Analítica III - Caso de Estudio Salud

Cesar Iván Ávila Díaz. cesar.avila@udea.edu.co
Verónica Andrea Morales González. veronica.moralesg@udea.edu.co

Repositorio GitHub: https://github.com/veronica1908/SALUD_E3_AIII

Resumen: El informe del caso aborda la implementación del Pago Global Prospectivo (PGP) en el Hospital Alma Mater de Antioquia, con el fin de predecir la hospitalización mensual de adultos mayores de 60 años con condiciones crónicas. A través de un exhaustivo proceso de limpieza y transformación de datos, seguido de un análisis exploratorio detallado, se identifican variables relevantes para el modelado predictivo. La selección de variables se lleva a cabo mediante la evaluación de diversos modelos de regresión, destacando GradientBoostingRegressor como el más efectivo en términos de Error Cuadrático Medio (MSE). Este enfoque sistemático proporciona una base sólida para desarrollar un modelo predictivo que pueda mejorar la atención médica y optimizar el uso de recursos en el hospital.

Palabras Clave: Clasificación funcional, Datos, Hospitalización, Modelado predictivo, Selección de variables.

1. Descripción del caso problema

En el contexto del Sistema General de Seguridad Social en Salud (SGSSS), el Hospital Alma Mater de Antioquia implementa el Pago Global Prospectivo (PGP) para pacientes crónicos. Se clasifican según su estado funcional en cinco categorías, determinando el tratamiento ambulatorio o domiciliario. La aseguradora envía mensualmente una lista de pacientes bajo PGP. El desafío es predecir la hospitalización mensual para adultos mayores de 60 años con estado crónico, según su clasificación funcional. Se utilizan bases de datos de egresos hospitalarios, crónicos y un listado de población.

2. Diseño de la solución para el caso de estudio

La solución propuesta se muestra gráficamente a continuación:

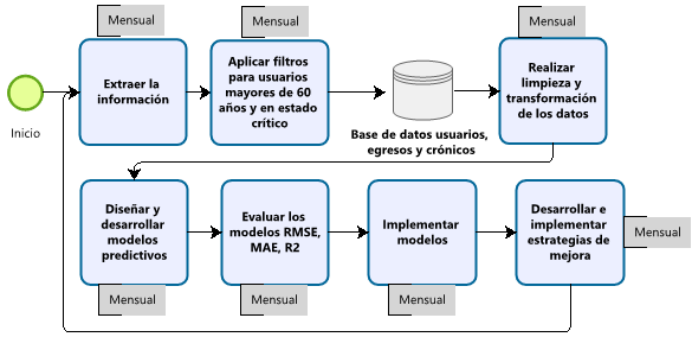


Ilustración 1. Solución teórica propuesta. Elaboración propia.

3. Desarrollo del caso de estudio

a. *Limpieza y transformación*

Durante esta etapa, se realizan operaciones de limpieza y transformación de datos para asegurar la coherencia y calidad de los datos. Esto incluye manejar valores nulos, eliminar duplicados y ajustar el formato de los datos según sea necesario. Además, se aplica el filtro previo de acuerdo con la información que se desea analizar. Para este caso, se tomó primero la base de usuarios y se aplicó un filtro por edad, para tomar solamente aquellos que tuvieran 60 años o más, pasando de 183.911 registros a 130.595. Luego se verificó si había registros duplicados, ya que de esta base interesan los usuarios únicos. Efectivamente; se encontró 123.665 registros duplicados, por lo que se extrae los valores únicos de acuerdo con su número de documento, quedando así con 6930 registros y 16 variables. Se verifica la cantidad de datos nulos por variable y se eligen las variables que cuentan con la mayor cantidad de datos y que son indispensables para el análisis. Aunque las variables asociadas con clase funcional tienen una gran cantidad de datos nulos, se toma "ÚLTIMA CLASE FUNCIONAL", ya que tiene menor cantidad y en el contexto del caso, es importante esta clasificación. Se unifica algunas categorías en cuanto a su escritura y finalmente la base usuarios queda con dimensiones de 6930 filas y 7 columnas.

