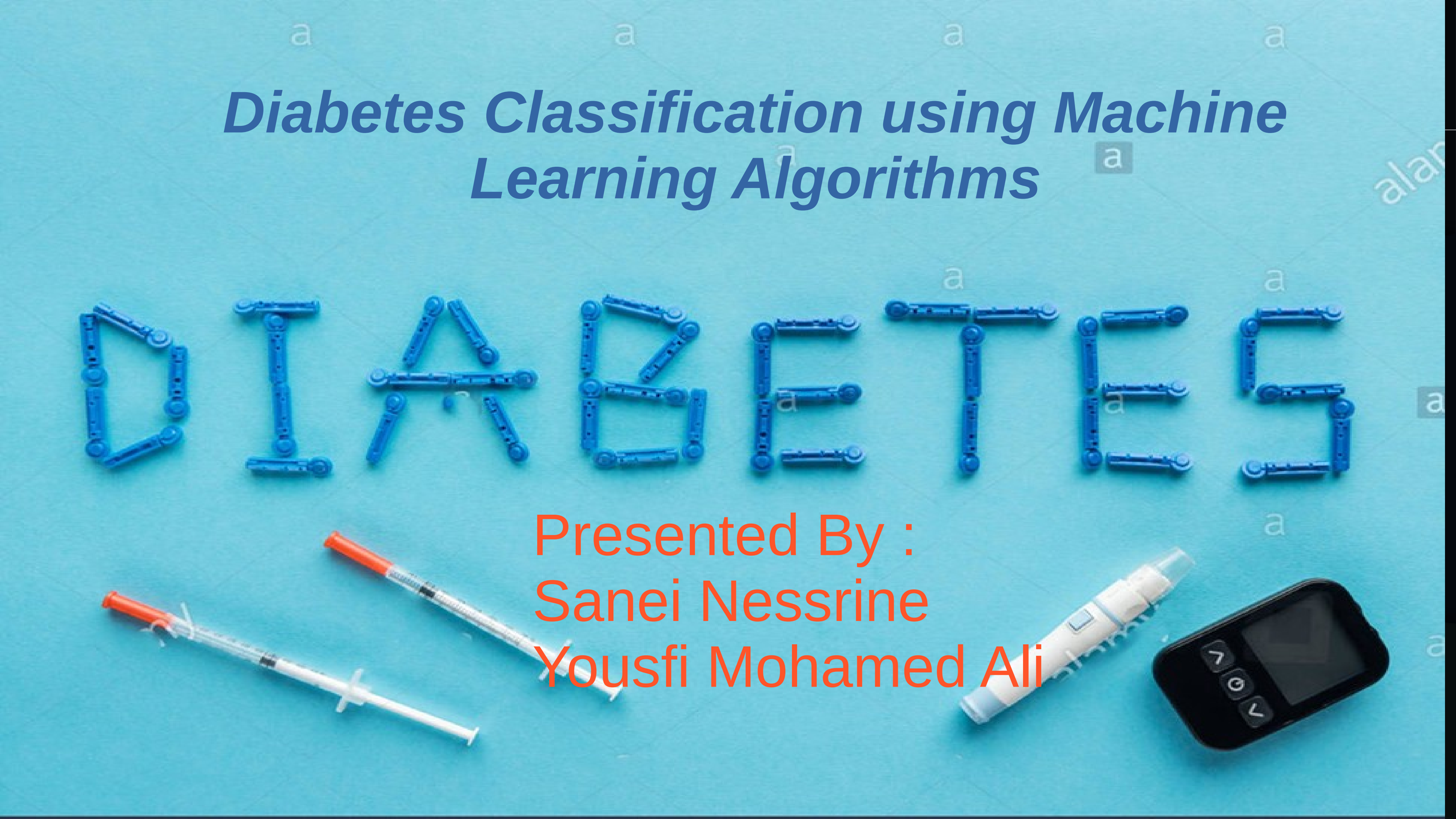
Diabetes Classification using Machine Learning Algorithms



Presented By :
Sanei Nessrine
Yousfi Mohamed Ali

PLan

Overview

Database - Pima Indians Diabetes Dataset

Methodology

Algorithms Comparison & Results Interpretation

Conclusion

Overview

The Pima indians (Akimel Oodham) of Arizona have the highest rate of diabetes of any population in the world.

