JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Machine Learning Algorithms Comparative Study

[1]Suppawit Glawkate, [2]Lakhshan Sarvanan, [3]Pragati Prashant Tolamatti

[1]Student, [2]Student, [3]Student
[1]School of Computer Science and IT,
[1]Jain Deemed-to-be University, Bangalore, India

Abstract : Water quality index (WQI) is widely used to assess and classify groundwater and surface water quality and plays an important role in water resource management. In recent decades, artificial intelligence (AI) in the form of machine learning models has been applied to solve various environmental engineering problems, such as water quality modeling. Machine learning models represent a major innovation in research for monitoring and controlling a variety of engineering processes, using their algorithms. The proposed models have been evaluated and tested using certain statistical parameters. The results have revealed that the Random Forest algorithm has achieved the highest accuracy level compared to other machine learning moles like decision tree, MLP and gradient boosting.

IndexTerms **– Artificial Intelligence, Machine Learning, Water Quality Index,**

I. INTRODUCTION

Water is the most important resource of life, essential for the survival of most organisms and humans. Living organisms require water of sufficient quality to sustain life. There are certain limits to the contaminants that a water species can tolerate. Exceeding these limits affects the existence of these organisms and is life-threatening.

Most surrounding bodies of water such as rivers, lakes and streams have specific quality standards that indicate their quality. There are also standards for water specifications for other uses. Water quality for industrial applications also requires different characteristics based on specific industrial processes. Some of the freshwater resources such as groundwater and surface water are natural water resources. However, such resources can be contaminated by human/industrial activities and other natural processes.

Access to safe drinking water is essential to good health, a fundamental human right and part of effective public health policy. This is an important issue at national, regional as well as local levels regarding health and development. In some regions, investments in water and sanitation have been shown to yield net economic returns. This is because the adverse health effects and reduced healthcare costs outweigh the costs of implementing the intervention.

Water Quality Index (WQI) is widely used to assess and classify groundwater and surface water quality and plays an important role in water resource management. This indicator integrates multiple physical and chemical factors into a single parameter to quantify the water quality state. Calculating this index provides an efficient approach to assessing water quality.

All these applications are based on the general concept of Water Quality Index proposed by the US National Sanitation Foundation (NSF), NSFWQI being the most widely used method worldwide.

In recent decades, artificial intelligence (AI) in the form of machine learning models has been applied to solve various environmental engineering problems, such as fresh water quality modeling. Machine learning models represent a major innovation in research for monitoring and controlling a variety of engineering processes, using their algorithms. to make accurate predictions without the need for complicated processes. Machine learning models are based on data mining, where a subset of a data set (training data) is used to build an algorithm and another subset of the data set (test set) is used to improve predictive performance.

II. LITERATURE WORK

Yang explored different imputations techniques to handle a large amount of missing data in the Water Quality dataset obtained from Kaggle.com. The comparison between the effectiveness of different imputation techniques is done by applying Artificial Neural Network (ANN) on the dataset. The imputation models implemented include KNN imputation, Mean/Median imputation, Arbitrary value imputation, as well as deleting the missing data [5]. Although it was concluded that KNN imputation is that best approach for imputation of missing value for a small dataset and generate the most accurate test results with 64% test accuracy [5], ANN models would not perform well with a small dataset, such as the Water Quality dataset [4] which was used in the paper which contains a little over than 3000 records. Hence, using machine learning algorithms would be a more appropriate approach.

Additionally, the presence of a large amount of outliers in the dataset and the nature of the dataset being imbalanced was not addressed, this could affect the overall performance of the proposed models.
Radhakrishnan & Pillai presented a comparative study and analysis of water quality classification models via 3 machine learning algorithms, SVM, Decision Tree and Naïve Bayes. With 4 significant features being used to identify the water quality, pH, BOD, DO and EC, the decision tree algorithm was found to outperform SVM and Naïve Bayes with an accuracy score of 98.50% [3]. However, the proposed models can be improved by training with a larger dataset and fine-tuning the hyperparameters of the models to obtain greater results [3].

Uddin et al., determine robust and reliable machine learning models to predict the WQI (Water Quality Index) in the coastal area of Cork Harbour [1]. Predictions of WQI were obtained using 8 different machine learning algorithms, this include Linear Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Extreme Gradient Boosting, Extra Tree, and Gaussian Naïve Bayes. The tree-based and ensemble-based models were found to be the most effective in reducing the uncertainty of WQI predictions and surpass the others when evaluating the models performance with MSE, MAE, RMSE, R2, and PREI metrics [1]. It was noted that the prediction of WQI with other ML algorithms is necessary and must be carried out in further studies using temporal data attributes [1].

