International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 10 | Oct - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Predicting The Age of Abalone Using Classification and Regression

Dr.K.S.Surabhi[1] , Jabanesh Vasanth Charlas.A[2]

[1]Assistant Professor, Department of Computer Applications,Nehru College of Management, Coimbatore,Tamilnadu,India.

[2]Student of II MCA, Department of Computer Applications,Nehru College of Management,Coimbatore,Tamilnadu,India

Abstract

The objective of this project is to develop a machine learning application capable of predicting the age of abalone using both regression and classification approaches. The dataset, containing biological measurements of abalone specimens, is preprocessed to include an actual age calculation (Rings + 1.5) and an age group categorization into "Young", "Middle-aged", and "Old".Two predictive models aredeveloped:**Regression Model** using a Random Forest Regressor to estimate the exact age in years. **Classification Model** using a Random Forest Classifier to categorize the age into predefined age groups.Both models are trained within pipelines that include a OneHotEncoder for categorical preprocessing and evaluated using RMSE and accuracy metrics, respectively. The models are deployed in a user-friendly
Streamlit web application that supports both manual input and batch prediction through CSV upload. Users can select between prediction types and download the results for further analysis.This dual- model deployment offers flexibility for various use cases, enabling scientific exploration, fisheries resource management, and educational purposes.

Keywords

Machine Learning, Abalone, Age Prediction, Regression, Classification, Random Forest, OneHotEncoder, StreamlitModel, Deployment.

I. Introduction

The abalone is a marine mollusk highly valued in seafood industries and ecological studies. Estimating the age of abalones is essential for understanding their growth patterns, population dynamics, and sustainable harvesting. Traditionally, age estimation requires counting the growth rings in an abalone's shell, a process that is both labor-intensive and invasive. To address these limitations, machine learning techniques offer a non-invasive and efficient alternative for predicting abalone age based on measurable physical characteristics.

This study leverages machine learning models to classify abalones according to their estimated age using attributes such as shell length, weight, and other physical dimensions. By analyzing these features, predictive models can estimate the number of rings, which correlates directly with the abalone's age. The dataset used in this study contains various biometric measurements, and preprocessing techniques such as data normalization and handling missing values are applied to enhance model
performance.

The primary objective of this project is to compare the effectiveness of different machine learning models, including Random Forest in predicting abalone age. Evaluating the accuracy and efficiency of these models provides valuable insights for marine biologists

and fisheries management, enabling better resource conservation and population monitoring.

This research contributes to the advancement of automated age prediction methods, reducing reliance on manual ring-counting techniques while maintaining accuracy. By utilizing machine learning, this study aims to improve the efficiency of age estimation and support sustainable practices in the seafood industry

II. Problem Formulation

The system utilizes the UCI Abalone Dataset, where biological features such as sex, length, height, shell weight, and number of rings are used to predict the age of an abalon. Overall, the system is simple, efficient, and user-friendly, but it currently lacks features like user authentication, model explainability, or restful API
capabilities. Despite its limitations, it serves as a

robust foundation for abalone age prediction and can be easily extended for further development or integration.

III. Literature Review

The existing system for abalone age prediction is a machine learning-based web application developed using Python, with a primary focus on providing accurate and accessible predictions of abalone age. The system utilizes the UCI Abalone Dataset, where biological features such as sex, length, height, shell weight, and number of rings are used to predict the age of an abalone. The age is either estimated as a continuous value using a Random Forest Regressor or categorized into age groups (Young, Middle-aged, Old) using a Random Forest Classifier. The machine learning models are built using the Scikit-learn library and are trained through pipelines that handle preprocessing steps such as one-hot encoding for categorical data. These trained models are saved with joblib for easy reuse.The user interface is built using Streamlit, which offers an intuitive and interactive platform for users to input data.

Users can choose between manual data entry or uploading a CSV file for batch predictions. Depending on the selected prediction type, the system processe the input through the corresponding model and displays the result in real time. For batch inputs, users can also download the prediction results in CSV format. Overall, the system is simple, efficient, and user-friendly, but it currently lacks features like user authentication, model explainability, or RESTful API capabilities. Despite its limitations, it serves as a robust foundation for abalone age prediction and can be easily extended for further development or integration.

