An approach for the prediction of water quality using Random Decision Forest model

Nithin Sai Jalukuru
School of Engineering and Applied Sciences
University at Buffalo, Buffalo, NY, USA
njalukur@buffalo.edu

Abstract：In this present work, Classification of water quality is done using Random decision forest classifier with fine-tuned hyper-parameters based on the various attributes of factors involved in water quality.

Keywords：Ensemble classification, Supervised machine learning, Random Forest classifiers

1.INTRODUCTION

Classification is one of the major machine learning methods. Classification can be regarded as the assignment of a label to an observation based on attributes.

A classifier that makes a decision based on inspecting one attribute at each internal node is known as a decision or classification tree. At each stage of construction, the classifier is built by anticipating all attributes and selecting one. This choice is made in accordance with the impurity standard.

Random Forest is an *Ensemble Supervised Machine Learning* technique that has emerged recently. Random Forest uses *decision tree* as base classifier. Random Forest generates multiple decision trees; the randomization is present in two ways: (1) random sampling of data for bootstrap samples as it is done in bagging and (2) random selection of input features for generating individual base decision trees. *Strength* of individual decision tree classifier and *correlation* among base trees are key issues which decide *generalization error* of a Random Forest classifier.

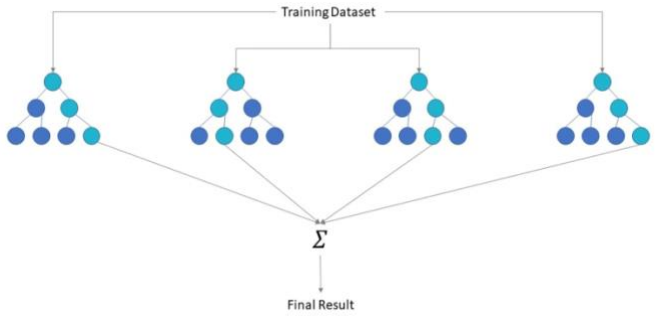

Fig1: Diagram of Random Forest Classifier (Source IBM)

Random Forest, according to Breiman [3], performs well on large databases, can handle thousands of input variables without variable deletion, provides estimates of significant variables, generates an internal, unbiased estimate of generalization error as the forest grows, has a useful method for estimating missing data and maintains accuracy when a significant portion of data are missing, and has methods for balancing class error in class population unbalanced data sets.

1.1 Random Forest

Decision forest is referred as Random Forest since a random vector is chosen for every tree classifier. The random vector determines a subset of the initial training set for each tree. This method helps to avoid overfitting and decrease the generalization error

Definition : *A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(x, \Theta_k), k = 1,2 \dots \dots\}$ *Where* Θ_k *are independent, identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.*

The key element is the independence of each random vector $\{\Theta_k\}$ from$\Theta_1 \dots \Theta_{k-1}$, but all vectors must share the same distribution.

The corresponding out-of-bag data are used to estimate the out-of-bag error. The distribution of data for a Random Forest is shown in Figure 2. First, training and test sets of the input data are created. The bootstrapping method is then used by each tree to choose a subset from the training set. The chosen data is used to train the tree, creating "*in-bag*" data. The remaining portion, referred to *as out-of-bag* data, differs for each tree.

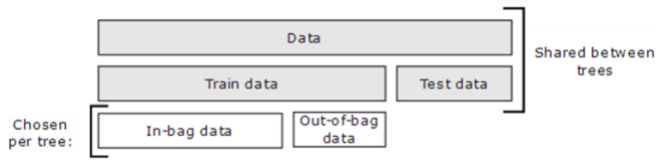

Fig2: The data partition for a Random Forest classifier.

The standard decision tree construction is modified by the Random Tree algorithm with randomization. From an initial set of features M, a subset of features *m* are chosen at each node. Then, using an impurity measure, the best attribute among m is chosen as a split point. In his initial proposal, Breiman used the Gini Index to select the ideal variable out of a group of variables that were chosen at random. The trees are not pruned and are built to their maximum depth. A new instance is processed through each

tree in the ensemble in order to be classified. The majority of "votes" determines the final class.

