# REDEFINING CANCER TREATMENT

# (Major Project Report)

Bachelor of Technology

In

(Computer Engineering)



JAMIA MILLIA ISLAMIA

Under The Supervision of Submitted By Dr. Waseem Ahmed Yousuf Ansari

16BCS010

Md Ikrar khan

16BCS044

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY

JAMIA MILLIA ISLAMIA, NEW DELHI - 110025

CERTIFICATE

This is to certify that the project report entitled “REDEFINING CANCER TREATMENT”, is an authentic work carried out by Yousuf Ansari and Md Ikrar Khan.

The work is submitted in complete fulfilment of the requirement for the award of the degree of Bachelor of Technology in Computer Engineering under my guidance. The matter embodied in this project work has not been submitted earlier for the award of any degree to the best of my knowledge and belief.

Place: New Delhi

Date: 22th June, 2020

Dr. Tanveer Ahmad Dr. Waseem Ahmed

Professor & Head,​​ Assistant Professor,

Dept. of Computer Engg. Dept. of Computer Engg.

Jamia Millia Islamia Jamia Millia Islamia

New Delhi New Delhi

ACKNOWLEDGMENT

We would like to thank our supervisor, Dr. Waseem Ahmed (Dept. of Computer Engineering) for giving us the opportunity to take up this project. We thank him for his invaluable guidance and overwhelming moral support. We are also very grateful to our HOD, Prof. Tanvir Ahmad for his valuable support throughout the project. We also thank our entire faculty for their teaching, guidance and support.

We are also thankful to our friends for their suggestions and support when required. We regret any inadvertent omissions.

Md Ikrar Khan Yousuf Ansari

16BCS044 16BCS010

Dept.of Computer Engg. Dept.of Computer Engg,

F/O Engg. and Tech F/O Engg. and Tech.

Jamia Millia Islamia Jamia Millia Islamia

New Delhi New Delhi

ABSTRACT

In the Medical field there are lots of diseases which are very dangerous for human beings because their treatment is very difficult, Cancer is also one of them.

Cancer is a disease which occurs due to genetic mutation.

Since cancer has more than hundreds types so identification of type of cancer is difficult and time consuming. That is why cancer is a dangerous disease all over the world.

Increasingly, clinical molecular laboratories are detecting novel sequence variants in the course of testing patient specimens for a rapidly increasing number of genes associated with genetic disorders. While some phenotypes are associated with a single gene, many are associated with multiple genes.

A model that defines the type of cancer on the basis of available data and predicts which type of cancer might be. To design this model, we use the Machine Learning Model because the model should learn from past results and data.

Our understanding of the clinical significance of any given sequence variant falls along a gradient, ranging from those in which the variant is almost certainly pathogenic for a disorder to those that are almost certainly benign.

Table of Content

Pages

Certificate……………………………………………………………………………………...2

Acknowledgment………………………………………………………………………………….3

Abstract………………………………………………………………………………………………….4

Contents

1: Introduction ……………………………………………………….…………………....6

1.1 About The Project ………………………………………….………..6

1.2 Challenges ……………………………………………………………….7

1.3 Need For REDEFINING CANCER TREATMENT….…9

1.4 Tools Used …………………………………………………………….12

2: Implementation …………………………………………………………………....14

2.1 Introduction ……………………………………….....................14

2.2 Basic Algorithm …………………………………………………….14

3: Experimental results ……………………………………………. …………….. 20

3.1 Dataset ………………………………………........................... 20

3.2 Results ………………………………………............................ 31

4: Conclusion and Future Work……….………………....................... 33

5: References …………………………………………….……….......................34

1.INTRODUCTION

Increasingly, clinical molecular laboratories are detecting novel sequence variants in the course of testing patient specimens for a rapidly increasing number of genes associated with genetic disorders. While some phenotypes are associated with a single gene, many are associated with multiple genes.

Our understanding of the clinical significance of any given sequence variant falls along a gradient, ranging from those in which the variant is almost certainly pathogenic for a disorder to those that are almost certainly benign.

While the previous ACMG recommendations provided interpretative categories of sequence variants and an algorithm for interpretation, the recommendations did not provide defined terms or detailed variant classification guidance.

This report describes updated standards and guidelines for classification of sequence variants using criteria informed by expert opinion and empirical data.

1.1 About Project

The American College of Medical Genetics and Genomics (ACMG) previously developed guidance for the interpretation of sequence variants. In the past decade, sequencing technology has evolved rapidly with the advent of high-throughput next generation sequencing.

By adopting and leveraging next generation sequencing, clinical laboratories are now performing an ever increasing catalogue of genetic testing spanning genotyping, single genes, gene panels, exomes, genomes, transcriptomes and epigenetic assays for genetic disorders. By virtue of increased complexity, this paradigm shift in genetic testing has been accompanied by new challenges in sequence interpretation.

