Personalized Medicine and cancer treatment

Background

A lot has been said during the past several years about how precision medicine and, more concretely, how genetic testing is going to disrupt the way diseases like cancer are treated.

But this is only partially happening due to the huge amount of manual work still required. Once sequenced, a cancer tumor can have thousands of genetic mutations. Currently this interpretation of genetic mutations is being done manually. This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

The project is to distinguish the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers). And to develop a Machine Learning algorithm that, using this knowledge base as a baseline, automatically classifies genetic variations.

Business problem

Automatically classifying genetic variations will decrease the time spent for classification and precision medicine for cancer treatment becomes from the potentiality to reality. It will increase the five years' survival rate of cancer patients, improve the life quality of the patients. It also will decrease the medical expense and relieve the public burden.

Client

Pharmaceutical company. Biotech company. Health provider.

Data wrangling

There are six records 'Text' variable are null and I delete the six rows.

Explanatory data analysis

I. Gene distribution overview

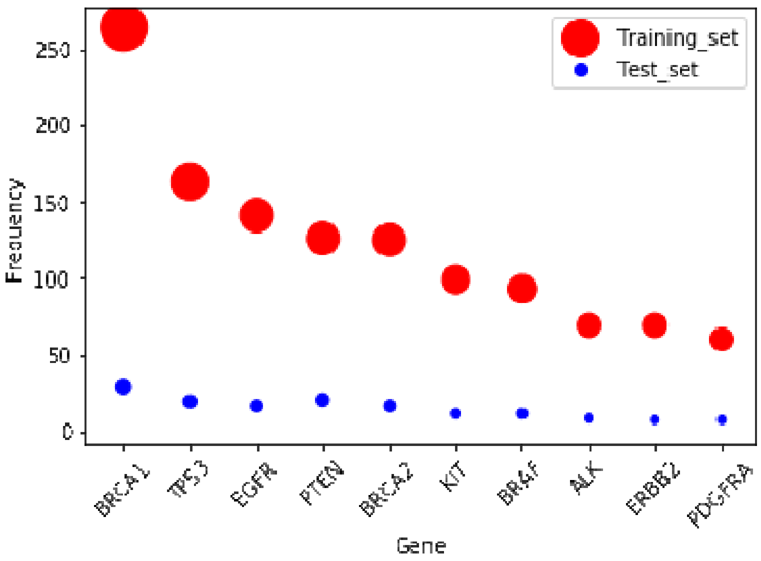
Top 10 occurred genes

```
Genes with maximal occurrences in training data Gene BRCA1 264 TP53 163
EGFR 141 PTEN 126 BRCA2 125 KIT 99 BRAF 93 ERBB2 69 ALK 69 PDGFRA 60 Name:
ID, dtype: int64
```

```
Genes with maximal occurrences in test data Gene F8 134 CFTR 57 F9 54
G6PD 46 GBA 39 PAH 38 AR 38 CASR 37 ARSA 30 BRCA1 29 Name: ID, dtype:
int64
```

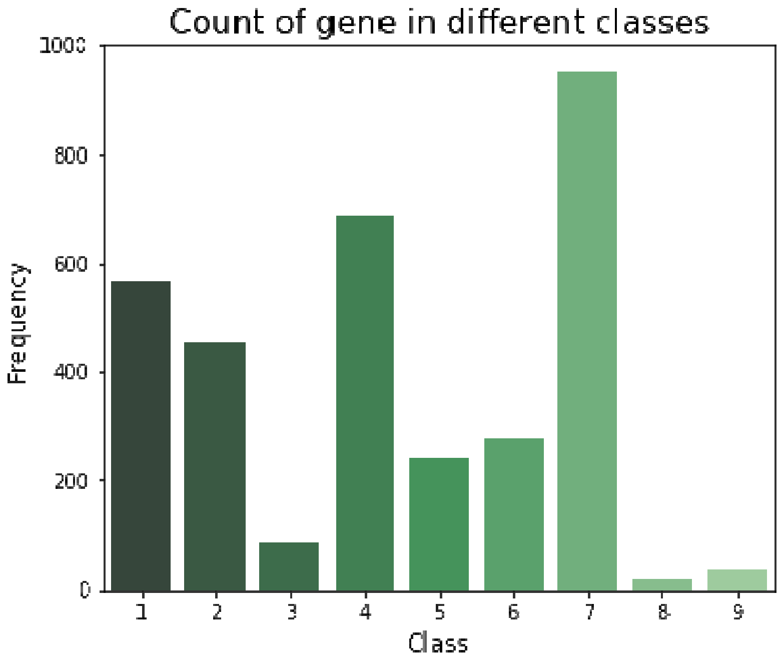
The top ten genes in training data are different from that of the test data. The only common top ten gene is BRCA1. It shows the gene distributions in the training and test data are not the same.

