Modelo predictivo de días de estancia hospitalaria como herramienta para la optimización de recursos

Juan Pablo Bertel Morales

Gustavo Adolfo Jerez Tous

Gustavo Andrés Rubio Castillo

Afiliación:
Maestría en ciencia de datos y analítica, Facultad de Ingeniería , Universidad EAFIT

Fecha:
26 de noviembre de 2024

1. Introducción.

En el contexto actual de los sistemas de salud, la gestión eficiente de los recursos hospitalarios es un desafío crucial. El aumento constante de la demanda de atención médica, junto con los costos operativos crecientes, ha puesto presión sobre hospitales para optimizar su funcionamiento. Entre las métricas más críticas para lograr esta optimización se encuentra la duración de la estancia hospitalaria (estancia_total), un indicador que mide el tiempo total que un paciente permanece hospitalizado, expresado en días completos a partir de las primeras 24 horas de ingreso. Este enfoque asegura la consistencia en el cálculo y facilita la comparación entre diferentes casos clínicos. La estancia_total tiene implicaciones directas en la ocupación de camas, la asignación de personal médico y los costos operativos.

El objetivo principal de este proyecto es desarrollar un modelo predictivo que permita estimar con precisión la duración de la estancia hospitalaria, utilizando datos clínicos y administrativos de pacientes. Este modelo se construye aplicando técnicas de aprendizaje supervisado, con especial énfasis en la identificación de patrones a partir de variables como el diagnóstico principal, la clasificación de GRD y los costos asociados.

La relevancia de esta investigación radica en su potencial para mejorar la planificación operativa hospitalaria, reducir costos y, en última instancia, optimizar la atención al paciente, impactando positivamente tanto en la experiencia del paciente como en la sostenibilidad financiera del hospital.

Para abordar este problema, se utilizó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), estructurada en seis etapas principales: entendimiento del problema, análisis exploratorio de datos, preparación de datos, modelado, evaluación y despliegue.

El presente documento se organiza de la siguiente manera: En el Capítulo 2 se presenta el marco teórico, proporcionando un análisis detallado del problema y los conceptos clave. En el Capítulo 3 se describe la metodología, incluyendo la preparación de los datos y la selección de modelos. El Capítulo 4 aborda los resultados y la evaluación del modelo seleccionado, mientras que el Capítulo 5 concluye con recomendaciones para futuras implementaciones.

2. Marco teórico.

La duración de la estancia hospitalaria es un indicador fundamental en la gestión operativa de las instituciones de salud, ya que influye directamente en la planificación de recursos clínicos, financieros y logísticos. En un sector caracterizado por cambios normativos constantes, avances tecnológicos acelerados y una creciente demanda de servicios médicos, la sostenibilidad de los hospitales depende en gran medida de su capacidad para optimizar la gestión de estos recursos. En este contexto, entender y gestionar la duración de la estancia hospitalaria se convierte en una prioridad estratégica, ya que su adecuada estimación permite anticipar necesidades operativas, garantizar una atención de calidad y optimizar los costos asociados.

Por definición, la duración de la estancia hospitalaria comienza a contabilizarse tras las primeras 24 horas desde la admisión del paciente y concluye al momento del alta médica, ya sea por mejoría, traslado o fallecimiento. Este indicador está determinado por múltiples factores, incluyendo el diagnóstico principal, las comorbilidades, la edad y los procedimientos realizados durante la hospitalización. Además, herramientas como los Grupos Relacionados por el Diagnóstico (GRD) juegan un rol esencial en la clasificación de la complejidad de los casos, proporcionando información crítica sobre costos, consumo de recursos y patrones de atención

clínica. Estos sistemas permiten a las instituciones hospitalarias analizar las variaciones en la duración de la estancia y fundamentar sus decisiones en datos precisos y estructurados.

El desarrollo de modelos predictivos para la estimación de la estancia hospitalaria ha sido objeto de numerosos estudios que combinan metodologías estadísticas y técnicas de aprendizaje automático. Por ejemplo, investigaciones recientes han analizado la correlación entre la experiencia del cirujano y la duración de la estancia mediante regresión lineal, destacando cómo las habilidades del personal médico impactan directamente los resultados clínicos (Analyzing the correlation between surgeon experience and patient length of hospital stay, 2019). Este enfoque refuerza la importancia de seleccionar variables relevantes para construir modelos predictivos precisos y robustos.

Asimismo, la integración de técnicas tradicionales con métodos de aprendizaje automático ha mostrado resultados prometedores. En Machine learning and regression analysis to model the length of hospital stay in patients with femur fracture (2020), se combinaron características clínicas y demográficas de los pacientes para mejorar la precisión de las predicciones. Este estudio resalta el valor de adoptar enfoques híbridos que aprovechen la capacidad de aprendizaje de patrones complejos por parte de los algoritmos.

El uso de registros electrónicos de salud también ha revolucionado la forma de abordar este problema. En Analysis of length of hospital stay using electronic health records: A statistical and data mining approach (2021), se utilizaron regresión múltiple y técnicas de minería de datos para predecir la duración de la estancia, identificando relaciones significativas en grandes volúmenes de datos estructurados.

En contextos quirúrgicos, los modelos multivariados han demostrado ser herramientas efectivas para analizar los factores que influyen en la estancia hospitalaria. Por ejemplo, Factors affecting

length of stay after elective posterior lumbar spine surgery (2021) identificó variables como la edad, las comorbilidades y la complejidad de los procedimientos como determinantes clave. De manera similar, en Analysis of perioperative outcomes, length of hospital stay, and readmission rate after gastric bypass (2022), se resaltó la importancia de incluir variables perioperatorias para optimizar la asignación de recursos.

Durante la pandemia de COVID-19, los modelos predictivos adaptados a condiciones específicas fueron esenciales para enfrentar los desafíos de recursos hospitalarios limitados. En Length of stay analysis of COVID-19 hospitalizations using a count regression model and quantile regression: A study in Bologna, Italy (2022), se emplearon regresiones de conteo y cuantílicas para analizar la duración de las hospitalizaciones en casos complejos, considerando la variabilidad en la gravedad de los pacientes.

En unidades de cuidados intensivos, donde la planificación es crítica, se han aplicado técnicas como la regresión logística para predecir la estancia hospitalaria con gran efectividad (Prediction of patient length of stay on the intensive care unit following cardiac surgery, 2021). Estos modelos han permitido anticipar demandas en áreas de alta presión operativa.

A partir de investigaciones previas y del contexto del Hospital San Vicente, este proyecto tiene como objetivo desarrollar un modelo predictivo basado en regresión para estimar la duración de la estancia hospitalaria. Utilizando datos históricos y clasificaciones GRD, se busca construir una herramienta precisa que permita optimizar la asignación de recursos y mejorar la sostenibilidad operativa de la institución. Este enfoque no solo fortalecerá la capacidad del hospital para adaptarse a un entorno dinámico, sino que también contribuirá a garantizar una atención eficiente y de calidad para los pacientes.

