgo o son relevantes para el estudio y tratamiento del paciente.

La BD2 se ha incluido en la fase final de esta Tesis Doctoral y se ha utilizado para crear el modelo predictivo final. Esta nueva base de datos se ha unido con la BD1 conformando un conjunto final de 11 126 registros de pacientes hospitalizados con esquizofrenia (5 412 pacientes). Por tanto, se han eliminado las variables que no coincidían con la BD1. Se creó la variable duración de la estancia calculando la diferencia entre la fecha de hospitalización y el alta del paciente. Se transformaron las variables de diagnósticos secundarios (D2-D20) del conjunto de datos original agrupándolas por tipo de comorbilidad, tal y como se explicó en el procesamiento de la BD1. Además, las variables de diagnósticos y de procedimientos se recodificaron al CIE-9. Esta recodificación del CIE-10 al CIE-9 se realizó porque todos los modelos predictivos fueron creados a partir de la BD1, y de esta forma mantener la misma codificación en el modelo final.

En ambas bases de datos, se realizó un análisis exploratorio general teniendo en cuenta el tipo de datos. Se analizaron los estadísticos básicos en cuanto a valores negativos, mínimos, máximos y medias. Se identificaron los valores ausentes, redundantes, ceros y los valores atípicos, para descartar registros o variables que no aporten información. Se eliminaron los dobles espacios en blanco y los caracteres

especiales para evitar falsas clasificaciones. Por último, se analizaron las variables con varianza próxima a cero.

Una vez transformadas las variables se obtuvo un conjunto de datos de 45 variables, de las cuales 32 corresponden a los diagnósticos del paciente. Estas variables se dividen en, el código del diagnóstico (16 variables) y la descripción del diagnóstico (16 variables). Las variables de procedimiento del paciente se dividen en, el código de procedimiento (3 variables) y la descripción del procedimiento (3 variables). Por tanto, se eliminan 19 variables con información redundante.

4.2 Selección de características

La selección de características en el desarrollo de modelos predictivos permite identificar las variables que proporcionan información irrelevante o causan sobreajuste. Por tanto, se incluyó una fase de selección de características que consta de dos partes. En una primera parte, se tuvo en cuenta la experiencia clínica del psiquiatra experto y los factores de riesgo de reingreso de los pacientes con esquizofrenia descritos en la literatura (Artetxe et al., 2018; Grudnikoff et al., 2019; Hung et al., 2017; Lorine et al., 2015; Morel et al., 2020; Portela et al., 2022; Rozin et al., 2019; Sugisawa et al., 2022; Thomsen et al., 2018). En la segunda parte de la selección de características, se aplica el algoritmo RF para obtener la importancia de las variables predictoras, y se utilizaron pruebas no paramétricas. Con las pruebas no paramétricas se analiza la relación entre las variables identificadas por el algoritmo RF y el reingreso de los pacientes con esquizofrenia, estableciendo un intervalo de confianza del 95%. En la sección 4.4 se describe el análisis estadístico aplicado.

RF es un método de ML que adopta *bootstrap*, una técnica de remuestreo aleatorio y un método de división de nodos para construir varios DT, obteniendo mediante votación el resultado final de clasificación o predicción (Breiman, 2001). RF es muy útil en la exploración de datos, ya que permite identificar rápida y eficazmente las variables predictoras del conjunto. Los datos utilizados para construir

cada árbol se muestrean sin reemplazo de los datos de entrenamiento originales y, en cada división, las variables candidatas son un subconjunto aleatorio de todas las variables. Para obtener árboles sesgados bajos, cada árbol crece completamente. Al mismo tiempo, el embolsado y la división aleatoria dan lugar a una baja relación de los árboles individuales. RF logra una alta precisión de predicción incluso para datos de alta dimensión con características correlacionadas y redundantes. Por tanto, se ha convertido en uno de los métodos de selección de características más utilizados en este campo.

Una de las medidas de importancia de las variables de este algoritmo, se basa en un criterio denominado índice de Gini (IG). El IG mide el grado de impureza de un conjunto de datos (Breiman, 2001). Cada nodo interno de un árbol de decisión se construye utilizando el criterio de Gini. Para dividir un nodo, la variable y el umbral se eligen de modo que, los subconjuntos divididos tengan un Gini combinado tan pequeño como sea posible (o la impureza más baja posible). En el contexto de la estimación de la importancia de las variables, el IG puede interpretarse como el grado de discriminación de una variable entre las clases.

Teniendo en cuenta un conjunto de datos *D* que cuenta con *m* muestras de *k* clases, el conjunto *D* es más puro a medida que *D* es más homogéneo o sea *k* es más pequeño. El IG se calcula con la ecuación 4.1.

$$G(D) = 1 - \sum_{k} f_k^2 \tag{4.1}$$

Donde *f* es un conjunto de frecuencias normalizadas ($f_1 + f_2, + \cdots + f_k$) en función de las clases *k*. En esta investigación, se extrajeron 1000 muestras bootstrap con reemplazo de los datos originales. Para cada una de las muestras bootstrap, se construyó un árbol de clasificación con el número de variables probadas en cada división fijado en 5. El árbol de clasificación obtenido se utilizó para predecir los

datos no incluidos en la muestra *bootstrap*. Teniendo en cuenta el error de estimación *out-of-bag* (*OOB*) el algoritmo calculó la importancia de cada variable.

Con el análisis de selección de las variables predictivas se obtuvo un total de 22 características, por lo que se han considerado como variables predictoras a utilizar en la predicción del riesgo de reingreso de los pacientes con esquizofrenia. La variable dependiente (*target*) se creó utilizando la ecuación 4.2.

$$target = \begin{cases} 0, & Si\ H = 0 \\ 1, & Si\ H > 0 \end{cases} \tag{4.2}$$

Donde *H* es el número de hospitalizaciones del paciente. Para determinar *H* se tuvo en cuenta el número de ingresos de cada paciente en el período de 2005 a 2015.

4.3 Algoritmos de Machine Learning

El ML es una rama de la inteligencia artificial basado en algoritmos computacionales, capaces de reconocer de forma automática patrones complejos en los datos. Se puede dividir en diferentes tipos, aprendizaje supervisado, no supervisado y por refuerzo (Veronese et al., 2013; Vieira et al., 2020). Teniendo en cuenta que nuestro conjunto de datos está etiquetado, en esta Tesis Doctoral usamos diferentes algoritmos de aprendizaje supervisado para predecir el reingreso de los pacientes hospitalizados con esquizofrenia. En esta sección, se describen los algoritmos clasificación de ML analizados durante todo el desarrollo de esta investigación.

4.3.1 Logistic Regression

LR es un método de regresión utilizado para la clasificación binaria, donde las observaciones se clasifican en un grupo u otro en función del valor de la variable utilizada como predictor (Hosmer & Lemeshow, 2002). Este algoritmo permite

estimar la probabilidad $p(Y_j|x_i)$ de que se produzca una de las dos opciones contempladas en la variable dependiente (Y_j, donde $j = 1, 2$) en función de la variable predictora x_i. La función del algoritmo viene dada por la ecuación 4.3.

$$p(Y_j|x_i) = \frac{e^{b_0+\sum_{i=1}^{n} b_i * x_i}}{1 + e^{b_0+\sum_{i=1}^{N} b_i * x_i}} \tag{4.3}$$

Donde *n* es el número de variables predictoras x_i, y los coeficientes del modelo son b_0 *y* bi. Para obtener un modelo adecuado es necesario definir y ajustar estos coeficientes, esto se puede hacer utilizando un estimador de mínimos cuadrados. Este estimador permite encontrar los valores de *bi* maximizando la probabilidad de obtener el conjunto de valores observados.

4.3.2 Naïve Bayes

NB es una técnica de clasificación que se encarga de resolver el problema de la relación no determinista entre la clase objetivo y las variables definidas, utilizando la teoría de probabilidades (Hastie et al., 2001). A partir de una muestra de datos de entrenamiento de *t* objetos clasificados, el algoritmo estima la probabilidad *p* (Y | *x*) de que una instancia $x = (x_1, \dots, x_t)$ pertenece a alguna clase Y. El clasificador óptimo de Bayes asigna instancia $x = (x_1, \dots, x_t)$ a la clase Y_j con la ecuación 4.4.

$$p(Y_j \,|x_1, \dots, x_t) = \frac{p(x_1, \dots, x_t|Y_j)p(Y_j)}{p(x_1, \dots, x_t)} \tag{4.4}$$

Teniendo en cuenta que $p(x_1, \dots, x_t)$ es constante para una instancia, la probabilidad condicional en el clasificador se puede expresar con la ecuación 4.5.

$$p(Y_j \mid x_1, \ldots, x_t) \propto \prod_{i=1}^{t} p(x_i|Y_j)\, p(Y_j) \tag{4.5}$$

Donde $p\,(x_i|Y_j)$ es la estimación de distribución para cada valor de *i* y *j*. Se pueden obtener estimaciones de las probabilidades mediante la normalización de todas las clases posibles, lo que permite al clasificador predecir no solo la clase sino también la probabilidad de cada una de las clases.

4.3.3 k-Nearest Neighbors

kNN es un algoritmo de clasificación supervisado, no paramétrico. Este algoritmo encuentra el promedio de los puntos de datos vecinos que comparten características más comunes con el nuevo punto de datos (Kuhn & Johnson, 2013). Para determinar la clase o el valor de un punto de datos, es necesario obtener la distancia entre el nuevo punto de dato y otros puntos de datos que se han usado para entrenar el algoritmo. Para medir la distancia de los puntos entre sí se usa la siguiente función euclidiana (Ecuación 4.6):

$$d(x_i, x_j) = \sqrt{\sum_{a=1}^{n} (x_{ai} - x_{aj})^2} \tag{4.6}$$

Una vez calculada la distancia de los puntos entre sí, es necesario determinar el número de vecinos *k*. En esta investigación el valor de *k*=5. Cuanto mayor sea el valor de *k* mejor será la precisión del modelo, ya que hay más puntos de datos involucrados en la votación de la clase o el valor del nuevo punto de datos. Sin embargo, para valores altos de *k* calcular la distancia de los puntos puede tener un alto coste computacional.

4.3.4 Decision Tree

DT es un algoritmo de aprendizaje supervisado utilizado para modelos de regresión y clasificación. Son modelos secuenciales que combinan de forma lógica una secuencia de pruebas sencillas, cada prueba compara un atributo numérico con un valor umbral o un atributo nominal con un conjunto de posibles valores (Kotsiantis, 2013). El algoritmo clasifica los elementos de datos exponiendo distintas preguntas sobre las características asociadas con los elementos. Cada una de las preguntas expuestas está incluida en un nodo, y cada nodo interno apunta a un nodo secundario para cada posible respuesta a su pregunta. De esta forma, las preguntas forman una jerarquía codificada como un árbol.

A lo largo de los años se han desarrollado diferentes algoritmos de DT. En esta investigación usamos el algoritmo de clasificación CART propuesto por (Breiman, 1984), el cual desarrolla reglas de decisión visualizadas para predecir una variable categórica. Tiene la capacidad de identificar las variables más significativas y eliminar las no significativas. La regla de división en la clasificación se mide por el IG para cuantificar la homogeneidad de los datos. El IG se calcula con la ecuación (4.1) planteada en la sección 4.2.

En el estudio se ha usado una profundidad máxima de 100 y el algoritmo deja de dividir cuando la mayoría de los nodos alcanzan el 95%. Aunque los algoritmos de DT por sí solos pueden ser excelentes clasificadores, a menudo se puede lograr una mayor precisión combinando múltiples árboles como el algoritmo RF.

4.3.5 Adaptive Boosting

AdaBoost es un algoritmo de ML adaptativo, que construye un clasificador fuerte combinando múltiples clasificadores débiles (*weak classifiers*) (Freund & Schapire, 1997). El algoritmo se logra cambiando la distribución de datos, de acuerdo con la corrección de clasificación de la muestra del conjunto de entrenamiento. El algoritmo en un inicio asigna el mismo peso a todas las

observaciones del conjunto de entrenamiento. Cuando se realiza la predicción del primer clasificador se actualizan las observaciones clasificadas correctamente disminuyendo su peso, mientras que las que son clasificadas de forma errónea son identificadas y se les asigna un mayor peso. En la segunda iteración el siguiente clasificador *weak* corrige las observaciones que han sido clasificadas erróneamente anteriormente. De esta forma, el nuevo clasificador se adapta a las observaciones minimizando el error ϵ_n asociado al clasificador. Como el proceso es iterativo el algoritmo sigue añadiendo clasificadores *weak* hasta alcanzar el número máximo establecido N. La predicción de la clasificación $H(x)$ se obtiene a partir del promedio ponderado de los *n* clasificadores h_n, utilizando la ecuación 4.7.

$$H(x) = sign\left[\sum_{n=1}^{N} \alpha_n h_n(x)\right] \tag{4.7}$$

Donde N es el número de clasificadores *weak* y α_n es el peso del clasificador *weak n*, que se calcula con la ecuación 4.8.

$$\alpha_n = {}^1\!/_2 \, log\left(\frac{1-\epsilon_n}{\epsilon_n}\right) \tag{4.8}$$

De esta forma el algoritmo AB es capaz de crear un clasificador robusto a partir de clasificadores *weak* usados en el proceso.

4.3.6 Random Forest

RF se describe como una técnica de aprendizaje de conjunto ya que combina los resultados de múltiples árboles de decisión que devuelven una única predicción (Breiman, 2001). Este algoritmo posee una gran capacidad para analizar las características de clasificación de datos complejos y multidimensionales, a través de

la rápida velocidad de aprendizaje, que también es robusta para conjuntos de datos con ruido y valores faltantes.

El RF es un algoritmo formado por un conjunto de clasificadores estructurados en forma de árbol $\{h\ (\boldsymbol{x},\ \theta_j),\ j\ =\ 1,\ \ldots\}$ donde $\{\theta_j\}$ son vectores aleatorios distribuidos de forma independiente y cada árbol emite un voto unitario para la clase más popular en la entrada x (Breiman, 2001).

Tal y como se muestra en la Figura 4.2 el algoritmo genera múltiples árboles y cada conjunto representa un árbol. En la medida que se van construyendo los árboles se van realizando cortes binarios, donde se seleccionan al azar j características del total p variables predictoras ($j < p$). El nodo d se calcula a través del mejor punto de división entre las j características, y se divide d en nodos hijos a través de la mejor división. Cada árbol clasifica una clase dando como resultado la clase con mayor número de votos.

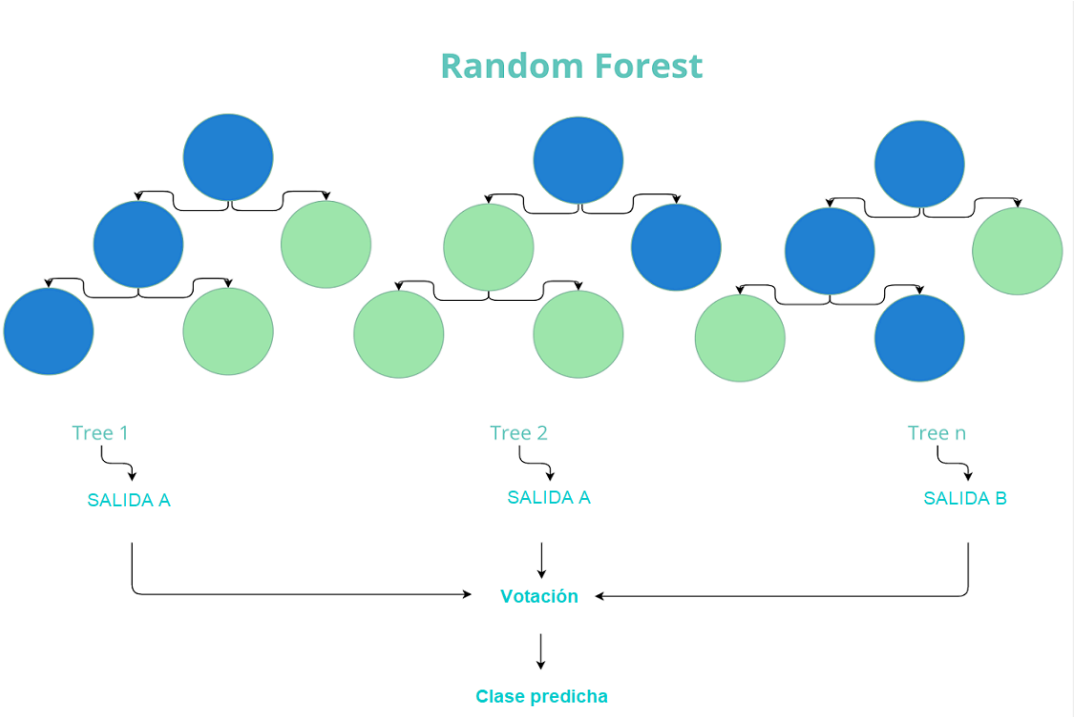

Figura 4.2 Diagrama del algoritmo Random Forest. Figura extraída de (Góngora Alonso et al., 2022)

Este algoritmo depende de dos parámetros principalmente, el número de variables p que se seleccionan en cada nodo (Mtree) y el número de árboles que forman el algoritmo (Ntree). Aunque el tipo de árbol cambia en el algoritmo, el parámetro de ajuste del número de predictores seleccionados aleatoriamente para elegir en cada división, es el mismo (se denomina *mtry*). Para los problemas de clasificación, (Breiman, 2001) se recomienda fijar mtry en la raíz cuadrada del número de predictores.

4.3.7 Extreme Gradient Boosting

XGBoost se basa en árboles de decisión y utiliza un marco de refuerzo de gradiente (Huberts et al., 2022). Permite utilizar una amplia gama de aplicaciones para resolver problemas de predicción, clasificación y regresión definidos por el usuario. El algoritmo usa un conjunto de datos $C = \{(x_j y_j)\}$, donde x son los datos de entrenamiento y la variable objetivo es y. XGBoost tiene un conjunto de árboles N cuyos resultados de predicción individuales se denotan como $f_n(x_j)$, el resultado final estimado se calcula con la ecuación 4.9.

$$\tilde{y}_j = \sum_{n=1}^{N} f_n(x_j) \tag{4.9}$$

El algoritmo cuenta con una función objetivo que se divide en dos términos, la función de pérdida diferenciable L y el término de regularización Ω, tal y como se muestra en la ecuación 4.10.

$$obj(f_k) = \sum_{j=1} L(\tilde{y}_j, y_j) + \sum_{n} \Omega(f_k) \tag{4.10}$$

L mide el error entre la salida estimada $\tilde{y}_j$ y la salida real y_j, mientras que Ω evita el sobreajuste. El término de regularización se calcula con la ecuación 4.11.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{4.11}$$

Donde *T* representa el número de hojas por árbol, *w* es el peso de las hojas de cada árbol, γ y λ son constantes para controlar el grado de regularización. La fórmula para calcular el valor predictivo $\tilde{y}_j^{(k)}$ después de k_{th} iteraciones es la siguiente (Ecuación 4.12):

$$\tilde{y}_j^{(k)} = \tilde{y}_j^{(k-1)} + f_k(x_j) \tag{4.12}$$

Utilizando la expansión de Taylor de segundo orden, la fórmula final de la función objetivo del algoritmo se obtiene con la ecuación 4.13.

$$obj(f_k) = \sum_{i=1}^{T}\left[\left(\sum_{j\in J_i} g_j\right)w_i + {}^1\!/_2\left(\sum_{j\in J_i} h_j + \lambda\right)w_i^2\right] + \gamma T \tag{4.13}$$

Donde $J_i = \{j|q(x_j) = i\}$ se define como el nodo hoja i_{th}, $g_j = \partial_{\tilde{y}_j^{(k-1)}} L(y_j, \tilde{y}_j^{(k-1)})$ y $h_j = \partial_{\tilde{y}_j^{(k-1)}}{}^2 L(y_j, \tilde{y}_j^{(k-1)})$ son estadísticos de gradientes de primer y segundo orden de la función de pérdida.

4.3.8 Multi-layer Perceptron

MLP es una red neuronal formada por múltiples capas (Rathbun et al., 1997). Su arquitectura se compone de múltiples capas interconectadas en las que cada capa consta de nodos. Cada neurona está conectada a todas las neuronas tanto de la capa

anterior como de la capa siguiente. Al combinar múltiples capas ocultas y funciones de activación no lineales, los modelos de red pueden aprender prácticamente cualquier patrón. Las interconexiones de capas se implementan como matrices de peso. Los nodos de la capa de salida representan el conjunto de etiquetas de clase, presentes en el conjunto de datos de entrenamiento. Las matrices de peso entre las capas de entrada y salida que forman las capas ocultas son responsables de aprender las representaciones latentes de los datos de entrada a través de transformaciones no lineales. La ecuación de un nodo en las capas ocultas es la siguiente (Ecuación 4.14):

$$h_{1j} = f\left(\sum_{i=1}^{n} w_{ij}\, x_i + \, b_j\right) \tag{4.14}$$

Donde h_{1j} se define como el nodo j de la capa oculta h_1, w_{ij} representa el peso de conexión entre la neurona de entrada i y la neurona oculta j, x_i es la característica de estrada i, y b_j es el sesgo asociado a la neurona oculta j. Estas capas identifican distribuciones de datos subyacentes y las asignan a uno de los tipos o clases, designados por un nodo en la capa de salida.

El aprendizaje del perceptrón cambia y adapta el parámetro de peso (w) del umbral. Esto significa que el algoritmo aprende automáticamente los coeficientes de peso óptimos. El principio y la base del aprendizaje en las redes neuronales se realiza por iteración, se introduce un conjunto de datos en el algoritmo varias veces, el algoritmo aprende ajustando los pesos y el sesgo puede reconocer diferencias en los datos de entrenamiento.

4.3.9 Support Vector Machine

SVM es un algoritmo supervisado de clasificación binaria, que representa los puntos de una muestra en el espacio, separando mediante un hiperplano de separación las clases en dos espacios (Bishop & Nasrabadi, 2006). Este algoritmo es

preciso en espacios de alta dimensión, utiliza en la función de decisión un subconjunto de puntos de entrenamiento, que se conoce como vectores de soporte.

El hiperplano de separación maximiza la distancia mínima entre los datos de diferentes clases en un nuevo espacio, que se obtiene al aplicar una función de kernel a los datos originales. Los datos de entrada son dos conjuntos de *n* vectores dimensionales.

