IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Water Quality Prediction and Analysis Using Machine Learning

Mrs. D. Leela Dharani[1], S. Jahnavi[2], Y. Yougender[3], M. Tanuja[4], M. Yaswanth[5]
Assistant Professor, Department of Information Technology[1]
B. Tech Students, Department of Information Technology[2,3,4,5]
Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: *Water pollution is a critical issue in India with a negative impact on water resources sustainability which can cause an inadequate water supply to all people even though a large number of water resources are available. Every year a large number of people are being affected with various diseases due to drinking polluted water. This project uses the python libraries to analyse and predict the water quality using Machine Learning based upon the dataset fed to the model. This project helps us to know whether the water is potable or not. The dataset consists of various parameters which impact the purity of the water like ph., conductivity, hardness, solids, turbidity, chloramines, sulphates, trihalomethanes, organic carbons.*

Keywords: Water Quality, Random Forest, Classification, Decision Tree, pandas, SK Learn, NumPy..

I. INTRODUCTION

Access to safe drinking water is essential to health, a basic human right and a component which effect the health. The data with affect the purity of the water is represented in csv file and it is represented in graphical form to analyse the attributes easily. The various libraries present in the python are used to analyse the data. Random Forest algorithm, which is a decision tree classification algorithm [3] is used to predict the output.
The objective of the project is to find whether the water is potable or not.[1][2][3]

II. PROPOSED SYSTEM

The idea is to find the potability of the water easily using Machine learning rather than IOT. The project uses the decision trees to predict the result. Random Forest Classifier [4] constructs the decision trees on the subsets of the dataset. The algorithm is a classification supervised learning algorithm. The dataset is first gathered and then the analysis of data is done. The barographs, histographs are used to analyse the dataset. This gives the information about the ranges on which the values of attributes lie pre-processed to fill the missing values in the data and the data is normalised using MinMaxScaler. Then the model is trained with the algorithm. The voting of the decision trees is considered as the final result of model.

III. TECHNOLOGIES USED

3.1 Anaconda and Python Notebook

Anaconda is an open-source distribution for python. It is used for data science, machine learning, deep learning, etc. More than 300 libraries are available for data science, it becomes fairly optimal for any programmer to work on anaconda for data science. Anaconda provides the Python notebook to perform the tasks. In python notebook the input is provided in the form of csv file.[7]

3.2 Python Libraries

The python notebook provides access to various python libraries. Pandas library provides the data analysis and it is used to get the input file from user. The NumPy library is used for operations on arrays and matrices. The Matplotlib and Seaborn are data visualization libraries which provides various graphs and plots to analyse the data.SKLearn library is used for data pre-processing and for splitting the data into training data and testing data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

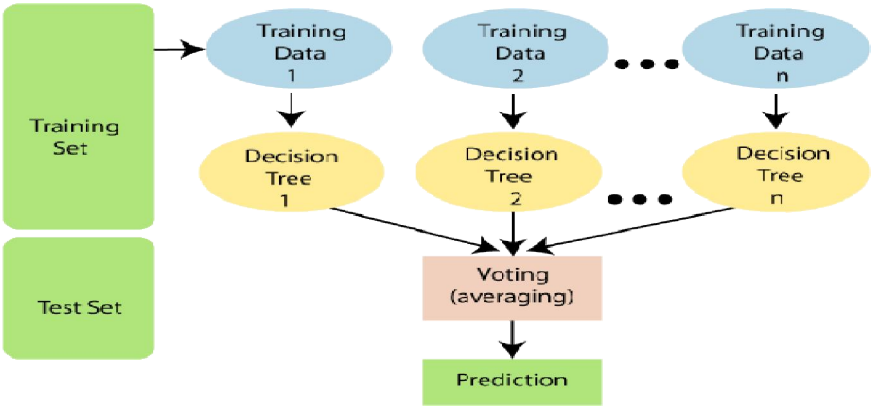
3.3 Data

The dataset is taken in the form of csv file which contains the data in the form of data frames. The dataset contains the attributes like ph., hardness, solids, chloramines, sulphates, conductivity, organic carbon, trihalomethanes, turbidity, potability. The potability is the result of the project.[5]

IV. IMPLEMENTATION

This project uses the Random Forest classification algorithm. The implementation of Random Forest algorithm is as follows.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.[4]



V. RESULTS

The below figure 1.2 shows the overview of the description of the attributes of the dataset.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690300	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833605	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762125	8.114887	359.950170	481.792305	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196010	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

