UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRICA

MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE EXTENSIÓN DEL PERIODO DE HOSPITALIZACIÓN EN PACIENTES CARDIACOS

POR

José Miguel Orellana Melo

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción
para optar al grado académico de Ingeniero Civil Biomédico

Profesor Guía
Rosa Figueroa Iturrieta

Comisión
Álvaro Saldaña V.
Pablo Aquevque N.

Enero 2026
Concepción
(Chile)

Agradecimientos

A mi profesora guía, por su orientación durante el desarrollo de esta Memoria de Título.

A mis amigos, que me acompañaron durante mi formación en la Universidad de Concepción.

Y a mi familia, por ser mi motivación y brindarme siempre su apoyo.

Resumen

El tiempo de estancia hospitalaria constituye un indicador clave para la gestión clínica, al estar directamente asociado al uso de recursos, los costos de atención y la calidad asistencial. En el contexto de las enfermedades cardiovasculares (una de las principales causas de hospitalización en población adulta mayor), la predicción temprana de estancias hospitalarias prolongadas resulta especialmente relevante para la planificación de camas y la optimización de la atención en sistemas de salud.

Esta Memoria de Título tiene como objetivo desarrollar y evaluar un enfoque metodológico para la predicción de estancias hospitalarias prolongadas en pacientes cardíacos, utilizando datos del *Cardiovascular Health Study* (CHS). El conjunto de datos integra información demográfica, signos vitales, resultados de laboratorio y el historial clínico del paciente, incluyendo códigos de diagnóstico y procedimientos codificados según el estándar CIE-9.

La metodología propuesta se estructura en varias etapas. En primer lugar, se realiza la preparación del set de datos y la definición de la variable objetivo *Length of Stay* (*LOS*), formulando el problema como una tarea de clasificación binaria. Posteriormente, se implementa un pipeline de preprocesamiento diferenciado para variables numéricas y categóricas, incorporando técnicas de imputación múltiple, tratamiento de valores extremos, transformaciones de distribución y estandarización. Para abordar la alta dimensionalidad de los códigos clínicos, se propone un modelado basado en tópicos latentes mediante factorización de matrices no negativas (NMF), que permite condensar la historia clínica del paciente en un conjunto reducido de variables interpretables. Adicionalmente, se diseñan variables derivadas orientadas a capturar la carga diagnóstica y la recurrencia hospitalaria.

El entrenamiento de los modelos se realiza mediante una comparación de distintos modelos de aprendizaje supervisado, utilizando validación cruzada estratificada y el F1-score como métrica principal. A partir de esta etapa, se selecciona y se optimiza un modelo *XGBoost* mediante una búsqueda aleatoria de hiperparámetros, cuyo desempeño final se evalúa sobre un conjunto de prueba independiente, obteniendo un valor AUC de 0,78 y exactitud de 0,74. El análisis se complementó con técnicas de interpretabilidad basadas en la importancia de características y valores SHAP.

Abstract

The Length of hospital stay is a key indicator in clinical management, as it is directly associated with resource utilization, healthcare costs, and quality of care. In the context of cardiovascular diseases (one of the leading causes of hospitalization among the older adult population), early prediction of prolonged hospital stays is particularly relevant for bed capacity planning and optimization of healthcare delivery.

This project aims to develop and evaluate a method for predicting prolonged hospital stays in cardiac patients using the Cardiovascular Health Study (CHS) dataset. CHS integrates demographic information, vital signs, laboratory results, and patients' clinical histories, including diagnosis and procedure codes encoded according to the ICD-9 standard.

The proposed methodology is structured in several stages. First, the dataset is prepared, and the target variable, Length of Stay (LOS), was defined, thereby formulating the problem as a binary classification task. Subsequently, a differentiated preprocessing pipeline was implemented for numerical and categorical variables. These pipelines incorporated multiple imputation techniques, outlier handling, distribution transformations, and standardization depending on the variable being preprocessed. The high dimensionality of clinical codes was addressed by using a latent topic modeling approach based on non-negative matrix factorization (NMF). NMF enabled patients' clinical histories to be summarized into a reduced set of interpretable variables. Additionally, derived features are designed to capture the diagnostic burden and recurrence of hospitalization.

The models are trained by comparing different supervised learning models, using stratified cross-validation and the F1-score as the main metric. From this stage, an *XGBoost* model is selected and optimized through a random search of hyperparameters, with its final performance evaluated on an independent test set, yielding an AUC value of 0.78 and an accuracy of 0.74. The analysis was complemented with interpretability techniques based on feature importance and SHAP values.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Capítulo 1. Introducción

1.1. Introducción General

Las enfermedades cardiovasculares son una de las principales causas de mortalidad a nivel mundial y tienen un impacto considerable en la salud de la población. En el año 2019, alrededor de 17,9 millones de defunciones fueron asociadas a alguna patología cardiovascular; esto representa el 32% de las muertes a escala mundial [1]. Particularmente en Chile, estas cifras alcanzaron las 102 defunciones por cada 100.000 habitantes en el año 2020 [2]. La atención a pacientes cardíacos implica altos costos, por ejemplo, se espera que la carga financiera causada por las enfermedades cardiovasculares supere los 20.000 millones de dólares en 2030 en algunos países [3]. Debido a esto, las enfermedades cardiovasculares constituyen una de las principales causas de utilización de los recursos sanitarios y aumento de costos asociados al sistema de salud. Luego, el poder contar con información a base de predicciones, como la estancia hospitalaria (LOS, por sus siglas en inglés *Length of Stay*), el reingreso y la mortalidad, se ha convertido en herramientas fundamentales para la gestión de costos y mejora de la eficiencia en la atención médica.

Particularmente, el tiempo de estancia hospitalaria se ha convertido en un parámetro crucial, tanto para referirse a costos de atención médica como para la calidad de atención de los pacientes. La estancia hospitalaria se define como el número de días que un paciente permanecerá en el hospital durante un único ingreso [4]. Una LOS prolongada requiere un alto consumo de recursos y genera mayores costos, impactando en la capacidad hospitalaria y pudiendo afectar la calidad de la atención [5]. Es por esto por lo que lograr predecirla de manera precisa es crítico para la gestión hospitalaria, la planificación de la capacidad de camas y la mejora de la eficiencia en la atención [6].

En el ámbito de la cardiología, estudios se centran frecuentemente en la predicción de LOS en admisiones por múltiples causas, como insuficiencia cardíaca, enfermedad coronaria y angina inestable, dada su alta relevancia e impacto en el uso de recursos [7]. Uno de los estudios más claros y amplios al respecto examinó la predicción de LOS por cualquier causa en cardiología, utilizando una variedad de modelos en 16.414 admisiones de un hospital en Arabia Saudita [3]. Además, factores como el diagnóstico de enfermedad arterial coronaria y ciertas categorías de medicamentos influyen significativamente en la LOS de pacientes cardíacos [5].

En este contexto, el set de datos del estudio de salud cardiovascular (CHS), constituye una fuente de datos útil para el desarrollo de un modelo capaz de predecir LOS en pacientes cardiacos, contando con la disponibilidad de una amplia variedad de parámetros de cada paciente, incluyendo

datos demográficos, exámenes clínicos y de laboratorio, códigos de diagnóstico y procedimientos durante la estancia, entre otros [8].

El presente trabajo tiene como objetivo desarrollar un modelo predictivo capaz de estimar la duración de la estancia hospitalaria en pacientes cardíacos mayores de 65 años, empleando el conjunto de datos CHS. Para ello, se propone una metodología basada en aprendizaje automático, orientada a evaluar y comparar distintos enfoques con el propósito de identificar aquel que ofrezca el mejor desempeño en la predicción de *LOS*.

Con todo esto se busca reforzar la importancia de la ciencia de datos como una herramienta fundamental y novedosa, con un alto impacto en la reducción de costos y mejora de la eficiencia en la atención médica.

1.2. Trabajos Previos

Esta sección presenta una revisión de trabajos previos relacionados con la identificación de factores de riesgo asociados a estancias hospitalarias prolongadas y con el desarrollo de modelos predictivos en contextos clínicos. La revisión bibliográfica abarca estudios estadísticos orientados a analizar los factores que influyen en la duración de la estancia, un artículo de revisión sistemática sobre la aplicación de técnicas de aprendizaje automático en este ámbito, así como diversos estudios de caso que implementan modelos predictivos basados en distintos enfoques metodológicos.

1.2.1 Factores asociados a estancias prolongadas en pacientes con falla cardiaca aguda

- *L. Arbeláez et al., "Factores de riesgo asociados a estancia hospitalaria prolongada en pacientes con falla cardiaca aguda", Rev. Colomb. Cardiol., vol. 28, nº 2, p. 6359, may 2022* [9].

Este artículo investiga los factores clínicos que influyen en la duración de la estancia hospitalaria en pacientes con falla cardiaca aguda (FCA), una de las principales causas de hospitalización en adultos mayores. En particular, el estudio se enfocó en identificar variables que permitan predecir estancias prolongadas, con el objetivo de optimizar el uso de recursos hospitalarios y diseñar estrategias diferenciadas de atención. El estudio se llevó a cabo en un hospital universitario de alta complejidad en Colombia, incluyendo a 776 pacientes adultos con diagnóstico confirmado de FCA. Se trató de un estudio retrospectivo, observacional y de cohorte, en el cual se analizaron variables demográficas, clínicas y de laboratorio. Se excluyeron pacientes con trasplantes, dispositivos

cardíacos, o remitidos dentro de las primeras 24 horas. La variable principal fue la estancia hospitalaria, evaluada en dos umbrales: más de 5 días y más de 10 días. Los datos se analizaron mediante regresión logística paso a paso para identificar los factores con mayor peso en la duración de la hospitalización.

Entre los resultados más relevantes, se identifican como factores significativamente asociados a estancias superiores a 5 días la edad avanzada, la troponina positiva, la hiperglucemia y niveles de albúmina inferiores a 3 g/dL. En el caso de estancias mayores a 10 días, se suman la presión arterial sistólica baja, la frecuencia cardíaca elevada y concentraciones de péptidos natriuréticos superiores a 500 pg/mL.

Uno de los aportes clave del estudio es que los factores asociados a estancias prolongadas, como biomarcadores cardíacos, parámetros vitales y perfil metabólico, son medibles al ingreso y permiten anticipar qué pacientes requerirán una estancia más larga. Esta información es útil para asignar camas, recursos clínicos, y definir si el paciente puede beneficiarse de alternativas como hospitalización en casa o unidades de corta estancia.

Este trabajo entrega información directamente aplicable al desarrollo de modelos predictivos de LOS hospitalaria. Además, valida la importancia de incorporar variables clínicas simples pero relevantes (edad, troponina, glucosa, presión arterial, etc.) como predictores robustos de LOS en pacientes cardiovasculares.

1.2.2 Factores asociados a la duración de la estancia en UCI

- *T. Peres, S. Hamacher, F. L. C. Oliveira, A. M. T. Thomé, y F. A. Bozza, "What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis", J. Crit. Care, vol. 60, pp. 183–194, dic. 2020* [10].

En este artículo se realizó una revisión sistemática y un metaanálisis de la literatura con el objetivo de identificar los factores que predicen la duración de la estancia en la Unidad de Cuidados Intensivos (UCI) en población adulta. Se consideraron un total de 113 estudios, de los cuales 28 fueron incluidos en el metaanálisis. El estudio realizó un análisis cuantitativo en el cual se utilizaron modelos de efectos aleatorios, midiendo tamaños de efecto mediante correlación parcial o diferencia de medias estandarizada (d de Cohen), dependiendo del tipo de estadística reportada en los estudios primarios.

Entre los factores más consistentemente asociados con estancias prolongadas se destacan las puntuaciones de gravedad como APACHE, SOFA o SAPS, las cuales mostraron una relación positiva con la LOS. Asimismo, variables clínicas como la ventilación mecánica, hipomagnesemia, delirio,

desnutrición e infecciones agudas (por ejemplo, sepsis o enfermedades parasitarias) fueron factores significativos en la mayoría de los estudios analizados. Otras variables relevantes incluyeron una baja relación PaO_2/FiO_2, valores elevados de RDW (amplitud de distribución eritrocitaria), la presencia de trauma, el diagnóstico de enfermedades cardiovasculares o pulmonares crónicas, la readmisión a la UCI y parámetros bioquímicos como la temperatura corporal elevada o niveles aumentados de MR-proANP (péptido natriurético auricular medio).

El metaanálisis demostró que muchos de estos factores, especialmente los relacionados con la severidad del cuadro clínico y la necesidad de intervenciones invasivas, tienen una influencia clara en la prolongación del tiempo de hospitalización en UCI. En contraste, variables como la edad o el sexo no mostraron asociaciones consistentes.

Entre las principales contribuciones de este trabajo se destaca la provisión de una lista consolidada de variables relevantes para el desarrollo de modelos predictivos de LOS. Esto la convierte en una referencia importante al momento de definir qué variables clínicas y demográficas podrían considerarse en la construcción de los modelos, especialmente en pacientes críticos como los que presentan patologías cardíacas.

1.2.3 Aprendizaje automático en la predicción de estancia hospitalaria

- *S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, y S. Koblar, "Machine learning in the prediction of medical inpatient length of stay", Intern. Med. J., vol. 52, nº 2, pp. 176–185, feb. 2022* [7].

En este trabajo se realizó una revisión sistemática de la literatura, explorando el uso de aprendizaje automático en la predicción de *LOS* en pacientes hospitalizados. El artículo consideró el análisis y evaluación de 21 estudios, destacando una variabilidad considerable en las metodologías utilizadas y resultados obtenidos. Se diferencia entre dos tipos de modelos de predicción*:* tareas de clasificación, donde la predicción sitúa a los individuos en categorías (como predecir la *LOS* mayor a 7 días o menor a 7 días) y tareas de regresión*,* si es que se predice un resultado continuo (por ejemplo, predecir la LOS como el número real de días que un paciente estará en el hospital). Cada tipo de modelo tiene sus propias métricas de rendimiento. Los estudios de clasificación suelen presentar métricas como la *exactitud*, el *valor predictivo positivo* y el *valor predictivo negativo*, así como métricas independientes de la prevalencia, tales como el AUC (del inglés *area under the receiver operator curve*), la *sensibilidad* y la *especificidad*. Por otra parte, los estudios de regresión presentan

métricas de rendimiento como el *error absoluto medio*, el *error cuadrático medio*, la *raíz del error cuadrático medio* y R^2.

Los modelos de aprendizaje automático empleados incluyeron máquinas de vectores de soporte (SVM), redes neuronales artificiales (ANN), redes bayesianas (BN), algoritmos de árboles de decisión (DT), algoritmos de bosques aleatorios (RF) y modelos de regresión logística (LR). Estos modelos se aplicaron típicamente a datos recopilados dentro de las primeras 12 a 48 horas de la admisión para hacer las predicciones. La mayoría de los estudios usaron una combinación de datos demográficos, administrativos, clínicos, de laboratorio y de tratamiento. Sin embargo, algunos tipos de datos menos frecuentes como el uso de imágenes médicas (por ejemplo, de radiología) y de lenguaje natural (por ejemplo, de las notas del paciente) se destacaron como opciones prometedoras para futuras investigaciones. El estudio también destaca la importancia tanto de especificar un enfoque particular para el manejo los datos faltantes, como de proporcionar información clara sobre la distribución de la LOS en los conjuntos de entrenamiento y prueba del modelo que se está evaluando para interpretar adecuadamente las métricas de rendimiento. También se destacó la importancia de validar externa y prospectivamente los modelos en nuevos conjuntos de datos, con el fin de asegurar la utilidad clínica real del modelo y la importancia de realizar estudios de impacto/implementación de los modelos en el mundo real, necesarios para demostrar un beneficio para el paciente o el sistema de salud. Por último, se menciona la posibilidad de generar predicciones continuas de *LOS* recurrentemente durante la hospitalización, en lugar de solo al inicio de la admisión, como una posible mejora para el rendimiento de los modelos.

En general, se destaca la predicción de la *LOS* en pacientes hospitalizados mediante aprendizaje automático como un área prometedora en términos de la práctica clínica actual y se enfatiza en la necesidad de estandarizar la metodología y mejorar la presentación de datos demográficos y clínicos para futuras investigaciones. Especificar el tratamiento de los datos faltantes y la distribución de la variable que se predice para mejorar la interpretación de las métricas de rendimiento. Así como la utilidad de utilizar combinaciones de datos como edad, género, estado del seguro, admisión en fin de semana, signos vitales, comorbilidades, datos de laboratorio (por ejemplo, niveles de creatinina, hemoglobina, entre otros) y medicamentos recetados, entre otros.

1.2.4 Predicción de tiempo de estadía en pacientes cardiacos

- *T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, y M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: A machine learning approach", Int. J. Cardiol., vol. 288, pp. 140–147, ago. 2019* [3].

Este trabajo se centró en el desarrollo de un modelo basado en aprendizaje automático para predecir la LOS hospitalaria en pacientes cardíacos utilizando datos de pacientes adultos admitidos entre 2008 y 2016 en el King Abdulaziz Cardiac Center (KACC) en Riyadh, Arabia Saudita. El conjunto de datos incluyó registros médicos electrónicos de 16.414 visitas hospitalarias consecutivas para 12.769 pacientes únicos. De los registros de pacientes cardíacos se extrajeron 70 atributos, entre los cuales se incluían información demográfica, factores de riesgo cardiovascular, signos vitales al ingreso, resultados de pruebas de laboratorio frecuentes y criterios de admisión.

La variable *LOS* se discretizó en tres categorías principales mediante agrupación de igual frecuencia: corta (0 a 2 días), intermedia (3 a 5 días) y larga (más de 5 días). En el análisis final de las 16.414 visitas, la distribución de las clases de LOS fue de 30,85 % para la categoría corta, 33,45 % para la intermedia y 35,71 % para la larga

Para seleccionar los atributos más relevantes para la predicción, se aplicó el algoritmo de ganancia de información (*Information Gain*), identificando un total de 21 variables principales. Posteriormente, se evaluaron cuatro técnicas de clasificación: *Random Forest*, Redes Neuronales Artificiales, Support Vector Machine y Redes Bayesianas, empleando validación cruzada de 10 pliegues (*10-fold cross-validation*). Las métricas de evaluación utilizadas incluyeron sensibilidad, precisión, F-score, AUROC y RMSE.

Lo que más se destaca es la determinación de los cinco atributos con mayor impacto en la predicción de la *LOS*: la frecuencia cardíaca al ingreso, la presión arterial sistólica y diastólica al ingreso, la edad y el estado del seguro. Otros atributos importantes seleccionados incluyeron género, área de superficie corporal, historial de diabetes, hipertensión, dislipidemia, tabaquismo, obesidad, creatinina sérica, lipoproteína de alta densidad, fracción de eyección, insuficiencia cardíaca congestiva, síndrome coronario agudo, infarto agudo de miocardio, insuficiencia renal, angina inestable, admisión estacional y experiencia del médico, tal como se observa en la Figura 1.

El estudio demostró que los métodos de aprendizaje automático, particularmente el modelo *Random Forest*, proporcionan una predicción precisa de la *LOS* para pacientes cardíacos utilizando datos disponibles al momento de la admisión. Se identificaron atributos clave como la frecuencia

cardíaca, la presión arterial sistólica y diastólica al ingreso, la edad y el estado del seguro como los de mayor influencia en la *LOS*.