Database - Pima Indians Diabetes Dataset

- This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.
- Pima Indian Diabetes dataset has 9 attributes in total.
- All the person in records are females.

Attribute	Description	Value
Pregnancies	Number of pregnancies	[0-17]
Glucose	Plasma glucose concentration	[0-199]
Blood Pressure	Diastolic blood pressure	[0-122]
SkinThickness	Triceps skin fold thickness	[0-99]
Insulin	2-Hour serum insulin	[0-846]
Body Mass	Body mass index	[0-67]
Pedigree	Diabetes pedigree function	[0-2.45]
Age	Age of an individual	[21-81]
Outcome	Tested +/-	{0,1}

Methodology

- Used 3 algorithms of supervised Learning :

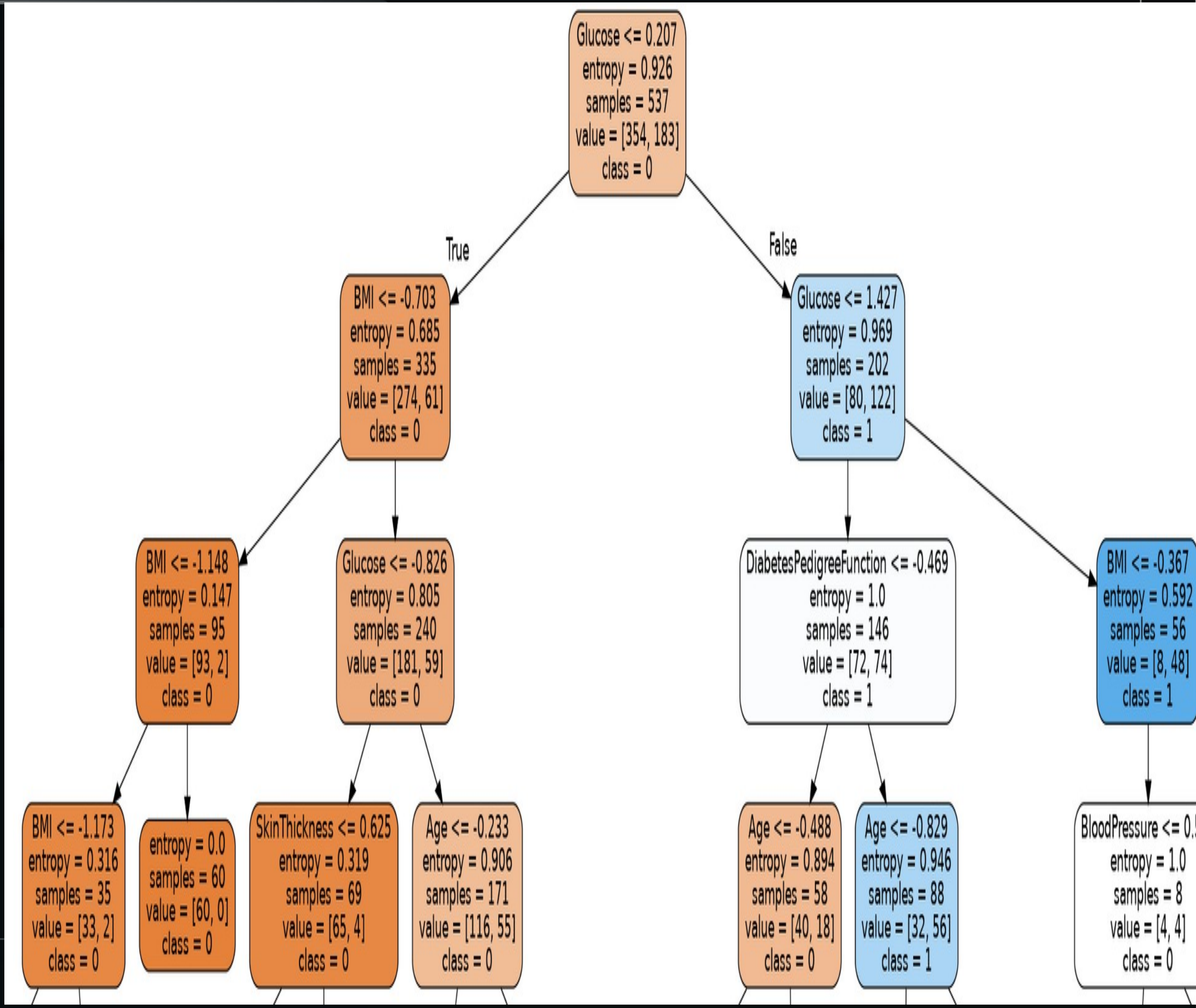
  * Decision Tree

  * Random Forest

  * K nearest neighbors

