

CSCE-5300 Introduction to Big Data and Data Science

Final Increment

Project Title: Water Potability Prediction

Team Members

Group – 1

1. Harsha Vardhana Buddana
2. Keerthi Priya Bankuru
3. Eswara Reddy Thimmapuram
4. Geetha Krishna Dodda

Motivation:

Maintaining good health is a significant factor for one's life to sustain and water plays a crucial role in achieving it. Apart from availability, it is important for the water to be safe and hygiene which makes it beneficial to use in our daily life. This has been the major health concern in many countries since decades. Today billions of individuals need admittance to securely oversee drinking water administrations. Hazardous cleanliness rehearses are far extensive, intensifying the consequences for individual's wellbeing. Many children die every day from various diarrheal diseases due to the poor sanitation and hygiene.

In certain areas of the world there is next to zero familiarity with great cleanliness and their part in diminishing the spread of disease. Investing in water supply and sanitation has been shown to have a positive impact on the economy in many areas. So, we came up with this project to check and report the quality of water whether it is portable or not for usage in one's daily life.

Objective:

The quality of water should be achieved through pollutant control measures. water quality that should be accomplished through pollutant control measures. Various objectives describe about the contaminant concentrations. These objectives are intended to address the most extreme measure of toxins that can stay in the water section without causing any unfriendly impact on organic entities involving the aquatic system as habitat, on individuals drinking those life forms or water, and on other possible advantageous uses.

There are nine different factors that WHO (World Health Organization) considers to decide whether the given sample of water is potable or not. The factors are

1. pHvalue
2. Hardness
3. Dissolved Solids
4. Chloramines
5. Turbidity
6. Trihalomethanes
7. Organic carbon
8. Conductivity
9. Sulphate.

After this analysis we create a Machine Learning model that can classify the given sample of water is fit for drinking or not by evaluating the above factors values.

Significance:

- From the very long existence of earth water has always played a crucial role in every living being's life. Now-a-days due to many changes in the environment the water purity has changed and started affecting the everyone's health causing illness like gastrointestinal illness, reproductive problems etc. Due to many chemicals and harsh elements release into waterbodies the values have been increased compared to the original values.
- Water is a worldwide issue. Many of the people in the world can't have the option to utilize safe drinking water and many individuals pass on because of water borne infections. So, using this project we can identify based on the range of values whether the water is portable or not for usage.

Features:

- Here we have used HIVE for storing the data and also used for analyzing the raw data.
- In shallow learning we have used the machine learning models like Logistic Regression, Random Forest, LightGBM, Support Vector Machines, XGBoost, Gaussian Naïve Bayes, Bernoulli Naive Bayes, KNN, Decision tree and Bagging Classifier.
- We used Keras for developing deep learning models which is a powerful open source library.

Dataset:

- We collected the dataset from Kaggle, which is scraped from the WHO website and the water samples are taken from wide range of sources from tap water, lake water to the Ocean water.

RELATED WORK

This section describes the background knowledge on several practical approaches and methods available to evaluate water quality. Usually, conventional lab analysis and statistical analysis are used in analysis to determine water quality, while certain researches involve machine learning techniques for finding an optimum solution for the water quality problem.

Conventional lab analysis helped us to get a better understanding of water quality problems in various areas. In one such research study[\[1\]](#), Shafaqat Ali et al collected several water samples from different areas and examined them against various parameters using manual lab analysis and found a high existence of chemicals due to sewage waste. Aamir et al[\[2\]](#) examined over 40 different samples from different areas using conventional lab analysis and found a high presence of toxic particles.

After getting adequate information about the water quality research, we then explored some research which is using machine learning techniques to determine water quality. In terms of utilizing machine learning algorithms, Shafi et al.[\[3\]](#) calculated water quality using machine learning algorithms like K Nearest Neighbors(KNN), Deep Natural Networks(DNN) and Support Vector Machines(SVM) with an accuracy of 93% using Deep NN. They have worked based on only three parameters pH, temperature and turbidity. By using only three parameters, it has been a limitation while predicting water quality.

Another such research is performed by Ali et al.[\[4\]](#) who used unsupervised techniques like the Average Linkage method of Hierarchical Clustering to examine the water quality index. However, they did not utilize any standard water quality index to predict the outcome and also they ignored some important parameters of the water quality index during the learning process.

A prediction based on analysis is presented by Krishnan et al.[\[5\]](#) in which Linear regression model is utilized to determine correlation among measured parameter values based on which, estimation of parameters for new sample is computed. In this research, they have ranked water quality using turbidity, pH and dissolved solids.

Similarly, Sun et al.[\[6\]](#) analyzed the water quality using T-S fuzzy neural network model on the data of water quality collected in three years in a certain city.

Abyaneh et al.[\[7\]](#) used two machine learning algorithms namely ANN and Multivariate linear regression and estimated chemical oxygen demand and biochemical oxygen demand. They used four parameters for prediction namely temperature, pH, total suspended solids(TSS) and total suspended(TS).

Sakizadeh[\[8\]](#) predicted the water quality index using three ANN's algorithms including ANN's with early stopping, Ensemble of ANN's and ANN's with Bayesian regularization on the data of 16 groundwater quality variables collected. His study resulted in correlation coefficients between observed and predicted values of 0.94 and 0.77, respectively.

Rankovic et al.[\[9\]](#) predicted the dissolved oxygen using a feedforward neural network(FNN) based on 10 parameters, which again defeats the purpose if it has to be used for a real time water quality index estimation with an IOT system.

Data Set:

The dataset which we are using consider of many different chemical values of different types of water collected by WHO. Based on these values, dataset will be determining using few known domains to check whether the water is potable for drinking purpose or not to mostly avoid the health illness.

Here we will be having 10 columns of chemical types present and 100+ different water samples data values have been collected. The chemicals which we consider has been listed in the objective section. As we know each chemical has it's own range value by which that can be considered as the potable one.

1. **PH Values:** PH is a measure of how acidic/basic water is. The range goes from 0 - 14, with 7 being neutral. pH's of less than 7 indicate acidity, whereas a pH of greater than 7 indicates a base. pH is really a measure of the relative amount of free hydrogen and hydroxyl ions in the water.

Water should contain ph value according to WHO (World Health Organisation):	7
No.of Data points before Pre-Processing	2785.0 (Count)
Average Value in Data Set	7.080795 (Mean)

2. **Hardness:** This is mainly caused due to the high found of calcium and magnesium salts and hard metals to soft metals and down to plastics and soft tissues.

Water should contain Hardness according to WHO (World Health Organization):	0 to 60 mg/L (milligrams per liter)
No.of Data points before Pre-Processing	3276.0 (Count)
Avg value in Data set	196.369496 (Mean)

3. **Solids:** Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfate's etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized.

Water should contain Solids according to WHO:	500 to 1000mg/L (milligrams per liter)
No.of Data points before Pre-Processing	3276.0 (Count)
Avg value in Dataset	22014.092526 (Mean)

4. **Chloramines:** Chloramines are disinfectants used to treat drinking water. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chloramines provide longer-lasting disinfection as the water moves through pipes to consumers.

Water should contain chloramines:	4 milligrams per liter (mg/L)
No.of Data points before Pre-Processing	3276.0 (Count)
Avg value in Dataset	7.122277 (Mean)

5. **Sulfate:** Sulfate is a salt that forms when sulfuric acid reacts with another chemical. It's a broader term for other synthetic sulfate-based chemicals you may be concerned about.

