Problem Statement: To develop algorithms to predict the outcome of a ticket and to further classify into categories and subcategories
#--
// loading the required libraries library(data.table) // to convert the object into data.table library(Matrix) // for conversion into matrix
library(caret) // for application of k-fold cross validation
library(tm) // To process text NLP(Natural Language Processing)
library(forcats) // To club the levels of a factor variable
library(e1071) // To build SVM classsifier
#--
// GENERATING THE DATA SETS USING THE GIVEN TEXT FILES

After loading all the required libraries into R environment, the working directory is set to the location where all the Data is present and the TextClassification_Data.csv file is imported
#--
// NATURAL LANGUAGE PROCESSING

The dataset is made to run though a series of data preprocessing steps in NLP Now, a corpus is build on the datasource followed by several preprocessing steps which includes:
-> Removal of white spaces
-> Conversion of the existing text to lowercase
-> Removing Punctiations
-> Removing Stopwords
-> Conversion of all the words into Base words by(Stemming)
-> Removing any numbers that are present in the corpus
-> Removal of any special characters that are present in the corpus using iconv function
-> Conversion of the corpus into Term Frequency-Inverse Document frequency to give weightage to the words present in the document If a particular words' frequency is High then the weightage is given low as the name suggests
-> To remove sparse terms from term document matrix Now, the DocumentTermMatrix is column binded to the original dataset which ends up at 200 variables

The above steps are performed on two variables 'Data' and 'SUMMARY'

#--

Removing the duplicate levels

The variables Categories, SubCategories and previous_appointment are turned into factors and then duplicate levels are removed.
Each of the variables is grouped wrt its own categories,subcategories and previous_appointment and then graphs are plotted

#---
Finding the dependency of the categorical variables by chi squared test:

The chi squared test is performed to find the dependency of those categotical variables and the p value is found to be less than a significant value of 0.5. So, the Null hypothesis 'Categorical variables are dependent on each other' is rejected.

#---
Determining the important variables:

Random.forest.importance function is used from Fselector package and made to run over the dataset and a mere 150 variable are selected after few iterations which lead to maximum accuracy.

Building the NAIVE BAYES classifier:

Initially the Naive Bayes classifier is build on the train dataset to predict the Class variables (Categories) and (SubCategories) while tested on train data lead to an accuracy of 44% and when tested on test data it lead to an accuracy of 32%

Building the SVM classifier:
#---------------------------
Finally the SVM classifier is build on the train dataset to predict the Class variables (Categories) and (SubCategories) while tested on train data lead to an accuracy of 98% and when tested on test data it lead to an accuracy of 95%

Naive Bayes v/s SVM :

- Computation speed of naive bayes classifier was fast when compared to that of SVM classifier
- When coming to accuracy SVM did a splendid job when compared to that of naive bayes
- The data used has as higher dimensions as 57000 * 195 both the algorithms handled the high dimension data

Observations:

It is observed that majority of the patients had not had any previous appointment and hence the category APPOINTMENTS has a higher count

The number of calls wrt the patients who has previous appointments are very less which induces the fact that they had recovered from their illness
Which clearly states the quality of doctors working there

Eventhough there are nearly 57085 patients who has no previous appointment (which means they are yet to meet the doctor for the first time) there are nearly 15095 records which relates to ‘prescription’ which is not desirable and the health care service should take appropriate actions to avoid such

Hoax calls are almost negligible which serves good

As most of the enquiries made are related to appointments and prescription it is advisable to appoint chemists who has better knowledge related to drugs