Personalised Medicine : Redefining Cancer Treatment

Nikhil Mehta (MT18043) Divya Garg (MT18095)

- ## Introduction:

Cancer is the leading cause of death all over the world and it is still increasing in alarming rate. The progress of cancer treatment has been quite slow. This is due to the complexity nature of cancer and its appropriate treatment.

Personalized medicine and treatment are studies that takes genetic makeup into account to maximize efficiency and minimize toxicity for each patient by using the right drug with the right dosage. We are provided with a dataset of genes and cancer-related mutations, a text file of scientific literature related to each mutation, and a 1-9 classification of each case, which had been hand-annotated by a scientist at MSK after reviewing the data.

Although we are not told in advance what each class represented, our goal is to correctly classify new test mutations after training machine learning models on the original datasets and text files. This task would ultimately help automate the work of distinguishing between driver (cancer-causing) and passenger (neutral) mutations.

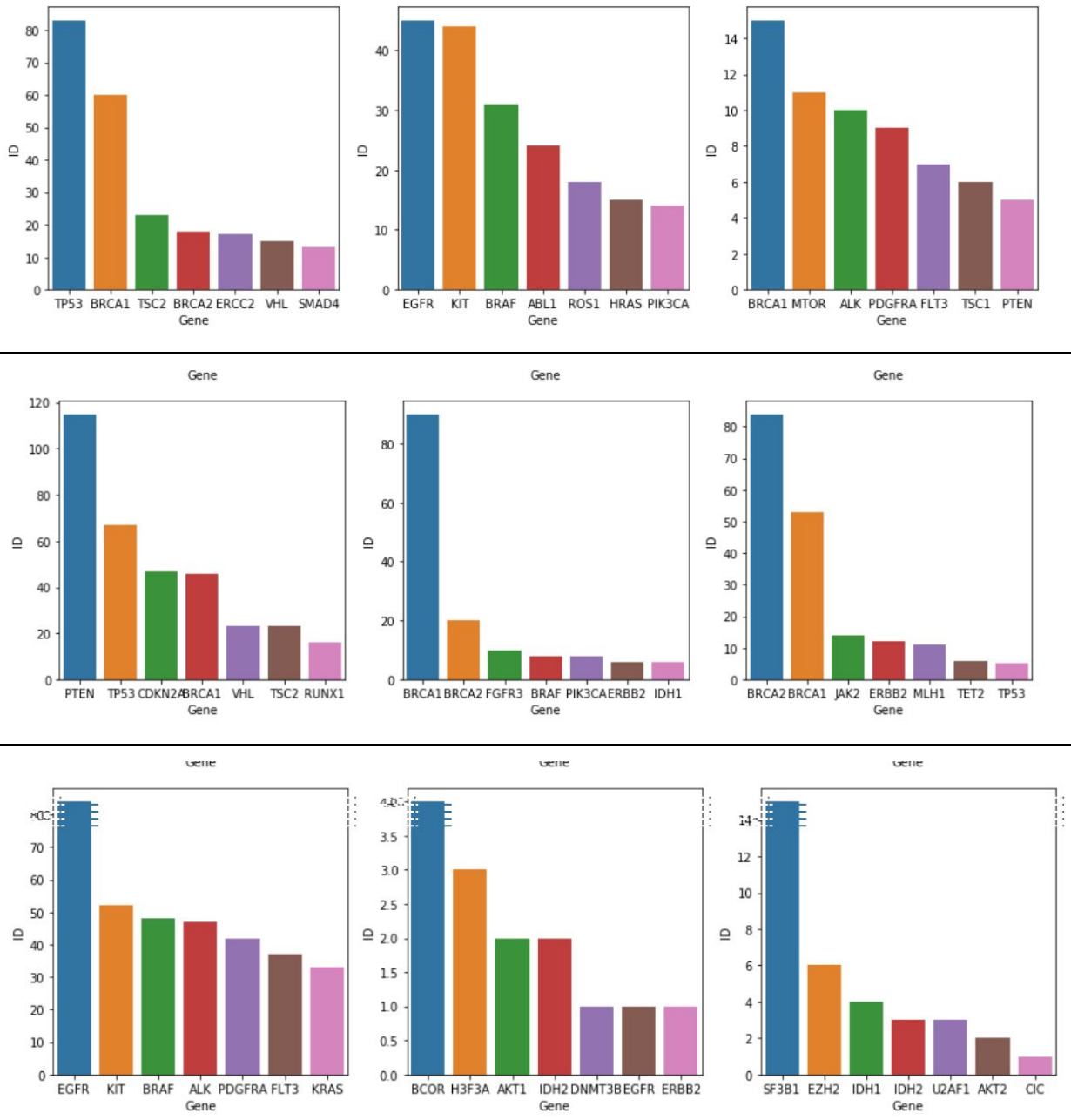
- ## Data Exploration:

We are provided with a dataset of genes and cancer-related mutations, a text file of scientific literature related to each mutation, and a 1-9 classification of each case, which had been hand-annotated by a scientist at MSK after reviewing the data

- **Gene data Summary:**

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

Genes with minimal occurences
 Gene
KLF4      1
FGF19     1
FANCC     1
FAM58A    1
PAK1      1
ERRFI1    1
PAX8      1
PIK3R3    1
PMS1      1
PPM1D     1
Name: Gene, dtype: int64
```

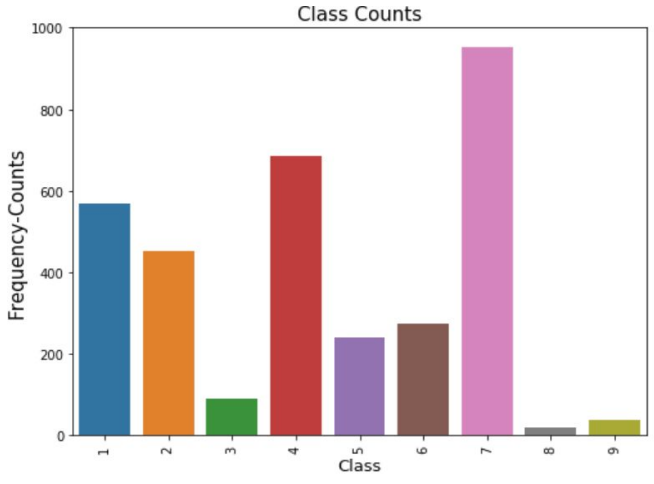


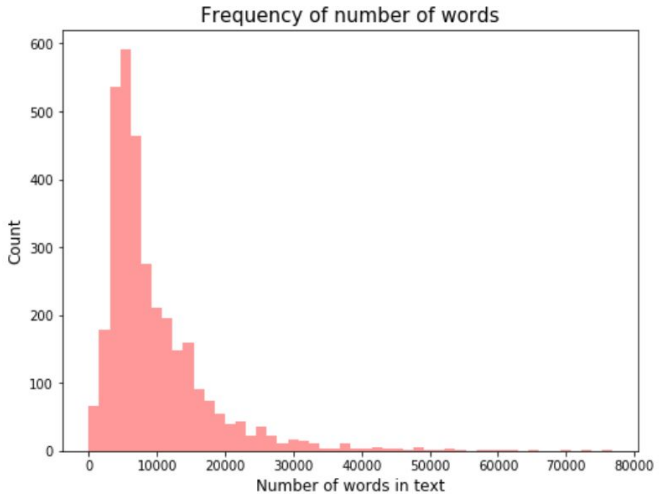
These 9 plots are for 9 different classes. First figures is corresponds to first class, Second figure corresponds to second class and respectively. From the plot, we can see that each class shares a different kind of relationship to the genes which is going to be quite useful in classification task.

