**EDWISOR- DATA SCIENCE PROJECT 2**

**Personalized Medicine: Redefining Cancer Treatment**

**Objective:**

To develop a Machine Learning algorithm that automatically classifies genetic variations.

**Introduction:**

Once sequenced, a cancer tumor can have thousands of genetic mutations.

But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers).

**Software requirements:**

**Language used:** R, **IDE used:** RStudio

**Hardware requirements:**

**RAM:** 8 GB, **OS: Windows 10, Processor:** 64 bit processor

**Given data:**

(after formatting)

training\_text: Observations: 3321, Variables: 2

test\_text: Observations: 5668, Variables: 2

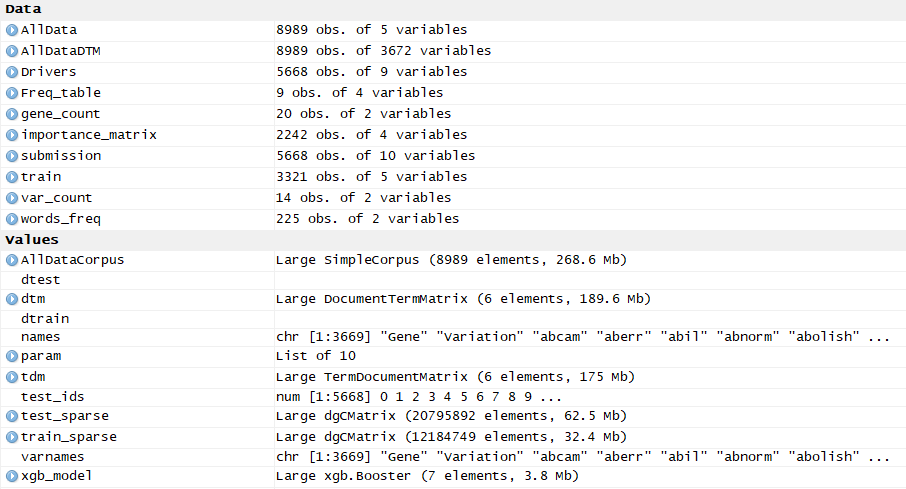
training\_variants: Observations: 3321, Variables: 4

test\_variants: Observations: 5668, Variables: 3

1. Combine test text and test variants into test
2. Combine training text and training variants into train
3. Combine 1 and 2 to get AllData by ID

AllData: Observations: 8989, Variables: 5

**All workspace objects, data frames and values:**



**Missing value analysis**

sum(is.na(AllData))

#No missing values in the data

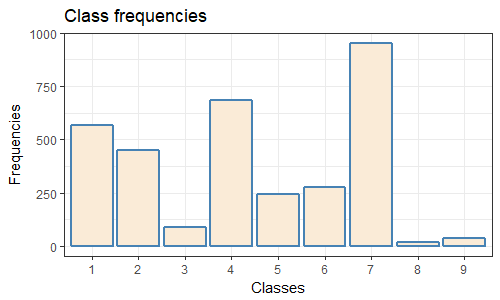
**Grubbs test for outliers**

grubbs.test(AllData$Class, type = 10)

# No outliers present

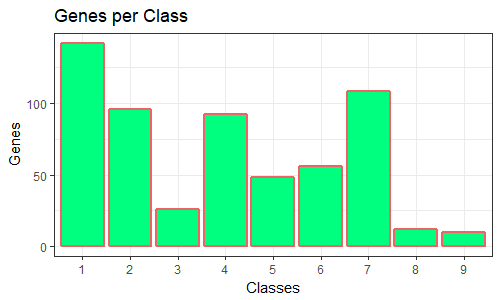
**Visualisatons:**

1. Frequencies of each class



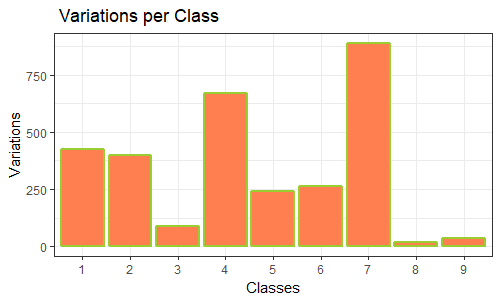
Most observations belong to class 7, 2nd most- class 4,3rd most- class 1