El proceso de entrenamiento de una función de decisión SVM equivale a identificar un hiperplano reproducible, que maximiza la distancia (margen) entre los vectores de soporte de ambas clases. Así, el hiperplano óptimo es aquel que "maximiza el margen" entre clases. Una SVM puede ser lineal o no lineal, va a depender del parámetro kernel que se seleccione. La gran ventaja de este algoritmo es que hace uso de kernel para aumentar el espacio vectorial de las variables predictoras, con el fin de establecer un límite no lineal entre grupos de observaciones. Teniendo en cuenta la lógica planteada en el estudio (James et al., 2021), el algoritmo se puede plantear con la ecuación 4.15.

$$f(x) = \beta_0 + \sum_{i=1}^{m} \alpha_i * K(x, x_i) \tag{4.15}$$

Donde existen m parámetros α_i, $i = 1, \dots, m$, por cada observación de entrenamiento. Para estimar los parámetros $\alpha_1, \dots, \alpha_m$ y β_0, necesitamos el producto interno $(x_i, x_{\hat{i}})$ entre todos los pares de observaciones de entrenamiento. Para evaluar la función $f(x)$, necesitamos calcular el producto interno entre el nuevo punto x y cada uno de los puntos de entrenamiento x_i. La función kernel de base radial se calcula con la ecuación 4.16.

$$K(x_i, x_{\hat{i}}) = e^{\left(-\gamma \sum_{j=1}^{k} (x_{ij} - x_{\hat{i}j})^2\right)} \tag{4.16}$$

Donde γ es un hiperparámetro que debe ajustarse para cada SVM. En esta investigación usamos validación cruzada para obtener los hiperparámetros que mejor ajustaban el modelo de SVM, penalización $C = 0.500$ y $\gamma = 0.125$.

4.4 Análisis estadístico

Para interpretar y evaluar los resultados obtenidos en el desarrollo de los modelos predictivos, en esta Tesis Doctoral, se aplicaron diferentes pruebas estadísticas, métricas de rendimiento y métodos de validación.

4.4.1 Pruebas estadísticas

Las pruebas estadísticas son herramientas que permiten evaluar si es posible inferir propiedades de una población a partir de los resultados observados en una muestra de datos. En esta investigación, se utilizaron diferentes pruebas estadísticas para evaluar la distribución normal de los datos y la homocedasticidad (igualdad de varianzas). Se aplicó la prueba de Lilliefors para evaluar la normalidad de cada una de las variables del estudio. El test de Levene se usó para evaluar la homocedasticidad. Se observó que las variables no se distribuyen normalmente y no son homocedásticas. Por tanto, se utilizan pruebas no paramétricas para determinar las diferencias estadísticamente significativas entre los grupos. Para analizar la asociación entre las variables predictoras y la variable de readmisión, se aplicó la prueba U de Mann-Whitney para las variables cuantitativas y la prueba X^2 (Chi-cuadrado) para las variables cualitativas. Se estableció un *p-value* < 0,05 para evaluar las diferencias significativas.

4.4.2 Métricas de rendimiento de los modelos

Las métricas de rendimiento de los modelos se derivan del número de sujetos clasificados correctamente y de forma errónea. En la clasificación estadística estas

medidas se derivan de la matriz de confusión, que es una herramienta que permite comparar los valores predichos con los valores reales.

- Verdaderos positivos (VP): Es el número de sujetos con reingreso que han sido clasificados correctamente.
- Verdaderos negativos (VN): Es el número de sujetos que no reingresan que han sido clasificados correctamente.
- Falsos positivos (FP): Es el número de sujetos positivos (reingreso) que han sido clasificados de forma errónea como un no reingreso.
- Falsos negativos (FN): Es el número de sujetos negativos (no reingreso) que han sido clasificados de forma errónea como un reingreso.

A partir de estos términos, en esta investigación se han calculado las siguientes métricas de rendimiento de los modelos.

- Exactitud o Accuracy (Acc): Es la proporción de sujetos clasificados correctamente por el modelo (Ecuación 4.17).

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.17}$$

- Precisión: Es la relación entre el número de resultados positivos identificados correctamente y el total de elementos positivos (reingresos) (Ecuación 4.18).

$$Precision = \frac{VP}{VP + FP} \tag{4.18}$$

- Sensibilidad o Exhaustividad (En inglés *recall*): Es la proporción de sujetos con reingresos clasificados correctamente (Ecuación 4.19).

$$Recall = \frac{VP}{VP + FN} \tag{4.19}$$

- Especificidad: Es la proporción de sujetos sin reingresos clasificados correctamente (Ecuación 4.20).

$$Especificidad = \frac{VN}{VN + FP} \tag{4.20}$$

- F1-score: Es la media armónica entre la precisión y el recall (Ecuación 4.21).

$$F1_{score} = 2 * \frac{(Precision * Recall)}{Precision + Recall} \tag{4.21}$$

- AUC: Una curva ROC es un gráfico de sensibilidad vs 1-especificidad de una prueba de diagnóstico (Zweig & Campbell, 1993). La AUC es una forma efectiva de resumir la precisión diagnóstica general de la prueba. Toma valores de 0 a 1, donde un valor de AUC de 0.5 sugiere que la prueba diagnóstica no tiene capacidad de discriminación mientras que con valores por encima de 0.5 se considera que la curva ROC tiene una capacidad de discriminación razonable para diagnosticar si el paciente reingresa o no (Mandrekar, 2010). Por tanto, un valor de AUC cercano a 1 indica que el modelo predictivo es más eficaz.

4.4.3 Método de validación y ajuste de hiperparámetros

Para validar los resultados obtenidos a lo largo de esta Tesis Doctoral se ha utilizado el método de validación cruzada *k*-fold, de acuerdo con los estudios de revisión sistemática (Artetxe et al., 2018; English et al., 2022; Huang et al., 2021) y los estudios planteados en la sección 2.2. La validación cruzada es un método de remuestreo de datos para evaluar la capacidad de generalización de los modelos predictivos y evitar el sobreajuste (Hastie et al., 2001).

El método de validación cruzada *k*-fold implica dividir aleatoriamente el conjunto de muestras en una serie de particiones de igual tamaño, donde *k* indica el número de particiones en que se divide el conjunto de datos. En esta investigación, se utiliza un valor de *k*=10, por tanto, el conjunto de datos se divide en diez particiones. El modelo se entrena utilizando *k*-1 subconjuntos, mientras que la partición restante se usa para datos de prueba. Este procedimiento se repite 10 veces hasta que cada uno de los *k* subconjuntos haya servido como conjunto de validación. Su valor final estimado será la media de los valores obtenidos en las *k* iteraciones (Kuhn, 2019).

El ajuste de los hiperparámetros de los modelos predictivos que se realizó en esta investigación tenía como objetivo mejorar el rendimiento de los algoritmos de ML. Para la selección de los hiperparámetros se utilizó la validación cruzada *k-fold* y se tuvo en cuenta la curva ROC, debido a que evalúa la capacidad discriminativa del reingreso de los pacientes hospitalizados con esquizofrenia. En la Tabla 4.1 se muestran los valores de cada algoritmo para construir el modelo, y los parámetros que mejor se ajustan a los datos de cada uno en las pruebas *k*=10 realizadas.

Los hiperparámetros ajustados de cada algoritmo se utilizan para construir el modelo. RF y MLP utilizan mtry = 36 y size = 3 respectivamente. El modelo XGBoost utiliza max_depth = 2, subsample = 1.00, colsample_bytree = 0.8, nrounds = 500, eta = 0.3, gamma = 0, y min_child_weight = 1. El algoritmo AB utiliza 100

iteraciones para desarrollar el modelo, mientras que SVMradial utiliza valores de *sigma* = 0.125 y C = 0.500

Tabla 4.1 Parámetros de ajuste con validación cruzada (*k*=10). Fuente: Tabla extraída de (Góngora Alonso et al., 2023).

Modelos predictivos	**Parámetros de ajuste**	**Valores**	**Mejores parámetros**
MLP	Size	1; 3; 5; 7;9 11; 13; 15; 17; 19	3
XGBoost	nrounds	50; 100; 150; 200; 250; 300; 350; 400; 450; 500	500
	max_depth	1; 2; 3; 4; 5; 6; 7; 8; 9; 10	2
	eta	0.3; 0.4	0.3
	colsample_bytree	0.6; 0.8	0.8
	subsample	0.500; 0556; 0.611; 0.667; 0.722; 0.778; 0.833; 0.889; 0.944; 1.0	1.0
SVMradial	Cost (C)	0.125; 0.177; 0.250; 0.353; 0.500	0.500
	sigma	0.0625; 0.088; 0.125; 0.177; 0.250; 0.354; 0.500; 0.707	0.125
AB	nIter	50; 100; 150	100
RF	mtry	2; 36; 681	36

4.5 Aplicación del modelo en la práctica clínica

En la fase final de esta Tesis Doctoral, se ha desarrollado una aplicación que permite calcular el riesgo de reingreso de un paciente hospitalizado con esquizofrenia cuando se le da el alta hospitalaria. El modelo utilizado en el desarrollo de esta

aplicación se construyó a partir del conjunto de datos total que conforman la BD1 y la BD2. Se ha utilizado una muestra mayor de observaciones, 11 126 registros que corresponden a 5 412 pacientes, y el algoritmo RF que presentó los mejores valores en sus métricas de rendimiento Acc y AUC. El modelo se desarrolló siguiendo el proceso metodológico de los modelos anteriores. Para la validación se ha usado validación cruzada k=10 y los hiperparámetros han sido ajustados en función de la métrica ROC.

La aplicación web se ha desarrollado en el entorno de programación R, utilizando la librería para interfaz gráfica: Shiny. Shiny es un paquete de R usado para crear aplicaciones web interactivas. Consta de dos componentes que funcionan de forma conjunta: un fichero de servidor y un fichero que contiene el diseño de la interfaz gráfica (*Shiny*, 2017). El archivo de servidor proporciona instrucciones al servidor permitiéndole crear salidas que responden a las entradas del usuario, y el archivo de interfaz de usuario actúa como intérprete de HTML (Beeley, 2016).

Partiendo de las variables de entrada, definidas en la aplicación como las variables del modelo entrenado, la aplicación calcula el riesgo de reingreso de los pacientes con esquizofrenia en los distintos hospitales de CyL. Se han representado gráficos del comportamiento de la variable abuso de sustancias por hospitales en el período de 2005-2020.

La aplicación se ha puesto en producción en el servidor web del grupo de investigación de Telemedicina y eSalud (GTe) donde se desarrolló esta Tesis Doctoral. Se puede acceder a la aplicación a través de un enlace web que se ha descrito en la Sección 5.5, y se ha desarrollado un manual de usuario que se muestra en el Apéndice D. La interfaz es de fácil acceso y arroja resultados en tiempo real, de esta forma permite al personal sanitario calcular el riesgo de reingreso de nuevos pacientes con trastornos de esquizofrenia.

Capítulo 5

Resultados

En este capítulo, se presentan los principales resultados obtenidos durante el desarrollo de esta Tesis Doctoral. Dado que el objetivo principal de esta investigación es desarrollar y evaluar modelos predictivos utilizando algoritmos de clasificación de ML, el capítulo se ha organizado en función de los objetivos específicos planteados en la sección 1.3: 5.1) análisis de las características de la población de estudio, 5.2) comparación de algoritmos de ML, 5.3) factores asociados al riesgo de reingreso en esta población, 5.4) modelos predictivos del riesgo de reingreso de los pacientes hospitalizados con esquizofrenia y 5.5) aplicación del modelo Random Forest en la práctica clínica.

5.1 Análisis de las características de la población de estudio

5.1.1 Características de los pacientes con reingreso de la BD1

En esta Tesis Doctoral, se analizaron un total de 3 065 pacientes con esquizofrenia de la BD1. Del total, 1 454 son registros de pacientes sin reingreso y 1 611 son pacientes con reingreso. La Tabla 5.1 muestra el análisis de los datos de esta investigación.

El grupo de edad más representativo de los pacientes con este trastorno es el comprendido entre los 31-50 años (p = 0.0171), con una media de 43 años en los pacientes con reingresos y una media de 47 años en los pacientes sin reingresos. Existen diferencias de género entre los pacientes, donde los hombres representan el 67.21% y las mujeres el 32.79%. La mayor incidencia de registros de pacientes que

reingresaron se muestra en el Complejo Asistencial de Burgos con un 19.68% ($p < 0.0001$), el Complejo Asistencial de León con un 12.17% ($p = 0.0002$), y el Hospital El Bierzo con un 11.85% ($p = 0.0001$).

Tabla 5.1 Análisis de las variables de la BD1. Fuente: Tabla extraída de (Góngora Alonso et al., 2023).

Variables	**Reingreso (N=1 611 pacientes)**		**No-Reingreso (N=1 454 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Grupos de Edad					< 0.0001
< 18 años	5	0.31	5	0.34	
18-30 años	334	20.73	186	12.79	
31-50 años	862	53.51	715	49.18	
51-65 años	302	18.75	374	25.72	
> 65 años	108	6.70	174	11.97	
Género					0.1659
Femenino	530	32.90	475	32.67	
Masculino	1 081	67.10	979	67.33	
Hospital					< 0.0001
Complejo Asistencial de Ávila	83	5.15	105	7.22	
Complejo Asistencial de Burgos	317	19.68	209	14.37	
Complejo Asistencial de León	196	12.17	229	15.75	
Complejo Asistencial de Palencia	102	6.33	108	7.43	
Complejo Asistencial de Salamanca	185	11.48	195	13.41	
Complejo Asistencial de Soria	79	4.90	80	5.50	
Complejo Asistencial de Zamora	177	10.99	164	11.28	
Complejo Asistencial de Segovia	100	6.21	78	5.37	
Hospital Clínico Universitario de Valladolid	130	8.07	79	5.43	

Variables	**Reingreso (N=1 611 pacientes)**		**No-Reingreso (N=1 454 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Hospital El Bierzo	191	11.85	139	9.56	
Hospital Universitario Rio Hortega	51	3.17	68	4.68	
Año de ingreso					0.4558
2005-2008	656	40.72	570	39.20	
2009-2011	408	25.33	350	24.07	
2012-2015	547	33.95	534	36.73	
Tipo_alta					< 0.0001
Domicilio	1 434	89.01	1 246	85.69	
Alta sin notificación documentada	2	0.12	0	0.00	
Traslado a otro hospital	100	6.21	155	10.66	
Traslado a centros de media y larga estancia	5	0.31	4	0.28	
Alta voluntaria	18	1.12	17	1.17	
Otros tipos de alta	52	3.23	32	2.20	
Tipo_Ingreso					0.4859
Ingreso urgente	1 554	96.46	1 407	96.77	
Ingreso programado	57	3.54	47	3.23	
Tipos Esquizofrenia					0.1058
Esquizofrenia simple	20	1.24	18	1.24	
Esquizofrenia de tipo desorganizado	75	4.66	38	2.61	
Esquizofrenia catatónica	4	0.25	5	0.34	
Esquizofrenia paranoide	1 052	65.30	956	65.75	
Esquizofrenia latente	6	0.37	3	0.21	
Esquizofrenia residual	244	15.15	231	15.89	
Otra esquizofrenia especificada	45	2.79	34	2.34	
Esquizofrenia no especificada	165	10.24	169	11.62	

Variables	**Reingreso (N=1 611 pacientes)**		**No-Reingreso (N=1 454 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Diagnósticos secundarios codificación CIE-9 (D2-D16):					
Trastornos mentales	1 124	69.77	538	37.00	< 0.0001
Abuso de sustancias	817	50.71	518	35.63	< 0.0001
Enfermedades infecciosas	111	6.89	64	4.40	0.2403
Neoplasias	27	1.68	29	1.99	0.2439
Enfermedades endocrinas	448	27.81	285	19.60	< 0.0001
Enfermedades de la sangre	69	4.28	30	2.06	0.3031
Enfermedades del sistema circulatorio	208	12.91	162	11.14	0.0015
Enfermedades del aparato respiratorio	106	6.58	64	4.40	0.0037
Enfermedades del aparato digestivo	153	9.50	82	5.64	0.0021
Enfermedades del aparato genitourinario	94	5.83	53	3.65	0.0901
Enfermedades de la piel y del tejido subcutáneo	60	3.72	30	2.06	0.2334
Enfermedades del sistema osteomioarticular	88	5.46	57	3.92	0.1476
Síntomas, signos y estados mal definidos	238	14.77	83	5.71	< 0.0001
Lesiones y envenenamientos	282	17.50	119	8.18	< 0.0016
Códigos de diagnóstico tipo V	1 259	78.15	913	62.79	< 0.0001
Procedimientos codificación CIE-9:					
Proc1	1 320	81.94	1 080	74.28	0.0006
Proc2	875	54.31	573	39.41	< 0.0001
Proc3	419	26.01	267	18.36	0.0198

Variables	**Reingreso (N=1 611 pacientes)**		**No-Reingreso (N=1 454 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Duración de la estancia (días)					< 0.0001
Media (SD)	18 (15.361)		17 (17.183)		

La estancia media de los pacientes con esquizofrenia en CyL es de 17 días. En este sentido, se ha obtenido la duración de la estancia del paciente < 30 días, que representa el 87.12% del total de hospitalizaciones. La tasa de reingreso de los pacientes analizados en la BD1 fue del 76.12%, lo que se traduce en costes sanitarios para los hospitales de CyL. La diferencia entre las variables edad ($p < 0.0001$), duración de la estancia ($p < 0.0001$) y hospital ($p < 0.0001$) respecto a los reingresos y no reingresos de los pacientes es significativa; mientras que la variable género ($p = 0.1659$) no muestra diferencias significativas. Las comorbilidades prevalentes en los pacientes con reingresos en CyL son los diagnósticos con códigos V02.52-V88.01 (78.15%, $p < 0.0001$), los trastornos mentales distintos de la esquizofrenia (69.77%, $p < 0.0001$) y el abuso de sustancias (50.71%, $p < 0.0001$).

5.1.2 Características de los pacientes con reingreso de la BD2

La BD2 incluye 5 037 registros correspondientes a 2 347 pacientes hospitalizados con esquizofrenia. El análisis de los datos clínicos muestra diferencias significativas en cuanto al género (Ver Tabla 5.2). Los hombres representan el 64.4% (1 512 de 2 347 pacientes), siendo los más afectados por este trastorno psiquiátrico, frente a las mujeres que representan el 35.6% (835 de 2 347 pacientes). En cuanto a la edad, el grupo de 31 a 50 ($p = 0.2206$) años es el más representativo con un 47.94% y un 49.71% en los grupos de reingreso y no-reingreso respectivamente. La mayor incidencia de pacientes con reingreso se registra en los complejos

asistenciales de Burgos (p = 0.0108) y León (p = 0.0073) con un total de 143 pacientes en el período de 2015-2020.

Tabla 5.2 Análisis de las variables de la BD2

Variables	**Reingreso (N=945 pacientes)**		**No-Reingreso (N=1 402 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Grupos de Edad					0.0035
< 18 años	6	0.63	13	0.93	
18-30 años	124	13.12	185	13.20	
31-50 años	453	47.94	697	49.71	
51-65 años	295	31.22	369	26.32	
> 65 años	67	7.09	138	9.84	
Género					0.2249
Femenino	348	36.83	487	34.74	
Masculino	597	63.17	915	65.26	
Hospital					< 0.0001
Complejo Asistencial de Ávila	40	4.23	72	5.14	
Complejo Asistencial de Burgos	143	15.13	244	17.40	
Complejo Asistencial de León	143	15.13	220	15.69	
Complejo Asistencial de Palencia	68	7.20	107	7.63	
Complejo Asistencial de Salamanca	109	11.53	120	8.56	
Complejo Asistencial de Soria	58	6.14	62	4.42	
Complejo Asistencial de Segovia	57	6.03	89	6.35	
Complejo Asistencial de Zamora	102	10.79	128	9.13	
Hospital Clínico Universitario de Valladolid	75	7.94	134	9.56	
Hospital El Bierzo	96	10.16	148	10.56	
Hospital Universitario Rio Hortega	54	5.71	78	5.56	

Variables	**Reingreso (N=945 pacientes)**		**No-Reingreso (N=1 402 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Año de ingreso					0.6444
2015-2016	185	19.58	342	24.39	
2017-2018	385	40.74	565	40.30	
2019-2020	375	39.68	495	35.31	
Tipo_alta					< 0.0001
Domicilio	781	82.65	1186	84.59	
Traslado a otro hospital	64	6.77	128	9.13	
Traslado a centros de media y larga estancia	11	1.16	12	0.86	
Alta voluntaria	7	0.74	8	0.57	
Otros tipos de alta	82	8.68	68	4.85	
Tipo_Ingreso					0.1344
Ingreso urgente	828	87.62	1 278	91.16	
Ingreso programado	117	12.38	124	8.84	
Tipos Esquizofrenia					0.0361
Esquizofrenia simple	20	2.11	17	1.21	
Esquizofrenia de tipo desorganizado	17	1.80	28	1.99	
Esquizofrenia catatónica	3	0.32	9	0.64	
Esquizofrenia paranoide	393	41.59	563	40.16	
Esquizofrenia residual	58	6.14	112	8.00	
Otra esquizofrenia especificada	323	34.18	408	29.10	
Esquizofrenia no especificada	131	13.86	265	18.90	
Diagnósticos secundarios codificación CIE-9 (D2-D16):					
Trastornos mentales	512	54.18	554	37.31	< 0.0001
Abuso de sustancias	458	48.47	521	35.08	< 0.0001

Variables	**Reingreso (N=945 pacientes)**		**No-Reingreso (N=1 402 pacientes)**		***p-value***
	n	**%**	**n**	**%**	
Enfermedades infecciosas	64	6.77	50	3.37	0.3245
Neoplasias	21	2.22	30	2.02	0.5480
Enfermedades endocrinas	356	37.67	394	26.53	< 0.0001
Enfermedades de la sangre	50	5.29	43	2.90	0.4841
Enfermedades del sistema circulatorio	145	15.34	187	12.59	0.0132
Enfermedades del aparato respiratorio	75	7.94	99	6.67	0.0899
Enfermedades del aparato digestivo	82	8.68	78	5.25	0.1566
Enfermedades del aparato genitourinario	77	8.15	82	5.52	0.3211
Enfermedades de la piel y del tejido subcutáneo	50	5.29	30	2.02	0.2917
Enfermedades del sistema osteomioarticular	73	7.72	54	3.64	0.3990
Síntomas, signos y estados mal definidos	212	22.43	183	12.32	0.0187
Lesiones y envenenamientos	163	17.25	95	6.40	0.0398
Códigos de diagnóstico tipo V	723	76.51	733	49.36	< 0.0001
Procedimientos codificación CIE-9:					
Proc1	854	90.37	1 070	72.05	< 0.0001
Proc2	401	42.43	366	24.65	0.0158
Proc3	174	18.41	144	9.70	0.1231
Duración de estancia (días)					< 0.0001
Media (SD)	31 (77)		44 (107)		

La estancia media de los pacientes que reingresan es de 31 días (SD = 77), con una tasa de reingreso (BD2) del 70.52%. Las variables edad (p = 0.0035), duración de la estancia ($p < 0.0001$) y hospital ($p < 0.0001$) son significativas para el conjunto de datos, mientras que la variable género (p = 0.2249) no muestra diferencias significativas. Las comorbilidades significativas de los pacientes con reingreso de la BD2 son los diagnósticos con códigos V (76.51%, $p < 0.0001$), el abuso de sustancias (48.47%, $p < 0.0001$) y los trastornos mentales distintos del diagnóstico principal (54.18%, $p < 0.0001$).

