CASO DE ESTUDIO DE SALUD

Piedrahita Allison; Ramírez Anyi; Vergara María

Departamento de Ingeniería Industrial, Universidad de Antioquia.

Medellín, Colombia.

RESUMEN ~ El Pago Global Prospectivo (PGP) representa un mecanismo de pago anticipado a la prestación de un servicio de salud de una institución prestadora de salud (IPS), dicha modalidad de pago es asignada una población dependiendo de la frecuencia de atenciones o cantidad de población. Con el fin de prestar los servicios de manera efectiva, es necesario asignar los recursos suficientes a los pacientes hospitalizados bajo esta modalidad de pago. Para ello, se propone una solución analítica por medio de modelos de machine Learning que permitan predecir la estancia hospitalaria de un grupo de pacientes de PGP para realizar una correcta distribución de recursos y establecer planes de acción que permitan reducir la estancia hospitalaria y a su vez, llevar a cabo las operaciones de la IPS de manera efectiva.

1. INTRODUCCIÓN

La prestación del servicio de salud se encuentra a cargo de las instituciones prestadoras de los servicios de salud (IPS), las cuales se encuentran sometidas a dos tipos de mecanismos de pago para la generación de ingresos, entre ellos, el Pago Global Prospectivo (PGP), el cual consiste en realizar el pago antes de la prestación del servicio para una cohorte poblacional determinada de acuerdo con la frecuencia de atenciones o a la cantidad de población. Por tal motivo, dicha modalidad de pago requiere una optimización de recursos para atender a la población asignada de manera efectiva.

En el Hospital Alma Máter de Antioquia, se realiza una atención previa que permite definir la modalidad de tratamiento (ambulatorio o domiciliario) para adecuar la atención a las necesidades, capacidades, habilidades y potencialidades del paciente para optimizar las condiciones de salud. Sin embargo, para mejorar la atención de los pacientes pertenecientes la modalidad de contrato PGP, es de gran utilidad anticiparse a los recursos que se deben destinar en la atención de las personas, por ello, es necesario conocer la estancia hospitalaria (uso de recursos en hospitalización) para hacer una mejor distribución de estos. Además "*la estancia hospitalaria constituye una preocupación mundial, ya que genera efectos negativos en el sistema de salud como, por ejemplo: aumento en los costos, deficiente accesibilidad a los servicios de hospitalización, saturación de las urgencias y riesgos de eventos adversos*" [1].

Teniendo en cuenta lo anterior, se analizan tres bases de datos que contienen información de los pacientes atendidos en el Hospital Alma Máter de Antioquia, con el fin de establecer una solución analítica que permita establecer la estancia hospitalaria de un grupo relevante de pacientes pertenecientes a la modalidad PGP con la finalidad de identificar y establecer una mejor distribución de recursos y planes de acción y mejora que permitan llevar a cabo las operaciones de la IPS de una manera efectiva.

2. METODOLOGÍA

En las siguientes secciones se describe cada uno de los pasos desarrollados.

2.1.DISEÑO DE LA SOLUCIÓN

Con las bases de datos proporcionadas por la IPS universitaria, se plantean los siguientes objetivos para desarrollar el caso de estudio:

- Segmentar los pacientes de las bases de datos para seleccionar aquellos que representen un porcentaje significativo de personas con modalidad de contrato PGP, con el fin de enfocar la solución analítica.
- Predecir la estancia hospitalaria del grupo segmentado por medio de un modelo analítico basado en variables significativas de 12 meses históricos.
- Identificar las variables más significativas para establecer planes de acción que permitan mejorar la asignación de recursos y reducir los tiempos de hospitalización.

Con base a lo anterior, se plantea la solución analítica de la Figura 1, en la cual se evidencian los objetivos anteriores y las condiciones de tiempo de la información.

