DATA 4380 – Tabbular Project

# Predicting Abalone Age Using Tabular Machine Learning Models

Alain Siddiqui & Sidhantaa Sarna

Advisor: Dr.Rostami

# Research Objectives

**How accurately can machine learning models predict the rings of abalone from physical measurements?**

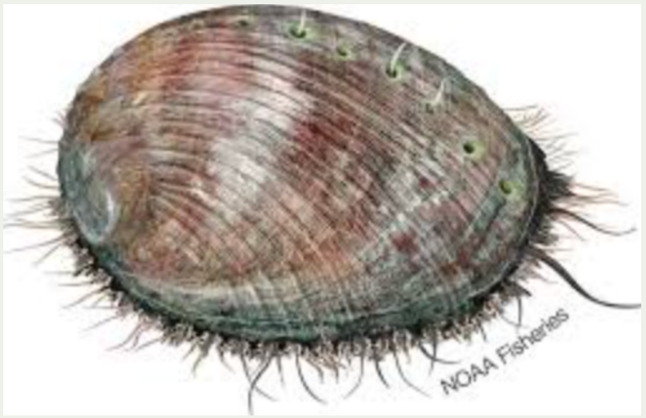
Which physical measurements (e.g., length, diameter, weight) are most important in predicting abalone age?

Does treating the age prediction as a regression task or as a classification task (young/middle/old age bins) lead to better predictive performance?

Do models systematically over- or under-predict the age of older abalones compared to younger ones?

# Introduction

- Abalone (Haliotis) is a type of marine snail. vegetarian, feeding mainly on algae.
- Use a large muscular foot to cling tightly to rocks.
- Have a single, flattened, ear-shaped shell with holes called respiratory pores.
- These pores are used for breathing, waste removal, and reproduction.
- Reproduce by releasing eggs and sperm into the water (broadcast spawning).
- Shells are beautiful and iridescent, used for jewelry, decoration, and currency.

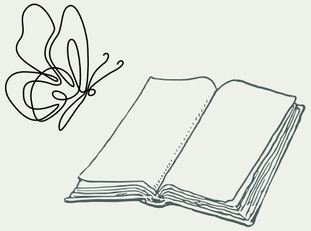



***Because counting the age rings on an abalone shell is slow and destructive, researchers wanted to estimate age using easily measurable features like weight and length, such as in this dataset.***

# Dataset Overview

Source: UCI Machine Learning Repository

- Tabular
- 4,177 rows
- 8 features

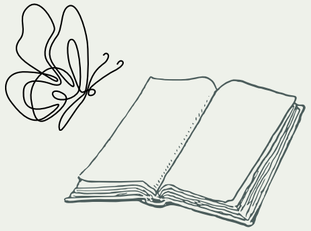


- No missing/null values
- Categorical
- Continuous
- Integer

# Literature Review

Data stream mining: having continuous data to make predictions and how regression and classification tasks play into that

Hybridizing methods to get the best results including using bagging with Forest models and using neural networks



Communication vs accuracy trade-off

# Research Methodology

The team will obtain the UCI Abalone dataset and calculate the target variable Age = Rings + 1.5 for all 4,177 records.

Exploratory analysis and visualizations will be conducted to examine feature distributions, identify correlations, and detect non-linear patterns between shell measurements and age.

The categorical feature sex will be encoded using OneHotEncoder, while numerical variables will be standardized using StandardScaler to ensure uniform scaling across all features.

Biologically relevant features such as weight_per_length, shell_to_whole_ratio, and shucked_to_whole_ratio will be created to capture shell composition and structural maturity.

A Linear Regression model will be developed to establish benchmark performance and evaluate its effectiveness in modeling linear growth relationships.

Ensemble learning algorithms, including Random Forest and XGBoost Regressors, will be trained and tuned using grid search to enhance predictive accuracy and generalization.

SHAP analysis will be applied to interpret model predictions, highlighting shell weight and tissue-to-shell ratios as the most influential features in determining abalone age.

SHAP analysis will be applied to interpret model predictions, highlighting shell weight and tissue-to-shell ratios as the most influential features in determining abalone age.

# EXPLORATORY DATA ANALYSIS

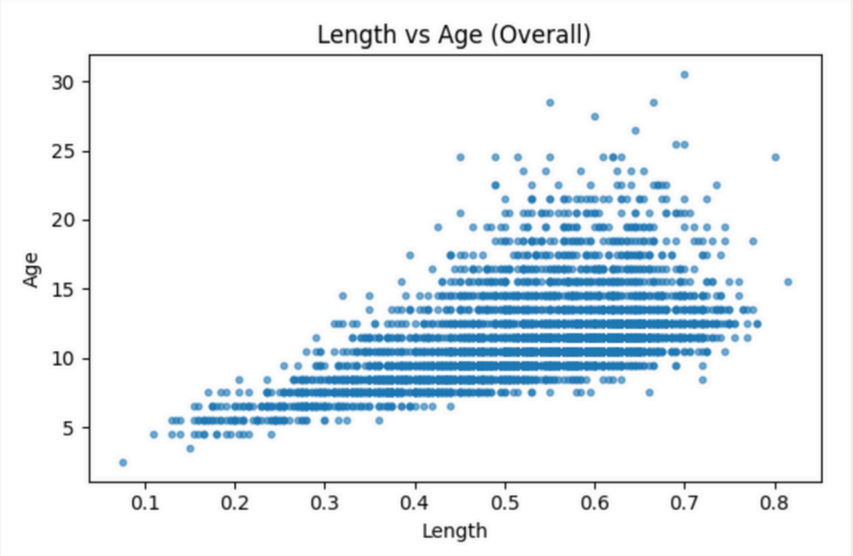
## Summary Statistics

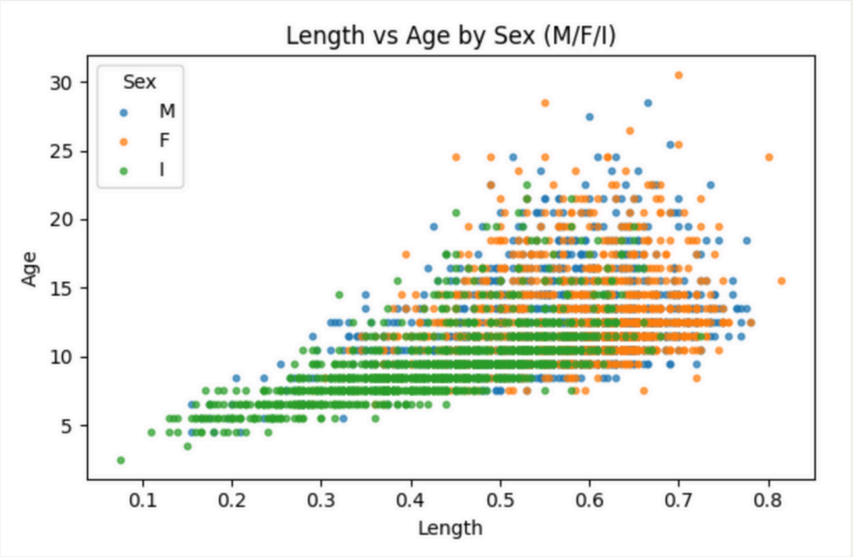
```
Summary statistics:
           length     diameter       height  whole_weight  shucked_weight  \
count  4177.000000  4177.000000  4177.000000   4177.000000     4177.000000
mean      0.523992     0.407881     0.139516      0.828742        0.359367
std       0.120093     0.099240     0.041827      0.490389        0.221963
min       0.075000     0.055000     0.000000      0.002000        0.001000
25%       0.450000     0.350000     0.115000      0.441500        0.186000
50%       0.545000     0.425000     0.140000      0.799500        0.336000
75%       0.615000     0.480000     0.165000      1.153000        0.502000
max       0.815000     0.650000     1.130000      2.825500        1.488000

       viscera_weight  shell_weight        rings          age
count     4177.000000   4177.000000  4177.000000  4177.000000
mean         0.180594      0.238831     9.933684    11.433684
std          0.109614      0.139203     3.224169     3.224169
min          0.000500      0.001500     1.000000     2.500000
25%          0.093500      0.130000     8.000000     9.500000
50%          0.171000      0.234000     9.000000    10.500000
75%          0.253000      0.329000    11.000000    12.500000
max          0.760000      1.005000    29.000000    30.500000

Sex distribution:
sex
M    1528
I    1342
F    1307
Name: count, dtype: int64
```

