1 Introduction

In this study, we investigate the analysis and prediction of water quality potability using a comprehensive dataset encompassing water quality metrics from 3276 distinct water bodies. Our methodology begins with an initial exploration involving standard descriptive statistics and visualization techniques, including variable distributions, histograms, and outlier detection. Furthermore, we delve into correlations among variables to discern any significant relationships. Subsequently, we undertake preprocessing steps to handle missing data, ensuring the dataset's readiness for modeling. Following this, we conduct a link analysis to probe potential similarities in pH levels across different water samples, aiming to uncover underlying patterns or relationships among water bodies. This additional investigation enriches our understanding of factors influencing water quality and potability, aiding in model refinement. We then proceed to develop predictive models using the training dataset, employing appropriate techniques such as data splitting for model training and evaluation. Finally, we evaluate the model's accuracy using the test dataset, providing insights into its predictive performance and potential implications for water quality management. This comprehensive approach contributes to the advancement of predictive modeling in water quality assessment, with implications for environmental health and resource management.

2 Dataset

1. **pH value**: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. **Hardness**: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. **Solids (Total dissolved solids - TDS)**: Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. **Chloramines**: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. **Sulfate**: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. **Conductivity**: Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μ S/cm.

7. **Organic_carbon**: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA ¡ 2 mg/L as TOC in treated / drinking water, and ¡ 4 mg/Lit in source water which is use for treatment.

8. **Trihalomethanes**: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. **Turbidity**: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. **Potability**: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

3 Descriptive Analysis

3.1 Null values

Upon inspecting the first 5 rows of the dataset using `data.head()` , we observed the presence of NULL values. Further examination revealed that NULL values are present in the columns: *pH*, *Sulfate*, and *Trihalomethanes* as showed in [Figure 1].

Notably, all attributes are numerical and of type float64, except for the target variable *potability*, which is of type int64.

3.2 Statistics Descriptions

	count	mean	std	min	25%	50%	75%	max
ph	2.785	7,08	1,59	0,00	6,09	7,04	8,06	14,0
Hardness	3.276	196,37	32,88	47,43	176,85	196,97	216,67	323,12
Solids	3.276	22.014,09	8768,57	320,94	15.666,69	20.927,83	27.332,76	61.227,20
Chloramines	2.495	7,12	1,58	0,35	6,13	7,13	8,11	12,13
Sulfate	3.276	333,78	41,42	129,00	307,70	333,07	359,95	481,03
Conductivity	3.276	426,21	80,82	181,48	365,73	421,88	481,79	753,34
Organic_carbon	3.114	14,28	3,31	2,20	12,07	14,21	16,56	28,30
Trihalomethanes	3.276	66,40	16,18	0,74	55,84	66,62	77,34	124,00
Turbidity	3.276	3,97	0,78	1,45	3,43	3,96	4,50	6,74
Potability	3.276	0,39	0,49	0	0	0	1	1

Table 1: Statistics Descriptions

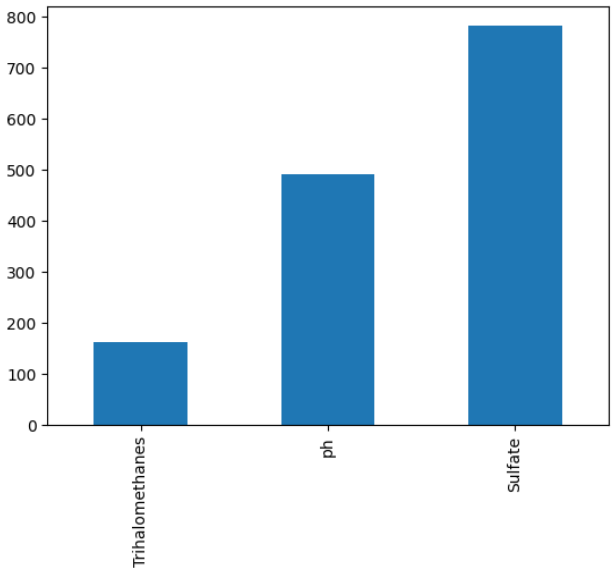

Figure 1: Null values

Upon examining the table, it becomes evident that the standard deviation of solids is notably high, suggesting potential anomalies within the dataset. Through graph plotting, it was observed that attributes such as pH, Hardness, Chloramines, Sulfate, and Organic Carbon adhere to a normal distribution. However, other attributes exhibit significant skewness and kurtosis, complicating the establishment of their normal distribution.

Remarkably, the dataset contains outliers, but their impact is minimal, allowing for their retention without compromising data integrity [Figure 2].

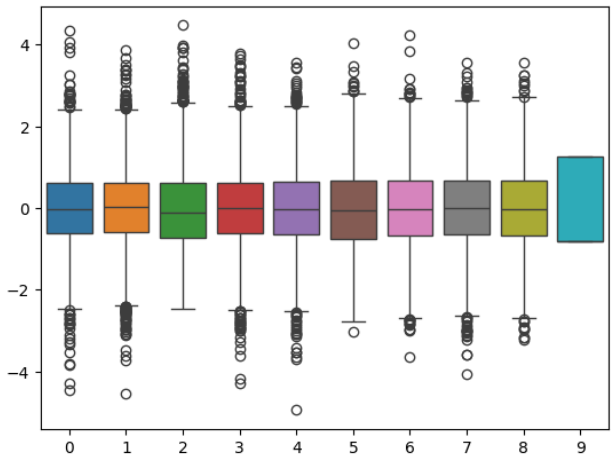

Figure 2: Outliers

Furthermore, a heatmap analysis was conducted to explore correlations between attributes. While correlations with the target variable, potability, are limited, positive and negative correlations exist among different columns, with an overall negligible average correlation [Figure 3].

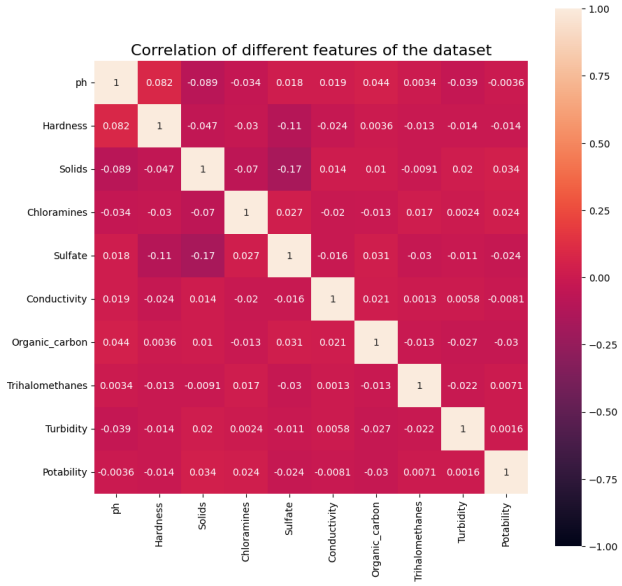

Figure 3: Correlations between the values

3.3 Problems

After analyzing the data from a statistical perspective, we intend to carefully examine whether they respect the prescribed standards:

- $6,5 < ph < 8,5$
- $Hardness < 100\,mg/L$
- $Solids < 1200\,mg/L$
- $Chloramines < 4\,mg/L$
- $Sulfate < 250\,mg/L$
- $Conductivity < 400\,\mu S/cm$
- $Organic\_carbon < 2\,mg/L$
- $Trihalomethanes < 80\,ppm$
- $Turbidity < 5,00\,NTU$

Through this script, we tried to verify whether water classified as potable actually met the strict standards set by the WHO. However, the results obtained did not meet expectations. In each category, 40% of the water considered potable was actually non-potable. These cases all represent false positives.

4 Link Analysis

In this section, we created a large graph where each node 'i' is connected to node 'j' if the pH value falls within the range of x to $x + 1$ (excluding $x + 1$). Null values for pH were removed

before constructing the graph. Our aim was to explore if the graph exhibited any resemblance to the Jaccard distance metric. However, upon analysis, it became apparent that the graph does not display any intersections between nodes. Consequently, the distance between all nodes remains equal to 1.

5 Predictive models

For our predictive modeling phase, we began with data preprocessing, which involved removing entries containing null values and splitting our dataset into an 80% training set and a 20% test set. Subsequently, we employed several supervised learning models, namely logistic regression, Support Vector Machine (SVM), decision tree, random forest, and sequential Artificial Neural Network (ANN). However, despite utilizing these models, our predictive accuracy remained disappointingly low across all methods. Specifically, our logistic regression model achieved an accuracy of 0.64, while SVM performed slightly better with an accuracy of 0.72. The decision tree model yielded an accuracy of 0.62, and the random forest model showed a slightly higher accuracy of 0.71. These results indicate the challenges encountered in accurately predicting water quality potability using the chosen models and dataset.

Despite achieving a 60% Accuracy, which is notably low, a closer examination of the confusion matrix reveals a troubling pattern: the model consistently misclassifies nearly all test points as false (0). This significant imbalance renders the model unsuitable for practical use.

SVM: Here again, even after getting an accuracy of 71%, by seeing the confusion matrix we can conclude that the model is not predicting even the 71% evenly. It is giving more importance to one side than the other.

Decision Tree: Similar to the output we got from SVM, the accuracy is 61%, but the model favours one side again.

Random forest: Again, the model is leaning towards one of the two possible states (true or false) and ignoring the rest despite getting a 70% accuracy.

In the ANN we have first set the hyperparameter values. We are using the sigmoid funtion as the output layer function and the tanh or the hyperbolic tangent activation function to yeield the optimum result. For the number of neural layers, we are keeping it low - three. We have found out that the epoch of 500 is giving the best output. We tried others, but all of them over train or under train the model. Here we can see that the accuracy is not that great (63%), we can see the predictions are somewhat correct for true positive and true negative values.

6 Conclusions

Our task was to implement a data streams algorithm, but given the time constraints and the fact that the water samples originate from 3176 distinct bodies of water, this endeavor proved unfeasible.

Throughout our analysis, indications surfaced suggesting potential issues with the dataset, prompting further investigation to confirm their presence. We attempted to conduct link analysis to identify similarities among the pH values of the water samples. However, our findings revealed no discernible patterns or commonalities.

Subsequently, we sought to leverage machine learning models to predict water potability. Regrettably, as with previous analyses, our efforts were thwarted by the dataset's inconsistencies and disorderliness. These challenges underscore the complexity of the data and the limitations of our current approaches in effectively modeling and predicting water quality potability.