Journal of Theoretical and Applied Information Technology
31st March 2022. Vol.100. No 6
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES ON GENETIC MUTATION BASED CANCER DIAGNOSIS DATA

ASHOK REDDY KANDULA[1], Dr. R. SATHYA [2] and Dr. S. NARAYANA[3]

[1]Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu-608002, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu-608002, India.

[3]Professor, Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Seshadri Rao Knowledge Village, Gudlavalleru-521356, India.

Email: [1]ashokreddy.gec@gmail.com, [2]sathya.aucse@gmail.com, [3]satyala1976@gmail.com

ABSTRACT

There are still several research studies about how precision medicine and advanced genetic testing are highly disrupted by how cancer-like diseases are treated. The major disadvantage is that identifying cancer disease by checking gene mutations is a manual process which leads to lot of misclassifications. The paper tries to intend several machine learning techniques to make the manual process into an AI-assisted process that makes the work much easier and efficient. The cancer dataset has a certain kind of complicated format. One such is text category; thus, to address the issue, the paper successfully established a complete data analysis that highly gave a detailed view about the data which is performed in the previous. The article applies several machine learning techniques like Naïve Bayes, K- Nearest Neighbours, and Logistic Regression to classify the data with and without class balancing to analyse the cancer dataset.

Keywords: *Genetic testing, Gene Mutation, Naïve Bayes, K-Nearest Neighbour, Logistic Regression.*

1. INTRODUCTION

Artificial Intelligence (AI) solves various kinds of problems in a broad range of areas, and one such is the medical industry. Machine learning, a part of AI, has likely different applications in medication and has been applied to a wide assortment of oncology problems and tasks, such as cancer predicting susceptibility, endurance and survival rates, and medicines. Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that permits machines to learn without or with the mediation of a human [1]. In AI, ML is perhaps the most well-known model carried out quickly to prepare machines and present the prescient model's ineffective decision-making part. Under classification and forecast or prediction issues, ML strategies are the primary alternative for acquiring a superior result. For example, ML strategies could identify the appropriate type in malignant growth research and forecast cancer disease. Furthermore, these ML techniques could anticipate whether the cancer that is predicted is malignant or benign.

As clinical writings and Image-based medical problem [3] classification considered as so much become so hype, while the significant focus goes on clinical reports that incorporate the most detailed data understood simply by field specialists, the paper chose to work with several machine learning (ML) models to fit for understanding its complex and explicit structure of the data. The data describes where clinical sections portray genetic changes, or mutations and contenders are needed to group or classify the given mutations. Presently, the understanding of shared mutations is being done manually by specialists, which takes a considerable amount of time and heavy resources [2].

The personalized cancer data's primary objective is to make automated genetic mutation classification depend on their commitment to tumour development. The dataset holds huge clinical documents as input and documents containing extra information about the malignant. The additional record has complete data about the gene, its mutation, and the appropriate id of the clinical content in the form of text and its type of cancer as input data in a folder format. From such whole data, we have done a vast data processing combining all the folders and files into one dataset and done a complete data analysis in the previous work. The target class from the data ranging from 1 to 9 addresses various dangerous levels or malignancy levels of mutations on explicit genes.

The work proceeds by applying Natural Language Processing (NLP) techniques like one-hot encoding and response coding to process the data into the format where appropriate machine learning techniques can be used to get the best results. The work's high focus is on processing the text document in the data. With the massive development of text archives on the web and in specific domains, necessary data retrieval has become a critical undertaking to fulfil the requirements of various end clients and many individual users. To this end, the programmed text category has arisen as an approach to adapt to such an issue. Programmed text or automated text as a category type in the data to be analyzed has endeavoured to replace and save human exertion that is highly needed in performing manual processing.

It comprises allotting and naming archives utilizing predefined classes and specific categories based on given document contents. Accordingly, one of the essential targets of programmed text categorization has been the upgrade and become a massive value in contributing to data retrieval tasks to handle problems in prediction, for example, data filtering and routing, a grouping of related documents, followed by classifying the documents into pre-determined subject topics.

