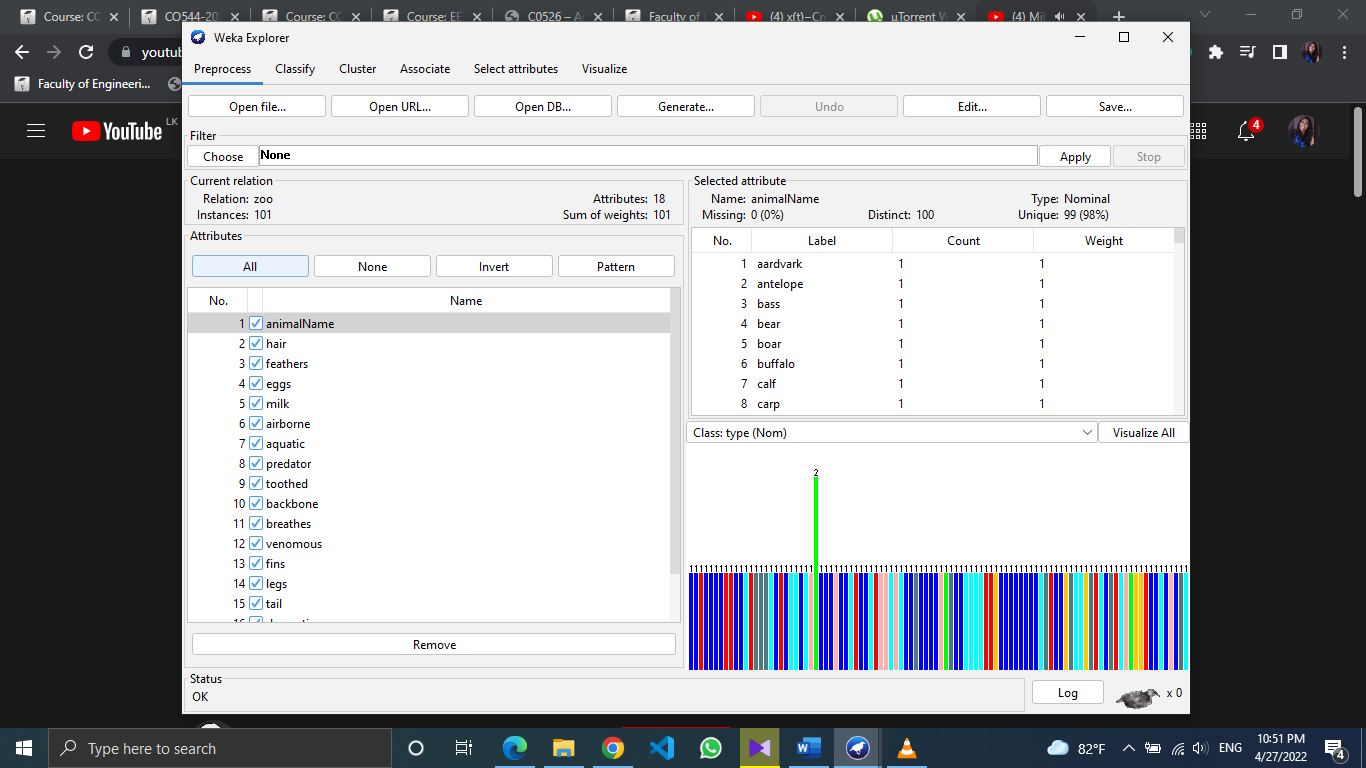
E/17/153 : Part 1 – Classification using WEKA

1.As it shows here there are 101 instances and 18 attributes in the “Zoo” dataset.



