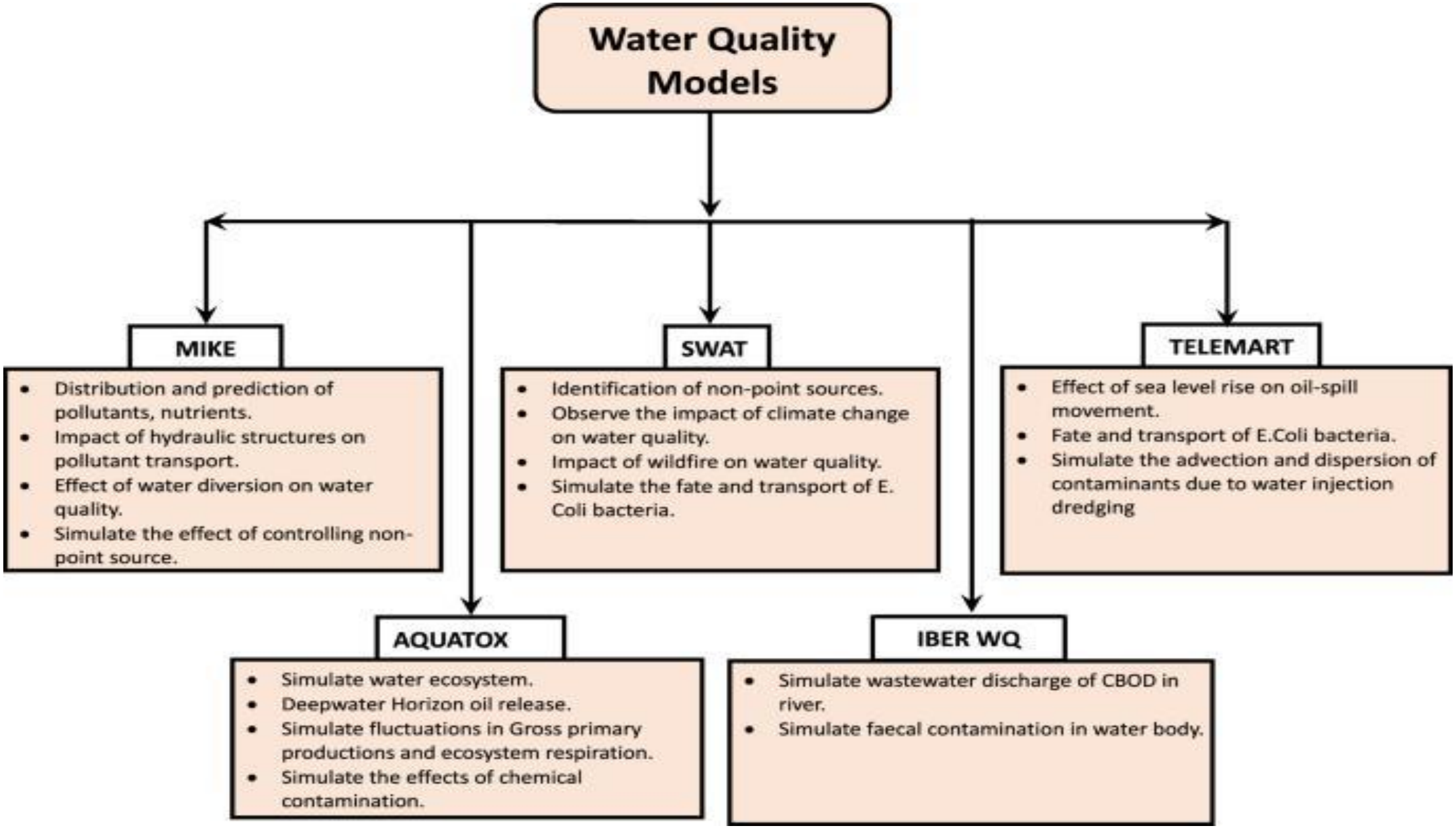
Performance and final submission phase

Model performance metrics:

- The management policy of water resource is a critical and systematic process that is associated with diverse components like as policy, law and regulations, institutional framework, advanced analytical facilities, skilled labour, well organization infrastructures, financial freedom etc.
- A number of framework used to implement the management policy for restoring good water quality status.
- The Water Framework Directive (WFD) is a useful tool that provides detailed guidelines for maintaining good water quality and a healthy aquatic ecosystem .

