Predicción de estancias hospitalarias en pacientes geriátricos en un hospital de cuarto nivel de complejidad de la ciudad de Medellín-Antioquia en los años 2021-2022

Vanessa Restrepo Correa
Astrid Viviana Sánchez Jiménez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
Maria Bernarda Salazar Sánchez, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Restrepo Correa & Sánchez Jiménez, 2023)
Referencia **Estilo APA 7 (2020)**	Restrepo Correa, V., & Sánchez Jiménez, A. V. (2023). *Predicción de estancias hospitalarias en pacientes geriátricos en un hospital de cuarto nivel de complejidad de la ciudad de Medellín-Antioquia en los años 2021-2022*. Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte V.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

Dedicamos este trabajo de grado a las siguientes personas, cuyo apoyo y aliento han sido esenciales en este viaje académico:

A nuestra familia, por su amor incondicional, paciencia y apoyo constante.

A nuestros profesores y asesora Maria Bernarda, por sus orientaciones para guiarnos.

A todos aquellos que, de una forma u otra, han contribuido a nuestro desarrollo académico y personal, les agradecemos de todo corazón.

.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

APA	American Psychological Association
MSE	Error Cuadrático Medio
Esp.	Especialista
UdeA	Universidad de Antioquia
PhD	Philosophiae Doctor

Resumen

Este trabajo se enfocó en definir una metodología de predicción de estancias hospitalarias en pacientes geriátricos de un hospital de cuarto nivel de complejidad en Medellín – Colombia. Se analizó un conjunto de datos que representa un total de 14385 egresos hospitalarios de población mayor de 59 años (75.3 ± 8.9 años) atendida durante el periodo comprendido entre el 1 de enero de 2021 al 31 de diciembre de 2022. Se analizaron 16 variables sociodemográficas y clínicas para un total de 14385 egresos; de los cuales 7727 (53.7%) son mujeres y el 46.3% restante corresponde a los hombres. Se utilizaron cuatro modelos de aprendizaje automático (Regresión Lineal, Regresión Logística, Arboles de decisión y Random Forest) evaluados con las métricas de error cuadrático medio, exactitud (Accuracy) y el coeficiente de determinación. Se obtuvo que el modelo de Random Forest es el más apropiado para este conjunto de datos, ya que supera una precisión de 0.41, explicando la mayor variabilidad en los datos. Tanto la Regresión Lineal como los modelos de Regresión Logística y Árboles de Decisión dado qué alcanzaron una precisión de 0.38, 0.12 y 0.40, respectivamente, lo que presenta desafíos significativos en el contexto de las estancias hospitalarias.

Repositorios GitHub:

- https://github.com/Vane1966/Clasificacion_estancias_geriatricas
- https://github.com/sanchezvivi01/Clasificacion_estancias_geriatricas

***Palabras clave*:** Estancia hospitalaria, Población geriátrica, Regresión lineal, Regresión logística, Arboles de decisión, Random Forest.

Abstract

This work focused on defining a methodology for predicting hospital stays in geriatric patients at a fourth level of complexity hospital in Medellín – Colombia. A set of data was analyzed that represents a total of 14,385 hospital discharges from a population over 59 years of age (75.3 ± 8.9 years) treated during the period from January 1, 2021 to December 31, 2022. 16 sociodemographic variables were analyzed. and clinics for a total of 14,385 discharges; of which 7727 (53.7%) are women and the remaining 46.3% correspond to men. Four machine learning models were used (Linear Regression, Logistic Regression, Decision Trees and Random Forest) evaluated with the metrics of mean square error, accuracy (Accuracy) and the coefficient of determination. It was found that the Random Forest model is the most appropriate for this data set, since it exceeds a precision of 0.41, explaining the greatest variability in the data. Both the Linear Regression and the Logistic Regression and Decision Tree models achieved a precision of 0.38, 0.12 and 0.40, respectively, which presents significant challenges in the context of hospital stays.

GitHub Repositories:

- https://github.com/Vane1966/Clasificacion_estancias_geriatricas
- https://github.com/sanchezvivi01/Clasificacion_estancias_geriatricas

***Keywords*:** Hospital stay, Geriatric population, Linear regression, Logistic regression, Decision trees, Random Forest.

1. Descripción del problema

La atención médica de la población geriátrica representa un desafío significativo en la actualidad debido al aumento constante de la longevidad en la sociedad. En este contexto, la gestión eficiente de los recursos hospitalarios se ha convertido en una prioridad, especialmente en lo que respecta a la planificación de las estancias hospitalarias de pacientes geriátricos.

La estancia hospitalaria se refiere al período de tiempo durante el cual un paciente permanece en hospitalización recibiendo atención médica y tratamiento, puede variar en duración según el diagnóstico de los pacientes, durante esta estancia hospitalaria los pacientes pueden recibir: atención médica y de enfermería, terapias especializadas, cirugías, administración de medicamentos y seguimiento constante de su estado de salud. Es fundamental destacar que la estancia hospitalaria tiende a extenderse significativamente debido a la complejidad clínica que estos individuos presentan. La población de adultos mayores se caracteriza por una mayor prevalencia de múltiples condiciones médicas crónicas, lo que implica un enfoque de atención más completo y a menudo más prolongado, la presencia de una enfermedad crónica puede influir en la progresión o el manejo de otra, lo que requiere una atención interdisciplinaria más intensiva por parte de un equipo médico.

Comúnmente la estancia hospitalaria se analiza a través de indicadores y factores que permiten evaluar la eficiencia y calidad de la atención médica en un hospital, tales como el porcentaje de ocupación o el giro cama. Estos indicadores permiten establecer la duración del tiempo que un paciente pasa en el hospital desde su admisión hasta su alta, para evidenciar las estancias prolongadas (mayor o igual a 15 días) o los pacientes hiperfrecuentadores.

Problema de negocio

La predicción de la duración de la estancia hospitalaria en pacientes geriátricos es un asunto de gran importancia en los ámbitos de la atención médica y la gerontología. Este desafío se enfoca en el desarrollo de un modelo predictivo que permita estimar la estadía de los pacientes de edad

avanzada en un entorno hospitalario de alta complejidad. La problemática se puede desglosar en los siguientes aspectos claves:

- **Envejecimiento de la población.** A medida que la población envejece, la cantidad de pacientes geriátricos que requieren atención hospitalaria aumenta, lo que genera una mayor presión sobre los recursos y la capacidad de los sistemas de salud.
- **Gestión eficiente de recursos.** La duración de la estancia hospitalaria es un factor crítico en la gestión eficiente de los recursos hospitalarios, como camas, personal médico y recursos financieros.
- **Calidad de atención y seguridad del paciente.** La duración excesiva de la estancia hospitalaria puede aumentar el riesgo de complicaciones, infecciones nosocomiales y otros problemas de salud. Por otro lado, dar de alta prematuramente a un paciente puede resultar en una atención deficiente y un mayor riesgo de reingreso.

En el contexto del envejecimiento de la población y la demanda de atención médica eficiente y de alta calidad, predecir la duración de las estancias hospitalarias en pacientes geriátricos se convierte en un desafío crucial. Este trabajo de monografía tiene como objetivo explorar en detalle estos desafíos y propone un enfoque analítico que considera múltiples factores. Se investigará la duración de la estancia hospitalaria teniendo en cuenta variables como la edad, el estado de salud, la presencia de comorbilidades y los procedimientos quirúrgicos.

1.1. Aproximación desde la analítica de datos

La investigación en este campo se centra en el desarrollo de modelos de predicción de la duración de las estancias hospitalarias en la población geriátrica. Para lograr esto, se utilizan técnicas avanzadas de aprendizaje automático, que incluyen, entre otras, árboles de decisión, bosques aleatorios, regresión lineal y regresión logística. Estas herramientas permiten analizar y predecir la duración de la estadía hospitalaria de manera más precisa y efectiva en el contexto de pacientes de edad avanzada. La búsqueda de modelos que permitan predecir la estancia hospitalaria tiene como objetivo generar un impacto significativo en diversas áreas de las instituciones de salud. Estos modelos ofrecen una serie de beneficios generales:

- **Gestión eficiente de recursos.** Permite asignar de manera óptima sus recursos, como camas, personal médico, quirófanos y suministros, al prever la duración de las estancias, los hospitales pueden programar procedimientos y admisiones de pacientes de manera más efectiva.
- **Reducción de costos y optimización de los recursos financieros.** Al optimizar la utilización de recursos y reducir las estancias prolongadas de pacientes geriátricos, los hospitales pueden disminuir los costos operativos y mejorar la eficiencia en la atención médica.
- **Planificación del alta hospitalaria.** Ayudar a los profesionales de la salud a planificar el alta de los pacientes geriátricos de manera más efectiva, coordinando la transición a cuidados post - hospitalarios, como cuidados en el hogar o centros de rehabilitación.

