Data Analysis, Prediction and Classification using Machine Learning Algorithms to Identify Abalone Gender

Carl Jeong, *M.Eng – Mechatronics Engineering Project Section*, *Toronto Metropolitan University*

***Abstract*—Taking a data set with a series of rows and columns, this research project is meant to design a subsection/portion of a neural network that can take in datasets and produce results as well as predict the probability of gender if specific conditions are decided. Now AI model is taught to predict a possible category and provide a probability of the data being a certain category. Taking the model and dataset from the UC Irvine Dataset Repository, this tabular source is used in the machine learning algorithm in order to produce comparisons between Female, Male, and Infant data. Zero Mean Score, Min Max, Robust methods, and Learning Categorization are all concepts used in data scaling. If there was more data to be updated and fed into the algorithm, this will cause the number of characteristics to improve and develop. There is an accuracy rating but it is still very low but a good start to learn more about how to take data and begin predicting with it.**

I. INTRODUCTION

THE science of Machine Learning is advancing and allows humans to come up with algorithms to create their own ideas in order to aid and solve challenging problems. The modern world has various problems, such as marine science, biomedical, entertainment industry, even space travel, and taking in data to identify or sort out the data is far more complex than one can think of.

In order to perform a prediction analysis using datasets, a data set containing data of abalones is used as an example. With this test, the process is to classify and sort out categories and determine information given. I also will need to perform prediction and determine decision making and train that to an AI. Abalones are giant marine mollusks, not like octopus, but shelled, like a clam. These fascinating simple yet practical creatures are used a lot in seafood, but also used in biotechnology and medicinal related ingredients. The features involve rings, length, height, whole weight, shell weight, and how to determine these characteristics depending on the gender of the abalones. We can deduce and come to a conclusion from categorizing and classifying the data.

Now we can also use Regression as there are different types of regression. Such as logarithmic, polynomial, multistep, these are all complex and nonlinear forms of regression with their own set of codes and functions used for calculations and predictions, but for the sake of this project, this research uses classification and determining types of characteristics depending on the three categories: Male, Infant, and Female.

Data scaling or normalization is an additional step between the two possibilities in the dataset to be tested. In this process, there are two possible ones: normal and non-normal distribution. The normal data distribution is ready to be processed, while the non-normal data needs to be normalized. The reason is that when new normalize data, we prefer the symmetry and differences will make the output more likely to be incorrect or false. Non-normal distribution focuses on chi-squared distributions, asymmetrical shapes and bell curves, these must be normalized in order to reach on outcome.

In general, normalization techniques or data scaling have an important role in data preprocessing. Min-max is more for normalization and Z-mean score is for standardization. And now for the performance.

Manuscript received December 20, 2023 Student – Carl Jeong. This is Selected Topics for Computer Engineering – Machine Learning. Completed tasks and program in Toronto Metropolitan University

II. Classification and Data Analysis

A. *Initial Data seen*

From what is required in this research project, the code opens up and there are files which contain the dataset to be analyzed. There are three groups, Female (F), Male (M), and Infant (I) and these three play a role in classification.
The first step would be to extract the dataset from this source:

Source: https://archive.ics.uci.edu/dataset/1/abalone

From UCI Education Sources, we can now use the dataset. Data collecting, processing, scaling, sharing, modeling, analysis, prediction, probability, and evaluation is all donne in this process.

B. *Preprocessing Data*

This research paper requires data balance

Before the initial setup with the data here are the following units as reference when working on them

Table 1. Abalone Characteristics and Units

Length	Mm
Diameter	Mm
Height	Mm
Whole Weight	Grams
Shucked Weight	Grams
Viscera Weight	Grams
Shell Weight	grams

For the rings you need add 1.5 to get the age in years.

One of the popular types of processing data is also Synthetic Minority Oversampling Technique method. Or Simply called SMOTE.

SMOTE takes the K from the KNN class and finds the closest neighbor of the data in the minority class.

However we need to do more steps than that. There are tests done using the Zero-Mean normalization method is based on the mean and standard deviation. Standardizing a dataset changes the value distribution scale, the mean is 0 and the standard deviation is 1. Standard deviation is calculated to get the average.