Decision Tree

- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
- The attribute/feature best for set is taken as root “Glucose”



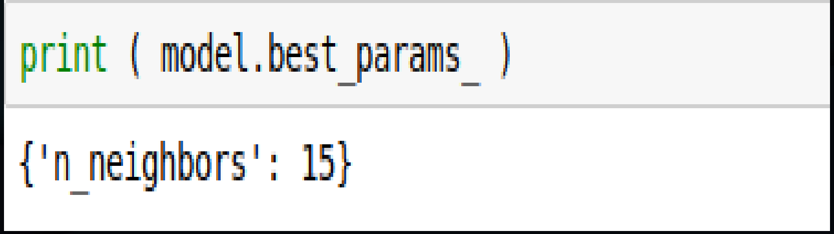
Random Forest

- Random Forest is a tree-based machine learning algorithm that leverages the power of multiple decision trees for making decisions

- Each node in the decision tree works on a random subset of features to calculate the output. The random forest then combines the output of individual decision trees to generate the final output.

K Nearest Neighbor

- KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest K neighbors.

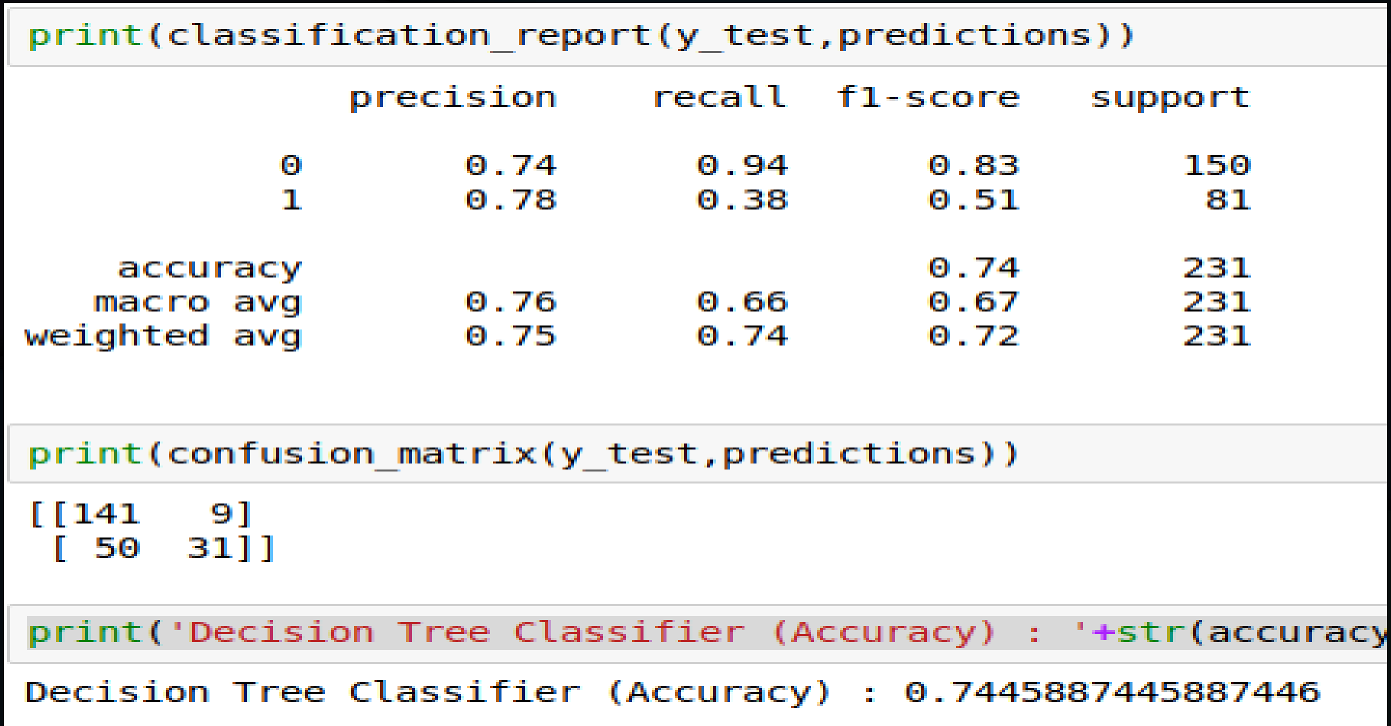


```
print ( model.best_params_ )
```

```
{'n_neighbors': 15}
```

