Water Potability Prediction Analysis

"In the domain of Water Quality Assessment, The project aims to develop a machine learning model to predict water potability based on various water quality parameters. This model will help in assessing whether water is safe for consumption, which is crucial for public health and water management."

Presented By: Milan Virash

Objectives :

- **Explore the Dataset**: Uncover patterns, distributions, and relationships within the data.
- **Conduct Extensive Exploratory Data Analysis (EDA)**: Dive deep into univariate and bivariate relationships against the target.
- **Preprocessing Steps**:
  - Address missing values
  - Treat outliers
  - Encode categorical variables
  - Transform skewed features to achieve normal-like distributions
- **Model Building**:
  - Establish models
  - Implement and tune classification models including Logistic Regression, SVM, Decision Trees, Random Forest and KNN
  - Emphasize achieving high recall for class 1, ensuring comprehensive identification potability
- **Evaluate and Compare Model Performance**: Utilize precision, recall, and F1-score to gauge models' effectiveness.

WorkFlow:

1. Introduction
2. Question
3. Dataset Overview
4. EDA
5. PreProcessing
6. Model Training
7. Model Selection
8. Conclusion

1 Introduction

- In this project, the aim is to develop a machine learning model to predict water potability based on various water quality parameters. This model will help in assessing whether water is safe for consumption, which is crucial for public health and water management.

Access to clean drinking water is a global challenge, with millions of people lacking reliable sources of potable water. By developing advanced predictive models, we can contribute to global efforts in improving water quality and ensuring safe drinking water for communities worldwide.

2 Question

- Is the water potable or not?

3 Dataset Overview

- **Dataset**
  - 3676 Rows
  - 10 Columns
- **Features (columns)**
  - **X** → **Input Features**: Ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity
  - **Y** → **Output Variable**: **Potability**
    - 1 = Potable
    - 0 = Not Potable

• All the columns given in the dataset are water components. Which I have tried to briefly introduce here in the table.

Parameter	Meaning
pH value	Indicates the acidity or alkalinity of water; standard range is 6.5 to 8.5.
Hardness	Measures the concentration of calcium and magnesium salts; no specific standard, but generally measured in mg/L.
Solids (Total dissolved solids - TDS)	Measures dissolved minerals in water; desirable limit is 500 mg/L, maximum limit is 1000 mg/L.
Chloramines	Disinfectants formed by adding ammonia to chlorine; safe levels are up to 4 mg/L.
Sulfate	Naturally occurring in minerals and soil; typical range in freshwater is 3 to 30 mg/L, can be higher in some areas.
Conductivity	Measures water's ability to conduct electricity; WHO standard is not to exceed 400 µS/cm.
Organic Carbon	Total amount of carbon in organic compounds; US EPA standard is < 2 mg/L in treated water, < 4 mg/L in source water.
Trihalomethanes	Byproducts of chlorination; safe levels are up to 80 ppm.
Turbidity	Measures the clarity of water; WHO recommended value is below 5 NTU.
Potability	Indicates if water is safe for consumption; 1 means potable, 0 means not potable.