El modelo propuesto en este proyecto pretende apoyar la toma de decisiones en la gestión de pacientes geriátricos y, por consecuente en la atención médica y gestión de los recursos de la institución de salud.

1.2. Origen de los datos

El conjunto de datos se recopiló de forma retrospectiva y representa el total de los egresos hospitalarios de la población mayor de 59 años de edad (75.3 ± 8.9) atendida en un hospital de cuarto nivel de complejidad de la ciudad de Medellín durante el periodo comprendido entre el 1 de enero de 2021 al 31 de diciembre de 2022. El conjunto de datos cuenta 16 variables sociodemográficas y clínicas para un total de 14385 egresos; de los cuales 7727 (53.7 %) son mujeres y el 46.3 % restante corresponde a los hombres.

1.3. Métricas de desempeño y de error

Se utilizó la biblioteca Scikit-learn, desarrollada en Python, la cual brinda una interfaz sencilla y coherente, permitiendo desarrollar tareas de machine learning como la clasificación y la regresión. Esta biblioteca ofrece métricas de desempeño y herramientas de validación cruzada que permiten evaluar la calidad de los modelos y realizar ajustes según sea necesario (Pedregosa, 2011).

1.3.1. Exactitud (Accuracy)

Esta métrica es una medida fundamental en problemas de clasificación como el que se está abordando, calcula la proporción de predicciones correctas hechas por cada modelo implementado en relación con el total de predicciones realizadas. En otras palabras, se trata de la fracción de muestras clasificadas correctamente. La fórmula para calcular la exactitud es:

$$Exactitud = \frac{Numero\ de\ predicciones\ correctas}{Total\ de\ predicciones\ realizadas}$$

Es importante tener en cuenta, que la exactitud en clases desequilibradas, donde una clase es mucho más común que la otra, puede dar una impresión engañosa del rendimiento del modelo.

1.3.2. Error Cuadrático Medio (MSE)

Se utiliza para evaluar la calidad de los modelos de regresión al medir la diferencia cuadrada entre los valores predichos por el modelo y los valores reales. El cálculo del MSE se basa en la diferencia entre las predicciones del modelo y los valores observados (también conocidos como valores reales) para cada punto de datos. Luego, se eleva al cuadrado esta diferencia para asegurarse de que los valores negativos no cancelen los valores positivos, y se promedian todas estas diferencias al cuadrado (Bishop, 2006). Cuanto menor sea el MSE, mejor será el rendimiento del modelo. Se representa a través de la siguiente ecuación:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(yi - \hat{y}i)^2$$

Dónde,

N Es el número total de observaciones o puntos de datos.

yi Representa el valor observado (o real) para la i-ésima observación.

$\hat{y}i$ Representa el valor predicho por el modelo para la i-ésima observación.

$\sum$ Representa la suma sobre todos los datos, desde i = 1 hasta n.

1.3.3. Coeficiente de Determinación (R^2)

Se utiliza para evaluar cuánto de la variabilidad de una variable dependiente puede explicarse por una o más variables independientes en un modelo. El R^2 varía entre 0 y 1, donde 0 indica que el modelo no explica ninguna de la variabilidad de la variable dependiente, y 1 indica que el modelo explica toda la variabilidad. Un valor de R^2 cercano a 1 implica que el modelo se ajusta bien a los datos, mientras que un valor cercano a 0 sugiere que el modelo no es adecuado para explicar la variabilidad (Johnson, 2007).

2. Objetivos

2.1.Objetivo general

Definir una metodología de predicción de estancias hospitalarias en pacientes geriátricos de un hospital de cuarto nivel de complejidad en Medellín - Colombia.

2.2.Objetivos específicos

1. Seleccionar la población objetivo y las variables representativas que permitan predecir con mayor exactitud la estancia hospitalaria de los pacientes geriátricos.
2. Proponer y evaluar una estrategia de predicción de estancias hospitalarias basada en técnicas de aprendizaje automático.
3. Validar la capacidad de predicción del modelo a partir de criterios de exactitud y desempeño, usando información de una institución prestadora de servicios de salud de cuarto nivel en Medellín- Colombia.

3. Datos

3.1. Datos originales

Este conjunto de datos contiene la información de los pacientes mayores de 59 años internados en un hospital de cuarto nivel de complejidad de la ciudad de Medellín desde el año 2021 hasta el año 2022. Contiene 16 variables de interés las cuales ayudarán con la predicción del número de días de estancia que estará el paciente internado desde el ingreso hasta el egreso de la hospitalización. El conjunto de datos abarca varios aspectos, como los datos generales del paciente, que incluyen edad y sexo, así como información clínica, que comprende diagnósticos y procedimientos tanto quirúrgicos como no quirúrgicos y la variable de salida que son los días de estancia del paciente. Las variables o características que contiene el conjunto de datos están relacionadas en la Tabla 1. Los datos no registraban información faltante en ninguna de las variables analizadas.

Tabla 1. Naturaleza de las variables del dataset

Nombre de la variable	**Descripción de la variable**	**Naturaleza/Tipo variable**
Episodio	Código único brindado al paciente al momento de la atención recibida.	Categórico/Nominal
Edad	Edad del paciente adulto mayor al momento del egreso hospitalario	Cuantitativa/Razón
Sexo	Genero del paciente atendido en la institución	Categórico/Nominal
Diagnóstico principal Egreso	Este es el diagnóstico principal con el que paciente egresa del hospital según codificación internacional de enfermedades CIE - 10	Categórico/Nominal
Cantidad comorbilidades	Corresponde al número de enfermedades o antecedentes que tiene el paciente adicional al diagnóstico principal por el cual está siendo atendido.	Cuantitativa/Razón
Tuvo cx	Denota si al paciente le realizaron un procedimiento quirúrgico durante su periodo de hospitalización.	Categórico/Nominal
Procedimiento	Esta característica nos indica si el paciente tuvo un procedimiento no quirúrgico durante su estancia en hospitalización	Categórico/Nominal
Ventilación Mecánica	Nos indica si el paciente requirió ventilación mecánica en su estancia.	Categórico/Nominal
Situación al alta	Es el estado del paciente al alta	Categórico/Nominal

UCI	Indica si el paciente estuvo un tiempo hospitalizado en una unidad de cuidado intensivo (UCI)	Categórico/Nominal
UCE	Indica si el paciente estuvo un tiempo hospitalizado en una unidad de cuidado especial (UCE)	Categórico/Nominal
Nombre Especialidad Egreso	Es el nombre de la especialidad médica que trata al paciente durante su estancia y que da el egreso administrativo al mismo.	Categórico/Nominal
Descripción empresa	Indica el nombre de la empresa o entidad que fue el principal pagador de la hospitalización del paciente.	Categórico/Nominal
Unidad Hospitalización	Es el nombre de la unidad de hospitalización de donde egreso el paciente al momento de su alta.	Categórico/Nominal
Estancia	Corresponde al número de días que el paciente pasa hospitalizado en alguna de las unidades. (variable respuesta)	Cuantitativa/Razón
Tipo Estancia	Clasificación de la estancia según el número de días que el paciente se quedó hospitalizado (Mayor o igual a 15 Estancia prolongada, de 0 a 15 estancia no prolongada	Categórico/Nominal

Es importante destacar que el conjunto de datos utilizado en este estudio es de naturaleza privada, ya que contiene información real de los pacientes atendidos. Por lo tanto, en el proceso de presentación de resultados, se enfatiza que en ningún momento se hará mención o referencia en nombre de los pacientes involucrados. Los resultados se presentarán a nivel global del modelo, sin revelar información personal o identificable de los pacientes.

3.2. Datasets

Para evitar sesgos en los conjuntos de entrenamiento y validación, se realizó una división aleatoria de los datos. Esta estrategia garantiza que los datos en ambos conjuntos sean representativos de la población total. La división de los datos se llevó a cabo asignando el 70 % de los datos al conjunto de entrenamiento y el 30 % al conjunto de prueba, utilizando la función *'train_test_split'* de la biblioteca scikit-learn.