Programmed text categories are highly used in web search engines, digital library frameworks, and several augment document management frameworks that include electronic email separations from spam, newsgroups characterization, and several survey-based data groups. As a continuation of the research work [4], addressed a complete data analysis that performed univariate analysis on individual features that found the importance of each feature, stability by applying machine learning models. The paper follows the second stage multivariate analysis by using models like K Nearest Neighbors, Naïve Bayes, Logistic regressions with and without class balancing, which was considered a statistical machine learning technique to figure out the way of classification on clinical texts, gene and Variation features. The primary objective is to evaluate the several classification models concerning the effectiveness and efficiency of each algorithm by classification through achieving test log loss, confusion, precision and recall matrix.

The paper is arranged in the structure, learning about cancer and gene mutation issues and several machine learning techniques in the Introduction section, followed by a review of the literature section on several above techniques following to materials and methodologies explained about the methods used in the paper. The results are well depicted with good discussions under the results and discussion section, and finally, the conclusion and future followings are mentioned in the conclusion section.

2. LITERATURE REVIEW

The patterns distinguished from the efficiently gathered molecular or genetic profiles of tumour patient samples, alongside clinical metadata, would help customize medicines for viable handling of malignancy patients with similar genetic types. There is a neglected need to create computational calculations for disease analysis like cancer and several other kinds in the medical field, visualization and therapeutics that can distinguish complex examples and help in groupings dependent on plenty of arising cancer disease research results in an open area [5]. Machine learning, a small part of the immense ocean of Artificial Intelligence (AI), holds extraordinary potential for pattern recognition or classifying and grouping the appropriate categories into one type. The literature explains the different studies related to cancer diseases and several machine learning techniques to know which is suitable for any problem.

In [6] presented a K-nearest Neighbourhood (KNN) based technique to analyse the gene expression data for classifying cancer malignancy in the oncogenomic field. It further utilized the counting quotient filter approach then Euclidean distance to class the data. The data generated is further classified with classifiers like J48, DNN and SVM classifiers. [7] investigated how AI performed with Cancer Genome Atlas of patient's data that holds lung adenocarcinoma (LUAD) to discover endurance explicit gene mutations that predict survival rates. To distinguish endurance rates that are specific to mutations as per different clinical variables, we had carried out a work that utilized four-component selection techniques (data acquiring, statistical methodologies, i.e., chi-squared test, most minor repetition with most excellent pertinence strategy, and connections among the data formula) has been kept basic to classify those clinical data.

In [8] presented non-invasive imaging-based biomarkers that highly demonstrate the histology followed by gene transformation status of brain metastasis from cancer cellular breakdown in the lungs. It used ANOVA-based statistical models and logistic regression analysis to determine the dissemination of weighted imaging boundaries as indicators of lung cancer's histology and gene transformations. In [9] explained progressively, the viability of emerging chemotherapy specialists for malignant breast growth has been affected by changes in the tumour genomic profile of patients. The work explored the correspondence between development and contractions between genes and gene duplicate number, changes/mutations, and articulation first in breast cancer malignancy cells that occur in patients. PCA (principal component analysis) is utilized that showed articulation was the most grounded marker of affectability for paclitaxel, and duplicate number and articulation were a handy feature for the classification of multi-factorial issues. The work was further tested with a Binary-based Support Vector Machine (SVM) [10] in a statistical tool performed in a MATLAB environment utilizing linear kernel function. In some places, cross-validation is performed before finding the appropriate fit. In [11] used Leave-one-out-cross-validation (LOOCV) technique.

In [12], chosen gene transformations are regularly used to direct the choice of cancer disease medications for any given tumour patient. Advanced AI techniques like regressions are utilized to explore how the malignant growth of the cell line affectability to appropriate drugs is anticipated relying upon the kind of tumour patient's profile. The uncovered gene articulation, gene expression in some contexts based on the given data, is considered the most prescient profile in the pan-cancer setting. Nonetheless, no examination has misused GDSC information to deliberately analyze machine learning models' working methodology and performances dependent on multi-gene articulation information against that of generally utilized single-gene quality markers that are highly dependent on genomics information. The work used systematic comparison using Random Forest [RF] classifiers [13] to address the issue.

