ASSIGNMENT 4 REPORT

—

Team Mates:

B Krupa Kiranmai, *2021201022*
N C S Jagannath, *2021201024*

Penguin Species Classification:

1.Building of classifier:

The task at hand is to build a machine learning model to classify the type of the penguin with the given features of the penguin.

Dataset:

The dataset consists of 9 attributes with the target variable “Species”.Out of the nine 6 were numerical and 3 were categorical attributes.

I. Island object
II. Clutch Completion object
III. Culmen Length (mm) float64
IV. Culmen Depth (mm) float6
V. Flipper Length (mm) float64
VI. Body Mass (g) float64
VII. Sex object
VIII. Delta 15 N (o/oo) float64
IX. Delta 13 C (o/oo) float64
X. Species object

The target attribute consists of three classes namely:

1. Gentoo penguin (Pygoscelis papua)
2. Adelie Penguin (Pygoscelis adeliae)
3. Chinstrap penguin (Pygoscelis antarctica)

Data Preprocessing:

- First analyzed the data for the potential outliers where no outlier was found.
- In the next step we filled up the missing values in the dataset.
- For the categorical attribute we added a value “missing” to the values that were null.

- For the numerical attribute we impute the mean of the particular attribute for the null values as all the values are uniformly distributed.
- Added noise data for the given data as part of the data augmentation.

Model construction and analysis:

- The data was being split into train, valis and test where the valid, test consists of 10% of data.
- The train data then is split as per the class it belongs as we have to build the all vs all classifier
- Then build three classifiers for each pair of the class that was possible with each of the three algorithms.
- Then calculated f1 scores and accuracy score on the test data that we kept aside and conducted the required analysis.
- For running the classifier.py file we have to give the paths of both train and test data.

2.Algorithm that performs better:

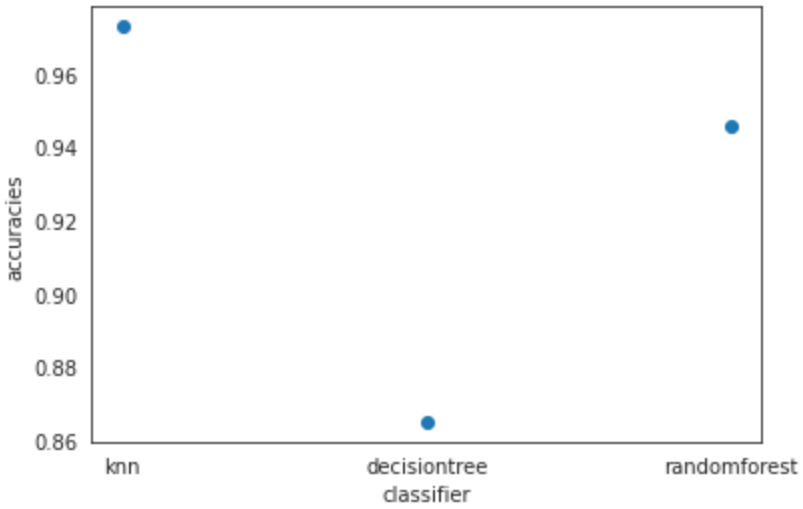
We build the all vs all multi label classifier of three classifiers namely knn ,decision tree and random forest. The accuracy scores on the test data were

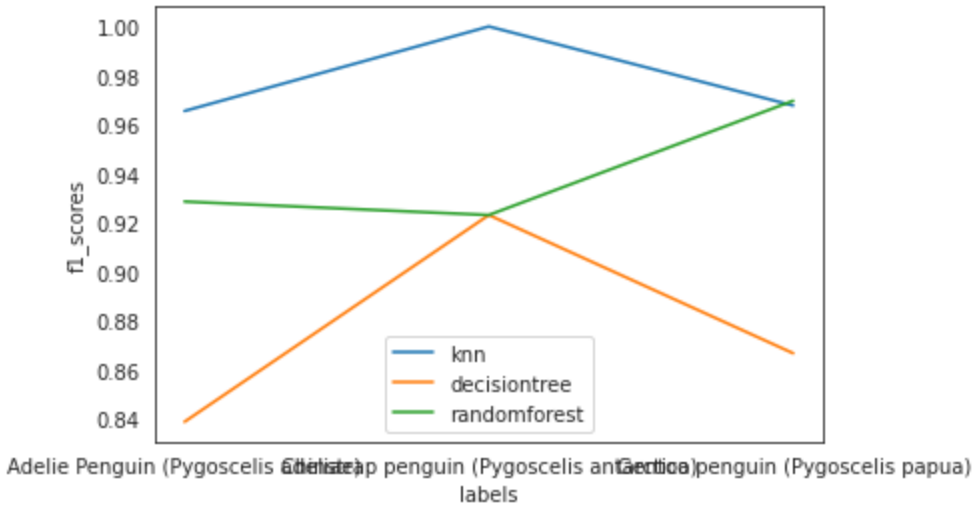
- KNN - 0.972972972972973
- Decision Tree - 0.8648648648648649
- Random Forest - 0.9459459459459459