3. Desarrollo Metodológico: CRISP-DM

El desarrollo del modelo predictivo se llevó a cabo siguiendo la metodología CRISP-DM, que proporcionó un marco estructurado para abordar cada etapa del proyecto. A continuación, se describen las fases clave del proceso, organizadas según las pautas establecidas en el entregable 6.

3.1. Entendimiento del Problema

El objetivo principal del proyecto es estimar la duración de la estancia hospitalaria de los pacientes en función de variables clínicas y operativas. La estancia hospitalaria es un indicador clave de la eficiencia hospitalaria y está asociada a costos, uso de recursos y calidad de atención. Este análisis tiene el potencial de optimizar la gestión hospitalaria y prever necesidades operativas.

3.1.1 Pregunta de Negocio

¿Cómo podemos predecir la duración total de la estancia hospitalaria de un paciente utilizando información clínica y operativa disponible durante las primeras 24 horas de la admisión del paciente?

3.1.2 Hipótesis

Las características clínicas (como diagnósticos, procedimientos principales y nivel de complejidad) y operativas (como costos operativos, peso estimado de los recursos e indicadores asociados a GRD) tienen un impacto significativo en la duración total de la estancia hospitalaria.

3.2. Análisis Exploratorio de Datos

3.2.1 Entendimiento de los Datos

El conjunto de datos original incluye información de pacientes hospitalizados con un total de 78,052 registros y 10 variables. Las variables incluyen información clínica, operativa y demográfica. A continuación, se describen las variables clave:

- **Variables predictoras:**
    - ***ir_cdm (Índice de Complejidad ajustado por Diagnóstico Médico):*** *Este indicador refleja la complejidad clínica de un paciente según su diagnóstico médico principal. Se calcula utilizando factores como la severidad de la enfermedad, la presencia de comorbilidades y las complicaciones asociadas. Es útil para categorizar a los pacientes según el nivel de recursos médicos necesarios para su atención.*
    - ***ir_grd_base (Índice basado en la clasificación GRD):*** *Corresponde a un índice derivado de los Grupos Relacionados por Diagnóstico (GRD), que agrupan a los pacientes en categorías homogéneas en términos de consumo de recursos. Este índice permite identificar casos clínicos similares en complejidad y costos, facilitando la evaluación operativa y financiera.*
    - ***nivel_de_complejidad (Complejidad clínica del caso):*** *Variable categórica que clasifica los casos en diferentes niveles de complejidad (por ejemplo: Baja, Media, Alta). Está determinada por la gravedad del diagnóstico, los procedimientos realizados y las comorbilidades. Este nivel impacta directamente en la planificación de recursos hospitalarios, como camas y personal médico.*

- ***procedimiento_principal (Procedimiento médico más relevante):*** *Indica el procedimiento médico principal realizado durante la hospitalización del paciente. Este procedimiento es seleccionado en base a su impacto clínico o en función de los recursos utilizados. Por ejemplo, puede incluir intervenciones quirúrgicas, tratamientos especializados o procedimientos diagnósticos complejos.*
- ***diagnostico_principal (Diagnóstico médico principal):*** *Representa la causa principal de hospitalización del paciente. Es una variable categórica que describe la condición médica que requirió atención hospitalaria y suele influir significativamente en la duración de la estancia y los costos asociados.*
- ***estancia_en_uci (Número de días en UCI):*** *Esta variable numérica indica los días que un paciente pasó en la Unidad de Cuidados Intensivos (UCI) durante su hospitalización. Es un indicador clave, ya que las estancias en UCI suelen ser costosas y están asociadas con casos clínicos de alta gravedad o procedimientos críticos.*
- ***edad (Edad del paciente):*** *Variable numérica que describe la edad del paciente en años al momento de la hospitalización. Es relevante porque la edad puede influir en la duración de la estancia, con patrones diferenciados entre pacientes jóvenes, adultos y adultos mayores.*
- ***costo_operativo_estimado (Estimación de los costos operativos asociados al paciente):*** *Representa la estimación de los costos operativos directos relacionados con la atención del paciente durante su hospitalización.*

Incluye costos de medicamentos, insumos médicos, procedimientos, diagnósticos y recursos humanos.

- ***peso_ir_estimado (Promedio del índice de recursos):*** *Es un indicador que mide el promedio del consumo de recursos hospitalarios asociado a la atención del paciente. Un valor alto puede indicar un caso clínico complejo o que requirió un uso intensivo de recursos.*

- **Variable objetivo (a predecir):**
  - ***estancia_total*****:** *Representa el número total de días que un paciente permanece hospitalizado desde su ingreso hasta el alta médica. Se cuenta en días completos, es decir, el período comienza a contabilizarse a partir de las primeras 24 horas desde el ingreso y concluye al momento del alta. Esta variable es fundamental para evaluar la eficiencia hospitalaria y planificar la gestión de recursos, ya que está influida por múltiples factores como el diagnóstico, las comorbilidades, la edad y los procedimientos realizados.*

3.2.2 Limpieza de Datos

Se realizaron los siguientes pasos para garantizar la calidad del dataset:

1. **Eliminación de columnas irrelevantes:**
   - Columnas como 'fecha', 'unidad_x_edad', 'año' y 'mes' fueron eliminadas debido a su irrelevancia para el modelo.
2. **Tratamiento de valores nulos y duplicados:**

- Se eliminaron registros incompletos y duplicados para asegurar la integridad del análisis, por lo tanto se redujo el tamañaño del data set en 76,258 registros.

3. **Transformación de variables categóricas:**

    - Las variables categóricas fueron limpiadas y normalizadas eliminando caracteres especiales, tildes y texto redundante.

3.2.3 Análisis Descriptivo:

El análisis descriptivo permitió comprender la naturaleza de las variables y su influencia en la variable objetivo (estancia_total). Este se estructuró en dos categorías principales: variables numéricas y categóricas, brindando una visión integral de los datos y destacando las características más relevantes para el modelado.

Variables Numéricas

Se analizaron las variables numéricas identificadas en el dataset: *costo_operativo_estimado*, *estancia_en_uci*, *edad* y *peso_ir_estimado*.

Resumen Estadístico

- **Costo Operativo Estimado:**

    - Rango: Desde valores mínimos cercanos a cientos hasta un máximo de más de 600 millones.

    - Media: 17,437,180 (aproximada), indicando una alta concentración de casos en valores relativamente bajos.

- Desviación estándar y Distribución: Desviación estándar de 2,576,580, reflejando una amplia dispersión entre los costos; Fuertemente sesgada hacia la izquierda, con valores extremos que representan casos clínicamente y operativamente complejos.

