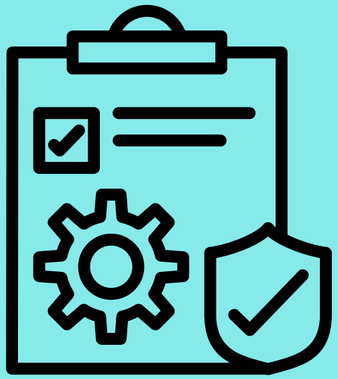


Water Quality Prediction Using Machine Learning

IT310[M] Course Project

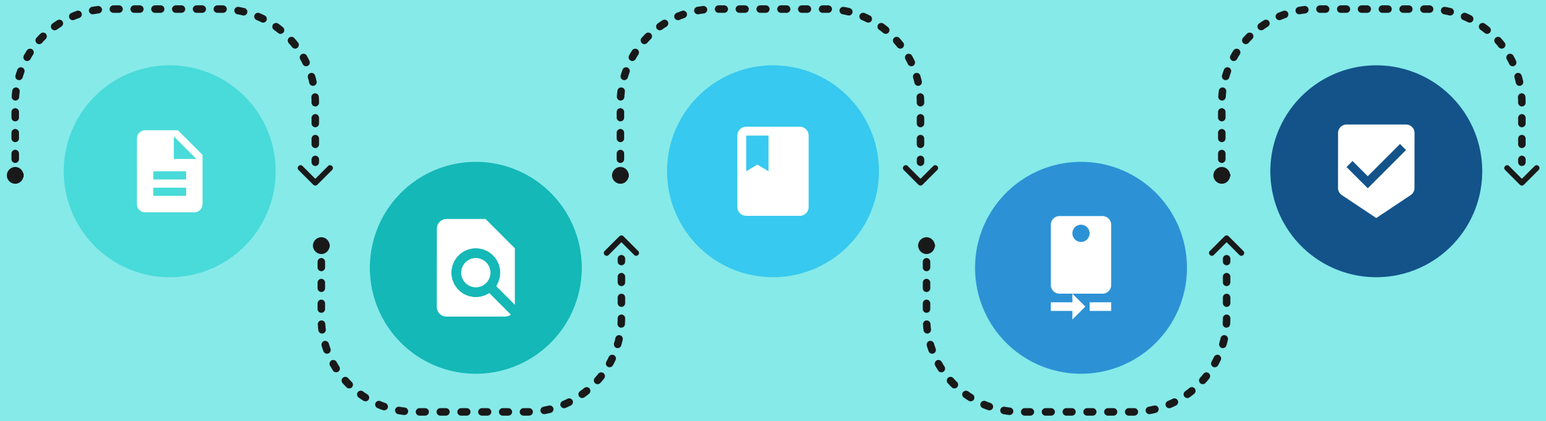
1. Kavana MS 211CV123
2. Prateek Rajput 211CV138
3. Sanket Babar 211IT015

DATASET

Features = ['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity', 'Organic_carbon', 'Trihalomethanes', 'Turbidity']

3275 observations

PROCESS FOLLOWED



1 - DATASET SEARCH

FINALISING DATASET AND FEATURES

2 - DATA PREPROCESSING

PREPROCESSING DECIDED DATASET

3 - TRAINING MODELS

TRAINING ON DATASET

4 - TESTING

TESTING ON DATASET

5 - COMPARISON

COMPARING ACCURACIES

MODELS USED

LOGISTIC REGRESSION

SVM

KNN

DECISION TREE

LOGISTIC REGRESSION

Supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class.
Logistic function - Sigmoid.
There is no change in accuracies after feature selection too.

Train Set Accuracy:60.53
Test Set Accuracy:62.80

K-NEAREST NEIGHBORS

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors.

Train Set Accuracy:71.95
Test Set Accuracy:63.25

SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a powerful machine learning algorithm
used for linear or nonlinear classification,
regression, and even outlier detection tasks.
Used RBF as kernel.

Train Set Accuracy:73.42
Test Set Accuracy:69.60

DECISION TREE

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks.
It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.

Train Set Accuracy:74.23
Test Set Accuracy:63.12

XGBOOST

XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm that builds an ensemble of weak learners, typically decision trees. The main idea behind XGBoost is to sequentially add trees to the ensemble, each correcting the errors of the previous ones. The final prediction is the sum of predictions from all the trees.

Test Set Accuracies:

0.8472

0.8139

0.8361

0.8444

0.8306

RESULTS COMPARISON