Predicting Abalone Age Using Machine Learning

Nana Chong

UCLA Department of Atmospheric & Oceanic Sciences

December 2025

Abstract

For this project, I used several machine learning models to predict abalone age from their physical measurements. This includes Linear Regression, Ridge Regression, Decision Trees, Random Forests, SVR, and a Neural Network (MLP). The linear models performed reasonably well, while the tree-based models did not capture the patterns in the data as effectively. I also tested different parameters for each model and looked at how scaling and dataset imbalance affected performance. Overall, the MLP model gave the highest accuracy, making it the best-fitting model for this dataset.

1. Introduction

The goal of this project is to use machine learning models to predict the age of abalone using their physical measurements. Since actual age isn't recorded, the dataset uses the number of rings as an age estimate, which turns this into a regression problem. The features include measurements like length, diameter, different weight categories, and the sex of the abalone. Throughout the project, I tried several regression models including Linear Regression, Ridge Regression, Decision Trees, Random Forests, Support Vector Regression (SVR), and a Neural Network (MLP). I trained all of them using the same train–test split so that their results were comparable. I also included the preprocessing steps, like one-hot encoding for the Sex feature, scaling the numeric data, and checking the balance of the Rings distribution. By comparing the performance of all the models, I wanted to see which approach predicts abalone age the most accurately and what the strengths and weaknesses of each method are.

2. Data

The dataset comes from the **UCI Machine Learning Repository** and contains 4,177 abalone samples. Each sample includes measurements such as:

- Length, Diameter, Height
- Whole, Shucked, Viscera, and Shell weight
- Sex (M = male, F = female, I = infant)
- Rings (target variable)

In the Abalone dataset, age is not measured directly. Instead, the number of shell rings is used as a proxy, where age (in years) is typically estimated as:

Age = Rings + 1.5

This means that "Rings" is an imperfect measurement of true age, because biological growth rates vary between individuals. As a result, there is natural noise in the target variable that limits how accurate any regression model can be.

	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

The dataset has no missing values, but the age distribution is imbalanced since most abalones are young. This imbalance can influence model accuracy, especially for older abalones.

Preprocessing the Data

To prepare the data for modeling, I applied the following steps:

1. **Loading & Cleaning**

```
data = pd.read_csv(filepath, names=column_names)
data.head()
```

This assigned column names from the UCI documentation and confirmed no missing values.

2. **Encode Categorical Features**

```
data_clean = pd.get_dummies(data, columns=['Sex'], drop_first=True)

print("\nFirst 5 rows after encoding 'Sex':")
display(data_clean.head())
```

This converts Sex to numerical variables.

3. **Feature & Target Split**

```
X = data_clean.drop('Rings', axis=1).values
y = data_clean['Rings'].values
```

This separated the dataset into input features X and the target variable y = Rings, which represents the abalone's age minus 1.5 years.

4. **Train-Test Split**

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

The dataset was split into training and testing subsets using an 80/20 split..

5. **Scaling**

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Numerical features were standardized using z-score normalization (StandardScaler), which helps models like Ridge Regression, SVR, and MLP perform more effectively.

These preprocessing steps create a clean numeric dataset that can be used with all of the machine learning models applied later in the project.

3. Explorating Data Analysis (EDA)

Before training any models, I analyzed the Abalone dataset to better understand the relationships between the physical measurements and the target variable, Rings, which is used as a proxy for age.

3.1 Distribution of the Target (Rings)

Figure 1 shows a right-skewed dataset with a strong imbalance: most abalones have between 6–10 rings, while older individuals (12+ rings) are relatively rare, which is important for model bias.

