EMAIL CLASSIFICATION

2023-04-17

DATA LOADING

```
data=read.csv("spam_ham_dataset.csv")
names(data)=c("col","label","text","label_num")
```

DATA PREPROCESSING

```
#Check the overview of the data
head(data)
```

```
##    col label
## 1  605   ham
## 2 2349   ham
## 3 3624   ham
## 4 4685  spam
## 5 2030   ham
## 6 2949   ham
##
## 1
## 2
## 3 Subject: neon retreat\nho ho ho , we ' re around to that most wonderful time of the year - - - neo
## 4
## 5
## 6
##   label_num
## 1         0
## 2         0
## 3         0
## 4         1
## 5         0
## 6         0
```

```
#Check for any missing values
sum(is.na(data))
```

```
## [1] 0
```

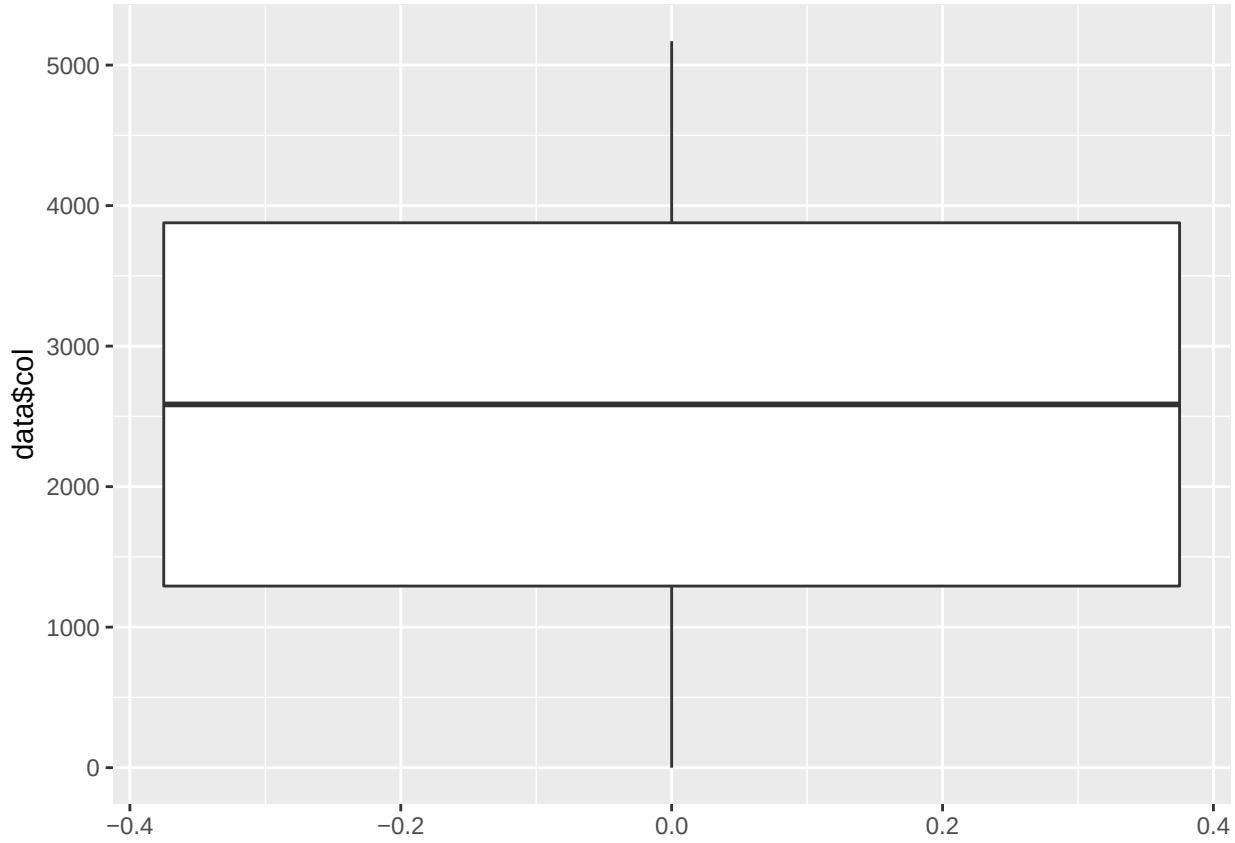
There are no missing values so there is no need of deleting any rows.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Plot a boxplot to check whether there are any outliers in the dataset.
library(ggplot2)
ggplot(data, aes(y = data$col)) +
  geom_boxplot()
```