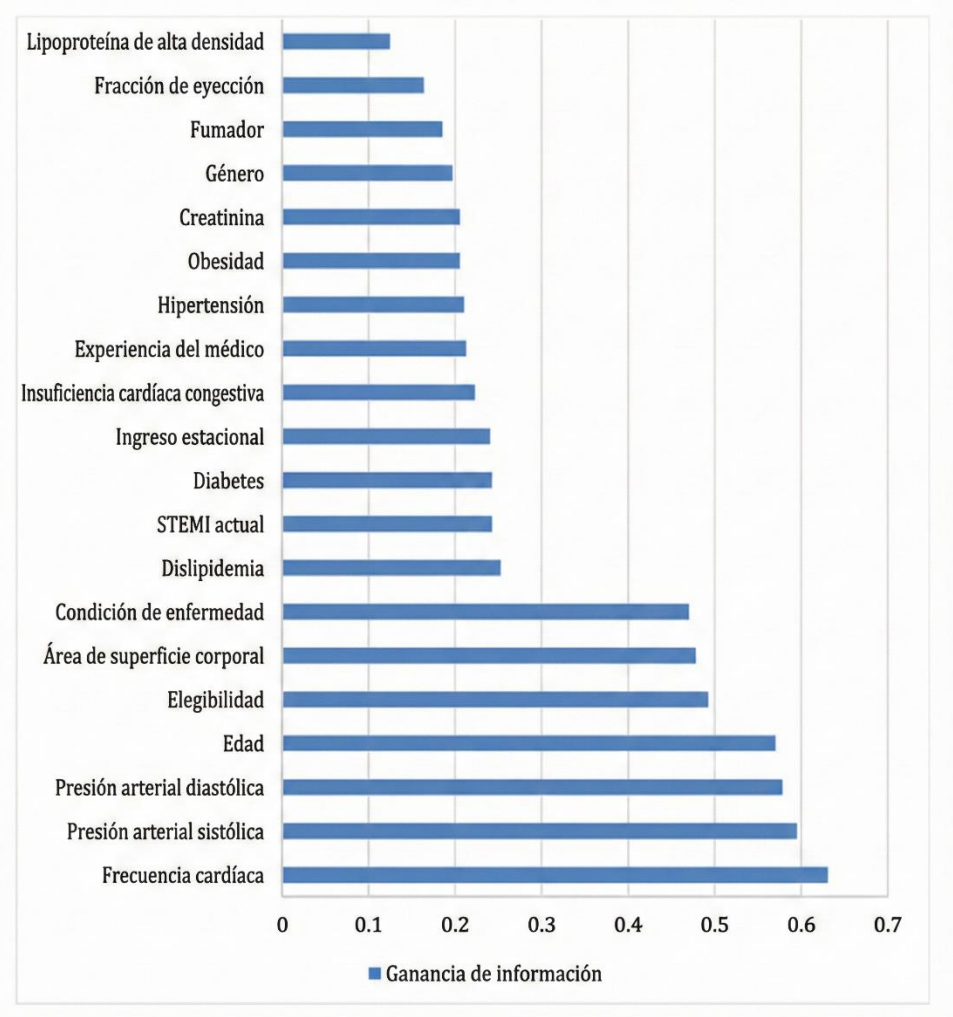

Figura 1. Características seleccionadas para construir el modelo de predicción de LOS, según el artículo [3].

1.2.5 Predicción de la duración de estancia en la UCI con aprendizaje automático

- *L. Hempel, S. Sadeghi, y T. Kirsten, "Prediction of Intensive Care Unit Length of Stay in the MIMIC-IV Dataset", Appl. Sci., vol. 13, nº 12, p. 6930, jun. 2023* [11].

En este estudio, se desarrollaron modelos predictivos de *LOS* en la UCI utilizando aprendizaje automático y datos clínicos disponibles durante las primeras 24 horas de hospitalización. Se utilizó el conjunto de datos MIMIC-IV, con una cohorte de 41.473 ingresos de pacientes adultos, excluyendo casos extremos o fallecidos.

Para desarrollar los modelos, se usaron variables demográficas, administrativas, diagnósticos médicos (agrupados según capítulos CIE-10), registros de signos vitales y exámenes de laboratorio. Para la clasificación, la LOS se binarizó en corta (menor a 4 días) y larga (mayor o igual a 4 días), y se entrenaron modelos de Regresión Logística, SVM, *Random Forest* y *XGBoost*. El proceso de evaluación incluyó métricas como F1, AUC, RMSE, MAE y R^2, utilizando validación cruzada con múltiples repeticiones.

En cuanto a los modelos implementados, *Random Forest* obtuvo los mejores resultados en clasificación, aunque con dificultades para detectar estancias largas. En regresión, el rendimiento general fue limitado debido al uso exclusivo de datos del primer día, mejorando al enfocarse solo en estancias cortas. Un enfoque por etapas (clasificación previa en estancias cortas y largas, seguida de regresión) mejoró significativamente la precisión, destacando SVM y *Random Forest*. Además, el estudio realizo un análisis de correlación entre características individuales y la LOS, para determinar como una sola característica afecta a la LOS y como se relacionan entre ellas, destacando la Escala de Coma de Glasgow (GCS) y signos vitales como la temperatura corporal y la frecuencia respiratoria. Se destacaron las 10 características con mayor porcentaje de importancia para la predicción, según el modelo *XGBoost*, tal como se observa en la Figura 2. La GCS verbal mostró la mayor importancia de rendimiento, en consonancia con su fuerte correlación con la LOS.

El estudio concluye que predecir LOS solo con datos tempranos es desafiante, y sugiere incorporar información de días posteriores para mejorar la predicción de estancias prolongadas. Como trabajo futuro, se propone explorar el uso de aprendizaje profundo y el análisis más detallado de los diagnósticos.

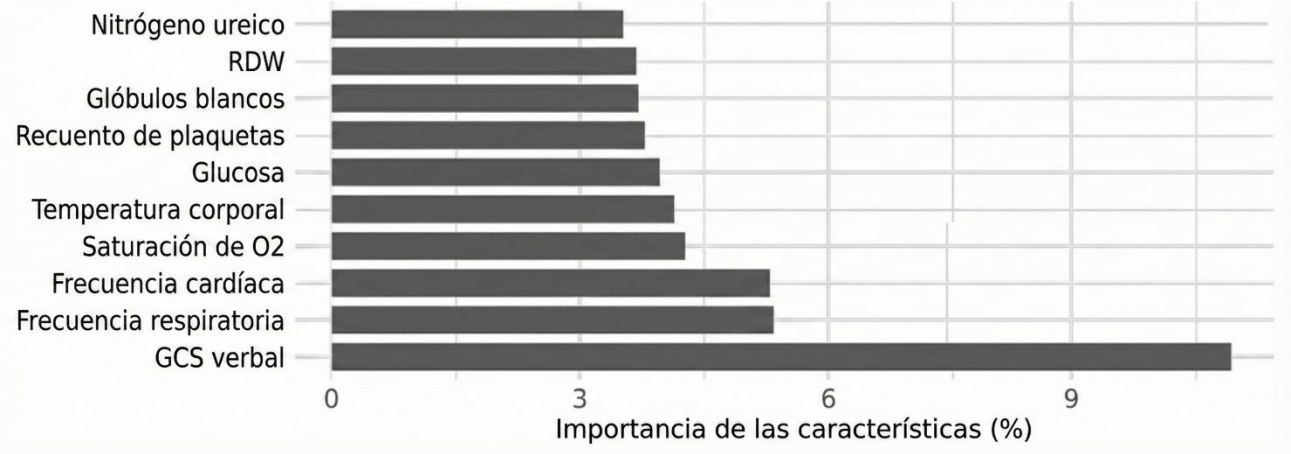

Figura 2. Importancia de características del modelo XGBoost utilizado para la clasificación en el paper [11].

1.2.6 Predicción de la estancia hospitalaria con un modelo de clasificación de dos niveles

- *I. E. Livieris, T. Kotsilieris, I. Dimopoulos, y P. Pintelas, "Decision Support Software for Forecasting Patient's Length of Stay", Algorithms, vol. 11, nº 12, Art. nº 12, dic. 2018* [12].

En este trabajo se propone un sistema de apoyo a la decisión para predecir la duración de la LOS utilizando un modelo de clasificación de dos niveles. Tuvo como principal objetivo ofrecer una herramienta que permita estimar el tiempo de hospitalización antes de la admisión, lo cual es clave para la planificación hospitalaria, la asignación de recursos y la detección temprana de posibles complicaciones clínicas.

El estudio se basó en datos de 2.702 pacientes mayores de 65 años hospitalizados en el Hospital General de Kalamata (Grecia) entre los años 2008 y 2012. Luego de limpiar y preprocesar los datos eliminando duplicados, valores atípicos, y registros con *LOS* igual a cero; se trabajó con 13 atributos que abarcaron: factores demográficos (edad, sexo, tipo de seguro), geográficos (altitud, distancia al hospital, tipo de zona de residencia), clínicos (día y mes de ingreso, número de ingresos diarios, sala de hospitalización), y el diagnóstico según la clasificación internacional de enfermedades (CIE-10).

Los pacientes fueron clasificados en seis grupos según la duración de su estancia hospitalaria: 1, 2, 3, 4, 5 y más de 5 días. Dado el desbalance natural del conjunto de datos, los autores diseñaron un esquema de clasificación jerárquico para mejorar la precisión. En una primera etapa (clasificador de nivel A), el modelo diferencia entre tres grupos amplios: 1 a 2 días, 3 a 5 días y más de 5 días. En la segunda etapa, clasificadores adicionales (niveles B1 y B2) refinan la predicción dentro de esos grupos: B1 distingue entre 1 y 2 días, mientras que B2 diferencia entre 3, 4 y 5 días. Para cada nivel, se evaluaron distintos algoritmos de clasificación, entre ellos Random Forest (RF), Naive Bayes, k-Nearest Neighbors (k-NN), Multilayer Perceptron (MLP) y Sequential Minimal Optimization (SMO), un algoritmo para entrenar máquinas de soporte vectorial. Se utilizó validación cruzada estratificada de 10 pliegues para evaluar el rendimiento. La mejor combinación se obtuvo con RF en el nivel A, k-NN en el nivel B1 y nuevamente RF en el nivel B2, alcanzando una precisión global del 78,53%. Esta cifra superó significativamente al mejor modelo individual (RF plano), que logró un 72,34%.

El análisis detallado mostró que el enfoque de dos niveles mejora sustancialmente la precisión en casi todas las clases. Por ejemplo, clasificó correctamente al 84,7% de los pacientes con estancias

de un día (vs. 74% con RF simple), y al 79,6% en el caso de dos días (vs. 72%). También hubo mejoras para estancias más largas, como en el grupo de más de cinco días (79,2% vs. 76,1%).

Como conclusión, se destaca que el esquema jerárquico supera consistentemente a los clasificadores tradicionales, especialmente en conjuntos de datos desbalanceados como el utilizado. Además, señalan como líneas futuras la incorporación de nuevos atributos clínicos (como signos vitales o resultados de laboratorio al ingreso), mejoras en la interfaz, integración de métodos de balanceo y el desarrollo de un sistema explicativo que justifique las predicciones generadas.

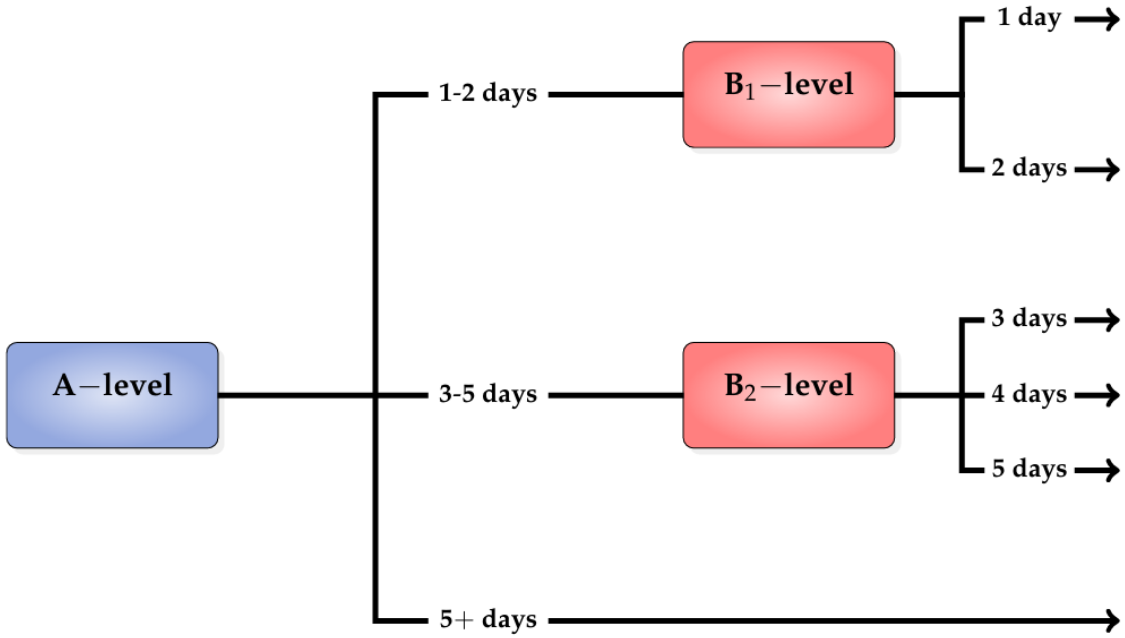

Figura 3. Esquema de clasificación en dos niveles [12].

1.2.7 Análisis de la duración de la estancia hospitalaria mediante registros médicos electrónicos: Un enfoque estadístico y de minería de datos

- *H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, y S. Yoo, "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach", PloS One, vol. 13, nº 4, 2018* [13].

Este estudio analizó factores asociados a la duración de la LOS utilizando registros electrónicos de salud (EHR) de 45.546 pacientes hospitalizados en un hospital universitario de Corea del Sur durante el año 2013. Su objetivo fue desarrollar una metodología para mejorar la gestión hospitalaria, reducir riesgos clínicos y costos, y optimizar el uso de recursos como camas y tratamientos. Para lograrlo, se empleó un marco de análisis dividido en tres fases: análisis descriptivo

y exploratorio, minería de procesos, y modelos predictivos. El análisis descriptivo identificó que la LOS tuvo una media de 7 días, una mediana de 4, y que aproximadamente el 3% de los pacientes permaneció hospitalizado más de 30 días. Diagnósticos como infartos cerebrales y trastornos del estado de ánimo se asociaron con estancias prolongadas, al igual que los pacientes transferidos a medicina de rehabilitación. Además, mediante minería de procesos se analizaron patrones de transferencia entre departamentos, encontrando que los pacientes transferidos tenían una LOS promedio 17 días mayor que los no transferidos. En la fase predictiva, se aplicaron modelos de Regresión Múltiple y *Random Forest*, identificando como variables más influyentes la frecuencia de transferencias, cirugías y diagnósticos, la gravedad del paciente y el tipo de seguro. El modelo de clasificación logró predecir con un 97,32% de precisión los casos de larga estancia, siendo la frecuencia de transferencia el predictor más importante.

El estudio destaca que, a diferencia de enfoques previos centrados en características inmutables del paciente, factores relacionados con el proceso hospitalario como tiempos de traslado, retrasos en el alta y complejidad de la atención ofrecen oportunidades concretas de mejora. Se propone que la optimización de rutas clínicas, una mejor coordinación entre servicios y la gestión racional de antibióticos podrían reducir la *LOS*, mejorar la recuperación funcional de los pacientes y disminuir los costos operacionales.

1.2.8 Modelo basado en Redes de Atención Profunda para Clasificación de LOS

- *G. Harerimana, J. W. Kim, y B. Jang, "A deep attention model to forecast the Length of Stay and the in-hospital mortality right on admission from ICD codes and demographic data", J. Biomed. Inform., vol. 118, p. 103778, jun. 2021*[14].

Este trabajo se centra en el desarrollo de un modelo avanzado basado en Redes de Atención Profunda (Deep Attention) denominado Hierarchical Attention Network for Length of Stay (HAN-LoS). Su objetivo principal es la clasificación temprana de la LOS, utilizando únicamente los datos disponibles del paciente en las primeras 24 horas de su admisión. Los datos utilizados en el estudio provienen de los registros médicos estructurados y de texto libre del conjunto de datos MIMIC-III.

El atributo de LOS fue discretizado en tres grupos principales para la clasificación: corta (menor a 10 días), media (de 10 a 30 días) y larga (mayor a 30 días). La distribución de las clases reveló un alto desequilibrio, con la clase de LOS larga representando solamente el 4,2% de las admisiones, lo cual requirió estrategias específicas de balanceo.

La principal contribución metodológica de este estudio radica en el uso de una arquitectura de atención jerárquica de dos niveles. Esta estructura permite al modelo ponderar automáticamente la importancia de los códigos clínicos específicos (CIE-9) dentro de una admisión (Atención a Nivel de Conceptos) y, posteriormente, ponderar la relevancia de las admisiones pasadas en el historial del paciente (Atención a Nivel de Paciente). Este enfoque busca activamente los patrones más influyentes para la LOS en la secuencia temporal de las interacciones del paciente con el sistema hospitalario.

Para la ingeniería de características y preprocesamiento, el estudio implementó una técnica avanzada de *Transfer Learning*. Se utilizaron meta-embeddings preentrenados (Cui2Vec y ClinicalBERT) para transformar los códigos médicos estructurados y el diagnóstico de texto libre en representaciones densas. Esta etapa se considera crítica para capturar las relaciones clínicas subyacentes. Dada la severidad del desequilibrio de clases, se aplicó la técnica SMOTE (Synthetic Minority Oversampling Technique) para balancear el conjunto de entrenamiento, lo cual fue fundamental para el rendimiento.

El rendimiento del modelo HAN-LoS (con SMOTE) se evaluó utilizando métricas robustas para conjuntos desequilibrados, como el AUROC y el F1-Score. Los resultados demostraron una mejora en el rendimiento sobre las líneas base existentes: el modelo alcanzó un AUROC Micro de 0,82 y un puntaje Micro-F1 de 0,24, superando significativamente a modelos baseline como la Red Neuronal Artificial sin mecanismos de atención.

Lo que más se destaca es que la superioridad del modelo no se debió solo a su profundidad, sino a la incorporación del mecanismo de atención, el cual aporta la transparencia necesaria para inspeccionar qué códigos CIE y admisiones históricas son los más cruciales para una *LOS* prolongada. El estudio demostró que la combinación de ingeniería de características avanzada (uso de meta-embeddings) con un manejo riguroso del desequilibrio (con SMOTE) es indispensable para lograr predicciones útiles en las clases minoritarias (*LOS* Prolongada).

1.2.9 Aprendizaje supervisado para la predicción de riesgo de mortalidad en personas mayores utilizando el set de datos CHS

- *J. P. Navarrete, J. Pinto, R. L. Figueroa, M. E. Lagos, Q. Zeng, y C. Taramasco, "Supervised Learning Algorithm for Predicting Mortality Risk in Older Adults Using Cardiovascular Health Study Dataset", Appl. Sci., vol. 12, no 22, p. 11536, nov. 2022* [15].

En este trabajo se presenta el desarrollo de un modelo de aprendizaje automático para predecir el riesgo de mortalidad en adultos mayores con enfermedades cardiovasculares y múltiples afecciones crónicas (MCC), utilizando la base de datos *Cardiovascular Health Study* (CHS). Su motivación surge

del aumento sostenido de personas con MCC, lo que representa un desafío para los sistemas de salud que suelen abordar enfermedades de forma individual. El objetivo fue crear una herramienta que apoye la toma de decisiones clínicas al identificar pacientes de alto riesgo de mortalidad.