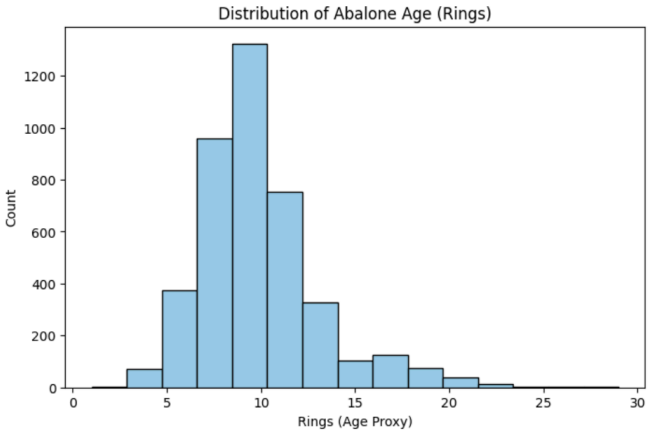

Figure 1: Histogram of Rings (Abalone Age Proxy)

This imbalance affects model performance because the regression models learn the patterns of the large “young abalone” group more effectively than the rare “older abalone” group. As a result:

- Predictions for younger abalones are more accurate
- Older abalones are consistently under-predicted
- Overall R^2 scores decrease due to the larger errors on the rare age group

3.2 Correlation Heatmap

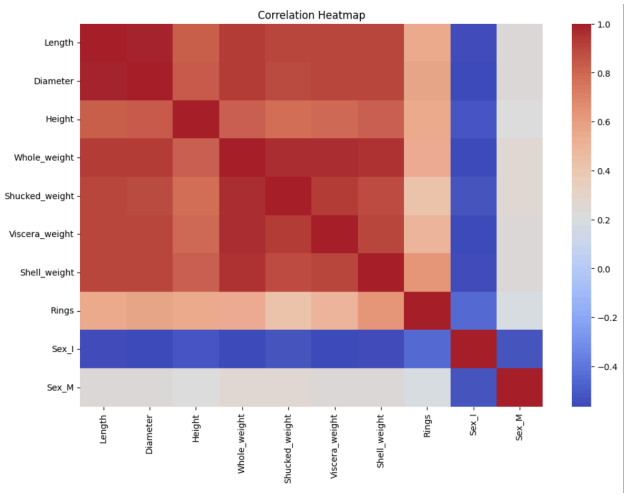

Figure 2: Correlation heatmap of all features including Rings

Figure 2 displays the strongest correlations with Rings:

- Shell Weight
- Viscera Weight
- Whole Weight

This is biologically reasonable since older abalones tend to be larger.

3.3 Scatter Plots

Figures 3–5 illustrate clear upward trends between Rings and each of the key size features: Length, Diameter, and Shell weight.

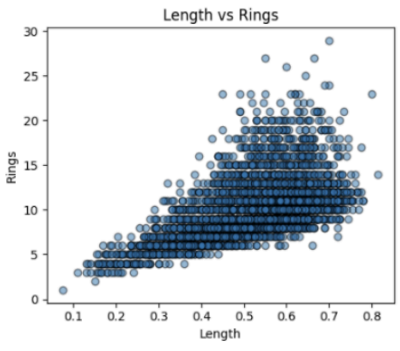

Figure 3: Length vs Rings

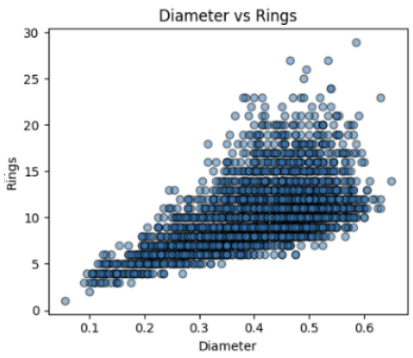

Figure 4: Diameter vs Rings

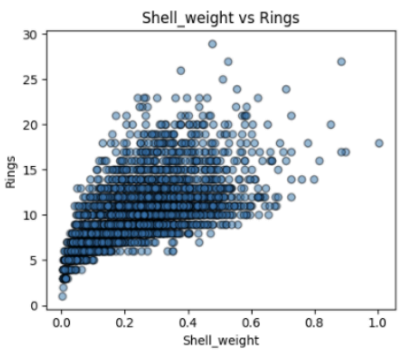

Figure 5: Shell Weight vs Ring

All three features show positive trends with Rings. The relationships are not perfectly linear, which supports trying nonlinear models. Increased scatter at higher ring counts reflects the imbalance dataset.

