# CS342: Machine Learning

Plants vs Animals Report

u1617935

November 4, 2018

## Methodology

In order to gauge which attributes provide the best insight into the classification of the organism, multiple approaches were taken to gain enough information to reach an accurate and informed conclusion.

The first approach was to plot the class assignments for each attribute whereby the values for this attribute are plotted on both the x and y axis seen in figure 3. This visual representation of the data is extremely useful, as it allows for quick and easy identification of the attributes which display promising properties. Such properties would be distinct regions on the graph in which only certain classifications fall under/display an overall trend in the areas in which certain classifications occur. A good example of an attribute which displays these features is 'tna', which from figure 3 it is clear to see that there is a distinct pattern displayed in the data. Lower values of tna are almost all the same classification, while as tna increases it begins to change the predominant classification until it reaches the large tna values where again almost all points are the same classification. The very clear presence of a predictable trend in the data makes tna a very useful attribute, particularly considering no other attributes display such clear-cut easily identifiable patterns.

After visually inspecting the attributes to estimate their levels of importance, a more concrete and accurate approach was needed. Two tests were used, one of which attempted to predict the classification of the organism based on a single attribute. This was carried out for all attributes to find the most useful and predictive one. The other approach was to use the built in feature importances feature of the RandomForrestClassifier, which assigns a value to each attribute based on how much of the observed variance in the target data is explained by that attribute, i.e. the importance of that attribute in respect to predicting the target variable.

## Feature Importance

The results were attained from the 'featureImportance.py' script included in the submission zip file. Figure 2 below shows the results from the feature importance taken from the best model produced (the one with the highest AUC on the validation set). Here it is clear that the initial conclusion made from the aforementioned visualisation of the attributes was correct; tna appears to be the most influential and important attribute, which is then followed by n2 and m2. These are the three stand-out attributes, with the remaining data displaying similar and considerably smaller influence on the classification.

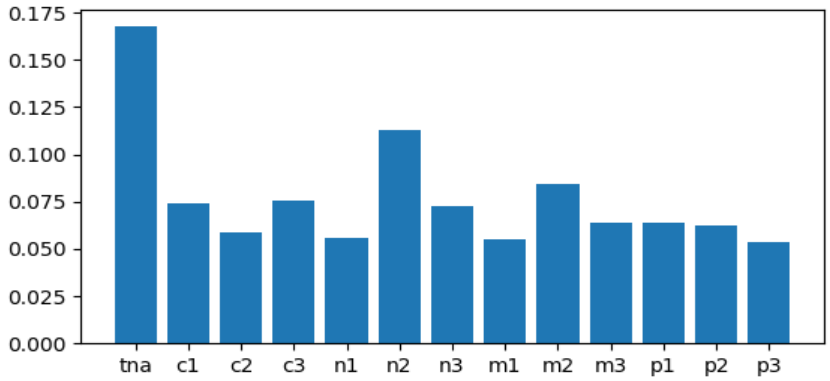


Figure 1: A graph representing the feature importance of the given attributes

## Single Attribute Based Predictions

In the figure below, the AUC found for the best model for each respective attribute has been calculated, scaled and then displayed. This aligns partially with the findings from the previous two results. Clearly here tna once again comes out on top, followed by n2. However, differences materialise from this point onward, with n1 taking 3rd spot opposed to m2 as is the case in the feature importance findings. Infact m2 produces the lowest AUC of all attributes, a suprising result indicating that neither one of these two tests is thorough enough to yield complete and telling results.

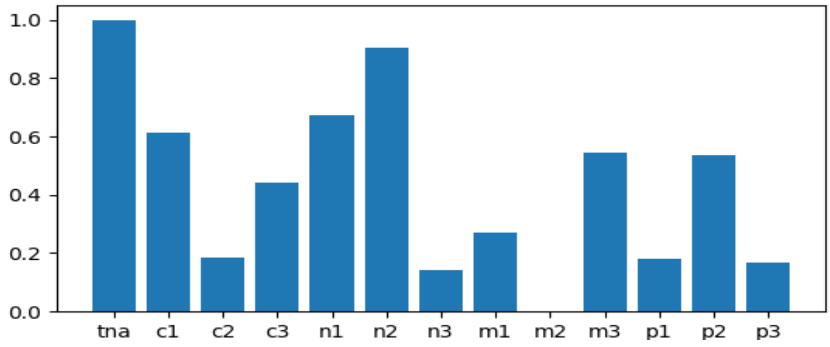


Figure 2: A graph representing the MinMaxScaled AUC results of predictions made by the best model using only the respective attribute

## Conclusion

Despite some findings from the different methods used find the best attributes for predicting the classification of the organism conflicting with each other, there is more than enough evidence found by all 3 measures employed to conclude with confidence that tna is the most useful attribute for predicting the class.