- The dataset contained 4,177 samples with three sex categories (M: 1,528, F: 1,307, I: 1,342).
- Length and age showed a clear positive correlation, varying slightly by sex.
- Data appeared well-distributed and free of missing values, supporting reliable model training.

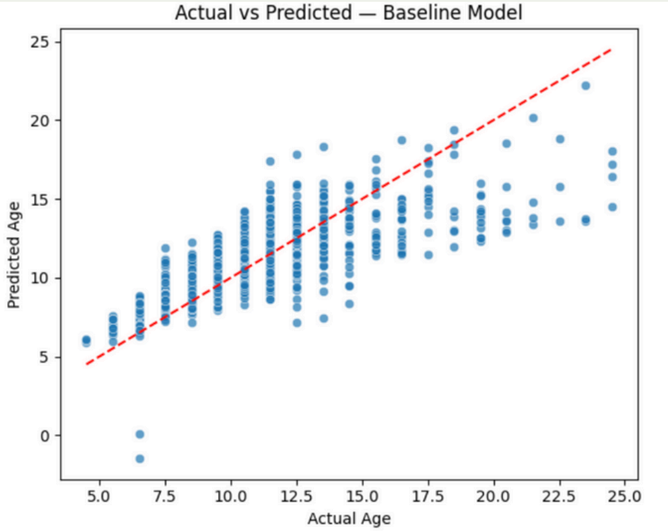


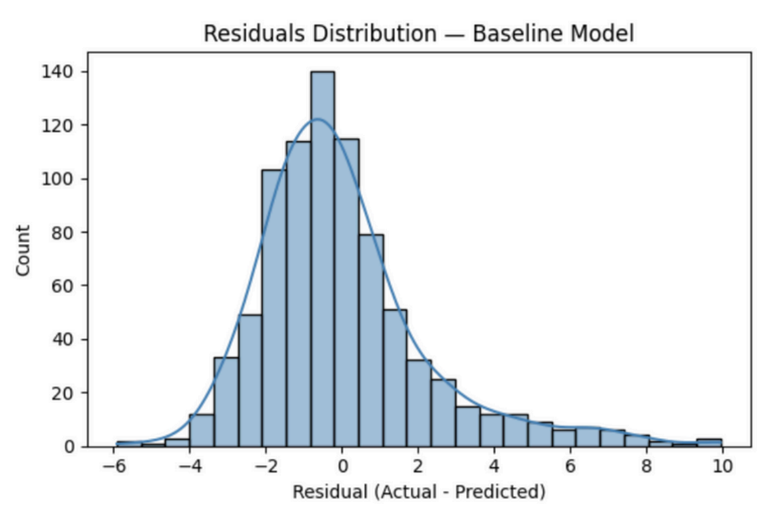
# Baseline Model

```
Baseline Model Performance:
RMSE: 2.248
MAE : 1.629
R²  : 0.533
```