Aldhyani et al., developed artificial Intelligence algorithms for predicting WQI and water quality classification (WQC). The WQI prediction, done using deep learning algorithms like NARNET and LSTM, and WQC forecasting, done using machine learning algorithms like SVM, KNN, and Naive Bayes, were applied on a dataset with 7 significant parameters [2]. The proposed models were able to accurately predict the WQI and classify the water quality, resulting in 96.17% regression coefficient for NARNET and 97.01% accuracy for the SVM algorithm. The proposed models were applied on the data of the water quality collected only in India from 2005 to 2014 [2]. Hence, the models can be trained and applied on data from other countries to prove its reliability and efficiency.

III. PROPOSED METHODOLOGY

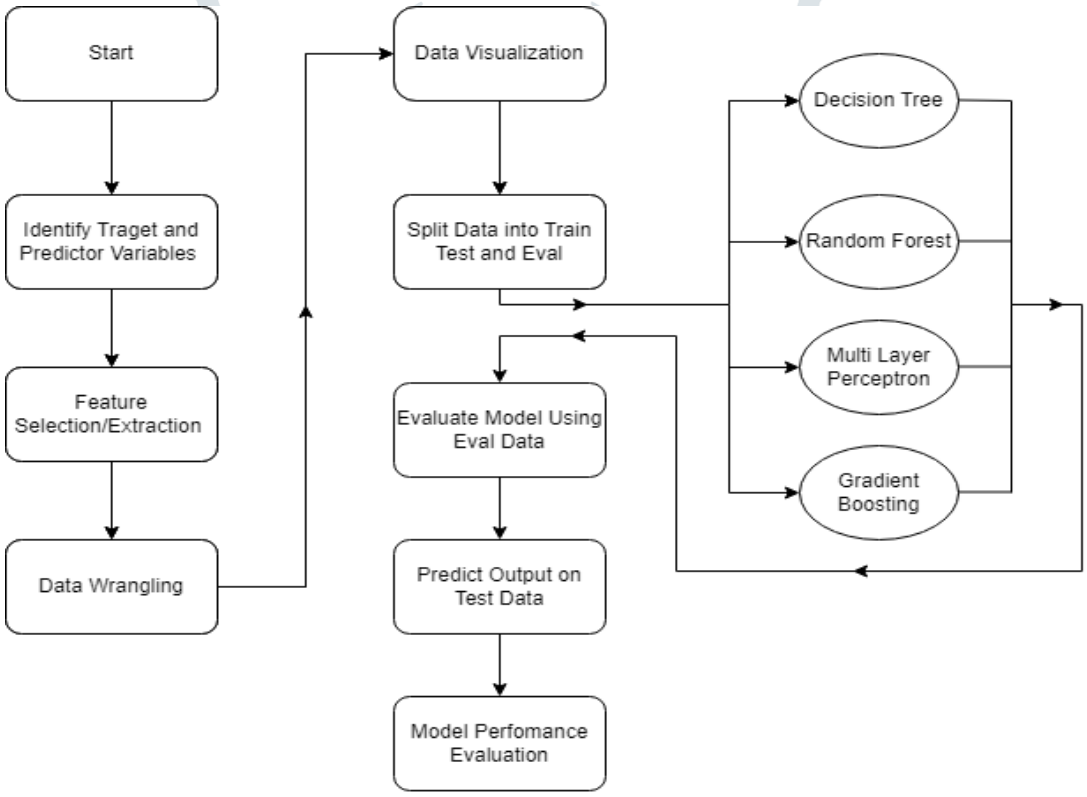

Figure 1. Machine Learning Pipeline

Experimental Dataset:
The dataset being used in this paper is the same used in Yang's work [5]. Water quality dataset is collected from Kaggle.com which consists of 10 columns and 3276 rows. The dataset constitutes 9 unique parameters used to assess the water quality from 3276 different water bodies and classify whether the water is safe for human consumption or not [4]. The 9 attributes and their ideal quantities and measurements are as followed:

Table 1. Permissible quantity/measurement of the features of the water quality dataset

Features	Ideal quantity/measurement in the freshwater supplies
pH value	6.52 - 6.83
Hardness	≤ 75 mg/L
TDS (Solids)	500 mg/L
Chloramines	4 mg/L
Sulfates	3 - 30 mg/L
Conductivity	≤ 400 µS/cm
Organic carbon	< 2 mg/L
Trihalomethanes	≤ 80 ppm
Turbidity	< 5.00 NTU