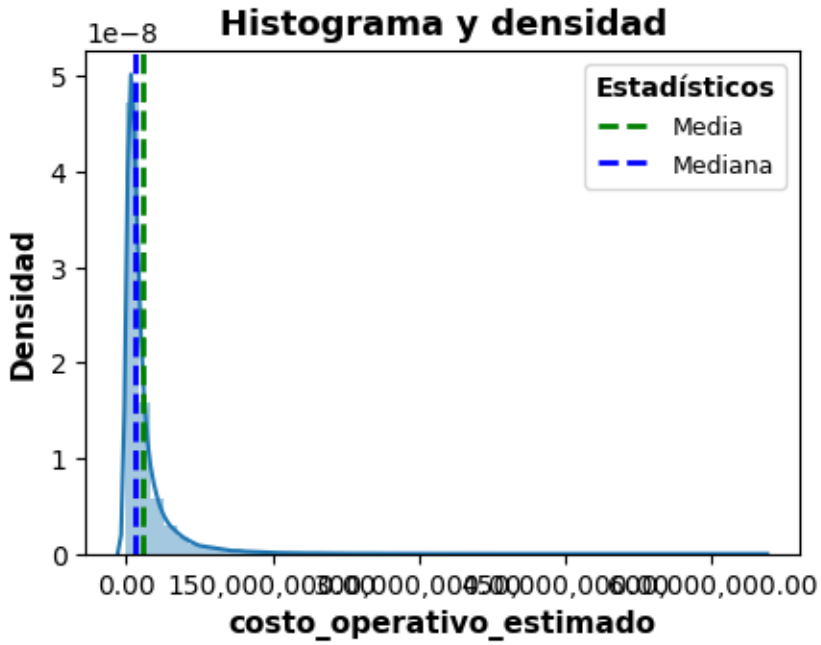

Figura 1Distribución del costo operativo estimado

- **Estancia en UCI:**
  - Media: 2 días, reflejando que la mayoría de los pacientes tienen estancias cortas en la UCI.
  - Percentil 90: Indica que menos del 10% de los pacientes permanecen más de 5 días en UCI.

- Distribución: Asimétrica, con un gran número de valores concentrados en 0 y 1 día. Valores superiores a 30 días son casos excepcionales.

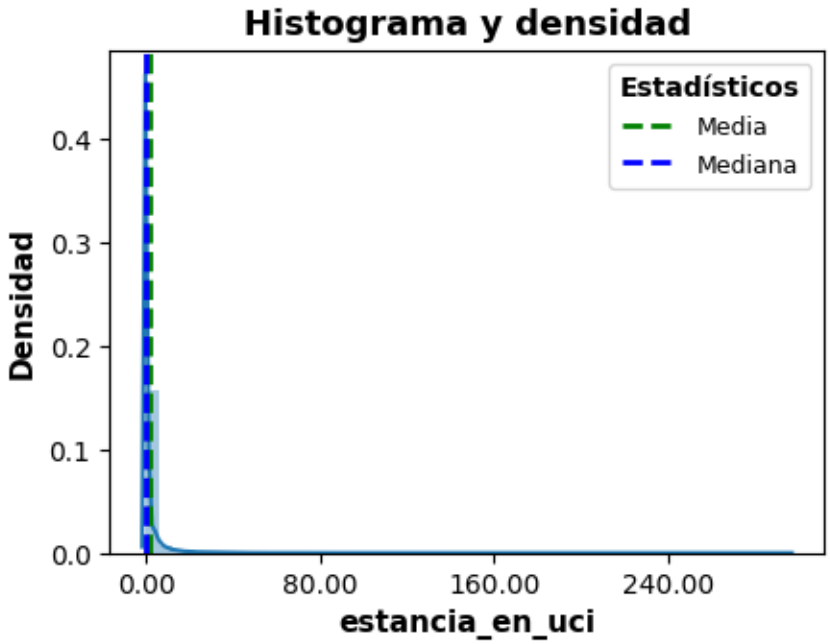

Figura 2 Distribución de la estancia en UCI

- **Edad:**
  - Media: 41 años, lo que sugiere que el dataset incluye tanto pacientes jóvenes como adultos mayores.
  - Rango: De 0 a 128 años.
  - Distribución: Bimodal, con picos en edades jóvenes (0-20 años) y adultos mayores (60-80 años).

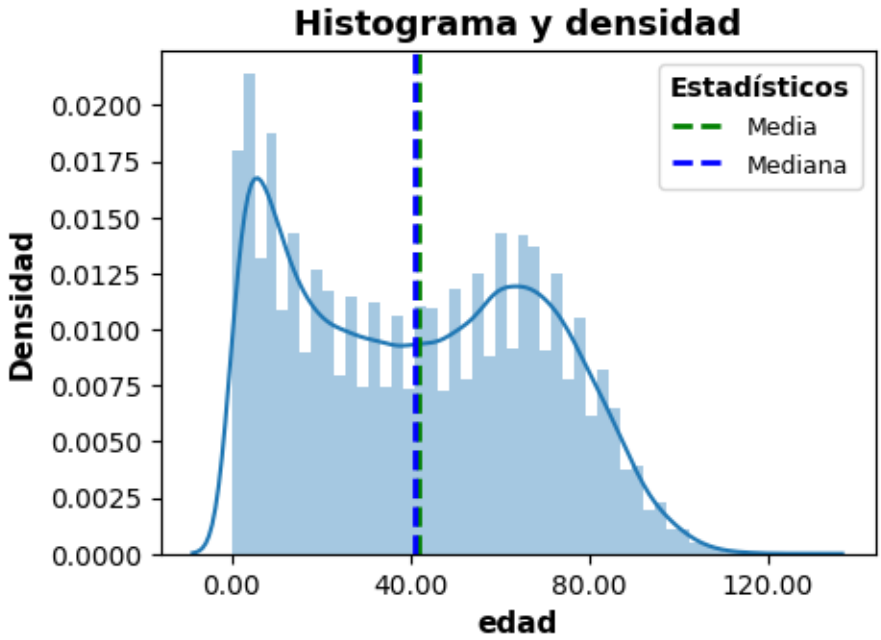

Figura 3 Distribución de la edad

- **Peso IR Estimado:**
    - Rango: La mayoría de los valores se concentran entre 0 y 5, con casos extremos llegando a 22.
    - Distribución: Sesgada hacia valores bajos, indicando que la mayoría de los pacientes tienen índices de recursos relativamente bajos.