1.2 Strength and Correlation

Breiman demonstrates that the generalization error converges as the number of forest trees increases. The upper bound for the generalization error PE for the Random Forest is defined as

$$PE \leq \frac{\rho(1-s^2)}{s^2}$$

where ρ is the correlation between the classifiers and s is strength of the classifiers. Strength is a fitness indicator that reveals how accurate each tree is. The *generalization error* will be the lowest if the ratio $\frac{\rho}{s^2}$ is the lowest. The objective is to maintain as high a level of strength for each classifier while limiting the degree of correlation between the classifiers.

1.3 Advantages of Random Forest Classifier

I. Higher accuracy compared to other classification models with *lower computational cost*
II. Only the initial subset of attributes are chosen for each tree's node as a potential split point. As a result, not all features are examined at each stage of the tree-building process. If a data set has a lot of attributes, this may be advantageous.
III. Every tree in Random Forest is constructed independently. Thus, the construction of a Random Forest can be done in *parallel*
IV. The classifier can assist in locating pathological data, such as instances that were incorrectly classified and outliers. With the help of the proximity

1.4 Limitations

I. Time consuming process
II. Requires more resources
III. More complex

2. EXPERIMENTATION

2.1 Methodology

For the Random Forest classifier model implementation, a data pipeline structure is constructed as illustrated in the accompanying figure 3.

Data collection and loading into the environment are both parts of data extraction After loading, the data has been labeled and cleaned so that the ML model can analyze it effectively. . In addition, feature extraction a few features/attributes have been extracted, and the most pertinent ones have been chosen to train the model.

The correct ML model has been selected, trained, and verified during the model selection, training, and validation processes by maximizing the more suitable assessment criteria

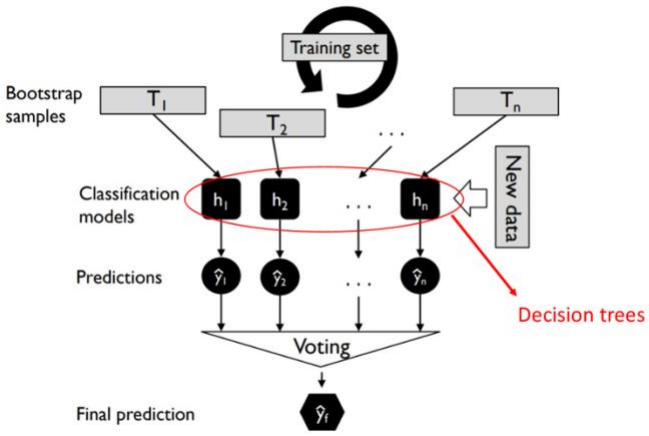

Fig: 3 random forests = Bagging classifier + Random feature subsets

2.2 Data Extraction

Access to clean drinking water is crucial for health, a fundamental human right, and a component of successful health protection policies. At the national, regional, and local levels . The Water quality data has been collected from Kaggle. This dataset has 3276 observations and 10 features. These features are

I. **PH** : In terms of water quality, this value is crucial. It turns acidic if the pH value is high and alkaline if the pH value is low (<7), *continuous attribute*
II. **Hardness**: Water hardness is measured in parts-per-million and is based on the milligrams of calcium carbonate per litre (ppm). According to studies, drinking hard water, *continuous attribute*
III. **Solids**: Numerous inorganic and some organic minerals or salts, including potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates, and others, can be dissolved by water. These minerals gave the water an undesirable flavor and diluted color. This is a crucial consideration when using water, *continuous attribute*.
IV. **Chloramines**: The primary disinfectants used in public water systems are chloramine. When ammonia is added to chlorine to treat drinking water, chloramines are most likely to form. Water that contains a chlorine concentration of up to 4 mg/L (4 ppm) is deemed safe for human consumption, *continuous attribute*
V. **Sulphates:** Sulfates are organic compounds that are naturally present in rocks, soil, and minerals. They can be found in the surrounding air, groundwater, vegetation, and food. Sulfate is primarily used in the chemical industry for commercial purposes, *continuous attribute*
VI. **Conductivity :** Pure water is a good insulator rather than a good conductor of electrical current. The electrical conductivity of water is increased as the concentration of ions rises., *continuous attribute*
VII. **Organic_Carbon :** 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, *continuous attribute*

VIII. **Trihalomethanes**: These are substances that may be present in chlorine-treated water. The amount of organic matter in the water, the quantity of chlorine needed to treat the water, and the temperature of the treated water all affect the concentration of THMs in drinking water. THM concentrations up to 80 ppm are regarded as safe for drinking water, *continuous attribute*

