

**PREDICCIÓN DE LA CANTIDAD DE HOSPITALIZACIONES EN EL HOSPITAL ALMA MÁTER**

**CATALINA MARÍA MAYA IREGUI**

**LAURA JASMIN ARIZA ALZA**

**JHAN CARLOS SÁNCHEZ SERNA**

**APLICACIONES DE LA ANALÍTICA**

**OLGA CECILIA USUGA MANCO**

**UNIVERSIDAD DE ANTIOQUÍA**

**FACULTAD DE INGENIERÍA**

**MEDELLÍN**

**2022**

1. **Introducción**

El Hospital Alma Máter de Antioquia recibe pacientes bajo una modalidad de pago global prospectivo (PGP), la cual requiere de una optimización de recursos por parte del hospital con el fin de atender la población de manera eficaz y eficiente. Los pacientes que llegan a diario son clasificados en una clase funcional de manera que se pueda definir una modalidad de tratamiento. A partir de lo anterior se realizará un modelo predictivo con el fin de conocer el uso mensual de recursos traducido en la cantidad de hospitalizaciones. Este trabajo se encuentra dividido en las siguientes secciones: metodología, con todo el proceso del desarrollo del modelo; resultados y conclusiones.

1. **Metodología**

Para la elaboración del trabajo se llevaron a cabo una serie de pasos que hacen parte de la construcción de un proyecto de Machine Learning. En esta sección se encontrará explicativamente lo realizado en el diseño de la solución, la limpieza y transformación de datos, el análisis exploratorio, la selección de variables, la selección de algoritmos y técnicas de modelado, el afinamiento de hiperparámetros, la evaluación y selección del modelo y despliegue del modelo. Toda la codificación y bases de datos se encuentran en [este repositorio](https://github.com/jhansanchez/caso_salud.git) de GitHub.

* 1. **Diseño de la solución propuesto**

Dentro del diseño de la solución se identifica el problema de negocio, el problema de analítica y se esquematiza cómo será la solución para dar respuesta a esos problemas.

* + 1. **Problema de negocio**

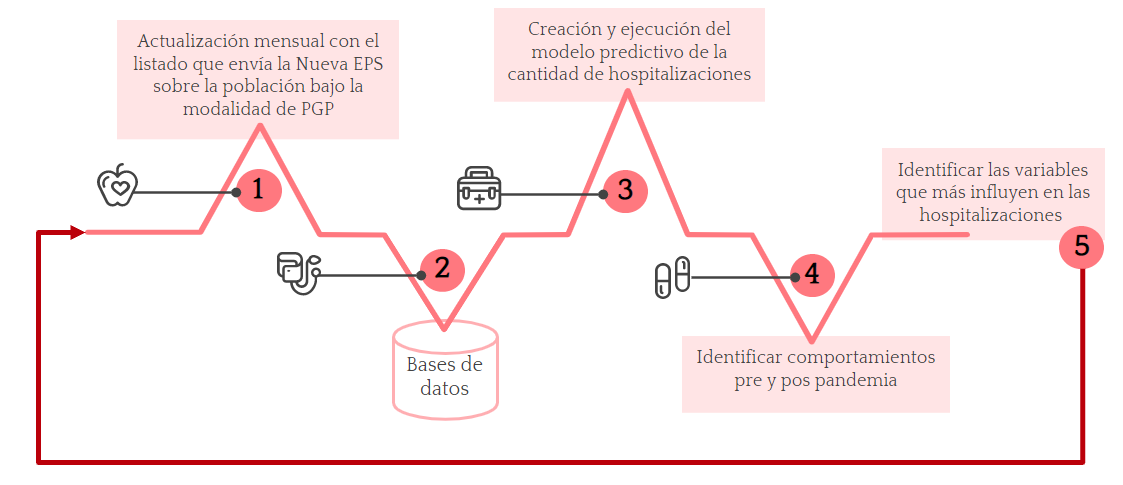
Estimar el uso mensual de recursos del Hospital Alma Máter de Antioquia de acuerdo con la cantidad de hospitalizaciones.

* + 1. **Problema de analítica**

Predecir la cantidad de hospitalizaciones de acuerdo con la clase funcional a la que pertenece el paciente.

* + 1. **Diseño de la solución**

En la Figura 1 se puede observar el diagrama construido para conocer los pasos que se seguirán para llevar a cabo la solución propuesta.



*Figura 1.* Diagrama del diseño de la solución. Fuente: elaboración propia

* 1. **Limpieza y transformación**

Para este caso de estudio donde se va a predecir la cantidad de hospitalizaciones, se inicia con 3 bases de datos: **RETO\_df\_usuarios.csv** (información general de los usuarios del hospital), **RETO\_df\_egresos.csv** (información de las hospitalizaciones de los usuarios), **RETO\_df\_cronicos.csv** (información médica de los usuarios).

En esta sección se separan las bases de datos iniciales para antes y después de pandemia, por lo tanto, quedan 6 bases de datos: cronicos\_pre, cronicos\_pos, usuarios\_pre, usuarios\_pos, egresos\_pre, egresos\_pos. Además, se lleva a cabo una exploración detallada en cada base de datos donde se eliminan variables innecesarias, se realiza tratamiento de nulos, homologación de variables, cambio de tipo de variable y se genera información de acuerdo con otras variables.

