**Capstone 2 Final Report:**

Problem statement: Classification is used across all fields and we use it every day in everything we do. If we can formulate a classification model to accurately predict and classify an animal, then similar models can be utilized in other classification tasks such as in the implementation of medical diagnoses, in the filtering of spam from ham, in the learning of associations between product purchases, in the determining of loan disbursement criteria, and in the recognition of speech and images. In addition, this animal classification project can be useful in identifying attributes associated with different animals in order to understand the animal kingdom better. For example, it can be used to classify venomous vs non-venomous animals for outdoor public safety, it can be used to identify attributes associated with endangered animals and it can be used to classify wild animals from domestic animals.

Clients: Clients that would be interested in this classification project would be scientists, zoologists, researchers, archeologists, park rangers, paleontologists, wildlife conservationists, microbiologists, veterinarians and educators. Also, since this classification problem can be applied to many other classification problems, as specified above, there are many clients that would benefit from the models implemented in this project.

**Data:** The data set for this project consists of 17 boolean valued attributes used to classify over 100 zoo animals. The acknowledgment for this data set is the: (UCI Machine Learning Repository, Dua, D. and Graff, C. (2019),[http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science)[1]. There were 99 unique animal names in the dataset with frog repeated twice since one frog was venomous and the other was not. Since the UCI dataset included ‘girl’ in the row of animal names, that row was dropped. The column name ‘tail’ was changed to ‘tails’ in order to not be confused with the Pandas function tail( ) and the number of animal species in class 1 was adjusted to 40. Once the response variable was set aside, there were a total of 16 different Boolean valued features. The response variable, ’Class\_Type’, consisted of 7 different Class Types which included: Mammal, Bird, Reptile, Fish, Amphibian, Bug and Invertebrate. The two datasets were checked for missing values and merged together on their common column,’Class\_Type’. Non-numeric columns, such as ‘Animal\_Name’, were removed before utilizing Scikit-learn machine learning models. Finally, the feature “Number of Animal Species In Class”, although numeric, was also removed since it made the machine learning models un-challenging.

Exploratory Data Analysis: As part of exploratory data analysis, the correlation heatmap found the following trends:

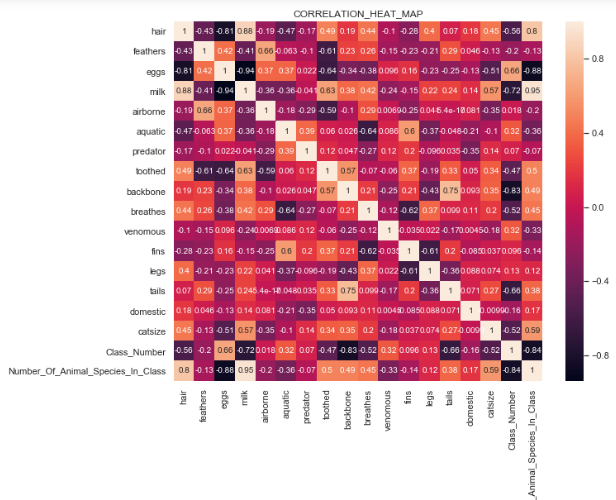
**1.)High correlation between teeth and backbone which is due to "Teeth being formed from the neural tube, the embryonic tissue that becomes the brain and spinal cord.[2]"**

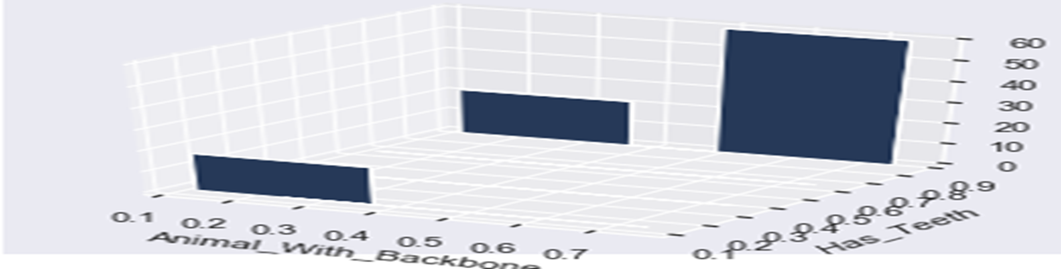
**2.)High correlation between aquatic and fins.**

**3.)High correlation between tail and backbone, which according to Carnegie Museum of Natural History[3], is because "the tail is part of the caudal vertebrae".**

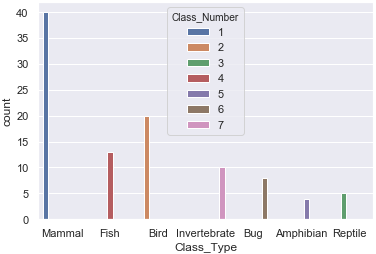
**4.)Low correlation between domestic animals and predators.**