Con la base de egresos, se inicia aplicando un filtro a los registros que correspondan al número de documento de los usuarios que fueron filtrados en la base anterior, que corresponden a edad =>60 años, esto a partir de la variable de número de documento, la cual permite relacionar las tablas. De esta manera se pasa de 6376 registros a 5781. Teniendo en cuenta que la base de usuarios es de 6930 registros, se puede inferir que no todos los usuarios tuvieron egresos.

Ser procede a verificar las 66 variables de esta base y de acuerdo con el diccionario proporcionado, se seleccionan 20 de ellas que son consideradas importantes para el análisis. Luego se verifica la cantidad de nulos por variables y justamente las variables con datos nulos son las que indican la fecha de ingreso a la clínica y al servicio, por lo que es imprescindible tener la información para calcular la estancia hospitalaria. Aquí se pasaría entonces a tener 3737 registros, es decir, se perdería el 35.3% de los registros. Se trabajaría entonces con el 64,7% de ellos, lo cual sigue siendo representativo y se procede a eliminar las filas nulas. Se convierten las variables DEMORA

ASIGNACION CAMA y DEMORA APLICACION MEDICAMENTO de horas a días. La variable DEMORA APLICACION MEDICAMENTO tiene varios valores negativos, por lo que se decide no considerarla para la base. Se ajusta el formato de las variables fecha a tipo datetime y de las numéricas a tipo integer. Luego se crea la variable TIEMPO ESTANCIA (DIAS) a partir de la resta entre FECHA SALIDA y FECHA INGRESO SERVICIO. Finalmente, se aplica un filtro para la MODALIDAD CONTRATO para el tipo PGP, ya que para la población que hace parte de este contrato se requiere predecir el uso mensual de recursos traducido en hospitalización de acuerdo con la clase funcional a la que pertenece el paciente. E4sta base queda entonces con 3635 filas y 19 columnas.

Para la tercera base correspondiente a los casos crónicos, lo primero fue filtrar los registros que correspondan al número de documento de los usuarios que fueron filtrados en la base egresos, que corresponden a edad =>60 años y que ingresaron a la clínica. Se pasó de tener 38.717 registros a 12.621. Luego se verifican las variables que en este caso eran 290 y como lo que interesa para este caso de estudio es evaluar la estancia hospitalaria para pacientes críticos de más de 60 años, se toma de aquí: la variable Diagnóstico principal para evaluarla con respecto a la estancia hospitalaria, así como la variable de cruce que es NRODOC y algunas otras como complemento para analizar el caso, los pacientes y sus diagnósticos, quedando con ocho variables. Se verificó la cantidad de diagnósticos asociados a cada número de documento y la cantidad de diagnósticos únicos, ya que, si bien un usuario puede tener varios diagnósticos principales, no debería tener el mismo diagnóstico más de una vez por lo que se eliminan los registros de los usuarios cuyo diagnóstico está repetido, dejando para cada usuario los diagnósticos asignados diferentes entre sí. Se realiza la verificación de datos nulos por variables y se tiene que: La variable de AMBITO SEGUN EL MEDICO, puede aportar una información complementaria y como es de tipo categórica, se reemplazará los datos nulos por la categoría "SIN INFORMACIÓN", igual proceso se realizará con la variable de DIAGNOSTICO PRINCIPAL, ya que si bien es importante saber el diagnóstico, se sabe que estos usuarios tienen alguna enfermedad crónica, y de todas maneras afectarán el indicador de estancia hospitalaria y los recursos. Si esta categoría llegara a tener por ejemplo un promedio alta de estancia hospitalaria, se recalcaría la importancia de registrar siempre la información completa, sobre todo de variables tan representativas. Se verifica los valores de las columnas PESO y TALLA para detectar valores negativos, de cero o valores demasiado altos. Se obtienen valores de cero en el mínimo, lo cual no es posible, por lo que estos valores se reemplazan con el promedio.