* + 1. **Tratamiento de datos nulos**

En la base de datos cronicos\_pre y cronicos\_pos se realizó un filtro donde se eliminan las variables que tienen más de un 30% de nulos debido a que tener tanta cantidad de nulos hace que esas variables no aporten información útil para nuestro estudio. Además, dentro de la medicina no es lógico imputar datos porque las pruebas o intervenciones de cada usuario son únicas por cada visita al hospital, debido a que sus datos cambian en cada ida al hospital por diferentes condiciones de salud del paciente. Por lo tanto, asignar la media, moda, mediana u otro valor que sea calculado estadísticamente con el total de usuarios no tendría sentido en este ámbito. Asimismo, se eliminaron las variables innecesarias y las filas de las variables que se salían de los rangos lógicos que se tienen estipulados en la medicina.

Para la base de datos usuarios\_pre y usuarios\_pos se elimina la variable departamento porque en prepandemia todos los datos son nulos y en pospandemia todos los datos son de Antioquia. Asimismo, para la base de datos usuarios\_pre se llenaron los datos nulos de la edad del usuario por medio de una función, donde se seleccionó de la fecha de nacimiento el mes y el año de nacimiento para realizar una comparación con el año y mes de registro teniendo en cuenta que si el mes de nacimiento era inferior al mes de registro se restaba un año y así poder obtener la edad del paciente. También, para la misma base se llenaron los datos nulos de la variable ciclo vital por medio de rangos teniendo en cuenta los datos que ya estaban en la base y la clasificación del ministerio de salud. Además, se eliminó las variables de fecha, primera y última clase funcional, quinquenio, barrio y municipio debido a tenían varios nulos y no eran significativas. En la base de datos usuarios\_pos se eliminaron las variables de fechas e información de la primera y última clase funcional y el barrio de los usuarios porque no se consideran relevantes. Además, se llenó los datos nulos de la variable municipio con la moda que es Medellín.

Por otro lado, en la base de datos egresos se eliminó las variables que involucran la eps debido a que para las hospitalizaciones no importa esa información al menos que sea para pagar o relacionado a la financiación; en las variables de diagnóstico y causa básica no se deja categoría o capitulo porque mencionan la misma información de diagnóstico de egreso y la causa basuca de muerte; las otras variables se eliminan porque no son importantes para el estudio.

* + 1. **Unión de bases de datos**

Para llevar a cabo la unificación de las bases de datos descritas anteriormente y depuradas, se comenzó con el cálculo de la variable respuesta ‘cantidad de hospitalizaciones'. Para ello se elaboró una función que dejará el registro más actualizado por usuario, puesto que médicamente la información más reciente es la que sirve para la toma de decisiones y los análisis del paciente. Este procedimiento se realiza para los datos prepandemia y pospandemia en donde se obtienen un total de filas y columnas como se muestra en la tabla 1. Es importante resaltar que existe un número diferente de variables en prepandemia respecto a pospandemia puesto que antes del Covid-19 había ciertas pruebas que no se realizaban o datos que no se registraban, pero después de la pandemia fue existiendo la necesidad de llevar un control y toma de datos correctamente.

*Tabla 1.* Total de datos

|  |  |  |
| --- | --- | --- |
| Base de datos | Total usuarios | Total variables |
| Prepandemia | 981 | 61 |
| Pospandemia | 873 | 69 |

* + 1. **Variables y tipos de datos**

Luego de todo el proceso de depuración, las variables que se conservan para las bases de datos unificadas se clasifican en algunas categorías de acuerdo con la funcionalidad que tengan. Las variables con las que se va a trabajar en prepandemia y la categoría se muestran en la tabla 2.

*Tabla 2*. Variables de la base completa prepandemia

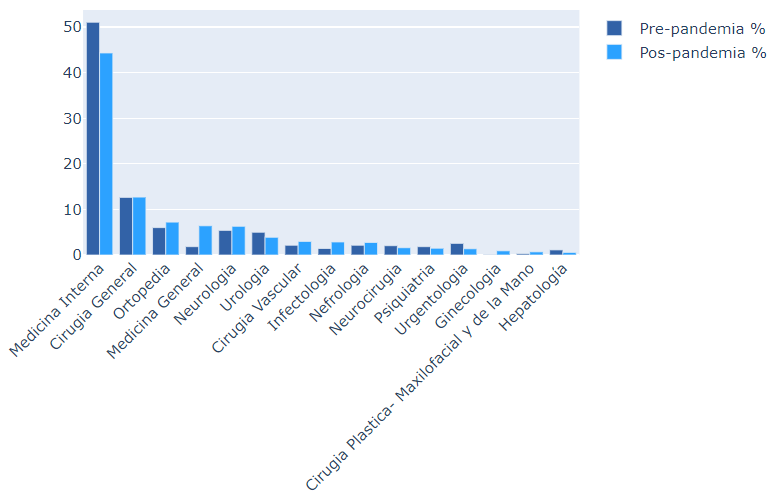
|  |  |
| --- | --- |
| Categoría | Variables |
| Datos generales del usuario | Nrdoc, Sexo, Edad, Ciclo vital, Peso, talla, Clasificación IMC, Clase funcional |
| Relacionados con el tiempo | Year, Mes, Fecha ingreso clínica, Fecha salida |
| Revisiones y controles médicos | Saturación de oxígeno (%), Perímetro muslo, Perímetro cintura, Pliegue triceps, Pliegue abdomen, Pliegue muslo, Sumatoria pliegues, Presión arterial sistólica, Presión arterial diastólica, Frecuencia cardíaca en reposo, Auto-calificacion nivel de ejercicio, Mets -índice metabólico, Vo2 - máxima cantidad de oxígeno, Índice de fragilidad groningen, Calificación (índice de fragilidad), Tiempo en segundos (apoyo monopodal), Calificación (apoyo monopodal), Velocidad (m/s), Calificación velocidad, Test findrisc, Diabetes mellitus, Glicemia, Hemoglobina glicada, Control diabetes, Tiene hta, Tiene epoc, Epoc (clasificación bodex), Enfermedad coronaria (en el último año), Insuficiencia cardíaca, Valvulopatía, Arritmia o paciente con dispositivo, Sufre de alguna enfermedad cardiovascular, Tabaquismo, Lipoproteina, Hdl, Colesterol total, Trigliceridos, Estadio de la enfermedad renal, Creatinina 2 consulta, Tasa de filtración glomerular tfg2, Úlcera de pie diabético, Diagnóstico principal, Ambito según el médico |
| Relacionados con hospitalización | Hospitalizaciones, Tipo egreso, Transfusión sangre, Antibiótico, Dxprincipal egreso, Profesional especialidad |