Text data summary:

	ID	Gene	Variation	Class	Text	Text_count
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...	6089
7	7	CBL	Deletion	1	CBL is a negative regulator of activated recep...	14683
16	16	CBL	Truncating Mutations	1	To determine if residual cylindrical refractiv...	8118
37	37	DICER1	D1709E	1	Sex cord–stromal tumors and germ-cell tumors a...	2710
38	38	DICER1	D1709A	1	Sex cord–stromal tumors and germ-cell tumors a...	2710

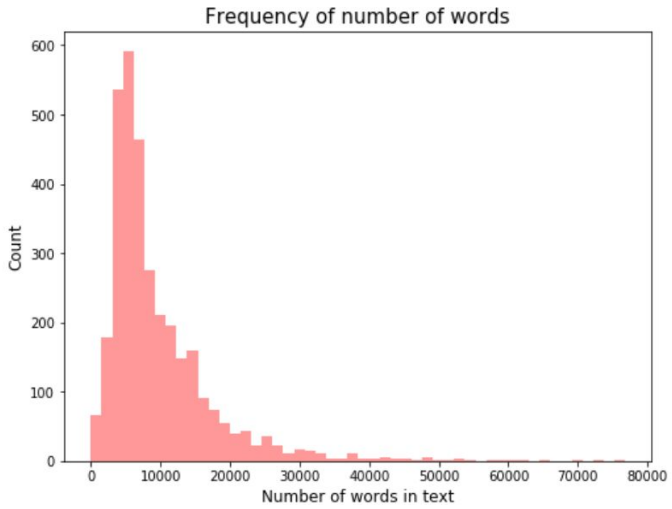
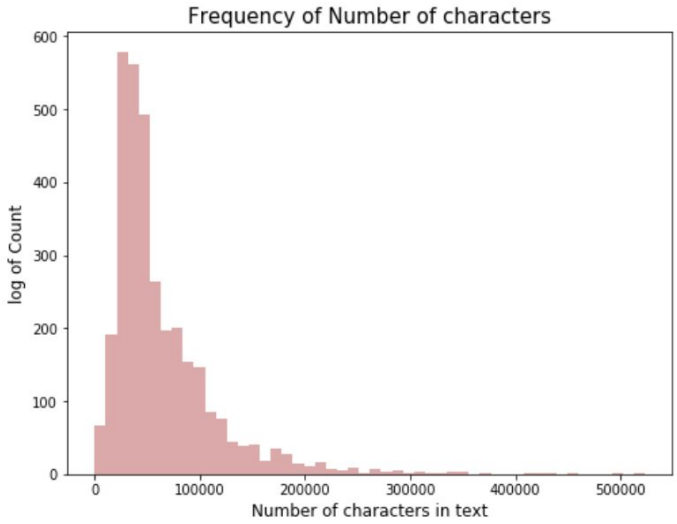
Class	count	mean	std	min	25%	50%	75%	max
1	566.0	9483.689046	6503.595573	183.0	4976.00	7317.0	12944.50	52972.0
2	452.0	9310.393805	7627.288722	116.0	4185.00	6810.0	12220.00	61957.0
3	89.0	6757.382022	3725.366918	1737.0	4283.00	5572.0	7415.00	27391.0
4	686.0	8983.390671	7280.220754	53.0	4566.00	6351.0	11537.00	43913.0
5	242.0	7517.049587	3902.868040	183.0	5245.00	6463.0	9513.50	24226.0





This figure shows the count of samples in each class. From above frequency histogram it is clear that there is a lot of variation between number of samples of different classes e.g. in class 7 there are 900+ samples whereas in class 8 there are even less than

This plot shows the number of words present in each sample. More than 80% of samples have less than 10000 words.Almost 15% of samples have >10000 words and almost 5% of samples have words close to 2000. So in case of calculation of words in a given

100 samples. So to handle this we are going to use oversampling so that each class have equal number of samples.	sample, we are going to divide by the length of words present in samples due to variation between the number of words present in the sample.