Se realiza ajustes de escritura en los valores del diagnóstico principal y además se detectan 155 diagnósticos diferentes, lo cual es un número muy alto al momento de aplicar modelos, por lo que se crea una variable llamada 'CATEGORIAS DIAGNOSTICOS' para agrupar los diagnósticos de acuerdo con su naturaleza, para ello, recurrimos a la inteligencia artificial a través de chat GPT para que nos sugiera agrupación de los diagnósticos y nombres de las categorías de acuerdo con la información y se obtuvo una clasificación de diagnósticos en 11 categorías. La base queda finalmente con 2556 filas y 9 columnas.

b. Análisis exploratorio

En el análisis exploratorio, se realizó para cada una de las bases de manera independiente para ver el comportamiento general de las variables, lo cual se incluyó en el código. Aquí se realizaron otros ajustes en los que se aplicaron filtros por lo que se eliminan las columnas de MODALIDAD CONTRATO, EPS VALIDADA, REGIMEN AFILIACION y TIPO DIAGNOSTICO PRINCIPAL, ya que se realizó anteriormente filtro por modalidad de contrato y directamente quedó solo una categoría en la EPS VALIDADA y en el REGIMEN AFILIACION. Además, se aplicó también el filtro para la categoría "Nuevo" en la variable TIPO DIAGNOSTICO PRINCIPAL, por lo que se retira de la base.

Se realizó la combinación de las tres tablas a partir de la variable de tipo de documento y se ajustó el tipo duplicados, quedando con una tabla de las siguientes dimensiones: 1247 filas y 29 columnas, se aplicó el siguiente análisis exploratorio con respecto a la variable objetivo:

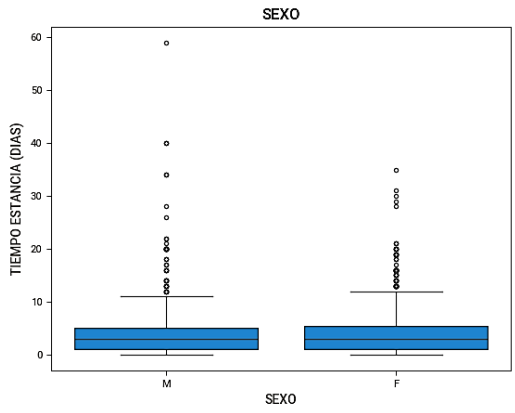


Ilustración 2. Estancia vs Sexo

La estancia hospitalaria es más alta para los usuarios de sexo femenino, en comparación con el sexo masculino en general.

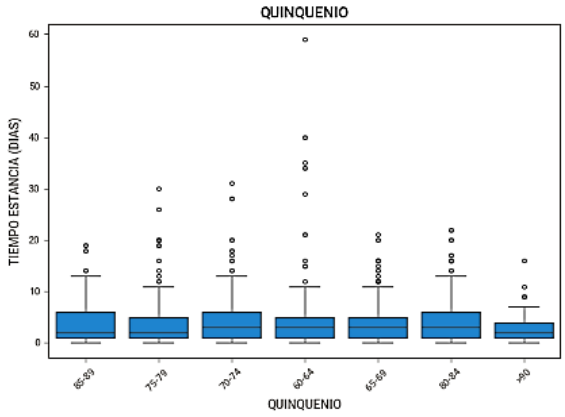


Ilustración 3. Estancia vs Quinquenio (edad)

Las personas con edades comprendidas entre los 70 - 74 años y 80-84 años, tienen mayor tiempo de estancia.

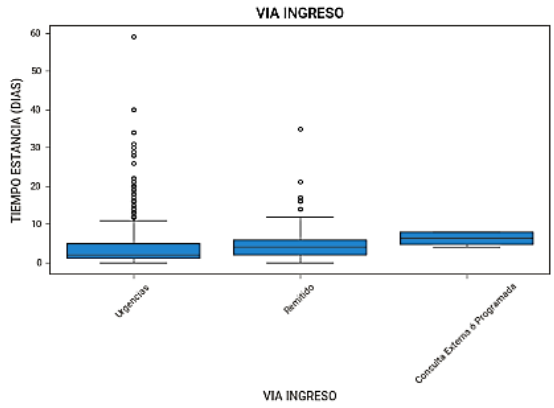


Ilustración 4. Estancia vs Vía ingreso

Revisando la vía de ingreso, se tiene que es la vía de consulta externa o programada, la que tiene mayor promedio de tiempo de estancia en comparación con urgencias que es la vía que tiene mayor cantidad de registros, aunque esta misma es la que posee mayor cantidad de datos atípicos que alcanzan hasta los casi 60 días.