Water should contain sulfate according to WHO:	250 milligrams per liter (mg/l)
No.of Data Points before Pre-processing	2495.0 (Count)
Avg Value in Dataset	333.775777 (Mean)

6. **Conductivity:** A measure of the ability of a substance to conduct electricity; the reciprocal of resistivity. in the case of a solution, the electrolytic conductivity is the current density divided by the electric field strength.

Water should contain conductivity according to WHO :	< 400 μ S/cm
No.of Data Points before Pre-processing	3276.0 (Count)
Avg Value in Dataset	426.205111 (Mean)

7. **Organic Carbon:** Total organic carbon is a measure of the carbon contained within soil organic matter. Continuous pasture builds organic carbon quicker than other rotations.

Water should contain organic Carbon according to WHO:	< 2 mg/L
No.of Data Points before Pre-processing	3276.0 (Count)
Avg Value in Dataset	14.284970 (Mean)

8. **Trihalomethanes:** Trihalomethanes (THMs) are the result of a reaction between the chlorine used for disinfecting tap water and natural organic matter in the water. At elevated levels, THMs have been associated with negative health effects such as cancer and adverse reproductive outcomes.

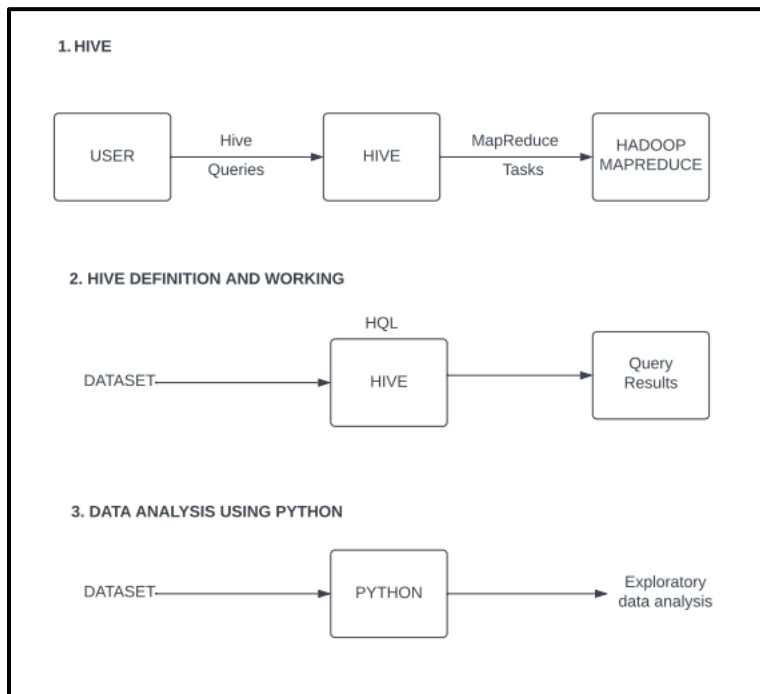
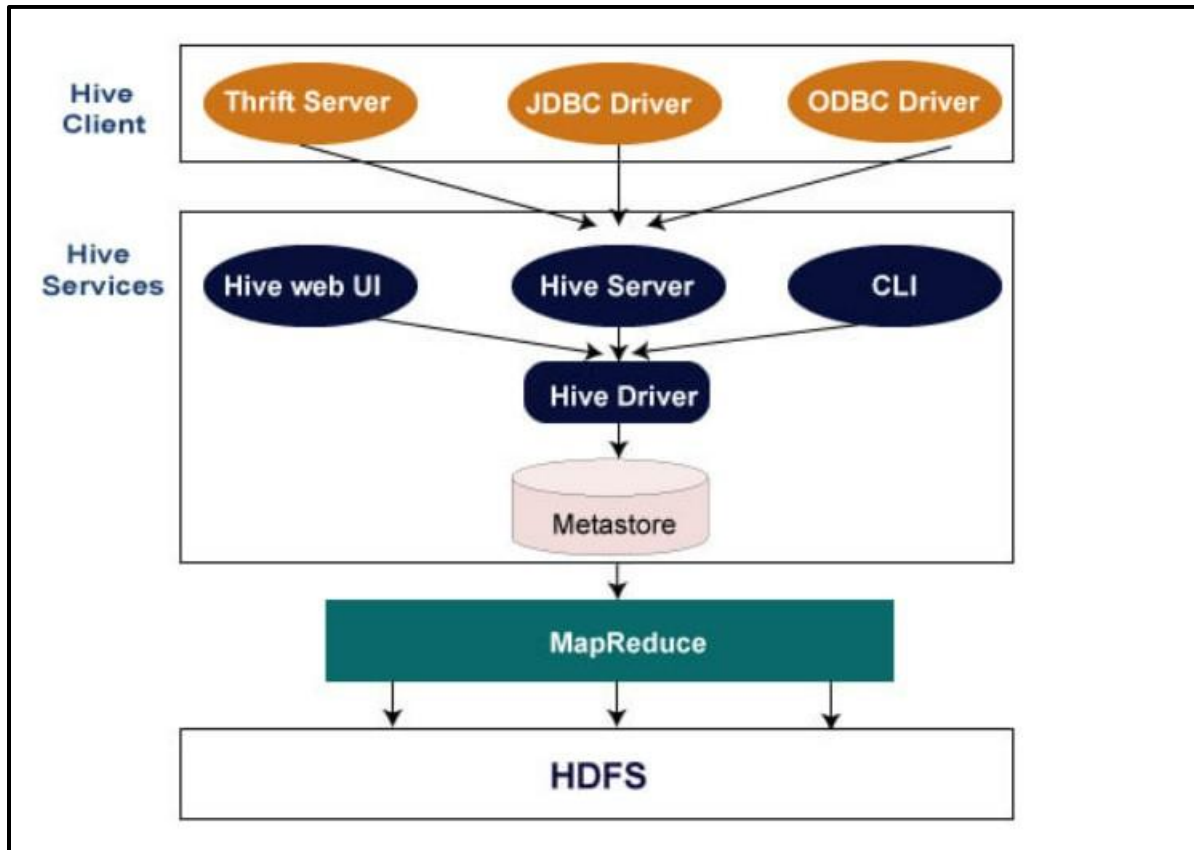
Water should contain Trihalomethanes according to WHO:	80 ppm
No.of Data points before Pre-Processing	3114.0 (Count)
Avg Value in dataset	66.396293 (Mean)

9. **Turbidity:** The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter.

Water should contain Turbidity according to WHO:	< 5 NTU
No.of Data Points before Pre-Processing	3276.0 (Count)
Avg Value in Dataset	3.966786 (Mean)

10. **Potability:** This is not a chemical, it will point whether the water is suitable for drinking or not using the binary values 0 & 1.

FEATURES AND IMPLEMENTATION-



Using HIVE to store the Data-

The screenshot shows the Hue Table Browser interface in a Mozilla Firefox browser. The URL is `quickstart.cloudera:8888/hue/metastore/table/default/water_potability`. The interface displays the table `water_potability` under the `default` database. The table has 10 columns: `ph`, `hardness`, `solids`, `chloramines`, and `sulfate`, all of type `double`. The properties section shows the table is managed by `cloudera` and was created on `04/02/2022 3:05 PM`. A task history panel on the right shows the creation of the table.

Name	Type	Comment
1 <code>ph</code>	double	Add a comment...
2 <code>hardness</code>	double	Add a comment...
3 <code>solids</code>	double	Add a comment...
4 <code>chloramines</code>	double	Add a comment...
5 <code>sulfate</code>	double	Add a comment...

