Data Analytics 1

Assignment 4 Report

P Shiridi Kumar
Karnati Jahnavi

2) We have Experimented with multiple models (SVM , Decision trees , Logistic regression , Random Forest , KNN) , and we noticed that all the models are equally performing well, but **SVM** being the one which has a slight higher accuracy most of the times (after experimenting several times) .
The accuracies for the models are: (test size =0.3 ,82 tuples)

DecisionTree Accuracy : 0.9512195121951219
SVM Accuracy : 0.975609756097561
RandomForest Accuracy : 0.975609756097561
LogisticRegression Accuracy : 0.9634146341463414
KNN Accuracy : 0.9512195121951219

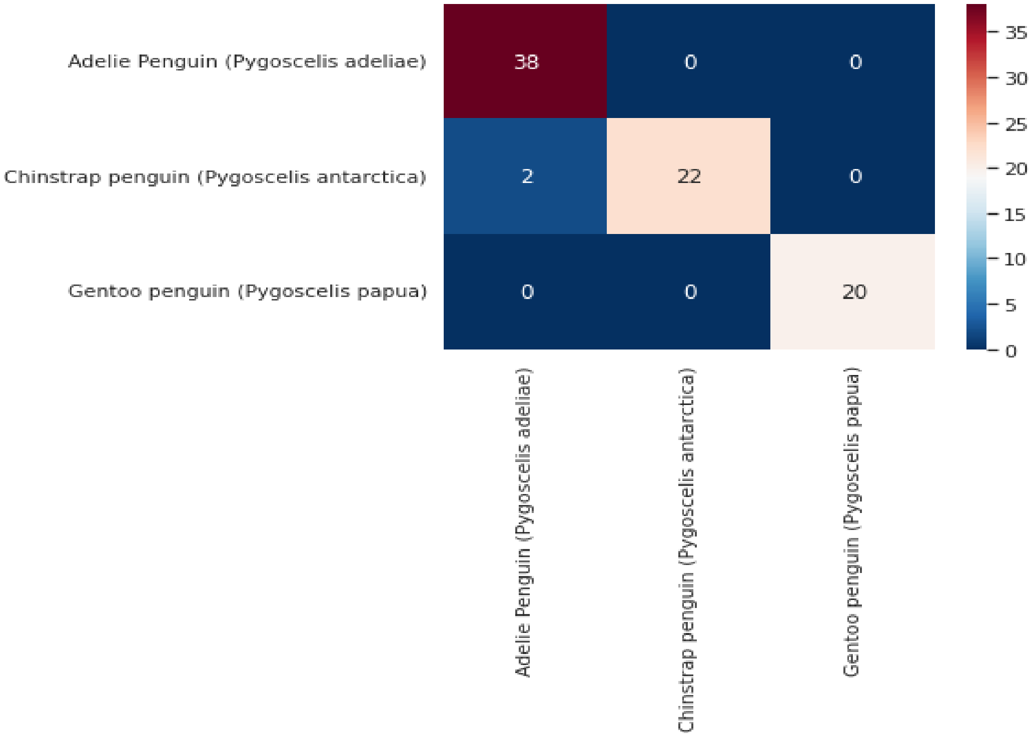
3) In One vs ALL we build n classifiers if there are n classes , each classifier is trained in such a way that it tries to classify between that one specific class and the rest of all the classes . After training n classifiers we just take voting of all classifiers and output the class with maximum votes. Where as in ALL vs ALL , we build a classifier between each pair of classes (i.e, if there are n classes we will have nC2 classifiers) , at the end we again take voting and return the class with maximum number of votes as output

4) SMOTE : We have experimented our dataset by augmenting the given data using SMOTE imbalanced dataset augmenting technique.
SMOTE (Synthetic Minority Oversampling Technique) is used for augmenting the data for minority classes in the dataset(in our dataset we found Chinstrap penguin (Pygoscelis antarctica) to be the minority class while experimenting) , by synthetically generating data using its neighbor data points of the same label class. The entire process , helps us to generate more data.

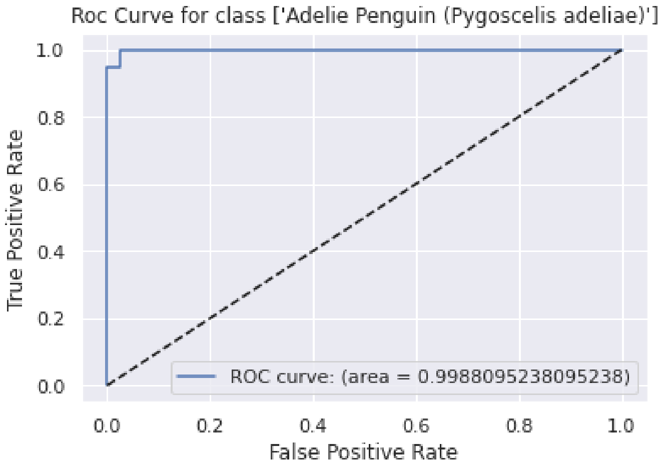
5) Apart from regular pre processing , which includes converting categorical to numeric by encoding them , we Normalized the continuous variable in the range of 0 to 1(feature

