1. **Assumption**:
2. Features are independent to each other.
3. Values associated to each class are normally distributed.
4. **Pre-processing:**
5. In this work, target column of dataset is deleted from training dataset and store in a new .csv file with target1 name.
6. Drop ‘id’ column from training and testing data
7. Missing value: In this dataset range of values is varies from 0 to 104 So, Median method is used to impute the missing values.
8. **Classifier**:

***Naive Bayes Classifier Algorithm***

Naive Bayes classifier is a powerful classification technique. Naive Bayes model is easy to build and it perform well when dataset is large. This model is based on Bayes theorem of probability to predict the label (target) of test data set.

Bayes theorem works when dataset is independent among predictors as it is already assumed in this work. Bayes theorem is based on conditional probability. Conditional probability is used to predict the probability of an occurrence on the basis of its prior knowledge. For example, purchasing butter when you already purchased the bread.

In this work Gaussian Naïve Bayes technique is used to classify the unknown test data. It is applied because each class is distributed according to Gaussian.

1. **Performance**:

Performance is estimated using cross validation. It is a statistical technique to validate the machine learning model. In this work 10-fold cross validation is applied. In which dataset divides in 10 parts whereas 9 parts used to train the model and 1 part is used to test the model.

Performance: 98% accurate.

Note: This work is compared with Logistic Regression (LR) that has cross validation score 96% that shows Naïve Bayes algorithm works well. (Source Code containing # on logistic regression function. You can check the results of LR via removing # from LR and placing # in front of naïve Bayes function)

1. **Software and Hardware used**:

Software: Eclipse IDE, Python 2.7, Microsoft office 10. Hardware: System with 4GB Ram and Linux operating system.

1. **Improvements if had more time:**
2. As work has done on 4GB RAM system so, the main bottleneck is the size of the data. In further processing either performs on 8GB RAM or use SPARK to make it speed up.
3. Multiple imputations will perform rather than median to handle the missing values of different type of features.
4. Neural network will apply to get the target values more accurately.
5. Following Steps for source code
6. Divide the training dataset in two park one park containing target label and second part containing remaining dataset.
7. Then remove ‘id’ column from test and train dataset.
8. Apply in classification model
9. Save results in .CSV file.