Using Regression and Neural Network Models for Predicting Abalone Age

Authors: Hannah Wah Day (z5480942), Nadav Dar* (z5259682), Xincheng Qiao (z5327961)

*Nadav Dar submitted code on Edstem

1. Abstract

This report details the development of models for predicting the age of abalone using the UCI Abalone dataset. The study applied linear regression, logistic regression and neural networks trained with Stochastic Gradient Descent (SDG) for both regression and classification tasks. Initial data processing revealed that shell weight and diameter are most highly correlated with abalone ring age, with coefficients of 0.63 and 0.57 respectively. Due to high variance in the dataset, performance of the linear regression model was limited, achieving R^2 score of 0.52 and root mean squared error (RMSE) of 2.17 on the test set. However, using logistic regression to classify abalone as older/younger than 7 years achieved an accuracy score of 94.0% and area under the Receiver Operating Characteristic (ROC) Curve of 0.94. The neural network approach was similarly effective, with 1 hidden layer containing 64 neurons and a learning rate of 0.01 facilitating a classification accuracy score of 0.933.

2. Introduction

Using neural network models for classification and prediction tasks has become increasingly popular in recent times. They have already been proven capable in the fields of agriculture, medical science, finance, education, security, engineering, trading commodity and art. [1] Neural classifiers have emerged as an alternative to conventional classification methods because they are self-adaptive and nonlinear models, meaning they are more flexible in modelling the complexity of real-world relationships. [2]

The multilayer perceptron (MLP) is a feedforward neural network model composed of multiple layers of which one or more are hidden. Each layer feeds into the next via a set of weights and a non-linear differential activation function. The model requires tuning of various hyper-parameters such as the number of hidden layers, number of neurons and learning rate to achieve optimal performance which are typically obtained through trial and error. [3]

A study published in 2022 in the interest of modelling data that was not built for binary classification purposes, similar to the abalone dataset, found that linear (R^2 score up to 0.54, RMSE down to 0.25) and logistic regression (accuracy up to 87.6%) models had consistent predictions for 89.9% of the dataset, suggesting relatively strong performance in both. [4] However, in an earlier comparative study regression performance was weaker than that of multi-layer perceptron, with areas under the ROC curves of 0.9508 and 0.9528 respectively, further demonstrating the potential of neural networks for classification problems. [5] Despite these contributions, there remains a significant knowledge gap in more recent literature regarding comparison of neural networks against regression models for classification problems. Hence, this study evaluates and compares the performance of regression and neural network models for classification.

In this study, we investigate the use of linear/logistic regression and neural network models for predicting abalone age given 8 features in the UCI Abalone dataset. We determine the most correlated features using a linear regression model, and use this to predict ring age. We then use a logistic regression model with a sigmoid activation function to classify abalone as older/younger than 7. Finally, we take a neural network approach for both the regression and classification problem, and discuss the comparative effectiveness against the conventional regression models.

3. Methodology

3.1 Data Overview

The Abalone dataset from (https://archive.ics.uci.edu/dataset/1/abalone) is a dataset containing information of the physical characteristics of the abalone such as Sex, Length, Diameter, Weight and more. Additionally, the dataset contains the target variable Rins, which when adding 1.5 to the number of rings determines the age. However, the process of counting rings is ‘boring and time consuming’. Thus, the motivation behind the modelling is to predict the rings present in an abalone based on its physical and measurable properties.

3.2 Data Processing and Cleaning

Part of the requirement to build high quality and accurate machine learning models is the requirement of clean data. In our abalone dataset we have the categorical variable of the abalone’s Sex, which is Male, Female or Infant. To use this variable we must convert it to a numerical feature that a model can interpret. Hence, by converting the categorical variable into a discrete numerical one it can be used for

modelling and analysis. In our dataset we converted the categorical variable into numerical based on the following dictionary: Male = 0, Female = 1, Infant = 2.

Each of the model's regression and classification requires a different representation of the target variable. In the case of regression, the study wants to predict the number of rings on the abalone, while the classification task wants to predict the binary outcome of whether the abalone has above or below 7 rings. The current dataset suits the regression task as the target variable rings contains the number of rings. However, the data is not suited to the classification problem as it is not represented by binary outcomes, meaning a transformation is required. Creating the following mapping function

If rings <= 7 then 0 else 1.