	
This plot shows the number of Characters presents in each sample. More than 80% of samples have less than 40000 to 70000 words.Almost 10% of samples have >100000 words and almost 7% of samples have Characters <40000.	This plot shows the variations of number of words present in each samples of different classes. According to this plot most of classes have samples with words varying from 0 to 30,000 And it is considering the outlier if a sample have greater than a certain threshold depends on the class. because there is very rare samples which have more than 20,000 words that's why it is considering them outlier in most of the classes.

- **Challenges in data acquisition and feature extraction**:

The data comes in 4 different files. Two csv files and two text files:

training/test variants: These are csv files which contains the details about mutations together with the target value / ground truth label. Feature variables are Gene, the specific gene where the mutation took place, and Variation and the nature of the mutation. On the other hand test data does not contain the target value. This is what we are going to predict through our models. These two files each are linked through an ID variable.

training/test text: These files contain the extensive clinical description of the evidence that was used to label the mutation classes of training/test variants.

File data samples is given below:

	ID	Text
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2
3	3	CBL	N454D	3
4	4	CBL	L399V	4

These are the steps which are used for transforming the data suitable for machine learning model:

Step1- Merge the data of these files (training variants with training text and test variants with test text) on the basis of their ids.

Data after Merging two files:

	ID	Gene	Variation	Class	Text
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	CBL	W802*	2	Abstract Background Non-small cell lung canc...
2	2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...
3	3	CBL	N454D	3	Recent evidence has demonstrated that acquired...
4	4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...

Step2- Remove the samples which does not contain text data. As there are a lot of the words in the text data of the sample and most of the samples contain more than 5000 words. So we are also removing the samples which contains less than 100 words as the information in them will provide very less evidence as compared to other samples.

Step3- As the files contains text data which can not be used for feeding in machine learning model, so we need to convert the text data into numerical form of data. For this, we are using different kind of models like tfidf, CountVectorizer, doc2vec etc. Description of all of these are as follows:

- **CountVectorizer**:- Convert a collection of text documents to a matrix of token counts. This implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix.
- **tf-idf matrix**:- For each word calculate term frequency (tf) and inverse document frequency(idf).Now take log of both tf and idf and multiply them for calculating tf idf.
    - tf:- the number of times a word comes into sample and divide it by the number of words in samples.
    - idf:- the number of samples in the dataset divided by the number of times a word comes into total samples.
- **doc2vec**:-Doc2Vec is used to compute a feature vector for each sample in the dataset. Doc2Vec is based on the concept that the sample representation should be good enough to predict the word present in particular sample or not. As all the classes might have different kinds of words in them, so it could help in differentiating them.
- Logistic regression and random forest classifier are used as estimator to determine the evaluation metrics like accuracy,precision,recall and f-score using k-fold cross validation to determine the probability of data points belonging to different classes.
- Naive Bayes is used for N grams probability where N lies between 1 to 3 an accuracy is calculated for data before preprocessing and after preprocessing.

- **Hypothesis**:

It is something that we guess by visualizing the properties of data, model or features given . It can proved to be true and false by implementing the algorithms.

Hypothesis1:

If the samples contain more stopwords then there are more chances that the prediction of the test data would be class 6 because it contains very large proportions of stop _words as can be visible from the below graph.