5.2 Comparación de algoritmos de Machine Learning

Para dar cumplimiento a uno de los objetivos específicos de esta investigación, en una primera fase se ha comparado el rendimiento de varios algoritmos de ML enfocados en la hospitalización de pacientes con esquizofrenia, para identificar el algoritmo que mejor clasifica a este tipo de pacientes con los datos que soportan esta Tesis Doctoral.

Los algoritmos de clasificación utilizados en la predicción de pacientes hospitalizados con esquizofrenia son: RF, AB, NB, k-NN, DT y SVM. La evaluación de los modelos predictivos se realizó mediante una validación cruzada estratificada k=10. Los resultados de las métricas de rendimiento de cada uno de los modelos han sido extraídos del estudio (Góngora Alonso et al., 2022), que se ha presentado como aval de calidad de esta Tesis Doctoral (Ver Tabla 5.3). Estos resultados se presentan en la Tabla A.1.2 y A.1.3 del Apéndice A, identificando los pacientes hospitalizados con esquizofrenia teniendo en cuenta las distintas clases de la variable dependiente.

En consecuencia, RF presenta los mejores valores en las diferentes métricas evaluadas, con un valor de Acc = 72.7%. Los algoritmos AB y DT muestran un 70.8% y un 68.2% de Acc, respectivamente. Estos valores son mejores en comparación con los algoritmos NB con una Acc = 67.0%, k-NN con Acc = 67.7% y SVM con el valor más bajo de Acc = 65.7%. Los algoritmos NB y k-NN presentan

mejores resultados en términos de AUC en comparación con DT, sin embargo, DT mejora en términos de Acc, precision, F1 y recall.

Tabla 5.3 Métricas de rendimiento aplicando validación cruzada estratificada *k=10*. Fuente: Tabla extraída de (Góngora Alonso et al., 2022).

Algoritmos	**AUC**	**Acc**	**Precision**	**F1-Score**	**Recall**
k-NN	0.729	0.677	0.676	0.676	0.676
Decision Tree	0.682	0.682	0.682	0.681	0.681
AdaBoost	0.765	0.708	0.708	0.708	0.708
SVM	0.641	0.657	0.657	0.657	0.657
Naïve Bayes	0.729	0.670	0.671	0.669	0.670
Random Forest	0.796	0.727	0.728	0.727	0.727

Las curvas ROC mostradas en las Figuras 5.1 y 5.2 muestran gráficos de sensibilidad frente a 1-especificidad a través de diferentes puntos de corte. Estas curvas se evaluaron para *target* = 0 y *target* = 1 con FP = 500, FN = 500 y probabilidad de *target* = 50.0%. De este modo, es posible evaluar visualmente el rendimiento general de cada algoritmo de clasificación. El AUC bajo la curva ROC proporciona un rendimiento medio normalizado del clasificador, considerando toda la gama de umbrales de decisión de salida en el plano de especificidad-sensibilidad.

Cada gráfico muestra las curvas ROC creadas por los valores de FP y FN con validación cruzada *k*=10. El algoritmo RF muestra el mejor valor de AUC = 0.796 (Ver Tabla A.1.2 y Tabla A.1.3 del Apéndice A) para la clase 0 (registros de no-esquizofrenia) y la clase 1 (registros de esquizofrenia).

Teniendo en cuenta las curvas ROC presentadas en las Figuras 5.1 y 5.2 se obtuvo un valor AUC de 0.796, 0.765, 0.682, 0.729, 0.729 y 0.641 para RF, AB, DT, k-NN, NB y SVM, respectivamente. Estos valores recomiendan el RF como el mejor

rendimiento promedio normalizado del clasificador, con un valor AUC mucho más alto que el resto de los algoritmos evaluados. Por tanto, al comparar las curvas ROC mostradas en las Figuras 5.1 y 5.2, los valores de sensibilidad y especificidad están equilibrados, y los algoritmos discriminan entre los pacientes hospitalizados con esquizofrenia y sin el trastorno con aproximadamente la misma probabilidad.

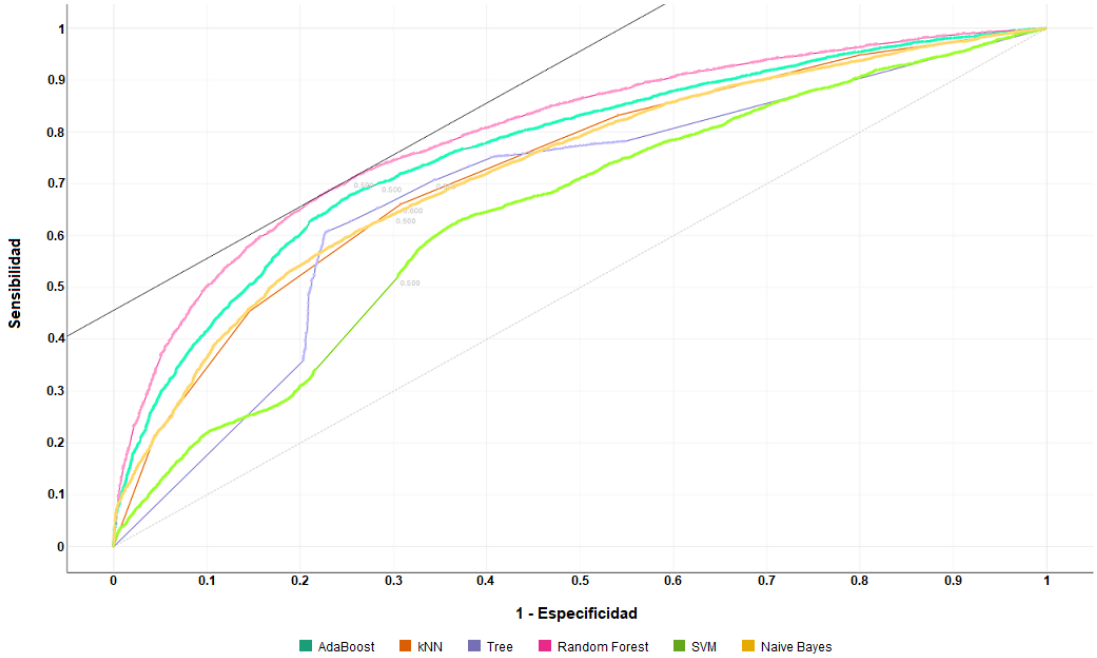

Figura 5.1 Curva ROC para *target* = 0 con FP = 500, FN = 500 y probabilidad de *target* = 50.0%. Fuente: Figura extraída de (Góngora Alonso et al., 2022).

5.3 Factores asociados al riesgo de reingreso en esta población

Sobre la base de los estudios encontrados en la literatura (Rojas et al., 2018); (Y. Zhao et al., 2020), la opinión del psiquiatra experto que colabora en esta investigación, y el método de conjunto RF, se seleccionaron 22 variables para el desarrollo de los modelos predictivos. Teniendo en cuenta que el conjunto de datos es limitado, se estableció un umbral de importancia de 0.5. Por tanto, se eliminaron las variables: tipo_reingreso y neoplasias con valores de importancia por debajo del

umbral. Estas variables mantienen un alto porcentaje de valores constantes, por lo que añaden al modelo más ruido que información. Además, se eliminó la variable año y el diagnóstico principal (esquizofrenia) teniendo en consideración el criterio del experto. Estas variables no aportan información en la predicción del reingreso de los pacientes con esquizofrenia.

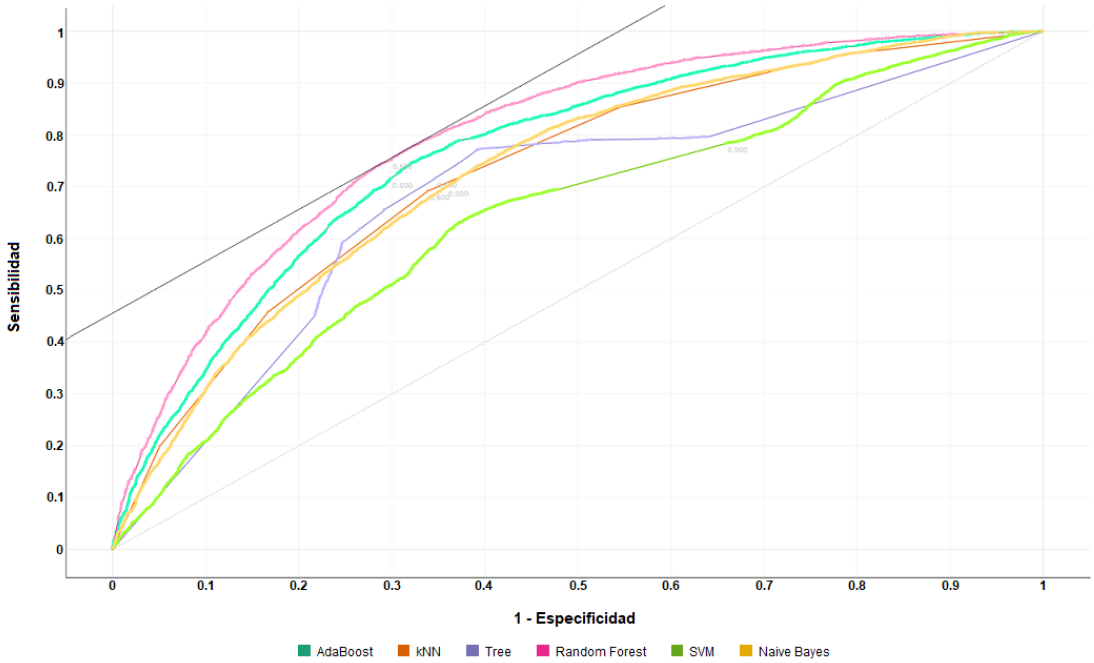

Figura 5.2 Curva ROC para *target* = 1 con FP = 500, FN = 500 y probabilidad de *target* = 50,0%. Fuente: Figura extraída de (Góngora Alonso et al., 2022).

El algoritmo RF, permite obtener la importancia de cada variable a partir de la disminución media del IG. La métrica IG se considera una medida de la pureza de los nodos, cuanto mayor sea la pureza de los nodos menor será el valor del IG. La Figura 5.3 muestra las características del conjunto de datos ordenadas de mayor a menor predicción en función del valor de IG.

RF identificó los diagnósticos clasificados como trastornos mentales ($p < 0.0001$), el hospital ($p < 0.0001$), los diagnósticos con códigos V02.52-V88.01 ($p <$

0.0001), la edad ($p < 0.0001$), el diagnóstico de abuso de sustancias ($p < 0.0001$) y la duración de la estancia ($p < 0.0001$) como las variables más predictivas.

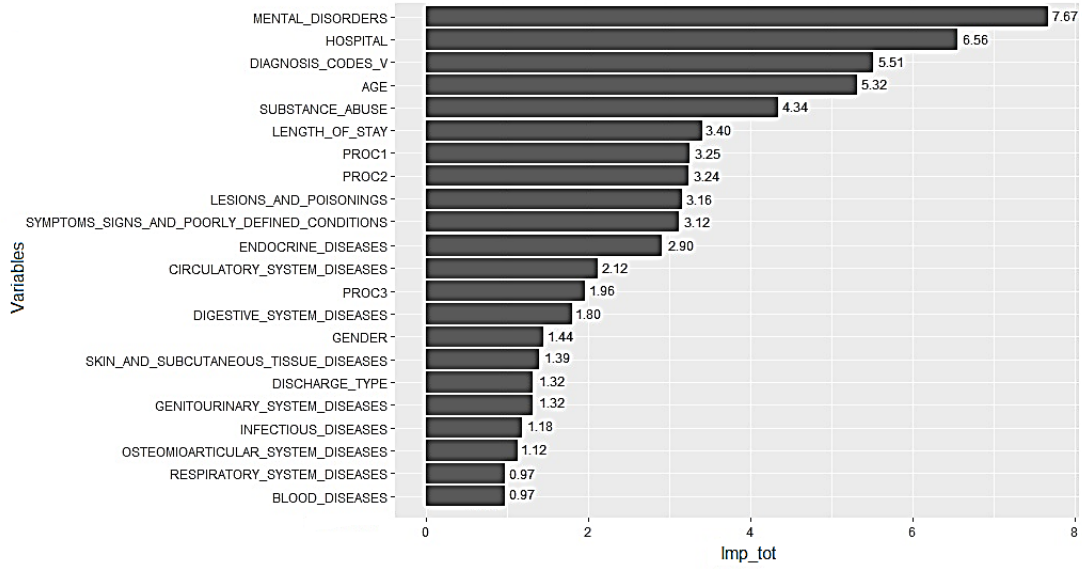

Figura 5.3 Importancia de las variables predictoras con el algoritmo RF (BD1). Fuente: Figura extraída de (Góngora Alonso et al., 2023).

En el aspecto clínico, la variable trastornos mentales presenta las patologías secundarias más prevalentes en la muestra de pacientes reingresados. Estas patologías son: la psicosis (3.37%, $p < 0.0001$), el trastorno delirante (2.74%, $p = 0.0004$) y el trastorno de personalidad (2.61%, $p < 0.0001$). En la variable hospital, el complejo asistencial de Burgos ($p < 0.0001$) presentó la categoría más significativa respecto a los datos. Los factores sociales que no constituyen una enfermedad o lesión del paciente influyen en el reingreso de los pacientes con esquizofrenia en CyL. Estos factores están incluidos en la clasificación de la variable códigos de diagnóstico V, los más representativos son: historial de incumplimiento del tratamiento médico (21.05%, $p < 0.0001$), historial familiar de enfermedad psiquiátrica (15.62%, $p < 0.0001$), y persona que vive sola (5.31%, $p < 0.0001$). La edad representativa en esta población es de 31-50 años y la tasa de reingreso en los

primeros 30 días del 66.32%. La variable abuso de sustancias es un factor muy representativo en el reingreso de un paciente con esquizofrenia. Los trastornos por abuso de tabaco (16.05%, $p < 0.0001$), el abuso de alcohol (10.29%, $p < 0.0001$) y el abuso continuo de cannabis (10.42%, $p < 0.0001$) son los diagnósticos más significativos en el conjunto de datos. Estas son las categorías de las variables con mayor prevalencia en la BD1 y con un alto valor significativo.

Para comparar los resultados obtenidos con la BD1 referente a los factores asociados al reingreso de estos pacientes, aplicamos la metodología propuesta en la sección 4.1 y 4.2 en la BD2. Tal y como se muestra en la Figura 5.4, se obtuvo la importancia de las variables predictoras de este nuevo conjunto de datos.

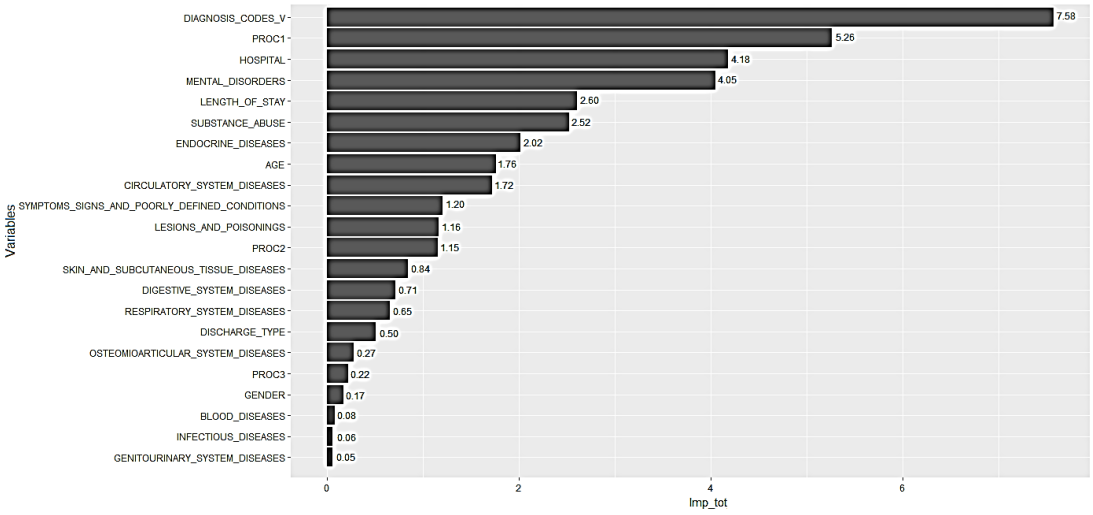

Figura 5.4 Importancia de las variables predictoras con el algoritmo RF (BD2). Fuente: Figura extraída de (Góngora Alonso et al., 2023).

Los resultados de importancia de las variables que se muestran en la Figura 5.4 tienen un comportamiento similar a los resultados que se muestran en la Figura 5.3. En este sentido, se han obtenido los códigos de diagnóstico V ($p < 0.0001$), la variable procedimiento 1 ($p < 0.0001$), el hospital ($p < 0.0001$), los trastornos mentales ($p < 0.0001$), duración de la estancia ($p < 0.0001$) y el abuso de sustancias ($p < 0.0001$), como las variables con mejores valores de importancia. Los resultados obtenidos con

la BD1 y BD2, demuestran que las variables predictivas identificadas en esta investigación están asociadas a los factores de riesgo del reingreso de pacientes con esquizofrenia en CyL.

5.4 Modelos predictivos del riesgo de reingreso

Centrando la atención en el objetivo principal de esta investigación y a partir de los resultados obtenidos en el estudio (Góngora Alonso et al., 2022), se desarrollan una serie de modelos para predecir el riesgo de reingreso de los pacientes hospitalizados con esquizofrenia en CyL. Teniendo en cuenta que los datos de la investigación son de carácter administrativo y no están basados en datos clínicos del paciente, los algoritmos basados en DT mejoran las métricas de rendimiento de los modelos desarrollados en este estudio (Góngora Alonso et al., 2022). Por tanto, sobre la base de los resultados obtenidos en la sección 5.2 y la revisión de la literatura planteada en la sección 2.2, utilizamos los algoritmos de clasificación RF, AB, SVM con *Radial Basis Function kernel* (SVMradial), MLP y LR.

Los resultados de evaluación de los diferentes modelos han sido extraídos del estudio (Góngora Alonso et al., 2023), que se ha presentado como aval de calidad de esta Tesis Doctoral y se muestran en la Tabla 5.4. La métrica más evaluada en los estudios de predicción de reingresos es el AUC (Artetxe et al., 2018; Huang et al., 2021). En esta investigación han sido ajustados los hiperparámetros de cada modelo en función de la métrica ROC, utilizando un remuestreo de validación cruzada k=10. Esta métrica indica el porcentaje de predicción del modelo para distinguir entre un paciente que reingresa y otro que no lo hace.

RF presenta los mejores valores en las diferentes métricas, Acc = 0.817, recall = 0.887, F1-score = 0.877 y AUC = 0.879. En cuanto a los clasificadores XGBoost y AB, presentan altos valores de precisión, recall, F1-score, y AUC en comparación con MLP; y muestra mejores valores con respecto al resto de clasificadores. LR y SVMradial tienen un comportamiento similar en sus valores. Teniendo en cuenta las

métricas de la Tabla 5.4, se puede afirmar que RF es el algoritmo óptimo para predecir el riesgo de reingreso de los pacientes con esquizofrenia en esta investigación, seguido de los clasificadores XGBoost y AB.

Tabla 5.4 Resultados de las métricas de rendimiento aplicando la validación cruzada *k*=10. Fuente: Tabla extraída de (Góngora Alonso et al., 2023).

Algoritmos	**Acc**	**Recall**	**F1-Score**	**AUC**
AdaBoost	0.798	0.881	0.873	0.789
Logistic Regression	0.782	0.874	0.858	0.756
SVMradial	0.770	0.861	0.840	0.642
XGBoost	0.804	0.869	0.859	0.804
MLP	0.791	0.876	0.863	0.779
Random Forest	0.817	0.887	0.877	0.879

La Figura 5.5 muestra las curvas ROC de los diferentes algoritmos de clasificación. Estas curvas se representan mediante un gráfico de sensibilidad frente a especificidad con diferentes puntos de corte. El AUC proporciona un rendimiento medio normalizado del clasificador, teniendo en cuenta toda la gama de umbrales de decisión de salida en el plano de especificidad-sensibilidad.

Los valores AUC obtenidos son 0.756, 0.804, 0.789, 0.779, 0.642, 0.879 para LR, XGBoost, AB, MLP, SVMradial y RF, respectivamente. La Figura 5.5 muestra los algoritmos RF y XGBoost con el mejor rendimiento medio normalizado del clasificador, AUC = 0.879 y AUC = 0.804 respectivamente. Para este valor de AUC, el punto de corte óptimo en el que se equilibran los valores de sensibilidad y especificidad es 0,814. Por tanto, este valor es donde el algoritmo discrimina mejor entre los pacientes con esquizofrenia con riesgo de reingreso y sin riesgo de reingreso.

