Name: Ahmed Aldawoud
Student ID No.: 18038024

Foundation of Information science

Assignment 1

Zoo Data set

This data set is about animals. It has 14 attributes as booleans(True or false), one attribute as a string of the animal's name, and two attributes as numeric; (from 1 to 7) which describes which class of animals the animal described belongs to, and (from 1 to 8) which describes the number of legs an animal has.

The set classifies each animal into a set of animals from 1 to 7 which according to my observation translates into a real animal classification, the classes are: **('mammal' 'bird' 'reptile' 'fish' 'amphibian' 'insects' 'invertebrates')**(they represent the numbers from 1 to 7 in the data set respectively.

In this data set there are 2 numerical values:
1- Number of legs the animal has:

- Minimum number is 0.
- Maximum number is 8
- Average number of legs: 2.841584158415841. (Does not mean that animals have 2.8 legs)

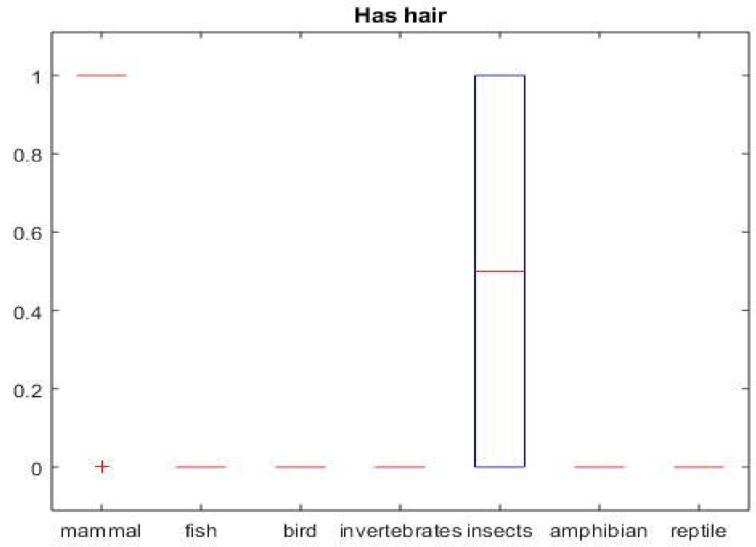
2- Animal type (the classification) which doesn't make sense to be a number or to get its average, minimum, or maximum, so I changed the value to the 7 strings mentioned above.

Name: Ahmed Aldawoud
Student ID No.: 18038024

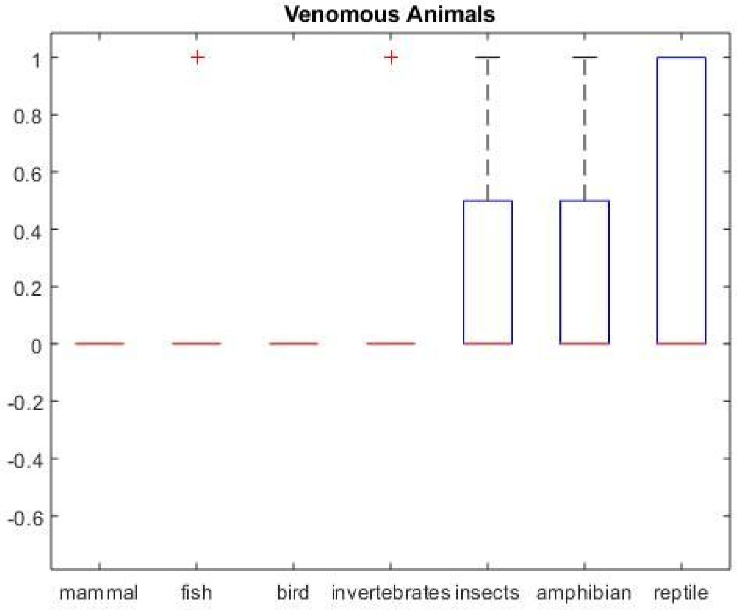
Graphs

Note: not all the graphs made were displayed below,

Box Plots:



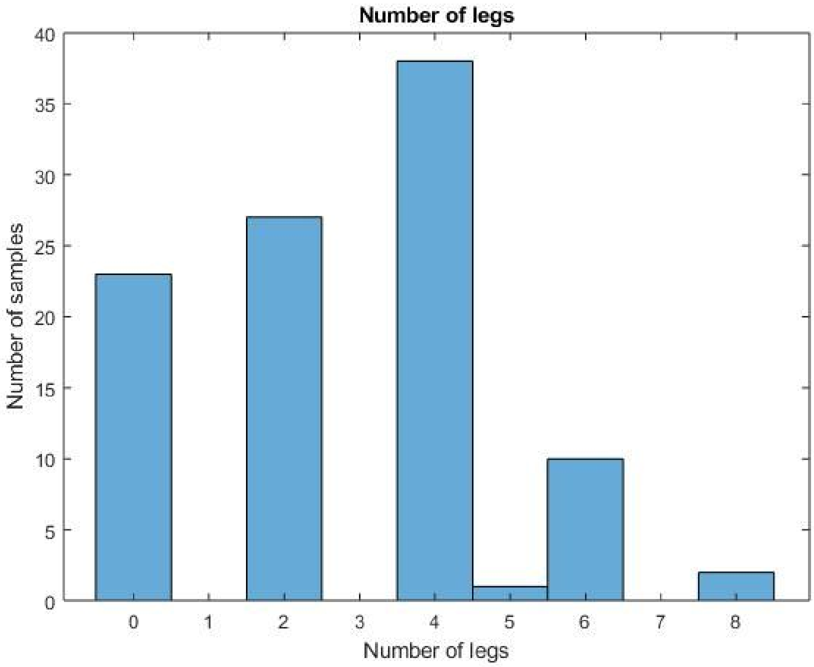
From the graph above we can notice that almost all mammals have hair, some insects have hair as well, other animal classes seem to be hairless.



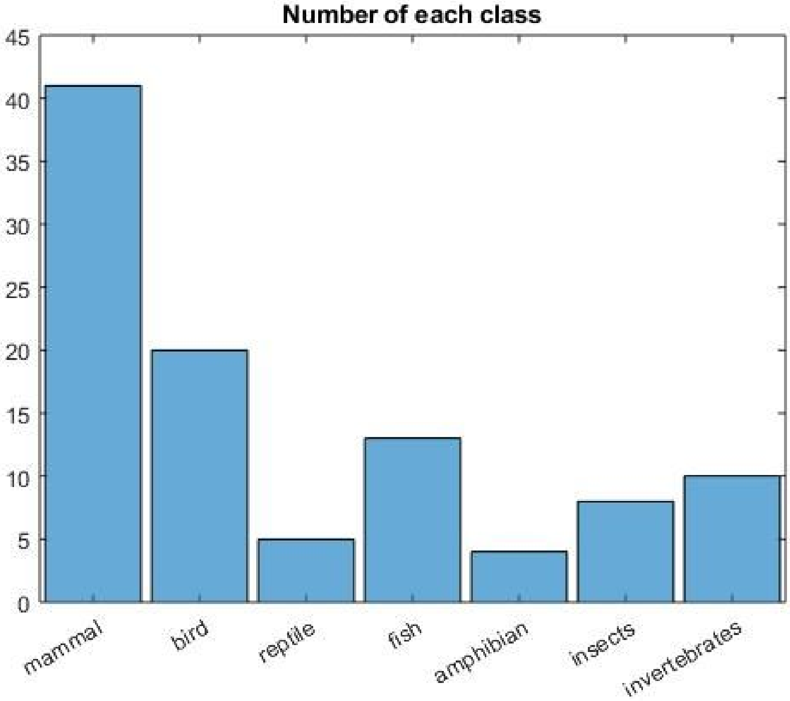
The Box plot above shows that there are numbers of venomous animals who are insects, reptiles, and amphibians. Other animals seem to be mostly not venomous with few exceptions.