Para la base de datos de pospandemia se mantienen las mismas variables excepto Test findrisc y se incluyen en los datos generales el municipio y en los datos de revisiones y controles médicos las siguientes variables: Índice tobillo/brazo, Control hta, Tiene riesgo de tener HTA, Cuantos cigarrillos día, Años de consumo, Clasificación de framinghan, Tiene próximo control y Requiere cita de morbilidad.

* 1. **Análisis exploratorio**

En esta sección se analizan los datos por medio de preguntas y se procede a su visualización.

* ¿Cuántos usuarios se atienden por cada grupo de especialidad antes y después de la pandemia (top 15)?

En la *Figura 2* se observa el top 15 por cada grupo de especialidad antes y después de pandemia, donde se evidencia que la medicina interna es la especialidad que tiene mayor cantidad de atenciones de usuarios antes y después de pandemia, esto se debe a que cuando un usuario es hospitalizado la primera revisión se la realiza un médico internista, el cual se encarga de hacer una atención integral, un diagnóstico y tratamiento, y la prevención de las enfermedades; seguido de esa especialidad está la cirugía general y ortopedia. Por otro lado, se identifica que las atenciones en la medicina general tuvieron un aumento relevante después de pandemia debido a que los usuarios pueden tener secuelas o ir a consulta por síntomas de COVID o alguna enfermedad ocasionada o alterada por la pandemia.

*Figura 2.* Cantidad de usuarios por especialidad antes y después de pandemia. Fuente: elaboración propia

* ¿Cuál es la etapa del ciclo vital que más requiere atención hospitalaria?

En la Tabla 3 se muestra las etapas del ciclo vital que más requieren atención hospitalaria antes y después de pandemia, donde se evidencia que la etapa de la vejez es la de más hospitalizaciones, siendo aproximadamente un 90% del total de los datos. Esto se da porque la vejez es la etapa que agrupa mayor cantidad de años y se encuentran las personas que necesitan mayor atención médica debido a las diferentes enfermedades que empiezan a adquirir o ya tienen.

*Tabla 3*. Ciclo vital que más requiere atención hospitalaria en prepandemia y pospandemia

|  |  |  |
| --- | --- | --- |
| Ciclo\_vital | Total prepandemia | Total pospandemia |
| Vejez | 898 | 803 |
| Adultez | 80 | 65 |
| Juventud | 3 | 5 |

* ¿Cuál es el sexo que tiene más atenciones hospitalarias según el ciclo vital?

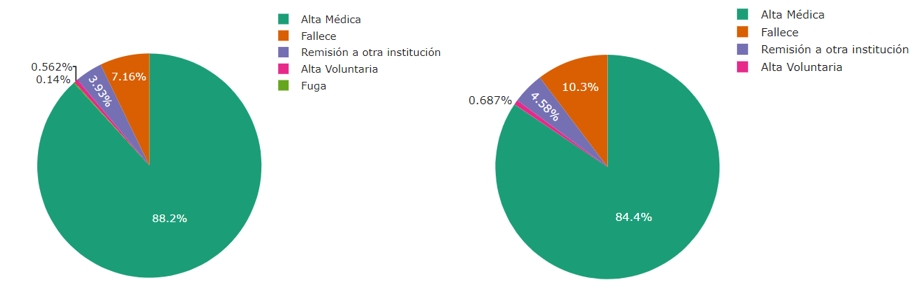
Gráfico, Gráfico de proyección solar

Descripción generada automáticamenteEn la *Figura 3*, se evidencia el sexo según las etapas del ciclo vital donde se observa que sexo femenino es el que tiene más atenciones hospitalarias antes y después de pandemia. Esto sucede porque las mujeres suelen manifestar mayores síntomas, sesión de dolor o presencia de alguna enfermedad que los hombres (Rosa-Jiménez et al., 2022), también se deben realizar chequeos médicos con el fin de prevenir enfermedades específicas que le dan al sexo femenino.

*Figura 3.* Sexo según las etapas del ciclo vital antes y después de pandemia. Fuente: elaboración propia

* ¿Cómo es el comportamiento del tipo de egreso por paciente antes y después de pandemia?

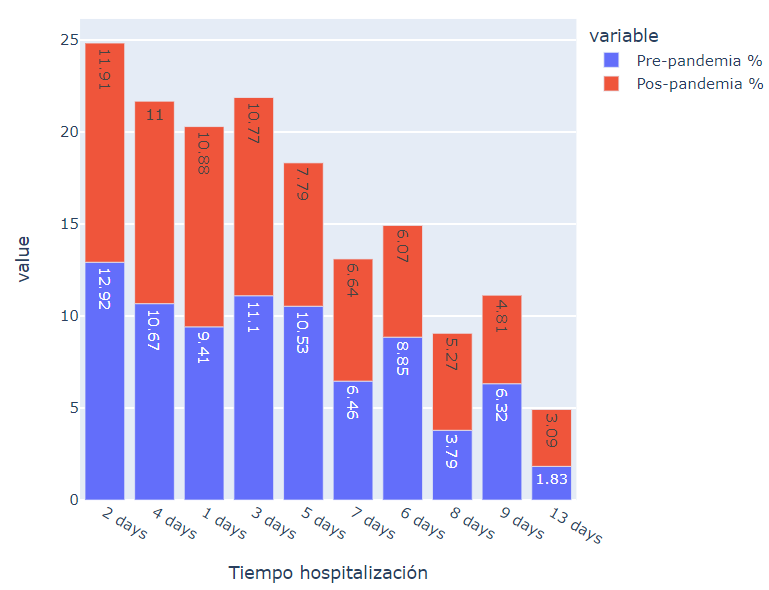
En la *Figura 4* se evidencia el tipo de egreso por paciente, donde se identifica que el alta médica es el egreso que más se da, siendo más de un 84% antes y después de pandemia. Lo anterior indica que se generan mayores salidas del paciente por su excelente o satisfactoria recuperación, ya que los otros tipos de egresos como fallecimiento del paciente, remisión a otra institución, alta voluntaria o fuga para el caso de prepandemia no superan un 16% de las salidas de los usuarios, por lo tanto, la mayoría de los pacientes logra recuperarse de la enfermedad.



*Figura 4.* Tipo de egreso por paciente antes y después de pandemia. Fuente: elaboración propia

* ¿Cuál es el tiempo de hospitalización antes y después de pandemia?

En la *Figura 5* se observa el tiempo de hospitalización de un usuario con un rango de 1 día hasta 13 días, donde se evidencia que la cantidad de días o estancia hospitalaria que mayor frecuentan los usuarios es de 2 días. Sin embargo, los días con mayor porcentaje de hospitalizaciones son de 1 día a 4 días, esto se puede dar porque los pacientes se recuperaron rápidamente. Por otro lado, con relación al tiempo de hospitalización de 13 días, es mayor el porcentaje en pospandemia y esto puede generar por consecuencias o efectos generados por el COVID o alguna enfermedad ocasionada o alterada por la pandemia.



*Figura 5.* Tiempo de hospitalización antes y después de pandemia. Fuente: elaboración propia

* ¿Cuál es la clase funcional antes y después de pandemia?

En la *Figura 6* se muestra la clase funcional, donde se puede evidenciar que antes de pandemia las 2 clases funcionales con mayor porcentaje de hospitalizaciones son la 2B con un 42,2% donde están los pacientes con un estado funcional controlado con uno o varios factores de riesgo que no son controlados y 2A con un 42,8% donde los pacientes tienen un estado funcional adecuado con uno o más factores de riesgo presentes; los pacientes con esas clase funcionales requieren chequeos o controles sobre esos factores de riesgo que se pueden presentar. Por otro lado, después de pandemia la clase funcional con mayor % es la 4 con un 43,2% la agrupa a los pacientes frágiles que se remiten a atención hospitalaria, indicando que la pandemia influencio las enfermedades de los pacientes.

Gráfico

Descripción generada automáticamente

*Figura 6.* Clase funcional antes y después de pandemia. Fuente: elaboración propia

* Histograma cantidad de hospitalizaciones prepandemia y pospandemia

En la *Figura 7* se evidencian los histogramas de la cantidad de hospitalizaciones para antes y después de pandemia, donde en ambos histogramas se observa el mismo comportamiento, un sesgo a la derecha, los datos no son asimétricos y tienen datos atípicos, además, el pico de los datos ocurre alrededor de 1 o 2 hospitalizaciones. Sin embargo, antes de pandemia se muestra una dispersión de los datos desde 1 hasta 16 hospitalizaciones y después de pandemia de 1 hasta 7 hospitalizaciones, siendo menor la cantidad de hospitalizaciones después de pandemia.

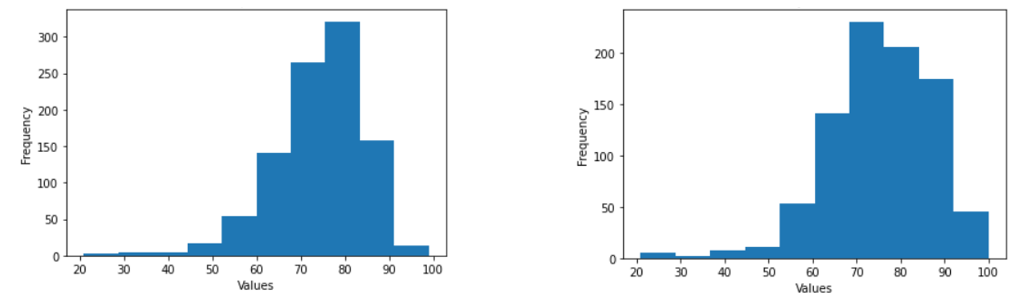
Gráfico, Histograma

Descripción generada automáticamente

*Figura 7.* Histograma cantidad de hospitalizaciones pre y pospandemia. Fuente: elaboración propia

* Histograma de la edad pre y pospandemia

En la *Figura 8* se observa el histograma de la edad donde se evidencia un sesgo a la izquierda y los datos no son asimétricos. Además, para el histograma de antes de pandemia el pico de los datos ocurre en la edad de 75 a 83 años y en el histograma después de pandemia el pico de los datos ocurre en la edad de 70 a 75 años. Igualmente, para ambos histogramas la concentración de los datos se genera en un rango de edad de 60 a 90 años donde hacen parte de la etapa de vejez, siendo esto lógico, debido a que esas personas se encuentran en la tercera edad y son los más propensos de presentar enfermedades y ser hospitalizados.



*Figura 8.* Histograma edad pre y pospandemia. Fuente: elaboración propia

* 1. **Selección de variables**

En esta sección se realiza la descripción del proceso que se llevó a cabo para la selección de variables y aquellas que fueron elegidas para desarrollar el modelo.

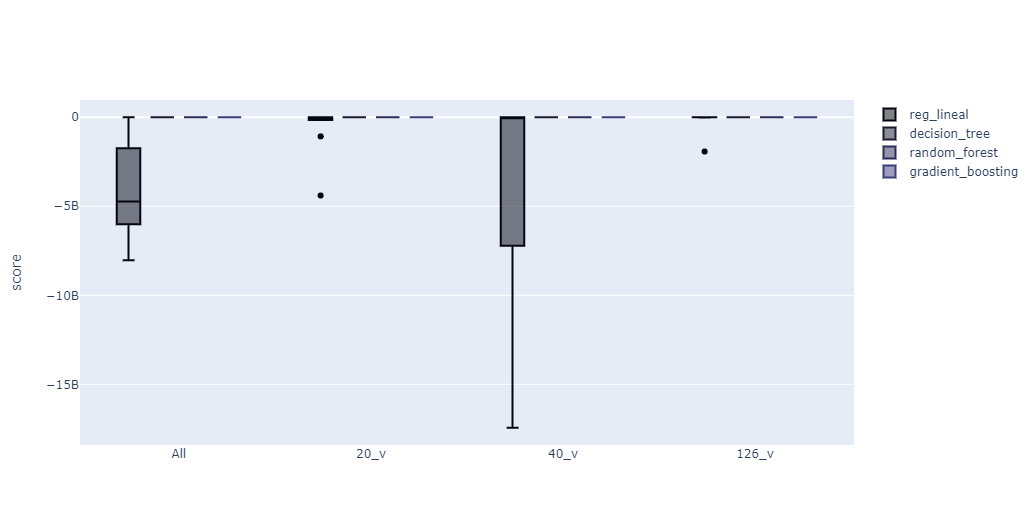
* Construcción de bodega de datos final asignando una clasificación numérica a las variables categóricas ordinales, convirtiendo las variables categóricas nominales a variables dummies y reescalando las variables numéricas. A partir de lo anterior se obtiene un total de 496 variables para prepandemia y 548 para pospandemia.
* Para la selección de variables se utilizan los métodos de ‘SelectKBest’ y ‘SelectFromModel’, construyendo cuatro modelos con diferente número de *features,* tanto para prepandemia como para pospandemia.

*Prepandemia*

A partir de la bodega final y los métodos utilizados se realizan los siguientes modelos:

* **Modelo 1:** teniendo en cuenta lo realizado por (Lo et al., 2021), en donde se desarrolla un modelo predictivo en el área de la salud y se utilizan alrededor de 20 features, se establece un k = 20 por medio del método ‘SelectKBest’.
* **Modelo 2:** se utiliza la función de ‘feature importance’ de Sckit Learn, en donde se obtiene que 40 variables son representativas. Por lo tanto, se utiliza este número para construir un nuevo modelo por medio de ‘SelectKBest’ con k = 40.
* **Modelo 3:** en este caso, se utiliza el método ‘SelectFromModel’ en donde el parámetro *threshold* es establecido con la media, obteniendo un total de 126 variables.
* **Modelo 4:** se utiliza el total de las variables, es decir 496 features.

Continuando con la evaluación de las mejores variables se comparan todos los modelos como se observa en la figura 9, donde los que obtuvieron los mejores valores en todos los casos son Decision Tree, Random Forest y Gradient Boosting. En este sentido, dado que el comportamiento de estos algoritmos es similar con las diferentes combinaciones de variables se decide seleccionar el que contiene 20 features, puesto que entre menos variables mejor el entendimiento y continúa presentando un buen comportamiento.



*Figura 9.* Boxplots modelos y variables - prepandemia. Fuente: elaboración propia

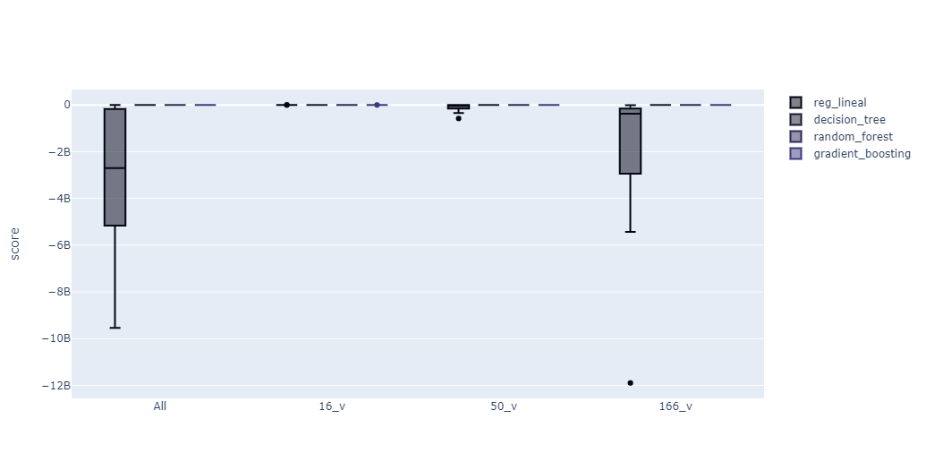
Las 20 variables seleccionadas son: 'Year', 'Antibiotico\_NO', 'Antibiotico\_SI', 'Dxprincipal egreso\_Embolia pulmonar con mencion de corazon pulmonar agudo', 'Dxprincipal egreso\_enfermedad cardiaca hipertensiva con insuficiencia cardiaca (congestiva) ', 'Dxprincipal egreso\_enfermedad cardiorrenal hipertensiva con insuficiencia cardiaca (congestiva)', 'Dxprincipal egreso\_enfermedad pulmonar obstructiva cronica con exacerbacion aguda no especificada', 'Dxprincipal egreso\_enfermedad toxica del higado con cirrosis y fibrosis del higado', 'Dxprincipal egreso\_linfoma de celulas grandes b difuso', 'Dxprincipal egreso\_tuberculosis del pulmon confirmada por hallazgo microscopico del bacilo tuberculoso en esputo con o sin cultivo', 'Profesional especialidad grd\_Cx trasplantes/hepatobiliopancreatrica', 'Profesional especialidad grd\_Hematología', 'Profesional especialidad grd\_Infectologia', 'Control diabetes\_Falta hemoglobina', 'Tiene epoc\_Si', 'Clase funcional\_Clase funcional 3', 'Diagnóstico principal\_J441 - enfermedad pulmonar obstructiva cronica con exacerbacion aguda no especificada', 'Diagnóstico principal\_J450 - asma predominantemente alergica', 'Diagnóstico principal\_J458 - Asma mixta', 'Diagnóstico principal\_M810 - osteoporosis postmenopausica sin fractura patologica'

*Pospandemia*

De la misma manera que en prepandemia, a partir de la bodega final y los métodos utilizados se realizan los siguientes modelos:

* **Modelo 1:** para la construcción del primer modelo que es basado en el criterio del artículo se lleva a cabo el método de ‘permutation\_importance’, para analizar la relevancia de las 20 variables seleccionadas. En este se encuentran cuatro variables con importancia nula, por lo que se decide construir un modelo con 16 variables por medio del ‘SelectKBest’
* **Modelo 2:** para este modelo, igual que en prepandemia, se utiliza la función de ‘feature importance’ de Sckit Learn. En este caso se obtiene que 50 variables son representativas. Por lo tanto, se utiliza este número para construir un nuevo modelo por medio de ‘SelectKBest’ con k = 50.
* **Modelo 3:** igualmente, se utiliza el método ‘SelectFromModel’ obteniendo con el parámetro *threshold = ‘mean’* un total de 164 variables.
* **Modelo 4:** se utiliza el total de las variables, es decir 548 features.

De acuerdo con la figura 10 se puede evidenciar que el desempeño de los algortimos es muy similar al caso presentado anteriormente de prepandemia. Por lo tanto, se decide seleccionar el modelo que contiene 16 features, con la misma justificación del buen comportamiento y menos número de variables.



*Figura 10.* Boxplots modelos y variables - pospandemia. Fuente: elaboración propia

Las 16 variables seleccionadas son: 'Mes', 'Dxprincipal egreso\_ulcera gastrica  aguda con hemorragia ', 'Diagnóstico principal\_T983 - secuelas de complicaciones de la atencion medica y quirurgica no clasificadas en otrasparte', 'Dxprincipal egreso\_otros transtornos del equilibro de los electrolitos de los líquidos no clasificados en otra parte, ‘Dxprincipal egreso\_malabsorcion intestinal  no especificada’, ‘Diagnótico principal\_R104 - otros dolores abdominales y los no especificados’, ‘Dxprincipal egreso\_epilepsia tipo no especificado, 'Ciclo\_vital', 'Dxprincipal egreso\_tumor maligno de la glandula parotida', ‘Diagnóstico principal\_I500 – insuficiencia cardiaca congestiva’, ‘Dxprincipal egreso\_trombocitopenia no especificada’ Dxprincipal egreso\_septicemia  no especificada ', ‘Diagnóstico principal\_F200 – esquizofrenia paranoide’, ‘Diagnóstico principal\_F209 – esquizofrenia no especificada’, ‘Diagnóstico principal\_M059 - artritis reumatoidea seropositiva sin otra especificación ’, ‘Profesional especialidad grd\_Hepatología’

* 1. **Selección de algoritmos y técnicas de modelado**

Para el desarrollo de los modelos predictivos (un modelo prepandemia y otro pospandemia) se decide seleccionar cuatro algoritmos. En primer lugar, Linear Regression, como uno de los estimadores básicos y comunes en un problema de regresión. También, Decisión Tree, como una de las clasificaciones que tiene la librería de Scikit Learn, siendo un algoritmo fácil de interpretar gracias a que no es caja negra. Por otro lado, se encuentra Random Forest y Gradiente Boosting, los cuales se seleccionan para tener en cuenta métodos de ensamble que pueden mejorar el rendimiento y robustez del modelo, puesto que tienen como objetivo combinar las predicciones de varios estimadores base construidos con un algoritmo de aprendizaje dado (Scikit Learn, 2022). En cuanto a la separación de train y test, implementados para la construcción de los modelos, se utiliza un criterio del tamaño del ‘test’ del 33%.

