ENTREGA FINAL DE PROYECTO

Por:

Paula Andrea Gil Vargas

Kevin Manuel Jaimes Ojeda

Alejandro Rivera Pérez

Asignatura:

Introducción a la Inteligencia Artificial

Docente:

Raúl Ramos Pollán

Universidad de Antioquia

Facultad de ingeniería

Medellín 2023

1. Introducción

Descripción del problema predictivo a resolver

Por lo general, los hospitales se enfrentan a un considerable flujo de pacientes, y bajo estas circunstancias, la permanencia prolongada de dichos pacientes se convierte en un parámetro crítico que desencadena la congestión de los servicios de urgencia. Esta situación genera una serie de problemas tanto para el personal médico como para los propios pacientes. En primer lugar, los largos tiempos de espera constituyen una fuente de preocupación y malestar para aquellos que requieren atención médica inmediata. Además, la capacidad del hospital se ve saturada, dificultando la correcta gestión de los recursos disponibles.

Esta congestión también tiene consecuencias negativas en términos de mortalidad. Cuando los pacientes permanecen en el hospital por períodos prolongados, aumenta el riesgo de complicaciones y deterioro de su estado de salud. Esto, a su vez, se traduce en una mayor tasa de mortalidad, lo cual es alarmante y preocupante tanto para los profesionales médicos como para los propios pacientes y sus familias.

Por otro lado, la congestión prolongada en los hospitales tiene implicaciones financieras significativas. Las pérdidas económicas resultantes de esta situación se deben a diversos factores. En primer lugar, se incrementan los costos operativos, ya que se requiere una mayor cantidad de recursos y personal para atender a los pacientes durante períodos prolongados. Además, las demoras en la atención y los tiempos de espera prolongados pueden llevar a una disminución en la satisfacción de los pacientes, lo que a su vez afecta la reputación del hospital y puede resultar en la pérdida de ingresos y oportunidades futuras.

En resumen, la gestión inadecuada del flujo de pacientes en los hospitales y la consiguiente permanencia prolongada de los mismos conlleva numerosas complicaciones. Estos problemas van desde el malestar y la preocupación de los pacientes hasta la saturación de la capacidad hospitalaria, un aumento en la mortalidad y pérdidas financieras significativas. Por lo tanto, es fundamental abordar esta situación mediante estrategias efectivas de gestión y planificación que permitan optimizar los recursos disponibles y garantizar una atención médica oportuna y eficiente.

La mencionada situación plantea la necesidad creciente de optimizar la estancia de los pacientes en las instalaciones hospitalarias con el fin de mejorar la eficiencia de la atención médica. Esto implica facilitar el flujo de usuarios y determinar la gravedad de cada paciente de manera precisa. Por esta razón, se propone la implementación de un modelo predictivo que, basándose en registros hospitalarios, pueda estimar la duración de la permanencia de los pacientes.

2. Exploración descriptiva del Dataset

Para el análisis de nuestro dataset se tienen 2 archivos:

Archivo 1: train.csv

En este archivo se encuentra el tren de datos registrados por los hospitales a los que les aplicará el modelo predictivo

case_id: Número de registro del paciente.

Hospital_code: Código que identifica el hospital.

Hospital_type_code: Código que identifica el tipo de hospital.

Department: Código que identifica la ciudad del hospital

Hospital_region_code: Código que identifica la región del hospital.

Available Extra Rooms in Hospital: Número de habitaciones adicionales disponibles en el hospital.

Department: El Departamento pasa por alto el caso.

Ward_type: Código del tipo de sala.

Ward_Facility_Code: Código de la instalación.

Bed Grade: Estado de la cama en el pabellón.

patientid: Id. único del paciente.

City_Code_Patient: Código de ciudad del paciente.

Type of Admission: Tipo de ingreso registrado por el Hospital.

Severity of Illness: Gravedad de la enfermedad registrada en el momento del ingreso.

Visitors with Patient: Número de visitantes con el paciente.

Age: Edad del paciente.

Admission_Deposit: Depósito en el momento del ingreso.