Allows for the classification modelling to occur as we are making the prediction of whether the rings are above or below 7.

When building the models we want to prevent underfitting and overfitting. One of the mechanisms to prevent overfitting is through splitting the data into a training set and testing set. By splitting the data, this study fits the model to training data and afterwards evaluates the performance of the model against testing data. The aim of this method of splitting is the model learns the generalisation of the training data so it can be applied to the unknown testing data.

3.3 Overview of the methods

In this study we will be utilising linear regression for the regression task and both logistic regression and neural networks for the classification. Linear regression is a modelling method used to fit a linear line of best fit to a given dataset X to predict a target Y, through the minimisation of error such as mean squared error. In this study after fitting a linear model to the Abalone dataset it can then be used to predict the number of rings which are present on the abalone. Additionally, regularisation such as Lasso or Ridge Regression can be used to manage overfitting however it was not considered relevant in this study.

The second modelling that this study follows is the classification task of predicting whether an abalone has more or less than 7 rings based on their measurable features. The binary classification task will be approached through the two methods: logistic regression and neural networks. Logistic regression is a form of a linear model, however used for binary classification tasks, which is training through

minimisation of a loss function. Due to its linear nature, it is a less complex model and will be quicker to train. Neural networks can be used for classification tasks however are significantly more complex models. Using multiple layers of interconnected neurons the network aims to train and adjust the weights of the neurons by learning from large amounts of data. Both classification models in this study are used through the feeding of training data and then evaluated against their testing dataset.

3.4 Software Suite

The programming language chosen in this study is Python due to its stable and well-maintained libraries for both data processing and machine learning related tasks.

- Numpy – A library used for numerical computing in python allowing for both data manipulation and calculations of metrics such as mean and standard deviation.
- Pandas – Allows for the use of DataFrame, an object which allows for easy manipulation, cleaning and wrangling of data.
- Matplotlib & Seaborn – Plotting libraries.
- Scikit-learn – Library with a wide range of functions, which is used to build linear regression, logistic regression and neural networks. Additionally, contains wide range of functions used to score and evaluate models.

3.5 Experiment Setting

When testing to find the optimisation of the hyper parameters in the neural network the following parameters were changed; layer size, hidden layers and learning rate. Using a random state, the study controls the splitting of the data into the training and testing sets allowing for fair experimentation on the hyper parameters. Looping through the dictionary of parameters and evaluating the results allows us to determine the optimal hyperparameters. This study tested the following hyperparameters

Hidden Layers : [1, 2, 3] – How many hidden layers the neural network will have
Hidden Layer Size : [16, 32, 64] – How many neurons each hidden layer will have.
Learning Rate : [0.001, 0.01, 0.1] – The learning rate used in the gradient descent.

4. Results

4.1 Heatmap

The analysis of correlation matrices and heatmaps is an important step in revealing the relationship between features and target variables (rings).

