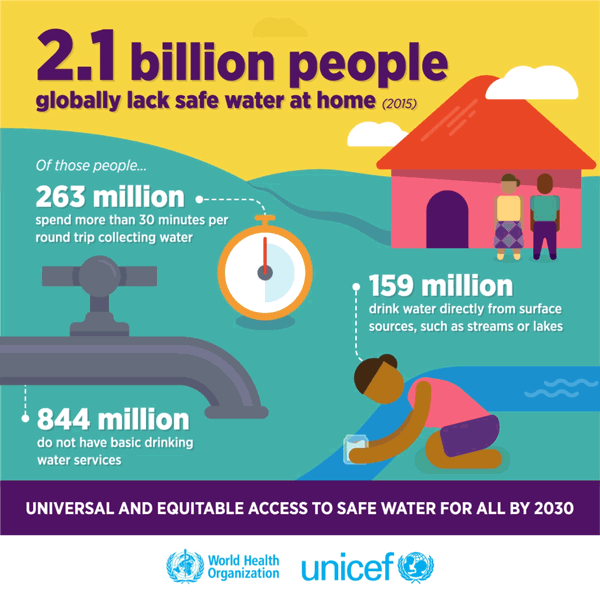
Water Quality Prediction using Machine Learning

# 1. Introduction

There has been an increase in the severity of water shortages in semi-arid regions in recent years. Recent estimates indicate that over 5 billion people may be affected by water shortages within the next decade. Because of the lack of fresh water and a lot of waste, water pollution problems are widespread in dry regions. As human and economic growth becomes more important, the quality of water in water bodies is increasingly being taken into consideration. This means that, in order for the growth of a society and the economy to be possible, it is necessary to evaluate and estimate water quality levels. Water is a very important natural resource on the planet, necessary for the survival of the vast majority of species, including humans. Water with sufficient purity is required by all living organisms to maintain their existence. Water species can endure only a certain amount of pollution. Excessive use of these resources has a negative impact on the environment and puts the lives of these creatures in jeopardy.

There are quality standards in place for most bodies of ambient water, including such rivers, lakes, and streams. Other water specifications/uses have their own criteria, too. So irrigation water, for example, must be overly salty or include hazardous elements that can be conveyed to plant or soil and endanger ecosystems by destroying them. Depending on the industrial process, water quality requirements change for industrial usage. Many of the low-cost fresh water sources, such as groundwater and surface water, come from the earth's natural supply. Human/industrial activity and other natural processes, however, can pollute these resources and degrade their quality.

Testing water quality samples in a lab involves a long and expensive process. As a result, it is impossible to forecast the quality of the water in advance. With the current advancements in Machine Learning, previous traditions have once again begun to emerge. In this project, we aim to use these Machine Learning algorithms to predict the quality of the water using various numerical variables.



## 1.1. Data

The dataset that is being used in this project to predict the water quality is available as secondary data on the Kaggle platform. Kaggle is a community platform for Data Scientists around the globe. The dataset consists of a total of over 3200 data points.

This data is classified based on nine numeric variables that describe the quality of water. The following are the variables in the dataset.

* pH Value: pH is a critical component in determining the acid-base balance in the water.
* Hardness: The most common cause of hardness in the water is because of the presence of calcium and magnesium ions in the water.
* Solids: Many different elements and compounds may be dissolved in water including potassium, calcium, sodium, magnesium, sulphates, etc. producing unwanted taste and diminished colour of the water.
* Chloramines: The main disinfectants utilised in public water supply are chlorine and chloramine.
* Sulphate: Sulphates are present in rocks, soil, and minerals as natural compounds.
* Conductivity: Water is an excellent insulating substance of electricity.
* Organiccarbon: Decaying natural organic matter (NOM) is a large contributor to total organic carbon (TOC) in water sources, along with manmade sources.
* Trihalomethanes: These compounds are present in water that has been treated with chlorine.
* Turbidity: Water's turbidity relies on the amount of solids suspended in the water.

## 1.2. Objectives

* Perform through background research on the similar topics and refer to the research papers.
* Learn and understand in depth of the algorithms that are being used as part of this project.
* Train the models on the data and produce effective results.
* Critically evaluate the models and compare the performances of the models.

## 1.3. Research Questions

## A research topic is a topic that a study tries to find answers for. This refers to something in the study that's addressed by examining the data and telling the storey it reveals. The research objective is typically formulated in a way that highlights a variety of aspects, such as the study's research population and factors, and also the study's main purpose. Research is commonly centred around scientific research. It's not surprising that researchers frequently revisit and revamp their research questions: Research questions tend to be evolving rather than static. Researchers must reassess and adjust questions as they conduct literatures and build a framework for the study. The research questions that we are trying to address in this project are:

* How does the use of feature engineering will help in improving the performance of the Machine Learning classifiers?
* What classifier shows promising results in classification of water quality?

## 1.4. Evaluation Metrics

The evaluation metrics that will be used in this project to evaluate the Machine Learning models are accuracy and confusion matrices.

Accuracy: The proportion of accurate predictions for test data is what we call accuracy. To determine the average percentage of accurate predictions, divide the total of correct predictions by the total number of predictions.

Confusion Matrix: A Confusion matrix is a N x N representation of a matrix that is used to evaluate the effectiveness of an algorithm, wherein N represents the size of target classes.

## 1.5. Tools

There are several tools employed as phase of this project to achieve its main objective. a few essential tools utilized during this Endeavour include:

NumPy: NumPy is a Library in python designed to help with array management. It has features for linear algebra, Fourier transforms, and matrices, as well.

