ANALÍTICA EN OPERACIONES - SALUD

Aplicaciones en analítica

Sistema para predecir el uso mensual de recursos traducido en hospitalización de acuerdo con la clase funcional a la que pertenece el paciente.

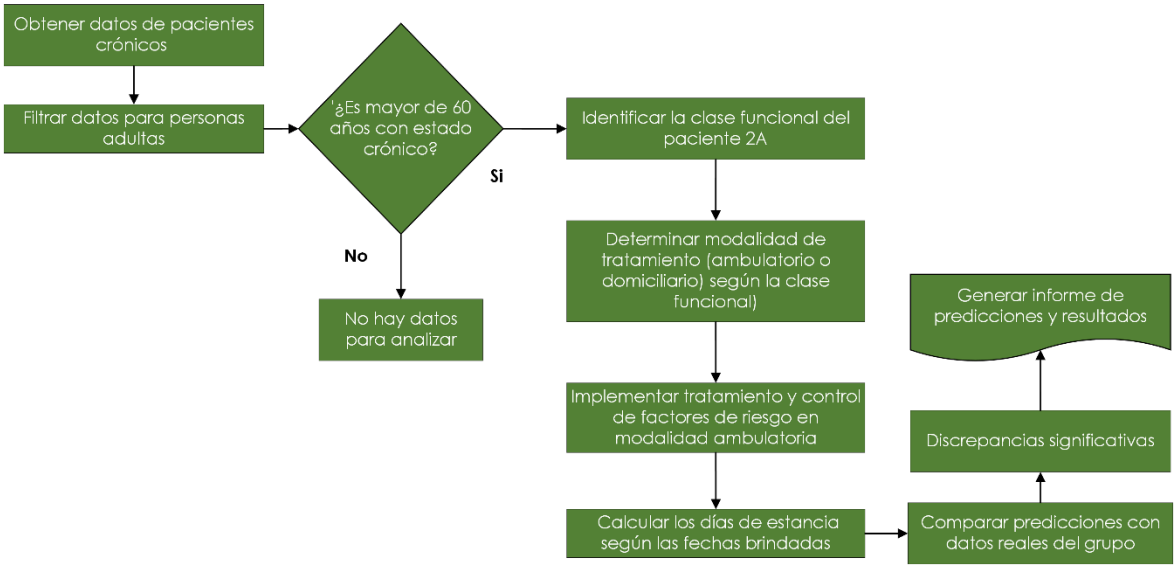
JAVIER BURGOS
CRISTHIAN ALEJO
SUSANA BARRIENTOS

UDEA

INTRODUCCION

El siguiente trabajo expondrá una propuesta de un modelo de machine learning, cuyo objetivo será la predicción del recurso de estancia hospitalaria para la población objetivo (población mayor de 60 años, crónica, asignada bajo la modalidad PGP del Hospital Alma Mater de Antioquia). Este modelo se desarrolla para optimizar la asignación de recursos y mejorar la calidad de la atención para este grupo poblacional.

1. DISEÑO DE SOLUCIÓN PROPUESTO



2. Preprocesamiento

Se inició con una exploración básica para entender la estructura de las bases de datos **egresos**, **usuarios**, y **crónicos**, utilizando métodos como **.info()** y **.shape** para revisar la cantidad de entradas y el tipo de datos en cada columna. Esta etapa fue crucial para identificar columnas con datos faltantes o nulos e identificar las columnas con un nivel informativo bajo (es decir que no se encuentran mucha variabilidad de datos).

2.1. Preprocesamiento de Datos

Al revisar las bases de datos se observa que hay datos repetidos y columnas con tildes en sus nombres, lo que dificulta la manipulación y comprensión de los datos; se cambiaron las vocales identificadas con tildes o caracteres especiales por vocales simples, y la unificación de variables repetidas, se eliminaron datos atípicos y se conservaron solo los del año 2021, ya que son los datos recientes. Finalmente se filtraron los datos según la población objetivo, es decir pacientes mayores

a 60 años con modalidad de contrato PGP y que fueran hospitalizados, este último filtro se obtuvo a partir de las clases funcionales y el ámbito dado por el médico.