1. Gene grouped by Class



Most genes belong to class 1

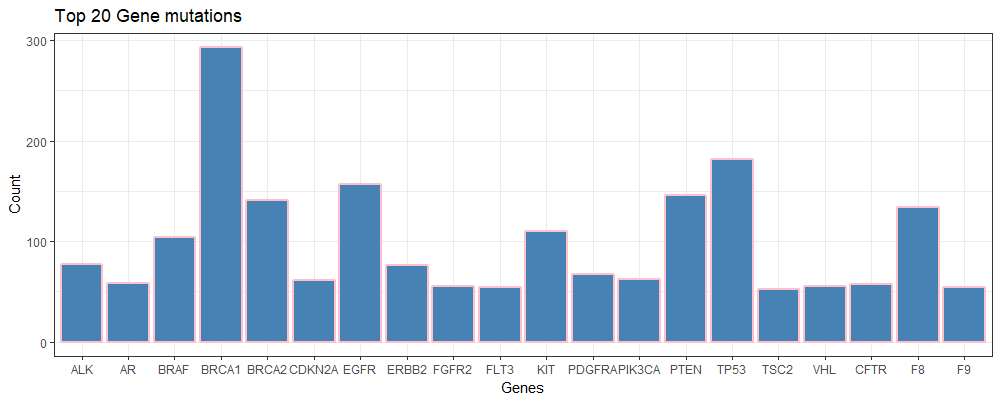
1. Variations grouped by Class



Most variations belong to class 7

1. Grouping All Data by gene

Plotting the top 20 gene mutations

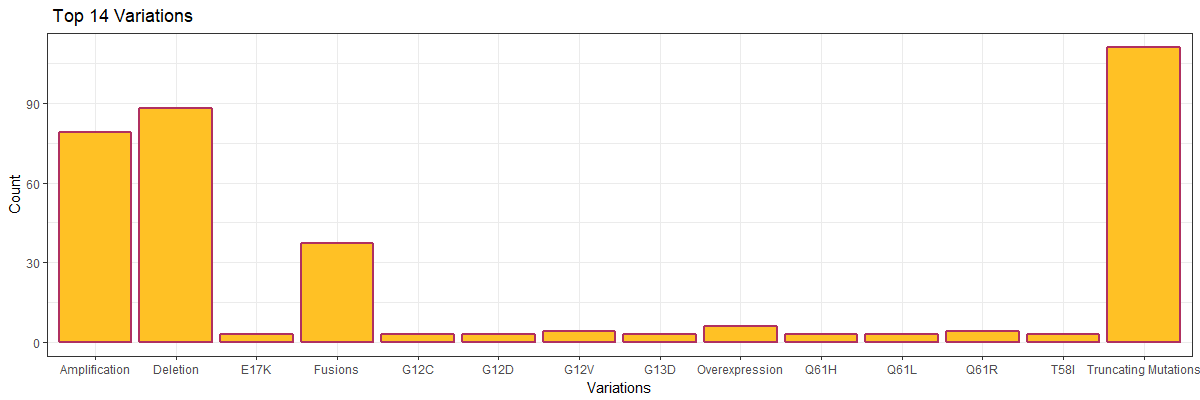


Most frequent Gene- BRCA1: 264 times

1. Grouping All Data by variation

Plotting the top 14 variations

Most frequent Variation- Truncating Mutations: 93 times



**Final steps:**

* **Text Mining**

AllDataCorpus = Corpus(VectorSource(AllData$Text))

#case folding

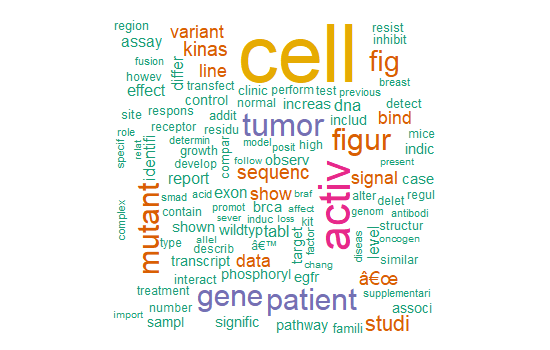
AllDataCorpus = tm\_map(AllDataCorpus, content\_transformer(tolower))

#remove punctuation marks

AllDataCorpus = tm\_map(AllDataCorpus, removePunctuation)

And so on..