- However, it has suggested a set of water quality measures for evaluating rivers and streams. In that case, it is frequently impractical and exceedingly costly for all parties involved, particularly those with minimal resources.
- Water quality index model is one them, its allows converting a vast water quality information into single numerical values more simplified way than traditional approaches .
- A number of WQIs have established by various countries/organizations in order to specific goals such as ground water quality index, surface quality water index etc.
- This technique has been criticized for a number of reasons, including (i) uncertainty issues, (ii) model reliability, (iii) transparency, and (iv) model sensitivity.


Water Quality Models
MIKE
Distribution and prediction of pollutants, nutrients.
Impact of hydraulic structures on pollutant transport.
Effect of water diversion on water quality.
Simulate the effect of controlling non-point source.
SWAT
Identification of non-point sources.
Observe the impact of climate change on water quality.
Impact of wildfire on water quality.
Simulate the fate and transport of E. Coli bacteria.
TELEMART
Effect of sea level rise on oil-spill movement.
Fate and transport of E.Coli bacteria.
Simulate the advection and dispersion of contaminants due to water injection dredging
AQUATOX
Simulate water ecosystem.
Deepwater Horizon oil release.
Simulate fluctuations in Gross primary productions and ecosystem respiration.
Simulate the effects of chemical contamination.
IBER WQ
Simulate wastewater discharge of CBOD in river.
Simulate faecal contamination in water body.

Performance metrics include:

- Dissolved oxygen
- PH
- Temperature
- Salinity
- Nutrients (nitrogen and phosphorus)

The above defined are some of the performance metrics of water quality management.

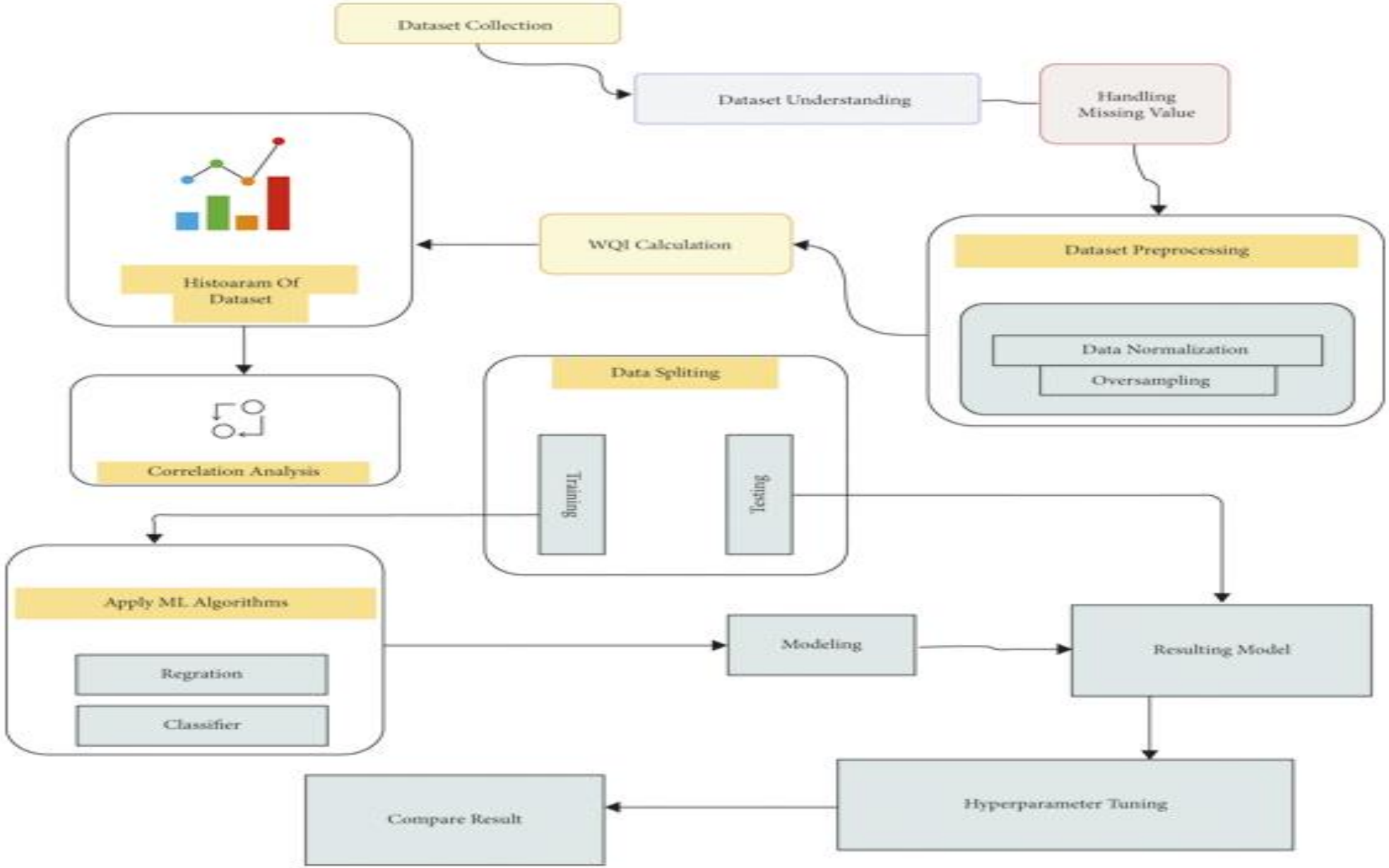
Project documentation

Abstract:

During the last few decades, the quality of water has deteriorated significantly due to pollution and many other issues. As a consequence of this, there is a need for a model that can make accurate projections about water quality. This work shows the comparative analysis of different machine learning approaches like Support Vector Machine (SVM), Decision Tree (DT), Random Forest, Gradient Boost, and Ada Boost, used for the water quality classification.Experiments results depict that Random Forest and Gradient Boost give the highest accuracy of 81%. One of the major issues with the machine learning model is lack of transparency which makes it impossible to evaluate the results of the model.Within the context of this investigation, Local Interpretable Model-agnostic Explanations (LIME) is utilized to ascertain the significance of the features.

Introduction:

- Whether it is utilized for drinking, household usage, food production, or leisure, safe and readily available water is critical for public health. Improving supplies of water, and also improved management of water resources, might help countries thrive and reduce poverty. There are many reasons why water is deteriorating because in our India there are many industrial areas so the release of pollutants in rivers is the main reason for water deteriorating. There are many other reasons for water deteriorating like people's garbage (plastics), the unwanted things in rivers, their nearest ponds, lakes, and also in sea, and due to plastics and unwanted garbage, there are some toxic occurrences. So, for all these reasons, water is deteriorating nowadays.Approximately 80% of the local population and 20% of the urban population do not have access to clean drinking water. Three-quarters of the nation's children's health issues are infectious diseases and environmental factors, mainly water supply and sanitation.


Dataset Collection
Dataset Understanding
Handling Missing Value
Dataset Preprocessing
Data Normalization
Oversampling
WQI Calculation
Histoaram Of Dataset
Correlation Analysis
Data Spliting
Training
Testing
Apply ML Algorithms
Regration
Classifier
Modeling
Resulting Model
Hyperparameter Tuning
Compare Result

Proposed system:

- The standards used to assess the sustainability of water resources are constantly evaluated as new factors are found. Standards and guidelines for contamination levels in drinking water are being developed by regulatory agencies. In response to the changing criteria, the water supply sector is creating new and improved operating and treatment procedures. All elements that affect water quality, as well as the public health relevance of components and available treatment technology, must be considered when developing drinking water quality guidelines.

- The initial task was to find out which factor would give a good indication of the quality of the water. Hardness, sulfate, solid, trihalomethanes, pH, turbidity, solids, organic carbon, and conductivity were chosen as parameters after extensive investigation. Water parameters delve into the logic behind these choices.As a resultwhich will determine if the water is potable or not.
- The second task was to deal with the dataset's missing values. The value of some factors may not be specified while defining the models, and the output may differ as a result. To solve this problem, we have included the mean value of the factor for which data is absentTo achieve our goal, we appropriately calculate the Water Quality Index (WQI) to analyze water quality.

Data Collection:

- The dataset used in this approach came from Kaggle's Water Quality Dataset. Some of the metrics employed in this investigation were hardness, sulfate, solid, trihalomethanes, pH, turbidity, solids, organic carbon, and conductivity. The description of all features is given in Table 1.

Parameters	WHO limits
Ph	6.5–8.5
Hardness	200 mg/L
Solids	1000 ppm
Chloramines	4 ppm
Sulfate	1000 mg/L
Conductivity	400 μS/cm
Organic carbon	10 ppm
Trihalomethanes	80 ppm
Turbidity	5 NTU

Algorithm for proposed model:

Input Data Water Portability Dataset from Kaggle

OutputYes (If water is portable), No

Data preprocessing

Normalization using Z-score

Oversampling using SMOTE

Calculate the WQI using equation (4).

