Water Potability Classification

1. Introduction

Water quality is a critical public health issue. Current methods for testing water safety often require manual sampling and laboratory analysis, which can be costly and time-consuming. In this study, I explore whether machine learning models can predict water potability using chemical and physical attributes from a dataset. Our goal is to identify patterns in the data that could help classify water as potable or non-potable. This project examines several modeling approaches, from basic models to advanced techniques like stacking, and addresses the key challenges associated with class imbalance and feature engineering.

2. Materials and Methods

2.1 Dataset

The dataset used contains various water quality indicators, such as pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. The target variable, Potability, indicates whether the water is safe to drink, with 0 representing non-potable and 1 representing potable. The dataset is imbalanced, with non-potable water being overrepresented. This imbalance prompted the need for special techniques to ensure a fair evaluation of model performance.

2.2 Data Preprocessing

Initially, missing values Ire handled by removing rows containing NaN values. The remaining features Ire then scaled using StandardScaler to normalize their ranges, allowing for more consistent model performance. To better understand the data, I computed a correlation matrix to assess the relationships between features and the target variable. However, the results indicated a lack of strong correlations between the original variables and water potability, which suggested the need for feature engineering to capture more complex interactions.

2.3 Feature Engineering

Given the absence of significant correlations in the original data, I introduced polynomial features to capture interactions between the variables. By generating polynomial features of degree two, I expanded the feature set from 9 to 45 variables, allowing the models to better capture non-linear relationships. Recursive Feature Elimination (RFE) was then applied to identify the most relevant features. After evaluating the model performance, I

found that selecting the top 15 to 20 features provided the best results, balancing model complexity and accuracy.

2.4 Addressing Class Imbalance

One of the major challenges in this project was the imbalance between potable and non-potable water samples. Early confusion matrix results showed that the models Ire biased towards predicting non-potable water, leading to a high number of false negatives for potable water. To address this, I used Synthetic Minority Over-sampling Technique (SMOTE), which artificially increased the number of potable water samples by generating synthetic data points. This helped the models learn to recognize potable water better and reduced the number of false negatives.

3. Model Development and Results

3.1 Initial Modeling

I began with Logistic Regression, balancing the dataset using the resample function. The Logistic Regression model achieved an accuracy of 50.37%, barely above chance, indicating that the linear model struggled to differentiate between potable and non-potable water. Next, I applied Random Forest, which yielded an improvement in accuracy to 58.81%. This model performed better due to its ability to handle non-linear relationships and its robustness to feature interactions. However, the accuracy was still not sufficient, and hyperparameter tuning via GridSearchCV did not result in any significant gains.

3.2 Introducing Polynomial Features

After analyzing the feature correlations and seeing little impact on potability prediction, I introduced polynomial features to capture more complex patterns in the data. This significantly improved the model performance, with accuracy rising to 68.49%. The additional interactions captured by the polynomial features allowed the Random Forest model to make more accurate classifications.

3.3 Hyperparameter Tuning and Threshold Adjustment

I then fine-tuned the Random Forest model, adjusting the n_estimators parameter, and found an optimal value of 230, which increased the accuracy to 69.73%. Although the overall accuracy improved, the recall for class 1 (potable water) was still low, at 0.48, while the recall for class 0 (non-potable water) was high, at 0.83. This imbalance in recall was concerning because the model was still prone to misclassifying potable water as non-potable, which is problematic in real-world applications where false negatives could lead

to unnecessary actions. To address this, I experimented with different decision thresholds. By adjusting the threshold from the default 0.5 to 0.4, I managed to balance the recall values better, resulting in F1-scores of 0.68 for non-potable water and 0.64 for potable water. Although this adjustment slightly decreased the overall accuracy to 66%, it improved the model's ability to identify potable water correctly.

3.4 Stacking Models

To further enhance the model's performance, I employed a stacking ensemble approach. I combined Random Forest and Logistic Regression as base models, with LightGBM included as an additional base model to capture more complex patterns. Gradient Boosting was used as the meta-model to aggregate the predictions from the base models. The stacked model produced balanced precision and recall values for both classes, maintaining an accuracy of 66% while reducing the disparity between the precision and recall for potable and non-potable water.

4. Discussion

The model development process demonstrated the challenges of predicting water potability, especially with an imbalanced dataset. The introduction of polynomial features played a critical role in improving model performance, as it allowed us to capture complex relationships that Ire not evident in the original data. However, even with these improvements, the model's overall accuracy and recall for potable water remain lower than what would be desirable for real-world applications. The use of SMOTE effectively addressed the issue of class imbalance, reducing false negatives and improving the model's ability to classify potable water. However, the fact that the final model still misclassifies a significant portion of the data—correctly identifying only 66% of instances—means it is not yet a viable solution for practical water safety monitoring. Future work should focus on further refining the feature engineering process, perhaps incorporating additional domain-specific variables such as geographic or seasonal data. More advanced models, such as XGBoost or CatBoost, may also offer improved performance. Additionally, hyperparameter tuning of the Gradient Boosting meta-model could yield further gains.

5. Conclusion

This study explored the application of machine learning to predict water potability based on a set of chemical and physical attributes. While the final stacked model achieved an accuracy of 66%, the model's limited ability to correctly classify potable water suggests that further improvements are needed before this approach can be used in real-world scenarios. Despite these limitations, the project highlighted the importance of feature

engineering and handling class imbalance, which Ire crucial to the gains in model performance. Future iterations of this work should explore more advanced models and further optimize feature selection and data augmentation techniques.