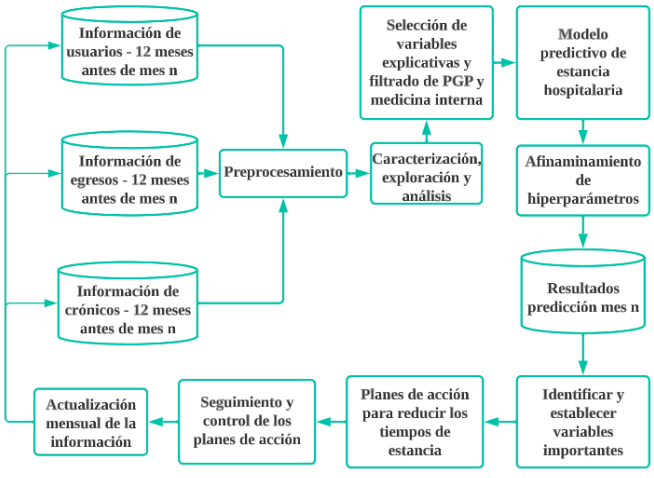


Figura 1. *Diseño de la solución analítica.*

2.2.LIMPIEZA Y TRANSFORMACIÓN

Para el desarrollo de la solución, se cuenta con tres bases de datos:

- *"RETO_df_egresos":* contiene la información administrativa y de gestión de los pacientes atendidos en la IPS de los años 2017, 2018 y 2021. Esta base de datos tiene 6.376 registros y 66 variables.
- *"RETO_df_cronicos":* contiene información de variables médicas que describen el estado de salud de los pacientes atendidos en la IPS de los años 2017, 2018, 2021 y 2022. Esta base de datos tiene 38.717 registros y 290 variables.
- *"RETO_df_usuarios":* contiene la información básica de los pacientes atendidos en la IPS de los años 2017, 2018, 2021 y 2022. Esta base de datos tiene 183.911 registros y 16 variables.

Para realizar la limpieza de la información, se realizan los siguientes pasos para cada una de las bases de datos:

- Separación de variables/columnas según el separador de datos (",")
- Eliminación de variables con más de 50% de registros nulos y selección de variables significativas que pueden aportar información según el análisis de las autoras.
- Eliminación de registros nulos de la base de datos reducida.
- Tratamiento de registros escritos diferente pero asociados a la misma categoría.
- Tratamiento del tipo de variable con el fin de asignar la clasificación correcta.

Para la base de datos de egresos, se realiza un procedimiento adicional, basado en el cálculo de la estancia hospitalaria para cada uno de los registros haciendo uso de la fecha de ingreso y salida del hospital. Con este procedimiento, se adiciona una variable más a la base de datos mencionada. Asimismo, aquellas variables con información de duración de tiempo se trataron con el fin de unificar la medida a horas.

Con los procedimientos anteriores, se obtienen tres bases reducidas del siguiente tamaño: usuarios con 6 variables y 125.225 registros, egresos con 21 variables y 4.103 registros y crónicos con 22 variables y 37.489 registros.

Después de tener cada una de las bases de datos limpias y uniformes, se realiza la unión de la información para obtener una base de datos general que condensa los datos de los pacientes. La unión se realiza por medio de la función "merge" bajo el método "inner" el cual lleva a cabo una intersección de los datos de las bases para extraer la información contenida en los tres conjuntos. La unión se hizo por medio de la variable "nrodoc", de manera adicional, se agregaron las claves "mes" y "año" con el fin de obtener al menos un registro por paciente para cada mes de los años informados. Teniendo en cuenta lo anterior, se obtiene una base de datos general con 43 variables y 777 registros, adicionalmente, se filtra la modalidad de contrato para seleccionar solo aquellos registros pertenecientes a PGP, obteniendo 758 registros y 42 variables.

