INTRODUCTION

The dataset Classification.csv - it shows two different groups, plotted against two explanatory variables. This is simulated data - the aim is to find a suitable method for classifying the 1000 datapoints into two groups from a selection of possible approaches

The classification.csv is a generalised dataset consisting of 1000 rows and 3 columns namely X1, X2 , group

The rows will be further separated ,based on the column ‘group’ ending them with the values 0 and 1

Naming X1 and X2” under one dataframe so itll be easy to locate them together

And naming them as X and why where X containing the X1 X2 columns and Y containing the group which will make will be used for split and train

Out of which 25 percentage will taken for test and 75%will be taken for train. X1's median is a little higher to X2 even though the range of X1 is wider to X2, and the mean of X1 is centred a little lower to X2 . The distribution of observations between the two groups is not equal.

Group 2 has only 50 whereas Group 1 has around 150 observations ,so it is understood that the observations of two groups are not split evenly .Among LDA,QDA,RANDOM FOREST, and the SVM ,the models KNN was found to be the best fit model.

```
[[ 19  59]
 [ 14 158]]
              precision    recall  f1-score   support

           0       0.58      0.24      0.34        78
           1       0.73      0.92      0.81       172

    accuracy                           0.71       250
   macro avg       0.65      0.58      0.58       250
weighted avg       0.68      0.71      0.67       250

Accuracy0.708
```

- LINEAR DISCRIMINANT ANALYSIS

Using axes that maximise the linear saliency between various classes of data, LDA is a type of supervised learning. For example, if there are x variables, LDA reduces the variable to ...y variable to reduce the dimensionality. LDA also removes variables that aren't independent or important

Before building the LDA model we FIT the standard scalar model

With StandardScaler, your data will be transformed to have a distribution with a standard deviation 1 and a mean value 0. This is carried out feature-by-feature.For multivariate data,

- There are many ways to figure out the best model once you get your models summary information,after fitting the standard scaler model , then we fit the Linear Discriminant model to conduct test and train,prediction for the fitted Lda model, proceeding with the confusion matrix for the predicted value,then we get the summary ,the fi-score of each model can be compared to finalize the best fit model,where the f1-score for LDA model is 0.34 and 0.81

Quadrative Discrimnant Analysis

```
[[ 29  49]
 [ 12 160]]
              precision    recall  f1-score   support

           0       0.71      0.37      0.49        78
           1       0.77      0.93      0.84       172

    accuracy                           0.76       250
   macro avg       0.74      0.65      0.66       250
weighted avg       0.75      0.76      0.73       250

Accuracy0.756
```

- On the classification.csv , we will now run Quadrative Discriminant Analysis.
- The Quadrative DiscriminantAnalysis() method of the sklearn library's d iscriminant analysis module can be used to fit an QDA model in Python
- The only significant difference between linear and quadratic discriminant analyses is the relaxation of the premise that the covariance and mean of all classes are equal. Consequently, we had to calculate it independently.
- After fitting the QDA and running the test and train , confusion matrix to check the accuracy score for the model we check the f1score to see if this model fits better than LDA
- Precision=0.71 and 0.77,recall= 0.37and 0.93,f1-score=0.49 and 0.84

- By comparing we could see that QDA model Has an high f1-score than LDA

K NEAREST NEIGHBOUR

```
[[ 38  40]
 [ 23 149]]
              precision    recall  f1-score   support

           0       0.62      0.49      0.55        78
           1       0.79      0.87      0.83       172

    accuracy                           0.75       250
   macro avg       0.71      0.68      0.69       250
weighted avg       0.74      0.75      0.74       250

Accuracy0.748
```

- To address the categorization model difficulties, this algorithm is utilised. The K-nearest neighbour technique, also known as K-NN, essentially draws an illogical boundary to categorise the data. The programme will attempt to anticipate additional data points to the nearest boundary line as they are received.
- The primary determining element is the number of neighbours..K is the quantity of closest neighbours in KNN. Generally speaking, if there are two classes, K will be an odd number. The algorithm is referred to as the nearest neighbour algorithm when K=1.
- Assume a dataset with K=5.Finding the vlaues for the brand new piece of data based on the weighted average of the five closest points is the task while performing classification.
- Precision=0.62 and 0.79,recall= 0.49 and 0.87,f1-score=0.55 and 0.85

