1.) Prepare a classification model using Naive Bayes for salary data.

Solution:-

**Business Problem:** To prepare classification model for salary data

**Dataset:**

**Salary.csv**

**14 variables**

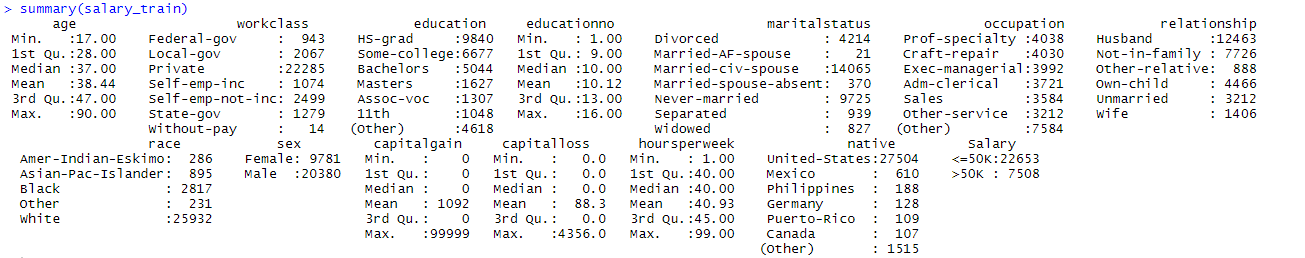
Dependent variable, y : salary

Independent variable, x : age, workclass, education, educationno, maritalstatus, occupation,relationship, race, sex, capitalgain, capitalloss, hoursperweek, native.

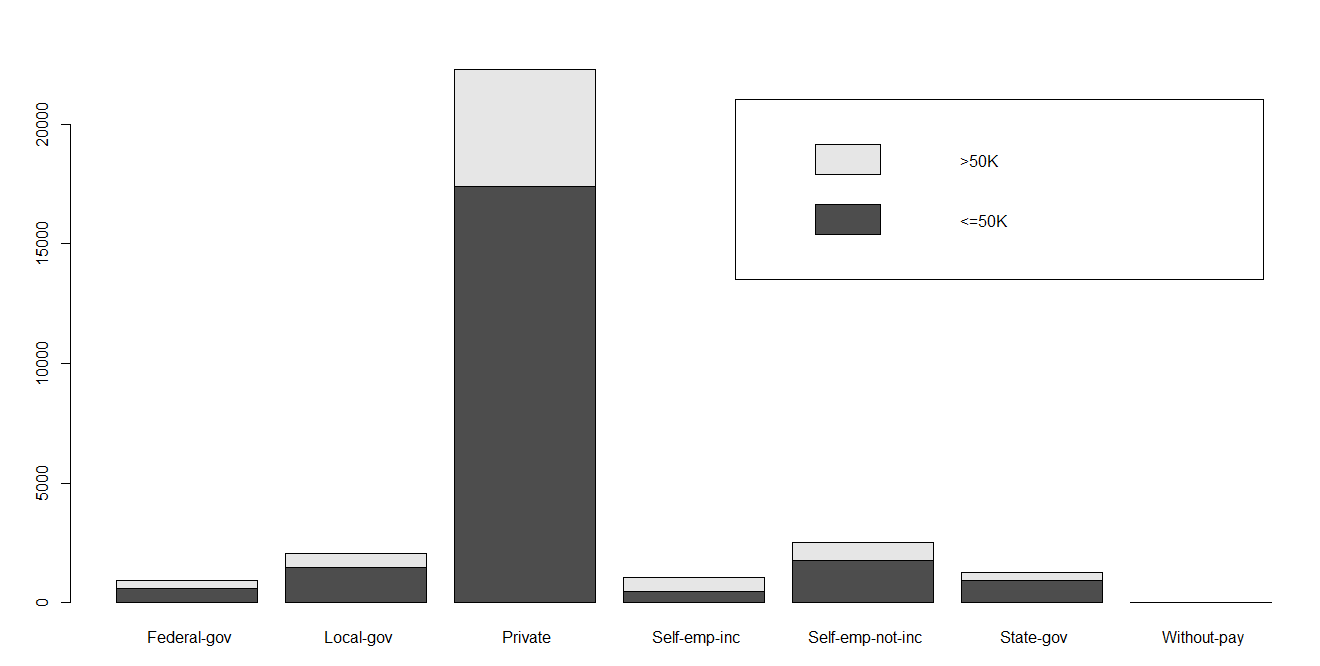
**EDA**

Here train and test data is given

Summary



Graphical Representation



The above graph represents about different work class category differenting the category of people getting salary >50K and <= 50k. Black colour indicates the number of people getting salary less than equal to 50K and grey colour indicates the people salary getting more than

50K.

**Model Building**

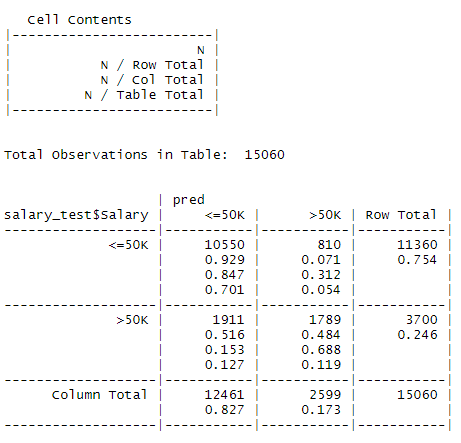
model<-naiveBayes(salary\_train$Salary~.,data=salary\_train)

A model is built with naiveBayes algorithm with train data

**Evaluation**

Once the model is prepared on train datasets we have to check the model performance on test data sets.