All features have been used for training different ML models. This dataset has one target variable, that is Potability, which indicates whether the water is safe for human consumption or not. The value of Potability is either 0 or 1, 0 means the water is not potable and 1 means the water is potable.

```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Figure 2. Water quality dataframe

Data pre-processing:

During the data exploration, missing values were found in 3 columns, pH, Sulfates and Trihalomethanes as shown by the below diagram. Sulfates attribute is found to have the most missing data with 23.84% of its data missing followed by pH and Trihalomethanes, with 14.99% and 4.95% respectively shown in figure 4.

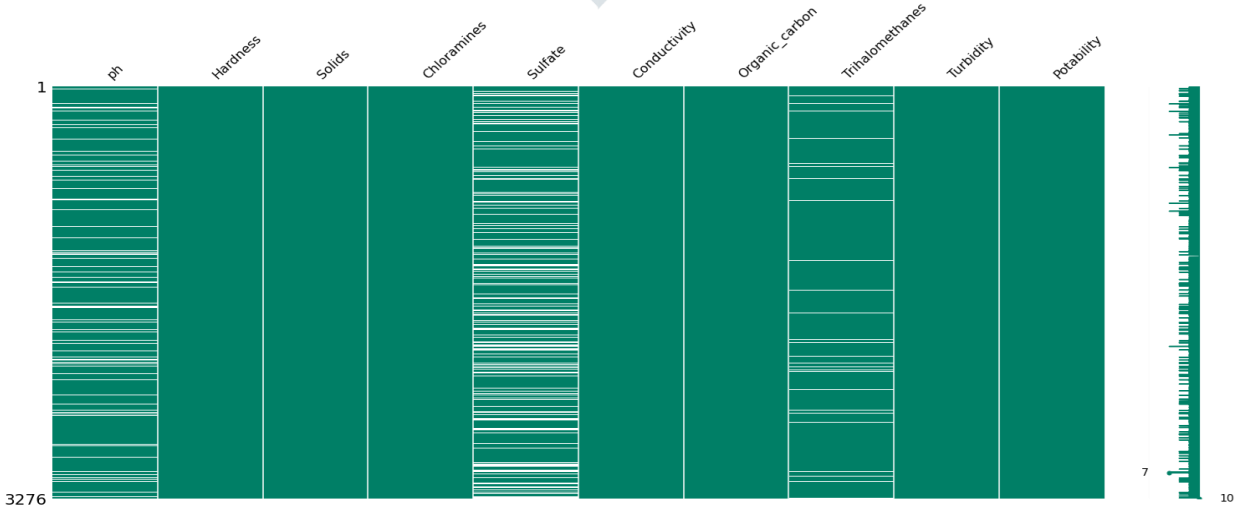

Figure 3. Missing no. matrix of water quality dataset

```
In [5]: df.isnull().sum()

Out[5]: ph                491
        Hardness            0
        Solids              0
        Chloramines         0
        Sulfate           781
        Conductivity        0
        Organic_carbon      0
        Trihalomethanes   162
        Turbidity           0
        Potability          0
        dtype: int64
```

```
In [6]: missing_col=['ph','Sulfate','Trihalomethanes']

In [7]: print('% of missing values\n')
        for col in missing_col:
            pct=(df[col].isnull().sum()/len(df))*100
            print(col+': {:.2f}%'.format(pct),'\n')

        % of missing values

        ph: 14.99%

        Sulfate: 23.84%

        Trihalomethanes: 4.95%
```

Figure 4. Number and percentage of missing data from the water quality dataset

To identify the best approach for handling the null values, the 3 features are visualized using box plots as shown below. Figure 5 shows the existence of a large number of outliers in all 3 features. Hence, the median of individual features were used for imputing the null values rather than using the mean as the mean value would be heavily influenced by the outliers.

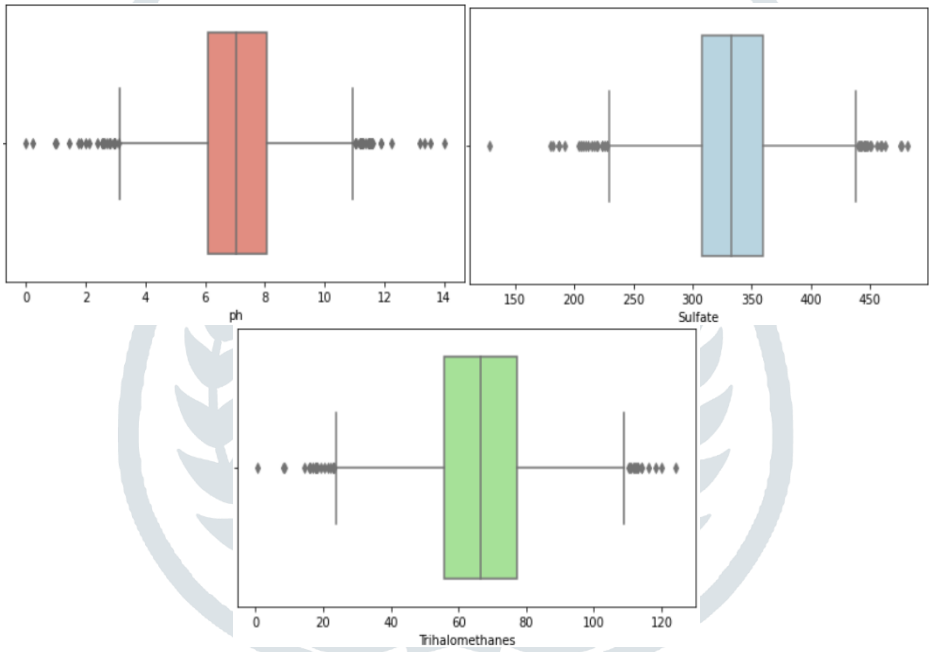

Figure 5. Box plots of pH, Sulphate and Trihalomethanes

Additionally, the medians of the features with null value are calculated according to the 2 different classes of "Potability", median of class 0 and of class 1 as shown by figure 6. This is done to avoid the data values of one class influencing the median of another class.