The f1-score is observed here to compare them with the Lda and Qda

- By far comparing the models KNN modle has the best f1 score 0.49 and 0.87

Standard Vision Method

```
[[ 26  52]
 [ 10 162]]
```

	precision	recall	f1-score	support
0	0.72	0.33	0.46	78
1	0.76	0.94	0.84	172
accuracy			0.75	250
macro avg	0.74	0.64	0.65	250
weighted avg	0.75	0.75	0.72	250

Accuracy0.752

- By selecting the boundary that measures the difference from the closest data sets of all classes, SVM differentiates from previous classification techniques. An SVM seeks the best decision boundary,it doesn't just discover a decision boundary
- .The choice boundary with the largest margin from all classes' closest points is the most ideal one. Support vectors are the closest points to the decision border that optimise the space between the points and decision boundary

- You must scikit-learn model class must be imported in order to train the first support vector machine model.The svm module of Scikit-Learn contains the SVC class.
- The SVC class's predict function is used to generate forecasts.
- The most commonly utilized metrics in terms of classification problems include confusion matrices, , F1,recall, precision, measures. The classification report and confusion matrix methods are available in the Scikit-Learn metrics library, making it simple to determine the datas for such crucial metrics
- Precision=0.72 and 0.76,recall= 0.33 and 0.96,f1-score=0.46 and 0.84 still less than the KNN model

Random Forest Classification

	precision	recall	f1-score	support
0	0.59	0.45	0.51	78
1	0.77	0.86	0.82	172
accuracy			0.73	250
macro avg	0.68	0.65	0.66	250
weighted avg	0.72	0.73	0.72	250

```
[[ 35  43]
 [ 24 148]]
Accuracy0.732
```

- A supervised machine learning technique that uses ensemble learning is known as random forest. In order to create a more effective prediction model, you can combine several kinds of algorithms or use the same technique more than once in ensemble learning.
- Fact that the random forest method mixes several algorithms of the same type, or different decision trees, into a forest of trees. Both classification tasks and regression can be performed using the random forest approach.
- Since there are several tree ,each and every trees are trained using a portion of data, the random forest approach is not biassed. In essence, the random forest depends on "the crowd's" power, which lessens the system's overall bias.The algorithm is remarkably reliable. Since it is exceedingly difficult for fresh data to have an impact on all the trees, if an updated data is added to the dataset, the overall method is not significantly changed.In situations where there are multiple numerical and categorical variables, the random forest approach performs well.
- We'll employ the RandomForestClassifier class from the sklearn.ensemble library to perform classification
- Precision=0.59 and 0.77,recall= 0.45 and 0.86,f1-score=0.51 and 0.82 The f1 –score observed here are 0.51 and 0.82

CONCLUSION

```
Best leaf_size: 1
Best p: 2
Best n_neighbors: 19
              precision    recall  f1-score   support

           0       0.60      0.44      0.50        78
           1       0.77      0.87      0.82       172

    accuracy                           0.73       250
   macro avg       0.68      0.65      0.66       250
weighted avg       0.72      0.73      0.72       250

Accuracy0.732
```

- Grid search cross-validation creates several models using various combinations of hyperparameters, then evaluates the performance of each combination.

- The challenging process of creating and assessing models with various combinations of hyperparameters is handled by Sklearn's GridSearchCV() algorithm.
- A parameter that cannot be estimated from data is referred to as a hyperparameter. Before a model starts its learning process, a hyperparameter's value must be set.
- A tool called grid search constructs a model for each combination of hyperparameters that we specify, then assesses each model to determines hyperparameters results whichever pairs in the best model
- Let's attempt using hyper-parameter tune to enhance the model performance as it is still low.
- GridSearch demonstrates p = 2, is the best distance algorithm, whereas the best leaf size = 1.So K should be as high as possible, which is 19
- Even after performing the Hyper parameters tunning for the best fit KNN model ,the model has only performed poor even before tunning, even though there is not much difference but still considering the poor performance after tunning , KNN model can concluded the best fit model