Evaluación Final
Modulo IV :
VI PDE Machine Learning con python

2020

Puntaje

Pregunta	**Puntaje**
1	5Ptos.
2	5Ptos.
3	5Ptos.
4	5Ptos.

1. Realice las siguientes regresiones lineales usando el conjunto de datos `Advertising.csv` . Los modelos a desarrollar son :

$$sales \approx \beta_0 + \beta_1\, TV \tag{1}$$
$$sales \approx \beta_0 + \beta_1\, TV + \beta_2\, radio + \beta_3\, newspaper \tag{2}$$

2. De respuesta a las siguientes cuestiones acerca del conjunto de datos `Titanic.csv`

   a) Utilice el método `shape` para saber cuantas filas tiene conjunto de datos.

   b) Encuentre las dos variables con una mayor cantidad de elementos faltantes. Debe obtener las siguientes dos variables (con sus respectivas cantidades de elementos faltantes) :

Variable	Num. de elementos faltantes
Age	177
Cabin	687

   c) Grafique el histograma de a variable `Age`

   d) Cree la variable `ViajaSolo` de la siguiente manera

`ViajaSolo`	`Sisbp + Parch`
0	mayor que cero
1	en otro caso

   e) Realice una regresion logistica entre la variable dependiente `Survived` y las variables independientes : `Age`, `ViajaSolo`, `Pcall`, `Embarked` y `Sex`.

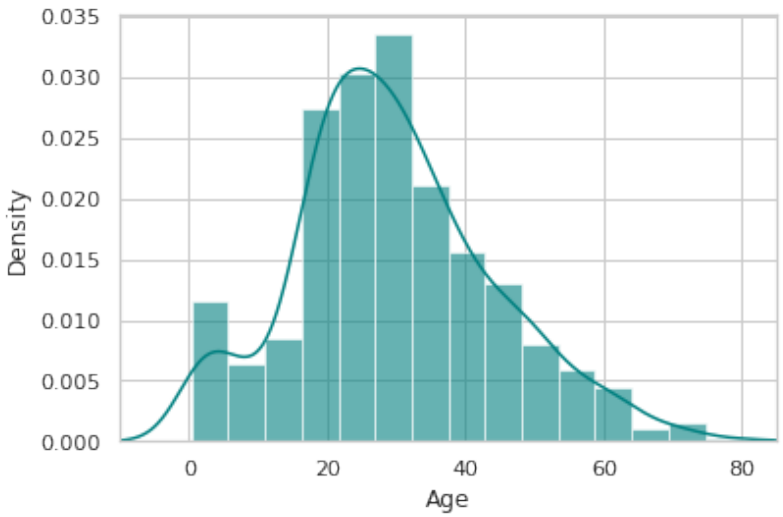


Figura 1: Resultado Aproximado de la pregunta 2c.

3. El **clustering** es un método de aprendizaje no supervisado que nos permite agrupar un conjunto de objetos en función de características similares. En general, puede ayudarlo a encontrar una estructura significativa entre sus datos, agrupar datos similares y descubrir patrones subyacentes.El problema a resolver en esta pregunta es la **Predicción de las especies del dataframe `IRIS`**.Uno de los métodos de agrupamiento más comunes es el algoritmo K-means. El objetivo de este algoritmo es dividir los datos en un conjunto de manera que se minimice la suma total de las distancias al cuadrado desde cada punto hasta el punto medio del grupo. El problema en cuestion se configura de la siguiente manera

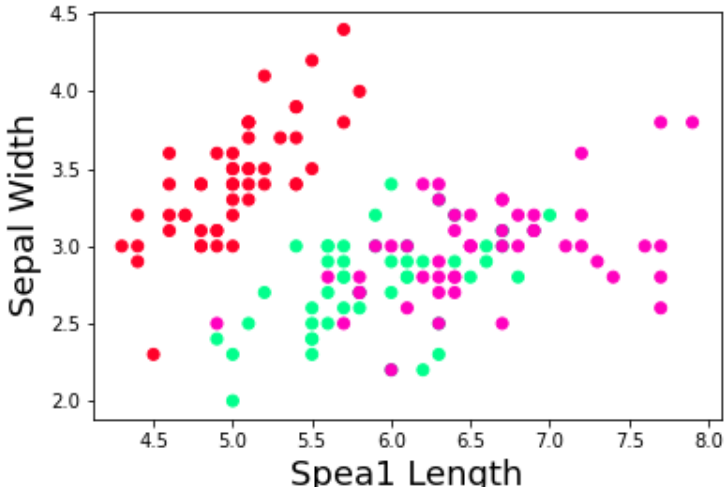
```
# Modulos
from sklearn import datasets
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.cluster import KMeans


# Cargamos la data
iris = datasets.load_iris()

# Definimos la variable target y las variables predictoras
X = iris.data[:, :2]
y = iris.target


plt.scatter(X[:,0], X[:,1], c=y, cmap='gist_rainbow')
plt.xlabel('Spea1 Length', fontsize=18)
plt.ylabel('Sepal Width', fontsize=18)
```

Figura 2: Pregunta 3



Diga usted cuales las coordenadas de los tres centros del dataframe `IRIS`.

4. Desarrolle un informe de investigacion sobre la tecnica PCA aplicada al conjunto de datos `load_digits()` del modulo `sklearn`.

```
from sklearn.datasets import load_digits
digits = load_digits()
print(digits.data.shape)

import matplotlib.pyplot as plt
plt.matshow(digits.images[1]) # modifique el 1 por el digito que desee ver
plt.show()
```

Figura 3: Pregunta 4 : Digito 1