Water Quality Prediction Report

-Bristi Halder

Introduction

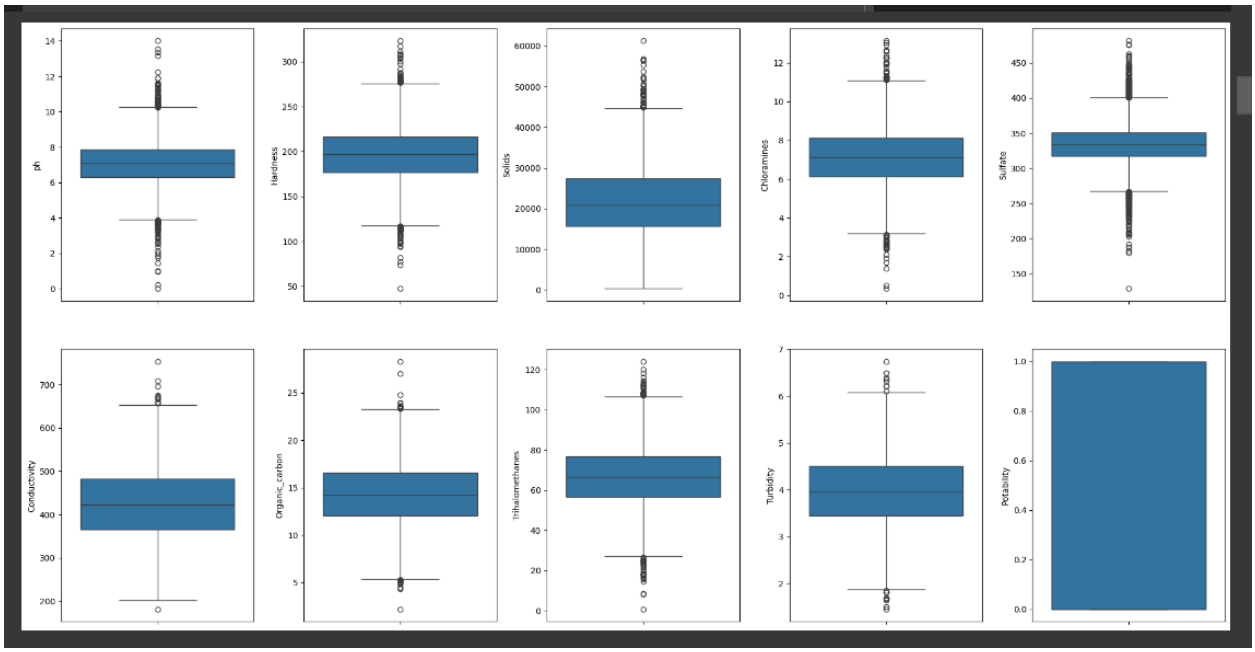
This report presents the findings and analysis of a machine learning project aimed at predicting water quality based on various properties of water.

