The Algorithms used and the output results

**Decision Tree:** Decision Tree is a flow chart type of supervised machine learning algorithm.It tries to predict the data in the sequence of features in the dataset based on some criteria like :gini index/information gain/Entropy.

Decision tree is self explainable. It has a problem of overfitting and to overcome this we use the Random Forest.

**Logistic Regression:** Logistic Regression is a supervised algorithm based on the logit function. It is a statictical algorithm. The goal of Logistic regression is to predict if the given outcome belongs to a given class or not. The output of the logistic regression looks similar to the Sigmoid function.

**Support Vector Machine:** SVM is a powerful supervised machine learning algorithm. The goal of SVM is to predict the best possible hyperplane which separates the classes and acts as the predictor. For non linear data, SVM tries to map the data to a higher dimensional dataspace using the ‘KERNEL TRICK’

**Random Forest:** Random Forest is a powerful supervised machine learning algorithm. It is an ensemble technique. It solves the problem of the overfitting of Decision tree by fitting multiple trees to subsets of the data, and combine the results by Majority Voting.

**AdaBoost:** AdaBoost is an ensemble method that trains and deploys trees in series. AdaBoost implements boosting, wherein a set of weak classifiers is connected in series such that each weak classifier tries to improve the classification of samples that were misclassified by the previous weak classifier.

**HistGradientBoosting:** Histogram based gradient boosting is an ensemble model which is designed to accelerate the training of gradient boosted trees. It achieves this by leveraging histograms and integer based data structures, rather than relying on sorted continuous valuse, which is the approach taken by traditional gradient boosting methods.

Here a dimensionality reduction is applied which is Principal Component Analysis.

MinMax Scaling : Min Max scaling is used to adjust the values between a particular range. The default values are 0 and 1 for the min max scaler. Here we used Min Max scaler instead of standard scaler as mist of the features are already following Gaussian Distribution.

Principal Component Analysis: PCA is a dimensionality reduction technique which is used to map the n dimesional feature space to m dimensional feature space where n>m. It converts the dependent features into independent features while computation.

|  |  |  |
| --- | --- | --- |
| Classifier Name | Test Accuracy Before PCA | Test Accuracy After PCA |
| Decision Tree | 98.5981308411215 | 100.0 |
| Logistic Regression | 85.04672897196261 | 84.57943925233646 |
| Support Vector Machine | 85.04672897196261 | 85.51401869158879 |
| Random Forest | 100.0 | 100.0 |
| Adaboost | 97.19626168224299 | 100.0 |
| HistGradient Boosting | 100.0 | 98.5981308411215 |

Sources :

<https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html>

<https://medium.com/chat-gpt-now-writes-all-my-articles/a-faster-ensemble-model-method-in-sklearn-histogram-based-gradient-boosting-7033ff170bc0>