Prediction is made on train data



**Crosstable**

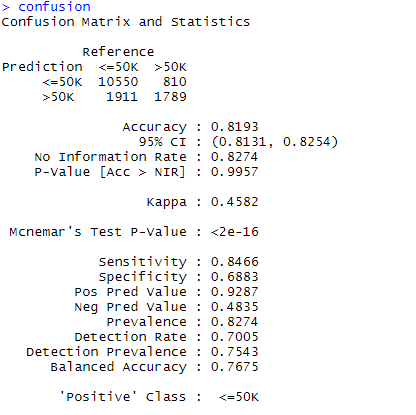
The above cross table summarizes the predicted data along with the original test data

There are Total 11360 “<=50K salary” data points in test datasets Out of which 10550 is classified as correctly and 810 is miss classified

Similarly, there is 3700 “>50k salary” as data points in test datasets. Out of which 1911 is classified correctly and 1789 is misclassified.

**Accuracy** = 81.9 %

**Confusion Matrix Summary**



2.) Build a naive Bayes model on the data set for classifying the ham and spam.

Solution:-

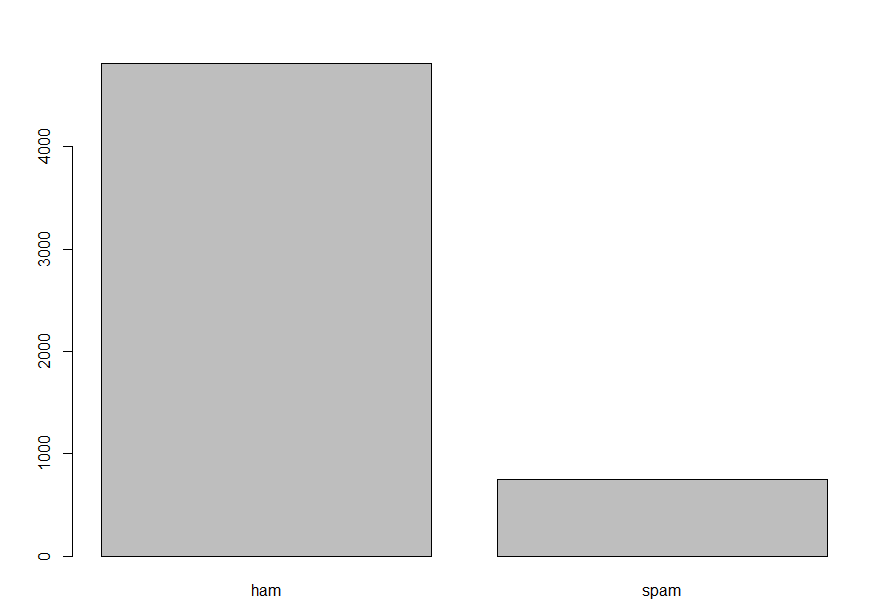
**Business Problem:** To build a model to classify message into ham and spam.

**Datasets:** Dependent variable: type

Independent variable: text

**EDA:**

**Graphical Representation**



There are 4812 ham and 747 spam in the datasets

Corpus is created and necessary data cleaning is done

Data cleaning consists, converting to lowercase letters, removing numbers, punctuation, stop words and white space

Document term matrix is created

**Data Partitioning**

Data is partitioned into train and test for sms datasets, corpus, and document term matrix

A dictionary is created of words repeated more than 5 times.

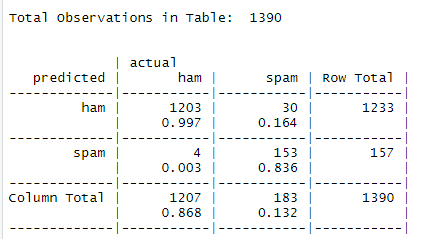
**Model Building:**

A model is created of train data based using NaiveBayes algorithm.

**Evaluation:**

The model performance is evaluated with test data.

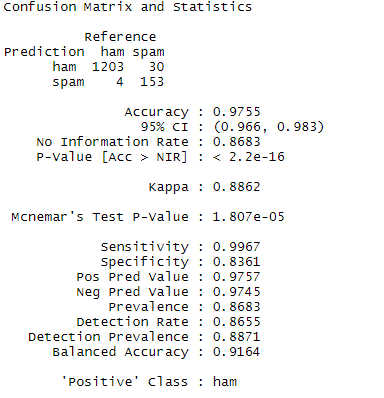
**Cross tabulation**



Out of total 1207 ham actual from test datasets, 1203 are classified correctly as ham and 4 are misclassified as spam after prediction

Out of total 183 spam actual from test datasets, 153 is classified correctly as spam and 30 are misclassified as ham after prediction

**Confusion matrix**



**Accuracy:** 97.55%