Comparison of Regression and Classification Models for Predicting Abalone Age

Team member:

Hangwei Liang(z5499015) Jiehao Chen(z5507255) Junru Hua(z5514957)

Abstract

The primary objective of this project is to predict and classify the age of abalone based on physical measurements, a task traditionally performed by counting the number of shell rings under a microscope — a time-consuming and labor-intensive process. In this study, we applied various predictive models, including linear regression, logistic regression, and neural networks, to evaluate their effectiveness in regression and classification tasks. Our results indicate that, for the regression task, the linear regression model achieved an average RMSE of 2.31 and an R-squared score of 0.50 on the test set, while the neural network model demonstrated comparable performance but with higher model complexity. For the classification task, the logistic regression model obtained an average accuracy of 0.93 and an ROC AUC of 0.95, highlighting its robustness and simplicity over the neural network.

1 Introduction

Neural networks and deep learning models have significantly transformed machine learning, particularly in applications requiring complex pattern recognition, such as image classification, speech recognition, and natural language processing **[1][2]**. These models excel in tasks with large datasets, leveraging high-dimensional feature spaces to learn nuanced relationships within data **[3]**. Foundational studies by McCulloch and Pitts on artificial neurons, followed by Rumelhart et al.'s introduction of

backpropagation, laid the groundwork for modern neural networks **[4][5]**. Despite their complexity, neural networks often outperform simpler linear models in non-linear tasks due to their capacity to model complex feature interactions. Linear models, including regression, however, retain relevance for interpretability, computational efficiency, and strong performance on linearly structured data **[6]**. Seminal work by Gauss and Legendre established their utility in statistical and predictive modeling **[7]**.

This project aims to explore the performance of neural networks versus linear models in the context of regression and classification tasks. Although neural networks are widely acknowledged for their performance on complex datasets, fewer studies have systematically compared their efficacy against linear models, particularly on structured datasets with both linear and non-linear elements. Given the recent emphasis on explainable AI and the continued applicability of linear models in numerous domains, comparing these models can provide valuable insights into their respective strengths and limitations, especially in age prediction tasks like those required for abalone datasets.

In this project, we investigate the predictive performance of linear regression, logistic regression, and neural networks for age estimation and classification of abalone specimens based on physical measurements. Using the publicly available Abalone dataset, we aim to compare these models in both regression and classification tasks, evaluating their effectiveness with metrics such as RMSE, R-squared, accuracy, and ROC AUC. In addition to model comparisons, we explore the effects of regularization on linear models to improve their generalizability. For the neural networks, we employed grid search to tune hyperparameters, such as learning rate, hidden layer size, and dropout rate, to find the optimal configuration for accurate age predictions. This project seeks to highlight the conditions under which neural networks provide significant advantages, as well as situations where simpler linear models may perform comparably or better.

2 Methodology

In this project, our main purpose is to compare the affections to prediction performance on the same dataset using different models, this part consists of three parts:

- Data processing
- Model choosing
- Experiment setting

2.1 Data processing

For this dataset, it has 8 features applied to predict the ring age, the following image show histograms of all the features (target value included):