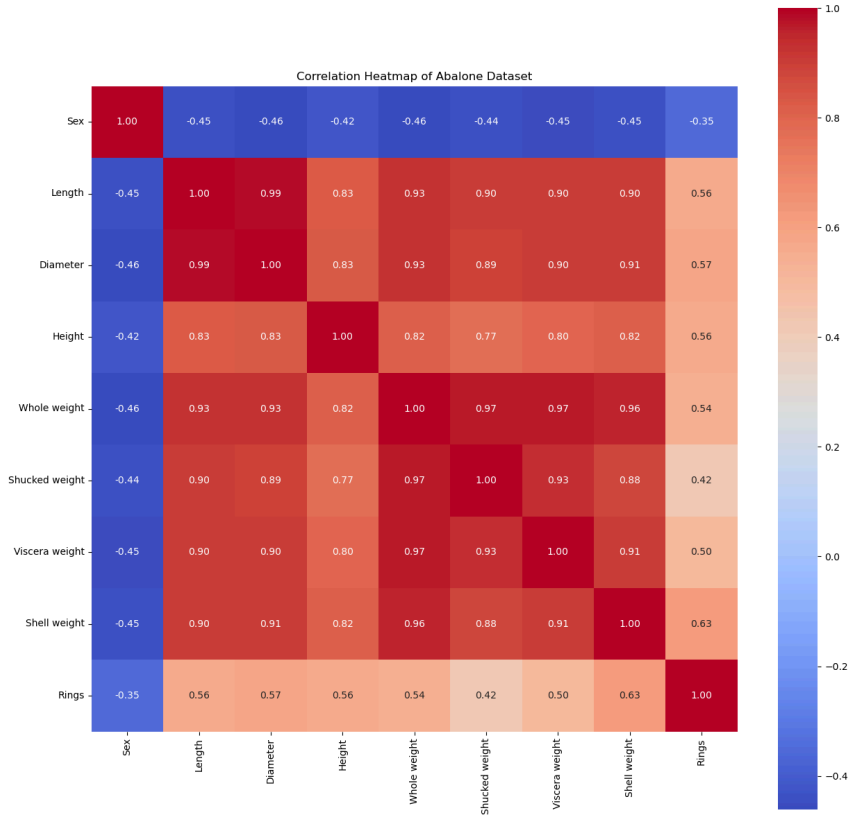

Figure 1.1 Correlation Heatmap of Abalone Dataset

The highest correlation between shell weight and Rings can be visualised in the heatmap Figure 1.1. The correlation coefficient is about 0.63, which suggests that shell weight is a very important feature in predicting the age of abalone (Rings), and therefore it can be prioritised in the subsequent development of the model.

The correlation between Diameter and Length with Rings was also relatively high (correlation coefficients of 0.57 and 0.56, respectively). This indicates that the

volumetric features (Length and Diameter) of abalone can effectively reflect its age information.

Notably, the lowest correlation between sex and the target variable (Rings) was -0.35, which suggests that sex information may not have a significant effect when predicting abalone age.

4.2 Scatter plot with ring-age

Based on the relationship between the features observed in the correlation heatmap and the target variable, shell weight and diameter can be selected for analysis. The reason for this is because they are the two most correlated with rings (correlation coefficients of 0.63 and 0.57 respectively)

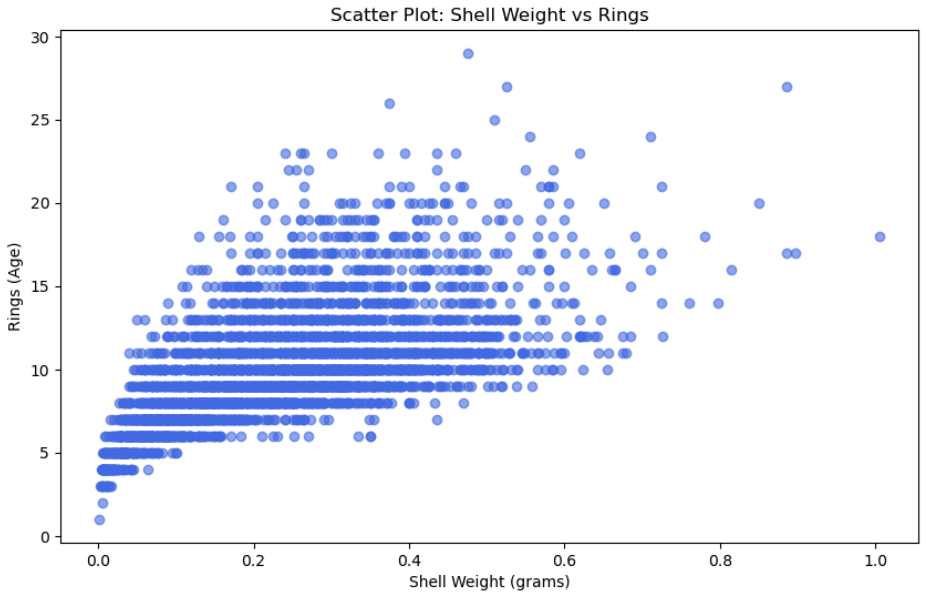

Figure 2.1 Scatter plot: Shell weight vs Rings

From the graph Figure 2.1, it can be observed that shell weight (x-axis) shows a positive correlation with the number of rings (y-axis). As the shell weight increases, the age also increases gradually. However, the relationship between shell weight and age is not completely linear, and age shows some fluctuations at higher weights.