2.3.ANÁLISIS EXPLORATORIO

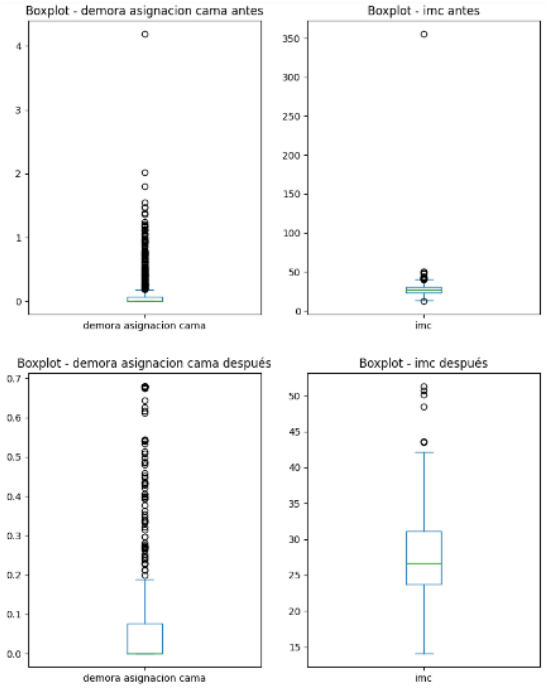
Después de realizar un análisis para comprender las características de los pacientes y de las variables de la base de datos general, se obtiene que:

- Los registros de edad presentan una media de 74 años y una mediana de 75 años. Además, existen registros de personas entre los 20 y 100 años.
- El IMC presenta registros entre 12.4 y 355, sin embargo, el 75% de los datos son iguales o inferiores a 30.22, lo que puede indicar un posible error de digitación en el registro del indicador para el registro máximo mencionado.
- La mayoría de los registros son de servicio habilitado general adultos, con el 75.46% de los 758 datos totales.
- El 68.07% de los 758 datos de la base ingresan a la unidad estratégica de internación por medio de urgencias.
- El 75.33% de los registros son dirigidos a la unidad estratégica de hospitalización adultos.
- La mayoría de los pacientes (90.24%) que salen de la clínica son debido a la alta médica.
- El 68.47% de los registros de la base de datos, presentan atención de un profesional de la especialidad de medicina interna.
- El 58.84% de registros obtenidos son pacientes de sexo femenino.
- El 45.12% de registros presentan una clasificación de clase funcional 4, es decir, presentan un estado de frágil según la prueba de fragilidad de Gröningen y son remitidos a atención domiciliaria.
- La estancia hospitalaria presenta un rango de variación entre 0.497 y 1395.667 horas, con un promedio de 157.987 horas. El 75% de los pacientes registran una estancia igual o inferior a 194.103 horas. Esta variable objetivo presenta una asimetría hacia la derecha en su distribución.
- Las variables demora en asignación de cama, demora en salida de la clínica, IMC, presión arterial diastólica, índice metabólico, máxima cantidad de oxígeno, hemoglobina glicada, lipoproteína, hormona estimulante de la tiroides (tsh) presentan valores extremos atípicos, que pueden deberse a errores de digitación o registro.
- Se evidencian niveles de correlación altos entre el IMC y el índice metabólico con un valor de -0.71, también entre el IMC y la máxima cantidad de oxígeno con un valor de -0.72, asimismo, entre el colesterol total y el hdl con un valor de 0.66 y entre el colesterol total con los triglicéridos de 0.57. De manera adicional, se encuentran dos exámenes de creatinina con una correlación media de 0.51
- Se observa una correlación directa fuerte entre la estancia hospitalaria y la demora de aplicación de medicamentos, con un índice de correlación de 0.87

Posteriormente, se identifican variables explicativas correlacionadas que podrían generar problemas de colinealidad, por tal motivo, se elimina una de las dos y permanece aquella que presente mayor

correlación con la variable objetivo. Por otra parte, se identifica que el 68,47% de los registros ha consultado la especialidad de medicina interna. Por tal motivo, se selecciona este grupo de pacientes para desarrollar la solución analítica, debido a que constituyen una mayoría que permite empezar a distribuir los recursos de una manera más efectiva. Teniendo en cuenta lo anterior, se obtiene un base de datos con 34 variables y 519 registros.

