Desarrollo de un Sistema de Predicción del Tiempo de Estancia en la Unidad de observación de Urgencias mediante Machine Learning

Para obtener el título de:

Ingeniero Industrial

Por:

Juan Pablo Cancelado Caro

Asesor:

Juan Fernando Pérez Bernal

Presentado a:

Universidad de los andes

Departamento de ingeniería Industrial

2024

Resumen:

Uno de los mayores desafíos para el personal administrativo de las Instituciones Prestadoras de servicios de Salud (IPS) en las unidades de observación de urgencias es gestionar eficientemente el tiempo de estancia de los pacientes, lo cual impacta directamente la disponibilidad de camas y la capacidad de respuesta ante nuevas emergencias. Este proyecto se enfoca en desarrollar un sistema predictivo basado en machine learning para estimar el tiempo de estancia de los pacientes en la unidad de observación de urgencias ocupando una cama hospitalaria, utilizando datos de 55 IPS de Medellín, obtenidos de MEDATA, que incluyen información demográfica, características clínicas, información administrativa y tiempos del paciente (Alcaldía de Medellín, 2023). Se entrenaron y evaluaron 4 diferentes modelos de machine learning: CatBoost, LightGBM, XGBoost y redes neuronales con embeddings. El modelo seleccionado se implementó en una herramienta web interactiva que permite los auxiliares administrativos y los coordinadores de recursos de las IPS ingresar datos del paciente y recibir predicciones tanto rápidas como precisas junto con estadísticas relevantes. Los resultados muestran que el modelo CatBoost fue el más preciso, con un MAE promedio en los datos de validación de 8.875 horas y una desviación estándar de 0.026 horas usando validación cruzada de 5 folds. Se logró reducir el número de variables sin afectar significativamente la precisión, optimizando la eficiencia del modelo y mejorando la experiencia del usuario.

Enlace del repositorio donde se puede ver la limpieza de los datos, la ingeniería de características, el análisis exploratorio de los datos y el modelamiento: https://github.com/Juanxtron/Proyectodegrado_2/tree/main

Contenido:

pág.

1. Introducción 4
   - 1.1 Descripción del problema 4
   - 1.2 Objetivos 5
     - 1.2.1 Objetivo General 5
     - 1.2.2 Objetivos Específicos 5
   - 1.3 Alcance, limitaciones y producto final 6
2. Marco de referencia 7
   - 2.1 Estado del arte 7
   - 2.2 Marco normativo 8

3. Metodología 8
3.1 Etapas desarrollo modelos de ML 8
3.2 Metodología análisis exploratorio de datos 9
4. Base de datos 10
4.1 Datos originales 10
4.2 Esquemas de codificación 11
4.2.1 Codificación prestador y EAPB 12
4.2.2 Codificación departamentos y municipios 12
4.2.3 Codificación CIE-10 12
4.3 Limpieza de los datos 12
4.4 Variable a predecir: Tiempo en observación 13
4.5 Ingeniería de características 14
4.6 Análisis exploratorio de los datos 15
5. Investigación experimental 21
5.1 Métrica evaluación de los modelos 21
5.2 Modelos ML y Calibración de hiperparametros 21
5.3 Resultados 22
5.3.1 Resultado de los modelos 22
5.3.2 Importancia de las variables 25
6. Implementación 31
6.1 Flujo de la Página web 31
6.2 Modelado de datos 32
6.3 Arquitectura del sistema 33
6.4 Seguridad y privacidad 34
7. Conclusiones y consideraciones finales 35
8. Referencias 36

9. Apéndice.....38

9.1 Apéndice A: Marco conceptual……38

1. Introducción

1.1 Descripción del problema

La gestión eficiente de recursos en entornos hospitalarios es esencial para garantizar atención de calidad y optimizar infraestructuras. Uno de los mayores desafíos que enfrenta el personal administrativo de las IPS en las unidades de urgencias, especialmente el auxiliar administrativo, encargado de registrar ingresos, actualizar datos de los pacientes y gestionar documentación, y los coordinadores de recursos, responsables de planificar la asignación de personal, garantizar la disponibilidad de insumos y coordinar traslados, es la gestión del tiempo de estancia de los pacientes. Este desafío impacta directamente en la disponibilidad de camas y la capacidad de respuesta ante emergencias (Restrepo et al., 2018). La saturación de camas, influida por factores internos y externos, genera congestión, tiempos de espera elevados e insatisfacción de usuarios, además de aumentar el estrés del personal de salud y comprometer la capacidad de atender nuevos casos (Restrepo et al., 2018).

En la actualidad, los sistemas de salud generan una gran cantidad de datos administrativos relacionados con pacientes, departamentos, costos de material médico, disponibilidad de camas, enfermedades, entre otros (Gutierrez et al., 2021). Esta vasta cantidad de datos, que se almacena en bases de datos computacionales, tiene el potencial de ser utilizada para mejorar la gestión interna de recursos, reducir costos, y optimizar la atención al paciente (Gutierrez et al., 2021). Sin embargo, muchas instituciones de salud no cuentan con los sistemas y modelos integrados necesarios para mejorar la toma de decisiones y la eficiencia operativa en los hospitales. En el caso de la gestión de camas en la unidad de observación de urgencias (Ver más información de esta unidad en el Anexo A), la falta de integración entre plataformas y la fragmentación de los datos pueden dificultar el seguimiento de métricas clave, como el tiempo promedio de estancia de los pacientes. Esto es particularmente relevante dado que, según la Resolución 3100 de 2019 del Ministerio de Salud y Protección Social (MSPS), todas las IPS deben cumplir con una serie de estándares de habilitación, que incluyen no solo talento humano y gestión de procesos, si no infraestructura, lo cual incluye camas hospitalarias (MinSalud, 2019).

Tradicionalmente, la predicción de la fecha de alta de un paciente se ha basado en la estancia media hospitalaria (ALOS), calculada al sumar los días de cama de cada paciente dado de alta durante un periodo y dividir el total entre el número de pacientes. Aunque este método permite una estimación general, es impreciso y deja un amplio margen de mejora (AltexSoft, 2022). Durante la pandemia, la Gobernación de Antioquia utilizó un enfoque más avanzado, analizando rutas de cama de pacientes con COVID-19 y estimando la duración de ocupación (LoS) mediante

modelos de Supervivencia Paramétrica (AFT) con distribuciones como Weibull y Log Normal. Estos cálculos consideraron factores como edad, sexo y región. Aunque precisos, se concluyó que el algoritmo no era adecuado para el contexto local, dada la diferencia en factores y tiempos de estancia frente a otros países.

En este contexto, el uso de modelos de machine learning se han convertido en herramientas valiosas en la gestión de servicios de salud, permitiendo predicciones clínicas y optimización operativa (Pedrero et al., 2021). Al analizar grandes volúmenes de datos, estos algoritmos identifican patrones que mejoran la eficiencia, la integración de servicios, las predicciones y la toma de decisiones informadas, contribuyendo a una gestión más efectiva de recursos y servicios de calidad (Pedrero et al., 2021).

1.2 Objetivos:

1.2.1 Objetivo general

Desarrollar un sistema predictivo basado en técnicas de machine learning para estimar el tiempo de estancia de los pacientes en la unidad de observación de urgencias, e implementarlo en una herramienta web interactiva para los auxiliares administrativos y los coordinadores de recursos de las IPS.

1.2.2 Objetivos específicos

- Realizar la limpieza y el preprocesamiento de los datos contenidos en el dataset, asegurando la calidad y la coherencia de la información para su uso en el modelo predictivo.

- Realizar un análisis descriptivo y exploratorio de los datos disponibles, en la base de datos con el fin de identificar patrones y tendencias significativas.

- Evaluar 4 diferentes métodos de machine learning para seleccionar el modelo más adecuado para predecir el tiempo de estancia de los pacientes en la unidad de observación de urgencias.

- Diseñar y desarrollar una herramienta web interactiva que permita a los auxiliares administrativos y los coordinadores de recursos de las IPS ingresar información del paciente y recibir predicciones sobre el tiempo de estancia utilizando el modelo de machine learning seleccionado, acompañadas de gráficos y reportes relevantes.

- Desplegar la página web y la base de datos en un entorno en la nube, garantizando un acceso seguro, eficiente y confiable para todos los usuarios.

1.3 Alcances, limitaciones y producto final.

Alcances del Proyecto:

1. **Modelos de Predicción con Machine Learning:** El proyecto contemplará el desarrollo de al menos cuatro algoritmos de machine learning. Aunque se podrían probar muchos más, el enfoque se centrará en aquellos que, según la bibliografía, han demostrado ser los mejores para la problemática específica, con el objetivo de limitar el alcance del trabajo y garantizar que los modelos seleccionados sean robustos estadísticamente.
2. **Enfoque usuario final:** El proyecto está diseñado específicamente para satisfacer las necesidades de los auxiliares administrativos y los coordinadores de recursos de las IPS, quienes son los usuarios finales de esta herramienta. Esto implica que todas las funcionalidades, desde la predicción de tiempos de estancia hasta la visualización de resultados, están enfocadas en proporcionar información relevante para la toma de decisiones administrativas, como la planificación de recursos y la gestión de camas en la unidad de observación de urgencias. Además, el modelo está optimizado para generar predicciones precisas con los datos limitados a los que los administrativos pueden acceder
3. **Simplicidad de la Página Web:** Aunque la página podría expandirse con más funcionalidades en el futuro, el enfoque principal será proporcionar de manera interactiva acceso al modelo desarrollado y ofrecer una interfaz intuitiva que facilite la gestión de las predicciones y estadísticas relevantes para la administración de camas en la unidad de observación de urgencias.
4. **Integración de una Base de Datos en la Nube:** La infraestructura se centrará en lo esencial para el funcionamiento de la página web, asegurando que los datos sean accesibles y gestionables en un entorno seguro. A pesar de que la infraestructura de una base de datos en la nube puede ser compleja, el proyecto se limitará a lo funcional, garantizando el almacenamiento seguro y eficiente de los datos críticos.

Limitaciones del Proyecto:

1. **Limitación de Generalización:** El proyecto está limitado por el tamaño del dataset, que incluye información de solo 55 IPS de Medellín, una fracción de las 1048 IPS en la ciudad. Esto impide generalizar los resultados a todas las IPS de Medellín o a nivel nacional, restringiendo la solución a las IPS incluidas en el dataset.
2. **Limitación por Complejidad de Factores Relevantes:** El modelo depende exclusivamente de los datos disponibles, sin incluir factores adicionales como incidencia de bacterias intrahospitalarias, recursos médicos o tecnológicos. Esto limita su capacidad para capturar todas las variables relevantes, reduciendo su precisión y alcance predictivo.

2.Marco de referencia

2.2 Estado del arte.

Peres et al. (2021) revisaron modelos para predecir la estancia en unidades de cuidados intensivos (Intensive Care Units, ICU), concluyendo que métodos basados en datos, como Random Forest y Support Vector Regression (SVR) (véase Apéndice A para definiciones detalladas), superan a los métodos estadísticos tradicionales, aunque la elección del modelo depende del conjunto de datos específico. Destacaron, además, la importancia del preprocesamiento de datos para mejorar las predicciones.

Otros estudios destacan la inclusión de datos no estructurados. Chrusciel et al. (2021) analizaron cómo estos datos, como registros médicos electrónicos, mejoran la predicción de la estancia hospitalaria. Utilizando modelos de Random Forest, encontraron que los datos no estructurados ofrecen una ligera ventaja en la precisión frente a los estructurados. En Colombia, Iglesias (2017) empleó el Registro Individual de Prestación de Servicios de Salud (RIPS) para desarrollar modelos de predicción de Length of Stay (LoS) al momento de la admisión. El modelo con mejor desempeño fue XGBoost (véase Apéndice A para definiciones detalladas), aunque se señaló que la ausencia de variables adicionales, como resultados de pruebas de laboratorio, limitó su precisión. Restrepo Correa y Sánchez Jiménez (2023) desarrollaron un modelo predictivo para estimar la estancia hospitalaria en pacientes geriátricos en Medellín, empleando Random Forest, que explicó gran parte de la variabilidad en los datos. Martínez (2021) implementó un sistema de inteligencia artificial para predecir el tiempo de estancia en ICU en Medellín, logrando una precisión del 69% con Random Forest y Naive Bayes (véase Apéndice A).

En la mayoría de estos trabajos, la duración de la estancia hospitalaria fue categorizada en niveles como "alto", "medio" o "bajo". Este proyecto adopta un enfoque diferente, tratando la variable Tiempo en observación como continua, calculada en horas a partir de las marcas de tiempo de entrada y salida de los pacientes. Además, este proyecto incorpora modelos más recientes de boosting, como CatBoost y LightGBM (véase Apéndice A para definiciones detalladas), así como técnicas avanzadas de redes neuronales, incluyendo capas de embeddings (véase Apéndice A para definiciones detalladas) para variables categóricas y redes neuronales densas.

2.3 Marco normativo.

En Colombia, la Ley Estatutaria 1581 de 2012 regula la protección de datos personales. Por ello, en este proyecto es esencial garantizar que la información de los usuarios recopilada en la página web (como nombres y contraseñas) se gestione de forma segura para proteger su privacidad (GOV, 2012). Así mismo, la Resolución 8430 de 1993 establece los lineamientos para la investigación en salud, exigiendo que el manejo de datos de pacientes cumpla con criterios éticos y procedimientos aprobados (MinSalud, 1993). Aunque los datos utilizados en este proyecto son públicos, la Ley 1437 de 2011 asegura que el acceso a información pública en proyectos con datos abiertos se realice conforme a las disposiciones legales (GOV, 2011).

El uso de información clínica, específicamente los diagnósticos de los pacientes extraídos de la historia clínica, es necesario para predecir el tiempo de estancia hospitalaria. Según la Resolución 3100 de 2019, la historia clínica es un documento privado y solo puede ser consultado con el consentimiento informado del paciente (MinSalud, 2019). En esta resolución, también se menciona en el estándar de dotación para los servicios de hospitalización, se establece que deben contar con un número de camas hospitalarias adecuado según la demanda y la oferta del servicio para no afectar la capacidad de respuesta de las IPS (MinSalud, 2019).

Los administrativos de las IPS, aunque tienen restricciones para acceder a la historia clínica completa, pueden acceder a información clave como el diagnóstico y los recursos utilizados en el tratamiento. Esto es necesario para procesos administrativos como facturación y auditoría. La Ley 100 de 1993 y la Resolución 5261 de 1994 permiten que las entidades encargadas de estos procesos accedan a los datos clínicos necesarios, garantizando siempre la protección de datos sensibles y la privacidad del paciente. (GOV, 1993; MinSalud, 1994).

Por último, la base de datos utiliza la Clasificación Internacional de Enfermedades (CIE-10), estándar global desarrollado por la OMS. En Colombia, su uso es obligatorio en todos los establecimientos de salud según la Resolución 3374 de 2000. (MinSalud, 2000).