4 EDA

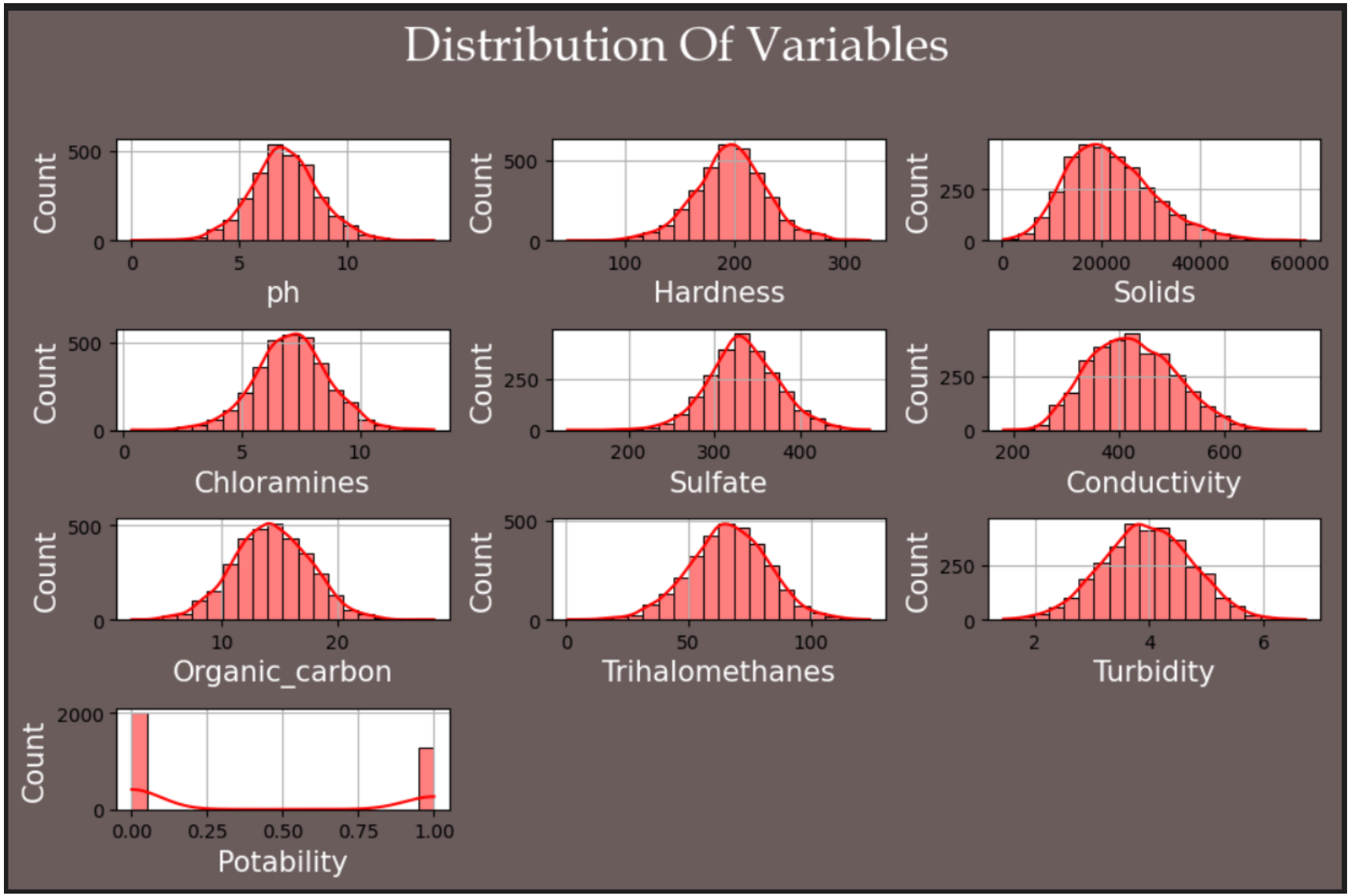
For **Exploratory Data Analysis (EDA)**, we'll take it in two main steps:

1. Univariate Analysis:
    - Here, we'll focus on one feature at a time to understand distribution and range.

2. **Bivariate Analysis:**
    - In this step, we'll explore the relationship between each feature and the target variable. This helps us figure out the importance and influence of each feature on the target outcome.