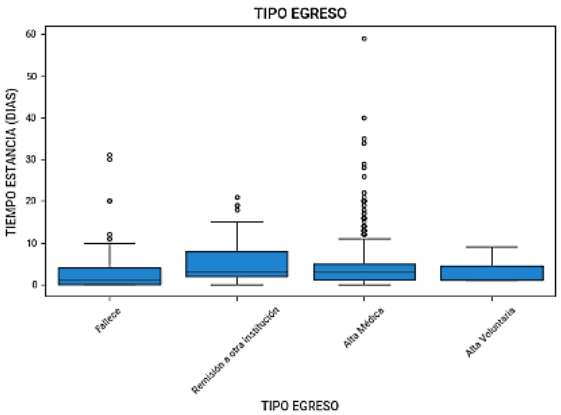


Ilustración 5. Estancia vs Tipo egreso

Para el tipo de egreso, es la remisión a otra institución, la que tiene una media de tiempo de estancia más alta y en segundo lugar, la alta médica.

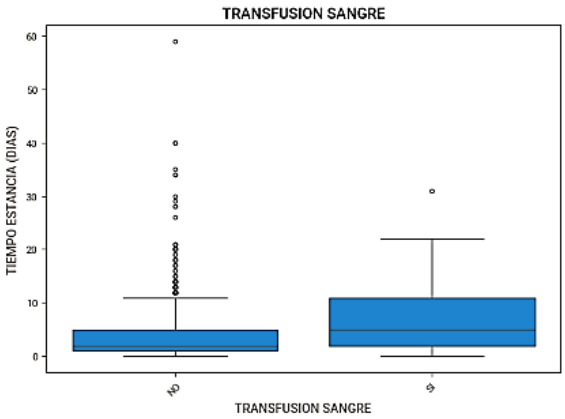


Ilustración 6. Estancia vs Transfusión de sangre

Los pacientes que han necesitado transfusión de sangre, son los que representan un promedio más alto de tiempo de estancia en comparación con los que no.

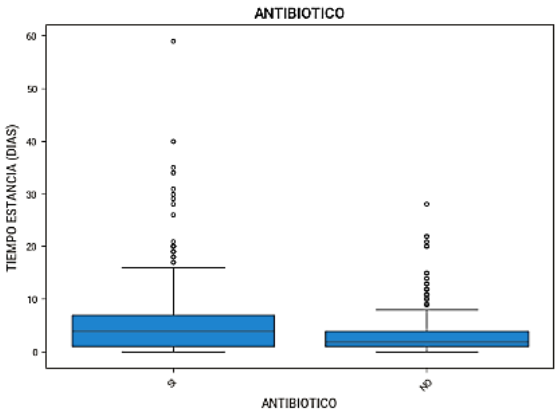


Ilustración 7. Estancia vs Antibiótico

Los pacientes que han requerido antibiótico tienen mayor tiempo de estancia.

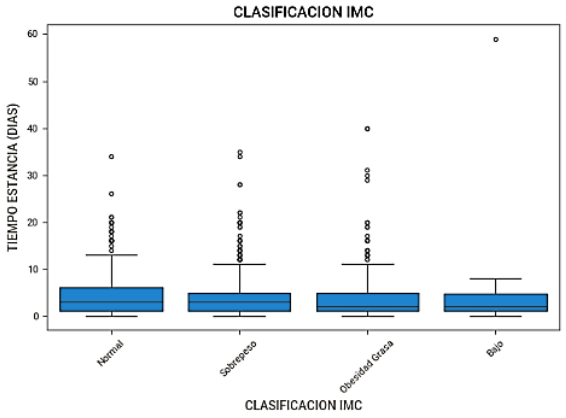


Ilustración 8. Estancia vs Clasificación IMC

La media entre pacientes con clasificación IMC normal y sobrepeso son muy similares, sin embargo, los pacientes con Clasificación IMC normal, abarcan un rango más amplio de tiempo de estancia.