IX. **Turbidity:** It measures the water's ability to emit light, and the test results are used to determine how well waste is discharged in terms of colloidal matter, *continuous attribute*

X. **Potability** : It tells whether the water is safe for drinking or not. 1 means potable and 0 is not potable.

2.2 Data cleaning and feature Engineering

I. Data contained missing values from ph , Sulfate, trihalomethanes.

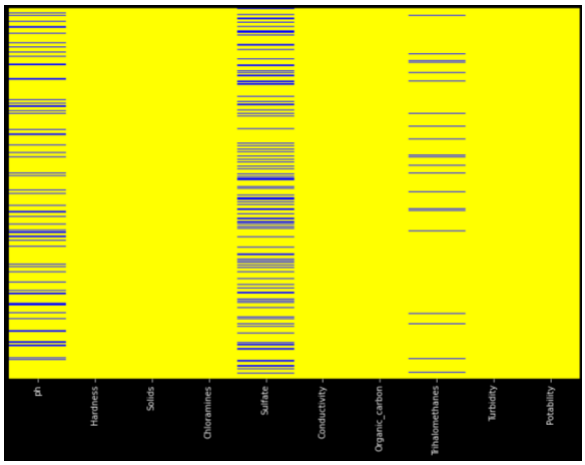

Fig4: Representation of Missing values in data

In Fig4 , horizontal blue lines indicate the missing/NAN in the features. So, without removing those observations. NaN's are treated by *KNN imputer*

II. Treating Outliers : It can be seen from the fig5, there are few outliers present in each attribute. But, important thing is **Random forest can handle outliers automatically (That's the beauty of random forest)**. The algorithm is very stable. If a new data point is introduced in the dataset, the overall algorithm is not impacted much .

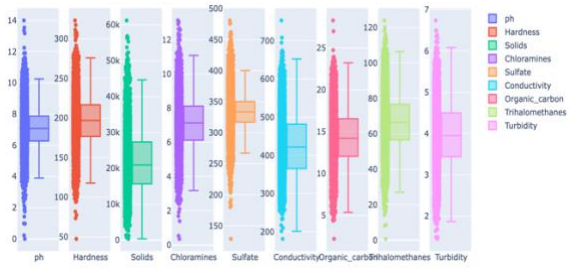

Fig5: Representation of Outliers (Plot in Plotly Express)

III. Performed feature standardization on the numerical data attributes where the feature values are re-scaled to have a mean of $\mu = 0$ and standard deviation $\sigma = 1$.

2.3 Exploratory Data Analysis

I. Distribution of potability of water : From Fig6, it is evident that in the present dataset 61% of the sample are non-potable (not safe for drinking) and 39% samples are safe for drinking (potable)

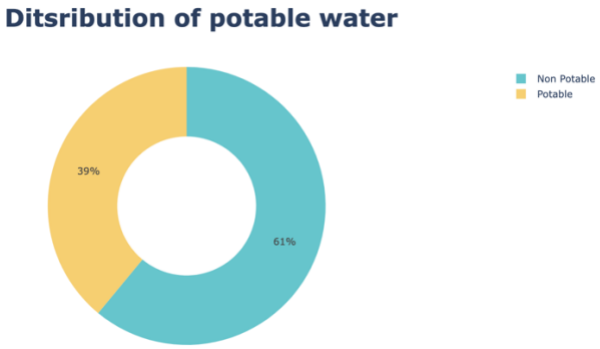

Fig6: Representation of potable water

II. Distribution of Numerical Attributes: It is evident that all the features follow gaussian like distribution. (no need for scaling the attributes)