Pandas: Pandas is a Python-based library that works with data analysis and manipulation. The focus is on the operations and data structures required to manipulate tables and time - series data.

Scikit-learn: Scikit-learn is a library for Python that helps with machine learning. It is open source and available to anyone. It includes several algorithms, such as support vector machines, among others.

Matplotlib: Matplotlib is a tremendous 2D plotting library in Python, perfect for visualizing array data. Matplotlib is a library built on the concept of NumPy arrays, and it is made to work with the other components of the Scipy stack. Matplotlib has numerous plots, including lines, bars, scatter plots, histograms, and more.

Seaborn: Seaborn is an example of a Python library that works with the matplotlib data visualization framework and integrates with pandas data structures. Seaborn is Seaborn's central visualization system, which is crucial in helping the exploration of data. See how the distribution is univariate and bivariate.

# 2. Ethical, Social and Legal issues

**1. Ethical Issues:**

The risk zones that the current project falls into are listed below:

Machine Ethics & Algorithmic Biases – Due to the imbalanced nature of the real-world data, the algorithms that are trained on such imbalanced data without proper precautions can behave biased and lead to misclassification of the data points.

Necessary actions such as checking the imbalance in the data, using the techniques to eliminate the imbalance in the data should always be encouraged.

Surveillance State – Though the project aims to predict the quality of the water, it is highly essential to track the individuals who rely on these predictions to monitor their health conditions regularly.

The individuals should be aware of them being tracked and should consent to their acknowledgement to monitor their health conditions.

**2. Social Issues:**

The social issues that can cause a problem as a result of the project are political-related issues and unemployment.

In some areas of developing countries, the automation of water quality can be obstructed by the political parties due to a lack of awareness.

Automating the prediction of the quality of water can be a threat to the individuals employed in this stream who perform the quality tests manually.

**3. Legal Issues:**

Any harm caused to the individuals who rely on the predictions of the algorithms can attract legal obligations from the supporters and other people. This will severely damage the belief in the technology among those people.

## 3. Methodology

## 3.1. Installing set-up

I have used Python 3.7, which I downloaded from the official Python website, for my project. I installed Python on my machine by setting it up on my hard drive and adding Python to my path. Additionally, I've installed Anaconda on my computer to get everything ready for running Python via Jupyter. After that, I was able to execute this project by using command prompt to install a few libraries.

Example : Pandas, Numpy, Scikit-learn etc.

## 3.2.Data Exploration

The data for this project is available at Kaggle [- https://www.kaggle.com/adityakadiwal/water-potability](-%20https:/www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes). The dataset includes records with 33 variables combination data kinds make up the information.

The dataset is further explored to identify Exploratory data Analysis.

## 3.3 Exploratory Data Visualization

Due to the greater number of useful features in the data, the data gives us a high score of exploratory data analysis on the data. Data Analysis will help us understand the patterns in each feature with respect to the dependent variable and helps us understand how each feature is contributing to the water portability prediction.

Take the example of your wolf pack watching a movie you've never heard of. There's no arguing with that. It'll leave you perplexed with a slew of questions that need answering before you can make a conclusion. As a good chieftain, one of the first things you'd want to know is who's among the cast and crew. You'd also see the movie's trailer on YouTube on a daily basis. You'd also learn about the movie's popularity based on audience feedback such as ratings and reviews. In data scientists' jargon, 'Exploratory Data Analysis' refers to any investigative measures you may take before buying snacks for your family at the movie theatre. Using statistics and graphical representations, exploratory data analysis identifies the crucial process of conducting early investigations on data to uncover patterns, spot anomalies, test hypotheses, and double-check assumptions. When working with data, it's important to first comprehend it so you can extract as many insights as possible. Using EDA, data in hand can be made sense of prior to getting messy.

## 3.3.1 Univariate Analysis

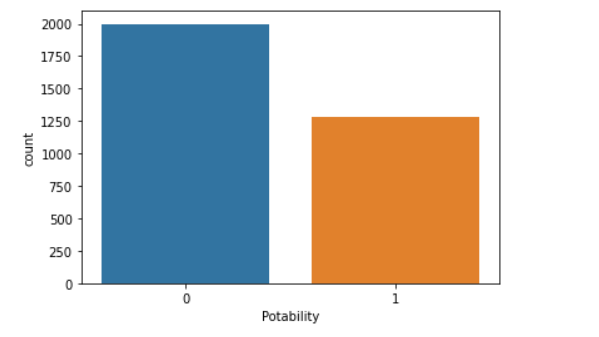
The technique of univariate analysis is for comparing and analyzing the relationship between a single feature and response variable. The prefix "uni" highlights the analysis only covering a single variable and its impact on a parameter.

For Example, the study can focus on a variable such as "gender," "height," or "weight."

However, only one variable is examined each time.