Top 10 genes of training set scatter plot



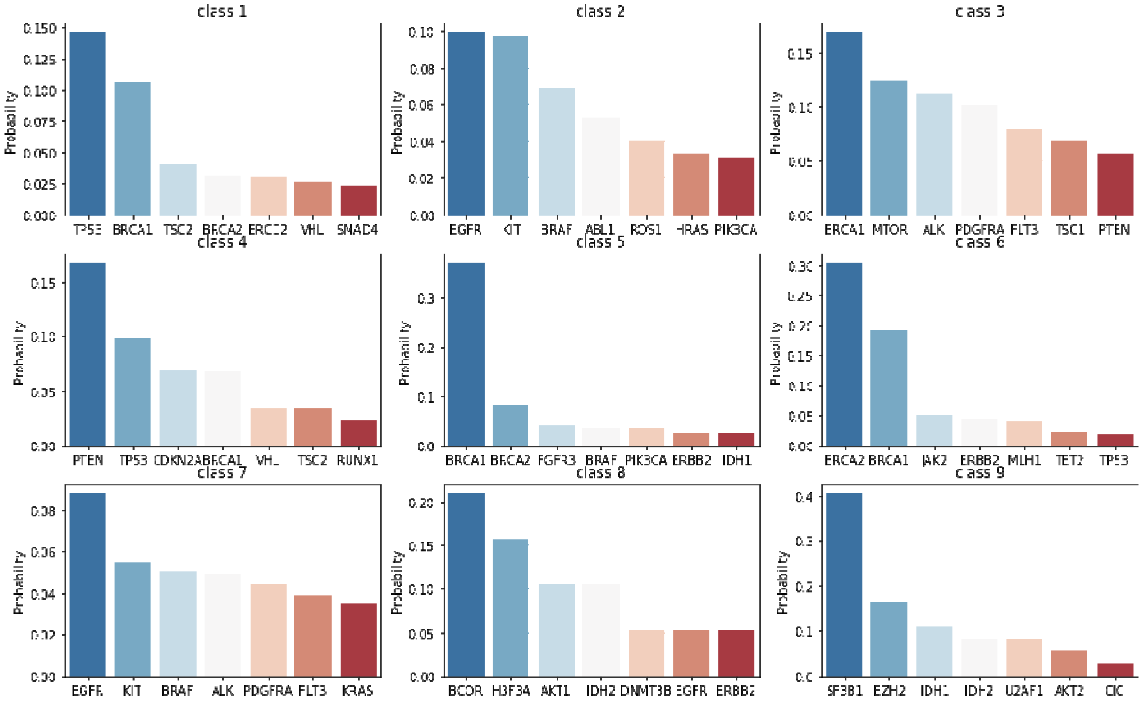
II. Gene distribution in classes

Gene frequency in nine mutation classes



Genes’ frequency peaks at class 7. Lowest gene occurred class is 8 and 9.

Top seven genes distribution in different classes



EGFR rank first in class 2 and 7.

BRCA1 ranks first in class 3, 5 and ranks second in class 6.

BRCA2 ranks first in class 6 and ranks second in 5.

And these three genes are among the top ten occurred genes in training data.