3.3. Analítica descriptiva

El dataset contiene la información de los pacientes mayores de 59 años internados en un hospital de cuarto nivel de complejidad de la ciudad de Medellín. Tiene los datos de 14385 egresos hospitalarios desde el 1 de enero de 2021 hasta el 31 de diciembre de 2022. Incluye 16 variables de interés las cuales ayudarán con la predicción de la estancia hospitalaria. Incluye datos demográficos del paciente como edad y sexo; información clínica como diagnósticos, procedimientos quirúrgicos y no quirúrgicos; y contiene una variable respuesta que representa la estancia, es decir, el tiempo que un paciente permaneció en internado en un servicio de hospitalización.

La edad promedio de los pacientes es de 75 años y el número de comorbilidades son en promedio 5 por paciente. Explorando la distribución de las variables categóricas se observa que el mayor porcentaje de pacientes pertenece al género femenino (53.7 %), aunque la diferencia no es muy significativa con relación a los pacientes del género masculino. Por otro lado, se evidencia que un porcentaje pequeño de los pacientes (28.9 %) tuvieron un procedimiento quirúrgico durante su estancia. Contrario a esto, cuando se habla de procedimientos no invasivos o no quirúrgicos se encuentra que el 94.3 % de los pacientes durante su estancia les realizaron por lo menos un procedimiento de este tipo.

La ventilación mecánica y las estancias en unidades de cuidado critico por su parte son importantes en la permanencia hospitalaria debido a que aumentan la complejidad del paciente. En este caso, el 9.4 % de los pacientes fueron sometidos a ventilación mecánica durante su proceso de hospitalización, siendo consecuente al momento de analizar las variables que indican si el paciente necesitó unidad de cuidados intensivos o unidad de cuidados especial en su estancia, se evidencia que también el porcentaje es pequeño (15.1 %) frente al total de los datos.

4. Proceso de analítica

4.1. Pipeline principal

A continuación, se abordan cada una de las fases del esquema presentado en la Figura 1.

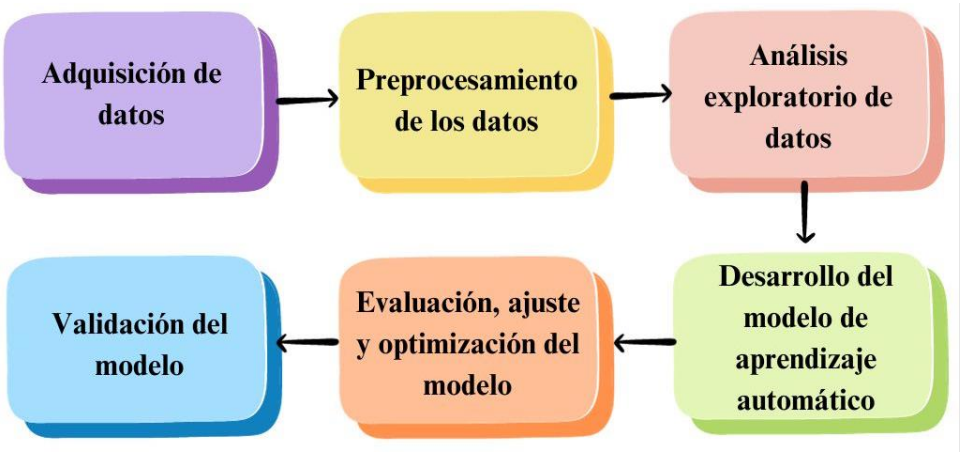


Figura 1. Esquema metodológico implementado durante la ejecución del proyecto

Adquisición de datos: En esta etapa se selecciona el conjunto de datos para el entrenamiento y la evaluación de un modelo que tiene como fin la predicción del tipo de estancias hospitalarias de un paciente geriátrico en un hospital de cuarto nivel de complejidad de la ciudad de Medellín. Este ejercicio se realiza como una necesidad interna de un área del hospital con un propósito académico que tiene como fin el aprendizaje, no la investigación clínica o el uso comercial de datos.

Preprocesamiento de los datos: En esta fase se realiza la preparación de los datos para su uso en el modelado de machine learning, lo que incluyó identificación de inconsistencias en los datos, como datos faltantes, valores atípicos y duplicados, selección de características relevantes, y la transformación de datos según aplique para variables categóricas.

Análisis exploratorio de datos: En esta etapa se busca comprender la distribución y características de la información, se resumieron los datos de manera concisa a través de medidas como la media, la mediana, la moda, la desviación estándar y otros resúmenes estadísticos que ayudan a entender la tendencia central, la dispersión y la forma de la distribución de los datos. Se analizaron la distribución de las variables, lo que ayuda a comprender características claves y proporcionar una visión general de la información contenida en los datos. También se realiza una exploración de la correlación de estas con la variable respuesta para determinar su significancia en el ejercicio planteado. Estas estadísticas son fundamentales para la toma de decisiones, la comunicación de resultados y la comprensión de patrones y tendencias en los datos.

Desarrollo del modelo de aprendizaje automático: Se configuran modelos de clasificación para la predicción del tipo de estancia hospitalaria, tales como: Regresión lineal, regresión logística, árboles de decisión y random forest; los cuales permiten a modelar relaciones entre variables y predecir resultados en función de los datos disponibles. Teniendo en cuenta que cada modelo tiene sus ventajas y desventajas, se selecciona el más adecuado para la predicción del tipo de estancias hospitalarias.

Evaluación, ajuste y optimización del modelo: En esta etapa se calculan métricas de evaluación, como; el error cuadrático medio (MSE) y el coeficiente de determinación R^2 teniendo en cuenta que el MSE se centra en la precisión de las predicciones individuales, mientras que el R^2 se enfoca en cuánta variabilidad se explica en los datos. Se lleva a cabo la fase de entrenamiento del modelo utilizando el conjunto de datos de entrenamiento, con el fin de permitir que el modelo adquiera la capacidad de realizar predicciones precisas. Además, se realiza la optimización de los hiperparámetros del modelo, lo que involucra un análisis y ajuste de los parámetros para mejorar significativamente el rendimiento y ajustarlos de manera óptima al conjunto de datos de la predicción del tipo de estancias hospitalarias. Este proceso de afinamiento de hiperparámetros se realiza con el objetivo de maximizar la capacidad predictiva del modelo y lograr que se adapte de manera óptima a las características específicas del conjunto de datos, lo que contribuye a su capacidad de generalización y, en última instancia, a la precisión de las predicciones.

Validación del modelo: En esta etapa se busca que el modelo sea capaz de generalizar y que se propicie el sobreajuste a los datos de entrenamiento. Mediante el conjunto de entrenamiento y validación se mide la capacidad del modelo para generalizar a datos independientes. La validación implica evaluación, pero va más allá al incluir técnicas para garantizar que el modelo no esté sobreajustado y sea adecuado para el uso en las situaciones del mundo real.
Por último, se interpretan los resultados y realizan recomendaciones basadas en los hallazgos obtenidos con el modelo.

4.2.Estrategias de preprocesamiento

En el proceso de preparación del conjunto de datos, se llevó a cabo una selección de características que no aportarían al modelo según conocimiento del negocio, y se continua con que aquellas que contribuyen a los objetivos de la investigación. Además, se realiza una búsqueda de caracteres especiales o datos nulos para determinar si era necesario un método de imputación de datos, para el caso no lo fue. Se realiza una recategorización de las variables en función de su naturaleza, lo que permitió una representación más precisa y adecuada para el análisis posterior. En el análisis exploratorio de los datos se calculan medidas de tendencia central para las variables numéricas y se grafica su comportamiento.

4.3. Modelos implementados

Antes de configurar los modelos, se realizó un proceso de codificación de variables categóricas como el diagnóstico, se utilizan dos enfoques clave: la técnica "*Get Dummies"* también conocida como codificación "*one-hot encoding"* y la técnica de "Label Encoder" o codificación de etiquetas. La codificación one-hot se emplea cuando se busca representar cada categoría de una variable como un vector binario, donde solo una posición es '1' (indicando la presencia de la categoría) y las demás son '0'. Esta técnica es útil para variables categóricas con múltiples categorías no ordenadas, ya que evita la introducción de un orden artificial. Por otro lado, la codificación de etiquetas asigna un valor numérico único a cada categoría de la variable. A menudo se utiliza cuando las categorías de la variable tienen un orden intrínseco o cuando el número de

categorías es grande. A diferencia de la codificación one-hot, esta técnica no crea dimensiones adicionales.