The screenshot shows the Hue Editor interface in a Mozilla Firefox browser. The URL is `quickstart.cloudera:8888/hue/editor?editor=64`. The query editor shows a SQL query: `SELECT available_ph, count(trihalomethanes) as Available_Trihalomethanes, count(turbidity) as Available_Turbidity, count(potability) as Available_potability from water_potability`. The query has been executed, and the results are displayed in a table with 8 columns: `available_ph`, `available_hardness`, `available_solids`, `available_chloramines`, `available_sulfates`, `available_conductivity`, `available_organic_carbon`, and `available_`. The results show 1 row of data.

	available_ph	available_hardness	available_solids	available_chloramines	available_sulfates	available_conductivity	available_organic_carbon	available_
1	2785	3276	3276	3276	2495	3276	3276	3114

The screenshot shows the Hive Impala query interface. The query entered is `Select max(ph),min(ph),avg(ph)FROM water_potability;`. The results table has three columns: `max(ph)`, `min(ph)`, and `avg(ph)`. The first row shows the values 13.999999999999998, 0, and 7.0807945042768186 respectively.

	max(ph)	min(ph)	avg(ph)
1	13.999999999999998	0	7.0807945042768186

The screenshot shows the Hive Impala query interface. The query entered is `Select max(hardness),min(hardness),avg(hardness)FROM water_potability;`. The results table has three columns: `max(hardness)`, `min(hardness)`, and `avg(hardness)`. The first row shows the values 323.12400000000002, 47.432000000000002, and 196.36949601730177 respectively.

	max(hardness)	min(hardness)	avg(hardness)
1	323.12400000000002	47.432000000000002	196.36949601730177

The screenshot shows the Hive Impala query interface. The query entered is `Select max(solids),min(solids),avg(solids)FROM water_potability;`. The results table has three columns: `max(solids)`, `min(solids)`, and `avg(solids)`. The first row shows the values 61227.196007712133, 320.94261127435902, and 22014.092526077111 respectively.

	max(solids)	min(solids)	avg(solids)
1	61227.196007712133	320.94261127435902	22014.092526077111

This is a very important stage in a research since from here we will draw our conclusion and recommendation based on the observed descriptive. At this stage we will also improve the quality of our data so as to get the desirable dataset which are free from outliers and is consistent, unbiased and sufficient for the study. In

this stage, the water quality index has been calculated from the most vital parameters or variables contained in the dataset. The samples of water to be used for this study have been categorized on the basis of the water quality index values. We are going to use the z score technique to normalize our data(El-Kowrany et al., 2016).

Computation of water quality index

At this stage, the seven parameters outlined in the dataset are used to calculate the water quality index in order to measure the quality of water. The published data obtained from Kaggle website is used to evaluate the projected model based on the parameters. Below is the formula for obtaining the water quality index. We are going to use the z score technique to normalize our data(El-Kowrany et al., 2016).

$$\text{Water quality index} = \left(\sum_{i=1}^N qi \times wi \right) / \left(\sum_{i=1}^N wi \right)$$

$$qi = 100 \times \left(\frac{Vi - V_{Ideal}}{Si - V_{Ideal}} \right)$$

$$wi = \frac{K}{Si}$$

$$K = \frac{1}{\sum_{i=1}^N Si}$$

N = Total number of Parameters

qi = Quality rating scale for each parameter *i* in the tested sample

V_{ideal} = Is the ideal value of parameter *i* in pure water.

S_i = Is the recommended standard value of parameter *i*.

k = Proportionality Constant

DATA ANALYSIS:

The analysis was carried out on the data set with index range of 0 to 3276. The shape of dataset is (3276,10). Before moving on to the analysis, we need to check for the missing data and handle it in an appropriate way so as it will not affect our analysis. Before moving on to the missing values, the following is a snapshot of the dataset before cleaning and its statistical overview.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

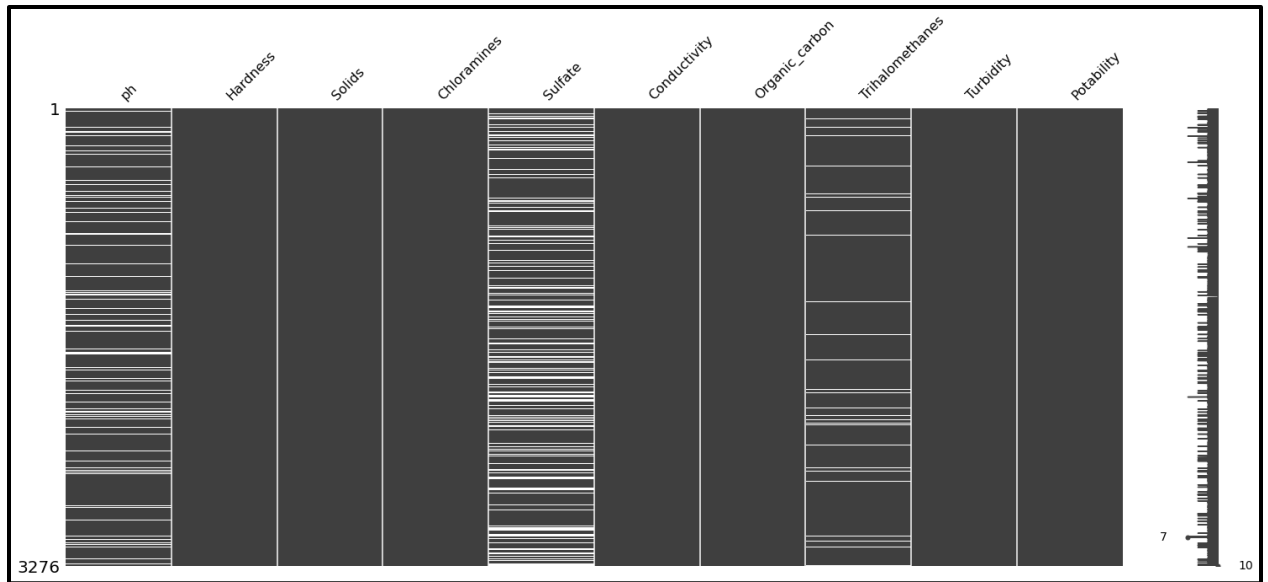
	count	mean	std	min	25%	50%	75%	max
ph	2785.0	7.080795	1.594320	0.000000	6.093092	7.036752	8.062066	14.000000
Hardness	3276.0	196.369496	32.879761	47.432000	176.850538	196.967627	216.667456	323.124000
Solids	3276.0	22014.092526	8768.570828	320.942611	15666.690297	20927.833607	27332.762127	61227.196008
Chloramines	3276.0	7.122277	1.583085	0.352000	6.127421	7.130299	8.114887	13.127000
Sulfate	2495.0	333.775777	41.416840	129.000000	307.699498	333.073546	359.950170	481.030642
Conductivity	3276.0	426.205111	80.824064	181.483754	365.734414	421.884968	481.792304	753.342620
Organic_carbon	3276.0	14.284970	3.308162	2.200000	12.065801	14.218338	16.557652	28.300000
Trihalomethanes	3114.0	66.396293	16.175008	0.738000	55.844536	66.622485	77.337473	124.000000
Turbidity	3276.0	3.966786	0.780382	1.450000	3.439711	3.955028	4.500320	6.739000
Potability	3276.0	0.390110	0.487849	0.000000	0.000000	0.000000	1.000000	1.000000

To see which column has the missing data, we can have overview of the dataset by using the `df.info()`. Following is the snapshot of the described.

```
<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
```

We can see that there are missing fields in the column of pH, Sulfate and Trihalomethanes. We explored how many fields were missing statistically and graphically.