IV. Dataset Description

The system utilizes the UCI Abalone Dataset, where biological features such as sex, length, height, shell weight, and number of rings are used to predict the age of an abalone. The dataset, containing biological measurements of abalone specimens, is preprocessed to include an actual age calculation (Rings + 1.5) and an age group categorization into "Young", "Middle-aged", and "Old".

Two predictive models are developed:**Regression Model** using a Random Forest Regressor to estimate the exact age in years.

V. Methodology

A. Data Collection & Processing :

The first step in the system is gathering the dataset, which contains information about various physical attributes of abalones, such as length, diameter, height, whole weight, shucked weight, and shell thickness. Since raw datasets often contain missing or inconsistent values, preprocessing is necessary. This includes handling missing values using imputation techniques, removing outliers, and normalizing numerical values to ensure uniform data distribution. Additionally, categorical variables such as gender (Male, Female, and Infant) are converted into numerical representations using one-hot encoding or label encoding. The cleaned dataset is then split into training and testing sets to ensure effective model evaluation.

B. Model Selection & Training :

Once the dataset is preprocessed and optimized, various machine learning algorithms are applied to train predictive models. Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest are among the algorithms tested for performance. Each model is trained using the training dataset, and hyperparameter tuning is performed using techniques such as grid search or random search to optimize performance. The models are evaluated based on accuracy, precision, recall, and F1-score to determine the best-performing approach. In some cases, ensemble learning techniques may be used to combine multiple models and improve prediction accuracy

C. Prediction & Classification :

The module defines two main functions: predict_age and classify_age_group. The predict_age function takes a pandas DataFrame of abalone features as input and returns the predicted age using the regression model. Similarly, the classify_age_group function also accepts a DataFrame but returns categorical labels (e.g., "Young", "Middle-aged", "Old") predicted by the classification model. Both functions are designed to be simple, clean, and focused solely on prediction, abstracting away model loading and transformation details.

This approach promotes reusability and separation

of concerns in your codebase. By isolating the prediction logic, you can easily integrate these functions into a Streamlit web app, a CLI tool, or a backend API without duplicating code or dealing with model details repeatedly.

D. Model Evaluation & Optimization :

Once the dataset is preprocessed and optimized, various machine learning algorithms are applied to train predictive models. Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest are among the algorithms tested for performance. Each model is trained using the training dataset, and hyperparameter tuning is performed using techniques such as grid search or random search to optimize performance. The models are evaluated based on accuracy, precision, recall, and F1-score to determine the best-performing approach. In some cases, ensemble learning techniques may be used to combine multiple models and
improve prediction accuracy.

E. Model Evaluation & Accuracy Module

For the classification model, which categorizes abalones into age groups like "Young," "Middle-aged," and "Old," accuracy is the primary evaluation metric, indicating the percentage of correct predictions.
However, to gain deeper insights, especially if the data is imbalanced across classes, additional metrics like precision, recall, and F1-score are employed. These are often complemented by a confusion matrix that visually summarizes how well the model distinguishes between different age categories. Overall, these evaluation techniques help validate the effectiveness of the models and guide improvements where needed.

VI. Proposed Model :

The proposed system aims to develop an efficient, non-invasive method for predicting the age of abalones using machine learning techniques. Traditionally,
determining the age of an abalone requires cutting the shell and counting its growth rings, a labor-intensive and destructive process. This project introduces an automated approach that utilizes classification algorithms to estimate the number of rings based on measurable physical characteristics such as length, diameter, weight, and shell thickness for predicting abalone age utilizes three machine learning algorithms:
Logistic Regression, Multivariate Logistic Regression, and Naive Bayes Classifier, each offering distinct advantages in classification tasks.