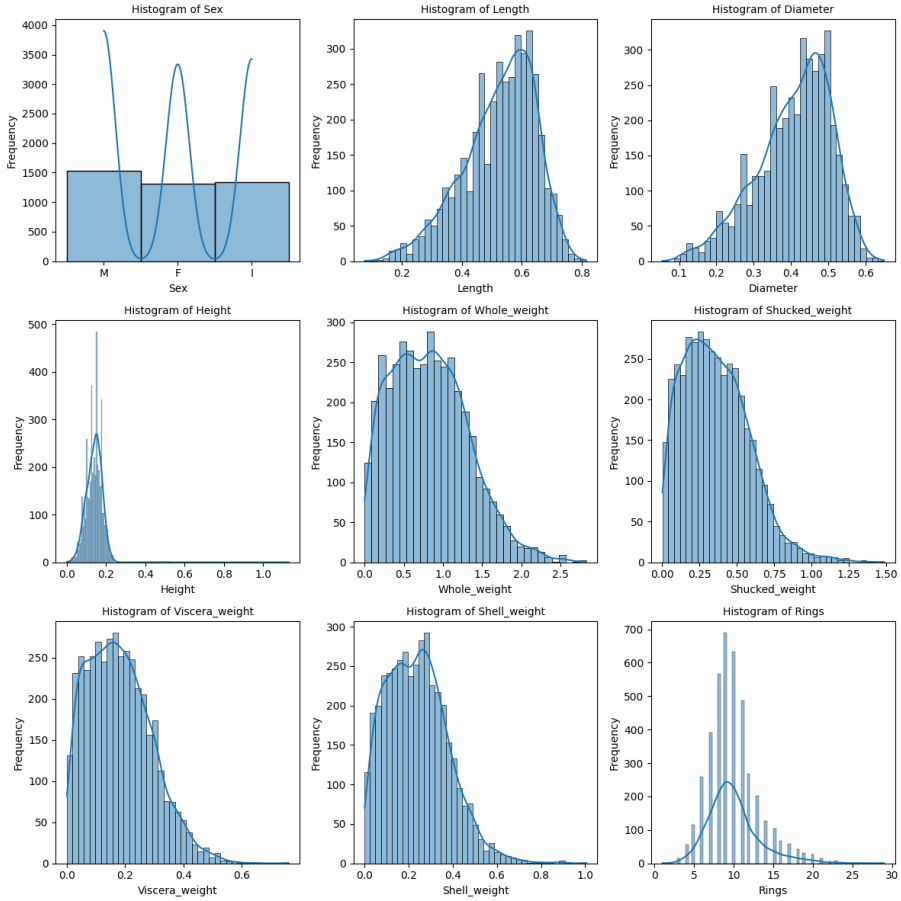


Figure1: Histograms of Abalone Data Feature Distributions

As we can see in Figure1 any of the distributions are unnormal, it can be usually seen in real-world data. For the convenience of subsequent predictions, we use the map method to map different genders to different numbers: male as 0, female as 1, and infant as 2. This step allows us to treat categorical data as numerical data for further processing.

In the next step, we use correlation matrix to explore the two most correlated features. The heatmap is shown in the following image. Based on the information from the heatmap, we can conclude that the two features most correlated with the target variable are Shell_weight (0.63) and Diameter (0.57). After selecting the two most relevant features, we plotted a scatter plot of these features to explore their distribution. From the scatterplot we can see that:

a. Diameter vs. Rings (Figer2(a)):

• The scatter plot shows a positive correlation between diameter and rings: as the diameter increases, the number of rings also tends to increase. This indicates that larger diameters are associated with a higher number of rings.

• The data is relatively dense in the middle range (with diameters between approximately 0.3 and 0.5), suggesting that most samples fall within this range of diameter and rings.

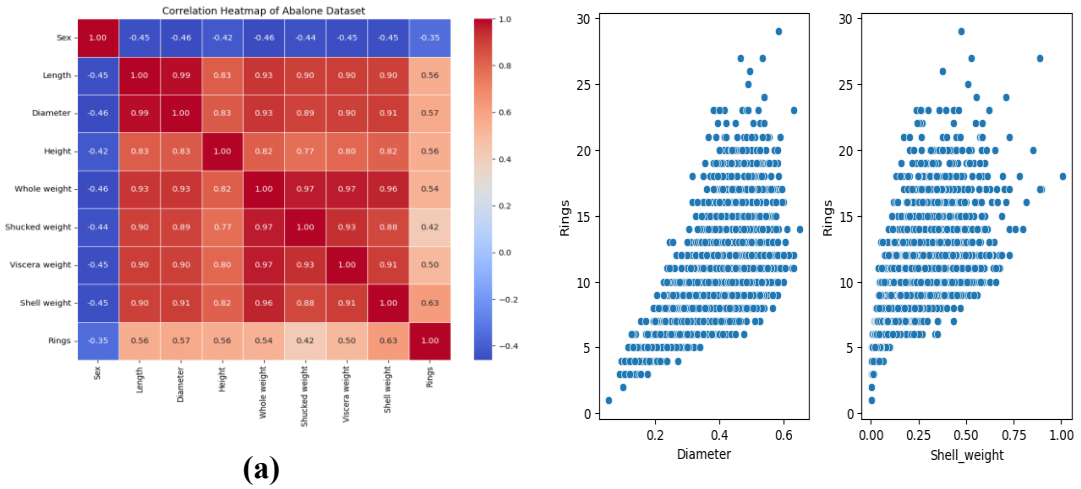
• As the diameter increases beyond approximately 0.5, the relationship between diameter and rings becomes slightly more scattered, with some points showing a higher number of rings.

b. Shell_weight vs. Rings (Figer2(b)):

• This plot also indicates a positive correlation: as shell weight increases, the number of rings generally increases as well. Heavier shells tend to have more rings.

