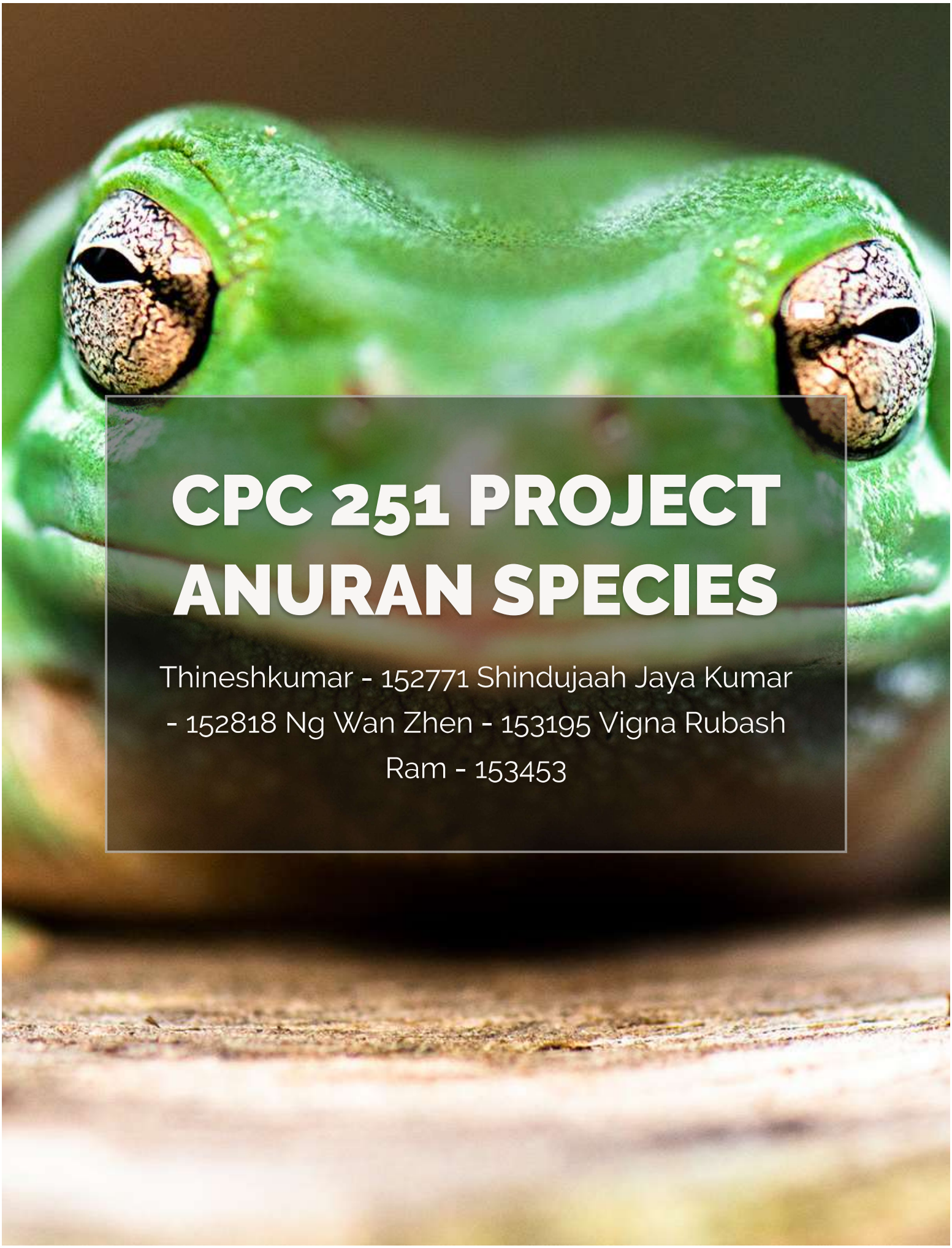

CPC 251 PROJECT ANURAN SPECIES

Thineshkumar - 152771 Shindujaah Jaya Kumar - 152818 Ng Wan Zhen - 153195 Vigna Rubash Ram - 153453

INTRODUCTION

Frogs are a group of small, tailless amphibians with a short body. They are mostly carnivorous and have a lot of different types. They are usually called Anura, which means "without a tail" in Ancient Greek. Frogs can be found all over the world, from the tropics to the subarctic. The most species can be found in tropical rainforests, but there are many more frogs in the subarctic. Frog populations worldwide have been quickly dropping. Habitat loss, alien species, and climate change are all being blamed for the decline.