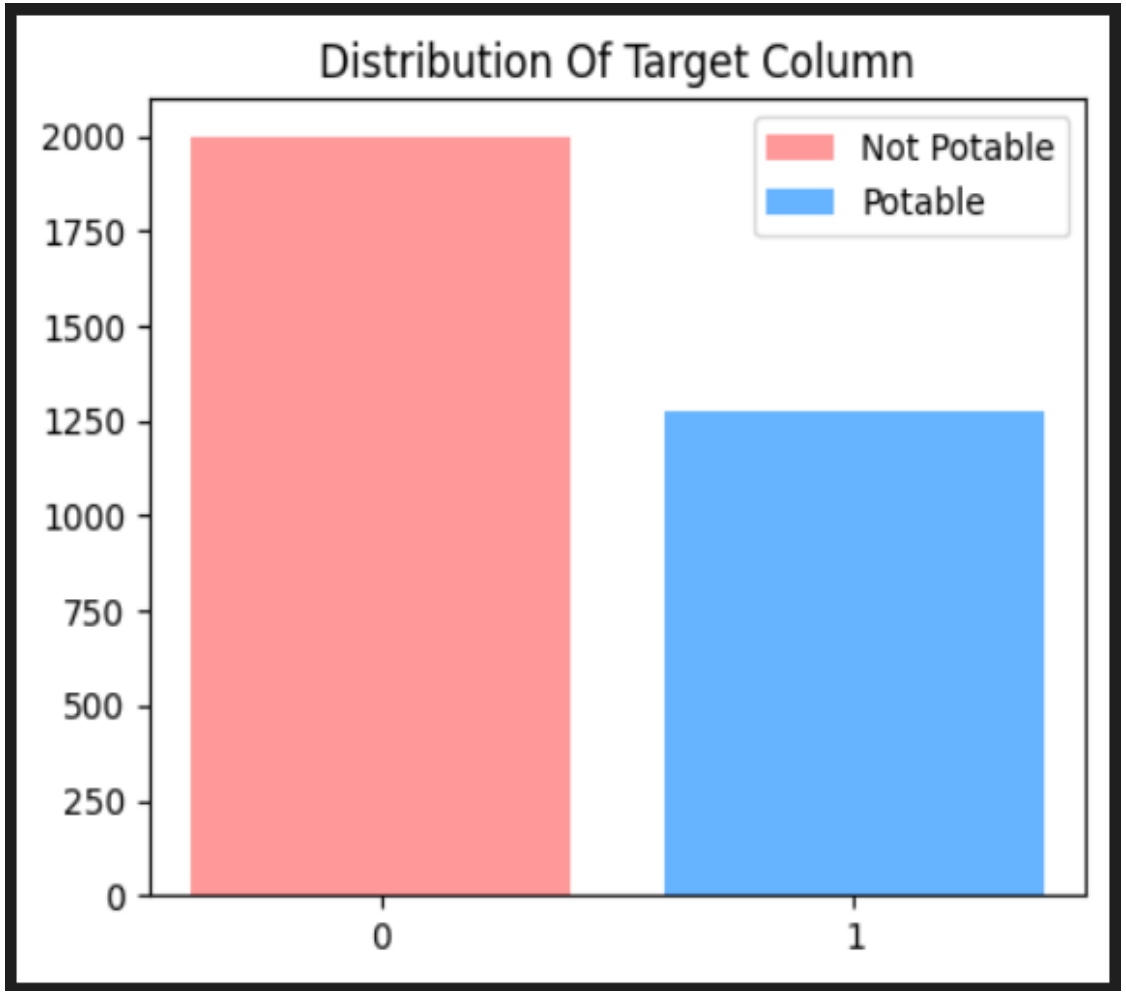
With these two steps, we aim to gain insights into the individual characteristics of the data and also how each feature relates to our main goal: **predicting the target variable.**

Univariate Analysis :



- All features have a very similar distribution of their values.
- In addition, they all show a very gaussian form so there is no need to normalize any feature.
- However here I have applied some algorithms (tree based algorithms) which require normalization or standardization so I have standardize the data further.