Another test that will be done is the Min-Max normalization technique. Data ranges are set and usisng pandas and pycaret we can get information with this:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4177 entries, 0 to 4176
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Sex             4177 non-null   object
 1   Length          4177 non-null   float64
 2   Diameter        4177 non-null   float64
 3   Height          4177 non-null   float64
 4   Whole weight    4177 non-null   float64
 5   Shucked weight  4177 non-null   float64
 6   Viscera weight  4177 non-null   float64
 7   Shell weight    4177 non-null   float64
 8   Rings           4177 non-null   int64
dtypes: float64(7), int64(1), object(1)
memory usage: 293.8+ KB
```

Figure 1. Categories and Classification with pandas

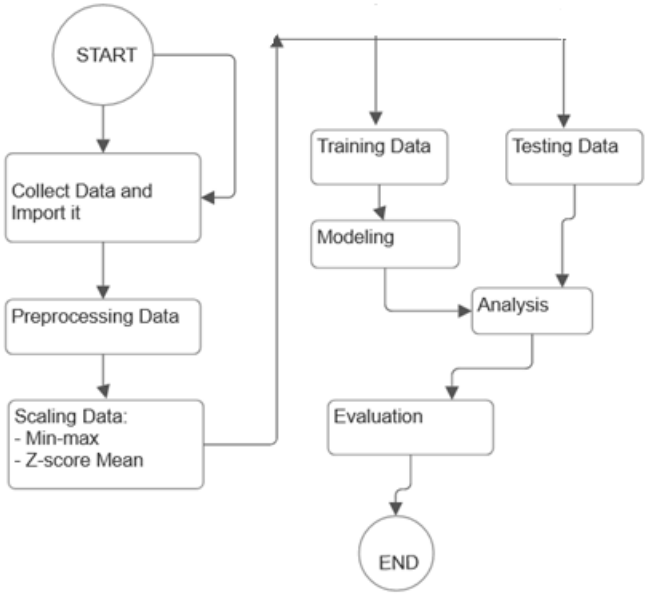

Figure 2. Research Stages

$$z_{std} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(z_i - z_{mean}\right)^2}$$

$$z_i' = \frac{z_i - z_{mean}}{z_{std}}$$

Figure 3. Z-score Formulae for Standardization

C. *Classification*

In this step, in order to classify data, functions and code from pycaret is used to demonstrate how to classify data from the abalone dataset.

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Ring
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.00000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.93368
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.22416
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.00000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.00000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.00000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.00000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.00000

Figure 4. Statistics by Describe Function

Now we have an idea on how much percentage there is for all of the characteristics. Although Rings are quite helpful, as they demonstrate and have some correlation to the gender between M/F and I. Using pycaret and functions provided for Machine Learning here's the training and testing model.

III. Train / Test and Decide Best Model