In countries like India, the most commonly occurring disease are breast malignant. There is an opportunity of fifty percentages for a casualty for a situation as one of two women determined to have breast cancer has the least survival rate in Indian women [14]. In [15] compared different and effective ML algorithms and methods like kNN (k-Nearest-Neighbor), Naïve Bayes and Random Forest for addressing malignant breast growth,

In [16], characterized breast malignancy types depended on the feature-engineered breast disease features by providing a particular design like crossover strategy, i.e., hybrid machine learning methods. They have created a CAD-based plan through the investigation by applying a blend of Wrapper and Naive Bayes techniques to increment the precision value. The Wrapper methodology is utilized at the feature extraction stage, while the Naive Bayes calculation is used at the grouping stage. In [17] to address colon cancer malignancy, it utilized Naïve Bayes classification method based on simplistic probability of Bayes theorem In [18] to address natural language problems that are text-based review processing the work compared several machine learning techniques like Naïve Bayes, Support Vector Machines, Logistic Regression and Random Forest that proved that comparing provides the best view in identifying the perfect model in approaching the research work.

In [19], we presented three commitments identified with learning the sparse classifiers. A work utilized a multinomial-based logistic regression approach is used to address the issue. Following joining a bound advanced optimized technique with a component-wise update method, the assignment approached the faster algorithmic strategy to address sparse based multi-class classifiers that could perform well scaling for high training and heavy feature dimensionality, even suitable to fit high dimensional data. [20], addressed the oral cancer issue and its classification by presenting a multivariate regression technique where the results are shown in odds ratio associated with a confidence level (CI).

3. MATERIALS AND METHODOLOGY:

The proposed approach in this paper is followed by the previous work, where the complete data analysis is performed in [4], addressing which features are useful and which are not.

3.1 Previous Work

Considering the Features Gene, Variation and Text features and their corresponding test data log-loss metric, Gene feature obtained 1.1940, Variation obtained 1.7143, and text feature obtained 1.1774. among the three features, we have concluded that text feature is handy, followed by Gene feature and Variation which is less valuable and less stable than gene and text feature. But still, every test data result of each feature gave values lesser than the Randomized Model; thus, in the upcoming machine learning operations, we planned to combine all features into one data frame for performing multivariate analysis.

3.2 Proposed Work

The work considers all features by combining them into a single data frame to perform multivariate analysis. On top of the combined data frame, we apply several machine learning models to test which model performs better to take over future work. The results are obtained in the Log-Loss metric, several misclassified points, and a confusion matrix.

4. DATA PREPARATION

The stacking of the features are performed in two ways, One hot encoding and response coding to represent the given data. The work considers the above two feature representation methods suitable for some specific machine learning model types. The stacking will be done by combining each feature for train, test, and CV. From the One-hot encoding stacking method, the data dimension obtained from the number of data points and several features is trained (2124,55772), test (665,55772), CV (532, 55772). From the response coding stacking method, the data dimension obtained from several data points with many features is a train (2124, 27), test (665, 27), CV (532, 27). The above stacking methods gave different dimensional as one-hot encoding gave high dimensional of 55772-dimensional data while response coding gave 27-dimensional data, 9-dimensions for each feature. The machine learning models are applied on two sets of data based on their handling type that is tree-based model could handle lesser dimensional data only while the Regression-based model could handle high dimensional data.

A. Naïve Bayes

Naive Bayes is an algorithmic technique frequently utilized in many machine learning problems, specifically text-based categorical data. The fundamental idea of the technique is to join the likelihood of words and several categories to gauge the probability of the category in the given data set or document. Naïve Bayes is a methodology that prompts Baye's hypothesis by joining past information with new information. Thus, this technique is considered one of the most straightforward classification techniques that could result in more accuracy. The Bayesian Classification technique depends on the Bayes hypothesis [21], which has comparative grouping capacities to the decision tree and neural networks. Bayesian Classification technique is demonstrated to have high precision and at higher speed even for large datasets works explicitly well for text classification [22].

Bayes theorem is calculated by using equation-1 and equation-2 given below Probability(A|B) from Probability(A),Probability(B) and Probability(B|A).

$$Probability(A|B) = \frac{Probability(B|A) * Probability(A)}{Probability(B)} \quad (1)$$

$$P(A|B) = P(A_1|B) \times P(A_2|B) \times \ldots .\times P(A_n|B) \times P(A) \quad (2)$$

Where P(A|B) is Posterior probability, P(B|A) is a likelihood, P(A) is prior probability, P(B) is predictor prior probability. The equation explains that calculating the posterior probability (A|B) is the probability of class (A, target) with a given predictor factor (B, attributes), while probability (A) denotes the prior probability for the given class variable. Thus, probability (B|A) represents the likelihood, that is, the probability for the predictor share class, where Probability (B) denotes the prior probability of the obtained predictor.

B. K Nearest Neighbor

The KNN classification technique is a speculation-based calculation technique for closest neighbour rules. Its inductive balance is the class label of the k-example with the class labels that must be tested generally with the closest one. Compared with the nearest neighbour, it varies in that it grows the closest neighbour to k as it goes to the decision-making stage. This generalized expansion permits the KNN technique to handle more extensive data. However, it excludes the way toward getting the hang of preparing comparative with other classification techniques with unique training stages.

The nearest neighbour approach of the KNN technique is considered the most established and oldest technique for classifying target classes. The decision-making during the classification depicts a straightforward process, where the given sample which needed to be tested is equivalent to the given sample category nearest to it. During the calculation, if the training set and corresponding distance metric when kept unaltered [23], the prediction result of the closest neighbour rule has been mainly decided for any data point to be tested.

Discussing the workings, In the given classification problem, the K-nearest neighbour approach does that for a given estimation of K that led to track down the K nearest neighbour of that of data point that is unseen in training data and afterwards specific class that has the highest number of distributions are assigned to unseen data point out of all classes of K neighbours. The Euclidean metric has been utilized to calculate the distance metric, given in equation number 3.

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + \cdots + (x_n - x_n')^2} \quad (3)$$

At most, the input data point x is assigned to some class with the highest probability.

$$P(y = j|X = x) = \frac{1}{K}\sum_{i\in A} I\left(y^{(i)} = j\right) \quad (4)$$

C. Logistic Regression:

Logistic Regression (LR) based techniques are perhaps the most mainstream approaches to fit models for clear-cut information in categorical data, particularly for binary paired data. It is the most significant (and likely generally utilized) individual from a class of models under generalized linear models. In contrast to linear regression-based regression models, LR can straightforwardly anticipate probabilities (values that are confined to the (0, 1) stretch); moreover, those probabilities are very much aligned when contrasted with the possibilities anticipated by some different classifiers, like Naive Bayes, K- Nearest Neighbours [24,25]. Thus, LR highly safeguards the minor probabilities inside the training set in the data. The coefficients of the model likewise give some trace of the general significance of each input variable. The LR for two case-based classification holding (0, 1) is derived by assuming the log odds of a given target y can be expressed as a linear function of a set of k input variables given in equation 5.

$$log\frac{P(x)}{1 - P(x)} = \sum_{j=0}^{K} b_j x_j \quad (5)$$

In the above equation 5, it is adding the constant b_0 by setting $x_0 = 1$. It provides (K+1) parameters. The equation on the left is called logit of P, derived from logistic regression in above equation 5, taking exponent in both sides to get as equation 6.

$$\frac{P(x)}{1 - P(x)} = \exp\left(\sum_{j=0}^{K} b_j x_j\right) \quad (6)$$

$$= \prod_{j=0}^{K} \exp(b_j x_j) \quad (7)$$

The above equation 7 explains that logistic models are multiplicative over their input sources (instead of additive models like linear models) that provides a practical approach to interpret the coefficients. The value $\exp(b_j)$ in the above equation explains how the response's chances increase or decrease for the input variable as x_i increases by one unit while all other expressions are equal.

For the Regression-based regression approach, the procedure tends to be similar, rather than the classes of the nearest neighbours where we tend to estimate the target class and discover the target cover for the inconspicuous datapoint, i.e., unseen data point. The procedure follows through utilizing a one-versus many-based approach in classifying the target class. The metric used to analyze and test the data is through log loss metric. Using calibrated models on top of each model, we tend to get log-based probabilities, which calculates the loss metric for each model. Further, the model is differentiated by the misclassification point rate. The misclassified points are calculated when a predicted y hat is not equivalent to the original target y, and then it is counted as a misclassification point. Where to log loss of CV is calculated from actual probabilities, this Percentage of misclassification point doesn't depend on the log of CV and test, and it doesn't change much while log-loss of test and CV shift much. The models are well hyper tuned before fitting the model with several sets of alpha values and tested well to get the best results.

5. Results and Discussions:

The data is ultimately converted into the data form to apply different machine learning techniques for classifying. The results are compared with log loss metric, confusion matrix, recall matrix and precision matrix.