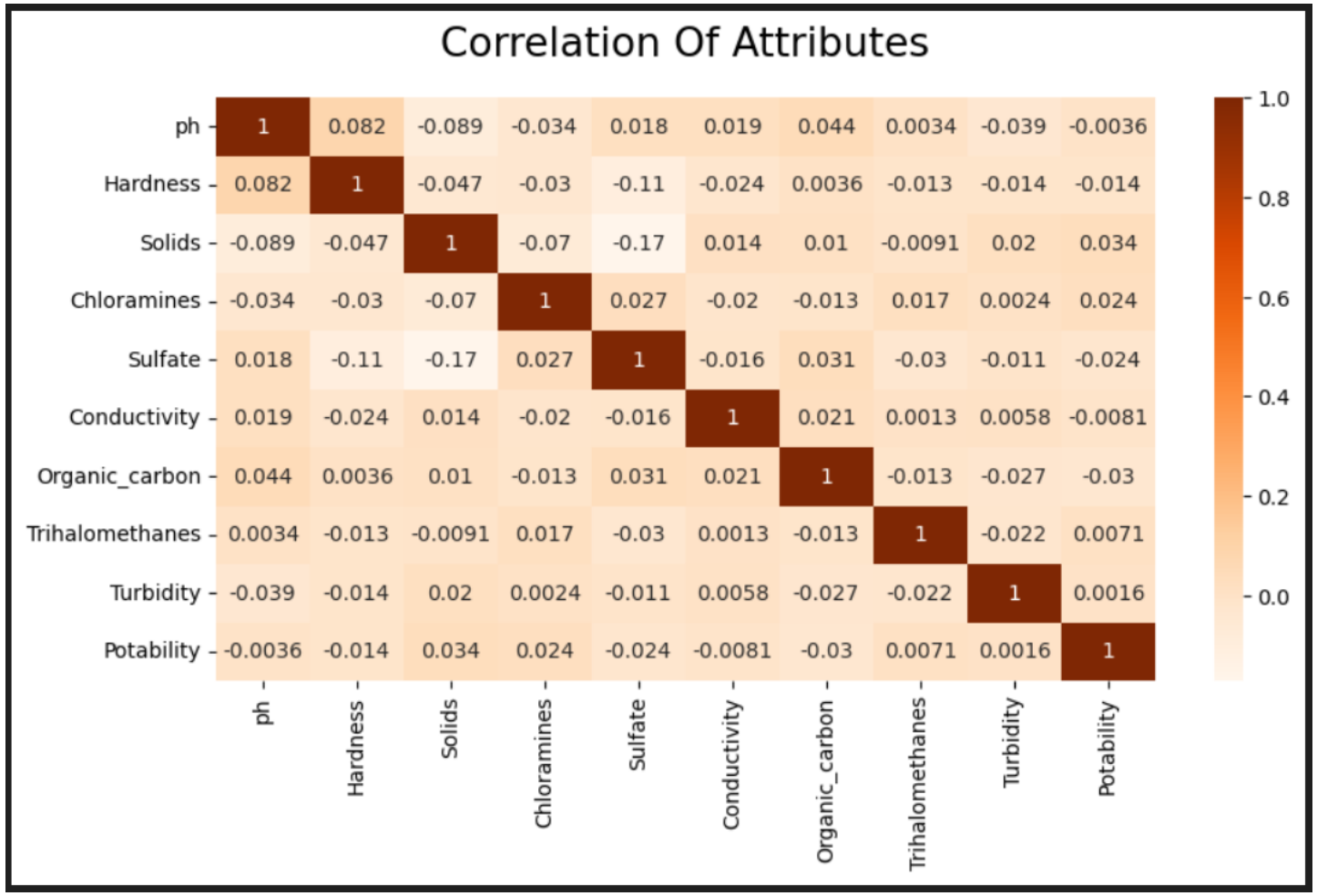
Distribution Of Target Column



- Explanation Visualization Purpose: The plot shows the count of each class (0 = Not Potable, 1 = Potable) in the dataset, allowing you to quickly assess whether the classes are balanced or imbalanced.

- Balanced vs. Imbalanced Data: A balanced dataset has an equal number of instances for each class, while an imbalanced dataset has a significant disparity between classes. Understanding this distribution is crucial, as an imbalanced dataset might require special handling to ensure the model is not biased towards the majority class.

Bivariate Analysis :



- No strong positive correlations exist between any variables and water potability.
- Sulfate, Trihalomethanes, and Conductivity show moderate negative correlations with potability.
- Factors like pH, Hardness, Solids, Chloramines, Organic_carbon, and Turbidity have weak or negligible correlations with potability.

5 PreProcessing

- **Irrelevant Feature Removal:**
  - All features in the dataset appears to be relevant based on **EDA**. We will retain all the columns, ensuring no valuable information is lost, especially given the dataset's small size.