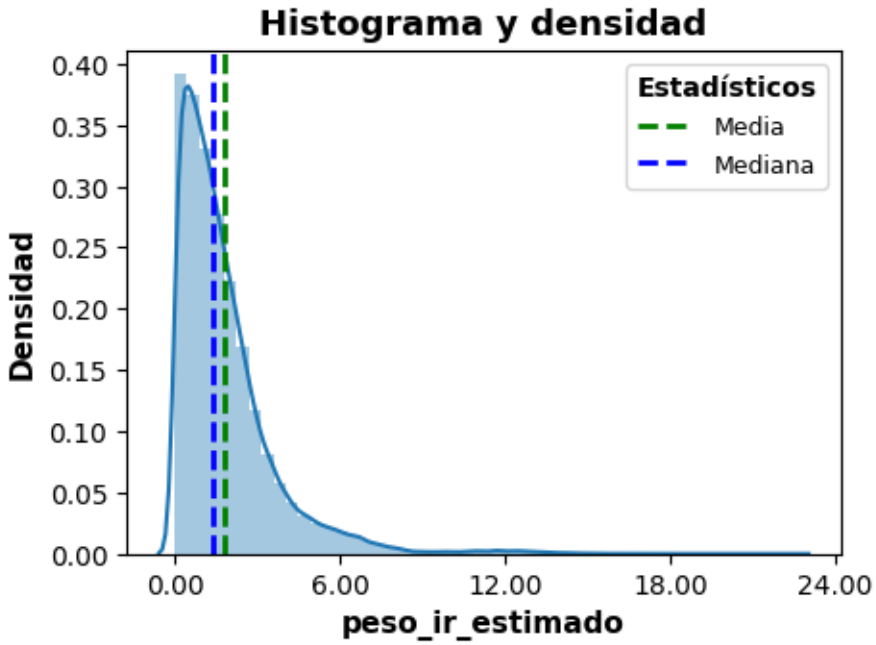

Figura 4 Distribución del peso IR

Relación con la Variable Objetivo

El análisis de la relación entre las variables predictoras y la variable objetivo se realizó utilizando el coeficiente de correlación de Pearson. Este coeficiente mide la fuerza y la dirección de la relación lineal entre dos variables, siendo particularmente útil para identificar patrones significativos en datos cuantitativos.

- **Costo Operativo Estimado vs. Estancia Total:**
    - Correlación positiva alta (r=0.82), lo que sugiere que un mayor costo operativo está asociado con una mayor estancia hospitalaria.
- **Estancia en UCI vs. Estancia Total:**

    - Correlación moderada (r=0.60), lo cual es intuitivo, ya que una mayor estancia en UCI contribuye directamente a la duración total de la estancia.

- **Edad vs. Estancia Total:**
    - Baja correlación (r=0.05), indicando que la edad no es un predictor fuerte por sí sola. Sin embargo, en combinación con otras variables, podría aportar valor analítico dentro del modelo.

- **Peso IR Estimado vs. Estancia Total:**
    - Relación débil (r=0.42) pero significativa. A medida que aumenta el peso IR estimado, la estancia tiende a incrementarse ligeramente-

Patrones de variabilidad:

Se generaron histogramas y diagramas de caja (boxplots) para explorar la distribución de las variables:

- **Costo Operativo Estimado:**
    - La mayoría de los casos se concentran en valores bajos, pero se observan outliers significativos.
    - El diagrama de caja muestra una gran dispersión con valores extremos que se deben analizar cuidadosamente.

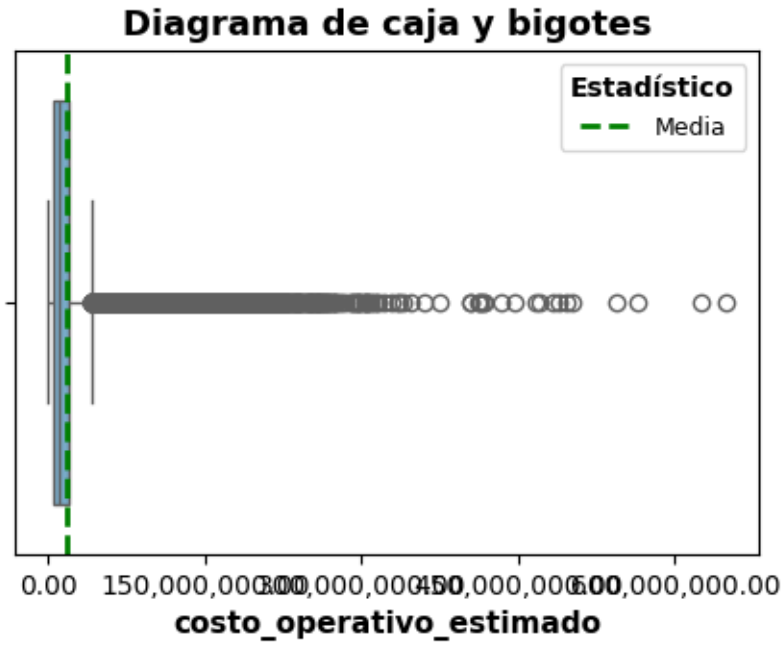

Figura 5 Diagrama de caja y bigote del costo operativo estimado

- **Estancia en UCI:**
    - El histograma muestra un gran número de valores en el rango de 0-5 días, con casos excepcionales por encima de los 100 días.

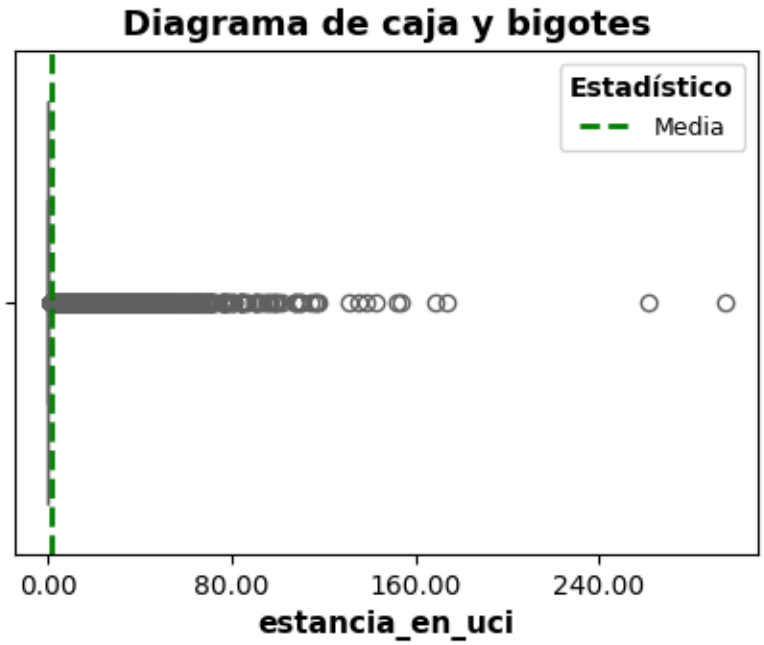

Figura 6 Diagrama de caja y bigote de la estancia en UCI

- **Edad:**
    - Distribución bimodal, con una amplia representación de pacientes jóvenes y mayores.