In addition, Figure 2.2 can be observed that the diameter (x-axis) shows a clear positive correlation with the rings-age (y-axis). As the diameter increases, the mean value of the rings also increases gradually. This is consistent with the correlation observed in the previous heat map, indicating that the diameter can reflect the age of abalone to some extent. Overall, the rate of increase in ring-age begins to slow down at higher diameters (rings fluctuating between 15 and 20), again suggesting that there may be some non-linear relationship between rings and diameter.

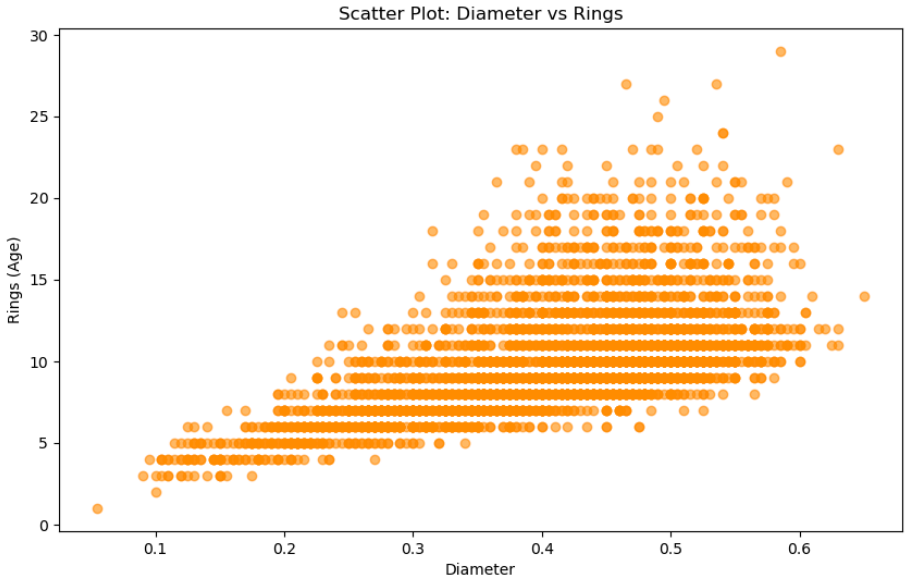

Figure 2.2 Scatter plot: Diameter vs Rings

4.3 Histograms of the two most correlated features, and the ring-age

Based on the previous discussion, we create histograms for these three variables (Shell weight, Diameter and Rings) and analyse them in further detail.

4.3.1 Histogram of shell weight distribution

From Figure 3.1, shell weight shows a clear right-skewed distribution, with the majority of samples clustered between 0.2 and 0.4 grams, with a peak number of samples (350+ samples), implying that the majority of abalone fall into the medium-sized category. In addition, a significant Decreasing Frequency was observed, with a sharp decrease in the number of samples with shell weights greater than 0.4 grams, suggesting that heavier abalone individuals are relatively rare. It is also worth noting the Long Tail Effect. In the region of weights greater than 0.6 grams, some extreme values (shell weights close to 1.0 grams) can be seen, which may represent older or abnormal abalone individuals.