Dataset Overview

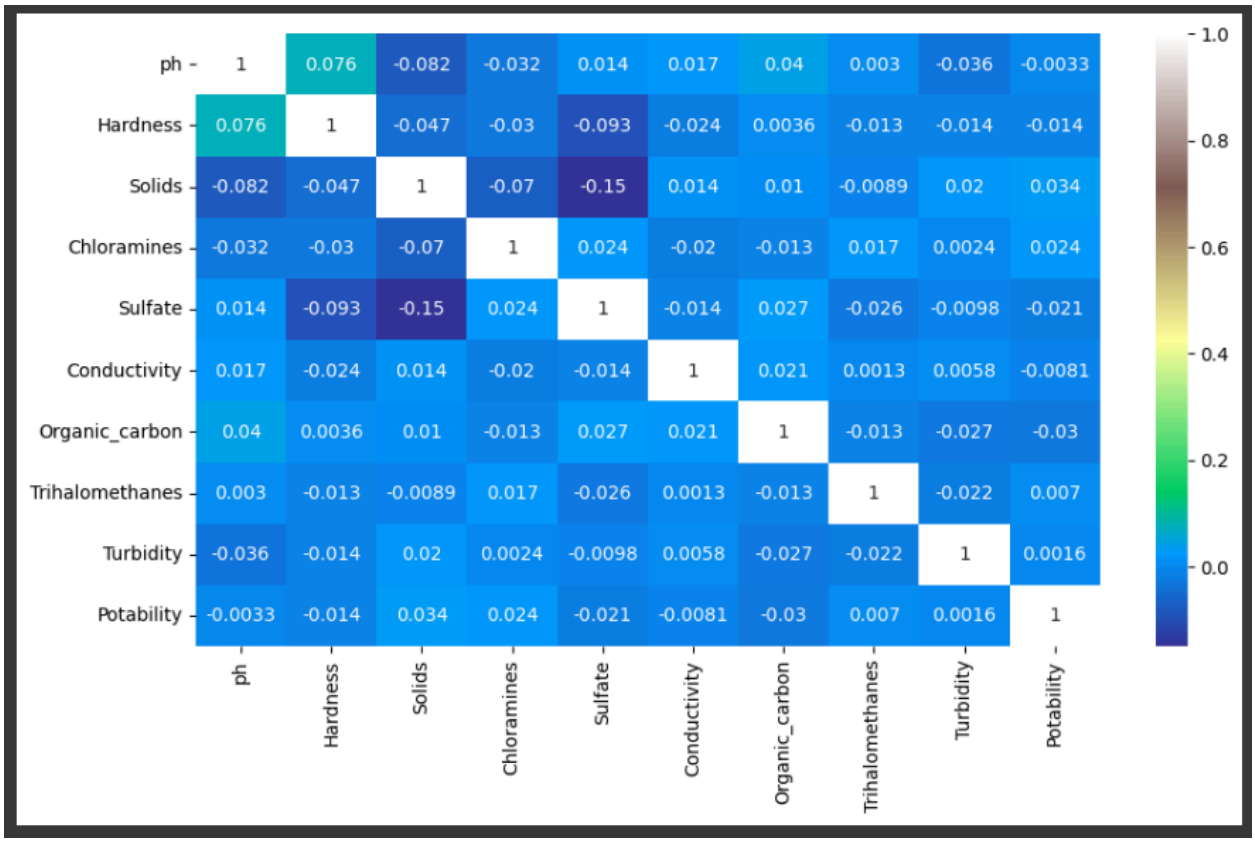
- The dataset has been taken from Kaggle.
- The target variable ‘Probability’ indicates whether the water samples are potable(safe for consumption) or not.

Exploratory Data Analysis (EDA)

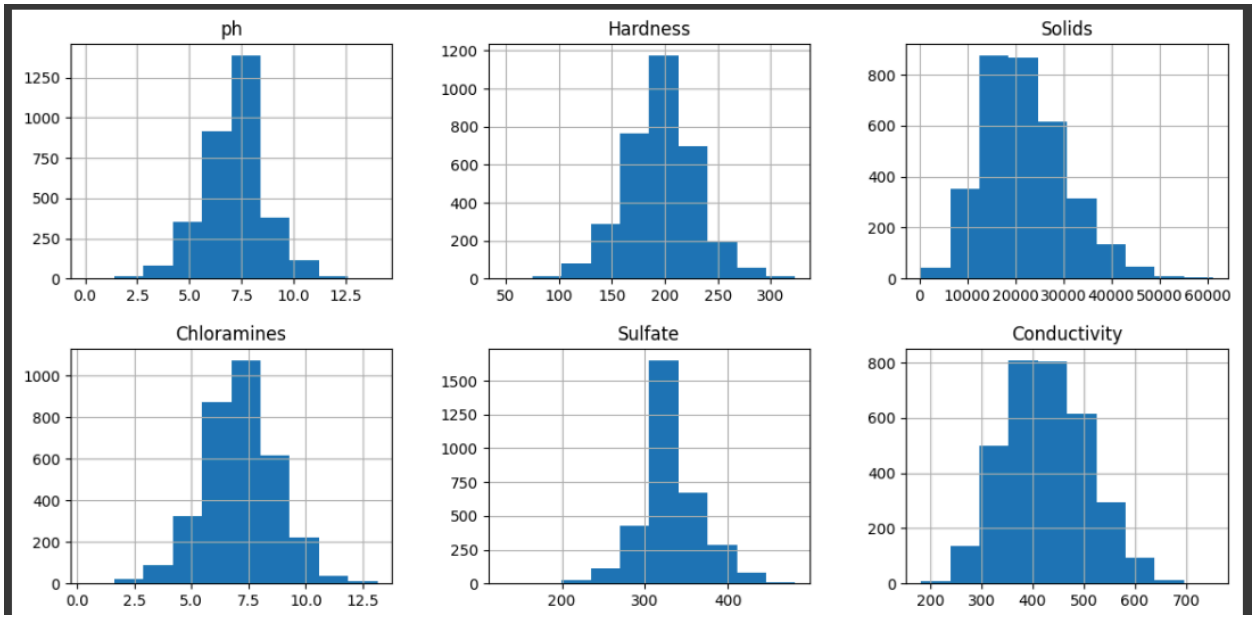
- **Data Cleaning:** The dataset was checked for missing values and outliers. Missing values were imputed using appropriate strategies, and outliers were addressed through visualization and statistical analysis.
  In our analysis, we identified an outlier in the solids parameter. We decided not to remove this outlier, as the elevated level of solids might indicate that the water is still good or drinkable.