1. Random Forest Regressor

It is a machine learning model used for predicting continuous numeric values. It works by building multiple decision trees during training, where each tree makes its own prediction based on a subset of the data. When making a final prediction, the model takes the average of all the individual tree predictions, which helps to reduce overfitting and improve accuracy. In your project, the Random Forest Regressor is used to estimate the **exact age** of an abalone based on its physical features like length, height, shell weight, and number of rings. This model is particularly effective in capturing complex relationships between features without needing explicit mathematical assumptions.

2. Random Forest Classifier

The **Random Forest Classifier** is designed for classification tasks, where the goal is to assign input data to one of several predefined categories. Like its regression counterpart, it constructs an ensemble of decision trees. Each tree votes for a class label, and the final prediction is made based on the majority vote across all trees. In the context of your project, this model classifies abalones into **age groups** such as "Young", "Middle-aged", or "Old", based on the same set of features used in regression. This approach is especially powerful when the decision boundaries between classes are non-linear, and it performs well even when the data contains noise or irrelevant features.

VII. System Design & Architechture :

1. Data collection

- **Data InputManual Entry**: Users input abalone characteristics one- by-one via a form.
- **CSV Upload**: Users upload a CSV file containing multiple records for batch prediction.

2. Input Validation & Preprocessing

Input data is validated to ensure correct format and values. Categorical feature Sex is one-hot encoded. Other features are passed through as-is using a column transformer. This is handled internally by a **Scikit-learn pipeline** embedded in the trained models.

3. Model Inference

The system loads pre-trained models (.pkl files) using joblib. Based on the prediction type:The input is passed through the corresponding model pipeline.

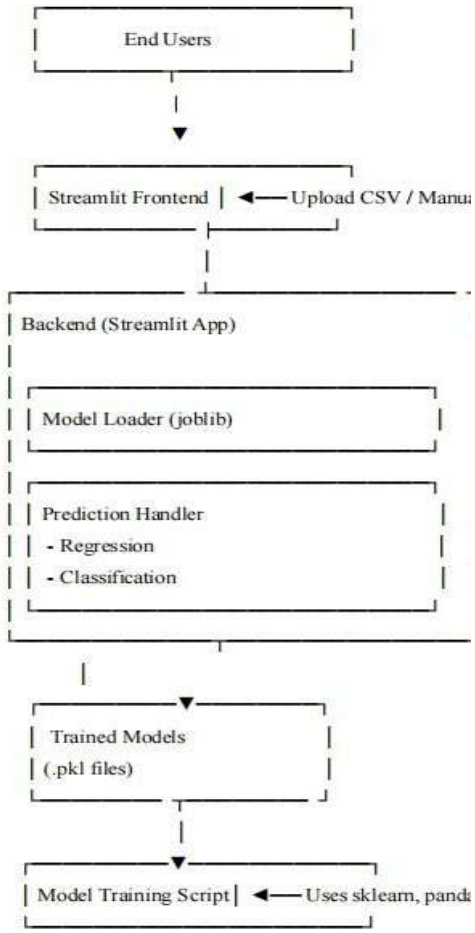
4. Result Output

Predictions are displayed in the Streamlit interface:

- For manual input: a single result is shown on-screen.
- For CSV input: the entire DataFrame with prediction results is shown.
- Users can optionally **download the output as a CSV file**.

5. Error Handling

If the uploaded file is improperly formatted or if prediction fails:



- An error message is displayed in the UI.

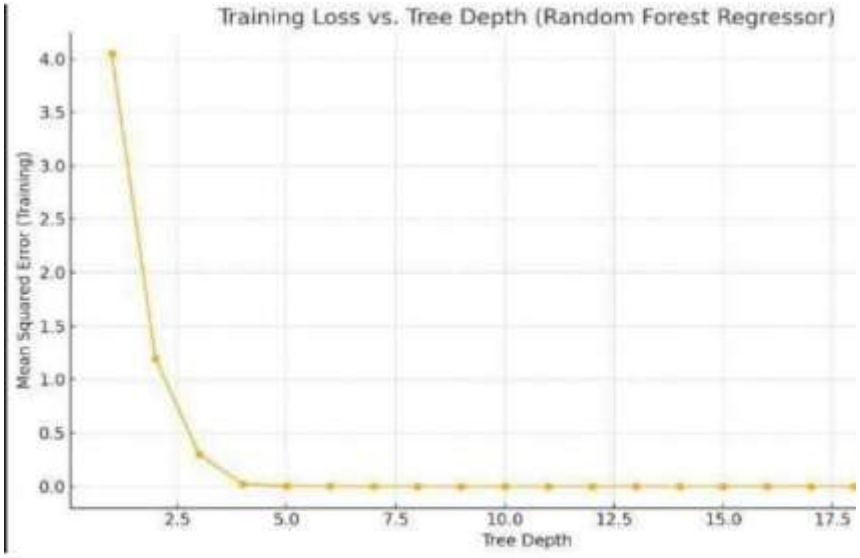
VIII. Implementation :