To study the development of the frog population and enhance its protection strategy, it is necessary to amass information about frogs and their environment. Various feature reduction approaches and classifiers were investigated in those research for frog classification. Mel-Frequency Cepstral Coefficients (MFCCs) are a well-known characteristic for identifying frog sounds. The dataset was utilized in a variety of classification tasks linked to the difficulty of recognizing anuran species based on their sounds. Some species are from the campus of Federal University of Amazonas, Manaus, others from Mata Atlântica, Brazil, and one of them from Córdoba, Argentina.

DATASET DESCRIPTION & AIM

The recordings were stored in wav format with 44.1kHz of sampling frequency and 32bit of resolution, which allows us to analyze signals up to 22kHz. In this project, we seek to find if machine learning can be a useful tool in classifying frog calls and also to find which feature reduction technique is used in machine learning that can reduce the number of features or reduce data dimensions before classification without losing majority of the data.

SAMPLES OF DATASET

This dataset was created segmenting 60 audio records belonging to 4 different families, 8 genus, and 10 species. Each audio corresponds to one specimen (an individual frog), the record ID is also included as an extra column. We used the spectral entropy and a binary cluster method to detect audio frames belonging to each syllable. The segmentation and feature extraction were carried out in Matlab. After the segmentation we got 7195 syllables, which became instances for train and test the classifier.

	MFCCs_1	MFCCs_2	MFCCs_3	MFCCs_4	MFCCs_5	MFCCs_6	MFCCs_7	MFCCs_8	MFCCs_9	MFCCs_10
0	1.0	0.152936	-0.105586	0.200722	0.317201	0.260764	0.100945	-0.150063	-0.171128	0.124676
1	1.0	0.171534	-0.098975	0.268425	0.338672	0.268353	0.060835	-0.222475	-0.207693	0.170883
2	1.0	0.152317	-0.082973	0.287128	0.276014	0.189867	0.008714	-0.242234	-0.219153	0.232538
3	1.0	0.224392	0.118985	0.329432	0.372088	0.361005	0.015501	-0.194347	-0.098181	0.270375
4	1.0	0.087817	-0.068345	0.306967	0.330923	0.249144	0.006884	-0.265423	-0.172700	0.266434

Figure 1: Dataset sample from MFCCs_1 to MFCCs_10

...	MFCCs_14	MFCCs_15	MFCCs_16	MFCCs_17	MFCCs_18
...	0.082245	0.135752	-0.024017	-0.108351	-0.077623
...	0.022786	0.163320	0.012022	-0.090974	-0.056510
...	0.050791	0.207338	0.083536	-0.050691	-0.023590
...	-0.011567	0.100413	-0.050224	-0.136009	-0.177037
...	0.037439	0.219153	0.062837	-0.048885	-0.053074

Figure 2: Dataset sample from MFCCs_14 to MFCCs_18

MFCCs_19	MFCCs_20	MFCCs_21	MFCCs_22	Species
-0.009568	0.057684	0.118680	0.014038	AdenomeraAndre
-0.035303	0.020140	0.082263	0.029056	AdenomeraAndre
-0.066722	-0.025083	0.099108	0.077162	AdenomeraAndre
-0.130498	-0.054766	-0.018691	0.023954	AdenomeraAndre
-0.088550	-0.031346	0.108610	0.079244	AdenomeraAndre

Figure 3: Dataset sample from MFCCs_19 to MFCCs_22 including the Species

DATA ANALYSIS

This section contains the relationship graph between randomly selected MCFFs audio recording values.

