Inspiring Excellence

CSE422: Artificial Intelligence

Project Name: Water Probability

Submitted By:

Group 6

Name	ID
TASHFIQ ALAM OVEY	20301299
MD. MINHAZUL ISLAM	20301433
MD. SHAFAYAT SADAT SAAD	20301457
AYESHA BINTEE ROB	23241079

Section: 10

Submitted To:

Mehran Hossain, Swattic Ghose

Lecturer

Introduction

Water quality is a critical global public health issue, and early detection of potability is key to preventing waterborne diseases. Machine learning algorithms, in particular, have shown significant potential for improving the accuracy of water potability classification using artificial intelligence (AI) techniques. We collect data on various indicators associated with water quality, preprocess and engineer the features, and train and evaluate machine learning models. Thus, this project provides a comprehensive analysis of water potability classification using AI techniques. By leveraging these advanced technologies, we aim to contribute to global efforts to ensure access to safe drinking water, a basic human right and a crucial component of effective public health policy.

Motivation

This study on water potability classification and the development of this report are driven by a variety of factors, including both practical and scientific considerations. The following are some major reasons:

Firstly, access to safe drinking water is a fundamental human right and a critical global health issue. The chance of preventing waterborne diseases and enhancing public health is significantly increased by early detection and an appropriate classification of water potability. By helping authorities identify potable water more precisely, the creation of an ML-based classification system can improve public health.

Secondly, to enhance public health outcomes and reduce costs associated with waterborne diseases, the public health sector is increasingly implementing data-driven strategies. This initiative aligns with the broader trend of using data analytics and AI in public health.

Thirdly, artificial intelligence has advanced remarkably, especially in the areas of machine learning and deep learning. AI algorithms can effectively analyze complex water quality data when used in the exciting and potential field of water quality assessment.

Then again, the implementation of machine learning theories acquired in an artificial intelligence course is part of this project. It enhances students' understanding of AI principles by giving them practical experience in data preparation, model selection, and evaluation.

In summary, the significance of the water potability classification effort stems from its potential to enhance public health through AI-driven assessment by enabling early detection and intervention, improving public health outcomes, and reducing healthcare costs. It addresses the critical need for precise water potability assessment, takes advantage of AI advances, provides educational value, promotes multidisciplinary collaboration, and is consistent with the broader trend of data-driven public health, making it an essential and crucial achievement.