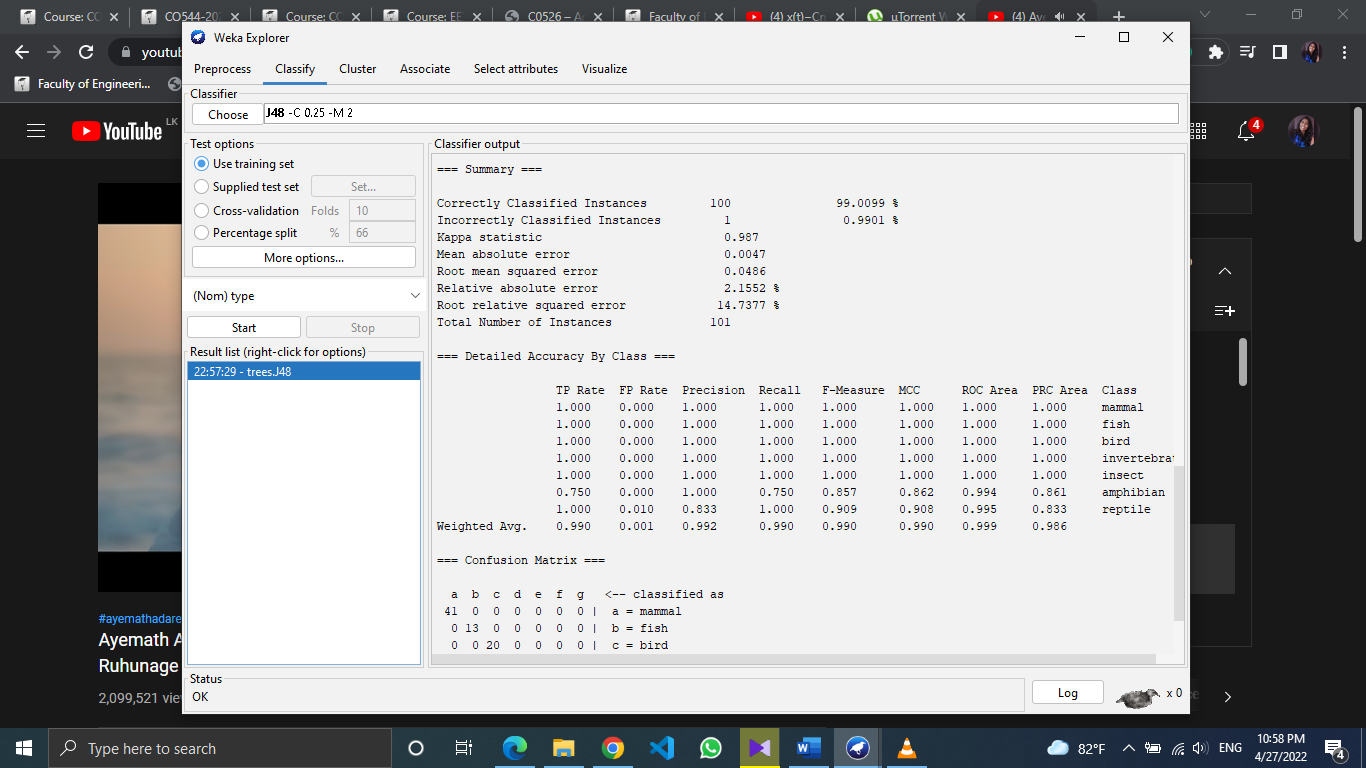
2. Output of the C4.5 algorithm:

Correctly classified instances = 100

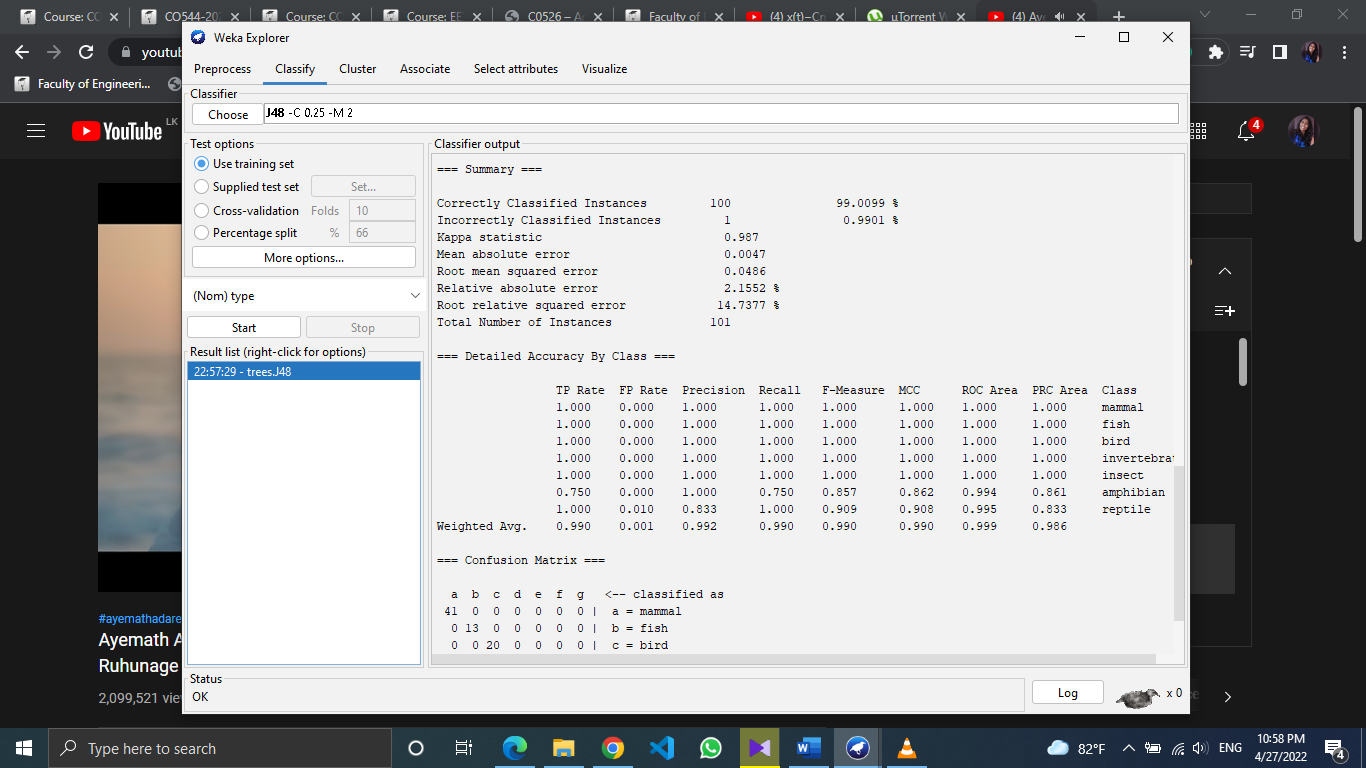
Incorrectly classified instances = 1

3. Accuracy = 99.0099%

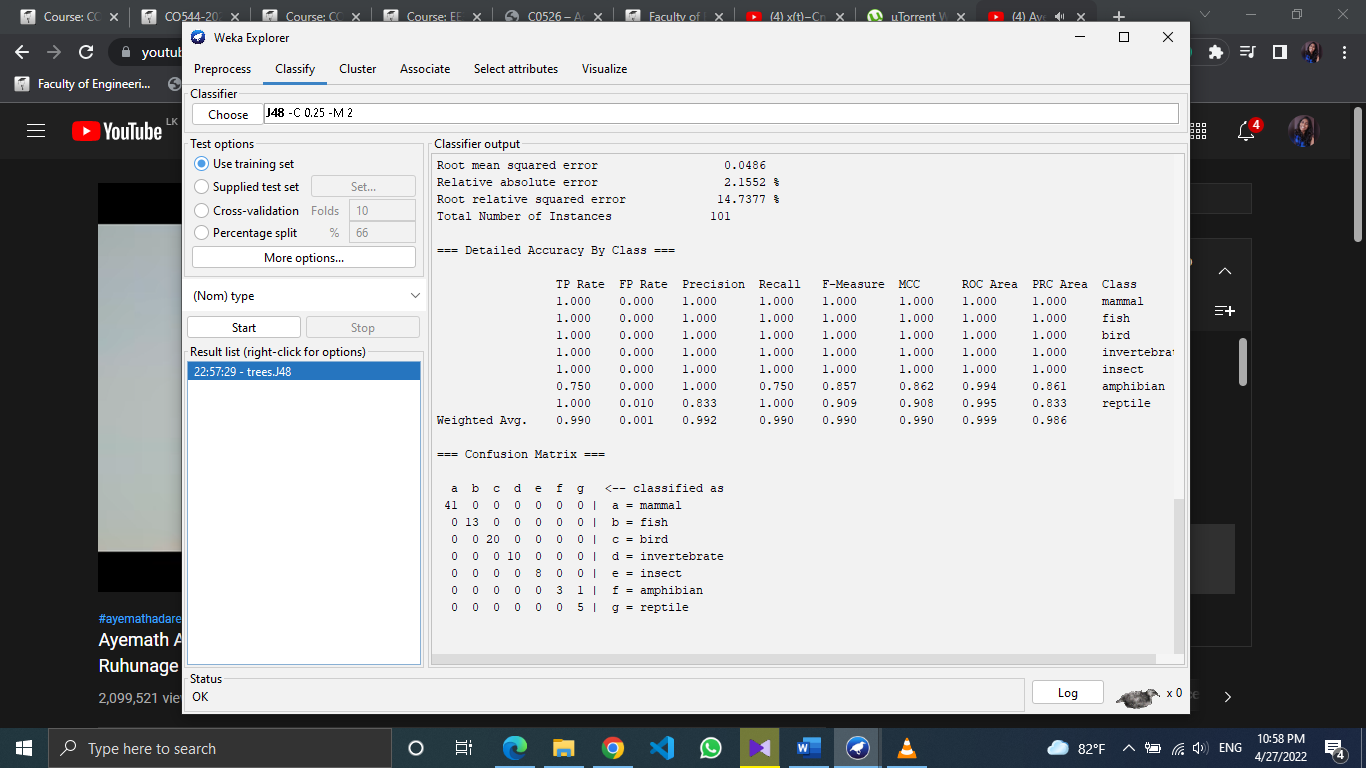
Mean absolute error = 0.0047



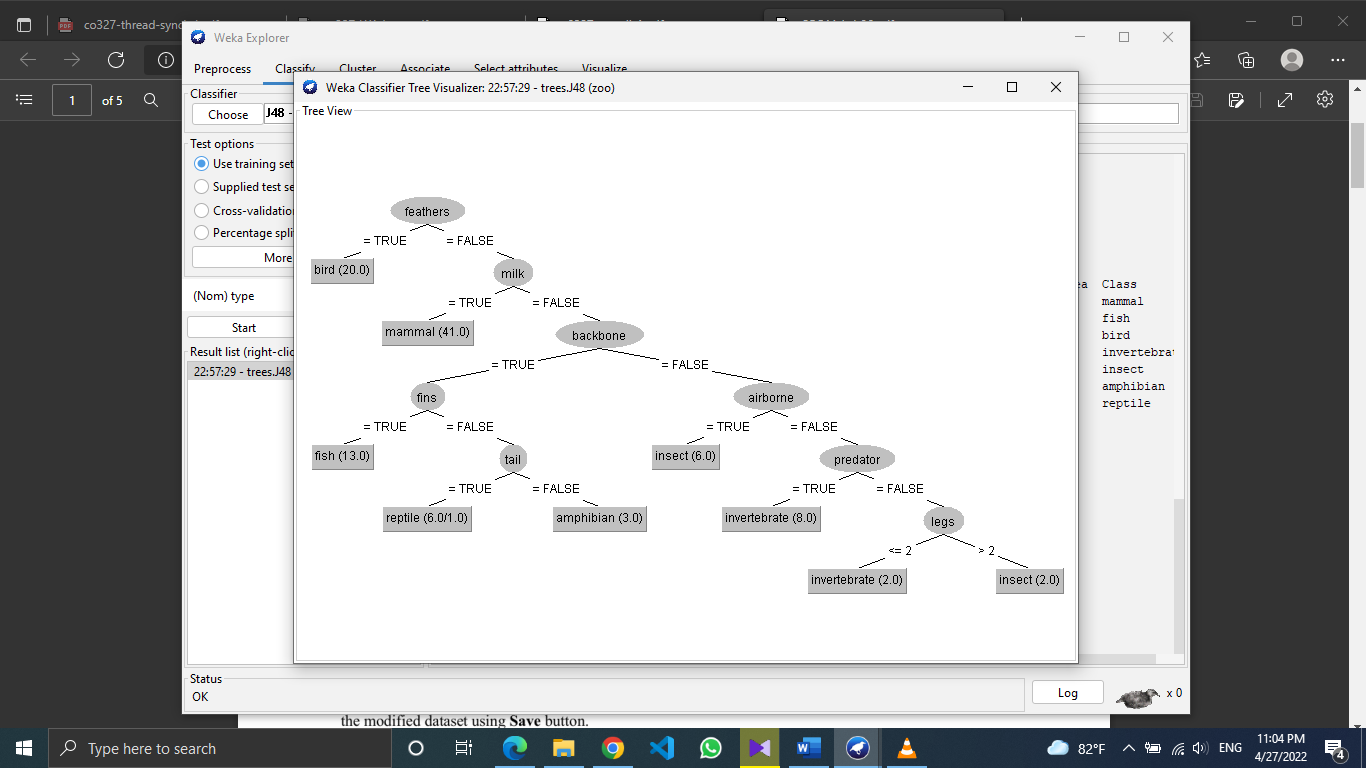
These are the True Positive and False Positive rates for each class. As it is observed, weighted average TP rate is 0.990 and weighted average FP rate is 0.001. The only FP happened involved in the ‘Reptile’ class.



This is the confusion matrix obtained from the classification. If the accuracy is hundred percent only diagonal elements are filled with the numbers in the confusion matrix. But here there is an element in (g,f) position in addition to the diagonal elements. This means 1 ‘f’ classed animal (amphibian) has classified as a ‘g’ classed (reptile) animal. This detail was partially obtained from the FP rate column in the previous table too.