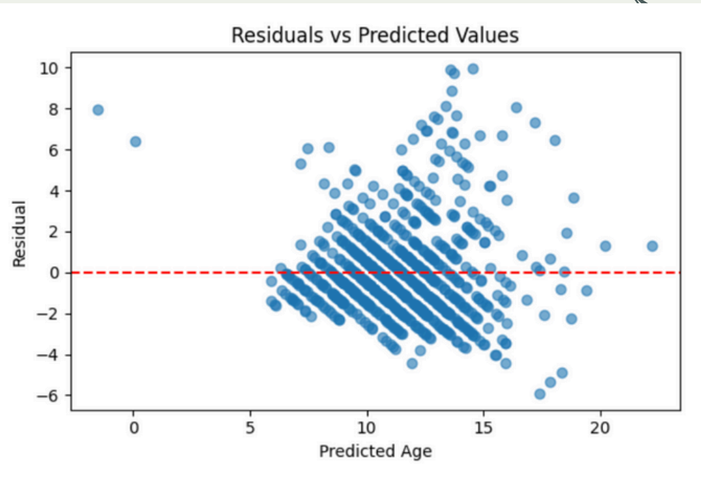
## Linear Regression Baseline




Actual vs Predicted Scatterplot




Residual Plot



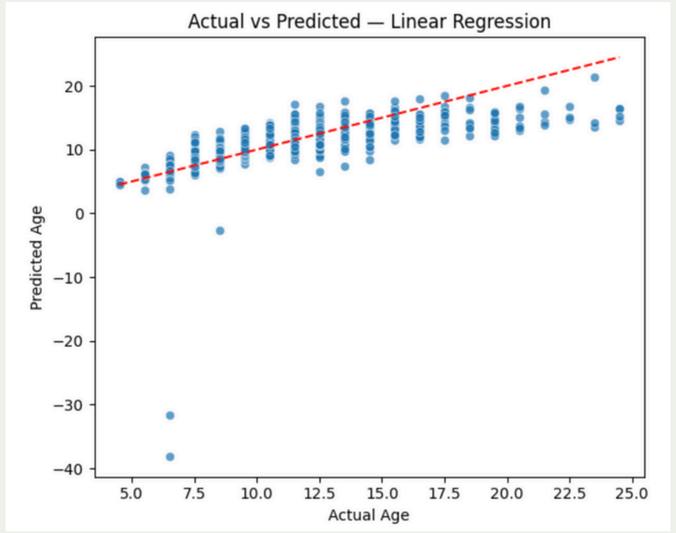

Homodescasity Check

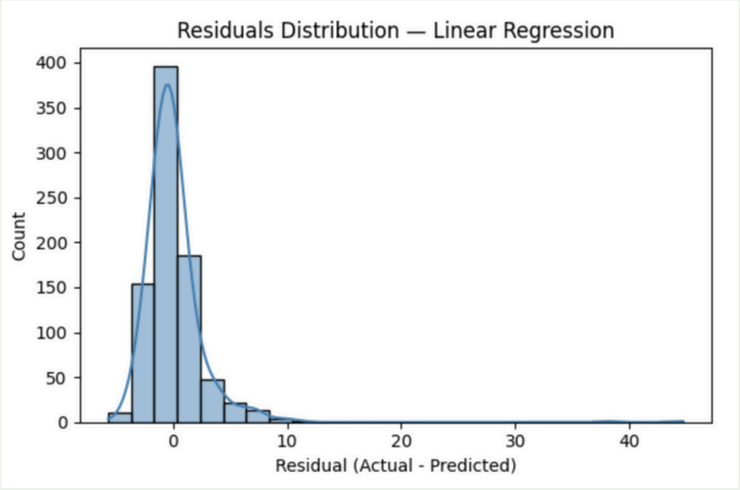
# DATA PREPROCESSING

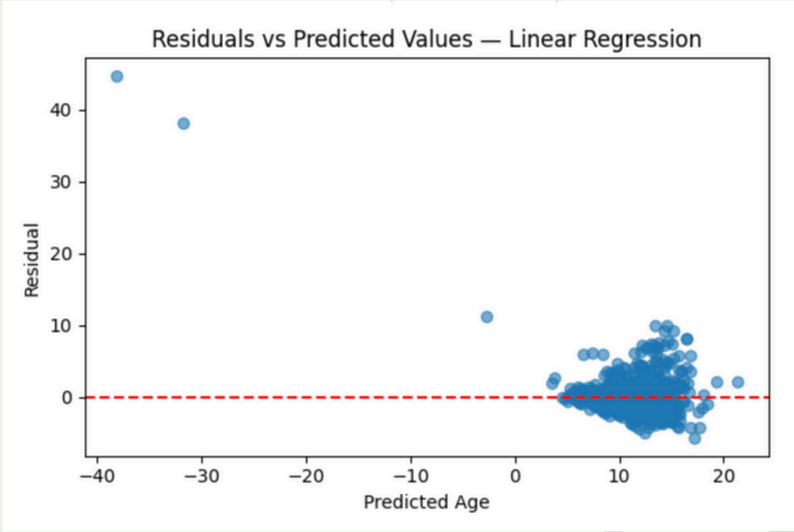
- Encoding Features
- Sex (M,F)
- Scaling Numerical Features

## After Preprocessing:

```
Linear Regression Performance:
RMSE: 2.992
MAE : 1.662
R²  : 0.173
```



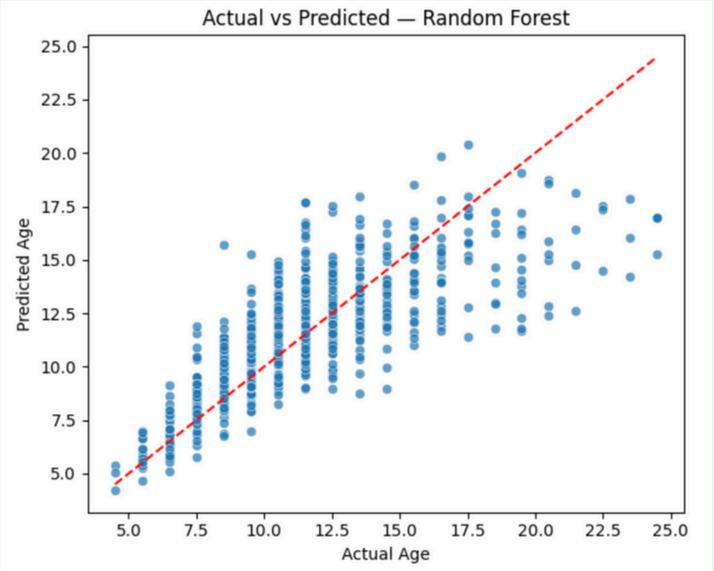



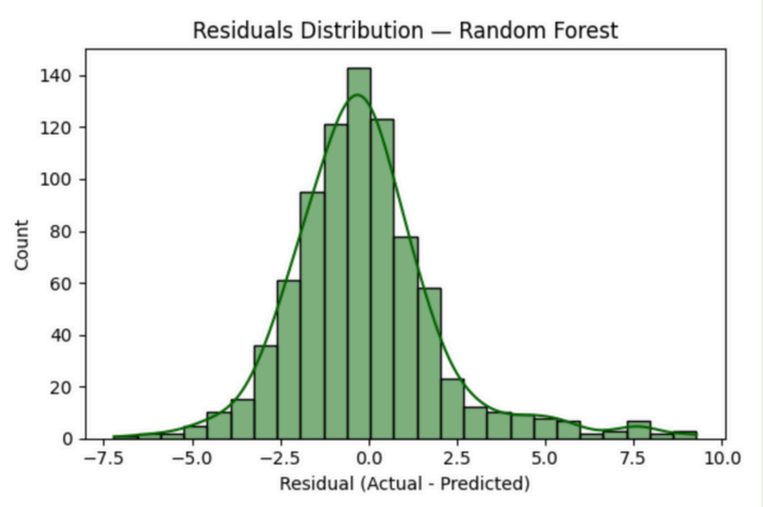
# RANDOM FOREST

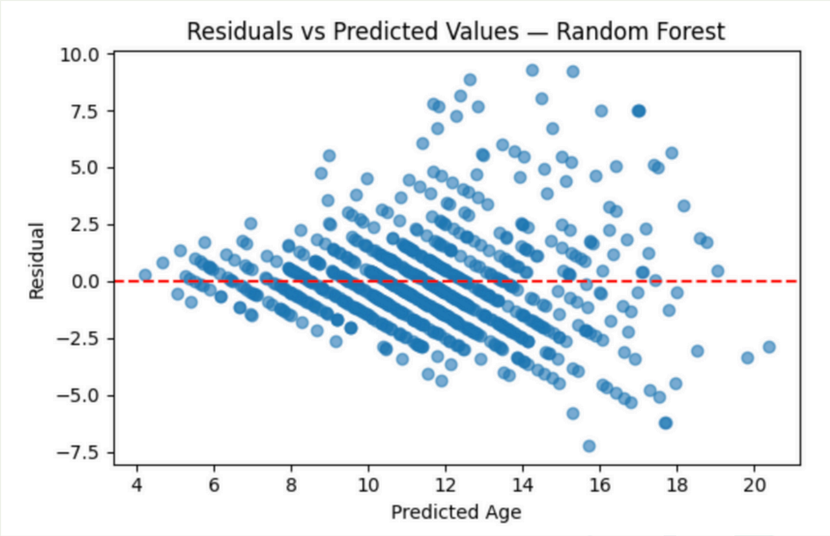
## After Preprocessing:

```
Random Forest Performance:
RMSE: 2.146
MAE : 1.532
R²  : 0.575
```

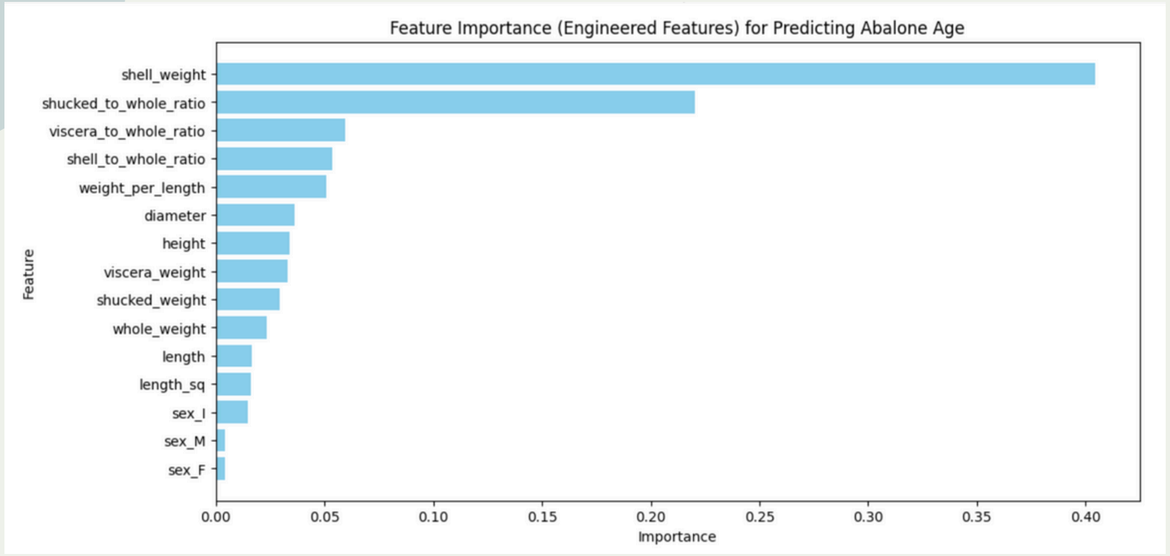
- The Random Forest achieved RMSE = 2.146, MAE = 1.532, and $R^2$ = 0.575, outperforming linear models.
- Residuals were centered near zero, indicating low bias and consistent variance across ages.
- This model demonstrated the best overall balance between accuracy and generalization, making it the baseline for tuning.



