Project Report
On

Prediction of water potability using ML models

Submitted as partial of fulfillment of course work required for the M. Tech program in Mehta Family School of Data Science

By

Satyam goje

M. Tech Data Science

Enrollment No-2156601

Under the guidance of

Dr. Kasiviswanathan K S

(Asst. Professor,WRDM)

Mehta Family School of Data Science Indian Institute of Technology Rookie, Roorkee-247667 (India)

May, 2022

Acknowledgement

I wish to express my most sincere appreciation and gratitude to **Prof. Kasivishwanathan K S**, Mehta family school of Data Science and Artificial Intelligence. Indian Institute of Technology - Roorkee, Roorkee-247667. For his valuable guidance and continuous encouragement during preparation of the project work.

I also extend my sincere thanks to HOD, **Prof. Dr. Durga Toshniwal**., for her support and also to the entire faculty and my friends who had helped directly or indirectly.

Above all, I thank the Almighty who gave me a firm platform for the successful completion of the report.

Date:8 May, 2022

Satyam Goje

Place: IIT Roorkee

Table of Contents

List of Figure

Abstract

"Potable water" simply refers to water that is safe to drink, yet it is becoming increasingly scarce across the world. Growing demand is putting a strain on freshwater supplies across the world, and an infinite array of impurities may convert once-potable water into a health hazard. In this project, we are trying to predict whether given water sample is potable or not using parameters Ph, Hardness, Solids, chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity using Machine learning models Ada boosting, Decision tree, XG boosting, Random Forest. Also we performed Exploratory data analysis to understand the dataset properly followed by hyperparameter tuning to improve the performance of machine learning models.

Dataset:

Parameters:

We have 10 columns in the dataset.

- Ph: PH is an important parameter in evaluating the acid-base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5.

- Hardness: Calcium and magnesium salts are the primary causes of hardness. These salts dissolve from geologic deposits that water passes through. The amount of hardness in raw water is determined by the length of time water encounters hardness generating material. Hardness was initially described as water's ability to precipitate soap due to calcium and magnesium.

- Solids: Water may dissolve a broad variety of inorganic and some organic minerals or salts, including potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulphates, and others. These minerals gave water an unpleasant flavor and a diluted tint. This is a critical metric for water use. A high TDS measurement implies that the water is heavily mineralized. The desirable limit for TDS is 500 mg/l, while the maximum level for drinking is 1000 mg/l.

- Chloramines: The most common disinfectants used in municipal water systems are chlorine and chloramine. When ammonia is added to chlorine to purify drinking water, chloramines are most typically generated. Chlorine levels in drinking water up to 4 milligrams per litre (mg/L or 4 parts per million (ppm)) are deemed safe.
- Sulfate: Sulfates are naturally occurring compounds found in minerals, soil, and rocks. They can be found in the air, groundwater, plants, and food. Sulfate's main commercial application is in the chemical industry. The quantity of sulphate in saltwater is around 2,700 milligrams per litre (mg/L). In most freshwater sources, it varies from 3 to 30 mg/L, while significantly greater amounts (1000 mg/L) are found in specific geographical areas.

- Conductivity: Pure water is an excellent insulator rather than a strong conductor of electric current. An increase in ion concentration improves water's electrical conductivity. Electrical conductivity is often determined by the quantity of dissolved particles in water. Electrical conductivity (EC) is a measurement of a solution's ionic mechanism that allows it to transfer electricity. According to WHO guidelines, the EC value should not exceed 400 S/cm.

- Organic carbon: Total Organic Carbon (TOC) in source waters is derived from both natural organic matter (NOM) breakdown and manmade sources. The total quantity of carbon in organic compounds in pure water is measured as TOC. The US EPA recommends 2 mg/L TOC in treated / drinking water and 4 mg/Lit in source water used for treatment.

- Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

- 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. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

- Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable. This is our target variable.

Libraries used:

Python programming language is used to analyze historical data. Some of the important Libraries used are:

-**Pandas:** Pandas is data manipulation and analysis software package created for the Python computer language. It provides data structures and functions for manipulating numerical tables and time series in particular. It includes tools for data analysis, cleansing, exploration, and manipulation. Wes McKinney invented the moniker "Pandas" in 2008 as a reference to both "Panel Data" and "Python Data Analysis." Pandas enable us to examine large amounts of data and draw conclusions based on statistical theory. Pandas can clean up messed-up data sets and make them more understandable and meaningful. Relevant data is critical in data science.