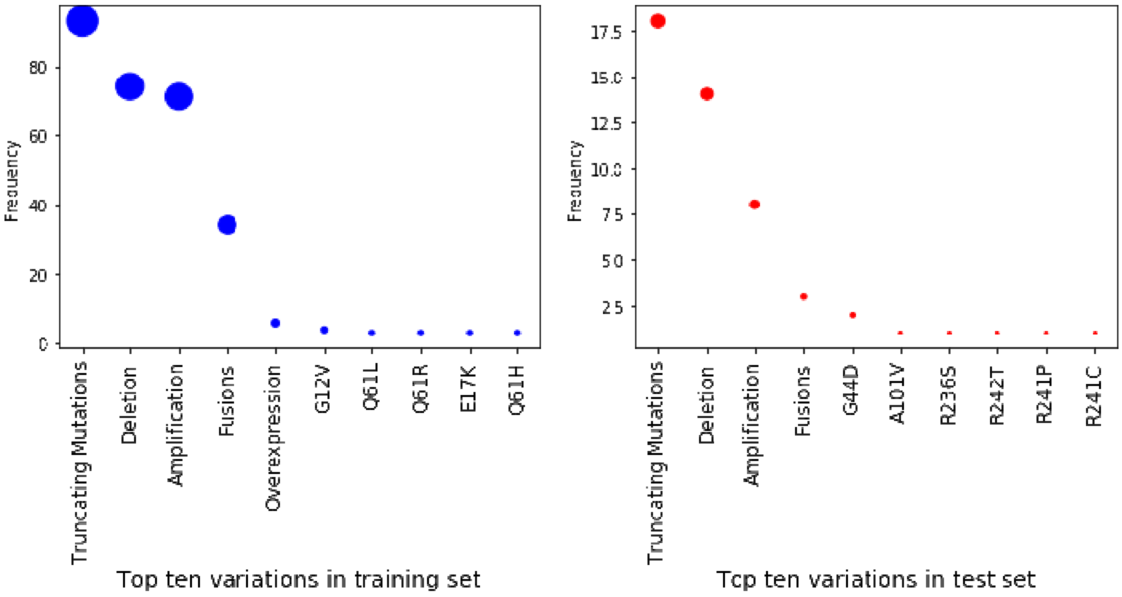
III. Variations overview

Top 10 variations

```
Variations with maximal occurrences Variation Truncating Mutations 93
Deletion 74 Amplification 71 Fusions 34 Overexpression 6 G12V 4 E17K 3
T58I 3 Q61L 3 Q61R 3 Name: Variation, dtype: int64
```

```
Variations with maximal occurrences Variation Truncating Mutations 18
Deletion 14 Amplification 8 Fusions 3 G44D 2 H1464P 1 H12Q 1 H132P 1 H136R
1 H137L 1 Name: Variation, dtype: int64
```

Top 10 variations' scatter plot



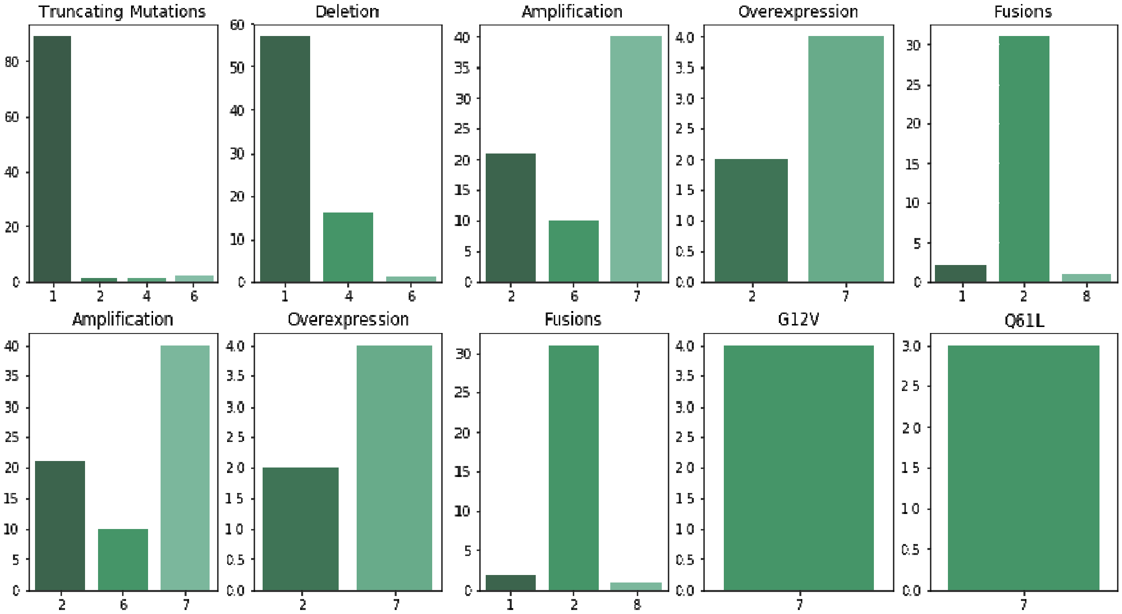
Top four variations are same in train and test data set. They are Truncating mutations, Deletion, Amplification and Fusions.