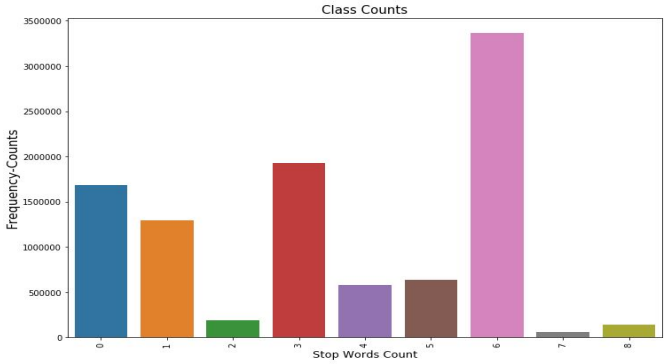
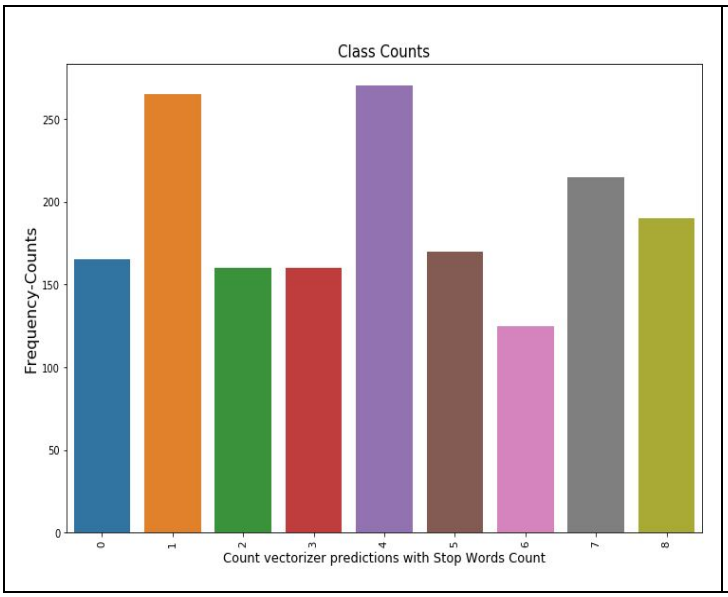
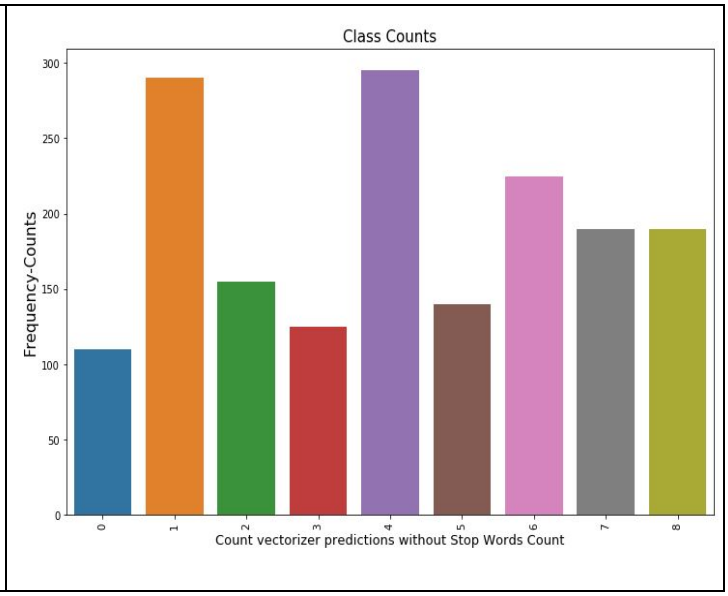


Fig: Above histogram shows the frequency distribution of number of stop words among classes.