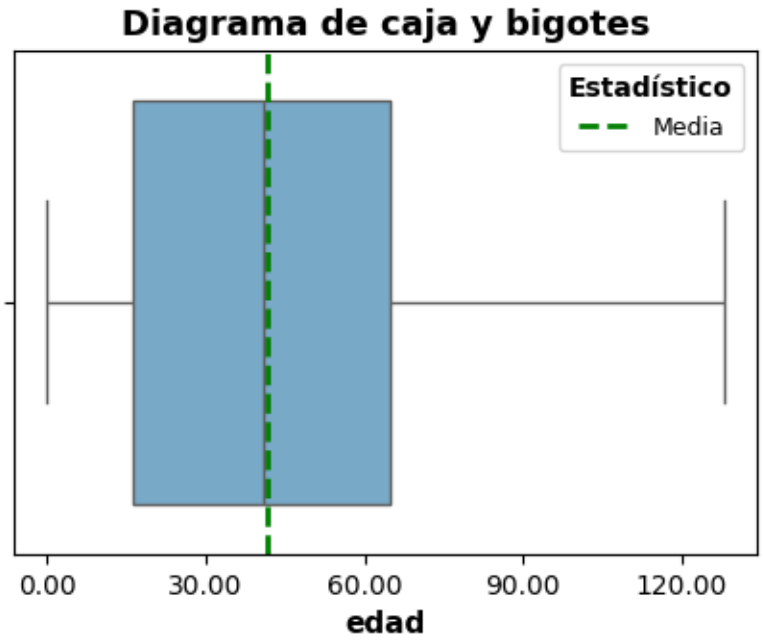

Figura 7 Diagrama de caja y bigote de la edad

- **Peso IR Estimado:**
    - Sesgo hacia valores bajos, aunque algunos casos extremos superan los 20.

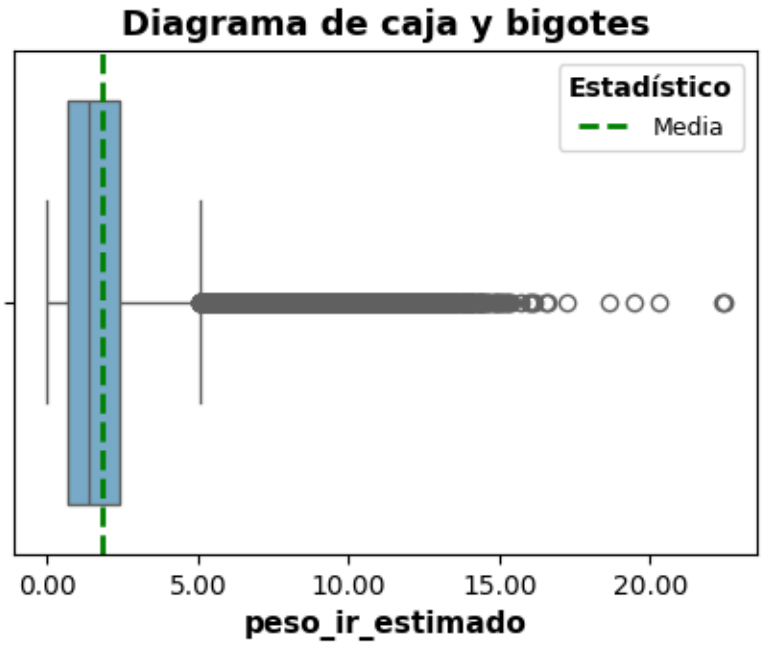

Figura 8 Diagrama de caja y bigote del peso IR estimado

Variables Categóricas

Las variables categóricas analizadas incluyen *ir_cdm*, *ir_grd_base*, *nivel_de_complejidad*, *procedimiento_principal* y *diagnostico_principal*. Estas variables representan aspectos clínicos y operativos que influyen directamente en la estancia hospitalaria.

Distribuciones

- **IR CDM:**
    - Gran cantidad de categorías poco frecuentes, con algunas categorías dominantes que representan casos clínicos recurrentes.
    - Gráficos de barras muestran que algunas categorías tienen una asociación notable con estancias prolongadas.
- **IR GRD Base:**

- Variación amplia en el promedio de estancia según la categoría. Esto refuerza su importancia como predictor.
- Las categorías con valores extremos representan casos de alta complejidad clínica.

- **Nivel de Complejidad:**
  - Distribución: La mayoría de los casos corresponden a "Baja Complejidad" (52.8%), seguidos de "Mediana Complejidad".
  - Impacto en la variable objetivo: Los casos de "Alta Complejidad" tienen un promedio significativamente mayor de estancia total, lo cual valida su relevancia para el modelo.
- **Procedimiento Principal:**
  - Se observan algunas categorías que destacan por su asociación con estancias prolongadas.
  - Los histogramas muestran que ciertos procedimientos tienen un impacto notable en la estancia hospitalaria.
- **Diagnóstico Principal:**
  - Distribución: Gran cantidad de categorías poco frecuentes. Las categorías más comunes están relacionadas con condiciones respiratorias y neurológicas.
  - Impacto en la variable objetivo: Los diagnósticos específicos están asociados con estancias prolongadas.

Relación con la Variable Objetivo

Se realizaron pruebas ANOVA para evaluar la significancia estadística de las diferencias en *estancia_total* entre las categorías de cada variable:

- **IR CDM:** Categorías con diferencias significativas en el promedio de estancia.

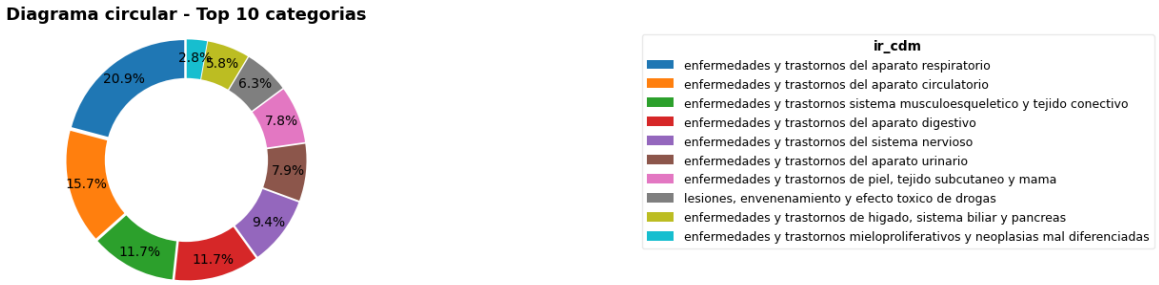

Figura 9 Grafico de anillo del IR CDM

- **IR GRD Base:** Amplias variaciones entre categorías, reflejando su relevancia clínica.

Figura 10 Grafico de anillo del IR GRD BASE

- **Nivel de Complejidad:** Las estancias aumentan significativamente con la complejidad.

Figura 11 Grafico de anillo del nivel de complejidad

- **Procedimiento Principal y Diagnóstico Principal:** Ambos factores muestran un impacto significativo en la estancia total, validando su inclusión en el modelo.