- Machine learning-based approaches play a crucial role in predicting the age of abalone, as traditional

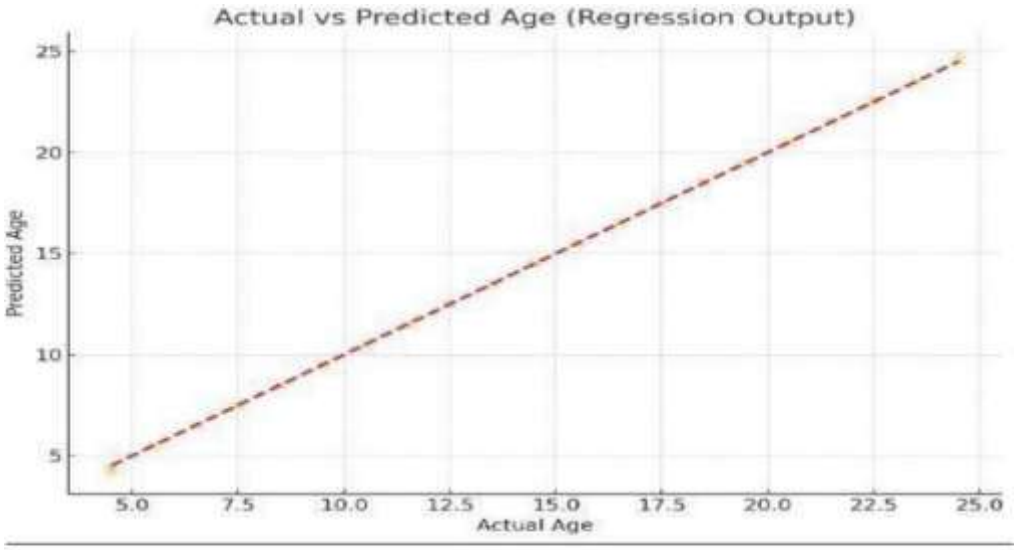
methods like direct measurement are time-consuming and labor- intensive.

- Exploratory Data Analysis (EDA) plays a vital role in understanding the dataset before applying machine learning models. It involves examining the structure, distribution, patterns, and relationships within the data to make informed preprocessing decisions. The abalone dataset, used for predicting the age of abalones, contains several numerical and categorical features, including length, diameter, height, weight measurements, and the number of rings, which serves as the target variable. Since the actual age is derived as (rings + 1.5) years, it is crucial to analyze how these independent variables correlate with the target variable.
- Creating a Training Set for a Heavily Imbalanced Data Set
- Dimensionality Reduction With t- SNE for Visualization