The f1 scores on the test data were:

- KNN - [0.96551724 1. 0.96774194]
- Decision Tree - [0.83870968 0.92307692 0.86666667]
- Random Forest - [0.92307692 0.93333333 0.96969697]

The graphs of the above data are below:





- Out of the three classifiers the decision tree performs worse than the random forest comparably performs equally.
- As the data is too small the complex algorithm like random forest tends to overfit the data so KNN will be the best classifier.

3.Multi Label Classification strategies:

- The task that was given at the hand is to classify penguin type among the three classes given with the help of its characteristics.
- Generally classification done on the binary valued target variables ,we follow certain strategies which use binary classifiers to classify the multiple labels.
- One vs all and all vs all were two strategies that were defined.
- One vs all:
    - The One-vs-all strategy splits a multi-class classification into one binary classification problem per class.
    - Here we build three classification problems:
        - Gentoo penguin (Pygoscelis papua) vs (remaining two as not)
        - Adelie Penguin (Pygoscelis adeliae) vs (remaining two as not)
        - Chinstrap penguin (Pygoscelis antarctica) vs (remaining two as not)
    - This approach requires that each model predicts a class membership probability or a probability-like score. The argmax of these scores (class index with the largest score) is then used to predict a class.
- All vs All:
    - The All-vs-All strategy splits a multi-class classification into one binary classification problem per each pair of classes.
    - Here we build three classification problems and the dataset that corresponds to the pair of class:
        - Gentoo penguin (Pygoscelis papua) vs Adelie Penguin (Pygoscelis adeliae)
        - Gentoo penguin (Pygoscelis papua) vs Chinstrap penguin (Pygoscelis antarctica)
        - Adelie Penguin (Pygoscelis adeliae) vs Chinstrap penguin (Pygoscelis antarctica)
    - Each binary classification model may predict one class label and the model with the most predictions or votes is predicted by the all-vs-all strategy.

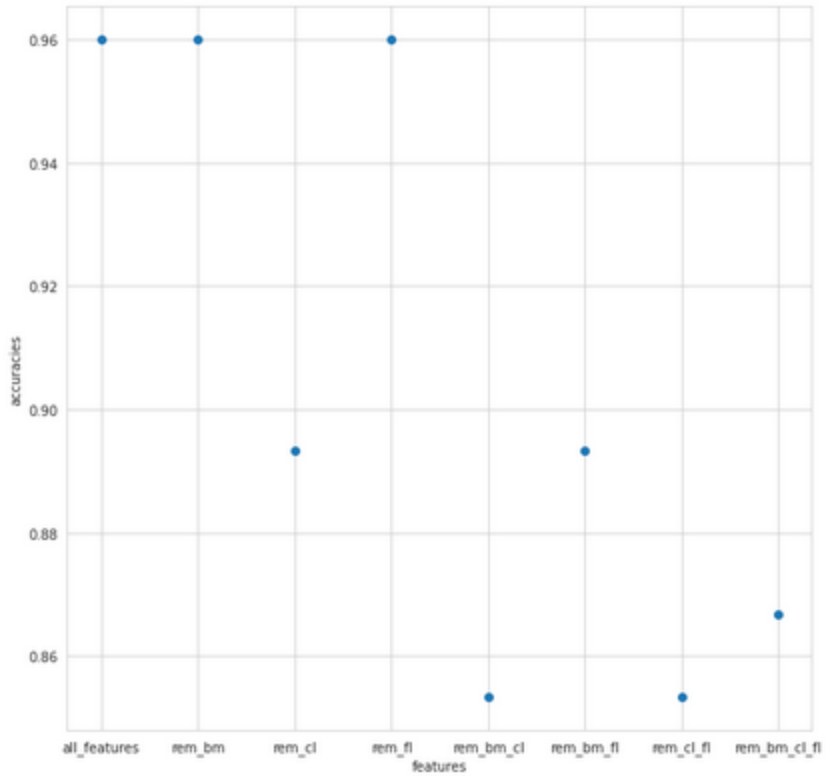
4.Ways to tackle the low Training Data :