Posterior, se configuraron los siguientes modelos para predecir los días de estancia hospitalaria: Regresión Lineal, Regresión Logística, Árboles De Decisión y Random Forest. Los modelos antes mencionados se iteraron con ambas técnicas de codificación de variables. A continuación, se describen cada uno de los modelos utilizados y posteriormente su configuración:

4.3.1. Regresión lineal

La regresión lineal es una técnica fundamental en estadística y análisis de datos que se utiliza para modelar la relación entre una variable dependiente (o variable objetivo) y una o más variables independientes. Su objetivo principal es predecir o estimar el valor de la variable dependiente en función de las variables independientes mediante una ecuación lineal, se utiliza un método que minimiza la diferencia entre los valores predichos por el modelo y los valores reales observados. Esto se logra ajustando los coeficientes de la ecuación lineal a través de técnicas como el método de los mínimos cuadrados. Un modelo de regresión lineal incluye la elección de variables independientes relevantes, la identificación de la variable dependiente y la estimación de los coeficientes de la ecuación lineal. También se deben considerar aspectos como la normalización de datos, la selección de características y la evaluación de la calidad del modelo mediante métricas como el R^2 y el MSE (Montgomery, 2012).

4.3.2. Regresión logística

Se basa en el principio de que la variable dependiente (o variable objetivo) sigue una distribución de probabilidad logística. A través de un modelo matemático, se estima la probabilidad de que una observación pertenezca a una categoría en función de una serie de variables independientes. El modelo utiliza la función logística para transformar una combinación lineal de las variables independientes en una probabilidad en el rango de 0 a 1. Es esencial seleccionar las variables independientes relevantes, que pueden ser numéricas o categóricas, y se deben preparar los datos para el análisis. El ajuste del modelo implica estimar los coeficientes que ponderan la contribución de cada variable independiente en la predicción de la probabilidad. Esto se logra mediante técnicas de optimización, como la máxima verosimilitud. Una vez configurado, el modelo

puede clasificar nuevas observaciones en una de las categorías de interés en función de la probabilidad estimada (Kleinbaum, 2010).

4.3.3. Arboles de decisión

Se utilizan para tomar decisiones basadas en un conjunto de reglas y atributos. En su forma más simple, un árbol de decisión se asemeja a un diagrama de flujo con nodos internos que representan pruebas sobre atributos, ramas que conectan los nodos y hojas que representan las decisiones. El proceso comienza en la raíz y se desplaza hacia abajo en el árbol siguiendo las ramas apropiadas según las reglas definidas por los atributos. Cada nodo y rama del árbol se basa en decisiones condicionales, lo que permite dividir el conjunto de datos en subconjuntos. La configuración de un árbol de decisión implica la selección de atributos relevantes y la definición de reglas que optimicen la capacidad del modelo para clasificar o predecir datos (Tan, 2006).

4.3.4. Random forest

La técnica de Random Forest genera bosques aleatorios, los cuales son un poderoso algoritmo de aprendizaje automático que combina múltiples árboles de decisión para mejorar la precisión y la generalización del modelo. Cada árbol en el bosque se construye utilizando un subconjunto aleatorio de datos y atributos, lo que reduce el sobreajuste y aumenta la robustez del modelo. Durante la predicción, cada árbol emite una predicción y, finalmente, se realiza una agregación de estas predicciones para obtener un resultado final. Los Random Forest son altamente efectivos para tareas de clasificación y regresión, y son capaces de manejar conjuntos de datos grandes y complejos. La configuración de un Random Forest implica la selección de hiperparámetros importantes, como la cantidad de árboles en el bosque y la profundidad máxima de cada árbol (Breiman, 2001).

4.3.5. Configuración de modelos con la técnica “Get Dummies”:

Se cuenta con un dataset que contiene las variables relevantes para predecir la duración de la estancia hospitalaria. Las variables independientes (características) se almacenan en X, y la variable dependiente (la duración de la estancia) se almacena en Y. Los datos se dividen en conjuntos de entrenamiento y prueba. El 70 % de los datos se utilizarán para entrenar los modelos

(X_train, y_train), y el 30 % restante se utilizará para probar los modelos (X_test, y_test). Los cuatro modelos definidos se configuran con los hiperparámetros por defecto (parámetros específicos del modelo) y se inicializan las puntuaciones para evaluar su rendimiento. Se ajusta el modelo a los datos de entrenamiento.

Se realizan predicciones en el conjunto de prueba utilizando validación cruzada de 5 divisiones (cross-validation) para evaluar el rendimiento del modelo y se calculan diversas métricas, como el error cuadrático medio (MSE), la precisión media y el coeficiente de determinación (R^2) y se comparan las métricas para cada modelo.

4.3.6. Configuración de los modelos con la técnica "Label Encoder":

Se crean dos conjuntos, X y Y, donde X contiene las características o variables independientes que se utilizarán para predecir la variable objetivo y, que en este caso representa la duración de la estancia hospitalaria. Los datos se dividen en conjuntos de entrenamiento y prueba usando la función train_test_split. En este caso, el 70 % de los datos se utilizarán para entrenar los modelos (X_train y y_train), y el 30 % restante se utilizará para probar los modelos (X_test y y_test). Para el entrenamiento de modelos, se itera a través de un diccionario llamado models que contiene diferentes modelos de regresión, como Regresión Lineal, Regresión Logística, etc. Para cada modelo, se hace lo siguiente:

- El modelo se ajusta a los datos de entrenamiento utilizando model["model"].fit(X_train, y_train).
- Se realizan predicciones en el conjunto de prueba (y_pred = model["model"].predict(X_test)).
- Se realiza una validación cruzada con 5 divisiones (cross-validation) para evaluar el rendimiento del modelo y se calculan métricas como el error cuadrático medio (MSE), la precisión media y el coeficiente de determinación (R^2).

Estas métricas se almacenan en el diccionario model["scores"]. Después de entrenar los modelos y calcular las métricas de desempeño, se generan dos gráficos de barras utilizando *Matplotlib* para comparar las métricas de MSE (error cuadrático medio) y R^2 (coeficiente de determinación) de los diferentes modelos. Estos gráficos permiten visualizar cuál de los modelos tiene el mejor desempeño en términos de estas métricas.

4.3.7. Ajuste de hiperparámetros para cada modelo

Se realiza una evaluación automatizada de varios modelos de aprendizaje automático a través de un proceso de búsqueda de hiperparámetros y validación cruzada. Para cada modelo; Regresión Lineal, Regresión Logística, Arboles De Decisión y Random Forest, se buscan los mejores hiperparámetros utilizando la función *GridSearchCV*, imprime los resultados, genera predicciones en un conjunto de prueba, calcula métricas de rendimiento y crea gráficos de dispersión para visualizar las predicciones en comparación con los valores reales.

5. Metodología

5.1. Baseline

En el primer ejercicio realizado se consideraron 42 características y una variable de respuesta llamada “estancia” (cantidad de días de estancia). Para este primer acercamiento se realiza un preprocesamiento de los datos y se aplica un modelo Random Forest realizando una iteración para el mismo.

Se evidenciaron algunos inconvenientes en la preparación de los datos; atípicos y variables con múltiples categorías. Los datos atípicos fueron corregidos a través de la técnica de valores atípicos de Tukey y con le técnica de variables Get Dummies otorga un peso estándar a todas las variables, con el objetivo de minimizar el error en la predicción del modelo.

5.2. Validación

Para la primera iteración realizada en el modelo de Random Forest se utilizó una partición de los datos de 80/20, se escalan las variables y se aplica el modelo con 100 árboles y 5 niveles de profundidad. Para la validación se calcula el RMSE y el R^2 donde se evidenció que los resultados no fueron los esperados (RMSE: 4.27). Por lo tanto, se realizó una prueba de significancia de características, donde se identificó que efectivamente varias de las características ingresadas no otorgaban valor a la predicción. Por lo tanto, se planteó para el ejercicio actual realizar una selección de características en el preprocesamiento y buscar los mejores hiperparámetos.

5.3. Herramientas utilizadas

Se empleó Google Colab como entorno de desarrollo y ejecución. Google Colab es una plataforma en línea que permite ejecutar código en Python de manera interactiva y colaborativa, lo que resultó ser una elección ideal para la implementación. En el proceso de desarrollo, se hizo uso de varias bibliotecas esenciales en el ámbito de la ciencia de datos y el aprendizaje automático. Entre las herramientas más destacadas, se incluyeron las bibliotecas de Pandas, Numpy, Sklearn y Matplotlib.