Para el tratamiento de los datos atípicos asociados a errores de digitalización o registro en las variables demora en asignación de cama, IMC, saturación de oxígeno, lipoproteína, microalbuminuria, TSH, se hace una imputación de valores extremos (aquellos que tienen valores superiores a 8 veces el rango intercuartílico de cada variable) con la finalidad de evitar sesgos en los modelos analíticos Para la imputación se toma el valor máximo o mínimo de los registros sin tener en cuenta los atípicos, y se reemplazan dichos valores extremos, se hace la imputación usando este método con la finalidad de no afectar las características que representan los registros de los pacientes. Después de realizar la imputación de datos para las variables mencionadas, se obtienen los resultados de las siguientes figuras.



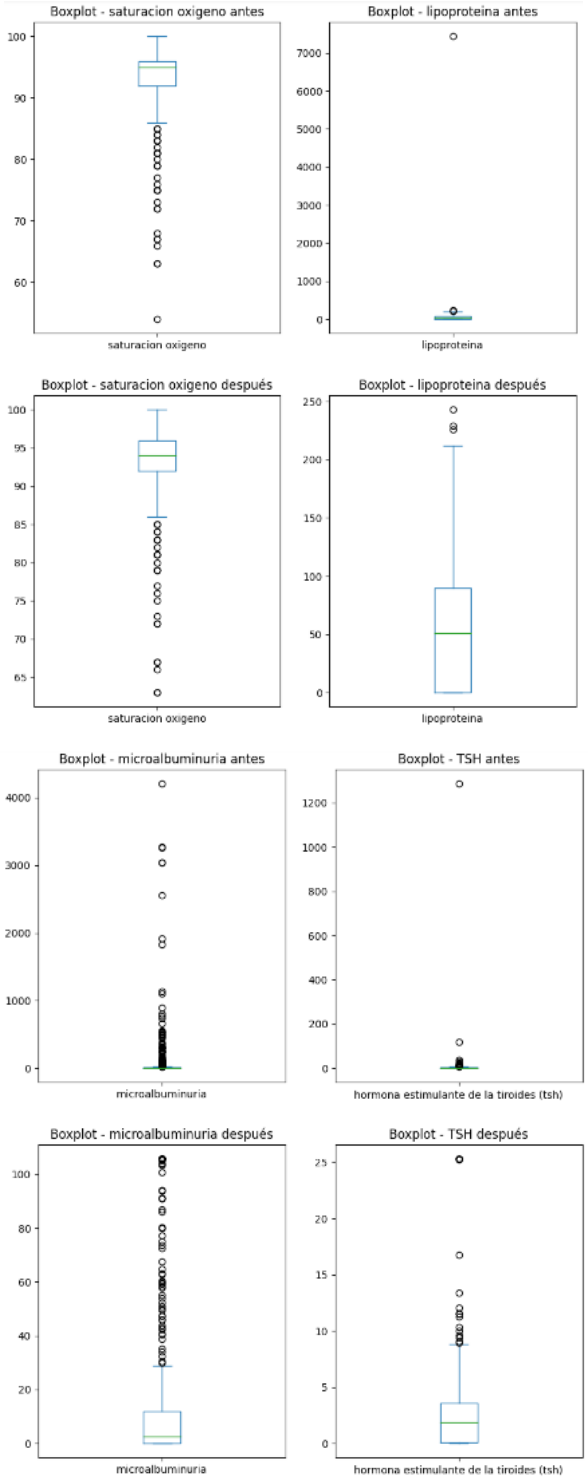


Figura 2. *Tratamiento de datos atípicos.*

Posteriormente, se observa que las variables explicativas presentan niveles de correlación inferiores a 0.52 absoluto, eliminando problemas de colinealidad en el desarrollo de la solución.