3.4 Boxplots of Rings by Sex

Comparison of the age distributions across the three sex categories (M, F, I) is plotted below:

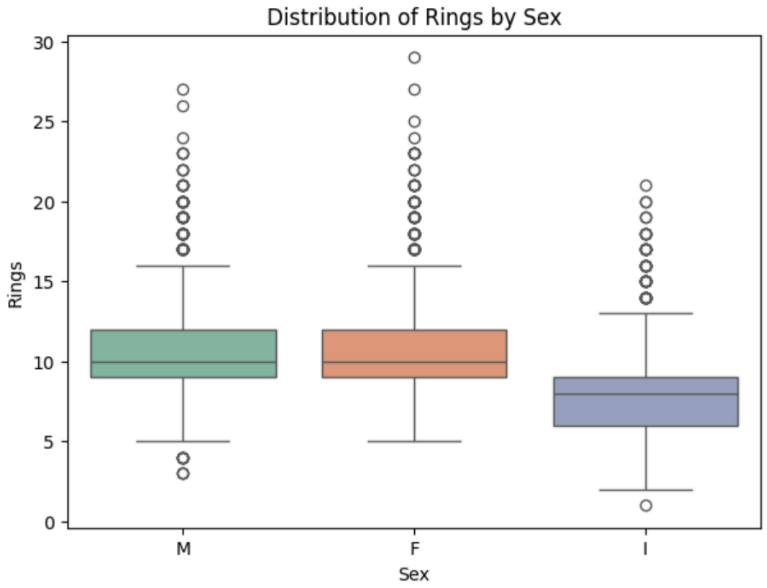

Figure 6: Distribution of Rings across Sex (M, F, I)

The plot reveals:

- Infant (I) abalones tend to be significantly younger
- Male (M) and Female (F) abalones have similar age distributions
- Each category contains multiple older outliers

Figure 6 shows only small differences among sexes, meaning Sex contributes but is not dominant.

4. Modeling

After preprocessing the data and examining its structure, I trained several supervised learning models to predict the number of rings (age) of each abalone. Each model was trained on standardized training data (for linear, SVR, and MLP models) and evaluated using R^2 and RMSE on the test set.

4.1 Linear Regression

Linear Regression served as a baseline model.

This model is useful for understanding whether Rings can be predicted from a purely linear combination of features.

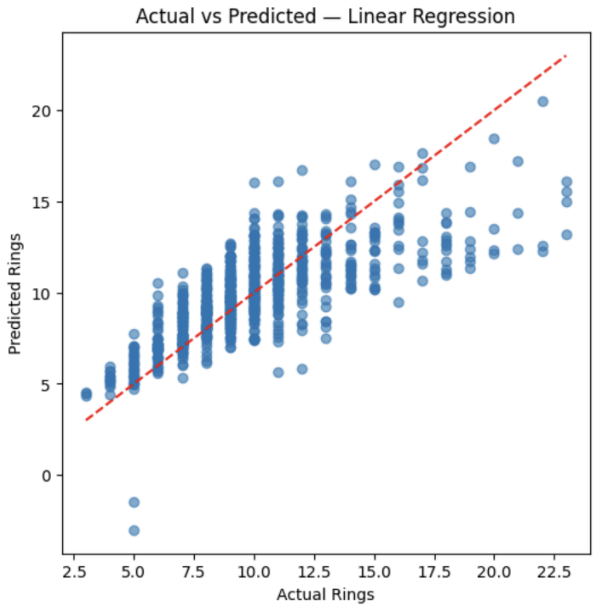

Figure 7: Actual vs Predicted Rings – Linear Regression

The model produced moderate performance, confirming that while feature size relates to age, the relationship is not perfectly linear.

4.2 Ridge Regression

Ridge Regression adds L2 regularization to prevent overfitting and stabilize coefficients.

This model performed very similarly to standard Linear Regression but slightly better in terms of generalization.