Pandas se utilizó por su eficiencia en la manipulación y análisis de datos, lo que facilitó la carga y la preparación del conjunto de datos. Numpy, por su parte, desempeñó un papel crucial en las operaciones numéricas y matriciales necesarias para las tareas de procesamiento y modelado de datos.

La librería Sklearn, también conocida como scikit-learn, proporciona una amplia gama de herramientas para el aprendizaje automático y la minería de datos, lo que permitió implementar y evaluar modelos predictivos de manera eficaz. Con Sklearn, se realizaron tareas de preprocesamiento de datos, entrenamiento de modelos y evaluación de su rendimiento de manera eficiente. Además, Matplotlib desempeñó un papel crucial en la visualización de resultados y la creación de gráficos informativos.

6. Resultados y discusión

El dataset cuenta con 14385 egresos de pacientes mayores a 59 años de un Hospital de cuarto nivel de complejidad de la ciudad de Medellín. Cuenta con 16 variables, las cuales están dividas entre numéricas y categóricas. A continuación, se presenta la tabla con el análisis general de las variables numéricas: edad, cantidad de comorbilidades y días de estancia (Ver Tabla 2).

Tabla 2. Descripción de las variables numéricas

Medidas	Edad	Cantidad de comorbilidades	Días de estancia
Promedio	75.4	9.8	9.1
Desviación estándar	8.9	5.2	10.5
Mínimo	60	0	0
Percentil 25	68	6	3.0
Mediana	74	9	6.0
Percentil 75	82	13	11.0
Máximo	106	20	179.0

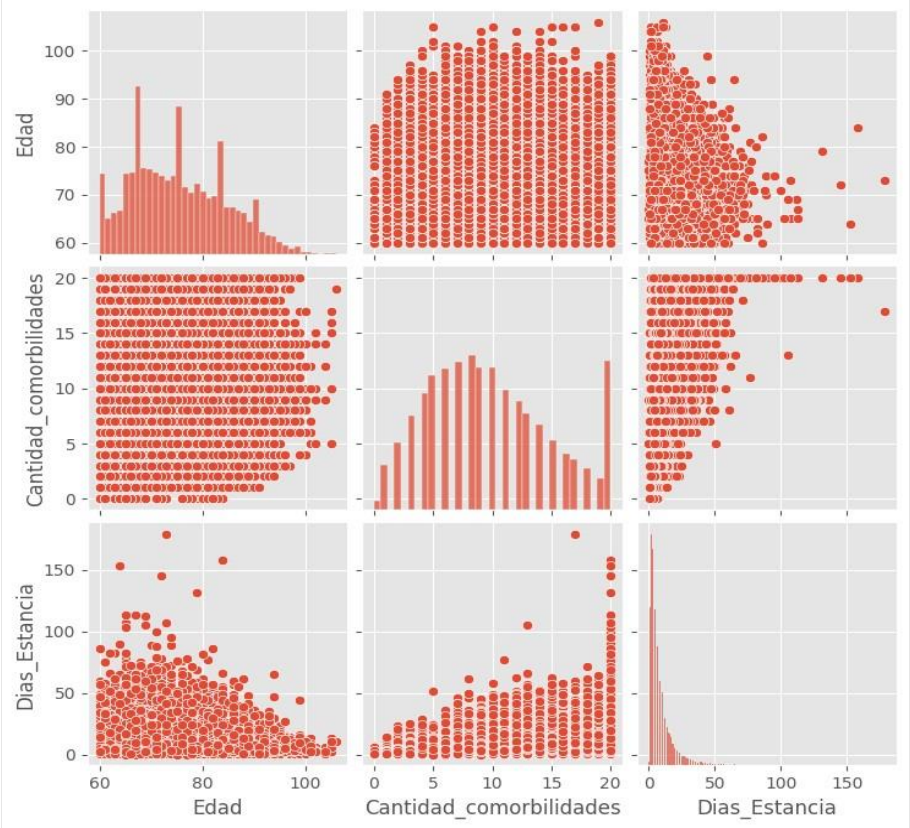
La variable edad muestra una distribución que oscila entre un valor mínimo de 60 años y un valor máximo de 106 años en la población estudiada. La edad promedio, representada por la media, es de aproximadamente 75.4 años. Los percentiles también son informativos: el 25% de los individuos tiene 68 años o menos, el 50% tiene 74 años o menos, y el 75% tiene 82 años o menos. Esto sugiere que la mayoría de la población tiene edades en el rango de 68 a 82 años.

La variable cantidad de comorbilidades refleja la presencia de condiciones médicas adicionales en cada individuo. La media de comorbilidades es de aproximadamente 9.8, lo que indica que, en promedio, las personas en este estudio tienen alrededor de 10 condiciones médicas adicionales. Sin embargo, es importante destacar que la desviación estándar es de aproximadamente 5.21, lo que sugiere una variabilidad significativa en la cantidad de comorbilidades. Esto podría indicar que algunas personas tienen menos comorbilidades, mientras que otras tienen un número considerablemente mayor.

La variable días de estancia representa la duración de la hospitalización; la media de esta variable es de alrededor de 9.1 días, con una desviación estándar de aproximadamente 10.5. Esto sugiere una variabilidad sustancial en la duración de las hospitalizaciones. Además, el rango de esta variable es bastante amplio, desde un mínimo de 0 días hasta un máximo de 179 días. Los percentiles indican que el 25% de las hospitalizaciones tienen una duración de 3 días o menos, el 50% de 6 días o menos, y el 75% de 11 días o menos.

Con relación a las variables anteriormente descritas, se puede observar que no existe una distribución normal de los datos, ni una correlación marcada entre ellas. Sin embargo, sí se observa una relación positiva entre el número de comorbilidades y el número de días de estancia del paciente; a medida que aumenta el número de comorbilidades, la cantidad de días estancia también incrementan (Ver **¡Error! No se encuentra el origen de la referencia.**).

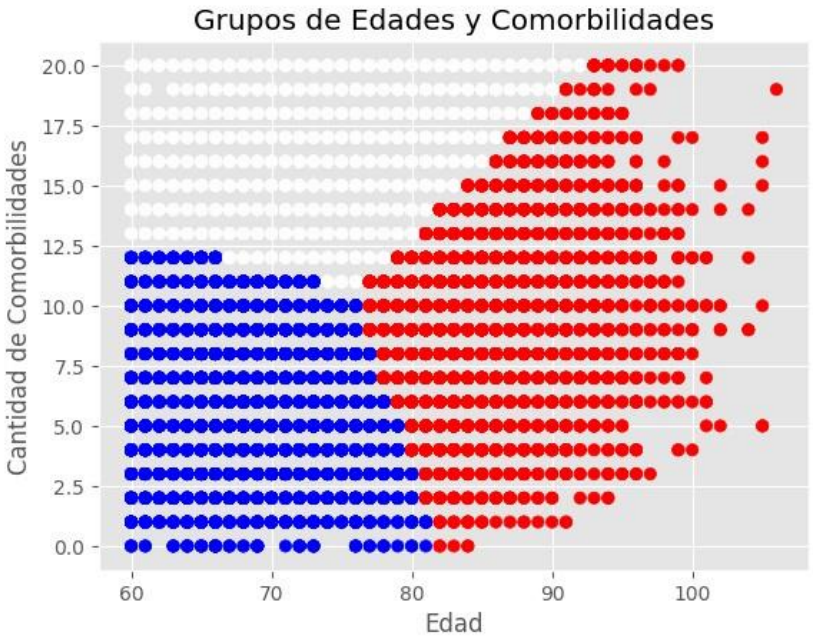
Figura 2. Correlación de variables numéricas



Analizando la relación entre la variable edad con cantidad de comorbilidades como se había mencionado no se observa una correlación clara entre ellas, pero si se puede observar cuando se agrupan los datos (ver Figura 3) que en general, parecería que la cantidad de comorbilidades aumenta con la edad. Los individuos de 70-80 años tienen, en promedio, más comorbilidades que los individuos de 60-70 años. Sin embargo, hay una excepción a esta tendencia. El grupo de 70-80 años con la menor cantidad de comorbilidades tiene un tamaño significativamente menor que los otros grupos. Esto sugiere que este grupo puede ser un grupo atípico.

