SMARTBRIDGE
Let's Bridge the Gap

Smart Internz

Project Title: Abalone Age Prediction using Machine Learning

Date	**3rd OCT 2025**
Team ID	**LTVIP2025TMIDS67772**
Project Name	**Abalone Age Prediction using Machine Learning**
Max Marks	**2 Marks**

Data Collection and Preprocessing Phase

Data Quality Report

1. Introduction

Data quality is the foundation of any reliable machine learning project. Poor data quality directly impacts model accuracy, decision-making, and overall system performance. In the **Abalone Age Prediction** project, ensuring the quality of data is essential because small inaccuracies in physical measurements (like shell length, weight, or height) can lead to significant prediction errors.

This phase focuses on assessing the **integrity, completeness, consistency, and validity** of the Abalone dataset, ensuring it meets the standards required for accurate machine learning predictions.

2. Importance of Data Quality

High-quality data ensures that:

- Models learn meaningful patterns instead of noise.

- Predictions are consistent and trustworthy.
- The system generalizes well to unseen data.

In the context of abalone age estimation, maintaining accurate and clean data is crucial because:

- Biological datasets often contain natural variations.
- Inaccurate readings can mislead model learning.
- Uniform feature representation enhances model interpretability.

3. Dataset Description

The dataset used for this analysis is the **Abalone Dataset** obtained from the **UCI Machine Learning Repository**. It consists of **4,177 samples** and **9 attributes** (8 features and 1 target variable).

Attribute	**Description**	**Type**
Sex	Gender of abalone (M, F, I)	Categorical
Length	Longest shell measurement (mm)	Continuous
Diameter	Perpendicular to length (mm)	Continuous
Height	Height with meat in shell (mm)	Continuous
Whole weight	Weight of whole abalone (grams)	Continuous
Shucked weight	Weight of meat (grams)	Continuous
Viscera weight	Gut weight after bleeding (grams)	Continuous
Shell weight	Weight after drying shell (grams)	Continuous
Rings	Number of rings (used to estimate age)	Integer

Target Variable: Age = Rings + 1.5

4. Data Quality Dimensions

The dataset was evaluated across six major dimensions of data quality:

Dimension	Definition	Evaluation Process	Findings
Accuracy	Data correctly represents real-world measurements.	Compared feature values to biological standards for abalone size and weight.	All values within realistic biological limits.
Completeness	The degree to which all required data is available.	Checked for missing or null values.	No missing values detected — dataset 100% complete.
Consistency	Uniform representation and logical consistency.	Checked for duplicate or contradictory records.	Dataset consistent, no duplicates found.
Validity	Compliance with defined data formats and ranges.	Verified that all numerical features are positive and within logical bounds.	All values valid and well-structured.
Timeliness	Data's relevance and applicability over time.	Verified dataset source credibility and continued relevance.	Dataset remains valid for modeling as biological data does not expire.
Uniqueness	Ensuring that each record is distinct.	Checked for duplicate entries using Pandas .duplicated().	No duplicate rows found.

5. Data Quality Assessment Techniques

To evaluate the above dimensions, the following techniques were used:

5.1 Missing Value Analysis

- Checked using:
- df.isnull().sum()
- Result: No missing values detected.
- Conclusion: Dataset is 100% complete.

5.2 Duplicate Record Detection

- Checked using:
- df.duplicated().sum()
- Result: 0 duplicates found.
- Conclusion: Dataset is unique and clean.

5.3 Data Type Consistency

- Ensured each feature had a valid data type:
    1. Numeric: float64 or int64
    2. Categorical: object
- No mismatched data types found.

5.4 Range Validation

Compared minimum and maximum values of each numeric column to standard biological references for abalone characteristics:

Feature	Min	Max	Expected Range	Status
Length	0.075	0.815	0.05 – 1.0	Valid
Diameter	0.055	0.65	0.05 – 0.8	Valid
Height	0.00	1.13	0.00 – 1.2	Valid
Whole Weight	0.002	2.8255	0.00 – 3.0	Valid
Shucked Weight	0.001	1.488	0.00 – 2.0	Valid
Viscera Weight	0.0005	0.76	0.00 – 1.0	Valid
Shell Weight	0.0015	1.005	0.00 – 1.5	Valid

All values are within realistic biological limits.

6. Summary of Data Quality Evaluation

Quality Aspect	Status	Remarks
Completeness	Excellent	No missing values
Accuracy	High	Consistent with real measurements
Consistency	Strong	Uniform data structure
Validity	Verified	Values within logical limits
Timeliness	Satisfactory	Still relevant and widely used
Uniqueness	Maintained	No duplicates detected

7. Data Quality Challenges (Observed & Resolved)

Issue Identified	Impact	Resolution
Mixed data types in some columns (object/numeric).	Could affect mathematical operations.	Converted all numerical columns to float64.
Slight outliers in weight and height attributes.	Might affect model training.	Verified and retained as biologically valid.
Categorical variable Sex not numeric.	Model cannot interpret text.	Applied One-Hot Encoding.

8. Data Quality Improvement Measures

To maintain high data quality throughout the project, the following steps were implemented:

- Used encoding and scaling to normalize feature distributions.
- Applied data validation rules before training.
- Documented data lineage and transformations for transparency.
- Stored cleaned data separately from raw data for reproducibility.

9. Tools Used

Tool/Library	Purpose
Python (Pandas, NumPy)	Data analysis and validation
Matplotlib/Seaborn	Data visualization and pattern identification
Scikit-learn	Feature encoding and preprocessing
Jupyter Notebook / VS Code	Interactive development and documentation

10. Conclusion

The Abalone dataset used in this project demonstrates excellent data quality across all critical dimensions.

The dataset is complete, accurate, consistent, and valid, ensuring a reliable foundation for machine learning modeling.

This ensures that subsequent phases — model development, optimization, and deployment — will produce accurate, meaningful, and scientifically valid results. Maintaining this level of data integrity enhances the credibility and effectiveness of the final predictive system.