Predicción de duración de internación para gestión de camas en hospital

Diplomatura en Ciencia de Datos y Análisis Avanzado - UTN.BA

Grupo N

Integrantes:

- Ignacio Gamez
- Lara Bordenave
- José Pollola Barón
- Diego Hernández

Fecha 4-10-2025

Definición del Problema y Relevancia

Contexto y Motivación

La disponibilidad de camas define la capacidad del hospital y hoy se gestiona de forma reactiva: se abren camas, se posponen cirugías o se derivan pacientes cuando la ocupación ya es crítica.

En Cardiología, la duración de internación es muy variable y, combinada con cirugías programadas, genera picos que saturan rápido. Proponemos pasar a un enfoque predictivo: usar datos clínicos y de laboratorio al ingreso para estimar días de estancia y riesgo de larga estadía. Con esa predicción y el calendario de cirugías, se proyecta la demanda de camas con anticipación y se disparan alertas tempranas para habilitar acciones como: altas seguras anticipadas, reasignación de camas entre servicios, reprogramación selectiva de cirugías, derivaciones planificadas.

Resultado esperado: uso más eficiente de recursos, menor presión sobre equipos y mejor experiencia del paciente.

Objetivo Concreto y Medible

Entrenar un modelo que, al ingreso, prediga duración de estadía y riesgo de prolongación para activar alertas que guíen reasignación de camas, altas y programación de intervenciones en Cardiología.

CRISP-DM en este proyecto

Business Understanding

Optimizar gestión de camas anticipando long-stay y LOS para reducir cuellos de botella, planificar dotaciones y coordinar altas/interconsultas. Métricas operativas clave: **recall@top-N** para alertas tempranas y **MAE de LOS** para planificación.

Data Understanding

Exploración de distribuciones (edad, diagnósticos, servicios), estacionalidad y outliers; correlaciones y tasas de long-stay por grupos etarios (p.ej., 50–79) y ventanas de LOS (5–14 días). Auditoría de calidad (faltantes, duplicados, rangos imposibles) y sesgos por servicio/estación.

Data Preparation

Imputación numérica (mediana/KNN según variable), codificación categórica (One-Hot/Ordinal), escalado cuando aplica e ingeniería mínima para evitar leakage. Exclusión de **LOS > 15 días** para estabilizar la distribución y enfocarse en procesos planificables. Partición con **CV estratificado (clasificación)** preservando proporciones.

Modeling

Modelos: **LGBM** y **XGBoost** para regresión (LOS) y clasificación (riesgo long-stay) + **DNN simple** como baseline. Pipelines con **5-fold CV**, búsqueda ligera de hiperparámetros y semilla fija. Selección por **MAE** (regresión) y **PR-AUC / recall@top-N** (clasificación).

Evaluation

Regresión (LOS): objetivo **MAE ≈ 1.5–2.0 días**; reportar MAE/RMSE y error por decil.

Clasificación (long-stay): PR-AUC, **ROC-AUC**, **Recall@Top-20%** y **Precision@Top-20%**. Incluye trade-offs (curvas P-R), interpretabilidad (importancias/SHAP) y pruebas de robustez.

Deployment & Operación

Se plantea desarrollar tablero operativo con:

- **Ranking diario de riesgo** (listas accionables).
- **Estimación de LOS por paciente** para altas y coordinación con quirófano.
- **Monitoreo de drift** y **recalibración mensual** (o bajo umbral de rendimiento), con logging de métricas y retraining controlado.

Datos y metodología

Análisis EDA:

El análisis exploratorio se llevó a cabo con el objetivo de comprender el dataset disponible, examinando las variables presentes, sus dimensiones y características. Se revisaron los datos para identificar posibles ajustes necesarios y determinar cuáles variables son relevantes para el estudio y cuáles pueden ser descartadas.

En la fase final del análisis, el enfoque se centra en extraer conclusiones que puedan aplicarse de manera práctica, con el objetivo principal de reducir la ocupación de camas dentro del hospital.

1. Carga e inspección inicial de los datos

- **15.757 admisiones** (2 años), **56 variables**: demografía, tipo de admisión, labs (Hb, creatinina, glucosa…), comorbilidades/diagnósticos y resultado (alta/fallecimiento).

Conclusión :volumen suficiente; heterogeneidad de tipos (numéricas/categóricas/binarias).