```
## Warning: Use of `data$col` is discouraged. Use `col` instead.
```



There are no outliers from the dataset. Now we will check the datatypes of the columns.

```
#Check the datatypes of each column.
str(data)
```

```
## 'data.frame':    5171 obs. of  4 variables:
```

```
##  $ col      : int  605 2349 3624 4685 2030 2949 2793 4185 2641 1870 ...
##  $ label    : chr  "ham" "ham" "ham" "spam" ...
##  $ text     : chr  "Subject: enron methanol ; meter # : 988291\nthis is a follow up to the note i ga
##  $ label_num: int  0 0 0 1 0 0 0 1 0 0 ...
```

Yes, all the columns are having the correct datatypes. Because if we see the label it should be categorical but a separate column has been created to represent it as a category. That is "label_num".

```
data$text<-gsub('[^[:alnum:] ]', ' ',data$text )
data$text<-gsub('Subject','',data$text)
head(data)
```

```
##    col label
## 1  605   ham
## 2 2349   ham
## 3 3624   ham
## 4 4685  spam
## 5 2030   ham
## 6 2949   ham
##
## 1
## 2
## 3   neon retreat ho ho ho   we   re around to that most wonderful time of the year       neon leader
## 4
## 5
## 6
##   label_num
## 1         0
## 2         0
## 3         0
## 4         1
## 5         0
## 6         0
```

The text attribute is modified to remove punctuation marks and other special characters within the text. Also the subject tag is removed as it remains the same. This helps identify key words in classification.

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 4.2.3
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.2.2
```

```
dat_corpus <- Corpus(VectorSource(data$text))

# Clean the corpus
dat_corpus <- tm_map(dat_corpus, tolower)
```

```
## Warning in tm_map.SimpleCorpus(dat_corpus, tolower): transformation drops
## documents
```

```
dat_corpus <- tm_map(dat_corpus, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(dat_corpus, removeNumbers): transformation drops
## documents
```

```
dat_corpus <- tm_map(dat_corpus, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(dat_corpus, removePunctuation): transformation
## drops documents
```

```
dat_corpus <- tm_map(dat_corpus, stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(dat_corpus, stripWhitespace): transformation
## drops documents
```

```
dat_corpus <- tm_map(dat_corpus, removeWords, stopwords("english"))
```

```
## Warning in tm_map.SimpleCorpus(dat_corpus, removeWords, stopwords("english")):
## transformation drops documents
```

```
dat_dtm <- DocumentTermMatrix(dat_corpus)
dat_dtm
```

```
## <<DocumentTermMatrix (documents: 5171, terms: 45045)>>
## Non-/sparse entries: 309558/232618137
## Sparsity           : 100%
## Maximal term length: 24
## Weighting          : term frequency (tf)
```

Removing terms which don't occur frequently

```
sdtm<-removeSparseTerms(dat_dtm,0.95)
sdtm
```

```
## <<DocumentTermMatrix (documents: 5171, terms: 117)>>
## Non-/sparse entries: 62655/542352
## Sparsity           : 90%
## Maximal term length: 11
## Weighting          : term frequency (tf)
```