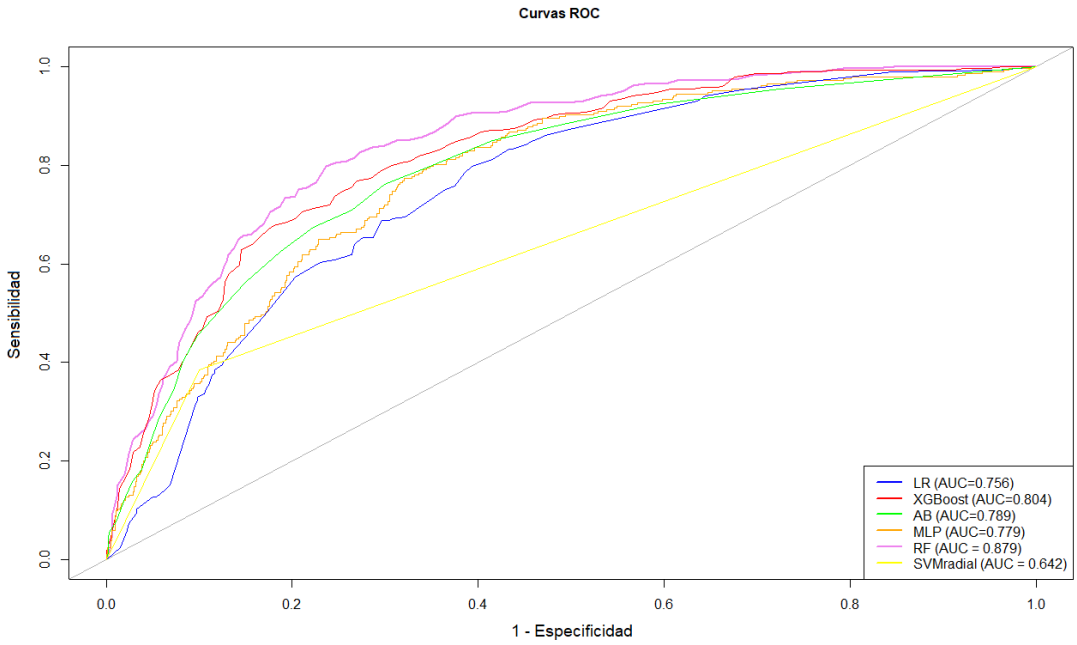

Figura 5.5 Curvas ROC de los algoritmos de ML utilizados en el estudio. Fuente: Figura extraída de (Góngora Alonso et al., 2023).

5.5 Aplicación del modelo Random Forest en la práctica clínica

El objetivo final de los modelos predictivos desarrollados en esta Tesis Doctoral es que ayuden a predecir el riesgo de reingreso de los pacientes con esquizofrenia, y en consecuencia sean aplicables en la práctica clínica. Por tanto, en la fase final de esta investigación, se desarrolló una aplicación web que permite calcular la probabilidad de reingreso de un paciente hospitalizado con esquizofrenia cuando se le da el alta. Para el desarrollo de dicha aplicación se construyó un modelo final a partir del conjunto de datos de la BD1 y la BD2 con el algoritmo RF.

5.5.1 Modelo con Random Forest

El algoritmo RF es el clasificador que presenta los mejores resultados en sus métricas de rendimiento tal y como se muestra en la Tabla 5.3 y 5.4. En la fase final

de esta Tesis Doctoral, se obtuvo la BD2 que se ha fusionado con la BD1 y se ha obtenido un conjunto de datos más amplio con un total de 11 126 registros (5 412 pacientes). Con una muestra mayor de observaciones, se ha desarrollado y validado un modelo final para predecir el riesgo de reingreso de los pacientes hospitalizados con esquizofrenia en CyL, utilizando el algoritmo RF. El enfoque metodológico que se ha seguido es el planteado en el Capítulo 4 de esta investigación. Con un mayor valor de la muestra el comportamiento de las variables es diferente, por tanto, se eliminaron 7 variables respecto a las 22 seleccionadas. Estas variables se eliminaron para evitar el sobreajuste del modelo. Los hiperparámetros han sido ajustados en función de la métrica ROC, utilizando un remuestreo de validación cruzada k=10. Los resultados de rendimiento del modelo se muestran en la Figura 5.6.

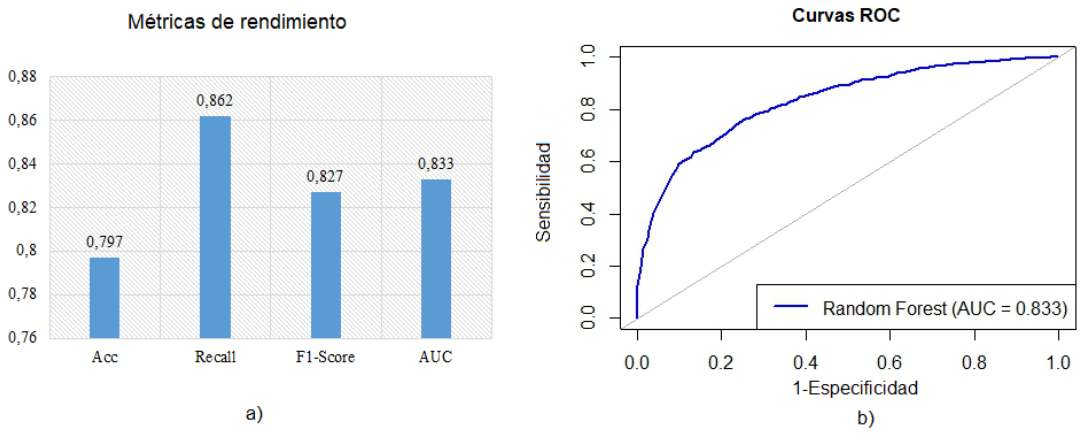

Figura 5.6 Modelo final de predicción de reingresos de pacientes hospitalizados con esquizofrenia en CyL: a) Métricas de rendimiento, b) Curvas ROC.

Las métricas de rendimiento obtenidas muestran valores por debajo del modelo de RF creado con la BD1 (Ver Tabla 5.4). Los valores de AUC varían de 0.879 a 0.833 y Acc de 0.817 a 0.797. Teniendo en cuenta las características de la población, estas diferencias de valores están dadas por el número de variables y la muestra utilizada.

5.5.2 Aplicación con Shiny

El modelo final entrenado se ha utilizado para desarrollar la aplicación web que calcula el riesgo de reingreso de un paciente con esquizofrenia. Se ha desarrollado en el entorno de programación R, con la librería de interfaz gráfica: Shiny.

Esta aplicación es un diseño inicial que puede ser manejado de forma sencilla por un profesional sanitario. Su interfaz es accesible y se muestra en la Figura 5.7.

Para usar la aplicación solo es necesario seleccionar los valores de cada variable y pulsar el botón "Calcular el riesgo de reingreso". Este botón devuelve una tabla con todos los valores seleccionados en cada una de las variables y la probabilidad de reingreso del paciente. De acuerdo con la opinión del psiquiatra experto que colabora en esta investigación, la interpretación del resultado final de este modelo está sujeta al criterio del profesional sanitario.

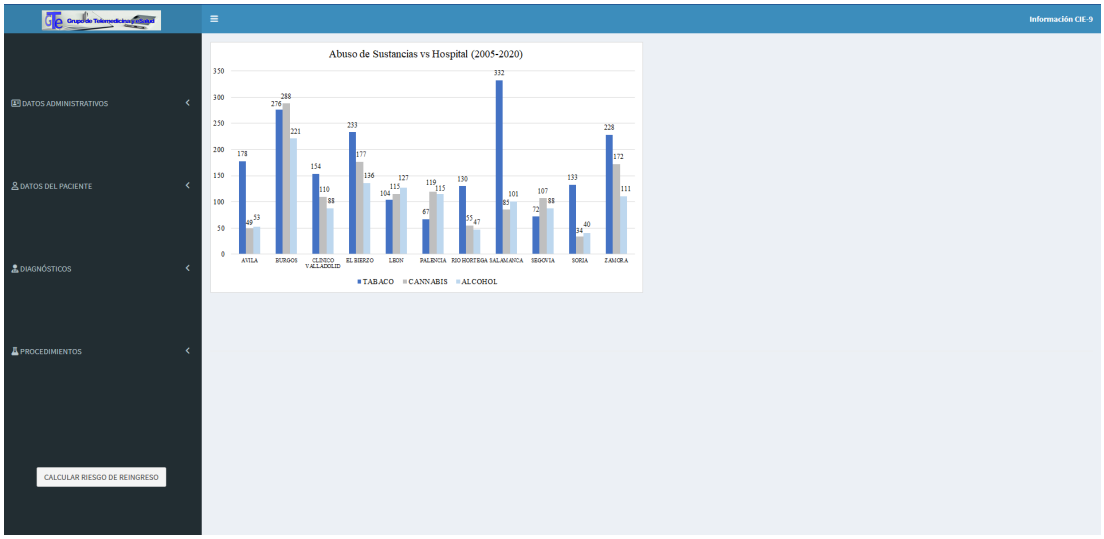

Figura 5.7 Interfaz gráfica de la aplicación web.

En la interfaz, se muestra también un gráfico general del comportamiento de la variable abuso de sustancias en cada hospital (Ver Figuras 5.8 y 5.9). Se identificaron las categorías de mayor prevalencia en estos pacientes y se compararon

entre sí. El abuso de tabaco ($p < 0.0001$), de cannabis ($p < 0.0001$) y de alcohol ($p < 0.0001$) son las más prevalentes en esta población.

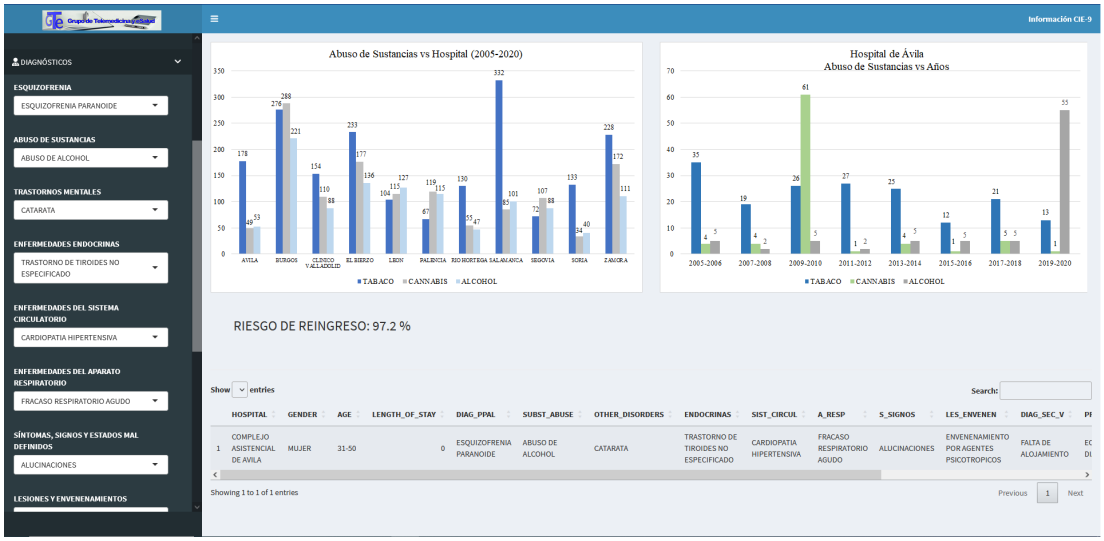

Figura 5.8 Interfaz de resultados, seleccionando el Complejo Asistencial de Ávila

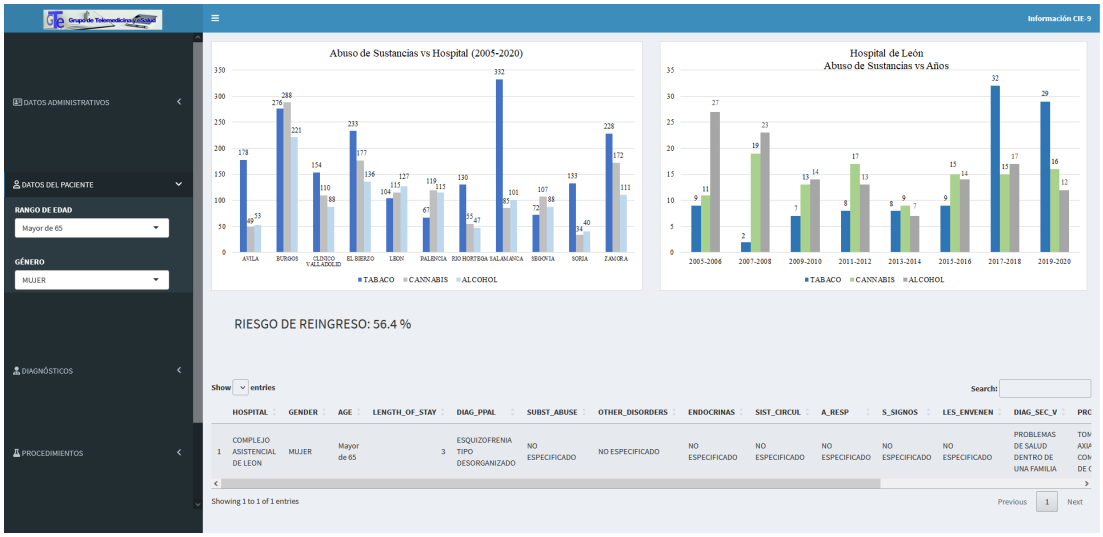

Figura 5.9 Interfaz de resultados, seleccionando el Complejo Asistencial de León.

Cuando se calcula el riesgo de reingreso del paciente, se muestra en la interfaz una tabla con todas las variables seleccionadas por el médico, el valor del riesgo de

reingreso y un gráfico que muestra el comportamiento de la variable abuso de sustancia por años en el hospital seleccionado (Ver Figuras 5.8 y 5.9).

Esta aplicación está disponible en línea y puede utilizarse ejecutando el siguiente enlace: Aplicación Web. Teniendo en cuenta que es un prototipo inicial, que se desarrolló para mostrar la aplicación práctica de los resultados de esta Tesis Doctoral, se pueden hacer mejoras e implementarla en el entorno hospitalario.

Capítulo 6

Discusión

La investigación presentada en esta Tesis Doctoral está enfocada en la predicción del riesgo de reingreso de pacientes hospitalizados con esquizofrenia. Para ello, se han aplicado diferentes técnicas de ML. En este sentido, se discuten en este capítulo los principales factores identificados en la investigación, asociados al riesgo de hospitalización de estos pacientes. Posteriormente, se discuten los resultados obtenidos en cuanto a las métricas de rendimiento de los modelos y se realiza una comparación con otros estudios del estado del arte. Finalmente, se exponen las principales limitaciones de la investigación.

6.1 Factores asociados al riesgo de hospitalización

En esta investigación, se identifican los principales factores asociados al riesgo de reingreso de pacientes con esquizofrenia en CyL. Se aplicó el algoritmo RF para identificar las variables más predictivas del estudio (Y. Zhao et al., 2020). La variable trastornos mentales presenta el mayor valor de significancia en el conjunto de datos. En cuanto al aspecto clínico, la psicosis (3.37%, $p < 0.0001$), el trastorno delirante (2.74%, $p = 0.0004$) y el trastorno de la personalidad (2.61%, $p < 0.0001$) son los diagnósticos secundarios que más afectan a los pacientes que reingresan. El grupo de edad de 31-50 años, compuesto mayoritariamente por el género masculino, presenta el mayor número de reingresos, por lo que se considera un factor de riesgo para tener en cuenta en esta población.

En cuanto a la variable código de diagnósticos V, los diagnósticos categorizados como: historial de incumplimiento del tratamiento médico (21.05%, $p < 0.0001$), historia familiar de enfermedad psiquiátrica (15.62%, $p < 0.0001$), y persona que vive sola (5.31%, $p < 0.0001$), influyen en el reingreso de los pacientes con esquizofrenia. Los estudios (J. Edgcomb et al., 2019; Y. Zhao et al., 2020) identifican la duración de la estancia como un factor de riesgo asociado al reingreso, esta investigación presenta la mayor tasa de reingreso en los primeros 30 días con un 66.32%.

El abuso de sustancias es otro factor de riesgo descrito en el estudio (Morel et al., 2020), se ha demostrado la asociación que existe entre pacientes que sufren trastornos mentales como la esquizofrenia y el abuso de sustancias. El comportamiento suicida es otro factor clínico que se asocia a los pacientes que sufren esquizofrenia, depresión/ansiedad y trastornos por consumo de sustancias (Costanza et al., 2021). Estos factores de riesgo clínicos y demográficos son causas frecuentes de ingreso en el servicio de urgencias, aumentando la tasa de reingreso en hospitales (Costanza et al., 2020). En este conjunto de datos, la variable abuso de sustancias es prevalente con trastornos como: abuso de tabaco (16.05%, $p < 0.0001$), abuso de alcohol (10.29%, $p < 0.0001$) y abuso de cannabis continuo (10.42%, $p < 0.0001$).

La BD2, obtenida en la fase final de esta Tesis Doctoral, ha permitido comparar las poblaciones incluidas en cada una. Los resultados que se muestran en la Figura 5.4 demuestran que el conjunto de variables predictoras obtenido en esta investigación, indica posibles factores de riesgo asociados al reingreso de pacientes con esquizofrenia en CyL.

6.2 Modelos predictivos del riesgo de reingreso

Las personas con trastorno de esquizofrenia presentan una alta tendencia a ser hospitalizadas con frecuencia, lo que impone un coste económico al sistema sanitario. Desde el punto de vista de esta investigación, la implementación de

modelos predictivos en los sistemas médicos puede ser una herramienta útil en la prevención de las hospitalizaciones de pacientes con esquizofrenia en CyL.

Los algoritmos de clasificación que se comparan en la sección 5.2 de esta Tesis Doctoral, tienen una Acc predictiva del 65.7-72.7% (Tabla 5.3). Estos resultados permiten identificar el algoritmo que mejor se ajusta a los datos de la investigación. Es necesario destacar que, la capacidad de un algoritmo para ajustar en mayor o menor medida va a depender de las características del conjunto de datos, en este sentido se recomienda el RF con un Acc = 72.7%.

Aunque existe un gran número de estudios que utilizan un enfoque de ML para la predicción del reingreso de pacientes en salud mental (Morel et al., 2020) y en unidades de agudos en general (Deschepper et al., 2019; Tong et al., 2016), se ha encontrado en la literatura un número limitado de estudios enfocados al reingreso en esquizofrenia. En (Thongkam & Sukmak, 2014), los autores comparan algoritmos basados en árboles de decisión y en (Huberts et al., 2022), identifican el riesgo de una crisis de salud mental en personas diagnosticadas con esquizofrenia. En el estudio (Ying et al., 2021), los autores exploran la asociación entre la administración de la terapia electroconvulsiva y las tasas de reingreso de este tipo de paciente, mientras que en (Fond et al., 2019) determinan los predictores de recaída psicótica en una muestra de sujetos con esquizofrenia sin hospitalizar. Por tanto, hasta donde sabemos, es la primera investigación que desarrolla modelos para predecir el riesgo de reingreso de los pacientes con esquizofrenia utilizando LR, SVMradial, MLP y clasificadores basados en DT.

Los resultados de esta Tesis Doctoral muestran que, los algoritmos que mejor predicen el riesgo de reingreso de los pacientes con esquizofrenia son RF, XGBoost y AB, con un AUC de 0.879, 0.804 y 0.789 respectivamente. RF muestra en las métricas evaluadas los mejores resultados, mientras que modelos como LR y SVMradial presentan valores inferiores al clasificador MLP. En consecuencia, se

puede recomendar el uso de algoritmos basados en DT, para desarrollar modelos predictivos con un conjunto de datos similares.

La aplicación web desarrollada en esta investigación puede aplicarse en diferentes tipos de hospitales (privados y/o públicos), para predecir el riesgo de reingreso de los pacientes hospitalizados con esquizofrenia. En consecuencia, puede ayudar en la gestión hospitalaria de CyL, con la prevención de hospitalizaciones y la reducción de costes asociados.

Una posible estrategia preventiva, de acuerdo con el paciente que reingresa, puede ser el modelo de gestión hospitalaria implementado en el Complejo Asistencial de Zamora en 2012. El objetivo de este modelo de gestión es reducir el número de hospitalizaciones de pacientes con trastornos mentales en CyL. Consiste en no admitir a los pacientes con enfermedades mentales en el hospital, en su lugar, son supervisados en pisos tutelados o centros para pacientes con este tipo de trastorno. En el estudio (Góngora Alonso et al., 2020), con la BD1 se analizó el comportamiento de las hospitalizaciones de pacientes con trastornos mentales en el período de 2005-2015. Los resultados del estudio muestran la reducción del número de hospitalizaciones de estos pacientes en Zamora a partir de 2012. Estos resultados sugieren que, una evaluación más profunda de este modelo de gestión, puede ser una solución para dar seguimiento a los pacientes con esquizofrenia, y evitar las altas tasas de reingreso hospitalario.

6.3 Comparación con estudios similares

En la predicción de reingreso en pacientes con trastornos mentales, estudios similares como (Morel et al., 2020) usan algoritmos como XGBoost (AUROC = 0.738) y GLMNet (AUROC = 0.697). Los valores de AUC de nuestra investigación oscilan entre 0.642-0.879 (Ver Tabla 5.4).

En estudios como (Deschepper et al., 2019) los autores predicen el riesgo de reingreso aplicando LR (AUC = 0.715), LR Penalizada (AUC = 0.736), GBM (AUC

= 0.731), y RF (AUC = 0.774); mientras que en (Tong et al., 2016) usan los algoritmos LACE (AUC = 0.655), STEPWISE logistic (AUC = 0.735), LASSO logistic (AUC = 0.737) y AB (AUC = 0.737). La Tabla 6.1 muestra los estudios similares encontrados en la literatura.

Tabla 6.1 Estudios relacionados con el reingreso hospitalario utilizando algoritmos de ML

Estudios	**Conjunto de datos**	**Método de Validación**	**Algoritmos de ML**	**AUC**
(Tong et al., 2016)	10 9421 pacientes	Validación cruzada estratificada $k \geq 3$	LACE	0.655
			STEPWISE logistic	0.735
			LASSO logistic	0.737
			AB	0.737
(Deschepper et al., 2019)	29 702 pacientes	Validación cruzada $k = 10$	LR	0.715
			LR Penalizada	0.736
			GBM	0.731
			RF	0.774
(J. Edgcomb et al., 2019)	552 pacientes	Validación cruzada $k = 10$	CART	0.880
(Morel et al., 2020)	65 426 pacientes	Validación cruzada $k = 10$	XGBoost	0.738
			GLMNet	0.697
(Ying et al., 2021)	2 131 pacientes	Validación cruzada $k = 5$	XGBoost	0.730
(Huberts et al., 2022)	75 000 personas con esquizofrenia	Validación cruzada	LR	0.613
			Hierarchical regression	0.586
			XGBoost	0.653

Comparando los resultados de esta investigación con los estudios enfocados en esquizofrenia (Huberts et al., 2022; Ying et al., 2021), que se muestran en la Tabla

6.1, obtenemos valores superiores de AUC. El algoritmo XGBoost evaluado en ambos estudios presenta valores de AUC = 0.730 para una muestra de 2 131 pacientes y de AUC = 0.653 para una muestra de 75 000 personas con esquizofrenia, mientras que en esta investigación se obtiene un AUC = 0.804 para una muestra de 3 065 pacientes. En el estudio (J. Edgcomb et al., 2019), identifican predictores modificables de reingreso psiquiátrico entre sujetos con trastorno bipolar y enfermedades no clasificadas como trastornos mentales. En este sentido, utilizan el algoritmo CART y obtienen un valor de AUC = 0.880, superando los resultados de los modelos desarrollados en esta investigación.