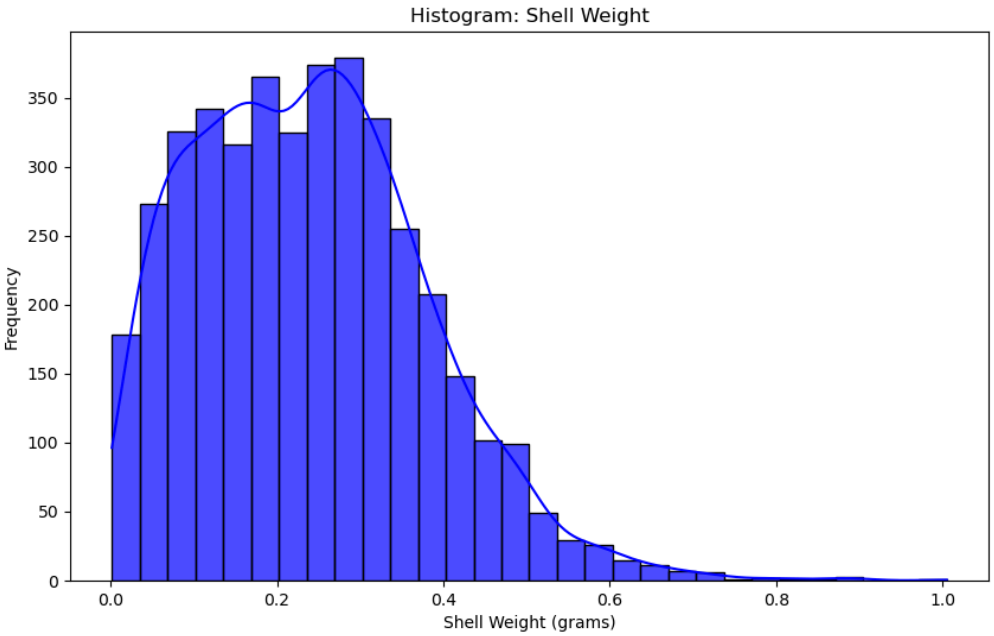

Figure 3.1 Histogram of shell weight

4.3.2 Histogram of Diameter distribution

The shape of the diameter distribution was close to normal, with the majority of abalone samples clustered between 0.3 and 0.5 mm in diameter and peaking around 0.4 mm (350+ samples), which suggests that the majority of abalone were of medium size. In addition, there was a significant decrease in the number of samples in areas with diameters less than 0.2 mm and greater than 0.5 mm, which could be juvenile or exceptionally large adult abalone.

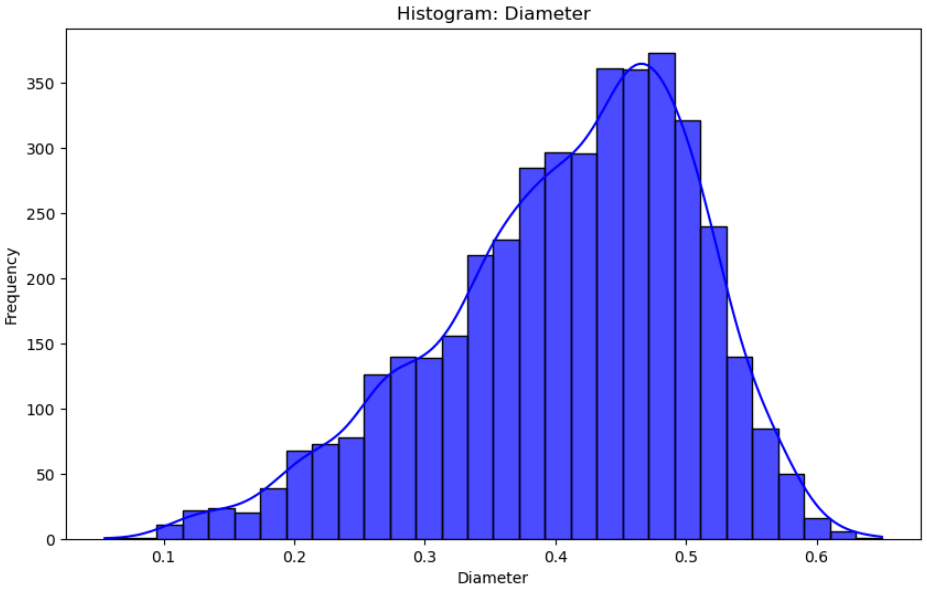

Figure 3.2 Histogram of shell weight

4.3.3 Histogram of Rings (Age) distribution

By looking at the shape of the distribution and the overall trend in Figure 3.3, the distribution of the number of rings shows a multi-peak pattern, with the most pronounced peak occurring at ring 10, suggesting that the age distribution in the dataset is not homogeneous, but rather is concentrated in specific age ranges (particularly between 7 and 12 years) . In general, the overall distribution of the samples is slightly to the right, with the number of samples decreasing as the number of rings increases. In the region of rings 20 to 30, there are some extreme values of samples, which may represent a small number of very old abalone individuals. These are consistent with previous observations.