Este estudio utilizó datos de 5.888 participantes del CHS, recolectados entre 1989 y 1999, que incluyen antecedentes médicos, mediciones clínicas, datos de laboratorio, electrocardiogramas, medicación y variables psicosociales como depresión. Se seleccionaron 123 variables clínicas relevantes mediante correlaciones estadísticas y conocimiento experto. Para capturar la evolución temporal, se construyeron dos matrices de características: una que resumía los ocho años de datos usando la moda de cada variable, y otra que agrupaba las modas cada tres años, permitiendo analizar los cambios en el tiempo. El preprocesamiento de los datos consistió en imputar valores faltantes con la moda, escalar las variables entre 0 y 1, y dividir los datos en conjunto de entrenamiento (80%) y prueba (20%). Se evaluaron tres modelos supervisados: *Random Forest*, *Support Vector Machine* (SVM) y Regresión Logística. El modelo *Random Forest* obtuvo los mejores resultados en todas las métricas: precisión, *F1-score*, *recall* y AUC. En particular, el uso de la matriz con modas trianuales mejoró significativamente el rendimiento de todos los modelos. *Random Forest* alcanzó un AUC de 0,80, SVM un AUC de 0,76 y regresión logística un AUC de 0,74. También se aplicó validación cruzada con 10 pliegues y análisis de componentes principales (PCA), lo que confirmó la estabilidad de los modelos y permitió mejorar levemente los resultados. *Random Forest* demostró una precisión cercana al 90% en el conjunto de prueba, junto con altos valores de *F1-score* y *recall*. Además, se identificaron las variables más importantes para la predicción, entre ellas: infarto de miocardio, angina, edad, insuficiencia cardíaca, cirugía de *bypass*, diabetes y debilidad.

El estudio concluye que los modelos supervisados, especialmente *Random Forest*, pueden predecir con buena precisión la mortalidad en adultos mayores con enfermedades cardiovasculares usando datos del CHS. Destaca además que incorporar información temporal (a través del resumen cada tres años) es clave para mejorar la capacidad predictiva.

Esta investigación es particularmente relevante para este trabajo, ya que demuestra el potencial del CHS como base de datos para desarrollar modelos predictivos en salud cardiovascular. Los métodos utilizados en este trabajo, como la construcción de matrices temporales y la validación cruzada, entregan una base metodológica útil para el desarrollo del modelo propuesto.

1.2.10 Discusión

De la revisión de la literatura, se observa que el problema de predicción de la LOS ha sido abordado en diferentes contextos, ya sea en pacientes cardiacos, en la UCI o en base a datos de EHR, entre otros. Sin embargo, muy pocos estudios se centran específicamente en pacientes con problemas cardiacos y comorbilidades, y ninguno de ellos aborda el problema de la predicción de LOS utilizando el set de datos CHS. Esto hace resaltar la necesidad de explorar más alternativas en el desarrollo de modelos predictivos en pacientes cardiacos, ya que, al ser un tema poco estudiado, enfocarnos en esto acota el problema y podría permitir un modelo más especializado, y mejor adaptado. Esto también permitirá encontrar patrones propios de datos asociados a pacientes cardiacos, que indudablemente pueden llegar a ser de utilidad en el contexto clínico y distintos a otros patrones que pudieran encontrarse en pacientes hospitalizados con otras patologías o en tipos atención más específicos (como la UCI).

De los estudios analizados también es posible determinar un conjunto de variables más incidentes en la predicción de la estadía, como la frecuencia cardíaca, la presión arterial sistólica y diastólica al ingreso, la edad avanzada, el estado del seguro, temperatura corporal y la frecuencia respiratoria y resultados de laboratorio como niveles de troponina, glucemia, albúmina y péptidos natriuréticos. Además, destaca el uso de algoritmos como *Random Forest*, *Support Vector Machine*, Regresion Logísticaca, *Naive Bayes*, k-NN, entre otros. Estos modelos tuvieron buenos resultados en cuanto a rendimiento y generalmente fueron evaluados con métricas como AUC, *recall*, exactitud y *F1-Score*. Uno de los estudios más destacados utilizó un enfoque por etapas para predecir LOS. Con esto se consigue separar mejor las predicciones de estancias largas de las cortas, destacando el esquema jerárquico como una mejora consistente respecto de los clasificadores tradicionales, especialmente en conjuntos de datos desbalanceados.

Este proyecto se centrará en el desarrollo de un modelo preciso y confiable capaz de predecir la LOS en pacientes cardiacos, para esto se considerarán algunas de las variables y modelos previamente mencionados junto a sus métricas, para determinar el que se adapte mejor a la base de datos de CHS.

1.3. Definición del Problema

La duración de la estancia hospitalaria es uno de los principales factores que determinan el costo de la atención hospitalaria [16]. En particular, los costos de hospitalización en pacientes con

afecciones cardiacas representan una carga económica importante tanto para los pacientes, como para los sistemas de atención sanitaria. Según estadísticas de egresos hospitalarios a nivel país en Chile, el 9,1% corresponden a pacientes ingresados por enfermedades del sistema circulatorio, entre las cuales se destacan las principales causas como enfermedades cerebrovasculares, enfermedades isquémicas del corazón e insuficiencia cardiaca [17], siendo estas causas las que afectan más en adultos mayores, ya que el 58% de los ingresados por diagnóstico de patologías cardiacas correspondían a adultos mayores de 65 años, con un promedio de LOS de 8,5 días de estadía [17]. Esto hace que se vuelva relevante contar con herramientas que ayuden a los equipos de salud a conocer los factores de riesgo y parámetros que afectan en el largo de estadía, para poder planificar el uso de camas y salas y poder gestionar mejores intervenciones que deban realizarse prioritariamente antes de que se presenten pérdidas monetarias importantes o se perjudique la calidad de atención del paciente.

Para poder determinar los factores de riesgo determinantes de una LOS prolongada se utilizará la base de datos Cardiovascular Health Study, el cual describe y cuantifica factores relacionados con la aparición de cardiopatías coronarias y accidente cardiovascular [8].

Para diseñar un modelo adecuado, se analizarán las variables de CHS, en busca de aquellas que tengan una alta influencia en la LOS. Para esto será clave analizar la base de datos en busca de variables que estén altamente relacionadas con la LOS, además de contrastar con estudios que determinen dichos factores.

1.4. Objetivos

1.4.1 Objetivo General

Desarrollar y validar un modelo predictivo basado en algoritmos de aprendizaje automático para la detección de estancia hospitalaria prolongada en pacientes cardíacos mayores de 65 años, utilizando el set de datos Cardiovascular Health Study.

1.4.2 Objetivos Específicos

- Formular una metodología respaldada por la literatura y sustentada en el conjunto de datos CHS, que permita el procesamiento riguroso de datos y la evaluación comparativa de modelos predictivos de estancia hospitalaria en pacientes cardíacos.

- Realizar un análisis exploratorio de la base de datos CHS, tendiente a conocer la información disponible, con el fin de identificar la presencia de las variables predictoras de *LOS* en pacientes cardiacos encontradas en la literatura y realizar los preprocesamientos necesarios.
- Desarrollar modelos para predecir *LOS* en pacientes cardiacos utilizando la base de datos CHS.
- Evaluar los modelos en términos de exactitud, AUROC y *Recall*, entre otras.

1.5. Alcances y Limitaciones

- El modelo será entrenado y probado con el set de datos CHS, disponible a través de autorización BioLINCC RMDA V02 1d20120806 y bajo autorización del comité de ética, bioética y bioseguridad de la Universidad de Concepción.
- La complejidad del modelo en cuanto a las variables y los algoritmos utilizados, estarán limitados por la capacidad computacional disponible.
- La efectividad del modelo puede verse disminuida por la disponibilidad y balance de los datos presentes en la base de datos CHS.
- Uno de los alcances de este trabajo consistirá en abordar el desbalance y ausencia de datos en ciertas variables de interés.

1.6. Metodología

Para el desarrollo del algoritmo de predicción de estancias hospitalarias prolongadas, se implementó y evaluó un conjunto de modelos de aprendizaje automático representativos de distintos enfoques de clasificación: *Random Forest*, Regresión Logística, *Support Vector Machine* (SVM) y *XGBoost*. El estudio se realizó utilizando un subconjunto del *Cardiovascular Health Study* (CHS), integrado por variables demográficas, clínicas, resultados de laboratorio y el historial médico completo de los pacientes, incluyendo diagnósticos y procedimientos codificados según el estándar CIE-9.

La metodología comienza con una etapa de análisis exploratorio de datos, orientada a caracterizar la estructura del conjunto de datos, identificar patrones relevantes y analizar la distribución de las variables. Posteriormente, se desarrolló un pipeline de preprocesamiento diferenciado para variables numéricas y categóricas, que incluyó limpieza de datos, imputación de valores faltantes, tratamiento de distribuciones asimétricas mediante transformaciones (Yeo-Johnson), estandarización con *StandardScaler* y manejo del desbalance de clases.

Dada la alta dimensionalidad asociada a los códigos clínicos, se incorporó una etapa de reducción de dimensionalidad mediante modelado de tópicos latentes basado en factorización de matrices no negativas (NMF), permitiendo representar el historial clínico de cada paciente en un espacio de menor dimensión y mayor interpretabilidad. Adicionalmente, se diseñaron variables derivadas para capturar información agregada sobre carga diagnóstica y recurrencia de hospitalizaciones.

La selección de características se abordó mediante un enfoque híbrido, combinando criterios estadísticos y métodos basados en importancia de variables obtenida a partir de modelos de ensamble. El entrenamiento y evaluación inicial de los modelos se realizó utilizando validación cruzada estratificada, empleando el F1-score como métrica principal debido al desbalance presente en la variable objetivo. A partir de esta comparación, el modelo *XGBoost* fue seleccionado y sometido a un proceso de optimización de hiperparámetros mediante búsqueda aleatoria. Finalmente, el desempeño del modelo optimizado se evaluó sobre un conjunto de prueba independiente, complementando las métricas de desempeño con análisis de interpretabilidad basados en importancia de características y valores SHAP.

En la Figura 4 se detalla un esquema de la metodología mencionada.

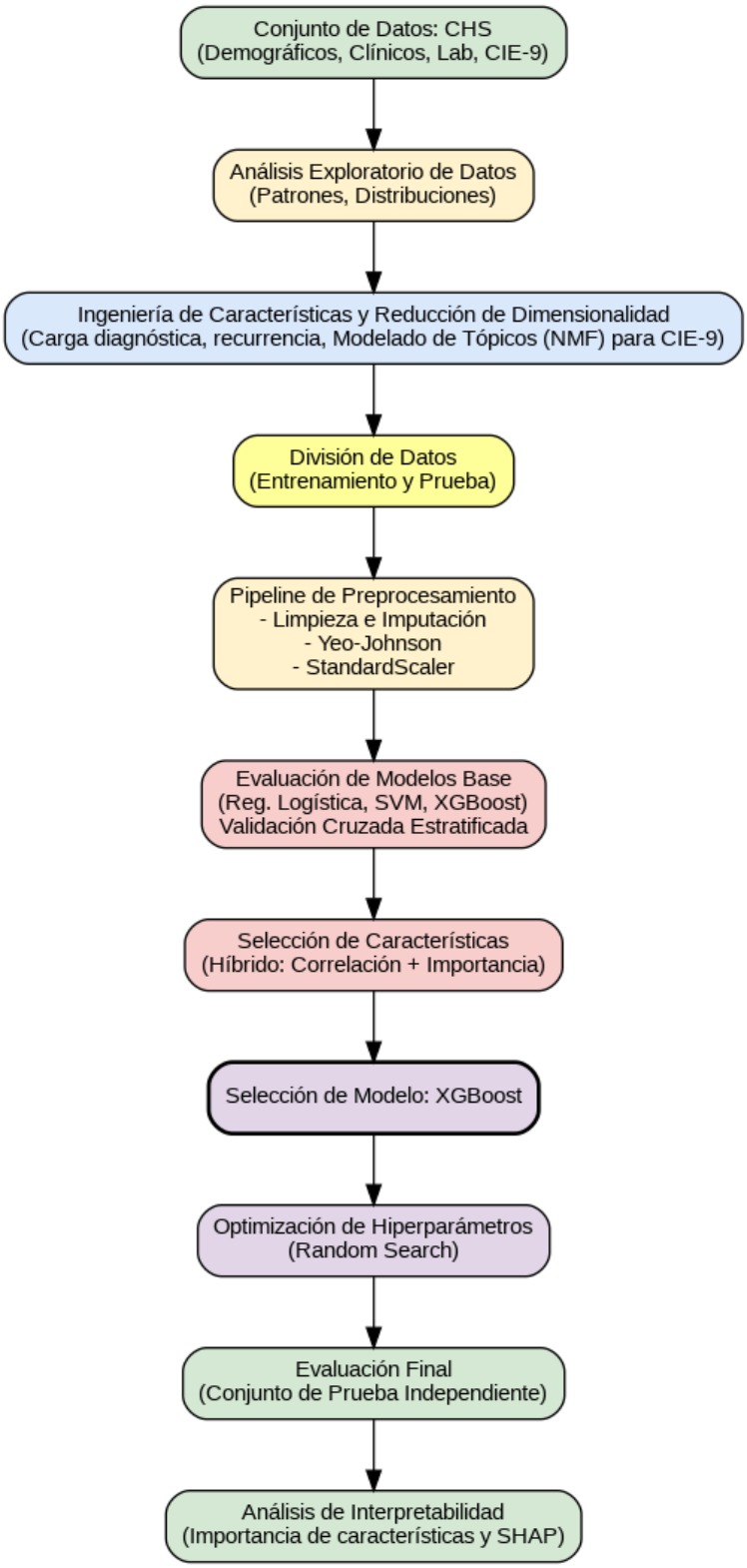

Figura 4. Esquema metodológico para el desarrollo del modelo predictivo.

1.7. Temario

(i) **Capítulo 1. Introducción:** Se presenta en forma general el problema abordado y se exponen los objetivos y alcances del proyecto, indicando la metodología para el cumplimiento de estas.

(ii) **Capítulo 2. Marco Teórico:** Resume los conceptos clínicos y metodológicos necesarios para el estudio, incluyendo fundamentos de estancia hospitalaria, aprendizaje automático y técnicas relevantes utilizadas.

(iii) **Capítulo 3. Desarrollo:** Describe la metodología propuesta, el preprocesamiento de datos, la ingeniería de características y el entrenamiento y optimización de los modelos predictivos.

(iv) **Capítulo 4. Resultados:** Expone el desempeño del modelo seleccionado, las métricas de evaluación y el análisis de interpretabilidad de los resultados obtenidos.

(v) **Capítulo 5. Conclusiones:** Sintetiza los principales hallazgos, se vinculan con los objetivos planteados y presenta líneas de trabajo futuro.

Capítulo 2. Marco Teórico

2.1. Introducción

En esta sección se detallan conceptos que serán necesarios para el desarrollo del modelo predictivo, como el concepto de Estancia Hospitalaria y conceptos necesarios sobre modelos de Aprendizaje Automático.

2.2. Estancia Hospitalaria

La estancia hospitalaria, definida como el número de días que un paciente hospitalizado permanecerá en el hospital durante un único evento de admisión [4], es una métrica clínica que otorga información respecto de los resultados de un hospital, una sala en específico o incluso de un enfermedad o procedimiento en concreto. Se considera un indicador importante del rendimiento de un centro de salud en términos de eficiencia, uso de recursos y calidad de atención, siendo una de las principales herramientas para el control de costos sanitarios y mejora de la gestión hospitalaria [16].

Debido al limitado número de recursos disponibles en hospitales como camas, personal y servicios de atención en cada sala, contar con la posibilidad de tener una herramienta capaz de predecir este parámetro impactaría positivamente en la capacidad de atención, disponibilidad de camas y personal, reducción de costos de hospitalización y de listas de espera. Sin embargo, la estancia hospitalaria, es una métrica compleja y multidimensional, asociada a una variedad de factores que pueden ser agrupados según si son factores asociados al paciente, datos de diagnóstico y tratamiento o factores a nivel institucional. Entre los factores asociados al paciente se encuentran la edad, el género, el estado civil, la presencia de comorbilidades, el tipo de admisión, el estado funcional y social, y la existencia de complicaciones clínicas. Desde el punto de vista del diagnóstico y tratamiento, influyen aspectos como la gravedad de la enfermedad, el tipo de procedimiento realizado, la duración de la anestesia o el grupo de diagnóstico relacionado (GRD). A nivel institucional, destacan variables como el número de camas, la ubicación geográfica del hospital, los recursos disponibles, la presencia de especialistas y las políticas internas de atención y alta médica [10].

Para predecir la LOS se han propuesto distintos enfoques. Se han utilizado herramientas de investigación operativa como modelos compartimentales y simulaciones de eventos discretos. Asimismo, se han empleado métodos estadísticos como el análisis de regresión y árboles de decisión. Sin embargo, en años recientes, se ha impulsado el uso de técnicas de aprendizaje automático, incluyendo redes neuronales profundas, máquinas de vectores de soporte y clasificadores

probabilísticos, que han mostrado un rendimiento superior en la predicción de estancias hospitalarias, especialmente en contextos clínicos complejos como unidades de cuidados intensivos [4]. En la Figura 5, se destacan los distintos enfoques para la predicción de LOS según la literatura.

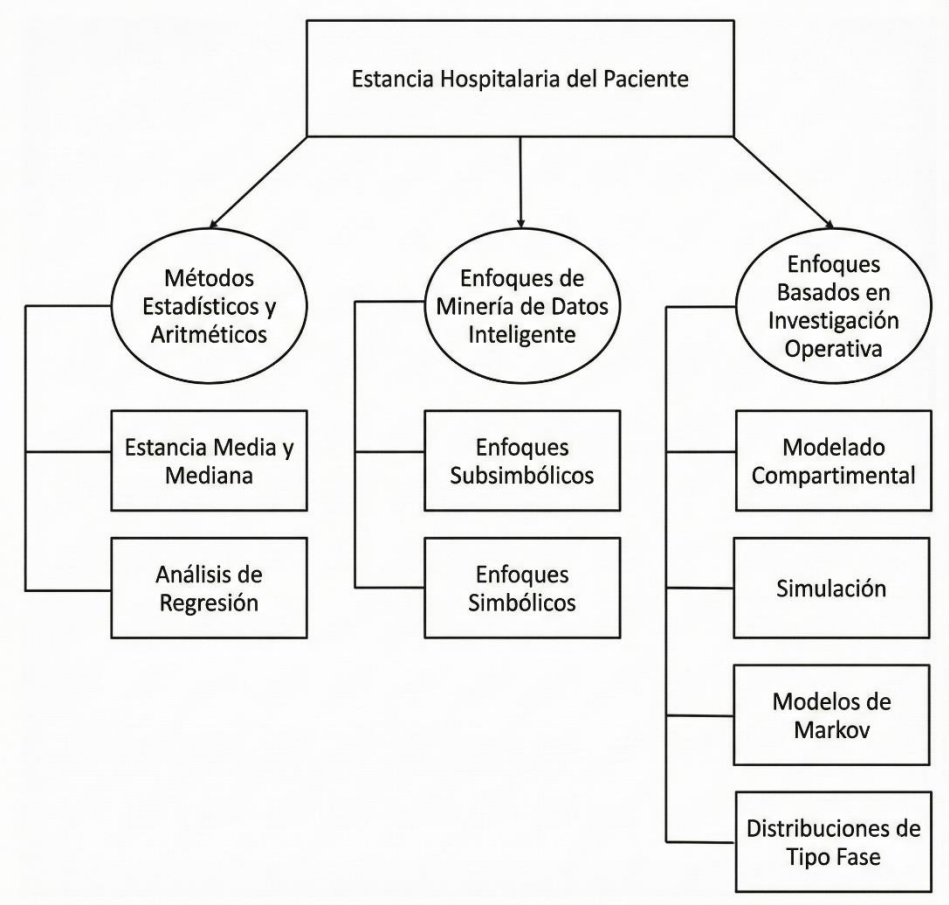

Figura 5. Clasificación de los distintos enfoques para predecir LOS [4].

2.3. Aprendizaje Automático

Dentro del aprendizaje automático existen múltiples enfoques diferentes, entre los cuales se destacan el aprendizaje supervisado y el aprendizaje no supervisado. Ambos son tipos de modelos que se basan en encontrar patrones dentro de los datos. Sin embargo, el aprendizaje supervisado aprende patrones subyacentes a partir de datos anotados, conocidos como etiquetas, mientras que el aprendizaje no supervisado, solo dispone de los datos sin etiquetar [18].

Entre estos dos enfoques, el aprendizaje supervisado es el más comúnmente utilizado en implementaciones prácticas de predicción de resultados clínicos donde la salida este bien definida (y etiquetada), como mortalidad o estancia hospitalaria.

El aprendizaje supervisado diferencia entre dos tipos de tareas de predicción: tareas de clasificación y tareas de regresión.

Los métodos de clasificación, tales como regresión logistica, arboles de decisión y Support Vector Machine (SVM), se utilizan cuando se requiere predecir respuestas discretas, mientras que los métodos de regresión, como la regresión lineal, son más adecuados para la predicción de valores continuos [18], [19].

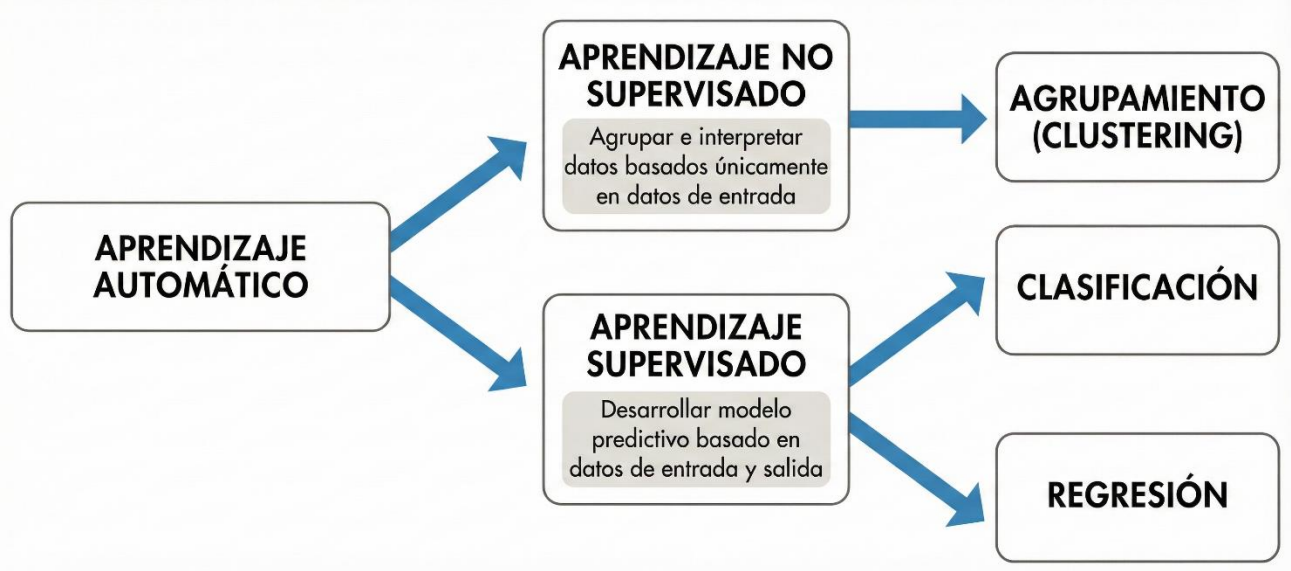

Figura 6. Técnicas de aprendizaje automático [19].

Particularmente, el problema de predicción de LOS puede ser definido como un problema de clasificación definiendo uno o más umbrales (por ejemplo, LOS corta, mediana o prolongada), o un problema de regresión, en donde lo que se busca es predecir el número exacto de días de estancia. No obstante, hay casos donde se usan ambos enfoques combinados para potenciar la capacidad predictiva de los modelos, como en [12].

Para implementar Modelos de Aprendizaje Automático en la predicción de resultados como la LOS, es necesario una serie de pasos que involucran el desde el preprocesamiento de los datos y la selección de características (ingeniería de variables), hasta la entrenamiento, prueba y validación del modelo.

2.3.1 Modelos de clasificación

A. Random Forest

Random Forest (RF) es un algoritmo de clasificación que funciona mediante la formación de múltiples árboles de decisión. El árbol de decisión extrae reglas de decisión simples de las características de los datos. Cuanto más profundo sea el árbol, más complejas serán las reglas y más ajustado el modelo. Los modelos RF superan el problema del sobreajuste de los árboles de decisión. Para clasificar un nuevo registro, cada árbol de decisión accede a un subconjunto aleatorio de los datos de entrenamiento, aumentando la diversidad del modelo en su conjunto y otorgándole robustez. Cuando el modelo realiza una predicción, el bosque aleatorio toma las salidas de todos los modelos de árboles de decisión individuales y emite la predicción con los votos más altos entre los modelos individuales, es decir, la etiqueta de clase final será la que tenga más votos. En la práctica, la RF es una técnica de clasificación rápida y capaz de tratar grandes conjuntos de datos desequilibrados [3], [18].

B. Support Vector Machine

Un algoritmo *Support Vector Machine* (SVM) es un clasificador cuyo objetivo es obtener un hiperplano óptimo en un espacio de N dimensiones, donde N es el número de características utilizadas para clasificar los datos. El objetivo es encontrar un plano que maximice el margen máximo, es decir, la distancia máxima entre puntos de datos que pertenecen a clases diferentes, de tal manera que cualquier punto de prueba pueda clasificarse con seguridad en una clase determinada [20]. En la Figura 7, se ilustra la división del set de datos utilizada en modelos SVM [18].

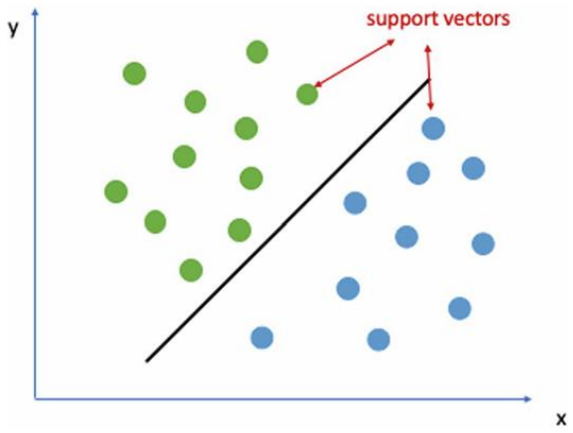

Figura 7. División de los datos por un hiperplano, utilizado en modelos SVM [18].

C. Extreme Gradient Boosting

Extreme Gradient Boosting (*XGBoost*) es un algoritmo de aprendizaje supervisado basado en árboles de decisión que implementa una versión optimizada del enfoque *gradient boosting* [21]. A diferencia de métodos de ensamble basados en bagging, como RF, *XGBoost* construye árboles de secuencialmente, donde cada nuevo árbol se entrena para corregir errores cometidos por los modelos previamente construidos. Con esto, se logra mejorar progresivamente el desempeño del modelo [22].

XGBoost optimiza una función objetivo que combina una función de pérdida con términos de regularización que penalizan la complejidad del modelo, controlando el número de hojas y el peso de las predicciones.

Desde el punto de vista computacional, el modelo *XGBoost* incorpora paralelización en la construcción de los árboles y un manejo nativo de valores faltantes, aprendiendo automáticamente la mejor dirección de división para datos incompletos. En aplicaciones prácticas, *XGBoost* ha demostrado un alto desempeño en tareas de clasificación con datos tabulares complejos y de alta dimensionalidad, siendo ampliamente utilizado en problemas clínicos y biomédicos [22].

2.3.2 Métricas de evaluación

Una vez construido un modelo de aprendizaje automático es fundamental evaluar el rendimiento de este. Para esto se utilizan, generalmente métricas como la exactitud, precisión (valor predictivo positivo, PPV), sensibilidad (también llamado *recall*), especificidad y *F1-score*. La matriz de confusión es utilizada para visualizar la relación entre estas métricas. Además, se detallan las fórmulas matemáticas para el cálculo de exactitud, precisión, sensibilidad y especificidad.

		Predicted		
		True	False	
Actual	True	True positive (TP)	False negative (FN) Type II error	Recall = Sensitivity = $\frac{TP}{TP+FN}$
	False	False positive (FP) Type I error	True negative (TN)	Specificity = $\frac{TN}{TN+FP}$
		Precision = $\frac{TP}{TP+FP}$		Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ F1 = $\frac{2\times Precision \times Recall}{Precision+Recall}$

Figura 8. Matriz de confusión y sus métricas [18]

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

$$Presición = \frac{VP}{VP + FP}$$

$$Sensibilidad = \frac{VP}{VP + FN}$$

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

Otra métrica ampliamente utilizada para evaluar el rendimiento de un modelo de aprendizaje supervisado es el área bajo la curva característica operativa del receptor (en inglés, *area under receiver operating curve*, AUROC), que corresponde a la suma del área bajo la curva en el gráfico con de la tasa de falsos positivos (FPR, especificidad) en el eje x y la tasa de verdaderos positivos (TPR) en el eje y, como se muestra en la figura [18].

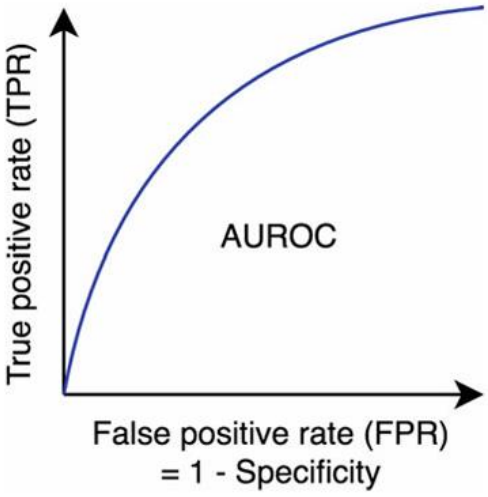

Figura 9. Ejemplo de gráfica de AUROC [18].

Otra técnica útil en el desarrollo y evaluación de modelos de aprendizaje automático es la validación cruzada (VC), especialmente cuando se busca mitigar el riesgo de sobreajuste. Entre las variantes más comunes se encuentra la VC *k-fold*, cuyo principio consiste en dividir el conjunto de datos en *k* subconjuntos o *folds*. El modelo es entrenado *k* veces, utilizando en cada iteración *k – 1* particiones para el entrenamiento y reservando la partición restante para la validación. De tal modo que cada subconjunto es utilizado una única vez como conjunto de validación. Al finalizar el proceso, se calcula el error promedio de las *k* ejecuciones, lo que proporciona una estimación más robusta del

rendimiento del modelo frente a nuevas muestras. En la práctica, los valores más utilizados para *k* suelen ser 5 o 10, aunque en su forma más extrema puede aplicarse la validación cruzada *leave-one-out* (LOOCV), en la que se entrena el modelo n veces, dejando una observación fuera en cada iteración (siendo n el número total de muestras) [18].

Figura 10. Esquema del modelo de validación cruzada [18].

Capítulo 3. Desarrollo

3.1. Introducción

Para abordar el problema de predicción de estancias hospitalarias prolongadas, se diseñó un flujo de trabajo que transforma datos clínicos heterogéneos en una estructura adecuada para el aprendizaje automático. El conjunto de datos se construyó a partir de un subconjunto del estudio CHS, integrando información demográfica, signos vitales, resultados de laboratorio y el historial clínico del paciente, incluyendo códigos de diagnóstico y procedimientos codificados según el estándar CIE-9.

Considerando la complejidad inherente de los datos clínicos (alta dimensionalidad, distribuciones asimétricas y desbalance de clases), se desarrolló un enfoque metodológico estructurado en etapas. Este enfoque comprende el preprocesamiento de variables, el modelado de información clínica mediante tópicos latentes, la selección de características y el entrenamiento y evaluación de modelos predictivos. En este capítulo se describen en detalle los procedimientos implementados en cada una de estas etapas.

3.2. Preparación del conjunto de datos

Esta sección describe el proceso de construcción del conjunto de datos base utilizado en el desarrollo del modelo predictivo. En primer lugar, se detalla la selección inicial de variables a partir de la base de datos CHS. Posteriormente, se define la variable objetivo correspondiente a la duración de la estancia hospitalaria y se establece el criterio utilizado para su binarización. Estas etapas permiten estructurar un conjunto de datos coherente y adecuado para las fases posteriores de preprocesamiento, ingeniería y selección de características.

3.2.1 Selección inicial de variables

Para la construcción del conjunto de datos base, se seleccionaron inicialmente 209 variables provenientes del estudio CHS, correspondientes a 50 variables numéricas y 158 variables categóricas. Este conjunto incluye características demográficas, resultados de exámenes físicos y de laboratorio, variables asociadas al historial médico del paciente (como hábitos de salud y medicamentos prescritos), así como registros de diagnósticos y procedimientos codificados según el sistema CIE-9.

Las variables consideradas corresponden a información registrada durante los dos primeros años del estudio. Para ello, se integraron datos provenientes del *baseline* del estudio CHS y de los

registros de hospitalizaciones, excluyendo aquellos ingresos que finalizaron en fallecimiento. En la Tabla 2 se presentan ejemplos representativos de las variables seleccionadas, junto con su descripción.

Tabla 1. Ejemplos de variables numéricas y categóricas seleccionadas del estudio CHS.

Nombre de Variable	Tipo de Dato	Descripción
GEND01	Categórica	Género
AGE2	Categórica	Grupo de edad
MARIT01	Categórica	Estado Civil
WBLD23	Numérica	Recuento de glóbulos blancos
LDLADJ	Numérica	LDL calculado
HEMOGL23	Numérica	Hemoglobina (g/dl)
DCODE1	Categórica	Código de diagnóstico principal (ICD-9)
PCODE1	Categórica	Código de procedimiento principal (ICD-9)
ACE06	Categórica	Inhibidores de la ECA sin diuréticos
BETA06	Categórica	Betabloqueantes sin diuréticos
STTN06	Categórica	Inhibidores de la HMG CoA reductasa (Estatinas)

A partir de este conjunto inicial, se realizaron ajustes y transformaciones orientadas a preparar los datos para la fase de entrenamiento de los modelos predictivos, los cuales se describen en las secciones siguientes.

3.2.2 Definición de la variable objetivo: LOS

Dado que el conjunto de datos de hospitalizaciones de CHS no incluye directamente una variable que represente la duración de la estancia hospitalaria, la variable objetivo fue calculada a partir de las variables *ADDAYS58* y *DIDAYS58*, que indican el número de días transcurridos desde el enrolamiento del paciente hasta la admisión y el alta hospitalaria, respectivamente.

Con el fin de reducir la influencia de valores extremos, los registros con valores de *LOS* superiores al percentil 99,5 fueron reemplazados por valores nulos. Este criterio corresponde a estancias mayores a 58 días y afectó a un total de 15 observaciones.

La Figura 11 muestra la distribución de la variable *LOS* calculada a partir de los registros de admisión y alta hospitalaria. La distribución resultante presenta un marcado sesgo positivo, característica común en datos clínicos, lo cual motiva un tratamiento específico de esta variable en etapas posteriores del modelado.

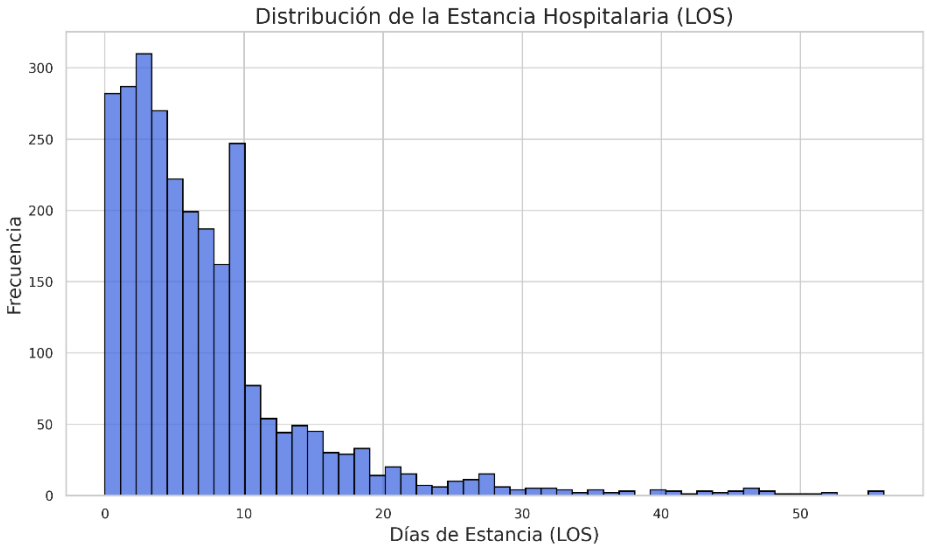

Figura 11. Distribución de la variable LOS.

3.2.3 Binarización de la variable objetivo

Con el objetivo de formular el problema como una tarea de clasificación binaria, se definió un criterio para discriminar entre estancias hospitalarias prolongadas y no prolongadas. Para ello, se consideraron distintos umbrales de corte basados en la distribución de la variable LOS, incluyendo criterios estadísticos como el percentil 75, el percentil 90 y el método del rango intercuartílico.

La Figura 22 (Anexo A) ilustra la posición de estos umbrales candidatos sobre la distribución de *LOS*, mientras que la Tabla 11 (Anexo A) resume los valores de corte asociados a cada método y la proporción de casos positivos resultante. Esta información se utilizó como apoyo para la selección del criterio de binarización.

Finalmente, se adoptó un umbral de 7 días para discretizar la variable objetivo. Este criterio permite identificar estancias de mayor complejidad clínica, manteniendo una proporción de clases adecuada para el entrenamiento de modelos predictivos. A partir de este umbral, la variable *LOS* fue categorizada en dos clases: estancia no prolongada y estancia prolongada.

3.3. Análisis exploratorio de los datos

Con el objetivo de seleccionar de manera fundamentada los métodos de procesamiento e ingeniería de características para el desarrollo de modelos predictivos, se realizó un análisis exploratorio inicial del conjunto de datos, considerando la presencia de valores nulos y la distribución de las variables incluidas.

3.3.1 Análisis de valores nulos

En primer lugar, se revisó la cantidad de valores nulos por columna en el set de datos. En las tablas 2 y 3, se muestran las cinco columnas categóricas con mayor cantidad de valores nulos y las cinco columnas numéricas con mayor cantidad de valores nulos, respectivamente.

Tabla 2. Columnas categóricas con mayor porcentaje de valores nulos.

Variable	Cantidad de Nulos	Porcentaje de Nulos
ESTBL	1391	51,4%
AFIB	855	31,6%
INSUL12	402	14,86%
APOE	372	13,75%
NCEPMS	198	7,32%

Tabla 3. Columnas numéricas con mayor porcentaje de valores nulos.

Variable	Cantidad de Nulos	Porcentaje de Nulos
CYSGFRBL	424	15,67%
IL6BL	302	11,16%
STDDIA16	246	9,09%
NETSCR03	243	8,98%
STDSYS16	238	8,8%

Las columnas con un porcentaje de valores nulos mayor al 20% fueron removidas del set de datos. Para el resto de las variables se implementó una estrategia diferenciada según el tipo de dato. En el caso específico de los códigos de diagnóstico y procedimientos, se optó por tratar los valores nulos como una nueva categoría explícita denominada *SIN_REGISTRO*. Esta decisión se fundamenta en la hipótesis de que la ausencia de registro posee valor predictivo para el modelo; por ejemplo, un código de procedimiento nulo indica la no realización de intervenciones, lo cual impacta directamente en la complejidad clínica y, consecuentemente, en la duración de la estancia hospitalaria.

3.3.2 Análisis de distribuciones

En las figuras se muestran las distribuciones de algunos ejemplos de variables numéricas continuas presentes en el set de datos. Tal como se observa en los histogramas de la Figura 13, hay variables que se distribuyen de forma simétrica. Sin embargo, como se presenta en la Figura 14, hay

otras distribuciones fuertemente sesgadas y con presencia de *outliers*. Este tipo de distribuciones y valores atípicos se trataron utilizando una transformación logarítmica y escalado de los datos, detallado en las secciones más adelante.

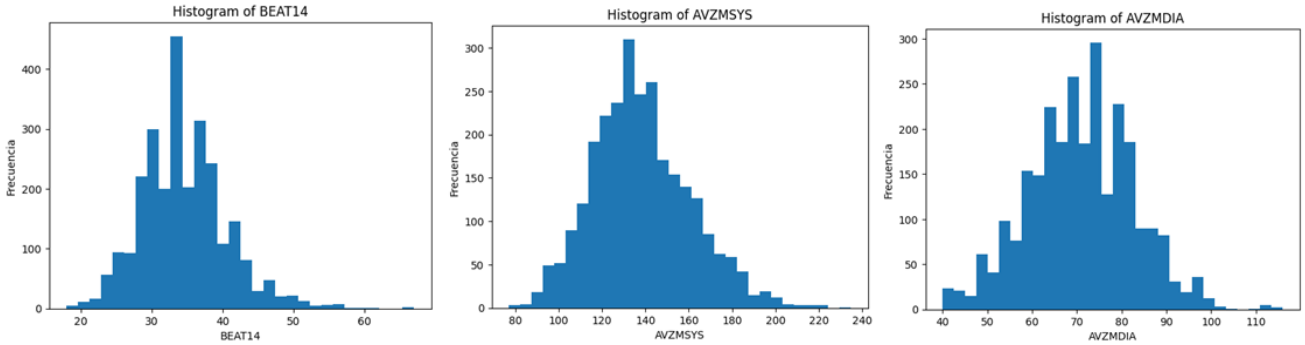

Figura 12. Distribuciones de variables numéricas del set de datos.

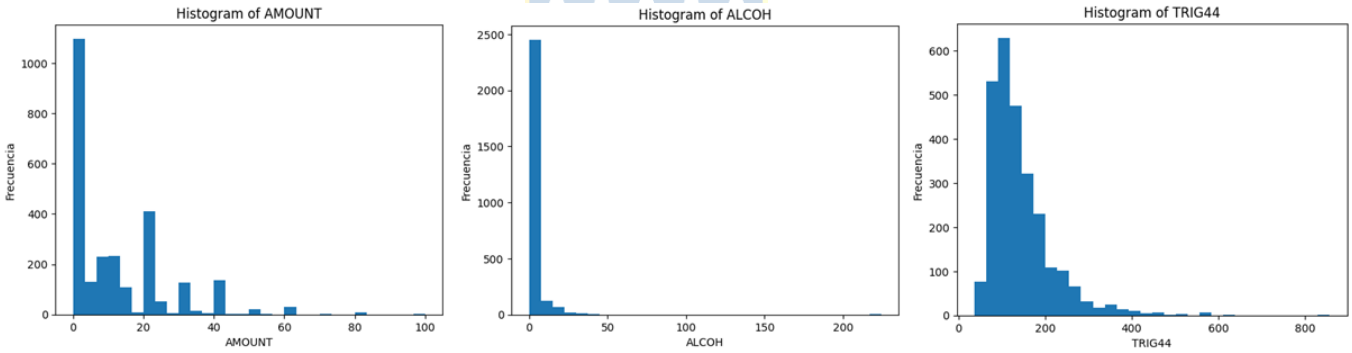

Figura 13. Distribuciones de variables numéricas del set de datos.

3.4. Procesamiento de los datos e Ingeniería de Características

La preparación de los datos para el modelo se estructuró mediante un flujo de trabajo que integra el preprocesamiento y la ingeniería de características. En una primera etapa, se definieron pipelines diferenciados para variables numéricas y categóricas, encargados de la imputación y codificación de los datos. Sobre esta base, se incorporaron transformaciones y nuevas variables predictivas, dando lugar a una matriz de diseño final adecuada para maximizar el desempeño del algoritmo en la estimación de LOS. La Figura 15 muestra el esquema global de procesamiento de los datos para el diseño de la matriz de entrada X lista para los modelos de clasificación.

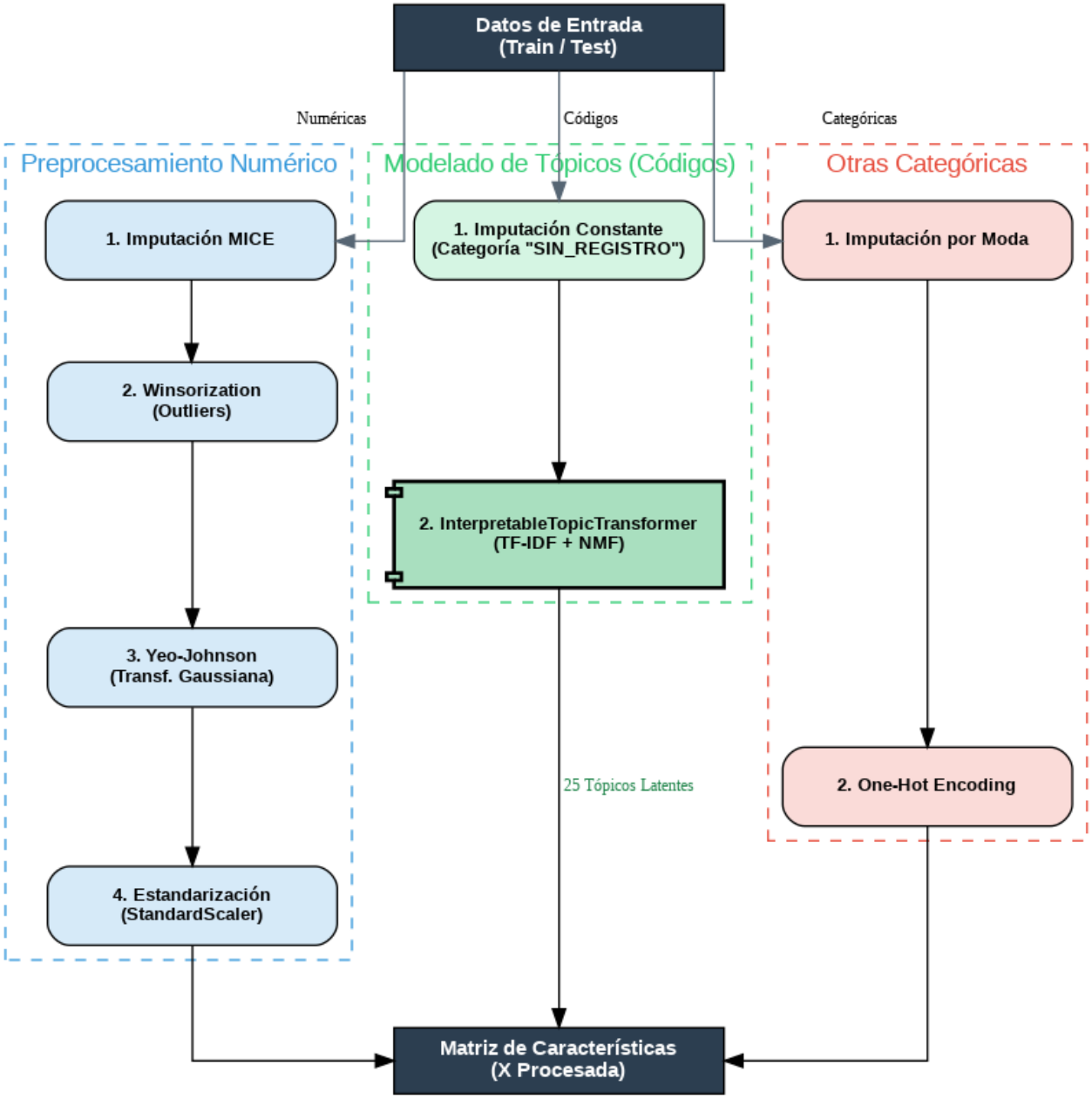

Figura 14. Esquema metodológico del pipeline de procesamiento diferenciado.

3.4.1 Preprocesamiento de variables Numéricas

Las variables numéricas fueron procesadas mediante un pipeline específico, diseñado para manejar valores faltantes, mitigar la influencia de valores extremos, corregir asimetrías en las distribuciones y homogenizar las escalas antes de su incorporación en los modelos predictivos.

En primer lugar, los valores faltantes se imputaron utilizando el método de imputación múltiple por ecuaciones encadenadas (MICE), preservando las relaciones multivariadas entre las variables.

Posteriormente, se aplicó un proceso de *winsorización* basado en cuantiles para limitar el impacto de valores atípicos extremos. A continuación, las variables fueron transformadas mediante la transformación de *Yeo–Johnson*, con el objetivo de reducir el sesgo y aproximar las distribuciones a la normalidad. Finalmente, las variables numéricas fueron estandarizadas mediante *StandardScaler*, asegurando media cero y desviación estándar unitaria.

La Figura 16 muestra un ejemplo representativo de la distribución de la variable numérica *WBLD23* antes y después de la transformación Yeo-Johnson, donde se observa una reducción de la asimetría de la distribución, preservando su estructura general. Complementariamente, la Figura 17 ilustra el efecto de la aplicación del *StandardScaler* sobre la variable *WEIGHT* del conjunto de datos, evidenciando el centrado y escalado de los valores.

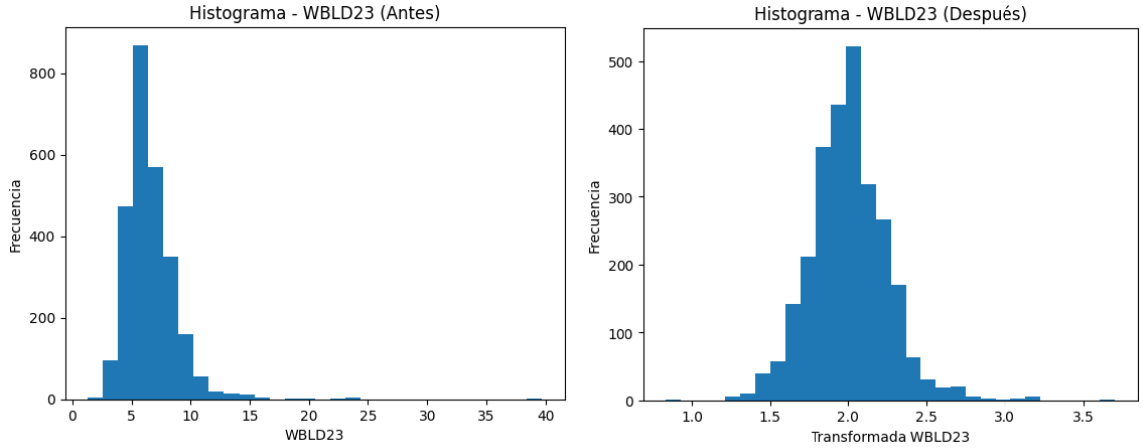

Figura 15. Histogramas de las distribuciones de WBLD23 antes y después de aplicar Yeo-Johnson.

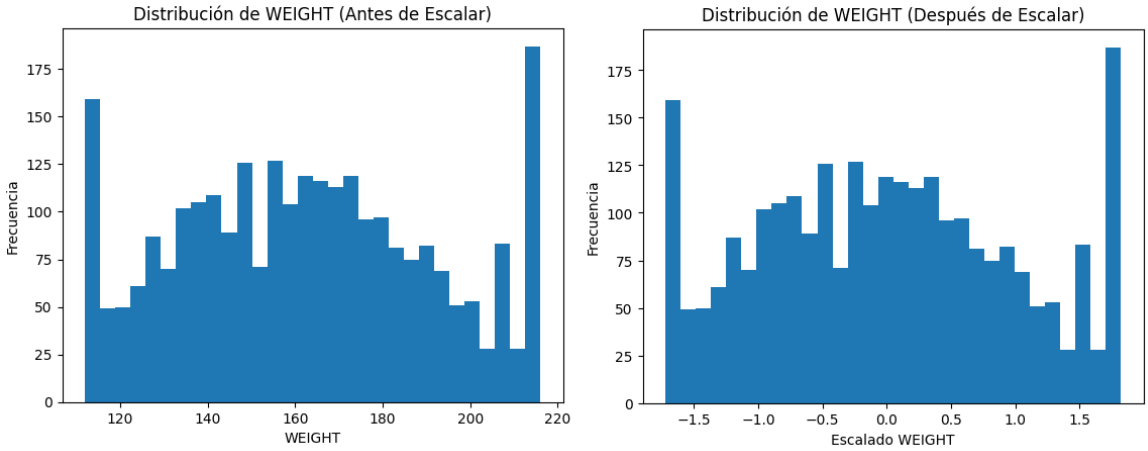

Figura 16. Efecto de la aplicación del StandardScaler sobre la variable WEIGHT.

3.4.2 Preprocesamiento de variables Categóricas

Las variables categóricas que no corresponden a códigos de diagnóstico ni de procedimientos fueron procesadas mediante un pipeline específico. En primer lugar, los valores faltantes se imputaron utilizando la moda de cada variable. Posteriormente, estas variables fueron transformadas mediante codificación one-hot, con el fin de obtener una representación numérica adecuada para su incorporación en los modelos predictivos.

3.4.3 Modelado de Códigos Clínicos mediante Transformer Interpretable

El historial clínico de un paciente, compuesto por códigos de diagnóstico, procedimientos realizados y medicamentos administrados, constituye una fuente de información relevante, pero altamente dispersa y de gran dimensionalidad. El tratamiento de cada código como una variable independiente mediante técnicas tradicionales, como la codificación one-hot, generaría un espacio de características con miles de dimensiones, lo que introduce ruido y dificulta el entrenamiento y la convergencia de los modelos predictivos.

Para abordar este problema, se diseñó e implementó un componente personalizado denominado *Interpretable Topic Transformer*, orientado a modelar de forma compacta e interpretable la información contenida en los códigos clínicos. Este enfoque combina técnicas de procesamiento de lenguaje natural y factorización matricial para condensar la historia clínica de cada paciente en un conjunto reducido de variables latentes.

El proceso de transformación se estructuró en dos etapas secuenciales dentro del pipeline de procesamiento:

A. Construcción del Documento Clínico

Para cada paciente, se concatenaron los códigos de diagnóstico, procedimientos y medicamentos asociados a una admisión hospitalaria en una única secuencia de texto. De este modo, cada episodio clínico se representó como un "documento", donde los términos corresponden a eventos médicos discretos.

B. Factorización de Matrices No Negativas (NMF)

A partir de los documentos clínicos, se construyó una matriz TF-IDF (Term Frequency–Inverse Document Frequency), sobre la cual se aplicó factorización de matrices no negativas (NMF), configurando un espacio latente de $k = 25$ componentes. A diferencia de otras técnicas de reducción

de dimensionalidad o de agrupación manual de códigos, NMF impone restricciones de no negatividad que favorecen la interpretabilidad clínica de los resultados. En este contexto, cada componente puede interpretarse como un patrón clínico latente o una comorbilidad subyacente de carácter aditivo.

De esta manera, la historia clínica de cada paciente se representa mediante un vector denso de 25 dimensiones, que captura información semántica relevante del estado de salud sin incurrir en los problemas asociados a la alta dimensionalidad y la dispersión de los códigos originales. En la Tabla 4 se describen a modo de ejemplo, 8 tópicos creados en la etapa de modelado, junto a sus descripciones sugeridas, basadas en los códigos más relevantes en el modelado de NMF. Además, en la Tabla del Anexo F, se muestran los elementos más relevantes que componen los 25 tópicos modelados, utilizados para describir cada tópico.

Tabla 4. Ejemplos de tópicos relevantes, junto a su descripción.

Tópico	Descripción
topic_0	Síndrome Cardiorrenal y Manejo de Fluidos
topic_8	Paciente con Comorbilidad Cardiovascular y Manejo Multidisciplinario
topic_9	Recuperación Funcional y Manejo de Anemia en Cirugía Ortopédica Mayor
topic_12	Salud Ginecológica y Terapia de Reemplazo Hormonal
topic_13	Paciente con Comorbilidad Psiquiátrica
topic_16	Hipotiroidismo y Mantenimiento Hormonal
topic_20	Tópico de Manejo Metabólico y Prevención del Riesgo Cerebrovascular
topic_22	Hemorragia Gastrointestinal Aguda y Procedimientos de Diagnóstico Endoscópico

3.4.4 Cálculo de Nuevas Características Predictivas

Con el objetivo de representar de forma adecuada la complejidad clínica y el historial asistencial de los pacientes, se generaron variables numéricas derivadas a partir de los datos originales. Estas características buscan capturar aspectos relacionados con la carga diagnóstica, la intensidad de las intervenciones médicas y la recurrencia hospitalaria, factores que han sido identificados en la literatura como relevantes para la estimación del tiempo de estancia hospitalaria. La Tabla 5 resume las variables derivadas consideradas en este estudio junto con su descripción.

Tabla 5. Variables derivadas junto a su descripción.

Variable	Descripción
num_diagnosticos_total	Número total de diagnósticos registrados.

num_procedimientos_total	Número total de procedimientos realizados.
NRO_HOSPITALIZACION	Número de hospitalizaciones previas del paciente.
DIAS_DESDE_ULTIMA_ALTA	Número de días transcurridos desde la última alta hospitalaria.

El comportamiento de las variables derivadas fue evaluado mediante análisis bivariante y visual para determinar su potencial predictivo respecto a la variable objetivo. Los resultados completos se reportan en la Tabla 13 y Figuras 23-26 del Anexo C.

3.5. Selección de Características

Con el fin de reducir la dimensionalidad del conjunto de datos y eliminar información redundante, se aplicó un proceso de selección de características en dos etapas, combinando criterios de correlación estadística y métodos basados en modelos de aprendizaje supervisado.

En una primera etapa, se evaluó la colinealidad entre las variables numéricas explícitas mediante el cálculo de la correlación lineal absoluta. Se utilizó un umbral de 0,75, eliminando una variable de cada par altamente correlacionado. Como resultado, se descartaron nueve variables numéricas, principalmente asociadas a mediciones clínicas estrechamente relacionadas, reduciendo el número total de características de 197 a 188.

En una segunda etapa, se aplicó un método de selección basado en la importancia de variables estimada por un clasificador *XGBoost* entrenado sobre el conjunto reducido. Se utilizó un umbral relativo de 1,25 veces la importancia media para conservar únicamente las variables con mayor contribución predictiva. Este procedimiento permitió seleccionar un subconjunto final de 89 variables relevantes para la predicción del tiempo de estancia hospitalaria.

En conjunto, este proceso permitió una reducción total del 54,8 % en el número de variables, manteniendo una representación compacta y adecuada para el entrenamiento de los modelos predictivos.

3.6. Entrenamiento y evaluación de modelos

Con el fin de definir una estrategia de modelado adecuada para la predicción de estancias hospitalarias prolongadas, se llevó a cabo una etapa de entrenamiento y evaluación comparativa de distintos algoritmos de clasificación. Esta fase permitió analizar el desempeño relativo de diferentes enfoques de aprendizaje automático bajo un protocolo homogéneo, estableciendo una base metodológica sólida para la selección del modelo y su posterior optimización.

3.6.1 Selección de Modelos de base

Se consideraron modelos con distintos enfoques de aprendizaje: Regresión Logística, SVM, *Random Forest*, y el modelo *XGBoost* basado en *gradient boosting*. Esta selección buscó cubrir arquitecturas lineales y no lineales, así como métodos de ensamble con diferente capacidad de modelado.

La evaluación se realizó mediante validación cruzada estratificada de 5 pliegues sobre el conjunto de entrenamiento, manteniendo la proporción de clases en cada partición. Todos los modelos fueron entrenados utilizando la misma matriz de características resultante del proceso de preprocesamiento e ingeniería de características descrito previamente.

Se reportaron métricas como precisión, *recall*, *F1-score* y exactitud. En base a estos resultados, se seleccionó el modelo con mejor desempeño global para la etapa posterior de optimización de hiperparámetros.

3.6.2 Optimización del Modelo

Se llevó a cabo un proceso de optimización de hiperparámetros con el objetivo de mejorar la capacidad de generalización del modelo frente al desbalance presente en la variable objetivo. Para esto, se utilizó un esquema de búsqueda aleatoria mediante *RandomizedSearchCV*, acoplado a validación cruzada estratificada con k=5 pliegues, preservando la proporción de clases en cada partición.

La optimización se realizó sobre un conjunto acotado de hiperparámetros relevantes del modelo, incluyendo el número de estimadores, la tasa de aprendizaje, la profundidad máxima de los árboles, el término de regularización *gamma* y el factor de ponderación de clases (*scale_pos_weight*), este último ajustado en función del grado de desbalance observado en el conjunto de entrenamiento.

El *F1-score* fue utilizado como métrica objetivo durante el proceso de optimización, dado su equilibrio entre precisión y sensibilidad en contextos con clases desbalanceadas. El conjunto de hiperparámetros que maximiza esta métrica fue seleccionado como configuración final del modelo para su evaluación posterior.

3.6.3 Evaluación Final del Modelo

El desempeño del modelo *XGBoost* optimizado fue evaluado utilizando el conjunto de prueba, previamente separado y no utilizado durante las etapas de entrenamiento ni de optimización de

hiperparámetros. Esta evaluación tuvo por objetivo estimar la capacidad de generalización del modelo en datos no vistos.

Se reportaron las métricas de desempeño definidas en la sección anterior, incluyendo F1-score, precisión, sensibilidad (*recall*) y exactitud. Adicionalmente, se obtuvo la matriz de confusión y la curva ROC, las cuales permiten analizar el comportamiento del clasificador frente a ambas clases y evaluar el compromiso entre sensibilidad y tasa de falsos positivos.

Por último, con el fin de facilitar la interpretabilidad del modelo, se analizaron la importancia de características y los valores SHAP, permitiendo identificar las variables con mayor contribución en las decisiones del modelo. Estos análisis proporcionan una base para la interpretación clínica del modelo y serán discutidos en detalle en el capítulo de resultados.

Capítulo 4. Resultados

4.1. Introducción

En este capítulo se presentan los resultados obtenidos a partir de la metodología descrita previamente para la predicción de estancias hospitalarias prolongadas. Se reporta el desempeño de los modelos evaluados, la comparación entre arquitecturas base y la evaluación final del modelo optimizado sobre un conjunto de prueba independiente. Asimismo, se incluyen análisis de interpretabilidad orientados a identificar las variables más relevantes en la toma de decisiones del modelo.

4.2. Descripción del conjunto de datos final

Tras el proceso de preparación, preprocesamiento e ingeniería de características descrito en el capítulo anterior, se obtuvo un conjunto de datos final compuesto por 2706 hospitalizaciones correspondientes a pacientes del estudio CHS. La variable objetivo *LOS* fue binarizada utilizando un umbral de 7 días, resultando en una proporción de *34,57%* de estancias prolongadas y *65,43%* de estancias no prolongadas.

La Figura 17 muestra la distribución final de la variable objetivo binarizada, evidenciando un desbalance moderado entre clases, condición considerada durante las etapas de entrenamiento y evaluación de los modelos.

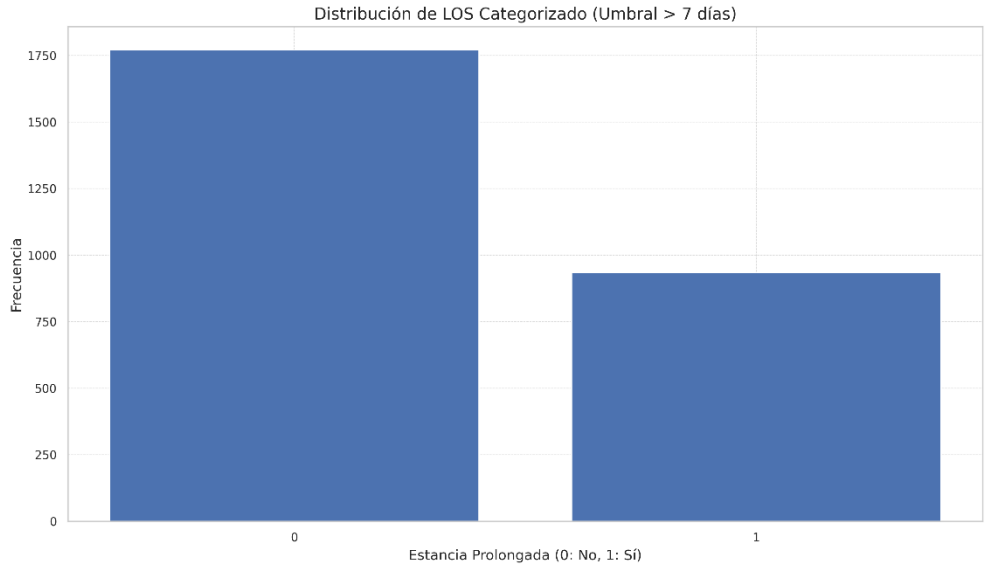

Figura 17. Distribución de clases de la variable objetivo.

4.3. Resultados de la comparación de modelos

El desempeño de los modelos base evaluados se resume en la Tabla 7, donde se reportan las métricas promedio obtenidas mediante validación cruzada estratificada sobre el conjunto de entrenamiento.

En general, se observaron diferencias entre las arquitecturas evaluadas. En particular, el modelo *XGBoost* obtuvo el mejor desempeño global, alcanzando el mayor valor de F1-score promedio, junto con un balance favorable entre precisión y sensibilidad.

En base a estos resultados, el modelo *XGBoost* fue seleccionado como arquitectura final para la etapa de optimización de hiperparámetros.

Tabla 6. Métricas de evaluación promedio obtenidas para cada modelo.

Modelo	AUC	Recall	F1	Precisión
AdaBoost	0,731	0,432	0,507	0,616
Regresión Logística	0,713	0,614	0,556	0,509
RandomForest	0,743	0,329	0,446	**0,696**
SVC	0,719	**0,615**	0,561	0,516
XGBoost	**0,747**	0,527	**0,566**	0,612

4.4. Optimización de hiperparámetros

La optimización de hiperparámetros del modelo *XGBoost* permitió mejorar su desempeño predictivo respecto a la configuración base. La Tabla 8 presenta los hiperparámetros seleccionados como óptimos para el modelo.

Tabla 7. Hiperparámetros óptimos para el modelo XGBoost.

Hiperparámetro	Valor Optimizado
colsample_bytree	0,82
gamma	2,02
learning_rate	0,02
max_depth	4
n_estimators	151
scale_pos_weight	1,89
subsample	0,88

4.5. Evaluación Final del Modelo en el conjunto de Prueba

El desempeño final del modelo *XGBoost* optimizado fue evaluado utilizando el conjunto de prueba independiente, obteniendo una exactitud de 0,74 y un valor AUC de 0,78. La Tabla 9 resume las métricas obtenidas, incluyendo F1-score, precisión y *recall*.

Tabla 8. Evaluación final del modelo XGBoost en el set de Prueba.

Clase	Precisión	*Recall*	f1-score
Corta	0,81	0,79	0,8
Larga	0,62	0,66	0,64

El modelo alcanzó un F1-score de 0,64 en la clase de *LOS* prolongada, reflejando dificultades para encontrar el equilibrio entre la identificación de estancias prolongadas y la reducción de falsos positivos. La Figura 18 muestra la matriz de confusión asociada, permitiendo visualizar la distribución de predicciones correctas e incorrectas para ambas clases. Adicionalmente, se presenta la curva ROC del modelo, obteniéndose un área bajo la curva (AUC) de 0,78, lo que indica una capacidad discriminativa del clasificador superior a los modelos de línea base en distintos umbrales de decisión.

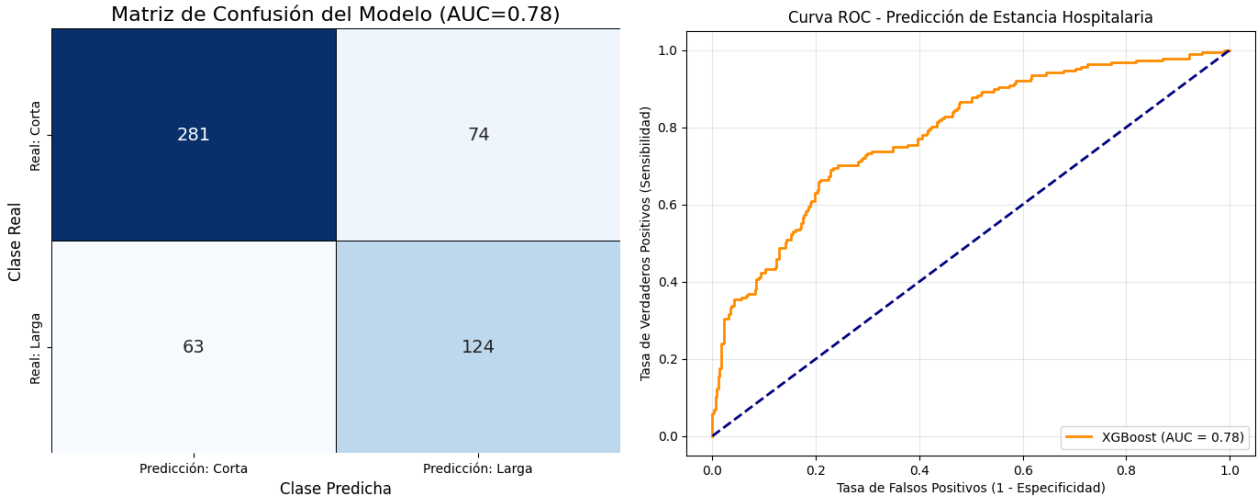

Figura 18. Matriz de Confusión y Curva ROC del modelo XGBoost.

4.6. Importancia de Características e interpretabilidad del modelo

La Figura 19 muestra la importancia global de las características, destacándose variables asociadas a la carga clínica del paciente (cantidad de diagnósticos y procedimientos), antecedentes médicos y patrones latentes derivados del modelado de tópicos latentes.

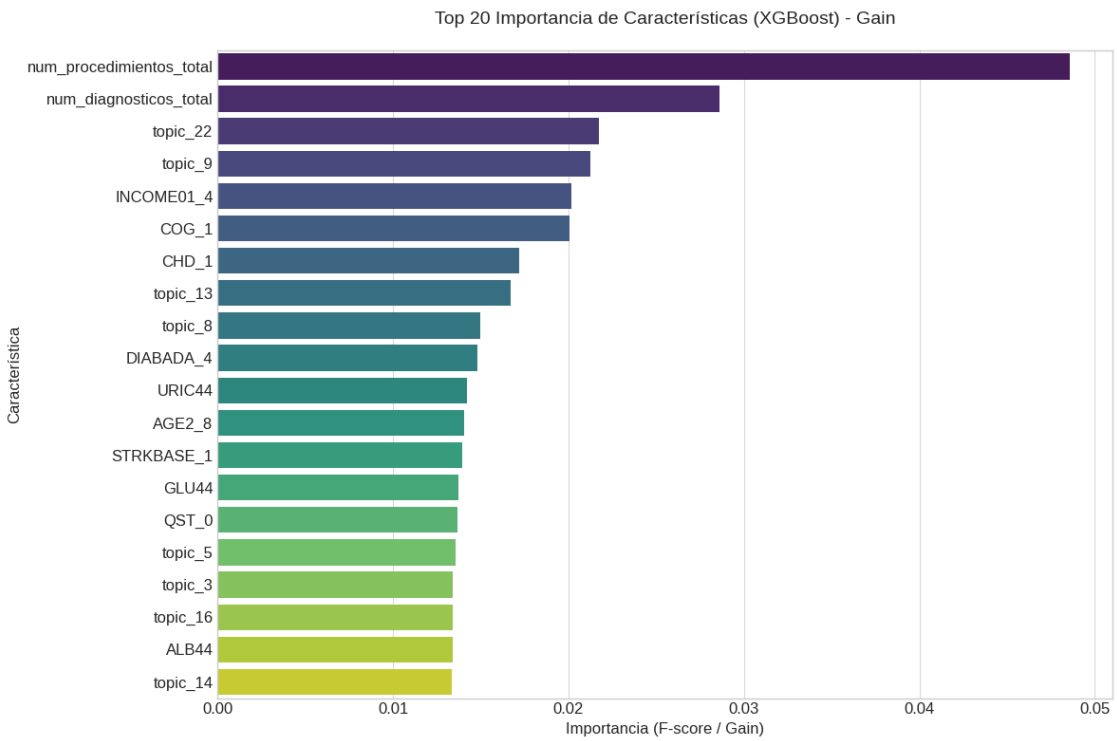

Figura 19. Importancia de Características del modelo XGBoost.

La Figura 20 presenta el resumen global del impacto de las variables, donde se observa que las características con mayor impacto en la predicción corresponden a *num_procedimientos_total* y *num_diagnosticos_total*, las cuales muestran una relación positiva clara con la probabilidad de estancia prolongada. En particular, valores elevados de estas variables incrementan de forma consistente la salida del modelo hacia la clase positiva, reflejando una mayor complejidad clínica y carga asistencial.

Adicionalmente, se identifican múltiples componentes derivados del *Modelado de Tópicos Latentes* (por ejemplo, *topic_22*, *topic_9*, *topic_8* y *topic_13*) entre las variables más influyentes. Estos tópicos capturan patrones latentes asociados a combinaciones recurrentes de diagnósticos, procedimientos y medicamentos, contribuyendo de manera significativa a la capacidad predictiva del modelo.

Las Tablas 10 y 11 muestran los códigos de diagnóstico y procedimientos más relevantes dentro de los tópicos 22 y 9, respectivamente. El tópico 22 representa un perfil de paciente complejo que combina eventos como la hemorragia digestiva con la necesidad de transfusiones y procedimientos invasivos. Por otra parte, el tópico 9 agrupa diagnósticos de osteoartrosis y anemia

aguda con procedimientos de reemplazo articular y rehabilitación. Este tópico captura la complejidad de la recuperación post-artroplastia.

Entre las variables clínicas tradicionales, destacan indicadores de laboratorio y evaluación funcional como *WBLD23*, *CIS42*, *AVZMDIA*, *LDLADJ* y *LTAAI*, cuyo impacto es menor en comparación con las variables de carga clínica, pero consistente en la dirección de la predicción.

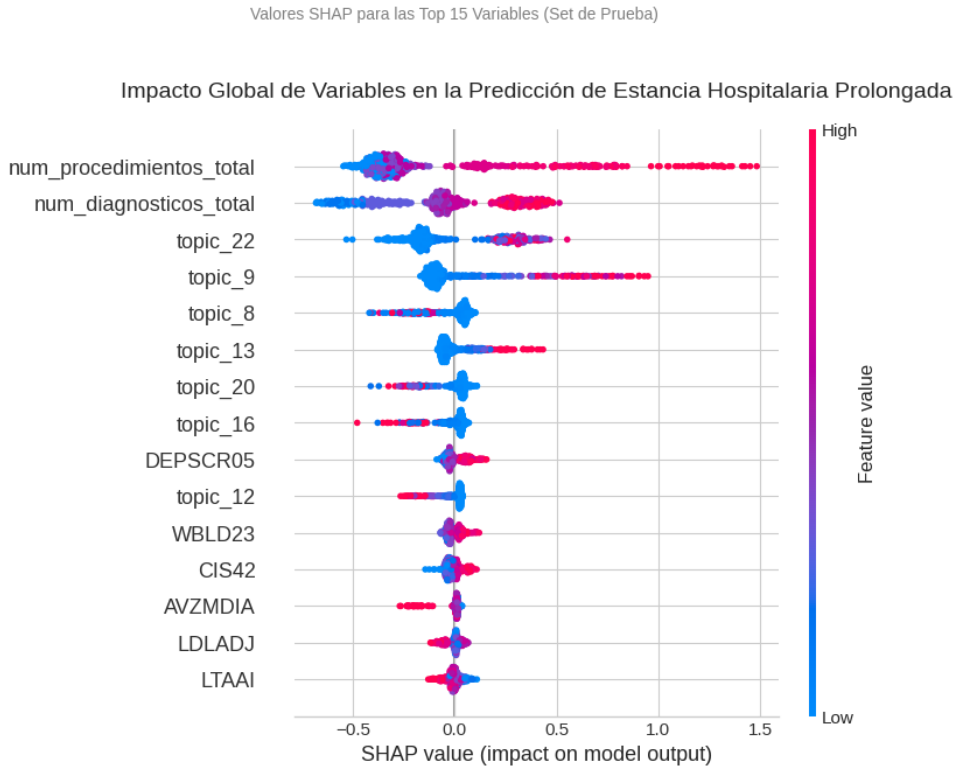

Figura 20. Impacto global de las variables en la salida del modelo.

Tabla 9. Códigos de Diagnóstico y Procedimientos más relevantes del Tópico 22.

Código	Peso	Descripción
d_285.1	0,642	Anemia posthemorrágica aguda
d_562.10	0,489	Diverticulosis de colon
d_578.9	0,480	Hemorragia gastrointestinal, no especificada
p_45.13	0,954	Endoscopía con biopsia cerrada
p_99.04	0,901	Transfusión de concentrado de glóbulos rojos
p_99.25	0,350	Inyección o infusión de quimioterapia contra el cáncer
p_45	0,341	Operaciones de incisión, escisión y anastomosis del intestino

Tabla 10. Códigos de Diagnóstico y Procedimientos más relevantes del Tópico 9.

Código	Peso	Descripción
d_715.36	0,955	Osteoartrosis localizada de la pierna
d_285.1	0,330	Anemia post-hemorrágica aguda
d_v43.6	0,176	Articulación sustituida por otros medios (Prótesis)
p_81.54	1,241	Sustitución total de rodilla
p_93.39	1,086	Otras terapias físicas
p_93.83	0,693	Terapia ocupacional para actividades de la vida diaria

A nivel individual, la Figura 21 muestra un ejemplo de explicación local para un paciente clasificado con estancia hospitalaria prolongada. En este caso, la predicción está dominada por un alto número de procedimientos (*num_procedimientos_total*) y diagnósticos (*num_diagnosticos_total*), junto con contribuciones adicionales de tópicos clínicos específicos, que desplazan la predicción desde el valor base hacia la clase positiva. De forma complementaria, algunas variables con valores bajos actúan reduciendo parcialmente la probabilidad predicha.

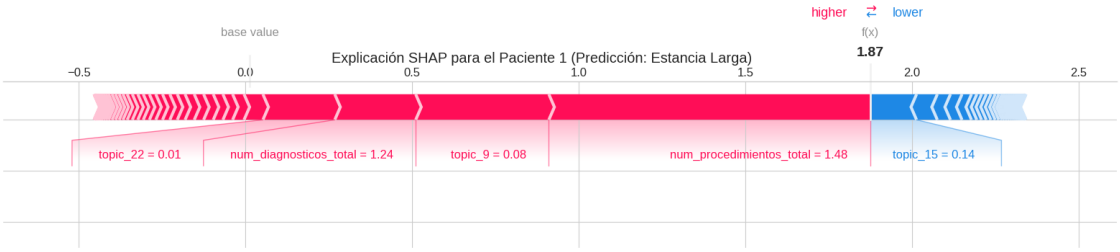

Figura 21. Interpretabilidad del modelo de predicción de estancia prolongada mediante SHAP.

En conjunto, estos resultados evidencian que el modelo prioriza información relacionada con la intensidad del manejo clínico, la complejidad diagnóstica y patrones latentes del historial médico, lo que refuerza la coherencia clínica y la interpretabilidad del modelo propuesto.