```
In [12]: #Imputing the null values with the median of the feature with respect to 'Potability'
         for col in missing_col:
             df[col].fillna(df.groupby(['Potability'])[col].transform('median'),inplace=True)
```

Figure 6. Handling null values

Furthermore, the distribution of the data of the rest of the features were also visualized using box plots and were found to have numerous outliers as well, as shown by figure 7.

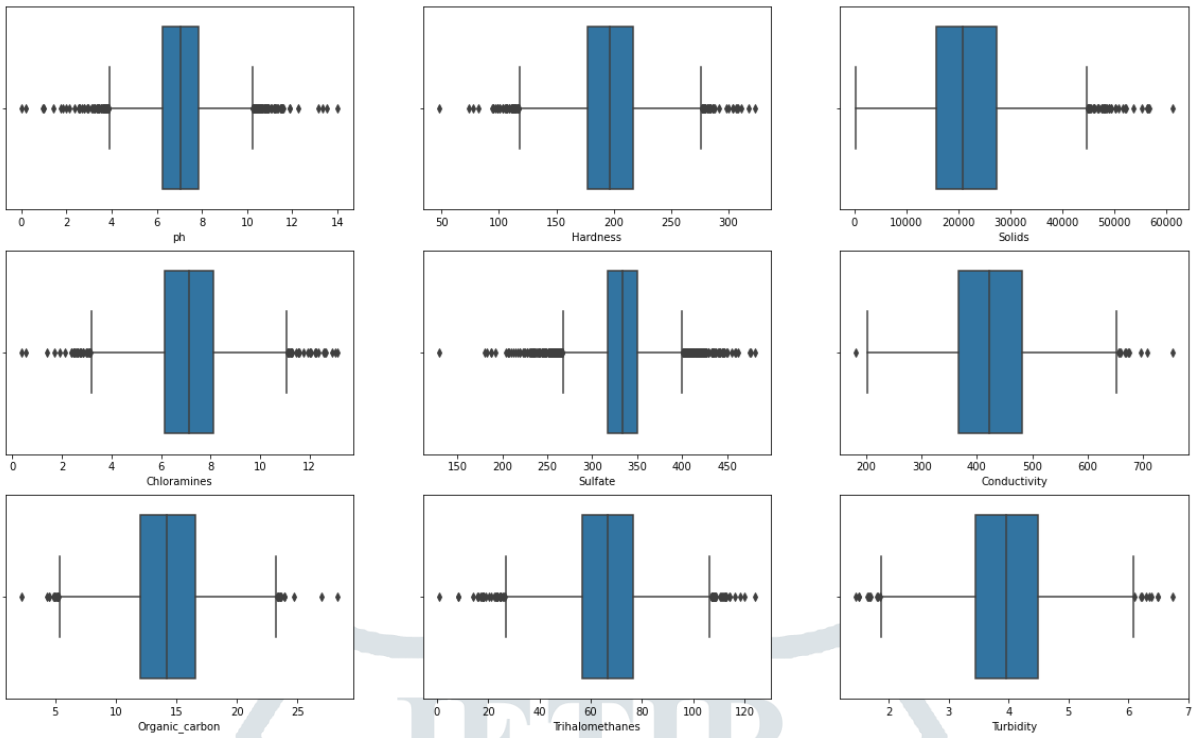

Figure 7. Box plots of all of the features in the water quality dataset

To tackle the problem, 2 methods were proposed - the following methods include:

1. Replacing the outliers with medians
2. Dropping the outliers

The 2 methods were, individually, applied on the original dataset to create 2 separate datasets, a dataset with all the outliers replaced by the medians and a dataset with all the outliers dropped. The outliers are detected using the IQR (Interquartile range) method. This is done by calculating the upper boundary and lower boundary using the upper quartile and the lower quartile respectively [9]. The values existing beyond the boundaries are considered as outliers.

$$IQR=Q_1-Q_3 \text{ [9]}$$

$$\text{(Lower boundary, Upper boundary)}=[(Q_1-(1.5 \times IQR)), (Q_3+(1.5 \times IQR))] \text{ [9]}$$

Moreover, a copy of the original dataset, consisting of the outliers, was created to be a controlled dataset for model performance comparison.

Figure 8 shows the percentage of the data from different water bodies classified as potable and not potable. It is clear that there is a significant difference between the number of data classified as class 1 (water is potable - safe for consumption) and class 0 (water is not potable - not safe for consumption) with 61% of class 0 constitute the majority of the dataset and only 39% are of class 1. Synthetic Minority Oversampling Technique (SMOTE) is proposed to handle the problem of imbalance data before training the ML models. SMOTE helps populate the dataset by resampling the existing data in the dataset [7]. This makes sure that the number of data belonging to the minority class, in this case class 1, is equal to the number of data classified as class 0.