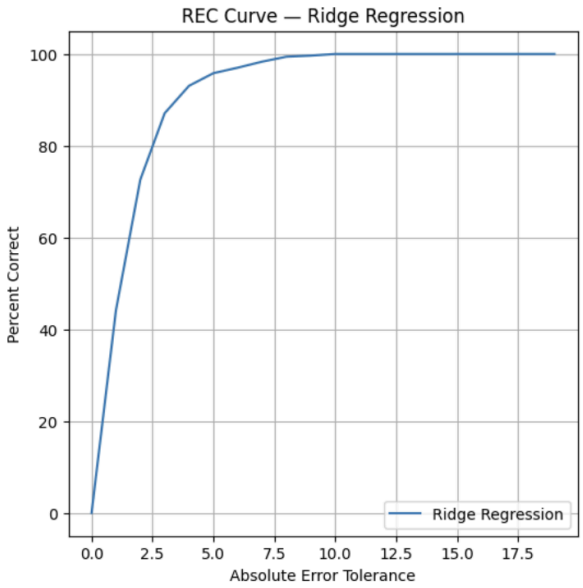

Figure 8: REC Curve – Ridge Regression

Ridge was helpful because several features (especially the different weight measurements) are correlated with each other, and regularization tries to stabilize the model and reduce coefficient inflation.

4.3 Decision Tree Regression

A Decision Tree was used to model nonlinear relationships between features and Rings.

Decision Trees can capture feature interactions and sharp splits in the data, though they tend to overfit, especially with small variations between samples.

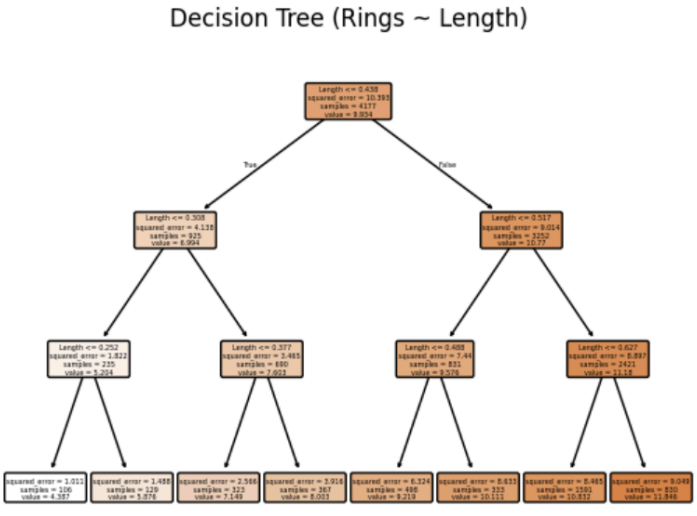

Figure 9: Visualization of the Decision Tree

In this dataset, the Decision Tree performed noticeably worse than linear models, suggesting that its structure was too rigid for predicting continuous age.

4.4 Random Forest Regression

The Random Forest model aggregates many decision trees to reduce overfitting and improve generalization.

Although it performed better than a single Decision Tree, it still did not outperform the linear or kernel-based models in this dataset.

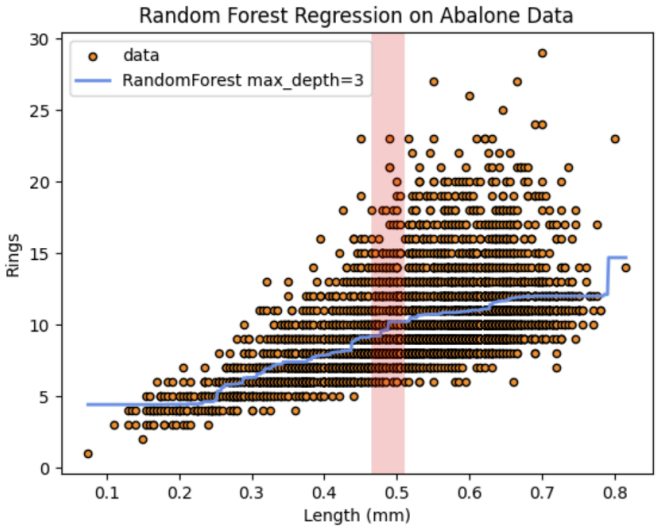

Figure 10: Random Forest Predictions vs Actual

This suggests that the relationships between features and Rings are smooth rather than sharply segmented, making simple averaging of tree splits less effective.

4.5 Support Vector Regression (SVR)

SVR with an RBF kernel performed significantly better than the tree-based models.

SVR handles smooth nonlinear relationships well and relies on support vectors rather than all points, it captures the structure of the data more effectively.

