Tutorial 7

Classification

You have been recruited by the Health Service to explore the relationship between different patient characteristics and risk of diabetes. Diabetes is a major concern for the Health Services and early detection is key to improving health outcomes.

The Health Service has provided you with the file "diabetes.csv". This file contains data on people who have recently been diagnosed with diabetes following a test.

You have been asked to develop a classification model to aid early detection of the disease. You should refer to the lecture notes for learning week 7 to help you complete the tasks.

TASK1

Create a new Google Collaborate notebook and load the dataset into a new Pandas data frame. Print out the first few rows of the data frame.

TASK2

For people diagnosed with diabetes, what is their median blood pressure?

TASK3

Identify the variable in the dataset that indicates whether a person was diagnosed with diabetes following a test. Is the data set balanced?

TASK4

Prepare your data for modelling by separating your dataset into two variables X and Y. X should contain features of the dataset used to predict diabetes. Y should contain the target variable.

TASK5

Explain the purpose of the following Python statements. What impact does "stratify=y" have?

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=
0.2, random_state=1, stratify=y)
```

TASK6

Adapt and modify the code from TASK5 so that a train-test split of 3:1 is obtained.

TASK7

Fit a DecisionTreeClassifier to your training dataset. Evaluate the performance of your trained model using the accuracy score. Comment on the result.

TASK8

Determine whether there is any evidence of model overfitting

TASK9

In the following screenshot of a classification report, what do the values “precision” and “recall” refer to?

```
Classification Report
              precision    recall  f1-score   support

           0       0.77      0.89      0.82       100
           1       0.71      0.50      0.59        54

    accuracy                           0.75       154
   macro avg       0.74      0.70      0.71       154
weighted avg       0.75      0.75      0.74       154
```

TASK10

Create a confusion matrix to analyse the performance of your DecisionTreeClassifier. What proportion of cases were false positives? How sensitive is your model to cases of diabetes?

TASK11

The following code can be used to create a visualisation of the decision tree classifier stored in the variable “dt”. Input this code into a new cell and generate a plot of the tree. **Hint:** you will need to modify the variable passed to plot_tree() so that it refers to the DecisionTreeClasssifier you created in TASK7

```
import matplotlib.pyplot as plt
from sklearn import tree
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (3,3), dpi=300)
tree.plot_tree(dt)
plt.show
```

TASK12

Use the output of the tree plot from task 10 to identify and rank patient characteristics that are most useful in predicting diabetes.

TASK13

Before your model can be used, the Health Service would like to compare its performance with an alternative model.

Fit a KNeighboursClassifier with 4 neighbours. Inspect the confusion matrix and/or classification report and make a recommendation as to which model performs best.

Clustering

In Learning Week 5, you were asked to advise US Geological Survey on weather and pollution patterns.

Having reviewed your work, the USGS are concerned that weather monitoring stations are not sufficiently geographically dispersed. Any insight you can offer will be invaluable.

As previously, you should use the file "pm25_2016_2020_v3.csv" to conduct your analysis.

TASK14

Load the weather pattern data into a new Pandas data frame. Identify **TWO** columns that could be used to analyse the geographic distribution of weather monitoring stations.

TASK15

Extract the two columns from the data frame and store them in a new data frame called X.

TASK16

Explain the purpose of the following Python statements. What will be the range of possible values in column "Cluster"?

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=6)
X["Cluster"] = kmeans.fit_predict(X)
X["Cluster"] = X["Cluster"].astype("category")
```

TASK17

Adapt the code in TASK16 so that the data you extract in TASK15 is clustered into 5 groups. How many observations are placed into each of the clusters?

TASK18

The following code can be used to print a graph of the clusters; however, it contains some errors. Modify the code, inputting values or x, y, hue and data to product a colour-coded visualisation of your clusters.

```
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
matplotlib.rcParams['figure.figsize'] = (12,8)

sns.relplot(x="", y="", hue="", data=, height=6)
```