• Compared to the diameter plot, the relationship between shell weight and rings is more scattered. Although a general positive trend is visible, the data points are more spread out, particularly at higher shell weights (around 0.75 to 1.0), where the number of rings is more variable.

• The middle range of shell weight (approximately 0.2 to 0.6) shows a relatively dense distribution of points, indicating that most data falls within this range.



(a) **(b)**

Figure2: (a) Correlation Heatmap of Abalone Dataset;(b) Scatter Plots of Rings vs Diameter and Shell Weight

2.2 Model choosing

2.2.1 Regression

a. Linear Regression model

Regression analysis is one of the most widely used techniques for analyzing multi-factor data.[9] Linear regression is employed to predict the age of abalones, which is a typical regression task since age is a continuous variable. By using 8 physical features of abalones (such as sex, length, weight, etc.) as input variables, the linear regression model learns the linear relationships between these features and the number of rings, allowing for an estimation of abalone age based on their physical measurements.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon$$

- y is the target variable, which in this study is the age of abalones (number of rings).
- $X_1 X_2 \cdots X_8$ are the independent abalone physical features
- $\beta_1 \beta_2 \cdots \beta_8$ are the regression coefficients, representing the influence of each feature on the target variable, ε is the error term.

b. Linear neural networks

Unlike linear regression, linear neural networks can enhance model expressiveness by

combining multiple linear layers. Even though each layer is linear, the combination of layers allows the model to capture more complex features.[10] The output of each layer is a linear combination of the input from the previous layer, and the model is represented as follows:

$$y = W_n W_{n-1} \cdots W_1 x$$

Here, x represents the abalone feature variables as input data, $W_n W_{n-1} \cdots W_1$ are the weight matrices for each layer, and y is the final output, which corresponds to the predicted abalone age. As we can see, a linear neural network is essentially a more complex linear transformation, but the final output remains a linear function of the input.

2.22 Classification

a. Logistic regression model

The logistic regression model maps the linear output of the model to a probability range between 0 and 1 using the Sigmoid function, thereby determining which class the data belongs to. The basic form of the model is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon$$

$$\delta(y) = \frac{1}{1 + e^{-y}}$$

Where $\beta_1 \beta_2 \cdots \beta_8$ are the regression coefficients and $X_1 X_2 \cdots X_8$ are the feature variables of the abalone. The model classifies based on a set threshold (usually 0.5): if the output is greater than 0.5, it is classified into one category; otherwise, it is classified into another category. The gradient descent method minimizes the error between the predicted value and the actual class, effectively improving classification performance and reducing the number of iterations. The iterative formula is:

$$\omega = \omega - \alpha \Delta Y$$

Where α is the learning rate and Y is the error function.

b. Logistic neural networks

The logistic neural network model is a multi-layer network based on neuron structures, where the input layer, hidden layer, and output layer perform a series of nonlinear transformations to map inputs to outputs. The output of the model can be expressed as:

$$h = \delta(W_1 x + b_1)$$

$$P(x = 1) = \delta(W_2 x + b_2)$$

- x represents the input data (features of the abalone), such as shell length, diameter, etc.
- $W_1 W_2$ are the weight matrices for the input-to-hidden and hidden-to-output layers, respectively. $b_1 b_2$ are the bias terms.
- h is the output of the hidden layer, generated after applying the nonlinear activation function δ .

By introducing hidden layers, the logistic neural network can capture complex nonlinear relationships within the data, improving classification flexibility and accuracy. [12] In predicting abalone age, the network processes the intricate features of the input through its multi-layer structure, allowing the model to handle both linear and nonlinear relationships while identifying deep interactions among the features.

The gradient descent method is used to optimize the parameters of the logistic neural network by iteratively adjusting the weights and biases in each layer, minimizing the error between predicted and actual values, thereby enhancing classification accuracy.

2.3 Experiment setting

To identify the best-performing models for both regression and classification tasks, the experiment is divided into two main parts. The first part examines the performance of the Linear model and the neural network model in the regression task, while the second part investigates the performance of the Logistic model and the neural network model in the classification task.