- **Missing Value Treatment:**
  - Since our dataset is already small, removing rows with missing values would reduce its size even further. Therefore, I have filled the missing values with the **mean**.

- **Outlier Treatment:**
  - In a water potability dataset, outliers could represent instances where water quality is dangerously high or low for certain contaminants. These extreme values could be crucial for identifying water sources that pose significant health risks.

- **Standardization:**
  - **here I have applied some algorithms (tree based algorithms) which require normalization or standardization so I have standardize the data further.**

• Splitting The Data Into X & Y:

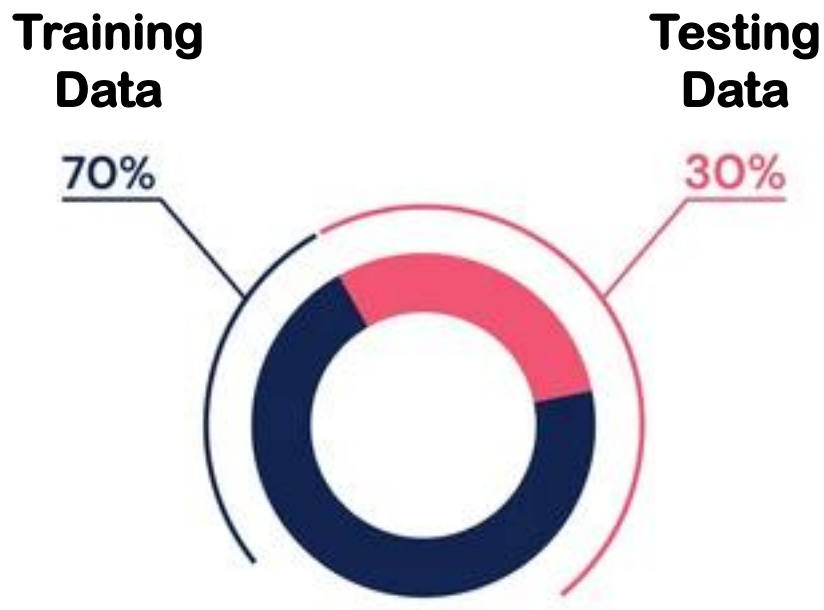
• We divide the dataset into two parts: X and Y.

"X" typically represents the Independent variables, and "Y" represents the Dependent (target variable) that we want to predict or understand

• Train Test Split:

• We divide the data into training(70%) and testing(30%) sets.

Setting a random state ensures consistent results and stratify=Y maintains a proportional distribution of the target variable in both sets.



6 Model Training

• Model Used:

- **Logistic Regression**: logistic Regression is commonly used for **binary classification problems**. it's preferred because it provides a **simple** an **efficient** way to **model** the **relationship** between the **independent variables** and the **probability** of a certain **outcome**.

- **Decision Tree**: Decision Tree algorithms are used for **classification** because they are **simple**, **computationally efficient**, and **effective** in handling **high-dimensional data**.
  Works best for categorical independent columns.

- **Random Forest Algorithm**: Random Forest: Random Forest is a robust supervised algorithm suitable for both regression and classification tasks.