**5.)High correlation between breathing and animals with hair and strong negative correlation between not breathing and fins since gills collapse out of water.**

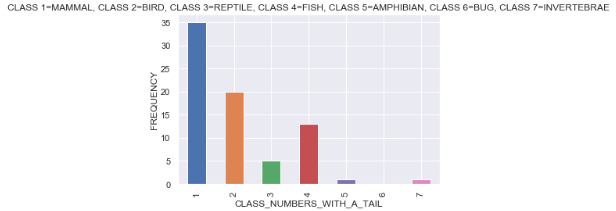




The 3D histogram above agrees with the correlation heatmap finding of there being a strong correlation between animals having a backbone and having teeth. The histogram shows that the reason the correlation is only 0.6 and not greater, could be due to the fact that there are animals without a backbone having teeth.



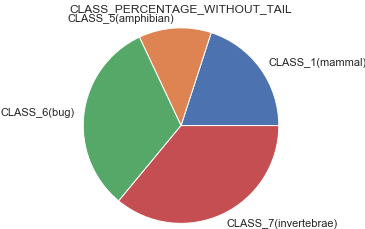
**The histogram above shows how the majority of animals are in class number 1(mammals) and include: aardvarks, antelopes, bears, and boars-just to name a few.**

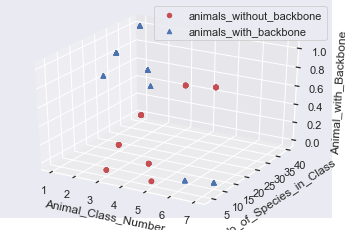


**The bar plot above shows that 75% of all animals in the dataset have tails, and 50% of those 75% are mammals (Class 1). According to the bar plot, none of the bugs (Class 6) have tails and very few amphibians and invertebrates (Class 5 & 7) have tails. The graph also shows that Birds, Reptiles and Fish (Classes 2, 3 & 4) all have tails. Since only one class of animals, bugs, which consists of only 8 out of the total 100 animals, were found to not have tails, it is plausible that tails may not be the strongest classification feature.**

**Other Visualizations and Trends seen:**

* **Since there was a 0.75 correlation on the heatmap between animals having a backbone and animals having a tail, there is a possibility that animals without a backbone are correlated with not having a tail. The pie chart below shows this exact finding and also shows that the highest percentage of animals without a backbone are invertebrates.**



* **The 3D scatterplot below showed that all animals in classes 1, 2, 3 ,4 and 5 have a backbone. The scatter plot also showed that none of the animals in classes 6&7 have a backbone which makes sense since those classes include invertebrates and bugs. Based on the scatter plot findings, backbone may not be the strongest feature for animal classification.** 
* **The correlation heatmap showed a negative correlation between animals that are predators and animals that are airborne. The seaborn scatterplot validated the heatmap finding by showing that the majority of predators are not airborne. It also showed that predators may not be the strongest feature in animal classification.**
* **According to the correlation heatmap, the attribute domestic had a low correlation with all of the other attributes. The seaborn scatterplot showed that domestic animals fell in the classes of 1, 2, 4 and 6 which corresponds to: mammals, birds, fish and bugs respectively. The scatterplot also showed that all of the classes contain animals that are not domestic. This may explain why domestic had such a low correlation with all of the other attributes. The scatterplot showed that the only domestic animals that were aquatic were class 4(fish). Finally, the scatterplot indicated that domestic may not be the strongest feature in animal classification.**
* **The seaborn scatterplot showed that finned animals belong to classes 1 & 2(mammals and fish). The scatterplot also showed that 6 out of the 7 total classes include animals that are aquatic and that 5 out of the 7 total classes include animals that are not aquatic. This agreed with the low correlation on the heatmap and indicated that aquatic may also not be the strongest feature for animal classification.**
* **The matplotlib barplot showed that there were no venomous animals in classes 1 & 2 (mammals and fish). It also showed that the majority of venomous animals are in classes 3, 6 and 7(reptiles, bugs and invertebrates). The bar plot showed that the attribute venomous may not be the strongest feature in animal classification.**
* **In conclusion, data exploration and visualization findings showed that the following features may not be the strongest predictors of animal class: 1.)domestic, 2.)tails, 3.)aquatic, 4.)backbone, 5.)venomous and 6.)predator.**

## **Hypothesis Testing #1:**

### null hypothesis: animals that can breathe that have fins - animals that can't breathe that have fins = 0

### alt hypothesis: animals that can breathe that have fins - animals that can't breathe that have fins != 0[¶](http://localhost:8888/notebooks/Desktop/cap_stone_project_2/Zoo%20Classification.ipynb#alt-hypothesis:--animals-that-can-breathe-that-have-fins---animals-that-can't-breathe-that-have-fins-!=-0)

Conclusion**:**

Since the p-value was greater than 0.05, we failed to reject the null hypothesis that there's no statistical difference. In other words, there was no statistical difference between animals that can breathe with fins and animals that can't breathe with fins. This statistical conclusion makes sense since our correlation heatmap showed a strong negative correlation between animals that can breathe atmospheric air and animals having fins.