But the counts of top four variations in training data are much higher than that in test data.

It shows the variation distributions in training data and test data are of much difference.

IV. Variation and class

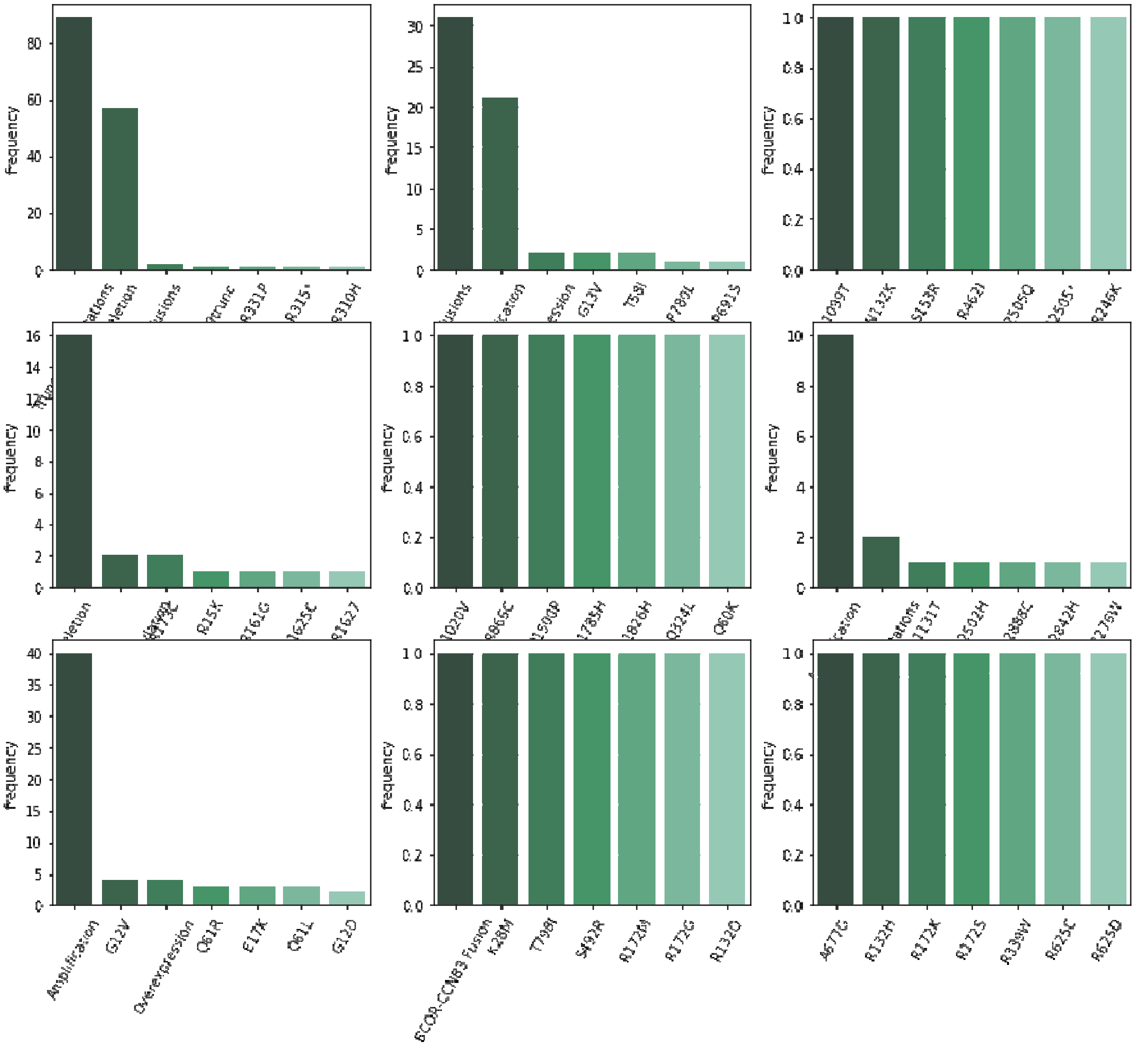
Top 10 variations' class distribution



Top one and top two variations are Truncating mutations and Deletion. They are both located in Class 1, 4 and 6. Truncating mutations is also in class 2.