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



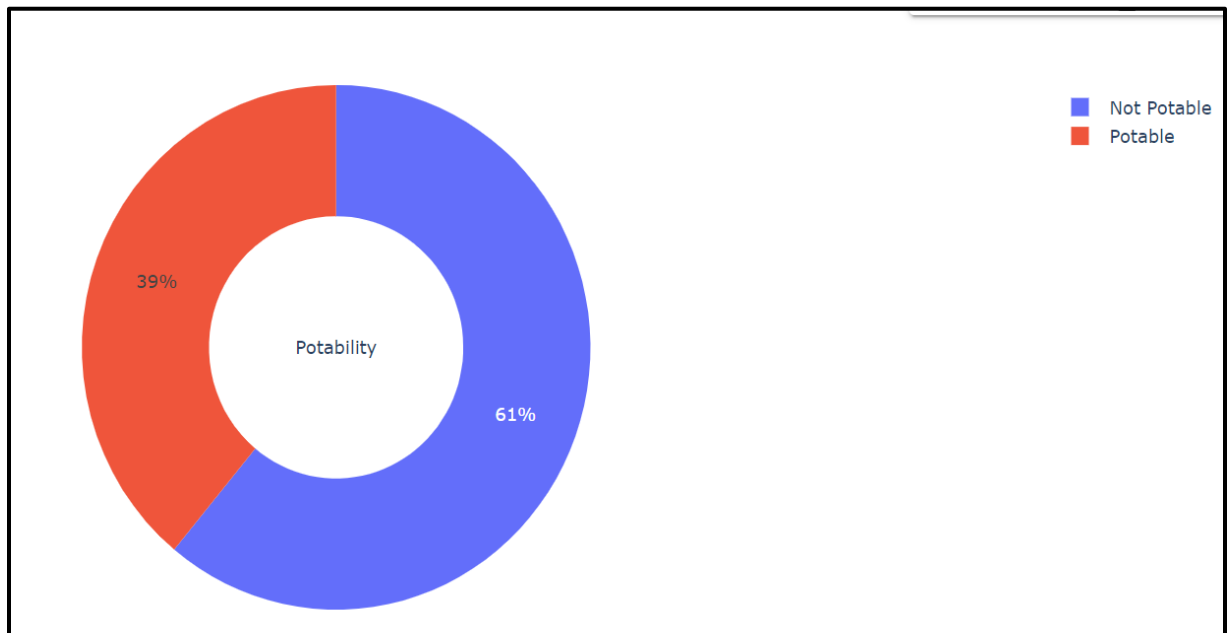
All the three columns which have the missing data are crucial in prediction. Hence, we cannot drop the missing values rather we choose to fill the missing values with respective median of the data according to their final output. After filling the missing values with their respective median, the missing fields

```

ph          0
Hardness    0
Solids      0
Chloramines 0
Sulfate     0
Conductivity 0
Organic_carbon 0
Trihalomethanes 0
Turbidity   0
Potability  0
dtype: int64

```

Based on the analysis the following are some of the graphs which were obtained.



The chart above was used to summarize the data in two categories, portable and not portable.

In our analysis we found that 61% of water was not portable and was considered unhealthy for human consumption. On the other hand 39% of the water was portable and deemed safe for human consumption.

The table below shows the range of measured water quality based on the parameters. The table was used to help us determine the quality of the sample.

	Name	Range
0	ph	0 to 14
1	Hardness	47 to 324
2	Solids	320 to 61228
3	Chloramines	0 to 14
4	Sulfate	129 to 482
5	Conductivity	181 to 754
6	Organic_carbon	2 to 29
7	Trihalomethanes	0 to 124
8	Turbidity	1 to 7

pH: From the analysis we found that the range of the pH, varies from 0 to 14 which is above the acceptable normal range of 6.5 to 8.5. This shows that the acidity or alkalinity of water was out of the optimal range.

Hardness: This is the amount of dissolved calcium and magnesium in water. The range of this parameter was between 47 to 372 which is within the acceptable range of up to 600mg/L. This shows that the index quality of this parameter was good and acceptable and could not cause any health issue to the consumers.

Dissolved solids: High number of dissolved solids in water makes the water unsafe for human consumption. This can lead to various infections such as rashes on the skin, dizziness, and chronic infections among many others. High level of dissolved solids in the water is an indication of a highly mineralized water. The standard limit of total dissolved solids in water should range between 500 to 1000ppm. In our analysis we found the range of Dissolved solids to be between 320-61228ppm which is not within the recommended range. Our analysis indicated a range of up to 14 which is above the recommended level.

Chloramines: The recommended level of chloramine in water is 4miligram per litre, research has indicated that that at 4mg/L the chances of harmful health effects are very unlikely.

Turbidity: This is a significant indicator of the quantity of suspended sediment in water, which may have a negative impact on human health. The acceptable limit of turbidity should be between 5 to 10NTU. In our study we found the level to be ranging between 1 to 7.

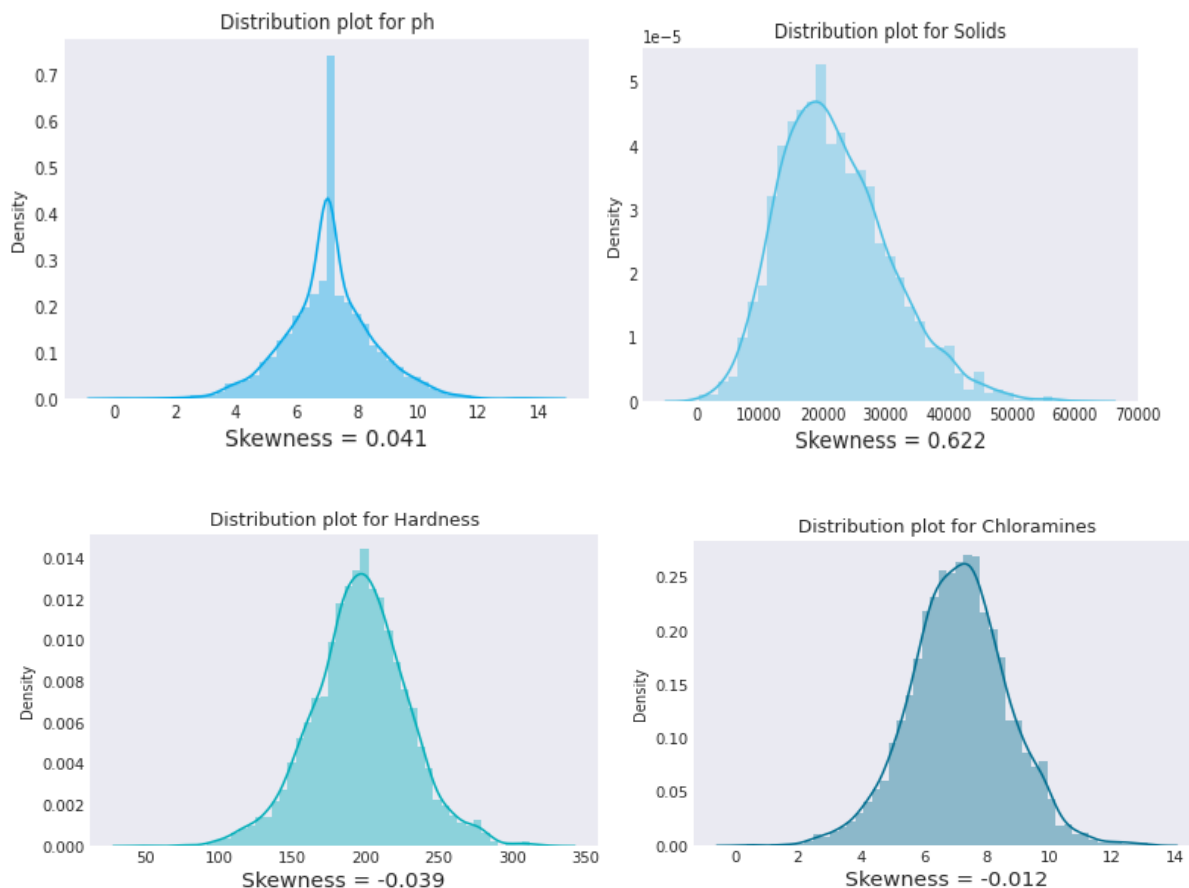
Trihalomethanes: The acceptable limit of Trihalomethanes should range up to 80ppm. In our research it was found that level of Trihalomethanes was up to 124ppm which is above the recommended limit.