Visualize and analyze the data

Correlation analysis

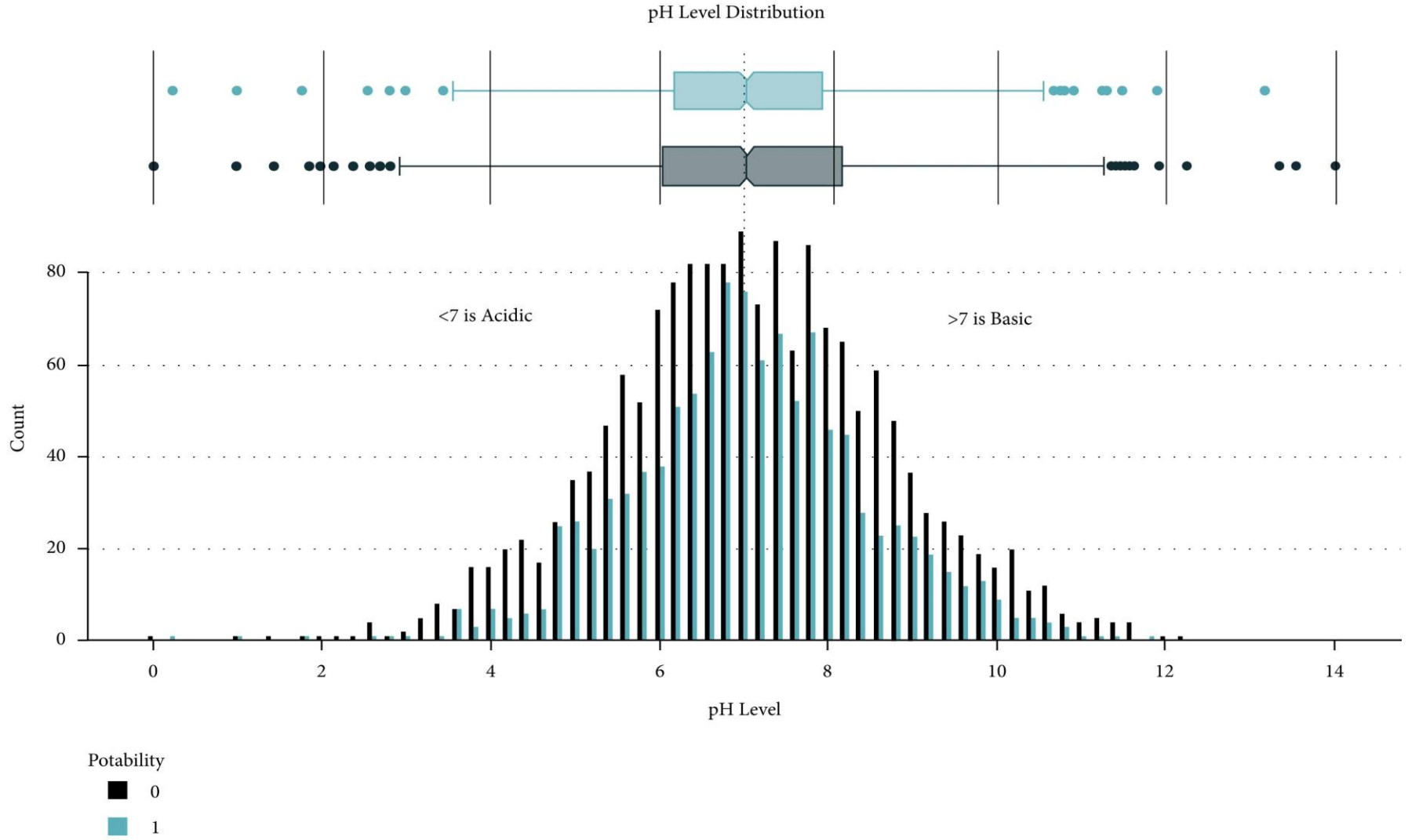
Data splitting