Name: Ahmed Aldawoud
Student ID No.: 18038024

Histograms:



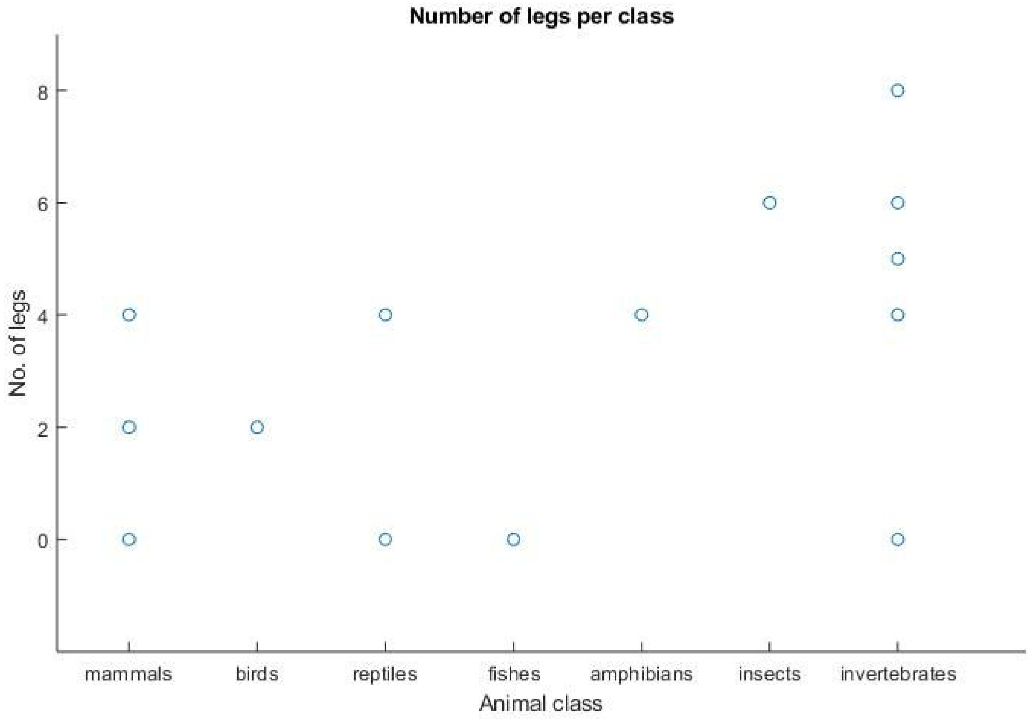
The histogram above classifies animals according to their number of legs.



The graph above explains the number of samples that belong to each class in the given data set.

Name: Ahmed Aldawoud
Student ID No.: 18038024

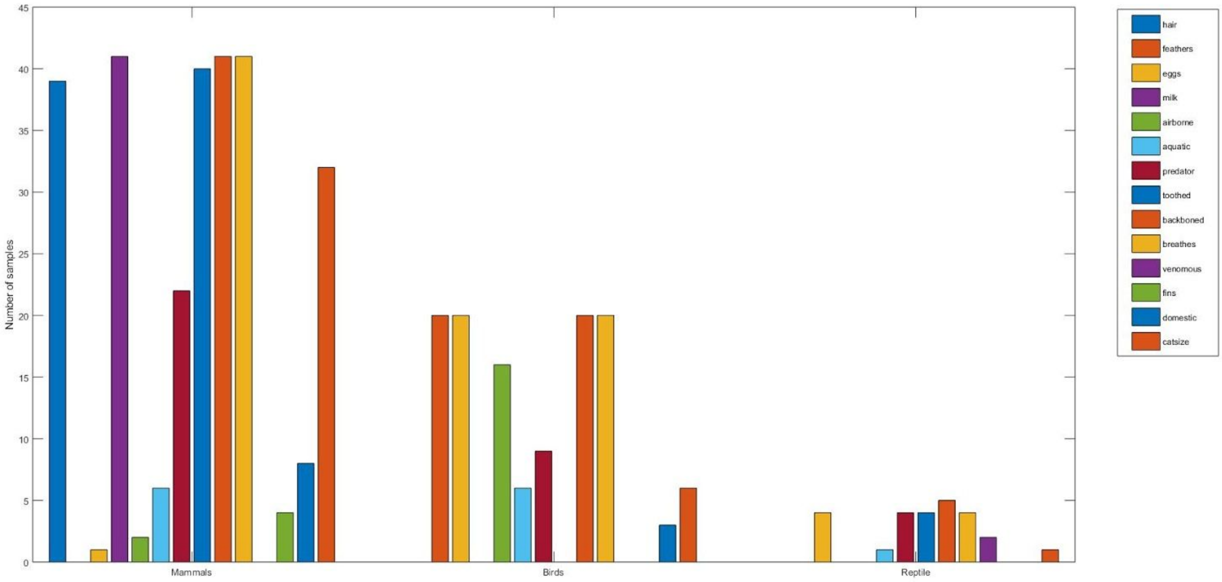
Scatter Plot:



The scatterplot above shows the correlation between the number of legs and what class the animal belongs to.

Name: Ahmed Aldawoud
Student ID No.: 18038024

Bar chart:



The bar graph above shows the number of animals from each class that have a certain feature in them. Features are represented by a color explained on the right side of the graph (I only added three classes to this graph because details will show too small if I would add all the 7 classes.

Summary about the data:

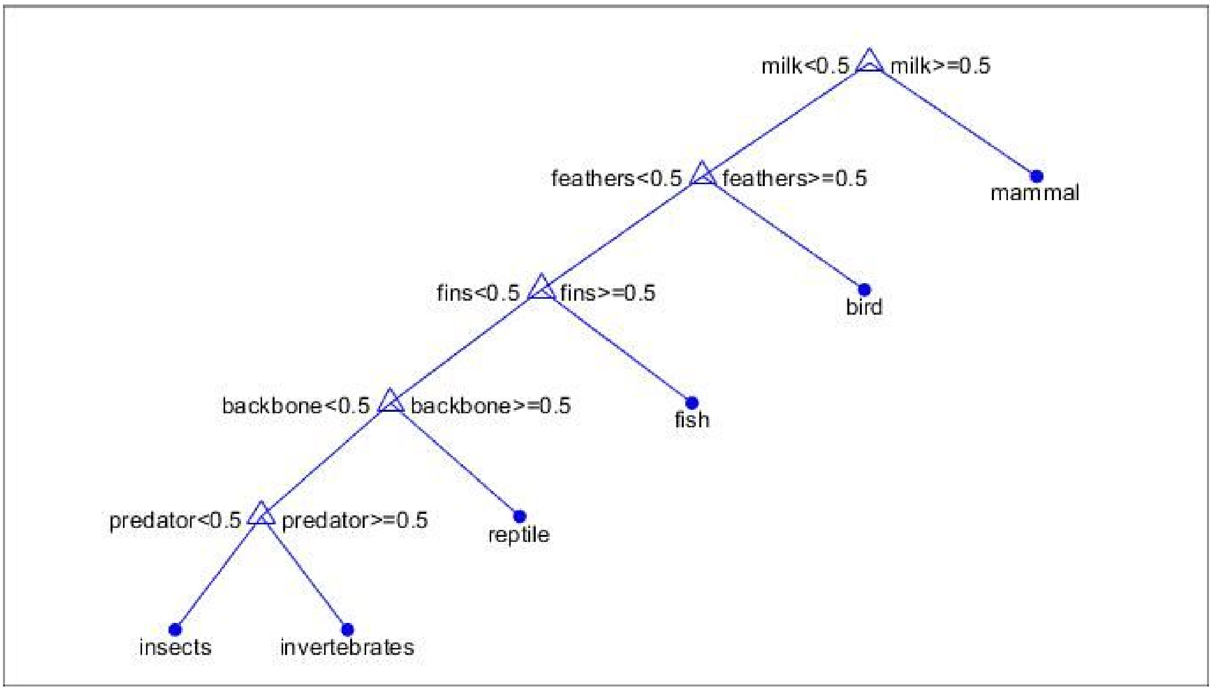
The attributes are majorly booleans with the exception of 2 numerical and one string attribute.
I decided to ignore the string attribute as it was a unique string (animal name) which wasn't going to help with the classification.
The only repeated animal was 'frog' with two occurrences which is why I did not use the animal name attribute to work with my classifiers later on.
The animal classes in the original data was represented by a numerical value, so I decided to replace it with strings using a for loop to show the name of the animal class which is a more useful information than an integer.