1. Word cloud for top 120 words



**Top 15 words:**

> head(arrange(words\_freq,desc(Frequency)), n = 15)

Words Frequency

1 mutat 818393

2 cell 817523

3 activ 425278

4 express 333165

5 protein 300213

6 tumor 297620

7 patient 285337

8 gene 283341

9 cancer 281748

10 mutant 270886

11 figur 268995

12 fig 258794

13 result 187007

14 studi 186787

15 domain 183667

**Frequent terms:**

> findFreqTerms(tdm, 100)

[1] "abcam" "aberr" "abil"

[4] "abnorm" "abolish" "absenc"

[7] "acceptor" "accord" "achiev"

[10] "acid" "act" "action"

[13] "activ" "addit" "address"

[16] "adult" "affect" "ago"

[19] "aim" "alanin" "albeit"

[22] "allel" "allow" "almost"

[25] "alon" "alreadi" "alter"

…

* **Prediction using XGBoot package**

# Params for xgboost

param = list(booster = "gbtree",

objective = "multi:softprob",

eval\_metric = "mlogloss",

num\_class = 9,

eta = .2,

gamma = 1,

max\_depth = 5,

min\_child\_weight = 1,

subsample = .7,

colsample\_bytree = .7)

# Train the model

xgb\_model = xgb.train(data = dtrain,

params = param,

watchlist = list(train = dtrain),

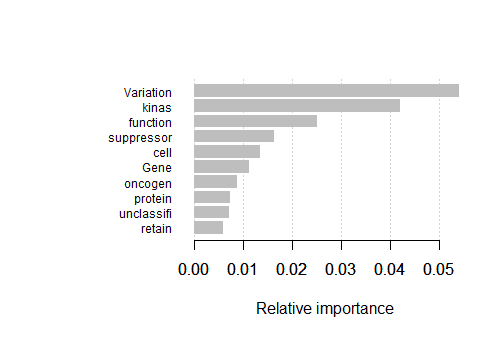
nrounds = 250,

verbose = 1,

print.every.n = 5)

Drivers = as.data.table(t(matrix(predict(xgb\_model, dtest), nrow=9, ncol=nrow(dtest))))

colnames(Drivers) = c("class1","class2","class3","class4","class5","class6","class7","class8","class9")



* **Write the output in a csv file**

write.table(data.table(ID=test\_ids, Drivers), "C:/MyFiles/Kshiti/Edwisor/PROJECT 2/submission.csv", sep=",", dec=".", quote=FALSE, row.names=FALSE)

submission <- read\_csv("C:/MyFiles/Kshiti/Edwisor/PROJECT 2/submission.csv")

> head(submission,5)

| **ID** | **class1** | | **class2** | | **class3** | | **class4** | | **class5** | **class6** | | **class7** | | **class8** | | **class9** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | |  | |  | |  | |  |  | |  | |  | |  |
| **1** | 0.0093872063 | | 0.1029391512 | | 1.218516e-03 | | 0.0066765398 | | 0.0003449027 | 0.0025391763 | | 0.8726164103 | | 0.0031370898 | | 1.140994e-03 |
| **2** | 0.1558232754 | | 0.0247223750 | | 8.109096e-04 | | 0.8096651435 | | 0.0003397389 | 0.0037860579 | | 0.0037673793 | | 0.0005014918 | | 5.836158e-04 |
| **3** | 0.0184464157 | | 0.3077886701 | | 5.807454e-03 | | 0.0812530443 | | 0.0023864775 | 0.0495482758 | | 0.5271501541 | | 0.0052772304 | | 2.342299e-03 |
| **4** | 0.0423361436 | | 0.4241456985 | | 2.266459e-03 | | 0.0200422630 | | 0.0016057910 | 0.0059727901 | | 0.4964525998 | | 0.0050693941 | | 2.108884e-03 |
| **5** | 0.6685518622 | | 0.0717946738 | | 1.438175e-03 | | 0.2025402486 | | 0.0020925642 | 0.0093217166 | | 0.0316102318 | | 0.0069399416 | | 5.710611e-03 |