After preprocessing and generating the outputs, now the final step is to generate classification output.

Topics learned in the class will demonstrate the use of numpy and scipy and pycaret algorithms and functions. Using these codes, the output is a confusion matrix, histogram, charts, beat mean and best categorizing calculations and the report of mathematics in the form of a .pkl model.

So starting with:

- Target: It is the class, which is the "Sex"
- Session ID is a randomly generated seed. The seed can range anywhere between 1 to 100, for randomness and chaos, I picked 45.
- Training = 80% and Testing = 20%
- Normalization is required
- Normaliization method : Min-max, however, Z-Score mean is also tested to compare differences and the results are shown.
- Polynomial degree above 3, so I wanted to try a 4th degree. Any number around 2, 3, or 4 works.

The Math used in AI Neural Networks mostly uses linear algebra and statistics, basing a lot of discrete math and computer science.

IV. Results and Analysis

The output code shows that there are three classes, (F/Female, M/Male, and I/Infant), as well as eight other features shown. Initial Data seen as histograms:

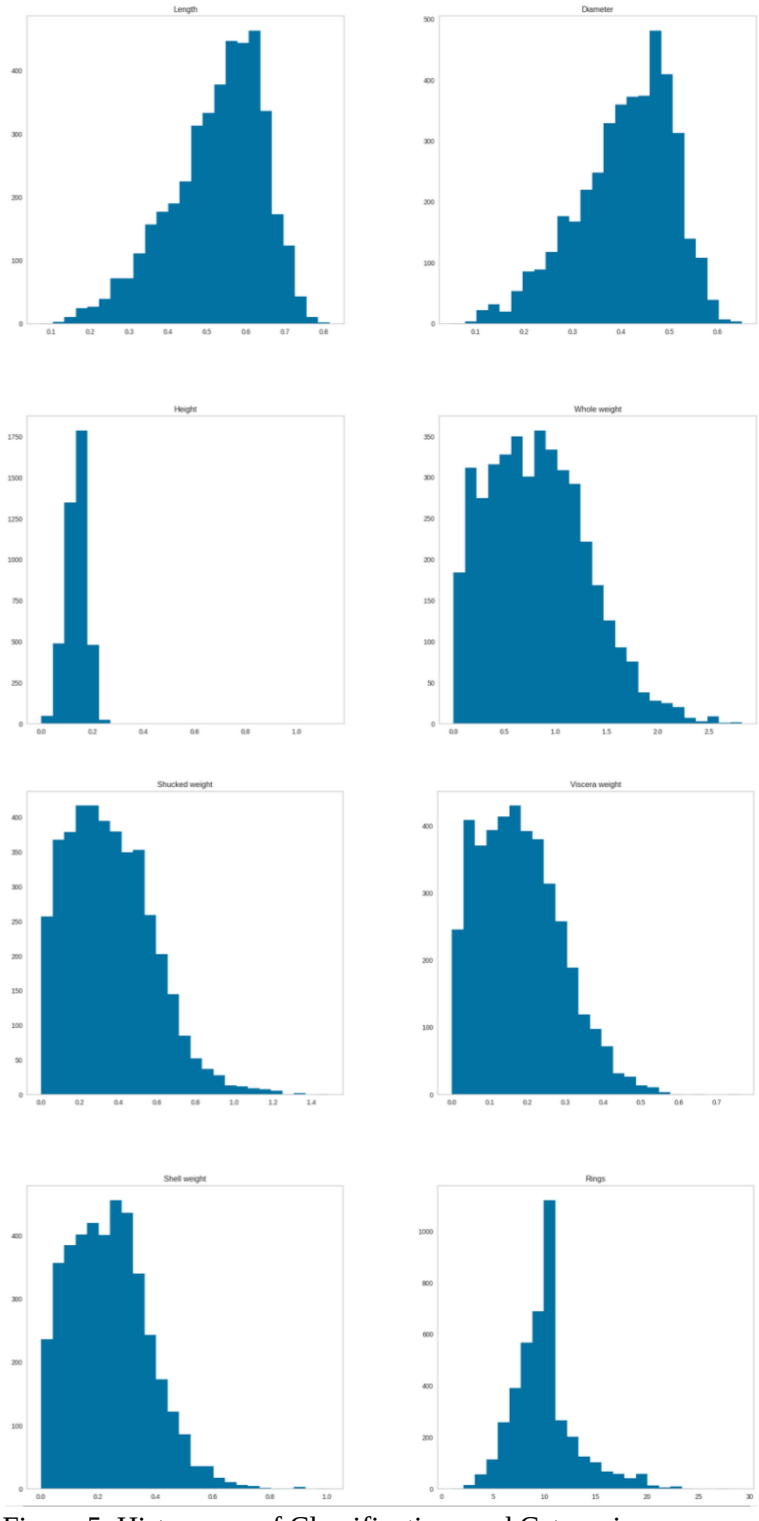

Figure 5. Histograms of Classifications and Categories

So we at least have gathered a lot of information.

A. Pyplot Charts and Results

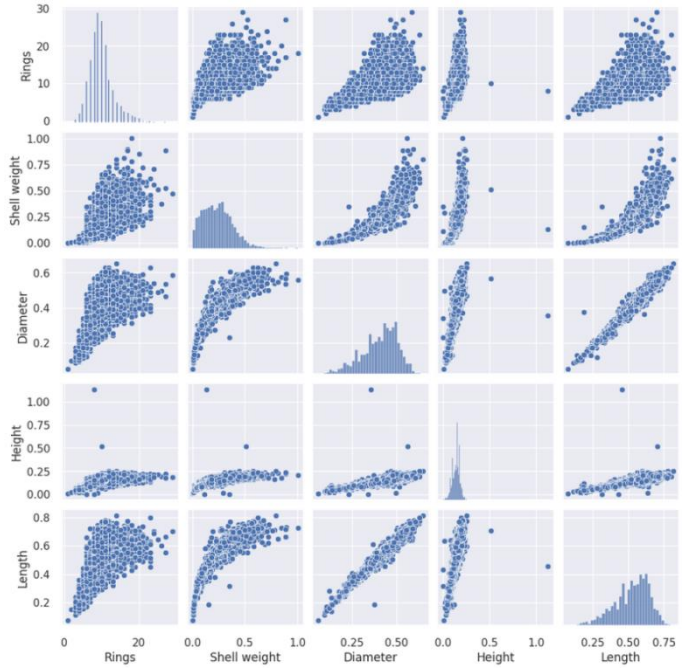

Figure 6. Seaborn PairPlot Of Features

From the information provided, we see that from Figure 5,6, and 7 shows Rings popular form of comparison and can be used to map out histograms and which data is the most interesting. We can also note that on Figure 7, Height, Diameter, and Length have the strongest correlation in differing the genders of abalones. There is a huge advantage in using these characteristics. However Rings are massively differing outlier. The Rings could be a useful experiment for analysis as differing rings can occur and continuous training will handle data better.

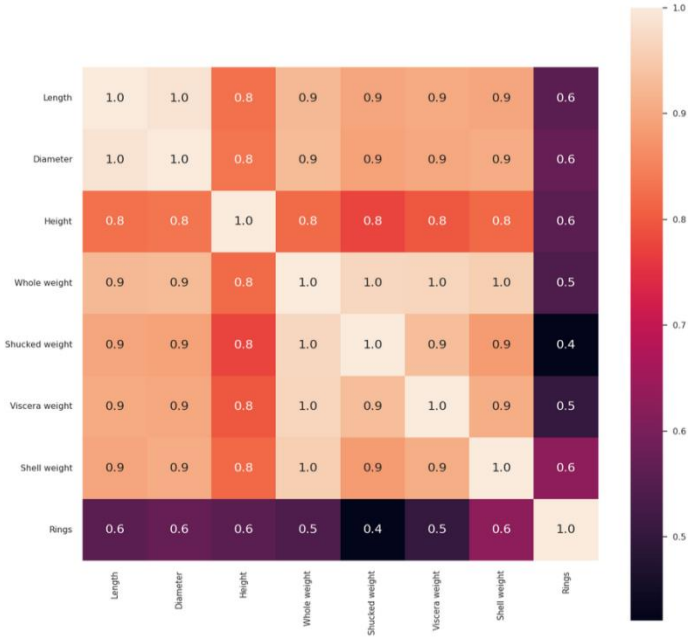

Figure 7. Correlation Matrix of all the Features

