PRIMERA ITERACIÓN SEMINARIO

Entregable II

Especialización en Analítica y Ciencia de Datos
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Colombia

Integrantes

Vanessa Restrepo Correa CC. 1017181948

Astrid Viviana Sánchez Jiménez CC. 1036621926

A. Comprensión del problema de aprendizaje automático:

1. **Descripción clara el problema de predicción que está abordando, su campo de aplicación y explique si corresponde a un problema de clasificación o de regresión:**

   Se requiere el desarrollo de un modelo de predictivo desde el ingreso de un paciente geriátrico al servicio de hospitalización hasta el egreso del mismo, con el fin de conocer cuánto tiempo permanecerá un paciente en función de la información que conocen cuando llegue ese paciente, la necesidad de camas hospitalarias y prever los recursos asistenciales necesarios para la atención. Para lo anterior, es importante definir dos conceptos:
   Estancia: está definida como el tiempo entre el ingreso hospitalario y el alta expresada en días.
   Paciente geriátrico: Está definido como la persona mayor de 59 años.
   El análisis corresponde a un problema de regresión.

2. **Realice una búsqueda de artículos que hayan abordado el mismo problema de aprendizaje que ustedes están trabajando. Incluya, en la medida de lo posible, trabajos que hayan empleado la misma base de datos. Describa brevemente:**
   ¿Qué técnica(s) de aprendizaje usan en los artículos?
   ¿Qué metodología de validación usaron?
   ¿Cuáles fueron los resultados obtenidos en cada uno de los trabajos citados?

   Antes de iniciar con la citación de los artículos, es importante aclarar que la base de datos es propia y privada. Por lo tanto, no se han realizado estudios con esta fuente de datos. Los artículos que se citarán fueron extraídos desde internet:

   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.
   La estancia hospitalaria prolongada 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. El presente trabajo presenta una investigación que enumera las causas más comunes encontradas en la prolongación de la estancia y las metodologías de logística hospitalaria más aplicadas para su estudio y mejoramiento. Resultados: los factores causales de estancias prolongadas más encontrados en la literatura son: demora en la realización de procedimientos quirúrgicos y diagnósticos, necesidad de atención en otro nivel de complejidad, situación sociofamiliar y edad del paciente. Se concluye que para analizar el problema de la estancia hospitalaria es conveniente realizar un "ajuste por riesgo", utilizando el método de grupo relacionado de diagnóstico y que una metodología adecuada es la simulación, para la identificación de cuellos de botella.

   Belletti GA, Enders J, Serra G, Yorio MA. Score predictor de días de estancia para sala común. Rev Fac Cien Med Univ Nac Cordoba [Internet]. 14 de

febrero de 2020 [citado 19 de junio de 2023];65(1):8-15. Disponible en: https://revistas.unc.edu.ar/index.php/med/article/view/27685

Aunque las internaciones más frecuentes son en sala, no existen scores para predecir días de estancia en ese sector. Los pacientes son clasificados según el diagnóstico de ingreso. Objetivo: elaborar un score para predecir días de estancia en sala común. Métodos: Estudiamos todos los pacientes ingresados al Hospital Italiano desde marzo del 2004 a mayo del 2005 en la ciudad de Córdoba (Argentina). Los criterios de inclusión fueron: pacientes mayores de 18 años, internados por más de 24 horas, no programados, por patologías médicas o quirúrgicas en sala común. Evaluamos 53 variables que incluyeron antecedentes patológicos y tóxicos, variables fisiológicas, datos demográficos, laboratorios, oxigenoterapia, datos sociales, servicio de cabecera, condición nutricional y funcional, al ingreso. Los que fallecieron durante la internación no fueron incluidos en el análisis del score. Resultados: Incluimos 1003 pacientes. Se consideró como internaciones cortas las hospitalizaciones de 4 días o menos y largas las de 5 días o más en sala común. Elaboramos un score con 11 variables, de acuerdo a percepción clínica. El análisis estadístico no fue significativo con cada variable por separado. Cuando analizamos el score con las 11 variables en conjunto mostraron significancia estadística. Subdividimos categorías y puntajes por cortes estadísticos. Puntaje mínimo: 11, máximo 33. Mostraron R2:0,77 (p: 0,06) entre puntaje y días de estancia. Conclusión: Con puntajes bajos, habla de alta probabilidad de egreso antes de 5 días. Este score puede ser una herramienta simple y factible para administración hospitalaria y para la predicción de camas disponibles en sala común.

B. Entrenamiento y evaluación de modelos

Experimentos:

1. **La metodología de validación usada y la base de datos que está usando para llevar a cabo el proyecto, incluyendo la fuente de la base de datos como referencia, el número de muestras, variables, la distribución de muestras por clase, etc.**

   La base de datos es propia y privada de un Hospital de cuarto nivel de la ciudad de Medellín, cuenta con 23 variables y 18.092 registros de egresos de pacientes mayores a 59 años. Lo primero que se realizó para la predicción de estancias fue incluir variables como la edad del paciente, el género, el diagnóstico principal y diagnósticos relacionados, estado vital al alta. Teniendo los datos consolidados se procedió a realizar análisis exploratorios para identificar posibles correlaciones entre las variables y la estancia. Así mismo, se verificaron valores atípicos y datos faltantes que posiblemente pueden afectar el modelo. Finalmente se buscará una relación lineal entre las variables independientes con la variable dependiente (estancia).

2. **Las medidas de desempeño que usarán para evaluar el sistema, indicando la medida principal.**

   Error cuadrático medio (MSE), calcula la diferencia promedio al cuadrado entre los valores predichos y los valores reales de la duración de la estancia hospitalaria. Esta medida penaliza los errores grandes y proporciona una idea de la magnitud de los errores.
   Error absoluto medio (MAE), calcula la diferencia promedio absoluta entre los valores predichos y los valores reales. Esta medida es fácil de interpretar y no penaliza los errores grandes de la misma manera que el MSE.
   Error porcentual absoluto medio (MAPE), el promedio de los errores porcentuales absolutos entre los valores predichos y los valores reales. Esta medida puede proporcionar una evaluación relativa del rendimiento del modelo en términos de porcentaje de error.
   Estas medidas de desempeño proporcionan diferentes perspectivas sobre el rendimiento del modelo de regresión en la predicción de estancias hospitalarias. El MSE y el MAE permiten evaluar la magnitud de los errores, mientras que el MAPE evalúa el rendimiento relativo en términos de porcentaje de error.

   La medida principal que utilizamos es el Error cuadrático medio. El MSE calcula la diferencia al cuadrado entre los valores predichos y los valores reales de la duración de la estancia hospitalaria, y luego toma el promedio de estos errores al cuadrado. Cuanto menor sea el valor del MSE, mejor será el desempeño del modelo, ya que indicará que las predicciones se acercan más a los valores reales.

   La elección del MSE como medida principal se debe a que enfatiza la precisión de las predicciones en términos de magnitud de los errores. Esto es especialmente relevante en el contexto de las estancias hospitalarias, donde se busca predecir con la mayor precisión posible la duración de la estancia de los pacientes.

C. Resultados y discusión:

Incluya un apartado en su informe donde exponga los resultados obtenidos con los modelos de predicción que considere más relevantes y en la cual

analice los resultados obtenidos y los compare con los resultados de los artículos consultados.

Para la Base de Datos lo primero que se realizó fue la limpieza de los datos, la cual incluyó: Identificación, revisión y recategorización de la naturaleza de las variables.

```
Edad                              int64
Sexo                             object
Diagnostico_principal_Egreso     object
Cantidad_comorbilidades           int64
Tuvo_cx                          object
Procedimiento_1                  object
Ventilacion_Mecanica             object
Situacion_al_alta                object
UCI                              object
UCE                              object
Nombre_Especialidad_Egreso       object
Estancia                          int64
dtype: object
```

Búsqueda de datos faltantes o caracteres especiales.

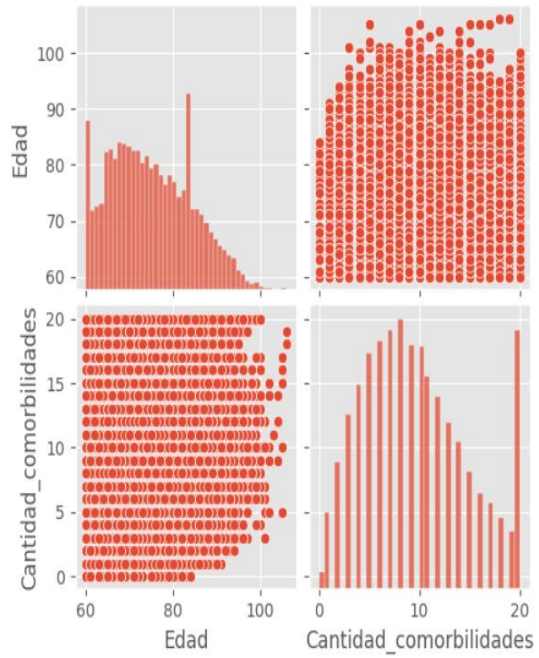
```
Edad = list(d['Edad'].value_counts().index)
Sexo = list(d['Sexo'].value_counts().index)
Diagnostico_principal_Egreso = list(d['Diagnostico_principal_Egreso'].value_counts().index)
Cantidad_comorbilidades = list(d['Cantidad_comorbilidades'].value_counts().index)
Tuvo_cx = list(d['Tuvo_cx'].value_counts().index)
Procedimiento_1 = list(d['Procedimiento_1'].value_counts().index)
Ventilacion_Mecanica = list(d['Ventilacion_Mecanica'].value_counts().index)
Situacion_al_alta = list(d['Situacion_al_alta'].value_counts().index)
UCI = list(d['UCI'].value_counts().index)
UCE = list(d['UCE'].value_counts().index)
Nombre_Especialidad_Egreso = list(d['Nombre_Especialidad_Egreso'].value_counts().index)
Estancia = list(d['Estancia'].value_counts().index)
```

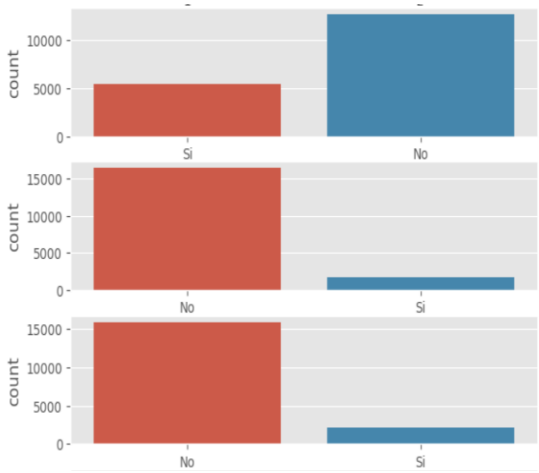
Tratamiento de datos atípicos.

```
#Calculo de valores atípicos


#Calculo de Q1 t Q3
Q1 = np.percentile(d['Estancia'], 25, interpolation = 'midpoint')
Q3 = np.percentile(d['Estancia'], 75, interpolation = 'midpoint')

#Cálculo del rango intercuartil
IQR = Q3 - Q1

#Cálculo de valor mínimo y máximo para los valores atípicos
VAInf = Q1 - 1.5*IQR
VASup = Q3 + 1.5*IQR

print(f'Valor atípico leve inferior:{VAInf}')
print(f'Valor atípico leve superior:{VASup}')
```

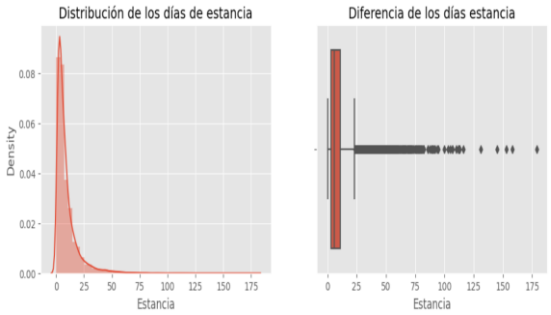
```
Valor atípico leve inferior:-9.0
Valor atípico leve superior:23.0
```

Se visualizan las variables numéricas y categóricas.





Se realiza una exploración de la variable de salida, para nuestro caso es la estancia:



Por último, se crearon los datos de entrenamiento y de prueba para realizar el modelo de regresión que escogimos que fue Random Forest:

Inicialmente se escogieron los datos de entrenamiento y de prueba, los cuales se realizaron con un 80% - 20%.

```
X_train, X_test, y_train, y_test = train_test_split(
                                       X,
                                       y.values.reshape(-1,1),
                                       train_size   = 0.8,
                                       random_state = 1234,
                                       shuffle      = True
                                   )
```

Posteriormente, se procedió a la creación del modelo Random Forest con 100 árboles y 5 niveles de profundidad

```
RandomForestRegressor
RandomForestRegressor(criterion='friedman_mse', max_depth=5,
                      max_features='auto', n_jobs=-1, random_state=1234)
```

El cual nos arrojó los siguientes valores para las medidas MSE y R^2 :

```
El error rmse de test es: 4.225308460469824
El error R^2 de test es: 0.32168981783234063
```

Los valores que nos entregó el modelo en el MSE nos indica que el modelo no tiene un buen desempeño debido a que el dato no está cerca de cero por lo tanto los valores de las predicciones no se acercan a los valores reales.

Así mismo, el R^2 como está cerca de 0 nos indica que no explica ninguna variabilidad.

Teniendo en cuenta lo anterior, es importante que sigamos explorando otras medidas de desempeño para realizar un análisis más exhaustivo para una evaluación completa del modelo de regresión.

Para futuras exploraciones también se deberá correr el modelo con las características que le aportan mayor importancia a la predicción de la estancia hospitalaria. Se creó inicialmente el código para determinar cuales son las que podrían ayudar más a predecir el modelo.

	predictor	importancia
1	Cantidad_comorbilidades	0.6475
1698	Tuvo_cx_Si	0.1348
1702	UCI_Si	0.0470
1728	Nombre_Especialidad_Egreso_MEDICINA CRITICA Y ...	0.0314
1699	Procedimiento_1_Si	0.0274
...	...	...
654	Diagnostico_principal_Egreso_I251	0.0000
653	Diagnostico_principal_Egreso_I250	0.0000
652	Diagnostico_principal_Egreso_I221	0.0000
651	Diagnostico_principal_Egreso_I220	0.0000
874	Diagnostico_principal_Egreso_J841	0.0000

1748 rows × 2 columns