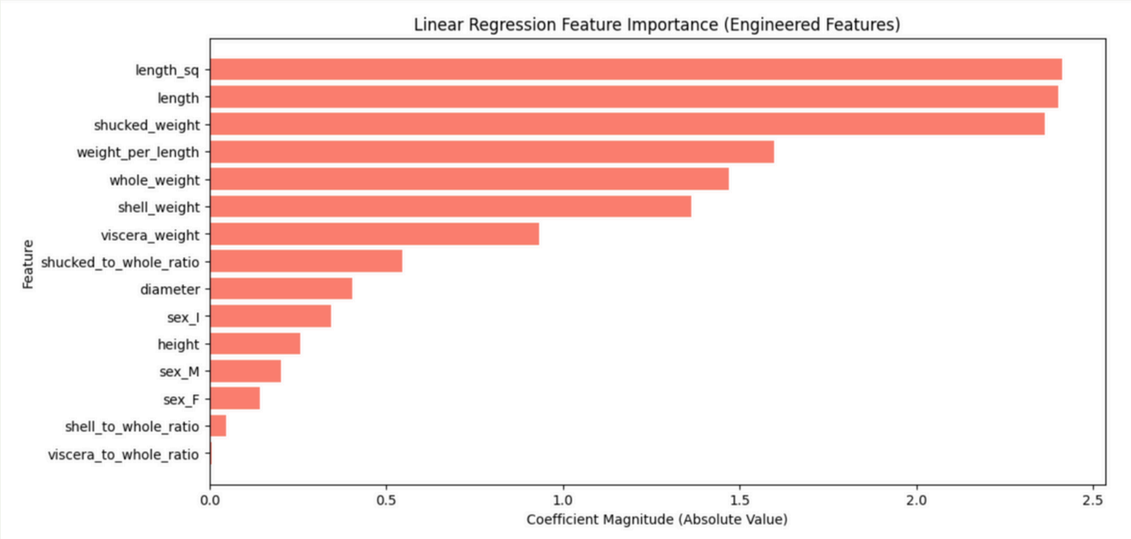



# FEATURE IMPORTANCE




- The Random Forest ranked shell_weight and shucked_to_whole_ratio as the top predictors of age.
- Engineered ratios and weight-based metrics contributed far more than basic dimensions.
- This confirms that mass-related traits best capture biological aging in abalone.

- Linear Regression emphasized length_sq, length, and shucked_weight as the strongest predictors.
- Magnitude differences suggest linear models rely more on size-driven relationships.
- Compared to RF, linear models underutilize the complex interactions present in the data.



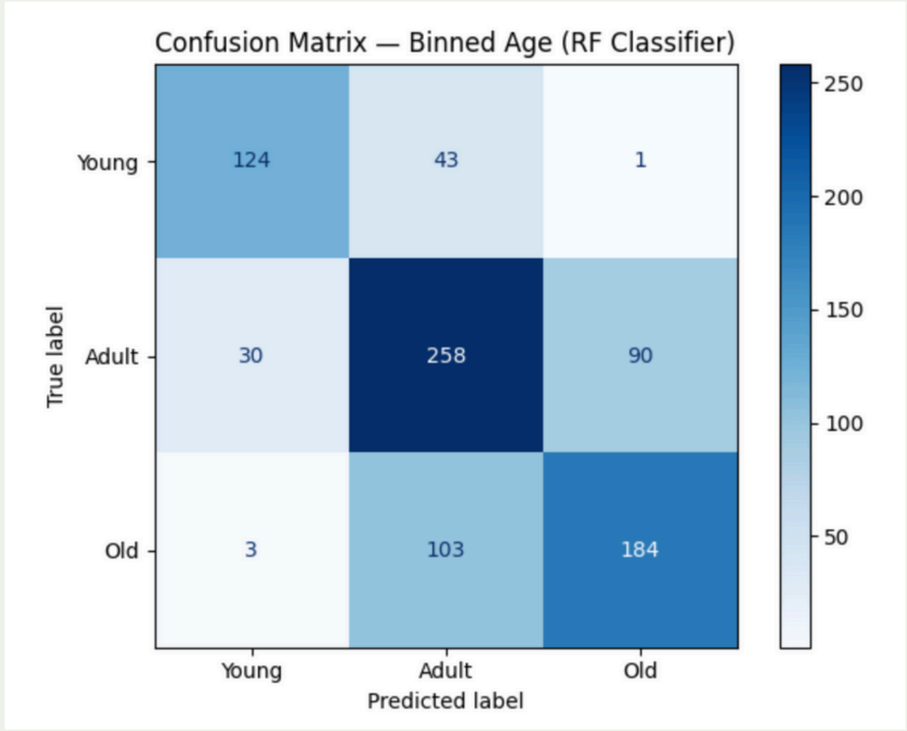
# XG Boost using engineered features

```
XGBoost Regressor Performance:
RMSE: 2.205
MAE : 1.571
R²  : 0.551
```

- The XGBoost Regressor achieved RMSE = 2.205, MAE = 1.571, and $R^2$ = 0.551 using engineered features.
- It performed closely to the Random Forest, confirming the dataset's variance limit (~58%).
- XGBoost showed strong generalization with slightly higher bias, validating tree-based models' reliability.

# CLASSIFICATION MODEL

- The tuned Random Forest Classifier achieved Accuracy = 0.677 and Macro F1 = 0.695 across three age groups.
- “Young” abalone were classified best (F1 = 0.763), while “Adult” and “Old” showed balanced trade-offs.
- The model demonstrates consistent multi-class performance, proving the regression-based insights generalize well to categorical age prediction.

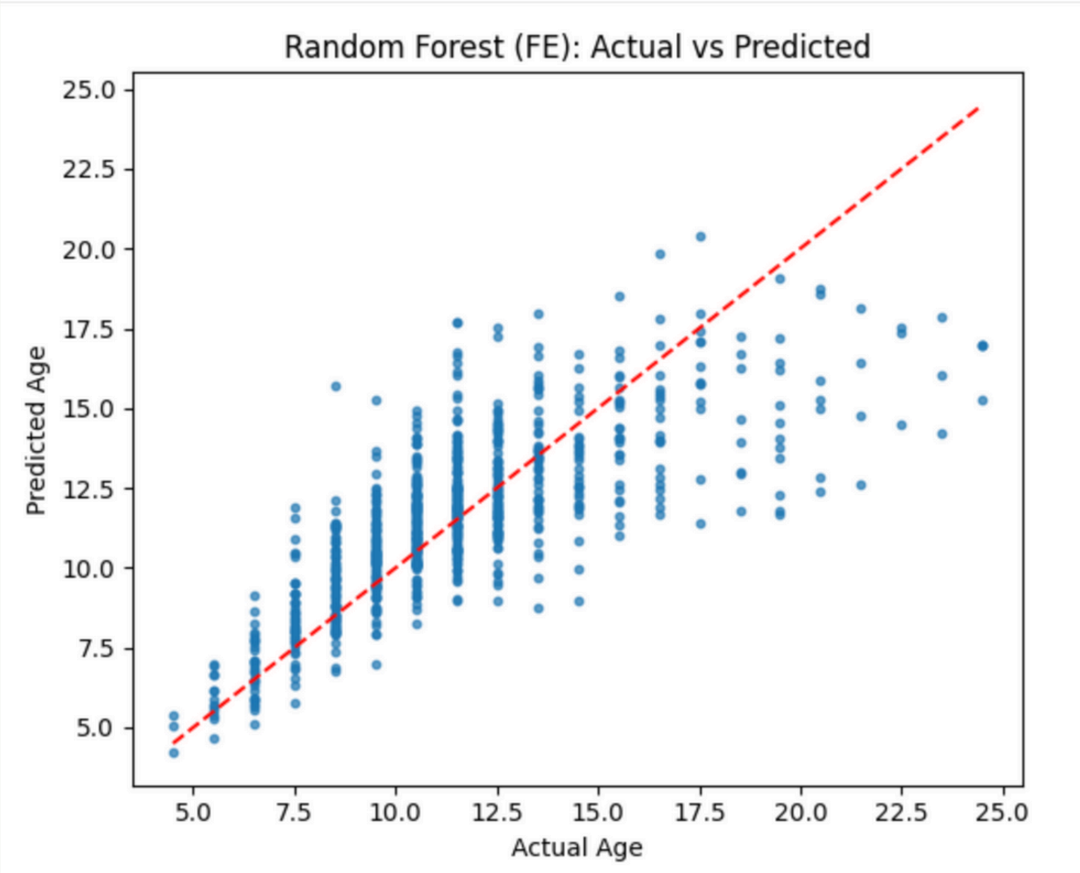



Accuracy: 0.677

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Young | 0.790 | 0.738 | 0.763 | 168 |
| Adult | 0.639 | 0.683 | 0.660 | 378 |
| Old | 0.669 | 0.634 | 0.651 | 290 |
| accuracy | | | 0.677 | 836 |
| macro avg | 0.699 | 0.685 | 0.691 | 836 |
| weighted avg | 0.680 | 0.677 | 0.678 | 836 |

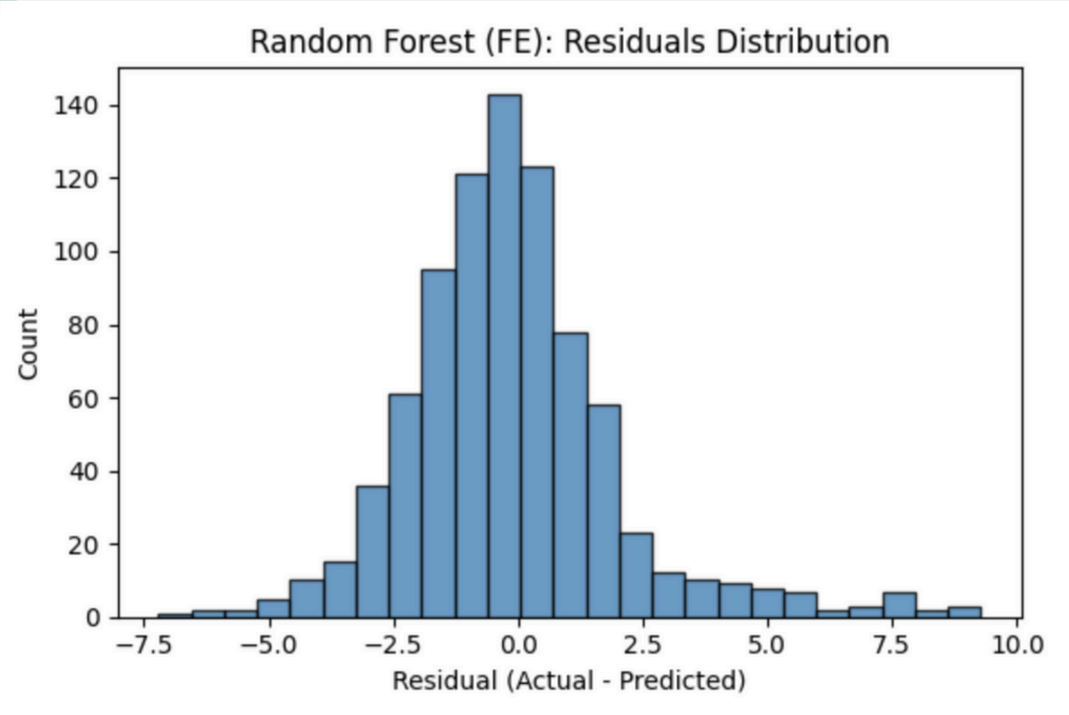
| | Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| 0 | Young | 0.790 | 0.738 | 0.763 | 168 |
| 1 | Adult | 0.639 | 0.683 | 0.660 | 378 |
| 2 | Old | 0.669 | 0.634 | 0.651 | 290 |

# RESIDUAL ANALYSIS
## post feature engineering

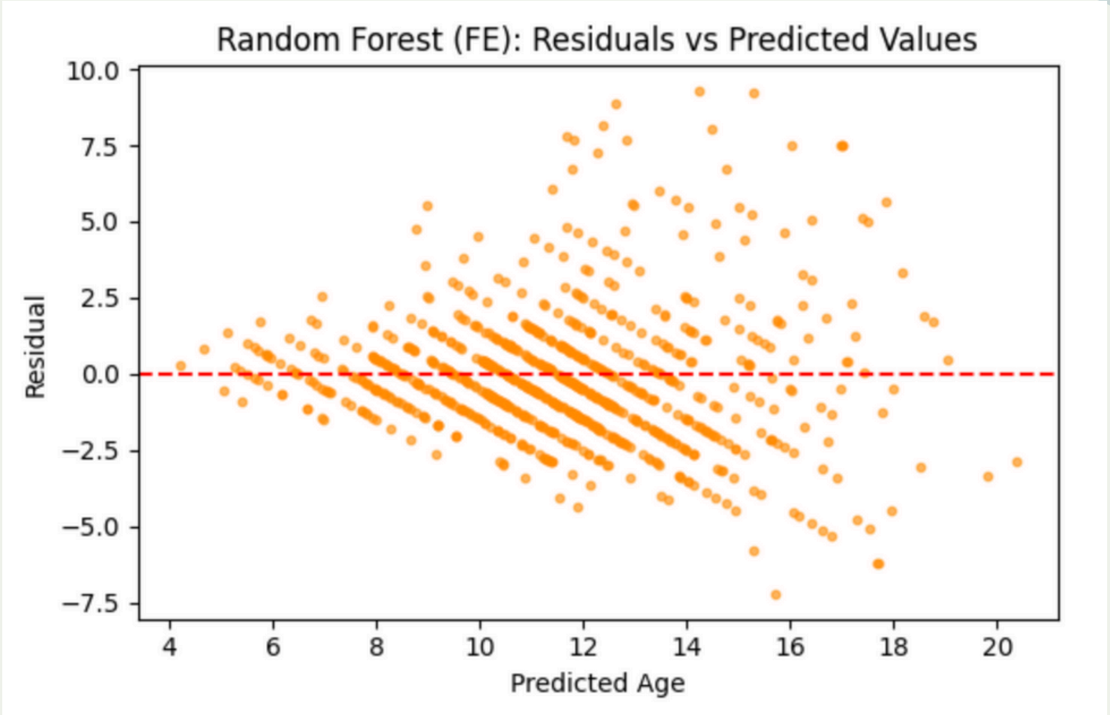



- The predicted vs actual plot shows strong alignment along the 45° line, confirming accurate regression fit.
- Slight underprediction occurs for older abalones, reflecting biological variability.
- Overall, the Random Forest model demonstrates reliable and unbiased predictions after feature engineering.

# RESIDUAL ANALYSIS
## post feature engineering




- The residuals form a near-perfect normal distribution centered at zero, confirming unbiased model predictions.
- Symmetry indicates the Random Forest neither over- nor under-predicts systematically.
- This distribution validates model stability and well-balanced error behavior across the dataset.

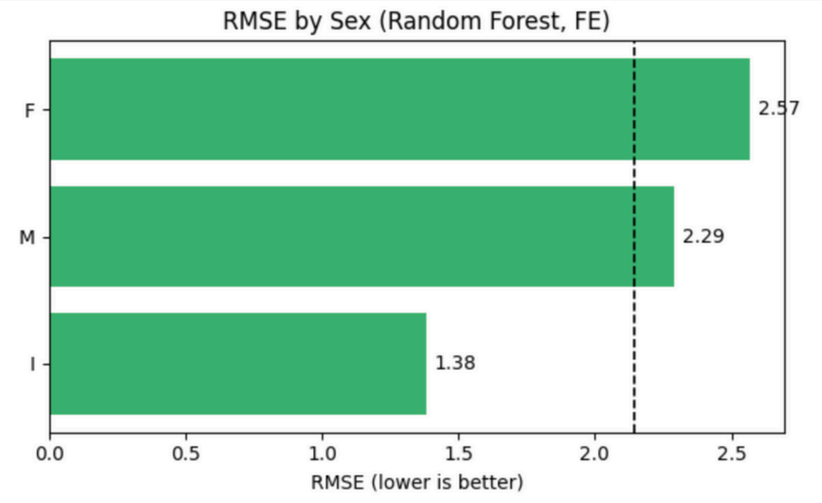


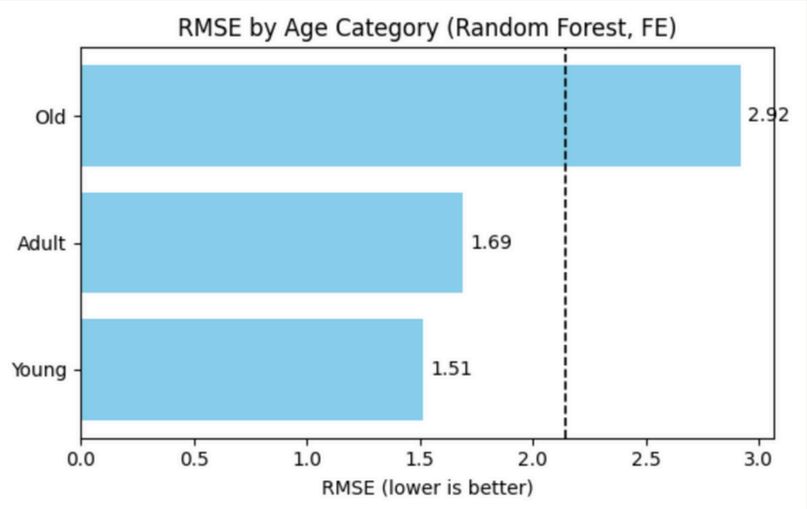

- Residuals are evenly scattered around zero, showing no clear trend or heteroscedasticity.
- This indicates that the model errors are random and independent of prediction magnitude.
- The pattern confirms good fit quality and absence of systematic bias after feature engineering.

# BIAS AND ERROR ANALYSIS
## using random forest

```
RMSE by Sex:
 sex
I    1.382319
M    2.291802
F    2.567146
Name: se2, dtype: float64
```

```
RMSE by Age Bin:
 bin
Young    1.514297
Adult    1.690714
Old      2.918152
```




- Error by sex showed best predictions for Infant abalone (RMSE = 1.38), while Female samples had the highest error (2.57).
- By age, Young (RMSE = 1.51) and Adult (1.69) were modeled more accurately than Old (2.92), indicating slight underfitting in older samples.
- Overall, results show consistent generalization with minor bias toward younger and smaller abalones.

# MODEL IMPORVEMENT

```
Ridge Regression Performance:
RMSE: 3.015
MAE : 1.665
R²  : 0.160
```

#Ridge Regression Summary

- The Ridge model achieved $R^2 = 0.160$, confirming poor fit even after regularization.
- Linear methods couldn’t capture nonlinear relationships in biological features.
- This motivated shifting to tree-based ensemble models for deeper pattern learning.

```
Best Parameters: {'model__max_depth': 10, 'model__min_samples_split': 10, 'model__n_estimators': 600}
Best CV RMSE: 2.114

Tuned Random Forest Performance:
RMSE: 2.116
MAE : 1.503
R²  : 0.586
```

# Ridge Regression Summary

- The Ridge model achieved $R^2 = 0.160$, confirming poor fit even after regularization.
- Linear methods couldn’t capture nonlinear relationships in biological features.
- This motivated shifting to tree-based ensemble models for deeper pattern learning.

```
Best XGBoost Parameters: {'model__gamma': 0.5, 'model__learning_rate': 0.05, 'model__max_depth': 3, 'model__min_child_weight': 1, 'model__n_estimators': 400}
Best XGBoost CV RMSE: 2.105

Tuned XGBoost Performance:
RMSE: 2.143
MAE : 1.521
R²  : 0.576
```

# Tuned XGBoost Summary

- The tuned XGBoost achieved RMSE = 2.143, MAE = 1.521, and $R^2 = 0.576$.
- Best parameters: max_depth = 3, n_estimators = 400, learning_rate = 0.05, gamma = 0.5.
- Nearly matched RF performance, confirming ensemble consistency and robustness.

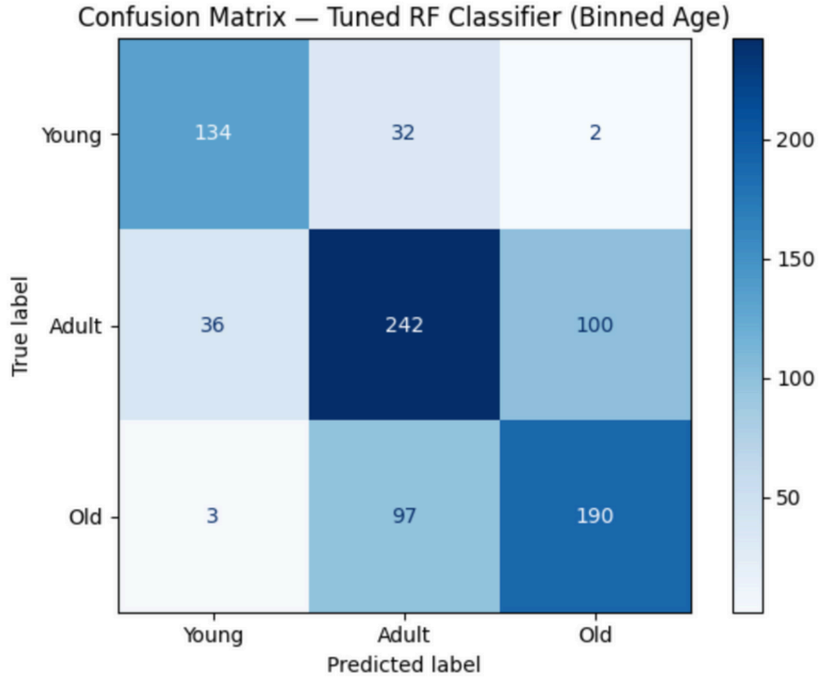
# MODEL IMPORVEMENT

```
Best Classifier Parameters: {'model__max_depth': 10, 'model__min_samples_split': 5, 'model__n_estimators': 300}
Best CV Macro-F1: 0.721

Tuned Classifier Test Performance:
Accuracy: 0.677
Macro-F1: 0.695
```

#Tuned Classifier Summary

- The tuned Random Forest Classifier (max_depth = 10, n_estimators = 300) reached Accuracy = 0.677 and Macro F1 = 0.695.
- Cross-validation improved Macro-F1 to 0.721, confirming enhanced stability and balance across age groups.
- The “Young” class maintained the highest F1 (0.786), showing clear benefit from hyperparameter tuning.




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Adult | 0.652 | 0.640 | 0.646 | 378 |
| Old | 0.651 | 0.655 | 0.653 | 290 |
| Young | 0.775 | 0.798 | 0.786 | 168 |
| accuracy | | | 0.677 | 836 |
| macro avg | 0.693 | 0.698 | 0.695 | 836 |
| weighted avg | 0.676 | 0.677 | 0.677 | 836 |

# MODEL IMPROVEMENT - cross validation

```
Random Forest 10-Fold Cross-Validation Results:
Average R² : 0.568 ± 0.047
Average RMSE: 2.096
```

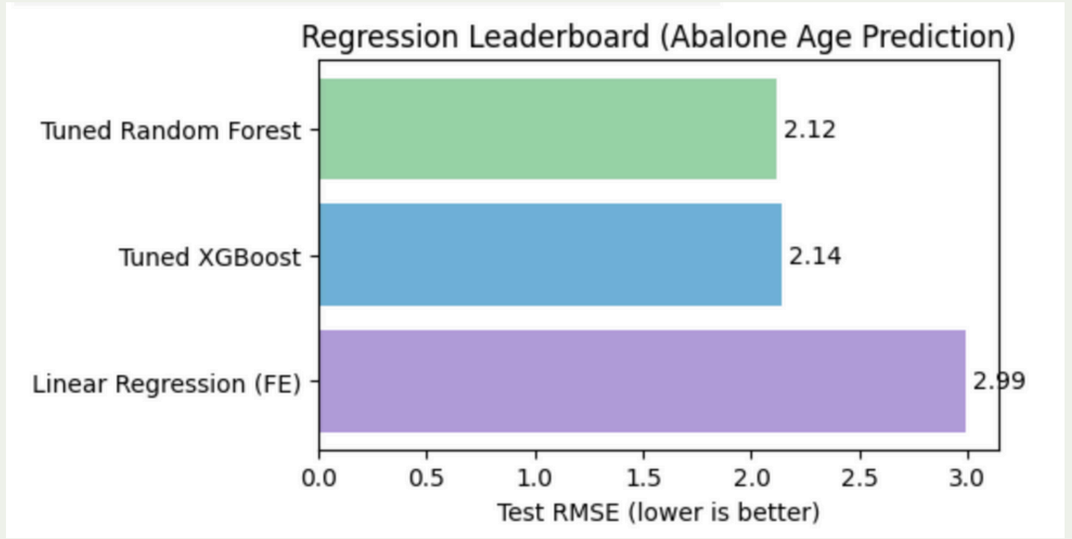
```
XGBoost 10-Fold Cross-Validation Results:
Average R² : 0.574 ± 0.047
Average RMSE: 2.081
```

- Both models showed stable generalization: RF ($R^2 = 0.568 \pm 0.047$) and XGBoost ($R^2 = 0.574 \pm 0.047$).
- Low RMSE values (~2.1) across folds confirm minimal overfitting and consistent predictive strength.
- The results validate that model performance is reliable beyond the test split, proving robustness.

# RESULTS COMPARISON

Regression Results:

| | Model | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| **0** | Tuned Random Forest | 2.116 | 1.503 | 0.586 |
| **1** | Tuned XGBoost | 2.143 | 1.521 | 0.576 |
| **2** | Linear Regression (FE) | 2.992 | 1.662 | 0.173 |




Classification Results:

| | Model | Accuracy | Macro F1 |
|---|---|---|---|
| **0** | Tuned RF Classifier (Binned Age) | 0.677 | 0.695 |

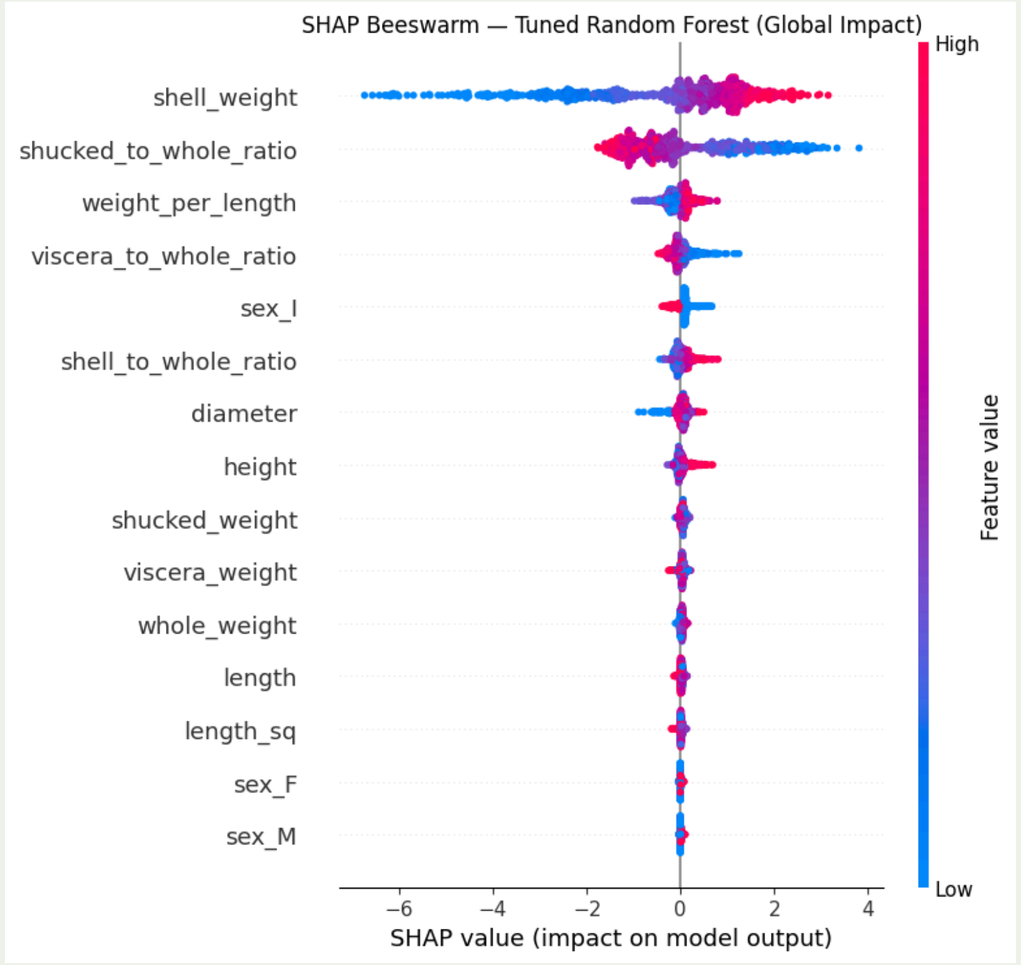
- The Tuned Random Forest achieved the best performance ($R^2$ = 0.586, RMSE = 2.116), closely followed by XGBoost ($R^2$ = 0.576).
- Linear Regression with feature engineering lagged behind ($R^2$ = 0.173), confirming nonlinearity in the data.
- The binned-age classifier (Accuracy = 0.677, F1 = 0.695) proved consistent with regression insights.