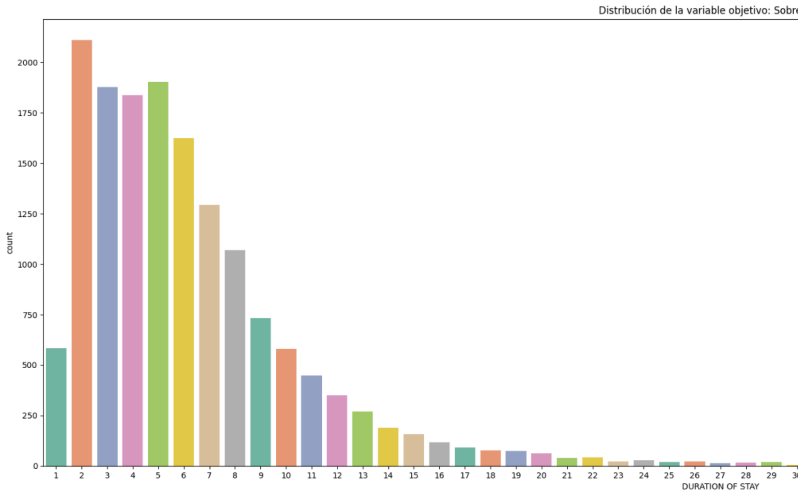
2. Valores faltantes

- BNP >50% faltantes → eliminada.
- Imputación numérica con mediana para EDA (<50% faltantes); para simulación se aplicará KNN.

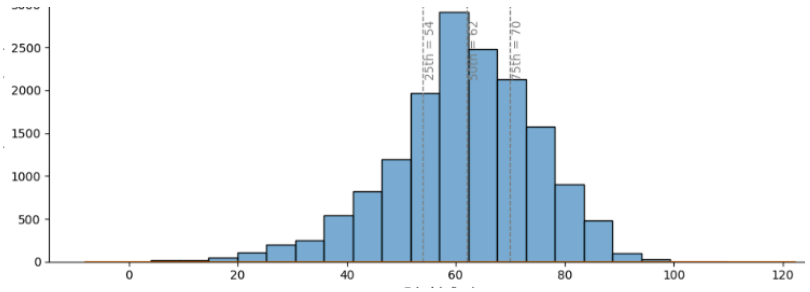
Conclusión: variables clínicas completas post-imputación, robustas ante outliers.

3. Análisis univariado

- LOS (días): pico en 2 días; cola relevante hasta ~12+; >7 días son numerosos → ocupación sostenida y riesgo de saturación pese a alta rotación de estancias cortas.
- Edad: mediana ~60; mayor densidad 45–75; pocos extremos. Relevante para el modelo, probablemente con efecto no lineal.



cantidad de pacientes vs. días hospitalizados



edad vs. días hospitalizados

El análisis de la distribución etaria muestra una población centrada en la mediana de **60 años.** La distribución es aproximadamente normal, con la mayor concentración entre 45 y 75 años y pocos casos extremos. Esto sugiere que la edad es una variable relevante para modelado, aunque su efecto sobre LOS puede ser no lineal; por tanto proponemos seguir evaluándose.

CONCLUSIÓN DEL ANÁLISIS UNIVARIADO:

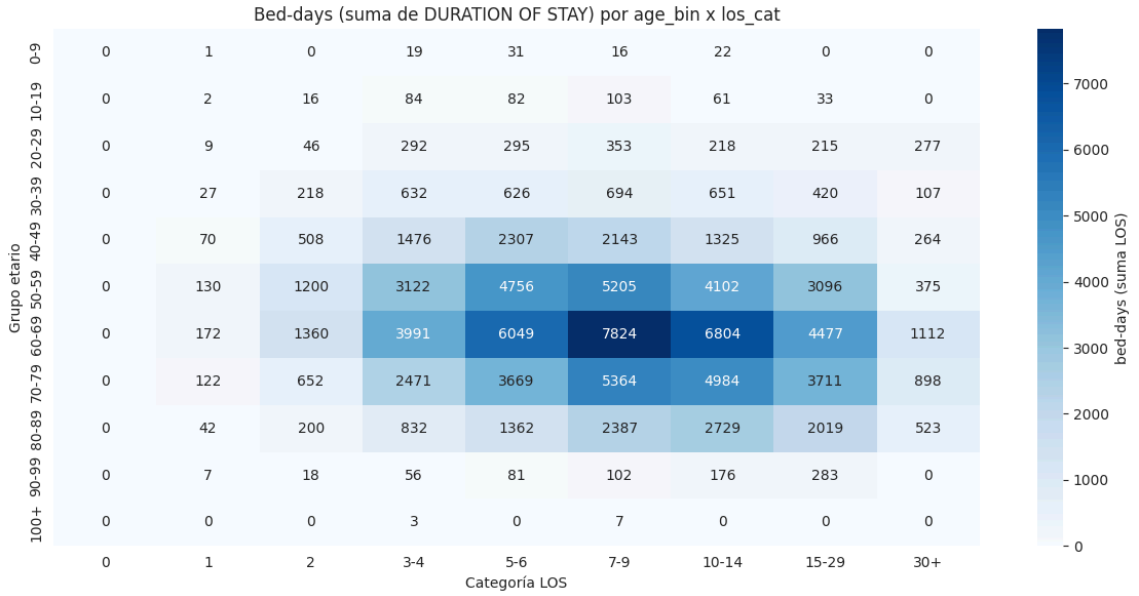
La mayoría de las admisiones corresponden a estancias cortas (pico en 2 días; mediana de 5 días) y la población atendida se concentra alrededor de los 60 años. Sin embargo, un pequeño grupo de pacientes con internaciones prolongadas (mayor o igual a 15 - 30 días) representa una proporción significativa del consumo total de camas.

Entonces proponemos evaluar una relación entre : Edad - Días hospitalarios - Cantidad de pacientes.

4. Análisis bivariado

En el siguiente mapa de calor puede observarse que:

- Cada celda contiene la **suma total de días-cama** que generaron los pacientes de ese grupo etario que tuvieron ese rango de duración de estancia.
- **Concentración central:** la mayor parte del consumo de camas está concentrado en los grupos etarios **50–79 años**.
- **Picos por categoría de LOS:** las columnas con más bed-days son típicamente las de **5–6, 7–9 y 10–14** días. En particular la combinación **60–69 / 7–9** es la celda con mayor bed-days
- **La población joven (<40)** aporta muy pocos bed-days en todas las categorías.



Como conclusión a este mapa de calor: La carga real sobre camas está en adultos de 50 - 79 años con estancias de 5 - 14 días. Esas son las “zonas calientes” donde se concentra la mayor proporción de bed-days y, por tanto, donde una mejora en gestión tendrá más impacto.

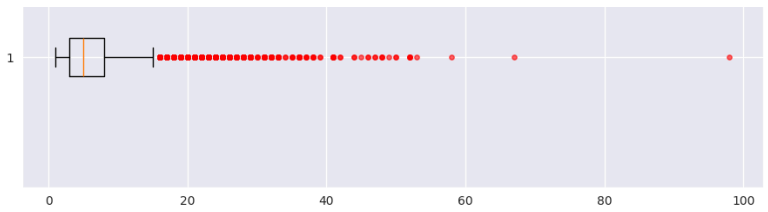
CONCLUSIÓN DE ANÁLISIS BIVARIADO**:**

Las estancias cortas son mayoría, por lo que conviene agilizar admisiones y altas para mejorar la rotación y evitar congestión.

El mayor consumo de camas proviene de pacientes de 50–79 años con internaciones de 5–14 días; identificarlos temprano y acortar su estancia de forma segura es clave para reducir la ocupación total.