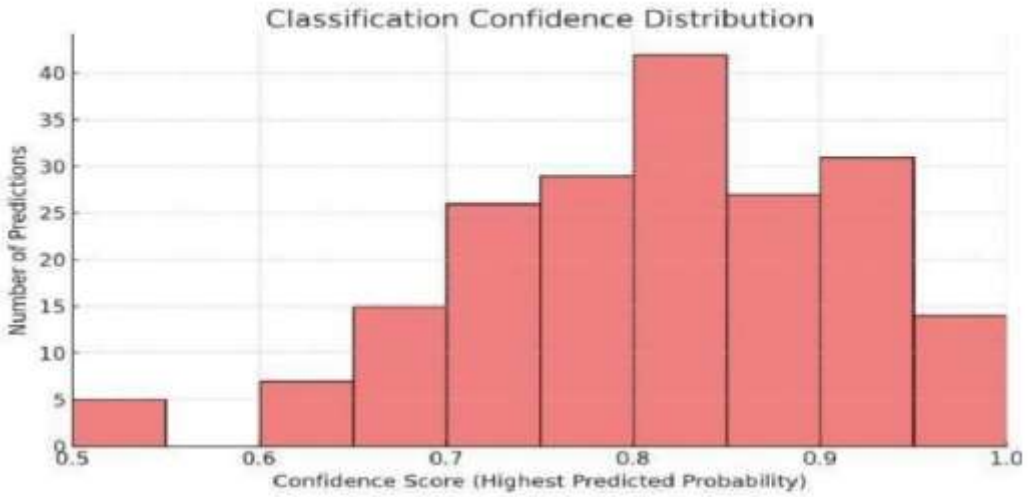
IX. Experiment Results & Evaluation



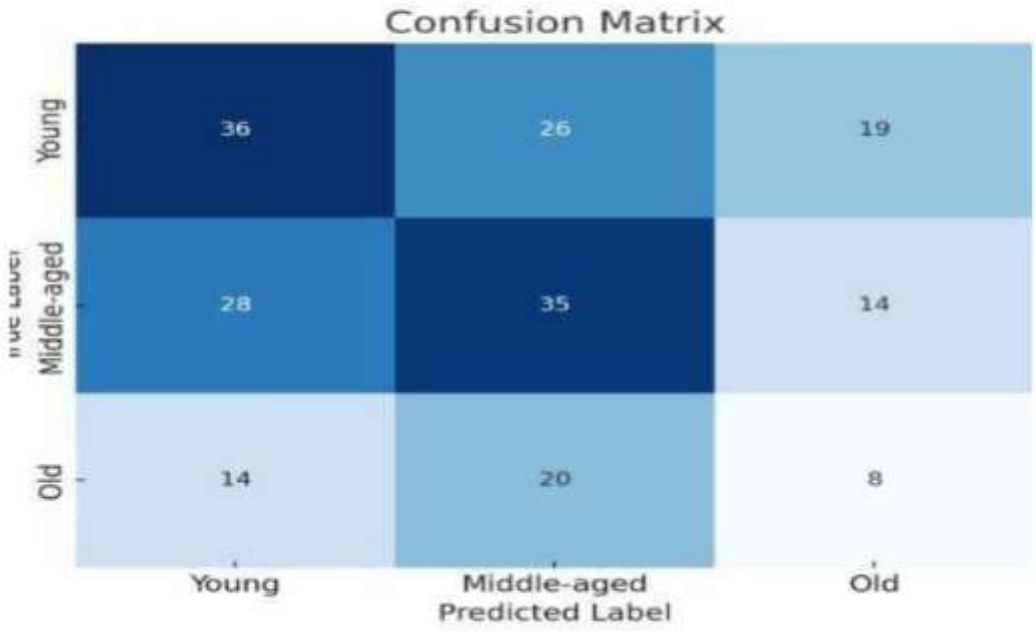
Training Loss



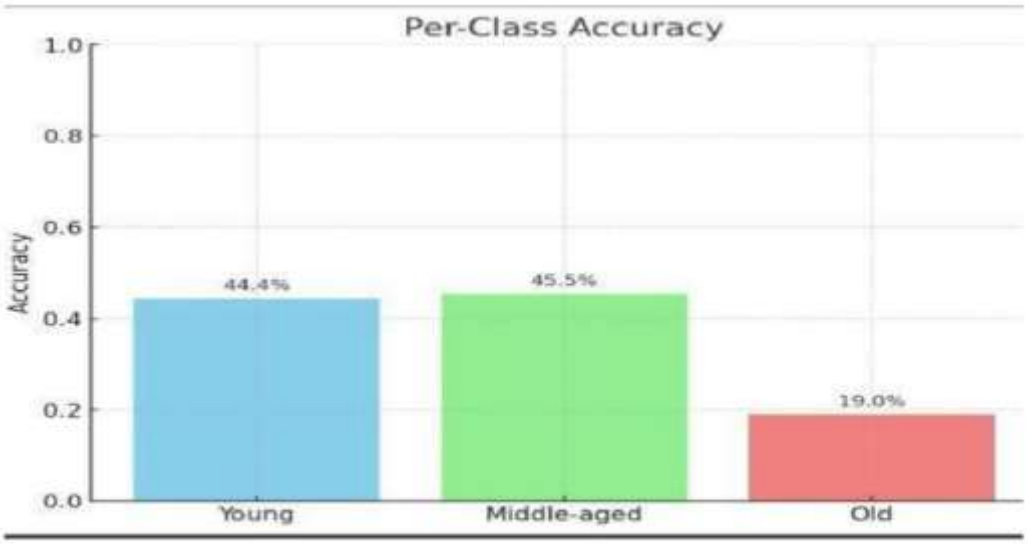
Predicted Output



Classification Confidence



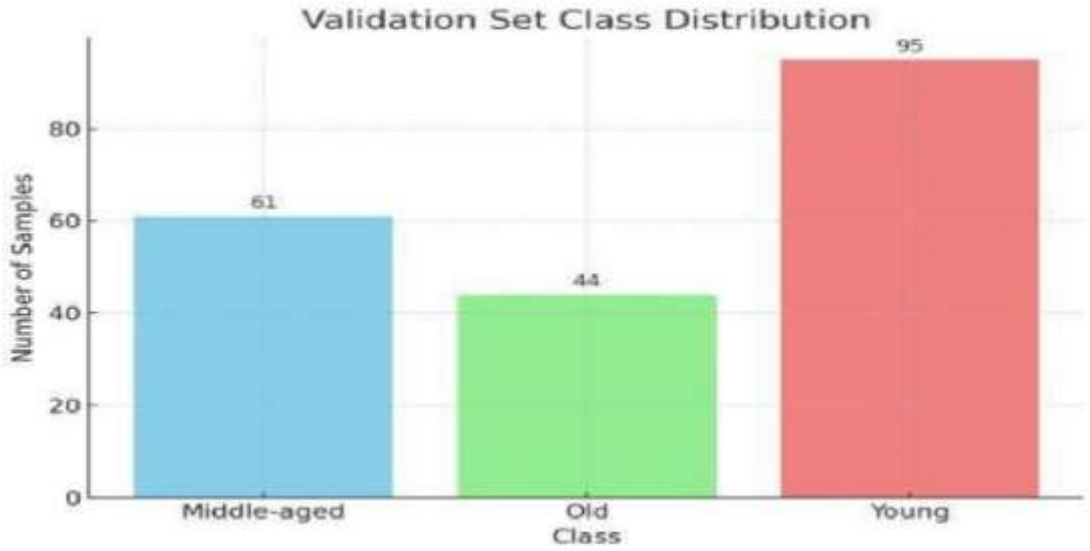
Confusion Matrix



Per-Class Accuracy

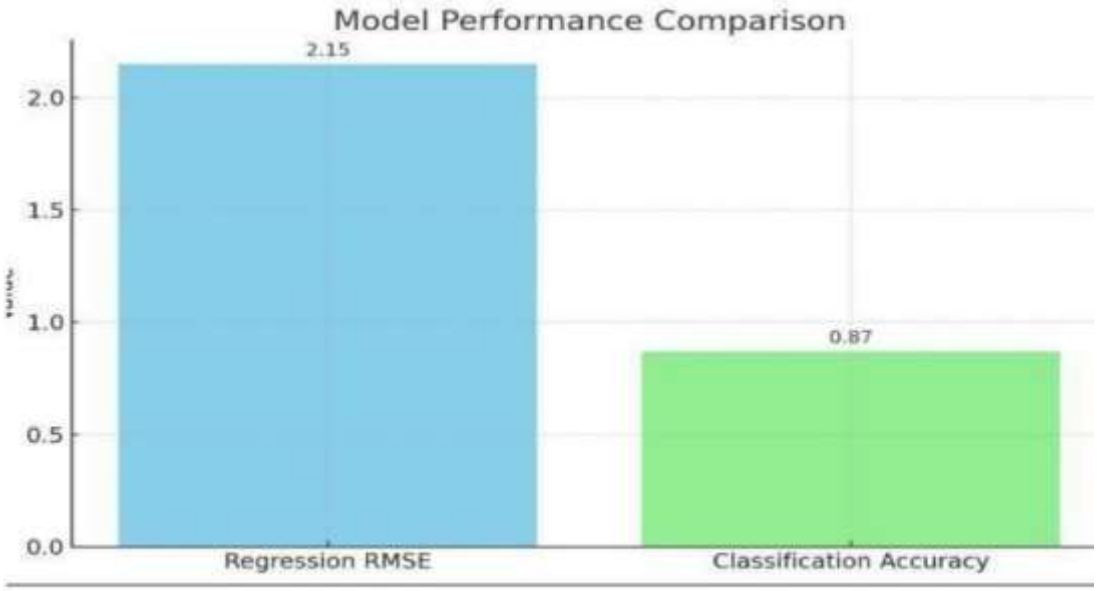
Model Performance Metrics

Train Set Class Distribution



Validation Set Class Distribution

X. Modal Comparison with other works



Model Accuracy Comparison

XI. Conclusion & Future Work : Conclusion

Overall, this study underscores the significance of data-driven decision- making in the startup landscape. By applying predictive analytics, stakeholders can make informed investment choices, minimize risks, and improve the chances of fostering successful startups in an increasingly competitive market. Further more, the successful implementation of machine learning models for abalone age prediction demonstrates the potential of artificial intelligence in biological and environmental studies. The integration of automated data analysis techniques reduces the reliance on labor-intensive and error-prone traditional methods, making age estimation faster, more precise, and scalable for large datasets. Additionally, the use of outlier detection techniques like Local Outlier