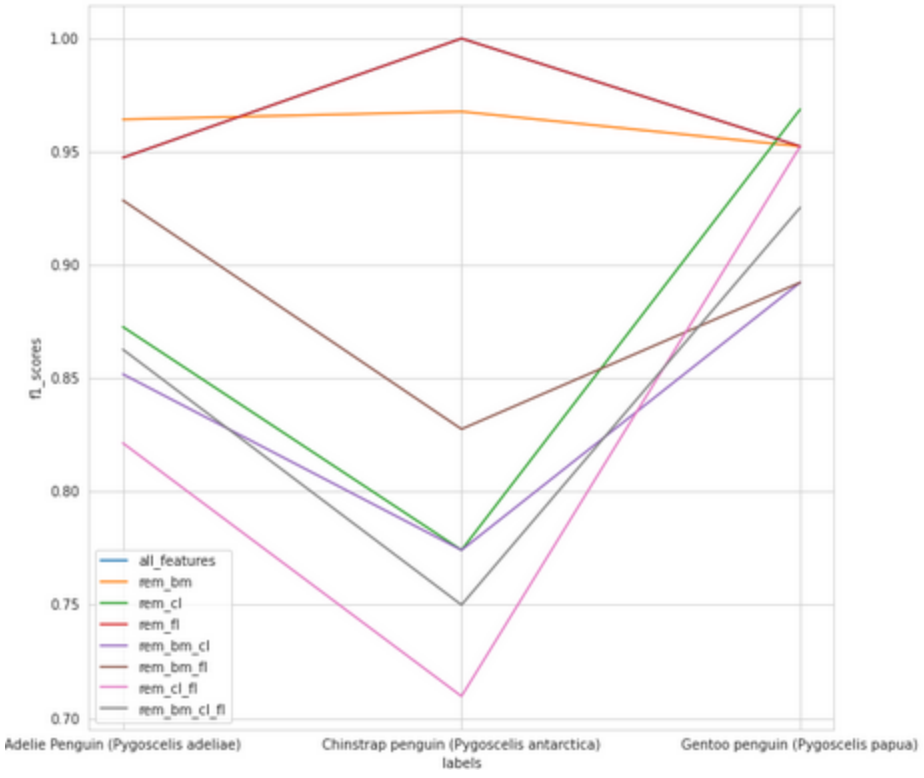
- We did one technique of data augmentation i.e generation of synthetic data.
- We added noise data(synthetic data) to the given dataset. That was we generate fake data that was near to the original data.
- The ensemble method like random forest may overfit compared with the knn.So use of KNN while training will be better.

5.Feature Selection/engineering :

- We analyzed the constructed KNN classifier for the setting of different number of features.
- As the dataset was small the attributes that are more correlated to the target allows the classifier to overfit the data.
- So we analyzed the effect of removal of such attributes.The attributes that were more correlated were body mass,culmen length,flipper length.
- The results were shown below.
- Accuracies:
    - With all attributes - 0.972972972972973
    - Removal of body mass - 0.96
    - Removal of culmen length - 0.8933333333333333
    - Removal of flipper length - 0.96
    - Removal of body mass and culmen length - 0.8533333333333334
    - Removal of body mass and flipper length - 0.8933333333333333
    - Removal of culmen length and flipper length - 0.8533333333333334
    - Removal of three attributes - 0.8666666666666667

Corresponding graph:



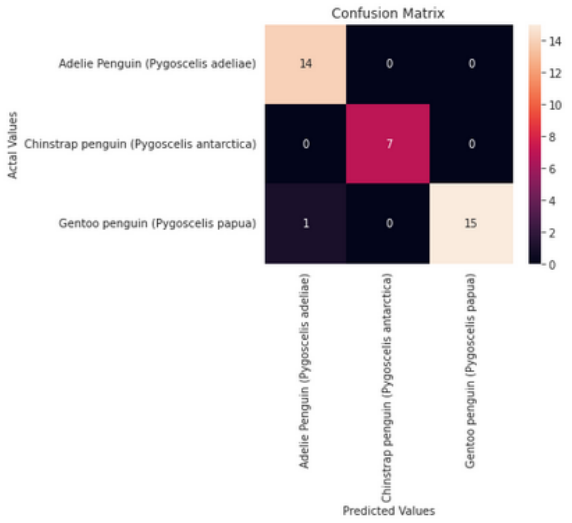


Above was the analysis of feature selection. We thought that there were no meaningful features that could be engineered for the dataset.