## 3.3.1.1 Potability

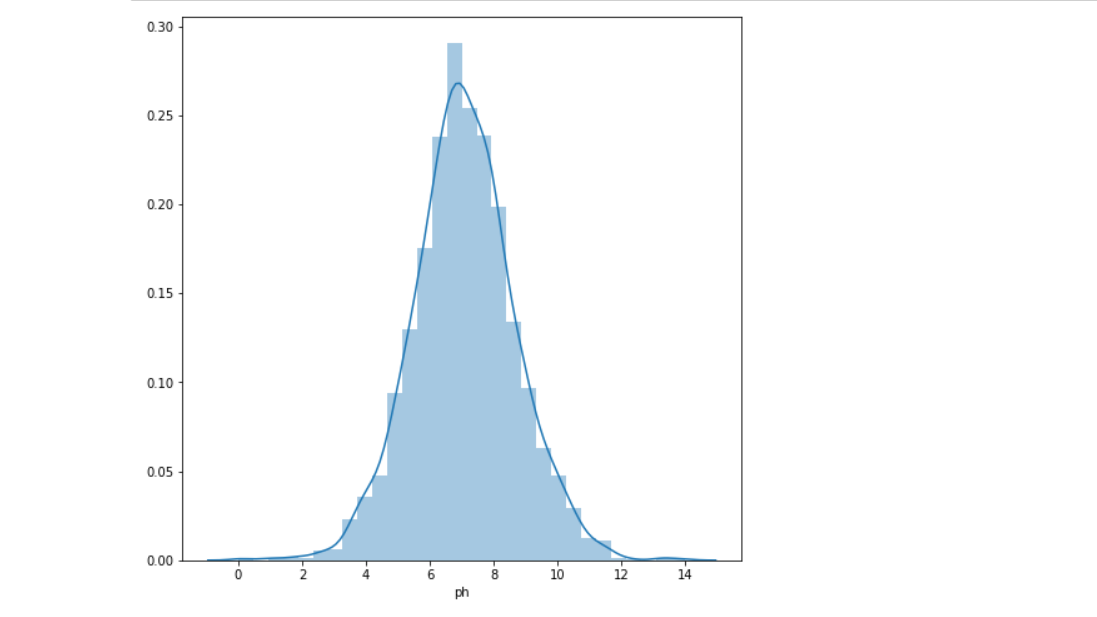
Potability is Our Target Variable. Potability is a dependent variable in our dataset which is categorical feature. This feature Indicates if safe for human consumption or not where 1 means potable and 0 means not potable .There are two categories 1 and 0. For counting the number of categories of potable and not potable i have plot the Countplot.



Potability Variable consist of 3276 records in which there are around 1998 records are not potable and 1278 records are from potable category.

## 3.3.1.2 Ph value

Ph value is a integer variable which describes about the measurement of acid balance. WHO has suggested maximum limit of pH from 6.5 to 8.5.



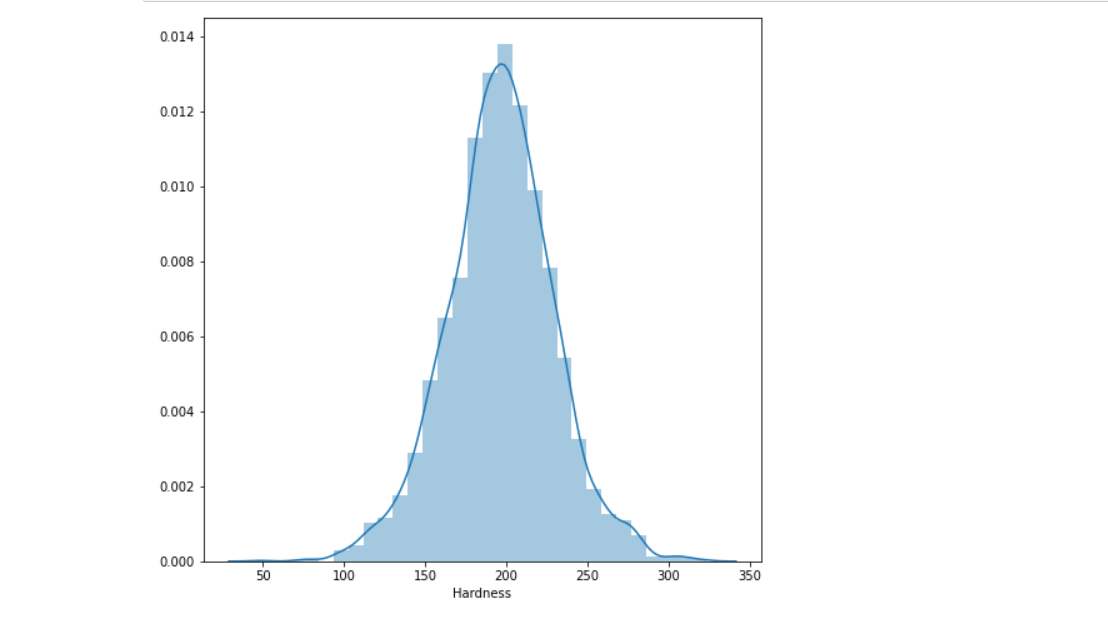
As we can see the above is normally distributed. So we can say that there is no outliers in this feature. The stats of this feature is :

|  |  |
| --- | --- |
| Count | 2785.000 |
| Mean | 7.0807 |
| Std | 1.59 |
| min | 0.00 |
| 25% | 6.0903 |
| 50% | 7.0367 |
| 75% | 8.062 |
| max | 14.000 |

The maximum value of ph is 14.00 and minimum value is 0.00.

## 3.3.1.3 Hardness

Hardness is a integer variable. Hardness in water caused by calcium and magnesium salts. Later on I have divided the hardness value in soft , medium, hard etc. The dist plot of this feature is given below:

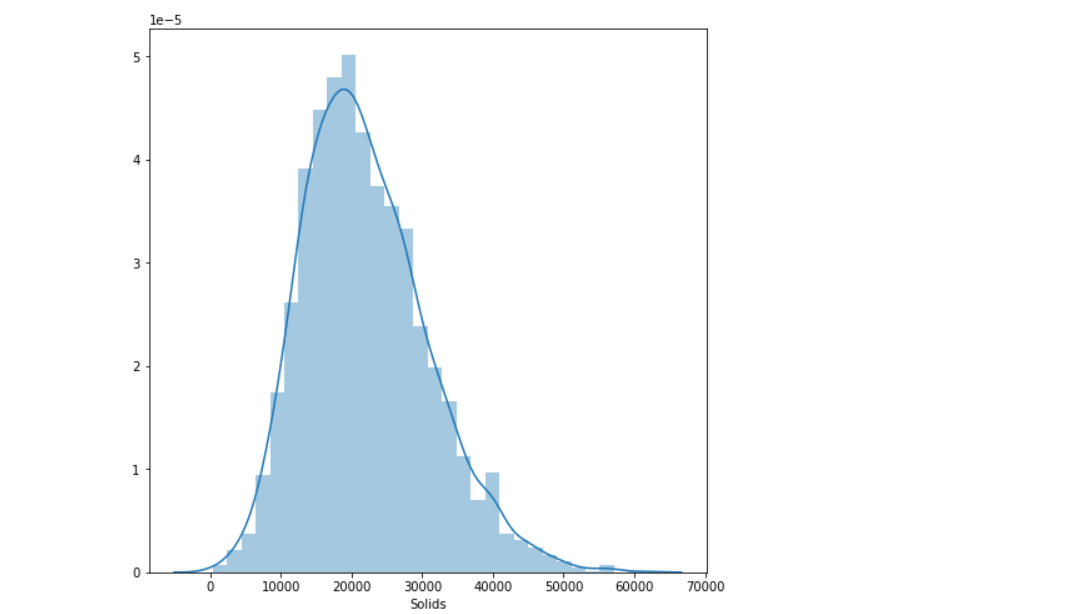


As we can see the above is normally distributed. So we can say that there is no outliers in this feature. The stats of this feature is :

|  |  |
| --- | --- |
| Count | 3276.000 |
| Mean | 196.36 |
| Std | 32.87 |
| min | 47.43 |
| 25% | 176.85 |
| 50% | 196.65 |
| 75% | 216.65 |
| max | 323.1400 |

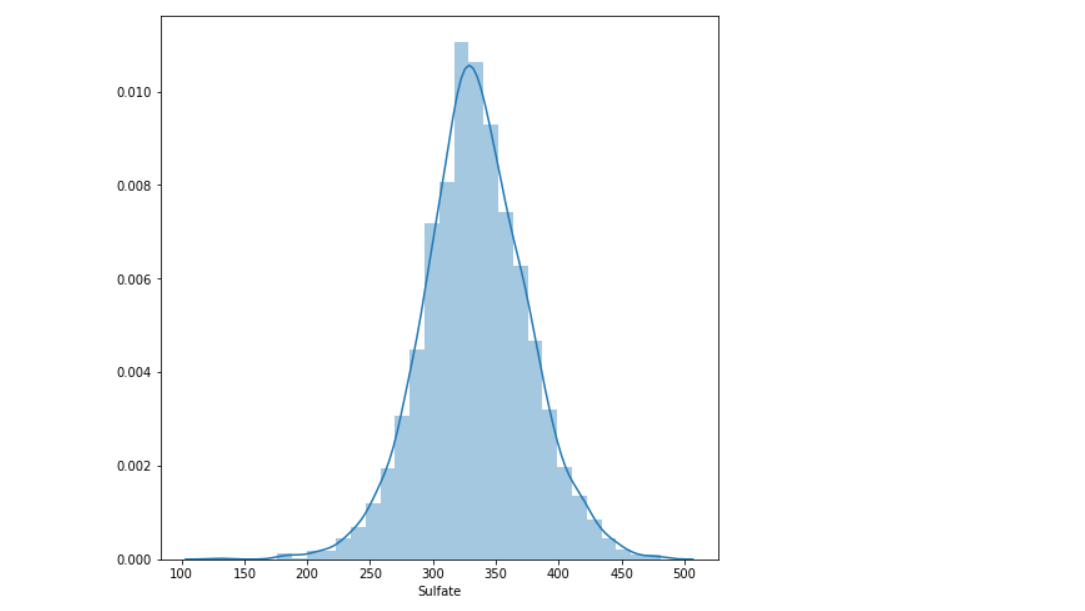
## 3.3.1.4 Solids

Solid is a integer variable. This feature indicates how much the water is mineralized or Total Dissolved Solids. The desired range for TDS is 500 up to 1000 mg/l. The dist plot of this feature is given below:



## 3.3.1.5 Sulfate

This is a integer variable. Sulfates are naturally occurring substances that are found in soil, minerals, and rocks.



As we can see the above is normally distributed. So we can say that there is no outliers in this feature. The stats of this feature is :

|  |  |
| --- | --- |
| Count | 2495.000 |
| Mean | 333.00 |
| Std | 41.4168 |
| min | 129.000 |
| 25% | 307.699 |
| 50% | 333.07 |
| 75% | 359.95 |
| max | 481.00 |

## 3.3.2 Multivariate Analysis

The technique of Multivariate analysis is for comparing and analyzing the relationship between a Multiple features and response variable. The prefix "Multi" highlights the analysis only covering a multi variable and its impact on a parameter.