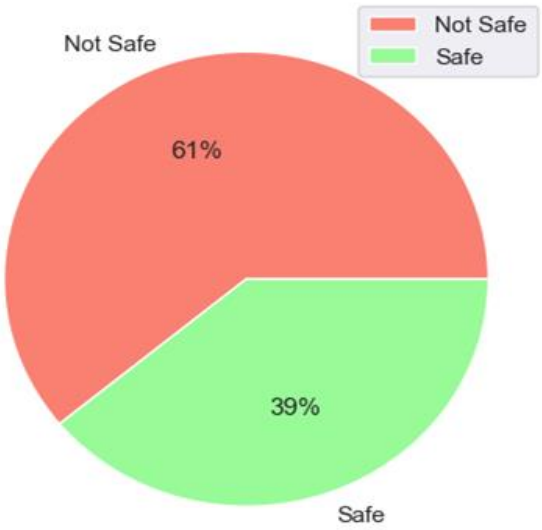

Figure 8. Proportion of class 0 (not safe for consumption - red) and class 1 (safe for consumption - green)

Although some features share the same unit of measurements, such as Hardness, TDS, Chloramines, Sulfates and Organic carbon all of which are measured in mg/L, others do not use the same scale, this includes pH, Conductivity, Trihalomethanes and Turbidity. Since most of the features seem to be normally distributed, as shown by the above diagram, the dataset is standardized to make all data scale-free and consistent [8].

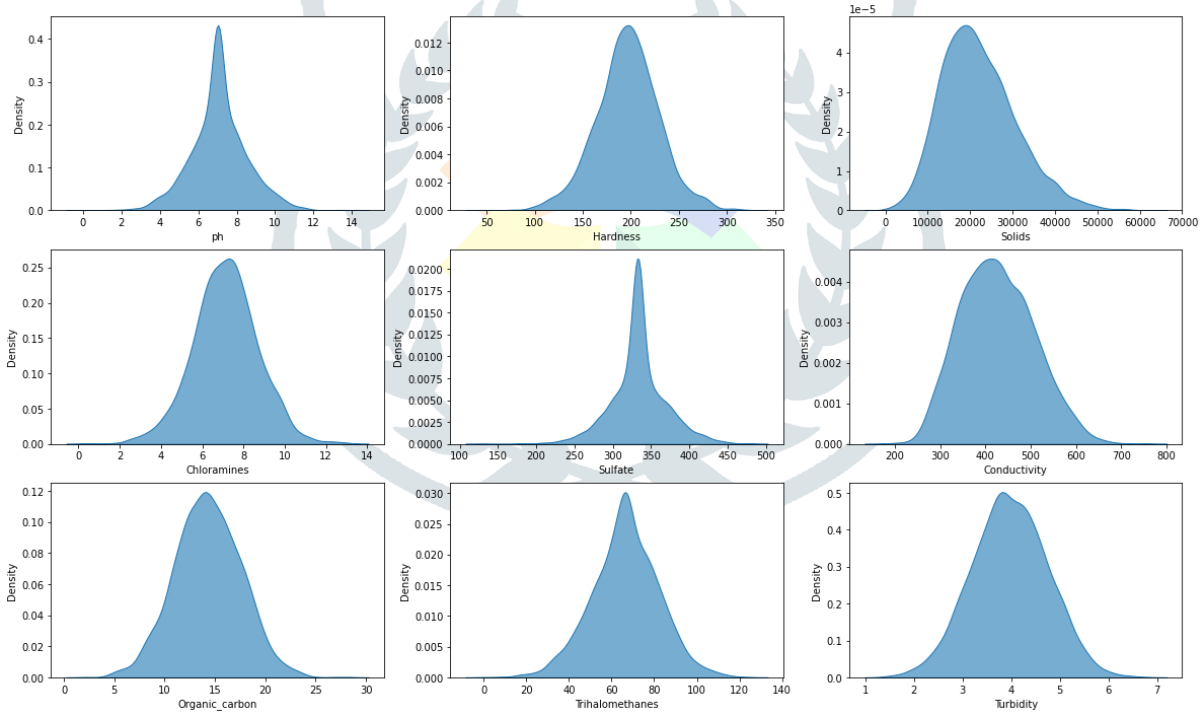

Figure 9. Distribution of all of the features of the water quality dataset

Figure 10 shows the correlation between the different features. It's noted that there's no strong correlation between any 2 specific features, suggesting that all of the features are independent of each other. The strongest positive correlation that exists within this dataset is Hardness and pH and the strongest negative correlation is of Sulfate and Solids.

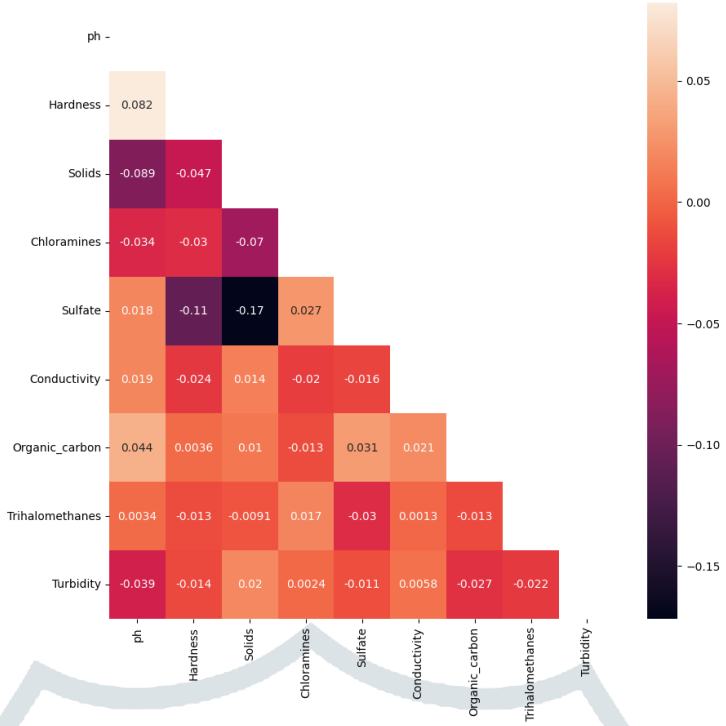

Figure 10. Heat map of the water quality dataset

Figure 11 shows a scatter plot of all the features against each other. It differentiates whether or not the water is safe or not (Blue - Not Safe, Orange - Safe) for consumption. It is shown that the features do not have significant correlation with each and there is no clear classification between the 2 classes. Hence, It would be difficult to apply distance-based models, such as KNN and SVM, to classify such kinds of data.

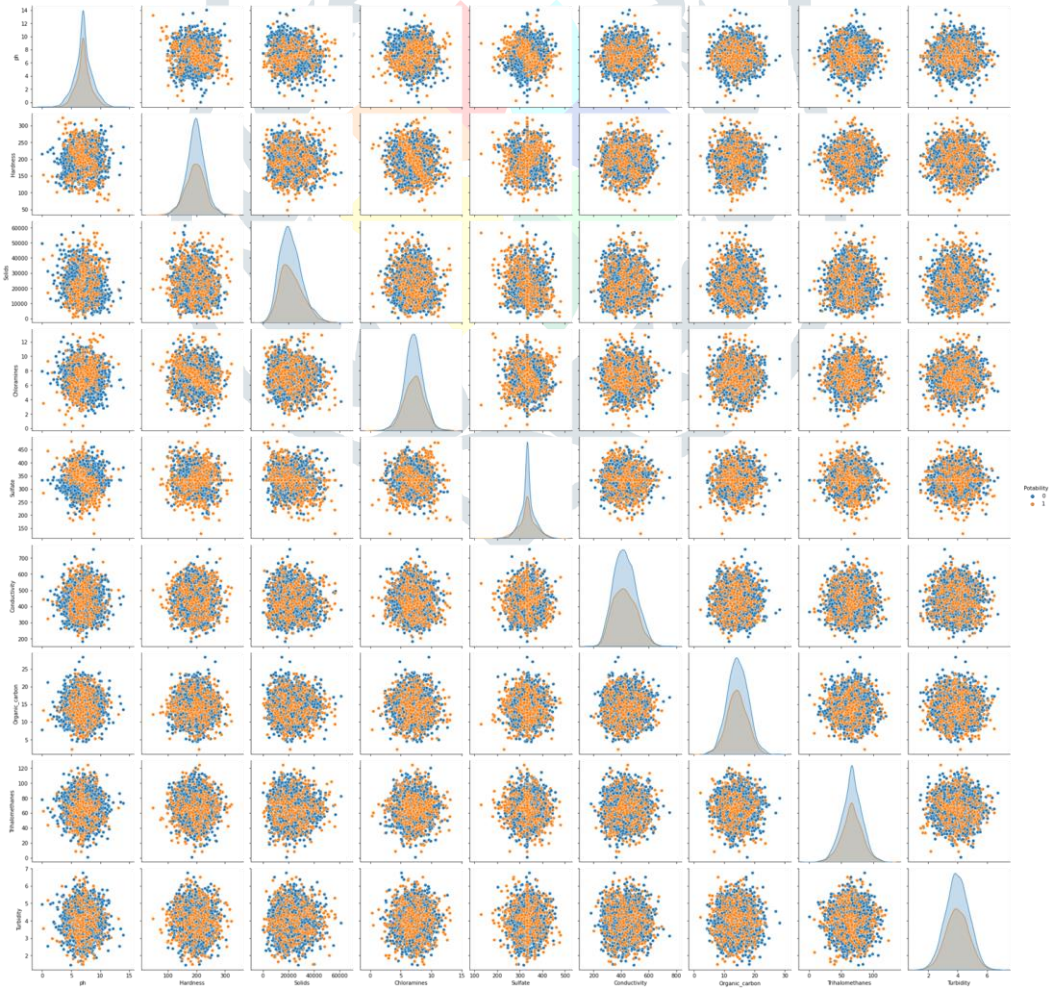

Figure 11. Correlation between all of the features of the water quality dataset with respect to the 2 classes

IV. EXPERIMENT

Data Splitting :