Name: Ahmed Aldawoud
Student ID No.: 18038024

Classification Results

Note: all numbers seen below might change when you run the code again because I am using a function to randomise the data before dividing it into training and testing sets.

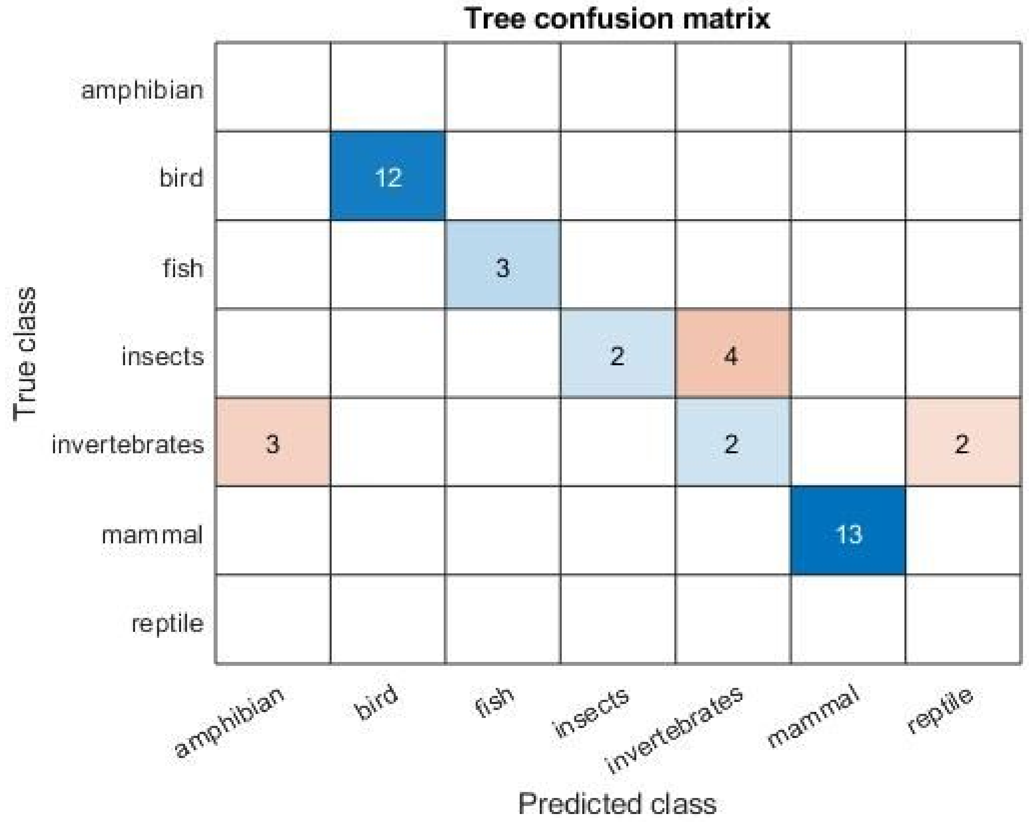
Tree Classification:



The graph above shows how the decision tree model work.
Note: the tree graph might change every time you run it because I have put a randomise function that randomises the data before dividing it into a 60 samples out of 101 samples for training and 41 for testing..

From the tree graph we can notice that the tree is unbalanced which is not good.