La Figura 3 muestra que existe una gran variabilidad en la cantidad de comorbilidades dentro de cada grupo de edad. Algunos individuos tienen un número considerable de comorbilidades, mientras que otros tienen pocas. La edad puede ser un factor importante que influye en la cantidad de comorbilidades.

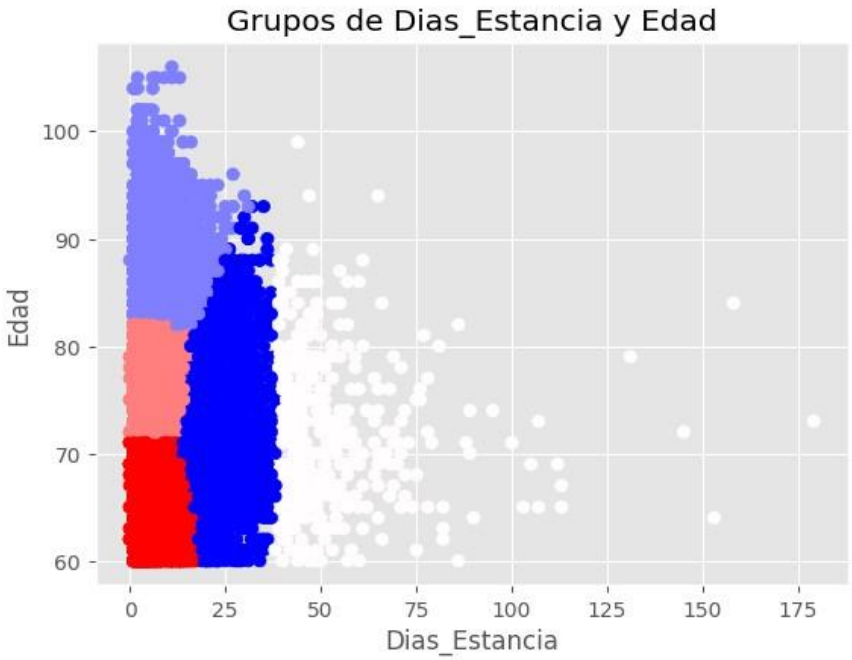
Figura 3. Correlación de grupos de edad y cantidad de comorbilidades



Al analizar la relación entre la edad y la cantidad de días de estancia (Ver Figura 4) se puede evidenciar que la mayor cantidad de pacientes están entre las edades de 60 y 90 años con un rango de días de estancia entre 0 y 30 aproximadamente. Además, se observa que pocos pacientes

tuvieron una estancia superior a 50 días y que la mayoría de los pacientes con una edad superior a 90 años tuvieron una estancia de 0 a 15 días.

Figura 4. Correlación de grupos de edad y días de estancia hospitalaria



Para continuar con el análisis del conjunto de datos y teniendo en cuenta los objetivos del proyecto, se aplicaron diversos modelos de regresión y clasificación con el objetivo de analizar y predecir los días de estancia hospitalaria, con base en dos técnicas; get dummies y label encoder para la variable categórica *Diagnósticos*. Los modelos utilizados incluyen regresión lineal, regresión logística, árboles de decisión y bosques aleatorios. Los resultados obtenidos con base en la configuración Get Dummies se pueden observar en la Tabla 3.

El modelo de regresión lineal muestra resultados sorprendentemente altos en términos de MSE y R^2, aunque esto podría ser indicativo de un sobreajuste o un problema con las características seleccionadas. Un MSE tan alto sugiere que la dispersión de los datos con respecto a la línea de regresión es extremadamente amplia, lo que puede ser problemático y la precisión media negativa es una señal de que el modelo no está realizando bien las predicciones, ya que una precisión negativa no tiene sentido en el contexto de la regresión. Esto sugiere que el modelo no es adecuado para el conjunto de datos o que se necesita una revisión exhaustiva del proceso de modelado.

Tabla 3. Resultados de las métricas para los modelos evaluados según la técnica Get Dummies

Modelos	**Error cuadrático medio**	**Exactitud (Accuracy)**	**Coeficiente de determinación**
Regresión lineal	5.95e+21	-4.72e+18	-5.18e+19
Regresión logística	76.50	0.14	0.33
Arboles de decisión	40.86	0.68	0.64
Random Forest	23.71	0.79	0.79

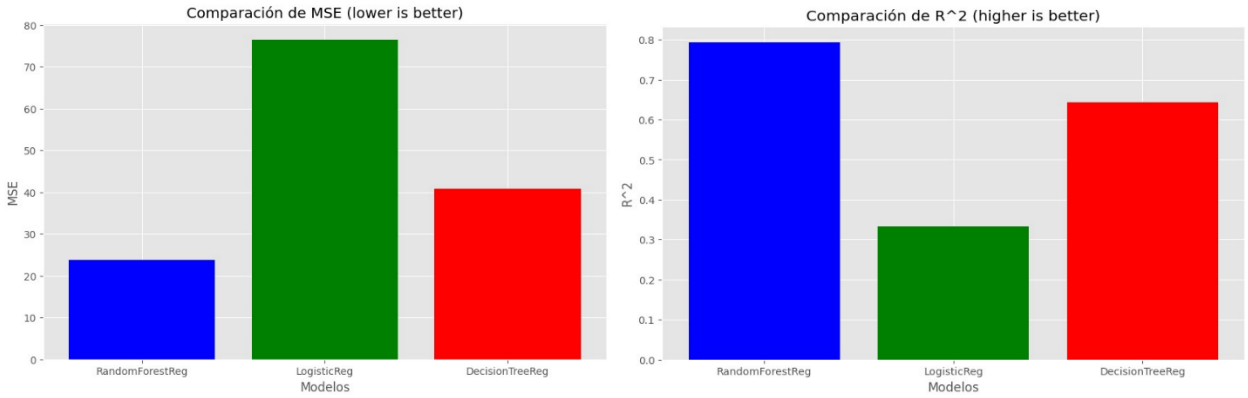
En contraste, el modelo de regresión logística muestra un MSE más bajo y una precisión media positiva. Esto sugiere que el modelo tiene un mejor rendimiento en comparación con la regresión lineal en este contexto. Sin embargo, el R^2 sigue siendo relativamente bajo, lo que indica que solo una fracción moderada de la variabilidad en los datos se puede explicar mediante este modelo. Por su parte, el modelo de árboles de decisión exhibe un mejor rendimiento, con un MSE moderado y una precisión media cercana a 0.68. Además, el R^2 sugiere que este modelo explica una parte significativa de la variabilidad en los datos. Los árboles de decisión son conocidos por su capacidad para capturar relaciones no lineales en los datos, lo que puede explicar su rendimiento mejorado en comparación con los modelos anteriores.

El modelo de random forest presenta mejores resultados con relación a los modelos anteriormente descritos, con un MSE considerablemente más bajo, una precisión media superior al 0.79 y un R^2 que supera el 0.79. Esto indica que el modelo de random forest es el más adecuado para predecir los días de estancia hospitalaria. El bajo MSE y el alto R^2 sugieren que este modelo es capaz de realizar predicciones precisas y explicar una gran parte de la variabilidad observada en los datos.

Considerando lo anterior, al representar gráficamente tanto el MSE como el R^2 de los modelos (ver Figura 6), se excluye el modelo de regresión lineal debido a la presencia de un valor extremadamente elevado en el MSE. Este valor atípico como se mencionó sugiere que el modelo

de regresión lineal podría no ser adecuado para el análisis de los datos en cuestión, ya que la magnitud de error es desproporcionadamente alta en comparación con los otros modelos evaluados.

Figura 5. Comparación de la capacidad de predicción de los modelos con la técnica Get Dummies. Lado izquierdo: MSE; lado derecho: R^2.



En este caso se encuentra que la regresión lineal genera un valor de MSE excesivamente alto, lo que indicaría que la adición de columnas a través de *GetDummies* es inconveniente para el funcionamiento de este modelo, pues agrega una cantidad de variables que son tenidas en cuenta y afectan los resultados predictivos.

A partir de los resultados obtenidos al evaluar los modelos, y teniendo en cuenta que la regresión lineal podría ser un modelo eficiente en caso de ejecutarse correctamente, se utilizará otra estrategia para la evaluación de los datos categóricos; utilizando *Label Encoder* se evitará la generación de columnas adicionales para los diagnósticos, en vez de eso se le asignará un valor numérico a cada uno de ellos. Se aplica esta técnica y se iteran nuevamente los modelos sin ajustar los hiperparámetros. Los datos se muestran en la Tabla 4.