5. Outliers

- Boxplot global



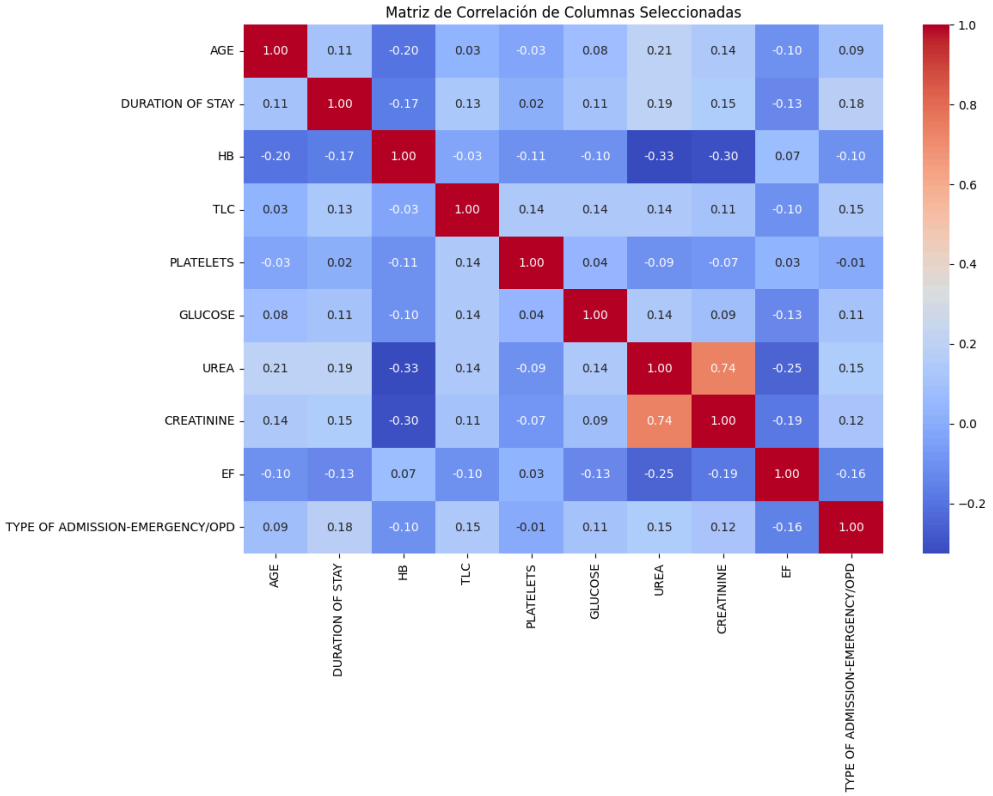
Outliers - Días de internación

La mayoría de las internaciones son cortas (mediana 5 días), pero existe un grupo reducido con estancias muy largas (30+ días) que concentra gran parte del uso de camas.

Estos casos deben validarse y gestionarse con métodos robustos, pero son clínicamente posibles y viables.

Como conclusión, el foco se centrará exclusivamente donde se encuentra la mayor concentración de ocupación de camas, con rotaciones frecuentes.

6. Correlación entre variables numéricas

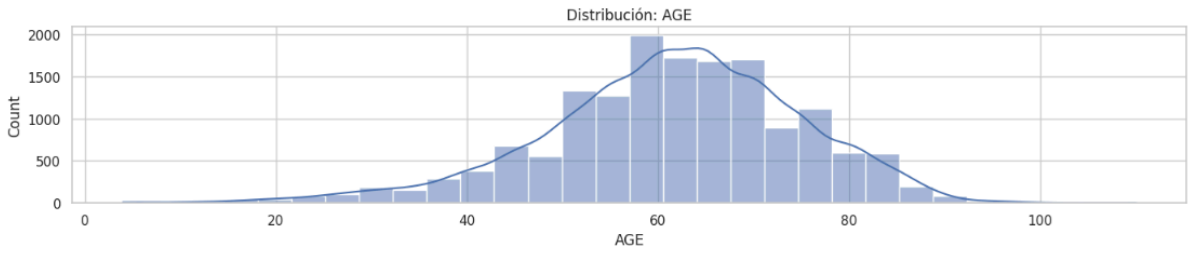


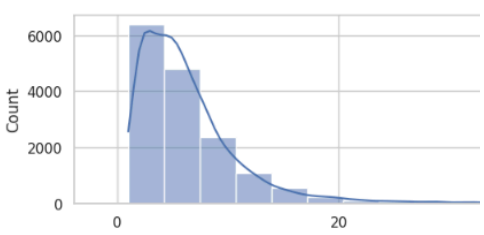
Heatmap (Pearson) — resumen: muestra correlaciones lineales entre numéricas clave (rojo=positiva, azul=negativa). Destaca **urea–creatinina** con alta correlación → captan la **función renal** y son **redundantes**

7. Resumen estadístico

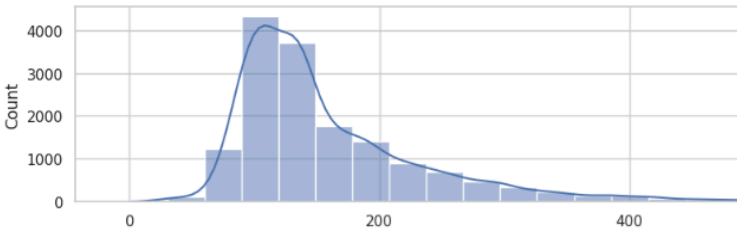
A continuación se presenta un resumen estadístico descriptivo de las principales variables numéricas del dataset (edad, duración de la internación y principales parámetros de laboratorio). El cuadro resume el número de observaciones válidas, medias, medianas, desviaciones estándar, mínimos y máximos.

	num_observ	promedio	desviación estándar	mínimo	25%	mediana	75%	máximo
AGE	15,757.00	61.43	13.42	4.00	54.00	62.00	70.00	110.00
DURATION OF STAY	15,757.00	6.42	5.01	1.00	3.00	5.00	8.00	98.00
HB	15,757.00	12.22	2.31	3.00	10.70	12.40	13.80	26.50
TLC	15,757.00	11.50	7.41	0.10	7.90	10.10	13.30	314.00
PLATELETS	15,757.00	238.38	102.87	0.58	173.00	226.00	287.00	1,179.00
GLUCOSE	15,757.00	161.29	82.05	1.20	107.00	136.00	190.00	888.00
UREA	15,757.00	49.93	42.21	0.10	25.00	35.00	57.00	495.00
CREATININE	15,757.00	1.34	1.19	0.07	0.78	1.00	1.40	15.63
EF	15,757.00	43.31	12.77	14.00	34.00	42.00	60.00	60.00

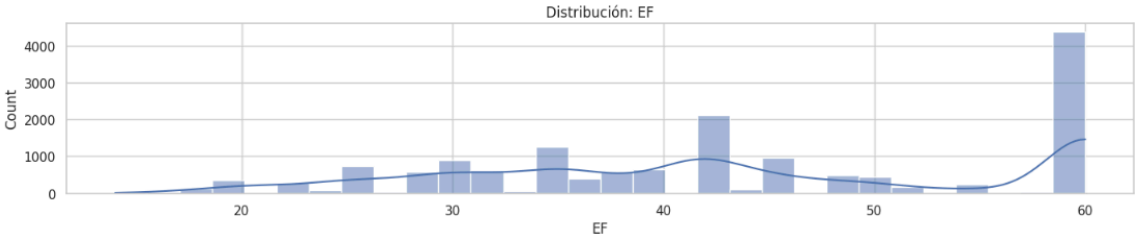




Días de internación vs. cantidad de ingresados



Cantidad de glucosa vs. cantidad de ingresados



Los datos muestran:

- **Edad:** media **61** (±**13**). Población mayormente adulta/avanzada.
- **LOS:** mediana **5** días; media **6,4** → **cola derecha**.
- **Labs:** alta **asimetría** y **outliers** → priorizar **mediana/IQR** sobre la media.

Conclusión: cohorte de mayor edad y con comorbilidades, lo que **explica la variabilidad** del LOS y refuerza el enfoque con **estadísticos robustos** y modelos tolerantes a colas.

8. **Síntesis final del EDA**:

El análisis muestra que el mayor volumen de pacientes permanece internado entre 2 y 5 días, lo que indica una oportunidad clara para optimizar la gestión a través de circuitos ambulatorios o altas más tempranas en casos menos críticos. Si bien las estancias prolongadas siguen siendo relevantes y deben preverse por su impacto en la ocupación, el verdadero margen de mejora está en reducir hospitalizaciones simples y repetitivas. Anticipar este comportamiento y diseñar estrategias diferenciadas para ambos escenarios permitiría una gestión de camas más eficiente y menos reactiva.

Preparación de datos

Imputación de nulos (solo numéricas): KNN Imputer (p. ej., k=10, distancia euclídea) ajustado en **training** y aplicado a **validación/test** con los mismos parámetros para evitar fuga de información.
Codificación y escalado: codificación estándar de categóricas y normalización de continuas según corresponda al algoritmo (sin ingeniería de atributos).
Sin feature engineering: en pruebas internas, **ensuciaba el modelo y empeoraba las métricas** respecto a la línea base.

Implicancias metodológicas

La exclusión de **LOS > 15 días** redefine explícitamente el **dominio del problema**: el modelo **no** cubre estancias muy prolongadas; se centra en la **gestión de rotación habitual** de Cardiología (planificación de camas, agenda de procedimientos y altas tempranas).
Esta decisión mejora la **validez operativa** para el servicio y reduce la varianza inducida por outliers, alineando el entrenamiento con el objetivo de negocio.

Modelado y Evaluación

Diseño experimental

Objetivo: predecir (a) **LOS en días** (regresión) y (b) **riesgo de long-stay** (clasificación binaria) para gestionar saturación de camas.

