CSE 445 (Section 07)
Project Report

Project Title: Water Potability Prediction Using Machine Learning

Submitted To:

Dr. Mohammed Shifat-E-Rabbi

Date:

15/11/2023

Group Members

ID	Name
2011657042	Zubair Mahmood Sowrab
2013412042	Md. Zakaria Khan
2014215642	Tanjim Imtial
2011312642	Maqsudul Mahmud Fahim
2012972642	Al-jubayer Pial

"Waves of Assurance: Machine Learning for Water Potability Prediction"

Introduction: This machine learning project aims to predict water potability based on a dataset of water quality features. The dataset contains various attributes such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, with the target value being water potability. In this report, we will provide a detailed overview of the project, including data preprocessing using Min-Max scaling, the choice of a Random Forest Classifier model, and a graphical representation of the predicted versus actual values.

Data Preprocessing: Min-Max Scaling Data preprocessing is a fundamental step in any machine learning project, as it helps ensure that the data is suitable for modeling. In this project, Min-Max scaling was employed to preprocess the dataset. Min-Max scaling is a feature scaling technique that transforms the features to a specific range, typically between 0 and 1, ensuring that all features have equal weight and preventing the dominance of certain features over others during model training.

Min-Max scaling is particularly well-suited for this project due to the varying scales of the water quality features. For example, pH values range from 0 to 14, while hardness values may be in the hundreds. Scaling the data to a common range helps the machine learning model perform optimally by preventing features with larger numerical values from disproportionately influencing the model's predictions. It also contributes to faster convergence during training and can lead to more accurate results.

Machine Learning Model: Random Forest Classifier The choice of the machine learning model is a critical decision in any predictive modeling project. In this project, a Random Forest Classifier was selected as the model of choice for predicting water potability.

The Random Forest Classifier is an ensemble learning algorithm that combines multiple decision trees to make predictions. Its strength lies in its ability to handle both numerical and categorical features, as well as its robustness against overfitting. The Random Forest model is known for its excellent performance in classification tasks and is often used when there are multiple features with varying importance levels.

The decision to use a Random Forest Classifier in this project is justified by its suitability for handling a diverse set of water quality features. It can capture complex relationships between the features and the target variable, allowing for accurate potability predictions. Additionally, the ensemble nature of the model helps reduce the risk of overfitting and ensures robust generalization to unseen data.

Data Splitting and Model Training: Before training the Random Forest Classifier, the dataset was split into a training set and a test set. The training set was used to train the model, while the test set served as an independent dataset for model evaluation.

The model was trained using the training set, and hyperparameters were fine-tuned to optimize its performance. Hyperparameters such as the number of trees in the forest, maximum tree depth, and minimum samples per leaf were adjusted to achieve the best possible results.

Model Evaluation: Model evaluation is a crucial step in assessing the performance of the machine learning model. In this project, several evaluation metrics were used to gauge the model's effectiveness, including accuracy, confusion matrix.

Accuracy: Accuracy is a fundamental metric for classification tasks and measures the percentage of correct predictions made by the model. In this project, the Random Forest Classifier achieved an accuracy of 68.75%. While accuracy provides a general indication of the model's performance, it may not be sufficient in cases where class distribution is imbalanced. Further metrics are needed to fully assess the model's effectiveness.

Confusion Matrix: The confusion matrix provides a more comprehensive view of a model's performance by breaking down the predictions into four categories: True Positives (TP): The number of correctly predicted potable water samples, True Negatives (TN): The number of correctly predicted non-potable water samples. False Positives (FP): The number of non-potable water samples incorrectly predicted as potable. False Negatives (FN): The number of potable water samples incorrectly predicted as non-potable. For this project, the confusion matrix is as follows:

	Predicted Potable	**Predicted Non-Potable**
Actual Potable	361	51
Actual Non-Potable	154	90

The confusion matrix provides a detailed breakdown of the model's performance, highlighting the trade-offs between false positives and false negatives. In this case, the model correctly predicted 361 instances of potable water and 90 instances of non-potable water. However, there were 51 false positives and 154 false negatives. These evaluation metrics collectively demonstrate the model's capability to accurately predict water potability based on the chosen features.

Graphical Representation: To provide a visual representation of the model's predictions, a scatter plot was created to compare the actual water potability values with the predicted values. The scatter plot allows for a direct visual assessment of how closely the model's predictions align with the actual potability values. Each point on the plot corresponds to a water sample.

Conclusion: Overall, this machine learning project, demonstrates the potential utility of machine learning in water quality assessment and represents a foundation for further enhancements and applications in the field. Future work may involve exploring additional modeling techniques, collecting more data, and conducting rigorous validation to ensure the model's reliability in real-world scenarios.

Dataset Link: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability?fbclid=IwAR3Eo3NI7marF1fMobYFSvlFiE4mJzqRmo9RR_WtgYEjzRMhe2UF3vd6sGw