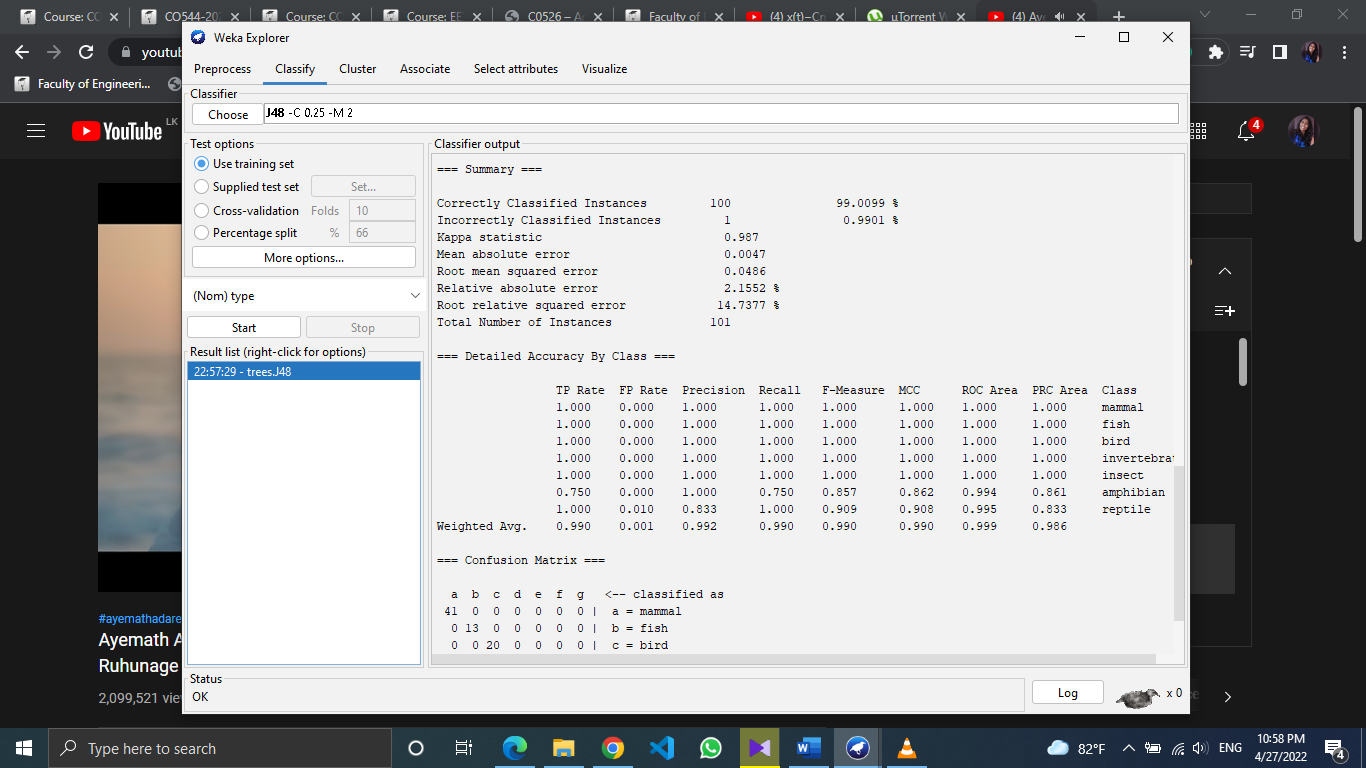


Decision tree

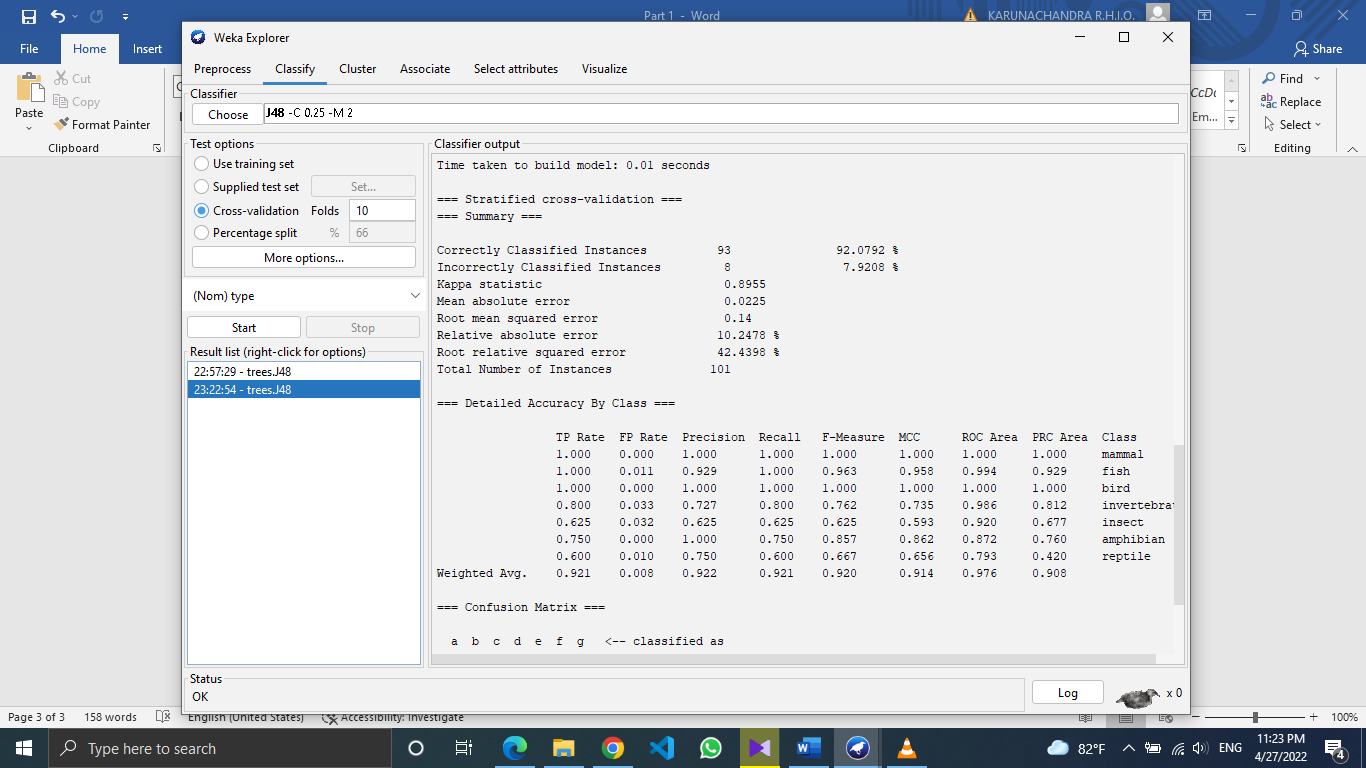


4.

