WATER QUALITY PREDICTION

Iraiyanbu S T
HemaKumar

Problem Statement

- Ensuring safe and clean water is a priority for public health and environmental sustainability.
- The project aims to classify water samples as “safe” or “unsafe” using machine learning algorithms based on water quality attributes.
- This classification helps authorities in monitoring and improving water safety

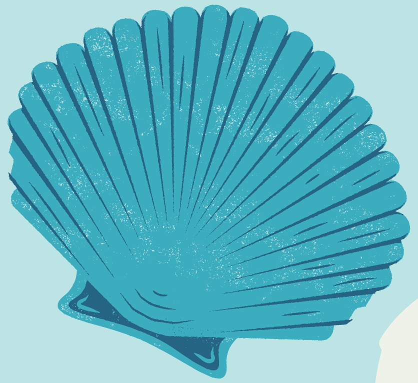
Introduction

- The goal of this project is to predict whether a given water sample is "safe" or "unsafe" for consumption based on its chemical properties.

- Machine learning provides a data-driven approach that can automate this process and offer predictions based on past water sample data, making it efficient for large-scale monitoring.

- The dataset contains various chemical properties such as aluminium, ammonia, arsenic, and more. The target variable is “is_safe,” which indicates whether the water sample is safe for drinking (1 for safe, 0 for unsafe).