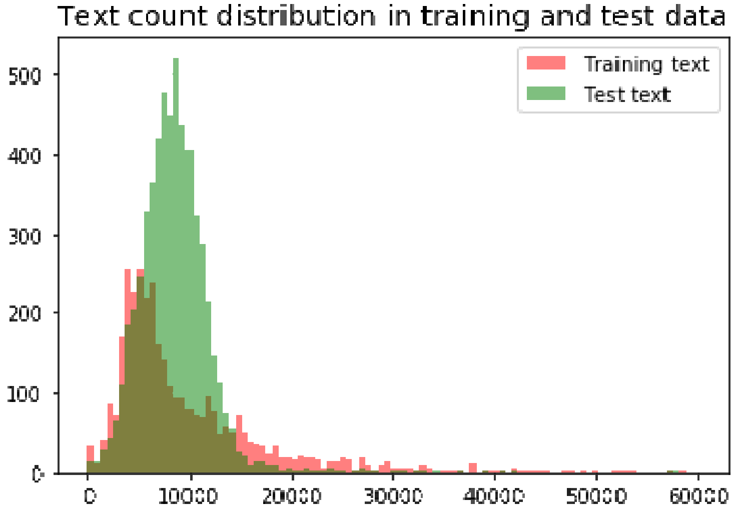
Top three and top four variations are Amplification and Fusions. They are both located in Class 2 and 7. And Amplification is also in class 6.

Variation distribution in different classes



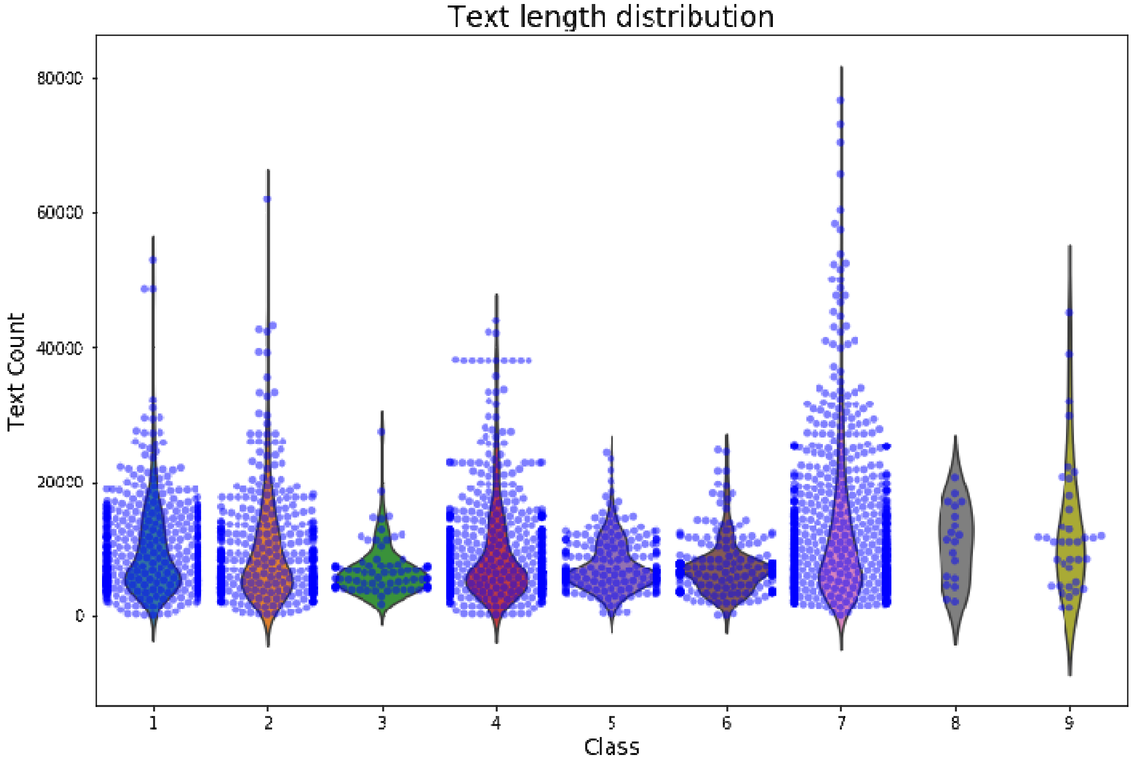
Class 3, 5, 8, 9 the maximum gene variation is 1.
Truncating mutations ranks first in class 1 and second in class 6 (class location 1, 2, 4, 6).
Deletion ranks first in class 4 and second in class 1 (class location 1, 4, 6).
Amplification ranks first in class 6 and 7. And it ranks second in class 2 (class location 2, 6, 7).
Fusions ranks first in class 2 (class location 2, 7).