In the next figure, Figure 7, the facet grid shows we can discriminate the abalone gender into different colors and map out different characteristics, which can help the model be able to calculate probability that a certain abalone with these characteristics given can tell how much statistically likely the abalone is male vs female.

Oddly, some anamoly dots are noted, for example, Female (Orange) for Height vs Rings, there is a data point where Height is massive and huge whereas it is smaller than Male(Blue), almost equaling the height of the Infant(Green) charts.

Shell weights appear to be more scattered and have more variable compared to the height. Rings have a range of 1-30 but the remaining features have a range of 0 – 5. So Rings make a great optimization variable.

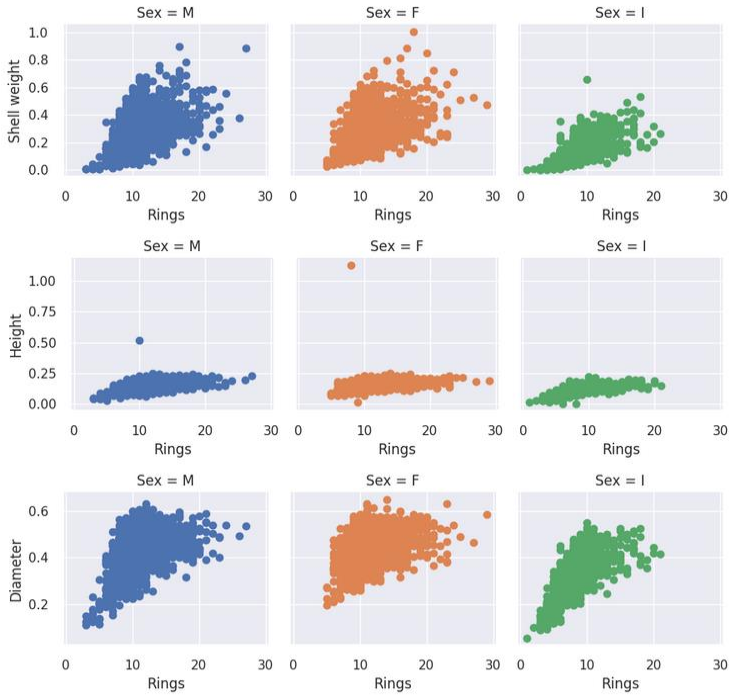

Figure 8. Facet Grid of Abalone Gender, comparing Shell Weight, height, and diameter.

Table 2. Abalone Gender Number Count

Female [0]	1307
Infant [1]	1342
Male [2]	1528

By a total of 4177 abalones accounted for, this shows that the most data received came from the males. The tests also show that for changing features, the probability of determining gender is not exactly predictable.

Diameter does show correlation and growth into larger diameter as abalones age so it makes sense if Infant levels are lower. It does note that Females have a lot of

larger diameter and weights as well compared to the males. There are more male abalones so the data will prefer to analyze more or learn more about the males as a result of the continuous training.

Using the Train and Test Planning, which includes the Feature Engineering provided, we can use a larger polynomial and
There is a 80% Train and 20% test idea so to maintain this similar testing, I was aiming more 70% Training, 15% Validation and 15% Testing.
I also worked on 80% Training, 10% Validation and 10% Testing, and some results did differ a little.

B. Best Model and Ensemble Models

Ultimately we have some data and comparison/conditions that differ for Male, Female, and Infant. The model demonstrates

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.5546	0.7497	0.5546	0.5487	0.5475	0.3337	0.3363	109.3530
ridge	Ridge Classifier	0.5539	0.0000	0.5539	0.5479	0.5352	0.3365	0.3458	0.2150