Proof :

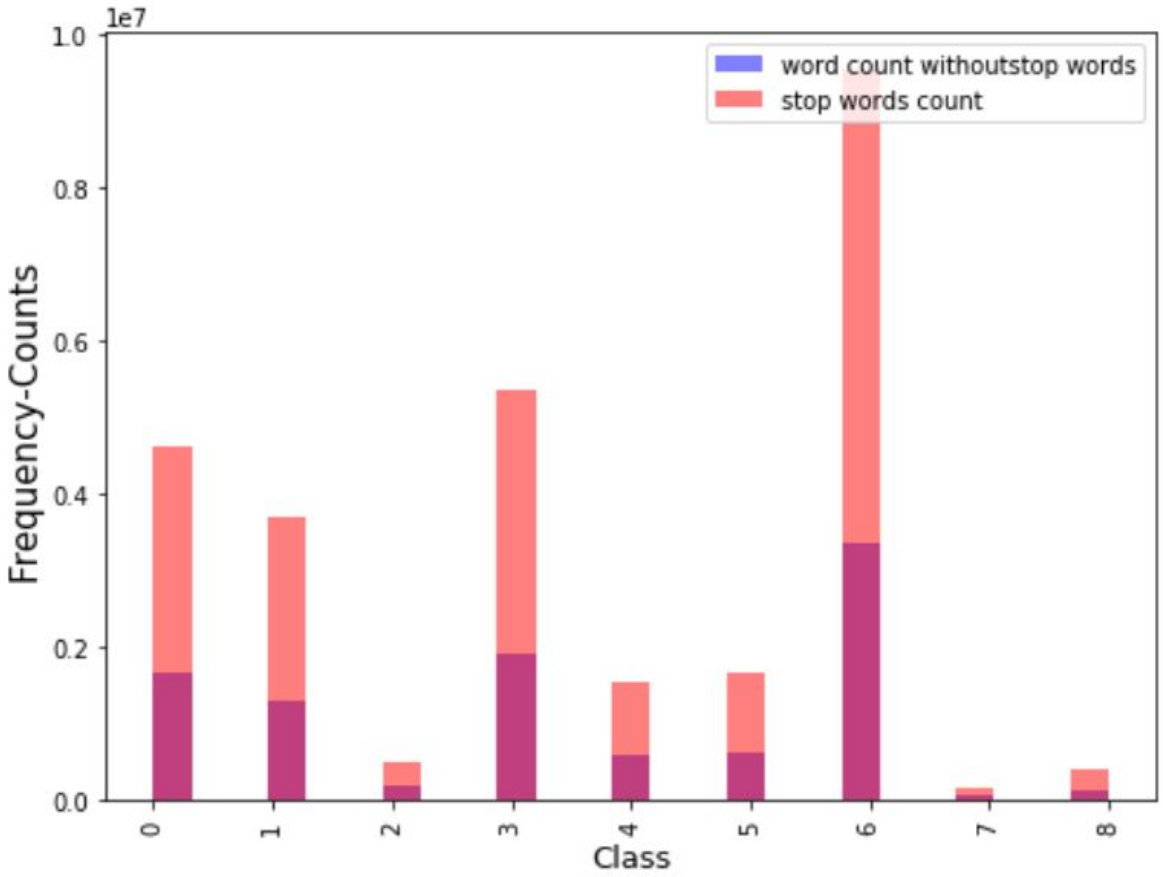
We are going to use countvectorizer models for predicting the classes for test data with or without using stop words.





Observation:

From the above plot, It is clear that there is not much difference in prediction with or without stop words. Hence there is no support for our hypothesis that stop word removal is effective. Our hypothesis is rejected .To support it ,we are now comparing number of stop words in class with a number of other words in a class.



From above histogram comparison it is clear that the ratio of number of stop words to other words is the same for each class. Hence there is no effect of removing stop_words. So, We are ***rejecting*** our hypothesis1.

Hypothesis 2:

As the number of samples in each class is not evenly distributed so we have to use stratified split otherwise it will going to predict the class with more samples most of the time.