Figura 12 Grafico de anillo de procedimiento principal

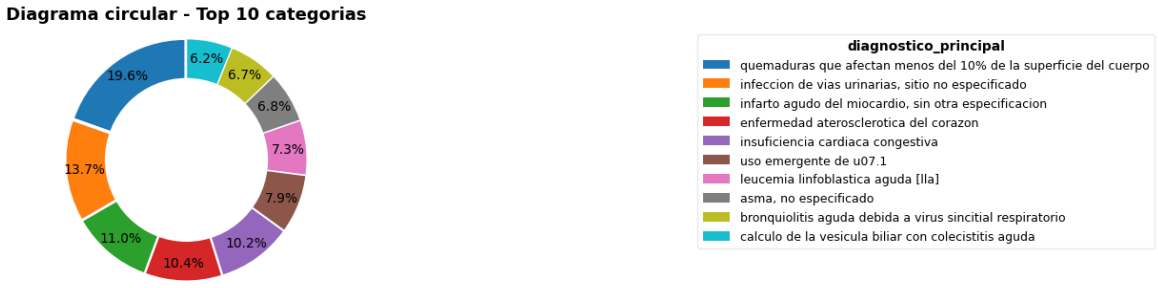

Figura 13 diagnostico principal

3.2.4 Insights Clave

1. **Variables Numéricas:**
    - *Costo operativo estimado* y *estancia en UCI* son los principales predictores de la duración total de la estancia hospitalaria.
    - La relación entre algunas variables, como *peso IR estimado*, es no lineal, lo que sugiere la necesidad de modelos capaces de capturar estas dinámicas.
2. **Variables Categóricas:**
    - Categorías de *nivel de complejidad* y *procedimiento principal* tienen un impacto significativo en la estancia total, siendo factores clave para el modelo predictivo.
    - La alta variabilidad en categorías como *IR GRD Base* y *diagnóstico principal* refuerza su relevancia para explicar las diferencias en estancias.
    - *diagnóstico principal y procedimiento_principal presentan una alta cardinalidad por lo que complejiza los análisis como reducción de dimensionalidad por temas de complejidad y costo computacional, por tanto técnicas como One Hot Encoding no son tan eficientes.*

4. Preparación de los Datos

4.1 Análisis de Outliers

Para garantizar la calidad del dataset y mitigar el impacto de valores atípicos en los modelos predictivos, se emplearon los siguientes métodos de detección de outliers:

- **Método de Mahalanobis:**

  Este método identificó un total de 4,450 registros como outliers en variables numéricas.

Sin embargo, debido a la suposición de linealidad en las relaciones entre variables, los resultados de este método no fueron utilizados para los modelos posteriores.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**
  Detectó 406 registros atípicos en las variables numéricas utilizando un enfoque de clustering basado en densidad. Este método fue seleccionado por su capacidad para identificar patrones no lineales en los datos.

- **Análisis de Correspondencia Múltiple (MCA):**
  Antes de aplicar este análisis, se realizó una agrupación de categorías con baja frecuencia en variables categóricas como diagnostico_principal, procedimiento_principal e ir_grd_base. Las categorías con menos de tres registros fueron reemplazadas por la categoría con mayor frecuencia (moda). Posteriormente, MCA detectó:
  - 2,092 outliers utilizando codificación One-Hot Encoding.
  - 2,561 outliers utilizando codificación por factorización**.**

Resultados finales:

Tras evaluar los métodos, se seleccionaron DBSCAN y MCA debido a su capacidad para manejar relaciones no lineales y datos categóricos con mayor flexibilidad. Como resultado, se eliminaron 5,030 registros (6.6% del dataset original), optimizando la calidad del conjunto de datos para los modelos predictivos.

4.2 Ajuste de la Variable Objetivo

Para minimizar el impacto de valores extremos en la variable objetivo (estancia_total), se aplicó

la técnica de transformación Winsorizing. Esta metodología reemplazó el 1% de los valores más altos con el percentil 99, estableciendo un máximo permitido de 54 días de estancia hospitalaria.

4.3 Particionamiento del Dataset

El dataset fue dividido en dos subconjuntos para garantizar un proceso de modelado y evaluación riguroso:

- Conjunto de entrenamiento (90%): Incluyó 64,105 registros, utilizados para entrenar los modelos.
- Conjunto de prueba (10%): Contenía 7,123 registros, destinados a la evaluación del desempeño predictivo.

4.4 Transformación de Variables

Se llevaron a cabo las siguientes transformaciones para preparar las variables del dataset:

- Codificación de variables categóricas:

  Se utilizó Target Encoding suavizado, el cual combina la media de cada categoría con la media general de la variable objetivo (estancia_total). Este enfoque permite capturar la relación entre las categorías y la variable objetivo, mientras se reduce el riesgo de sobreajuste.

- Normalización de variables numéricas:

  Se aplicó StandardScaler para estandarizar las variables numéricas. Este método ajusta

los valores para que tengan una media de 0 y una desviación estándar de 1, mejorando la estabilidad y el desempeño de los modelos.

Estas etapas garantizaron un dataset limpio, balanceado y listo para su uso en el modelado predictivo, asegurando que los resultados fueran representativos y generalizables.

5. Selección de Modelos y Entrenamiento

Con el conjunto de datos preparado y libre de outliers, el siguiente paso consistió en seleccionar y evaluar los modelos más adecuados para la predicción de la duración de la estancia hospitalaria. Este proceso incluyó la comparación de varios algoritmos de regresión, tanto lineales como no lineales, para identificar aquel que ofreciera el mejor balance entre precisión y capacidad de generalización.

La selección de modelos consideró no solo el desempeño en términos de métricas de error, sino también la interpretación de los resultados y su aplicabilidad en un entorno hospitalario. A continuación, se presentan los modelos evaluados junto con sus características clave.

5.1 Modelos Evaluados

Se evaluaron los siguientes algoritmos de regresión:

1. Random Forest Regressor.
2. K-Nearest Neighbors (KNN).
3. Linear Regression.

4. Ridge Regression.
5. Lasso Regression.
6. Decision Tree Regressor.

5.2 Resultados del Torneo de Modelos Inicial (Sin optimización de hiperparametros):

Se evaluaron diversos modelos de regresión con el conjunto de datos preparado, priorizando métricas de desempeño como el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE). A continuación, se detallan los resultados obtenidos:

- **Random Forest Regressor:**

  Este modelo alcanzó el mejor desempeño general, con un R^2 de 0.7489, un RMSE de 5.008 y un MAE de 3.2918. Su capacidad para manejar relaciones no lineales y captar interacciones complejas entre variables lo convierte en una opción poderosa, aunque su complejidad puede dificultar la interpretabilidad.
- **K-Nearest Neighbors (KNN):**

  KNN demostró un desempeño competitivo con un R^2 de 0.6927, un RMSE de 5.4071 y un MAE de 3.5773. Su simplicidad y la facilidad para entender su funcionamiento lo posicionan como una alternativa práctica, especialmente en entornos donde la interpretabilidad es prioritaria.
- **Linear Regression:**

  El modelo de regresión lineal presentó un R^2 de 0.6838, un RMSE de 5.5400 y un MAE de 3.6818. Aunque es un modelo sencillo y rápido, no logra capturar la complejidad de los datos, lo que limita su capacidad predictiva.