Luego de filtrar y limpiar las bases de datos se realizó la unificación de estas mediante “Inner joins” teniendo como llave principal el número de documento del paciente (NRODOC), luego de unificar de las bases se realizó otra limpieza, eliminando las columnas con un porcentaje de nulos mayor al 90%.

Luego de esta segunda limpieza se extrajo la variable objetivo (DIAS HOSPITALIZADO) a partir de la resta de las fechas de ‘Fecha ingreso servicio’ y ‘Fecha alta médica’.

2.2. Exportación de Datos

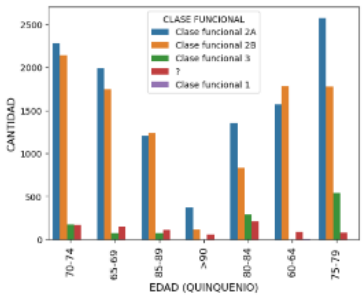
El DataFrame final, preparado y limpio, se exportó a un archivo CSV para facilitar su uso en análisis posteriores o compartirlo con otros sistemas o partes interesadas.

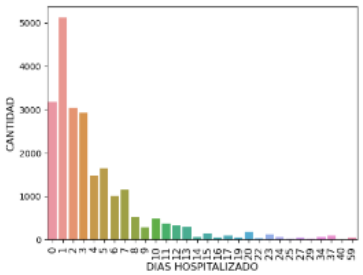
Este proceso de limpieza y transformación de datos es importante para asegurar la integridad y utilidad de los datos en análisis de salud. Se han aplicado criterios de selección y filtrado rigurosos para centrarse en los subconjuntos de datos más relevantes y actuales, mientras que la transformación y normalización de formatos garantiza la coherencia a través de las bases de datos combinadas. Este enfoque meticuloso mejora significativamente la calidad de los insights que pueden ser extraídos en análisis futuros, ayudando a soportar decisiones basadas en datos en el ámbito de la atención médica.