Algorithms Comparison & Results interpretation

- DECISION TREE



```
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       0.74      0.94      0.83       150
           1       0.78      0.38      0.51        81

    accuracy                           0.74       231
   macro avg       0.76      0.66      0.67       231
weighted avg       0.75      0.74      0.72       231
```

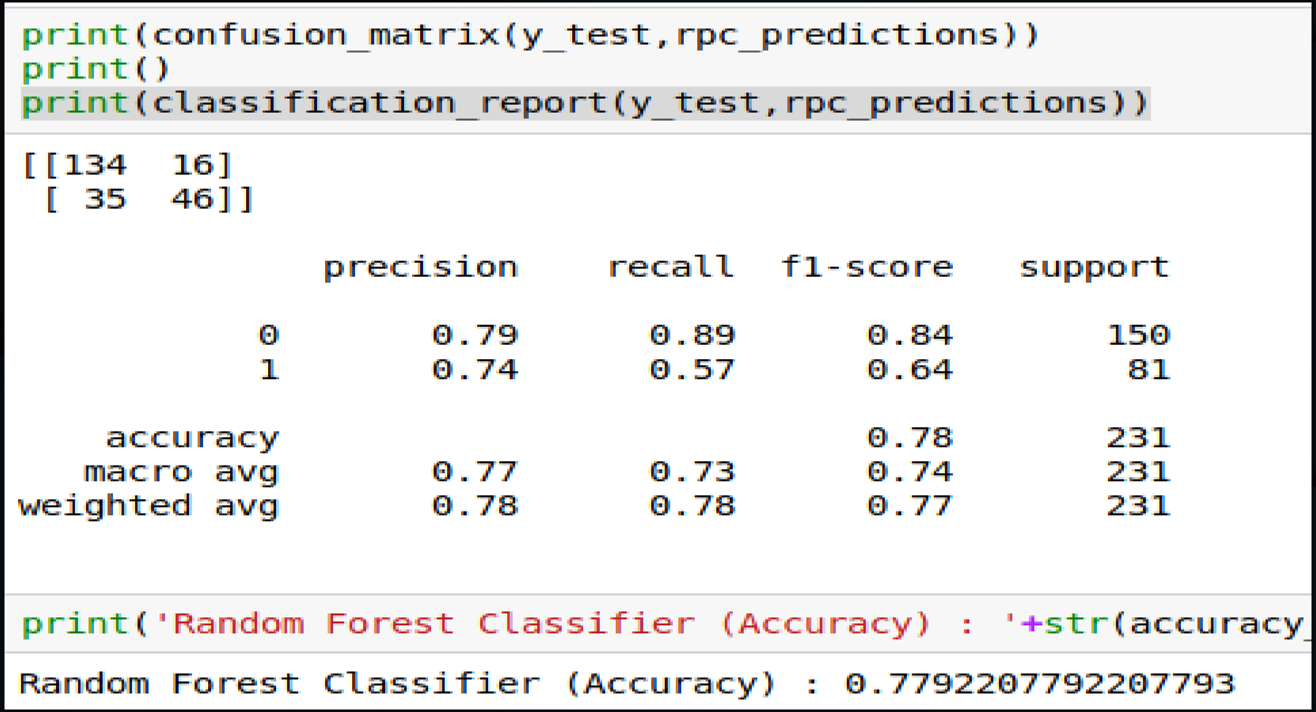
```
print(confusion_matrix(y_test,predictions))
```

```
[[141   9]
 [ 50  31]]
```

```
print('Decision Tree Classifier (Accuracy) : '+str(accuracy
```

```
Decision Tree Classifier (Accuracy) : 0.7445887445887446
```