lr	Logistic Regression	0.5513	0.7564	0.5513	0.5439	0.5316	0.3331	0.3426	1.6800
rf	Random Forest Classifier	0.5475	0.7343	0.5475	0.5449	0.5444	0.3218	0.3230	4.6070
et	Extra Trees Classifier	0.5408	0.7337	0.5408	0.5378	0.5380	0.3105	0.3113	1.1710
lightgbm	Light Gradient Boosting Machine	0.5386	0.7308	0.5386	0.5359	0.5361	0.3072	0.3078	56.6120
lda	Linear Discriminant Analysis	0.5370	0.7144	0.5370	0.5306	0.5323	0.3041	0.3050	0.8830
xgboost	Extreme Gradient Boosting	0.5367	0.7308	0.5367	0.5345	0.5350	0.3039	0.3043	18.2970
svm	SVM - Linear Kernel	0.5363	0.0000	0.5363	0.5590	0.4678	0.3130	0.3537	0.5350
ada	Ada Boost Classifier	0.5236	0.7158	0.5236	0.5169	0.5122	0.2894	0.2942	7.4820
knn	K Neighbors Classifier	0.5221	0.7050	0.5221	0.5200	0.5130	0.2877	0.2926	0.4420

Figure 9. Best Model Analysis

From SVMs to KNNs, the classification using these techniques has resulted in that min-max is very useful. The paper that mentions Data Scaling Performance states that the min-max has better results in training the model rather than the zero-score mean comparison.

The idea is to use ensemble to combine and put together data like a multiplexer on a digital circuit. In the Best model, Area Under Curve (AUC) works best under logistic regression at a level of 77%. Gradient boosting works well for Accuracy and Recall and F1. The Ensemble will repeat the fold, exactly 10-fold. So the results are demonstrate iteratively from 0 to 9. Linear Kernel SVMs did well for precision but it is still less than 70% for analysis. KNN didn't work too well.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.5896	0.0000	0.5896	0.5864	0.5871	0.3844	0.3850
1	0.5373	0.0000	0.5373	0.5255	0.5275	0.3062	0.3084
2	0.5468	0.0000	0.5468	0.5461	0.5455	0.3204	0.3210
3	0.5955	0.0000	0.5955	0.5909	0.5880	0.3960	0.3998
4	0.4944	0.0000	0.4944	0.4916	0.4926	0.2375	0.2377
5	0.5655	0.0000	0.5655	0.5630	0.5626	0.3492	0.3503
6	0.5543	0.0000	0.5543	0.5418	0.5424	0.3336	0.3372
7	0.5768	0.0000	0.5768	0.5827	0.5771	0.3663	0.3683
8	0.5543	0.0000	0.5543	0.5514	0.5524	0.3312	0.3316
9	0.5768	0.0000	0.5768	0.5694	0.5719	0.3642	0.3649
Mean	0.5591	0.0000	0.5591	0.5549	0.5547	0.3389	0.3404
Std	0.0279	0.0000	0.0279	0.0291	0.0280	0.0431	0.0436