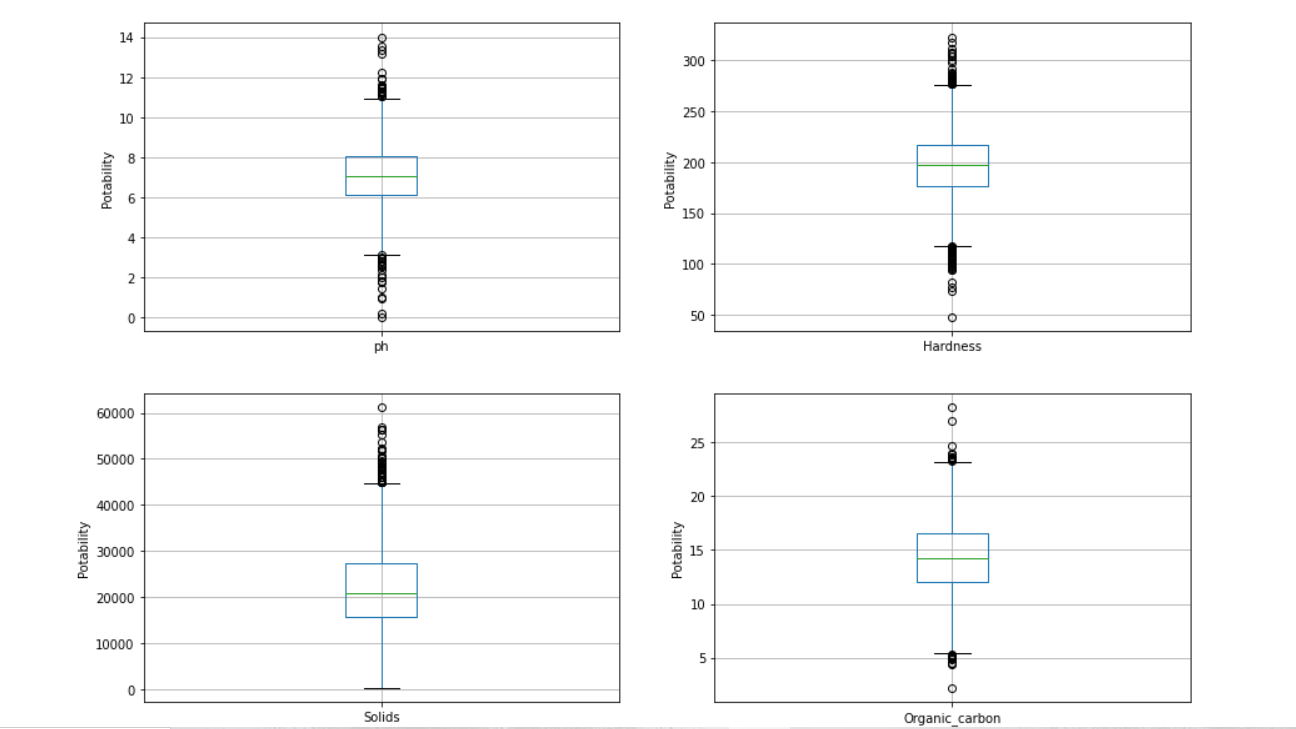
* An important step in EDA is to discover patterns and relationships between variables in the dataset.
* I have used heat map to discover the patterns and relationships in the dataset.



From the heatmap, we can observe from the heatmaps that there is no correlation between the features.

## 3.3.3. Outliers detection

We can see the Outliers in the numerical columns with the help of boxplots. I have plotted the boxplots in numerical columns for outlier detection.



## 3.4 Feature Engineering

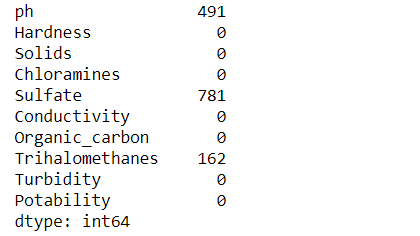
feature engineering makes data easier to examine. The world we live in is imperfect, therefore the data we use in it can be messy and unpredictable. Regardless of the type of data source (e.g., linear SQL database, Excel file, etc.), The data, which is normally formed as a table with a new row representing each sample and each column showing a trait, may be difficult to comprehend and handle.

To better understand our machine learning models' data and hence help them perform better, we need execute feature engineering. The task of converting data into a simpler form of comprehension belongs to feature engineering in learning algorithms. In this case, we are working to make the information clearer for a trained model, but feature can be generated such that data visualizations more approachable for non-data professionals can be created. But understanding the concept of clarity in ml algorithms is complex since the technique varies according on the type of data used.

In this project feature Engineering is most important part because the dataset contains lots of missing values in the features and feature extraction is also require for this project because In this project there are some features like hardness and ph value. From these features we can extract new features. I have perform these following steps in feature Engineering:

## 3.4.1 Missing Values Treatment

In this project some features contains missing value and we cannot remove these missing values because the number of missing values is very high if we remove these rows from the dataset then for sure there is some information loss.



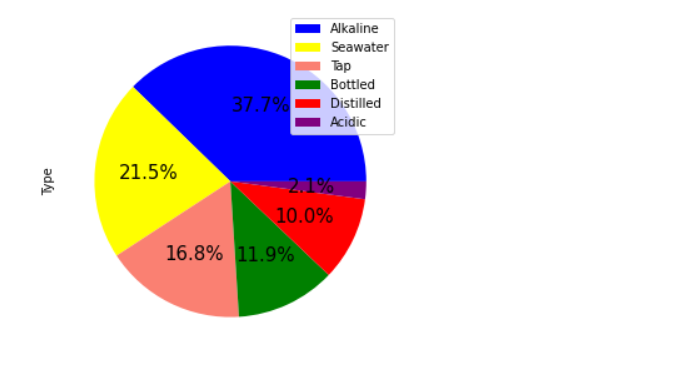
As we can see In the above diagram ph, Sulfate and Trihalomethanes contains missing values. I have imputed these missing values with the median values. Because Mean will deviate if the dataset contains any outlier.