- **Ridge Regression:**

  Este modelo, que incluye regularización para mitigar problemas de sobreajuste, obtuvo resultados idénticos a los de la regresión lineal con un R^2 de 0.6838, un RMSE de 5.5400 y un MAE de 3.6818. Esto sugiere que la regularización no tuvo un impacto significativo en este caso.

- **Lasso Regression:**

  Al igual que Ridge, Lasso combina regresión lineal con regularización, pero además realiza selección de características. Aun así, obtuvo un desempeño menor con un R^2 de 0.6648, un RMSE de 5.7040 y un MAE de 3.8024, indicando una menor capacidad para ajustar los datos.

- **Decision Tree Regressor:**

  Este modelo presentó el peor desempeño, con un R^2 de 0.5006, un RMSE de 6.9629 y un MAE de 4.5544. Esto evidencia su tendencia a sobreajustar los datos de entrenamiento y a ser menos efectivo sin optimización de hiperparámetros.

5.3 Modelo seleccionado.

Aunque el Random Forest Regressor presentó el mejor desempeño, se optó por el modelo K-Nearest Neighbors (KNN) debido a su simplicidad y facilidad de interpretación. KNN permite una implementación más directa y comprensible para usuarios no técnicos, lo que lo hace más práctico en un entorno hospitalario donde la claridad y la adaptabilidad son esenciales. Además, su desempeño es competitivo, logrando un equilibrio adecuado entre precisión y facilidad de uso.

5.4 Ajuste de hiperparametros para la optimización del modelo KNN.

Una vez seleccionado el modelo **K-Nearest Neighbors (KNN)** como el candidato más adecuado debido a su equilibrio entre simplicidad y desempeño, se procedió a la optimización de sus hiperparámetros con el fin de mejorar su precisión y ajuste al conjunto de datos.

Se utilizó un enfoque de validación cruzada con cinco pliegues para evaluar diferentes combinaciones de hiperparámetros clave. Estos incluyeron:

- **weights**: Peso asignado a los vecinos (uniforme o en función de la distancia).
- **n_neighbors**: Cantidad de vecinos utilizados para realizar las predicciones.
- **metric**: Métrica utilizada para calcular la distancia (e.g., euclidean, manhattan).
- **algorithm**: Método utilizado para realizar las búsquedas de vecinos más cercanos (e.g., auto, ball_tree).

Tras evaluar 100 combinaciones de parámetros, el modelo óptimo presentó los siguientes valores:

- **weights**: 'distance' (ponderación basada en la distancia de los vecinos).
- **n_neighbors**: 18 (cantidad óptima de vecinos).
- **metric**: 'euclidean' (distancia euclidiana como métrica principal).
- **algorithm**: 'auto' (selección automática del algoritmo más eficiente según el tamaño y estructura de los datos).

El mejor desempeño del modelo tras la optimización se reflejó en los siguientes resultados:

- **RMSE (Error cuadrático medio en el conjunto de prueba)**: 5.2528
- **R-squared (Coeficiente de determinación en el conjunto de prueba)**: 0.7139

- **MAE (Error absoluto medio) 3**.4886

Estos ajustes permitieron refinar el modelo, mejorando su capacidad de generalización y asegurando un desempeño más robusto en escenarios reales.

5.5 Conclusiones.

El modelo K-Nearest Neighbors (KNN) se posicionó como una solución adecuada para la predicción de la duración de la estancia hospitalaria, destacándose por su simplicidad y facilidad de interpretación. Aunque el modelo Random Forest Regressor mostró métricas superiores como un mayor R^2 y menor RMSE, la claridad y la facilidad de uso de KNN lo convierten en una opción más práctica para entornos hospitalarios. En estos contextos, donde las decisiones deben ser rápidas y comprensibles para los usuarios no técnicos, la transparencia del modelo resulta clave para respaldar la toma de decisiones.

Además, la capacidad de KNN para capturar relaciones significativas entre variables clínicas y operativas permite una predicción precisa y confiable de la duración de la estancia. Esto es fundamental para optimizar la asignación de recursos hospitalarios, como camas, personal médico y costos operativos, contribuyendo así a la sostenibilidad y eficiencia del sistema de salud. Su equilibrio entre desempeño y simplicidad lo hace particularmente valioso en escenarios donde la interpretabilidad es tan importante como la precisión.

De cara al futuro, se recomienda explorar enfoques híbridos que combinen la simplicidad de KNN con la robustez de modelos más complejos, como Random Forest. También sería beneficioso implementar pruebas en tiempo real para evaluar su desempeño en condiciones

dinámicas, y ampliar las características del modelo con datos adicionales que podrían mejorar aún más su capacidad predictiva y su aplicabilidad en entornos reales.

6. Tecnología.

La selección y diseño de la tecnología empleada en este proyecto se realizaron con un enfoque práctico, orientado a su eventual implementación en el entorno hospitalario. Este enfoque asegura que las herramientas y estrategias adoptadas no solo respondan a los requerimientos técnicos del análisis, sino que también se integren eficientemente en los flujos operativos y de gestión del hospital. A partir de datos hospitalarios extraídos periódicamente de sistemas como SAP y AGFA, se trabajó principalmente con datos estructurados provenientes de tablas relacionales que detallan información clave como registros de pacientes, procedimientos, costos y diagnósticos, garantizando una base sólida para el desarrollo y la operación del modelo predictivo.

6.1. Fuentes de Datos y Naturaleza.

El proyecto se basa en datos hospitalarios provenientes de diversas fuentes que abarcan diferentes naturalezas y formatos:

- **Batch:** Datos históricos extraídos periódicamente de sistemas hospitalarios (e.g., ERP hospitalarios, sistemas de información clínica y SAP).
- **Estructurada:** Tablas de datos relacionales que contienen registros de pacientes, procedimientos médicos, costos y diagnósticos definidos. (e.g. AGFA)

6.2. Ingesta de Datos

La arquitectura propuesta para la implementación del proyecto se basa principalmente en mecanismos de **batch**:

- **Batch:**
    - Extracción diaria de datos históricos de estancia hospitalaria desde sistemas hospitalarios (SAP y AGFA) en formato CSV.

6.3. Almacenamiento

El almacenamiento se organiza en un sistema híbrido que combina un **Data Lake** y bases de datos relacionales:

- **Lago de Datos (Data Lake):**
    - **Tecnología:** Se uso de **Amazon S3** para almacenar datos no estructurados y semiestructurados en formatos escalables, sin embargo, en la implementación se usará la infraestructura interna de almacenamiento propio del hospital San Vicente.
    - **Organización:** Estructura jerárquica con carpetas para datos en bruto (raw), y transformados (trusted).
    - **Propósito:** Archivar información histórica completa para análisis avanzados o auditorías.
- **Bases de Datos Relacionales:**
    - **Tecnología:** Uso de **PostgreSQL** o **Amazon RDS** para almacenar datos estructurados utilizados en modelado predictivo.

- **Propósito:** Facilitar la consulta eficiente de registros categóricos y numéricos utilizados en modelos de predicción de estancia.

