Project Title: Abalone Age Prediction using Machine Learning

Date	**2nd OCT 2025**
Team ID	**LTVIP2025TMIDS67772**
Project Name	**Abalone Age Prediction using Machine Learning**
Max Marks	**6 Marks**

Data Collection and Preprocessing Phase

Data Preprocessing

1. Introduction

Data preprocessing is one of the most critical stages in any machine learning project. It ensures that the dataset used for training and testing the model is clean, consistent, and suitable for predictive analysis.

In the **Abalone Age Prediction** project, the dataset contains measurements of abalone shells and related attributes. The main objective of this phase is to **prepare the data** for model training by handling missing values, encoding categorical features, scaling numerical values, and dividing the dataset into training and testing subsets.

This process enhances the **accuracy**, **efficiency**, and **reliability** of the final model.

2. Dataset Description

The project uses the **Abalone Dataset** from the **UCI Machine Learning Repository**, a well-known public dataset commonly used for regression tasks.

Dataset Overview

Attribute	Description	Data Type
Sex	Gender of abalone (M: Male, F: Female, I: Infant)	Categorical
Length	Longest shell measurement (in mm)	Continuous
Diameter	Perpendicular to length (in mm)	Continuous
Height	Height with meat in shell (in mm)	Continuous
Whole Weight	Whole abalone weight (in grams)	Continuous
Shucked Weight	Weight of meat only (in grams)	Continuous
Viscera Weight	Gut weight after bleeding (in grams)	Continuous
Shell Weight	Weight after drying shell (in grams)	Continuous
Rings	Number of rings; used to determine age (Age = Rings + 1.5)	Integer

Dataset Size: 4,177 samples × 9 attributes

Target Variable: Rings (converted to Age = Rings + 1.5)

3. Data Import and Inspection

The dataset (abalone.csv) was imported using the **Pandas** library in Python:

```
import pandas as pd

df = pd.read_csv("abalone.csv")
```

Initial inspection using:

```
df.info()

df.describe()

df.head()
```

helped identify data types, missing values, and outlier patterns.

4. Data Cleaning Steps

Step 1: Handling Missing Values

- Checked for missing or null entries using:
- df.isnull().sum()
- Result: **No missing values** detected.
  Hence, the dataset is **complete and ready** for processing.

Step 2: Encoding Categorical Variables

- The Sex column contains non-numeric values (‘M’, ‘F’, ‘I’) that need to be encoded.
- Applied **One-Hot Encoding** to convert categorical data into numerical format:
- df = pd.get_dummies(df, columns=['Sex'], drop_first=True)
- New columns generated: Sex_F, Sex_I, Sex_M (binary encoded).

Step 3: Feature Scaling

Since the dataset contains measurements with different scales (mm and grams), scaling was applied to standardize all features using **StandardScaler** from Scikit-learn.

```
from sklearn.preprocessing import StandardScaler


scaler = StandardScaler()
scaled_features = scaler.fit_transform(df.drop('Rings', axis=1))
```

This ensures uniform contribution of each variable during model training.

Step 4: Target Variable Transformation

- The target column Rings was converted to **Age** using the standard formula:
- Age = Rings + 1.5
- This transformation provides a more realistic biological representation of abalone age.

Step 5: Outlier Detection

Outliers were checked visually using box plots for numerical features (Length, Diameter, Height, etc.). Minor outliers were retained as they represent **natural biological variation** rather than data errors.

5. Data Splitting

To ensure unbiased model evaluation, the dataset was split into training and testing subsets.

```
from sklearn.model_selection import train_test_split


X = df.drop('Age', axis=1)

y = df['Age']


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

- **Training Data:** 80% (3,341 samples)
- **Testing Data:** 20% (836 samples)

6. Final Preprocessed Dataset Summary

Property	Details
Total Records	4,177
Numeric Features	8 (after encoding)
Categorical Features	1 (Sex → One-hot encoded)
Missing Values	None
Scaled	Yes (StandardScaler)
Target Variable	Age = Rings + 1.5

7. Tools and Libraries Used

Library	Purpose
Pandas	Data import and manipulation
NumPy	Numerical computation
Matplotlib / Seaborn	Visualization for outlier detection
Scikit-learn	Encoding, scaling, and splitting data

8. Results of Preprocessing

- The dataset is **clean, complete, and numerically encoded**.
- Features are **scaled and normalized** for optimal model performance.
- Dataset is **divided** for unbiased model training and testing.

This pre-processed dataset forms the **foundation** for the next phase — model development and evaluation.

9. Conclusion

The preprocessing phase successfully prepared the Abalone dataset for machine learning modeling. All required transformations — encoding, scaling, and splitting — were completed without any data loss.
The dataset is now in a standardized format, ensuring accurate and efficient training of predictive models in the subsequent phase.