-**NumPy:** NumPy is a Python package that allows you to interact with arrays. It also includes functions for working with linear algebra, the fourier transform, and matrices. Travis Oliphant developed NumPy in 2005. It is a free and open-source project. NumPy is an acronym that stands for Numerical Python. Lists in Python serve the same purpose as arrays, although they are slower to process. NumPy's goal is to produce array objects that are up to 50 times quicker than ordinary Python lists. In NumPy, the array object is named ndarray, and it comes with a slew of helper methods that make dealing with ndarray . Arrays are often employed in data science, where speed and resource availability are critical.

-**Matplotlib:** Matplotlib is a fantastic Python visualization package for 2D array charts. Matplotlib is a multi-platform data visualization package based on NumPy arrays and designed to operate with the SciPy stack. It was first presented in 2002 by John Hunter. One of the most significant advantages of visualization is that it provides us with visual access to massive volumes of data in simply consumable representations. Matplotlib has a variety of plots such as line, bar, scatter, histogram, and so on.

- **Scikit-learn :** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**.

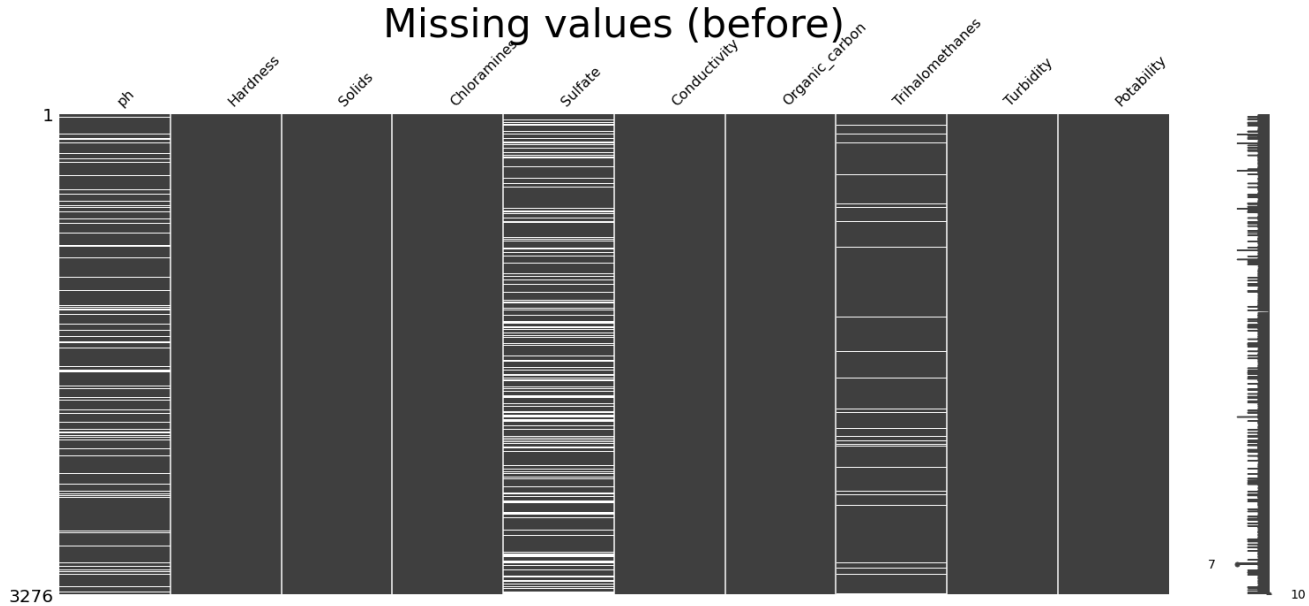
-**Missingno:** Missingno is a fantastic and easy-to-use Python tool that provides a set of visualizations to help you understand the existence and distribution of missing data inside a pandas dataframe. This can be represented as a bar plot, matrix plot, heatmap, or dendrogram.