- RANDOM FOREST



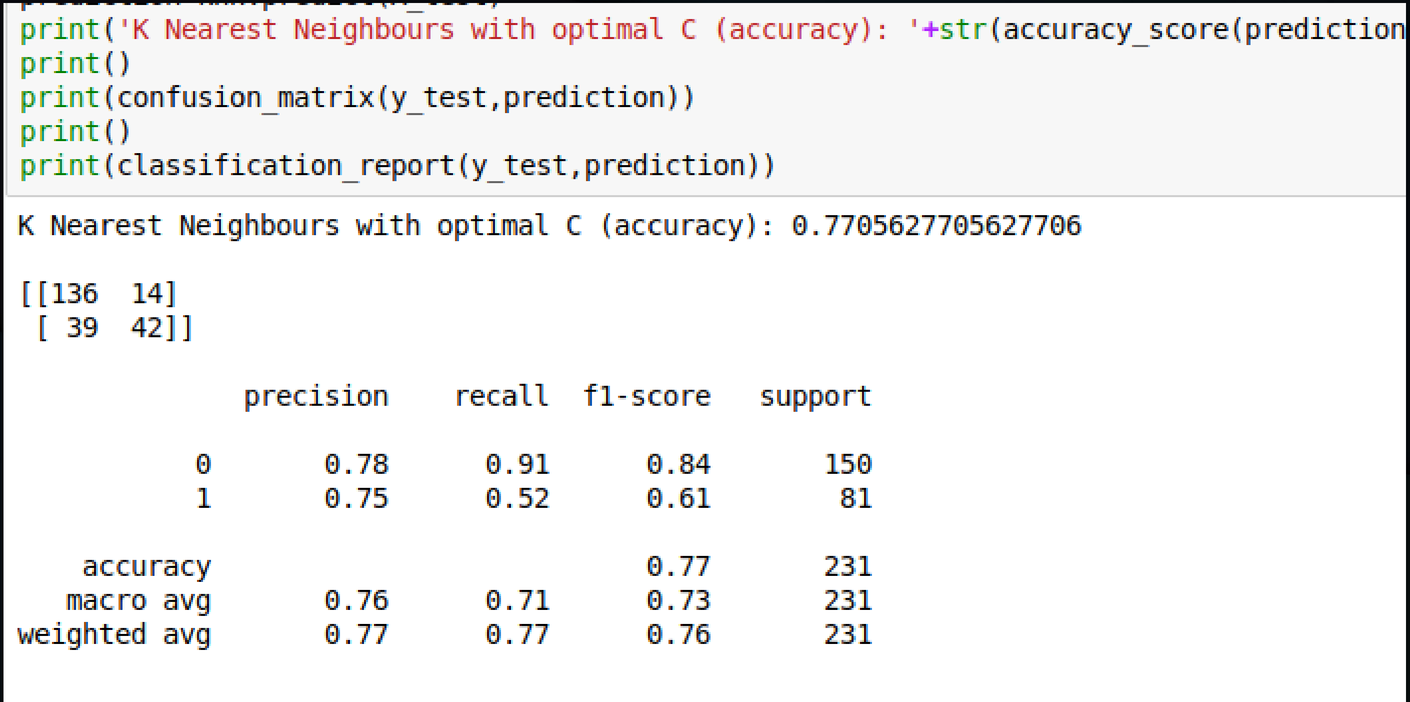
```
print(confusion_matrix(y_test,rpc_predictions))
print()
print(classification_report(y_test,rpc_predictions))
```

```
[[134  16]
 [ 35  46]]

              precision    recall  f1-score   support

           0       0.79      0.89      0.84       150
           1       0.74      0.57      0.64        81

    accuracy                           0.78       231
   macro avg       0.77      0.73      0.74       231
weighted avg       0.78      0.78      0.77       231
```

```
print('Random Forest Classifier (Accuracy) : '+str(accuracy_
```