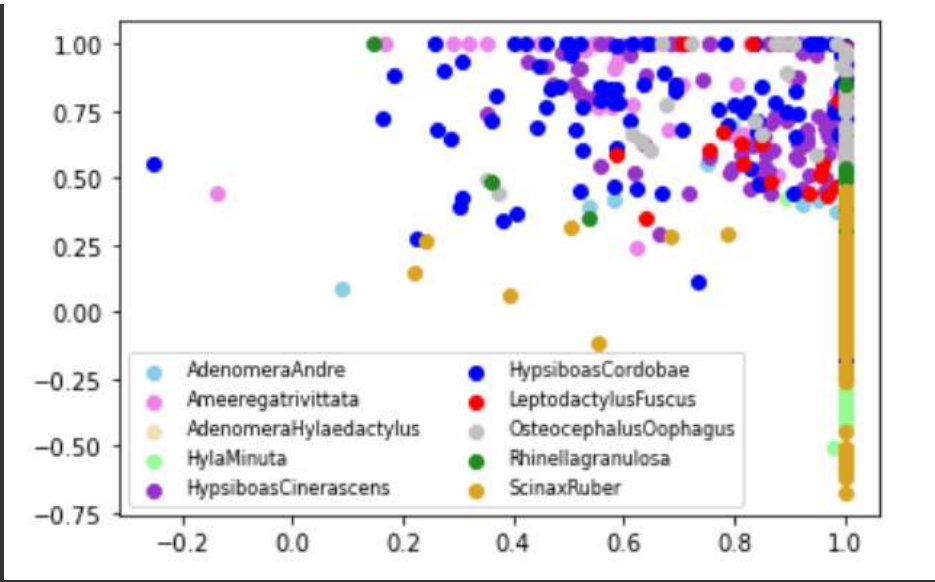

Figure 4: Scatter plot of MFCCs_1 and MFCCs_2

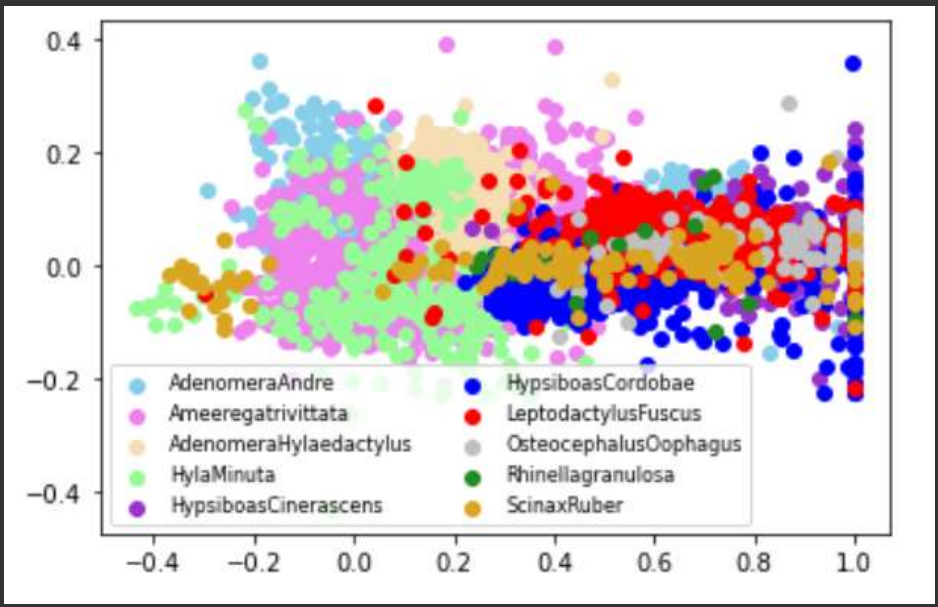

Figure 5: Scatter plot of MFCCs_3 and MFCCs_21

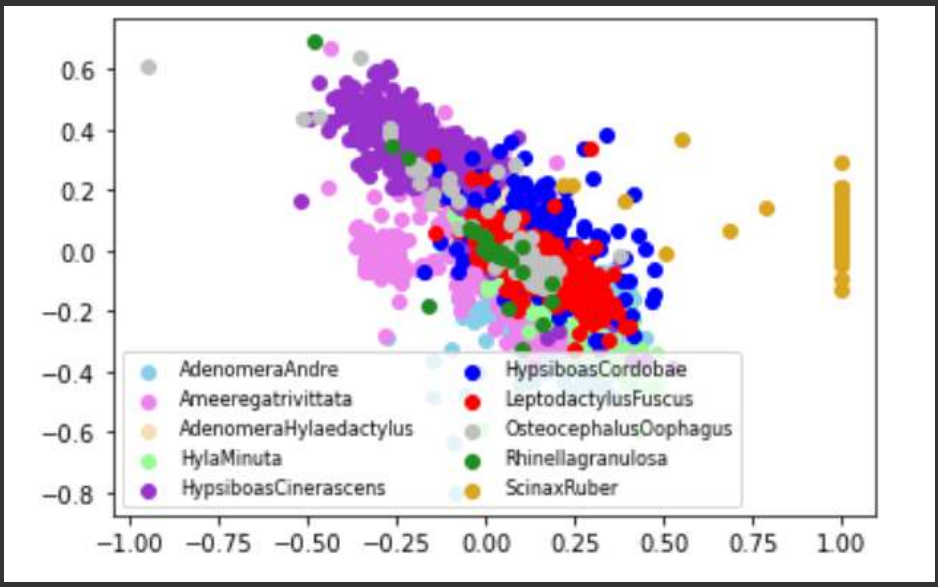

Figure 6: Scatter plot of MFCCs_10 and MFCCs_12

After running the Decision Tree algorithm, a decision tree model was built as shown below:

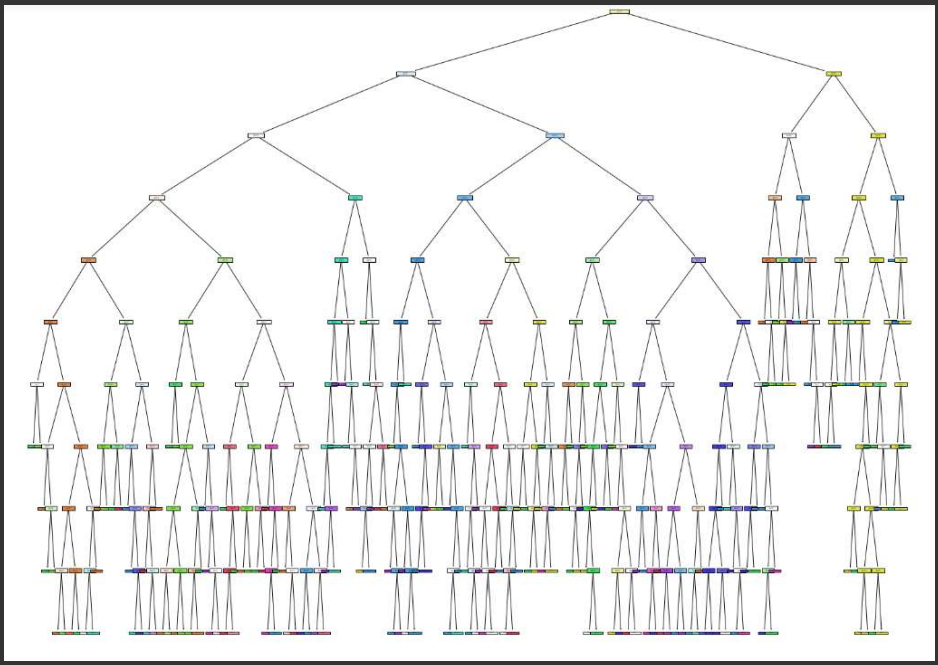

Figure 7: Decision Tree Diagram of the model

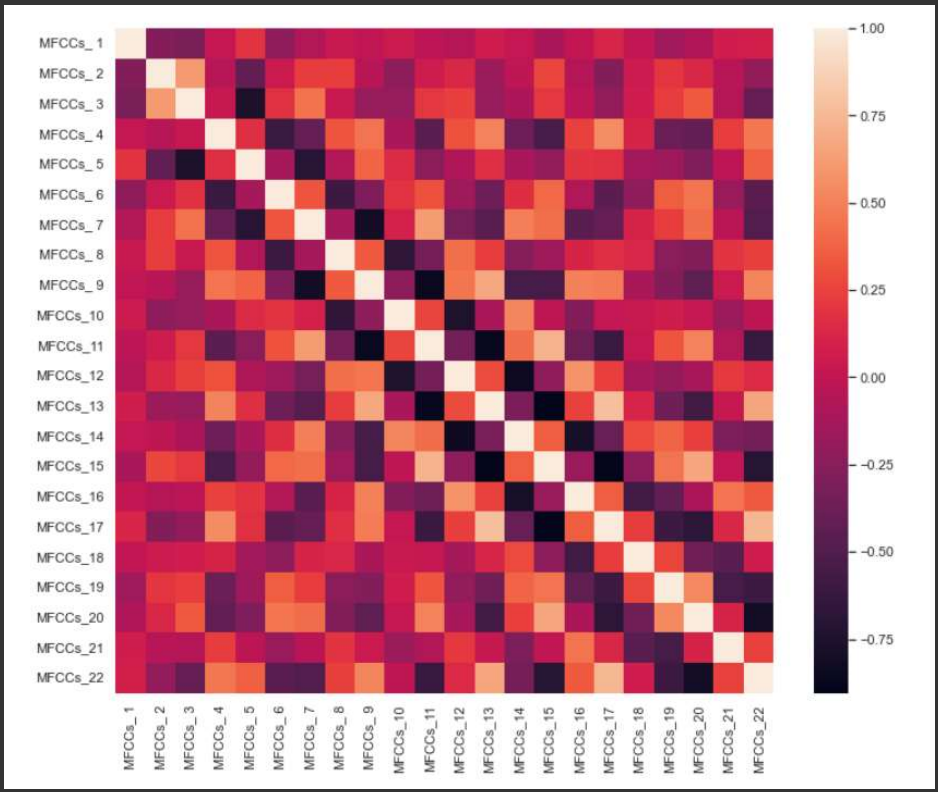

Figure 8: Heatmap of MCFFs

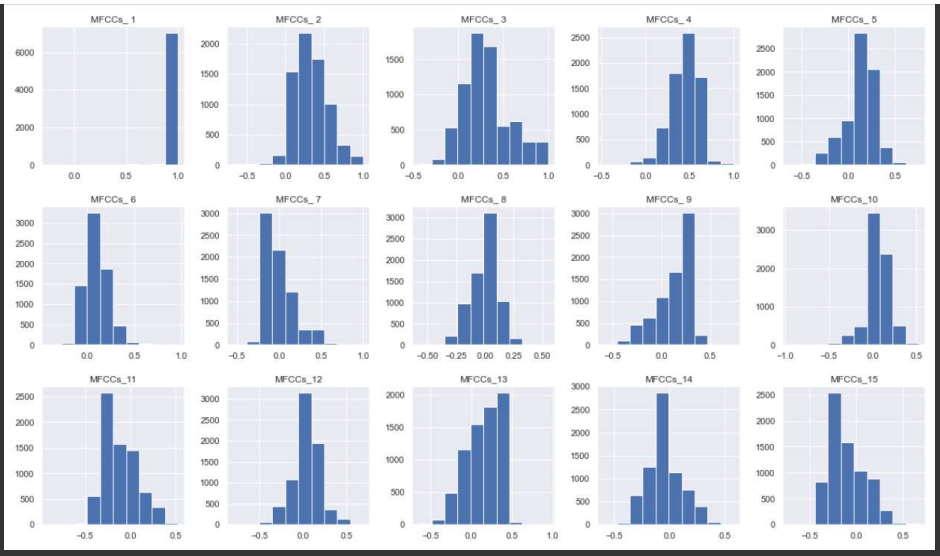

Figure 9: Individual analysis of each MFCCs (Part 1)

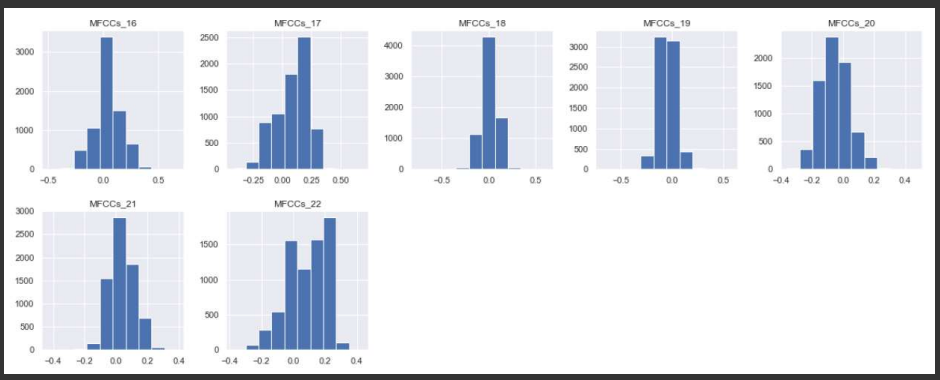

Figure 10: Individual analysis of each MFCCs (Part 2)

PART 1 ANALYSIS

DATA MODELLING - PART 1

Two predictive models were built using Decision Tree and K-Nearest Neighbor algorithm. The ratio of split is 80% training set and 20% testing set. The parameters of the model are given below.

ALGORITHM	PARAMETERS
DECISION TREE	CRITERIA : GINI MAX DEPTH : 10 MIN SAMPLES IN LEAF : 7
K-NEAREST NEIGHBOR	K : 1

Table 1: Parameters of the predictive models

The results of the classification of each predictive models are as below :

DECISION TREE MODEL

```
Analysis for Decision Tree:
-----------------------------------
Accuracy: 0.9305072967338429

                        precision    recall  f1-score   support

          AdenomeraAndre       0.89      0.93      0.91       139
AdenomeraHylaedactylus       0.99      0.98      0.99       696
     Ameeregatrivittata       0.84      0.96      0.90        95
            HylaMinuta       0.83      0.66      0.74        68
   HypsiboasCinerascens       0.95      0.94      0.94        98
      HypsiboasCordobae       0.93      0.92      0.92       226
    LeptodactylusFuscus       0.74      0.87      0.80        53
  OsteocephalusOophagus       0.61      0.55      0.58        20
      Rhinellagranulosa       0.92      0.85      0.88        13
            ScinaxRuber       0.76      0.71      0.73        31

               accuracy                           0.93      1439
              macro avg       0.85      0.84      0.84      1439
           weighted avg       0.93      0.93      0.93      1439


Confusion Matrix:
[[129   0   7   1   0   0   0   2   0   0]
 [  0 685   1   6   0   3   0   0   0   1]
 [  3   0  91   1   0   0   0   0   0   0]
 [  5   3   7  45   0   1   4   0   1   2]
 [  1   0   0   0  92   2   0   3   0   0]
 [  1   2   2   0   2 207   9   2   0   1]
 [  1   1   0   0   1   4  46   0   0   0]
 [  0   0   0   0   2   4   2  11   0   1]
 [  0   0   0   0   0   0   0   0  11   2]
 [  5   1   0   1   0   1   1   0   0  22]]
```

Figure 11: Results of classification using Decision Tree Model

K-NEAREST NEIGHBOUR MODEL

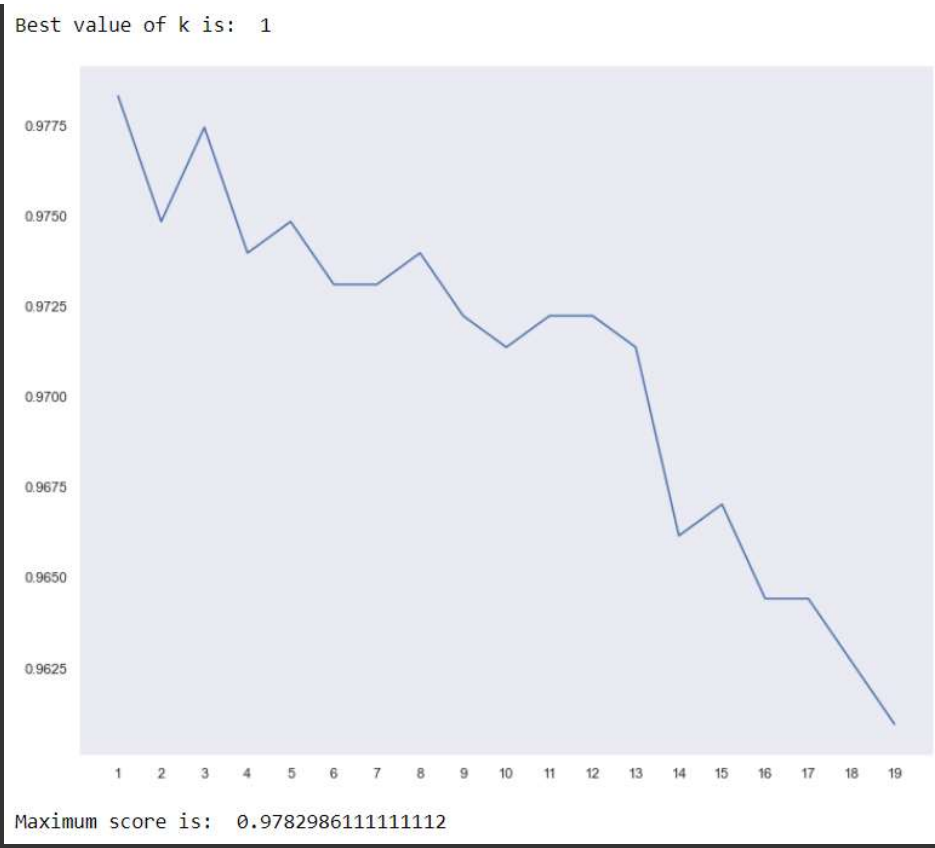

Figure 12: Graph that shows the best value of k and the maximum score the model can achieve

```
Accuracy score is  0.9770674079221682


Confusion Matrix
[[133   0   2   2   0   1   0   1   0   0]
 [  0 696   0   0   0   0   0   0   0   0]
 [  0   0  94   1   0   0   0   0   0   0]
 [  1   4   0  63   0   0   0   0   0   0]
 [  0   0   0   0  97   0   0   1   0   0]
 [  0   2   0   1   4 213   5   1   0   0]
 [  0   0   0   0   1   1  51   0   0   0]
 [  1   0   0   0   3   1   0  15   0   0]
 [  0   0   0   0   0   0   0   0  13   0]
 [  0   0   0   0   0   0   0   0   0  31]]


Accuracy, Recall, Precision and F1-score
                        precision    recall  f1-score   support

        AdenomeraAndre       0.99      0.96      0.97       139
AdenomeraHylaedactylus       0.99      1.00      1.00       696
    Ameeregatrivittata       0.98      0.99      0.98        95
           HylaMinuta        0.94      0.93      0.93        68
  HypsiboasCinerascens       0.92      0.99      0.96        98
     HypsiboasCordobae       0.99      0.94      0.96       226
   LeptodactylusFuscus       0.91      0.96      0.94        53
 OsteocephalusOophagus       0.83      0.75      0.79        20
     Rhinellagranulosa       1.00      1.00      1.00        13
           ScinaxRuber       1.00      1.00      1.00        31

              accuracy                           0.98      1439
             macro avg       0.96      0.95      0.95      1439
          weighted avg       0.98      0.98      0.98      1439
```

Figure 13: Result of classification with K-Nearest Model using default value of k = 5

```
Accuracy score is  0.9784572619874913


Confusion Matrix
[[134   0   0   3   0   1   0   1   0   0]
 [  0 696   0   0   0   0   0   0   0   0]
 [  0   0  94   1   0   0   0   0   0   0]
 [  0   3   0  65   0   0   0   0   0   0]
 [  0   0   0   0  97   0   0   1   0   0]
 [  0   3   0   1   7 211   3   1   0   0]
 [  0   0   0   0   1   1  51   0   0   0]
 [  0   0   0   0   3   1   0  16   0   0]
 [  0   0   0   0   0   0   0   0  13   0]
 [  0   0   0   0   0   0   0   0   0  31]]


Accuracy, Recall, Precision and F1-score
                       precision    recall  f1-score   support

        AdenomeraAndre       1.00      0.96      0.98       139
AdenomeraHylaedactylus       0.99      1.00      1.00       696
    Ameeregatrivittata       1.00      0.99      0.99        95
           HylaMinuta        0.93      0.96      0.94        68
  HypsiboasCinerascens       0.90      0.99      0.94        98
     HypsiboasCordobae       0.99      0.93      0.96       226
   LeptodactylusFuscus       0.94      0.96      0.95        53
 OsteocephalusOophagus       0.84      0.80      0.82        20
     Rhinellagranulosa       1.00      1.00      1.00        13
           ScinaxRuber       1.00      1.00      1.00        31

              accuracy                           0.98      1439
             macro avg       0.96      0.96      0.96      1439
          weighted avg       0.98      0.98      0.98      1439
```

Figure 14: Result of classification with K-Nearest Neighbor Model using best value of k = 1

Discussion - Part 1

Based on the results of classification, we can observe that the precision of K-Nearest Neighbor model is higher compared to the precision of Decision Tree. This can be concluded as the precision of the K-Nearest Neighbor model is more precise with the higher average value. Moreover, from the aspect of recall, K-Nearest Neighbor also contains a higher average value of recall compared to Decision Tree model. Therefore, we can conclude that K-Nearest Neighbor algorithm is the best predictive model to be used.

PART 2 ANALYSIS

DATA MODELLING - PART 2

Two predictive models were built for the second part which are Multinomial Logistic Regression and Neural Network. The ratio of split is 80% training set and 20% testing set.

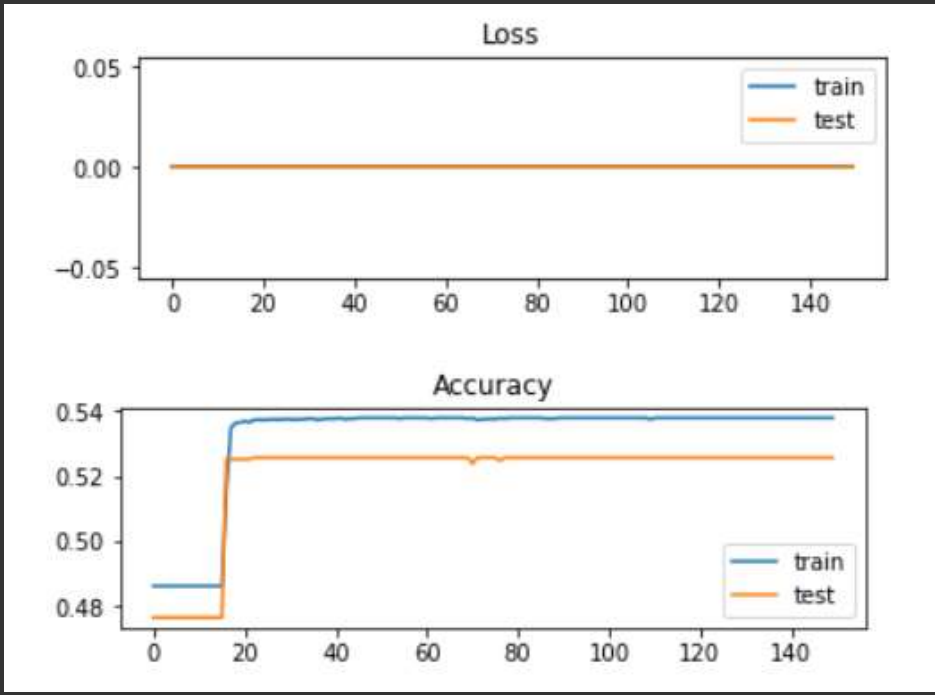

Figure 15: The graph of loss and accuracy during model training

MULTINOMIAL LOGISTIC REGRESSION

```
0.9027102154273802
[[124   0  12   0   0   1   0   2   0   0]
 [  0 694   1   0   0   1   0   0   0   0]
 [  1   0  93   1   0   0   0   0   0   0]
 [ 11  12  10  28   0   7   0   0   0   0]
 [  2   0   0   0  94   1   0   1   0   0]
 [  0   2   0   0   3 218   1   2   0   0]
 [  3   0   0   1   0   9  40   0   0   0]
 [  0   0   0   0   6  13   1   0   0   0]
 [  0   0   0   0   0  13   0   0   0   0]
 [  0   0   2   2   2  16   0   1   0   8]]
```

Figure 16: Confusion matrix generated by Multinomial Logistic Regression Model

	precision	recall	f1-score	support
AdenomeraAndre	0.88	0.89	0.89	139
AdenomeraHylaedactylus	0.98	1.00	0.99	696
Ameeregatrivittata	0.79	0.98	0.87	95
HylaMinuta	0.88	0.41	0.56	68
HypsiboasCinerascens	0.90	0.96	0.93	98
HypsiboasCordobae	0.78	0.96	0.86	226
LeptodactylusFuscus	0.95	0.75	0.84	53
OsteocephalusOophagus	0.00	0.00	0.00	20
Rhinellagranulosa	0.00	0.00	0.00	13
ScinaxRuber	1.00	0.26	0.41	31
accuracy			0.90	1439
macro avg	0.72	0.62	0.63	1439
weighted avg	0.89	0.90	0.89	1439

Figure 17: Result of analysis using Multinomial Logistic Regression Model

NEURAL NETWORK

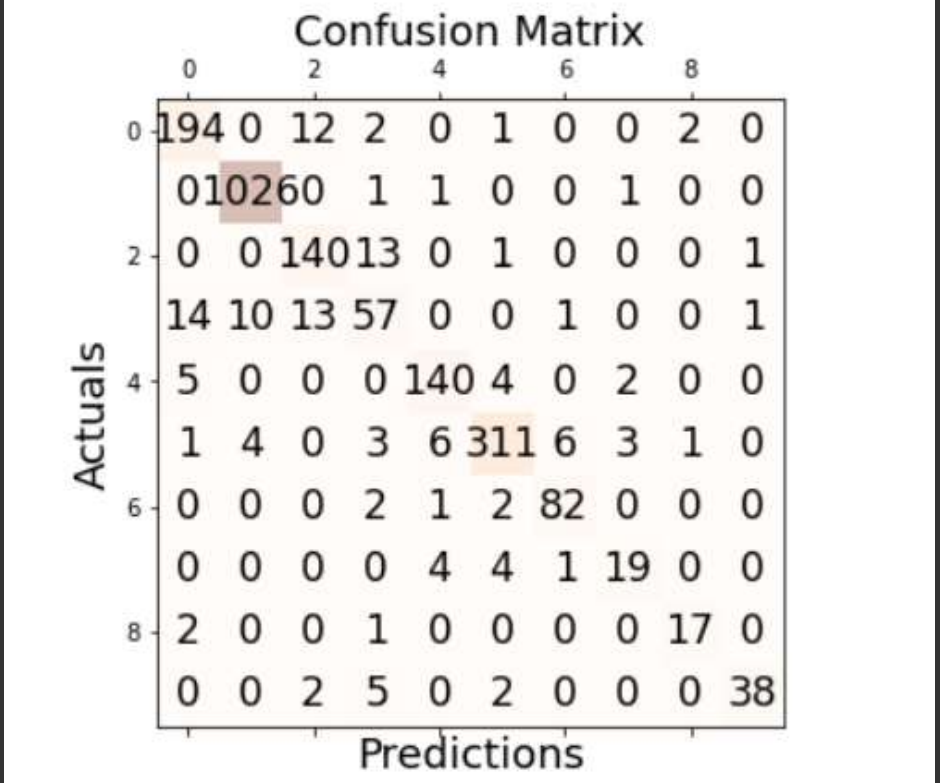

Figure 18: Confusion matrix generated by Neural Network Model

Analysis for Neural Network:

Accuracy: 0.937
Precision: 0.875
Recall: 0.855
F1 Score: 0.864

	precision	recall	f1-score	support
0	0.90	0.92	0.91	211
1	0.99	1.00	0.99	1029
2	0.84	0.90	0.87	155
3	0.68	0.59	0.63	96
4	0.92	0.93	0.92	151
5	0.96	0.93	0.94	335
6	0.91	0.94	0.93	87
7	0.76	0.68	0.72	28
8	0.85	0.85	0.85	20
9	0.95	0.81	0.87	47
accuracy			0.94	2159
macro avg	0.88	0.85	0.86	2159
weighted avg	0.94	0.94	0.94	2159

Figure 19: Result of analysis using Neural Network Model

DISCUSSION - PART 2

Based on the result analysis, we can extract some calculation data. For the Multinomial Logistic Regression Model, the calculation data are as follows :

Accuracy score - 0.90, Precision - 0.72, Recall - 0.62, f1-Score - 0.635

For the Neural Network Model, the calculation data are as follows :

Accuracy score - 0.94, Precision - 0.88, Recall - 0.86, f1-Score - 0.86

Based on these data, we can conclude that Neural Network Model is the better predictive model compared to Multinomial Logistic Regression Model because of the overall higher score of average for accuracy, precision, recall and f1-Score.