Este proyecto de grado se regirá bajo todas estas normativas para garantizar el cumplimiento legal y ético en el manejo de datos de salud y la implementación de una solución técnica alineada con los estándares nacionales e internacionales.

3. METODOLOGIA

3.2 Etapas desarrollo modelos de ML.

En el presente trabajo se utilizará el marco de trabajo propuesto por Nascimento et al. (2019), que proporciona una estructura clara para el desarrollo de sistemas de aprendizaje automático (ML) en entornos empresariales. Este marco ha sido diseñado específicamente para abordar los desafíos que enfrentan los desarrolladores al implementar sistemas de ML, enfocándose en el cumplimiento de métricas empresariales. Su aplicación en este contexto es especialmente relevante, ya que permitirá no solo el desarrollo efectivo del modelo, sino también su implementación dentro de una plataforma web para ser utilizada por usuarios finales. El proceso incluye cuatro etapas principales:

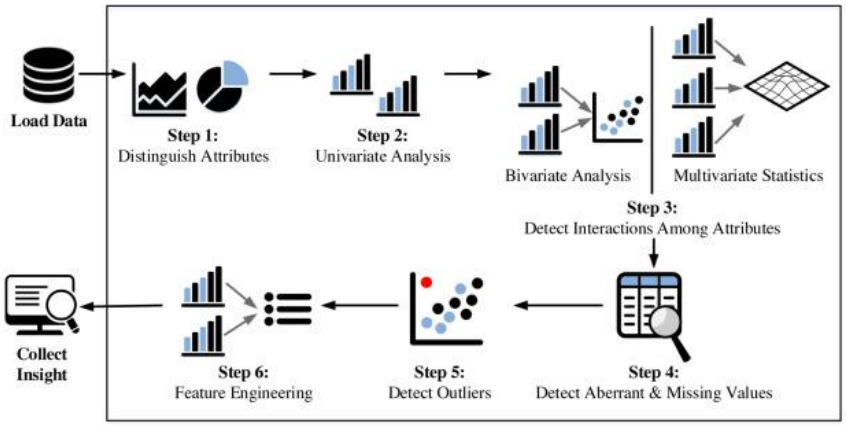
1. **Comprensión del problema**: Se define claramente el problema a resolver y las métricas de éxito, alineando el modelo con los objetivos del proyecto.
2. **Manejo de datos**: Implica la adquisición, estructuración y preparación de los datos para el entrenamiento del modelo. La calidad en esta etapa es crucial para el éxito del sistema.

3. **Construcción del modelo**: Se entrenan y evalúan varios modelos, seleccionando el más adecuado según las métricas definidas y preparándolo para su despliegue en producción.
4. **Monitoreo del modelo**: Una vez en producción, el modelo es monitoreado y ajustado para mantener su rendimiento a lo largo del tiempo, adaptándose a nuevos datos.

3.3 Metodología análisis exploratorio de los datos.

Se detalla el proceso de análisis exploratorio de datos (EDA) (véase Apéndice A para definiciones detalladas), para tratar los datos, analizar relaciones entre variables explicativas y la variable de respuesta, y preparar los modelos de machine learning. Este análisis se basa en el marco de trabajo de Ghosh et al. (2018), orientado a la comprensión de los datos y la detección de patrones clave. Tal y como se observa en la figura 1 el proceso de EDA se incluye análisis univariado, bivariado y multivariado, tratamiento de valores atípicos y la ingeniería de características para los modelos. Se realizará un cambio a esta metodología y es que primero se identificarán y manejarán valores aberrantes y faltantes, se tratarán los valores atípicos y se hará la ingeniería de características. Posteriormente, si se realizar un análisis bivariado y multivariado para explorar relaciones significativas entre variables.

Figura 1: Los pasos fundamentales del proceso de análisis exploratorio de datos.



Nota: Adaptado de *A comprehensive review of tools for exploratory analysis of tabular industrial datasets*, por A. Ghosh, M. Nashaat, J. Miller, S. Quader, y C. Marston, 2018, *Visual Informatics, 2*(4), 235-253. Copyright 2018 por los autores.

4. BASES DE DATOS

4.1 Base de datos original.

El conjunto de datos utilizado en este proyecto registra la estancia de pacientes en la unidad de observación de urgencias, incluyendo solo aquellos que ocuparon una camilla para observación continua tras la atención inicial. Se excluyen los casos que solo recibieron atención inicial sin observación prolongada. Los datos contienen 1’291.374 registros de un total de 55

IPS. El conjunto incluye 23 variables clave: datos administrativos (como número de factura y código del prestador), detalles clínicos (diagnósticos y estado al salir) y datos demográficos (edad, sexo, zona de residencia). Excepto por la edad y el año, todas las variables son categóricas. A continuación, se muestra una descripción general de cada una de las variables:

NumeroFactura: Número de factura.

CodigoPrestador: Código del prestador. Valor único asignado a la IPS en el proceso de habilitación por el ente territorial (SDS). Es un número de habilitación generado por el ente territorial correspondiente, de 12 caracteres, incluye sedes.

FechaIngreso: Fecha de ingreso.

HoraIngreso: Hora de ingreso.

CausaExterna: Causa externa. Código que representa el tipo de causa externa, como accidentes de trabajo, accidentes de tránsito, accidentes rábicos, accidentes ofídicos, otros tipos de accidentes, eventos catastróficos, lesiones por agresión, lesiones auto infligidas, sospechas de maltrato físico, sospechas de abuso sexual, sospechas de violencia sexual, sospechas de maltrato emocional, enfermedades generales, enfermedades laborales u otras causas.

CodigoDiagnosticoPrincipalSalida: Código diagnóstico principal de salida. Código del diagnóstico según la tabla de referencia CIE 10.

CodigoDiagnosticoRelNISalida: Código diagnóstico relacionado 1 de salida. Código del diagnóstico según la tabla de referencia CIE 10.

CodigoDiagnosticoRelNasalida: Código diagnóstico relacionado 2 de salida. Código del diagnóstico según la tabla de referencia CIE 10.

CodigoDiagnosticoRelNaSalida: Código diagnóstico relacionado 3 de salida. Código del diagnóstico según la tabla de referencia CIE 10.

DestinoUsuario: Destino del usuario después de la atención, indicando si fue dado de alta en urgencias, remitido a otro nivel de complejidad o hospitalizado.

EstadoSalida: Estado del paciente al salir, que puede ser vivo(a) o fallecido(a).

CausaBasicaMuerteUrgencias: Causa básica de muerte en urgencias. Código del diagnóstico según la tabla de referencia CIE 10.

FechaSalida: Fecha de salida.

HoraSalida: Hora de salida.

CodigoEAPB: Código EAPB. Código de la Entidad Administradora de Planes de Beneficios, máximo de 6 caracteres.

TipoUsuario: Tipo de usuario. Indica el tipo de afiliación del usuario, como contributivo, subsidiado, vinculado, particular, otros, víctimas con afiliación al Régimen Contributivo (R.C), víctimas con afiliación al Régimen Subsidiado (R.S) o víctimas no aseguradas.

Edad: Edad del paciente.

UnidadMedidaEdad: Unidad de medida de la edad, que puede estar en años, meses o días.

Sexo: Sexo del paciente, masculino (M) o femenino (F).

CodDepto: Código de departamento de residencia. Código correspondiente al departamento según la tabla de Departamentos.

CodMunicipio: Código del municipio de residencia. Código correspondiente al municipio según la tabla de Municipios.

ZonaResidencia: Zona de residencia, que puede ser urbana (U) o rural (R).

Ano: Año del servicio.

Cabe resaltar que este dataset es totalmente público y fue obtenido de la página de datos abierto de Medellín MEDATA (Alcaldía de Medellín, 2023).

4.2 Esquemas de codificación.

Durante el preprocesamiento de datos, se realizaron mapeos específicos para transformar las columnas codificadas en descripciones más comprensibles, facilitando así el análisis e interpretación. Utilizando la información del archivo JSON disponible en la página de MEDATA, se decodificaron columnas como CausaExterna, que originalmente contenía códigos numéricos, para reflejar descripciones detalladas como "Accidente de trabajo" o "Evento catastrófico". De manera similar, se mapeó la columna DestinoUsuario para representar opciones claras como "Alta de urgencias" o "Remisión a otro nivel de complejidad". La columna TipoUsuario también fue transformada para incluir descripciones como "Contributivo" o "Subsidiado", y la columna UnidadMedidaEdad fue convertida de códigos numéricos a descripciones como "años", "meses" o "días" (Alcaldía de Medellín, 2023).

4.2.1 Codificación prestador de servicios de salud (IPS) y Entidad Administradora de Planes de Beneficios (EAPB).

Para el mapeo de las Entidades Administradoras de Planes de Beneficios (EAPB), que inicialmente solo contaban con códigos, se utilizó una base de datos oficial proporcionada por la ADRES, la cual contiene información detallada sobre las entidades del Sistema General de Seguridad Social en Salud (SGSSS) (ADRES, 2022). Para el mapeo de los prestadores de servicios de salud (IPS), se empleó la base de datos pública de "Datos Abiertos Colombia", que

registra los prestadores habilitados en el país, permitiendo crear las variables "NombrePrestadorSede" y "NombreEAPB" con los nombres respectivos (GOV, 2024).

4.2.2 Codificación departamentos y municipios

Para realizar este mapeo, se utilizaron los datos proporcionados por la Secretaría Distrital de Salud, disponibles en un documento oficial titulado "Codificación de Municipios por Departamento" (saludcapital, s.f). Este documento presenta la estructura de los códigos de la siguiente manera: para cada municipio del país, se listan el código del departamento, el código del municipio, el nombre del municipio y el nombre del departamento correspondiente.

4.2.3 Codificación CIE-10

Para este mapeo, se utilizaron seis datasets de CIE-10, obtenidos de la página oficial del Centers for Medicare & Medicaid Services (CMS), cubriendo los años 2019 a 2024 (CMS, 2024). Dado que los registros del dataset abarcan varios años, era esencial utilizar datasets de diferentes años para capturar cualquier cambio o actualización en los códigos de diagnóstico.

Para llevar a cabo el mapeo, se tuvo en cuenta solo la descripción general de la enfermedad. Por lo tanto, si el código en la base de datos original correspondía a una variante específica de una enfermedad, se aproximaba a la descripción de la enfermedad general. Este enfoque permitió reducir la cantidad de categorías que quedarían con un solo registro, lo que ayudó a evitar problemas de dispersión en los datos. Además, esta estrategia redujo la dimensionalidad de las columnas, lo cual es crucial para evitar la maldición de la dimensionalidad en los modelos de machine learning.

Como parte de este proceso, se crearon nuevas variables, tales como DescripcionDiagnosticoPrincipal, DescripcionDiagnosticoRelacionado_uno, DescripcionDiagnosticoRelacionado_dos y DescripcionDiagnosticoRelacionado_tres, a partir de las columnas de códigos de diagnóstico correspondientes en el dataset original. Los valores nulos de estas últimas 3 columnas fueron reemplazados por "No tuvo", especificando que esos registros en blanco indicaban que el paciente no presentó diagnósticos relacionados 1, 2 o 3.

4.3. Limpieza de datos.

Primero, se transformaron las edades que estaban codificadas en diferentes unidades de medida (años, meses, días) utilizando la columna UnidadMedidaEdad. Las edades en meses fueron convertidas a años dividiendo el valor por 12, y las edades en días fueron convertidas a años dividiendo el valor por 365. Este proceso permitió uniformizar la columna Edad, haciendo que todos los registros fueran comparables. Adicionalmente, se identificaron y eliminaron registros con valores de edad claramente incorrectos. Específicamente, se encontraron 14 registros con edades superiores a 110 años, incluyendo un registro con una edad de 185 años, la cual no es posible en la realidad. Dado que solo se trataba de un pequeño número de registros, su eliminación no afectó significativamente las estadísticas globales de la columna Edad.

En segundo lugar, se llevaron a cabo varias operaciones de conversión de tipos de datos, asegurando que todas las columnas relevantes fueran tratadas como cadenas de texto (string) en las columnas CodigoPrestador, DestinoUsuario, TipoUsuario y UnidadMedidaEdad.

En tercer lugar, se eliminaron filas con valores inválidos o formatos incorrectos en la columna NombreMunicipio. Además, se eliminaron los registros donde el sexo del paciente estaba marcado como "I", ya que esta no es una categoría válida. Asimismo, se eliminaron registros con valores nulos, vacíos o iguales a "0" en la columna CodDepto. Posteriormente, se eliminaron los valores nulos en las columnas DescripcionDiagnosticoPrincipal y NombrePrestadorSede, ya que la cantidad de filas afectadas era baja en comparación con el tamaño total del conjunto de datos, por lo que su eliminación no comprometió la representatividad ni la diversidad de los datos finales.

Finalmente, para las columnas resultantes del mapeo mencionado anteriormente, se aplicaron filtros adicionales para asegurar que los datos estuvieran en el formato adecuado, eliminando aquellos registros que no cumplían con los criterios especificados como paréntesis adicionales o inconsistencias en los valores.

4.4. Variable a predecir: Tiempo en observación.

Existen varias medidas para analizar la saturación de los sistemas de salud, como la tasa de alta hospitalaria, que indica el número de pacientes que abandonan un hospital después de pasar al menos una noche (OECD, 2021). Sin embargo, este estudio se enfocará en la variable tiempo en observación, la cual indica cuanto tiempo se demoró el paciente en la unidad de observación de urgencias desde que registro su entrada hasta su salida. Esta se calculó a partir de las marcas de tiempo de ingreso y salida de cada paciente. Para ello, se combinaron las columnas FechaIngreso y HoraIngreso en una única columna de tipo datetime, llamada FechaHoraIngreso, y de forma similar, se unieron FechaSalida y HoraSalida en la columna FechaHoraSalida.

La fórmula utilizada para calcular esta variable fue la siguiente:

$$TiempoenObservación = \frac{FechaHoraSalida \; - \; FechaHoraIngreso}{3600 \; segundos \; por \; hora}$$

Este cálculo convierte la diferencia de tiempo entre la entrada y la salida de segundos a horas, lo cual es más útil para el análisis hospitalario. Por lo tanto, esta variable se va a manejar siempre en horas a lo largo de este documento.

4.5 Ingeniería de características.

Se estableció un umbral mínimo de media hora para la variable "TiempoenObservación" basándose en el estudio *Machine learning techniques to predict hospital length of stay (LOS) at the time of patient admission* (2017), donde médicos en Colombia recomendaron que 30 minutos es un límite inferior adecuado para el tiempo de estancia. Por lo tanto, valores extremadamente bajos, como menos de un minuto, son poco realistas debido a los procesos administrativos

hospitalarios. Dado que el mínimo inicial era de 0.01667 horas (1 minuto), se eliminaron las filas donde los datos fueran menores a media hora para reflejar mejor la realidad operativa. Se eliminaron el total 75468 filas, esto es el 4.86% de los datos.

Un aspecto clave en la mejora de los datos fue la creación de una nueva columna llamada "HospitalOccupancy", la cual refleja el número de pacientes que estuvieron presentes simultáneamente en un hospital determinado durante un día específico. Esto se calculó considerando el rango de días de hospitalización de cada paciente, generando una fila por cada día de estancia. Luego, los datos se agruparon por hospital y día para sumar el total de pacientes presentes diariamente. Para evitar duplicados por facturas con múltiples servicios en el mismo día, se consolidaron en un solo registro. Finalmente, se unieron los datos hospitalarios con la ocupación diaria, eliminando las columnas intermedias utilizadas en el cálculo.

Se realizó una limpieza de datos eliminando columnas sin valor directo para el análisis, incluidas todas las que contenían "Codigo", ya que estas ya habían sido decodificadas. También se eliminaron variables como 'DestinoUsuario', 'EstadoSalida', 'CausaBasicaMuerteUrgencias', y las fechas y horas de ingreso y salida. Estas columnas fueron consideradas irrelevantes para la predicción, ya que algunas solo se conocen después del evento y otras se usaron únicamente para calcular "TiempoenObservación".

4.6. Análisis exploratorio de los datos

Distribución de la Variable Objetivo: Tiempo en la unidad de observación

La variable objetivo, TiempoenObservación tiene un promedio de 25.53 horas, con una mediana de 5.15 horas y una desviación estándar de 101.96 horas, lo que indica una distribución altamente sesgada hacia valores bajos. Los valores extremos oscilan entre un mínimo de 1 minuto (0.0167 horas) y un máximo de aproximadamente 664 días (15,938 horas). Se puede observar en la figura 2 que la mayoría de los pacientes permanecen menos de 5 horas, mientras que un menor porcentaje supera las 50 horas. Además, se observó que el 19.36% de los registros corresponden a pacientes cuyo tiempo en observación excedió las 24 horas.

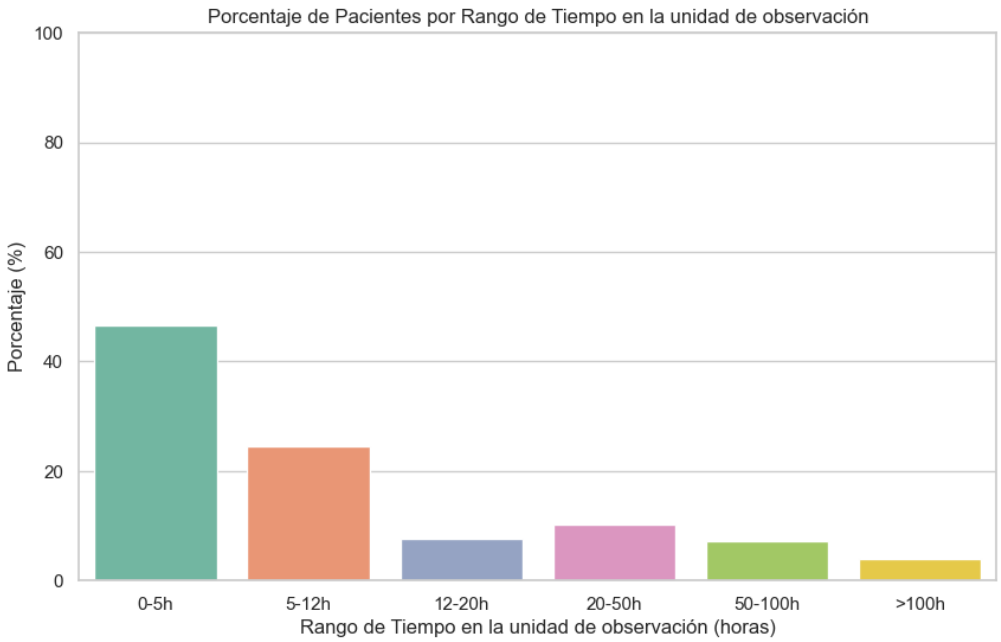


Figura 2: Porcentaje de pacientes por rango de tiempo en la unidad de observación

Relación entre Tiempo en la unidad de observación y Rango de Edad

Como se observa en la figura 3, el tiempo promedio de estancia en la unidad de observación de urgencias muestra una tendencia creciente con la edad, alcanzando su punto máximo en la categoría de Gran Vejez (85+ años). Esto se podría asociar con la complejidad médica y la mayor fragilidad de esta población, lo que a menudo resulta en tiempos prolongados de atención. Esta observación resalta la necesidad de analizar cómo las enfermedades más frecuentes en esta población afectan su tiempo de permanencia, lo que será explorado en la sección del impacto de las variables.

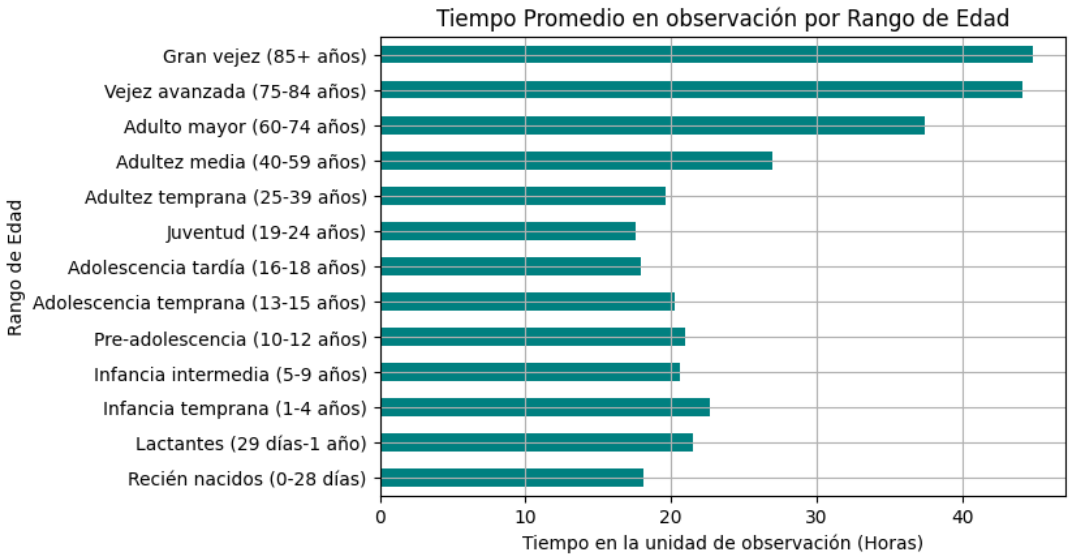


Figura 3: Tiempo promedio en la unidad de observación por rango de edad.

En la población de gran vejez, como se detalla en la figura 4 las enfermedades crónicas del tracto respiratorio inferior, enfermedades cardiacas y las enfermedades del aparato urinario son las causas más comunes de consulta. Estas patologías no solo son frecuentes, sino que también implican un manejo clínico más complejo, lo que explica en parte los tiempos prolongados observados. Este hallazgo es consistente con el perfil epidemiológico de esta población, donde las enfermedades crónicas y las complicaciones asociadas al envejecimiento son predominantes.

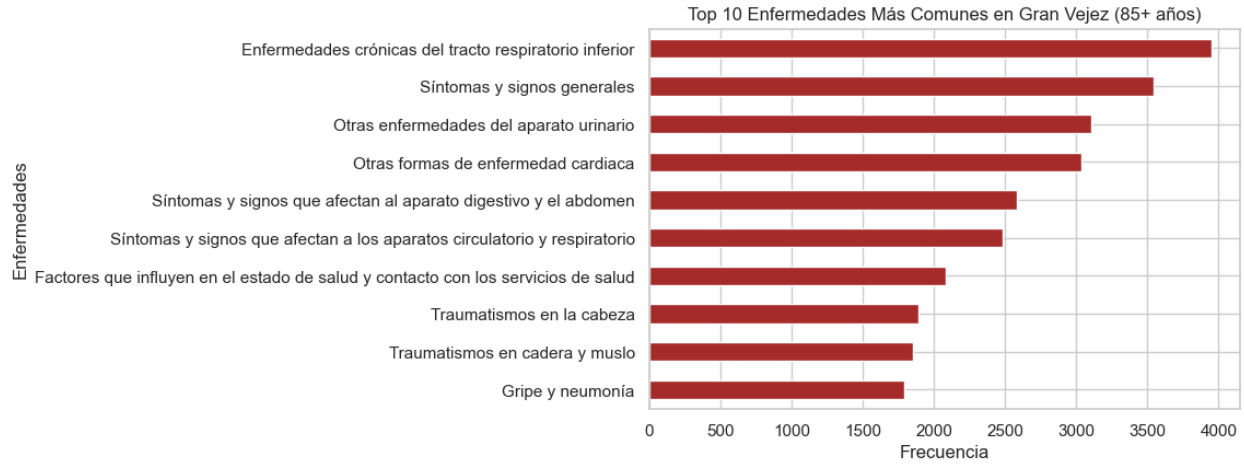


Figura 4: Top 10 enfermedades más comunes para el grupo etario Gran vejez

Análisis por Causas Externas, Edad y Género

Respecto a el análisis por causas externas revela diferencias significativas tanto por rango de edad como por género. En la categoría de Enfermedad General, las mujeres tienen una mayor proporción de casos en todos los grupos etarios, especialmente a partir de los 40 años, donde se observa un incremento notable. Por otro lado, los Accidentes de Tránsito y las Lesiones por Agresión son más prevalentes en hombres jóvenes y adultos, destacando patrones de riesgo diferenciales entre géneros. Además, en el caso de Sospecha de Abuso y Violencia Sexual, las mujeres representan la mayoría de los registros, especialmente en las edades más tempranas. En el rango de 0-18 años, los casos de Sospecha de Violencia Sexual y Abuso Sexual en mujeres son significativamente superiores a los observados en hombres, lo que evidencia la vulnerabilidad de este grupo frente a estas situaciones. En la población de Gran Vejez (85+ años), las Enfermedades Generales continúan siendo la causa predominante, con una distribución significativamente mayor en mujeres en comparación con hombres.

Análisis por Causa Externa y Tipo de Usuario

En el siguiente grafico se muestra un análisis del tiempo promedio en la unidad de observación por Causa Externa y Tipo de Usuario:

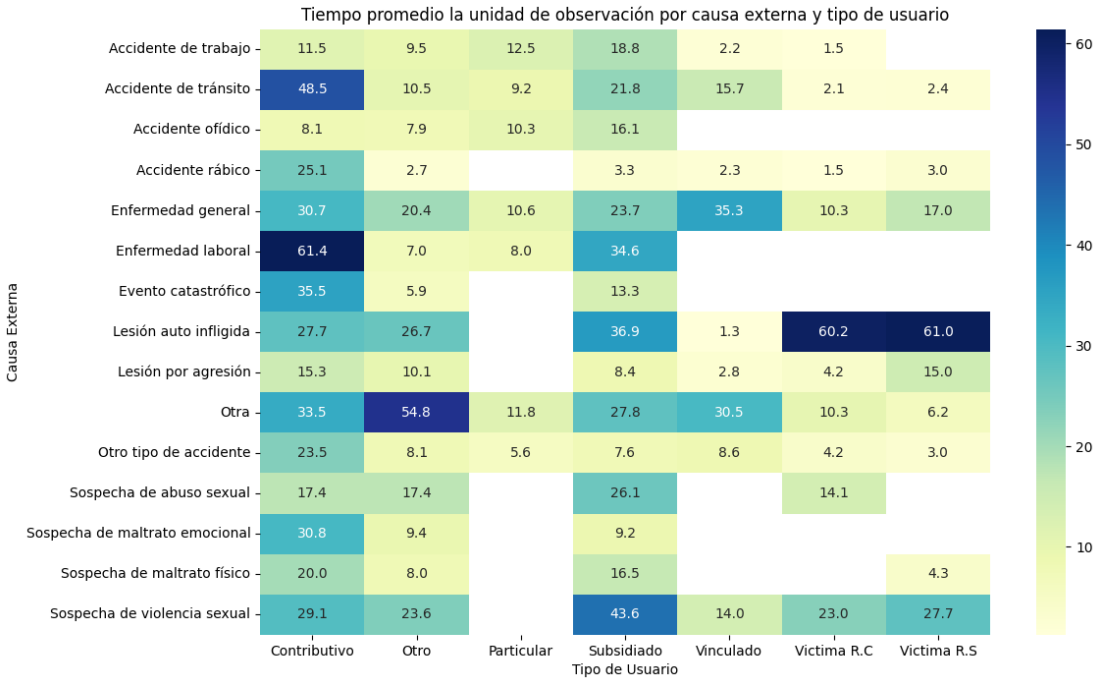


Figura 5: Tiempo promedio la unidad de observación por causa externa y tipo de usuario

Como se observa en la figura 5, en el caso de Lesión Auto Infligida y pacientes clasificados como Víctimas con afiliación al Régimen Contributivo (Víctimas con afiliación al RC), se identifican tiempos promedio elevados de alrededor de 60 horas. Las patologías que más afectan este grupo son los Trastornos del Estado de Ánimo alcanzando un promedio de 135.68 horas, seguidos por Causas Externas de Morbilidad y Mortalidad con 72.28 horas. Este resultado refleja la complejidad del manejo de trastornos mentales y el tiempo requerido para estabilizar a estos pacientes en un entorno de urgencias.

Por otro lado, para la causa externa Otra y el tipo de usuario clasificado como Otro, las enfermedades relacionadas con el Aparato Circulatorio presentan los mayores tiempos promedio en observación, alcanzando 294.2 horas, seguidas por trastornos como Trastornos Sistémicos del Tejido Conectivo (216.66 horas) y Enfermedades Isquémicas Cardiacas (204 horas). Este grupo incluye patologías de alta complejidad clínica, como Enfermedades Cerebrovasculares y Neoplasias Malignas, que requieren un manejo prolongado y especializado.

Dinamismo y Desempeño de las IPS

Se describen algunas de las IPS con mayor tiempo promedio de estancia en la unidad de observación de urgencias, y con mayor ocupación promedio.

EMPRESA SOCIAL DEL ESTADO METROSALUD - Unidad de Salud Mental

Esta unidad se especializa en la atención de trastornos mentales y del comportamiento, ofreciendo servicios como hospitalización psiquiátrica, consultas especializadas en psiquiatría y psicología, y terapia ocupacional. Las enfermedades más comunes atendidas incluyen trastornos del estado de

ánimo, esquizofrenia y trastornos de ansiedad. La naturaleza compleja de estas patologías y la necesidad de evaluaciones exhaustivas contribuyen a estancias más prolongadas en la unidad de observación. Como se observa en la figura 6 en promedio un paciente en esta IPS pues estar más de 400 horas, lo que es alrededor de 16 días utilizando una cama.

PROMOTORA MÉDICA LAS AMÉRICAS S.A - Clínica Las Américas

La Clínica Las Américas es un centro hospitalario de alta y mediana complejidad, reconocido como una de las instituciones líderes en América Latina en la prestación de servicios de salud especializados. Esta clínica tiene un promedio de estancia de 236 horas y una ocupación promedio de 177 pacientes. Además, registra un alto número de casos con 70,321 registros en total, atendiendo principalmente patologías como traumatismos y enfermedades cardiovasculares, lo que explica los tiempos prolongados y su elevada demanda.

HOSPITAL ALMA MÁTER DE ANTIOQUIA - Sede Principal

Este hospital universitario ofrece una amplia gama de especialidades médicas y quirúrgicas, incluyendo cirugía general, ginecológica, ortopédica y oftalmológica. Las enfermedades más comunes atendidas incluyen síntomas y signos que afectan al aparato digestivo, síntomas generales y otras infecciones agudas del tracto respiratorio. Registra un promedio de estancia de 165 horas, una ocupación promedio de 471 pacientes y un total de 65,577 casos, reflejando la alta presión sobre sus recursos y la complejidad de los tratamientos que ofrece.

VIRREY SOLIS I.P.S S.A - sede VIRREY SOLIS I.P.S S.A SAN DIEGO

Esta IPS tiene una ocupación promedio destacable de 1141 pacientes, lo que, sumado a un tiempo promedio de 72 horas, resalta la alta afluencia de pacientes que manejan. Este comportamiento también se observa en su sede Virrey Solís IPS. S.A Tranvía Plaza, con 71 horas promedio y 238 pacientes en ocupación, lo que evidencia la importancia de su papel en el sistema de atención en la región.

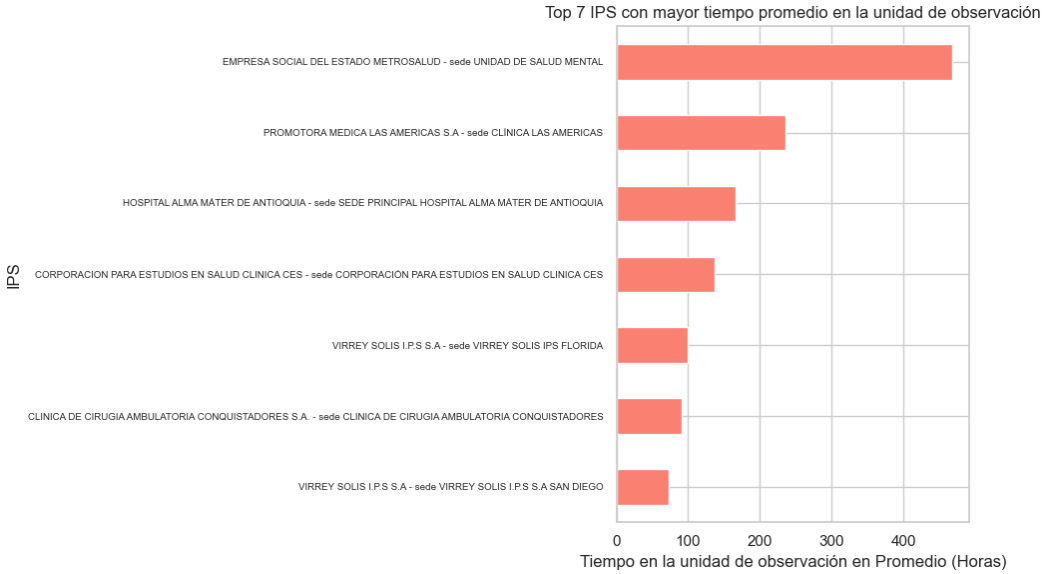


Figura 6: Top 7 IPS con mayor tiempo promedio en la unidad de observación

Análisis IPS publicas vs privadas:

El análisis de los datos por tipo de IPS revela diferencias significativas en el comportamiento de las instituciones privadas y públicas. El dataset tiene 55 IPS de las cuales 15 son de naturaleza pública y pertenecen a la red pública de Metrosalud. Esta red, conformada por 9 unidades hospitalarias, 41 centros de salud y otras 3 sedes, está distribuida estratégicamente en todas las comunas y corregimientos de Medellín, lo que permite un acceso amplio a los servicios de salud en la ciudad (Secretaria de salud de Medellín, 2018). En promedio, los tiempos de estancia en observación son notablemente mayores en las IPS públicas (37.28 horas) en comparación con las privadas (19.98 horas). Esto puede reflejar diferencias en la complejidad de los casos atendidos, la disponibilidad de recursos o la capacidad de respuesta en cada tipo de institución.

Por otro lado, las IPS privadas manejan un promedio de registros por institución más alto (14,220 registros) en comparación con las públicas (9,259 registros). Además, el volumen total de registros muestra una ligera preponderancia hacia las IPS privadas, con 782,116 registros en comparación con 509,291 en las públicas. Esto sugiere una mayor fragmentación en los registros públicos, lo que podría indicar diferencias en la capacidad o cobertura entre estas instituciones.

En cuanto a la ocupación promedio, tanto las IPS públicas como las privadas presentan valores similares, con un promedio de 208 camas ocupadas en las públicas y 199 en las privadas. Sin embargo, al analizar la máxima ocupación registrada, las privadas destacan significativamente con un pico de 1,832 camas ocupadas, frente a 682 camas en las públicas. Este dato podría estar relacionado con la capacidad instalada de las instituciones privadas y su posibilidad de manejar mayor flujo de pacientes en momentos de alta demanda.

Distribución de la Ocupación por Día, Mes y Hora

Como se observa en la figura 7, los días de la semana, las diferencias son mínimas, lo que sugiere que la ocupación se mantiene relativamente constante de lunes a domingo. Sin embargo, se observa un ligero descenso durante los fines de semana en comparación con los días entre semana.

En cuanto a los meses, marzo y abril destacan como los periodos con mayor ocupación promedio. Esto puede estar relacionado con un aumento en los casos de enfermedades respiratorias, como la tuberculosis, que se incrementan en estas fechas debido a condiciones climáticas que favorecen la transmisión de este tipo de patologías. Además, es importante resaltar que estos dos meses son también los periodos con mayor cantidad de registros para la IPS Virrey Solís San Diego, una de las instituciones con tiempos de estancia más prolongados en la unidad de observación. Esto podría contribuir significativamente a la alta ocupación observada en estos meses.

Respecto a la distribución por horas del día, se evidencia un pico notable en la ocupación a las 2:00 p.m de alrededor de 700 pacientes. Este no se muestra en la gráfica para poder ver el

comportamiento de las demás horas de una manera más precisa. El aumento a las 2 de la tarde no se debe a un aumento real en la cantidad de pacientes, sino a un fenómeno particular: varias IPS, como Virrey Solís Tranvía Plaza, Virrey Solís San Diego y Virrey Solís Florida, tienden a registrar sus ingresos en la unidad de observación en este horario, lo que genera un aparente aumento en la ocupación.

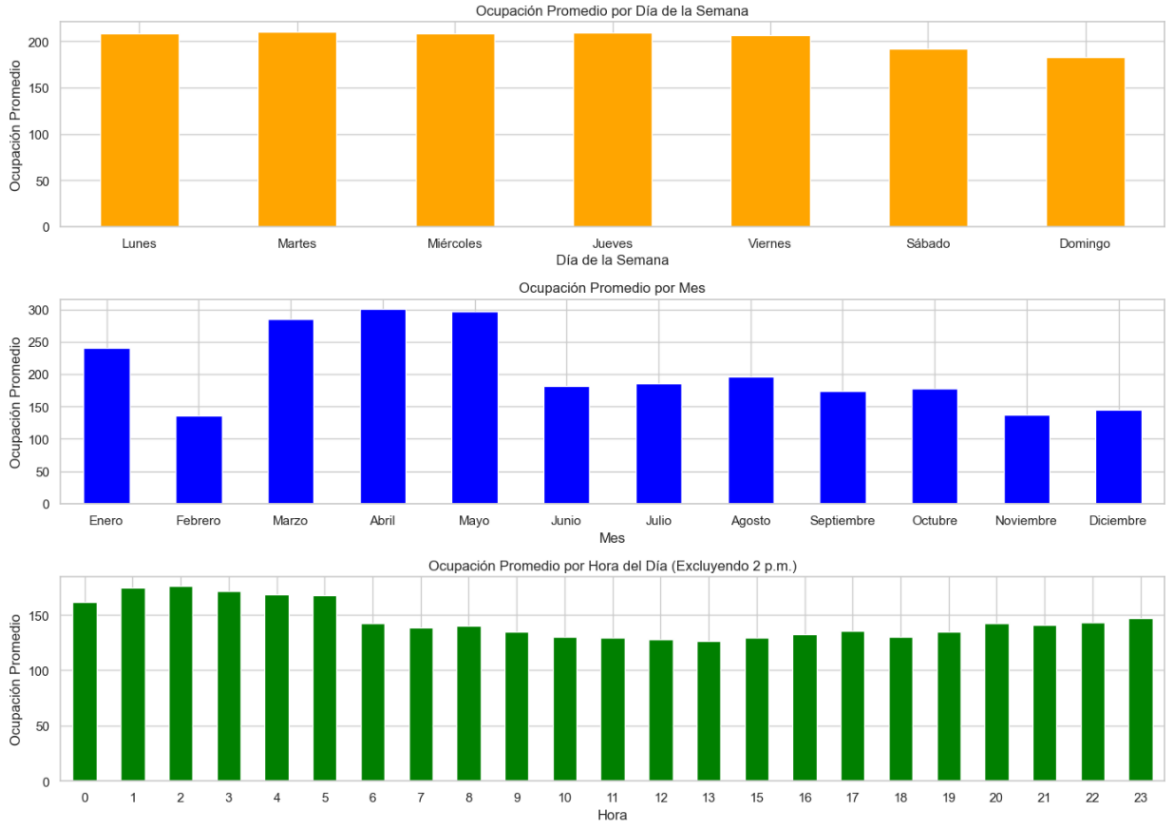


Figura 7: Ocupación promedio de todas las IPS por dia de la semana, mes y hora

Análisis del Tiempo Promedio en observación según el Destino de los Pacientes

El análisis del tiempo promedio en la unidad de observación según el destino de los pacientes revela un comportamiento coherente con las dinámicas operativas de los servicios de urgencias. Los pacientes dados de alta tienen un tiempo significativamente menor, lo cual es lógico, ya que sus casos suelen ser menos graves y requieren menos tiempo de estabilización. Por otro lado, los tiempos promedio similares entre hospitalización y remisión a otro nivel de complejidad, que rondan las 60 horas, pueden explicarse por la naturaleza crítica de estos casos, donde los pacientes suelen necesitar un manejo prolongado para estabilizarlos antes de su traslado o ingreso. Además, estos procesos están sujetos a trámites administrativos y logísticos. En términos de frecuencia, el 78.77% de los pacientes son dados de, mientras que un 20.57% requiere hospitalización. Solo el 0.65% de los casos son remitidos a otro nivel de complejidad, lo que indica que la mayoría de los pacientes reciben atención resolutiva en esta unidad de urgencias.

5. Investigación Experimental

5.1 Metrica Evaluación de los modelos y Partición de los datos.

Para evaluar los modelos predictivos, se utilizó la métrica MAE (Mean Absolute Error) junto con una validación cruzada de 5-folds (véase Apéndice A para definiciones detalladas). El MAE, al medir el error absoluto promedio, ofrece una interpretación directa de la precisión del modelo y es menos sensible a grandes discrepancias entre los valores predichos y observados. La elección de 5-folds balancea el trade-off entre bias y varianza. Según *An Introduction to Statistical Learning* (2013), entre 5 y 10 pliegues proporcionan un equilibrio óptimo, permitiendo estimaciones confiables del error sin alta varianza ni bias considerable (p. 184). Además, K-fold cross-validation es computacionalmente eficiente comparado con LOOCV, ya que solo requiere ajustar el modelo cinco veces, mejorando la representatividad y estabilidad de las estimaciones de error (Shulga, 2018).

5.3 Modelos machine learning y calibración hiperparametros.

Para preparar los datos categóricos para los modelos de machine learning, fue necesario analizar cuántas categorías únicas había en cada una de las variables categóricas. Este análisis es crucial para entender la magnitud del espacio de características una vez que estas variables sean transformadas mediante one-hot encoding (véase Apéndice A para definiciones detalladas). El análisis mostró que al convertir todas las variables categóricas remanentes a dummies, quedarían en total 1855 columnas, siendo la variable de municipio la que más dummies generaría. Teniendo en cuenta que el dataframe final tiene 1,291,374 registros, este proceso haría el tratamiento de los datos computacionalmente inmanejable. Por este motivo, se optó por utilizar algoritmos de machine learning capaces de manejar eficazmente grandes conjuntos de datos con múltiples categorías sin la necesidad de expandir excesivamente la matriz de características mediante técnicas como one-hot encoding. Además, se seleccionaron algoritmos que abordan de manera eficiente dos problemas comunes en variables categóricas: primero, el impacto desproporcionado que pueden tener las categorías con pocos registros, pero valores extremos en el modelo, y segundo, la capacidad del modelo para realizar predicciones precisas incluso en categorías con datos limitados, evitando que estas sean subrepresentadas. A continuación, se describen la calibración de hiperparámetros para los cuatro modelos desarrollados para la predicción del tiempo de estancia, cada uno seleccionado por su capacidad para trabajar con datos categóricos de manera eficiente y robusta.

Algoritmo XGBoost: El primer modelo implementado fue XGBoost (véase Apéndice A para una definición detallada y explicación de como maneja variables categoricas). Se seleccionaron cuidadosamente los hiperparámetros para mitigar el sobreajuste, como una tasa de aprendizaje baja (learning_rate=0.005), una profundidad reducida de los árboles (max_depth=2), y una regularización más agresiva tanto en términos de L2 (lambda=20) como de L1 (alpha=20).

Algoritmo CatBoost: A continuación, se probó el modelo CatBoost (véase Apéndice A para una definición detallada y explicación de como maneja variables categoricas). Se utilizaron hiperparámetros como una profundidad de 6 capas y una tasa de aprendizaje de 0.05, en conjunto con una regularización L2 para evitar el sobreajuste. En cada pliegue, se aplicó una técnica de *capping* a la variable objetivo para manejar los valores atípicos, lo que ayudó a estabilizar las predicciones.

Algoritmo LightGBM: El tercer modelo que se desarrolló fue LightGBM (véase Apéndice A para una definición detallada y explicación de como maneja variables categoricas). Para optimizar el rendimiento, se llevó a cabo una búsqueda de hiperparámetros utilizando *Grid Search* con validación cruzada, lo que permitió identificar los mejores ajustes para el problema, como un número de hojas de 50 una profundidad máxima de 7 y una regularización L2 con un valor de 10.

Red neuronal densa con capas de embedding: Finalmente, se exploró una red neuronal densa con capas de embeddings (véase Apéndice A para una definición detallada y explicación de como maneja variables categoricas) Dado que muchas de las variables categóricas tenían miles de valores únicos (por ejemplo, más de 600 municipios), se decidió usar capas de *embedding* para convertir estas variables en representaciones de menor dimensión, lo que permite al modelo capturar relaciones y patrones de manera más eficiente. Además, se incluyeron técnicas de regularización como L2 y *dropout* en las capas densas para evitar el sobreajuste. El modelo fue entrenado usando *early stopping* para garantizar que no se entrenara más allá del punto óptimo, lo que ayudó a prevenir el sobreajuste, especialmente dado el tamaño del conjunto de datos.

5.4 Resultados.

5.4.1 Resultados de los modelos.

El mejor modelo obtenido con todas las variables fue Catboost con un Mae de 8.7851 en entrenamiento y 8.8320 horas en validación. El procedimiento para el modelo CatBoost comenzó analizando el impacto de las variables en la predicción del tiempo de estancia de los pacientes, utilizando el modelo SHAP (véase Apéndice A para definiciones detalladas) para identificar la relevancia de cada una de ellas. En la figura 8 se muestra la importancia de las diferentes características utilizadas en el modelo de CatBoost para predecir el tiempo de estancia. De acuerdo con este gráfico, la variable 'NombrePrestadorSede' tiene el mayor impacto, con un incremento de hasta 8 horas en promedio en el tiempo de estancia estimado, lo que sugiere que el centro de atención puede influir significativamente en la duración de la estancia de los pacientes. La variable 'HospitalOccupancy' también tiene un impacto considerable, con un efecto promedio de alrededor de 6 horas. Por el contrario, variables como 'Sexo', 'ZonaResidencia', 'DescripcionDiagnosticoRelacionado_tres' y 'NombreMunicipio' tienen impactos mínimos, con efectos menores a 0.5 horas. Esto sugiere que su contribución al modelo es marginal, y podrían omitirse para optimizar el proceso de recolección de datos sin afectar la precisión en la predicción.

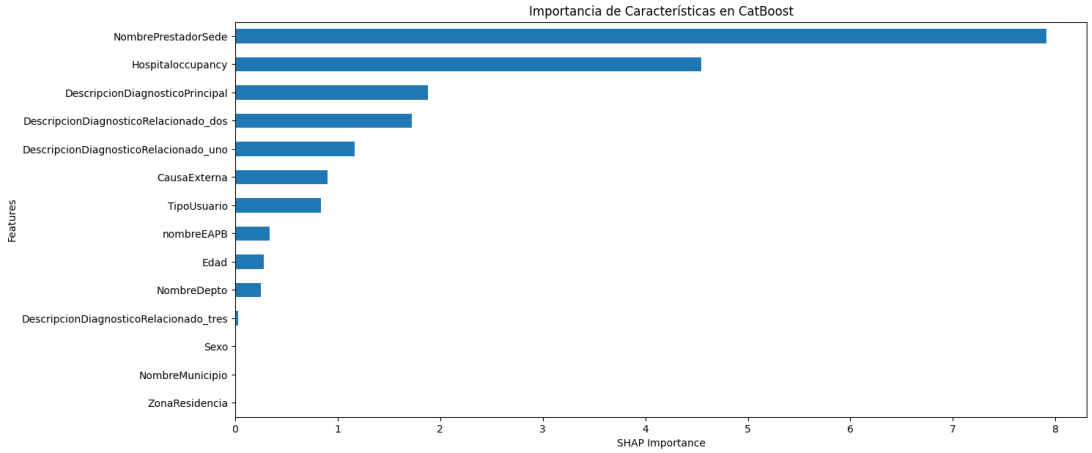


Figura 8: Importancia promedio de las variables predictoras en el modelo Catboost con todas las variables.

A pesar de que los algoritmos de boosting ya cuentan con mecanismos de regularización que penalizan automáticamente las variables menos relevantes, se decidió realizar un proceso adicional de selección de variables. Se estableció un umbral de 0.5 en el MAE (aproximadamente 30 minutos), Por lo tanto, si al quitar una variable el modelo empeoraba más de eso, ya no se consideraba como opción quitar esas variables y se volvía al modelo que se tenía. Se estableció este umbral en específico dado que este es el tiempo mínimo de estancia que se había establecido. De esta forma, se realizaron pruebas eliminando algunas de las variables menos importantes para observar su efecto en el rendimiento del modelo de CatBoost. A continuación, se muestran los resultados:

Modificaciones en el Modelo de CatBoost	**MAE 5 fold cross validation Entrenamiento**	**MAE 5 fold cross validation Validación**
Sin DescripcionDiagnosticoRelacionado_tres, Sexo, ni ZonaResidencia	8.775	8.822
Sin DescripcionDiagnosticoRelacionado_tres, Sexo, Municipio, ni ZonaResidencia	8.802	8.841
Sin DescripcionDiagnosticoRelacionado_tres, Sexo, ni ZonaResidencia	8.780	8.827
Sin DescripcionDiagnosticoRelacionado_tres sexo, municipio, ni departamento.	8.802	8.835
Sin DescripcionDiagnosticoRelacionado_tres sexo, municipio, departamento, ni ZonaResidencia	8.776	8.825
Sin DescripcionDiagnosticoRelacionado_tres sexo, municipio, departamento, ZonaResidencia, CausaExterna, nombreEAPB ni Edad.	8.863	8.900
Con todas las variables.	8.785	8.832
Sin DescripcionDiagnosticoRelacionado_tres sexo, municipio, departamento, ZonaResidencia, CausaExterna, nombreEAPB, Tipo de usuario ni edad.	8.875	8.915
Sin DescripcionDiagnosticoRelacionado_tres sexo, municipio, departamento, ZonaResidencia, CausaExterna, nombreEAPB, Tipo de usuario, DescripcionDiagnosticoRelacionado_uno, ni edad.	9.701	9.725

Tabla 1: MAE en entrenamiento y validación para cada una de las combinaciones de variables probadas

Por lo tanto, aunque el modelo de CatBoost que dio el menor MAE fue aquel sin las variables 'DescripcionDiagnosticoRelacionado_tres', 'Sexo' y 'ZonaResidencia', finalmente se seleccionó el modelo sin las variables 'DescripcionDiagnosticoRelacionado_tres', 'Sexo', 'NombreMunicipio', 'NombreDepto' , 'ZonaResidencia', CausaExterna, nombreEAPB, tipo de usuario ni Edad. A pesar de que esta configuración aumentó ligeramente el MAE, el cambio fue muy bajo y casi imperceptible, con una variación de 0.088 (equivalente a 5 minutos y 17 segundos) en la predicción. Esta pequeña pérdida en precisión se consideró aceptable, ya que, al reducir el número de variables, se optimiza la eficiencia del modelo y se mejora la experiencia del usuario al requerir menos información de entrada.

Aunque ya se sabía que con las variables completas el mejor modelo era Catboost, se quiso evaluar si quitando variables que no aportaran a los otros modelos se encontraba un modelo mejor. En la tabla siguiente se muestran las variables utilizadas en cada modelo, indicando las que fueron finalmente seleccionadas:

Variable	CatBoost	XGBoost	LightGBM	Red Neuronal
CausaExterna		✓	✓	✓
TipoUsuario		✓	✓	✓
Sexo		✓	✓	
ZonaResidencia		✓		
NombrePrestadorSede	✓	✓	✓	✓
nombreEAPB		✓	✓	✓
NombreDepto		✓		✓
NombreMunicipio		✓	✓	✓
DescripcionDiagnosticoPrincipal	✓	✓	✓	✓
DescripcionDiagnosticoRelacionado_uno	✓	✓	✓	✓
DescripcionDiagnosticoRelacionado_dos	✓	✓	✓	✓
DescripcionDiagnosticoRelacionado_tres		✓	✓	
Edad		✓	✓	✓
Hospitaloccupancy	✓	✓	✓	✓

Tabla 2: Variables usadas en los modelos que dieron el mejor MAE en validación

A continuación, se presenta una tabla que resume los resultados del Error Absoluto Medio (MAE) para cada uno de los modelos finales en los cinco folds de validación cruzada, así como el promedio y la desviación estándar para cada modelo. Esto permite comparar de forma clara y cuantitativa el rendimiento de cada método.

Fold / Métrica	Red Neuronal con Embeddings	LightGBM	CatBoost	XGBoost
Fold 1 (Entrenamiento)	17.261	19.785	8.874	11.481
Fold 1 (Validación)	17.219	20.160	8.887	20.514
Fold 2 (Entrenamiento)	16.990	19.764	8.887	11.478
Fold 2 (Validación)	17.340	20.309	8.948	20.712
Fold 3 (Entrenamiento)	17.088	19.723	8.870	11.481
Fold 3 (Validación)	17.128	20.087	8.885	20.512
Fold 4 (Entrenamiento)	17.136	19.800	8.861	11.464
Fold 4 (Validación)	17.104	20.280	8.934	20.449
Fold 5 (Entrenamiento)	17.075	19.756	8.885	11.481
Fold 5 (Validación)	17.353	20.330	8.921	20.711
MAE Promedio (Entrenamiento)	17.110	19.766	8.875	11.477
MAE Promedio (Validación)	17.229	20.233	8.915	20.580
Desviación Estándar (Validación)	0.099	0.084	0.026	0.102

Tabla 3: Resultados del MAE para los 4 modelos realizados

Los resultados de la tabla 3 muestran que el modelo CatBoost presenta el menor MAE promedio tanto en entrenamiento como en validación, con un MAE de 8.875 y 8.915 respectivamente. Esto indica un mejor ajuste y una capacidad de predicción más precisa en comparación con los otros modelos. Es importante mencionar que en todos los modelos la desviación estándar entre los folds es muy baja, lo que indica un rendimiento confiable y estable en cada uno de ellos.

Por lo tanto, el modelo seleccionado para realizar el análisis de la importancia de las variables y desarrollar la herramienta computacional fue Catboost con las variables mencionadas en la tabla 2.

5.4.2 Importancia de las variables.

En esta sección se utilizó la técnica de SHAP (*SHapley Additive exPlanations*) para interpretar la importancia de las variables en el modelo de CatBoost (véase Apéndice A para una definición detallada). Un valor SHAP positivo indica que la variable aumenta la predicción, mientras que un valor negativo indica que la disminuye. Por ejemplo, un valor SHAP de +2 para "HospitalOccupancy" sugiere que esta variable incrementa en promedio 2 horas el tiempo de estancia.

El impacto de cada característica sobre el tiempo de estancia en el modelo de CatBoost, medido a través de valores SHAP. La variable 'NombrePrestadorSede' tiene el mayor valor SHAP promedio (9.20), indicando que ciertos centros de atención pueden prolongar significativamente la estancia, con contribuciones individuales de hasta 60 horas en algunos casos. Le sigue 'HospitalOccupancy', con un valor SHAP promedio de 5.40, que puede tanto incrementar como reducir el tiempo de estancia, dependiendo de su saturación, con variaciones extremas de 60 horas. Después, sigue 'DescripcionDiagnosticoPrincipal' con un impacto promedio de 2.14, seguido por 'DescripcionDiagnosticoRelacionado_dos' (1.58) y

'DescripcionDiagnosticoRelacionado_uno' (1.36), que, aunque tienen un menor valor promedio, pueden generar incrementos de hasta 20 horas en el tiempo de estancia.

Importancia de las Enfermedades por edad.

Para este análisis, se agruparon todas las enfermedades registradas en cada grupo etario y se calculó el promedio de sus impactos según los valores SHAP. Este agrupamiento incluye tanto las enfermedades comunes a múltiples grupos etarios como aquellas específicas de cada grupo. Este análisis revela que el impacto promedio de las enfermedades presentes en cada grupo etario sobre el tiempo de estancia incrementa a medida que aumenta la edad, alcanzando su punto máximo en el grupo de Gran Vejez (85+ años), con un valor de 1.12. Este resultado es consistente con lo observado en el análisis exploratorio, donde se destacó que las personas mayores en promedio enfrentan mayores tiempos de estancia. En particular, el grupo de Vejez Avanzada (75-84 años) también presenta un impacto elevado, con un valor SHAP de 0.94. Esto contrasta notablemente con los valores observados en grupos más jóvenes, como los Recién nacidos (0-28 días) o los Lactantes (29 días-1 año), donde el impacto promedio es negativo (-0.23 y -0.19, respectivamente), reflejando tiempos de estancia considerablemente menores.

A continuación, se presenta una gráfica que muestra las combinaciones de enfermedades con mayor impacto en el grupo de Gran Vejez:

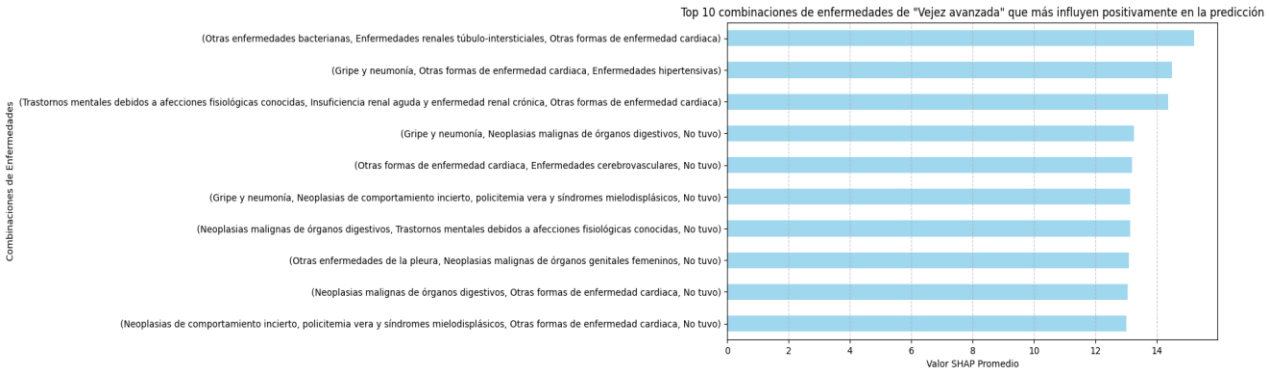


Figura 9: Top 10 de enfermedades del grupo etario Vejez avanzada que más impacto positivo tienen en la predicción

De acuerdo con la figura 9 entre estas, destacan combinaciones como "Otras enfermedades bacterianas", "Enfermedades renales túbulo-intersticiales" y "Enfermedades cardíacas", junto con condiciones como "Gripe y neumonía" y "Neoplasias malignas de órganos digestivos". Estas enfermedades no solo son frecuentes en esta población, sino que también implican un manejo clínico complejo y prolongado, justificando el incremento del tiempo en observación para estos pacientes. Este hallazgo refuerza lo descrito en el análisis exploratorio, donde se señaló que las patologías crónicas y las complicaciones asociadas al envejecimiento son factores determinantes en los tiempos de estancia.

Impacto de las enfermedades agrupadas por Causas Externas y por Género

Para este análisis, se agruparon las enfermedades registradas por causas externas para cada género y se calculó su impacto según los valores SHAP, considerando tanto las enfermedades compartidas entre causas y géneros como las específicas de cada uno. En las mujeres, las categorías con las enfermedades de mayor impacto incluyen Sospecha de Abuso y Violencia Sexual, lo que refuerza su vulnerabilidad en las edades más tempranas. Este resultado complementa el comportamiento identificado en el rango de 0-18 años, donde estas causas externas son significativamente más frecuentes en mujeres. De este análisis, se concluye además que las enfermedades asociadas a estas causas tienen un mayor impacto en el tiempo en observación para mujeres que para hombres.

El análisis también revela nuevas observaciones. En las mujeres, las enfermedades en causas como Sospecha de maltrato físico y Lesión auto infligida tienen un impacto promedio elevado, posiblemente debido a la necesidad de un manejo integral que incluya evaluaciones psicológicas y médicas. Para los hombres, las enfermedades de categorías como Accidente de trabajo, Accidente ofídico y Accidente de tránsito tienen impactos promedios negativos, lo que podría reflejar procesos específicos en urgencias que permiten resolver estos casos con mayor eficiencia. Esto puede estar relacionado con las observaciones del análisis exploratorio, donde estas causas son más prevalentes en hombres jóvenes y adultos, sugiriendo la existencia de protocolos específicos que optimizan el tratamiento de estas enfermedades en hombres.

Impacto de las Enfermedades y el Desempeño de las IPS

En este análisis, las enfermedades se agruparon según las IPS donde fueron atendidas, considerando tanto las patologías compartidas entre instituciones como aquellas específicas de cada una. Además, se evaluó el impacto SHAP de las IPS en sí mismas. Esto permite identificar no solo el papel de las enfermedades en los tiempos de atención, sino también cómo los factores institucionales contribuyen a la variabilidad observada.

Tal como se muestra en la figura 10 y en la figura 11, la EMPRESA SOCIAL DEL ESTADO METROSALUD - sede UNIDAD DE SALUD MENTAL encabeza tanto el top de análisis de impacto SHAP de las enfermedades como el de impacto SHAP general de la IPS. Esto evidencia que su enfoque en el manejo de trastornos mentales como esquizofrenia y ansiedad contribuye significativamente a los tiempos de estancia prolongados. La naturaleza compleja de estas patologías y las evaluaciones exhaustivas que requieren justifican estos resultados, confirmando que los tiempos extendidos están directamente relacionados con la gravedad y complejidad de los casos atendidos en esta institución.

Por otro lado, la VIRREY SOLIS I.P.S S.A y sus diferentes sedes (Tranvía Plaza, Florida, y San Diego) destacan por tener un impacto SHAP significativo en ambos análisis. Esto confirma su relevancia en la atención de urgencias, con una alta afluencia de pacientes y una gran diversidad de casos clínicos que explican sus resultados. Este hallazgo complementa las observaciones sobre su papel crucial dentro del sistema hospitalario de la región, evidenciando su importancia para la cobertura y resolución de urgencias.

Finalmente, el HOSPITAL ALMA MÁTER DE ANTIOQUIA - sede SEDE PRINCIPAL se posiciona con un impacto SHAP promedio de 5.74 en enfermedades y un impacto SHAP general alto como institución. Este resultado refuerza su posición como un hospital de referencia en la región. Su atención a casos graves relacionados con síntomas digestivos y respiratorios, sumada a su alta ocupación promedio, pone de manifiesto la complejidad de los pacientes atendidos y la presión constante sobre sus recursos hospitalarios.

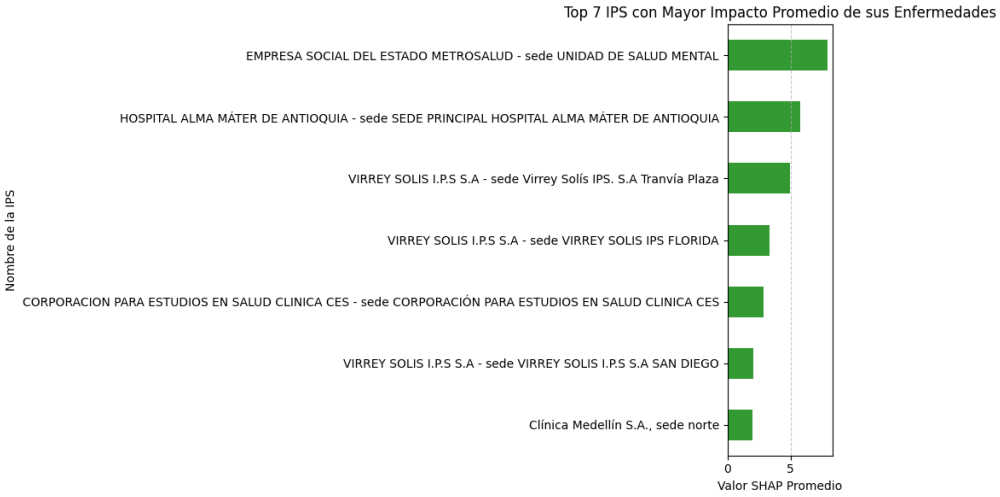


Figura 10: Top 7 IPS que mayor impacto positivo promedio tienen sus enfermedades en el modelo

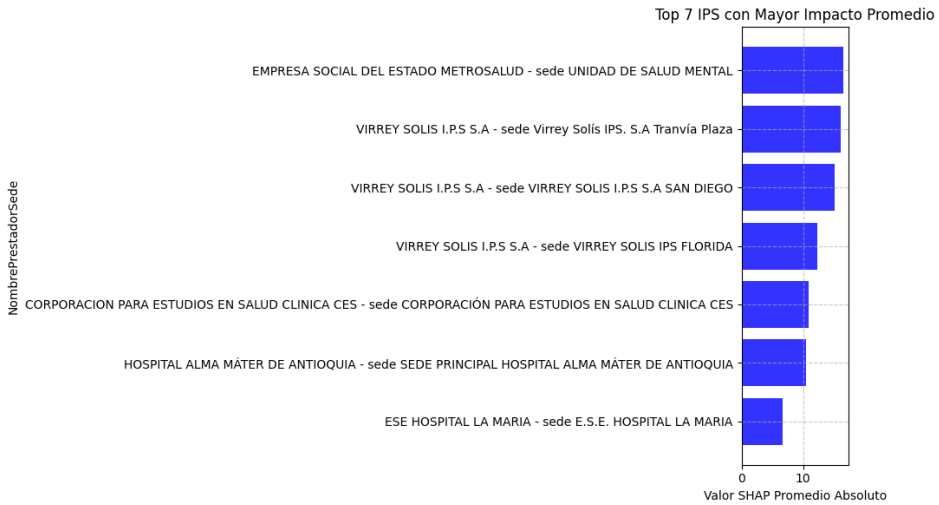


Figura 11: Top 7 IPS que mayor impacto positivo promedio tienen en el modelo

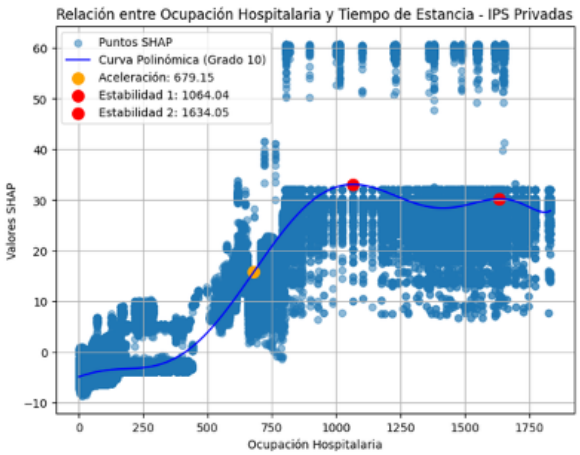
Análisis del impacto de las IPS Públicas y Privadas

Este análisis revela diferencias clave en la dependencia de la ocupación hospitalaria entre IPS públicas y privadas. Primero se realizó un análisis del punto de aceleración y saturación tanto para las ips públicas como para las privadas. El análisis de los puntos de aceleración y

estabilidad en la ocupación hospitalaria de IPS públicas y privadas es crucial para entender cómo estos dos tipos de instituciones manejan su capacidad operativa. Conocer el punto de aceleración permite entender en qué nivel de ocupación los tiempos de espera comienzan a crecer exponencialmente, permitiendo diseñar estrategias preventivas para evitar colapsos en el sistema. Asimismo, el punto de estabilidad revela la capacidad máxima funcional antes de que los efectos de la sobrecarga hospitalaria se estabilicen, lo cual puede indicar un punto de saturación en el sistema

Se analizó la relación entre la ocupación hospitalaria y los valores SHAP de la variable HospitalOccupancy para identificar los puntos de aceleración y estabilidad. Se ajustaron polinomios de distintos grados a estos valores, seleccionando dinámicamente el grado óptimo que mejor equilibrara el ajuste, evitando sobreajuste y sesgo, según el código proporcionado. La primera derivada del polinomio mide la tasa de cambio de los valores SHAP respecto a la ocupación. El punto de aceleración se identifica donde esta derivada alcanza su máximo, indicando el nivel de ocupación a partir del cual el impacto en los tiempos de espera aumenta rápidamente. El punto de estabilidad se determina como el punto o los puntos donde la derivada más se aproxima a cero o se vuelve directamente 0, indicando que el crecimiento del impacto se estabiliza.

Según la figura 12, el punto de aceleración se encuentra en 679 camas para las IPS privadas y en 587 camas para las públicas, mientras los puntos de estabilidad son 1,064 y 1,634 camas en las privadas, y 653 camas en las públicas, indicando mayor capacidad de atención antes de llegar al punto de saturación en las IPS privadas. Se observa que el shap disminuye luego de la de nivelación, este comportamiento explicarse por los protocolos específicos que implementan las IPS al llegar a un nivel crítico de saturación, como movilizar recursos adicionales desde otras sedes. Por otro lado, las IPS públicas tienen un SHAP promedio absoluto de -1.19 y las privadas de 1.82. Aunque las enfermedades en públicas tienen mayor impacto promedio (0.15 vs. 0.09) y tienden a ser más graves, las privadas presentan un mayor impacto general, posiblemente por factores administrativos o mayor demanda. Esto refleja que las públicas enfrentan limitaciones en recursos o procesos, afectando su eficiencia.



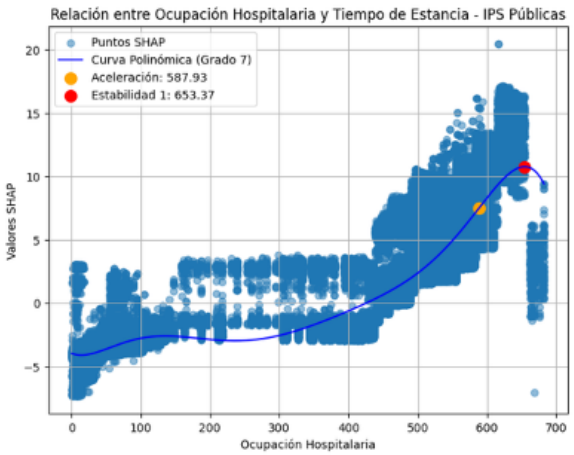


Figura 12: Relación entre Ocupación Hospitalaria y Valores SHAP en IPS Públicas y Privadas, con el punto de saturación y de aceleración

Impacto promedio de las enfermedades por mes

En este análisis, se evaluó el tiempo promedio de uso de camas hospitalarias en la unidad de observación de urgencias considerando exclusivamente el impacto promedio de las enfermedades presentes en cada mes. Solo se incluyeron enfermedades con registros durante el mes específico para garantizar un análisis representativo.

En marzo y abril, meses de mayor ocupación, se observaron impactos positivos significativos atribuidos a las enfermedades (0.303 y 0.272, respectivamente), lo que respalda la hipótesis de que el aumento de enfermedades respiratorias, como la tuberculosis, durante estos meses incrementa el tiempo promedio de hospitalización. En contraste, febrero y noviembre mostraron impactos negativos en las enfermedades (-0.048 y 0.006, respectivamente), reflejando una menor incidencia de patologías que requieren hospitalización prolongada. Por otro lado, en septiembre y octubre, el impacto promedio de las enfermedades fue positivo (0.166681 y 0.198416), lo que indica que el uso de camas hospitalarias en estos meses estuvo principalmente influenciado por la gravedad y complejidad de las enfermedades tratadas

6. Implementación

6.1 Flujo de la Página Web

La página web desarrollada tiene como objetivo permitir al auxiliar administrativo y a los coordinadores de recursos de la unidad de observación de urgencias gestionar y predecir el tiempo de estancia de los pacientes.

El recorrido de los usuarios dentro de la página sigue un flujo lógico, guiando al personal a través de secciones específicas según sus necesidades de uso. Tras iniciar sesión, el usuario se encuentra con un panel de navegación lateral, donde puede acceder a las funcionalidades principales del sistema. El proceso mas detallado se puede observar en la figura 13. A continuación, se muestran las principales funcionalidades de la página:

1. **Predicción del Tiempo de Estancia:** Permite ingresar información de un paciente (diagnósticos, ocupación actual, prestador de salud) y obtener una predicción del tiempo de estancia. Es especialmente útil para los auxiliares administrativos porque les ayuda a estimar cuánto tiempo permanecerá un paciente en urgencias, facilitando el registro y la gestión de camas. El resultado se muestra en horas y minutos, junto con interpretaciones visuales basadas en SHAP que destacan los factores más influyentes.
2. **Predicción del Tiempo de Estancia (Múltiple):** Permite cargar un archivo CSV con datos de varios pacientes. Es útil para gestionar grupos de pacientes simultáneamente, generando predicciones individuales, especificando un "Top N" de pacientes con mayor o

menor tiempo estimado, y ofreciendo estadísticas resumen (promedio, mínimo, máximo). Además, incluye gráficas interactivas para análisis conjunto, ayudando a optimizar procesos de ingreso y egreso.

3. **Estadísticas Pacientes:**

   Estadísticas de Enfermedades: Analiza el impacto de los diagnósticos en el tiempo de estancia mediante gráficos. Permite visualizar combinaciones de diagnósticos o uno específico, con opciones de análisis por mes, día, año, grupo etario, sexo, tipo de usuario y causa externa. Esto permite a los coordinadores de recursos identificar patrones relacionados con ciertas patologías, planificar recursos médicos y anticipar necesidades según la estacionalidad o características demográficas

   Estadísticas de Ocupación: Muestra el impacto de la ocupación hospitalaria en el tiempo de estancia, graficando los puntos donde la estancia comienza a incrementarse rápidamente (aceleración) y donde el impacto se estabiliza (saturación). Esta información es clave para los coordinadores en la planificación de recursos, optimizando la asignación de camas y personal en función de la demanda proyectada.

