Methodology: This project aims at implementing different machine learning classification algorithms on a selected dataset and analyzing the results in terms of comparison among the performance of those algorithms. After selecting a dataset, four classifications algorithm namely Decision Tree Induction, Random Forest Classifier, Naïve Bayes Classifier, and Support Vector Classifier were implemented to predict the class level. After implantation, the report containing accuracy has been generated for all the algorithm. In addition, we will try to visualized confusion matrix, also some graphical implementation like ROC, FPR, TPR, Precision, Recall. All the process will be using both *Weka* tool and different *Python* libraries in *Jupyter Notebook*.

Decision Tree Induction: Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or *terminal node*) holds a class label. The topmost node in a tree is the root node. Geometrically, the split describes a partition orthogonal to one of the coordinates of the decision space. The technical details of a decision tree are in how the questions about the data are formed. In the CART algorithm, a decision tree is built by determining the questions (called splits of nodes) that, when answered, lead to the greatest reduction in Gini Impurity. The Gini Impurity of a node is the probability that a randomly chosen sample in a node would be incorrectly labeled if it was labeled by the distribution of samples in the node.

Random Forest Classifier: A random forest is a data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. This type of algorithm helps to enhance the ways that technologies analyze complex data. In general, decision trees are popular for machine learning tasks. In a random forest, engineers construct sets of random decision trees to more carefully isolate knowledge from data mining, with different applied variable arrays. One way to describe the philosophy behind the random forest is that since the random trees have some overlap, engineers can build systems to study data redundantly with the various trees and look for trends and patterns that support a given data outcome.

Support Vector Classifier: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. The algorithm basically implies that only support vector are important whereas other training examples are ‘ignorable’.

Naïve Bayes Classifier: Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. The dataset is divided into two parts, namely, feature matrix and the response vector. Feature matrix contains all the vectors (rows) of dataset in which each vector consists of the value of dependent features. Response vector contains the value of class variable (prediction or output) for each row of feature matrix.