Figure 1.2: Overview of Attributes

The below figure 1.3 shows the count of values of potability attribute, the dataset we considered as more not potable water data than the potable water data in the form of bar graph. 0 means the water is not drinkable and 1 means water can be drinkable.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

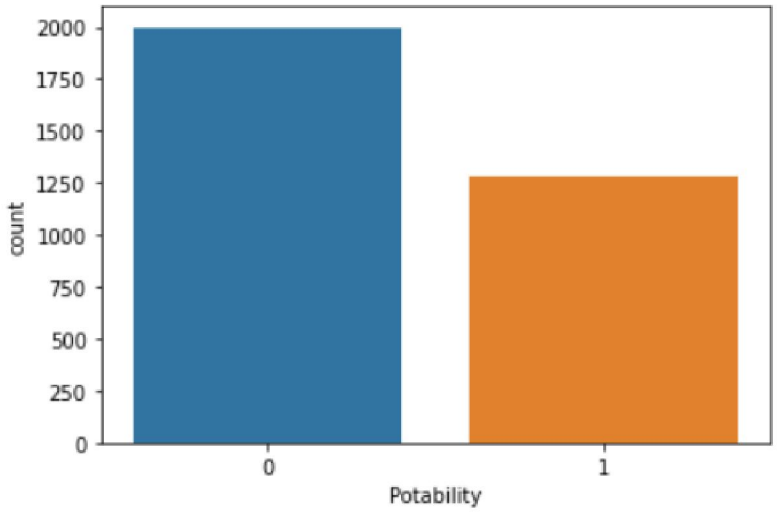


Figure 1.3: Potability graph

The below figure 1.4 is the histograms which shows the analysis of the attributes in graphical form.

