**Milestone 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 learning associations between product purchases, in determining loan disbursement criteria, and also in recognizing images and speech. This capstone 2 animal classification project can also be helpful 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, identifying attributes associated with the endangerment of animals and classifying wild animals from domestic animals.

Clients: Clients that would be interested in this classification problem 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 other clients that could also benefit from the classification models implemented in this project.

**Data:** The dataset for this project consists of 101 animals from a zoo from the UCI machine learning repository (Dua, D. and Graff, C. (2019). The acknowledgment for the data source is: UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml],( Irvine, CA: University of California, School of Information and Computer Science)[1]. There were 99 unique animal names with frog repeated twice since one frog was classified as venomous and the other was not. Since the UCI dataset included human in the row of animal names, that row was dropped, the column name tail was changed to tails, and the number of animal species in class 1 was adjusted to 40. There were a total of 16 boolean column values assigned to different animal attributes. Both datasets included the Class\_Type column. There were 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. The common column was the response predicted, or in this case the animal class, from the classification of 16 different Boolean valued animal attributes. Non-numerical columns were removed before beginning sci-kit learn machine learning classifiers. The variable “Number of Animal Species In Class” , although numerical, was also dropped since it made the machine learning models too easy.

Exploratory Data Analysis: The correlation heatmap found the following trends:

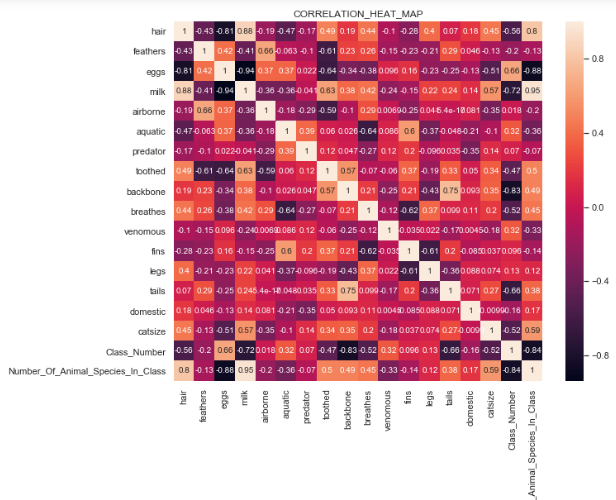
**1.)High correlation between teeth and backbone which was due to "Teeth being formed from the neural tube, the embryonic tissue that becomes the brain and spinal cord.[2]"**

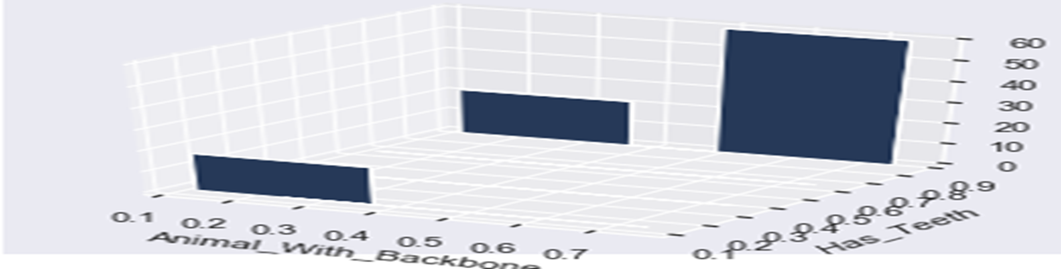
**2.) High correlation between aquatic and fins which, of course, makes sense.**

**3.)High correlation between tail and backbone which also makes sense**

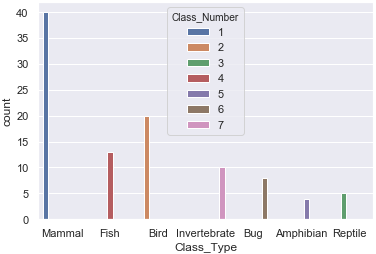
**4.)Low correlation between domestic animals and predators which also makes sense**

**5.)There was high correlation between breathing and animals with hair and strong negative correlation between not breathing and fins**

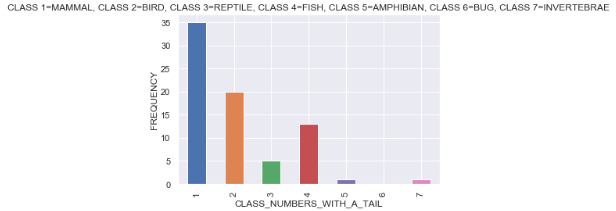




The heat map visualization showed how the correlation between teeth and backbone was one of the highest correlations at close to 60%. The 3d bar plot above showed why it is not higher than 60%. Specifically, it showed that there are a small number of animals without a backbone that have teeth even though the majority of animals with a backbone have teeth.