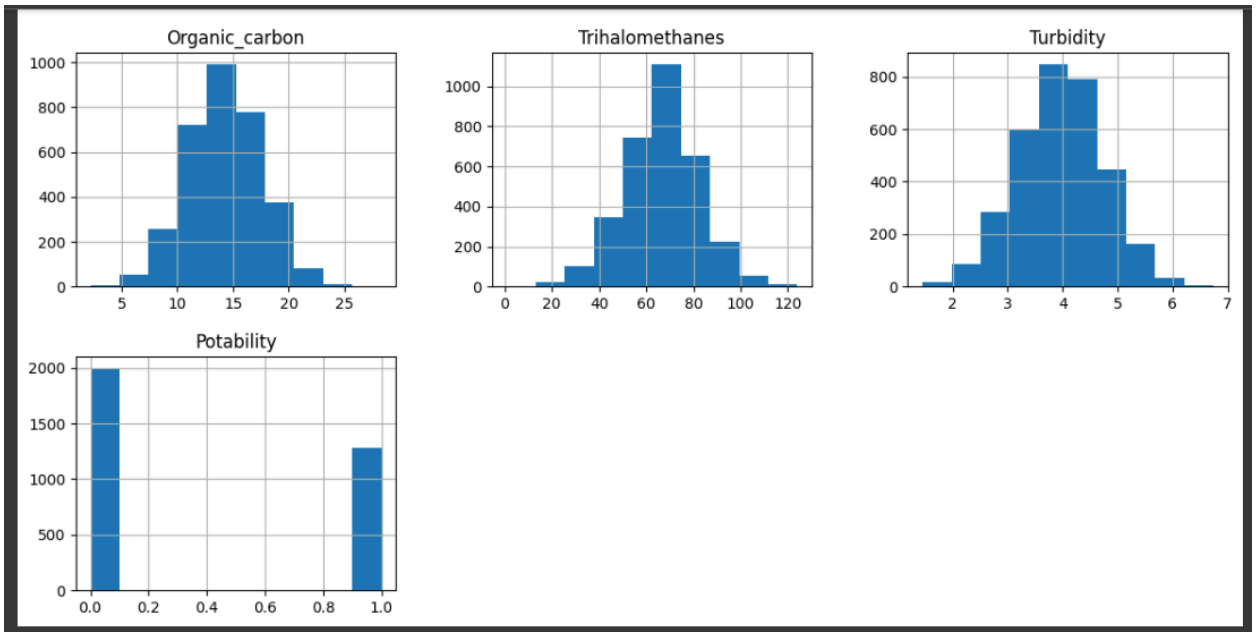


- For **dimensionality reduction** we need to check which feature is less important.
  By examining the data correlations using a heatmap, we found that none of the features had correlations exceeding 50%. Therefore, we chose not to perform dimensionality reduction and retained all nine parameters for our analysis.



- We have analyzed data using histogram and found no need of normalization as it is already normalized.





- **Data Visualization:** Various plots such as histograms, scatter plots and box plots were created to explore the relationships between different variables and their distributions.

Model Building:

8 machine learning algorithms are used

1. Decision Tree Classifier
2. K-Nearest Neighbors (KNN) Classifier
3. Logistic Regression
4. Random Forest Classifier
5. XGBoost Classifier
6. Gaussian Naive Bayes
7. Support Vector Machine (SVM) Classifier
8. AdaBoost Classifier

Model Evaluation

- Each model was evaluated using cross-validation and performance metrics such as accuracy, precision, recall and F1-score.
- Hyperparameter tuning was performed using techniques such as grid search to optimize model performance.
- Did model optimization for Decision Trees and KNN, for other models we already had used the best parameters.

	Model	Accuracy_score
4	XGBoost Classifier	66.543438
5	Gaussian Naive Bayes	63.863216
7	AdaBoost	63.401109
2	Logistic Regression	62.846580
3	Random Forest	62.846580
6	SVM	62.846580
1	KNN	60.073937
0	Decision Tree	58.780037

Before optimization

	Model	Accuracy_score
0	Decision Tree	58.040665
1	KNN	60.073937
2	Logistic Regression	62.846580
3	Random Forest	62.846580
4	XGBoost Classifier	66.543438
5	Gaussian Naive Bayes	63.863216
6	SVM	62.846580
7	AdaBoost	63.401109

After optimization

Results

Based on the evaluation metrics, the **XGBoost Classifier** emerged as the best performing model with the highest accuracy.

Conclusion

The XGBoost Classifier demonstrated superior performance in predicting water quality compared to other models.