Existen otros estudios como (Fond et al., 2019), que determinan los predictores de recaída psicótica en personas con esquizofrenia utilizando el algoritmo CART. Este estudio no se ha incluido en la Tabla 6.1 porque evalúa otras métricas de rendimiento: Acc = 0.638, sensibilidad = 0.710, especificidad = 0.448. Cuando comparamos los resultados de recall y F1-score de esta investigación (Ver Tabla 5.4) con el estudio (Thongkam & Sukmak, 2014), nuestros resultados tienen valores más bajos, excepto el algoritmo AB con F1-score = 0.873 y recall = 0.881.

La Figura 6.1 muestra la comparación del algoritmo RF y AB de ambos estudios, ya que han sido los modelos con mejores valores en sus métricas. El estudio (Thongkam & Sukmak, 2014) obtiene valores de recall del 85.12-96.81% y F1-score = 77.68–95.94% para algoritmos como Random Tree, RF, DT, AB, y Baggin. Además, desarrollan modelos híbridos con el algoritmo AB y Bagging para mejorar los resultados, obteniendo un recall = 94.7-98.11% y F1-score = 94.30-9.41%. Si comparamos el conjunto de datos y las variables predictoras de ambos estudios, nuestra investigación presenta un conjunto de datos mayor (3 065 pacientes) con diferentes variables predictoras. Por tanto, se considera que la calidad de los datos evaluados es un factor fundamental para obtener mejores valores de precisión en los modelos predictivos.

En general, los resultados presentados en esta Tesis Doctoral muestran un enfoque fiable para predecir el riesgo de reingreso de los pacientes con esquizofrenia en esta población, utilizando diferentes técnicas de ML.

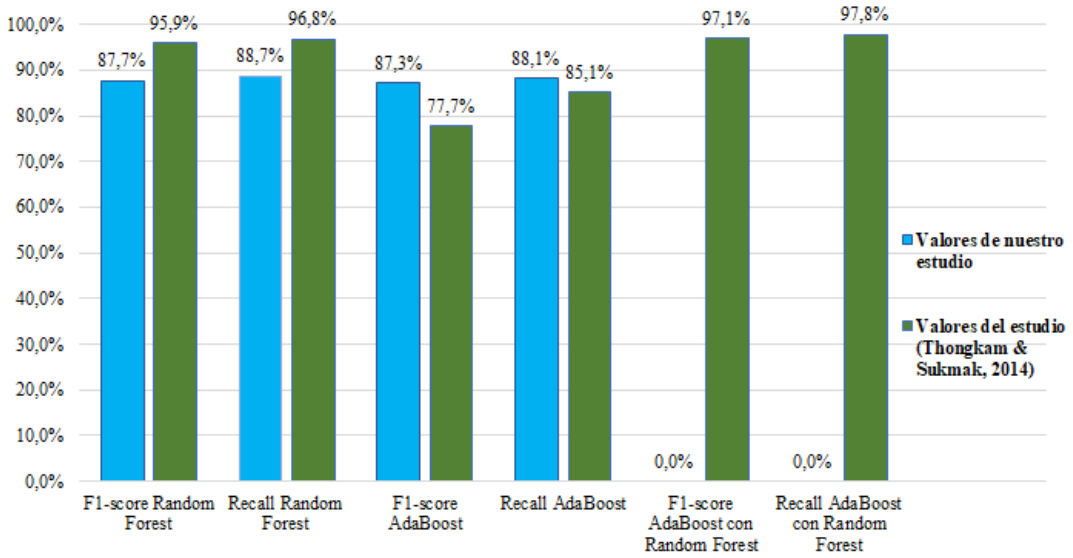

Figura 6.1 Comparación de las métricas de rendimiento F1-score y recall de esta investigación con el estudio (Thongkam & Sukmak, 2014). Fuente: Figura extraída de (Góngora Alonso et al., 2023)

6.4 Limitaciones de la investigación

La presente Tesis Doctoral, ha demostrado la utilidad de los algoritmos de ML aplicados a la predicción del reingreso de pacientes con esquizofrenia en CyL. En este sentido, es necesario señalar algunas limitaciones de la investigación que podrían haber condicionado los resultados alcanzados.

Una limitación relevante es que se centra en una comunidad específica de España, por lo que las variables predictivas asociadas al riesgo de reingreso de esta población no son generalizables a otros conjuntos de datos, aunque coinciden con los factores de riesgo existentes en la literatura para este tipo de pacientes. Es importante

tener en cuenta el tipo de población y los datos que se analizan en este estudio. Según los datos registrados en el INE, CyL tiene una población del 62.29% en el grupo de edad de 15-64 años y un 26.31% de personas mayores de 64 años. La mayor cantidad de personas de esta región se encuentra en el grupo (31-50 años) más representativo de las personas con esquizofrenia que reingresan. Sin embargo, los algoritmos utilizados para el desarrollo de los modelos pueden ser aplicados por otros investigadores en conjuntos de datos similares.

Otra limitación relacionada con los datos de esta investigación es que sólo se ha trabajado con datos administrativos, que corresponden a los registros de hospitalización de los pacientes. Estos registros son datos globales, no están basados en la psicopatología clínica del paciente, incluyen información demográfica, características de episodios de hospitalización, diagnósticos y procedimientos referentes al paciente hospitalizado. Por tanto, sería útil cruzar este conjunto de datos con registros clínicos, para incluir nuevas variables que ayuden a mejorar el modelo de predicción. Además, es necesario mencionar que en el estudio sólo se han evaluado pacientes que están hospitalizados, no fue posible incluir personas con esquizofrenia que no hayan sido hospitalizadas.

Los datos utilizados fueron introducidos manualmente en el sistema por los profesionales sanitarios, lo que conlleva la existencia de campos vacíos en las variables de comorbilidad. Estos datos ausentes pueden influir en el valor de significancia de las variables predictoras y en la evaluación de las métricas de rendimiento.

Los principales resultados alcanzados en esta investigación se han obtenido con los datos analizados en el período comprendido entre 2005 y 2015, con sólo 22 variables predictoras para desarrollar los modelos. El uso de 22 variables predictoras tiene la ventaja de utilizar menos requisitos computacionales en el desarrollo del modelo, sin embargo, se recomienda añadir otras variables al conjunto de datos y evaluar el comportamiento de los reingresos de estos pacientes con esquizofrenia en

los años posteriores. La BD2 obtenida en la fase final de esta Tesis Doctoral, sólo ha sido utilizada para analizar las características de los nuevos pacientes y conformar el conjunto de datos que se utilizó para entrenar el modelo final.

Capítulo 7

Conclusiones

Esta Tesis Doctoral se centra en el reingreso de pacientes con esquizofrenia en los hospitales públicos de CyL. Para abordar este problema, se desarrollan modelos predictivos utilizando un enfoque de ML, y se identifican los factores asociados al riesgo de reingreso de este tipo de paciente. En este capítulo, se exponen las principales contribuciones alcanzadas en la sección 7.1. En la sección 7.2 se plantean las conclusiones de la investigación y se sugieren posibles líneas de investigación futuras en la sección 7.3.

7.1 Contribuciones

A continuación, se destacan las principales aportaciones de esta Tesis Doctoral en este campo de investigación.

- Características de los pacientes hospitalizados con esquizofrenia que reingresan en los hospitales públicos de CyL. Se analizaron las características del conjunto de datos para obtener un perfil del paciente que tiende a reingresar en los hospitales de esta región.
- Identificación del algoritmo de ML que mejor se ajusta a los datos de hospitalización de pacientes con esquizofrenia en CyL. Se compararon diferentes algoritmos de ML obteniendo mejores valores de rendimiento, en comparación con los estudios similares encontrados. Lo que demuestra que el enfoque de ML es adecuado para desarrollar modelos predictivos en esta población.

- Identificación de factores de riesgo asociados al reingreso de pacientes con esquizofrenia. Se aplicó el método de envoltura para determinar la importancia de las variables. Se demostró que, las variables más predictivas del conjunto de datos están asociadas a los factores de riesgo del reingreso descritos en la literatura para este tipo de paciente.
- Modelo de aplicación clínica para predecir el riesgo de reingreso de pacientes con esquizofrenia en esta población. Este modelo fue validado con una muestra de 5 412 pacientes hospitalizados de dos conjuntos de datos diferentes, mostrando una alta capacidad de predicción del riesgo de reingreso en comparación con estudios similares.
- Prototipo de aplicación web que calcula el riesgo de reingreso de los pacientes con esquizofrenia al alta hospitalaria. Hasta donde sabemos, es la primera aplicación web que se desarrolla en CyL para calcular el riesgo de reingreso de esta población.

Estas contribuciones pretenden respaldar la toma de decisiones clínicas, ayudar en el manejo hospitalario y reducir las hospitalizaciones en los hospitales públicos de CyL.

7.2 Principales conclusiones

A partir de los resultados obtenidos a lo largo de esta investigación, se plantean las siguientes conclusiones:

- De acuerdo con el conjunto de datos administrativos y los algoritmos que utilizamos en el estudio, se ha llegado a la conclusión que los algoritmos que mejor se ajustan a este tipo de datos en esta población, son los algoritmos basados en DT. Los mejores valores de rendimiento obtenidos a lo largo de esta investigación han sido obtenidos a partir

de clasificadores como: RF, XGBoost y AB, en comparación con algoritmos como SVM y MLP.

- El enfoque de ML ayuda a la identificación de los factores de riesgo asociados con el reingreso de pacientes con esquizofrenia en esta población. Tras aplicar un método de envoltura para determinar la importancia de las variables, en esta investigación, se identificaron las variables más predictivas del conjunto de datos. Estas variables son: la edad, la duración de la estancia, los diagnósticos clasificados como trastornos mentales, abuso de sustancias, y los diagnósticos tipo V. En este sentido, se puede concluir que estas variables predictoras indican posibles factores de riesgo asociados al reingreso de pacientes con esquizofrenia en CyL.
- Sobre la base de los resultados obtenidos de cada uno de los modelos comparados en esta investigación, se puede concluir que, el modelo que mejor predice el riesgo de reingreso de los pacientes con esquizofrenia en CyL es el RF, con un intervalo de AUC = (0.833-0.879) y Acc = (0.797-0.817). Estos valores de rendimiento del modelo muestran un enfoque fiable para predecir el riesgo de reingreso de esta población, en comparación con los estudios similares existentes.
- La aplicación web desarrollada en esta investigación, permite implementar en la práctica clínica el modelo que mejor predice el riesgo de reingreso de los pacientes con esquizofrenia en CyL. Se considera que esta herramienta puede servir de apoyo al personal sanitario para la toma de decisiones.

De forma general, con los resultados obtenidos con esta Tesis Doctoral, se ha llegado a la conclusión que: predecir el riesgo de reingreso de estos pacientes y los

factores asociados, mediante un enfoque de ML, puede conducir al desarrollo de estrategias de intervención preventiva. Estas estrategias pueden ayudar a mejorar la calidad de la atención sanitaria, la seguridad del paciente, el bienestar y reducir los costes sanitarios en el área de salud mental.

7.3 Líneas futuras

A lo largo de esta Tesis Doctoral han surgido varias cuestiones que pueden ser abordadas en estudios futuros, para complementar los resultados de esta investigación. A continuación, se plantean las principales líneas futuras:

- Cruzar las BD1 y BD2, con datos clínicos de los pacientes con esquizofrenia que hospitalizan en estos centros sanitarios. Esto permite que la identificación de los factores de riesgo de este tipo de paciente no esté basada solo en datos de carácter administrativo, sino que se analicen datos clínicos que puedan influir en esta población.
- Analizar la BD2 que contiene un mayor número de variables que la BD1. Sobre la base de este análisis se pueden plantear nuevos modelos predictivos utilizando la metodología propuesta.
- La BD2 incluye unas variables de desagregación, que no se han podido analizar en esta Tesis Doctoral. Estas variables se encargan de agrupar diferentes grupos de diagnósticos que sean clínicamente homogéneos, y se asocian a los niveles de severidad y riesgo de mortalidad del paciente. En este sentido, es interesante plantear un estudio del comportamiento de los pacientes hospitalizados con esquizofrenia respecto a esas variables de desagregación.
- Recolectar datos de CyL, de pacientes con esquizofrenia que no hospitalizan y poder hacer comparaciones entre los pacientes que hospitalizan y los que no. Este tipo de comparaciones permiten

obtener un mejor perfil del paciente que reingresa en los hospitales públicos de esta región.

- Combinar algoritmos diferentes a los utilizados en esta investigación, para desarrollar nuevos modelos predictivos sobre la base de la metodología propuesta. En este sentido, se puede utilizar un enfoque diferente para la selección de características, con métodos integrados que incluyen métodos de envoltura y filtrado.
- El abuso de sustancias es uno de los factores de riesgo más influyentes en la hospitalización de los pacientes con esquizofrenia. En consecuencia, se puede analizar el comportamiento de esta variable en los pacientes que reingresan, y plantear modelos predictivos, utilizando las bases de datos de esta investigación.
- Otra línea futura es la validación de la metodología propuesta en bases de datos más amplias de otras regiones del país. Esto podría aumentar la fiabilidad y la generalización de los resultados.
- Las bases de datos obtenidas para esta investigación incluyen diferentes registros de hospitalización de pacientes con otros trastornos de salud mental. En este sentido, se pueden plantear investigaciones futuras que analicen otros tipos de trastornos relacionados con la esquizofrenia como son la psicosis y los trastornos de personalidad.
- Por último, es interesante mejorar la aplicación que calcula el riesgo de reingreso de los pacientes con esquizofrenia, e implementarla en el servicio de psiquiatría del Complejo Asistencial de Zamora, al que pertenece el psiquiatra experto que colabora en esta Tesis Doctoral.

Capítulo 8

Conclusions

This Doctoral Thesis focuses on the readmission of patients with schizophrenia in public hospitals in Castilla y León. To address this problem, predictive models are developed using a ML approach, and the factors associated with the readmission risk of this type of patient are identified. In this chapter, the main contributions achieved in section 8.1 are presented. Section 8.2 presents the conclusions of the research and suggests possible future lines research in section 8.3.

8.1 Contributions

The main contributions of this Doctoral Thesis in this field of research are highlighted below.

- Characteristics of hospitalized patients with schizophrenia who are readmitted to public hospitals in CyL. The characteristics of the data set were analyzed to obtain a profile of the patient who tends to be readmitted to hospitals in this region.
- Identification of the ML algorithm that best fits the hospitalization data of patients with schizophrenia in CyL. Different ML algorithms were compared obtaining better performance values, compared to similar studies found. This shows that the ML approach is suitable for developing predictive models in this population.
- Identification of risk factors associated with readmission of patients with schizophrenia. The wrapper method was applied to determine the

importance of the variables. It was shown that the most predictive variables in the dataset are associated with the risk factors for readmission described in the literature for this type of patient.

- Clinical application model to predict the readmission risk of patients with schizophrenia in this population. This model was validated with a sample of 5 412 hospitalized patients from two different datasets, showing a high predictive ability of readmission risk compared to similar studies.
- Web application prototype that calculates the readmission risk of patients with schizophrenia at hospital discharge. To the best of our knowledge, this is the first web application developed in CyL to calculate the risk of readmission in this population.

These contributions are intended to support clinical decision-making, help in hospital management, and reduce hospitalizations in public hospitals of CyL.

8.2 Main conclusions

Based on the results obtained throughout this research, the following conclusions can be reached:

- According to the administrative dataset and the algorithms we used in the study, it has been concluded that the algorithms that best fit this type of data in this population are the DT-based algorithms. The best performance values obtained throughout this research, have been obtained from classifiers such as: RF, XGBoost and AB, compared to algorithms such as SVM and MLP.
- The ML approach helps in the identification of risk factors associated with the readmission of patients with schizophrenia in this population.

After applying a wrapper method to determine the significance of variables, in this research, the most predictive variables were identified from the dataset. These variables are: age, length of stay, diagnoses classified as mental disorders, substance abuse, and type V diagnoses. In this sense, it can be concluded that these predictor variables indicate possible risk factors associated with the readmission of patients with schizophrenia in CyL.

- Based on the results obtained from each of the models compared in this research, it can be concluded that the model that best predicts the readmission risk of patients with schizophrenia in CyL is the RF with AUC = (0.833-0.879) y Acc = (0.797-0.817). These model performance values show a reliable approach to predict the risk of readmission in this population, compared to existing similar studies.
- The web application developed in this research allows the implementation in clinical practice of the model that best predicts the readmission risk of patients with schizophrenia in CyL. It is considered that this tool can be used to support health personnel in decision-making.

Overall, with the results obtained in this Doctoral Thesis, it has been concluded that: predicting the readmission risk of these patients and the associated factors, through a ML approach, can lead to the development of preventive intervention strategies. These strategies can help to improve the quality of health care, patient safety, well-being and reduce health care costs in the mental health area. In general, with the results obtained with this Doctoral Thesis

8.3 Future lines

Throughout this Doctoral Thesis, several questions have arisen that can be addressed in future studies to complement the results of this research. The main future lines of this Thesis are the following:

- To cross the DB1 and DB2 with clinical data on patients with schizophrenia hospitalized in these health centers. This allows the identification of risk factors for this type of patient to be based not only on administrative data, but also on the analysis of clinical data that may influence this population.
- Analyze BD2 that contains a larger number of variables than BD1. Based on this analysis, new predictive models can be developed using the proposed methodology.
- The BD2 includes some disaggregation variables, which could not be analyzed in this Doctoral Thesis. These variables are responsible for grouping different groups of diagnoses that are clinically homogeneous, and are associated with the levels of severity and mortality risk of the patient. In this sense, it is interesting to propose a study of the behavior of hospitalized patients with schizophrenia with respect to these disaggregation variables.
- To collect data from CyL on patients with schizophrenia who are not hospitalized and to be able to make comparisons between patients who are hospitalized and those who are not. This type of comparison allows us to obtain a better profile of the patient who is readmitted to public hospitals in this region.
- Combine algorithms different from those used in this research, to develop new predictive models based on the proposed methodology.

In this sense, a different approach for feature selection can be used, with integrated methods that include wrapping and filtering methods.

- Substance abuse is one of the most influential risk factors in the hospitalization of patients with schizophrenia. Consequently, the behavior of this variable in patients who are readmitted can be analyzed, and predictive models can be proposed, using the databases of this research.
- Another interesting future line is the validation of the proposed methodology in larger databases from other regions of the country. This could increase the reliability and generalization of the results.
- The databases obtained for this research include different hospitalization records of patients with other mental health disorders. In this sense, future research may consider analyzing other types of disorders related to schizophrenia such as psychosis and personality disorders.
- Finally, it is interesting to improve the application that calculates the readmission risk of patients with schizophrenia, and to implement it in the psychiatry service of the Complejo Asistencial de Zamora, to which the expert physician collaborating in this Doctoral Thesis belongs.

Apéndice A

Resultados de la BD1

A.1.1 Transformación de las variables de la BD1

La Tabla A.1.1 muestra las transformaciones realizadas en la BD1. A lo largo de esta Tesis Doctoral se utilizó el entorno de desarrollo integrado RStudio de R. Las variables de diagnósticos secundarios (D2-D10) del conjunto original, contienen diferentes códigos de enfermedades que siguen el estándar CIE-9. Cada una de estas variables incluye las siguientes enfermedades: 1) trastornos de salud mental (Códigos 290.0-319), 2) sistema nervioso (Códigos 323.9-389.9), 3) infecciosas (Códigos 041.4-138), 4) neoplasias (Códigos 141-239.6), 5) sistema circulatorio (Códigos 394.1-459.81), 6) endocrinas (Códigos 240-279.11), 7) aparato respiratorio (Códigos 461.1-519.8), 8) aparato digestivo (Códigos 521.09-579.8), 9) aparato genitourinario (Códigos 584.8 626.6), 10) de la sangre (Códigos 280-289.81), 11) piel y tejido subcutáneo (Códigos 681.02-709.01), 12) lesiones y envenenamientos (Códigos 801.30-999.2), 13) anomalías congénitas (Códigos 743.61-759.89), y 14) artropatías y trastornos relacionados (Códigos 710.1-737.39). La BD1 también contiene códigos de diagnóstico V, que se registran como “diagnósticos” o “problemas” cuando el paciente presenta una circunstancia que influye en su estado de salud, pero que no constituye una enfermedad o lesión clasificada en las categorías 001-999 de la CIE-9. Las variables se trasformaron teniendo en cuenta el tipo de comorbilidad. La variable días de estancia ha sido creada a partir de la fecha de ingreso y de alta del paciente.

Tabla A.1.1 Variables de la base de datos de los hospitales públicos de CyL en el período de 2005-2015. Fuente: Tabla extraída de (Góngora Alonso et al., 2023).