Before applying the Machine Learning algorithms on the dataset, Dataset is divided into two subsets, training and testing, where training set is used to train the model and test set is used to test the model on the unseen data to measure its prediction performance. The water quality dataset is split into 80% training and 20% testing sets.

Algorithms :

Decision Tree:

Decision Tree is a Supervised Machine Learning algorithm which can be used for regression and classification. Through training the model it makes decisions based on the given parameters. It uses Entropy and Information Gain to choose its root node and forms a top-to-down tree-like pattern of decision based on the input parameters [6].

Random Forest:

Random Forest is also a Supervised Machine Learning algorithm which can be used for regression and classification.It is a type of ensemble learning where it is a collection of decision tree classifiers of a given input and get the result by averaging out the best trees [12].

Multi-Layer Perceptron:

MLP or Multi-Layer Perceptron is one of the popularly used feed forward Artificial Neural Network (ANN) that generates output from the given input. It works with the help of different layers of neurons which takes data and does different computations [11].

Gradient Boosting Classifier:

Gradient Boosting Classifier is an ensemble learning which means it combines different weak learning models to create a better model. Gradient Boosting usually uses Decision Trees to train the model [10].

V. RESULTS

This paper uses Python version 3.10.2 along with a commonly used machine learning package sklearn version 1.0.2 to apply the machine learning algorithms on the dataset.

The results of the four models are listed in Table 2.

Different Machine Learning algorithms are trained using the water quality training set ignoring the outliers, replacing them by their median and removing them from the dataset shown in Table 2. The Multi-Layer Perceptron model has an accuracy of 62% when outliers are removed and 60% when the outliers are replaced by median. Compared to MLP model, Decision Tree model and Gradient Boosting model achieves high accuracy level.The Random Forest model has the highest accuracy