In this context, the ACMG convened a workgroup in 2013 composed of representatives from the ACMG, the Association for Molecular Pathology (AMP) and the College of American Pathologists (CAP) to revisit and revise the standards and guidelines for the interpretation of sequence variants.

The group consisted of clinical laboratory directors and clinicians. This report represents the expert opinion of the workgroup with input from ACMG, AMP and CAP stakeholders.

These recommendations primarily apply to the breadth of genetic tests used in clinical laboratories including genotyping, single genes, panels, exomes and genomes. This report recommends the use of specific standard terminology: ‘pathogenic’, ‘likely pathogenic’., ‘uncertain significance’, ‘likely benign’, and ‘benign’ to describe variants identified in Mendelian disorders.

1.2 Challenges

A lot has been said during the past 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.

But the challenge is distinguishing the mutations that contribute to tumor growth from the neutral 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.

We will develop a Machine Learning algorithm that, using this knowledge base as a baseline, automatically classifies genetic variations.

Understanding the genetic mutations that really matter in a cancer tumor is a really challenging task with a potential huge impact on millions of lives.

The given fig (1.1) shows genetic mutation in genes which causes fatal diseases like cancer.

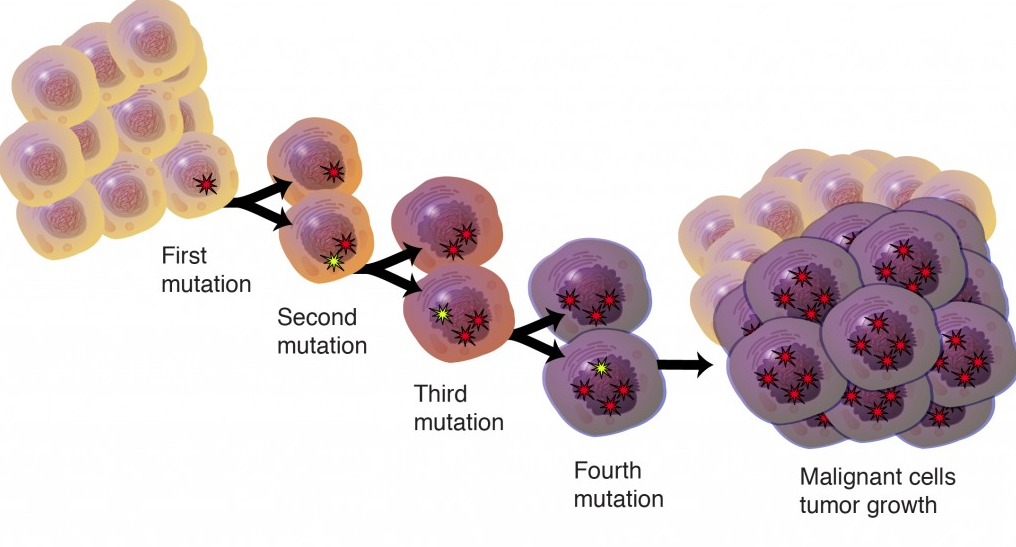


Fig. 1.1: Simple Diagram of Mutation

1.3 Need for REDEFINING CANCER TREATMENT

Many cancer treatments are available. Your treatment options will depend on several factors, such as the type and stage of your cancer, your general health, and your preferences. Together you and your doctor can weigh the benefits and risks of each cancer treatment to determine which is best for you.

Cancer treatment options include:

1.3.1 Surgery:-

The goal of surgery is to remove the cancer or as much of the cancer as possible.

1.3.2 Chemotherapy:-

Chemotherapy uses drugs to kill cancer cells.

1.3.3 Radiation therapy :-

Radiation therapy uses high-powered energy beams, such as X-rays or protons, to kill cancer cells. Radiation treatment can come from a machine outside your body (external beam radiation), or it can be placed inside your body (brachytherapy).

1.3.4 Bone marrow transplant :-

Your bone marrow is the material inside your bones that makes blood cells from blood stem cells. A bone marrow transplant, also known as a stem cell transplant, can use your own bone marrow stem cells or those from a donor.  
A bone marrow transplant allows your doctor to use higher doses of chemotherapy to treat your cancer. It may also be used to replace diseased bone marrow.

1.3.5 Immunotherapy:-

Immunotherapy, also known as biological therapy, uses your body's immune system to fight cancer. Cancer can survive unchecked in your body because your immune system doesn't recognize it as an intruder. Immunotherapy can help your immune system "see" the cancer and attack it.

1.3.6 Hormone therapy :-

Some types of cancer are fueled by your body's hormones. Examples include breast cancer and prostate cancer. Removing those hormones from the body or blocking their effects may cause the cancer cells to stop growing.

1.3.7 Targeted drug therapy :-

Targeted drug treatment focuses on specific abnormalities within cancer cells that allow them to survive.

1.3.8 Cryoablation :-

This treatment kills cancer cells with colds. During cryoablation, a thin, wand-like needle (cryoprobe) is inserted through your skin and directly into the cancerous tumor. A gas is pumped into the cryoprobe in order to freeze the tissue. Then the tissue is allowed to thaw. The freezing and thawing process is repeated several times during the same treatment session in order to kill the cancer cells.

1.3.9 Radiofrequency ablation :-

This treatment uses electrical energy to heat cancer cells, causing them to die. During radiofrequency ablation, a doctor guides a thin needle through the skin or through an incision and into the cancer tissue. High-frequency energy passes through the needle and causes the surrounding tissue to heat up, killing the nearby cells.

1.3.10 Clinical trials :-

Clinical trials are studies to investigate new ways of treating cancer. Thousands of cancer clinical trials are underway.

Other treatments may be available to you, depending on your type of cancer. In these types of Cancer Treatment there is a lot of Manual work which takes a lot of time. Due to this time consumption, the number of deaths increases. To solve this problem we need a new treatment method which takes less time to identify cancer( type of cancer also ) .

We need to design a model which learns from past data and identifies the type of cancer in a patient .

1.4 Tools Used

1.4.1 The Jupyter Notebook :

Jupyter notebook is a web-based interactive development environment for Jupyter notebooks, code and data.

JupyterLab is a flexible configure and arrange the user interface to support a wide range of workflows in data science, scientific computing and machine learning. JupyterLab is extensible and modular:write plugins that add new components and integrate with existing ones. Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code,equations, visualizations and narrative text.

Uses include: data cleaning and transformation, numerical simulation, statistical, modeling, data visualization, machine learning and much more.

1.4.2 PANDA :

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

You should have a basic understanding of Computer Programming terminologies. A basic understanding of any of the programming languages is a plus. Pandas library uses most of the functionalities of NumPy.

1.4.3 NUMPY:

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided using random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. We will be using Logistic regression from this library.

1.4.4 Scikit-learn:

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

2. Implementation

2.1 Introduction:-

We need to develop a Machine Learning algorithm that, using this knowledge base as a baseline, automatically classifies genetic variations.

2.2 Basic Algorithm

2.2.1 Naive Bayes:-

It is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. Bayes’ theorem is stated mathematically as the following equation.

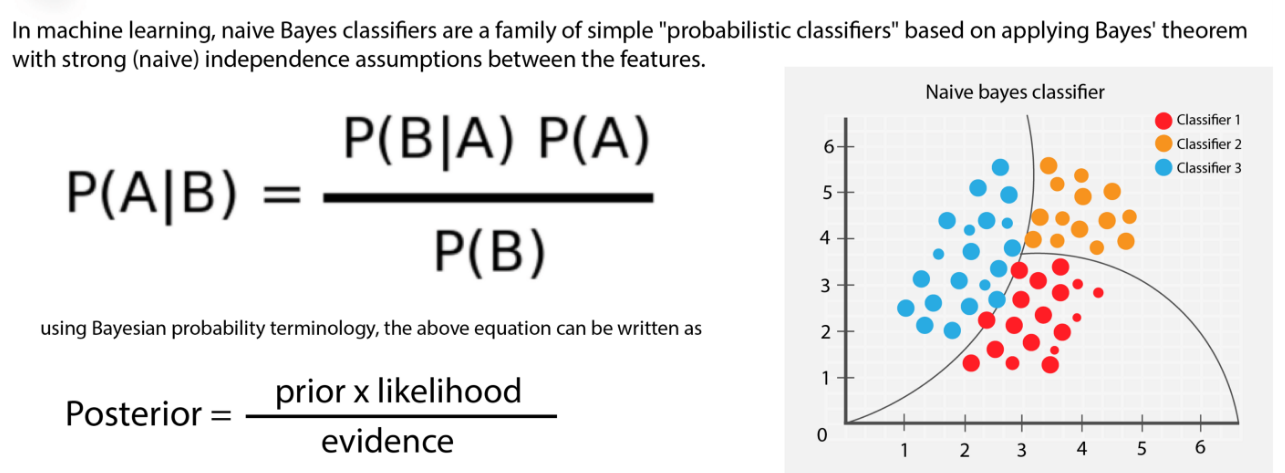


Fig (2.1) Naive Bayes Classifiers

where A and B are events and P(B) = 0.

* Basically, we are trying to find the probability of event A, currently this interpretation of genetic mutations is being done manually.
* We will have given the event B is true. Event B is also termed as evidence.
* P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
* P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

2.2.2 KNN Classifier:

KNN is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

The K-Nearest Neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph. There are other ways of calculating distance, and one way might be preferable depending on the problem we are solving. However, the straight-line distance (also called the Euclidean distance) is a popular and familiar choice.

It is widely disposable in real-life scenarios since it is non-parametric, meaning it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

2.2.3 Support Vector Machine :

In machine learning, Support vector machines(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features), with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that best differentiates the two classes.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces.

2.2.4 Random Forest Classifier:

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn’t depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform. Sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

2.2.5 Stacking model:

If you have ever competed in a Kaggle competition, you are probably familiar with the use of combining different predictive models for improved accuracy which will creep your score up in the leaderboard. While it is widely used, there are only a few resources that I am aware of where a clear description is available (One that I know of is [here](http://mlwave.com/kaggle-ensembling-guide/), and there is also a [caret package extension](https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html) for it). Therefore, I will try to workout a simple example here to illustrate how different models can be combined. The example I have chosen is the [House Prices](https://www.kaggle.com/c/house-prices-advanced-regression-techniques) competition from Kaggle. This is a regression problem and given lots of features about houses, one is expected to predict their prices on a test set. I will use three different regression methods to create predictions (XGBoost, Neural Networks, and Support Vector Regression) and stack them up to produce a final prediction. I assume that the reader is familiar with R, Xgboost and caret packages, as well as support vector regression and neural networks.

The main idea of co-constructing a predictive model by combining different models can be schematically illustrated as below:

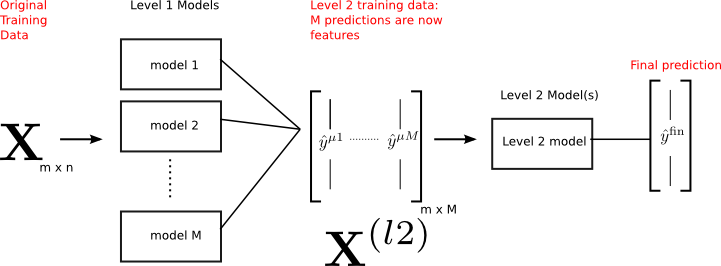


Fig 2.2 , Stacking Model

* Initial training data (X) has *m* observations, and *n* features (so it is *m x n*).
* There are M different models that are trained on X (by some method of training, like cross-validation) beforehand.
* Each model provides predictions for the outcome (y) which are then cast into a second level training data (Xl2) which is now *m x M*. Namely, the M predictions become features for this second level data.
* A second level model (or models) can then be trained on this data to produce the final outcomes which will be used for predictions.

There are several ways that the second level data (Xl2) can be built. Here, I will discuss stacking, which works great for small or medium size data sets. Stacking uses a similar idea to k-folds cross validation to create out-of-sample predictions.

2.2.6 Maximum Voting Classifier :

We use VotingClassifier() to combine Logistic Regression , Linear SVM, and Naive Bayes together as one model.

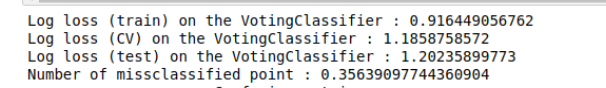
A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

It simply aggregates the findings of each classifier passed into the Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each of them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

2.3 Voting Classifier supports two types of voting:-

1.Hard Voting

2.Soft Voting



3 . Dataset And Results

3.1 Dataset :

Input dataset:-

Data Field - There are three categories of data ;

Gene ,Variation and Text Data

3.1.1 Gene :-

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA. The Human Genome Project estimated that humans have between 20,000 and 25,000 genes. The most commonly mutated gene in people with cancer is p53 or TP53. More than 50% of cancers involve a missing or damaged p53 gene. Most p53 gene mutations are acquired. Germline p53 mutations are rare, but patients who carry them are at a higher risk of developing many different types of cancer.

3.1.2 Variation :-

The genetic variation is how one individual of a species variates from another, variation can be due to changes in the nucleotide sequence like insertions, deletions, any genetic rearrangements or any environmental factors.Variations are seen in groups or populations of an individual.

3.1.3 Text Data :-

There is lots of data available in the text form from past years treatment.

3.1.4 Data Property :

● Training Samples ~ 3k

● Unique Genes : 264

● Unique variations ~ 3k

● Test Samples ~370

● Multiclass classification (1 -9)

● The Peak is around 4000 words

One of the biggest challenges of this Dataset is to process these high -dimensional word

features into something meaningful and try to reduce overfitting because of few number of

training samples

# 3.1.5 Data file’s information:

3.1.5.1 training\_variants (ID , Gene, Variations, Class)

3.1.5.2 training\_text (ID, Text)

3.1.6 Performance Metric(s):

3.1.6.1 Multi class log-loss

Log Loss takes into account the uncertainty of your prediction based on how much it varies from the actual label. This gives us a more nuanced view into the performance of our model*.*

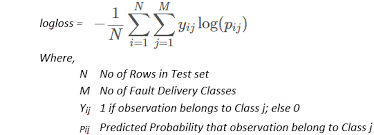
**

Fig 3.1, Multi class log-loss

The graph below shows the range of possible log loss values given a true observation (isDog = 1). As the predicted probability approaches 1, log loss slowly decreases. As the predicted probability decreases, however, the log loss increases rapidly. Log loss penalizes both types of errors, but especially those predictions that are confident and wrong!

The given fig (3.1) shows formula for Multi class log-loss.

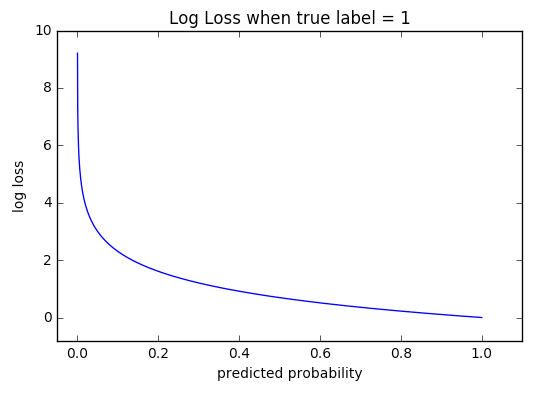
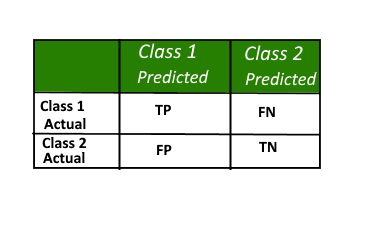


Fig 3.2, Graph of log-loss

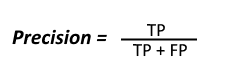
3.1.7 Confusion matrix:

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

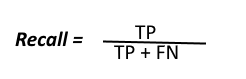
Table 3.1, Table of Confusion Matrix



Precision:



Recall:



# 3.1.8 One Hot Encoding :

# It refers to splitting the column which contains numerical *categorical data* to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.

# For example :

# Consider the data where fruits and their corresponding categorical value and prices are given.

Table 3.2, Dataset of Table

| FRUIT | CATEGORICAL VALUE OF FRUIT | PRICE |
| --- | --- | --- |
| Apple | 1 | 5 |
| Mango | 2 | 10 |
| Apple | 1 | 15 |
| Orange | 3 | 20 |

# The output after one hot encoding the data is given as follows,

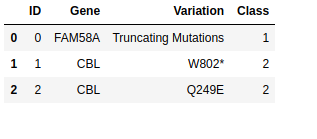
Table 3.3, Table of Output

| APPLE | MANGO | ORANGE | PRICE |
| --- | --- | --- | --- |
| 1 | 0 | 0 | 5 |
| 0 | 1 | 0 | 10 |
| 1 | 0 | 0 | 15 |
| 0 | 0 | 1 | 20 |

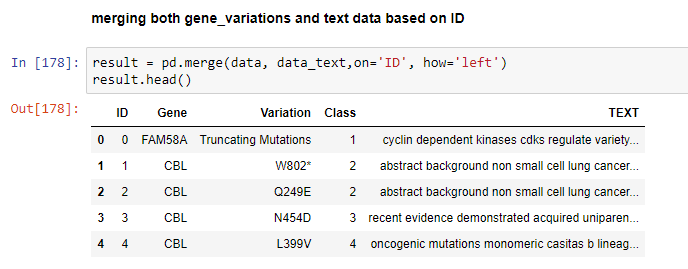
3.1.9 Reading the Data from files

After Removing Stop Words from the text fields, We combine the two training and text variants file to single data frame to work on, the header of the data frame looks like:-

Table 3.4: Table for data frame



# Table 3.5, Table of merged\_data



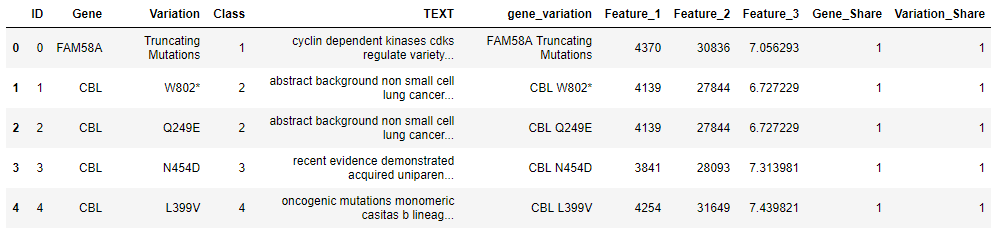
# 

# 3.1.10 Feature Engineering

Following are The Features of the data frame after feature engineering:-

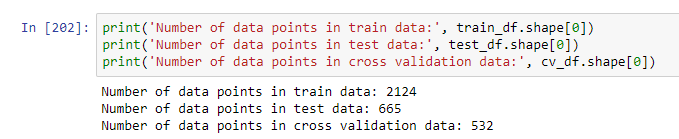
1. Gene.
2. Variation.
3. TEXT.
4. gene\_vartiation:- combination of gene and variation.
5. Feature\_1:- Number of Words in Text.
6. Feature\_2:-Number of Characters in Text.
7. Feature\_3:-Average length of words used in statement.
8. Gene\_Share:- Number of Gene per data Point.
9. Variation\_Share:-Number of Variation per data Point.

Table 3.6: Feature Table

**

Features in the Data Frame

# 3.1.11 Train, Test, and Cross-Validation Split



# Fig 3.3: Splitted Data

# 3.1.12 Encoding Feature

# We Use One\_hot Encoding for Gene , Variation And gene\_variation features whereas we use er for text encoding.

# 3.1.13 Univariate Analysis

## **3.1.13.1 Gene feature**

By just using the Gene feature and applying Logistic Regression on it, the best alpha we get is 0.001 .

Train Error: 1.041

Test Error: 1.179

CV Error: 1.208

3.1.13.2 Variation feature

By using just variation features and applying Logistic Regression on it, the best alpha we get is 0.001.

best alpha = 0.001 The train log loss: 0.617

best alpha = 0.001 The cv log loss is: 1.100

best alpha = 0.001 The test log loss is: 1.067

## **3.1.13.3 Gene\_Variation feature**

By combining just gene and variation Feature and applying Logistic Regression on it, the best alpha we get is 0.001.

best alpha = 0.001 The train log loss: 0.521

best alpha = 0.001 The cv log loss is: 1.102

best alpha = 0.001 The test log loss is: 1.135

## 

## 

## **3.1.13.4 text feature**

By combining just gene and variation Feature and applying Logistic Regression on it, the best alpha we get is 0.001 .

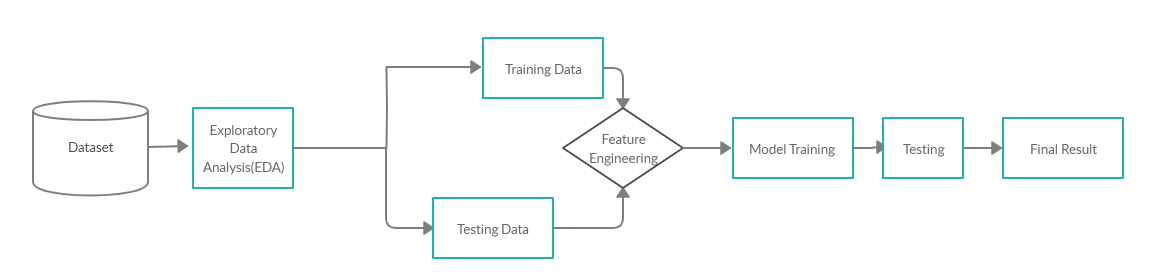
best alpha = 0.0001 The train log loss: 0.7467683268790596

best alpha = 0.0001 The cv log loss is: 1.0964906815887412

best alpha = 0.0001 The test log loss is: 1.1013255697297029

# 

FlowChart :



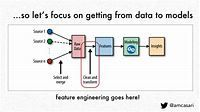
3.1.14 Exploring Data Analysis:

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, [such as R](https://searchbusinessanalytics.techtarget.com/definition/R-programming-language).

Training data :-80% of data is kept separate to train our model Testing Data :-20 % of data is kept separate to test our model

3.1.15 Feature Engineering :

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself.



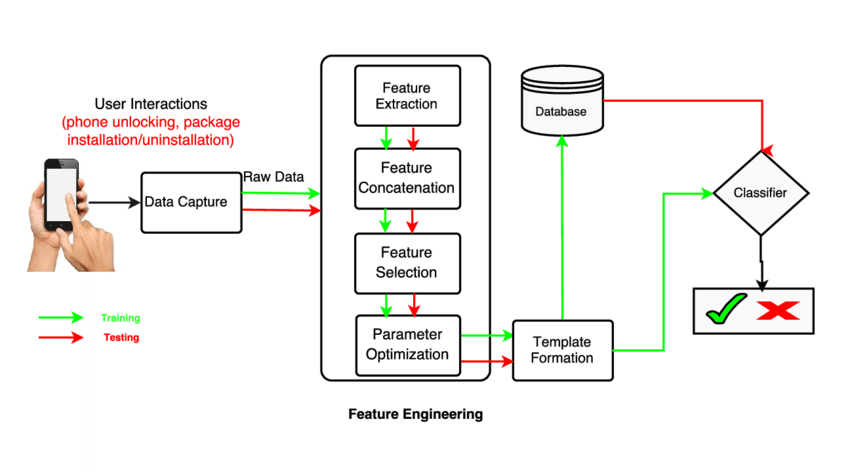


Fig 3.4: Feature Engineering block diagram

3.1.16 Model Training:-

Training sets are used to fit and tune your models. Test sets are put aside as "unseen" data to evaluate your models.

We should always split our data before doing anything else.

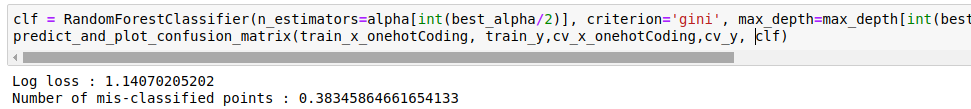
This is the best way to get reliable estimates of your model’s

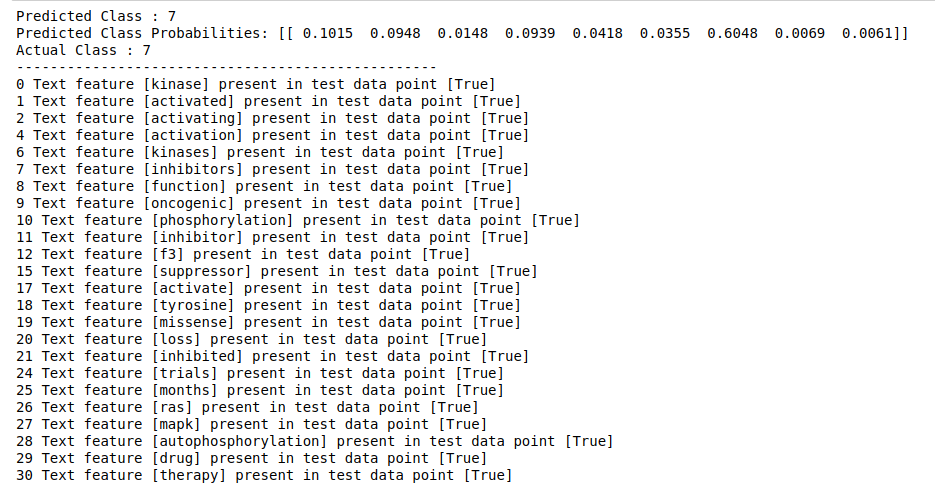
Performance. After splitting our data, we do not touch our test set until we are ready to choose our final model.

Model Training:- after Train our Dataset with separate dataset(80%) Testing of Model is done with test data (20 % of our dataset).

Comparing test vs training performance allows us to avoid overfitting if the model performs very well on the training data but poorly on the test data, then it’s overfit.

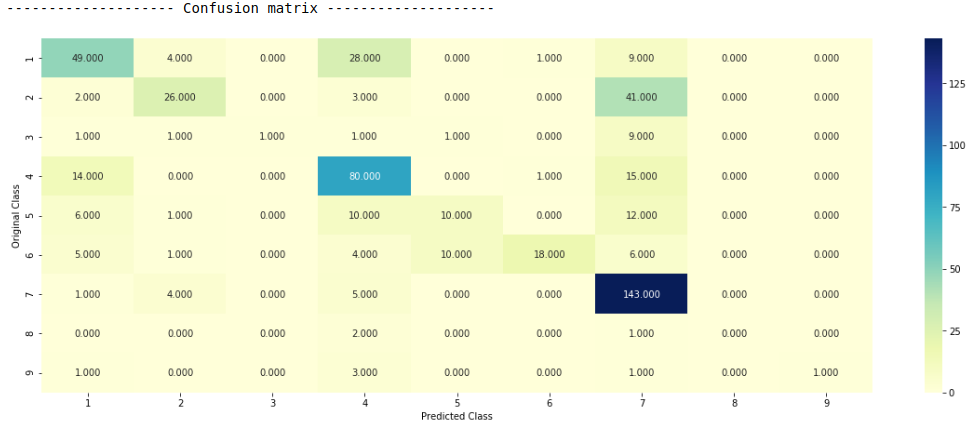
3.2 Result



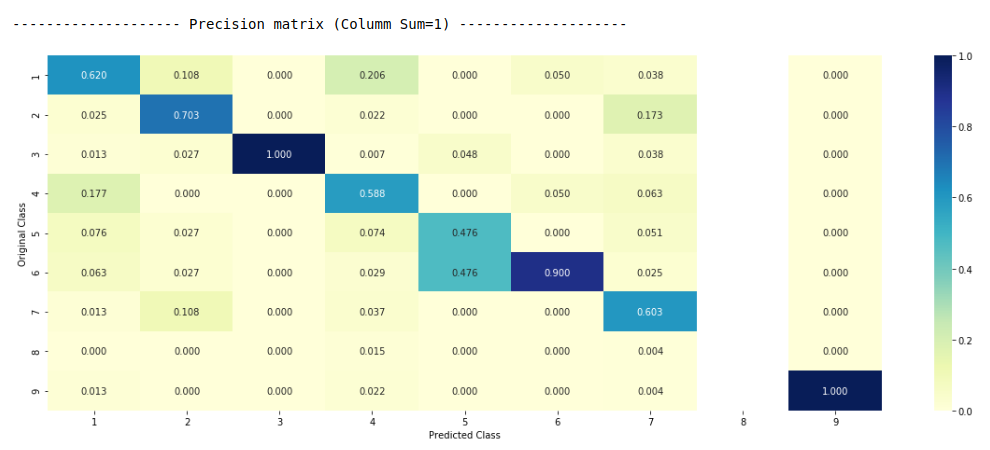


Random forest Model Predicted class And actual class are Same.