Variables	**Descripción**	**Tipo de Variable**	**N (3 065 pacientes) /%**
Edad	Edad del paciente	Categórica:	
		< 18 años	10 (0.33)
		18-30 años	520 (16.96)
		31-50 años	1 577 (51.45)
		51-65 años	676 (22.06)
		> 65 años	282 (9.20)
Género	Género del paciente	Categórica:	
		Femenino	1 005 (32.79)
		Masculino	2 060 (67.21)
Hospital	Nombre del complejo asistencial donde el paciente es hospitalizado	Categórica:	
		Complejo Asistencial de Ávila	188 (6.13)
		Complejo Asistencial de Burgos	526 (17.16)
		Complejo Asistencial de León	425 (13.87)
		Complejo Asistencial de Palencia	210 (6.85)
		Complejo Asistencial de Salamanca	380 (12.40)
		Complejo Asistencial de Soria	159 (5.19)
		Complejo Asistencial de Zamora	341 (11.12)
		Complejo Asistencial de Segovia	178 (5.81)
		Hospital Clínico Universitario de Valladolid	209 (6.82)
		Hospital El Bierzo	330 (10.77)
		Hospital Universitario Río Hortega	119 (3.88)
Año	Año de hospitalización del paciente	Categórica:	
		2005-2008	1 226 (40.00)
		2009-2011	758 (24.73)
		2012-2015	1081 (35.27)

Variables	**Descripción**	**Tipo de Variable**	**N (3 065 pacientes)/%**
Tipo_alta	Tipo de alta hospitalaria	Categórica:	
		Domicilio	2 680 (87.44)
		Alta sin notificación documentada	2 (0.07)
		Traslado a otro hospital	255 (8.32)
		Traslado a centros de media y larga estancia	9 (0.29)
		Alta voluntaria	35 (1.14)
		Otros tipos de alta	84 (2.74)
Tipo_ingreso	Tipo de ingreso hospitalario	Categórica:	
		Ingreso urgente	2 961 (96.61)
		Ingreso programado	104 (3.39)
D1_Código	Código de diagnóstico principal: Esquizofrenia	Categórica:	
		295.0	38 (1.24)
		295.1	113 (3.69)
		295.2	9 (0.29)
		295.3	2 008 (65.51)
		295.5	9 (0.29)
		295.6	475 (15.50)
		295.8	79 (2.58)
		295.9	334 (10.90)
D1_Descripción	Nombre del diagnóstico principal: Esquizofrenia	Categórica:	
		Esquizofrenia simple	38 (1.24)
		Esquizofrenia de tipo desorganizado	113 (3.69)
		Esquizofrenia catatónica	9 (0.29)
		Esquizofrenia paranoide	2 008 (65.51)
		Esquizofrenia latente	9 (0.29)
		Esquizofrenia residual	475 (15.50)
		Otra esquizofrenia especificada	79 (2.58)
		Esquizofrenia no especificada	334 (10.90)

Variables	Descripción	Tipo de Variable	N (3 065 pacientes) /%
D2_Código	Código del diagnóstico secundario: Trastornos mentales	Categórica: 53 categorías	1 662 (54.23)
D2_Descripción	Nombre del diagnóstico secundario: Trastornos mentales	Categórica: 53 categorías	1 662 (54.23)
D3_Código	Código del diagnóstico secundario: Abuso de sustancias	Categórica: 20 categorías	1 335 (43.56)
D3_Descripción	Nombre del diagnóstico secundario: Abuso de sustancias	Categórica: 20 categorías	1 335 (43.56)
D4_Código	Código del diagnóstico secundario: Enfermedades infecciosas	Categórica: 52 categorías	175 (5.71)
D4_Descripción	Nombre del diagnóstico secundario: Enfermedades infecciosas	Categórica: 52 categorías	175 (5.71)
D5_Código	Código del diagnóstico secundario: Neoplasias	Categórica: 53 categorías	56 (1.83)
D5_Descripción	Nombre del diagnóstico secundario: Neoplasias	Categórica: 53 categorías	56 (1.83)
D6_Código	Código del diagnóstico secundario: Enfermedades endocrinas	Categórica: 26 categorías	733 (23.92)
D6_Descripción	Nombre del diagnóstico secundario: Enfermedades endocrinas	Categórica: 26 categorías	733 (23.92)
D7_Código	Código del diagnóstico secundario: Enfermedades de la sangre	Categórica: 37 categorías	99 (3.23)
D7_Descripción	Nombre del diagnóstico secundario: Enfermedades de la sangre	Categórica: 37 categorías	99 (3.23)
D8_Código	Código del diagnóstico secundario: Enfermedades del sistema circulatorio	Categórica: 37 categorías	370 (12.07)
D8_Descripción	Nombre del diagnóstico secundario: Enfermedades del sistema circulatorio	Categórica: 37 categorías	370 (12.07)
D9_Código	Código del diagnóstico secundario: Enfermedades del aparato respiratorio	Categórica: 52 categorías	170 (5.55)
D9_Descripción	Nombre del diagnóstico secundario: Enfermedades del aparato respiratorio	Categórica: 52 categorías	170 (5.55)
D10_Código	Código del diagnóstico secundario: Enfermedades del aparato digestivo	Categórica: 38 categorías	235 (7.67)

Variables	Descripción	Tipo de Variable	N (3 065 pacientes)/%
D10_Descripción	Nombre del diagnóstico secundario: Enfermedades del aparato digestivo	Categórica: 38 categorías	235 (7.67)
D11_Código	Código del diagnóstico secundario: Enfermedades del aparato genitourinario	Categórica: 29 categorías	147 (4.80)
D11_Descripción	Nombre del diagnóstico secundario: Enfermedades del aparato genitourinario	Categórica: 29 categorías	147 (4.80)
D12_Código	Código del diagnóstico secundario: Enfermedades de la piel y del tejido subcutáneo	Categórica: 45 categorías	90 (2.94)
D12_Descripción	Nombre del diagnóstico secundario: Enfermedades de la piel y del tejido subcutáneo	Categórica: 45 categorías	90 (2.94)
D13_Código	Código del diagnóstico secundario: Enfermedades del sistema osteomioarticular	Categórica: 34 categorías	145 (4.73)
D13_Descripción	Nombre del diagnóstico secundario: Enfermedades del sistema osteomioarticular	Categórica: 34 categorías	145 (4.73)
D14_Código	Código del diagnóstico secundario: Síntomas, signos y estados mal definidos	Categórica: 40 categorías	321 (10.47)
D14_Descripción	Nombre del diagnóstico secundario: Síntomas, signos y estados mal definidos	Categórica: 40 categorías	321 (10.47)
D15_Código	Código del diagnóstico secundario: Lesiones y envenenamientos	Categórica: 52 categorías	401 (13.08)
D15_Descripción	Nombre del diagnóstico secundario: Lesiones y envenenamientos	Categórica: 52 categorías	401 (13.08)
D16_Código	Código del diagnóstico secundario: Códigos de diagnóstico tipo V	Categórica: 53 categorías	2 172 (70.86)
D16_Descripción	Nombre del diagnóstico secundario: Códigos de diagnóstico tipo V	Categórica: 53 categorías	2 172 (70.86)

Variables	Descripción	Tipo de Variable	N (3 065 pacientes) /%
PROC1_Código	Código del procedimiento 1	Categórica: 52 categorías	2 400 (78.30)
PROC1_Descripción	Nombre de procedimiento 1	Categórica: 52 categorías	2 400 (78.30)
PROC2_Código	Código del procedimiento 2	Categórica: 52 categorías	1 448 (47.24)
PROC2_Descripción	Nombre del procedimiento 2	Categórica: 52 categorías	1 448 (47.24)
PROC3_Código	Código del procedimiento 3	Categórica: 43 categorías	686 (22.38)
PROC3_Descripción	Nombre del procedimiento 3	Categórica: 43 categorías	686 (22.38)
Duración_estancia	Días de estancia hospitalaria	Numérico continuo	Media: 17 días SD: 16.26

A.1.2 Resultados de los modelos de predicción de hospitalización

En la Tabla A.1.2 y A.1.3 se muestran los resultados de las métricas de rendimiento de cada modelo, a partir de cada clase de la variable dependiente. Estos modelos son los que se han comparado en la primera fase de esta Tesis Doctoral, para identificar el que mejor se ajusta al conjunto de datos

Tabla A.1.2 *Scores* con *target* = 0. Fuente: Tabla extraída de (Góngora Alonso et al., 2022).

Modelo	AUC	Acc	F1	Precision	Recall
Random Forest	0.79593	0.72736	0.72165	0.73375	0.70994
AdaBoost	0.76800	0.70818	0.70525	0.70923	0.70132
Tree	0.68197	0.68176	0.68883	0.67105	0.70757
kNN	0.72839	0.67654	0.67044	0.68024	0.66092
Naïve Bayes	0.72864	0.67023	0.65925	0.67878	0.64080
SVM	0.66229	0.65727	0.65908	0.65279	0.66548

Tipo de muestreo: Validación cruzada estratificada k=10

Clase objetivo: 0 - no-esquizofrenia

Tabla A1.3 *Scores* con *target* = 1. Fuente: Tabla extraída de (Góngora Alonso et al., 2022).

Modelo	**AUC**	**Acc**	**F1**	**Precision**	**Recall**
Random Forest	0.79595	0.72736	0.73285	0.72143	0.74464
AdaBoost	0.76801	0.70818	0.71105	0.70716	0.71498
Tree	0.68197	0.68176	0.67436	0.69359	0.65617
kNN	0.72839	0.67654	0.68242	0.67308	0.69202
Naïve Bayes	0.72864	0.67023	0.68052	0.66264	0.69940
SVM	0.66229	0.65727	0.65544	0.66188	0.64913

Tipo de muestreo: Validación cruzada estratificada k=10

Clase objetivo: 1 - esquizofrenia

Apéndice B

Resultados de la BD2

B.1 Transformación de las variables de la BD2

La Tabla B1.1 muestra las transformaciones realizadas en la BD2. A lo largo de esta Tesis Doctoral se utilizó el entorno de desarrollo integrado RStudio de R. Las variables de diagnósticos secundarios (D2-D20) del conjunto de datos original, contienen diferentes códigos de enfermedades que siguen el estándar CIE-10. En un inicio se trasformaron las variables de igual manera que en la BD1. Cada una de estas variables incluye las siguientes enfermedades: 1) infecciosas (Códigos A00-B99), 2) sistema nervioso (Códigos G00-G99), 3) trastornos mentales (Códigos F01-F99), 4) neoplasias (Códigos C00-D49), 5) aparato circulatorio (Códigos I00-I99), 6) endocrinas (Códigos E00-E89), 7) aparato respiratorio (Códigos J00-J99), 8) aparato digestivo (Códigos K00-K95), 9) aparato genitourinario (Códigos N00-N99), 10) de la sangre (Códigos D50-D89), 11) piel y tejido subcutáneo (Códigos L00-L99), 12) lesiones y envenenamientos (Códigos S00-T88), 13) anomalías congénitas (Códigos Q00-Q99), y 14) artropatías y trastornos relacionados (Códigos M00-M99). Además de los códigos de diagnóstico V (Códigos Z00-Z99). Las variables se trasformaron teniendo en cuenta el tipo de comorbilidad, posteriormente, se codificaron del CIE-10 al CIE-9. La variable días de estancia ha sido creada a partir de la fecha de ingreso y de alta del paciente.

Tabla B1.1 Variables de la base de datos de los hospitales públicos de CyL en el período de 2015-2020

Variables	**Descripción**	**Tipo de Variable**	**N (2 347 pacientes) /%**
Edad	Edad del paciente	Categórica:	
		< 18 años	19 (0.81)
		18-30 años	309 (13.17)
		31-50 años	1 150 (49.00)
		51-65 años	664 (28.29)
		> 65 años	205 (8.73)
Género	Género del paciente	Categórica:	
		Femenino	835 (35.58)
		Masculino	1 512 (64.42)
Hospital	Nombre del complejo asistencial donde el paciente es hospitalizado	Categórica:	
		Complejo Asistencial de Ávila	112 (4.77)
		Complejo Asistencial de Burgos	387 (16.49)
		Complejo Asistencial de León	363 (15.47)
		Complejo Asistencial de Palencia	175 (7.46)
		Complejo Asistencial de Salamanca	229 (9.76)
		Complejo Asistencial de Soria	120 (5.11)
		Complejo Asistencial de Segovia	146 (6.22)
		Complejo Asistencial de Zamora	230 (9.80)
		Hospital Clínico Universitario de Valladolid	209 (8.90)
		Hospital El Bierzo	244 (10.40)
		Hospital Universitario Río Hortega	132 (5.62)
Año	Año de hospitalización del paciente	Categórica:	
		2015-2016	527 (22.45)
		2017-2018	950 (40.48)
		2019-2020	870 (37.07)

Variables	**Descripción**	**Tipo de Variable**	**N (2 347 pacientes)/%**
Tipo_alta	Tipo de alta hospitalaria	Categórica:	
		Domicilio	1 967 (83.81)
		Traslado a otro hospital	192 (8.18)
		Traslado a centros de media y larga estancia	23 (0.98)
		Alta voluntaria	15 (0.64)
		Otros tipos de alta	150 (6.39)
Tipo_ingreso	Tipo de ingreso hospitalario	Categórica:	
		Ingreso urgente	2 106 (89.73)
		Ingreso programado	241 (10.27)
D1_Código	Código de diagnóstico principal: Esquizofrenia	Categórica:	
		295.0: Esquizofrenia simple	37 (1.58)
		295.1: Esquizofrenia de tipo desorganizado	45 (1.92)
		295.2: Esquizofrenia catatónica	12 (0.51)
		295.3: Esquizofrenia paranoide	956 (40.73)
		295.6: Esquizofrenia residual	170 (7.24)
		295.8: Otra esquizofrenia especificada	731 (31.15)
		295.9: Esquizofrenia no especificada	396 (16.87)
D2_Código	Código del diagnóstico secundario: Trastornos mentales	Categórica: 47 categorías	1 066 (43.87)
D3_Código	Código del diagnóstico secundario: Abuso de sustancias	Categórica: 18 categorías	979 (40.29)
D4_Código	Código del diagnóstico secundario: Enfermedades infecciosas	Categórica: 23 categorías	114 (4.69)
D5_Código	Código del diagnóstico secundario: Neoplasias	Categórica: 28 categorías	51 (2.10)
D6_Código	Código del diagnóstico secundario: Enfermedades endocrinas	Categórica: 23 categorías	750 (30.86)

Variables	Descripción	Tipo de Variable	N (2 347 pacientes) /%
D7_Código	Código del diagnóstico secundario: Enfermedades de la sangre	Categórica: 22 categorías	93 (3.83)
D8_Código	Código del diagnóstico secundario: Enfermedades del sistema circulatorio	Categórica: 29 categorías	332 (13.66)
D9_Código	Código del diagnóstico secundario: Enfermedades del aparato respiratorio	Categórica: 29 categorías	174 (7.16)
D10_Código	Código del diagnóstico secundario: Enfermedades del aparato digestivo	Categórica: 28 categorías	160 (6.58)
D11_Código	Código del diagnóstico secundario: Enfermedades del aparato genitourinario	Categórica: 21 categorías	159 (6.54)
D12_Código	Código del diagnóstico secundario: Enfermedades de la piel y del tejido subcutáneo	Categórica: 27 categorías	80 (3.29)
D13_Código	Código del diagnóstico secundario: Enfermedades del sistema osteomioarticular	Categórica: 26 categorías	127 (5.23)
D14_Código	Código del diagnóstico secundario: Síntomas, signos y estados mal definidos	Categórica: 35 categorías	395 (16.26)
D15_Código	Código del diagnóstico secundario: Lesiones y envenenamientos	Categórica: 37 categorías	258 (10.62)
D16_Código	Código del diagnóstico secundario: Códigos de diagnóstico tipo V	Categórica: 42 categorías	1 456 (59.92)
PROC1_Código	Código del procedimiento 1	Categórica: 32 categorías	1 924 (79.18)
PROC2_Código	Código del procedimiento 2	Categórica: 25 categorías	767 (31.56)
PROC3_Código	Código del procedimiento 3	Categórica: 25 categorías	318 (13.09)
Duración_estancia	Días de estancia hospitalaria del paciente	Numérico continuo	Media: 42 días SD: 100

Apéndice C

Logros científicos

C.1.1 Publicaciones para defender la Tesis Doctoral

En el desarrollo de esta Tesis Doctoral se han generado las siguientes producciones científicas:

1. **Góngora Alonso, S.**, Marques, G., Agarwal, D., De la Torre Díez, I., & Franco-Martín, M. (2022). Comparison of Machine Learning Algorithms in the Prediction of Hospitalized Patients with Schizophrenia. *Sensors*, *22*(7), 2517. https://doi.org/10.3390/s22072517

2. **Góngora Alonso, S.**, Herrera Montano, I., Ayala, J. L. M., Rodrigues, J. J., Franco-Martín, M., & de la Torre Díez, I. (2023). Machine Learning Models to Predict Readmission Risk of Patients with Schizophrenia in a Spanish Region. *International Journal of Mental Health and Addiction*, 1-20. https://doi.org/10.1007/s11469-022-01001-x

C.1.2 Publicaciones relacionadas con la Tesis Doctoral

1. **Alonso, S. G.**, De la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. *Journal of Medical Systems*, *42*(9), 161. https://doi.org/10.1007/s10916-018-1018-2

2. Barreno, D. C., **Alonso, S. G.**, De la Torre Díez, I., Coronado, M. L., & Franco, M. (2020). A New Software Tool for Analyzing Mental Health Data in a Spanish Region. *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, *76*, 898–906. https://doi.org/10.1007/978-3-030-31635-8_109

3. **Góngora Alonso, S.**, Sainz-De-Abajo, B., De la Torre-Díez, I., & Franco-Martin, M. (2020). Health Care Management Models for the Evolution of Hospitalization in Acute Inpatient Psychiatry Units: Comparative Quantitative Study. *JMIR Mental Health*, *7*(11), e15776. https://doi.org/10.2196/15776

4. **Góngora Alonso, S.**, de Bustos Molina, A., Sainz-De-Abajo, B., Franco-Martín, M., & De la Torre Díez, I. (2021). Analysis of Mental Health Disease Trends Using BeGraph Software in Spanish Health Care Centers: Case Study. *JMIR Medical Informatics*, *9*(6), e15527. https://doi.org/10.2196/15527

C.1.3 Otras publicaciones indexadas en JCR

1. **Alonso, S. G.**, De la Torre Díez, I., Rodrigues, J. J. P. C., Hamrioui, S., & López-Coronado, M. (2017). A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector. *Journal of Medical Systems*, *41*(11), 183. https://doi.org/10.1007/s10916-017-0832-2

2. De la Torre Díez, Isabel, **Góngora Alonso, S.**, Hamrioui, S., López-Coronado, M., & Motta Cruz, E. (2018). Systematic Review about QoS and QoE in Telemedicine and eHealth Services and Applications. *Journal of Medical Systems*, *42*(10), 182.

3. **Góngora Alonso, S.**, Hamrioui, S., De la Torre Díez, I., Motta Cruz, E., López-Coronado, M., & Franco, M. (2018). Social Robots for People with

Aging and Dementia: A Systematic Review of Literature. *Telemedicine and E-Health*, *00*(00), tmj.2018.0051. https://doi.org/10.1089/tmj.2018.0051

4. De la Torre Díez, Isabel, **Alonso, S. G.**, Hamrioui, S., Cruz, E. M., Nozaleda, L. M., & Franco, M. A. (2019). IoT-Based Services and Applications for Mental Health in the Literature. *Journal of Medical Systems*, *43*(1), 11. https://doi.org/10.1007/s10916-018-1130-3

5. **Alonso, S. G.**, Arambarri, J., López-Coronado, M., & De la Torre Díez, I. (2019). Proposing New Blockchain Challenges in eHealth. *Journal of Medical Systems*, *43*(3), 64. https://doi.org/10.1007/s10916-019-1195-7

6. García, J. S., **Alonso, S. G.**, De la Torre Díez, I., Garcia-Zapirain, B., Castillo, C., Coronado, M. L., & Salvador, J. C. (2019). Reviewing Mobile Apps to Control Heart Rate in Literature and Virtual Stores. *Journal of Medical Systems*, *43*(4), 80. https://doi.org/10.1007/s10916-019-1202-z

7. **Alonso, S. G.**, De la Torre Díez, I., & Zapiraín, B. G. (2019). Predictive, Personalized, Preventive and Participatory (4P) Medicine Applied to Telemedicine and eHealth in the Literature. *Journal of Medical Systems*, *43*(5), 140. https://doi.org/10.1007/s10916-019-1279-4

8. De la Torre Díez, I., **Alonso, S. G**., Cruz, E. M., & Franco, M. A. (2019). Measuring QoE of a Teleconsultation App in Mental Health Using a Pentagram Model. *Journal of Medical Systems*, *43*(7), 213.

9. **Góngora Alonso, S.**, Fumero Vargas, G., Morón Nozaleda, L., Sainz de Abajo, B., De la Torre Díez, I., & Franco, M. (2020). Usability Analysis of a System for Cognitive Rehabilitation, "Gradior", in a Spanish Region. *Telemedicine and E-Health*, *26*(5), 671–682. https://doi.org/10.1089/tmj.2019.0084

10. **Góngora Alonso, S.**, Toribio Guzmán, J. M., Sainz de Abajo, B., Muñoz Sánchez, J. L., Martín, M. F., & De la Torre Díez, I. (2020). Usability evaluation of the eHealth Long Lasting Memories program in Spanish elderly people. *Health Informatics Journal*, *26*(3), 1728–1741. https://doi.org/10.1177/1460458219889501

11. Montano, I. H., Marques, G., **Alonso, S. G.**, López-Coronado, M., & De la Torre Díez, I. (2020). Predicting Absenteeism and Temporary Disability Using Machine Learning: a Systematic Review and Analysis. *Journal of Medical Systems*, *44*(9), 162. https://doi.org/10.1007/s10916-020-01626-2

C.1.4 Conferencias internacionales

1. **Alonso, S. G.**, de Bustos Molina, A., Hamrioui, S., Coronado, M. L., Martín, M. F., Khanna, A., & De la Torre Díez, I. (2021). Analyzing Mental Health Diseases in a Spanish Region Using Software Based on Graph Theory Algorithms. In *Advances in Intelligent Systems and Computing* (Vol. 1165, pp. 701–708). https://doi.org/10.1007/978-981-15-5113-0_57

2. **Alonso, S. G.**, De Abajo, B. S., De La Torre Diez, I., Vargas, G. F., & Martin, M. F. (2021). Using a Computer-Based Program to Treat Neurocognitive Deficit Disorders in Spanish Population. *Iberian Conference on Information Systems and Technologies, CISTI*, (June), 23–26. https://doi.org/10.23919/CISTI52073.2021.9476514

3. **Alonso, S. G.**, De Abajo, B. S., De La Torre Diez, I., Toribio Guzman, J. M., Munoz Sanchez, J. L., Martin, M. F., & Rodrigues, J. J. P. C. (2021). An Experience with Mental Health Professionals using Long Lasting Memories Program. *2021 IEEE Global Communications Conference (GLOBECOM)*, 01–06. https://doi.org/10.1109/GLOBECOM46510.2021.9685479

C.1.5 Prácticas internacionales

Prácticas de investigación en el Instituto de Telecomunicações, Delegação da Covilhã, Portugal, durante tres meses. El objetivo de la estancia estuvo enfocado en colaborar con el Instituto de Telecomunicações, para aplicar algoritmos de ML a una base de datos de pacientes con trastornos de esquizofrenia, en CyL. Con el fin de analizar y predecir resultados de hospitalización y reingresos de estos pacientes.