```
Fitting 10 folds for each of 10 candidates, totalling 100 fits
```

Figure 10. Best and Tuned Model by 10-Fold

To tune the model, the model is put together using a blend-model function with a parameter of 10-fold. This will generate 10 iterative outputs and tests of the model after being put together multiple times. The Mean and standard deviation are somewhat low so there are some problems faced as the algorithm is put together.

The Process starts off with Raw data, with whatever raw data we choose, we need to give it label and input the labeled organized data into rows and columns and send them to the train and testing:

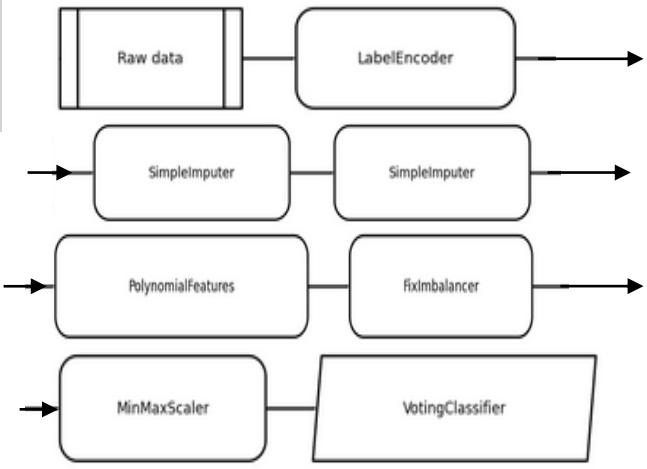

Figure 11. Pipeline demonstrating Plot type and flow process

Pycaret and the Machine Learning algorithms use the best model and ensemble models to be used for neural network learning.

After inputting the results, the confusion matrix has a result based on the inputs, however the model doesn't

work too well as shown in Figure 12.

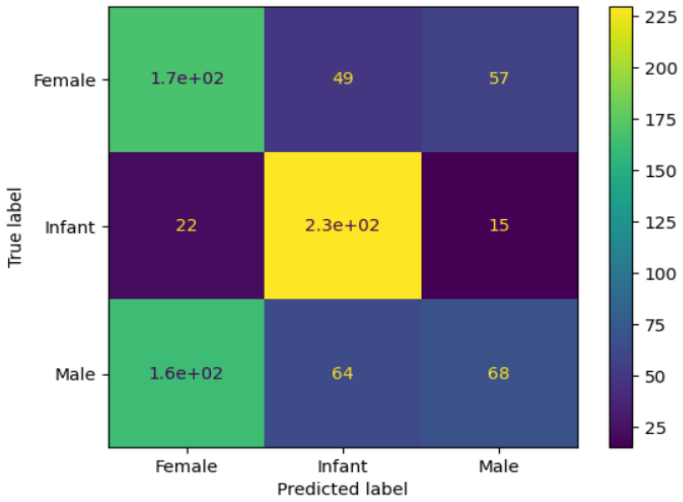

Figure 12. Confusion Matrix of the Abalone Gender

Once the results have been 10-fold and repeated until we reach Figure 10,

Ultimately this will now generate the final .pkl model. This uses Ensemble and Best Modeling and after all the tuning, the output is produced. Some changes can be done but this is how the output is generated after all the module programming.