-**Seaborn:** Seaborn is a great Python visualization tool for statistical graphics charting. It has nice default styles and colour palettes to make statistics charts more appealing. It is designed on top of the matplotlib software and is tightly connected with pandas data structures. Seaborn's goal is to make visualization the primary means of exploring and comprehending data. It provides dataset-oriented APIs, allowing us to switch between several visual representations for the same variables in order to have a better knowledge of the dataset.

Preprocessing dataset:

There are 3276 records in the dataset having 10 features.

We will use missingno library for identifying null values in the dataset.



There are 491, 781, 162 null values in ph, sulphate and trihalomethanes respectively

```
df.isna().sum()
```

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

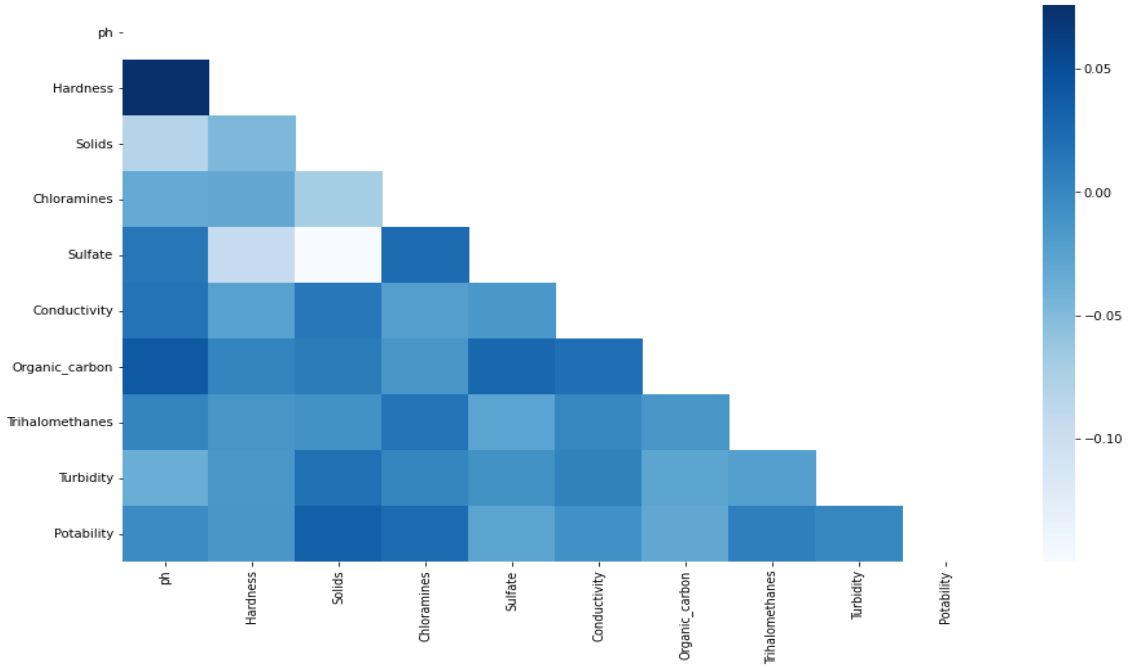
We will fill the null values with the mean values of respective columns.

Exploratory data analysis:

Correlation analysis:

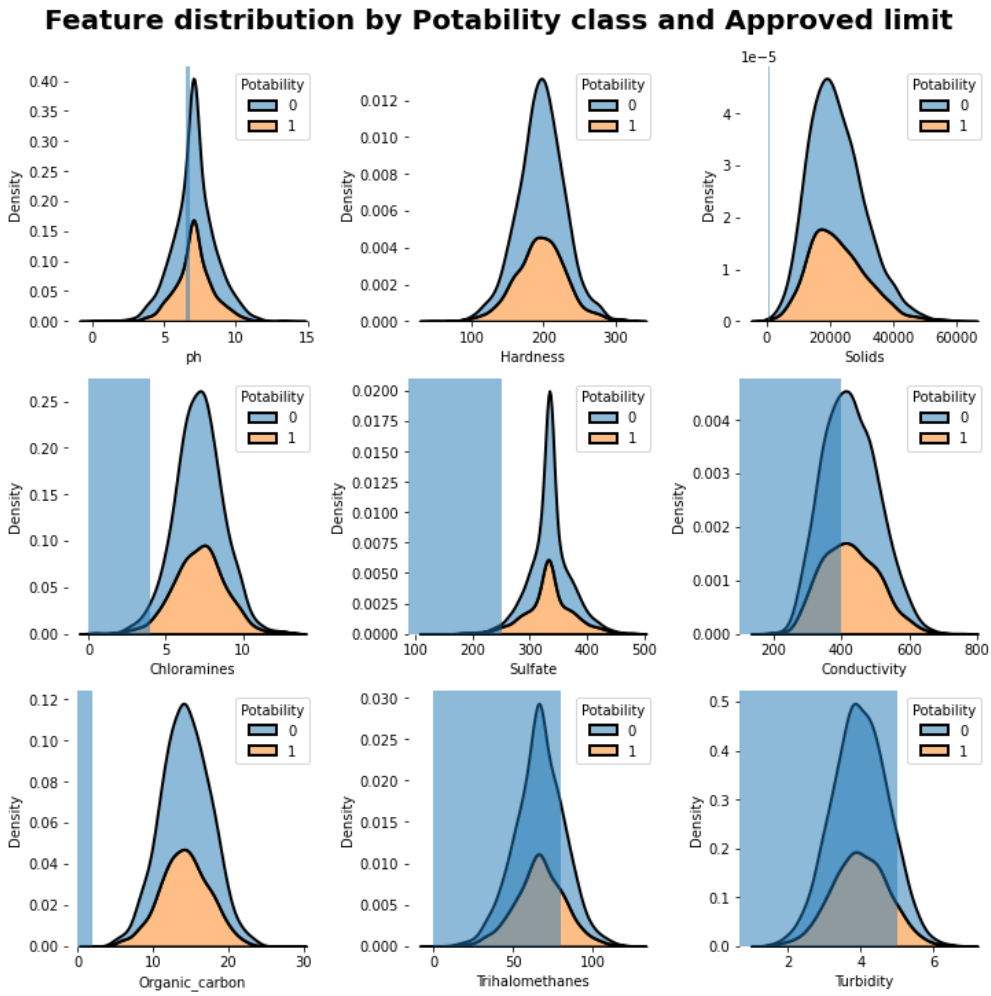
Correlation analysis in research is **a statistical method used to measure the strength of the linear relationship between two variables and compute their association**. Simply put -

correlation analysis calculates the level of change in one variable due to the change in the other.



Histogram Analysis:

-A dataframe is created by defining the safe limits of parameters, which will be used to for plotting the histograms



-Ph is distributed widely as compared to the safe range standards.

-There seems to be anomaly in the dataset for the Solids values as all the values in dataset are outside the safe limits.

-most of the values for Chloramines, sulphate and organic carbon are higher than safe limits.

- most of the values for trihalomethane and turbidity are within the safe limits.

****Blue shaded area specifies the safe limits.**

Model Development:

Six models have been developed for potability prediction

1. Logistic regression: The method of modelling the likelihood of a discrete result given an input variable is known as logistic regression. The most prevalent logistic regression models produce a binary result. Multinomial logistic regression can be used to describe events with more than two discrete outcomes. Logistic regression is a helpful analytical tool for classification issues, such as determining if a new sample belongs in a specific group.

2. K-nearest neighbor: K-Nearest Neighbors is a simple yet important classification technique in Machine Learning. It is a supervised learning algorithm that is widely used in pattern recognition, data mining, and intrusion detection. It is extensively applicable in real-world circumstances since it is non-parametric, which means it makes no underlying assumptions regarding data distribution (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).
We are provided some previous data (also known as training data) that categories locations into groups based on a characteristic.

3. Decision Tree: The Decision Tree algorithm is a member of the supervised learning algorithm family. The decision tree approach, unlike other supervised learning algorithms, may also be utilized to solve regression and classification issues. The purpose of employing Decision Trees is to build a training model that can predict the class or value of a target variable by learning basic decision rules from past data (training data). In Decision Trees, we begin at the root of the tree to forecast a class label for a record. We compare the values of the root attribute with the attribute of the record. Based on the comparison, we follow the branch corresponding to that value and proceed to the next node.