Stay: Días de estancia del paciente. ➔ Archivo 2: test.csv En este archivo se encuentran los datos de prueba para realizar las predicciones y posterior comparación de los resultados.

Archivo 2: test.csv

En este archivo se encuentran los datos de prueba para realizar las predicciones y posterior comparación de los resultados.

Métricas

La métrica de evaluación de los resultados que se aplicará en el desarrollo del modelo es el error logarítmico cuadrado medio, el cual se define como se muestra a continuación:

$$\sqrt{[\frac{1}{n}]\sum_{i=1}^{n} (log\,(x_i+1) - log\,(y_i+1))^2}$$

La implementación de esta métrica proporcionará un valioso indicador: el error resultante entre el valor predicho x y el valor real y. Es esencial destacar que la expresión mencionada anteriormente generará un resultado mayor si el valor entregado por el modelo es inferior al valor real. Esta característica es adecuada, ya que en los resultados esperados se considera que la sobreestimación es menos perjudicial que la subestimación.

Esta distinción es de vital importancia en diversos contextos, especialmente en aquellos donde se busca minimizar los riesgos asociados a una estimación inexacta. Al evaluar el desempeño del modelo, es preferible que se cometan errores que conduzcan a una sobreestimación del valor real, en lugar de subestimar su magnitud. La razón detrás de esto radica en las implicaciones prácticas de ambos tipos de errores.

Cuando se produce una subestimación, es decir, cuando el modelo predice un valor inferior al valor real, existe un riesgo potencial de no asignar los recursos o las medidas necesarias para abordar adecuadamente la situación. Esto puede conducir a consecuencias negativas, como una atención inadecuada o insuficiente, retrasos en la toma de decisiones críticas o una asignación ineficiente de recursos limitados. En cambio, una sobreestimación tiende a ser menos perjudicial, ya que implica una precaución adicional que garantiza la disponibilidad de los recursos adecuados y la implementación de las acciones necesarias.

Es importante destacar que esta preferencia por la sobreestimación no debe interpretarse como una invitación a realizar estimaciones excesivamente infladas de manera sistemática. El objetivo es encontrar un equilibrio adecuado que permita minimizar los errores y maximizar la precisión de las predicciones. El uso de esta métrica específica, que considera la importancia relativa de la sobreestimación y la subestimación, brinda una herramienta valiosa para evaluar y ajustar los modelos en función de los riesgos y las implicaciones prácticas involucradas.

3. Iteraciones de desarrollo

Desempeño

La culminación de este modelo requerirá contar con la capacidad de prever de manera precisa la duración de la estancia de un grupo específico de pacientes en hospitales, ya sea en entornos con un flujo de personas bajo o alto. Asimismo, se contempla la posibilidad de adaptarlo de manera más amplia para abordar diversas situaciones que involucren el manejo de datos en intervalos de tiempo específicos. Esto permitiría su

aplicabilidad en escenarios variados que demanden una gestión eficiente basada en pronósticos confiables.

Exploración del dataset y preprocesado

En primer lugar, se realizó el preprocesado y exploración de los datos con el fin de identificar las variables que están en el encabezado del archivo, la cantidad de datos en los dataframe y los datos faltantes que pueda haber en los mismos.

1. En primera instancia se carga la base de datos y se exploran los encabezados de las columnas para el dataframe de prueba y el de entrenamiento. También se explora el tamaño de ambos dataframe.

2. Posteriormente se exploran los datos nulos presentes en ambos dataframes y se procede a eliminarlos.

3. También se realizó la identificación del tipo de variables en los dataframes.

4. Posteriormente se realizó el análisis de las variables categóricas presentes en el conjunto de datos, estas variables categóricas son las posibilidades de estados que pueden tomar cada una de las columnas del conjunto de datos. Para el caso los conjuntos de variables categóricas para cada una de dichas columnas son los siguientes.

```
#Variables categóricas
ccols = [i for i in data_train.columns if not i in data_train._get_numeric_data()]
for c in ccols:
    print ("%10s"%c, np.unique(data_train[c].dropna()))
```

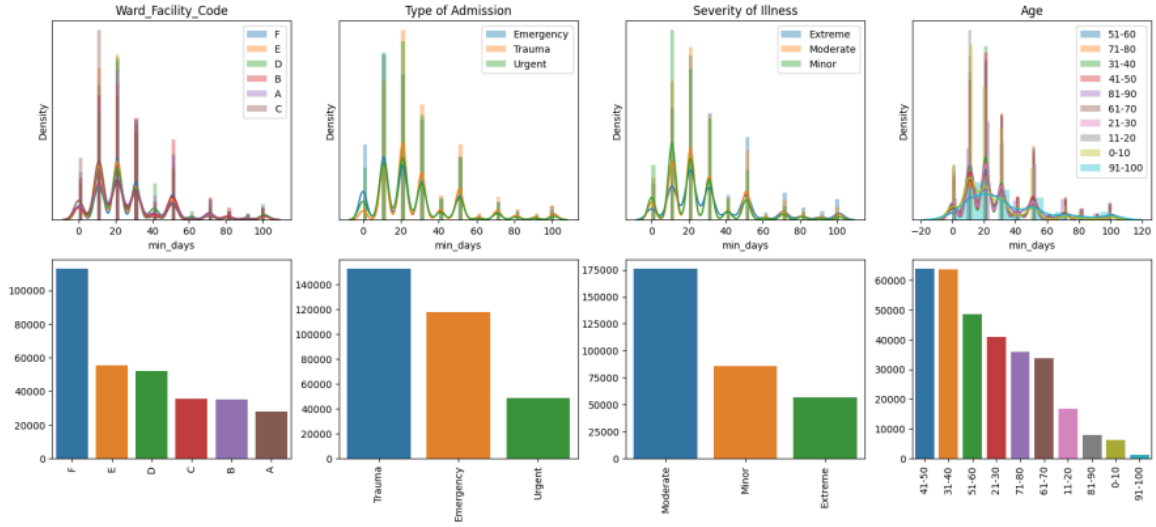
```
Hospital_type_code ['a' 'b' 'c' 'd' 'e' 'f' 'g']
Hospital_region_code ['X' 'Y' 'Z']
Department ['TB & Chest disease' 'anesthesia' 'gynecology' 'radiotherapy' 'surgery']
 Ward_Type ['P' 'Q' 'R' 'S' 'T' 'U']
Ward_Facility_Code ['A' 'B' 'C' 'D' 'E' 'F']
Type of Admission ['Emergency' 'Trauma' 'Urgent']
Severity of Illness ['Extreme' 'Minor' 'Moderate']
       Age ['0-10' '11-20' '21-30' '31-40' '41-50' '51-60' '61-70' '71-80' '81-90'
 '91-100']
     Stay ['0-10' '11-20' '21-30' '31-40' '41-50' '51-60' '61-70' '71-80' '81-90'
 '91-100' 'More than 100 Days']
```

En la imagen anterior se puede observar que variables como el código de hospital toman un conjunto de variables entre las letras a y g, así como el código de región, el tipo de sala y el código de facilidad de sala. También se puede observar variables como el departamento del hospital en el que se encuentra cada paciente, el tipo de admisión y la severidad de la enfermedad cuyas variables categóricas responden a nombres particulares.

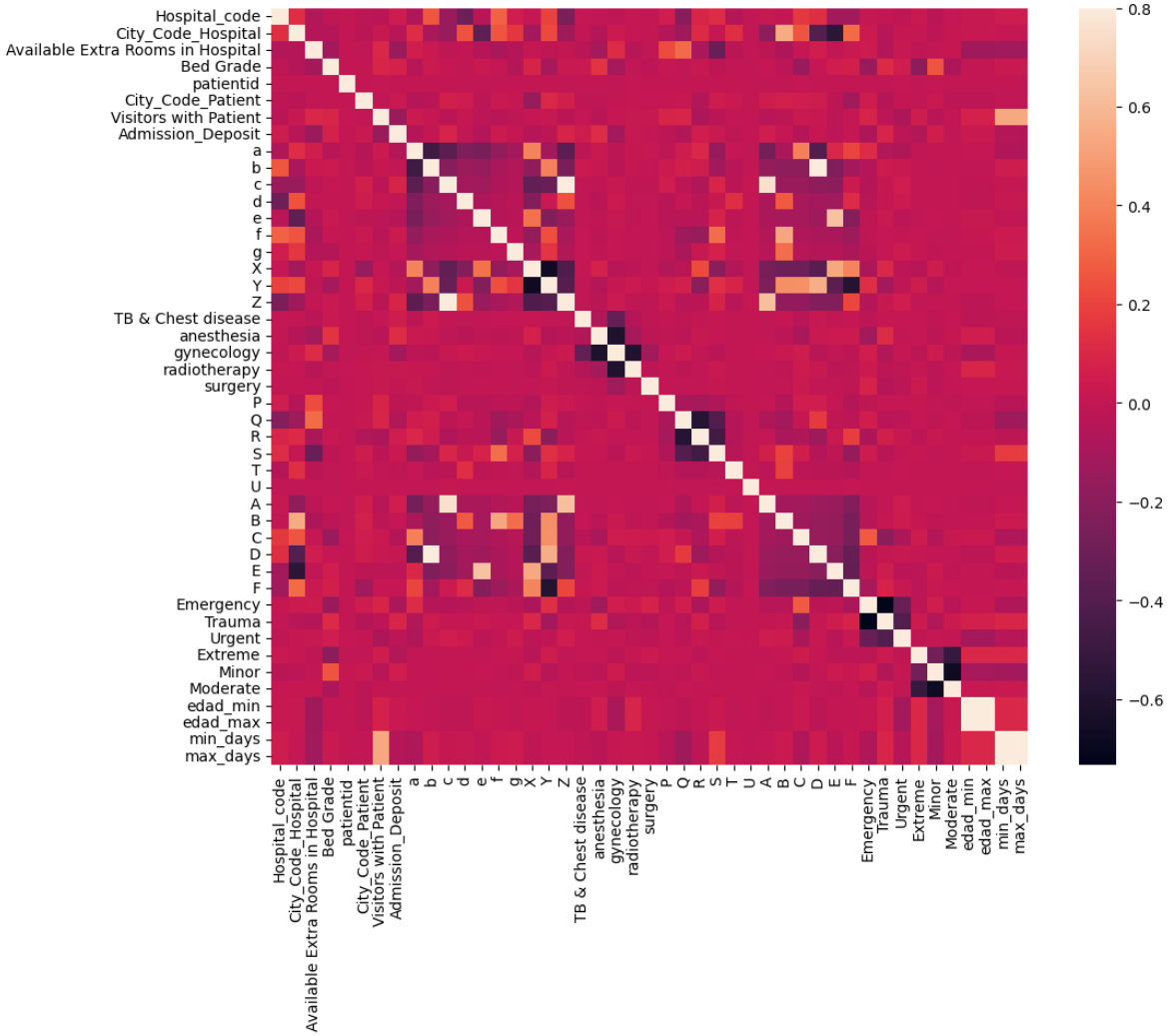
5. Con el análisis anterior se puede tomar la decisión de eliminar algunas columnas que no representan gran cantidad de información o cuya información no es tan relevante para análisis posteriores del modelo, estas columnas eliminadas fueron, el código de región del hospital, el tipo de código, y el código de facilidad de las salas.

6. También se le asignó un valor numérico a cada variable categórica.

7. Conociendo la variable objetivo, que para el caso como se definió en el entregable uno es el tiempo de estadía de los pacientes en el hospital, se realizó la distribución de densidades de dicha variable.



8. Por último se realizó una matriz de correlación que muestra cómo están relacionadas entre sí una a una las variables del conjunto de datos. De esta matriz se pueden ver resultados importantes, como la alta interrelación que tiene variables como Visitantes con pacientes y el tiempo de estadía del paciente.

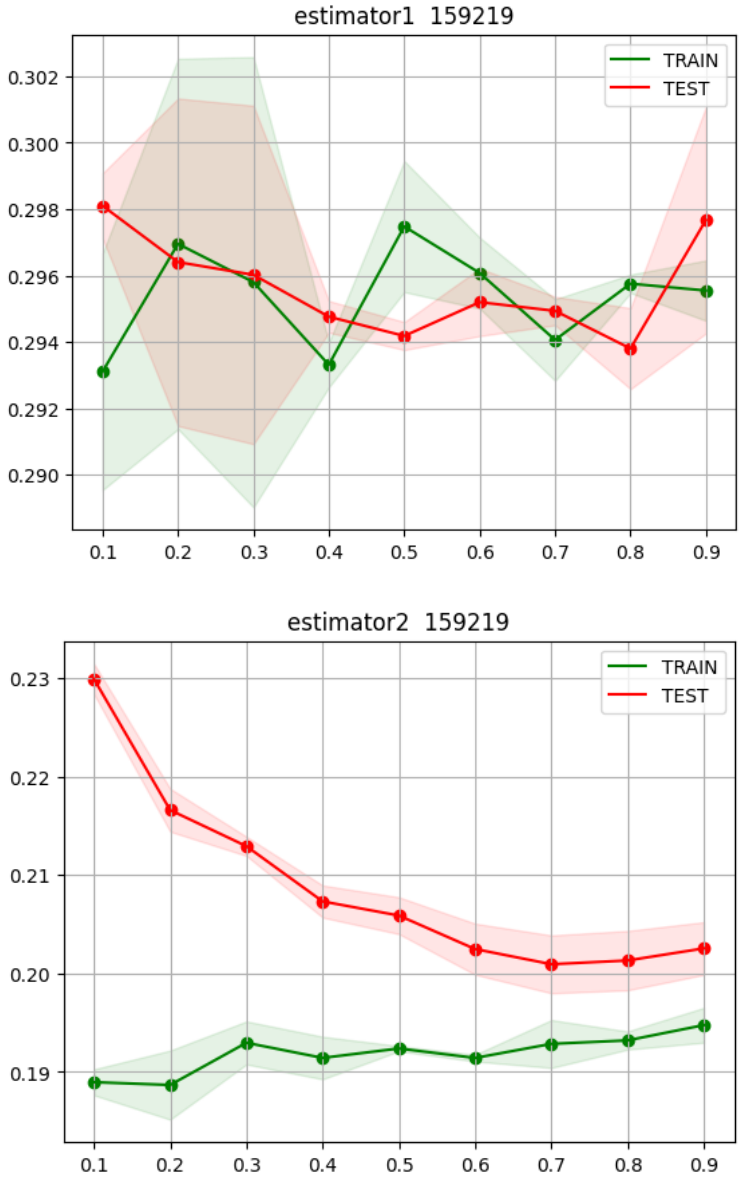


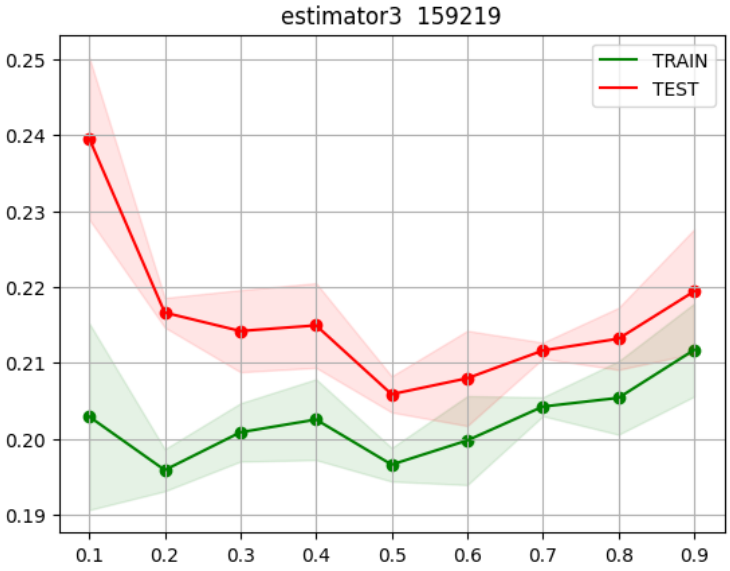
Generación del modelo

En el proceso de evaluación, se compararon y analizaron los resultados de los tres modelos utilizados: el árbol de decisión, la regresión logística y el random forest. Se aplicó un bucle for para realizar pruebas exhaustivas y evaluar el rendimiento de cada modelo en términos de precisión, exactitud y otros indicadores relevantes. Este enfoque permitió seleccionar el modelo más adecuado para el problema específico, garantizando así resultados óptimos y confiables en el análisis de los datos.

```
"""se crean varios modelos, cada uno con diferente grado de complejidad"""

estimator1 = LogisticRegression()
estimator2 = DecisionTreeClassifier(max_depth=10)
estimator3 = RandomForestClassifier(n_estimators=3, max_depth=9)


estimadores = [(estimator1,'estimator1'),(estimator2,'estimator2'),(estimator3,'estimator3')]

test_size_eval = np.array([(perc/10) for perc in range(1,10)])
```

Se generó una gráfica que muestra la curva de aprendizaje para los modelos de regresión logística, árbol de decisión y random forest. Esta gráfica permite variar los parámetros de complejidad y el número de datos para observar posibles casos de sobreajuste y sesgo. Así podemos evaluar cómo estos modelos se desempeñan y si es necesario corregir el sobreajuste o sesgo en función de los resultados obtenidos.