```
[836 rows x 10 columns]
Accuracy: 0.5574162679425837
Precision: 0.5415381357049063
Recall: 0.5574162679425837
F1 Score: 0.5268022526766994

Confusion Matrix:
[[168  49  57]
 [ 22 230  15]
 [163  64  68]]

Classification Report:
              precision    recall  f1-score   support

           F       0.48      0.61      0.54       274
           I       0.67      0.86      0.75       267
           M       0.49      0.23      0.31       295

    accuracy                           0.56       836
   macro avg       0.54      0.57      0.53       836
weighted avg       0.54      0.56      0.53       836
```

Figure 13. Abalone.pkl Final Results

We have the Final output but the averages and recall all rate at levels below 60%. Even after using pycaret, some uncertainty exists so it is still difficult to determine a higher precision on the model. Min-max appears to have a 53% on Naïve Bayes and 57% on decision making. Given how small the dataset is, it could be difficult if the data is insufficient or limited.

It shows that in the Accuracy rating:

Female = 65.19%
Infant = 82.06%
Male = 64.23%

The next tasks would train the model to be more precise however this is a good start to reach a conclusion after multiple tests.

The .pkl file is made and all of the resources, datasets, pynb files, all are in the GitHub. Uploading documents and software is now complete and there are certain conditions that could improve from this first draft. Confusion Matrix has demonstrated some problems and requires changes to minimize the False positives and negatives, however this is a good start to work on.

V. Literary Research and Anaysis

A lot of the results are similar to that of the sources [6], [8], [9], and [10]. Source [7] is another dataset just to see what happened to see any differences in hopes to increase higher mean and better model to raise a percentage better than 58%.

Distribution of Min-Max Normalization does in fact aid in the increase of accuracy. A lot of the papers use forms of classification models and Z-score in order to model the algorithm. The papers read over also use logarithmic and polynomial regression in order to reach a far better result. And it shows. Regression based analysis reaches an F1 Score = 0.90 and ROC AUC that of 0.86. Based on the results, biological features are now great indicators and allow a probability of deciding which abalone is a male or female or even an infant. Minimizing False Negatives and False Positives is the task that could've made the model better. Other Papers use class handling methods such as SMOTE and MAE analysis to map out a scatter plot.

VI. Conclusion

The idea of programming machine learning algorithms is to design learning algorithms that can gather data and learn on it's own and develop without the need for human intervention. Classification, Regression, and Reinforcement Learning are topics used for the AI to develop and gather knowledge, and requires continuous data in order to learn itself. Abalone dataset is one of the

practices where training AI is done and feeding the AI data is how it develops.

It's not the best design and the best mean output is still low and approximately around 60%, but it is at least a start and can grow in the future.

To predict information about the gender of an abalone may be bizarre in the form of designing learning neural networks however the physical characteristics trains the AI to understand differing images and understand how slight differences are everywhere. Robust Regression or polynomial regression traces out how far

There are techniques such as SMOTE and Cross Validation which can also aid in development. Naïve Bayes also performs more consistent calculations than SVMs. This shows that normalization boosts improvement in the abalone dataset.

However, it is gained that the Characteristics and features can determine the probability of gender of certain abalones. Using characteristics and seeing patterns is done in both human logic as well as a machine that gathers the knowledge and tests itself. Future work will be considered now that there are other AI tasks that can be focused from this training.

REFERENCES

[1] https://archive.ics.uci.edu/dataset/1/abalone

[2] *Abalone Age Prediction Problem: A Review*. (2019, September 19). https://www.ijcaonline.org/archives/volume178/number50/mehta-2019-ijca-919425.pdf..

[3] https://www.kaggle.com/code/princeashburton/abalone-analysis-supervised-learning

[4] *Intro to Machine Learning via the Abalone Age Prediction Problem*. (2020, July 23).

[5] *CLASSIFICATION METHOD FOR ESTIMATING THE NUMBERS OF RINGS OF ABALONE*. (2021, March 8). Retrieved November 5, 2023, from https://medium.com/analytics-vidhya/classification-method-for-estimating-the-numbers-of-rings-of-abalone-bb13264dd186

[6] *Exploratory data analysis of Abalone characteristics, and the development and comparison of various machine learning models to predict the age of the abalone*. (2020, May 16).

[7] https://datahub.io/machine-learning/abalone

[8] Intro to Machine Learning via the Abalone Age Prediction Problem. (2020, July 23). (2021). https://bisa.ai/portofolio/detail/ODUx

[9] W. A. Arifin, I. Ariawan, A. A. Rosalia, L. Lukman, and N. Tufailah, "Data scaling performance on various machine learning algorithms to identify abalone sex," Jurnal Teknologi dan Sistem Komputer, vol. 10, no. 1, pp. 26-31, 2022. doi: 10.14710/jtsiskom.2022.14105, v

[10] *Abalone Age Classification Project Report — Abalone Age Classifier*. (n.d.).https://ubcmds.github.io/abalone_age_classification/Project_report_milestone2.html