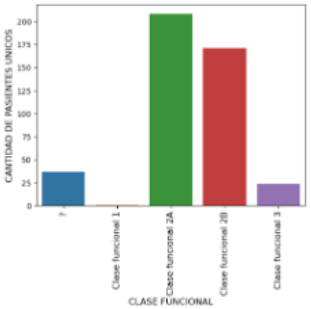
3. ANÁLISIS EXPLORATORIO

3.1. IDENTIFICACIÓN DE COLINEALIDAD

Se tienen 396 pacientes únicos en los datos

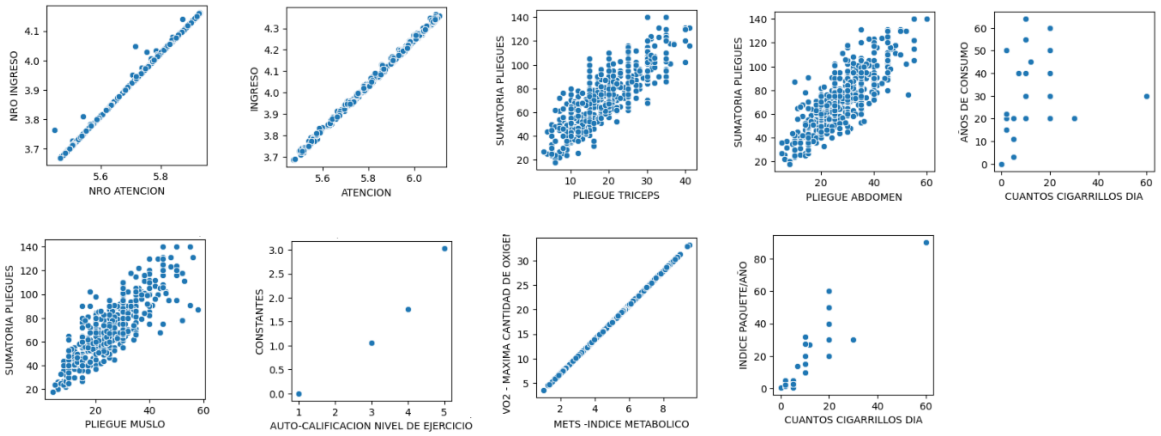






Basándonos en la cantidad de usuarios en cada clase funcional, se ha decidido continuar trabajando solo con los datos asociados a la clase funcional 2. Esta decisión se toma en función de la cantidad de datos, ya que tiene la mayor cantidad de datos, es decir, hay más información disponible para analizar, permitiendo una comprensión más completa y la construcción de modelos más precisos, y al trabajar solo con los datos de clase funcional 2, se reduce el riesgo de sesgos y errores en el

análisis. Esto se debe a que los datos de esta clase funcional son más homogéneos y provienen de una fuente más confiable.



Adema el análisis mediante matriz de correlación y chi-cuadrado se reveló la existencia de correlaciones extremadamente altas entre ciertas variables. Eliminando estas variables permitirá reducir la multicolinealidad y mejorar la estabilidad del modelo, disminuir la dimensionalidad del conjunto de datos, simplificando el análisis y la interpretación y la posibilidad de obtener un modelo más parsimonioso y con relaciones más claras entre las variables restantes. Por lo tanto, se elimina las siguientes variables redundantes:

VARIABLES	CORRELACIÓN
NRO ATENCION y NRO INGRESO	99.61%
ATENCION y INGRESO	99.90%
PLIEGUE TRICEPS y SUMATORIA PLIEGUES	83.49%
PLIEGUE ABDOMEN y SUMATORIA PLIEGUES	83.41%
PLIEGUE MUSLO y SUMATORIA PLIEGUES	88.46%
AUTO-CALIFICACION NIVEL DE EJERCICIO y CONSTANTES	99.05%
METS -INDICE METABOLICO y VO2 - MAXIMA CANTIDAD DE OXIGENO	100.00%
CUANTOS CIGARRILLOS DIA y AÑOS DE CONSUMO	82.64%
CUANTOS CIGARRILLOS DIA y INDICE PAQUETE/AÑO	87.32%

3.2. DEFINICIÓN DE VARIALES

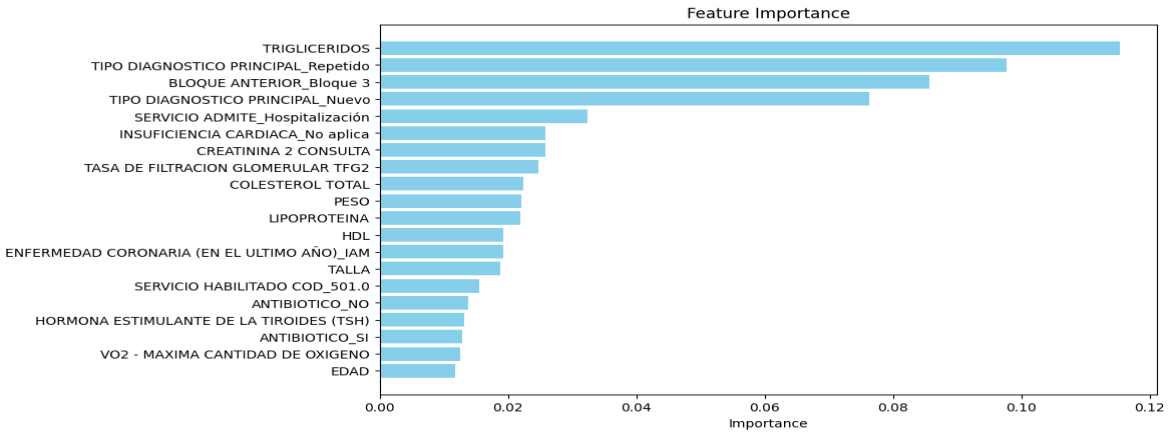
Se define la variable objetivo 'DIAS HOSPITALIZADO', así mismo se selecciona y limpia las variables relevantes, eliminando redundancias. Por su parte la conversión de las variables categóricas en numéricas mediante dummies es fundamental para continuar con el desarrollo del proceso, también se normalizó las variables numéricas para garantizar una escala comparable.