Con el cambio de técnica en la codificación de variables, el modelo de Regresión Lineal muestra una mejoría con un MSE moderado y un R^2 positivo (ver Figura 6). El MSE indica que el modelo tiene una cantidad razonable de error en sus predicciones, y el R^2 sugiere que alrededor del 39 % de la variabilidad en los datos puede ser explicada por el modelo. Esto indica un rendimiento prometedor en términos de regresión lineal, lo que podría estar relacionado con una relación lineal razonable entre las variables independientes y dependientes. Por su parte, el modelo de Regresión

Logística muestra un MSE más elevado y un R^2 negativo. Estos resultados sugieren que este modelo no se ajusta adecuadamente a los datos en cuestión.

Tabla 4. Resultados de las métricas para los modelos evaluados según la técnica Label Encoder

Modelos	**Error cuadrático medio**	**Exactitud (Accuracy)**	**Coeficiente de determinación**
Regresión lineal	69.54	0.38	0.39
Regresión logística	135.03	0.12	-0.17
Arboles de decisión	119.50	-0.01	-0.04
Random Forest	68.22	0.41	0.40

El MSE alto indica una discrepancia significativa entre las predicciones del modelo y los valores observados. El R^2 negativo sugiere que el modelo tiene un rendimiento deficiente en la explicación de la variabilidad en los datos. El modelo de Árboles de Decisión también muestra resultados problemáticos, con un MSE alto y un R^2 negativo. Estos indicadores sugieren que este modelo no es adecuado para el conjunto de datos y no está realizando predicciones precisas a la estancia hospitalaria. Por último, el modelo de Random Forest presenta resultados más prometedores, con un MSE moderado y un R^2 positivo, esto indica que el modelo es capaz de realizar predicciones más precisas y explicar una parte significativa de la variabilidad en los datos, superando a los modelos anteriores.

Al realizar una nueva iteración de los modelos, pero esta vez buscando los mejores hiperparámetros, se procede a la validación y ajuste de los modelos previamente mencionados, incorporando la optimización de los hiperparámetros. Los resultados de dicha validación son los relacionados en la Tabla 5.

Figura 6. Comparación de la capacidad de predicción de los modelos con la Label Encoder. Lado izquierdo: MSE; lado derecho: R^2.

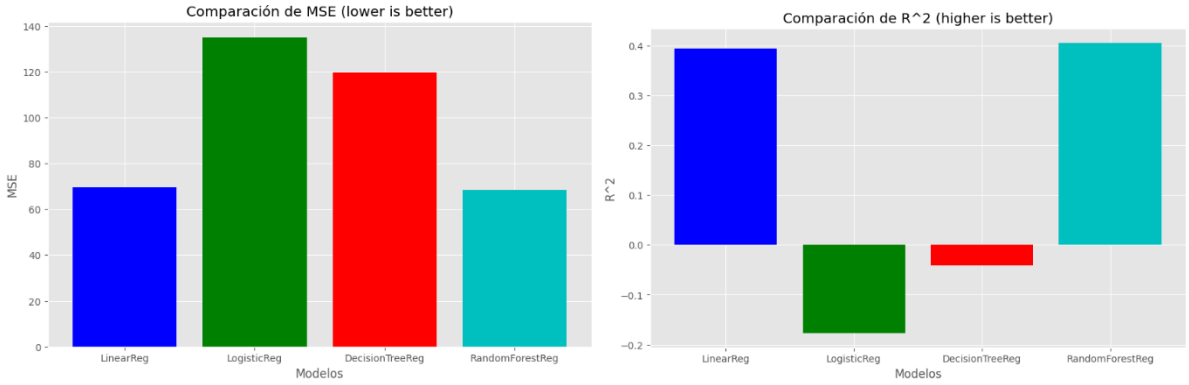


Tabla 5. Resultados de las métricas para los modelos evaluados con los mejores hiperparámetros

Modelos	**Error cuadrático medio**	**Exactitud (Accuracy)**	**Coeficiente de determinación**
Regresión lineal	69.55	0.380	0.394
Regresión logística	136.71	0.128	-0.191
Arboles de decisión	64.16	0.404	0.441
Random Forest	67.89	0.410	0.408

El modelo de Regresión Lineal, después de la validación de hiperparámetros, continúa mostrando un MSE moderado y un R^2 positivo. Estos resultados indican un rendimiento razonable en términos de regresión lineal, lo que podría sugerir una relación lineal apropiada entre las variables independientes y dependientes (Ver Figura 7.

En el caso del modelo de Regresión Logística, la validación de hiperparámetros ha generado un aumento en el MSE y un R^2 negativo, lo cual sugiere que este modelo no se adapta adecuadamente a los datos. Los valores altos de MSE y R^2 negativo indican discrepancias significativas entre las predicciones del modelo y los valores observados, lo que lleva a la conclusión de un rendimiento deficiente en la explicación de la variabilidad en los datos. (Ver Figura 8. .

Figura 7. Grafica de ajuste de entrenamiento y prueba del modelo Regresión Lineal

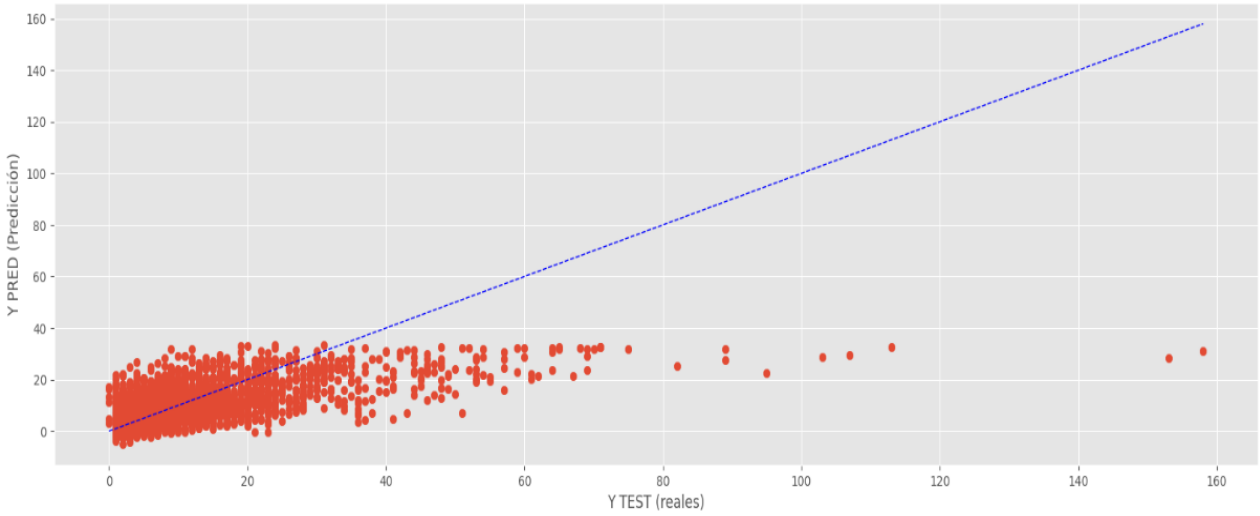
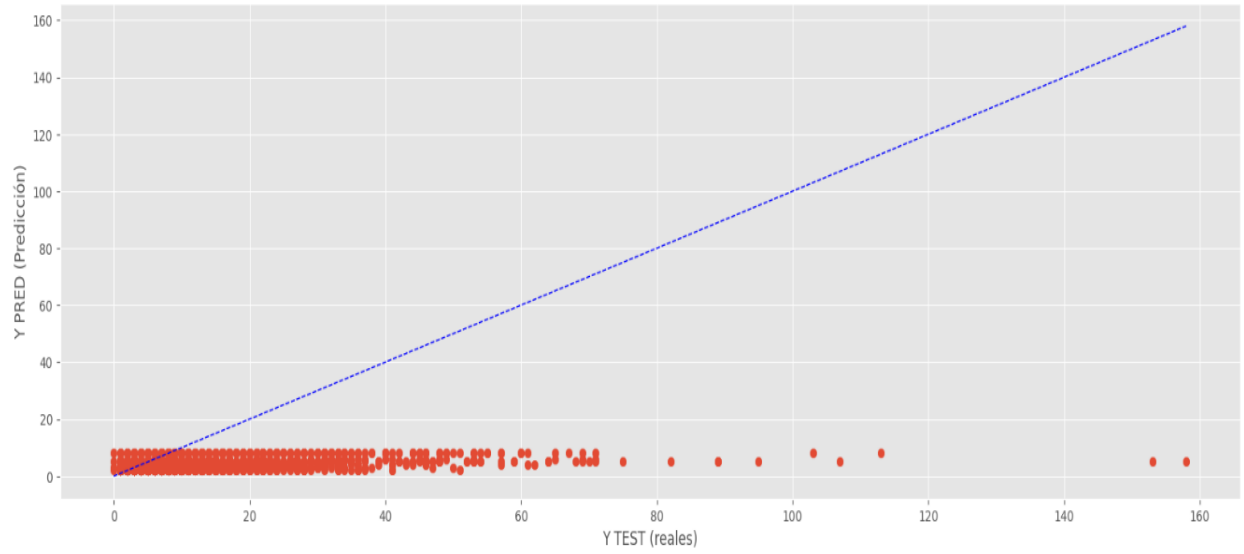
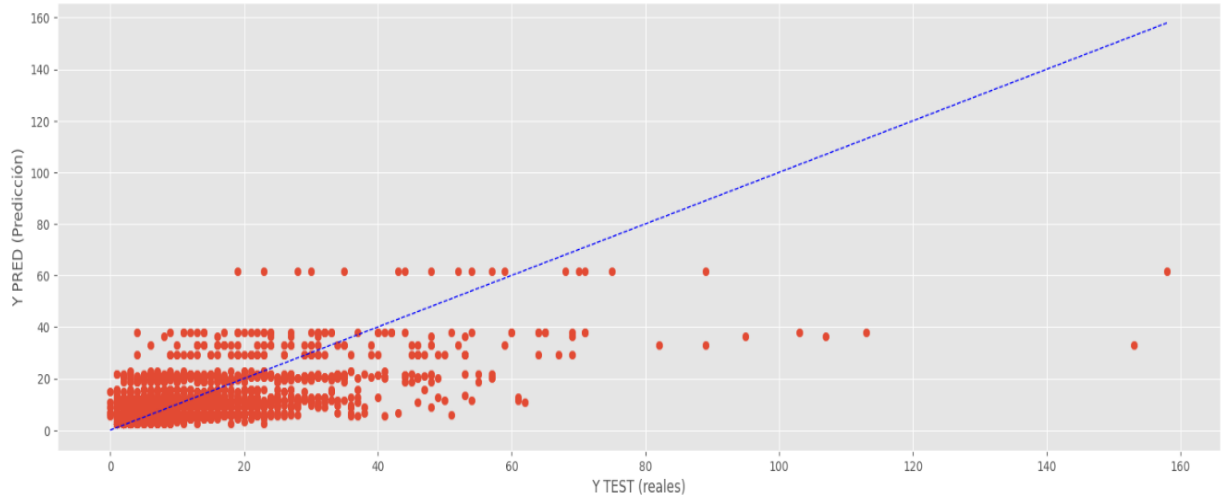