Training set



10-fold cross validation



When comparing accuracy values, it can be seen that,

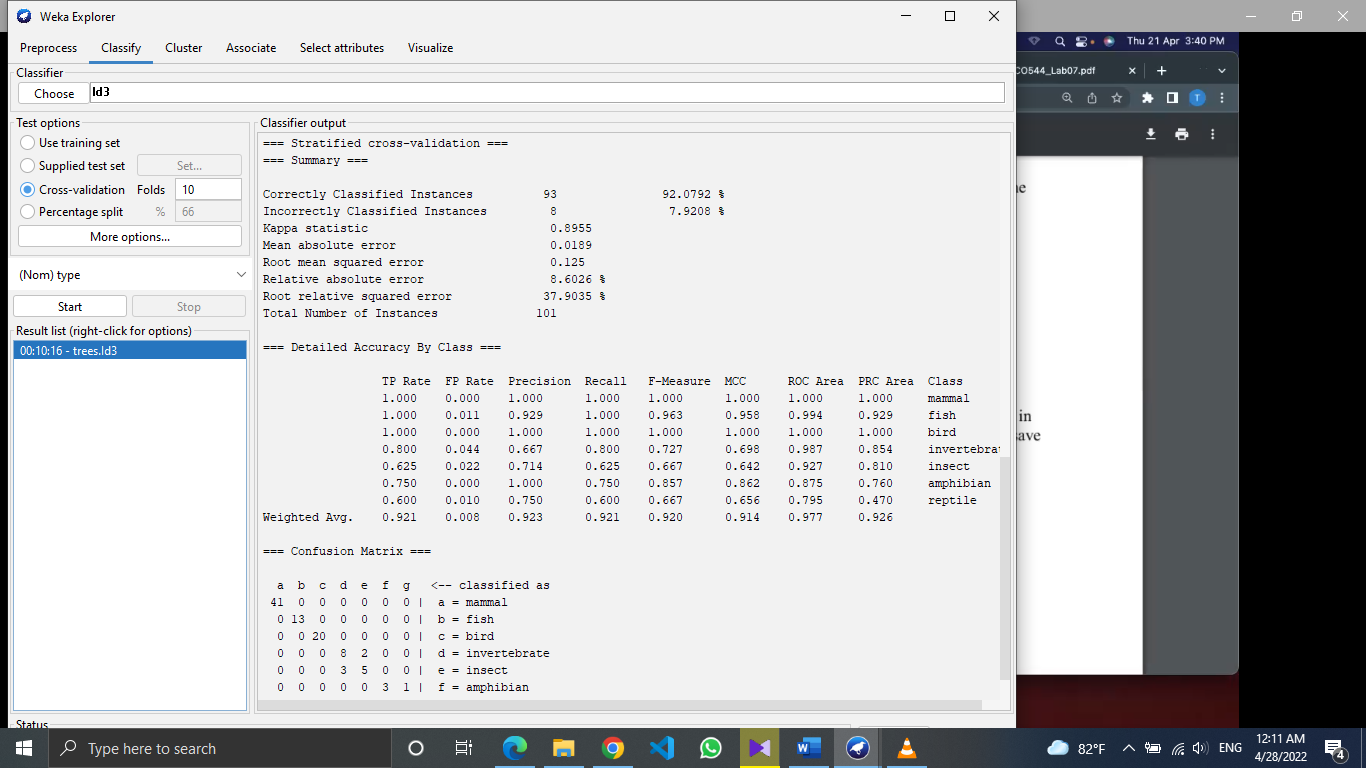
Accuracy (Training set) > Accuracy (10-cross validation)

Hence, training test model will give a better future performance with compare to 10-fold cross validation model.