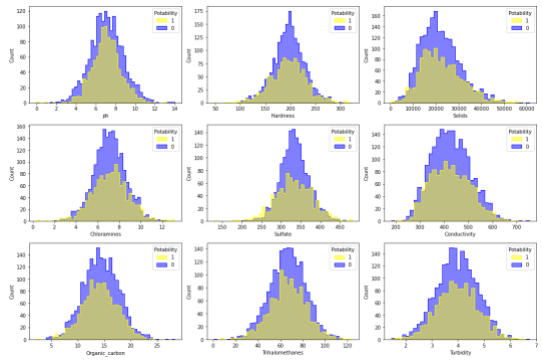

Fig7:Distribution of Attributes

III. Correlation Plot: It is evident from the correlation plot that none of the features are highly correlated with

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
ph	1.000000	0.075833	-0.081884	-0.031811	0.014403	0.017192	0.040061	0.002994	-0.036222	-0.003287
Hardness	0.075833	1.000000	-0.046899	-0.030054	-0.092766	-0.023915	0.003610	-0.012690	-0.014449	-0.013837
Solids	-0.081884	-0.046899	1.000000	-0.070148	-0.149840	0.013831	0.010242	-0.008875	0.019546	0.033743
Chloramines	-0.031811	-0.030054	-0.070148	1.000000	0.023791	-0.020486	-0.012653	0.016627	0.002363	0.023779
Sulfate	0.014403	-0.092766	-0.149840	0.023791	1.000000	-0.014059	0.026909	-0.025605	-0.009790	-0.020619
Conductivity	0.017192	-0.023915	0.013831	-0.020486	-0.014059	1.000000	0.020966	0.001255	0.005798	-0.008128
Organic_carbon	0.040061	0.003610	0.010242	-0.012653	0.026909	0.020966	1.000000	-0.012976	-0.027308	-0.030001
Trihalomethanes	0.002994	-0.012690	-0.008875	0.016627	-0.025605	0.001255	-0.012976	1.000000	-0.021502	0.006960
Turbidity	-0.036222	-0.014449	0.019546	0.002363	-0.009790	0.005798	-0.027308	-0.021502	1.000000	0.001581
Potability	-0.003287	-0.013837	0.033743	0.023779	-0.020619	-0.008128	-0.030001	0.006960	0.001581	1.000000

Fig 8: Correlation Plot

IV. Scatterplot for all the variables.

From the Fig 9 , it is evident that all the numerical attributes follow gaussian or gaussian like distributions.

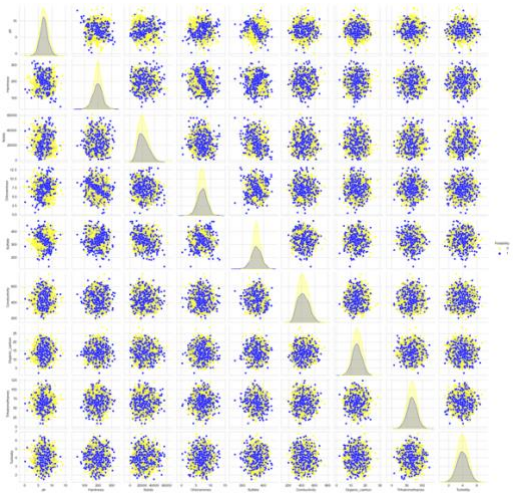

Fig 9 : Scatter plot

The data is split into 75% of training and 25% into testing and is ready for model training

2.3 Random Forest Modelling

Random Forest classifier is trained with 100 number of estimators, entropy criterion and max depth of 3. Just to visualize how the random forest works. Let's consider the first decision tree ,Tree has total 15 nodes at 3 depths

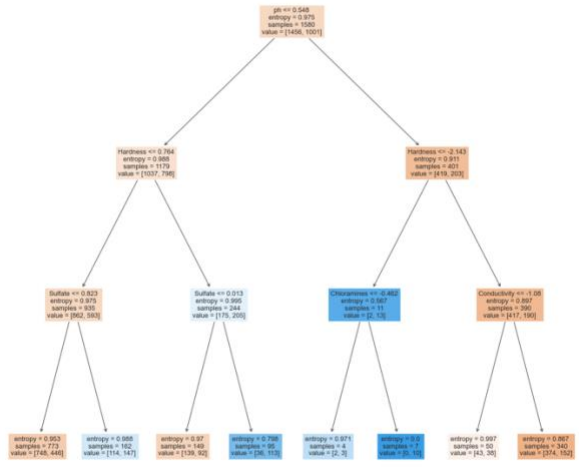

Fig 10 : A decision tree from RF classifier

In this Tree, *Ph value and Hardness* decides the potability of the water for this Tree.

The predictions are not accurate with this basic model, has a accuracy of 64% it can be further improved by tuning hyper parameters of the model.