Dataset Description

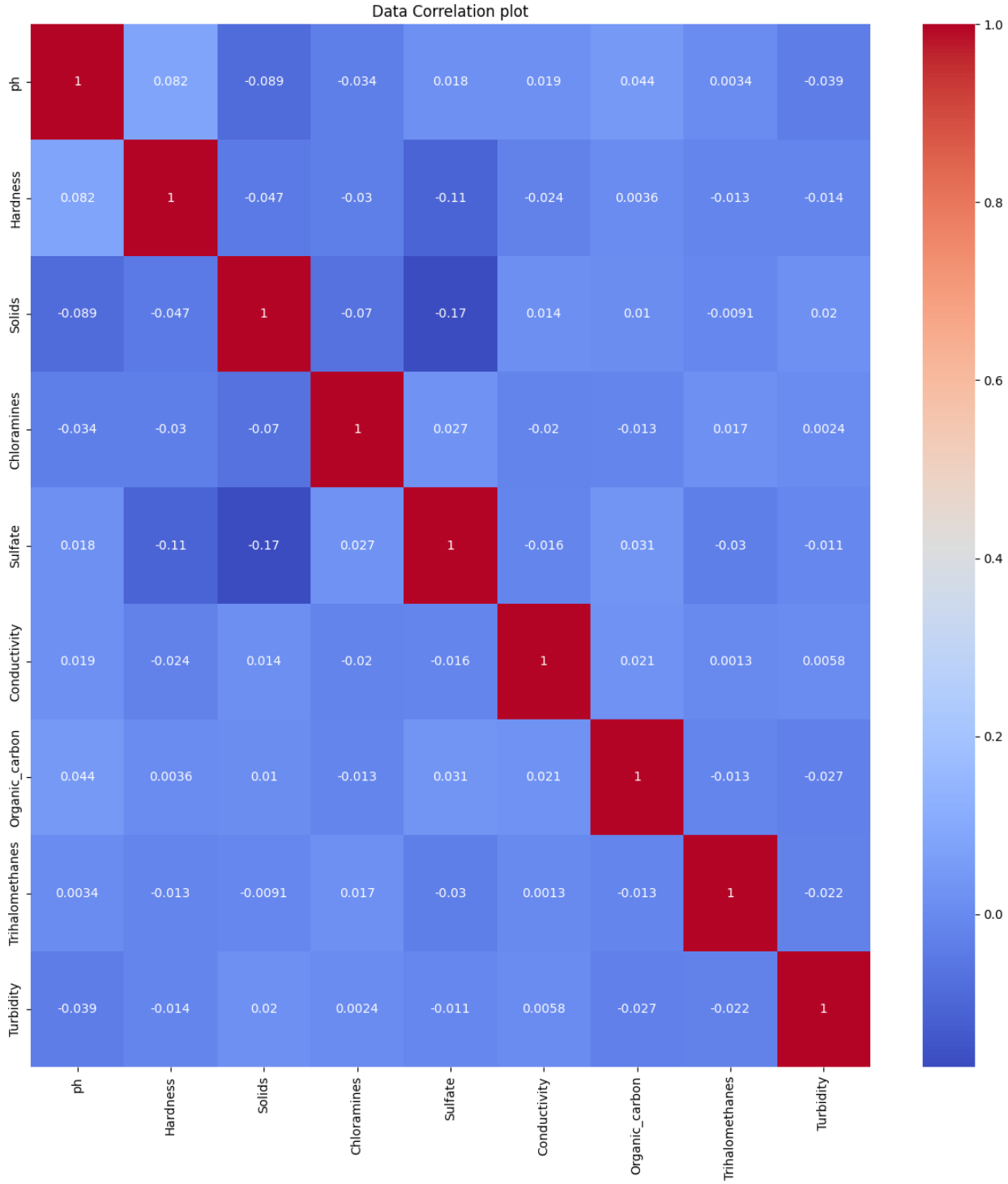
Link:
https://www.kaggle.com/datasets/adityakadiwal/water-potability/data

Number of columns: 10
Number of features: 9

The dataset used in this study, which was collected from Kaggle, has a total of 3276 rows, each of which contains details on different types of water samples. Ten columns, each containing a variety of features that shed light on the properties of the water sources, are used to represent these samples. The dataset contains nine unique variables with a focus on forecasting water potability that are important in establishing the caliber and safety of the water. The dataset's

size—3276 instances—indicates a sizable amount of data that allows for thorough analysis and model training. This dataset's label is binary, comprising two classes denoted by the numbers 0 and 1. This dataset's label is binary, comprising two classes denoted by the numbers 0 and 1. This labeling practice mimics the binary classification challenge involved in determining the potability of water, where models are used to discriminate between samples of water that are drinkable (1) and those that are not. In order to assure the safety of drinking water sources, precise and trustworthy prediction models must be developed. The dataset's diversity, quantity, and explicit binary classification targets all contribute to this. The preprocessing procedures involve several steps, including handling missing values, randomizing the dataset, and dividing it into training and testing subsets. we imputed the missing values with the most frequent values of that feature.