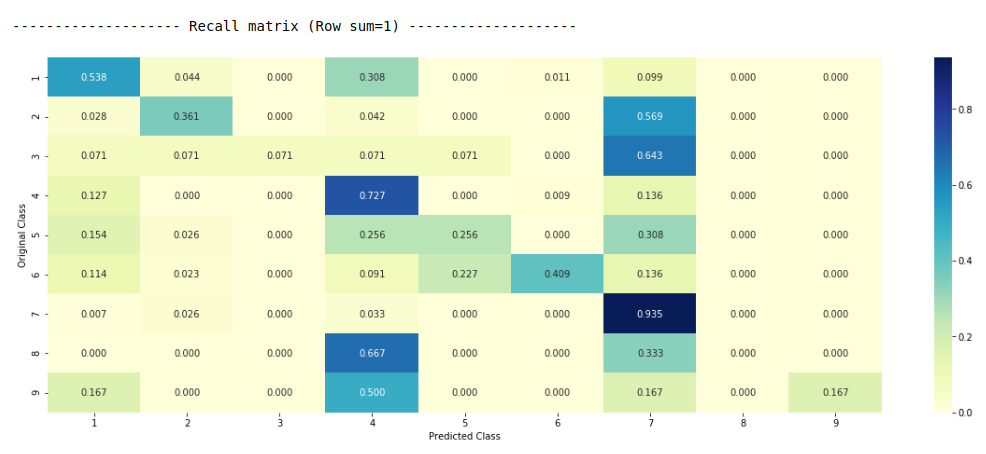
3.2.1 Confusion Matrix of random Forest Model



3.2.2 Precision Matrix of random Forest Model



3.2.3 Recall Matrix of random Forest Model



4. Conclusion and Future Work

4.1 Conclusion

1. Performed Exploratory data analysis on train and test datasets.

2. Performed checks for missing values and duplicate data.

3. Observed various plots of density, distribution of mean, skewness, kurtosis and both train and test .

4. Observed if there was any correlation between the features

Performed feature engineering and created features.

5. Created a random model using Log-Loss to compare other models.

6. Random Forest performed better than KNN model.

7. Applied machine learning model Maximum Voting Classifier machine on these features and observed accuracy of 68 percent.

8. To further improve the model’s score, use a unique value count in a row to identify synthetic samples and differentiated real and synthetic samples. Created magic features in both train and test sets using this data.

9. As a data augmentation step, performed data oversampling as the data is imbalanced.

10. Applied machine learning model Random Forest and Multi

class Log-loss again with newly created features and it helped to improve accuracy to 0.68.

4.1.1 Best Model for this Dataset is: -

Logistic Regression with Class Balancing

1. Train Log-Loss: 0.4246

2. CV Log-Loss: 0.9572

3. Test Log-Loss: 0.9994

4. Percentage Misclassified: 0.3270

5.Accuracy: 68.00%

4.2. Future Work:-

1. This project is a general approach for multi class classification to identify cancer class. In future we try to improve accuracy by using Deep Learning. Our model here has accuracy of about 68 percent.

2. We will try to use such a model to build for other genetic problems like Albinism, Angelman syndrome, Apert syndrome.

3. In the future, Our effort will be to build a model for Covid-19 which can easily detect symptoms of a man who is suspected.

5. References

1. Bao, Yujia, et al. "Using machine learning and natural language processing to review and classify the medical literature on cancer susceptibility genes." *JCO clinical cancer informatics* 1 (2019): 1-9.
2. Wadhwa, Somin, et al. "Semi-Automating Knowledge Base Construction for Cancer Genetics." *arXiv preprint arXiv:2005.08146* (2020).
3. Patel, Roma, et al. "Syntactic patterns improve information extraction for medical search." *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2018. No. Short Paper. NIH Public Access, 2018.
4. Deng, Zhengyi, et al. "Validation of a Semiautomated Natural Language Processing–Based Procedure for Meta-Analysis of Cancer Susceptibility Gene Penetrance." *JCO clinical cancer informatics* 3 (2019): 1-9.
5. Deng, Zhengyi, et al. "Validation of a Semiautomated Natural Language Processing–Based Procedure for Meta-Analysis of Cancer Susceptibility Gene Penetrance." *JCO clinical cancer informatics* 3 (2019): 1-9.
6. Braun, Danielle, et al. "A clinical decision support tool to predict cancer risk for commonly tested cancer-related germline mutations." *Journal of genetic counseling* 27.5 (2018): 1187-1199.

7.Aphinyanaphongs, Yindalon, et al. "Text categorization models for high-quality article retrieval in internal medicine." *Journal of the American Medical Informatics Association* 12.2 (2005): 207-216.

8. Webster, Rachel, et al. "Hereditary Cancer Syndromes: Risk Assessment and Genetic Counseling." *Holland‐Frei Cancer Medicine* (2016): 1-27.

9. Kastrinos, Fay, Gregory Idos, and Giovanni Parmigiani. "Prediction Models for Lynch Syndrome." *Hereditary Colorectal Cancer*. Springer, Cham, 2018. 281-303.

10. Hughes, Kevin S., et al. "Natural language processing to facilitate breast cancer research and management." *The breast journal* 26.1 (2020): 92-99.