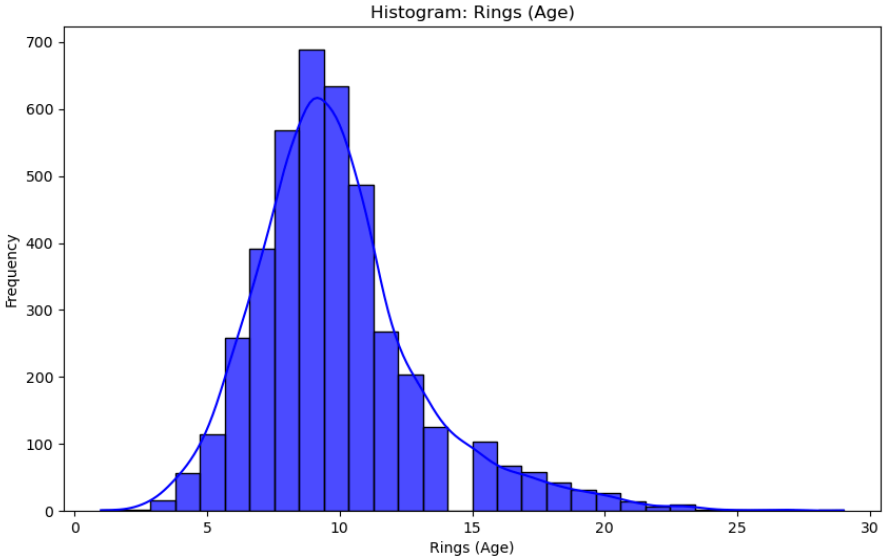

Figure 3.3 Histogram of Rings (age)

4.4 Linear Regression Model (All Features)

The linear regression model using all features for ring-age achieved an R^2 score of 0.5182 and RMSE of 2.1834 on the training set, and an R^2 score of 0.5182 and RMSE of 2.1834 on the test set. There was no evidence of overtraining since performance on train and test sets was consistent. The model is visualised in Figure 4.1, which shows the desirable linear trend, but with a relatively high variance.

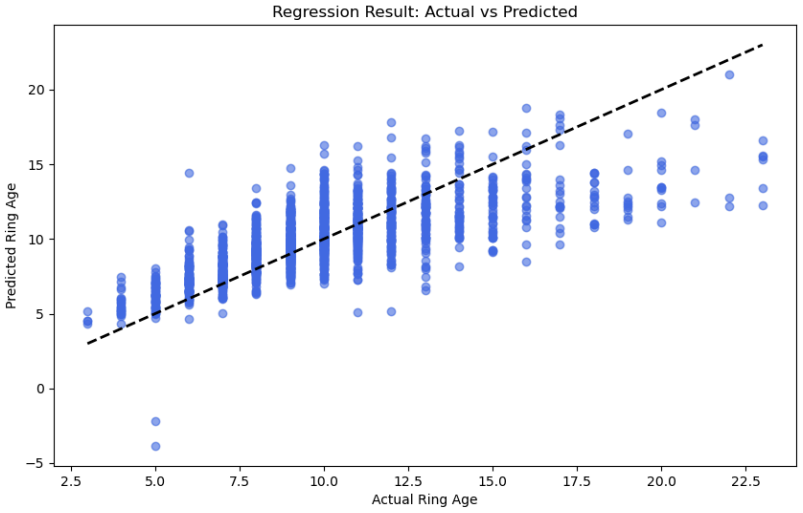

Figure 4.1 : Performance of linear regression model using all features for ring-age

4.5 Logistic Regression Model (All Features)

Using a logistic regression model for binary classification of abalone ring-age as older/younger than 7 years was effective. It is necessary to first note the uneven class distribution of the dataset which is shown by the confusion matrix in Figure 4.2. This Figure exemplifies the effect on the model of the imbalanced dataset. The true positive rate was 97%, while the false positive rate was 44%, which shows good recognition of positive outcomes (ring-age greater than 7) but poor recognition of negative outcomes (ring-age less than 7). This was expected as the dataset contains far fewer negative than positive cases.

