21K-3326 Syed Hadi Arshad

Project: Water Quality Analysis

1. Introduction

Access to clean drinking water is a fundamental human need and a critical global health concern. Water quality assessment traditionally relies on laboratory testing of various physicochemical parameters, which can be time-consuming and resource intensive. This project applies machine learning techniques to predict water potability based on measurable water quality metrics, potentially offering a faster preliminary assessment tool for water safety.

The dataset used in this study contains 3,276 water samples with 9 features: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, and Turbidity. The target variable, Potability, indicates whether water is safe for human consumption (1) or not (0). The dataset presents several challenges, including missing values in pH (491), Sulfate (781), and Trihalomethanes (162) columns, as well as a class imbalance with approximately 39% potable samples.

2. Methodology

2.1 Data Preprocessing

Our preprocessing approach addressed several key challenges in the dataset:

- **Missing Value Imputation**: Mean values were used to fill missing data points in pH, Sulfate, and Trihalomethanes columns. This approach was selected after experimental comparison with median imputation, which yielded lower model performance.
- **Feature Standardization**: We applied StandardScaler to normalize all features to a common scale, essential for distance-based algorithms like KNN and for improving the convergence of Logistic Regression.
- **Train-Test Split**: The dataset was divided into training (70%) and testing (30%) sets with random_state=1 for reproducibility.

2.2 Model Implementation

We implemented and compared five machine learning approaches:

1. **K-Nearest Neighbors**: Configured with n_neighbors=2 and Manhattan distance metric (p=1) to capture local patterns in the water quality data.
2. **Decision Tree**: Implemented with entropy criterion and max_depth=12 to balance complexity capture and overfitting prevention.
3. **Naive Bayes**: Applied Gaussian Naive Bayes, assuming normal distribution of features.

4. **Logistic Regression**: Configured with max_iter=120 and n_jobs=20 for efficient computation.
5. **Ensemble Learning**: Created a Voting Classifier combining all four models with soft voting (probability-based) to leverage the strengths of multiple approaches.

3. Results and Analysis

3.1 Model Performance

The performance comparison of implemented models revealed:

Model	Training Accuracy (%)	Testing Accuracy (%)
K-Nearest Neighbors	79.76	61.95
Decision Tree	78.63	63.58
Naive Bayes	63.24	61.95
Logistic Regression	61.58	59.51
Voting Classifier	85.48	63.28

The Decision Tree model demonstrated the highest testing accuracy at 63.58%, closely followed by Naive Bayes and KNN. All models showed relatively modest accuracy, highlighting the challenging nature of water potability prediction based solely on physicochemical parameters.

3.2 Feature Importance

Analysis of feature importance based on the Decision Tree model revealed that pH is the most influential parameter in determining water potability, followed by Sulfate and Chloramines. The relatively even distribution of importance scores across features suggests that water potability depends on a complex interaction of multiple parameters rather than being dominated by a single factor.

3.3 Model Evaluation

For the best-performing Decision Tree model, additional evaluation metrics were calculated:

- **MSE**: 0.3642
- **RMSE**: 0.6035
- **MAE**: 0.3642
- **Precision**: 0.6190
- **Recall**: 0.2613
- **F1 Score**: 0.3675

The confusion matrix analysis indicated better performance in identifying non-potable water than potable water, which is advantageous from a public health perspective as it reduces the risk of classifying unsafe water as potable.

4. Discussion and Conclusions

- **Model Performance**: Decision Tree achieved the highest accuracy (63.58%) among the implemented models, suggesting that tree-based methods can effectively capture the non-linear relationships in water quality data.
- **Feature Significance**: pH emerged as the most important predictor of water potability, aligning with established water quality standards that emphasize pH as a critical parameter.
- **Prediction Challenges**: The moderate accuracy of all models indicates the complex nature of water potability prediction and suggests potential limitations in the current feature set.