4.Random Forest: Random Forest is a Supervised Machine Learning Algorithm commonly used in classification and regression problems. It constructs decision trees from several samples and uses their majority vote for classification and average for regression. One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. It is more effective at classification tasks.

5.AdaBoost: AdaBoost is an ensemble learning approach (sometimes known as "meta-learning") that was developed to improve the performance of binary classifiers. AdaBoost employs an iterative strategy to improve poor classifiers by learning from their mistakes.

6.XGboost: This technique generates decision trees in a sequential fashion. Weights are very significant in XGBoost. All of the independent variables are given weights, which are subsequently put into the decision tree, which predicts results. The weight of factors that the tree predicted incorrectly is raised, and these variables are subsequently put into the second decision tree. These various classifiers/predictors are then combined to form a more powerful and precise model. It can solve issues including regression, classification, ranking, and user-defined prediction.

The results obtained for the algorithms are:

	Accuracy
Logistic	57.012195
Knn	65.243902
Decision Tree	71.341463
Random Forest	75.609756
Ada boosting	70.731707
XG Boosting	71.951220

Now we will do hyperparameter tuning for the top 4 algorithms to improve their performance.

Hyperparameter Tuning:

We used GridSearchCV for hyperparameter tuning. GridSearchCV from scikit-learn is an excellent tool for adjusting hyperparameters. It goes through all of the parameters that are provided into the parameter grid and finds the optimal combination of parameters based on a scoring criterion of your choosing (accuracy, f1, etc). The "best" parameters identified by GridSearchCV are theoretically the best that could be created, but only by the parameters in the given parameter grid.

After applying GridSearchCV we got the following parameters:

- Best parameters for Random Forest: {'min_samples_leaf': 2, 'n_estimators': 200}.
- Best parameters for Decision Tree: {'criterion': 'entropy', 'max_depth': 9, 'min_samples_leaf': 40}
- Best parameters for XGBoost: {'n_estimators': 250, 'learning_rate': 0.2}
- Best parameters for AdaBoost: {'learning_rate': 0.5, 'n_estimators': 100}

Results:

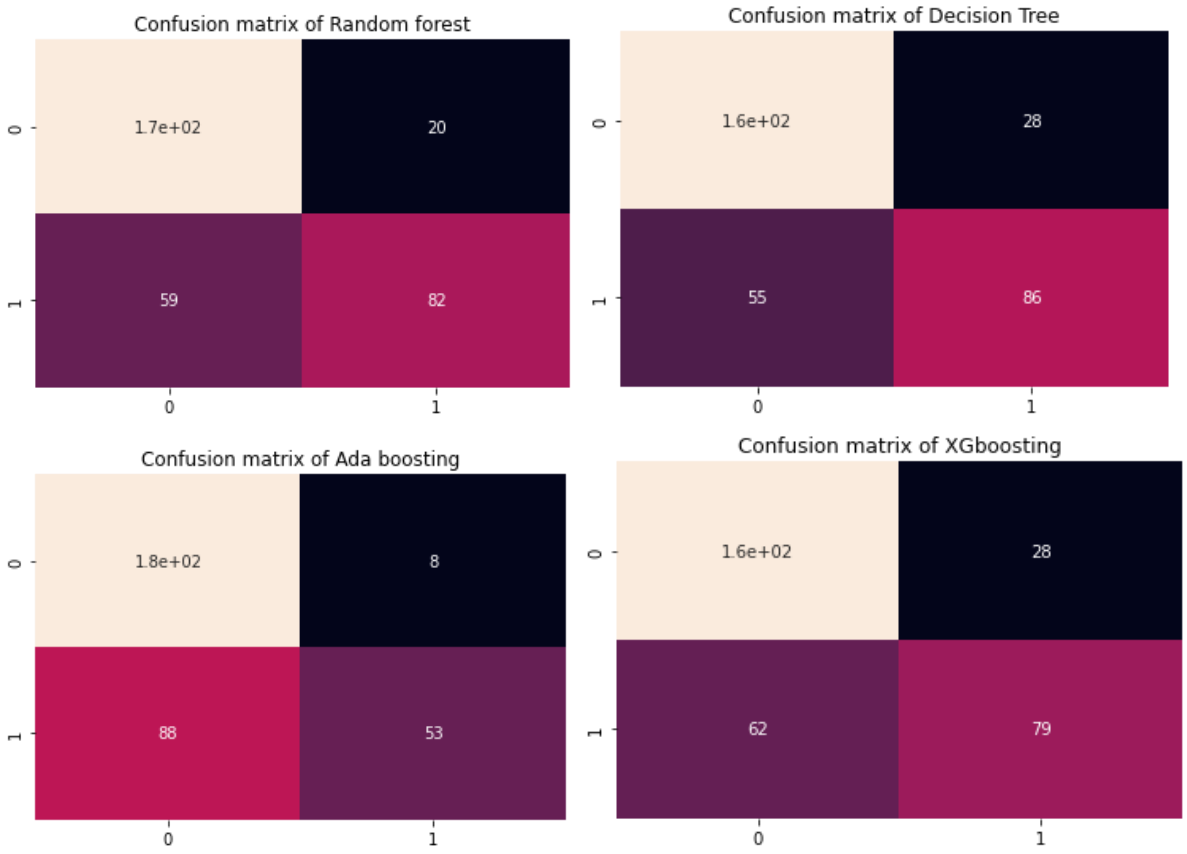
We are using confusion matrix to convey our results, confusion matrix is an immensely popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problems.