2.3.1 Regression Task

a. Linear Model

Since the linear model has relatively few tunable parameters, we directly use the LinearRegression method from the sklearn library for training. The model is trained on three different versions of the dataset: the raw, unprocessed data, the scaled data, and the data containing only the most relevant features. For each experimental setup, the process is repeated 30 times, with a different random seed used for splitting the dataset into training and testing sets each time. This ensures variability in the results and guarantees the reproducibility of the experiments.

b. **Neural Network Regression**

This section involves using a neural network model to make predictions. Since the optimal parameters were not known beforehand, grid search was used during training to find the best parameters for the model. The results show that the optimal configuration includes 64 neurons in each layer, the ReLU activation function, a learning rate of 0.05, and the SGD optimizer. The model was trained three times, with 500 epochs for each run, and the experiment was repeated three times.

2.3.2 Classification Task

a. **Logistic Model**

In this experiment, the base model we used is Logistic Regression, which is specifically designed for classification tasks. Since the original target values are numerical, they need to be converted into categorical data for the classification task. Therefore, a threshold value is required to distinguish between the categories. In this experiment, the value of ring_age = 7 (as recommended on Ed) was chosen as the standard. Values less than 7 were mapped to 0, while values greater than or equal to 7 were mapped to 1. This ensures that the experiment can proceed smoothly.

b. Neural Network

In this section, the MLPClassifier method from the sklearn library was used, and grid search was employed to find the optimal parameters. The best configuration was determined to be 64 neurons in each hidden layer, 500 epochs, a learning rate of 0.05, the SGD optimizer, and the logistic function as the activation function.

3 Results

3.1 Regression

Table1: Regression result of training data:

Model	Mean RMSE	Std RMSE	Mean R2	Std R2
Linear (full features)	2.1835	0.0322	0.5399	0.0113
Normalized Linear (full features)	2.1835	0.0322	0.539	0.0113
Linear (2 features)	2.5021	0.0393	0.3959	0.0127
Neural network	2.0565	0.061	0.587	0.0162

Table2: Regression result of testing data:

Model	Mean RMSE	Std RMSE	Mean R2	Std R2
Linear (full features)	2.234	0.051	0.520	0.022
Normalized Linear (full features)	2.234	0.051	0.520	0.022

Linear (2 features)	2.521	0.060	0.390	0.019
Neural Network	2.2864	0.0754	0.5053	0.0472

According to the results from Table 1 and Table 2, we can observe that the Linear Regression model (with full features) has a Mean R^2 of 0.5399 on the training set and 0.520 on the test set, indicating that the model can only explain about half of the variance, which may be related to the fact that the data is not normally distributed.

The performance of the Linear Regression model with two features shows a significant drop, suggesting that the linear model is sensitive to feature selection. And we can see that the RMSE and R^2 before and after normalization are almost identical. This may be because the model itself is robust to different feature scales.

However, on the test set, the performance of the neural network is lower than that of the linear model, and its higher Std RMSE and Std R^2(**Figure3** in appendix) indicate that the model's performance fluctuates more during training. This could be due to its higher sensitivity to noise or details in the training data. Linear neural network is essentially still a complex nonlinear model. Even if only a few hidden layers are used, its expression ability is still far superior to Linear regression. If there is not enough nonlinear structure, a simple Linear regression model may be more suitable than a complex neural network model.

3.2 Classification

Table3: Classification result of training data:

Model	Mean accuracy	Std accuracy	Mean AUC	Std AUC
Logic (full features)	0.9353	0.0028	0.9543	0.0032

Normalized Logic (full features)	0.9353	0.0028	0.9543	0.003
Logic (2 features)	0.9341	0.0026	0.9514	0.0036
Neural Network	0.9363	0.0021	0.9601	0.0036

Table4: Classification result of testing data:

Model	Mean accuracy	Std accuracy	Mean AUC	Std AUC
Logic (full features)	0.9363	0.0043	0.9540	0.0051
Normalized Logic (full features)	0.9363	0.0043	0.9540	0.0051
Logic (2 features)	0.9357	0.0045	0.9518	0.0047
Neural Network	0.9354	0.0013	0.9546	0.0066

According to the results from Table 1 and Table 2, we can see that the Logistic Regression (full features) model has a Mean Accuracy of 0.9363 and a Mean AUC of 0.9540 on the test set, indicating that the model has strong predictive capability, performing similarly to the training set. Additionally, the very small standard deviation suggests that the Logistic Regression model is highly stable and exhibits strong generalization ability. The impact of normalization on the model remains negligible. However, the reduction in the number of features leads to a slight decline in model performance, highlighting the importance of feature selection.

The Neural Network model shows a Mean Accuracy and Mean AUC of 0.9363 and 0.9601, respectively, on the training set, which is a slight improvement over the Logistic Regression model, particularly in terms of the AUC value. This demonstrates the Neural Network's advantage in capturing nonlinear features. There may be the following

reasons. Logistic neural networks can automatically learn feature representation through hidden layers. It will perform multiple linear transformations on input features and add nonlinear activation to build better features. Logistic regression is a single-layer linear model with relatively few model parameters and limited model capacity.

3.3 Neural Network trained with SGD

Although the evaluation criteria for regression and classification models differ, in terms of model performance, classification clearly outperforms regression in predicting abalone age. The overall AUC and accuracy of the classification model are excellent, whereas the highest explanatory power of the regression model is less than 60%, indicating its predictive capability is far inferior to classification. Therefore, we use the full-featured logistic neural network with the SGD method to explore the parameters. (**Figure4:** ROC Curve for Optimized MLP Model (10 Neurons, lr=0.01)
The optimal parameter results obtained are as follows:

Table5: Best parameter of neural network trained with SGD

Neurons	Learning Rate	Test Accuracy	AUC
10	0.01	0.9366	0.9538
30	0.001	0.9354	0.9517
30	0.01	0.9354	0.9520

4.Discussion

4.1 Our goals

After analyzing the results, we observed that in the regression task, using the dataset with all features performed significantly better. Regularizing all the features led to a slight improvement compared to the unprocessed data, but the difference was not substantial. However, the use of a neural network provided a noticeable improvement in handling the task.

In the classification task, the Logistic model showed almost identical performance across the three different datasets, whereas the neural network achieved significantly

better classification results. Therefore, in both the regression and classification tasks, the neural network demonstrated superior performance.

Overall, neural networks exhibit superior performance in both regression and classification tasks. This advantage arises from their ability to capture complex, nonlinear relationships between features during the training process. In contrast to traditional linear models that rely on individual features, neural networks leverage the full spectrum of feature interactions, leading to more effective data utilization and enhanced predictive accuracy.

4.2 Limitations

The methods used in this experiment have certain limitations. For the linear model, only individual features were used as input without applying feature interactions. As a result, the linear model failed to capture hidden relationships between features during training. For the neural network model, the experiment was limited by the choice of optimization method, and more effective optimizers were not explored. This led to inefficient use of computational resources during training, unnecessarily prolonging the convergence time.

5. Conclusions

This experiment confirmed the effectiveness of neural networks in both regression and classification tasks. Under the same data processing conditions, neural networks consistently produced superior results. The directions for future research are as follows:

- Process the dataset by incorporating more interaction features to explore the relationships between different features.
- In the classification task, the division of the target can lead to an imbalance between positive and negative samples [13]. In future work, a midpoint could be selected as the basis for division to achieve a better balance between positive and negative samples.

- For neural networks, future research could explore the use of different optimizers to test and analyze how they impact the training process. Additionally, varying the number of neurons and hidden layers is another potential direction for further investigation.

References

[1] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115-133.

[2] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[5] Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press.

[6] Gauss, C. F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Frid. Perthes et IH Besser.

[7] Legendre, A. M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes.Friedman J, Popescu B E. Gradient directed regularization for linear regression and classification[R]. Technical Report, Statistics Department, Stanford

University,2003.

[8] Bottou, L.(2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier, Y., Saporta, G. (eds) Proceedings of COMPSTAT'2010. Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2604-3_16

[9] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

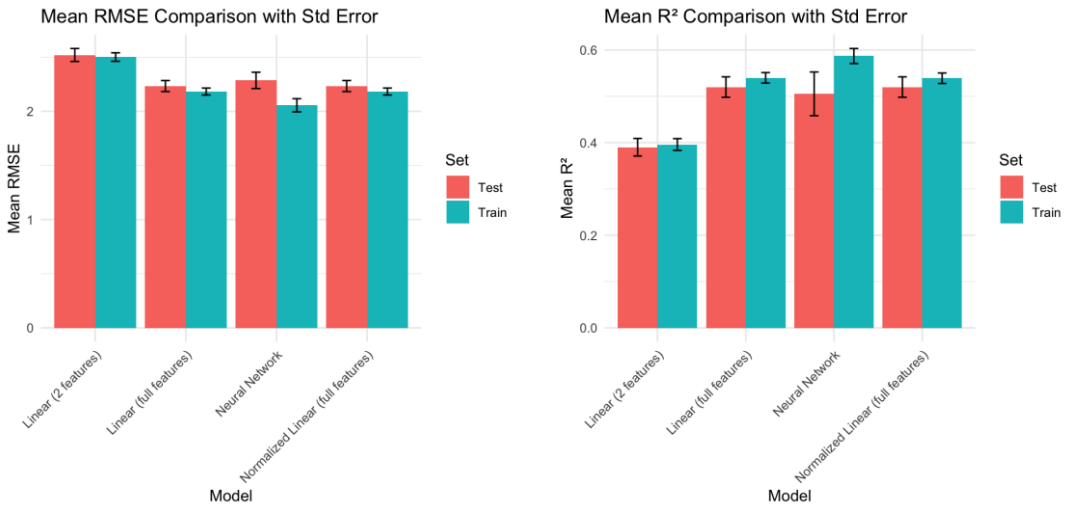
[10] Arora, Sanjeev, et al. "A convergence analysis of gradient descent for deep linear neural networks." arXiv preprint arXiv:1810.02281 (2018).