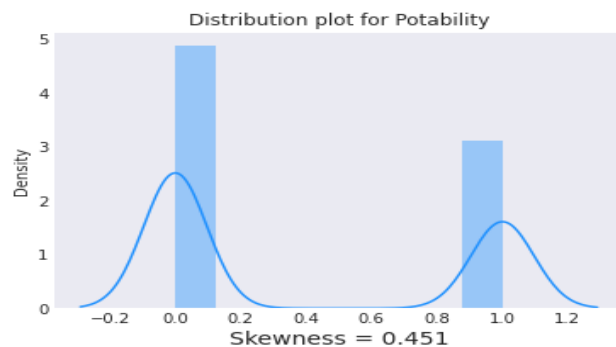
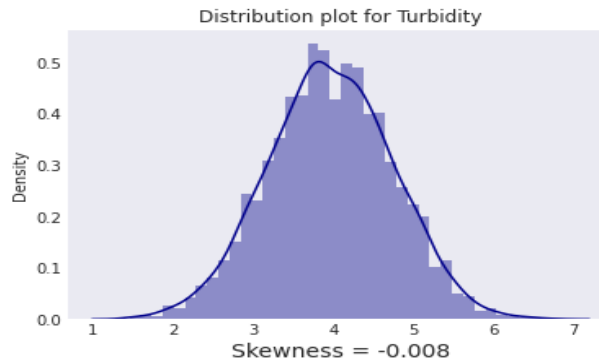
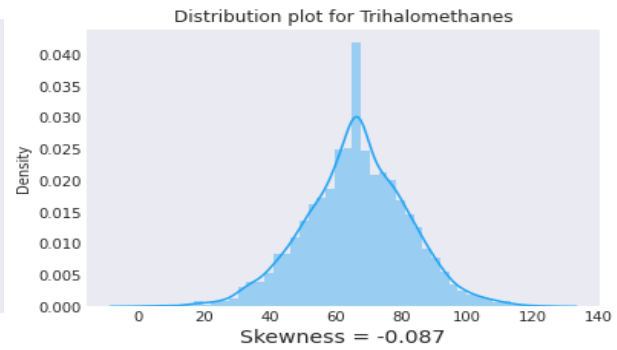
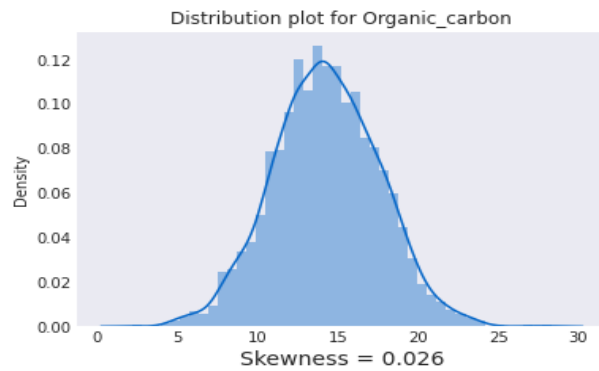
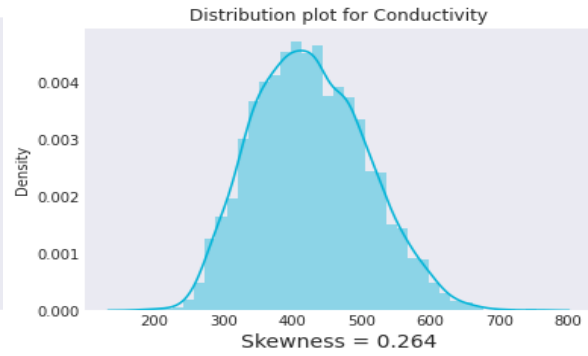
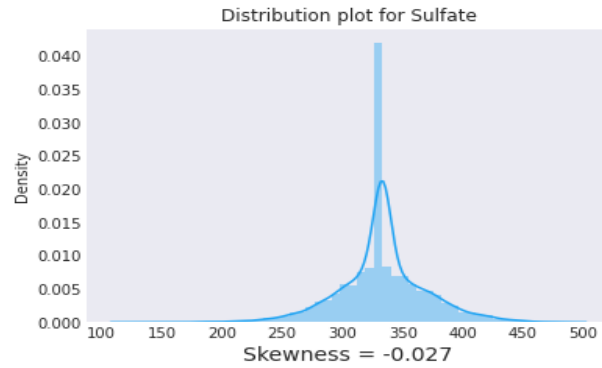
Organic Carbon: This the amount of carbon in organic compounds measured in ppm. The acceptable limit of organic carbon in water should be less than 2mg/L. Our analysis found this level to be between 2 to 29 which is above the acceptable limits.

Conductivity: the electrical conductivity of a safe water should be up to 400 $\mu\text{S}/\text{cm}$. In our study we found conductivity to be between 181 to 754 $\mu\text{S}/\text{cm}$.

Sulfate: The acceptable amount of sulfate dissolved in water is 400 mg/L. From our analyzing we found the range to be between 129 to 482 mg/L.

