Pontificia Universidad Javeriana
Proyecto de Inteligencia artificial
Predicción-de-Stroke
Realizado por: Laura Sofía Gómez Lizarazo

El siguiente proyecto tiene como objetivo la predicción de Stroke en diferentes pacientes, a partir del dataset mostrado a continuación:
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

En dicho Dataset aparecen las siguientes características:

1) "id: unique identifier"
2) "gende"r: "Male", "Female" or "Other"
3) "age": age of the patient
4) "ypertension": 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) "heart_disease": 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) "ever_married": "No" or "Yes"
7) "work_type": "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) "Residence_type": "Rural" or "Urban"
9) "avg_glucose_level": average glucose level in blood
10) "bmi": body mass index
11) "smoking_status": "formerly smoked", "never smoked", "smokes" or "Unknown"
12) "stroke": 1 if the patient had a stroke or 0 if not

Dado lo anterior se procedió a realizar el procesamiento de los datos, que en este caso incluía la limpieza de los mismos y la creación de un nuevo archivo con los nuevos datos. Así mismo se normalizó y se realizaron los siguientes modelos de predicción con su respectiva validación.

* SVM (Máquinas de soporte vectorial)
* Regresión logística
* KNN (K vecinos más cercanos)

A continuación se muestran los resultados de los métodos utilizados, a partir de dos métricas, f1 y AUC_ROC:

Para SVM

```
[ ] G_S_SVM.best_params_
```

```
{'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
```

```
[ ] G_S_SVM.best_score_
```

```
0.7249953439248363
```

```
[ ] G_S_SVM.cv_results_['mean_test_AUC']
```

```
array([0.72499534, 0.66757912, 0.62290126, 0.72499534, 0.68552306,
       0.61737969, 0.60054986, 0.71586584, 0.57155518, 0.60054986,
       0.71774609, 0.56928645, 0.60933523, 0.71615423, 0.57443636,
       0.60933523, 0.71013258, 0.57348895, 0.67527076, 0.708794  ,
       0.58116792, 0.67527076, 0.71368685, 0.58103773, 0.57671502,
       0.70539322, 0.5854412 , 0.57671502, 0.69159467, 0.58545303,
       0.56664256, 0.68927337, 0.60002402, 0.56664256, 0.69266413,
       0.60049146, 0.56901276, 0.69353771, 0.60575298, 0.56901276,
       0.69610966, 0.6060824 ])
```

```
[ ] G_S_SVM.cv_results_['mean_test_f1']
```

```
array([0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.02081081,
       0.        , 0.        , 0.02081081, 0.        , 0.        ,
       0.02005013, 0.        , 0.        , 0.02005013, 0.        ,
       0.        , 0.02802283, 0.        , 0.        , 0.02802283,
       0.        , 0.        , 0.04375758, 0.        , 0.        ,
       0.04375758, 0.        , 0.        , 0.05248019, 0.        ,
       0.        , 0.05248019])
```

Para regresión logística:

```
[ ] G_S_LR.best_params_
```

```
{'C': 0.029}
```

```
[ ] G_S_LR.best_score_
```

```
0.837600425490289
```

```
[ ] G_S_LR.cv_results_['mean_test_AUC']
```

```
array([0.81963887, 0.82265387, 0.82517373, 0.82723319, 0.82898367,
       0.83012401, 0.83141698, 0.83245314, 0.83308015, 0.83349721,
       0.83405518, 0.83453207, 0.83495103, 0.83536228, 0.83566369,
       0.83583336, 0.83609857, 0.83640708, 0.83666426, 0.83673064,
       0.83673936, 0.83681422, 0.83704217, 0.83701387, 0.8372321 ,
       0.83731905, 0.83732095, 0.83742462, 0.83760043])
```

```
[ ] G_S_LR.cv_results_['mean_test_f1']
```

```
array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])
```

Para KNN:

```
[ ] G_S_k.best_params_
```

```
{'algorithm': 'auto', 'n_neighbors': 50, 'weights': 'uniform'}
```

```
[ ] G_S_k.best_score_
```

```
0.8091997710175098
```

```
[ ] G_S_k.cv_results_['mean_test_AUC']
```

```
array([0.52760562, 0.52760562, 0.56523318, 0.56507457, 0.62600452,
       0.6255411 , 0.71344893, 0.70863992, 0.79276095, 0.78619002,
       0.80919977, 0.80481029, 0.80864477, 0.80764015, 0.52760562,
       0.52760562, 0.56523318, 0.56507457, 0.62600452, 0.6255411 ,
       0.71344893, 0.70863992, 0.79276095, 0.78619002, 0.80919977,
       0.80481029, 0.80864477, 0.80764015, 0.52760562, 0.52760562,
       0.56523318, 0.56507457, 0.62600452, 0.6255411 , 0.71344893,
       0.70863992, 0.79276095, 0.78619002, 0.80919977, 0.80481029,
       0.80864477, 0.80764015, 0.52760562, 0.52760562, 0.56523318,
       0.56507457, 0.62600452, 0.6255411 , 0.71344893, 0.70863992,
       0.79276095, 0.78619002, 0.80919977, 0.80481029, 0.80864477,
       0.80764015])
```

```
[ ] G_S_k.cv_results_['mean_test_f1']
```

```
array([0.0977839 , 0.0977839 , 0.01864499, 0.0977839 , 0.01882614,
       0.04708114, 0.        , 0.01      , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.0977839 ,
       0.0977839 , 0.01864499, 0.0977839 , 0.01882614, 0.04708114,
       0.        , 0.01      , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.0977839 , 0.0977839 ,
       0.01864499, 0.0977839 , 0.01882614, 0.04708114, 0.        ,
       0.01      , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.0977839 , 0.0977839 , 0.01864499,
       0.0977839 , 0.01882614, 0.04708114, 0.        , 0.01      ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        ])
```

Conclusiones

A partir de las diferentes métricas y modelos de predicción utilizados, es posible evidenciar que el mejor método para predecir si un paciente presenta Stroke es el de regresión logística con una métrica de AUC_ROC.

Algo a resaltar es que la métrica que daba más alto en todos los casos fue la de AUC_ROC. Por otra parte el resultado más bajo fue el obtenido con máquinas de soporte vectorial.

En cuanto a los tiempos de ejecución es muy importante destacar que el método más demorado fue el de SVM, tardándose aproximadamente 1 hora.

Otro aspecto importante que vale la pena mencionar es que a lo largo del proyecto se utilizó Gridserch cv, evaluando cada método con dos métricas f1 y AUC-ROC, por lo que es importante que se especifique el refit dentro de los parámetros del Gridserch-cv, pues en caso contrario saldrá error.