Por tanto, para lograr el objetivo principal de la estancia doctoral se realizaron las siguientes actividades:

1. Revisión exhaustiva de artículos científicos que usan técnicas de ML para la predicción de reingresos y factores de riesgo de pacientes con esquizofrenia.
2. Evaluación del riesgo de reingresos en pacientes de CyL con trastornos de esquizofrenia. Para evaluar el riesgo de reingresos se han aplicado y desarrollado modelos predictivos usando los algoritmos de ML obtenidos en la tarea 1 (empleando R).
3. Se han obtenido resultados clínicos referente a los factores de riesgo en pacientes con esta enfermedad.
4. Validación de los resultados usando validación cruzada *k-fold*.
5. Con los resultados obtenidos en la investigación publicamos el artículo: **Góngora Alonso, S.**, Herrera Montano, I., Ayala, J. L. M., Rodrigues, J. J., Franco-Martín, M., & De la Torre Díez, I. (2023). Machine Learning Models to Predict Readmission Risk of Patients with Schizophrenia in a Spanish Region. *International Journal of Mental Health and Addiction*, 1-20. https://doi.org/10.1007/s11469-022-01001-x
6. Como continuación de la investigación y teniendo en cuenta el estudio anterior, identificamos el abuso de sustancias como uno de los principales factores de riesgo en esta población. Por tanto, se ha realizado una revisión

exhaustiva de artículos científicos referente a la influencia del uso de sustancias en pacientes con esquizofrenia.

7. En colaboración con el Instituto de Telecomunicações, se presentó en IEEE Global Communications Conference (GLOBECOM-2021), un estudio que generó la siguiente publicación:

   Alonso, S. G., De Abajo, B. S., De La Torre Diez, I., Toribio Guzman, J. M., Munoz Sanchez, J. L., Martin, M. F., & Rodrigues, J. J. P. C. (2021). An Experience with Mental Health Professionals using Long Lasting Memories Program. *2021 IEEE Global Communications Conference (GLOBECOM)*, 01–06. https://doi.org/10.1109/GLOBECOM46510.2021.9685479

Apéndice D

D.1 Manual de usuario de la aplicación web

1. Interfaz gráfica principal de la aplicación web para calcular el riesgo de reingreso de pacientes con esquizofrenia en CyL (Figura C.1.1). En la interfaz, se muestra el menú principal y un gráfico general del comportamiento de la variable abuso de sustancias en cada hospital, teniendo en cuenta las categorías de mayor prevalencia en estos pacientes: tabaco, cannabis y alcohol.

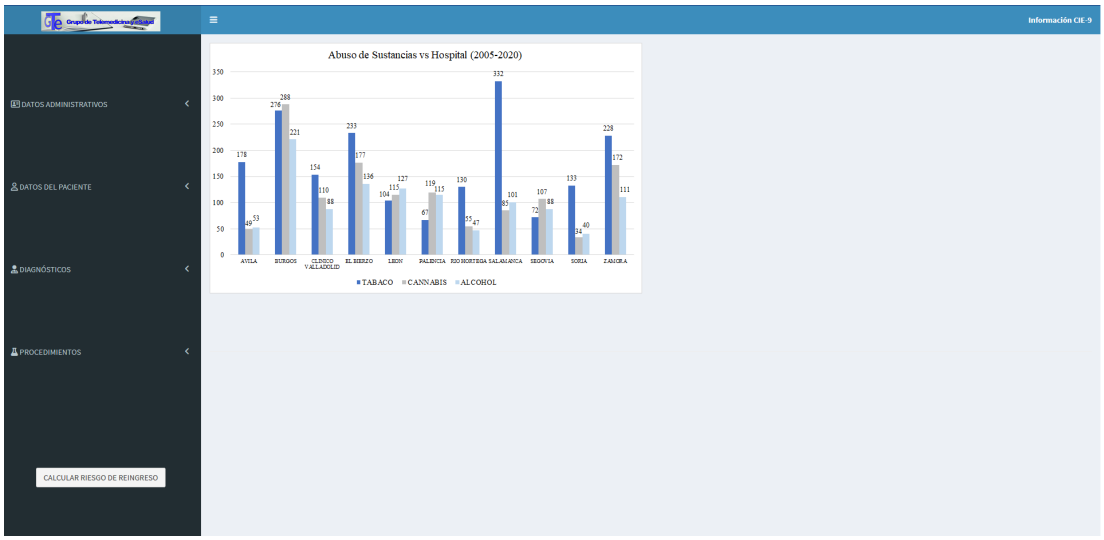

Figura C.1.1 Interfaz gráfica de la aplicación.

2. Menú principal de la aplicación para seleccionar los distintos tipos de variables del paciente (Ver Figura C.1.2).

   Datos Administrativos: contiene los 11 hospitales públicos de CyL y la variable días de estancia.

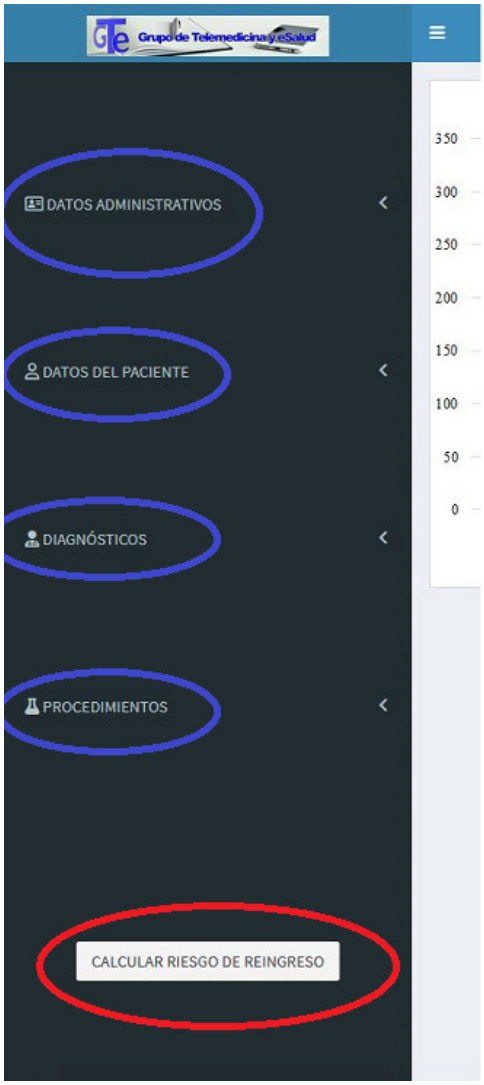

Figura C.1.2 Menú principal.

Datos del Paciente: contiene el rango de edad de la persona y el género**.**

Diagnósticos: incluye el diagnóstico principal de esquizofrenia y los diagnósticos secundarios que presenta el paciente a lo largo de la estancia hospitalaria.

Procedimientos: Estas variables incluyen los procedimientos realizados en el hospital que requieren recursos materiales y humanos especializados, que

implican un riesgo para el paciente o que son relevantes para su estudio o tratamiento.

Calcular Riesgo de Reingreso: Este botón te permite calcular el riesgo de reingreso del paciente al alta hospitalaria. Para calcular, el usuario tiene que seleccionar todas las variables del menú. Cuando el paciente no presente esa enfermedad se selecciona "No especificado".

3. Resultados de selección de los datos. Al seleccionar el botón "Calcular Riesgo de Reingreso", se muestra la interfaz con el resultado de riesgo del paciente (Ver Figura C.1.3)

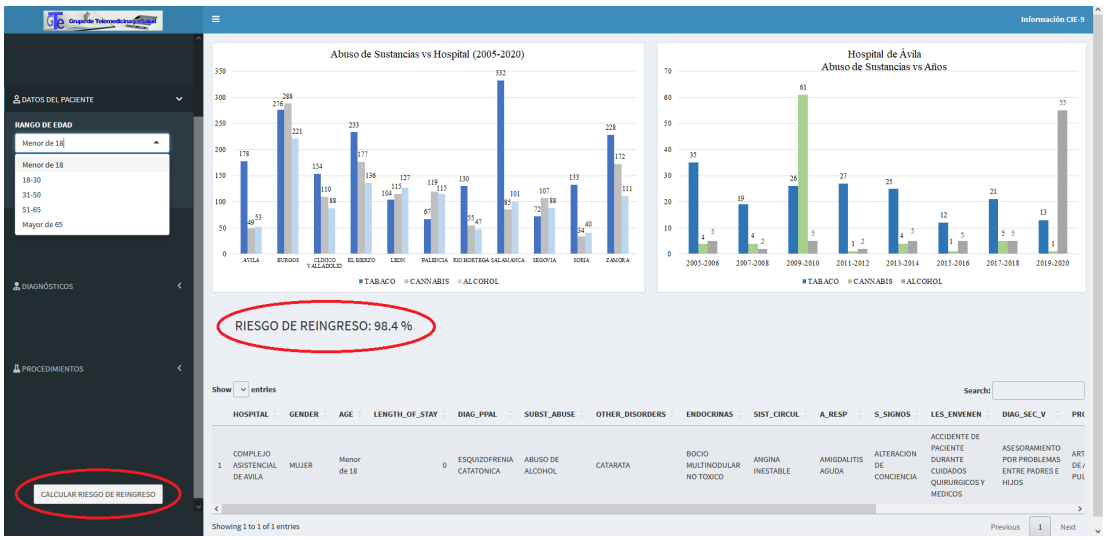

Figura C.1.3 Resultados del riesgo de reingreso en el Complejo Asistencial de Ávila.

4. El botón de calcular devuelve una tabla con todos los valores seleccionados en cada una de las variables y la probabilidad de reingreso del paciente (Ver Figura C.1.4).

5. Una vez que se calcula el valor de riesgo de reingreso, se muestra en la interfaz, un gráfico del comportamiento de la variable abuso de sustancia por años (2005-2020), en el hospital seleccionado (Ver Figura C.1.5). En la parte

superior derecha se visualiza un enlace, que permite acceder en línea a los códigos de clasificación de las enfermedades CIE-9.

Figura C.1.4 Tabla de registro del paciente.

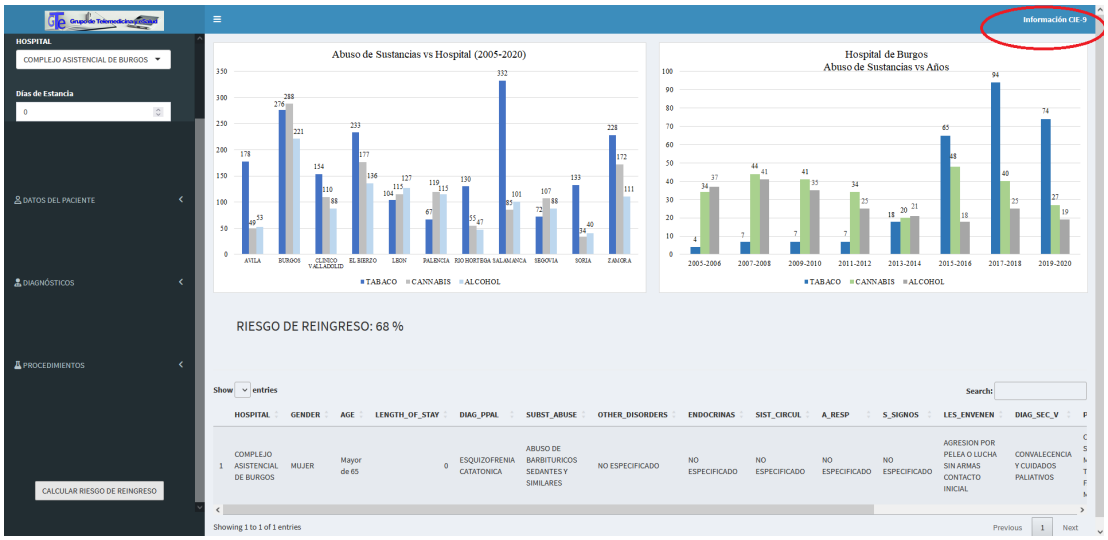

Figura C.1.5 Resultados del riesgo de reingreso en el Complejo Asistencial de Burgos.

Bibliografía

Adebiyi, M. O., Mosaku, S. K., Irinoye, O. O., & Oyelade, O. O. (2018). Socio-demographic and clinical factors associated with relapse in mental illness. *International Journal of Africa Nursing Sciences*, *8*, 149–153. https://doi.org/10.1016/j.ijans.2018.05.007

Akdeniz, C., Tost, H., & Meyer-Lindenberg, A. (2014). The neurobiology of social environmental risk for schizophrenia: an evolving research field. *Social Psychiatry and Psychiatric Epidemiology*, *49*(4), 507–517. https://doi.org/10.1007/s00127-014-0858-4

Alonso, S. G., de la Torre Díez, I., Rodrigues, J. J. P. C., Hamrioui, S., & López-Coronado, M. (2017). A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector. *Journal of Medical Systems*, *41*(11). https://doi.org/10.1007/s10916-017-0832-2

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders, 5th edition. In *Washington, DC: American Pychiatric Association*. Elsevier.

Arranz, B., Garriga, M., García-Rizo, C., & San, L. (2018). Clozapine use in patients with schizophrenia and a comorbid substance use disorder: A systematic review. *European Neuropsychopharmacology*, *28*(2), 227–242. https://doi.org/10.1016/j.euroneuro.2017.12.006

Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine*, *164*, 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006

Awad, A., Bader-El-Den, M., McNicholas, J., & Briggs, J. (2017). Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International Journal of Medical Informatics*, *108*, 185–195. https://doi.org/10.1016/j.ijmedinf.2017.10.002

Bae, Y. J., Shim, M., & Lee, W. H. (2021). Schizophrenia Detection Using Machine Learning Approach from Social Media Content. *Sensors*, *21*(17), 5924. https://doi.org/10.3390/s21175924

Baeza, F. L. C., da Rocha, N. S., & Fleck, M. P. de A. (2018). Readmission in psychiatry inpatients within a year of discharge: The role of symptoms at discharge and post-discharge care in a Brazilian sample. *General Hospital Psychiatry*, *51*, 63–70. https://doi.org/10.1016/j.genhosppsych.2017.11.008

Beeley, C. (2016). Web application development with R using Shiny. In *Packt Publishing Ltd*.

Berardelli, I., Rogante, E., Sarubbi, S., Erbuto, D., Lester, D., & Pompili, M. (2021). The Importance of Suicide Risk Formulation in Schizophrenia. *Frontiers in Psychiatry*, *12*, 1–13. https://doi.org/10.3389/fpsyt.2021.779684

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., & Kane, J. M. (2017). A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research*, *19*(8), e289. https://doi.org/10.2196/jmir.7956

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (1984). *Classification And Regression Trees* (1st editio). Routledge. https://doi.org/10.1201/9781315139470

Breiman, L. (2001). Random Forest. In *Machine Learning* (Vol. 45, pp. 5–32). https://doi.org/10.1023/A:1010933404324

Castillo-Sánchez, G., Acosta, M. J., Garcia-Zapirain, B., De la Torre, I., & Franco-Martín, M. (2022). Application of Machine Learning Techniques to Help in the Feature Selection Related to Hospital Readmissions of Suicidal Behavior. *International Journal of Mental Health and Addiction*, *0123456789*. https://doi.org/10.1007/s11469-022-00868-0

Charlson, F. J., Ferrari, A. J., Santomauro, D. F., Diminic, S., Stockings, E., Scott, J. G., McGrath, J. J., & Whiteford, H. A. (2018). Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophrenia Bulletin*, *44*(6), 1195–1203. https://doi.org/10.1093/schbul/sby058

Chi, M. H., Hsiao, C. Y., Chen, K. C., Lee, L.-T., Tsai, H. C., Hui Lee, I., Chen, P. S., & Yang, Y. K. (2016). The readmission rate and medical cost of patients with schizophrenia after first hospitalization — A 10-year follow-up population-based study. *Schizophrenia Research*, *170*(1), 184–190. https://doi.org/10.1016/j.schres.2015.11.025

Christensen, M. K., Lim, C. C. W., Saha, S., Plana-Ripoll, O., Cannon, D., Presley, F., Weye, N., Momen, N. C., Whiteford, H. A., Iburg, K. M., & McGrath, J. J. (2020). The cost of mental disorders: a systematic review. *Epidemiology and Psychiatric Sciences*, *29*, e161. https://doi.org/10.1017/S204579602000075X

Commission on Professional and Hospital Activities. (2014). The International Classification of Diseases, 9th Revision, Clinical Modification.

Costanza, A., Mazzola, V., Radomska, M., Amerio, A., Aguglia, A., Prada, P., Bondolfi, G., Sarasin, F., & Ambrosetti, J. (2020). Who Consult an Adult Psychiatric Emergency Department? Pertinence of Admissions and Opportunities for Telepsychiatry. *Medicina*, *56*(6), 295. https://doi.org/10.3390/medicina56060295

Costanza, A., Rothen, S., Achab, S., Thorens, G., Baertschi, M., Weber, K., Canuto, A., Richard-Lepouriel, H., Perroud, N., & Zullino, D. (2021). Impulsivity and Impulsivity-Related Endophenotypes in Suicidal Patients with Substance Use Disorders: an Exploratory Study. *International Journal of Mental Health and Addiction*, *19*(5), 1729–1744. https://doi.org/10.1007/s11469-020-00259-3

Cronin, R. M., Hankins, J. S., Byrd, J., Pernell, B. M., Kassim, A., Adams-Graves, P., Thompson, A., Kalinyak, K., DeBaun, M., & Treadwell, M. (2019). Risk factors for hospitalizations and readmissions among individuals with sickle cell disease: results of a U.S. survey study. *Hematology*, *24*(1), 189–198. https://doi.org/10.1080/16078454.2018.1549801

de Pedro Cuesta, J., Saiz Ruiz, J., Roca, M., & Noguer, I. (2016). Mental health and public health in Spain: Epidemiological surveillance and prevention. *Psiquiatria Biologica*, *23*(2), 67–73. https://doi.org/10.1016/j.psiq.2016.03.001

Deng, Y., Hung, K. S. Y., Lui, S. S. Y., Chui, W. W. H., Lee, J. C. W., Wang, Y., Li, Z., Mak, H. K. F., Sham, P. C., Chan, R. C. K., & Cheung, E. F. C. (2019). Tractography-based classification in distinguishing patients with first-episode schizophrenia from healthy individuals. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *88*, 66–73. https://doi.org/10.1016/j.pnpbp.2018.06.010

Deschepper, M., Eeckloo, K., Vogelaers, D., & Waegeman, W. (2019). A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Computer Methods and Programs in Biomedicine*, *173*, 177–183. https://doi.org/10.1016/j.cmpb.2019.02.007

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., & Meyer, D. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS ONE*, *11*(2), 1–23. https://doi.org/10.1371/journal.pone.0148195

Edgcomb, J. B., Thiruvalluru, R., Pathak, J., & Brooks, J. O. (2021). Machine Learning to Differentiate Risk of Suicide Attempt and Self-harm After General Medical Hospitalization of Women With Mental Illness. *Medical Care*, *59*(2), S58–S64. https://doi.org/10.1097/MLR.0000000000001467

Edgcomb, J., Shaddox, T., Hellemann, G., & Brooks, J. O. (2019). High-Risk Phenotypes of Early Psychiatric Readmission in Bipolar Disorder With Comorbid Medical Illness. *Psychosomatics*, *60*(6), 563–573. https://doi.org/10.1016/j.psym.2019.05.002

El Naqa, I., Li, R., & Murphy, M. J. (2015). What Is Machine Learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology*. Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3

El-Missiry, A., Aboraya, A. S., Manseur, H., Manchester, J., France, C., & Border, K. (2011). An Update on the Epidemiology of Schizophrenia with a Special Reference to

Clinically Important Risk Factors. *International Journal of Mental Health and Addiction*, *9*(1), 39–59. https://doi.org/10.1007/s11469-009-9241-1

English, M., Kumar, C., Ditterline, B. L., Drazin, D., & Dietz, N. (2022). *Machine Learning in Neuro-Oncology, Epilepsy, Alzheimer's Disease, and Schizophrenia* (pp. 349–361). https://doi.org/10.1007/978-3-030-85292-4_39

Fleury, M.-J., Fortin, M., Rochette, L., Grenier, G., Huỳnh, C., Pelletier, É., & Vasiliadis, H.-M. (2019). Assessing quality indicators related to mental health emergency room utilization. *BMC Emergency Medicine*, *19*(1), 8. https://doi.org/10.1186/s12873-019-0223-8

Fond, G., Bulzacka, E., Boucekine, M., Schürhoff, F., Berna, F., Godin, O., Aouizerate, B., Capdevielle, D., Chereau, I., D'Amato, T., Dubertret, C., Dubreucq, J., Faget, C., Leignier, S., Lançon, C., Mallet, J., Misdrahi, D., Passerieux, C., Rey, R., … Llorca, P. M. (2019). Machine learning for predicting psychotic relapse at 2 years in schizophrenia in the national FACE-SZ cohort. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *92*(December 2018), 8–18. https://doi.org/10.1016/j.pnpbp.2018.12.005

Freund, Y., & Schapire, R. E. (1997). A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *5*(1), 119–139. https://doi.org/10.1007/3-540-59119-2_166

Góngora Alonso, S., Herrera Montano, I., Ayala, J. L. M., Rodrigues, J. J. P. C., Franco-Martín, M., & de la Torre Díez, I. (2023). Machine Learning Models to Predict Readmission Risk of Patients with Schizophrenia in a Spanish Region. *International Journal of Mental Health and Addiction*, *0123456789*. https://doi.org/10.1007/s11469-022-01001-x

Góngora Alonso, S., Marques, G., Agarwal, D., De la Torre Díez, I., & Franco-Martín, M. (2022). Comparison of Machine Learning Algorithms in the Prediction of Hospitalized Patients with Schizophrenia. *Sensors*, *22*(7), 2517. https://doi.org/10.3390/s22072517

Góngora Alonso, S., Sainz-De-Abajo, B., De la Torre-Díez, I., & Franco-Martin, M. (2020). Health Care Management Models for the Evolution of Hospitalization in Acute

Inpatient Psychiatry Units: Comparative Quantitative Study. *JMIR Mental Health*, *7*(11), e15776. https://doi.org/10.2196/15776

Grudnikoff, E., McNeilly, T., & Babiss, F. (2019). Correlates of psychiatric inpatient readmissions of children and adolescents with mental disorders. *Psychiatry Research*, *282*, 112596. https://doi.org/10.1016/j.psychres.2019.112596

Harris, L. W., Guest, P. C., Wayland, M. T., Umrania, Y., Krishnamurthy, D., Rahmoune, H., & Bahn, S. (2013). Schizophrenia: Metabolic aspects of aetiology, diagnosis and future treatment strategies. *Psychoneuroendocrinology*, *38*(6), 752–766. https://doi.org/10.1016/j.psyneuen.2012.09.009

Hastie, T., Friedman, J., & Tibshirani, R. (2001). The Elements of Statistical Learning. In *The Elements of Statistical Learning* (Vol. 27, Issue 2). Springer New York. https://doi.org/10.1007/978-0-387-21606-5

Hettige, N. C., Nguyen, T. B., Yuan, C., Rajakulendran, T., Baddour, J., Bhagwat, N., Bani-Fatemi, A., Voineskos, A. N., Mallar Chakravarty, M., & De Luca, V. (2017). Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach. *General Hospital Psychiatry*, *47*, 20–28. https://doi.org/10.1016/j.genhosppsych.2017.03.001

Higgins, N., Meehan, T., Dart, N., Kilshaw, M., & Fawcett, L. (2018). Implementation of the Safewards model in public mental health facilities: A qualitative evaluation of staff perceptions. *International Journal of Nursing Studies*, *88*, 114–120. https://doi.org/10.1016/j.ijnurstu.2018.08.008

Holderness, E., Miller, N., Cawkwell, P., Bolton, K., Meteer, M., Pustejovsky, J., & Hall, M. H. (2019). Analysis of risk factor domains in psychosis patient health records. *Journal of Biomedical Semantics*, *10*(1), 19. https://doi.org/10.1186/s13326-019-0210-8

Hor, K., & Taylor, M. (2010). Suicide and schizophrenia: a systematic review of rates and risk factors. *Journal of Psychopharmacology*, *24*(4_suppl), 81–90. https://doi.org/10.1177/1359786810385490

Hosmer, D. W., & Lemeshow, S. (2002). *Applied logistic regression* (J. W. & Sons, Ed.).