4.7. Discusión

Los resultados obtenidos confirman que el enfoque metodológico propuesto es adecuado para abordar la predicción de estancias hospitalarias prolongadas en pacientes cardíacos. En la etapa de comparación de modelos base, el modelo *XGBoost* mostró un desempeño superior y más consistente en comparación con los demás modelos evaluados. Esto evidencia su mayor capacidad para modelar

relaciones no lineales y manejar interacciones complejas entre variables clínicas. Este resultado respalda la elección de este modelo para la etapa final de optimización y evaluación.

La optimización de hiperparámetros permitió mejorar la capacidad de generalización del modelo, particularmente mediante el ajuste de parámetros asociados a la profundidad de los árboles, la tasa de aprendizaje y el manejo explícito del desbalance de clases a través de *scale_pos_weight*. El uso de validación cruzada estratificada durante este proceso contribuyó a obtener configuraciones estables y robustas, evitando sobreajuste al conjunto de entrenamiento.

En el conjunto de prueba, el modelo *XGBoost* optimizado alcanzó un desempeño aceptable, con un equilibrio adecuado entre precisión y sensibilidad. En particular, el *F1-score* obtenido para la clase de estancia prolongada refleja una capacidad razonable para identificar correctamente pacientes de mayor complejidad clínica, lo que resulta relevante desde el punto de vista operativo y de gestión hospitalaria. La matriz de confusión mostró que, si bien persisten errores de clasificación, el modelo logra capturar una proporción significativa de casos de estancia larga, manteniendo un compromiso aceptable con los falsos positivos.

No obstante, los resultados obtenidos indican que el desempeño del modelo aún tiene margen de mejora considerable, especialmente en la identificación de estancias prolongadas. Los valores moderados de precisión y *recall* para la clase positiva (0,62 y 0,66, respectivamente) reflejan la dificultad inherente del problema, influida tanto por el desbalance de clases como por la complejidad clínica de los pacientes. Este comportamiento sugiere que factores no observables en el conjunto de datos, como la evolución intrahospitalaria, decisiones clínicas dinámicas o limitaciones operativas, pueden desempeñar un rol relevante en la duración de la estancia y no son capturados por el modelo actual.

Desde la perspectiva de interpretabilidad, el análisis SHAP proporcionó evidencia de que las variables derivadas *num_procedimientos_total* y *num_diagnosticos_total,* asociadas a la carga asistencial del paciente, desempeñaron un rol central en la predicción. Estas variables se posicionaron como las variables con mayor impacto global en el modelo, lo que es coherente con la hipótesis de que pacientes con mayor complejidad diagnóstica y mayor intervención terapéutica tienden a requerir hospitalizaciones más prolongadas.

Adicionalmente, los tópicos latentes extraídos a partir de los códigos clínicos aportaron información relevante y complementaria al modelo. Los tópicos con mayor impacto en la predicción se asocian a escenarios clínicos de alta complejidad y consumo de recursos. En particular, el tópico 22 representa pacientes con eventos agudos graves, como hemorragia digestiva y anemia, que

requieren transfusiones y procedimientos invasivos, frecuentemente en el contexto de patologías crónicas subyacentes, lo que se traduce en trayectorias clínicas prolongadas. Por su parte, el tópico 9 agrupa diagnósticos relacionados con osteoartrosis y anemia, junto con procedimientos de reemplazo articular y rehabilitación, capturando procesos de recuperación postquirúrgica compleja que demandan estancias hospitalarias extendidas. Asimismo, el tópico 8 se asocia a intervenciones cardiovasculares de alta complejidad, como cirugías multivalvulares y control de arritmias.

La relevancia de estos tópicos refuerza la validez clínica del enfoque propuesto y demuestra que el modelado de tópicos mediante NMF, además de solucionar el problema de la alta dimensionalidad de los códigos de procedimiento y diagnósticos, permite capturar patrones clínicos complejos que no son evidentes a partir de variables individuales.

En conjunto, estos resultados sugieren que la combinación de ingeniería de características clínicas, modelos de *gradient boosting* e interpretabilidad basada en SHAP constituye una estrategia sólida para la predicción de la estancia hospitalaria prolongada, ofreciendo no solo un desempeño predictivo prometedor, sino también una base interpretable para su potencial aplicación en entornos clínicos reales.

Capítulo 5. Conclusiones

5.1. Discusión General

Este trabajo se centró en abordar el problema de la predicción de estancia hospitalaria prolongada en pacientes cardíacos mayores de 65 años, utilizando técnicas de aprendizaje automático aplicadas al conjunto de datos Cardiovascular Health Study (CHS). Para lograr esto, se diseñó e implementó una metodología que permitió transformar datos clínicos heterogéneos y de alta complejidad en una representación adecuada para el entrenamiento de modelos predictivos.

Los resultados obtenidos evidenciaron que fue posible capturar patrones clínicos relevantes asociados a la duración de la hospitalización combinando técnicas de ingeniería de características, reducción de dimensionalidad basada en tópicos latentes y modelos de clasificación no lineales. En particular, el uso de modelos de *gradient boosting* permitió modelar interacciones complejas entre variables clínicas, superando el desempeño de modelos lineales y enfoques más simples.

Además, el análisis de interpretabilidad confirmó que el modelo no solo alcanzó un desempeño predictivo razonable, sino que además el modelo trabajó en base a variables y patrones clínicamente coherentes, reforzando la validez del enfoque propuesto. No obstante, los resultados también reflejan la dificultad inherente del problema, influida por el desbalance de clases y por factores clínicos y operativos no capturados en el conjunto de datos, lo que limita el desempeño máximo alcanzable bajo el enfoque actual.

5.2. Conclusiones

A partir de los resultados obtenidos, se pueden establecer las siguientes conclusiones en relación con los objetivos planteados en esta Memoria de Título:

Se logró desarrollar y validar un modelo predictivo basado en aprendizaje automático para la detección de estancia hospitalaria prolongada en pacientes cardíacos mayores de 65 años, utilizando datos del estudio CHS. El modelo seleccionado, basado en *XGBoost*, demostró un desempeño superior y consistente en comparación con otros modelos de clasificación evaluados.

Se definió e implementó una metodología de análisis de datos clínicos alineada con la literatura científica, que incluyó etapas de análisis exploratorio, preprocesamiento avanzado, ingeniería de características y selección de variables. Esta metodología permitió abordar desafíos comunes en datos clínicos, tales como la alta dimensionalidad, distribuciones asimétricas y desbalance de clases.

El análisis exploratorio de la base de datos CHS permitió identificar variables clínicas y patrones potencialmente asociados a la duración de la estancia hospitalaria, proporcionando una base sólida para la posterior ingeniería de características y el modelamiento predictivo.

Se desarrollaron y evaluaron distintos modelos de clasificación aplicados al conjunto de datos CHS, utilizando métricas de desempeño apropiadas para el contexto clínico. El modelo *XGBoost* optimizado alcanzó un equilibrio adecuado entre precisión y sensibilidad para la clase de estancia prolongada, evidenciando fortalezas en la identificación de pacientes de mayor complejidad, aunque con márgenes de mejora en su desempeño global.

El análisis de interpretabilidad basado en valores SHAP permitió identificar variables claves asociadas a la carga asistencial del paciente representado en variables contadoras de cantidad de procedimientos y diagnósticos, así como también, tópicos latentes clínicamente coherentes derivados de los códigos de diagnóstico y procedimientos, tales como el tópico 22, que representa pacientes con eventos agudos graves, como hemorragia digestiva y anemia, el tópico 9, que representa diagnósticos relacionados con osteoartrosis y anemia, junto a procedimientos de reemplazo articular y rehabilitación, capturando procesos de recuperación postquirúrgica compleja que demandan estancias hospitalarias extendidas; y el tópico 8, asociado a intervenciones cardiovasculares de alta complejidad, tales como cirugías multivalvulares y control de arritmias.

Esto demuestra que el enfoque propuesto no solo es predictivo, sino también interpretable, lo cual es fundamental para su potencial aplicación en contextos clínicos reales.

5.3. Trabajo Futuro

Este trabajo abre algunas líneas de investigación y mejora que podrían explorarse a futuro, tales como la incorporación de información clínica longitudinal, como evolución intrahospitalaria, eventos adversos o cambios en el tratamiento, lo que podría mejorar significativamente la capacidad predictiva del modelo al capturar dinámicas temporales no consideradas en este estudio. Explorar arquitecturas de modelamiento adicionales, incluyendo modelos basados en aprendizaje profundo o enfoques híbridos que integren explícitamente información temporal y secuencial. Validar el modelo desarrollado en conjuntos de datos externos o en contextos clínicos distintos, con el objetivo de evaluar su capacidad de generalización y su robustez frente a variaciones poblacionales y operativas. Y, por último, profundizar en el análisis clínico de los tópicos latentes identificados, incorporando la colaboración de expertos médicos para refinar su interpretación y evaluar su utilidad como herramientas de apoyo a la toma de decisiones clínicas.

En conjunto, estas líneas de trabajo permitirán avanzar desde un enfoque exploratorio hacia el desarrollo de herramientas predictivas con mayor impacto real sobre la gestión hospitalaria y la atención clínica.

Capítulo 6. Glosario

LOS : Length of Stay.
RF : Random Forest.
SVM : Support Vector Machine.
ANN : Artificial Neural Network.
BN : Bayesian Network.
DT : Decision Tree.
CHS : Cardiovascular Health Study.
FCA : Falla Cardiaca Aguda.
UCI : Unidad de Cuidados Intensivos.
SOFA : Sequential Organ Failure Assessment.
SAPS : Simplified Acute Physiology Score.
AUROC : Area Under Reciever Operator Curve.
KACC : King Abdulaziz Cardiac Center.
MCC : Multiple Cardiac Conditions.

ANN : Artificial Neural Network (Red Neuronal Artificial).
APACHE : Acute Physiology and Chronic Health Evaluation (Sistema de evaluación de gravedad clínica).
AUROC : Area Under Reciever Operator Curve (Área bajo la curva ROC).
BN : Bayesian Network.
CHS : Cardiovascular Health Study.
EHR : Electronic Health Records (Registros Clínicos Electrónicos).
FCA : Falla Cardiaca Aguda.
CIE-9 : Clasificación Internacional de Enfermedades, novena revisión.
KACC : King Abdulaziz Cardiac Center.
KDE : Kernel Density Estimation (Estimación de densidad por kernel).
LOS : Length of Stay (Duración de la estancia hospitalaria).
MCC : Multiple Cardiac Conditions.
MICE : Multiple Imputation by Chained Equations (Imputación múltiple por ecuaciones encadenadas).

NMF : Non-negative Matrix Factorization (Factorización de matrices no negativas).

RF : Random Forest.

SAPS : Simplified Acute Physiology Score (Puntaje fisiológico simplificado).

SHAPS : SHapley Additive exPlanations (Método de interpretabilidad basado en valores de Shapley).

SOFA : Sequential Organ Failure Assessment (Evaluación secuencial de falla orgánica).

SVM : Support Vector Machine

UCI : Unidad de Cuidados Intensivos.

Capítulo 7. Referencias

[1] "Enfermedad pulmonar obstructiva crónica (EPOC)". Accedido: 23 de junio de 2025. [En línea]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] "La Carga de Enfermedades Cardiovasculares - OPS/OMS | Organización Panamericana de la Salud". Accedido: 23 de junio de 2025. [En línea]. Disponible en: https://www.paho.org/es/enlace/carga-enfermedades-cardiovasculares

[3] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, y M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: A machine learning approach", *Int. J. Cardiol.*, vol. 288, pp. 140–147, ago. 2019, doi: 10.1016/j.ijcard.2019.01.046.

[4] K. Stone, R. Zwiggelaar, P. Jones, y N. Mac Parthaláin, "A systematic review of the prediction of hospital length of stay: Towards a unified framework", *PLOS Digit. Health*, vol. 1, nº 4, p. e0000017, abr. 2022, doi: 10.1371/journal.pdig.0000017.

[5] A. Awad, M. Bader–El–Den, y J. McNicholas, "Patient length of stay and mortality prediction: A survey", *Health Serv. Manage. Res.*, vol. 30, nº 2, pp. 105–120, may 2017, doi: 10.1177/0951484817696212.

[6] V. Lequertier, T. Wang, J. Fondrevelle, V. Augusto, y A. Duclos, "Hospital Length of Stay Prediction Methods: A Systematic Review", *Med. Care*, vol. 59, nº 10, pp. 929–938, oct. 2021, doi: 10.1097/MLR.0000000000001596.

[7] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, y S. Koblar, "Machine learning in the prediction of medical inpatient length of stay", *Intern. Med. J.*, vol. 52, nº 2, pp. 176–185, feb. 2022, doi: 10.1111/imj.14962.

[8] L. P. Fried *et al.*, "The cardiovascular health study: Design and rationale", *Ann. Epidemiol.*, vol. 1, nº 3, pp. 263–276, feb. 1991, doi: 10.1016/1047-2797(91)90005-W.

[9] L. Arbeláez *et al.*, "Factores de riesgo asociados a estancia hospitalaria prolongada en pacientes con falla cardiaca aguda", *Rev. Colomb. Cardiol.*, vol. 28, nº 2, p. 6359, may 2022, doi: 10.24875/RCCAR.M21000022.

[10] I. T. Peres, S. Hamacher, F. L. C. Oliveira, A. M. T. Thomé, y F. A. Bozza, "What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis", *J. Crit. Care*, vol. 60, pp. 183–194, dic. 2020, doi: 10.1016/j.jcrc.2020.08.003.

[11] L. Hempel, S. Sadeghi, y T. Kirsten, "Prediction of Intensive Care Unit Length of Stay in the MIMIC-IV Dataset", *Appl. Sci.*, vol. 13, nº 12, p. 6930, jun. 2023, doi: 10.3390/app13126930.

[12] I. E. Livieris, T. Kotsilieris, I. Dimopoulos, y P. Pintelas, “Decision Support Software for Forecasting Patient’s Length of Stay”, *Algorithms*, vol. 11, nº 12, Art. nº 12, dic. 2018, doi: 10.3390/a11120199.

[13] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, y S. Yoo, “Analysis of length of hospital stay using electronic health records: A statistical and data mining approach”, *PloS One*, vol. 13, nº 4, p. e0195901, 2018, doi: 10.1371/journal.pone.0195901.

[14] G. Harerimana, J. W. Kim, y B. Jang, “A deep attention model to forecast the Length Of Stay and the in-hospital mortality right on admission from ICD codes and demographic data”, *J. Biomed. Inform.*, vol. 118, p. 103778, jun. 2021, doi: 10.1016/j.jbi.2021.103778.

[15] J. P. Navarrete, J. Pinto, R. L. Figueroa, M. E. Lagos, Q. Zeng, y C. Taramasco, “Supervised Learning Algorithm for Predicting Mortality Risk in Older Adults Using Cardiovascular Health Study Dataset”, *Appl. Sci.*, vol. 12, nº 22, p. 11536, nov. 2022, doi: 10.3390/app122211536.

[16] G. D. Castillo, “Length of stay”, en *Elgar Encyclopedia of Healthcare Management*, F. Lega, Ed., Edward Elgar Publishing, 2023, pp. 305–306. doi: 10.4337/9781800889453.ch105.

[17] “Egresos Hospitalarios, últimos 4 años - SAS® Visual Analytics”. Accedido: 27 de junio de 2025. [En línea]. Disponible en: https://informesdeis.minsal.cl/SASVisualAnalytics/?reportUri=%2Freports%2Freports%2F23138671-c0be-479a-8e9d-52850e584251&sectionIndex=0&sso_guest=true&reportViewOnly=true&reportContextBar=false&sas-welcome=false

[18] L. A. Celi, *Leveraging Data Science for Global Health*. Cham: Springer International Publishing AG, 2020.

[19] “Introducción a Machine Learning - MATLAB & Simulink”. Accedido: 26 de junio de 2025. [En línea]. Disponible en: https://la.mathworks.com/discovery/machine-learning.html

[20] C. Cortes y V. Vapnik, “Support-Vector Networks”, *Mach. Learn.*, vol. 20, nº 3, pp. 273–297, sep. 1995, doi: 10.1023/A:1022627411411.

[21] T. Chen y C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ago. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[22] “XGBoost: Explicación de Extreme Gradient Boosting | Ultralytics”. Accedido: 22 de diciembre de 2025. [En línea]. Disponible en: https://www.ultralytics.com/es/glossary/xgboost

Anexo A. Elección del umbral de LOS prolongada

Tabla 11. Valores de corte por cada método de selección de umbral.

Método	Umbral (días)	Proporción de Casos Positivos
Percentil 75	9	23,4%
Tukey	18	6,8%
Percentil 90	15	9,5%
Dist. Log-Normal	26	3,0%
Umbral de Arbitrario	7	34,7%

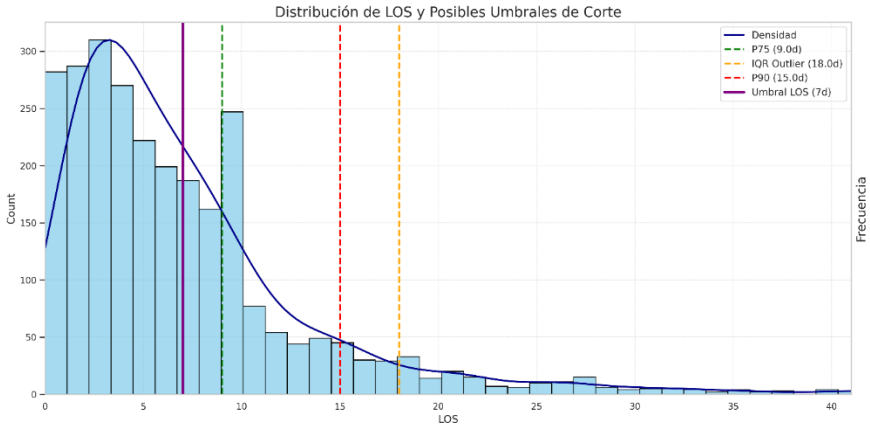

Figura 22. Distribución de LOS y posibles umbrales de corte.

Anexo B. Sesgo calculado de las variables numéricas del conjunto de datos

Tabla 12. Variables sesgadas del set de datos antes y después de aplicar transformación Yeo-Johnson.

Variable	Skewness (antes)	Skewness (después)
ALCOH	16,14745293	0,258652295
INS44	9,949098041	1,774677666
CRE44	6,50369307	2,738769639
GLU44	4,149949913	2,130950729
WBLD23	3,20517527	0,535181991

Anexo C. Evaluación de las variables derivadas del conjunto de datos.

Las Figuras 23-26 muestran el análisis bivariado por variable derivada.

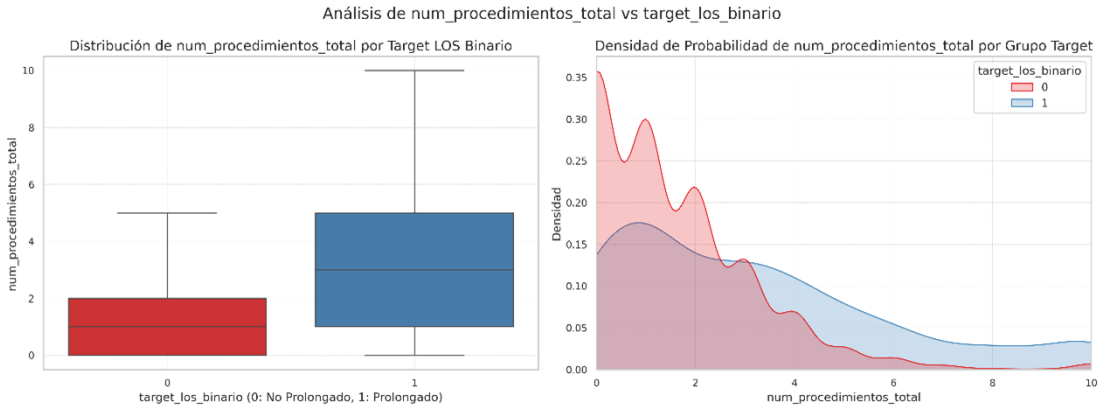

Figura 23. Análisis de num_procedimientos_total respecto a la variable objetivo.