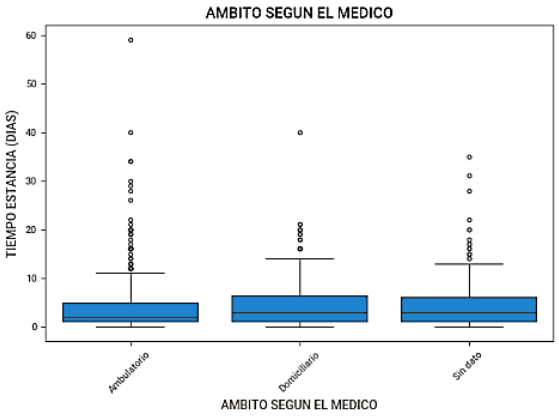


Ilustración 9. Estancia vs Ámbito según el médico

El ámbito según el médico de tipo ambulatorio, es el que presenta la media más alta de tiempo de estancia.

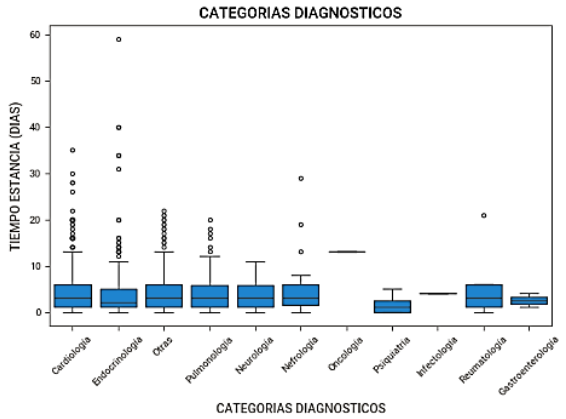
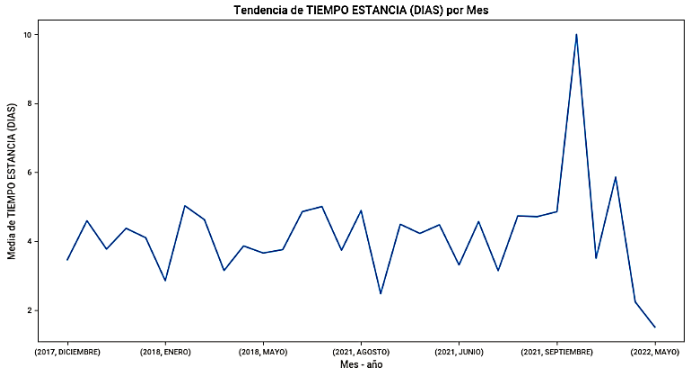


Ilustración 10. Estancia vs Categorías diagnósticos

Finalmente, la categoría de diagnósticos con promedio más alto de estancia es Oncología con 13 días de tiempo de estancia promedio. El diagnóstico principal que más cantidad de tiempo de estancia tiene es la ENFERMEDAD CARDIOVASCULAR NO ESPECIFICADA con un promedio de 22 días, en segundo lugar, está OTROS TIPOS DE HIPERTENSION SECUNDARIA con 16 días y en tercer lugar TUMOR MALIGNO DE OTRAS PARTES ESPECIFICADAS DEL PANCREAS con 13 días de tiempo de estancia.



Se observa cómo se ha mantenido en el tiempo, sin embargo, se evidencia un pico alto después de septiembre de 2021, alcanzando un promedio de 10 días de estancia.

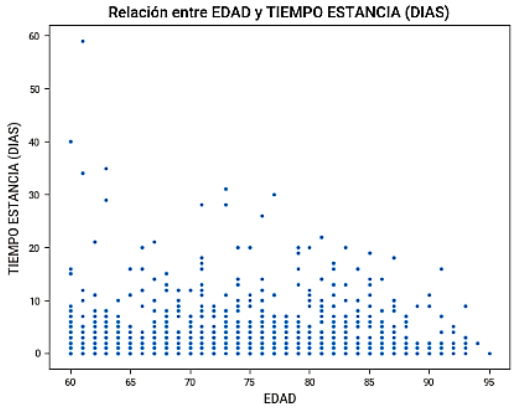


Ilustración 11. Estancia vs Edad

En cuanto a la edad, se ve un ligero sesgo a la izquierda, mostrando que las edades más cercanas a 60 años son las que muestran mayor tiempo de estancia.