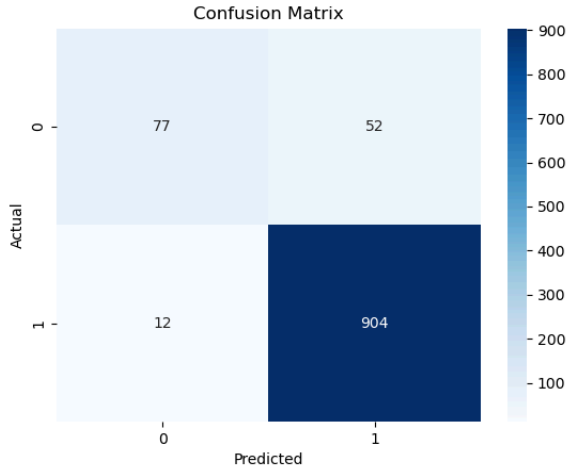

Figure 4.2: Confusion matrix from logistic regression model using all features for binary classification

The model achieved an accuracy of 0.93, but this was likely limited by the class distribution; accuracy favours classifiers that always predict a negative outcome for rare events. As such, we also considered the area under the ROC curve (AUC) shown in Figure 4.3. The reported AUC was 0.95, which shows performance of the logistic regression model is far superior to that of the no skill model.

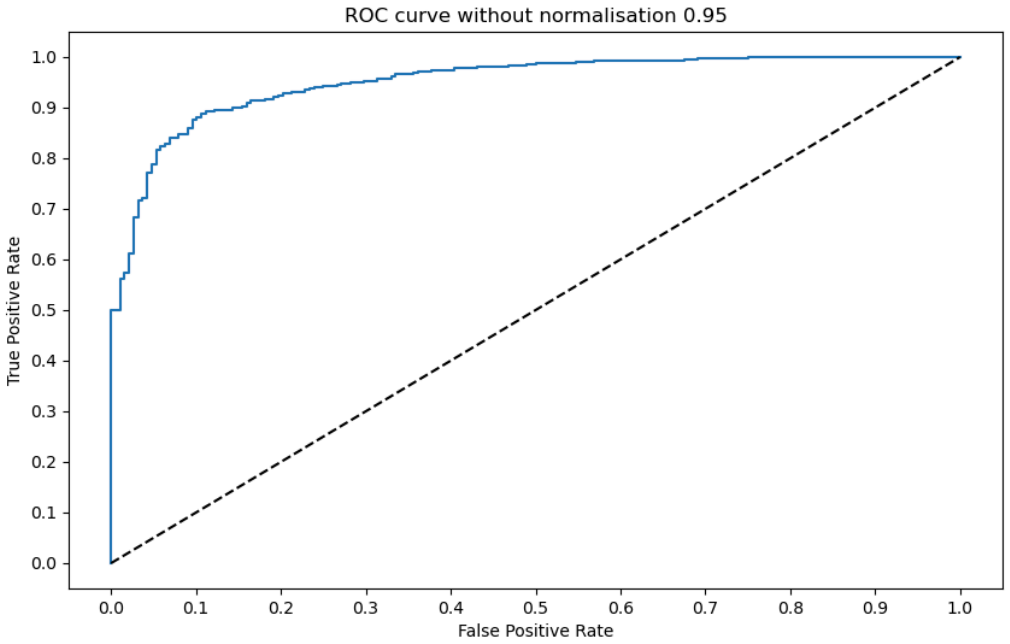

Figure 4.3: ROC Curve from logistic regression model using all features for binary classification

4.6 Neuron Network Model

As referenced in Section 3.5 the study will try find the optimal hyperparameters for the neural network based on the following options

Hidden Layers : [1, 2, 3] – How many hidden layers the neural network will have
Hidden Layer Size : [16, 32, 64] – How many neurons each hidden layer will have.
Learning Rate : [0.001, 0.01, 0.1] – The learning rate used in the gradient descent.

All the models with different hyper parameters were trained on the same training data and then evaluated against the same testing data. Across all the models the performance of accuracy score was high ranging from 0.923-0.935. Additionally, there was a high AUC score ranging from 0.950-0.954. The best performing model had both the highest AUC score of 0.954 and accuracy score of 0.935 with the following parameters

Hidden Layers: 1
Hidden Layers Neurons: 64
Learning Rate: 0.01

With these parameters 30 experiments were run to evaluate the performance of the neural network. In each of the experiments the exact same training and testing datasets were used.