Este enfoque mejorará la precisión del análisis de correlación, facilitará la interpretación de los resultados y permitirá utilizar el conjunto de datos en los modelos que se propondrán más adelante.

4. SELECCIÓN DE ALGORITMOS Y TÉCNICAS DE MODELADO

4.1. Selección de variables mediante Random Forest

La selección de las variables se tomó mediante la creación de una regresión de RandomForest para obtener la importancia de las variables independientes frente a la variable objetivo, de esta regresión tomamos las 20 variables con mayor importancia para finalmente usarlas en el modelo predictivo.



4.2. Selección de Modelos

Para la selección de modelos predictivos usamos la herramienta cross validation usando como métricas MSE y MAE, y los modelos que comparamos en el cross validation fueron: Decision Tree Random Forest y Gradient Boosting, de los cuales se obtuvieron las medias de sus respectivas métricas.

Modelo	MSE_Average	MAE_Average
decision_tree	37.981117	3.689259
random_forest	31.514575	3.437463
gradient_boosting	27.741217	3.222100

Como se observa en la imagen el modelo que muestra menor error tanto en el MSE como en el MAE es el Gradient Boosting, por ello se lo elige, para trabajar en base a este.

5. AFINAMIENTO DE HIPERPARÁMETROS

Partiendo de la asesoría, se iteró varias combinaciones de los parámetros definidos en este modelo, bajo el criterio de que dichos valores estén simétricamente escalonados entre sí. El cambio a pesar de ser pequeño, si influyo en el desempeño del modelo, como se observa en los resultados a continuación.

Parametros iniciales	Parametros finales
`param_grid = { 'n_estimators': [100, 200, 300, 400], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [None, 3, 5, 10], 'min_samples_split': [2, 5, 8], 'min_samples_leaf': [1, 2, 3], 'subsample': [0.5, 0.7, 1.0] }`	`param_grid = { 'n_estimators': [100, 200, 300, 400], 'learning_rate': [0.5, 0.7, 0.9], 'max_depth': [None, 3, 5, 7], 'min_samples_split': [2, 5, 8], 'min_samples_leaf': [1, 2, 3], 'subsample': [0.5, 0.7, 0.9] }`
MSE = 3.795622495 **MAE** = 1.462416727	**MSE** = 0.5832093268 **MAE** = 0.1609763087

6. EVALUACIÓN Y SELECCIÓN DEL MODELO

Con los parámetros seleccionados anteriormente, se entrenó un nuevo modelo Gradient Boosting y se evaluó su rendimiento en un conjunto de prueba independiente. El modelo final obtuvo un MSE de 0.5832093268592665 y un MAE de 0.16097630876541147, (como se muestra en la tabla anterior), una mejora significativa en comparación con el modelo original.

7. CONCLUSIONES

La predicción del uso mensual de recursos de hospitalización para este grupo de pacientes es posible mediante el uso de modelos de aprendizaje automático. El modelo Gradient Boosting con hiperparámetros ajustados demostró ser el más efectivo para esta tarea, obteniendo un menor error absoluto medio (MAE).

Así mismo, las aseguradoras y las instituciones prestadoras de servicios de salud pueden utilizar los resultados de este estudio para optimizar la asignación de recursos y mejorar la atención de los pacientes con clase funcional 2A.

Sin embargo, es importante investigar factores que influyen en la hospitalización de pacientes con clase funcional 3 para identificar posibles intervenciones que puedan reducir el riesgo de hospitalización.