2.4.SELECCIÓN DE ALGORITMOS Y VARIABLES

Teniendo en cuenta que en la base de datos general pueden existir registros del mismo paciente, se descarta el uso del algoritmo de regresión lineal múltiple, debido a los problemas asociados a la teoría de construcción de este, pues este método plantea que los datos deben ser independes.

Por otro lado, teniendo en cuenta la importancia de la interpretabilidad de la solución analítica, se plantea el uso de los siguientes algoritmos iniciales para evaluar su desempeño: árbol de decisión, bosque aleatorio y Extreme Gradient Boosting (XGB).

Antes de la construcción de los modelos, se eliminan aquellas variables que aportan información administrativa pero no generan valor en la solución analítica, tales como nrodoc, mes y año. Asimismo, se evidencia que las variables de diagnóstico principal y el número de cama están constituidos por más de 200 categorías, lo cual, aumenta la dificultad en la construcción del modelo y, por lo tanto, se eliminan de la base de datos.

Posteriormente, se hace una conversión de las variables categóricas a dummies con el fin de tratarlas como variables booleanas que permitan obtener un mejor desempeño de los modelos, asimismo, se escalan las variables numéricas con el fin de normalizar los datos y evitar sesgos en los resultados.

Teniendo en cuenta lo anterior, se construyen los modelos descritos para dos condiciones especiales: usando todas las variables de la base de datos general o usando el método de selección de variables bajo la función “SelectFromModel” y una significancia de 0.3 veces la media para reducir la cantidad de variables que aportan información a los modelos, de tal manera que se facilite su interpretación.

Para evaluar el desempeño de los modelos construidos, se tienen en cuenta cuatro métricas de desempeño: el MAE para conocer el promedio de los errores en la predicción y para abarcar los resultados de valores atípicos, el MSE para dimensionar la variabilidad de los residuales de las predicciones, el RMSE para dimensionar la desviación estándar de los errores de las predicciones y el MAPE para identificar el porcentaje de error en las predicciones. En el análisis, se le da prioridad a la última métrica mencionada, debido a que permite obtener una mayor interpretabilidad de los resultados de los modelos.

Para la construcción de los modelos, se tiene en cuenta una división de datos de 80% entrenamiento y 20% validación, debido a que solo se tienen 519 registros y el modelo debe tener la mayor cantidad de ejemplos posibles para realizar una buena predicción. Teniendo en cuenta lo anterior, se obtienen las siguientes métricas de desempeño en validación para los modelos inicialmente construidos.

Método	**Sin selección de variables (60 variables)**			**Con selección de variables (14 variables)**		
Modelo	**Árbol**	**Bosque**	**XGB**	**Árbol**	**Bosque**	**XGB**
MAE [h]	34,1	22,8	26,7	32,7	23,6	27,9
MSE [h^2]	4243,8	1431,2	1631,5	4073,6	1617,7	1936,3
RMSE [h]	65,1	37,8	40,4	63,8	40,2	44,0
MAPE[%]	29,1	24,0	29,6	29,5	24,1	30,5

Tabla 1. *Desempeño de los modelos sin selección y con selección de variables (test).*

En los resultados anteriores, se evidencia que, a pesar de la reducción significativa en la cantidad de variables para la construcción del modelo, las métricas obtenidas son similares en la mayoría de los modelos. Además, se observa que, para el árbol de decisión y para el bosque aleatorio el MAPE presenta mejores resultados con el método de selección de variables, mientras que, según el MAE, para el bosque aleatorio y el XGB el modelo sin selección presenta mejor desempeño. No obstante, los modelos con variables reducidas facilitan la interpretación.

Entre las 15 variables seleccionadas se encuentran: demora aplicación medicamento, edad, hemoglobina glicada, TSH, IMC, lipoproteína, microalbuminuria, piso urgencias, presión arterial diastólica, presión arterial sistólica, servicio admite_cirugia, servicio habilitado_general adultos, tasa de filtración glomerular tfg, triglicéridos y unidad estrategica_hospitalización adultos.