* 1. **Afinamiento de hiperparámetros**

En el camino de seleccionar el mejor modelo para predecir la cantidad de hospitalizaciones es necesario realizar un afinamiento de hiperparámetros por el método de validación cruzada, utilizando los parámetros y valores que se muestran en la tabla 4. Estos mismos parámetros se utilizan para los modelos de prepandemia y pospandemia.

*Tabla 4*. Hiperparámetros

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | | | **Decision Tree** | | |
| **Parámetro** | **Descripción** | **Valores** | **Parámetro** | **Descripción** | **Valores** |
| n\_estimators | Número de árboles incluidos en el modelo | 3, 100, 150, 500 | min\_samples\_split | Número mínimo de observaciones que debe de tener un nodo para que pueda dividirse. | 2, 3, 4, 5 |
| max\_features | Número de predictores considerados en cada división | 5, 7, 9, 20 | min\_samples\_leaf | Número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se produzca la división. | 1, 2, 3 |
| max\_depth | Profundidad máxima que pueden alcanzar los árboles | None, 3, 10, 20 | max\_depth | Profundidad máxima que puede alcanzar el árbol. | None, 3, 10, 20 |

Los mejores hiperparámetros por validación cruzada se muestran en la tabla 5, tanto para prepandemia como para pospandemia y para los modelos de Random Forest y Decision Tree.

*Tabla 5.* Hiperparámetros seleccionados

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random forest** | | | **Decision Tree** | | |
| **Parámetro** | **Pre** | **Pos** | **Parámetro** | **Pre** | **Pos** |
| n\_estimators | 150 | 150 | min\_samples\_split | 2 | 2 |
| max\_features | 5 | 9 | min\_samples\_leaf | 1 | 3 |
| max\_depth | 3 | 10 | max\_depth | 3 | 3 |

* 1. **Evaluación y selección del modelo**

Luego del afinamiento de hiperparámetros se analizan las métricas de entrenamiento y evaluación para verificar inexistencia de subajuste o sobreajuste y comparar los modelos. Para esto se implementa Cross Validate para train y test.

Para comparar los modelos se realiza el score medio de la prueba, los cuales se pueden observar en la tabla 6. Se puede evidenciar que, aunque ambos modelos no presentan una diferencia muy significativa, el Random Forest es el estimador que mejor rendimiento presenta, por lo tanto, es el modelo seleccionado para realizar las predicciones de la cantidad de hospitalizaciones, tanto para prepandemia como pospandemia. Además, se puede identificar que los modelos de pospandemia generan un score mejor (más cercano a cero) que el de prepandemia, siendo entonces más preciso y con mejor rendimiento.

*Tabla 6.* Score medio de prueba de los modelos

|  |  |  |
| --- | --- | --- |
|  | **Random Forest** | **Decision Tree** |
| **Prepandemia** | -1.4453 | -1.4674 |
| **Pospandemia** | -0,7679 | -0.7936 |

Luego de seleccionar el modelo del Random Forest se analiza el sobreajuste o subajuste, por medio de los puntajes de entrenamiento y de prueba (ver tabla 7), donde se evidencia que no existe una diferencia muy relevante entre train y test a partir de la segunda iteración, lo cual confirma el buen rendimiento de los modelos (prepandemia y pospandemia).

*Tabla 7.* Puntajes de train y test de los modelos

|  |  |  |  |
| --- | --- | --- | --- |
| **Prepandemia** | | **Pospandemia** | |
| Train score | Test score | Train score | Test score |
| -0.485373 | -3.733954 | -0.315641 | -1.784055 |
| -1.612641 | -0.391459 | -0.717691 | -0.522609 |
| -1.569619 | -0.967866 | -0.724067 | -0.519530 |
| -1.552626 | -1.087784 | -0.720557 | -0.536656 |
| -1.559954 | -1.045903 | -0.729040 | -0.477013 |

A partir de las figuras 11 y 12, a pesar de que se esperaba encontrar una distribución del error con un comportamiento normal para los dos modelos, la mayoría de los errores están cercanos a cero. Además, se encuentra un sesgo marcado a la izquierda tanto para prepandemia como pospandemia, pero con un sesgo menos marcado para este último. Cabe mencionar, que hay mayor número de errores atípicos, que se alejan de cero para prepandemia, donde hay errores que llegan hasta 15 puntos. Con relación a pospandemia los puntos atípicos no se encuentran tan lejanos y se presentan en menor cantidad. Esto posiblemente ratifique, que el modelo de pospandemia presenta un mejor comportamiento que el modelo prepandemia.

Gráfico, Histograma

Descripción generada automáticamente Gráfico, Histograma

Descripción generada automáticamente

Figura 11. Distribución del error de la cantidad de hospitalizaciones prepandemia

Figura 12. Distribución del error de la cantidad de hospitalizaciones pospandemia.