Proof:

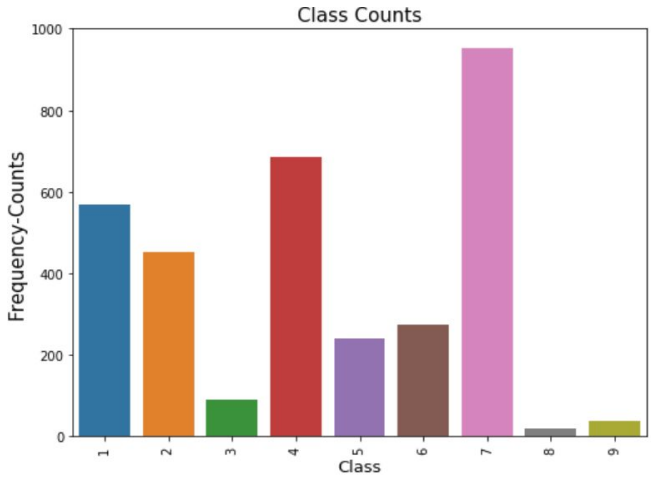
	Samples split by train test without stratify	**Expected count after stratifying**
Class 1	**491**	**474**
Class 2	**359**	**381**
Class 3	**69**	**91**
Class 4	**552**	**568**
Class 5	**217**	**213**
Class 6	**211**	**240**
Class 7	**759**	**782**
Class 8	**9**	**35**
Class 9	**25**	**49**

Observations:

So, from our results, we could see that the classes which have more samples for them will not affect results much. But for classes like class 8 there are very less samples coming after split than expected. Which will create problem while predictive modelling and can give false results.
Hence, we are **accepting** the **hypothesis 2.**

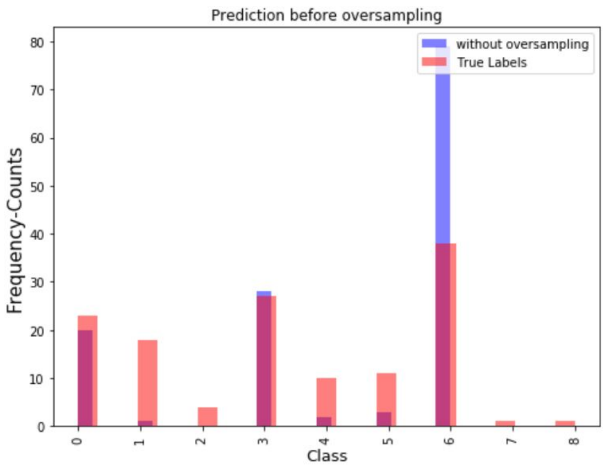
Hypothesis 3:

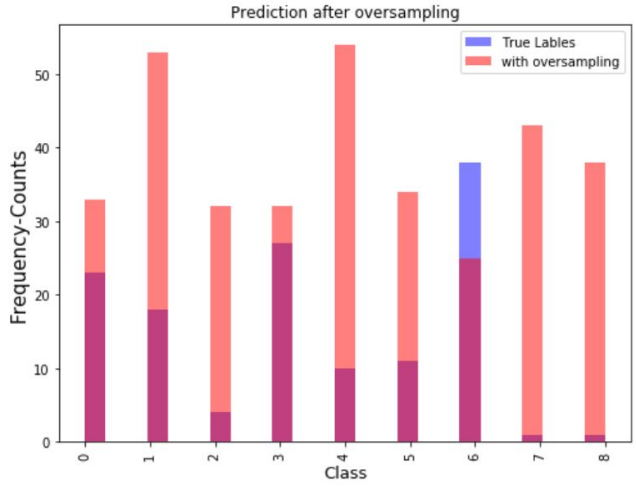
As the number of samples in each class is not evenly distributed to we need to do oversampling or undersampling otherwise most dominating class will cause discrepancy.



Proof:

To prove hypothesis, we have applied techniques like tf_idf vectorizer and count_vectorizer with sampling and without sampling and found how much accurate is our prediction for each class.





Observation:

As observed from the figure, prediction before oversampling is dominating towards class 6. Most of the labels are predicted from class 6 while after oversampling, it has reduced and samples are labelled equally in all classes. Hence we can say that our hypothesis is true .

Predictive Modelling:

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results.

Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software.

In our case , we have used tf_idf and simple count of tokens from clinical text given in data as features or predictors. Doc2Vec model from gensim is also used as a predictor to determine the features.