Tras evaluar cada modelo en base a los resultados obtenidos, se seleccionó el modelo de Random Forest debido a su destacado desempeño en comparación con los otros modelos. La evaluación exhaustiva reveló que el modelo de Random Forest logró una mayor precisión, exactitud y capacidad de generalización en la clasificación de los datos. Sus resultados consistentes y confiables en diversas métricas de evaluación lo convirtieron en la elección preferida para abordar el problema específico en cuestión. Al optar por el modelo de Random Forest, se busca obtener los mejores resultados y maximizar la calidad de las predicciones en el análisis de datos.

```
for estimator in estimators:
    print("--")
    z = cross_validate(estimator, Xtv, ytv, return_train_score=True, return_estimator=False,
                       scoring=rel_mrae, cv=ShuffleSplit(n_splits=10, test_size=val_size))
    report_cv_score(z)
    zscores.append(np.mean(z["test_score"]))
best = np.argmin(zscores)
#print ("selecting ", best)
best_estimator = estimators[best]
print ("\nModelo elegido")
print (best_estimator)
```

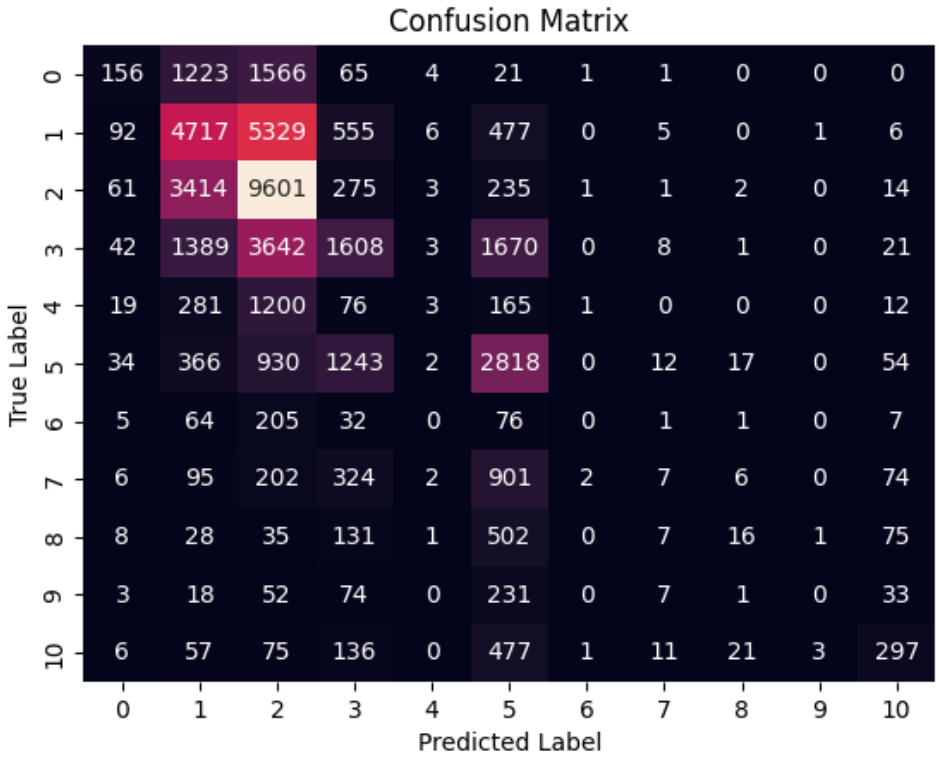
```
test score   0.295 (±0.0029) with 10 splits
train score  0.295 (±0.0021) with 10 splits
--
test score   0.209 (±0.0029) with 10 splits
train score  0.192 (±0.0017) with 10 splits
--
test score   0.209 (±0.0044) with 10 splits
train score  0.199 (±0.0045) with 10 splits

Modelo elegido
RandomForestClassifier(max_depth=9, n_estimators=3)
```