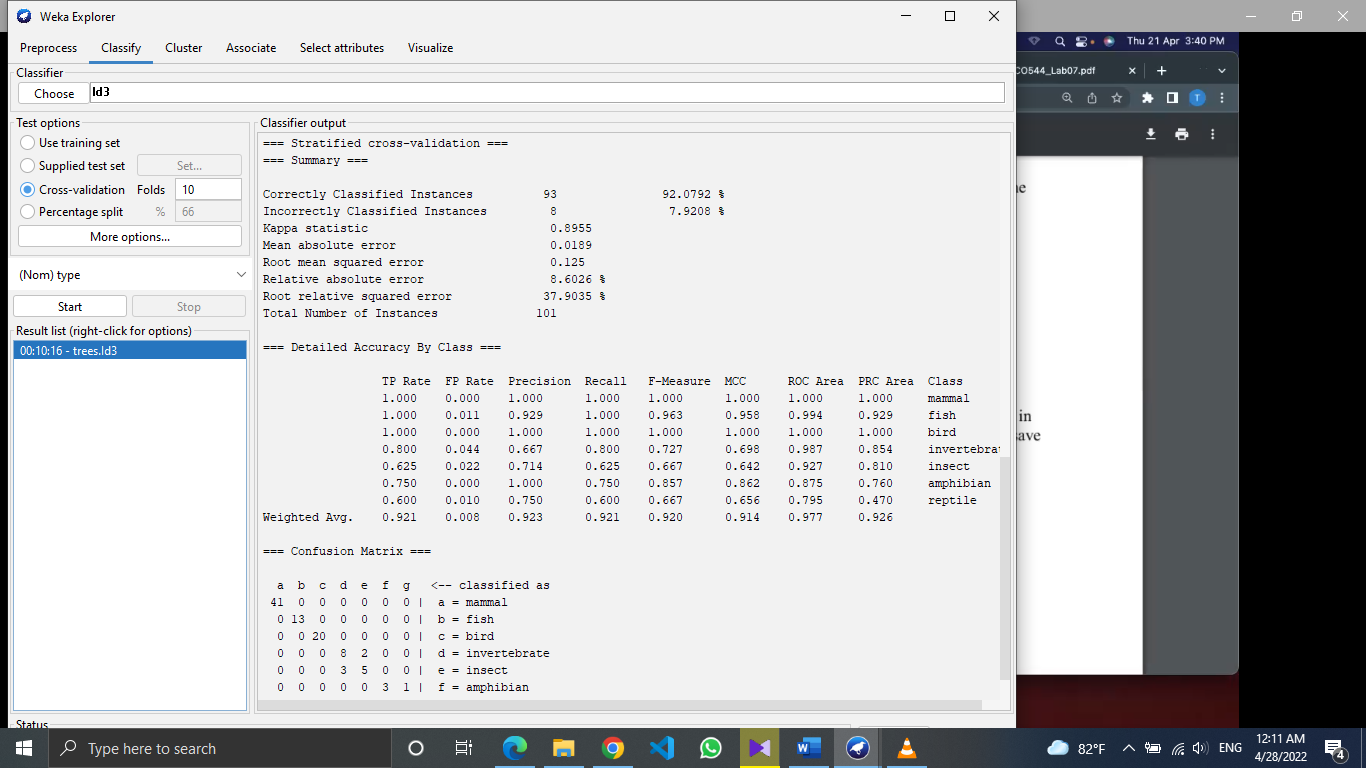
Cross validation is usually used for small datasets. Here, it randomly divides the set of observations into 10 folds and one of them is treated as test set. This is run 10 times. Since this run 10 times shown performance is the average across 10 times. So, the incorrectly classified instances might be higher than in the training set.

5. We can’t apply ID3 learning algorithm on this dataset since this algorithm only deals with nominal attributes. Here we got an attribute which has numerical values. So, the original dataset does not support the ID3 algorithm.

7. After removing animal name and legs attributes, ID3 decision tree was built. As the summary shows, 93 instances were classified correctly and 8 were incorrect. So, the accuracy was 92.0792% and this is still a lesser accuracy than the training set model.