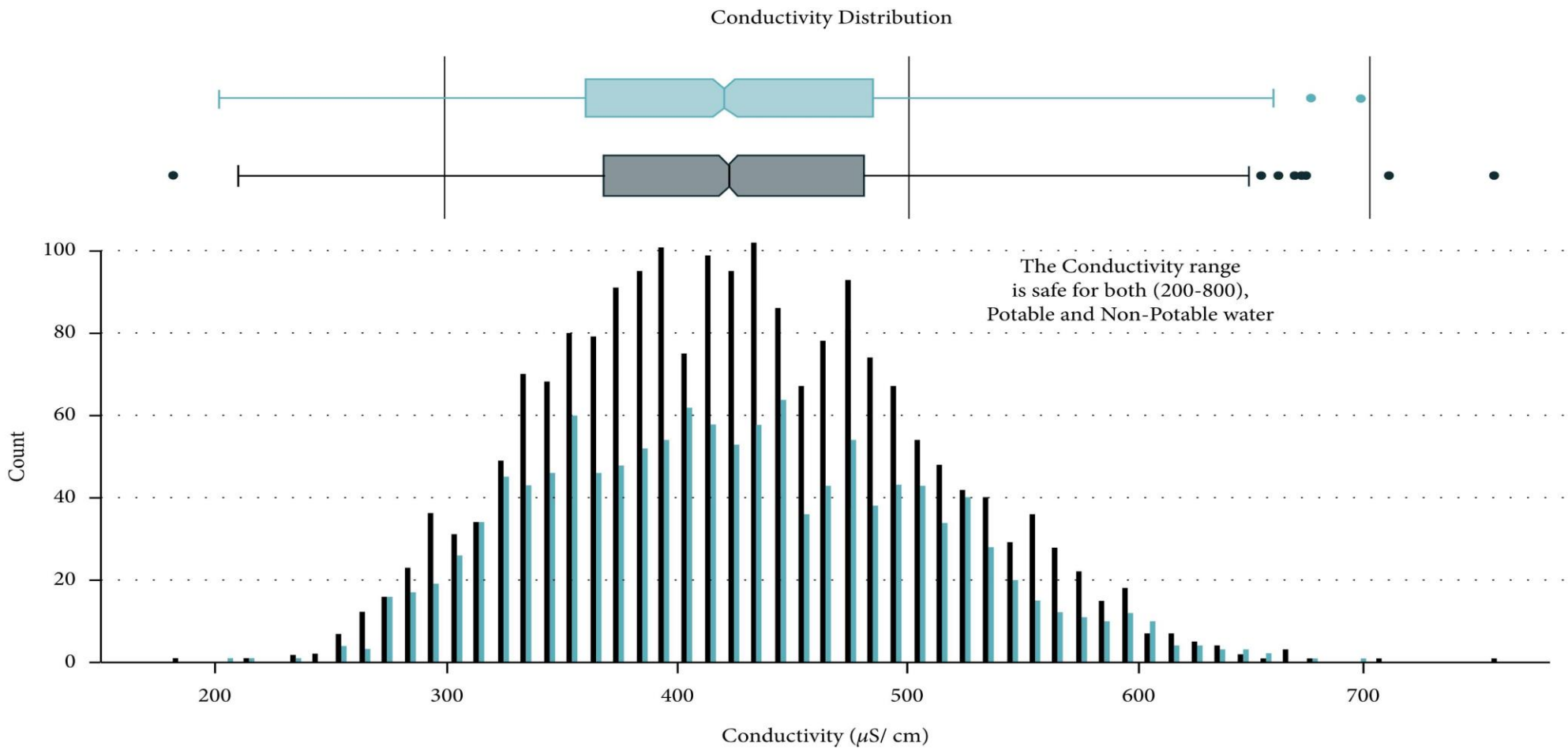
Apply different Machine Learning Model for the water quality prediction

