

UNIVERSIDAD MAYOR DE SAN ANDRÉS  
FACULTAD DE CIENCIAS PURAS Y NATURALES  
CARRERA DE INFORMÁTICA



## Examen Parcial - Pregunta 2

---

**Nombres:** EDUARDO MEDRANO AYARDE CI: 6989411

---

**Materia:** Inteligencia Artificial (INF-354)

---

**Docente:** M.Sc. Moises Martin Silva Choque

---

**Fecha:** 27 de julio de 2020

---

# 1. Realice un proceso completo de análisis usando pipeline

## 1.1. Descripción del Dataset

Para el caso de estudio usando Pipeline, se hizo uso del dataset Haberman, del siguiente enlace <https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival>, el cual describe, sobre la evolución y supervivencia de mujeres que hayan sido operadas, para tratar el cáncer de mama, las columnas del dataset son los siguientes:

- Edad: Edad de las mujeres, cuando fueron sometidas a la operación.
- Año de intervención quirúrgica: restando 2000-1900, la columna muestra el año en el que se realizaron la operación.
- Ganglios: Es un número entero, el cual nos señala, cuantos ganglios cancerosos fueron detectados.
- Supervivencia: 1 si sobrevivió más de 5 años después de la operación, 2 si murió entre los 5 años.

Para resolver el problema, se efectuó los siguientes procedimientos

- El dataset fue dividido en X\_train y Y\_train, donde el primero representa a la edad, año de operación y ganglios cancerosos, el segundo hace referencia a la supervivencia de las mujeres.
- Ambas matrices la primera matriz fue normalizada y estandarizada, posterior a esto se volvió a fraccionar en la matriz de entrenamiento y la de prueba, donde el porcentaje es 80 y 20.
- Las pruebas fueron realizadas usando pipeline en conjunto con la regresión logística y una prueba singular con la regresión logística, esto con el objetivo de comparar los resultados.

Todo el análisis anterior, está desarrollado bajo el siguiente código:

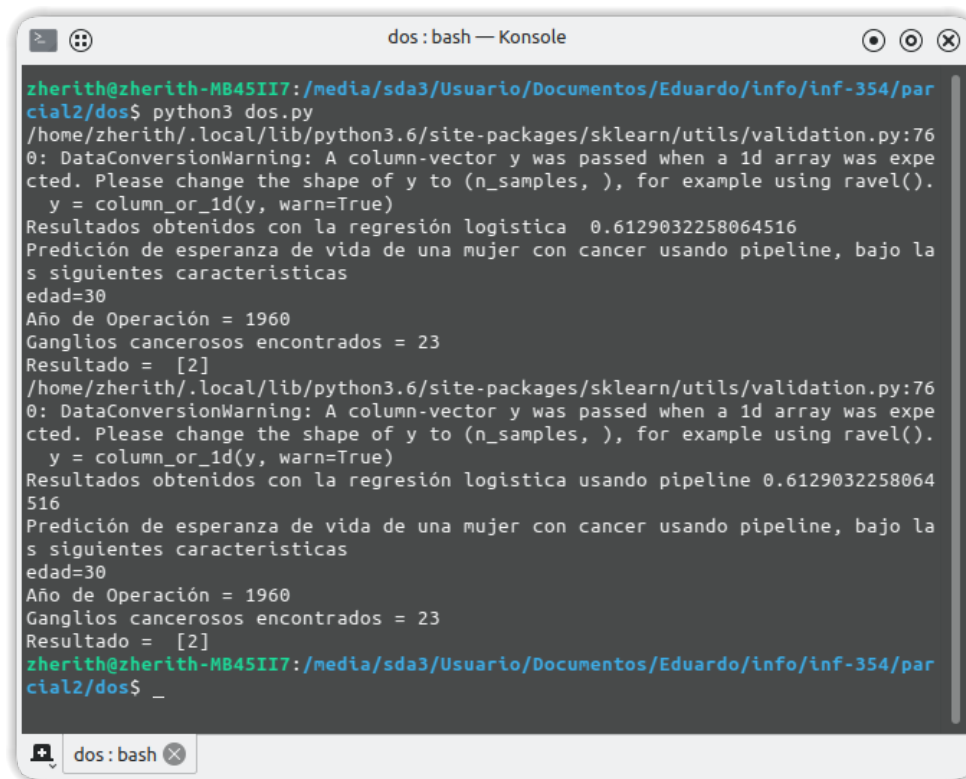
```
1  import numpy as np
2  import pandas as pd
3  from sklearn import preprocessing as pre
4  from sklearn import linear_model as lm
5  from sklearn.preprocessing import Normalizer as nm
6  from sklearn.preprocessing import StandardScaler as sc
7  from sklearn.model_selection import train_test_split
8  from sklearn.preprocessing import StandardScaler
9  from sklearn.pipeline import Pipeline
10 from sklearn.svm import SVC
11 from sklearn.linear_model import LogisticRegression
12
13 # cargamos el dataset
14 dato = pd.read_csv("haberman.csv")
15
16 aux = np.array(dato[['edad', 'year', 'auxilia']])
17 norm = nm().fit(X=aux).fit_transform(X=aux)
18 standar = sc().fit(X=norm).fit_transform(X=norm)
19 aux2 = np.array(dato[['supervivencia']])
20 norm2 = nm().fit(X=aux2).fit_transform(X=aux2)
21 X_train, X_test, y_train, y_test = train_test_split(standar,
22 aux2,
23 test_size=0.2,
24 random_state=0
25 )
26 pipe = Pipeline([('scaler', StandardScaler()),
27 ('svc', LogisticRegression(random_state=42))])
28 doge = lm.LogisticRegression(random_state=42)
29 doge.fit(X_train, y_train)
30 print("Resultados obtenidos con la regresión logistica ",
```

```

31 doge.score(X_test, y_test))
32 print("Predicción de esperanza de vida de una mujer con cancer usando pipeline
33 , bajo las siguientes características\edad=30\Año de Operación = 1960 ",
34 "Ganglios cancerosos encontrados = 23 \nResultado = ", doge.predict([[30, 60, 23])
35 pipe.fit(X_train, y_train)
36 print("Resultados obtenidos con la regresión logística usando pipeline",
37 pipe.score(X_test, y_test))
38 print("Predicción de esperanza de vida de una mujer con cancer usando pipeline,
39 bajo las siguientes características\edad=30\Año de Operación = 1960 ",
40 "Ganglios cancerosos encontrados = 23 \nResultado = ", pipe.predict([[30, 60, 23])

```

De donde se obtiene los siguientes resultados:



```

zherith@zherith-MB45II7:/media/sda3/Usuario/Documentos/Eduardo/info/inf-354/par
cial2/dos$ python3 dos.py
/home/zherith/.local/lib/python3.6/site-packages/sklearn/utils/validation.py:76
0: DataConversionWarning: A column-vector y was passed when a 1d array was expe
cted. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
Resultados obtenidos con la regresión logística 0.6129032258064516
Predicción de esperanza de vida de una mujer con cancer usando pipeline, bajo la
s siguientes características
edad=30
Año de Operación = 1960
Ganglios cancerosos encontrados = 23
Resultado = [2]
/home/zherith/.local/lib/python3.6/site-packages/sklearn/utils/validation.py:76
0: DataConversionWarning: A column-vector y was passed when a 1d array was expe
cted. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
Resultados obtenidos con la regresión logística usando pipeline 0.6129032258064
516
Predicción de esperanza de vida de una mujer con cancer usando pipeline, bajo la
s siguientes características
edad=30
Año de Operación = 1960
Ganglios cancerosos encontrados = 23
Resultado = [2]
zherith@zherith-MB45II7:/media/sda3/Usuario/Documentos/Eduardo/info/inf-354/par
cial2/dos$ _

```

Figura 1: Resultados Obtenidos del Dataset Haberman, usando Pipeline

Se puede ver que en ambos casos se obtuvo un resultado cerca de 0.61, y una predicción correcta.