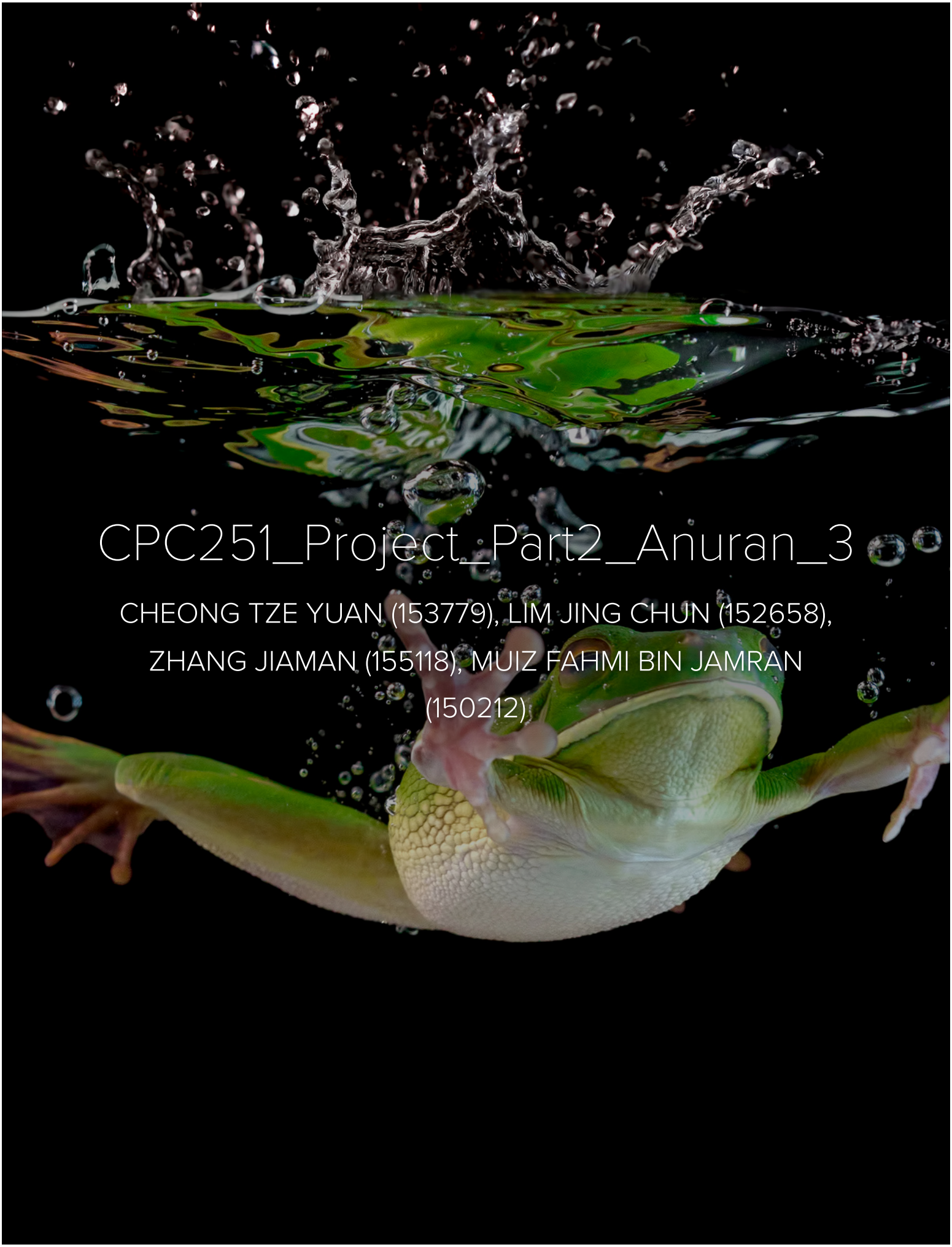


CPC251_Project_Part2_Anuran_3

CHEONG TZE YUAN (153779), LIM JING CHUN (152658), ZHANG JIAMAN (155118), MUIZ FAHMI BIN JAMRAN (150212)

Anuran is any of the vertebrates of the order Anura, characterized by the absence of a tail and very long hind legs specialized for hopping: class Amphibia (amphibians), which is commonly known as the frogs and toads. Recently, recognition of anuran species through their calls has received a lot of attention because of its potential applicability in ecological studies.

However, most of the recorded anuran species are considered to be monotypic based on various research presented. Hence, the classification of numerous anuran species would be a challenge for researchers.

Aim: To develop an interpretable and trustworthy predictive model that can classify various anuran species accurately and effectively.

In this project, we are using the given dataset about the Anuran species, to create two predictive models, to test and predict based on the dataset.

Data Description

The Anuran Species dataset is the dataset assigned to our group for project part 2. Based on the two tables below, we can notice that the dataset contains 22 columns of features and a single column of the target. There is not a very unique name for all the features, they only differ in terms of the numbering, which represents different animals without clarifying what the actual animal is. The target is the 'Species', in which we are expected to get the outcome of different types of Anuran species (10 different species).

	MFCCs_1	MFCCs_2	MFCCs_3	MFCCs_4	MFCCs_5	MFCCs_6	MFCCs_7	MFCCs_8	MFCCs_9	MFCCs_10	MFCCs_11
1727	1.0	0.177345	0.021361	0.591013	0.241141	-0.022545	-0.139830	-0.040433	0.368914	0.129189	-0.205062
5360	1.0	0.847515	0.882058	0.394028	-0.072178	0.353863	0.138748	0.144420	0.282237	-0.207722	-0.000795
3163	1.0	0.017307	-0.001150	0.451407	0.263967	0.103866	-0.103720	-0.074888	0.203453	0.139746	-0.247675
4110	1.0	0.228557	0.133194	0.474253	0.138945	0.024468	-0.112011	0.075913	0.323418	-0.028529	-0.407417
1052	1.0	0.120809	0.204131	0.415838	0.113902	0.059072	0.033450	-0.160150	-0.175745	0.014239	0.085644
2356	1.0	0.163049	0.148034	0.547341	0.222415	-0.051954	-0.153605	0.106623	0.268728	0.033979	-0.173320
5305	1.0	0.845180	0.706394	0.323032	-0.041422	0.228876	0.058695	0.071713	0.261275	0.001419	0.070003
3244	1.0	0.137138	0.023468	0.449341	0.181889	0.044411	-0.097628	0.076887	0.288678	0.111262	-0.270028
436	1.0	0.434591	0.592396	0.543714	-0.133993	-0.105860	0.343019	0.123359	-0.224352	0.052577	0.227083
507	1.0	0.478178	0.687712	0.600840	-0.222338	-0.149847	0.441022	0.071143	-0.346294	0.142127	0.336633

Table 1: Dataset sample 1

MFCCs_12	MFCCs_13	MFCCs_14	MFCCs_15	MFCCs_16	MFCCs_17	MFCCs_18	MFCCs_19	MFCCs_20	MFCCs_21	MFCCs_22	Species
0.048536	0.396620	-0.081590	-0.352294	0.014482	0.236941	0.078685	-0.148126	-0.178317	0.019327	0.220908	AdenomeraHylaedactylus
0.446579	-0.056946	-0.239524	0.213079	0.116030	-0.191531	-0.033739	0.092026	0.148543	0.099929	-0.111212	HypsiboasCinerascens
-0.000491	0.298866	-0.079190	-0.332492	0.028229	0.268636	0.102729	-0.092891	-0.196381	-0.037426	0.232715	AdenomeraHylaedactylus
0.063844	0.357171	-0.110503	-0.261927	0.086037	0.192672	0.013936	-0.133675	-0.212496	0.051246	0.224806	AdenomeraHylaedactylus
0.141874	0.042747	-0.021757	-0.035412	-0.139353	-0.138391	0.070786	0.040059	-0.093455	-0.095817	-0.002407	Ameeregatrivittata
0.142113	0.221791	-0.145545	-0.164296	0.065103	0.063280	-0.100817	-0.094602	-0.013419	0.141306	0.118285	AdenomeraHylaedactylus
0.281590	-0.217807	-0.187214	0.331285	0.108411	-0.097495	0.103971	0.039770	0.005668	0.022353	-0.088819	HypsiboasCinerascens
-0.057629	0.269756	-0.013931	-0.249028	-0.010789	0.233975	0.105894	-0.126081	-0.195791	-0.042994	0.226574	AdenomeraHylaedactylus
-0.073593	-0.116745	0.145202	0.075406	-0.108440	0.040087	0.096547	-0.047415	0.021464	0.122501	0.006756	AdenomeraAndre
-0.111968	-0.164393	0.194708	0.108305	-0.139689	0.057584	0.136974	-0.046876	-0.099424	0.021916	0.021012	AdenomeraAndre

Table 2: Dataset sample 2

No	Anuran species
0	AdenomeraAndre
1	AdenomeraHylaedactylus
2	Ameeregatrivittata
3	HylaMinuta
4	HypsiboasCinerascens
5	HypsiboasCordobae
6	LeptodactylusFuscus
7	OsteocephalusOophagus
8	Rhinellagranulosa
9	ScinaxRuber

Table 3: List of targets

Data modeling

Two predictive models are built to predict the target variable of the Anuran Calls (MFCCs) dataset using K-Nearest Neighbors and Neural Network algorithms. These models are analysed using the hold-out approach, which separates the dataset into the

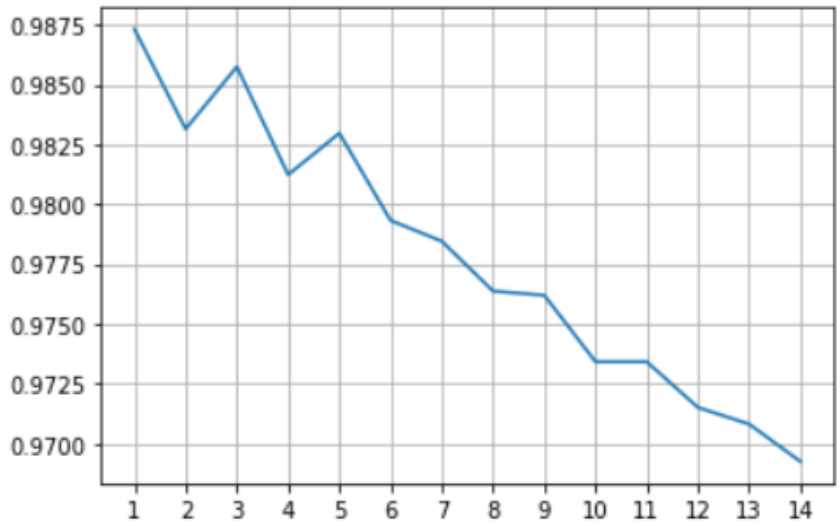
training set and test set. The dataset splitting ratio is 80% training set and 20% test set. The parameters of the predictive models are provided in Table 4.

Algorithm	Value/Statistics
K-Nearest Neighbors	K: 1 Distance Metric: Minkowski
Neural Network	Batch Size: 32 Dropout: 0.25 L2 Regularization: 0.01

Table 4: Parameters of the predictive models

Fine-tuning

In the k-nearest neighbors model, the best value of k is determined by selecting the k with the maximum score achieved. Initially, the model is fit with k ranging from 1 to 15. Subsequently, the score is calculated using a cross-validation approach. After the iterations are completed, the k with the highest score is selected as the best value of k. Next, it is passed into the parameters of KNeighborsClassifier and then the k-nearest neighbors model is trained with the respective parameters. Figure 1 shows the values of k ranging from 1 to 14 corresponding to their scores.




```
The best value of k is 1 with score 0.9873175816539264
```

Figure 1: Graph of scores against the values of k

In the neural network, the results may vary due to the stochastic nature of the neural network algorithm each time the program is executed. Therefore, the program is executed 10 times and the best result with the highest accuracy will be selected in order to compare with the k-nearest neighbors model. In addition, early stopping regularization is implemented to stop the training when the training loss is not improving. The parameter of the early stopping regularization, 'patience' is set to 5, representing the number of epochs without improvement. The advantage of using the early stopping method is to reduce the problem of overfitting and improve the generalization of the neural network model. Based on Figure 2, the model stop training at epoch 49 with a 0.2584 loss value and 0.9581 accuracy score, which represents the best result. Also, Figure 3 below shows the graph of training loss against epochs.

```
180/180 [==============================] - 0s 2ms/step - loss: 0.2734 - accuracy: 0.9574
Epoch 41/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2681 - accuracy: 0.9557
Epoch 42/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2688 - accuracy: 0.9545
Epoch 43/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2687 - accuracy: 0.9571
Epoch 44/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2573 - accuracy: 0.9590
Epoch 45/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2640 - accuracy: 0.9538
Epoch 46/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2592 - accuracy: 0.9555
Epoch 47/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2680 - accuracy: 0.9512
Epoch 48/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2597 - accuracy: 0.9548
Epoch 49/100
180/180 [==============================] - 0s 2ms/step - loss: 0.2584 - accuracy: 0.9581
```

Figure 2: Process of model fitting

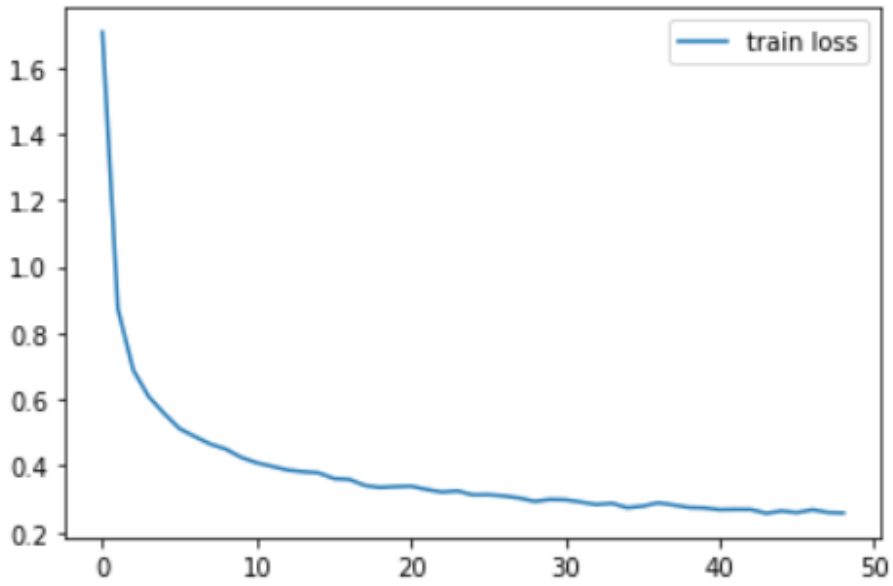


Figure 3: Graph of training loss against epochs

Selection of performance metrics

Furthermore, the performance metrics are selected including accuracy, recall, precision, F1-score measures and the confusion matrix to evaluate the decision tree as well as support vector machine model so that the results can be clearly compared and the best suited predictive model can be identified accurately. The results of classification of anuran species using these 2 models are provided below.

Results of classification

****

Accuracy: 0.9861014593467686

Confusion matrix:

```
[[137   0   0   0   0   1   1   0   0   0]
 [  0 696   0   0   0   0   0   0   0   0]
 [  0   0  95   0   0   0   0   0   0   0]
 [  1   0   0  67   0   0   0   0   0   0]
 [  0   0   0   0  96   1   0   1   0   0]
 [  0   2   0   1   2 217   2   2   0   0]
 [  0   0   0   0   1   0  52   0   0   0]
 [  0   0   0   0   3   1   1  15   0   0]
 [  0   0   0   0   0   0   0   0  13   0]
 [  0   0   0   0   0   0   0   0   0  31]]
```

Classification report:

	precision	recall	f1-score	support
AdenomeraAndre	0.99	0.99	0.99	139
AdenomeraHylaedactylus	1.00	1.00	1.00	696
Ameeregatrivittata	1.00	1.00	1.00	95
HylaMinuta	0.99	0.99	0.99	68
HypsiboasCinerascens	0.94	0.98	0.96	98
HypsiboasCordobae	0.99	0.96	0.97	226
LeptodactylusFuscus	0.93	0.98	0.95	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			0.99	1439
macro avg	0.97	0.96	0.96	1439
weighted avg	0.99	0.99	0.99	1439

Figure 4: Results of classification of anuran species using k-nearest neighbors

****

```
Accuracy: 0.9610840861709521

Confusion matrix:
[[126   0   5   5   0   1   0   1   0   1]
 [  0 692   0   3   1   0   0   0   0   0]
 [  1   0  90   2   0   1   0   0   0   1]
 [  6   5   1  56   0   0   0   0   0   0]
 [  0   0   0   0  96   1   1   0   0   0]
 [  0   2   0   1   0 216   5   2   0   0]
 [  0   0   0   1   0   1  50   1   0   0]
 [  0   0   0   0   4   2   1  13   0   0]
 [  0   0   0   0   0   0   0   0  13   0]
 [  0   0   0   0   0   0   0   0   0  31]]

Classification report:
              precision    recall  f1-score   support

           0       0.95      0.91      0.93       139
           1       0.99      0.99      0.99       696
           2       0.94      0.95      0.94        95
           3       0.82      0.82      0.82        68
           4       0.95      0.98      0.96        98
           5       0.97      0.96      0.96       226
           6       0.88      0.94      0.91        53
           7       0.76      0.65      0.70        20
           8       1.00      1.00      1.00        13
           9       0.94      1.00      0.97        31

    accuracy                           0.96      1439
   macro avg       0.92      0.92      0.92      1439
weighted avg       0.96      0.96      0.96      1439
```

Figure 5: Results of classification of anuran species using neural network

****

Discussion

According to the results above, the correct predictions that fall on the diagonal line of the confusion matrix in the k-nearest algorithm are more than in the neural network algorithm. For instance, there are 11 more correct predictions of

AdenomeraAndre species in k-nearest neighbors (137 are predicted correctly) than in neural networks (126 are predicted correctly). The performance metrics such as accuracy, precision, recall as well as f1-score are calculated based on the confusion matrix.

First, the accuracy score in the k-nearest neighbors algorithm (98.61%) is higher than in the neural network algorithm (96.11%), which indicates that the number of correct predictions made in the k-nearest neighbors algorithm is greater than in the neural network algorithm by 2.5%. One of the possible reasons for this result is the training dataset is not large enough for model fitting to achieve a sufficient accuracy score in the neural network model, thus the accuracy score is lower than in k-nearest neighbors.

Moreover, the precisions for each class in the k-nearest neighbors algorithm are higher than in the neural network algorithm, which means that the k-nearest neighbors algorithm returns more relevant results than the neural network algorithm. For instance, when the k-nearest neighbors model predicts an anuran which is actually classified as the Ameeregatrivittata species, it is correct around 100% of the time. Conversely, it is correct around 94% of the time, which has decreased by 6% when the neural network model is used to predict the anuran species.

Furthermore, the recalls for each class in the k-nearest neighbors algorithm are greater than in the neural network algorithm, which proves that the k-nearest neighbors algorithm returns most of the relevant results than the neural network algorithm. For example, for all anurans that actually belong to the AdenomeraAndre species, the k-nearest neighbors model correctly identifies around 99% as belonging to the AdenomeraAndre species. However, the neural network model only correctly identifies around 91% as belonging to the AdenomeraAndre species, which has decreased by 8%.

Last but not least, the f1-scores for each class in the k-nearest neighbors algorithm are higher than in the neural network algorithm. This indicates that the precisions and recalls evaluated in the k-nearest neighbors algorithm are greater than in the neural network algorithm.

In conclusion, **k-nearest neighbors is the best suited predictive model** for the multiclass classification of the anuran species as it has higher accuracy, precision, recall as well as f1-score when compared to the neural network model.

Credits:

Created with images by kuritafsheen - "Whitelipped frog in the water, swimming frog, Whitelipped frog swimming" • kuritafsheen - "Flying frog on red flower, beautiful tree frog on red flowe, animal closeup"