The following are the distribution plots for the parameters



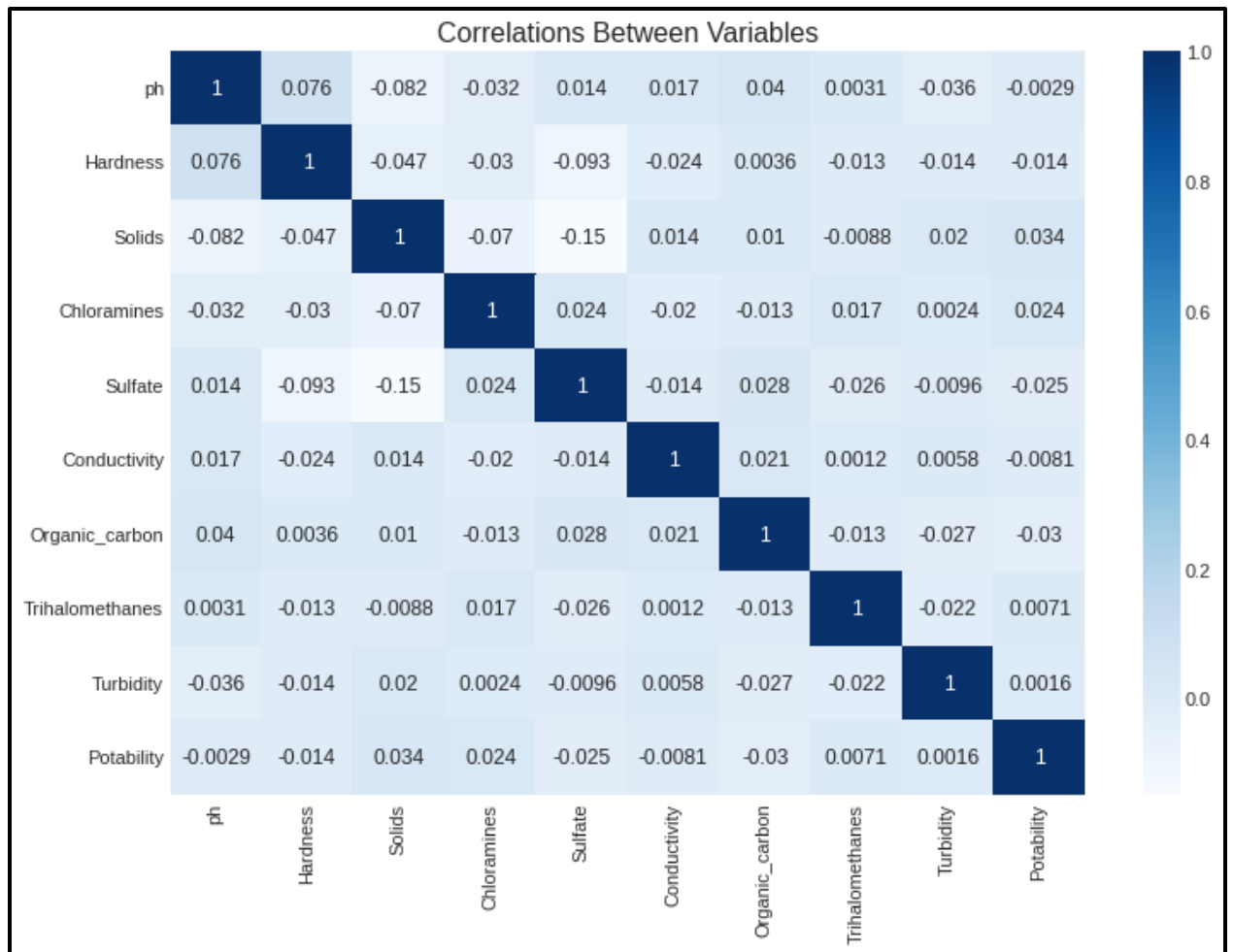


Correlation Heat Map –

Visualizing the linear correlations between variables using Heatmap Visualization.

The measure used for finding the linear correlation between each variable is

Pearson Correlation Coefficient.



From the heatmap we can conclude that there is no strong correlation between any two variables.

IMPLEMENTATION AND RESULTS:

After analysing the data, we have trained a range of Machine Learning algorithms with the processed data. We split the data in training and testing data in the ratio of 80:20. The range of Machine Learning models we used are:

- Logistic Regression
- Random Forest
- LightGBM
- Support Vector Machines
- XG Boost
- Gaussian Naive Bayes
- Bernoulli Naive Bayes
- KNN
- Decision Tree
- Bagging Classifier

1. Logistic Regression

The method of predicting the probability of a discrete result given an input variable is known as logistic regression. The outcome of a logistic regression would be binary in nature. It is a supervised machine learning technique. Logistic regression is mainly used in classification and regression problems.

Accuracy on this dataset – 62.80

2. Decision Tree

Decision tree is a graphical representation for an attribute test. It is generally used for categorization and prediction. For instance, in our dataset, if the sulphate value is greater than 30 mg/L, then we consider that particular sample as Not fit for drinking. In a decision tree, the attribute test is the root node and from there on the further tests are stored in internal nodes and coming to the leaf nodes we have our prediction. Decision trees are majorly used in classification problems.

Accuracy on this dataset – 73.47

3. Random Forest

Random forest at its root is collection of number of decision trees. Random forest is generally used for Classification and regression problems. For classifications problems, the output is picked by the class that is supported by maximum number of trees. For regression tasks, the class will be selected by what the average number of trees select.

Accuracy on this dataset – 80.79

4. XG Boost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is nothing but decision trees with boosting features so as to boost the performance of the model. It is used for both regression and classification.

Accuracy on this dataset – 80.33

5. LightGBM

LightGBM is a decision tree-based gradient boosting framework that improves model efficiency and minimize memory utilization. It is used for both regression and classification.

Accuracy on this dataset – 79.11

6. Support Vector Machines

SVM are used for both regression and classification tasks. The objective of the SVM is to produce a hyperplane between clusters of classes and diving them into different classes.

Accuracy on this dataset – 69.66

7. Gaussian Naïve Bayes

It is based on the Bayes theorem and the conditional probability. It supports continuous valued features. The model fits by simply finding the mean and standard deviation of the points within the label. Generally used in NLP applications.

Accuracy on this dataset – 62.95

8. Bernoulli Naive Bayes

Bernoulli Naïve Bayes is specifically developed for binary classification problems.

Accuracy on this dataset – 62.80

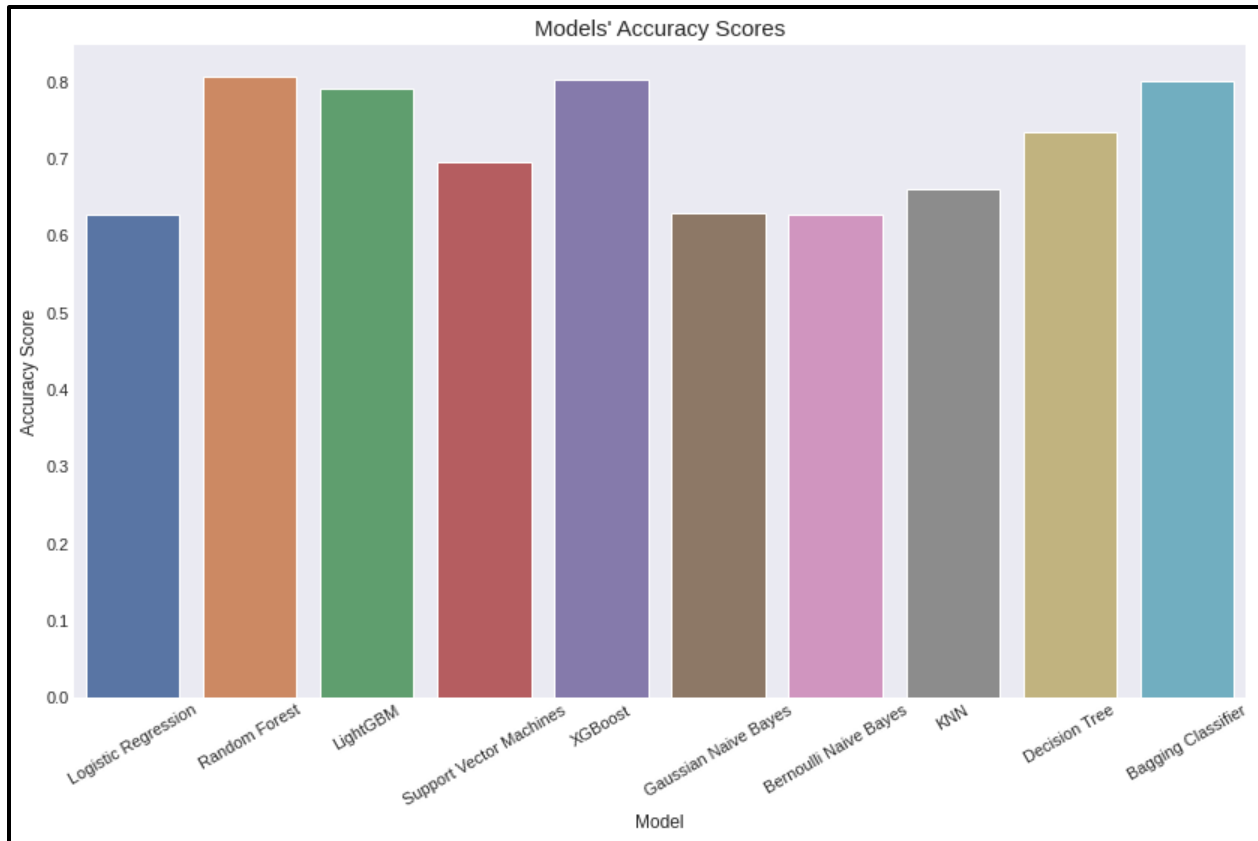
8. Bagging Classifier

Bagging classifiers are ensemble meta-estimators that fit base classifiers to random subsets of the original dataset and then combine their individual predictions to generate a final prediction.