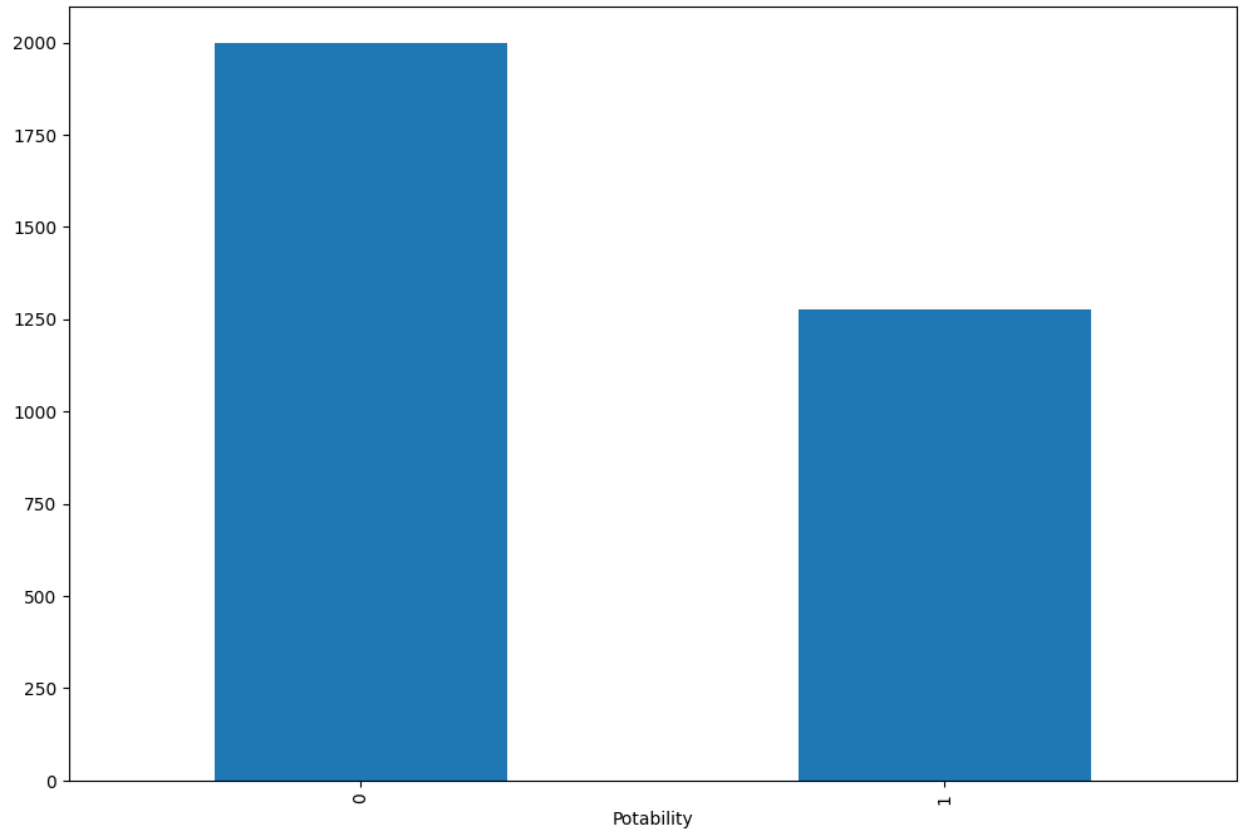
Correlation of the features along with the label/class:



The image presents a correlation matrix heatmap, a visual tool that elucidates the statistical relationship between various water quality parameters. Each square's color intensity and tone reflect the correlation magnitude between the water features. The deepest red signifies a perfect positive correlation of "1.0," where two features increase in unison, indicating a direct and proportional relationship. In contrast, the shades of blue depict negative correlations; for example, a value of "-0.17" implies an inverse relationship, where an increase in one feature may correspond to a decrease in the other. This heatmap is particularly insightful for identifying which water quality factors are interrelated, thus enabling a more informed selection of features for models aimed at predicting water potability. It's a pivotal step in data analysis, ensuring that the models used are not only accurate but also based on meaningful and non-redundant inputs, thereby enhancing their predictive capability.

Biased/Balanced:

As we can see in this bar chart, the data set is biased towards one class over the other.



Data Preprocessing:

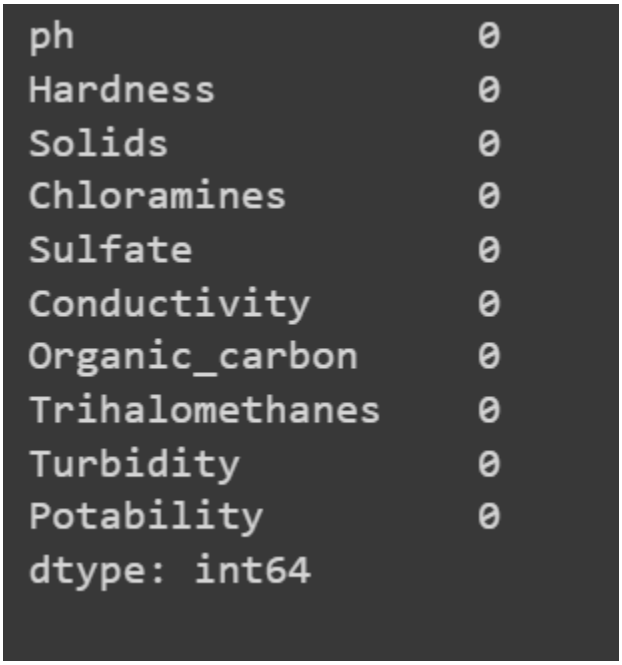
Problem 1: There are 491 null values in a column named ‘ph’, 781 null values in the column 'Sulfate',’ and 162 null values in the column 'Trihalomethane’ .