Hence to improve this model, a Grid search has been on random classifier. Best Grid is selected from [50,100,200,300] *n_estimators*,[Gini, entropy, log loss] *criterion*, [5,6,7,8,9,10,20,50,100] *max depth* and *oob score* set to True.

```
In [33]: param_grid={
             'n_estimators':[50,100,200,300],
             'criterion':['gini','entropy','log_loss'],
             'max_depth':[5,6,7,8,9,10,20,50,100],
             'oob_score':['True']
         }
```

Fig 11: Grid parameters

After performing Grid search on random Forest classifier , with given grid and with 10-fold cross validation. It is found that Best hyper parameters are n_estimators = 300 , criterion = "gini", max_depth = 20. Let's fit this classifier for testing data and find out the results.

2.4. Model Evaluation Metrics

Overall accuracy, recall, precision, f1 score, and AUC of the ROC curve are the most common evaluation measures for classification models.

False positives (FP) and false negatives (FN) are outcomes that were mistakenly classified by the model, while true positives (TP) and true negatives (TN) are outcomes of the positive class and negative class, respectively.

a) Overall Accuracy (OA): This is defined by the following the equation

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

The model could attain almost perfect overall accuracy if it consistently predicts the majority of classes.

b) This issue is more severe the more unbalanced the data. So, we require additional measurements. include Recall. It measures the proportion of accurately predicted positive classes to all positively categorized items.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Recall is important when we believe False Negatives are more important than False Positives

c) Precision: It is the ratio of correctly predicted positive classes to all items predicted to be positive

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

It tells us how correct or precise that our model's positive predictions are. When we think False Positives are more significant than False Negatives, precision is crucial.

d) F1- Score : The F1-score is a single performance statistic that considers both recall and precision. It is also frequently

referred to as the F-Measure. It is calculated by averaging the two metrics harmonically.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

with values closer to one indicating better performance, and values closer to zero indicating poorer performance

e) Out-Of-Bag (OOB) Error: This metric is the accuracy of examples x_i using all the trees in the random forest ensemble for which it was omitted during training. Thus, it kind of acts as a semi-testing instance.

f) ROC Curve: The ROC Curve is helpful since it not only provides a summary of our model's performance but also makes it simple to compare the effectiveness of other classifiers.

1.1 Test Results

The classifier score for the training data is 0.998 nearly 100%, That is accuracy of the model evaluating the training data is very high. This might be due to the overfitting the training data by the model. To avoid the metric , OOB error is introduced, this is the accuracy of the model when evaluating the training data only for the trees which model has omitted. This model got an *OOB score of 0.66*

Let's Look at classification report,

	precision	recall	f1-score	support
0	0.68	0.91	0.78	510
1	0.67	0.31	0.42	309
accuracy			0.68	819
macro avg	0.68	0.61	0.60	819
weighted avg	0.68	0.68	0.64	819

Fig 12 : Classification report

From above classification report , it is evident that accuracy of the model is 0.68 which is very similar to OOB. It thus follows through the theory that the oob accuracy is a better metric by which to evaluate the performance of your model rather than just the score.

Also lets look at confusion matrix,

Fig 13: Confusion Matrix

The model has predicted 47 non potable water samples as safe drinking water. Which is a huge concern, when this type of predictions happens in real time.

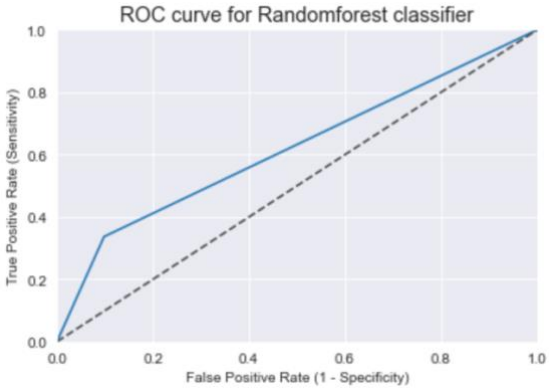

Fig 13: ROC Curve

And the AUC (Area under the curve) is around 63% (>50%) . This can be improved further.

3. CONCLUSION

Random forest classifier did moderate work in binary classification of water potability with accuracy of 68% and AUC of 63%. This score can be improved by ensemble Methods (Boosting and Bagging).

4. REFERENCES

1) CSE 474/574 Lecture slides, *Prof Chen*
2) Random Forest Classifier Tutorial, Kaggle
3) *Leo Breiman*, RANDOM FORESTS
4) Random Forest Blog by IBM