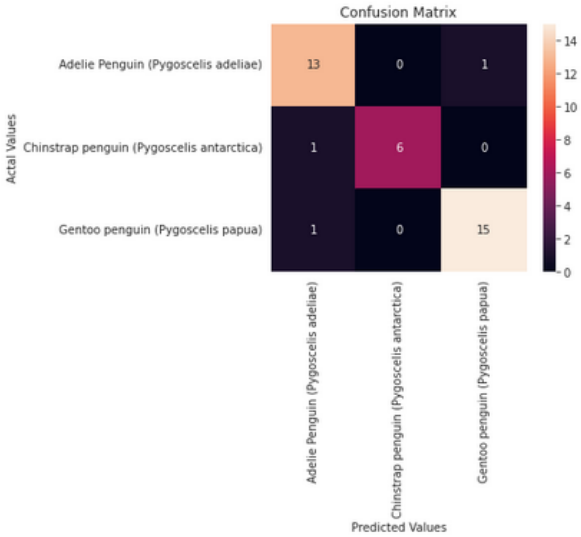
6.Error Metrics :

Confusion Matrix: The confusion matrix provides more insight into not only the performance of a predictive model but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made.

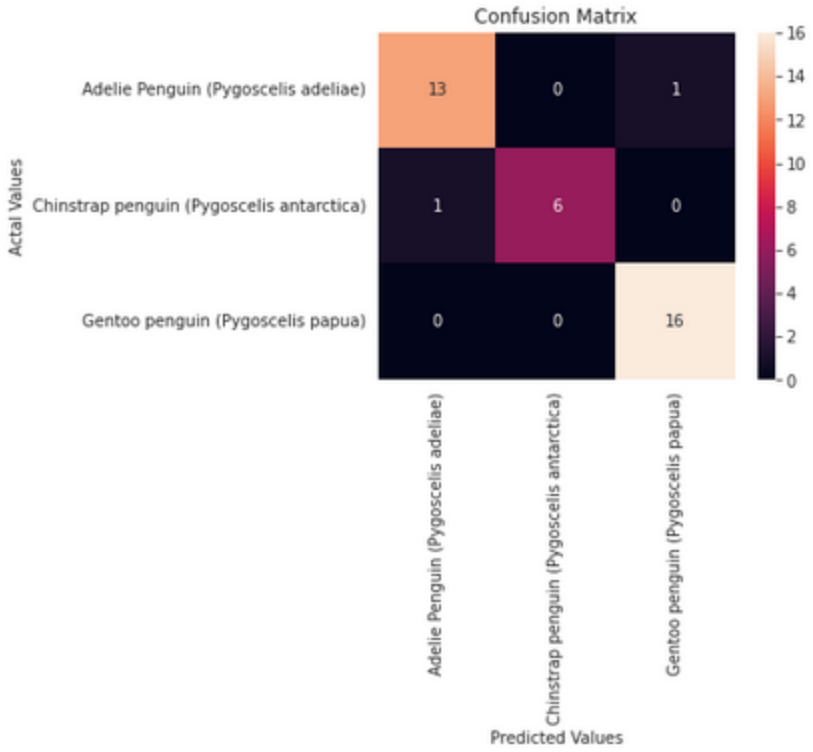
KNN :



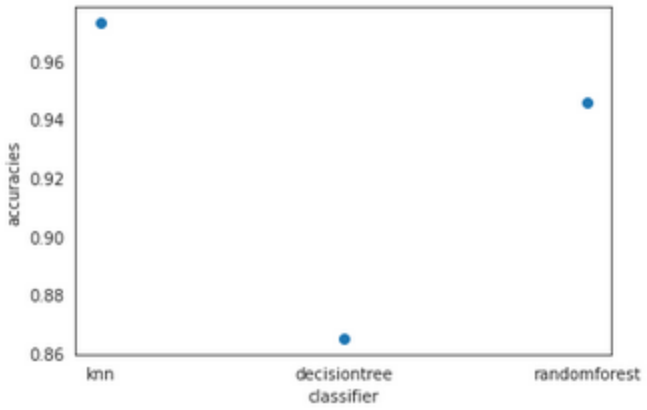
Decision Tree :