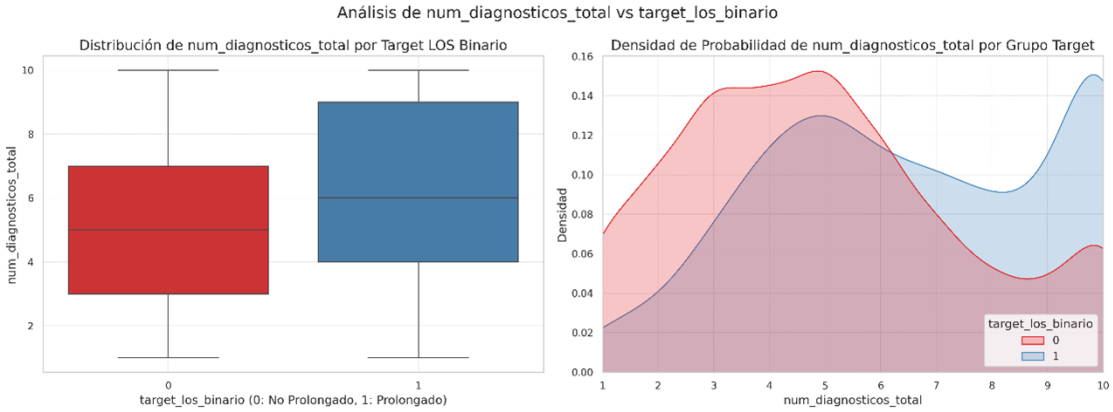

Figura 24. Análisis de num_diagnosticos_total respecto a la variable objetivo.

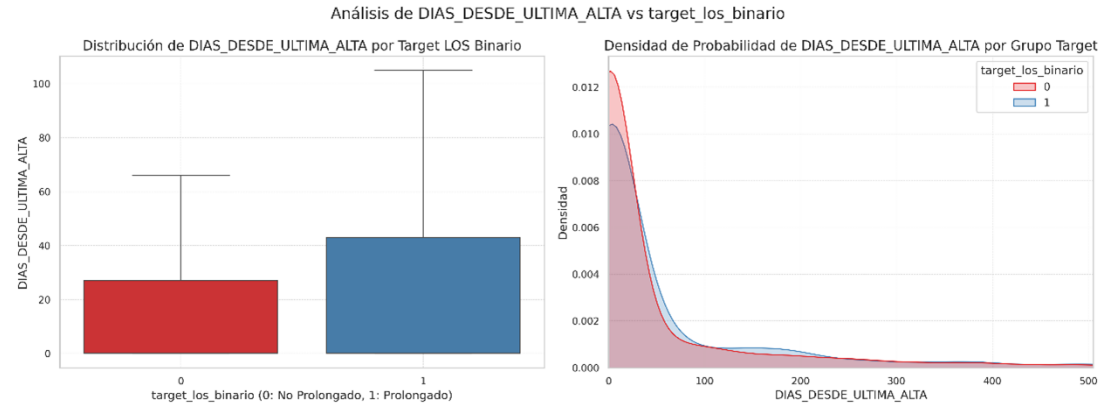

Figura 25. Análisis de la variable DIAS_DESDE_ULTIMA_ALTA respecto a la variable objetivo.

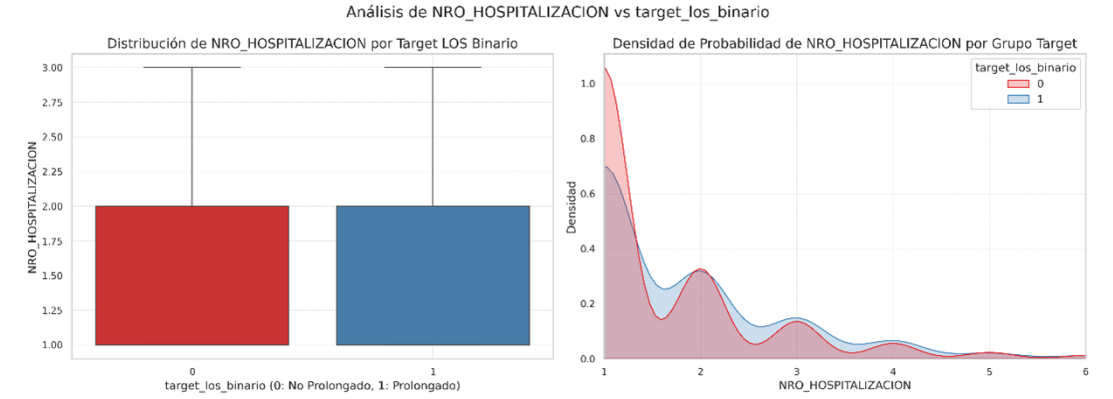

Figura 26. Análisis de la variable NRO_HOSPITALIZACION respecto a la variable objetivo.

Tabla 13. Análisis de significancia para las variables derivadas.

Variable	P-Value	Significancia (p<0.05)	Correlación con Target
NRO_HOSPITALIZACION	< 0,001	Significativa	0,083
DIAS_DESDE_ULTIMA_ALTA	0,0011	Significativa	0,027
num_diagnosticos_total	< 0,001	Significativa	0,283
num_procedimientos_total	< 0,001	Significativa	0,344

Anexo D. Variables numéricas más correlacionadas del conjunto de datos

Tabla 14. Variables numéricas más correlacionadas entre sí.

Variable 1	Variable 2	Correlación Absoluta
CRPBLORG	CRPBLADJ	1,000
LDLADJ	CHOLADJ	0,915
HEMOGL23	HEMATO23	0,903
STDSYS16	SUPSYS16	0,826
STDPUL16	SUPPUL16	0,812
BMI	WEIGHT	0,796
RTAAI	LTAAI	0,784
SUPSYS16	AVZMSYS	0,739
STDSYS16	AVZMSYS	0,736
STDDIA16	SUPDIA16	0,727

Anexo E. Lista completa de Variables Seleccionadas del CHS.

A. Variables Demográficas y datos administrativos

Tabla 15. Variables demográficas y administrativas seleccionadas del CHS.

Nombre de Variable	Descripción
IDNO	Número de identificación del paciente
GEND01	Género
AGE2	Edad en categorías
RACE01	Raza
MARIT01	Estado civil
EDUC	Educación
INCOME01	Grupo de ingresos
HSHOLD01	Número de identificación del hogar
WEIGHT	Peso (libras)
STHT	Estatura de pie (cm)
BMI	Índice de Masa Corporal (kg/m²)
WGT50	Peso usual a los 50 años
SEASON	Temporada de inscripción
ADJDTH	Fallecimiento
ADDAYS58	Días desde enrolamiento hasta la admisión
DIDAYS58	Días desde enrolamiento hasta el alta

B. Historial Médico y Condiciones Clínicas

Tabla 16.Variables del historial médico del paciente, seleccionadas del CHS.

Nombre de Variable	Descripción
MIBASE	Estatus de Infarto de Miocardio al inicio.
ANGBASE	Estatus de Angina al inicio.
CHFBASE	Estatus de Insuficiencia Cardíaca Congestiva al inicio.
STRKBASE	Estatus de Accidente Cerebrovascular (Stroke) al inicio.
TIABASE	Estatus de Ataque Isquémico Transitorio (TIA) al inicio.
CLDBASE	Estatus de Enfermedad Pulmonar Crónica al inicio.
COPDBL	EPOC (COPD) al inicio.
CHD	Reporte propio de Infarto, Angina, CABG o Angioplastia.
CBD	Reporte propio de Stroke, TIA o Endarterectomía.
AFIB	Ha tenido Fibrilación Auricular.
THROMB	Trombosis venosa profunda.
CHSTPN	Alguna vez tuvo dolor o molestia en el pecho.
PNEUMON	Neumonía diagnosticada por médico.

EMPHYSEM	Enfisema diagnosticado por médico.
CURASTHM	Diagnóstico actual de asma por médico.
DIABADA	Estatus de Diabetes ADA.
HYPER	Estatus de Hipertensión calculado.
IHYPER	Hipertensión sistólica aislada calculada.
FALL22	Caídas frecuentes.
HEALT01	Salud general.
APOE	Tipo de ApoE (Genotipo).
NCEPMS	Definición NCEP de síndrome metabólico.
RISKGRP	Grupo de riesgo de plaquetas

C. Variables sobre tabaquismo, alcohol y hábitos

Tabla 17. Variables sobre hábitos del paciente, seleccionadas del CHS.

Nombre de Variable	Descripción
SMOKE	Estatus de fumador.
EVERSM	Alguna vez fumó.
PRESSM	Fumador actual
AMOUNT	Cigarrillos fumados por día.
SMKAMT	Cantidad de fuma.
ALCOH	Bebidas alcohólicas por semana

D. Signos Vitales, Exámenes Físicos y ECG

Tabla 18. Variables de Signos vitales y exámenes, seleccionadas del CHS.

Nombre de Variable	Descripción
BEAT14	Frecuencia cardíaca (30 segundos).
SUPPUL16	Frecuencia cardíaca en supino (30 seg).
SUPSYS16	Presión sistólica en supino (mm Hg).
SUPDIA16	Presión diastólica en supino (mm Hg).
STDPUL16	Frecuencia cardíaca de pie (30 seg).
STDSYS16	Presión sistólica de pie (mm Hg).
STDDIA16	Presión diastólica de pie (mm Hg).
AVZMSYS	Promedio sistólico "Zero Mud" (mm Hg).
AVZMDIA	Promedio diastólico "Zero Mud" ajustado (mm Hg).
LTAAI	Índice tobillo-brazo izquierdo (%).
RTAAI	Índice tobillo-brazo derecho (%).
QT42	Intervalo Q-T, milisegundos.
CIS42	Puntuación de lesión cardíaca (Cardiac Injury Score).

QST	Q/QS menor con ST-T.
RAM_AVBL	Amplitud de onda R, por derivación.
SAM_BL	Amplitud de onda S, por derivación.

E. Exámenes de Laboratorio y Biomarcadores

Tabla 19. Variables de exámenes de Laboratorio, seleccionadas del CHS.

Nombre de Variable	Descripción
CHOLADJ	Colesterol (ajustado).
LDLADJ	LDL calculado.
HDL44	HDL (mg/dl).
TRIG44	Triglicéridos (mg/dl).
GLU44	Glucosa basal (mg/dl).
INS44	Insulina basal (IU/ml).
CRE44	Creatinina (mg/dl).
ALB44	Albúmina (g/dl).
K44	Potasio (mMol/dl).
URIC44	Ácido úrico (mg/dl).
IL6BL	Interleucina-6 al inicio (pg/ml).
CRPBLADJ	Proteína C reactiva, valores ajustados (mg/L).
CRPBLORG	Proteína C reactiva, valores originales (mg/L).
FIB44	Fibrinógeno (mg/dl).
HEMATO23	Hematocrito (%).
HEMOGL23	Hemoglobina (g/dl).
WBLD23	Conteo de glóbulos blancos.
PLATE23	Conteo de plaquetas.
CYSGFRBL	Tasa de filtración glomerular estimada (cistatina C).
CLOT12	Trastorno relacionado con la coagulación.
F744	Factor VII (%).

F. Puntuaciones de salud mental, cognitiva y social

Tabla 20. Puntuaciones de salud mental, cognitiva y social seleccionadas del CHS.

Nombre de Variable	Descripción
DEPSCR05	Escala de depresión.
LESCR05	Puntuación de eventos de vida.
SCORE03	Puntuación de apoyo social.
NETSCR03	Puntuación de red social.
COG	Categoría de función cognitiva.

ROSEIC	Claudicación intermitente (Cuestionario Rose).
ROSEANG	Angina (Cuestionario Rose).

G. Medicamentos agurpados por Clase Farmacológica

Tabla 21. Medicamentos agrupados, utilizados para construir la matriz modelada por NMF.

Variables	Clase Farmacológica
ACE06, ACED06, ANYACE06	Inhibidores ECA
A2A06, A2AD06	ARA II (Antagonistas R-Angiotensina II)
BETA06, BETAD06, ANYBETA06	Beta-Bloqueadores
CCB06, CCBT06, AMLOD06, NIFIR06, NIFSR06, DIHIR06, DIHSR06, DLTIR06, DLTSR06, VERIR06, VERSR06, CCBIR06, CCBSR06	Bloqueadores de Canales de Calcio
DIURET06, HCTZ06, HCTZK06, LOOP06, ANYDIUR06	Diuréticos
KCL06, KSPR06	Ahorradores de Potasio
STTN06	Estatina
FIBR06, BASQ06, NIAC06, MLPD06, PROB06, LIPID06	Otros Reductores de lípidos
WARF06, HPRNS06, ADPI06, ASA06	Anticoagulantes/Antiplaquetarios
ANAR1A06, ANAR1B06, ANAR1C06, ANAR306, DIG06	Antiarrítmicos
INSLN06	Diabetes - Insulina
OHGA06, BGND06, SLF106, SLF206, AGDI06, THZD06, DPP4I06, KBLKR06	Diabetes - ADO Orales
NTG06, PVDL06, VASO06, VASOD06, ANYVASO06, HTNMED06	Vasodilatadores
NSAID06, COX206	Antiinflamatorios
OSTRD06, ISTRD06, PRGSTN06, ESTRGN06, PRMRN06	Corticosteroides/Hormonas

SSRIS06, SNRIS06, TCA06, NTCA06, MAOI06, NDRIS06, SARIS06, TCAP06	Antidepresivos
ANTPSY06, TCAP06	Antipsicóticos
SYMPTH06, IPRTR06, OAIA06	Asma/Respiratorio
H2B06, OTCH2B06, PPI06	Gastrointestinal
EDD06, PDEI06	Disfunción Eréctil
ALZH06, PRKNSN06, BENZOD06	Neurológico/Demencia/Parkinson
GOUT06, URCOS06, XOI06	Gota/Uricosúricos
THRY06, WTLS06	Otros
UAZQTD06, UAZQTP06	Riesgo QT

Anexo F. Valores más relevantes que componen los 25 tópicos modelados.

Tabla 22. Elementos más relevantes dentro de cada tópico definido por el modelado de NMF.

Tópico	Diagnósticos (Top 2)	Procedimientos (Top 2)	Medicamentos (Top 2)
topic_0	d_428.0; d_425.4	p_39.95; p_93.92	m_loop06; m_diuret06
topic_1	d_440.9; d_414.9	p_88.74; p_80.51	m_ccbir06; m_ccb06
topic_2	d_401.9; d_414.9	p_85.43; p_81.51	m_anybeta06; m_beta06
topic_3	d_411.1; d_414.0	p_88.53; p_88.56	m_ntca06; m_ntg06
topic_4	d_496; d_466.0	p_93.94; p_96.04	m_pdei06; m_sympth06
topic_5	d_414.0; d_412	p_36.01; p_37.22	m_ntg06; m_h2b06
topic_6	d_250.00; d_250.90	p_86.22; p_88.48	m_ohga06; m_slf206
topic_7	d_401.9; d_428.0	p_87; p_79.36	m_anyace06; m_ace06
topic_8	d_401.9; d_424.1	p_14; p_13	m_anyvaso06; m_vaso06
topic_9	d_715.36; d_285.1	p_81.54; p_93.39	m_tca06; m_h2b06
topic_10	d_401.9; d_276.1	p_39; p_89.65	m_ccbsr06; m_ccb06
topic_11	d_600; d_185	p_60.2; p_57.32	m_h2b06; m_vasod06
topic_12	d_618.0; d_625.6	p_70; p_51.22	m_estrgn06; m_prmrn06
topic_13	d_mendis; d_276.5	p_03.31; p_87.03	m_tca06; m_tcap06
topic_14	d_440.9; d_429.2	p_88.48; p_38.12	m_uazqtd06; m_anar1a06
topic_15	d_434.0; d_780.2	p_87.03; p_88.72	m_anar1b06; m_htnmed06
topic_16	d_244.9; d_780.2	p_12; p_37	m_thry06; m_ntg06
topic_17	d_427.31; d_428.0	p_88.72; p_99.62	m_dig06; m_warf06
topic_18	d_401.9; d_276.8	p_38.12; p_03.09	m_hctz06; m_diuret06
topic_19	d_401.9; d_276.8	p_86.22; p_03.09	m_hctzk06; m_diuret06
topic_20	d_272.0; d_272.4	p_38.12; p_89.41	m_lipid06; m_fibr06
topic_21	d_715.90; d_v10.3	p_81.51; p_03.09	m_nsaid06; m_h2b06
topic_22	d_285.1; d_562.10	p_45.13; p_99.04	m_h2b06; m_htnmed06
topic_23	d_250.01; d_332.0	p_87; p_51.22	m_benzod06; m_tca06
topic_24	d_274.9; d_550	p_53; p_37	m_gout06; m_xoi06

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA
RESUMEN DE MEMORIA DE TITULO

Departamento : Departamento de Ingeniería Eléctrica
Carrera : Ingeniería Civil Biomédica
Nombre del memorista : José Miguel Orellana Melo
Título de la memoria : Modelos de aprendizaje automático para la predicción de extensión del periodo de hospitalización en pacientes cardíacos.
Fecha de la presentación oral : 23 de enero de 2026

Profesor(es) Guía : Rosa Figueroa
Profesor(es) Revisor(es) : Alvaro Saldaña V. y Pablo Aqueveque N.
Concepto :
Calificación :

Resumen

El tiempo de estancia hospitalaria constituye un indicador clave para la gestión clínica, al estar directamente asociado al uso de recursos, los costos de atención y la calidad asistencial. En el contexto de las enfermedades cardiovasculares (una de las principales causas de hospitalización en población adulta mayor), la predicción temprana de estancias hospitalarias prolongadas resulta especialmente relevante para la planificación de camas y la optimización de la atención en sistemas de salud.

Esta Memoria de Título tiene como objetivo desarrollar y evaluar un enfoque metodológico para la predicción de estancias hospitalarias prolongadas en pacientes cardíacos, utilizando datos del *Cardiovascular Health Study* (CHS). El conjunto de datos integra información demográfica, signos vitales, resultados de laboratorio y el historial clínico del paciente, incluyendo códigos de diagnóstico y procedimientos codificados según el estándar CIE-9.

La metodología propuesta se estructura en varias etapas. En primer lugar, se realiza la preparación del set de datos y la definición de la variable objetivo *Length of Stay* (*LOS*), formulando el problema como una tarea de clasificación binaria. Posteriormente, se implementa un pipeline de preprocesamiento diferenciado para variables numéricas y categóricas, incorporando técnicas de imputación múltiple, tratamiento de valores extremos, transformaciones de distribución y estandarización. Para abordar la alta dimensionalidad de los códigos clínicos, se propone un modelado basado en tópicos latentes mediante factorización de matrices no negativas (NMF), que permite condensar la historia clínica del paciente en un conjunto reducido de variables

interpretables. Adicionalmente, se diseñan variables derivadas orientadas a capturar la carga diagnóstica y la recurrencia hospitalaria.

El entrenamiento de los modelos se realiza mediante una comparación controlada de distintas arquitecturas de clasificación, utilizando validación cruzada estratificada y el F1-score como métrica principal. A partir de esta etapa, se selecciona y optimiza un modelo *XGBoost* mediante una búsqueda aleatoria de hiperparámetros, cuyo desempeño final se evalúa sobre un conjunto de prueba independiente, complementando el análisis con métricas estándar de clasificación y técnicas de interpretabilidad basadas en importancia de características y valores SHAP.