**Hypothesis Testing #2:**

**null hypothesis: animals that are domestic that are predators - animals that are not domestic that are predators = 0**

**alt hypothesis: animals that are domestic that are predators - animals that are not domestic that are predators != 0**

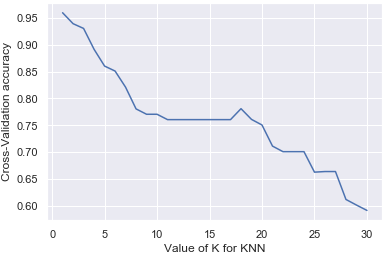
**Conclusion:**

**Since the p-value was less than 0.05, we rejected the null hypothesis and concluded that there was a statistical difference between domestic animals that are predators and non-domestic animals that are predators. This statistical conclusion makes sense since our correlation heatmap showed a low correlation between domestic and predator. This also agreed with the scatterplot finding that the attribute domestic may not be the strongest feature in animal classification.**

**Machine Learning Algorithms:**

1. **Decision Tree Classifier- utilized all 16 features initially but then used tuning of hyperparameters and sci-kit learn feature selection to give a test accuracy of 0.97, cross validation score of 0.97, precision of 0.97, recall of 0.97 and F1 score of 0.97.**
2. **SVM Classifier-utilized all 16 features to get test accuracy of 0.97, cross validation score of 0.95, precision of 0.94 , recall of 0.97 and F1 score of 0.95.**
3. **Gaussian NB Classifier-utilized all 16 features to get test accuracy of 0.97, cross validation score of 0.95, precision of 0.94, recall of 0.97 and F1 score 0.95.**
4. **Random Forest Classifier-utilized 9 features from ski-kit learn feature selection to get test accuracy of 0.97, cross validation score of 0.98, precision of 0.94, recall of 0.97 and F1 score of 0.95.**
5. **KNN classifier-utilized chi square feature selection in order to pick the best 10 features and also plotted and calculated the optimum ‘k’ value that gave the highest accuracy. The test accuracy was 0.97, cross validation score was 0.97, precision was 0.94, recall was 0.97 and F1 score was 0.95.**

**\*The Scikit-learn feature selection method as well as the Chi squared feature selection method dropped features that were synonymous with those features found to be weak classification attributes during data exploration and visualization.\***



1. **Keras Sequential Deep Learning model utilizing SGD optimizer, single layer perceptron, mse loss function and 50 epochs gave near 90% test accuracy.**
2. **Tensorflow Deep learning model with single layer perceptron and one hot encoding in order to binarize categorical data values and allow for a multi-class classification response gave a best test accuracy of 95%.**

**In conclusion, the decision tree classifier was the best at animal classification. Even though, all the Scikit-learn machine learning models had the same test accuracy, the decision tree classifier was chosen as the best classifier as it also had the highest precision, recall and F1 score.**

**The deep learning models gave a lot of valuable info about loss evaluation metrics. Namely, the loss metrics showed that the ANN model had difficulty classifying class number 7(invertebrates). This could be due to a combination of there being only 10 animals in class number 7 or due to the data exploration and visualization findings showing that backbone was not the strongest predictor of animal class. The deep learning models' evaluation metrics showed that individually dropping the features of ‘domestic’, ‘aquatic’ or ‘tail’ did not change the test accuracy even though dropping any one of the other attributes decreased the model's overall test accuracy.**

**This was an interesting finding as it matched our earlier data exploration and visualization observations that the attributes of ‘domestic’, ‘aquatic’ and ‘tail’ were weak predictors of animal class. This project was unique in that it had a multi-class classification target and an imbalanced dataset which affected accuracy. However, tuning the hyperparameters in the models and choosing the best model surmounted those dataset hurdles and made for a successful classification algorithm.**

Acknowledgments:

[1] UCI machine learning repository(Dua, D. and Graff, C. (2019). UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science)

[2]<https://carnegiemnh.org/what-is-in-a-tail/>

[3]https://www.ncbi.nlm.nih.gov/m/pubmed/15269893/

By: Roshnee Raval

July 2019