## 3.4.2 Feature Extraction

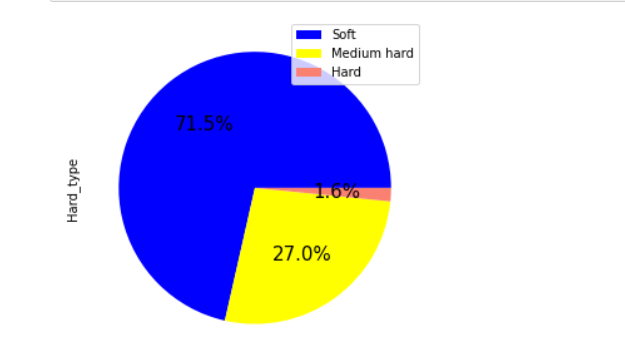
An initial set of raw data is separated and reduced to smaller, more manageable groupings as part of the dimension reduction process, which includes feature extraction. As a result, the processing will be simpler when the time comes. The fact that these datasets contain a significant number of variables is critical. To process all of these variables, you'll need a lot of computational power. By selecting and combing variables into features, feature extraction aids in obtaining the finest feature from large data sets while also successfully decreasing the amount of information. These features are simple to use, but they accurately and uniquely define the real data set.

In this project I have extracted the features from the given features ph, hardness.

For Example : if ph value is greater that 8 then this class belongs from Alkaline. If ph value is greater than 7.5 and less than 8 then this data point belongs to Seawater category. These are the categories which I divided from ph variable. Then type becomes new feature in our dataset. In which Alkaline, Seawater, Tap, Battled etc are the sub categories.



These are the categories which I divided from hardness variable.



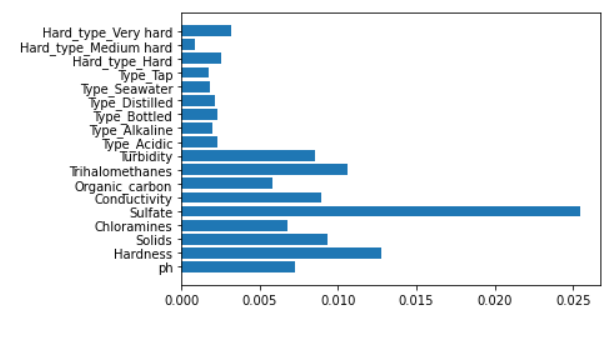
These are the categories which I divided from Hardness variable. Then hardness\_type becomes new features in our dataset. In which Soft, Medium hard, hard are the new features.

## 3.4.3. Encoding of Categorical Features

All inputs and Output variables must be numeric for a machine learning model. To fit and evaluate a model, we have to encode your category data into numbers first. I have used One hot Encoding in this project.

## 3.4.4 Feature Importance

When we talk about feature importance, we're talking about methods that rate the usefulness of input information in terms of predicting a target variable. Although notable examples include rank correlation scores, coefficients derived in linear regression, decision trees, and permutations importance scores, feature significance scores can take many different forms and be derived from many different places. Ratings on feature importance are critical in predictive modelling projects, since these scores provide information about both the problem and its solution. They also serve as the basis for dimension reduction and feature selection, both of which increase the predictive model's performance.



All features are important for the model building because none of the feature have value 0.

## 3.5 Balancing the Dataset

As we know the given dataset is imbalance dataset. So, we have to balance it before giving it to ML Model. I have used Smote for balancing the dataset.

#### SMOTE ( Synthetic Minority Oversampling Technique) :

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling.

## 3.6. Algorithms and Techniques

I have used Machine learning Algorithms for portability prediction.

## 3.6.1 Machine Learning

Machine learning can be addressed as a subset of artificial intelligence that allows machines to learn and develop automatically without any supervision of the humans. Machine learning is concerned with the development of computer programmes capable of having access to data and use it to train themselves. The learning phase starts with insights or facts, such as instructions to search for correlations in data and make more informed potential judgments based on the examples provided. The aim of the machine learning is to allow computers to train on the data automatically and change their activities as per the requirements without any human involvement or assistance.

Machine learning allows the processing of enormous amounts of data. Although it usually produces more reliable and timely outcomes for identifying lucrative opportunities or risky threats, correctly training it may take additional time and effort. Combining machine learning and artificial intelligence with variety of technologies have the potential to render it much more efficient at analyzing vast amounts of data. Within the real of machine learning, there are several algorithms that are published every day and they are organized based on whether they focus on supervised or unsupervised learning or feature similarities. Machine learning algorithms can be classified as two types. Namely, supervised machine learning algorithms and unsupervised machine learning algorithms.

For this project I have used Machine Learning Algorithms like Decision Tree , Random Forest , Xg-boost and Gradient boost Algorithms etc.

## 3.6.1.1 Decision Tree

Algorithms that use Decision Trees fall under the category of learning algorithms. The decision tree approach, in contrast to other supervised learning methods, can also be utilized to solve regression and classification issues. By observing and learning from prior data, the purpose of utilizing a Decision Tree would be to develop a training models that can be used to determine the class or values of the target attribute (training data).

In Decision Trees, we start from of the tree root to forecast a label for a record. We do a value comparison between the root attribute and the record's root attribute to see which one is greater. We take the branch that corresponds to that number and then move on to the next component based on the comparison. Machine learning algorithms can be classified as two types.

## Types of Decision Trees

There are two popular types of Decision tree:

## 1. CART (Classification and Regression Tree ):

This type of decision tree can be use for both classifications as well regression. These decision trees use gini method to create the split points.