* 1. **Despliegue del modelo**

El modelo se guarda en un archivo pickle, con el cual es posible realizar las predicciones de las cuales se discutirán en la sección 3 de resultados.

Además, se exporta en un archivo de Excel con los resultados de las predicciones para cada usuario. Este se encuentra dentro del [repositorio de GitHub](https://github.com/jhansanchez/caso_salud.git) donde está el desarrollo del proyecto.

1. **Resultados**

A partir de la selección del modelo Random Forest se realizan predicciones, con las cuales se conoce la cantidad de veces que una persona bajo la modalidad de pago global prospectivo (PGP) será hospitalizada en las instalaciones médicas por medio de la variable ‘Hospitalizaciones’, donde esta es el vector Target dentro del modelo de aprendizaje supervisado construido en este trabajo.

´

Así mismo, se la realizaron 873 predicciones a los usuarios de los servicios médicos para prepandemia y pospandemia. De esta forma, para prepandemia se encontró un promedio de hospitalización de 1.98 y para pospandemia 1.43, lo que podría dar a entender que después de la pandemia del covid-19 existen menos cantidad de hospitalizaciones. Pero si se analiza con mayor detalle, las predicciones para prepandemia indican que para los años 2017 y 2018, el 1.17% de las personas fueron hospitalizadas una o dos veces. A pesar de esto, existen personas que actúan como usuarios atípicos, donde son ingresados por hospitalización unas siete veces, inclusive hasta 10 veces, hecho que genera que el promedio de hospitalización prepandemia tenga un valor un poco mayor. Por otro lado, con relación a pospandemia el 1,60% de las personas fueron hospitalizadas una o dos veces y donde los casos más altos de cantidad de hospitalizaciones son dos personas con cuatro y cinco ingresos. De este modo, se puede concluir que no hay mayor diferencia en la cantidad de hospitalizaciones antes y después de pandemia, donde el promedio de hospitalización prepandemia se debe particularmente a tres casos atípicos que hacen que la media de ingresos a hospitalización aumente.

1. **Conclusiones**

* La limpieza y depuración de las bases de datos fue un proceso exigente dado los múltiples datos nulos y registros erróneos en muchas de las variables suministradas. Por los tanto, las bodegas de datos finales (prepandemia y pospandemia) fueron reducidas en el número de registros para poder realizar los modelos de manera adecuada.
* El modelo seleccionado y que mejor rendimiento tuvo según el score medio del error fue el del Random Forest, tanto para prepandemia como para pospandemia.
* El modelo de prepandemia fue construido con 20 variables y el de pospandemia con 16. En ambos la mayoría de las variables son relacionadas con enfermedades por las que posiblemente hospitalizarían a una persona. Sin embargo, se resalta de pospandemia el ciclo vital y el mes y, de prepandemia, la clase funcional 3 y grupos de especialistas que no se tienen en cuenta en el otro modelo. Además, aunque los modelos de prepandemia y pospandemia implementen variables diferentes, el comportamiento y rendimiento es muy similar.
* Medicina interna es la especialidad que mayor se requiere en el proceso de hospitalización sin importar si es antes o después de pandemia, debido a la necesidad de revisión diaria y asignación de tratamientos.
* El tipo de egreso de alta médica tiene el mayor porcentaje superando el 84% antes y después de pandemia, indicando que la mayoría de los pacientes logra recuperarse de la enfermedad.
* El ciclo vital que más hospitalizaciones presenta es la vejez, puesto que es el rango de edades en el que más enfermedades pueden aparecer y requieren mayor atención.
* Pese a que se esperaba mayor cantidad de hospitalizaciones después de pandemia, debido al aumento del sedentarismo, disminución del nivel de atención hospitalario durante pandemia, estrés, agravamiento de enfermedades por covid-19, entre otros factores y variables; las predicciones del modelo arrojan que no hay mayor diferencia en la cantidad de hospitalizaciones antes y después de pandemia. Esta última afirmación puede ocasionarse dado los avances en tecnología e investigación generados por la pandemia (vacunas, aumento de la capacidad hospitalaria, etc.) y el control que ya se tiene en este ámbito, al punto de generar una atención vía telefónica o chat, para indicar los pasos a seguir. Seguramente en los años donde el mundo se encontraba en plena pandemia esta cantidad de hospitalizaciones eran mucho más altas.
* Por último, modelo predictivo desarrollado puede generar información sobre la cantidad de hospitalizaciones, lo cual es un insumo para el Hospital Alma Máter, dada la estimación del uso mensual de recursos que tendrían de acuerdo con ese número por usuario.

1. **Referencias**

Lo, Y. T., Liao, J. C. hen, Chen, M. H., Chang, C. M., & Li, C. Te. (2021). Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC Medical Informatics and Decision Making*, *21*(1), 1–11. https://doi.org/10.1186/s12911-021-01639-y

Rosa-Jiménez, F., Montijano Cabrera, A. M., Ília Herráiz Montalvo, C., & Zambrana García, J. L.. (2005). ¿Solicitan las mujeres más consultas al área médica que los hombres?. Anales de Medicina Interna, 22(11), 515-519. Recuperado en 05 de septiembre de 2022, de http://scielo.isciii.es/scielo.php?script=sci\_arttext&pid=S0212-71992005001100003&lng=es&tlng=es.

Scikit Learn. (2022). *Ensemble methods*. https://scikit-learn.org/stable/modules/ensemble.html