Etiquetado long-stay: umbral fijo = 7 días sobre **TARGET_REG** (días).

Esquema de evaluación:

- **Validación externa (hold-out): división inicial train/test (p.ej., 80/20).:**
  - **El test quedó intacto hasta la etapa final.**
  - **La selección de modelos/hiperparámetros y cualquier decisión de umbral se realizó exclusivamente con el train.**
- **Validación interna (solo en train): KFold 5-fold para comparar baselines vs. modelos y estabilizar estimaciones.**

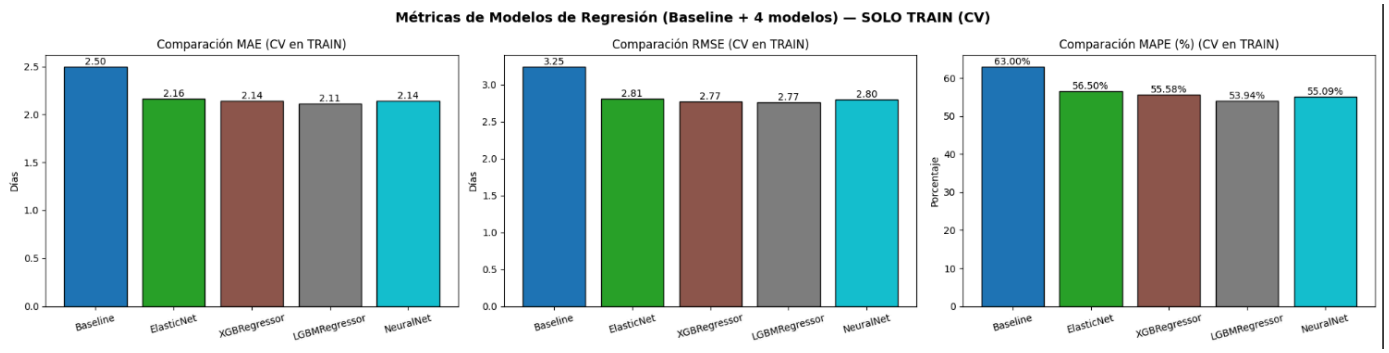
Métricas:Regresión: MAE, RMSE y MAPE. y **Clasificación: AP** (métrica primaria por desbalance), **ROC-AUC** y **Recall@20%** (recuperación en el **top-20%** de pacientes con mayor riesgo, proxy directo para priorización operativa).

Pipeline: preprocess común en todos los modelos (imputación/escala/one-hot) acoplado en **Pipeline** para evitar leakage.

Modelos: ElasticNet (con GridSearch), XGBoost, LightGBM y Red Neuronal (Keras) para **regresión**; Regresión Logística (GridSearch), XGBoost y Red Neuronal para **clasificación**

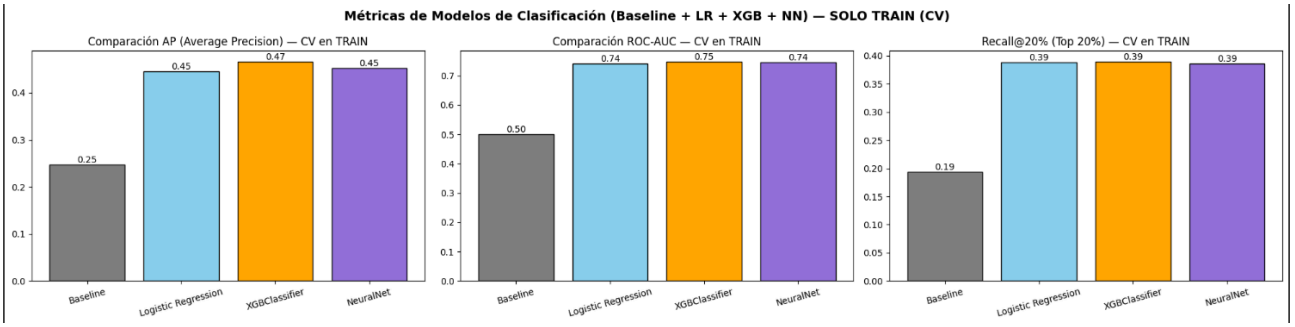
Elección del modelo

Regresión (LOS en días)



Decisión: con menor MAE, RMSE y MAPE adoptar LGBM Regressor como modelo principal.

