**DATA-2206-01 – CAPSTONE**

**FINAL DATA ANALYSIS REPORT**

WATER POTABILITY ANALYSIS

**A glass of water being poured into a glass

Description automatically generated with low confidence**

IS YOUR WATER SAFE TO DRINK?

SUBMITTED TO:

DAVID BABALOLA

SUBMITTED BY:

ARSHIYA MUSLIM NOOR (100839688)

DIMPLE DINESH PARMAR (100852388)

PRADEEP RAJASHEKARAN (100848443)

SUBMISSION DATE:

24TH July 2022

TABLE OF CONTENTS

[INTRODUCTION 3](#_Toc109620781)

[DATA ANALYSIS GOAL 3](#_Toc109620782)

[DATA DESCRIPTION 3](#_Toc109620783)

[DATA PREPARATION 4](#_Toc109620784)

[DATA PRE-PROCESSING 4](file:///C:\Users\coolb\Desktop\Final%20Project%20Files\Final_Project_Report_Team14.docx#_Toc109620785)

[EDA 4](#_Toc109620786)

[HANDLING OUTLIERS 4](#_Toc109620787)

[NULLS VALUES 4](file:///C:\Users\coolb\Desktop\Final%20Project%20Files\Final_Project_Report_Team14.docx#_Toc109620788)

[DUPLICATED VALUES 5](#_Toc109620789)

[DATA MODELLING AND METHODOLOGY 5](#_Toc109620790)

[1. DATA SCALING 5](#_Toc109620791)

[3. MODELING 5](#_Toc109620792)

[4. COMPARISON OF MODELS 5](#_Toc109620793)

[5. MODEL SELECTION 6](#_Toc109620794)

[VISUALIZATION ON AN APP 7](#_Toc109620795)

[CONCLUSION 8](#_Toc109620796)

[REFERENCES 8](#_Toc109620797)

## INTRODUCTION

**Safe water is important** for public health, whether it is used for drinking, food production or recreational purposes. Absent, inadequate, or inappropriately managed water and sanitation services expose individuals to preventable health risks. A key UN report indicates that water shortages will affect 2.3 billion people or 30% of the world's population in major nations by 2025. Already, the crisis of potable water in most developing countries is creating public health emergencies of staggering proportions. Whether this is achievable within the stated time is debatable, but it clearly delineates the state of the world we live in.

We took some inspiration from this to use this **Water Quality dataset** to understand what constitutes to safe drinking water and apply machine learning approaches to it to distinguish between **Potable and Non-Potable water**. The dataset is synthetically generated and has various variables like Ph, hardness, solids, or chemicals that are affecting the level of water potability. The analytics of machine learning techniques are very suitable for modeling and understanding the internal relation between the water quality components, and modeling to predict if water is safe to drink. This notebook will explore the different features related to water potability, modeling, and predicting water potability. (KONSTANTIN, 2021) (JAY, 2021)

## DATA ANALYSIS GOAL

The goal of this project is to perform exploratory data analytics and machine learning prediction analytics on a variety of water quality indicators, such as pH, hardness, carbon, solids etc. The aim of this study is the prediction of water potability using machine learning algorithms including Logistic regression, SVM, Decision tree, Adaboost classifier, K-neighbors, XGB, and random forest and finally using the most accurate model to use for future predictions. The significant sources of water contamination can be determined using various factors and the findings may serve as useful predictors for evaluating if the **water quality is fit for human consumption**. This analysis will provide us the following outcomes –

* Predict whether the water quality is feasible for human consumption.
* Anticipate the relevant measures to be taken if the water is contaminated.
* Find features that contribute the most for water to be potable.

## DATA DESCRIPTION

The dataset used for this analysis is a synthetically generated one which is taken from the Kaggle website.

The dataset consists of 3276 rows of recorded instances and has 9 columns of independent variables which are affecting the potability results or dependent variable noted in the 10th column. A description of these features are listed below. (KADIWAL, 2021)

* **ph**: pH of 1. water (0 to 14).
* **Hardness**: Capacity of water to precipitate soap in mg/L.
* **Solids**: Total dissolved solids in ppm.
* **Chloramines**: Amount of Chloramines in ppm.
* **Sulfate**: Amount of Sulfates dissolved in mg/L.
* **Conductivity**: Electrical conductivity of water in μS/cm.
* **Organic carbon**: Amount of organic carbon in ppm.
* **Trihalomethanes**: Amount of Trihalomethanes in μg/L.
* **Turbidity**: Measure of light emitting property of water in NTU.
* **Potability**: Indicates if water is safe for human consumption. Potable - 1 and not potable - 0

## DATA PREPARATION

The initial step in data preparation is **data cleansing**. We have detected and corrected corrupt or inaccurate records from our dataset following the steps below.

Graphical user interface

Description automatically generated Graphical user interface

Description automatically generated

### DATA PRE-PROCESSING

By WHO's standard, the TDS should be between 50 to 300 ppm, but the dataset contains value that goes unto 50,000 ppm for TDS. As the dataset is synthetically generated, we assume the values were generated with incorrect decimal placement. Hence, we shifted the decimal to correct it for our analysis for TDS column data to fix the structural errors.

EDA **-** To explore the features further, we performed a univariate, bivariate and multivariate analysis of our dataset. We found that most features are normal distributed with very little skewness. There is minimal to no correlation between any of the features. Hardness and Ph seems to be the most significant features in comparison to the rest while determining the potability. However, there is no strong correlation of the features with portability.

**Graphical user interface

Description automatically generated**

**Chart, bar chart

Description automatically generated**

HANDLING OUTLIERS **-** An outlier is an observation that lies at an abnormal distance from other values in a random sample of a population. We plotted a boxplot to notice that there are few outliers in the data. Hence, removed the outliers to further process the data. After removing the outliers, the dimensions of our data are 2951 rows and 10 columns. The below image shows that 2951 rows are found after removing the 325 rows of outliers.

Graphical user interface, text

Description automatically generated

NULLS VALUES **-** In this dataset we have found that there are 1434 rows of missing value which constitutes 4.4% of overall data. So, we have changed all the missing values to the median of their values, so the overall value doesn’t affect.

* PH - 14.98% | 7.0367 (median) 7.0807 (mean)
* Sulphate - 23.84% | 333.0735 (median) 333.7757 (mean)
* Trihalomethanes - 4.94% | 66.6224 (median) 66.3962 (mean)

Graphical user interface, text, application

Description automatically generated

**DATASET AFTER CLEANSING:**

**A picture containing graphical user interface

Description automatically generated**

DUPLICATED VALUES **-** There are no duplicate values that have been detected in our data. Hence, we did not have to remove any for our analysis.

## DATA MODELLING AND METHODOLOGY

1. DATA SCALING **-** This is important because scaling can ensure that one factor will not impact the model just because of its large magnitude.

Table

Description automatically generated

1. **SPLITTING THE TRAIN AND TEST** - This must be done to prevent our model from overfitting. We have assigned 20% of the data for testing. The dimension of the train data is (2360, 9) and the test data is (591, 9) after split.
2. MODELING**:** The purpose of this study is to predict the potability of water utilizing machine learning approaches, such as **Logistic regression, SVM, Decision tree, Adaboost classifier, K-neighbors, XGB, and random forest**. Once the splitting of the train and test is done our next step is to build each model and look at the output results and compare between the models to find which one suits the best for the dataset. Each of these models has its own advantages and we have selected these since they are all good at binary classification. So, performed an analysis using each of the models and come up with a better accurate model which can be used for future use.

### COMPARISON OF MODELS

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application

Description automatically generated

The table shows us the precision and accuracy of each model results.

### MODEL SELECTION

From the above results, we can clearly see that support vector machines and Random Forest have outperformed all the models. However, we will consider SVM as our final model.

SVM has given 67% in precision which is a good result. we got 67%, which says that among the predictions of the water being potable the model has predicted 67% of the trueness of being potable. The overall accuracy of the model is 66% which shows the model has performed well. In our analysis, we do have to consider the precision because our aim is to find the trueness among the actual potable water. The results below shows the classification report and confusion matrix for the SVM model. In confusion matrix, each outcome is explained below –

* **True Negative** - 58% of water is non-potable and shows non-potable.
* **False Positive** - 3.72% of water is non-potable but predicted to be potable. In precision, FP is in the denominator, and it inversely affects the precision. The lower the FP is, the higher the precision is. Hence, precision is what we will evaluate our model with.
* **False Negative** - 30.46% of water is potable but shows non-potable.
* **True Positive** - 7.61% of water is potable and shows potable.

Table

Description automatically generated

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

## VISUALIZATION ON AN APP

Using streamlit, we build an app to provide an interface to predict whether the water is potable and safe for human consumption. This will help any person to enter the values and receive an output without knowing the technical part of the backend. We have used the SVM algorithm for modeling and providing outputs on this interface.

Graphical user interface, application

Description automatically generated

The files used for this app have been saved on github repo in the following link - [*https://github.com/arshii-anjum/Capstone*](https://github.com/arshii-anjum/Capstone)

## CONCLUSION

In conclusion, we can say that support vector machines have performed well than the other models by giving out 66% of accuracy. So, we can use this model for future predictions in finding the water fit for drinking. The advantage of the support vector machine is it works comparably well when there is an understandable margin of dissociation between classes. And it is more productive in the high dimensional spaces. Also, it has effective in instances where the number of dimensions is larger than the number of specimens. It is also memory efficient because it only needs a small fraction of training points for the decision function (known as support vectors). The kernel trick is the real strength of SVM.

The limitations are a few and stated below.

* This dataset is a synthetically generated and not a real dataset and due to this there are some limitations. This can be solved using a real dataset.
* For huge data sets, the support vector machine approach is not suitable.
* When the target classes overlap and the data set has more sound, it does not operate very well.
* The support vector machine will perform poorly when there are more attributes for each data point than there are training data specimens.
* There is no probabilistic explanation for the classification because the support vector classifier places data points above and below the classifying hyperplane.
* To build the most sophisticated model, we can perform hyper-parameter tuning.

# REFERENCES

JAY. (2021, July). *Water Quality: Analysis (Plotly) and Modelling*. Retrieved from Kaggle: https://www.kaggle.com/code/jaykumar1607/water-quality-analysis-plotly-and-modelling

KADIWAL, A. (2021). *Kaggle*. Retrieved from Water Quality: https://www.kaggle.com/datasets/adityakadiwal/water-potability

KONSTANTIN, T. (2021, September). *Exploring and Predicting Drinking Water Potability*. Retrieved from Kaggle: https://www.kaggle.com/code/thomaskonstantin/exploring-and-predicting-drinking-water-potability

Leech, J. (2020, June 30). *7 Science-Based Health Benefits of Drinking Enough Water*. Retrieved from Healthline: https://www.healthline.com/nutrition/7-health-benefits-of-water

*Week 5: Machine Learning Fundamentals*. (n.d.). Retrieved from DC Connect: https://durhamcollege.desire2learn.com/d2l/le/content/430948/viewContent/5745823/View