# SHAP ANALYSIS

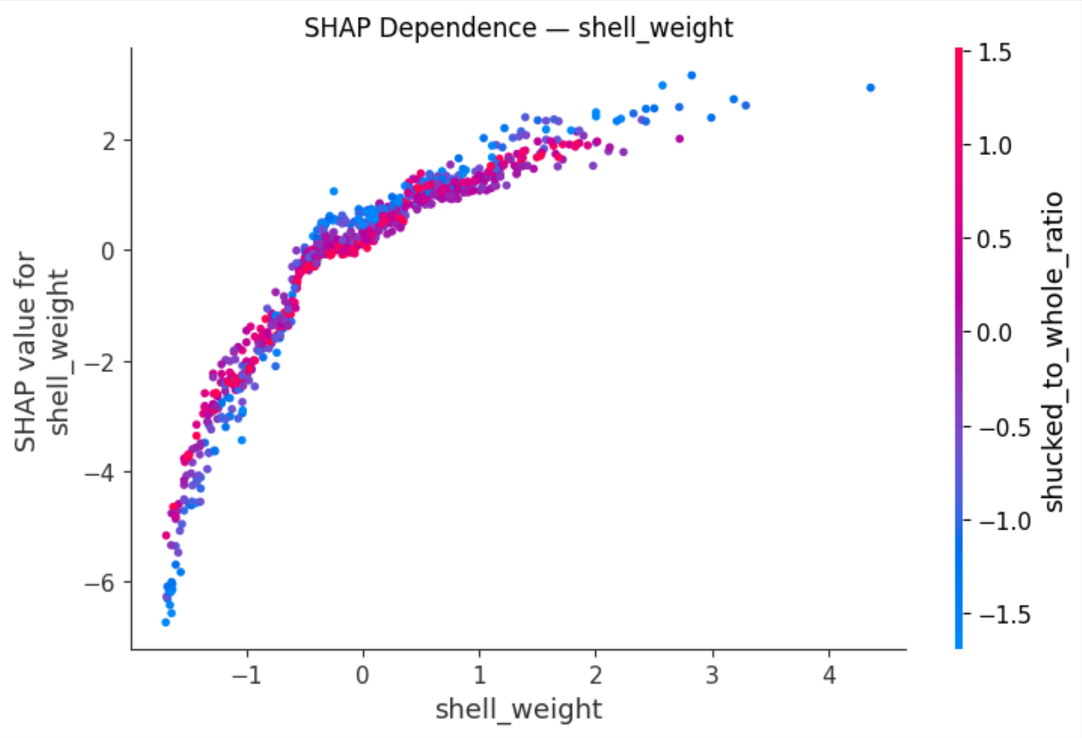
- The SHAP analysis revealed shell_weight and shucked_to_whole_ratio as the most impactful predictors.
- Weight- and ratio-based features dominated over basic size variables, reflecting biologically realistic aging trends.
- Tree-based models consistently captured these nonlinear feature effects effectively.

Top 15 features by mean |SHAP|:

| | Feature | Mean\|SHAP\| |
|---|---|---|
| 0 | shell_weight | 1.429928 |
| 1 | shucked_to_whole_ratio | 1.015536 |
| 2 | weight_per_length | 0.181017 |
| 3 | viscera_to_whole_ratio | 0.141726 |
| 4 | sex_I | 0.125550 |
| 5 | shell_to_whole_ratio | 0.107590 |
| 6 | diameter | 0.097566 |
| 7 | height | 0.081670 |
| 8 | shucked_weight | 0.054749 |
| 9 | viscera_weight | 0.051591 |
| 10 | whole_weight | 0.032339 |
| 11 | length | 0.028485 |
| 12 | length_sq | 0.024098 |
| 13 | sex_F | 0.008498 |
| 14 | sex_M | 0.007715 |



- The dependence plot showed a strong positive relationship between shell_weight and predicted age.
- Higher shucked_to_whole_ratio values amplified this effect, confirming interactive influence between shell and tissue mass.
- Overall, SHAP validated the model's focus on meaningful physical indicators rather than noise.



# Results & Interpretation

- The **tuned Random Forest model** achieved the **best overall performance ($R^2$ = 0.586, RMSE = 2.116),** with XGBoost performing similarly.
- Classification results reached **Accuracy = 0.677** and **Macro-F1 = 0.695**, showing **balanced class predictions**.
- Cross-validation confirmed strong generalization and stability across data folds.
- SHAP analysis revealed that **shell weight and tissue-to-shell ratios** are the s**trongest predictors of age.**
- The model **learned biologically meaningful patterns,** focusing on shell composition rather than size alone.
- Overall, ensemble models effectively captured the **non-linear relationship** between abalone features and age, achieving both accuracy and interpretability.

# Limitations & Future Work

- Dataset includes only shell measurements (no environmental data).
- Age label (Rings + 1.5) is an approximation.
- Small, region-specific dataset limits generalization.
- Linear models struggled with strong non-linear patterns.
- Slight class imbalance affected classification accuracy.

- Expand dataset with environmental and regional factors.
- Use advanced tuning (Bayesian / Randomized search).
- Test additional models (LightGBM, CatBoost, Neural Nets).
- Explore stacked or hybrid regression–classification models.
- Deepen explainability with SHAP and permutation analysis.
- Validate on new regional or temporal datasets.

# REFERENCES


Chai, Zhilei, et al. "Taking Advantage of Hybrid Bioinspired Intelligent Algorithm with Decoupled Extended Kalman Filter for Optimizing Growing and Pruning Radial Basis Function Network." Royal Society Open Science, vol. 5, no. 9, 2018, p. 180529. https://doi.org/10.1098/rsos.180529.t

Sahu, Anit Kumar, Dusan Jakovetić, and Soummya Kar. "Communication Optimality Trade-offs for Distributed Estimation." arXiv, 12 Jan. 2018, Chai, Zhilei, et al. "Taking Advantage of Hybrid Bioinspired Intelligent Algorithm with Decoupled Extended Kalman Filter for Optimizing Growing and Pruning Radial Basis Function Network." Royal Society Open Science, vol. 5, no. 9, 2018, p. 180529. https://doi.org/10.1098/rsos.180529.t

Rad, Radin Hamidi, and Maryam Amir Haeri. "Hybrid Forest: A Concept Drift Aware Data Stream Mining Algorithm." arXiv, 10 Feb. 2019, https://doi.org/10.48550/arXiv.1902.03609

# Thank You