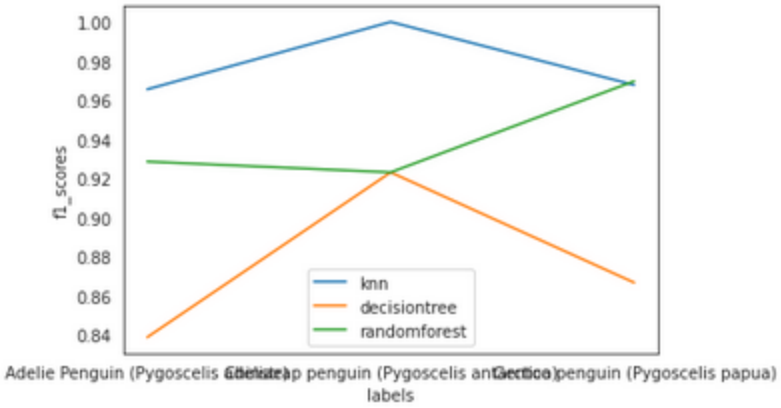


Random forest :



F1-score : The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. F1 score sort of maintains a balance between the precision and recall for your classifier. If your precision is low, the F1 is low and if the recall is low again your F1 score is low.



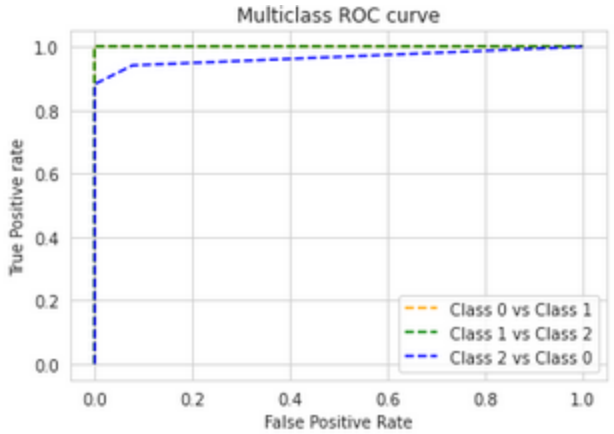


AUC - ROC curve : A ROC curve is a diagnostic plot for summarizing the behavior of a model by calculating the false positive rate and true positive rate for a set of predictions by the model under different thresholds. The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. When 0.5 < AUC < 1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.

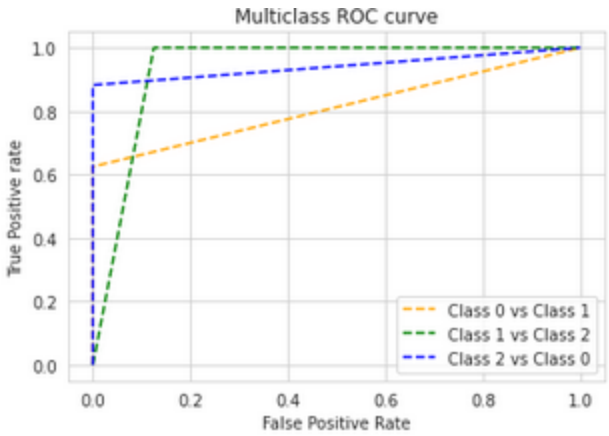
KNN :

Areas : 1.0, 1.0, 0.9660633484162896



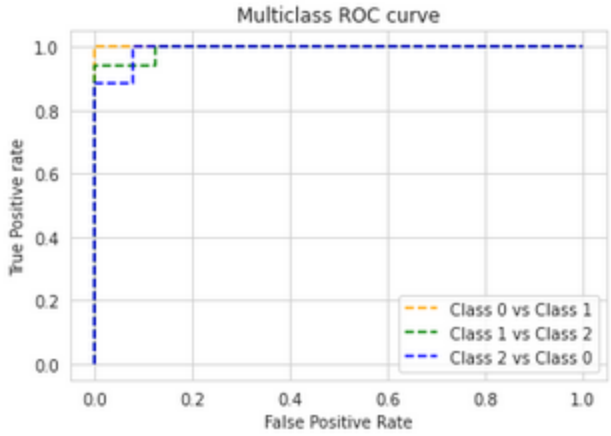
Decision Tree :

Areas : 0.9375, 0.9375, 0.9411764705882353



Random forest :

Areas: 0.9926470588235294, 0.9926470588235294, 0.9909502262443439



Order of best metrics that can be used for this problem

1. **ROC - AUC score** can be used to **select the best classifier** as it gives the score whether the classifier is correctly able to distinguish the positive class and negative class.
2. **F1 - Score :** We need **all the classes to be predicted correctly**. As f1-score maintains a balance between precision and recall it can be used as one of the error metrics for this problem.
3. **Confusion Matrix :** It can be **used to visualize TP, TN, FP, FP.**