level compared to all the 4 models with an accuracy of 79% when outliers are removed and 72% when the outliers are replaced by median. All four machine learning models exhibited an improved performance compared to Yang's model. This comparison is shown in Table 1. Here we see that Random Forest Algorithm gives 13% more accuracy compared to the best Artificial Neural Network (ANN) methods used, that is Deleting the missing values.

Table 2. Our models accuracy comparison

	Accuracy		
Algorithm	Outliers Ignored	Outliers Replaced by Median	Outliers Removed
Decision Tree	73%	70%	72%
Random Forest	**77%**	72%	**79%**
Multi-Layer Perceptron	65%	60%	62%
Gradient Boosting	76%	**73%**	71%

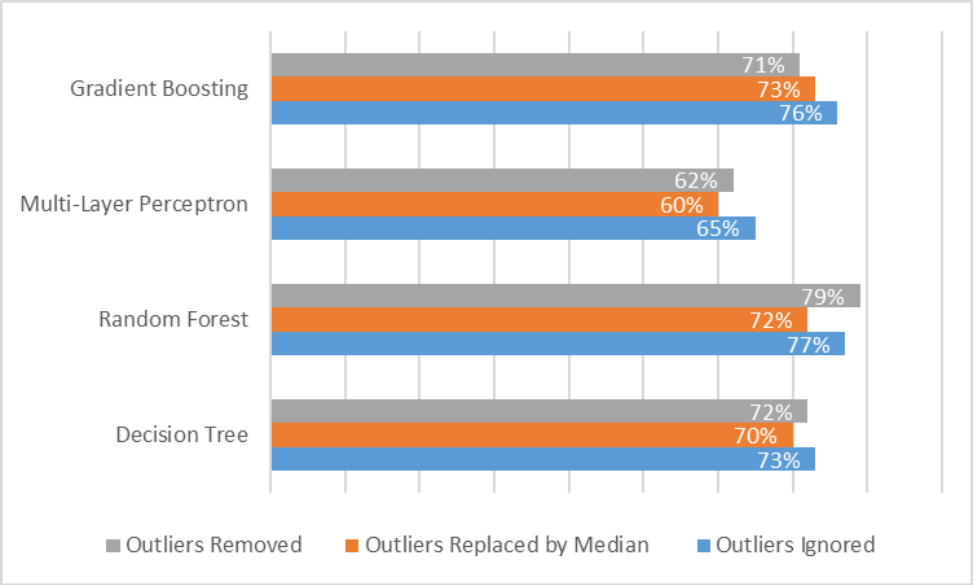

Figure 12. Accuracy comparison between our models

Table 3. Comparison of our model accuracy with other models

Model	Accuracy
Random Forest Algorithm Used in This Paper	**79%**
Yang's Model - Removing Missing Values	66%
Yang's Model - Median Imputation	64%
Yang's Model - Arbitrary Value Imputation	55%
Yang's Model - KNN Imputation	64%

VI. Conclusion

Modeling and forecasting water quality is very important for environmental protection. Model development using advanced artificial intelligence algorithms can be used to measure water quality in the future. Machine learning algorithms such as Decision Tree, Random Forest, MLP and Gradient Boosting were used to classify the freshwater quality dataset. The proposed models have been evaluated and tested using certain statistical parameters. The results have revealed that the Random forest algorithm has achieved the highest accuracy level compared to other machine learning moles like decision tree, MLP and gradient boosting. Also, this model performs well with respect to other Yang's models. In the future work, others tree-based models and ensemble models can be fine-tuned and applied on the same dataset.

References

[1] Uddin, M. D. G., Nash, S., Diganta, M. T., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for Predicting Coastal Water Quality index. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4135697