6.4. Ambiente de Procesamiento

El ambiente de procesamiento está diseñado para manejar grandes volúmenes de datos, integrar diferentes formatos y realizar análisis complejos:

- **Procesamiento Batch:**
  - **AWS Glue:** Para la limpieza, transformación y análisis de grandes conjuntos de datos hospitalarios. Este servicio completamente gestionado de AWS permite procesar datos de manera eficiente y escalable, aprovechando un motor subyacente basado en Apache Spark. Además, facilita la integración con otras herramientas del ecosistema de AWS, como S3, Redshift, y Athena, eliminando la necesidad de gestionar infraestructuras complejas.
  - **Pandas y PySpark:** Procesamiento de datos en notebooks locales o en entornos como **Google Colab**.

6.5. Aplicaciones y Visualización

El proyecto contempla la entrega de resultados y análisis a través de aplicaciones intuitivas y accesibles:

- **Dashboards:**
  - Uso de herramientas como **Power BI** para visualización interactiva de métricas clave como promedios de estancia, costos asociados y ocupación hospitalaria.
  - Presentación de datos y reportes para directivos hospitalarios y médicos.

- **Archivos:**
    - Exportación periódica de predicciones y métricas en formato Excel o CSV para distribución a equipos operativos.
- **Notificaciones Automatizadas:**
    - Integración con plataformas de mensajería hospitalaria (e.g.,Mensajero digital) para alertas de estancias críticas o predicciones fuera de rango esperado.

6.6. Despliegue del Proyecto en un Escenario Hipotético

En un escenario hipotético de implementación real en el hospital San Vicente fundación, el despliegue del proyecto se realizaría de la siguiente manera:

1. **Infraestructura en la Nube:**
    - Uso de **AWS** o **Google Cloud** para hosting del lago de datos, bases de datos, y servicios de procesamiento.
    - Escalabilidad automática para manejar aumentos en el flujo de datos o carga computacional.
2. **Pipeline Automatizado:**
    - Construcción de un pipeline de datos end-to-end utilizando herramientas como **Apache Airflow** para orquestación de tareas y monitorización.
3. **Integración con Sistemas Existentes:**
    - Conexión con sistemas hospitalarios locales (e.g. SAP) para la ingesta continua de datos estructurados y no estructurados.

4. **Capacitación y Validación:**
    - Capacitación a personal médico y administrativo en el uso de dashboards y predicciones.
    - Validación continua del modelo con datos y métricas, para garantizar su precisión.

7. Conclusiones generales del proyecto.

El desarrollo e implementación de este modelo predictivo para estimar la duración de la estancia hospitalaria representa un avance significativo hacia la optimización de los recursos hospitalarios y la mejora de la gestión operativa en instituciones de salud como el Hospital San Vicente Fundación. A través de un enfoque sistemático, estructurado bajo la metodología CRISP-DM, y con el uso de herramientas tecnológicas avanzadas, se logró construir una solución robusta y adaptada a las necesidades específicas del entorno hospitalario.

1. **Impacto Operativo y Clínico**

   El modelo desarrollado proporciona predicciones confiables sobre la estancia hospitalaria, permitiendo una planificación más efectiva de los recursos críticos, como camas, personal médico y equipos operativos. Al utilizar variables relevantes como diagnósticos principales, procedimientos y costos operativos, se maximizó la capacidad del modelo para reflejar dinámicas reales del entorno clínico. Esto contribuye a mejorar la eficiencia hospitalaria y a garantizar una atención de calidad para los pacientes.

2. **Simplicidad y Aplicabilidad**

   La selección del modelo K-Nearest Neighbors (KNN) destacó por su simplicidad, interpretabilidad y capacidad para capturar relaciones significativas entre variables. Aunque no fue el modelo con el mejor desempeño absoluto, su facilidad de

implementación y uso en contextos no técnicos lo posicionó como una solución práctica para entornos hospitalarios. Esto resalta la importancia de balancear el rendimiento técnico con la usabilidad en escenarios del mundo real.

3. **Preparación de Datos y Metodología Rigurosa**

   La preparación meticulosa del dataset, que incluyó la detección y eliminación de valores atípicos, la transformación de variables categóricas y la normalización de datos numéricos, garantizó la calidad del conjunto de datos para el modelado predictivo. Estas etapas críticas aseguraron que el modelo fuera representativo y generalizable, destacando la importancia de una ingeniería de datos sólida para el éxito de proyectos de esta naturaleza.

4. **Capacidades Tecnológicas y Despliegue**

   El diseño tecnológico adoptado para un escenario hipotetico para este proyecto integró tecnologías avanzadas como AWS Glue y bases de datos relacionales como PostgreSQL, lo que permite escalabilidad y procesamiento eficiente de grandes volúmenes de datos. Además, el uso de herramientas de visualización y notificaciones automatizadas facilita la interpretación de los resultados y su integración en los flujos operativos del hospital, reforzando su aplicabilidad en entornos reales.

5. **Sostenibilidad y Proyección Futura**

   Este modelo tiene el potencial de contribuir no solo a la sostenibilidad financiera del hospital, sino también a la mejora de la experiencia del paciente. A futuro, se recomienda explorar enfoques híbridos que combinen la simplicidad de KNN con la robustez de modelos más avanzados como Random Forest, así como implementar pruebas en tiempo

real para validar su desempeño dinámico. Asimismo, la integración de nuevas fuentes de datos y variables podría ampliar aún más la precisión y alcance del modelo.

En resumen, este proyecto no solo responde a una necesidad actual en la gestión hospitalaria, sino que establece un marco de referencia para futuras iniciativas de análisis predictivo en el sector salud, posicionando al Hospital San Vicente como un referente en la innovación tecnológica aplicada a la atención médica.

8. Referencia bibliográficas.

Rajpal, S., Shah, M., Vivek, N., & Burneikiene, S. (2020). Analyzing the correlation between surgeon experience and patient length of hospital stay. Cureus.

Ricciardi, C., Ponsiglione, A. M., Scala, A., & Borrelli, A. (2022). Machine learning and regression analysis to model the length of hospital stay in patients with femur fracture. Bioengineering.

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLOS ONE.

Gruskay, J. A., Fu, M., Webb, M. L., & Grauer, J. N. (2015). Factors affecting length of stay after elective posterior lumbar spine surgery: A multivariate analysis. The Spine Journal.

Dallal, R. M., & Trang, A. (2012). Analysis of perioperative outcomes, length of hospital stay, and readmission rate after gastric bypass. Surgical Endoscopy.

Zeleke, A. J., Moscato, S., Miglio, R., & Chiari, L. (2022). Length of stay analysis of COVID-19 hospitalizations using a count regression model and quantile regression: A study in Bologna, Italy. International Journal of Environmental Research and Public Health.

Meadows, K., Gibbens, R., Gerrard, C., & Collaborators. (2018). Prediction of patient length of stay on the intensive care unit following cardiac surgery: A logistic regression analysis based on the cardiac operative mortality risk. Journal of Cardiothoracic and Vascular Anesthesia.