	Accuracy	F1	AUC	Precision	Recall
Mean	0.933	0.963	0.953	0.947	0.98
Std	0.003	0.002	0.0004	0.008	0.01

After the 30 experiments the results recorded are still strong, by exhibiting high scores of metrics with the low amounts of variance.

5. Discussion

5.1 Performance of linear regression models

Both the normalised and normalised linear regression models yielded an R-squared of 0.5182 and RMSE of 2.1834. These results indicate that the model explained about 52% of the variance in the dataset, which suggests that linear regression was not able to capture the relationship between the features and the target variable (Rings).

5.2 Performance of logistic regression models

In the classification task, logistic regression performed well with an AUC score of 0.9508 for the unnormalised model and 0.9528 for the normalised model. The normalised model performed slightly better, indicating that feature scaling improved the classification performance. Moreover, the confusion matrix revealed a high true positive rate, which showed the model's effectiveness in classifying abalone age groups.

5.3 Performance of Neural Network models

The neural network with the hyper parameters of;

Hidden Layers: 1
Hidden Layers Neurons: 64
Learning Rate: 0.01

The neural network had strong performance after running the 30 experiments, with an average accuracy score of 0.933 and average AUC score of 0.953. Additionally,

along with these strong metric scores the 30 experiments also had very low amounts of variance. Hence, the neural network with the given hyper parameters is able to classify the amount of rings on the abalone to a high degree of accuracy.

6. Conclusion

In this study, we have successfully automated the age prediction of abalones, traditionally a manual and labour-intensive process, using advanced machine learning techniques including linear regression, logistic regression, and neural networks. These models have significantly enhanced the accuracy of predictions, showcasing their potential utility in the field of marine biology, particularly within aquaculture practices. Our comparative analysis reveals that while linear regression provides a solid baseline, logistic regression and neural networks offer substantial improvements in accuracy, handling the complexities inherent in biological datasets with greater efficacy.

A major breakthrough of our research has been the identification of shell weight and diameter as key indicators of abalone age. This insight is invaluable, not only for advancing biological research but also for informing effective conservation strategies and breeding programs. By revealing these critical relationships, our study highlights the sophisticated capabilities of statistical analysis in biological applications.

Looking to the future, the integration of environmental and genetic data could further refine our predictive models, enhancing their accuracy and applicability across different species and environmental conditions. Moreover, exploring cutting-edge machine learning approaches such as deep learning and ensemble methods could uncover more complex patterns in the data, potentially leading to breakthroughs in how we understand and predict biological ageing.

Implementing these models in real-world settings is a crucial next step. This would not only test their practical effectiveness but also allow for iterative improvements, ensuring that the models are robust and reliable. Additionally, it’s essential to consider the ethical and ecological impacts of deploying such technologies in natural settings, ensuring that our technological advances do not outpace our ecological responsibilities.

In General, this project stands at the intersection of technology and ecology, offering promising new tools for marine biology and beyond. As we continue to refine these models and expand their applications, the potential benefits for biological research and environmental conservation are profound and far-reaching.

7. References

[1] Abiodun, O.I. *et al.* (2018) ‘State-of-the-art in artificial neural network applications: A survey’, *Heliyon*, 4(11), p. e00938. Available at: https://doi.org/10.1016/j.heliyon.2018.e00938.

[2] *Classification with neural networks: a performance analysis* | *IEEE Conference Publication* | *IEEE Xplore* (no date) *ieeexplore.ieee.org*. Available at: https://ieeexplore.ieee.org/abstract/document/48672.

[3] Farooq Anjum, M., Tasadduq, I. and Al-Sultan, K. (1997) ‘Response surface methodology: A neural network approach’, *European Journal of Operational Research*, 101(1), pp. 65–73. Available at: https://doi.org/10.1016/S0377-2217(96)00232-9.

[4] Huang, K. and Zhang, H. (2022) ‘Classification and Regression Machine Learning Models for Predicting Aerobic Ready and Inherent Biodegradation of Organic Chemicals in Water’, 56(17), pp. 12755–12764. Available at: https://doi.org/10.1021/acs.est.2c01764.

[5] Kurt, I., Ture, M. and Kurum, A.T. (2008) ‘Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease’, *Expert Systems with Applications*, 34(1), pp. 366–374. Available at: https://doi.org/10.1016/j.eswa.2006.09.004.