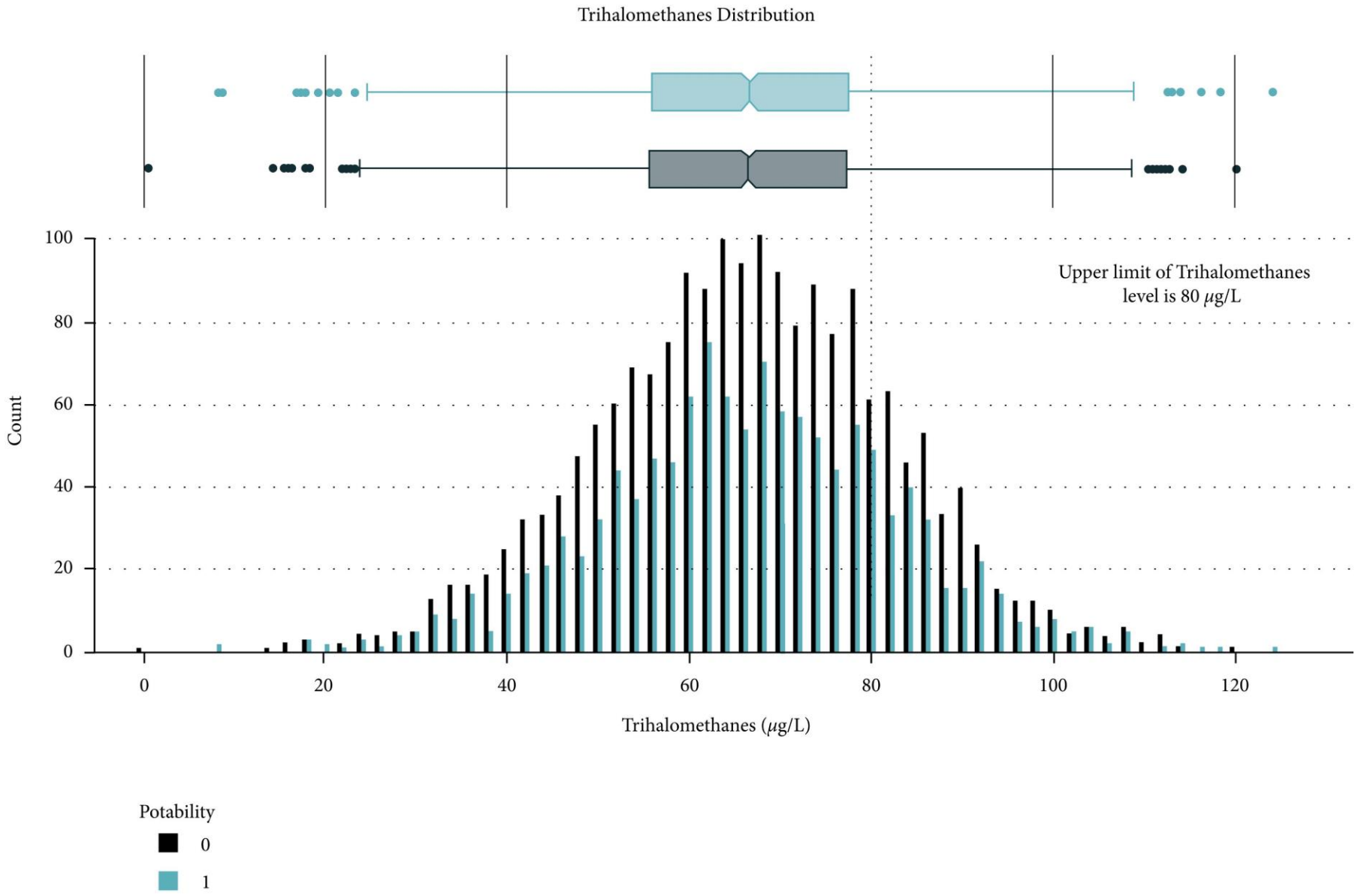
Evaluate the performance of the different model

Apply hyper parameter tuning to improve the performance of the model

Data Visualization:




Conductivity Distribution
The Conductivity range
is safe for both (200-800),
Potable and Non-Potable water
Count
100
80
60
40
20
0
200
300
400
500
600
700
Conductivity (μS/ cm)
Potability
0
1


Trihalomethanes Distribution
Upper limit of Trihalomethanes level is 80 μg/L
Count
0
20
40
60
80
100
0
20
40
60
80
100
120
Trihalomethanes (μg/L)
Potability
0
1

Results for algorithms:

Classifiers	Accuracy
Random Forest	0.80
Gradient Boost	0.76
Decision Tree	0.73
Support Vector	0.69
AdaBoost	0.68
Support Vector	0.67
KNeighbors	0.65
BernouliNB	0.61
GaussianNB	0.57
Passive aggressive	0.54
Nearest centroid	0.52
Logistic regression	0.52
Ridge	0.52
Stochastic gradient descent	0.51
Perceptron	0.51

Final Model and Results:

- Water quality is traditionally measured using water quality criteria obtained through time-consuming laboratory examination. We looked at different machine learning approaches for estimating it and discovered various research that used them. The model was evaluated using ten water quality parameters in these experiments.
- We were using cross validation to evaluate final model. Cross validation divides the splits of population among k segments and propagates over each one, with k-1 fragments serving as well as one substantial number of the training datasets serving as proving ground. A conventional approach to assessing the performance of automation is the k-fold cross-validation procedure. We were using repeated stratified K-fold in which it gives a method for improving a machine learning model's projected performance. Simply repeat the provided mean result throughout all layers by all iterations using the cross-validation routine for several rounds.

Cross validation:

Model	Parameters	Accuracy
Gradient Boosting	n_estimators = 500, max_features = log2	0.79
Random Forest	n_estimators = 100, max_features = auto	0.81

Conclusion:

- Water samples with acidic and basic pH levels are approximately evenly distributed in the data:
- (1) In the data that was considered hard of 91.73%
- (2) The water samples safe for chloramines are only 2.72%
- (3) The water samples safe for sulfate are only 1.77%
- (4) The water samples safe for carbon (in 10 ppm) are 90.57%
- (5) The water samples safe for trihalomethane are 81.62%
- (6) The water samples safe for turbidity are 90.42%
- (7) The correlation coefficients between the features were very low

Project demonstration

Introduction:

- Water is one of the largest resources on earth. People need water to sustain life, including drinking water. The core of the water quality online monitoring system is the online monitoring equipment. The system uses many technologies, including sensors, automatic measurement and control, computer applications, et al. By using these technologies, the system can sort out and analyze the detection data and output the maximum and minimum monitoring data in any time period. The average value of monitoring data in any time period can also be output by calculation [3, 4]. By analyzing the data, the system can automatically synthesize the data map reflecting the water quality, and finally store the information to the system data control center for research. The system has many functions.