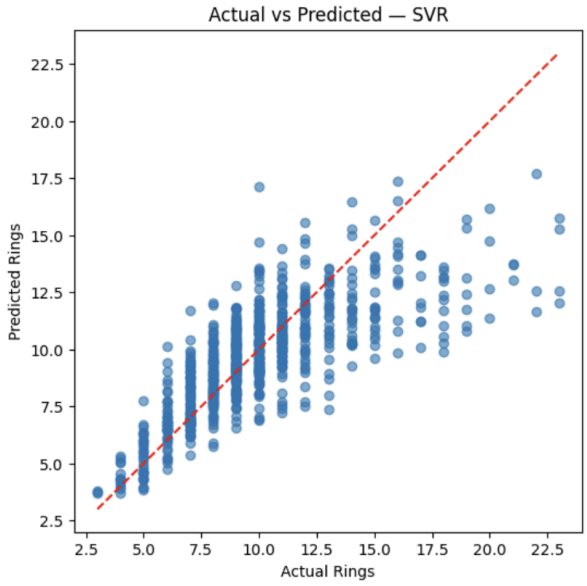

Figure 11: SVR Predictions vs Actual

SVR was one of the top-performing models, confirming that the underlying patterns in the dataset are nonlinear but smooth.

4.6 Neural Network (MLPRegressor)

A Multi-Layer Perceptron (MLP) was trained using two hidden layers.

After scaling the data, the neural network learned more complex interactions between features and produced the highest R^2 score among all models tested.

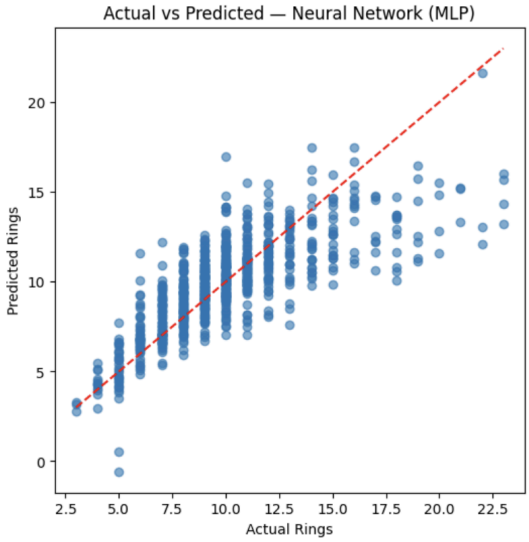

Figure 12: MLP Predictions vs Actual

MLP performed best likely because:

- It models nonlinear interactions between multiple features simultaneously
- Weight-based features and size measurements benefit from learned nonlinear transformations
- The dataset is large enough for the network to generalize well

This diversity in performance indicates that abalone age prediction benefits from models capable of learning smooth nonlinear relationships, especially when multiple interacting features are involved.

5. Results

To evaluate model performance, I compared all six regression models using two metrics:

- R^2 Score (higher is better): proportion of variance in Rings explained by the model
- RMSE (lower is better): average prediction error in units of rings

5.1 Performance Summary Table

Table 1. Test Set Performance of All Models

Model	R^2 Score	RMSE
MLP	0.585039	2.119443
SVR	0.562632	2.175913
Ridge	0.548180	2.211572
Linear	0.548163	2.211613
RandomForest	0.324037	2.650498
DecisionTree	0.318888	2.660573

5.2 R^2 Comparison Plot

This bar chart visually compares how well each model explains the variability in abalone age.

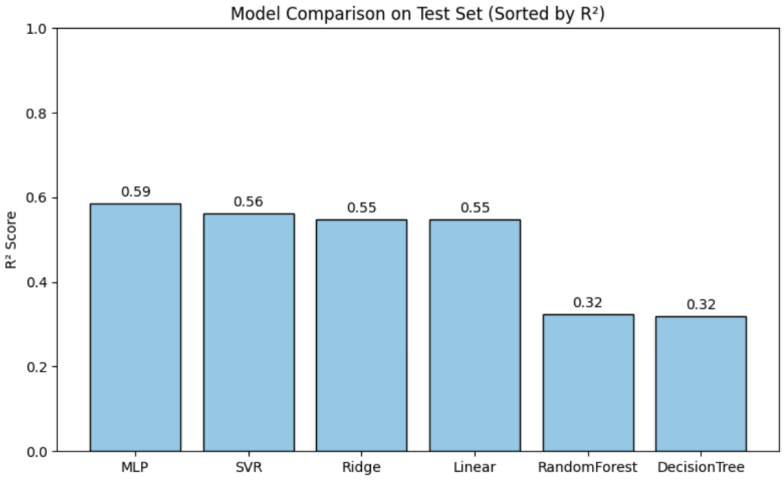

Figure 13: R^2 Scores for All Models

Figure 13 compares the R^2 values for all models, showing that the MLP model performed best overall.