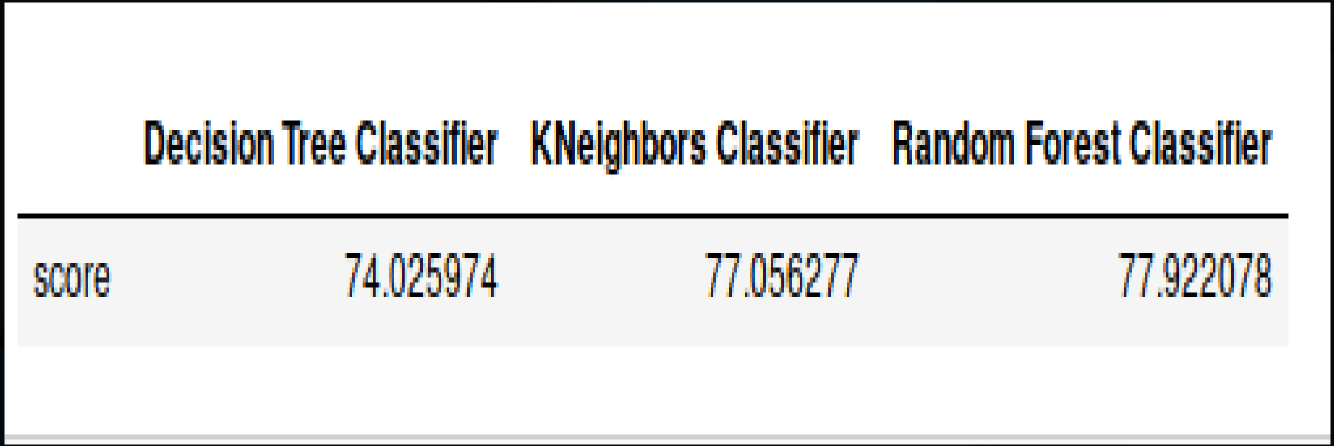
```
Random Forest Classifier (Accuracy) : 0.7792207792207793
```

- K-NN



```
print('K Nearest Neighbours with optimal C (accuracy): '+str(accuracy_score(prediction
print()
print(confusion_matrix(y_test,prediction))
print()
print(classification_report(y_test,prediction))
```

```
K Nearest Neighbours with optimal C (accuracy): 0.7705627705627706

[[136  14]
 [ 39  42]]

              precision    recall  f1-score   support

           0       0.78      0.91      0.84       150
           1       0.75      0.52      0.61        81

    accuracy                           0.77       231
   macro avg       0.76      0.71      0.73       231
weighted avg       0.77      0.77      0.76       231
```



	Decision Tree Classifier	KNeighbors Classifier	Random Forest Classifier
score	74.025974	77.056277	77.922078

Conclusion

- 77,92% accuracy rate provided with Random Forest .
- Random Forest (at the right parameters) can be a good choice and practical to classify a medical data .

Thank You

For Your Atten