2. Methodology

2.1. Machine Learning Models

In statistics, the binary logistic model is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). In this paper, we perform the binary logistic regression to each single variable. Specifically, we compute Hosmer-Lemeshow Goodness-of-Fit (HLGOF), and the significance level is set to be equal to or less than 0.05. For those features which do not pass the HL test, this paper considers it having no predictive power in the logistic model. Then the binary logistic regression is performed to those features which pass the HL test, and the HLGOF, the Exp(B), and the regression coefficient is computed [9].

In statistics, the k-nearest neighbor algorithm (KNN) is a non-parametric supervised learning method. The input consists of the k closest training examples in a data set, and in K-NN classification, the output is a class membership [10]. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In this paper, as for those parameters having no predictive power in the logistic model, we perform the K-Nearest Neighbor model and use K-fold Cross Validation to find the best K.

2.2. Data

The dataset used in this study was downloaded from Kaggle. A total of 3276 samples were collected and analyzed for 9 important hydro-chemical parameters, which are pH value, hardness, total dissolved solids (TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes and turbidity. The potability of each sample, which indicates if water is safe for human consumption, is given, where 1 means Potable and 0 means Not potable. This dataset is randomly divided in such a way that 2457 (75%) samples are used for the training, whereas the remaining 819 (25%) were used for testing the models. This paper uses single imputation to deal with the missing value, i.e., all missing values that are labeled "potable" will be imputed using the mean of all non-missing "potable" samples, and the same action will be applied to "non-potable" samples with missing values.

2.3. Training process

This paper computes Pearson Correlation Coefficient between any two features to find out the degree to which these variables are correlated. Images of the distribution of all features divided by target label are shown in figure. Judging from whether the sample is potable, this paper divides all the samples into two populations. We perform a two-tailed t-test to check if there is any significant difference between the two samples' means, considering the sample size differences and unequal variance. The significance level alpha is set to be equal to or less than 0.1. This paper performs the binary logistic model and K-NN to all the features.

This paper uses the trained model above to treat the dataset used for prediction and find the false positive rate. In this process, if all the features show that the sample is potable, then output 1. Otherwise output 0.

3. Results and Discussion

3.1. The establishment of simulation model

From figure 1, the overall data clearly shows that water resources are 61% non-potable, while only 39% are potable. This contrast is a strong indication that the amount of water currently potable is under serious threat. More and more pollution which breeds bacteria occupies large space. Such bacteria pollute the groundwater and air, and then threaten safety of drinking water. At the same time, because of over-exploitation of groundwater, some areas have produced phenomena like shrinkage of lakes and disappearance of the mud flats. These all would cause the reduction of water storage capacity and water self-purification. This will deteriorate the status of water potability.

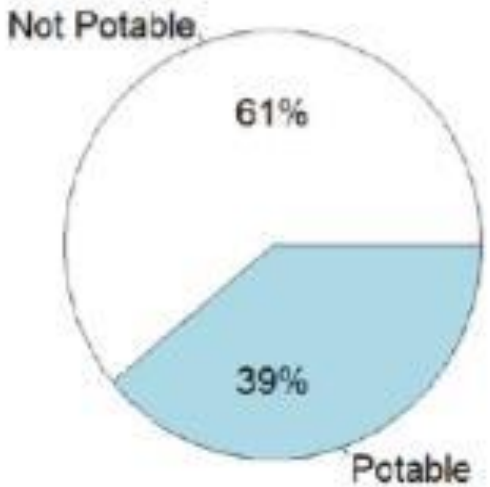


Figure 1. Pie chart of the water potability

3.2. The density distribution of variables

In this data, the nine variables which will affect water resources are considered. The data is divided into two groups by potability and their corresponding density distribution curves will be produced by analyzing nine variables. In this part, nine variables which affect water potability are split into three parts to describe. All density distributions are approximately normal distributions because their density curves are similar to normal distribution curves. The differentiation conditions are: (1) means are same but two curves are not; (2) the mean values are different; (3) two curves are almost the same.