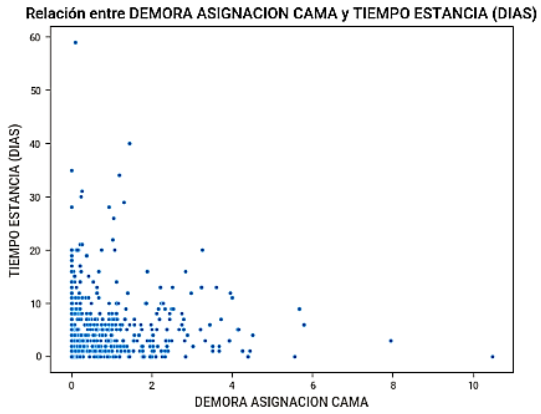


Ilustración 12. Estancia vs Demora asignación de cama

Para el tiempo de demora en asignación de cama, se tiene que cuanto más rápido se asigna una cama al paciente, mayor tiempo de estancia tendrá, esto asociado quizá a la gravedad del evento, se muestra un sesgo marcado hacia la izquierda.

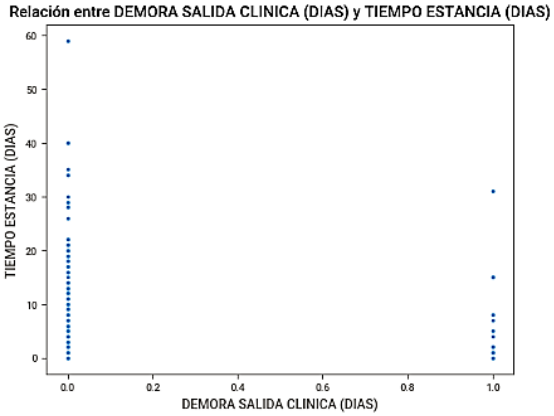


Ilustración 13. Estancia vs Salida clínica

Para la demora de salida de la clínica, también se tiene mayor estancia para los menores tiempos.

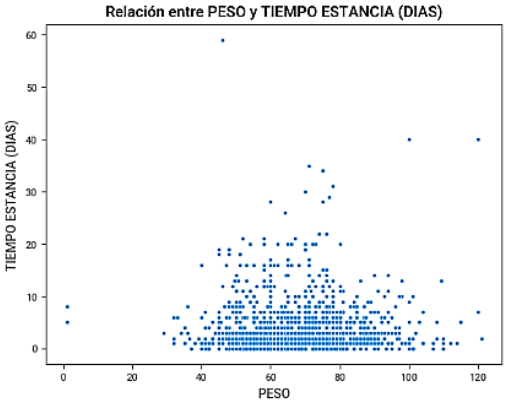


Ilustración 14. Estancia vs Peso

En cuanto al peso, no se ve una relación directa, y se observa que los pesos alrededor de los 70 y 80 kilos, son los que presentan mayor tiempo de estancia.

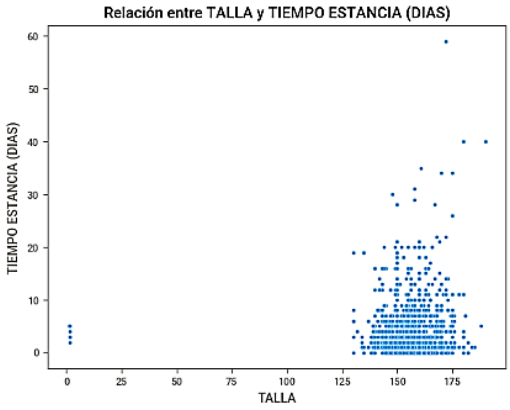


Ilustración 15. Estancia vs Talla

Finalmente, para la variable de talla, tampoco se evidencia una correlación.

c. Selección de variables

En el proceso de selección de variables, se llevó a cabo una serie de pasos para identificar y determinar qué características eran más relevantes para el modelo predictivo.