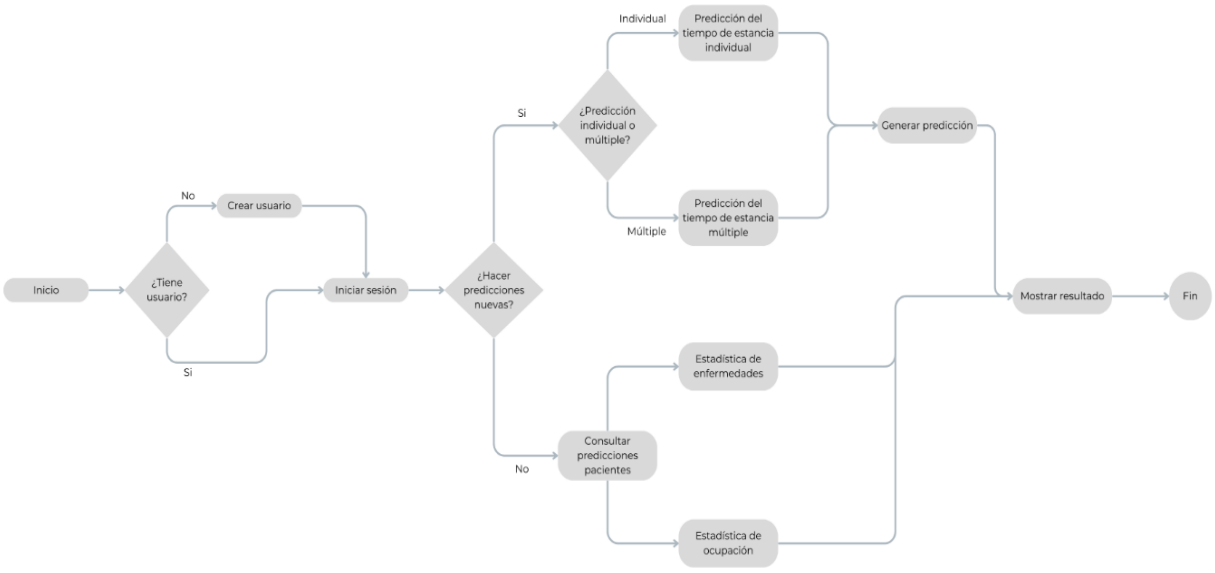


Figura 13: Diagrama de flujo de funcionamiento de la página web

6.2 Modelado de datos

A continuación, se muestra un diagrama Entidad-Relación (ER) con tres entidades principales: USERS, IPS, REGISTROSIPS. Este diagrama ilustra cómo se conectan las diferentes colecciones y campos de la base de datos para soportar las funcionalidades de "Estadísticas Pacientes”.

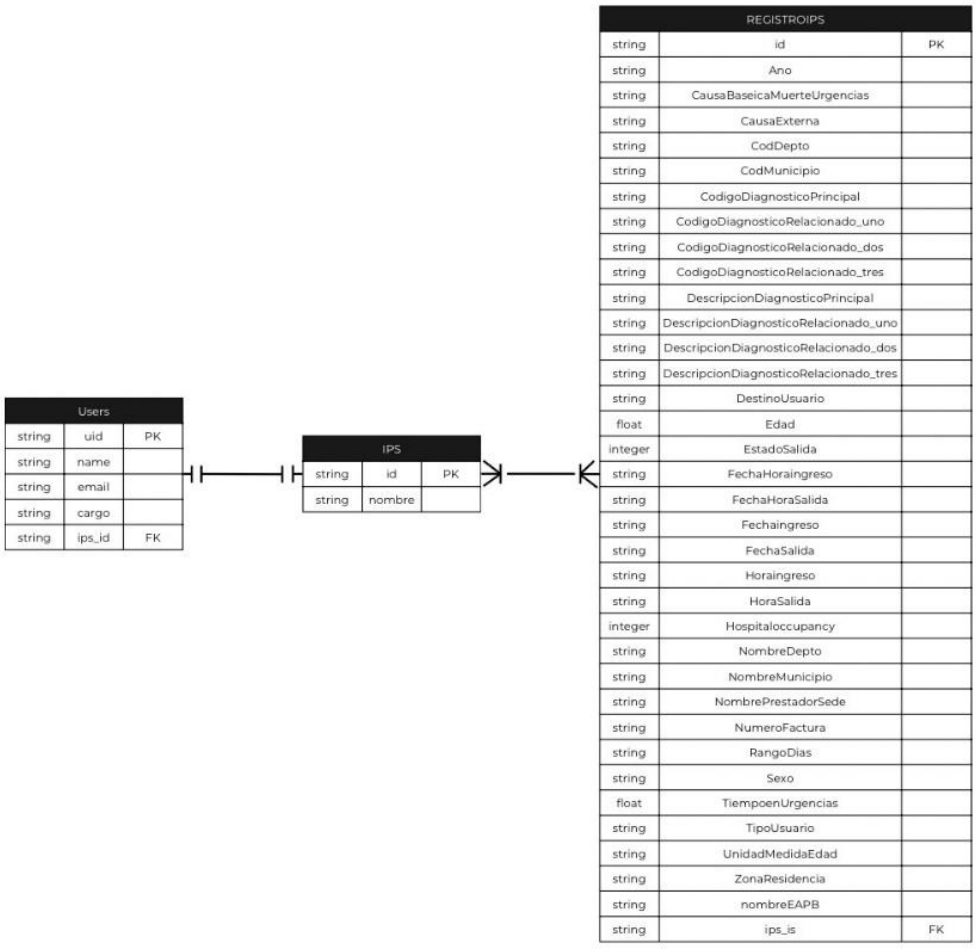


Figura 14: Diagrama Entidad relación de las bases de datos de la página web

Se observa en la figura 14 que la entidad USERS representa a los usuarios del sistema, identificados de manera única por el campo uid. Cada usuario tiene atributos como name, email y una referencia a la entidad IPS mediante el campo ips_id, estableciendo una relación de pertenencia donde muchos usuarios pueden estar asociados a una única IPS. Esta entidad permite gestionar múltiples IPS dentro de la colección REGISTROSIPS, facilitando la escalabilidad y organización de los datos.

La entidad REGISTROSIPS almacena registros detallados de pacientes, con numerosos campos que capturan datos clínicos y administrativos como diagnósticos, códigos, fechas, tiempos de estancia, y características demográficas. Esta entidad está relacionada con IPS a través del campo ips_id, que actúa como una clave foránea referenciando el id de la entidad IPS. Esto indica que cada registro en REGISTROSIPS está asociado a una IPS específica, permitiendo así la

trazabilidad y el análisis de los datos por entidad de salud. Por lo tanto, solo se puede acceder a esa base de datos si el usuario realmente pertenece a esa IPS.

En la funcionalidad de "Predicción del Tiempo de Estancia", se utilizan las bases de datos de "CodigosEnfermedades" para llenar las barras desplegables que permiten al usuario seleccionar y completar la información de los pacientes. Estas bases de datos proporcionan los códigos y descripciones necesarios para los diagnósticos principales y relacionados para poder llenar, mediante casillas desplegables, la información del paciente para la predicción. En el registro del usuario se usa la base de datos "IPS" para que el usuario le salga una casilla desplegable donde pueda buscar la IPS a la cual pertenece y hacer el correspondiente registro sin errores de typing. Por otro lado, en la funcionalidad de "Predicción del Tiempo de Estancia (Múltiple)", el CSV que se carga es de manera local.

6.3 Arquitectura del sistema

El sistema combina tecnologías distribuidas para front-end, back-end, y servicios de almacenamiento y autenticación. El front-end utiliza HTML, CSS y JavaScript, mientras que el back-end se desarrolla en Flask. La autenticación se gestiona con Firebase, y las bases de datos y el back-end están alojados en una instancia de AWS EC2. El front-end se despliega en AWS S3.

El flujo de colaboración del sistema comienza cuando el cliente solicita la página web, que está alojada en AWS S3. Este servicio responde cargando la interfaz de usuario, compuesta por archivos HTML, CSS y JavaScript, permitiendo que el usuario interactúe con el sistema. Primero, el usuario inicia sesión a través de Firebase para la autenticación y selección de la IPS correspondiente. Una vez autenticado y seleccionada la IPS, el usuario puede ingresar los datos necesarios para realizar una predicción. Estos datos se envían al backend, que está alojado en AWS EC2 utilizando Flask. En este entorno de Flask se carga el modelo serializado, que permite realizar las predicciones en base a los datos ingresados. Además, en Flask se implementa la lógica de generación de gráficos y análisis visuales que se muestran en cada funcionalidad del frontend. El servidor en EC2 se conecta con una base de datos alojada en la misma instancia de EC2 para obtener información adicional de la paciente almacenada. Esta base de datos responde con los datos necesarios, que el servidor utiliza para ejecutar el modelo de predicción. Este modelo devuelve un resultado al servidor en EC2, que finalmente envía el resultado de la predicción y los análisis generados al cliente, mostrando la información solicitada en la interfaz del usuario. A continuación, se muestra un diagrama de secuencia que ilustra las conexiones entre los distintos componentes.

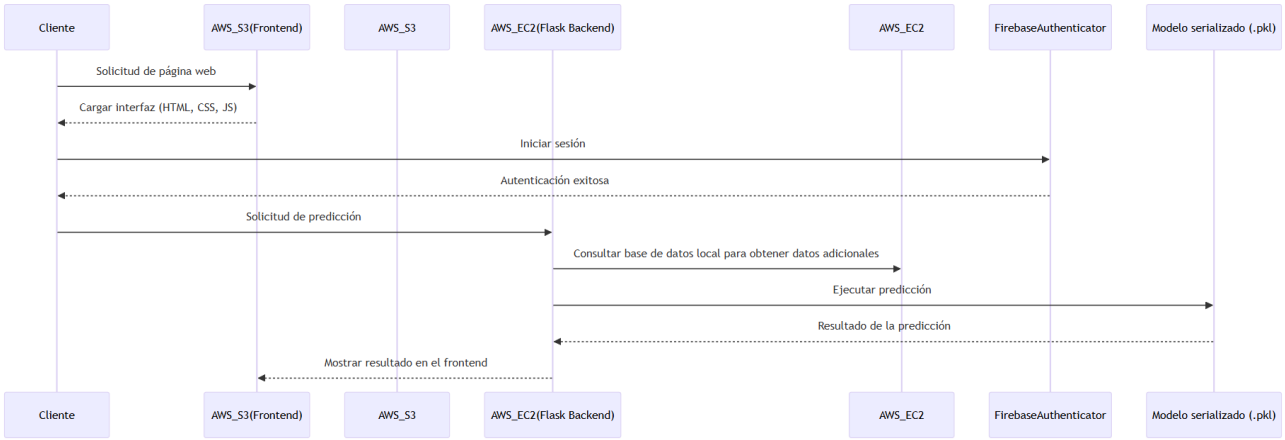


Figura 15: Diagrama de secuencia de colaboración de los componentes del sistema.

6.4 Seguridad y privacidad.

Para garantizar la privacidad y seguridad de los datos médicos de los pacientes, el sistema incorpora múltiples medidas de seguridad:

- **Autenticación de Usuarios**: Se utiliza Firebase Authentication para garantizar que solo los usuarios registrados puedan acceder al sistema. Cada usuario puede acceder y modificar únicamente su propia información, sin la posibilidad de editar los datos de otros usuarios.
- **Cifrado y Almacenamiento Seguro:** Amazon EC2 garantiza la seguridad de los datos mediante cifrado tanto en tránsito, mediante una red privada que cifra los datos entre servicios de AWS, como en reposo. Además, se han implementado grupos de seguridad y políticas de control de acceso que limitan la entrada a la información exclusivamente a los usuarios autorizados. Los grupos de seguridad de EC2 permiten definir reglas de tráfico de entrada y salida, controlando quién puede acceder a las instancias y desde qué ubicaciones
- **Seguridad en back end y front end**: AWS EC2 y S3 como infraestructura añade una capa adicional de seguridad. AWS implementa medidas avanzadas, como firewalls, monitoreo continuo y cifrado de datos en reposo y en tránsito. Estos mecanismos protegen el entorno donde está alojado el servidor y el front-end, asegurando que solo usuarios y servicios autorizados puedan interactuar con los recursos de la infraestructura.
- **Seguridad llaves de las bases de datos**: Las credenciales para acceder a Firebase y otros servicios sensibles se almacenan como variables de entorno en el servidor EC2. Esto garantiza que no se expongan públicamente en el código fuente, evitando que terceros no autorizados puedan acceder a las claves de acceso. Las variables de entorno proporcionan un método seguro de gestión de credenciales, contribuyendo a la integridad y confidencialidad del sistema.

7. Conclusiones y consideraciones finales

El desarrollo de un sistema predictivo basado en técnicas de machine learning para estimar el tiempo de estancia en la unidad de observación de urgencias ha demostrado ser una herramienta valiosa para mejorar la gestión de recursos en las IPS de Medellín. A través del análisis de datos provenientes de 55 IPS y la evaluación de cuatro modelos de machine learning, se identificó que CatBoost ofreció la mayor precisión, con un MAE promedio de 8.875 horas en los datos de validación. Además, La implementación del modelo en una herramienta web interactiva permite al personal administrativo ingresar datos del paciente y obtener predicciones rápidas y precisas, acompañadas de estadísticas sobre la importancia de cada variable en la predicción. La reducción del número de variables sin comprometer significativamente la precisión optimizó la eficiencia del modelo y mejoró la experiencia del usuario.

Para potenciar el impacto y la utilidad del sistema desarrollado, es recomendable su integración dentro de un Hospital Management System (HMS). Esta integración facilitaría el acceso a información actualizada y relevante, permitiendo que el modelo utilice datos en tiempo real y se beneficie de las funcionalidades existentes en los HMS, como la gestión integral de pacientes, recursos y procesos hospitalarios. Además, ofrecería un entorno más seguro y controlado para el manejo de datos sensibles, cumpliendo con las normativas de protección de datos personales. Al mismo tiempo, es crucial reconocer que los modelos de machine learning requieren actualizaciones periódicas para mantener su precisión y relevancia. El fenómeno de "drift" puede ocurrir debido a cambios en las variables, patrones de comportamiento de los pacientes, apertura o cierre de IPS, entre otros factores. Por lo tanto, se debe implementar un plan de monitoreo continuo del rendimiento del modelo y establecer procesos para su reentrenamiento con datos actualizados. Esto garantizará que el sistema se adapte a las dinámicas cambiantes del entorno hospitalario y mantenga su efectividad a largo plazo.

Una limitación significativa del proyecto es la representatividad de los datos, ya que se utilizaron registros de solo 55 IPS de Medellín. Esto limita la generalización de los resultados a todas las IPS de la ciudad o del país. Se recomienda ampliar el conjunto de datos incluyendo más instituciones y considerar variables adicionales que puedan influir en el tiempo de estancia, como recursos disponibles, protocolos internos y factores socioeconómicos.

Futuras investigaciones podrían explorar la incorporación de modelos multimodales que integren datos clínicos más detallados, resultados de laboratorio y otros indicadores relevantes. Asimismo, la colaboración con instituciones de salud para validar y ajustar el sistema en diferentes contextos contribuiría a mejorar su precisión y adaptabilidad.

8. Referencias

- ADRES (2022). *Entidades SGSSS.* Disponible en: https://www.adres.gov.co/entidades-territoriales/bdua/Entidades%20SGSSS/ENTIDADES_SGSSS_2022_ADRES.pdf
- Alcaldía de Medellín. (2023). *Registro Prestación Servicios Médicos en Urgencia con observación*. MEDATA. Disponible en: https://medata.gov.co/dataset/1-026-22-000128

- Char Iglesias, S. E. (2017). Machine Learning Techniques to Predict Hospital Length of Stay (LOS) at the Time of Patient Admission (Undergraduate thesis). Universidad de los Andes.
- Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.
- Chrusciel, J., Girardon, F., Roquette, L., et al. (2021). The prediction of hospital length of stay using unstructured data. BMC Medical Informatics and Decision Making. https://doi.org/10.1186/s12911-021-01722-4
- CMS (2024). *ICD-10 codes.* Disponible en: https://www.cms.gov/medicare/coding-billing/icd-10-codes
- David Martínez, C. C., Bonet Cruz, I., & Camacho Cogollo, J. E. (2021). Modelo predictivo para el pronóstico de tiempos de estancia de pacientes en unidades de cuidados intensivos. Universidad EIA.
- Ghosh, A., Nashaat, M., Miller, J., Quader, S., & Marston, C. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets. Visual Informatics, 2(4), 235-253. https://doi.org/10.1016/j.visinf.2018.12.004
- GOV. (1981). Ley 23 de 1981. Diario Oficial No. 35.704. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=68760
- GOV. (2011). Ley 1437 de 2011. Modificada por Ley 2220 de 2022, Ley 2195 de 2022, Ley 2080 de 2021, y otros. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=41249
- GOV. (2012). Ley Estatutaria 1581 de 2012. Reglamentada por Decreto 1081 de 2015, Decreto 886 de 2014, Decreto 1377 de 2013. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981
- GOV. (2020). Ley 2015 de 2020. Diario Oficial No. 51.238. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=105472
- GOV (2024). Registro Especial de Prestadores y Sedes de Servicios de Salud. Datos abiertos. Disponible en: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Registro-Especial-de-Prestadores-y-Sedes-de-Servic/c36g-9fc2/about_data
- Gutierrez, J., Sicilia, M., Sanchez-Alonso, S., & Barriocanal, E. (2021). Predicting length of stay across hospital departments. IEEE Access, PP(1), 1-1. https://doi.org/10.1109/ACCESS.2021.3066562
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics. Springer.
- Microsoft (2016) *LightGBM.* Disponible en: https://www.microsoft.com/en-us/research/project/lightgbm/
- Ministerio de Salud y Protección Social. (1993). Resolución 8430 de 1993. Por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud. Disponible en:

https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/RESOLUCION-8430-DE-1993.PDF

- Ministerio de Salud y Protección Social. (2019). resolución 3100 de 2019. Disponible en: https://www.minsalud.gov.co/normatividad_nuevo/resoluci%C3%B3n%20no.%203100%20de%202019.pdf
- Ministerio de Salud y Protección Social. (2000). Resolución 3374 de 2000. Por la cual se adopta la Clasificación Internacional de Enfermedades CIE-10. Disponible en: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/Resoluci%C3%B3n_3374_de_2000.pdf
- Ministerio de Salud y Protección Social (1994). *Resolución 5261 de 1994.* Disponible en: https://www.icbf.gov.co/cargues/avance/compilacion/docs/resolucion_minsalud_r5261_94.htm
- Nascimento, E. D., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., & Conte, T. U. (2019). Understanding development process of machine learning systems: Challenges and solutions. 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 1-6.
- OECD. (2021). Hospital discharges and average length of stay. In Health at a Glance 2021: OECD Indicators. OECD Publishing, Paris.
- Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., & Cortez-Pinto, E. (2021). Generalidades del machine learning y su aplicación en la gestión sanitaria en servicios de urgencia. Revista Médica de Chile, 149(2), 248-254.
- Peres, I. T., Hamacher, S., Oliveira, F. L. C., Bozza, F. A., & Salluh, J. I. F. (2021). Prediction of intensive care units length of stay: A concise review. Revista Brasileira de Terapia Intensiva, 33(2), 183-187. https://doi.org/10.5935/0103-507X.20210025
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features*. arXiv. https://doi.org/10.48550/arXiv.1706.09516
- 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. Facultad de Ingeniería, Especialización en Analítica y Ciencia de Datos.
- Restrepo-Zea, J. H., Jaén-Posada, J. S., Espinal Piedrahita, J. J., & Zapata Flórez, P. A. (2018). Saturación en los servicios de urgencias: Análisis de cuatro hospitales de Medellín y simulación de estrategias. Revista Gerencia y Políticas de Salud, 17(34), 130-144. https://doi.org/10.11144/javeriana.rgps17-34.ssua
- Secretaria de salud de Medellín. (2018). *Prestadores de Servicios de Salud habilitados en Medellín.* Disponible en: https://www.medellin.gov.co/irj/go/km/docs/pccdesign/medellin/Temas/Salud_0/IndicadoresEstadisticas/Shared%20Content/Observatorio/Archivos%20PDF/Prestadores_SS_habilitados_Medell%C3%ADn_Jun2018.pdf

- Shulga, D. (2018). 5 Reasons why you should use Cross-Validation in your Data Science Projects. Towards Data Science. Disponible en: https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-projects-6d80d678e731
- Saludcapital (s.f). *Codificación de Municipios por Departamento*. Disponible en: https://www.saludcapital.gov.co/Biblioteca%20de%20Documentos%20DPS%20RIPS/Codificaciones/Codificaci%C3%B3n%20de%20Municipios%20por%20Departamento.pdf

9. Apéndice

9.1 Apéndice A: Marco Conceptual

Machine Learning (ML): Campo de la inteligencia artificial que utiliza algoritmos para que los sistemas aprendan y hagan predicciones sin ser programados explícitamente. Aquí, ML se emplea para predecir el tiempo de estancia.

La Unidad de Observación de Urgencias: Es un área ubicada dentro del servicio de urgencias que brinda asistencia y supervisión a los pacientes después de su atención inicial. En esta unidad, los pacientes permanecen bajo observación mientras se determina el siguiente paso en su tratamiento, ya sea la realización de un examen especializado, su admisión para hospitalización o el regreso a su domicilio tras unas horas de vigilancia para asegurar su estabilidad. Este espacio permite que el equipo médico evalúe de forma segura la evolución del paciente antes de decidir el curso más adecuado para su situación.

Exploración de datos (EDA): Es el proceso de análisis inicial de un conjunto de datos con el objetivo de descubrir patrones, detectar anomalías, probar hipótesis y verificar supuestos utilizando estadísticas descriptivas y visualizaciones gráficas. El EDA permite comprender la estructura de los datos, identificar relaciones entre variables y preparar los datos para análisis más avanzados, como la construcción de modelos predictivos.

One-Hot Encoding: Es una técnica de codificación utilizada en machine learning para convertir variables categóricas en una representación numérica. Crea una columna binaria (con valores 0 y 1) por cada categoría de la variable, asignando un "1" a la columna correspondiente a la categoría presente en cada registro y "0" a las demás. Esto permite que los algoritmos procesen datos categóricos sin asumir un orden implícito.

Árboles de decisión: Modelo de machine learning para clasificación y regresión, donde los datos se dividen jerárquicamente mediante reglas de decisión. Cada nodo es una pregunta sobre una variable, y cada hoja representa una predicción. Son propensos al sobreajuste si no se controla su profundidad, pero técnicas como boosting pueden ayudar a mejorar su precisión.

Random Forest: Algoritmo basado en múltiples árboles de decisión donde cada árbol vota por la predicción final. Es robusto para datos no lineales y se usa tanto para clasificación como para regresión.

Boosting: Es una técnica de machine learning que combina múltiples modelos (normalmente árboles de decisión) para formar un modelo más robusto. El boosting mejora tanto la precisión del modelo como su capacidad para generalizar a nuevos datos, al enfocarse en los errores cometidos por árboles anteriores y ajustarlos de manera secuencial. Algoritmos como XGBoost, CatBoost, y LightGBM son ejemplos de técnicas de boosting.

CatBoost: CatBoost es un algoritmo de machine learning basado en boosting de árboles de decisión, diseñado para manejar de manera eficiente variables categóricas, especialmente aquellas con alta cardinalidad. Utiliza técnicas innovadoras como la codificación de estadísticas objetivo-ordenadas (Ordered Target Statistics) y el boosting ordenado (Ordered Boosting), las cuales mitigan problemas como el target leakage y el prediction shift, comunes en otros algoritmos de boosting (Prokhorenkova et al., 2019). Implementa árboles de decisión oblivios, lo que mejora la eficiencia en predicción y reduce el riesgo de sobreajuste (Prokhorenkova et al., 2019). Además, CatBoost incorpora el Bayesian Bootstrap, que asigna pesos probabilísticos a los datos durante el entrenamiento, mejorando la robustez del modelo al evitar la dependencia estricta de muestras específicas (Prokhorenkova et al., 2019). Para categorías con pocos datos, utiliza suavizado bayesiano en las estadísticas objetivo, reduciendo el impacto del ruido y mejorando la generalización del modelo.

Target Leakage: Ocurre cuando información del objetivo (target) influye en las variables predictoras durante el entrenamiento, generando un modelo con desempeño irrealmente alto en entrenamiento, pero pobre en generalización.

Prediction Shift: Problema en modelos de boosting donde las predicciones en los datos de entrenamiento se desvían de las predicciones en datos de prueba debido al uso repetido del mismo conjunto de datos para calcular gradientes y ajustar los modelos

Árboles de Decisión Oblivios: Árboles de decisión balanceados donde cada nivel utiliza la misma condición de división en todas las ramas, lo que simplifica la interpretación, reduce el riesgo de sobreajuste y mejora la eficiencia computacional.

Bayesian Bootstrap: Método que asigna pesos probabilísticos a los datos de entrenamiento mediante una distribución bayesiana, mejorando la robustez del modelo al reducir la dependencia de muestras específicas y aumentando la estabilidad de las predicciones.

Codificación de medias de categorías: Este método reemplaza cada categoría con la media de la variable objetivo de esa categoría. Esto permite al modelo capturar la relación entre la categoría y la variable a predecir, sin la necesidad de generar múltiples variables dummy.

Transformación por orden: Esta técnica asigna valores a cada categoría basándose en la estadística de los registros previos, sin utilizar el registro actual. De este modo, se evita el sobreajuste y se mejora la capacidad del modelo para generalizar a nuevos datos.

LightGBM: Light Gradient Boosting Machine (LightGBM) es un framework de boosting desarrollado por Microsoft, diseñado para mejorar la eficiencia y velocidad en el entrenamiento de modelos de machine learning, especialmente en datasets grandes. Utiliza un enfoque basado en histogramas para agrupar valores continuos, lo que reduce significativamente el uso de memoria y acelera la búsqueda de divisiones en los árboles. LightGBM emplea un método de construcción de árboles leaf-wise, que prioriza la expansión de las hojas con mayor ganancia, logrando modelos más precisos, pero con mayor riesgo de sobreajuste si no se regula adecuadamente. Además, LightGBM maneja variables categóricas directamente mediante gradientes y agrupa las categorías con pocos datos en un único clúster, mitigando el impacto de estas categorías en el modelo, pero sacrificando algo de información detallada. Ofrece excelente escalabilidad y paralelización para sistemas distribuidos.

Árboles Leaf-Wise: Son árboles de decisión que expanden primero las hojas con mayor ganancia en cada iteración, en lugar de dividir todos los niveles uniformemente (level-wise). Este enfoque permite una construcción más eficiente y precisa del árbol, aunque puede aumentar el riesgo de sobreajuste en datasets pequeños si no se regula adecuadamente.

XGBoost: Extreme Gradient Boosting (XGBoost) es un algoritmo de boosting altamente eficaz y utilizado ampliamente en competencias de machine learning. Este modelo se caracteriza por su capacidad de manejar relaciones no lineales entre variables mediante árboles de decisión estándar, apoyado por técnicas de regularización avanzadas como L1 y L2 para evitar el sobreajuste. Aunque originalmente utiliza One-Hot Encoding para variables categóricas, soporta divisiones categóricas directas (Categorical Split Transformation) en versiones recientes (Chen & Guestrin, 2016). Sin embargo, no incluye mecanismos para evitar problemas de target leakage, como los implementados en CatBoost. XGBoost también se destaca por su implementación eficiente, velocidad de entrenamiento, y capacidad para capturar patrones complejos en datasets grandes, aunque carece de técnicas como el Bayesian Bootstrap que aporta mayor robustez en escenarios con datos limitados o desbalanceados (Chen & Guestrin, 2016).

Categorical Split Transformation: Método para dividir variables categóricas en árboles de decisión directamente, sin convertirlas en variables binarias (One-Hot Encoding), lo que reduce la complejidad y mantiene la eficiencia del modelo.

Red Neuronal: Modelo inspirado en el cerebro humano, compuesto por capas de neuronas interconectadas. Cada neurona aplica una función matemática y transmite el resultado a las siguientes, útil para problemas de clasificación y regresión.

Red Neuronal densa con Embeddings: Una red neuronal densa es un tipo de red neuronal artificial compuesta por capas completamente conectadas, lo que significa que cada neurona de

una capa está conectada a todas las neuronas de la capa anterior y la capa siguiente. Las capas de embeddings permiten representar categorías en un espacio de menor dimensión, lo cual mejora la eficiencia del modelo y su capacidad para captar relaciones entre las categorías especialmente cuando estas tienen alta cardinalidad.

5-Fold Cross Validation (Validación Cruzada de 5 Pliegues): Método que divide los datos en 5 partes, usando 4 para entrenar y 1 para validar, repitiendo el proceso cinco veces. Esto reduce la varianza y proporciona una estimación más robusta del rendimiento del modelo.

MAE (Mean Absolute Error - Error Absoluto Medio): Métrica que mide la precisión de los modelos de regresión calculando la media de las diferencias absolutas entre los valores predichos y los reales. Un MAE menor indica mayor precisión, y es fácil de interpretar ya que está en la misma escala que la variable objetivo. Por ejemplo, un MAE de 8.82 horas indica que, en promedio, las predicciones se desvían 8.82 horas del valor real.

Support Vector Regression (SVR): Es un método de machine learning utilizado para resolver problemas de regresión, diseñado para encontrar una función que prediga los valores de una variable continua dentro de un margen de tolerancia definido. Ha demostrado ser más preciso que los métodos estadísticos tradicionales en aplicaciones como la predicción de estancias en unidades de cuidados intensivos.

ICD-10 (Clasificación Internacional de Enfermedades): Clasificación estandarizada de enfermedades desarrollada por la OMS. En este proyecto, se usaron códigos ICD-10 para entender las bases de datos codificadas.

SHAP (SHapley Additive exPlanations): Técnica para interpretar la importancia de las variables en un modelo. Se basa en la teoría de valores de Shapley para asignar una contribución a cada característica en la predicción. Un valor SHAP positivo indica que la variable aumenta la predicción, mientras que uno negativo indica que la disminuye. Por ejemplo, un valor SHAP de +2 para "HospitalOccupancy" indica que esta variable aumenta el tiempo de estancia en promedio 2 horas.