Accuracy on this dataset – 80.18

The accuracies achieved by each ML algorithm is listed below:

	Model	Accuracy Score
1	Random Forest	0.807927
4	XGBoost	0.803354
9	Bagging Classifier	0.801829
2	LightGBM	0.791159
8	Decision Tree	0.734756
3	Support Vector Machines	0.696646
7	KNN	0.661585
5	Gaussian Naive Bayes	0.629573
0	Logistic Regression	0.628049
6	Bernoulli Naive Bayes	0.628049



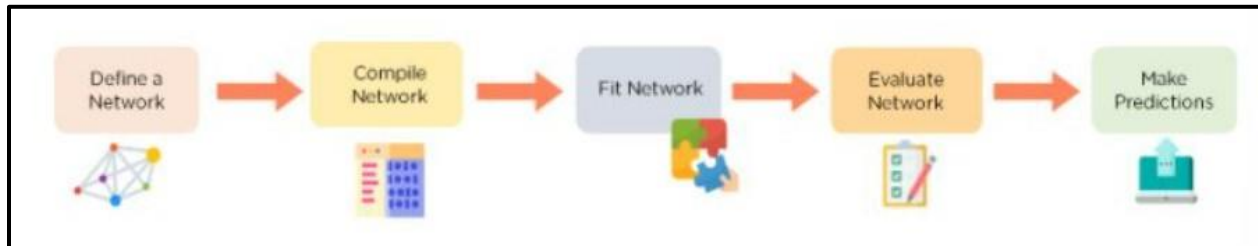
From the range of ML models, we trained with the dataset. Random Forest, XG Boost and Bagging Classifier works best with accuracy of 80%.

After using the shallow machine learning techniques, we went forward with applying deep learning techniques to the dataset. We used Keras for evaluating the deep learning models.

Keras:

- Keras is a powerful open source library for developing and evaluating deep learning models and neural networks.
- It was developed by Francois Chollet, a Google engineer using four principles Modularity, Minimalism, Extensibility and Python.
- The frameworks supported by keras are Tensorflow, Theano, MX Net, PlaidML and CNTK.

- It provides simple API's, minimizes the actions required to implement common code and also provides various deployment choices depending on user needs.



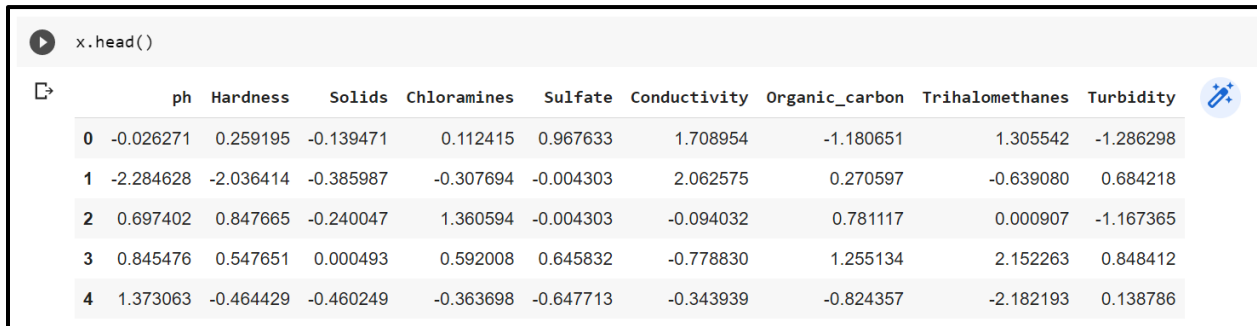
1. **Define a network:** In this first step, we describe about various layers in our model and the relationship between them. There are two types of models in Keras namely Sequential and Functional models. We can choose the type of model according to our needs and then characterize the dataflow between them,
2. **Compile a network:** Compilation means converting the code into machine understandable form. In Keras, we have a function named `model.compile()` to perform this action. In compilation, we define the metrics which finds accuracy, loss function which finds losses and optimizer which reduces the loss of the model.
3. **Fit the network:** In this step, we fit our model to the data and then start training the model on our data.
4. **Evaluate the network:** In this step, we perform evaluation of errors in our model.
5. **Make Predictions:** In this step, we make predictions on new data and this action is performed using `model.predict()` function.

Features:

- Keras provides a lot of prelabeled datasets which the user can import and load directly without much hassle.
- It has multiple backend support so that user can choose from Tensorflow, CNTK and Theano as their backend for their projects based on their needs.
- Keras supports data parallelism so that it can process large amounts of data and fastens the time for training the model.

- Keras has a high scalability of computation and it supports both recurrent and convolutional networks.
- Keras is modular and it has a wider community and developer support.
- It runs seamlessly on both CPU and GPU.

We have imported all the required libraries and standardized the range of values in our dataset. The data is scaled to the unit variance with standard scaler. After standardizing, the first five lines of the data looks like this



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
0	-0.026271	0.259195	-0.139471	0.112415	0.967633	1.708954	-1.180651	1.305542	-1.286298
1	-2.284628	-2.036414	-0.385987	-0.307694	-0.004303	2.062575	0.270597	-0.639080	0.684218
2	0.697402	0.847665	-0.240047	1.360594	-0.004303	-0.094032	0.781117	0.000907	-1.167365
3	0.845476	0.547651	0.000493	0.592008	0.645832	-0.778830	1.255134	2.152263	0.848412
4	1.373063	-0.464429	-0.460249	-0.363698	-0.647713	-0.343939	-0.824357	-2.182193	0.138786

After standardizing the data, we have split the dataset into training and testing in the ratio of 80:20 respectively. Next, we created a deep learning model with Relu activation function. We used Adam optimizer and binary cross entropy. Following is the summary of the deep learning model.

```
[ ] model.summary()
```

Model: "sequential_2"

Layer (type)	Output Shape	Param #
dense_8 (Dense)	(None, 64)	640
batch_normalization_2 (Batch Normalization)	(None, 64)	256
dense_9 (Dense)	(None, 32)	2080
dense_10 (Dense)	(None, 16)	528
dropout_2 (Dropout)	(None, 16)	0
dense_11 (Dense)	(None, 1)	17

```
=====  
Total params: 3,521  
Trainable params: 3,393  
Non-trainable params: 128  
=====
```