5.3 RMSE Comparison Plot

The corresponding RMSE comparison (shown previously in Figure 12) provides complementary information:

- MLP and SVR had the lowest RMSE values, confirming they produced the most accurate predictions.
- Although Linear and Ridge Regression had moderate RMSE values, they still underperformed compared to nonlinear models.

Summary of Model Performance

Across both evaluation metrics:

- Linear models underfit the nonlinear biological relationships.
- SVR and MLP performed well but required tuning and scaling.

The models capable of learning smooth nonlinear relationships are best fit for predicting abalone age.

6. Discussion

The results from my models showed a few clear patterns about both the Abalone dataset and how different algorithms handled it. Overall, the models that were able to capture smoother nonlinear trends, especially SVR and the MLP neural network, performed noticeably better than the linear and tree-based models. This matched what I saw in the EDA: features like shell weight and whole weight increased with age, but the relationship wasn't perfectly straight. The scatterplots (Figures 3–5) showed curved patterns, which explains why Linear Regression and Ridge Regression stayed around an R^2 of about 0.55.

Tree-based models performed the worst. Both the Decision Tree and Random Forest underfit the data, with R^2 values around 0.32. This lines up with what we know, since decision trees create piecewise-constant predictions that don't work well for a smooth biological relationship like growth and age. Random Forests usually help, but the dataset is small and the trend is gradual, so they didn't improve much.

The best-performing models were the MLP ($R^2 \approx 0.58$) and SVR ($R^2 \approx 0.56$). These models can learn nonlinear relationships more naturally: SVR through its kernel and MLP through activation functions. Because the patterns in the data were nonlinear but not extremely complex, these models were a better fit.

Even the best model didn't go above about 0.60 R^2, which shows that the dataset has limitations. The Rings variable has a lot of biological variability, and abalones of different ages can sometimes look very similar in size. The dataset is also imbalanced, with many more young abalones than older ones, which makes accurate predictions harder.

The imbalanced age distribution plays a major role in model performance. Because older abalones appear much less frequently in the dataset, the models do not have enough examples to learn their patterns well. All models tended to under-predict the age of older abalones, which contributed to lower R^2 values. This dataset imbalance also explains why even the strongest models (MLP and SVR) only achieved around 0.57–0.58 R^2.

Overall, these results showed that the Abalone dataset works better with models that can capture smooth nonlinear trends, and that tree-based models were not a great match. It also reinforced how important the EDA was, because the structure of the data basically predicted which models would perform well

7. Conclusion

From this project, several main takeaways emerged from applying multiple machine learning models to predict abalone age:

- The MLP neural network produced the strongest overall performance in this dataset, giving the highest R^2 score among all models tested.
- Shell weight, whole weight, and viscera weight appeared to be the most influential features, consistently showing strong predictive value across models.
- Although SVR also performed well, its results indicate that the model still had difficulty distinguishing age differences in some regions of the feature space.
- Decision Trees and Random Forests performed noticeably better when their depth was limited, but even then, they struggled to capture the smooth nonlinear growth patterns in abalones.
- Linear and Ridge Regression were the weakest performers because the relationship between age (Rings) and size-related features is not strictly linear.

Overall, nonlinear models provided the best predictive ability for this dataset, while simpler models were less suited to the gradual biological growth patterns present in abalone measurements.

8. References

1. Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). Abalone [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W
2. Mehta, T., Patel, K., & Patel, N. (2019). A review on abalone age prediction using machine learning techniques. International Journal of Computer Applications, 178(50), 1–5. https://doi.org/10.5120/ijca2019919425
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830