Figura 8. Grafica de ajuste de entrenamiento y prueba del modelo Regresión logística



El modelo de Árboles de Decisión, tras la validación de hiperparámetros, también presenta resultados más prometedores. El MSE y el R^2 indican que este modelo es capaz de realizar predicciones más precisas y explicar una parte significativa de la variabilidad en los datos. Este desempeño puede estar relacionado con la capacidad inherente de los Árboles de Decisión para capturar relaciones no lineales en los datos (Ver Figura 9).

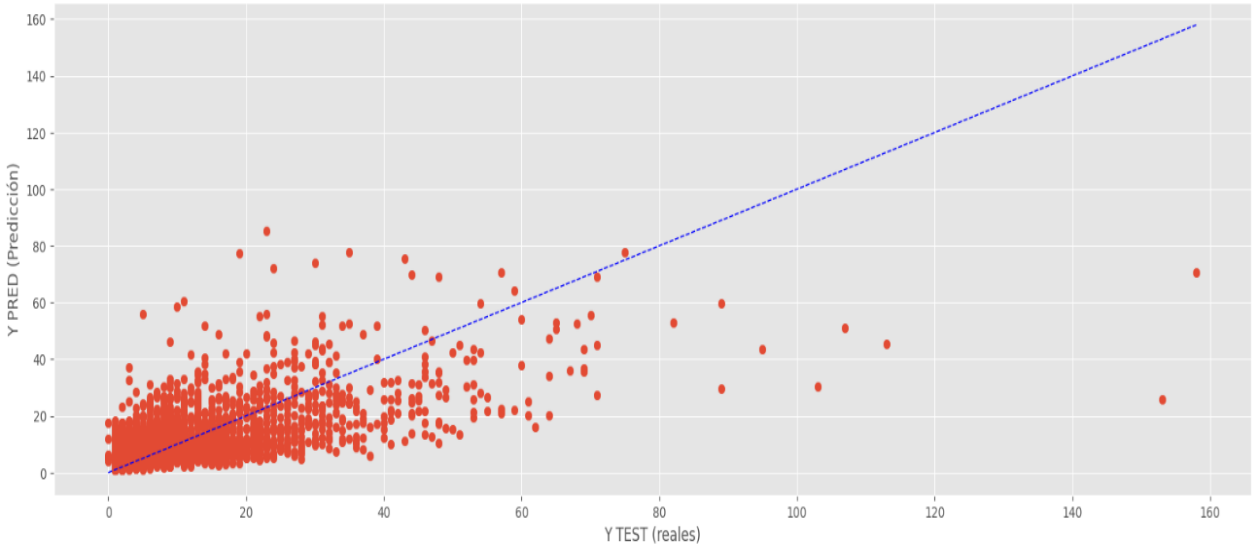
Figura 9. Grafica de ajuste de entrenamiento y prueba del modelo Arboles de decisión



El modelo de Random Forest después de la validación de hiperparámetros, continúa mostrando resultados prometedores. El MSE moderado y el R^2 positivo indican que este modelo es el más adecuado para la predicción de estancias hospitalarias, ya que es capaz de realizar predicciones precisas y explicar una parte significativa de la variabilidad en los datos (Ver Figura 10).

La validación de hiperparámetros ha permitido una evaluación más precisa de los modelos. Se concluye que el modelo de Bosques Aleatorios es el más apropiado para este conjunto de datos, ya que supera a los demás modelos en términos de precisión y capacidad para explicar la variabilidad en los datos. Tanto la Regresión Lineal como los modelos de Regresión Logística y Árboles de Decisión presentan desafíos significativos en este contexto y pueden no ser adecuados para la predicción de estancias hospitalarias.

Figura 10. Grafica de ajuste de entrenamiento y prueba del modelo Random Forest



7. Conclusiones

La población en estudio es en su mayoría de edad avanzada, con una cantidad significativa de comorbilidades, lo que sugiere un estado de salud generalmente complejo. Además, las hospitalizaciones varían en duración, lo que podría estar relacionado con la gravedad de las comorbilidades y otras condiciones clínicas. Este análisis proporciona información valiosa para futuras investigaciones en el campo de la salud, permitiendo una comprensión más profunda de las características demográficas y clínicas de la población estudiada.

Los resultados obtenidos pueden ser utilizados para mejorar la atención médica y la gestión hospitalaria, adaptándose a las necesidades específicas de los pacientes. Es importante destacar que estos hallazgos son el resultado de una investigación descriptiva inicial y podrían ser la base para estudios posteriores que exploren relaciones más complejas y profundas entre estas variables y otros factores de salud.

Referencias

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Vol. 4, No. 4, p. 738).* New York: Springer.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. DOI:10.1023/A:1010933404324.

Johnson, R. A. (2007). *Applied Multivariate Statistical Analysis (6th ed.).* New Jersey: Pearson Prentice Hall.

Kleinbaum, D. G. (2010). *Logistic Regression: A Self-Learning Text (3rd ed.).* New York: Springer.

Montgomery, D. C. (2012). *Introduction to Linear Regression Analysis (5th ed.).* New Jersey: Wiley.

Pedregosa, F. V. (2011). Scikit-learn: Machine Learning in Python (2nd ed.). *Journal of Machine Learning Research*, 2825-2830.

Tan, P. N. (2006). *Introduction to Data mining, Introduction to Data mining.* Pearson Education.