		Predicted	
Actual		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

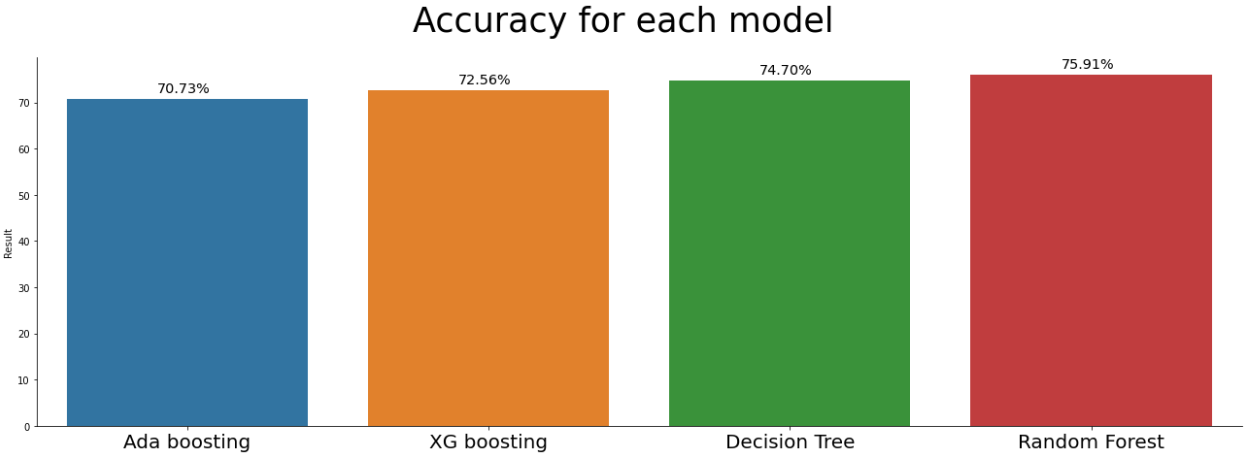
Confusion matrices represent counts from predicted and actual values. The output "TN" stands for True Negative which shows the number of negative examples classified accurately. Similarly, "TP" stands for True Positive which indicates the number of positive examples classified accurately. The term "FP" shows False Positive value, i.e., the number of actual negative examples classified as positive; and "FN" means a False Negative value which is the number of actual positive examples classified as negative. One of the most commonly used metrics while performing classification is accuracy. The accuracy of a model (through a confusion matrix) is calculated using the given formula below.

$$\text{Accuracy} = \frac{\text{TN}+\text{TP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}}$$

After reapplying the hyperparameters we obtained and then re-training our model, we obtained the following results:



After hyperparameter tuning, there is a significant increase in the accuracy of the decision tree, which is a close second. Results for all the models are shown below:



So, from the above results, we can say that **Random Forest** algorithm works better for given dataset.

**confusion matrix and accuracy value may change if we run the code again but after running the code multiple time , conclusion is that random forest is best most of the time.

भारतीय प्रौद्योगिकी संस्थान रुड़की
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE