Introducción a Inteligencia Artificial para ciencias e ingenierías

Segunda Entrega

Estudiantes:

Omar Alberto Torres CC 91220873
Carlos Alfredo Pinto Hernández CC 1100953378

Profesor:
Raúl Ramos Pollan

Faculta de ingeniería
Departamento de Ingeniería de Sistemas
Ude@
2023

I. Introducción

1. Problema predictivo

Dadas unas características físicas, enfermedades y resultados de laboratorios de una persona se predecirá la probabilidad de presentar enfermedad cerebro vascular o stroke.

2. Dataset elegido

Se utilizará el dataset de kaggle de la competición denominada “Playground Series - Season 3, Episode 2” disponible en el siguiente enlace: https://www.kaggle.com/competitions/playground-series-s3e2/ el cual contiene más de 15.000 muestras y divididas en 12 columna incluyendo: edad, género, índice de masa corporal, promedio de nivel de glucosa, tipo de residencia, tipo de trabajo, si ha estado casado, si presenta hipertensión o enfermedad cardíaca.

Se obtienen dos archivos uno denominado “train.csv - the training dataset”: donde se especifica el target “stroke” en formato binario y el otro documento “test.csv - the test dataset” con el cual se probará el modelo para predecir la probabilidad de presentar un stroke.

3. Métrica de desempeño

Como métrica de machine learning vamos a utilizar el área bajo la curva ROC entre la probabilidad del modelo vs el resultado (stroke); esta métrica fue definida en la competencia. Como métrica de negocio se podría plantear el incremento del número de personas que ingresan a controles de riesgo cardiovascular detectados por el modelo.

4. Criterio sobre el cual sería el desempeño deseable en producción

Se espera tener un recall (sensibilidad) mayor al 90% dado que son los valores de los modelos de prevención y tamizaje en el ámbito de salud.

II. Exploración descriptiva del Dataset

1. Variable objetivo (y)

La variable stroke es la variable objetivo del dataset y se representa de forma binaria, 0 significa que la persona no presentó stroke y 1 que si lo presentó. En el grafico 1 se muestra la frecuencia de las categorías de stroke en las muestras del dataset.

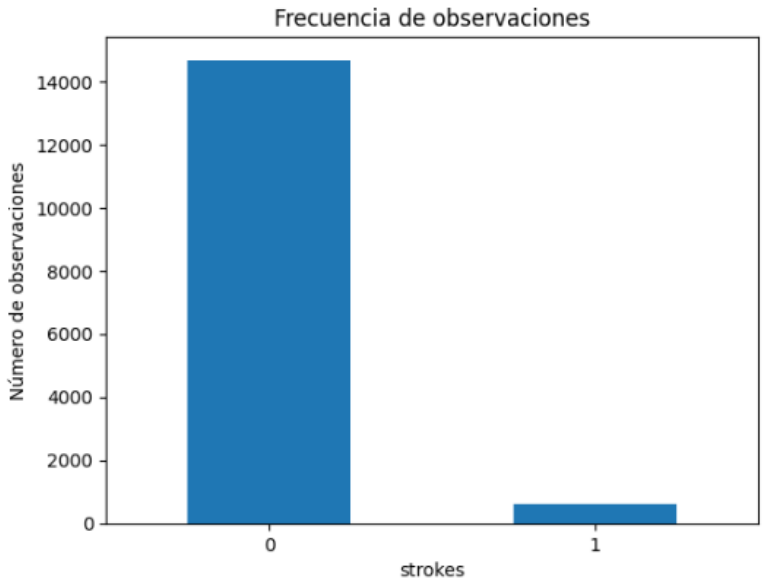

Gráfico 1*: Frecuencia de la variable stroke*

Se evidencia un dataset muy desbalanceado, dado el número mayor de la categoría 0, pero se considera que es algo habitual dado que el dominio del problema está relacionado con el área de la salud.

2. Características

Usando el método head() de pandas, se obtienen la tabla con las primeras 5 muestras con todas las características y variable de salida u objetivo. Con el método shape() se evidencia la cantidad de muestras o filas y las características/salida como columnas. En este caso el database contiene 15304 columnas y 12 columnas.

index	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	0	Male	28.0	0	0	Yes	Private	Urban	79.53	31.1	never smoked	0
1	1	Male	33.0	0	0	Yes	Private	Rural	78.44	23.9	formerly smoked	0
2	2	Female	42.0	0	0	Yes	Private	Rural	103.0	40.3	Unknown	0
3	3	Male	56.0	0	0	Yes	Private	Urban	64.87	28.8	never smoked	0
4	4	Female	24.0	0	0	No	Private	Rural	73.36	28.8	never smoked	0

Tabla 1: Visualización de database.head()

Luego al utilizar el método info() podemos visualizar el tipo de variable de cada columna y la cantidad de nulos que tiene el dataset. Para efectos académicos a esta BD se insertaron valores nulos del 5% para poder realizar manejo de estos, dado que originalmente no traían datos nulos. Se realizó la inclusión de valores nulos a las variables age y avg_glucose_level.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15304 entries, 0 to 15303
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 15304 non-null  int64
 1   gender             15304 non-null  object
 2   age                15304 non-null  float64
 3   hypertension       15304 non-null  int64
 4   heart_disease      15304 non-null  int64
 5   ever_married       15304 non-null  object
 6   work_type          15304 non-null  object
 7   Residence_type     15304 non-null  object
 8   avg_glucose_level  15304 non-null  float64
 9   bmi                15304 non-null  float64
 10  smoking_status     15304 non-null  object
 11  stroke             15304 non-null  int64
dtypes: float64(3), int64(4), object(5)
```

Gráfico 2*: Tipo de variables y datos nulos en BD original*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15304 entries, 0 to 15303
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 15304 non-null  int64
 1   gender             15304 non-null  object
 2   age                14556 non-null  float64
 3   hypertension       15304 non-null  int64
 4   heart_disease      15304 non-null  int64
 5   ever_married       15304 non-null  object
 6   work_type          15304 non-null  object
 7   Residence_type     15304 non-null  object
 8   avg_glucose_level  14552 non-null  float64
 9   bmi                15304 non-null  float64
 10  smoking_status     15304 non-null  object
 11  stroke             15304 non-null  int64
dtypes: float64(3), int64(4), object(5)
```

Gráfico 3*: Tipo de variables y datos nulos en BD con inclusión de valores nulos*

2.1 Descripción estadística de variables numéricas

Se tomaron las variables numéricas y se realizó análisis con método describe, con el fin de obtener valores como número total, media, desviación estándar, valor mínimo, valor máximo y los percentiles 25, 50 y 75. En la tabla 2, se evidencian estos resultados para las variables age, avg_glucose_level y bmi.

index	age	avg_glucose_level	bmi
count	14556.0	14552.0	15304.0
mean	41.41943391041495	88.97298584387025	28.112720857292214
std	21.45553096417527	25.48028637456789	6.722315422227762
min	0.08	55.22	10.3
25%	25.0	74.8175	23.5
50%	43.0	85.07	27.6
75%	57.0	96.92	32.0
max	82.0	267.6	80.1

Tabla 2: Descripción estadística de variables numéricas

A través de boxplot se pueden visualizar distribución de los datos, simetría y datos atípicos de variables numéricas. En el siguiente figura se presenta un boxplot de la variable avg_glucose_level.

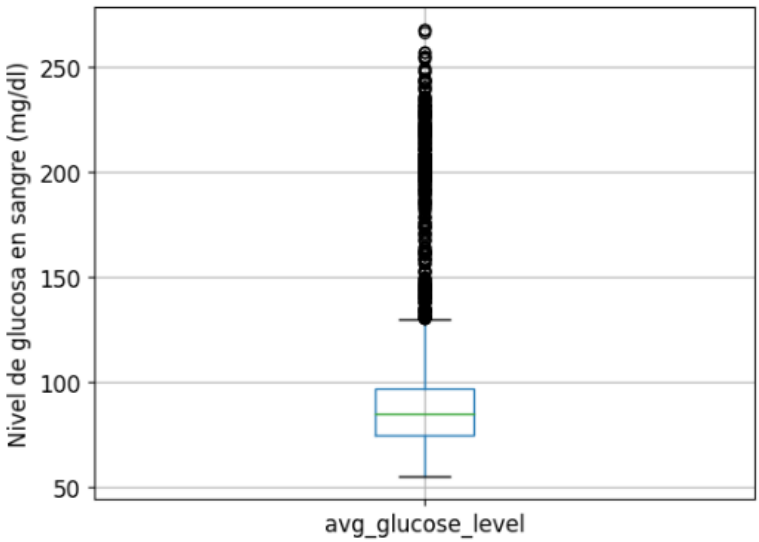

Gráfico 4*: Boxplot de distribución de la variable avg_glucose_level*

2.2 Variables categóricas

Para cuantificar las variables categóricas utilizamos el método value_counts() y con esto se pudieron realizar algunas gráficas de distribución. En el caso de la variable work_type se encontraron las frecuencias de sus 5 categorías las cuales fueron graficadas en plot tipo pie.

```
Private          9752
children         2038
Self-employed    1939
Govt_job         1533
Never_worked       42
Name: work_type, dtype: int64
```

Gráfico 5*: Frecuencias de las categorías de la variable work_type*

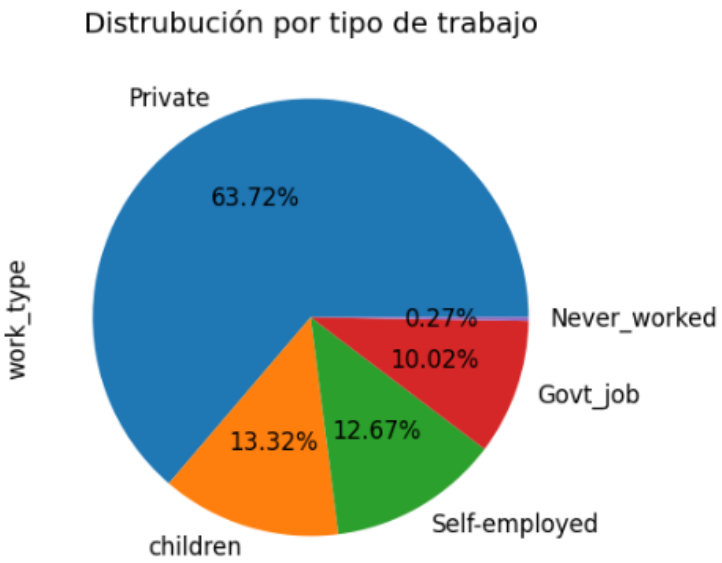

Gráfico 6*: Distribución por tipo de trabajo*

III. Avances del proyecto

1. Progreso

Se realizó inicialmente una exploración de los datos como está descrito en el numeral 2 además de la inclusión de análisis de correlación entre las variables y la variable objetivo.

Como parte del proceso de formación generamos 5% de nulos en dos variables en el dataset original para realizar el proceso de ajuste y ejecución de métodos para tratar estos valores. Además, realizamos la conversión de variables categóricas a numéricas.

2. Modelos implementados

El primer modelo implementado fue el Decision Tree utilizando el método DecisionTreeClassifier() importado de sklearn.tree. Este método presentó un AUC de 0.817 y la siguiente curva ROC:

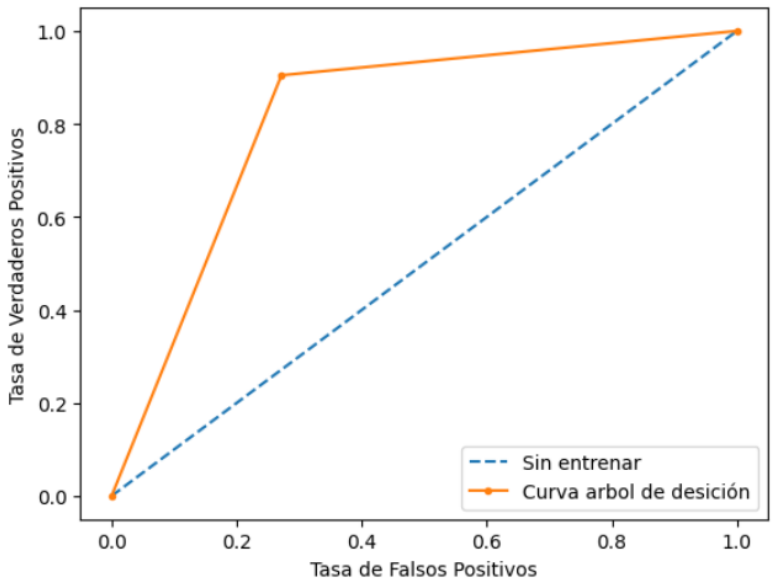

Gráfico 7*: Curva ROC del Modelo Árbol de decisión*

El segundo modelo que ejecutamos fue el de maquina de soporte vectorial quien mostró un resultado de AUC del 0.731 con la siguiente curva ROC:

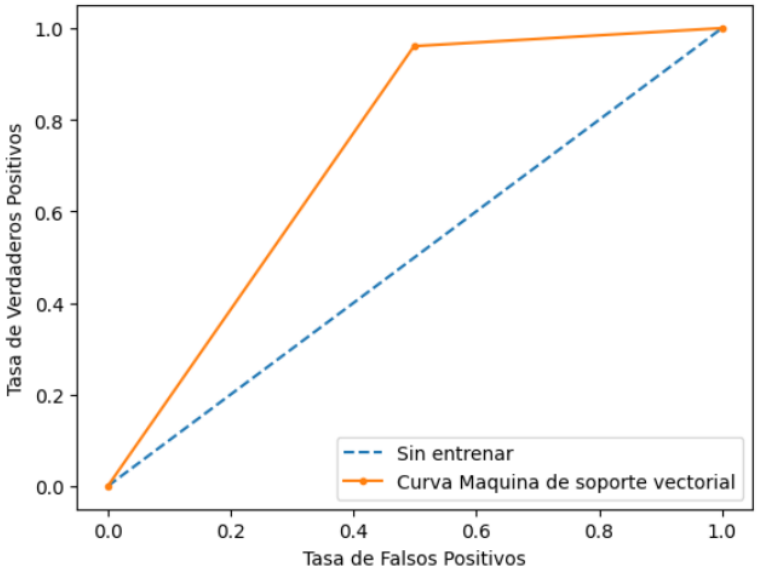

Gráfico 8*: Curva ROC del Modelo Maquina de soporte vectorial*

Para estos modelos se utilizamos 90% del dataset para entrenamiento y 10% restante para verificación. Pero tenemos que usar el otro dataset (test) para hacer la verificación y usar el 100% de esta BD (train) en los modelos planteados.

Hemos observado que es muy desbalanceado el dataset pero encontramos que es muy común en el área de la salud, por lo tanto, no sabemos si hacer el balanceado previo a los modelos o dejarlos tal como son el vida real y generar los modelos así.