V. Text length overview



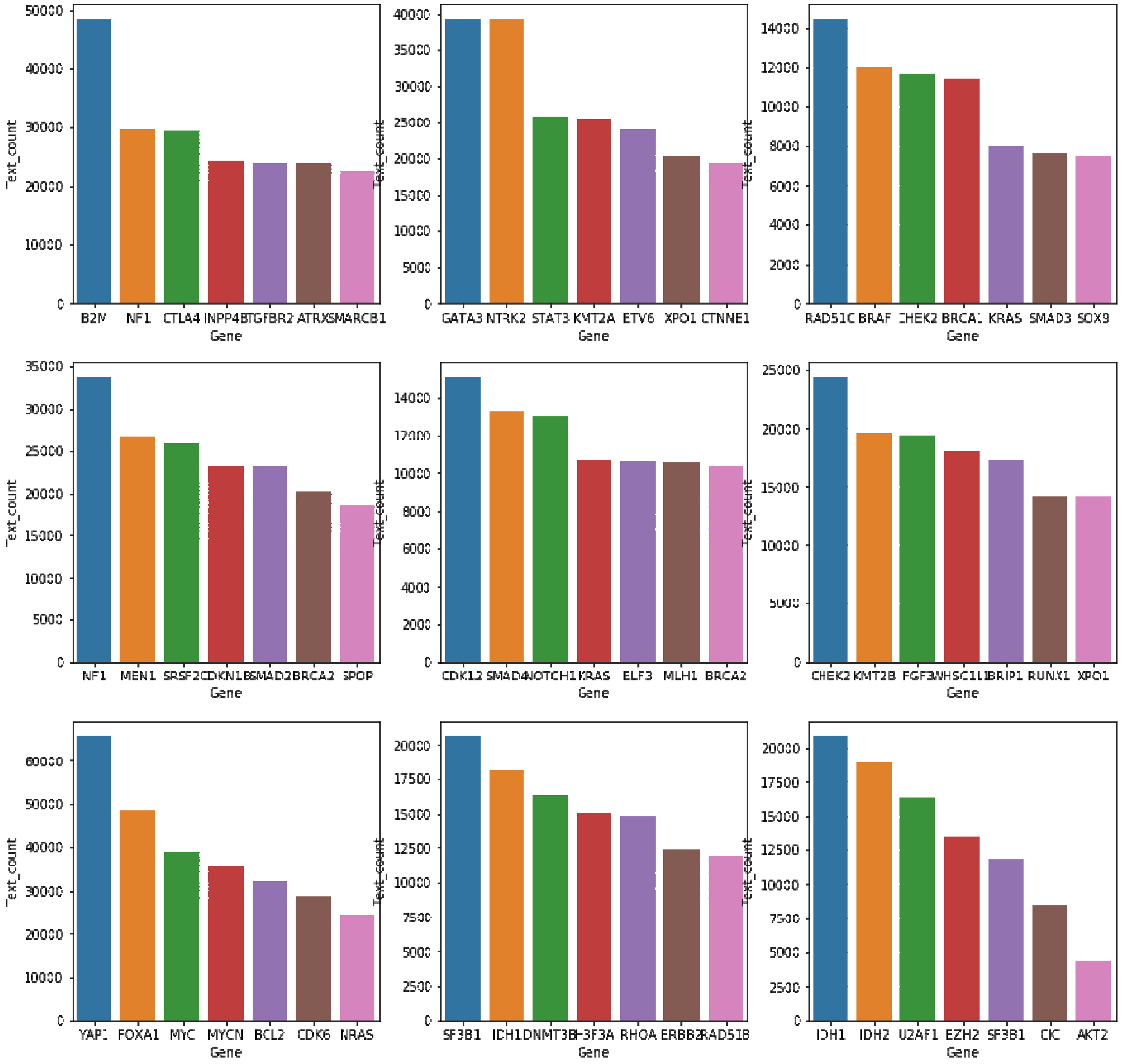
The most frequent text count in training set is about 5000 while that in test set is about 10000. The text amount distribution in both training and test data set are close to normal distribution.