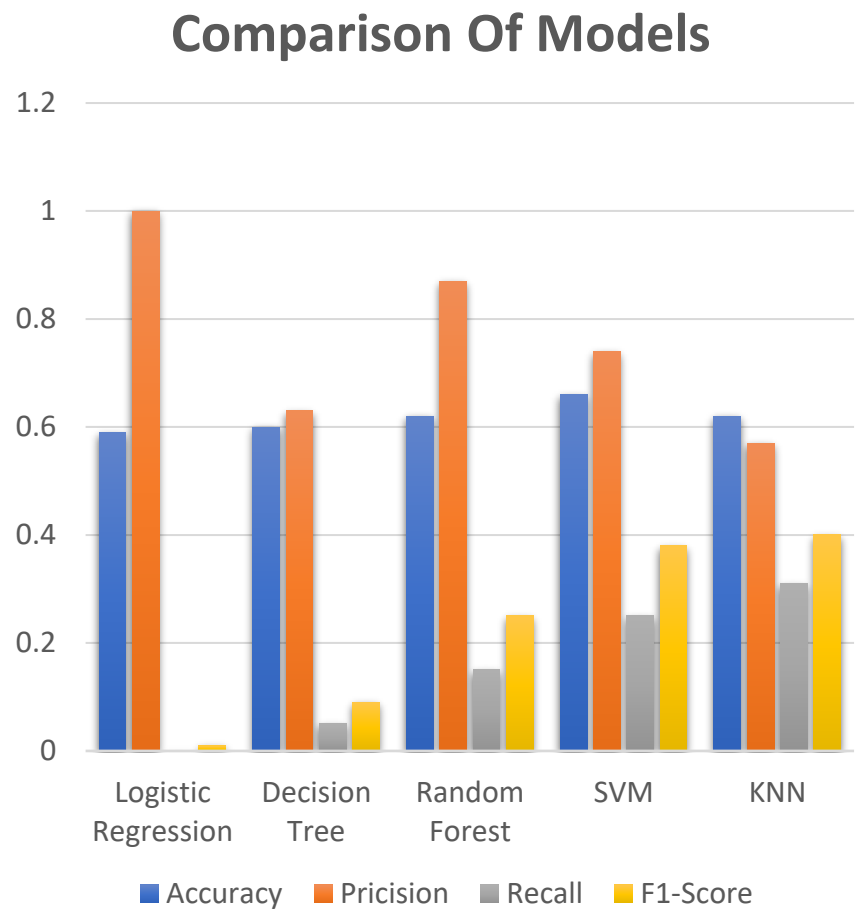
- **Support Vector Machine**: SVM is a **powerful supervised algorithm** that works best on **smaller datasets** but on **complex ones**. Support Vector Machine(**SVM**)can be used for both **regression** and **classification** tasks, but generally, they **work best** in **classification problems**.

- **K nearest Neighbours(KNN)**: KNN is a versatile algorithm used for classification, especially when data lacks a clear boundary. It works by considering the k nearest neighbors to a data point and classifying it based on their majority class.

7 Model Selection

• Model Comparison:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.59	1.00	0.00	0.01
Decision Tree	0.60	0.63	0.05	0.09
Random Forest	0.62	0.87	0.15	0.25
SVM	0.66	0.74	0.25	0.38
KNN	0.62	0.57	0.31	0.40



• Understanding:

- **Accuracy:** The accuracy of the model when it claims to have found something.
- **Precision:** it refers to the quality of being exact and accurate.
- **Recall:** The ability of a model to find all the relevant cases.
- **F1-Score:** A balance between recall and precision, useful when both false positives and false negatives need to be minimized.

• Best Model: **SVM**

• **Reason**: The SVM model strikes the best balance between precision and recall for class 1, leading to a higher F1-score (0.38). While it has slightly lower accuracy than Random Forest, the improved performance on class 1 (higher recall and F1-score) makes it a better choice for tasks where identifying the minority class (class 1) is important.

• Why Other Models Are Not Best?:

• Logistic Regression: Fails to identify class 1 (recall of 0.00), making it unsuitable for tasks needing balanced performance across classes.

• Decision Tree: Slightly better than Logistic Regression but still very poor recall for class 1 (0.05), leading to low F1-scores.

• Random Forest: High precision but low recall for class 1, suggesting potential overfitting or an inability to generalize well to minority class instances.

• KNN: Better recall for class 1 than some models, but overall performance metrics are still lower compared to SVM.

8 Conclusion

- Various machine learning algorithms were applied to predict water potability based on the given dataset.
- The models tested include Logistic Regression, Decision Tree, Random Forest, SVM, and KNN classifiers.
- Among these, the Support Vector Machine (SVM) classifier demonstrated the highest performance in terms of accuracy and robustness. This suggests that SVM is particularly effective in capturing the complex relationships within the data, making it a suitable choice for this classification task.
- The superior performance of SVM highlights its capability to handle the nuances of water quality prediction, which could be vital for ensuring safe and reliable water sources.

• Future Work:

- Future work could explore further optimization of the SVM model or the integration of additional features to enhance predictive accuracy.

Thank You