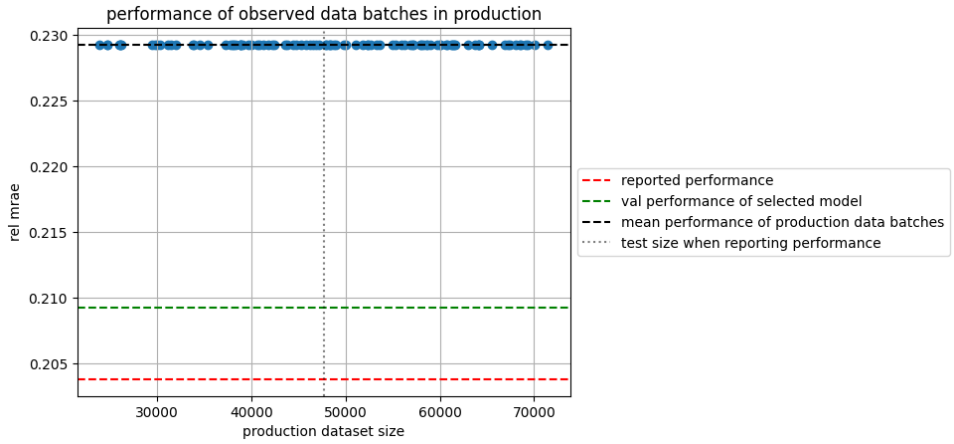
Además de la evaluación basada en diferentes métricas de desempeño, también se analizó la matriz de confusión para el modelo seleccionado, en este caso, Random Forest. La matriz de confusión nos permite observar y comprender los errores cometidos por el modelo durante la clasificación de los datos.

Al analizar la matriz de confusión, es posible identificar las clases o categorías que presentaron mayor dificultad para ser clasificadas correctamente por el modelo. Estos errores pueden ser tanto falsos positivos (clasificaciones incorrectas como positivas) como falsos negativos (clasificaciones incorrectas como negativas).

El análisis de la matriz de confusión nos brinda información valiosa sobre las limitaciones y los patrones de error del modelo. Esto permite identificar posibles áreas de mejora y afinar aún más el modelo para reducir los errores y mejorar su rendimiento general.

Es importante tener en cuenta que ningún modelo es perfecto y es normal que cometan errores en la clasificación de datos. La clave está en comprender y minimizar estos errores para obtener resultados lo más precisos y confiables posible.





En la grafica de a línea negra representa el desempeño real del modelo en situaciones de la vida cotidiana, mientras que las otras dos líneas muestran los desempeños con los que se seleccionó originalmente el modelo. Está claro que el modelo presenta más errores en la realidad de lo que se había anticipado inicialmente.

4. Retos y consideraciones de despliegue

Durante estos últimos cuatro meses de dedicación al proyecto, hemos enfrentado una variedad de desafíos que, con perseverancia y esfuerzo, hemos logrado superar de manera exitosa. Entre estos retos, se pueden resaltar algunos que han destacado por su importancia y complejidad.