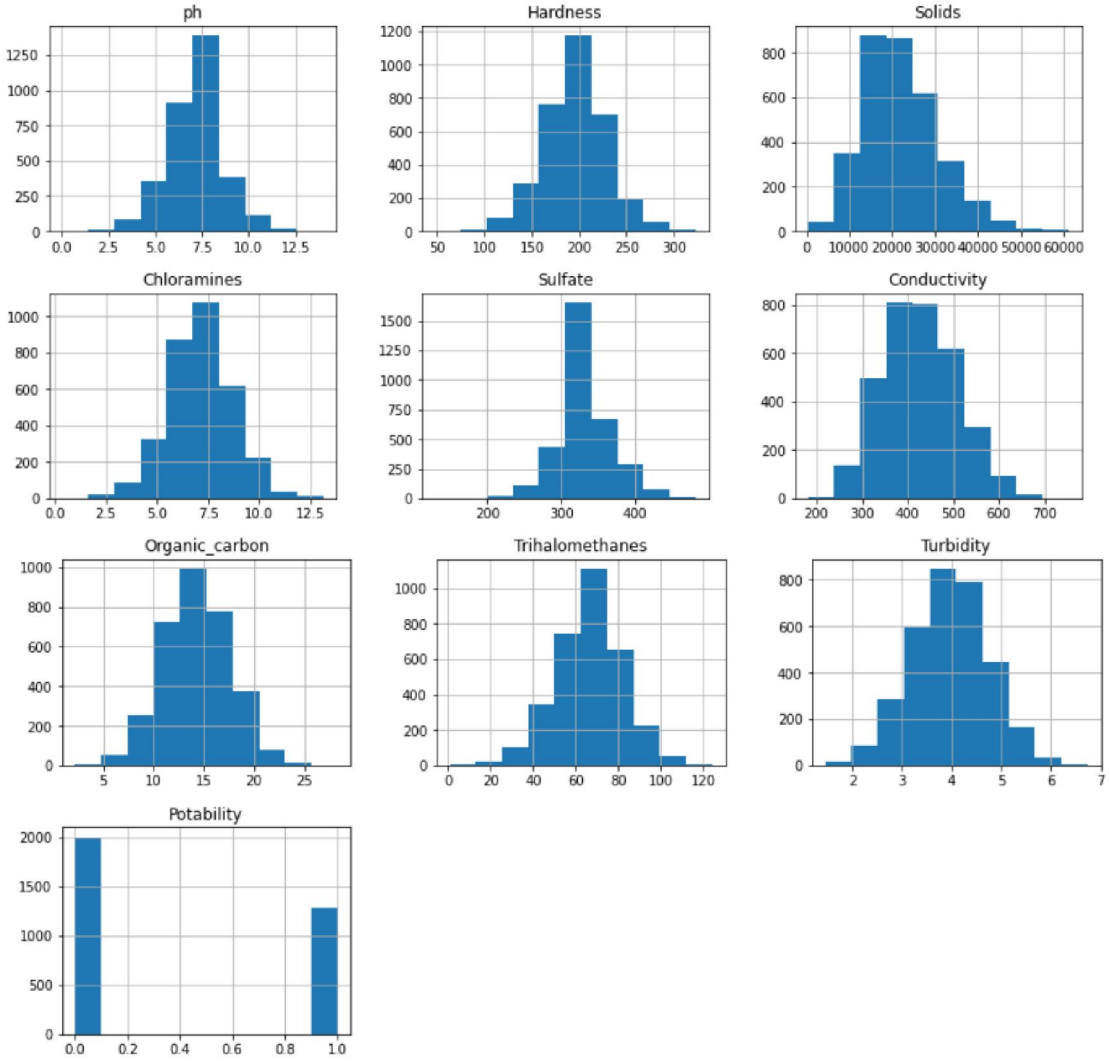


Figure 1.4: Analysis of Attributes

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

The below figure 1.5 is a heatmap graphical representation of the attributes with colour coded system.

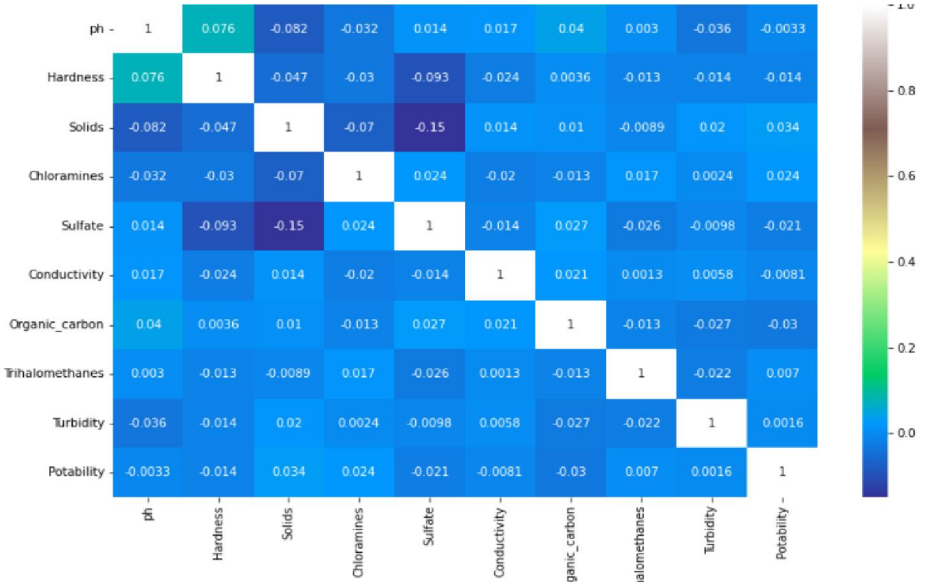


Figure 1.5: Heatmap representation

The accuracy of the model is increased by using Random Forest algorithm as compared to other classification algorithms.

	Model name	Accuracy
0	Random Forest Classifier	0.668852
1	Decision tree Classifier	0.603279

Figure 1.6: Accuracy of Model.

VI. SCOPE OF FUTURE USE

The project can be improved further by using better classification algorithm than Random Forest. By choosing more better algorithm the accuracy can also be increased. Water may contain many other impurities in the form of microorganisms like E.coli etc, which may impact the purity of water greatly. The microorganisms attributes can be added to the dataset to give better results.

VII. CONCLUSION

Thus, we can get to know how the attributes impact the purity of water. This can be used in any areas across the country. It can also be enhanced further. The model helps the people to know whether the water they are consuming is safe to drink are not.

REFERENCES

[1]. https://www.researchgate.net/publication/352907194_An_Introduction_to_Water_Quality_Analysis
[2]. https://environmentalsystemsresearch.springeropen.com/articles/10.1186/s40068-016-0053-6
[3]. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
[4]. https://www.javatpoint.com/machine-learning-random-forest-algorithm
[5]. Dataset Kaggle kernels output imakash3011/water-quality-prediction-7-model -p /path/to/dest
[6]. https://journals.sagepub.com/doi/full/10.1177/11786221221075005
[7]. https://www.anaconda.com/products/distribution