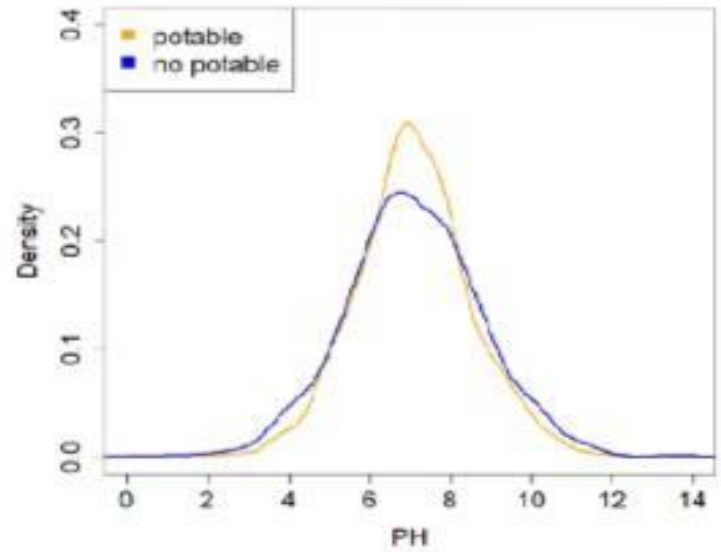


Figure 2. The density distribution curves of PH

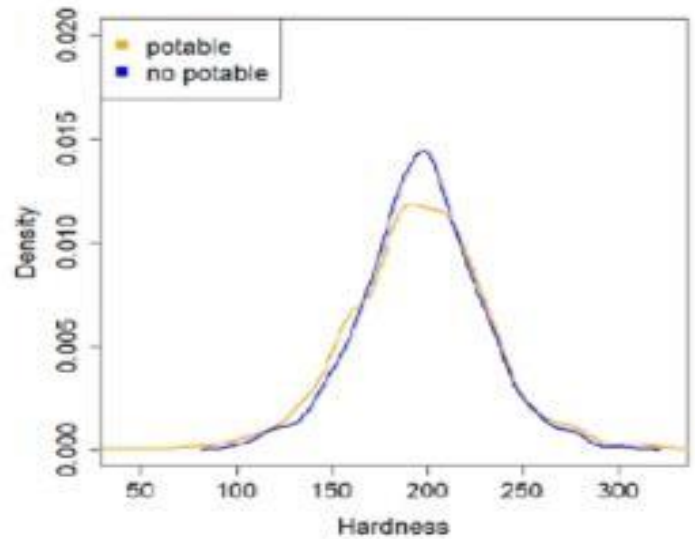


Figure 3. The density distribution curves of Hardness

Correlation of variables:

Correlation	pH	Hard	Solids	Chlor	Sulfate	Conduct	Organic	Trihal	Turbi
pH	1	0.109	-0.088	-0.025	0.011	0.014	0.028	0.018	-0.036
Hard	0.109	1	-0.053	-0.023	-0.109	0.012	0.013	-0.015	-0.035
Solids	-0.088	-0.053	1	-0.052	-0.163	-0.005	-0.005	-0.016	0.019
Chlor	-0.025	-0.023	-0.052	1	0.006	-0.028	-0.024	0.015	0.013
Sulfate	0.011	-0.109	-0.163	0.006	1	-0.016	0.027	-0.023	-0.01
Conduct	0.014	0.012	-0.005	-0.028	-0.016	1	0.016	0.005	0.012
Organic	0.028	0.013	-0.005	-0.024	0.027	0.016	1	-0.006	-0.015
Trihal	0.018	-0.015	-0.016	0.015	-0.023	0.005	-0.006	1	-0.02
Turbi	-0.036	-0.035	0.019	0.013	-0.01	0.012	-0.015	-0.02	1

Table 2. Results of relevant data fitted by KNN

Water prediction	Test label	Test label	Row Total
Train label	0	1	
0	391	103	494
1	209	108	317
Column Total	600	211	811

4. Conclusion

Water quality relates to everyone's life. The adequacy of water resources not only affects people's life safety, but also deeply affects the development and stability of society. Also, the government of each region should predict the sustainability of their water resources. Water used by humans should be guaranteed to be enough and safe. Different standards apply to different uses of water. When it comes to drinking water, the standards need to be particularly strict. Not every water resource meets these standards. In addition to the special substances contained in some specific type of water, there are nine main factors that affect the potability of water, including, PH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes and Turbidity.

According to the result of the research, the features of water are hardly related to each other. There are no general rules between each distribution of the factors. For example, the range of the non-potable distributions of PH value and organic are larger than the range of the potable distributions, while for the other factors except Trihalomethanes (the distribution of potable and non-potable is almost the same), the range of potable distribution is larger than the non-potable distributions. Qualification of one feature does not increase the possibility of the qualification of another feature. Only if every factor meets the potable standard, the water can be drunk.