Then Logistic Regression,Naive Bayes and Random Forest models are used to determine the prediction based on predictors formed from above mentioned techniques.

- ## Results:

Accuracy of Naive Bayes before and after preprocessing

N grams	Before Preprocessing	After Preprocessing
Uni gram	5.4	15.5
Bi gram	15.7	22.4
Tri gram	18.3	25.6

Result of Random Forest on different feature selection method

Metric	CountVectorizer	TfIdfVectorizer	DOC2VEC
Accuracy	50	50.2	48.8
kappa	0.41	0.402	0.386
f1-score	0.507	0.506	0.493
precision	0.53	0.534	0.531

recall	0.508	0.502	0.488

Result of Logistic regression on different feature selection method

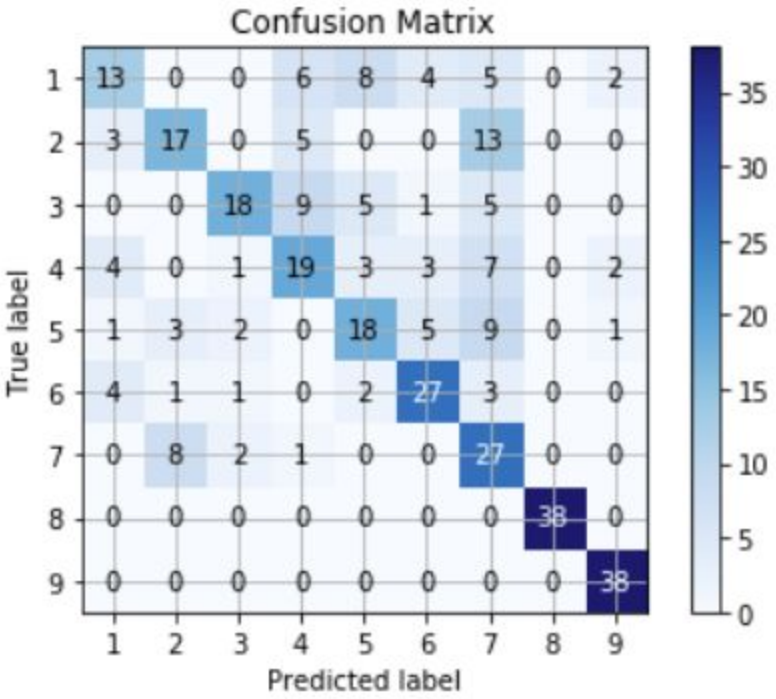
Model	CountVectorizer	TfIdfVectorizer	DOC2VEC
Accuracy	49.2	52.2	54.4
kappa	0.348	0.378	0.398
f1-score	0.455	0.469	0.479
precision	0.472	0.549	0.610
recall	0.492	0.522	0.552

After RandomSampling:

Result of Random Forest Classifier on different feature selection methods:

Metric	CountVectorizer	TfIdfVectorizer	DOC2VEC
Accuracy	67.7	61.3	62.5
kappa	0.636	0.56	0.578
f1-score	0.677	0.62	0.621
precision	0.709	0.668	0.641
recall	0.677	0.613	0.625

Confusion Matrix plot for Doc2Vec :



- ## Conclusion:

This report addressed the problem of classification of different kinds of cancer such that the task of personalised medicine can be improved . Naive bayes, Logistic regression, Random Forest and some text preprocessing techniques are used for the problem, where logistic regression has better performance than other techniques. This model is quite useful in case if you have gene mutation and to find out which class it belongs to as it is very difficult to classify with that much size of text data. Proposed techniques achieve an accuracy of 50.2% precision and 0.54 f1 score before sampling and 67.7% accuracy and 0.701 precision after sampling.

- ## GitHub Link:

https://github.com/divya1garg/data_science_project95_43.git

- ## References:

(1) Scikit-learn: Machine Learning in Python Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.