Solution: To handle the missing data in the dataset, we have implemented imputation by substituting the most frequent values in place of null values.

The number of null values in each column is:

```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

Before pre-processing



After pre-processing

Dataset splitting:

To train the model, data splitting was done at 70% for the training model and 30% for testing.

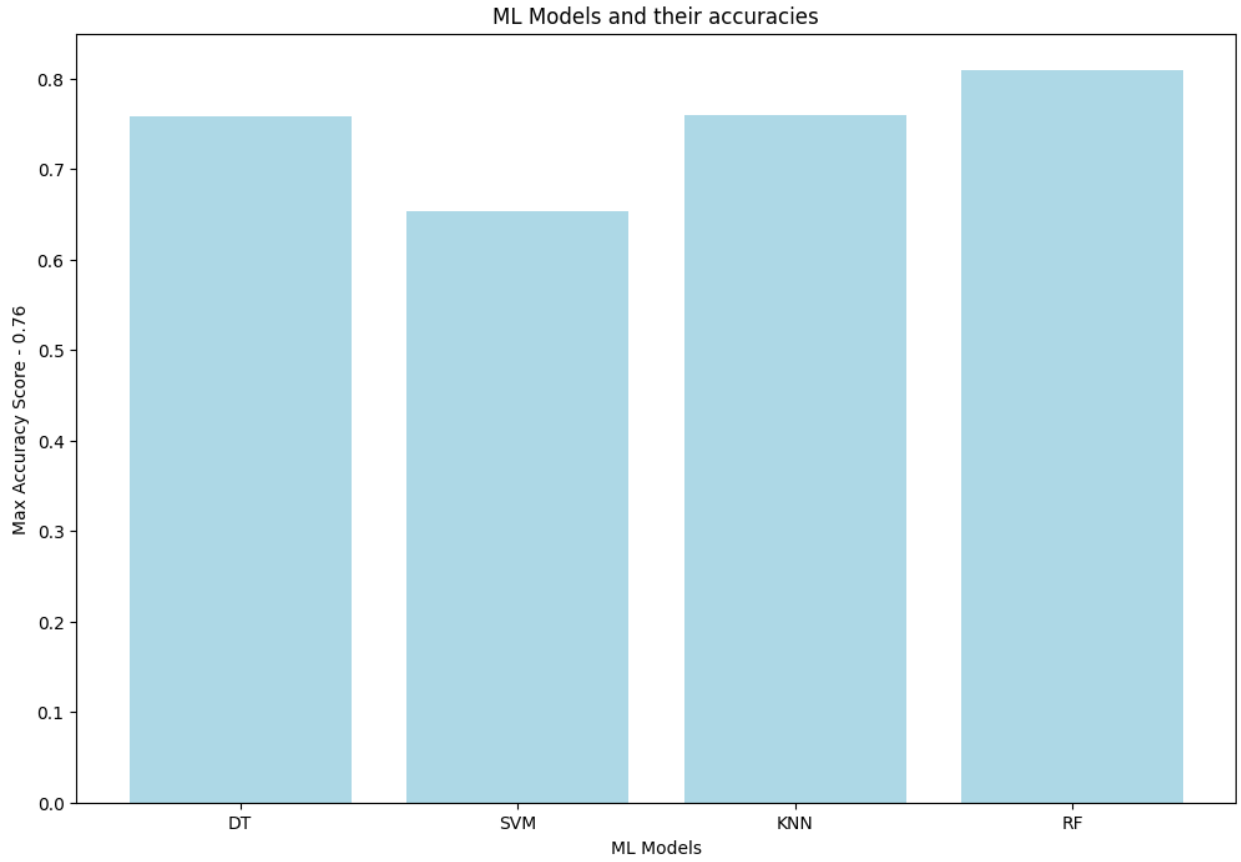
On this basis- the training data set - 2294 and the testing data set - 983.

Model Training:

Model Name	**Accuracy (%)**	**Error(%)**
Support Vector Machines (SVM)	65.38%	34.62%
K- nearest neighbors (KNN)	76%	24%
Decision Tree Classifier	75.75%	24.25%
Random Forest	80.90%	19.1%

From the table, in comparison to the Decision Tree (75.75%), KNN (76%), and Random Forest (80.90%) models, the SVM model demonstrates a comparatively lesser accuracy of 65.38%.

Model selection/Comparison analysis:



The support vector machine (SVM) model displays a comparatively lowered accuracy of 65.38% in contrast to the decision tree (75.75%), KNN (76%), and random forest (80.90%) models.

Model Testing:

To taste the model, the following data was used-

Result for SVM