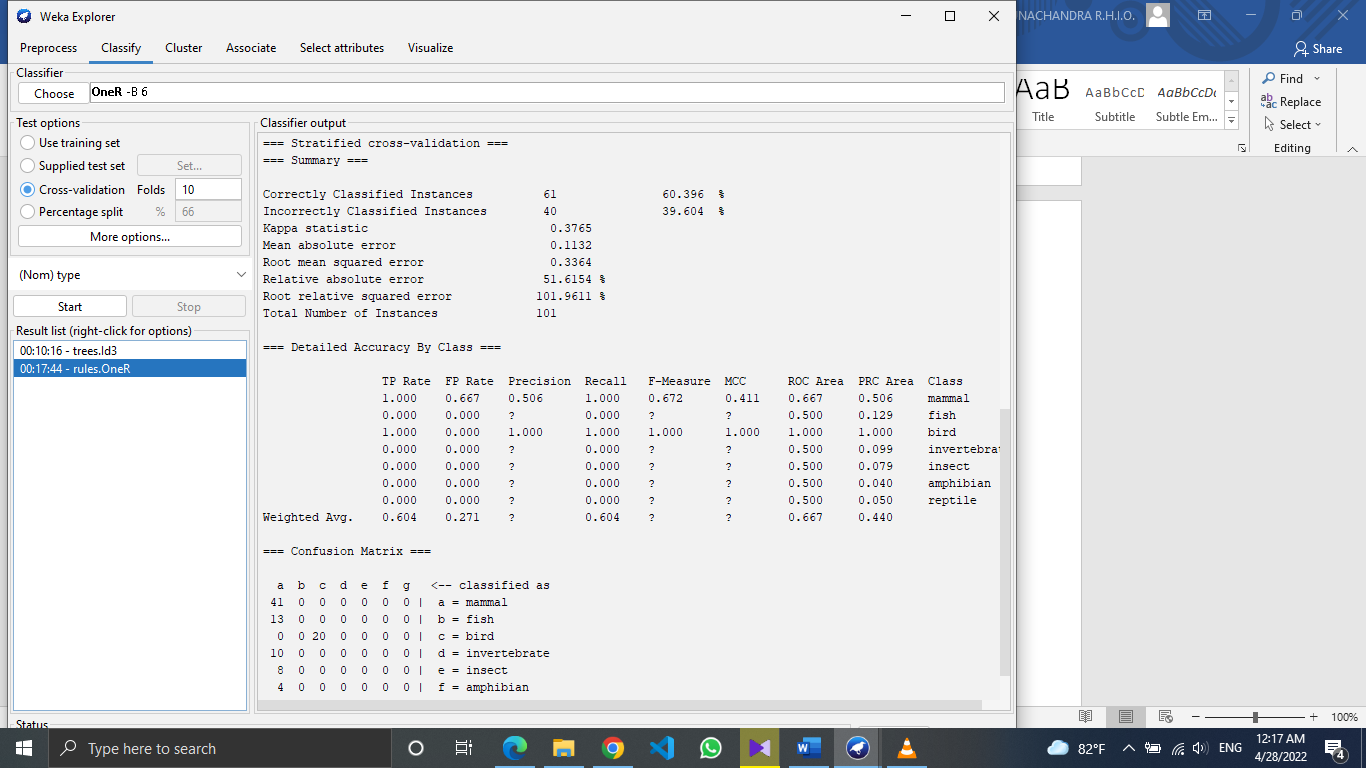


By studying the confusion matrix, incorrectly classified data could be identified.

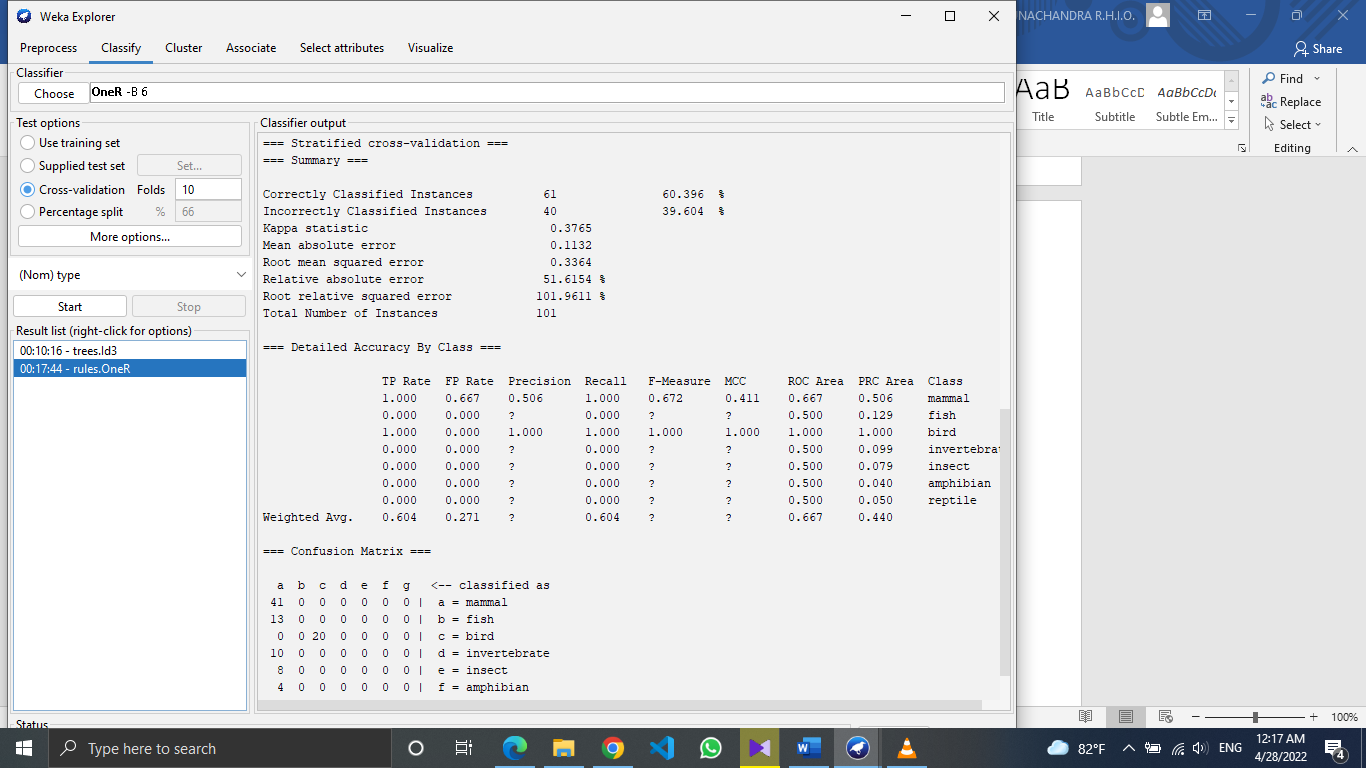


8. OneR algorithm

Only 61 instances were classified correctly and it decreased the accuracy into 60.396%



From the confusion matrix we can see that only mammals and birds have classified into their true category and all the others are mis classifies.



9. Prism algorithm

PRISM is a separate and conquer algorithm based on ID3’s cons.

ID3 doesn’t consider whether an attribute might be highly relevant to only one classification and irrelevant to the others. In PRISM a branch could be considered as an attribute-value pair. It considers the relevance between an attribute-value pair and the specific classification.

Since there is no big effect of the relevance in this dataset, PRISM algorithm doesn’t show much difference in the classification compared to ID3. But it obviously gives a better performance than OneR algorithm in this case.