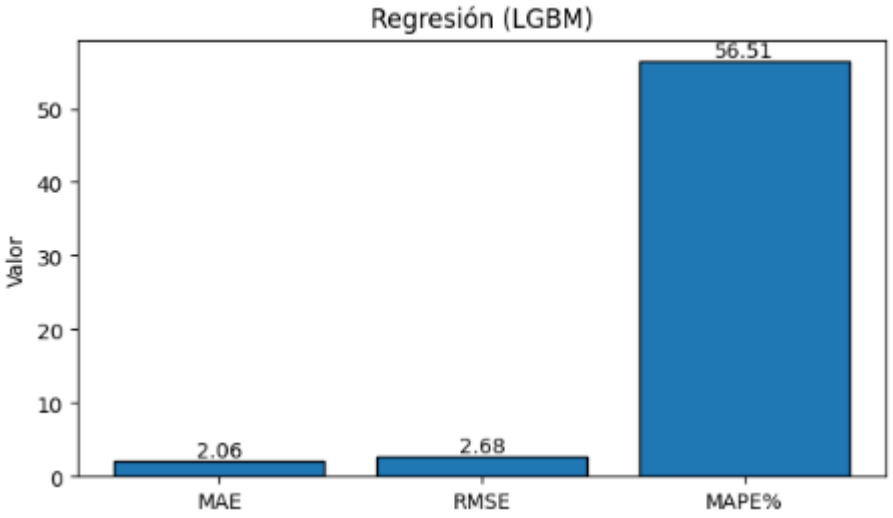
Clasificación (long-stay)



Decisión: con la AP más alta adoptar XGBClassifier como modelo principal

Resultados

Regresión - LGBM (LOS en días)

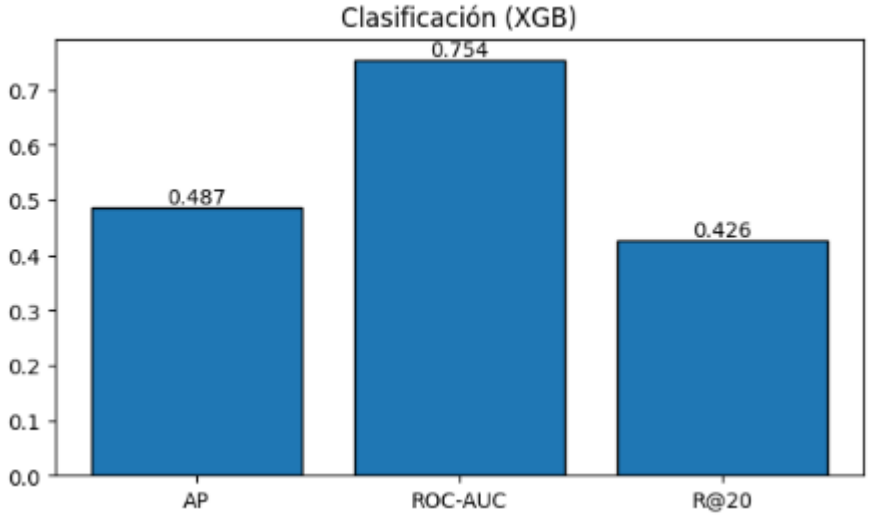


Lectura (y KPI):

KPI declarado: **MAE ≤ 1,5–2,0 días**. El **LGBM** queda **muy cercano** (2,06) y con **RMSE** competitivo; se considera **aceptable para uso operativo** por su interpretabilidad en días.

Decisión: para planificación de camas (proyección de ocupación y **bed-days**), elegimos **LGBM** como **modelo de regresión recomendado**.

Clasificación - XGBM (long-stay)



Selección operativa (y KPI):

- KPI declarados: **PR-AUC > baseline** (cumplido) y **Recall@Top-20% ≥ 60%** (**no cumplido**; mejor valor = **0,426** con XGB).

- Aunque la **RN** maximiza **AP/ROC-AUC**, para alerta temprana priorizamos **Recall@20%** (a quiénes revisamos primero). Bajo ese criterio, seleccionamos **XGBClassifier** (Recall@20% = 0,426).

- **Estrategia de uso:** dado que el **Recall@20%** está **por debajo del 60%**, la **clasificación se operará como *recomendador*** (ranking de riesgo) que guía la auditoría del **top-k** (20% parametrizable). Sigue siendo **útil** para dirigir acciones (altas tempranas, reprogramaciones, derivaciones) mientras trabajamos para alcanzar el KPI.

Implicancias para la gestión

- **Regresión (LGBM):** estimaciones de LOS con **error absoluto bajo y estable** → soporte para **proyección de ocupación** y **cálculo de bed-days** esperados.
- **Clasificación (XGB):** ranking de riesgo sobre el universo internado; el equipo **audita el top-20%** (o el percentil que defina la capacidad) para priorizar intervenciones.
- **Umbral/Top-k ajustables:** el **20%** es **parametrizable** según disponibilidad real de camas y recursos, manteniendo el mismo pipeline y métricas.

Controles y limitaciones

- **Calidad y trazabilidad del dato.** La falta de conocimiento directo del **data entry** dificulta la **selección robusta de variables** y la interpretación causal de señales. Puede haber **inconsistencias y sesgos** no documentados

- **No estacionariedad.** Cambios estacionales, de protocolos clínicos, mezcla de casos y disponibilidad de camas pueden desplazar la distribución de **LOS** y del umbral de **long-stay (7 días)**. Además, al tratarse de un **dataset externo (Kaggle)** con **geolocalización y estacionalidad desconocidas**, existe riesgo de que el patrón observado no refleje la operación local.

- **Generalización limitada.** El modelo fue entrenado sobre **Cardiología**; su desempeño puede degradarse en otros servicios o hospitales si no se **recalibra** y **reentrena** con datos propios de esos ámbitos.

Decisiones finales

- **Regresión (LOS): LightGBM** (mejor MAE/RMSE; **MAE cercano al KPI → se adopta**).

- **Clasificación (long-stay): XGBClassifier** (mejor Recall@20%) **como recomendador** hasta alcanzar el KPI de **60%** en Recall@Top-20%.

Impacto en el negocio

El modelo de duración de estancia (LOS) y la proyección de ocupación se aplica exclusivamente a las camas de Cardiología de internación). Su valor es operativo y medible: optimiza recursos del servicio, mejora tiempos de atención y entrega insumos para decisiones tácticas del día a día.

1) Eficiencia financiera y uso de recursos

- Días-cama evitables ↓: prioriza pacientes de alto riesgo de larga estancia para activar planes de alta temprana segura e interconsultas dirigidas.

- Dotaciones alineadas a la demanda: anticipa picos de ocupación del sector de Cardiología/UCO, ajustando turnos de enfermería y guardias médicas del servicio.

- Coordinación con procedimientos: cruza la ocupación prevista con agenda de Hemodinamia/Electrofisiología, evitando sobre-reservas y reprogramaciones.

2) Calidad y seguridad del paciente

- Menos espera a cama de Cardiología: asignación más oportuna desde Guardia/traslados internos.
- Gestión proactiva de long-stay: disparadores para objetivos diarios

- Menos cancelaciones por falta de cama en procedimientos cardiológicos programados.

Conclusiones

- El análisis confirma que la ocupación de camas puede anticiparse combinando predicción de LOS (regresión) y priorización de casos (clasificación) al momento del ingreso.
- El modelo de regresión alcanza desempeño útil para planificar días-cama esperados por paciente y servicio, suficiente para alimentar una proyección operativa de ocupación.
- El modelo de clasificación no sustituye a la regresión, pero añade valor como ranking para enfocar la gestión en los pacientes con mayor probabilidad de estancias prolongadas (top-k ajustable según capacidad).
- Las decisiones metodológicas (exclusión de LOS >15 días, imputación KNN, sin FE agresivo) mejoran estabilidad y alinean el modelo con el objetivo de rotación habitual de Cardiología.
- Principales límites: posible no-estacionariedad (cambios estacionales/protocolos), generalización restringida a Cardiología y calidad de dato heterogénea en algunas variables.

Recomendaciones

- Operativizar la proyección diaria de ocupación (días-cama estimados) e incorporar un ranking de pacientes de mayor riesgo para auditoría clínica y coordinación de altas.
- Mantener recalibraciones periódicas y monitoreo de drift (alertas si cambia la distribución de LOS o la mezcla de casos).
- Fortalecer la calidad de datos: trazabilidad de laboratorios y eventos críticos (p. ej., ingreso a UTI), diccionario de variables “confiables” y criterios de completitud.
- Antes de escalar, validar externamente en otro servicio/unidad con re-entrenamiento local y medir impacto en indicadores (días-cama evitados, cancelaciones y tiempos de espera).