```
Accuracy: 0.65375
Precision: 0.6890756302521008
Recall: 0.5970873786407767
```

Here, accuracy is about 65.375%.

Result for the Decision Tree

```
Accuracy: 0.7575
Precision: 0.7261410788381742
Recall: 0.8495145631067961
```

Here, accuracy is about 75.75%.

Result for KNN

```
AccuracyKNN: 0.76
Precision: 0.7330508474576272
Recall: 0.8398058252427184
```

Here, accuracy is about 76%

Result for Random Forest

```
AccuracyRandomForest: 0.8090075062552127
PrecisionRandomForest: 0.8285229202037352
RecallRandomForest: 0.7922077922077922
```

Here, accuracy is about 80.90%.

Conclusion:

We have rigorously developed and assessed a multitude of machine learning models in the course of our investigation into water potability. By employing algorithms including Support Vector Machines (SVM), k-Nearest Neighbours (KNN), Random Forest, and Decision Trees, we have successfully attained a variety of accuracy levels, thereby emphasizing the models' efficacy in the assessment of water safety.

With a notable accuracy rate of 80.90%, the RandomForest model showed its remarkable durability in the domain of potable water classification. Completing at 76% and 75.75% precision, respectively, were the KNN and Decision Tree models, which also performed admirably. Despite displaying a marginally diminished accuracy of 65.38%, the SVM model remains a potentially effective predictive instrument that could be enhanced further.

The implications of these results underscore the potential of machine learning methods in utilizing critical water quality indicators to forecast potability. Although the initial findings show promise, further investigation is required to achieve greater precision and applicability. For the purpose of improving the accuracy and practical applicability of the models, it is recommended that future research incorporate larger datasets and employ more advanced validation techniques.

As a result of the insights gained from this study, machine learning in water quality assessment can be implemented practically in the future. This development holds substantial ramifications in terms of enhancing global public health's most vital goal, namely, the provision of safe drinkable water. An essential element of effective public health policy and a fundamental human right, access to potable water is something that AI has the potential to facilitate on a global scale. This is demonstrated by the findings of this study.

Future Work:

To enhance the effectiveness of machine learning models employed in water potability classification, future investigations should give precedence to the collection of more comprehensive and diverse datasets that accurately reflect the entire spectrum of water sources. In addition, innovative methodologies for feature extraction could potentially have a significant impact on enhancing the model's accuracy by detecting nuanced fluctuations in water quality parameters. While advanced machine learning approaches, like deep learning, demonstrate promise in tackling complex classification challenges, they present considerable barriers attributable to their computational demands and dependence on large datasets.

Ensemble learning techniques, which connect insights from numerous models, offer a potentially efficacious strategy for enhancing the precision of water potability predictions. The implications of these models' potential to enhance clinical decision support systems and offer valuable insights extend beyond the realm of public health. Pertaining to this matter, these systems may provide assistance in the detection of water sources that present a potential hazard to public health, propose strategies for remediation, and track the efficacy of interventions over time.

In addition to its immediate consequences, the implementation and progression of machine learning models possess the capacity to substantially enhance the field of water quality

management. The utilization of these models in practice holds the ability to significantly increase the dependability and precision of potability assessments, thereby playing a crucial role in safeguarding public health and ensuring the accessibility of potable water.