**The majority of animals fell in class number 1 (mammals ) which included: aardvarks, antelopes, bears, and boars-just to name a few. The visualization above shows how the different classes were distributed.**



**What this shows is that 75% of the animals in the dataset have tails and 50% of those 75% are mammals(Class 1). It is important to note, however, that not all mammals have tails. NO bugs(Class 6) have a tail and very few amphibians and invertebrates(Class 5 & 7) have a tail. Finally, 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, do not have tails, it is plausible that tails may NOT be the strongest feature in our classification algorithms.**

**Other Visualizations and Trends seen:**

* **The pie chart showed that the highest percentage of animals without a tail also did not have a backbone ( invertebrates). Even though the correlation heatmap showed a high correlation between having a backbone and having a tail, the pie chart showed that there might also be a lot of animals without a backbone that do not have a tail.**
* **The 3D scatterplot showed that class 1 animals(mammals) have a backbone and also showed that the majority of animals in the dataset have a backbone and thus backbones may not be the strongest feature for animal classification. Finally, the 3D plot showed that many animals in classes 6&7 do not have a backbone which makes sense since those classes are invertebrates and bugs.**
* **The seaborn scatterplot showed that the majority of predators are not airborne.**
* **The seaborn scatterplot showed that domestic animals fall in the classes of 1, 2, 4 and 6 which corresponds to: mammals, birds, fish and bugs respectively but that all the classes contain animals that are not domestic. This along with the low correlation on the heatmap may indicate that domestic is not the strongest feature for animal classification. Finally, the scatterplot showed that the only domestic animals that are aquatic are class 4(fish).**
* **The seaborn scatterplot showed 6 out of the 7 classes included aquatic animals, with the exception of bugs, and that 5 out of the 7 classes have animals that are not aquatic. This visualization, along with the low correlation with other features on the heatmap, also showed that aquatic may not be the strongest feature in our machine learning classification models. Finally, this showed that finned animals belong to classes 1 & 2(mammals and fish).**
* **The seaborn bar plot showed that there were no venomous animals in classes 1 & 2(mammals and fish). It also showed that the majority of venomous animals were in classes 3, 6 and 7(reptiles, bugs and invertebrates). Overall, venomous animals were seen to be pretty evenly distributed among the different animal classes and thus may also not be the strongest predictor for animal classification.**

## **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 is greater than 0.05, we fail to reject the null hypothesis that there's no statistical difference. In other words, there's 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 the 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**

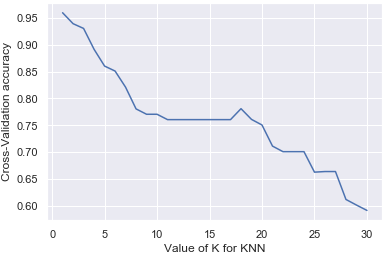
**Conclusion:**

**Since the p-value is less than 0.05, we reject the null hypothesis and conclude there is a statistical difference between animals that are domestic that are predators and animals that are not domestic that are predators. This statistical conclusion makes sense since our correlation heatmap showed a very low correlation between domestic animals and animals that are predators. In addition, the attribute domestic does not appear to be a strongly weighted 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 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 sci-kit learn feature selection method as well as the chi square feature selection method dropped features that agreed with data exploration/visualization determinations of weak classification predictors .\***



1. **Keras Sequential Deep Learning model utilizing SGD optimizer-near 90% test accuracy.**
2. **Tensorflow Deep learning model with single layer perceptron using one hot encoding due to categorical data- A multi-classification target response with 93% test accuracy.**

**In conclusion, the decision tree classifier was the best at animal classification. The data-set was unique in that it had a multi-classification response (7 classes of animals). Even though, all the sci-kit learn ML 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 the loss evaluation metrics. Namely, the metrics showed that the classifiers had difficulty with classifying animal class number 4 (amphibians) and this could be due to there only being four animals in class number 4 . The deep learning model evaluation metrics also showed that individually dropping the features domestic, aquatic or tail did not change the deep learning test accuracy which was interesting since it exactly matched our data exploration conclusions.**

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)