Converting the word sparse matrix to a dataframe

```
word_sparse=data.frame(as.matrix(sdtm),email_class=data$label)
head(word_sparse)
```

```
##   can change daily daren enron flow gas meter please volume attached file hpl
## 1   1      1     2     1     1    1   1     1      1      1        0    0   0
## 2   0      0     0     0     0    0   0     0      0      0        1    1   1
## 3   4      1     0     0     0    0   0     0      0      0        0    0   0
## 4   0      0     0     0     0    0   0     0      0      0        0    0   0
## 5   1      0     0     0     0    0   0     0      0      0        0    0   0
## 6   0      4     0     0     1    0   0     0      0      0        0    0   0
##   january nom see xls also available back com date email first following get
## 1       0   0   0   0    0         0    0   0    0     0     0         0   0
## 2       1   1   1   2    0         0    0   0    0     0     0         0   0
## 3       3   0   0   0    1         1    1   1    1     1     1         1   4
## 4       0   0   0   0    0         0    0   0    0     0     0         0   0
## 5       0   0   0   0    0         0    0   0    0     0     0         0   0
## 6       0   0   0   0    0         0    0   1    0     0     0         0   0
##   houston just know let like need now one think time week will www deal price
## 1       0    0    0   0    0    0   0   0     0    0    0    0   0    0     0
## 2       0    0    0   0    0    0   0   0     0    0    0    0   0    0     0
## 3       1    1    2   3    3    1   1   2     4    7    2    2   1    0     0
## 4       0    0    0   0    0    0   0   0     0    0    0    0   0    0     0
## 5       0    2    1   0    0    1   0   0     0    0    0    0   0    3     1
## 6       0    0    0   0    0    0   0   0     0    0    0    0   0    0     0
##   http made message needs click day don free mail march new purchase sale set
## 1    0    0       0     0     0   0   0    0    0     0   0        0    0   0
## 2    0    0       0     0     0   0   0    0    0     0   0        0    0   0
## 3    0    0       0     0     0   0   0    0    0     0   0        0    0   0
## 4    0    0       0     0     0   0   0    0    0     0   0        0    0   0
## 5    0    0       0     0     0   0   0    0    0     0   0        0    0   0
## 6    1    1       1     1     0   0   0    0    0     0   0        0    0   0
##   take today use best make net want forwarded mmbtu subject texas april
## 1    0     0   0    0    0   0    0         0     0       0     0     0
## 2    0     0   0    0    0   0    0         0     0       0     0     0
## 3    0     0   0    0    0   0    0         0     0       0     0     0
## 4    0     0   0    0    0   0    0         0     0       0     0     0
## 5    0     0   0    0    0   0    0         0     0       0     0     0
## 6    0     0   0    0    0   0    0         0     0       0     0     0
##   business call company days forward help information look may month next.
## 1        0    0       0    0       0    0           0    0   0     0     0
## 2        0    0       0    0       0    0           0    0   0     0     0
## 3        0    0       0    0       0    0           0    0   0     0     0
## 4        0    0       0    0       0    0           0    0   0     0     0
## 5        0    0       0    0       0    0           0    0   0     0     0
## 6        0    0       0    0       0    0           0    0   0     0     0
```

```
##   north number send actuals thanks system bob contract corp ect farmer fyi hou
## 1     0      0    0       0      0      0   0        0    0   0      0   0   0
## 2     0      0    0       0      0      0   0        0    0   0      0   0   0
## 3     0      0    0       0      0      0   0        0    0   0      0   0   0
## 4     0      0    0       0      0      0   0        0    0   0      0   0   0
## 5     0      0    0       0      0      0   0        0    0   0      0   0   0
## 6     0      0    0       0      0      0   0        0    0   0      0   0   0
##   line mary production robert sitara ticket volumes contact two find original
## 1    0    0          0      0      0      0       0       0   0    0        0
## 2    0    0          0      0      0      0       0       0   0    0        0
## 3    0    0          0      0      0      0       0       0   0    0        0
## 4    0    0          0      0      0      0       0       0   0    0        0
## 5    0    0          0      0      0      0       0       0   0    0        0
## 6    0    0          0      0      0      0       0       0   0    0        0
##   questions sent per energy last well work nomination order thank effective
## 1         0    0   0      0    0    0    0          0     0     0         0
## 2         0    0   0      0    0    0    0          0     0     0         0
## 3         0    0   0      0    0    0    0          0     0     0         0
## 4         0    0   0      0    0    0    0          0     0     0         0
## 5         0    0   0      0    0    0    0          0     0     0         0
## 6         0    0   0      0    0    0    0          0     0     0         0
##   america give list deals email_class
## 1       0    0    0     0         ham
## 2       0    0    0     0         ham
## 3       0    0    0     0         ham
## 4       0    0    0     0        spam
## 5       0    0    0     0         ham
## 6       0    0    0     0         ham
```