Name: Ahmed Aldawoud
Student ID No.: 18038024



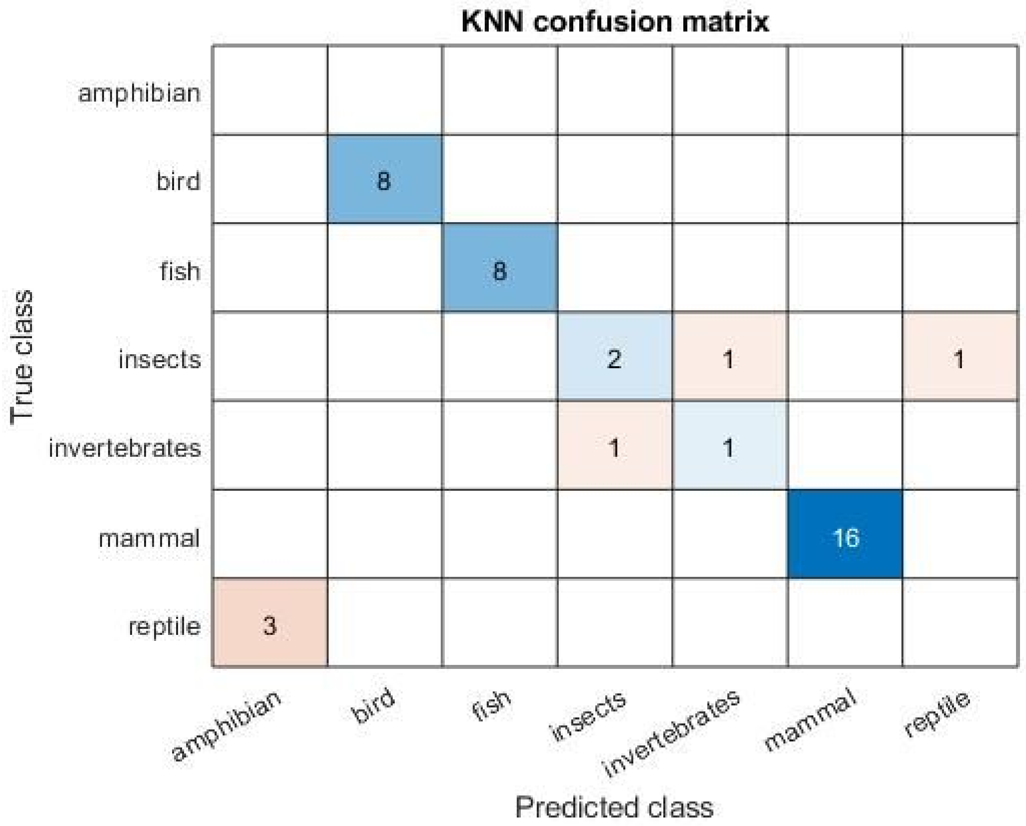
	amphibian	bird	fish	insects	invertebrates	mammal	reptile
amphibian							
bird		12					
fish			3				
insects				2	4		
invertebrates	3				2		2
mammal						13	
reptile							

Accuracy: 0.914 **or** 91.4% (rounded)
Precision: 0.7738 (rounded)
Recall: 0.6785 (rounded)
Error: 0.086 **or** 8.6% (rounded)

The accuracy rate is around 91 percent which is not bad.
The confusion matrix above shows that the decision tree made some mistakes when it came to invertebrates and classified 2 samples as reptiles and 3 as amphibians.
I can imagine that those mistakes were made because there are some characters shared between those 3 classes of animals.

K-Nearest Neighbor:

For this classifier, I used 2 neighbors because I found out that this was the best number to get the best results from this classifier.