Hou, C., Zhong, X., He, P., Xu, B., Diao, S., Yi, F., Zheng, H., & Li, J. (2020). Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. *JMIR Medical Informatics*, *8*(6), e17364. https://doi.org/10.2196/17364

Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021). Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC Medical Research Methodology*, *21*(1), 96. https://doi.org/10.1186/s12874-021-01284-z

Huberts, L. C. E., Does, R. J. M. M., Ravesteijn, B., & Lokkerbol, J. (2022). Predictive monitoring using machine learning algorithms and a real-life example on schizophrenia. *Quality and Reliability Engineering International*, *38*(3), 1302–1317. https://doi.org/10.1002/qre.2957

Hung, Y.-Y., Chan, H.-Y., & Pan, Y.-J. (2017). Risk factors for readmission in schizophrenia patients following involuntary admission. *PLOS ONE*, *12*(10), e0186768. https://doi.org/10.1371/journal.pone.0186768

Innes, H., Lewsey, J., & Smith, D. J. (2015). Predictors of admission and readmission to hospital for major depression: A community cohort study of 52,990 individuals. *Journal of Affective Disorders*, *183*, 10–14. https://doi.org/10.1016/j.jad.2015.04.019

Jahmunah, V., Lih Oh, S., Rajinikanth, V., Ciaccio, E. J., Hao Cheong, K., Arunkumar, N., & Acharya, U. R. (2019). Automated detection of schizophrenia using nonlinear signal processing methods. *Artificial Intelligence in Medicine*, *100*, 101698. https://doi.org/10.1016/j.artmed.2019.07.006

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Jin, H., & Mosweu, I. (2017). The societal cost of schizophrenia: a systematic review. *Pharmacoeconomics*, *35*(1), 25–42.

Jo, Y. T., Joo, S. W., Shon, S. H., Kim, H., Kim, Y., & Lee, J. (2020). Diagnosing schizophrenia with network analysis and a machine learning method. *International Journal of Methods in Psychiatric Research*, *29*(1), 1–11. https://doi.org/10.1002/mpr.1818

Johannesen, J. K., Bi, J., Jiang, R., Kenney, J. G., & Chen, C.-M. A. (2016). Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric Electrophysiology*, *2*(1), 3. https://doi.org/10.1186/s40810-016-0017-0

Jørgensen, M., Mainz, J., Egstrup, K., & Johnsen, S. P. (2017). Quality of Care and Outcomes of Heart Failure Among Patients With Schizophrenia in Denmark. *The American Journal of Cardiology*, *120*(6), 980–985. https://doi.org/10.1016/j.amjcard.2017.06.027

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. *Jama*, *306*(15), 1688–1698. https://doi.org/10.1001/jama.2011.1515

Kendler, K. S. (2016). Phenomenology of Schizophrenia and the Representativeness of Modern Diagnostic Criteria. *JAMA Psychiatry*, *73*(10), 1082–1092. https://doi.org/10.1001/jamapsychiatry.2016.1976

Khan, U., Habibur Rahman, M., Salauddin Khan, M., Hossain, M. S., & Morsaline Billah, M. (2022). Bioinformatics and network-based approaches for determining pathways, signature molecules, and drug substances connected to genetic basis of schizophrenia etiology. *Brain Research*, *1785*, 147889. https://doi.org/10.1016/j.brainres.2022.147889

Kirkpatrick, B., Buchanan, R. W., Ross, D. E., & Carpenter Jr, W. T. (2001). A Separate Disease Within the Syndrome of Schizophrenia. *Archives of General Psychiatry*, *58*(2), 165–171. https://doi.org/10.1001/archpsyc.58.2.165

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kotzeva, A., Mittal, D., Desai, S., Judge, D., & Samanta, K. (2023). Socioeconomic burden of schizophrenia: a targeted literature review of types of costs and associated drivers across 10 countries. *Journal of Medical Economics*, *26*(1), 70–83. https://doi.org/10.1080/13696998.2022.2157596

Kovács, G., Almási, T., Millier, A., Toumi, M., Horváth, M., Kóczián, K., Götze, Kaló, Z., & Zemplényi, A. T. (2018). Direct healthcare cost of schizophrenia – European overview. *European Psychiatry*, *48*(1), 79–92. https://doi.org/10.1016/j.eurpsy.2017.10.008

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Lange, C., Deutschenbaur, L., Borgwardt, S., Lang, U. E., Walter, M., & Huber, C. G. (2017). Experimentally induced psychosocial stress in schizophrenia spectrum disorders: A systematic review. *Schizophrenia Research*, *182*, 4–12. https://doi.org/10.1016/j.schres.2016.10.008

Lee, J., Chon, M. W., Kim, H., Rathi, Y., Bouix, S., Shenton, M. E., & Kubicki, M. (2018). Diagnostic value of structural and diffusion imaging measures in schizophrenia. *NeuroImage: Clinical*, *18*, 467–474. https://doi.org/10.1016/j.nicl.2018.02.007

Lin, E., Lin, C.-H., & Lane, H.-Y. (2022). A bagging ensemble machine learning framework to predict overall cognitive function of schizophrenia patients with cognitive domains and tests. *Asian Journal of Psychiatry*, *69*, 103008. https://doi.org/10.1016/j.ajp.2022.103008

Loreto, M., Lisboa, T., & Moreira, V. P. (2020). Early prediction of ICU readmissions using classification algorithms. *Computers in Biology and Medicine*, *118*, 103636. https://doi.org/10.1016/j.compbiomed.2020.103636

Lorine, K., Goenjian, H., Kim, S., Steinberg, A. M., Schmidt, K., & Goenjian, A. K. (2015). Risk Factors Associated With Psychiatric Readmission. *Journal of Nervous & Mental Disease*, *203*(6), 425–430. https://doi.org/10.1097/NMD.0000000000000305

Lurie, I., Shoval, G., Hoshen, M., Balicer, R., Weiser, M., Weizman, A., & Krivoy, A. (2021). The association of medical resource utilization with physical morbidity and premature mortality among patients with schizophrenia: An historical prospective population cohort study. *Schizophrenia Research*, *237*, 62–68. https://doi.org/10.1016/j.schres.2021.08.019

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Manual de Procedimientos del Conjunto Mínimo Básico de Datos. Castilla y León. (2019).

McGrath, J., Saha, S., Chant, D., & Welham, J. (2008). Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. *Epidemiologic Reviews*, *30*(1), 67–76. https://doi.org/10.1093/epirev/mxn001

Mizrahi, R. (2016). Social Stress and Psychosis Risk: Common Neurochemical Substrates? *Neuropsychopharmacology*, *41*(3), 666–674. https://doi.org/10.1038/npp.2015.274

Morel, D., Yu, K. C., Liu-Ferrara, A., Caceres-Suriel, A. J., Kurtz, S. G., & Tabak, Y. P. (2020). Predicting hospital readmission in patients with mental or substance use disorders: A machine learning approach. *International Journal of Medical Informatics*, *139*, 104136. https://doi.org/10.1016/j.ijmedinf.2020.104136

National Health System Annual Report 2020-2021. (2022). https://www.sanidad.gob.es/estadEstudios/estadisticas/sisInfSanSNS/tablasEstadisticas/InfAnSNS.htm

National Institute of Statistics. (2020). Hospital Morbidity Survey-2020. https://www.ine.es/

Neto, C., Senra, F., Leite, J., Rei, N., Rodrigues, R., Ferreira, D., & Machado, J. (2021). Different Scenarios for the Prediction of Hospital Readmission of Diabetic Patients. *Journal of Medical Systems*, *45*(1), 11. https://doi.org/10.1007/s10916-020-01686-4

Olfson, M., Gerhard, T., Huang, C., Crystal, S., & Stroup, T. S. (2015). Premature Mortality Among Adults With Schizophrenia in the United States. *JAMA Psychiatry*, *72*(12), 1172. https://doi.org/10.1001/jamapsychiatry.2015.1737

Orrico-Sánchez, A., López-Lacort, M., Muñoz-Quiles, C., Sanfélix-Gimeno, G., & Díez-Domingo, J. (2020). Epidemiology of schizophrenia and its management over 8-years period using real-world data in Spain. *BMC Psychiatry*, *20*(1), 149. https://doi.org/10.1186/s12888-020-02538-8

Pachange, S., Joglekar, B., & Kulkarni, P. (2015). An Ensemble classifier approach for Disease Diagnosis using Random Forest. *Annual IEEE India Conference (INDICON)*, 1–5.

Pirooznia, M., Seifuddin, F., Judy, J., Mahon, P., Potash, J., & Zandi, P. (2012). Data mining approaches for genome-wide association of mood disorders. *Psychiatric Genetics*, *22*(2), 55–61. https://doi.org/10.1097/YPG.0b013e32834dc40d.Data

Portela, R., Wainberg, M. L., Castel, S., de Oliveira, H. N., & Ruas, C. M. (2022). Risk factors associated with readmissions of patients with severe mental disorders under treatment with antipsychotics. *BMC Psychiatry*, *22*(1), 189. https://doi.org/10.1186/s12888-022-03794-6

Rantala, M. J., Luoto, S., Borráz-León, J. I., & Krams, I. (2022). Schizophrenia: The new etiological synthesis. *Neuroscience & Biobehavioral Reviews*, *142*, 104894. https://doi.org/10.1016/j.neubiorev.2022.104894

Rathbun, T. F., Rogers, S. K., DeSimio, M. P., & Oxley, M. E. (1997). MLP iterative construction algorithm. *Neurocomputing*, *17*(3–4), 195–216. https://doi.org/10.1016/S0925-2312(97)00054-4

Registro de altas-CMBD estatal. Manual de definiciones y glosario de términos. (2017).

Rojas, J. C., Carey, K. A., Edelson, D. P., Venable, L. R., Howell, M. D., & Churpek, M. M. (2018). Predicting Intensive Care Unit Readmission with Machine Learning Using Electronic Health Record Data. *Annals of the American Thoracic Society*, *15*(7), 846–853. https://doi.org/10.1513/AnnalsATS.201710-787OC

Rozin, E., Vanaharam, V., D'Mello, D., Palazzolo, S., & Adams, C. (2019). A retrospective study of the role of long-acting injectable antipsychotics in preventing rehospitalization in early psychosis with cannabis use. *Addictive Behaviors Reports*, *10*, 100221. https://doi.org/10.1016/j.abrep.2019.100221

Ruetsch, C., Un, H., & Waters, H. C. (2018). Claims-based proxies of patient instability among commercially insured adults with schizophrenia. *ClinicoEconomics and Outcomes Research*, *Volume 10*, 259–267. https://doi.org/10.2147/CEOR.S149519

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, *6*(10), e921. https://doi.org/10.1038/tp.2015.182

Salem, H., Ruiz, A., Hernandez, S., Wahid, K., Cao, F., Karnes, B., Beasley, S., Sanches, M., Ashtari, E., & Pigott, T. (2019). Borderline personality features in inpatients with bipolar disorder: Impact on course and machine learning model use to predict rapid readmission. *Journal of Psychiatric Practice*, *25*(4), 279–289. https://doi.org/10.1097/PRA.0000000000000392

Santos-Mayo, L., San-Jose-Revuelta, L. M., & Arribas, J. I. (2017). A computer-aided diagnosis system with EEG based on the p3b wave during an auditory odd-ball task in schizophrenia. *IEEE Transactions on Biomedical Engineering*, *64*(2), 395–407. https://doi.org/10.1109/TBME.2016.2558824

Schaefer, J., Giangrande, E., Weinberger, D. R., & Dickinson, D. (2013). The global cognitive impairment in schizophrenia: Consistent over decades and around the world. *Schizophrenia Research*, *150*(1), 42–50. https://doi.org/10.1016/j.schres.2013.07.009

Shadmi, E., Gelkopf, M., Garber-Epstein, P., Baloush-Kleinman, V., Doudai, R., & Roe, D. (2018). Routine patient reported outcomes as predictors of psychiatric rehospitalization. *Schizophrenia Research*, *192*, 119–123. https://doi.org/10.1016/j.schres.2017.04.049

Shim, M., Hwang, H. J., Kim, D. W., Lee, S. H., & Im, C. H. (2016). Machine learning based diagnosis of schizophrenia using combined sensor level and source level EEG features. *Schizophrenia Research*, *176*(2–3), 314–319. https://doi.org/10.1016/j.schres.2016.05.007

Shiny. (2017). RStudio. https://shiny.rstudio.com/

Situación de la salud mental en España. Psiquiatría Clínica. (2017).

Steeg, S., Emsley, R., Carr, M., Cooper, J., & Kapur, N. (2018). Routine hospital management of self-harm and risk of further self-harm: propensity score analysis using

record-based cohort data. *Psychological Medicine*, *48*(2), 315–326. https://doi.org/10.1017/S0033291717001702

Sugisawa, S., Kurihara, T., Nakano, Y., Tsuneoka, T., Koya, H., Nagai, T., Ikeda, T., Fujisawa, N., Inamoto, A., & Iwanami, A. (2022). Risk factors for readmission in schizophrenia treated with combined psychoeducation and standard therapy. *Neuropsychopharmacology Reports*, *42*(1), 77–83. https://doi.org/10.1002/npr2.12229

Tai, A. M. Y., Albuquerque, A., Carmona, N. E., Subramanieapillai, M., Cha, D. S., Sheko, M., Lee, Y., Mansur, R., & McIntyre, R. S. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artificial Intelligence in Medicine*, *99*, 101704. https://doi.org/10.1016/j.artmed.2019.101704

Teixeira, J., Alexandre, S., Cunha, C., Raposo, F., & Costa, J. P. (2022). Impact of clozapine as the mainstay therapeutical approach to schizophrenia and substance use disorder: A retrospective inpatient analysis. *Psychiatry Research Communications*, *2*(3), 100056. https://doi.org/10.1016/j.psycom.2022.100056

Thomsen, K. R., Thylstrup, B., Pedersen, M. M., Pedersen, M. U., Simonsen, E., & Hesse, M. (2018). Drug-related predictors of readmission for schizophrenia among patients admitted to treatment for drug use disorders. *Schizophrenia Research*, *195*, 495–500. https://doi.org/10.1016/j.schres.2017.09.026

Thongkam, J., & Sukmak, V. (2014). Enhancing decision tree with adaboost for predicting schizophrenia readmission. *Advanced Materials Research*, *931*, 1467–1471. https://doi.org/10.4028/www.scientific.net/AMR.931-932.1467

Tong, L., Erdmann, C., Daldalian, M., Li, J., & Esposito, T. (2016). Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. *BMC Medical Research Methodology*, *16*(1), 26. https://doi.org/10.1186/s12874-016-0128-0

Tsoi, D. T. Y., Hunter, M. D., & Woodruff, P. W. R. (2008). History, aetiology, and symptomatology of schizophrenia. *Psychiatry*, *7*(10), 404–409. https://doi.org/10.1016/j.mppsy.2008.07.010

van Mens, K., Elzinga, E., Nielen, M., Lokkerbol, J., Poortvliet, R., Donker, G., Heins, M., Korevaar, J., Dückers, M., Aussems, C., Helbich, M., Tiemens, B., Gilissen, R., Beekman, A., & de Beurs, D. (2020). Applying machine learning on health record data from general practitioners to predict suicidality. *Internet Interventions*, *21*, 100337. https://doi.org/10.1016/j.invent.2020.100337

Veronese, E., Castellani, U., Peruzzo, D., Bellani, M., & Brambilla, P. (2013). Machine learning approaches: From theory to application in schizophrenia. *Computational and Mathematical Methods in Medicine*, *2013*. https://doi.org/10.1155/2013/867924

Vieira, S., Lopez Pinaya, W. H., & Mechelli, A. (2020). Introduction to machine learning. In *Machine Learning* (pp. 1–20). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00001-8

Wang, K. Z., Bani-Fatemi, A., Adanty, C., Harripaul, R., Griffiths, J., Kolla, N., Gerretsen, P., Graff, A., & De Luca, V. (2020). Prediction of physical violence in schizophrenia with machine learning algorithms. *Psychiatry Research*, *289*, 112960. https://doi.org/10.1016/j.psychres.2020.112960

Wolff, P., Graña, M., Ríos, S. A., & Yarza, M. B. (2019). Machine Learning Readmission Risk Modeling: A Pediatric Case Study. *BioMed Research International*, 1–9. https://doi.org/10.1155/2019/8532892

World Health Organization-Schizophrenia. (2022). https://www.who.int/news-room/fact-sheets/detail/schizophrenia

Xue, Y., Liang, H., Norbury, J., Gillis, R., & Killingworth, B. (2018). Predicting the risk of acute care readmissions among rehabilitation inpatients: A machine learning approach. *Journal of Biomedical Informatics*, *86*, 143–148. https://doi.org/10.1016/j.jbi.2018.09.009

Ying, Y., Jia, L., Wang, Z., Jiang, W., Zhang, J., Wang, H., Yang, N., Wang, R., Ren, Y., Gao, F., Ma, X., Tang, Y., & McDonald, W. M. (2021). Electroconvulsive therapy is associated with lower readmission rates in patients with schizophrenia. *Brain Stimulation*, *14*(4), 913–921. https://doi.org/10.1016/j.brs.2021.05.010

Zhao, P., & Yoo, I. (2021). Potentially modifiable risk factors for 30-day unplanned hospital readmission preventive intervention—A data mining and statistical analysis. *Health Informatics Journal*, *27*(1). https://doi.org/10.1177/1460458221995231

Zhao, Y., Hoenig, J. M., Protacio, A., Lim, S., & Norman, C. C. (2020). Identification of risk factors for early psychiatric rehospitalization. *Psychiatry Research*, *285*, 112803. https://doi.org/10.1016/j.psychres.2020.112803

Zhao, Z., Zhang, X., Li, W., Hu, X., Qu, X., Cao, X., Liu, Y., & Lu, J. (2019). Applying Machine Learning to Identify Autism with Restricted Kinematic Features. *IEEE Access*, *7*, 157614–157622. https://doi.org/10.1109/ACCESS.2019.2950030

Zhu, L., Wu, X., Xu, B., Zhao, Z., Yang, J., Long, J., & Su, L. (2021). The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood. *Neuroscience Letters*, *745*, 135596. https://doi.org/10.1016/j.neulet.2020.135596

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://doi.org/10.1093/clinchem/39.4.561

La esquizofrenia es un trastorno mental grave que se caracteriza por síntomas como las alucinaciones, delirios, trastornos del pensamiento y la conducta. Las personas con esquizofrenia se asocian con un mayor riesgo de abuso de sustancias, suicidio y mortalidad en comparación con la población general. Presentan tasas de hospitalización de un 20-40% en un año, lo que deriva en altos costes en el sistema sanitario y afecta la calidad de vida de los pacientes y los familiares. En España, la estancia hospitalaria corresponde al 37.6% de los costes sanitarios totales.

El uso de técnicas de Machine Learning (ML), permite analizar patrones de los datos mediante métodos estadísticos, y crear modelos que aprenden y generalizan el comportamiento de los datos. En Castilla y León (CyL), reducir el número de hospitalizaciones y de reingresos es de suma importancia para los servicios de psiquiatría. En consecuencia, esta Tesis Doctoral está enfocada en desarrollar y evaluar nuevos modelos predictivos utilizando algoritmos de ML, con el fin de ayudar en la predicción de hospitalizaciones y reingresos de pacientes con esquizofrenia en CyL.Para alcanzar este objetivo, se utilizaron 11 126 registros administrativos que corresponden a 5 412 pacientes hospitalizados con esquizofrenia, de 11 hospitales públicos de CyL, en dos períodos de tiempo diferentes.

Los resultados obtenidos en esta Tesis Doctoral sugieren que algoritmos de ML como RF, tienen la capacidad de aprender características complejas de los datos y predecir el riesgo de reingreso de pacientes hospitalizados con esquizofrenia, en CyL. Se considera que los modelos desarrollados pueden ayudar a la toma de decisiones, mejorando la calidad de la atención al paciente y desarrollando tratamientos preventivos en función de reducir el número de hospitalizaciones. Además, la implementación de la aplicación web desarrollada en esta investigación, en los hospitales públicos de CyL, puede ser de gran utilidad al personal sanitario en función de reducir los costos asociados a estas hospitalizaciones.

TESIS DOCTORAL
MENCIÓN INTERNACIONAL

UVa