[2] Aldhyani, T. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics, 2020*, 1–12. https://doi.org/10.1155/2020/6659314

[3] Radhakrishnan, N., & Pillai, A. S. (2020). Comparison of water quality classification models using machine learning. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. https://doi.org/10.1109/icces48766.2020.9137903

[4] Kadiwal, A. (2021, April 25). *Water quality*. Kaggle. Retrieved January 16, 2023, from https://www.kaggle.com/datasets/adityakadiwal/water-potability

[5] Yang, R. (2022) "Analyses of approaches to deal with missing data in water quality data set," *Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)* [Preprint]. Available at: https://doi.org/10.2991/aebmr.k.220405.184.

[6] Quinlan, J.R. (1990) "Decision trees and decision-making," *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2), pp. 339–346. Available at: https://doi.org/10.1109/21.52545.

[7] Brownlee, J. (2021) *Smote for imbalanced classification with python*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (Accessed: January 23, 2023).

[8] Sawtell-Rickson, J. (2022) *When and why to standardize your data*, *Built In*. Available at: https://builtin.com/data-science/when-and-why-standardize-your-data (Accessed: January 20, 2023).

[9] Bonthu, H. (2022) *Detecting and treating outliers: How to handle outliers*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/ (Accessed: January 20, 2023).

[10] Nelson, D. (2023) *Gradient boosting classifiers in python with scikit-learn*, *Stack Abuse*. Stack Abuse. Available at: https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/ (Accessed: January 29, 2023).

[11] Bento, C. (2021) *Multilayer Perceptron explained with a real-life example and python code: Sentiment Analysis*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141 (Accessed: January 29, 2023).

[12] *Machine learning random forest algorithm - javatpoint* (no date) *www.javatpoint.com*. Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm (Accessed: January 29, 2023).