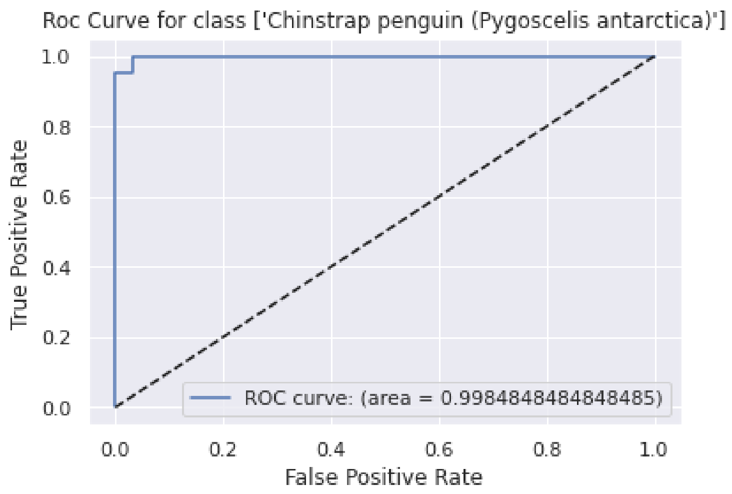
engineering) , before normalizing the data we have found the accuracy to be ~60% but after normalizing we found it to be 98%(approx). We have used Principal component analysis(5 Components) for feature selection , but we didn't find any further improve since the results were already good even before using PCA

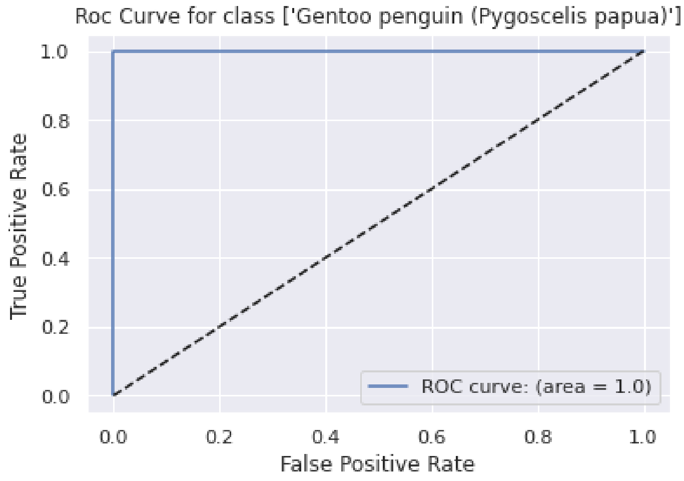
6) Error Plots :

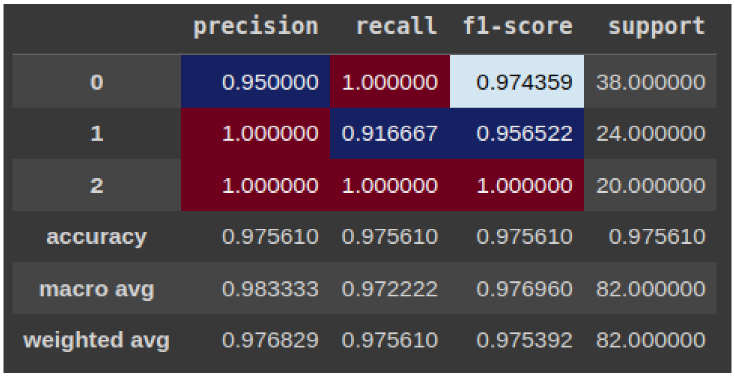


Confusion Matrix (SVM classifier on Test data)









	precision	recall	f1-score	support
0	0.950000	1.000000	0.974359	38.000000
1	1.000000	0.916667	0.956522	24.000000
2	1.000000	1.000000	1.000000	20.000000
accuracy	0.975610	0.975610	0.975610	0.975610
macro avg	0.983333	0.972222	0.976960	82.000000
weighted avg	0.976829	0.975610	0.975392	82.000000

F1 scores and classification report

The Best Error metric for the given data seems to be f1-score which captures the notion of both recall and precision. And since the number of data points are less , precision becomes important , especially for minority classes , as they are already less (Eg : if there is only 2 data points which are of class 2 and even if we don't predict one of the to be correct our recall falls to 0.5 for that specific class).And since our output number of classes are so less there might be a chance that model tries to classify the given sample always into one class(in that case precision might be 1 for that class) and we need to analyze recall as well in this case to maintain balance . So in order to inculcate and capture both precision and recall information we can use F1 score.