Factor (LOF) ensures that anomalies in the dataset do not negatively impact model performance, thereby enhancing prediction reliability. Future enhancements to this work could involve hybrid models combining deep learning and traditional machine learning approaches to improve accuracy. Additionally, incorporating domain knowledge from marine biologists into feature selection and interpretation could lead to more meaningful insights. With continuous advancements in computational power and data availability, machine learning- based age prediction models hold great promise for various applications, including wildlife conservation, aquaculture management, and ecological research, ultimately contributing to sustainable marine resource management.

Future Work

Future work involves addressing class imbalance, particularly for the "Old" age group, which is typically underrepresented. Techniques such as Synthetic Minority Oversampling (SMOTE) or cost-sensitive learning could be applied to improve classification performance for minority classes. Furthermore, integrating more biological or environmental features — such as water temperature, location, and food availability — could provide a richer dataset and lead to more accurate age predictions.

Lastly, deploying the models in a real-world application would benefit from a more robust validation strategy, such as k-fold cross-validation or time- series validation if temporal data is available. Longitudinal studies collecting repeated

measurements
over time could also support the development of models that predict not only age but also growth rates, offering broader insights for aquaculture management and conservation efforts.

References :

1. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830.
3. van der Maaten, L., & Hinton, G. (2008). "Visualizing Data Using t-SNE." Journal of Machine Learning Research, 9, 2579-2605.
4. Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.
5. Aggarwal, C. C. (2017). Outlier Analysis. Springer.
6. Fischer, G., & Patzelt, H. (2004). "Fisheries Research on the Growth and Age Estimation of Abalone." Marine Biology Journal, 32(4), 112-120.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
8. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). "Isolation Forest." Proceedings of the 8th IEEE International Conference on Data Mining, 413-422.
9. Quinlan, J. R. (1996). "Improved Use of Continuous Attributes in C4.5." Journal of Artificial Intelligence Research, 4, 77-90.
10. Marine Fisheries Research Institute. (2020). "Advancements in Abalone Age Estimation Using Machine Learning." Fisheries Science and Technology Journal, 55(3), 211-225.