2.5.SELECCIÓN DEL MODELO

Para seleccionar entre los modelos de árbol de decisión (A), bosque aleatorio (B) y XGB, se tiene en cuenta la interpretabilidad, complejidad y métricas de desempeño.

En primer lugar, se lleva a cabo un análisis de validación cruzada para observar el promedio de cada una de las métricas al variar los datos de entrenamiento y validación, obteniendo los resultados de la tabla y la figura mostradas a continuación.

	Entrenamiento			Validación		
Modelo	**Árbol**	**Bosque**	**XGB**	**Árbol**	**Bosque**	**XGB**
MAE [h]	0,26	10,89	15,24	32,67	23,64	27,91
MSE [h^2]	7,73	1017,09	595,80	4073,59	1617,74	1936,31
RMSE [h]	2,78	31,89	24,40	63,82	40,22	44,00
MAPE[%]	0,25	10,27	18,12	29,53	24,06	30,51

Tabla 2. *Desempeño de los modelos con selección de variables.*

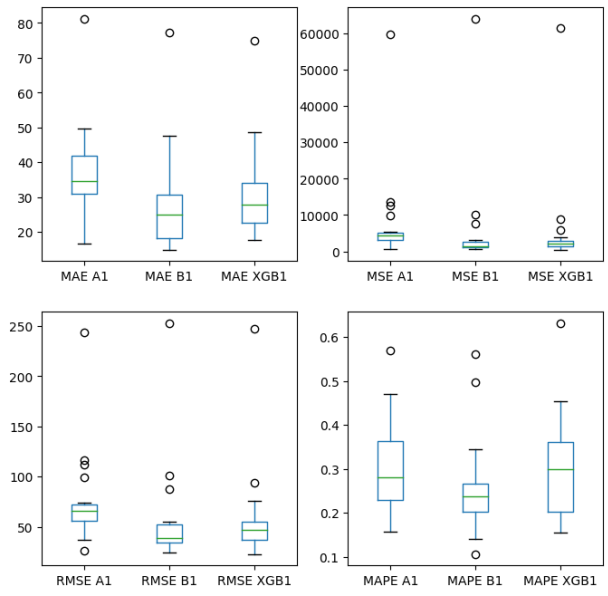


Figura 3. *Desempeño de los modelos en gráficos de bigotes.*

En los resultados anteriores, se observa que, en todos los modelos se presenta un sobreajuste en los datos de entrenamiento que puede ser mejorado por medio del afinamiento de hiperparámetros. Basados en las métricas de validación, se observa que el modelo XGB presenta el peor MAPE y de manera adicional, es un modelo con poca interpretabilidad debido a las bases teóricas de construcción de este, por tal motivo, es descartado para la solución analítica.

Por otra parte, se evidencia que el árbol de decisión presenta peores métricas de desempeño que el bosque aleatorio para todos los indicadores analizados, adicionalmente, se evidencia mayor sobreajuste en este modelo, y, por ende, se selecciona el bosque aleatorio para continuar con el desarrollo de la solución analítica.

2.6.AFINAMIENTO DE HIPERPARÁMETROS

Para mejorar el modelo seleccionado, se hace un afinamiento de hiperparámetros por medio del método de búsqueda aleatoria. La cuadricula de hiperparámetros implementada para la búsqueda contiene los siguientes atributos:

- Profundidad máxima de los árboles que conforman el bosque (“max_depth): con variaciones entre 15 y 40 niveles en pasos de 5.
- Máximo de nodos hoja en los árboles que conforman el bosque (“max_leaf_nodes): con variaciones entre 300 y 600 hojas en pasos de 50.
- Número de estimadores/árboles que conforman el bosque (“n_estimators”): con variaciones entre 25 y 200 estimadores en pasos de 25.
- La función para medir la calidad de división de la rama de los árboles (“criterion”): squared_error, absolute_error, Friedman_mse, poisson.
- Cantidad de características a considerar al buscar la mejor división (“max_features”): 35, auto, sqrt, log2, none.

- El número mínimo de muestras necesarias para estar en un nodo hoja de los árboles del bosque ("min_samples_leaf"): con variaciones entre 2 y 20 muestras en pasos de 2.

Con dicha cuadricula, se afinan los hiperparámetros teniendo en cuenta 20 iteraciones, la optimización de la métrica MAPE y una división de validación cruzada de 30, obteniendo los siguientes resultados:

Hiperparámetros	Bosque aleatorio (15 variables)
n_estimators	150
criterion	absolute_error
max_depth	35
max_leaf_nodes	350
max_features	auto
min_samples_leaf	14

Tabla 3. *Afinamiento de hiperparámetros (Parámetros seleccionados).*

Con base a los resultados anteriores, se construye el bosque aleatorio, obteniendo las siguientes métricas de desempeño para una distribución de datos 80% entrenamiento – 20% validación.

Etapa	Entrenamiento	Validación
MAE [h]	30,39	28,46
MSE [h²]	8653,41	3982,78
RMSE [h]	93,02	63,11
MAPE [%]	21,99	24,07

Tabla 4. *Métricas del modelo con hiperparámetros.*

Los resultados obtenidos, muestran que con el afinamiento de hiperparámetros no se evidencian problemas de sobreajuste. De manera adicional, se evidencia un mejoramiento general en las métricas de desempeño del modelo construido.

2.7.ANÁLISIS DEL MODELO

Con el modelo construido, se evalúan las métricas por medio de la validación cruzada con un divisor CV de 100, y se obtienen los resultados de la Tabla 5 y la Figura 3.

Etapa	Entrenamiento	Validación
MAE [h]	27,79	31,49
MSE [h²]	7373,87	8090,40
RMSE [h]	85,84	51,71
MAPE [%]	19,21	22,70

Tabla 5. *Desempeño del modelo bosque aleatorio.*

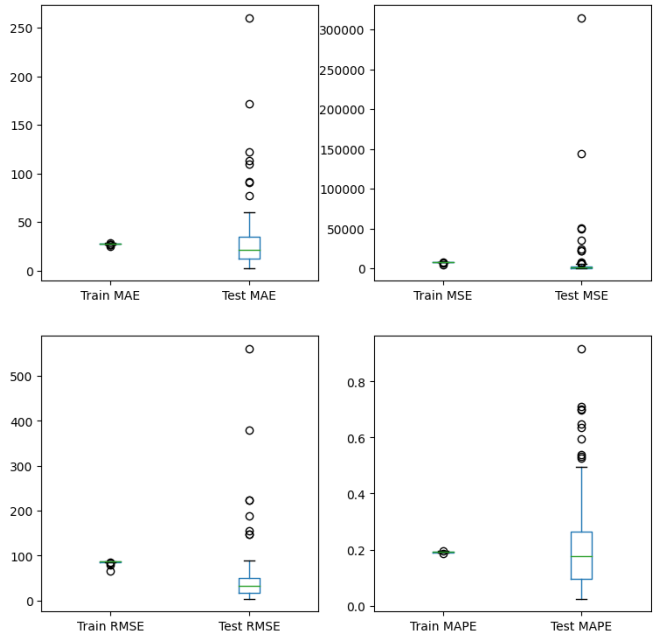


Figura 4. *Desempeño del modelo bosque aleatorio en gráficos de bigotes.*

En los resultados se evidencia un desempeño del modelo con tasas de entrenamiento y validación cercanas y una distribución de las métricas acorde a los resultados esperados. No se evidencia sobreajuste debido a que las métricas son similares en entrenamiento y validación.

No obstante, se evidencia que el modelo tiene un error de predicción cercano al 22%, lo que equivale a un promedio de error de 31 horas de la estancia hospitalaria, asimismo la variación del error en el cálculo de la variable objetivo es significativa, lo cual se refleja en una desviación estándar de 51 horas aproximadamente.

2.8.DESPLIEGUE DEL MODELO

Para la implementación del modelo construido para la estancia hospitalaria, se tienen en cuenta los siguientes aspectos:

- El modelo se entrenará con los 12 meses anteriores para predecir el mes 13 y de ahí en adelante, la actualización de los datos será de forma mensual, tomando los 12 meses anteriores con el nuevo mes incluido que contiene la información de los pacientes del mes anterior y

la información de estancia del mes anterior, con el fin de hacer predicciones para el próximo mes para los futuros pacientes de medicina interna con Pago Global Prospectivo.

- Se entregará un reporte mensual con las variables que más impactan en la estancia hospitalaria, de manera que se construyan planes de acciones preventivos enfocados a dichas variables y se planteen políticas internas para el mejoramiento de estas, por lo menos con aquellas variables que son de influencia directa del hospital.

3. RESULTADOS

Para el período de evaluación del modelo, se obtienen el siguiente nivel de importancia de las variables:

Variable	Importancia
demora aplicacion medicamento	85,38%
piso_Urgencias	0,9836%
servicio habilitado_General adultos	0,0717%
unidad estrategica_Hospitalizacion Adultos	0,0686%
lipoproteina	0,0614%
microalbuminuria	0,0490%
trigliceridos	0,0440%
hemoglobina glicada	0,0398%
imc	0,0356%
tasa filtracion glomerular tfg	0,0322%
hormona estimulante de la tiroides (tsh)	0,0232%
edad	0,0218%
presion art sistolica	0,0197%
presion art diastolica	0,0102%
servicio admite_Cirugia	0,0001%

Tabla 6. *Peso de las variables importantes*

4. CONCLUSIONES

De los resultados obtenido, se evidencia que:

- Es necesario establecer políticas de priorización para los medicamentos, puesto que desde la etapa de exploración se evidencia que el 25% de los usuarios pueden tener demoras iguales o superiores a 2.6 horas para la aplicación de medicamento.
- El piso de urgencia puede dar indicios que el hospital hace una clasificación según la evaluación de los pacientes y esta clasificación puede afectar su estancia.
- Los adultos en hospitalización es de los servicios que más presta la IPS Alma Máter, además que por el hecho de ser hospitalización los tiempos de estadía pueden ser muy largos.
- La edad es otra variable de influencia debido que desde el análisis exploratorio se encontró que en promedio los usuarios tienen 74 años (personas de la tercera edad), lo cual justifica la variable anterior.
- Se puede observar que otra variable de influencia en este caso es el servicio de cirugía, aunque para la cantidad de variables esta no presenta un peso tan significativo, puede ser una variable de estudio para analizar la estancia hospitalaria.
- Por último, se dejan las variables que tienen que ver con el estado de salud, estas variables indican estados puntuales de los pacientes, en estas variables no se tiene una influencia directa, sin embargo, se deben pesar en alternativas para mantener estas variables lo más estables posibles.
- Es importante destacar que, dadas las condiciones particulares de las personas y la complejidad de los datos, es difícil encontrar un modelo que explique en su totalidad la estancia hospitalaria, sin embargo, al haber similitud en los tipos de pacientes se pueden pensar en estrategias para que la estancia hospitalaria no sea tan prolongada y que contribuyan a prestar un mejor servicio.
- Aunque el modelo construido presente porcentajes de error promedio de 22%, se logra identificar una de las variables que impactan de manera significativa la estancia, lo cual ayuda a plantear acciones de mejora.

5. REFERENCIAS

[1]. *Ceballos-Acevedo T, Velásquez-Restrepo PA, Jaén-Posada JS. Duración de la estancia hospitalaria. Metodologías para su intervención. Rev. Gerenc. Polít. Salud. 2014; 13(27): 274-295. http://dx.doi.org/10.11144/Javeriana.rgyps13-27.dehm*