VI. Text length distribution by class



Text count in Class 7 ranges from 1 to nearly 80000. It is the largest range among the nine classes. Class 8 and 9 have the smallest ranges. The text count distribution is similar to the gene occurrence distribution in different classes.

VII. Mean text count by genes in different classes

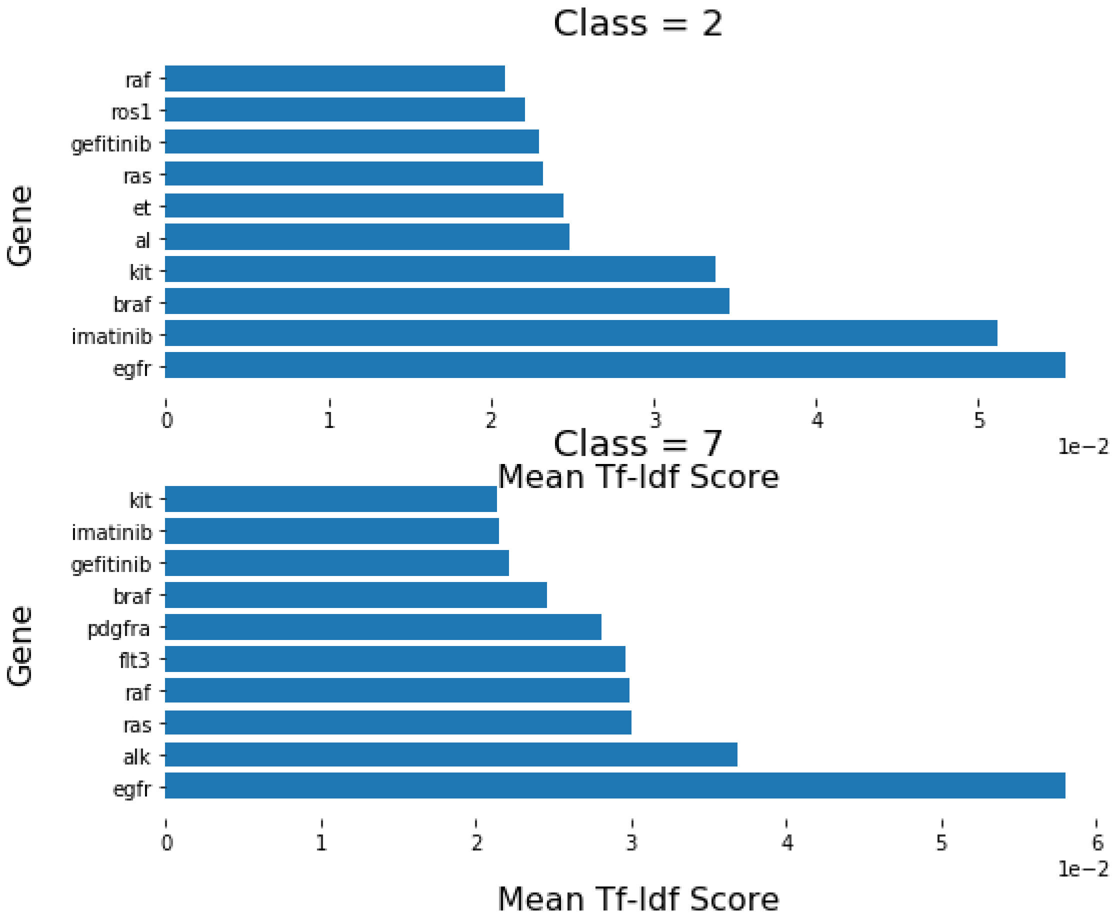


This plot shows the top seven mean text count for genes in different classes.
BRCA1 and BRCA2 are among the top ten occurred genes. And they are also in the top seven text count genes in some classes.

VIII. Top 10 TF-IDF features in class 2 and 7

Gene distribution analysis shows top one gene EGFR ranks first in class 2 and 7.
Variation distribution analysis shows Amplification and Fusions (top 2 variations) mainly located in class 2 and 7. What's more, Amplification ranks first in class 7 and Fusions ranks first in class 2.
The above fact give us an impression that class 2 and class 7 are highly similar in both gene occurrence and variation distribution. What are about their top 10 text features?

Tf-idf is known as one good technique to use for text transformation and get good features out of text for training our machine learning model. Let's see the top 10 text features in class 2 and class 7.



There are seven common text features in the two classes. There are 70% similarity for the top 10 text features between class 2 and class 7!