Fitting the model and setting the batch size as 32 and running it for 150 epochs

```
history=model.fit(X_train,y_train, batch_size=32, epochs=150, validation_data =(X_test,y_test))
```

```
Epoch 1/150  
82/82 [=====] - 1s 6ms/step - loss: 0.6836 - accuracy: 0.5920 - val_loss: 0.6783 - val_accuracy: 0.6280  
Epoch 2/150  
82/82 [=====] - 0s 4ms/step - loss: 0.6505 - accuracy: 0.6179 - val_loss: 0.6606 - val_accuracy: 0.6479  
Epoch 3/150  
82/82 [=====] - 0s 4ms/step - loss: 0.6235 - accuracy: 0.6626 - val_loss: 0.6385 - val_accuracy: 0.6845  
Epoch 4/150  
82/82 [=====] - 0s 4ms/step - loss: 0.6119 - accuracy: 0.6695 - val_loss: 0.6336 - val_accuracy: 0.6738  
Epoch 5/150  
82/82 [=====] - 0s 4ms/step - loss: 0.6013 - accuracy: 0.6798 - val_loss: 0.6106 - val_accuracy: 0.6890  
Epoch 6/150  
82/82 [=====] - 0s 4ms/step - loss: 0.5966 - accuracy: 0.6836 - val_loss: 0.6064 - val_accuracy: 0.6860  
Epoch 7/150  
82/82 [=====] - 0s 4ms/step - loss: 0.5861 - accuracy: 0.6962 - val_loss: 0.6060 - val_accuracy: 0.6860  
Epoch 8/150  
82/82 [=====] - 0s 4ms/step - loss: 0.5925 - accuracy: 0.6885 - val_loss: 0.6061 - val_accuracy: 0.6753  
Epoch 9/150  
82/82 [=====] - 0s 4ms/step - loss: 0.5885 - accuracy: 0.6939 - val_loss: 0.6151 - val_accuracy: 0.6524  
Epoch 10/150
```

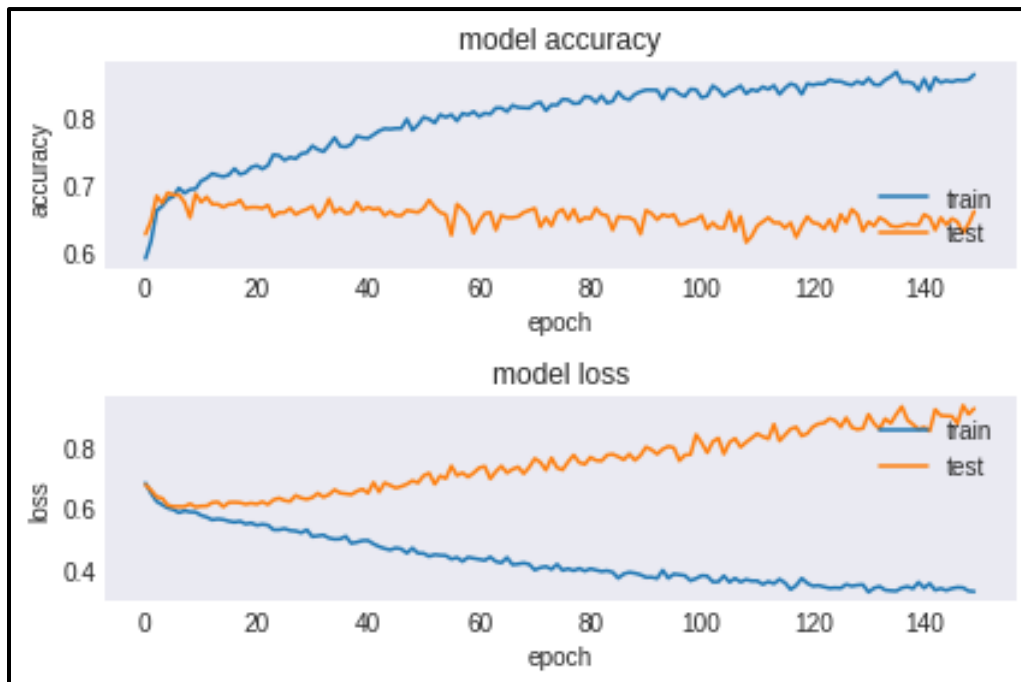
Evaluating the model

```
#Evaluating the model

eval_model=model.evaluate(X_train, y_train)
eval_model

82/82 [=====] - 0s 2ms/step - loss: 0.2386 - accuracy: 0.8992
[0.23857414722442627, 0.8992366194725037]
```

The deep learning model manages to get an overall accuracy of 89.92 % and with a loss of 0.23. Following are the evaluation plots of accuracy and loss with respect to increasing epochs.



Conclusion

Hive Integration Part:

We have used HIVE for the storage purpose and stored the raw dataset and performed basic analysis. As we have uploaded a raw dataset the performance of the analysis was not up to the mark in HIVE. Hence, we moved forward with data cleaning and Pre-when processing using pandas in python later employed Machine Learning techniques and when compared to the HIVE we have seen better performance level.

Hence instead of Integrating the HIVE with SPARK we have continued without integrating based on the performance level.

Machine Learning and Deep Learning:

We have used 10 Machine Learning Algorithms and achieved the maximum of 80% accuracy, later we used Keras-Deep Learning framework and achieved the accuracy of 90% by tweaking the Hyper-parameters and Epochs. From our analysis we have seen that Deep Learning has given better accuracy compared to the Machine Learning Algorithms for our project.

Project Management:

Implementation Status Report –

Work Completed

- Loading data into HIVE and running queries against the tables stored in HIVE: Done by Eshwar
- Detail design of features: Done by Geetha
- Data set, analysis and data visualization using Python libraries: Done by Harsha and Keerthi
- Model training and prediction: Done by Harsha and Keerthi

References-

- [1]<https://pubmed.ncbi.nlm.nih.gov/22925610/>
- [2]https://kipdf.com/public-health-quality-of-drinking-water-supply-in-orangi-town-karachi-pakistan_5ac959f01723dde349028715.html
- [3]https://www.researchgate.net/publication/329316037_Surface_Water_Pollution_Detection_using_Internet_of_Things
- [4]https://www.academia.edu/26326194/Data_analysis_quality_indexing_and_prediction_of_water_quality_for_the_management_of_rawal_watershed_in_Pakistan
- [5]<https://ieeexplore.ieee.org/document/8300197>
- [6]<https://ieeexplore.ieee.org/document/8328641>
- [7]https://www.researchgate.net/publication/259880906_Evaluation_of_Multivariate_Linear_Regression_and_Artificial_Neural_Networks_in_prediction_of_Water_Quality_parameters
- [8]https://www.infona.pl/resource/bwmeta1.element.springer-doi-10_1007-S40808-015-0063-9
- [9]https://www.researchgate.net/publication/223471289_Neural_network_modeling_of_dissolved_oxygen_in_the_Gruza_reservoir_Serbia
- [10]https://www.researchgate.net/publication/289379524_Water_pollution_in_the_Middle_Nile_Delta_Egypt_An_environmental_study
- [11]https://www.kaggle.com/adityakadiwal/water-potability?select=water_potability.csv
- [12] https://www.researchgate.net/publication/334710343_An
- [13]<https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras>
- [14]<https://www.guru99.com/keras-tutorial.html>
- [15]https://www.tutorialspoint.com/keras/keras_introduction.htm
- [16]<https://www.javatpoint.com/hive-architecture>
- [17] <https://data-flair.training/blogs/hive-data-model/>
- [18] [Tap Water and Trihalomethanes: Flow of Concerns Continues - PMC \(nih.gov\)](#)
- [19] [Indicators: Conductivity | US EPA](#)
- [20] [Basic Information about Chloramines and Drinking Water Disinfection | US EPA](#)