## 2.ID3 (Iterative Dichotomiser) :

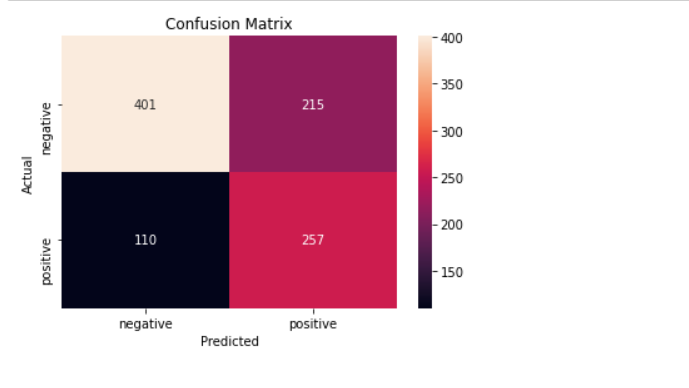
The splitting points are determined by the ID3 decision tree algorithm using Information Gain. We can compute the heterogeneity of a sample using entropy to see how much information we've gained

## Hyperparameter Tuning in Decision Tree

I have applied Hyperparameter tuning on Decision Tree. After Appling Hyperparameter tuning we can see that

{'criterion': 'entropy', 'max\_depth': 5, 'min\_samples\_leaf': 5, 'min\_samples\_split': 2}

These are the best parameters. For Training data the algorithm gives 68.08% accuracy value and with testing data the algorithm gives 66.93% accuracy. Clearly we can see that there is no overfitting. The Confusion Matrix for Decision tree is :



## 3.6.1.2 Random Forest Algorithm

One of the most prominent algorithms for supervised learning is Random Forest, which uses an approach known as regression. ML applications can make use of it to solve classification and regression difficulties. Ensemble learning refers to the methodology of amalgamating several classifiers to deal with an elaborate problem and to better the model performance.

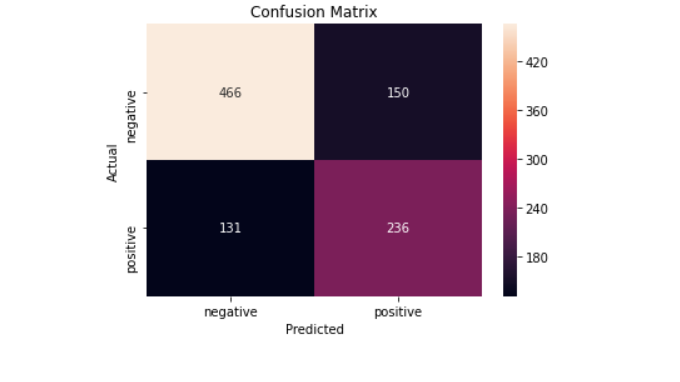
Its name is self-explanatory, as it uses a number of different decision trees on various subsets of the given dataset to take an average and improve the dataset's predictive accuracy. The random forest doesn't only rely on one decision tree. It instead bases its forecasts on majority voting amongst all the predictions and then predicts the ultimate result.

The below diagram explains the working of the Random Forest algorithm:



## Hyper parameter Tuning in Random Forest Classifier

After applying hyper parameter tuning in Random forest Algorithm we can see that for 100 number of estimators and for depth of 70 the algorithm gives higher accuracy The model gives around 76.44% accuracy on training data and 71.41% accuracy for testing data. So there is no overfitting in the model. The Confusion Matrix for Random forest is :



## 3.6.1.3 Xg- Boost

XGBoost is a variation of the ensemble learning technique. The findings of a single machine learning model may not be enough to rely on sometimes. Ensemble learning is a solution to combining the predictive potential of several learners, and it is more orderly than many other alternatives. The end result is a model that puts together all of the results from multiple models.

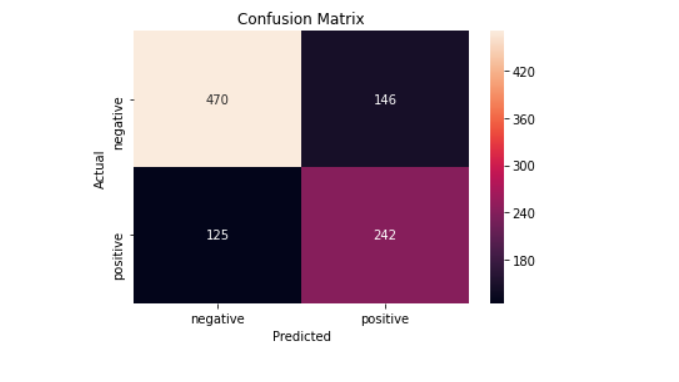
Base learners are generally split between multiple algorithms, though occasionally the same method will be replicated for simplicity. Two kinds of ensemble learners with a lot of traction are boosting and bagging. Decision trees are the most common statistical methodology to which these two methods have been applied.

## Hyperparameter Tuning in Xg-Boost Classifier

I have Applied Hyperparameter tuning on Xg-boost classifier for finding the best parameters. After Applying Hyperparameter tuning we can see that the best parameters are:

Base learner=100, Optimal depth=5

For these parameters the accuracy on training data is 76.23% and for testing data the accuracy is 71.43%. hence we can say that there is no overfitting in the model. The confusion matrix for Xg-Boost Classifier is :



## 3.6.1.4 Gradient Boost Decision Tree

Gradient boosting is a machine learning technique in which a model's performance is enhanced by combining the efforts of several weak learners. Decision trees tend to be bad students because they don't learn anything. Weak learners work in a sequential manner when using gradient boosting. Each iteration seeks to eliminate the flaws introduced by the preceding iteration. Many weak learners are combined into one powerful one in gradient boosted trees. Individual decision trees seem to be poor students in this case.