The 2nd best attribute, (based on evidence from the feature importance and single attribute predictions), is clearly n2. Past this point it is difficult to accurately and confidently say which attributes are better, as the attribute based predictions and feature importance data fail to align in their findings.

## Evaluation

As stated earlier in the report, the testing methods used to analyse the data are not thorough enough to yield complete results. The visualisation step is particularly unreliable and arbitrary as only very distinct and clear patterns, (such as those observed in tna), will be visible to the human eye when displayed in such a way. As a result, more testing would be needed to accurately rank the attributes any further.

An area failed to be addressed in the methodology used in the approaches which have been presented is the combination of attributes with each other to form new ones. Such testing could uncover a further level of relations between data abstracted away from and therefore invisible to the approaches used for this report. After combining these attributes together the tests used in this report would be applicable, and allow for a fuller and more informed insight into the data and its relation to the classification of the organism, in addition to providing new attributes with which to use for training and modelling, thereby allowing for more accurate models to be produced.

## Appendix

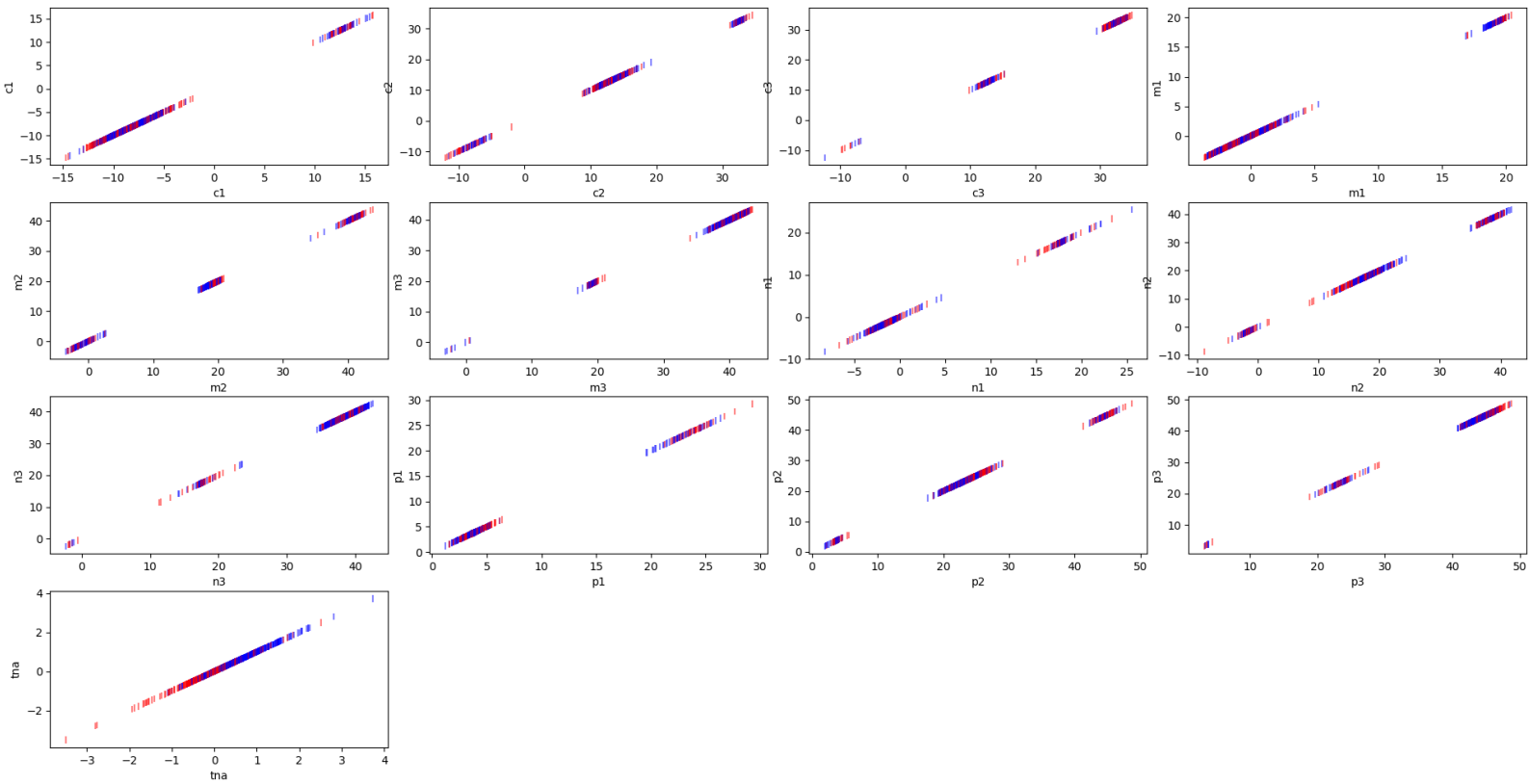


Figure 3: A graph whereby the values of the named attributes are plotted along the axis with the classifications plotted. Red representing classification of 0 and blue representing a classification of 1