DATA SPLITTING

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Set seed
set.seed(123)



# Splitting the dataset into training and testing sets
```

```
train_index <- createDataPartition(word_sparse$email_class, p = 0.7, list = FALSE)
train_data <- word_sparse[train_index, ]
test_data <- word_sparse[-train_index, ]
```

```
train_data$email_class <- factor(train_data$email_class, levels = c("spam", "ham"))
test_data$email_class <- factor(test_data$email_class, levels = c("spam", "ham"))
```

BUILDING NAIVE BAYES CLASSIFICATION MODEL

```
# Building the Naive Bayes Classifier Model
library(e1071)
nb_model <- naiveBayes(email_class ~ ., data = train_data)
```

```
# Use the classifier to make predictions on the test data
predicted_labels = predict(nb_model, newdata = test_data[,-ncol(test_data)])
```

Our model has been built. Now we will check the accuracy of the model.

```
levels(predicted_labels)
```

```
## [1] "spam" "ham"
```

```
levels(test_data$email_class)
```

```
## [1] "spam" "ham"
```

```
library(caret)
confusionMatrix(predicted_labels, test_data$email_class)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction spam ham
##       spam  442 446
##       ham     7 655
##
##                Accuracy : 0.7077
##                  95% CI : (0.6844, 0.7303)
##     No Information Rate : 0.7103
##     P-Value [Acc > NIR] : 0.6009
##
##                   Kappa : 0.4493
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9844
##             Specificity : 0.5949
##          Pos Pred Value : 0.4977
```

```
##          Neg Pred Value : 0.9894
##              Prevalence : 0.2897
##          Detection Rate : 0.2852
##    Detection Prevalence : 0.5729
##       Balanced Accuracy : 0.7897
##
##        'Positive' Class : spam
##
```

The accuracy of the model is 70%.

Logistic Regression Model

```
X_train <- data.matrix(train_data[, -ncol(train_data)])
Y_train <- (train_data$email_class)
X_test <- data.matrix(test_data[, -ncol(test_data)])
Y_test <- (test_data$email_class)
```

Training the logistic regression model

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.3
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-7
```

```
lr<-cv.glmnet(X_train,Y_train, family = "binomial")
```

Predicting test data

```
prediction_lr<-predict(lr,newx=X_test,type="class")
```

Accuracy details

```
library(caret)
confusionMatrix(as.factor(prediction_lr),as.factor(Y_test))
```

```
## Warning in confusionMatrix.default(as.factor(prediction_lr), as.factor(Y_test)):
## Levels are not in the same order for reference and data. Refactoring data to
## match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction spam  ham
##       spam  423   81
##       ham    26 1020
##
##                Accuracy : 0.931
##                  95% CI : (0.9172, 0.9431)
##     No Information Rate : 0.7103
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8381
##
##  Mcnemar's Test P-Value : 1.786e-07
##
##             Sensitivity : 0.9421
##             Specificity : 0.9264
##          Pos Pred Value : 0.8393
##          Neg Pred Value : 0.9751
##              Prevalence : 0.2897
##          Detection Rate : 0.2729
##    Detection Prevalence : 0.3252
##       Balanced Accuracy : 0.9343
##
##        'Positive' Class : spam
##
```