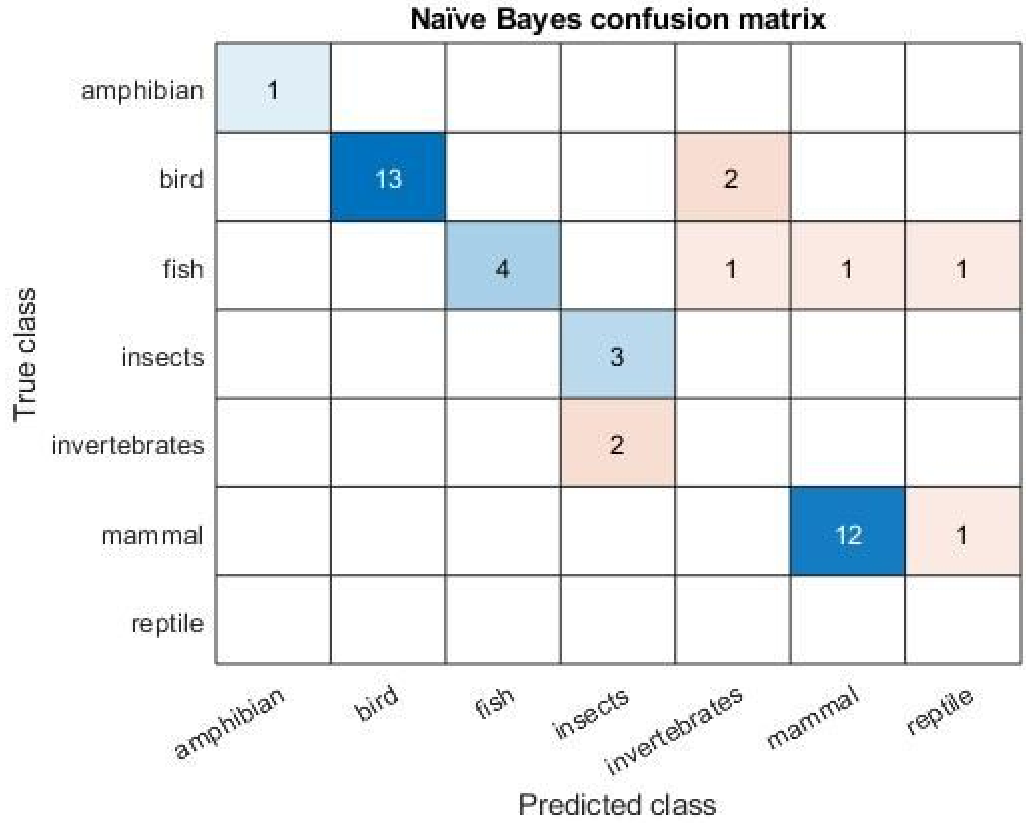
KNN confusion matrix

True class \ Predicted class	amphibian	bird	fish	insects	invertebrates	mammal	reptile
amphibian							
bird		8					
fish			8				
insects				2	1		1
invertebrates				1	1		
mammal						16	
reptile	3						

Accuracy: 0.9084 **or** 90.84% (rounded)
Precision: 0.8214 (rounded)
Recall: 0.7976 (rounded)
Error: 0.0916 **or** 9.16% (rounded)

The accuracy rate is around 90.8 percent which is close to the decision tree accuracy but a bit less.
From looking at the confusion matrix above, we can notice that the classifier has failed to classify 3 reptiles and it thought they were amphibians. This is probably because the reptiles and the amphibians fall close when it come to this classifier.

Naive Bayes:



Naïve Bayes confusion matrix

True class \ Predicted class	amphibian	bird	fish	insects	invertebrates	mammal	reptile
amphibian	1						
bird		13			2		
fish			4		1	1	1
insects				3			
invertebrates				2			
mammal						12	1
reptile							

Accuracy: 0.8038 **or** 80.38% (rounded)
Precision: 0.5992 (rounded)
Recall: 0.5417 (rounded)
Error: 0.1962 **or** 19.62% (rounded)

The accuracy rate is around 80.38 percent which is worse than the two previous classifiers. From the confusion matrix above we can see that it predicted 2 invertebrates to be insects. It is not the best classifier for this kind of data because it checks the probabilities which doesn’t work with animals’ characteristics because most of the data is boolean; which means an animal can either have that character or not have it.