Inicialmente, se identificaron las variables categóricas en el conjunto de datos y se almacenaron las categorías únicas de cada columna. Luego, se aplicó la codificación LabelEncoder a las columnas categóricas que tenían solo dos categorías, convirtiéndolas en variables numéricas.

Posteriormente, se generaron variables dummy para las columnas categóricas con más de dos categorías, lo que permitió incluir esta información en el análisis sin asignar un orden numérico a las categorías.

El siguiente paso crucial fue la selección de variables, que se realizó utilizando varios modelos de regresión, incluyendo Decision Tree Regressor, Random Forest Regressor y Gradient Boosting Regressor. En este proceso, se evaluó el rendimiento de los modelos utilizando todas las variables disponibles en comparación con un conjunto seleccionado de variables.

Para determinar qué conjunto de variables producía los resultados más óptimos, se calculó el Error Cuadrático Medio (MSE) para cada modelo con ambas configuraciones. Esto proporcionó una evaluación cuantitativa del rendimiento de los modelos con diferentes conjuntos de variables.

Además, se llevó a cabo una búsqueda exhaustiva de hiperparámetros para los modelos Random Forest Regressor y Gradient Boosting Regressor utilizando la técnica GridSearchCV. Esto implicó ajustar diferentes combinaciones de hiperparámetros y evaluar el rendimiento de cada modelo utilizando la métrica de MSE.

d. Selección de algoritmos y técnicas de modelado

La selección de algoritmos y técnicas de modelado en este caso se realizó de manera sistemática, utilizando una variedad de modelos de regresión para abordar el problema de predicción de estancias hospitalarias. Se optó por modelos ampliamente utilizados en el ámbito de la ciencia de datos y el aprendizaje automático, como DecisionTreeRegressor, RandomForestRegressor y GradientBoostingRegressor.

El proceso de selección de modelos implicó varios pasos:

Definición de Modelos Candidatos: Se seleccionaron modelos candidatos basados en su idoneidad para resolver problemas de regresión y su capacidad para manejar conjuntos de datos complejos.

Evaluación de Modelos: Se evaluó el rendimiento de cada modelo utilizando la técnica de validación cruzada y la métrica de Error Cuadrático Medio (MSE) y de R2. Esta evaluación proporcionó una medida cuantitativa del desempeño predictivo de cada modelo en el conjunto de datos.

Comparación de Resultados: Se compararon los resultados de los diferentes modelos utilizando tanto todas las variables disponibles como un conjunto seleccionado de variables. Esto permitió determinar qué modelos funcionaban mejor con qué conjunto de características.

Afinamiento de Hiperparámetros: Se realizó un ajuste fino de los hiperparámetros de los modelos seleccionados mediante la técnica GridSearchCV. Esto implicó explorar diferentes combinaciones de valores de hiperparámetros para optimizar el rendimiento de los modelos.

Selección del Modelo Ganador: Se seleccionó el modelo con el mejor rendimiento predictivo, basándose en la métrica de MSE y R2, tanto para el conjunto de datos con todas las variables como para el conjunto de datos con variables seleccionadas. Esta selección se realizó para cada modelo candidato, lo que permitió identificar cuál era el más adecuado para el problema específico de predicción de estancias hospitalarias.

```
MSE para todos los modelos con todas las variables:
DecisionTreeRegressor: 39.24574138598529
RandomForestRegressor: 23.513896993805655
GradientBoostingRegressor: 22.757998023947078

MSE para todos los modelos con variables seleccionadas:
DecisionTreeRegressor: 41.22948122338367
RandomForestRegressor: 24.48529757865428
GradientBoostingRegressor: 23.842416891579912
```

El modelo GradientBoostingRegressor tiene el MSE más bajo tanto para el conjunto de datos con todas las variables como para el conjunto de datos con variables seleccionadas. Por lo tanto, si solo consideramos el MSE como métrica de evaluación, podríamos concluir que el modelo GradientBoostingRegressor es el mejor de los tres para este conjunto de datos y condiciones particulares de entrenamiento.