All of the trees are linked together in a chain, with each successive tree attempting to reduce the error introduced by the one before it. Boosting algorithms take a long time to train because of this sequential relationship, but they are also quite accurate. Models that learn more slowly do better in statistical learning.

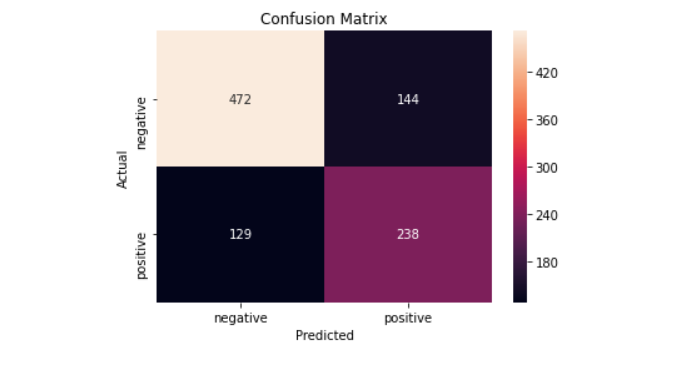
To enhance the model, the weak learners are re-fit so that each individual learner matches the residuals from the prior stage. The final model combines all of the results from the previous steps, resulting in a very capable learner.

## Hyperparameter Tuning in GBDT Classifier

I have Applied Hyperparameter tuning on GBDT classifier for finding the best parameters. After Applying Hyperparameter tuning we can see that the best parameters are:

Base learner=500, Optimal depth=7

For these parameters the accuracy on training data is 73.66% and for testing data the accuracy is 72.27%. hence we can say that there is no overfitting in the model. The Confusion matrix for Gradient boost classifier is :



## 

## 3.7 Performance Evaluation Metrics

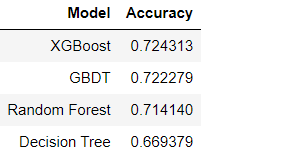
While it is important to validate an evaluation of a model's skill, using metrics is necessary to evaluate the performance of a model. Metrics are employed to determine the most suitable problem-solving method. These metrics are what matter most.

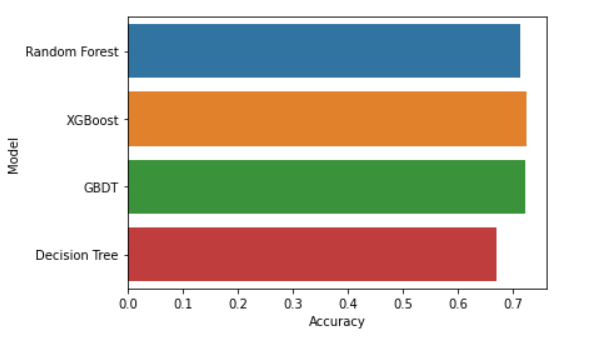
## 1. Accuracy

Accurately identifying the ratio of correctly predicted instances to the total amount of instances evaluated is a metric for measuring accuracy. While the accuracy might be adequate as a measurement for model performance, it may not be good enough because it fails to include incorrect predictions. If someone treats a fake post as a real one, it could cause a serious issue. False positives and false negative issues that accommodate for misclassification should be considered because of this.



The Accuracy of each model is :





2. Confusion Matrix:

A classification algorithm's performance can be summarized using a confusion matrix. A dataset with unequal numbers of samples in each class or more than two classes can be misled solely by classification accuracy.

It is possible to acquire a better understanding of your classification model's accuracy by computing a confusion matrix.

1. True Positive Rate : The true positive rate, also known as sensitivity or recall, is a metric used in machine learning to assess the proportion of real positives that are accurately detected.

2. False Positive Rate : There are two ways to compute the false positive rate: as FP/FP+TN and as FP/FP+TN (FP+TN being the total number of negatives). It's the likelihood that a false alarm will be sounded, resulting in a positive result when the underlying value is actually negative.

3. False Negative Rate : A false negative result that is actually positive. It's an example of a false negative when the results of a cancer test come back negative even though the person has cancer.

4. True Negative Rate : They don't have the sickness, which is exactly what we thought would be the case. Although we projected that they would be positive, they don't have the condition.... When a test results in a false negative (FN), it means that we projected the patient did not have the condition. A "Type II mistake" is another name for this phenomenon.)

## 

## 4. Conclusion

Water is a very important natural resource on the planet, necessary for the survival of the vast majority of species, including humans. Water with sufficient purity is required by all living organisms to maintain their existence. Water species can endure only a certain amount of pollution. Excessive use of these resources has a negative impact on the environment and puts the lives of these creatures in jeopardy.

In this paper, I have worked with different models of Machine Learning to predict the Portability of the water. from the given data to identify the better performing models. Firstly I have applied exploratory data Analysis on the given dataset to gain the insights of the dataset. After EDA I have applied feature Engineering methods. I have applied feature extraction on the given features (PH and Hardness) . from these features I have derived features like Acidic, Alkaline etc. After that I have checked feature importance of each column. The given dataset Is imbalanced dataset so I have applied SMOTE for balancing the dataset. After that I have applied some machine learning models (Random forest, Decision Tree, GBDT, XGBoost) in order to gain the better accuracy.

Based on the analysis my findings from answering the two research questions:

These are some feature Engineering methods that can be used in the data to extract more useful information:

* Null Value Imputation
* Feature Extraction
* One hot Encoding
* Feature Scaling

These Feature Engineering methods helps to increase the accuracy of the machine learning models. If our dataset contains null values then Machine learning models will not be able to perform well .So it is required to use feature engineering before giving the data to machine learning model. Random Forest and Xg-Boost gives better accuracy with testing data and also with the training data and the model is also not overfitted.

## 