[11] Zou, Xiaonan, et al. "Logistic regression model optimization and case analysis." 2019 IEEE 7th international conference on computer science and network technology (ICCSNT). IEEE,2019.

[12] Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5-6 (2002):352-359.

[13] Japkowicz, N. & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-449.

Appendix



(a) **(b)**

Figure3: (a) Mean RMSE Comparison with Standard Error Across Models;(b) Mean RMSE Comparison with Standard Error Across Models

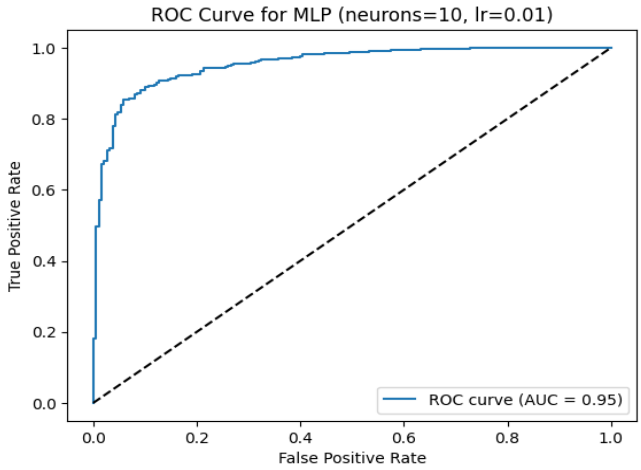


Figure4: ROC Curve for Optimized MLP Model (10 Neurons, lr=0.01）

Task Allocation Table

Task Module	Sub-task Description	Responsible Person
Report Writing	All members contribute to writing the final report.	Hangwei Liang, Jiehao Chen, Junru Hua
Data Processing		
Data Cleaning	Clean the data (e.g., convert	Junru Hua

	'M' and 'F' to 0 and 1). This can be done using code or simple find-and-replace.	
Correlation Heatmap	Develop a correlation heatmap and discuss major observations.	Jiehao Chen
Scatter Plot with Ring-Age	Pick the two most correlated features (positive and negative) and create a scatter plot with ring-age. Discuss major observations.	Junru Hua
Histograms	Create histograms of the two most correlated features and ring-age. Discuss major observations.	Hangwei Liang
Additional Visualization (OPTIONAL)	Add any other visualizations of the dataset you find appropriate.	Jiehao Chen
Linear Regression Model	Develop a linear regression model using all features to predict ring-age, using 60% of the data for training and 40% for testing. Visualize predictions, report RMSE, R-squared, AUC score, and create an ROC plot.	Hangwei Liang
Linear/Logistic Regression Comparison	Compare linear/logistic regression with and without normalizing input data (from step 1).	Junru Hua
Two-Feature Regression Model	Develop a linear/logistic regression model using two selected input features from the data processing step.	Jiehao Chen

Neural Network Comparison	Compare the best approach with a neural network trained with SGD. Experiment with hyperparameters such as the number of hidden layers/neurons and learning rate. Discuss observations.	Hangwei Liang
Discussion	Compare neural network and linear model results. Discuss how to further improve the model.	Hangwei Liang, Jiehao Chen, Junru Hua
Experimental Runs	Run 30 experiments (or 3 if faced with wall-time issues). Report the mean and std of RMSE and R-squared for train and test datasets.	Hangwei Liang, Jiehao Chen, Junru Hua

Contribution:

Hangwei Liang – 100%

Jiehao Chen – 100%

Junru Hua – 100%