e. Afinamiento de hiperparámetros

El afinamiento de hiperparámetros se realizó utilizando la técnica GridSearchCV. Esta técnica permite explorar exhaustivamente un espacio de hiperparámetros predefinido para encontrar la combinación óptima que maximice el rendimiento del modelo. En el proceso de ajuste de hiperparámetros, se identificaron los valores óptimos para los modelos RandomForestRegressor y GradientBoostingRegressor. Para RandomForestRegressor, los mejores parámetros encontrados fueron: max_depth: 10, min_samples_leaf: 1, min_samples_split: 10, n_estimators: 50, con un puntaje de MSE de 22.778560322711368. Este puntaje refleja el error cuadrático medio promedio durante la validación cruzada con los parámetros optimizados, indicando la precisión del modelo en la predicción de la variable objetivo "TIEMPO ESTANCIA (DIAS)" en el conjunto de datos de prueba. Por otro lado, para GradientBoostingRegressor, los mejores parámetros fueron: learning_rate: 0.05, max_depth: 3, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 50, subsample: 0.8, con un puntaje de MSE de 22.089599729040987. Aunque este puntaje es ligeramente mejor que el obtenido con RandomForestRegressor, la diferencia no es significativa. Además, se considera que el costo computacional del modelo RandomForestRegressor es mucho menor, lo que lo hace más práctico para su implementación con el conjunto de datos disponible. En consecuencia, se opta por utilizar el modelo RandomForestRegressor debido a su rendimiento comparable y eficiencia computacional superior.

f. Evaluación y selección del modelo

En la fase de evaluación del modelo RandomForestRegressor, se llevaron a cabo análisis tanto en el conjunto de datos de entrenamiento como en el conjunto de datos de evaluación después de sintonizar los

hiperparámetros del modelo. Se calculó el Error Cuadrático Medio (MSE) para medir la idoneidad del ajuste del modelo a los datos de entrenamiento y evaluación, resultando en valores de 10.994934 y 15.858835 respectivamente. Estos valores indican la precisión del modelo en ambos conjuntos de datos, reflejando un ajuste adecuado y una capacidad de generalización aceptable a datos no vistos. Además, se evaluó el Error Absoluto Medio (MAE), revelando valores de 2.2308552633420766 para el conjunto de datos de entrenamiento y 2.771020396376993 para el conjunto de datos de evaluación. Estos resultados sugieren que el modelo tiene un buen rendimiento en la predicción de estancias hospitalarias, con una precisión promedio de aproximadamente 2.23 días en el conjunto de entrenamiento y 2.77 días en el conjunto de evaluación.

4. Conclusiones y recomendaciones

RandomForestRegressor se destacó como la mejor opción por su capacidad predictiva y eficiencia computacional, lo que lo hace idóneo para predecir estancias hospitalarias en adultos mayores con enfermedades crónicas.

La implementación de modelos predictivos como RandomForestRegressor puede optimizar la asignación de recursos hospitalarios, mejorando la gestión de camas y reduciendo los tiempos de espera, lo que resulta en una atención más eficiente y personalizada.

La evaluación de modelos reveló que RandomForestRegressor tiene un rendimiento satisfactorio en la predicción de estancias hospitalarias, demostrando una precisión aceptable en la estimación de la duración de la hospitalización para adultos mayores con enfermedades crónicas. Sin embargo, es esencial continuar monitoreando su rendimiento en un entorno operativo para garantizar su eficacia continua.

Se recomienda integrar el modelo RandomForestRegressor en los sistemas de gestión hospitalaria para mejorar la planificación de recursos y la asignación de camas, lo que permitirá una atención más eficiente y oportuna para los pacientes crónicos. Además, se recomienda continuar monitoreando y evaluando el rendimiento del modelo en tiempo real, utilizando datos actualizados para ajustar y mejorar su precisión.

Se invita a implementar políticas y procedimientos para garantizar la calidad y completitud de los datos registrados, lo que mejorará la confiabilidad y efectividad de los modelos predictivos en futuros análisis.

Por último, se sugiere promover la capacitación del personal de salud en el uso y comprensión de las herramientas de análisis de datos y modelos predictivos, lo que facilitará su adopción y aprovechamiento óptimo en la práctica clínica diaria. Estas medidas combinadas pueden contribuir significativamente a la mejora continua de la atención médica y la satisfacción del paciente en el Hospital Alma Mater y otros entornos de atención médica.