- Desde el inicio del proyecto, nos hemos encontrado con el desafío de cargar de manera eficiente los archivos en la plataforma de evolución. En este sentido, es importante destacar que la información suministrada se presenta en archivos con formato de extensión .xlsx. La carga exitosa de estos archivos es fundamental para garantizar una correcta gestión y análisis de los datos.
- Durante el análisis del modelo, se pudo observar que las variables abordadas abarcaban una amplia gama de situaciones en el campo de la medicina. Sin embargo, se hizo evidente que los niveles de precisión alcanzados por el modelo en general no fueron completamente satisfactorios. Dado que se trata de un tema tan crucial como la salud y la gestión de los servicios médicos y hospitalarios, resulta evidente la necesidad de escalar hacia modelos más robustos.

  Es fundamental reconocer que los modelos más sofisticados permitirán abordar de manera más profunda las variables de interés, lo cual nos brindará una comprensión más clara y precisa de las relaciones entre los tiempos de respuesta y los resultados reales. Al optar por un enfoque más exhaustivo y detallado, podremos obtener resultados más significativos y valiosos que respalden la toma de decisiones en el ámbito de la salud.

  Es crucial resaltar que la implementación de modelos más robustos requerirá una cuidadosa consideración y análisis de las variables pertinentes, así como la adopción de técnicas avanzadas de análisis y modelado. Estos modelos deberán tener en cuenta

la complejidad y la interdependencia de las variables involucradas, permitiendo así una representación más precisa de las situaciones médicas.

En última instancia, el objetivo es lograr un mayor grado de precisión y confiabilidad en la predicción y gestión de los tiempos de respuesta en el ámbito de la salud. Esto no solo mejorará la eficiencia y la calidad de los servicios médicos y hospitalarios, sino que también proporcionará una base sólida para la toma de decisiones informadas y el diseño de estrategias efectivas en beneficio de los pacientes y el personal médico.

- Durante la implementación de los algoritmos predictivos, se observó la importancia de planificar de manera escalonada el procesamiento y análisis de un gran volumen de información. Esta planificación escalonada es crucial para lograr tiempos de ejecución más cortos y reducir la carga de recursos en la máquina utilizada. Sin embargo, al finalizar el proceso iterativo, surgieron problemas de ejecución al intentar graficar las curvas de aprendizaje correspondientes al número de variables y los datos asociados a ellas. Estas representaciones gráficas requerían una considerable cantidad de tiempo de ejecución, lo que en ocasiones afectaba negativamente el rendimiento del entorno y la máquina utilizada en el proyecto.
- Al abordar el proceso de eliminación de columnas de datos, es fundamental ejercer un discernimiento detallado y reflexivo. La selección precisa de las columnas a eliminar debe llevarse a cabo con el objetivo de preservar la coherencia en la distribución de la variable objetivo.

  Es importante tener en cuenta que cada columna de datos puede contener información relevante y contribuir de alguna manera al análisis. Sin embargo, en ocasiones, ciertas columnas pueden resultar redundantes, irrelevantes o tener un impacto negativo en la calidad de los resultados. Por tanto, la eliminación de estas columnas se convierte en una tarea esencial para obtener una representación más precisa y significativa de la variable objetivo.

5. Conclusiones

- El modelo desarrollado en este proyecto tiene un amplio potencial de aplicación en diversas áreas de la vida cotidiana, como el transporte, la disponibilidad de vuelos en aeropuertos y otras actividades relacionadas con la asignación de turnos. Además, es posible expandir su implementación en áreas de cobertura complejas dentro del sistema de salud, para facilitar actividades de baja demanda, no necesariamente limitadas al área de urgencias.
- Como estudiantes de ingeniería eléctrica (y de ingeniería en general), hemos encontrado una gran utilidad en la programación de eventos predictivos de diversos tipos. Estos conocimientos y experiencias pueden ser aplicados para desarrollar un modelo de demanda y generación de energía eléctrica por zonas o en un territorio específico. Esto demuestra cómo la experiencia adquirida en este proyecto, que no es convencional en el sector energético, puede ser transferible y adaptada a otras áreas, brindando soluciones innovadoras y eficientes.