GOOD TO KNOW



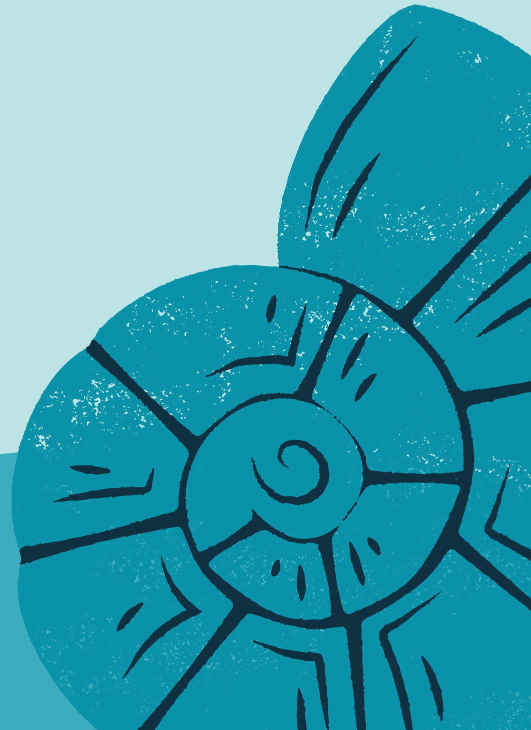
DATA COLLECTION

DATA PREPROCESSING

MODEL SELECTION

MODEL TRAINING

MODEL TESTING



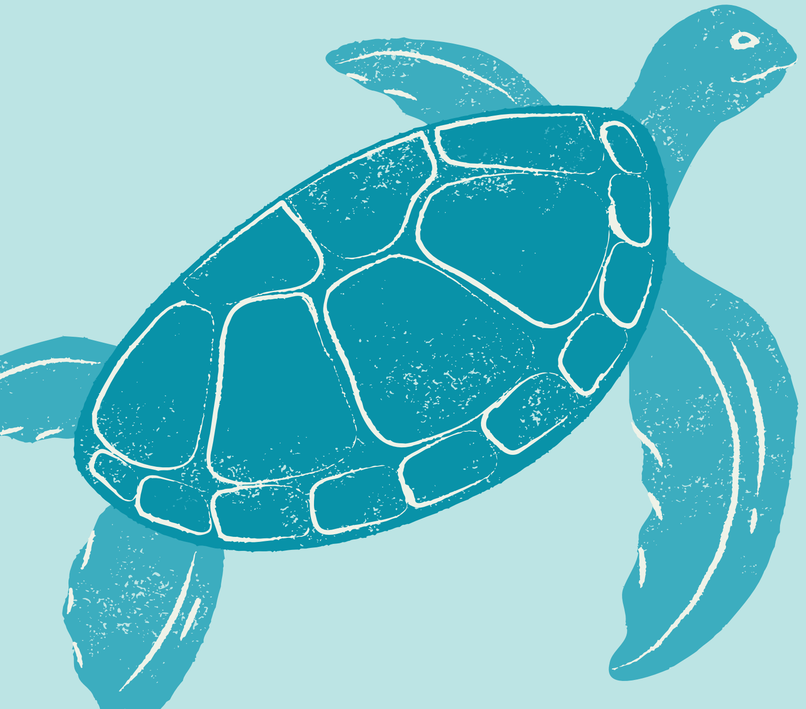
STEPS IN THE ANALYSIS

1.Data Collection:

- Use the dataset “water.csv” which includes 21 attributes (aluminium, ammonia, arsenic, etc.).

2.Data Preprocessing

- Handle missing data (fill missing values or drop incomplete rows).
- Convert categorical variables (e.g., ‘is_safe’ column) into numerical form.
- Normalize or scale data to ensure features contribute equally to the model



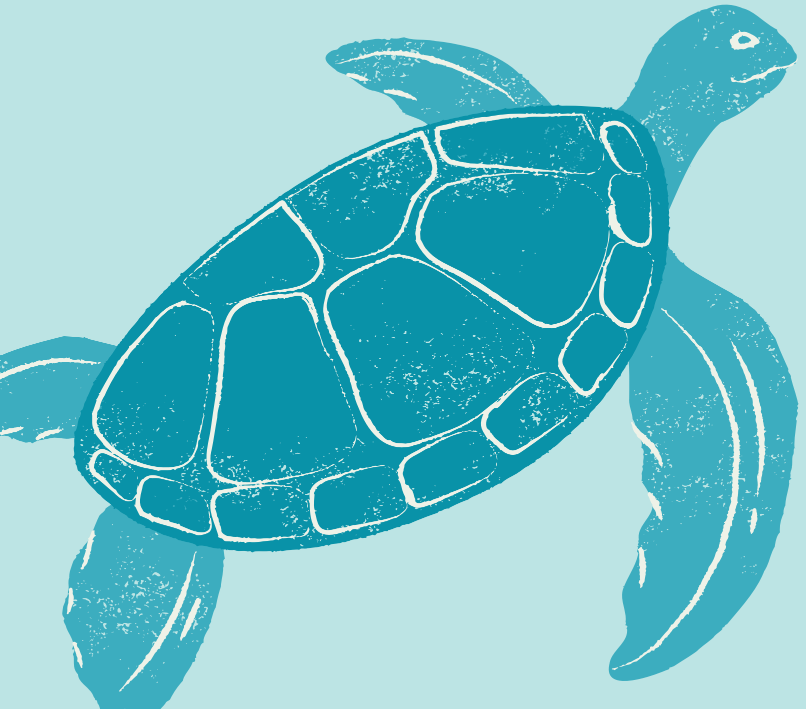
z STEPS IN THE ANALYSIS

3.Model Selection

- Consider algorithms like Logistic Regression, Decision Trees, and Random Forests for classification.

4.Model Training:

- Use 70% of the data to train the machine learning model.



5.Model Testing:

- Use the remaining 30% to test the model and evaluate performance.

MATHEMATICAL CALCULATION

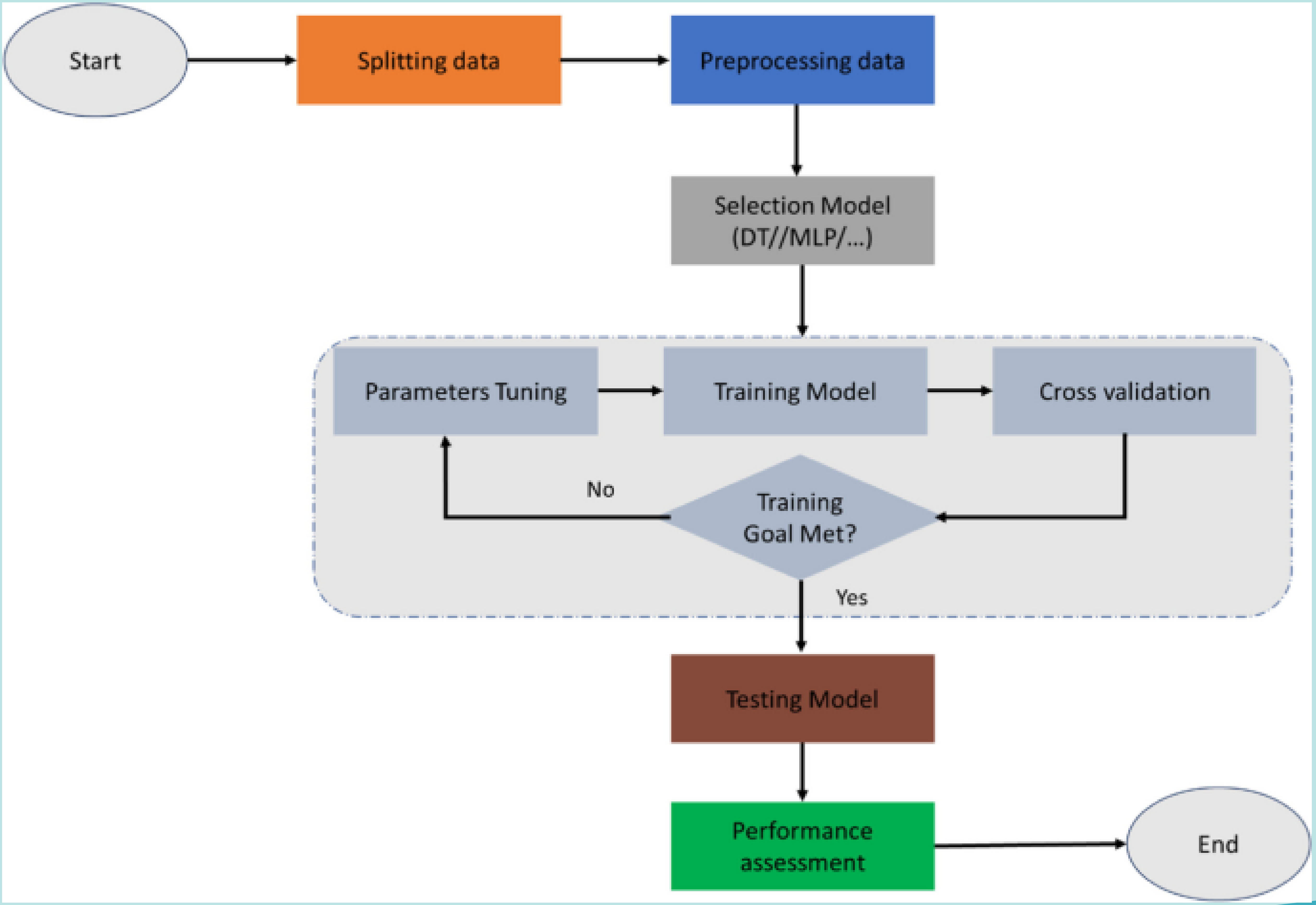
Logistic Regression Formula:

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

Example:

- For a sample test with Aluminium (Al) = 0.5 mg/L and Ammonia (NH3) = 0.2 mg/L, the formula can be used to predict the probability of the water being safe.

ARCHITECTURE DIAGRAM



Start

Splitting data

Preprocessing data

Selection Model (DT//MLP/...)

Parameters Tuning

Training Model

Cross validation

Training Goal Met?

No

Yes

Testing Model

Performance assessment

End

CONCLUSION

Model Performance:

After training and testing various models, the Logistic Regression model achieved an accuracy of 90%. The Random Forest model performed better, achieving a higher accuracy of 95%.

User Interface Integration:

The project includes an interface connected to the notebook, allowing users to input water sample data and receive real-time predictions on whether the water is safe or unsafe. This interface provides a practical application for field use, enabling authorities to assess water quality quickly and efficiently.

REFERENCES

- Dataset source: https://www.kaggle.com/datasets/mssmartypants/water-quality
- Libraries used: Pandas, NumPy, Scikit-learn, etc.