A. Multivariate analysis with Naïve Bayes Classification:

For Naïve Bayes (NB) model, one-hot encoding feature conversion is utilized where the dimensional will be higher than response coding because the data set holds text category as one variable. Thus, Naïve Bayes works well for text-based data. The model is tuned to perfect Laplace smoothing by the grid search approach. From figure1, the alpha value = 1 gave the minor error measure: log loss metric as 1.2570 for CV data and train log loss as 0.9258; thus, the naïve Bayes gave alpha = 1 as a correct measure of best fit to fit the model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

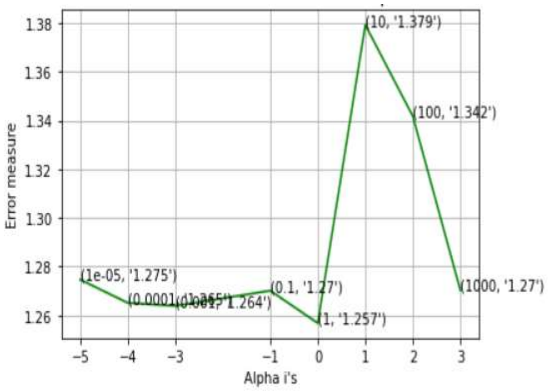


Figure1: Cross-Validation Error For Each Alpha In Naïve Bayes.

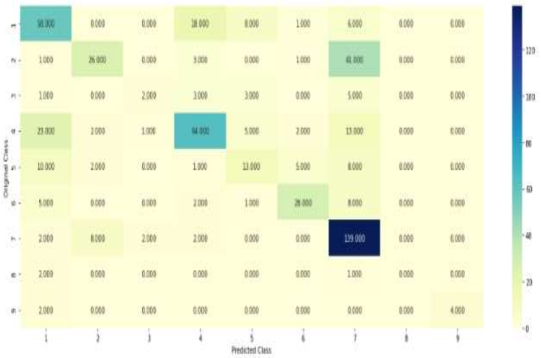
Testing the model with best hyperparameters:

With the best fit of alpha = 1, the model is trained with a multinomial Naïve Bayes classifier and on top of it calibrated model is applied to get a log-based loss metric measure. The results of the NB classifier of train log-loss is 0.9258, CV log loss is 1.2570, test log-loss is 1.2567, and accuracy is 83.83%, precision as 0.830, recall as 0.822 and F1 measure as 0.826, while the number of misclassified point rates is 37.21%. The highlighted note is in the previous work of the research paper [4] the text featured univariate analysis got CV log-loss as 1.1137, test log loss as 1.1774 while in NB CV log loss as 1.2570, test log loss as 1.2567, which is highest, this statement explains text feature alone that is classified with Logistic Regression could give better result compared to NB with all features, thus the data model is further tested with different classifiers.

From figure 2, the confusion matrix of NB clearly explains that the target class is distributed more under classes 7, 2, 1 and 4, while in other classes, the data distributed is less. Thus it becomes less effective in NB's confusion matrix. In a precision matrix where the diagonal should be equal to 1, the points are 100% correctly classified. From figure 3, considering a few classes, one such is class 4, the original class and predicted class obtained is 68% that tells, of all the points that are predicted to be class 4, 68% are only predicted as class 4, which is not 100% to be noted. While for class 7 and class 3, it predicted 40% of points are class 3, but it belongs to class 7, which is not good. Considering class 7 and class 2, of all points to be predicted as class 2, out of 21% of points predicted as class 2, it originally belongs to class 7.

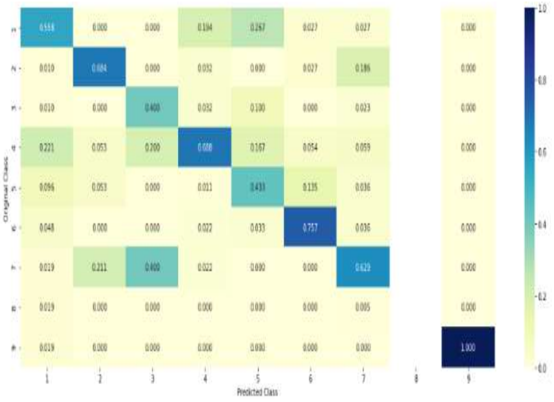
Similarly, for classes 4 and 1, almost 22% of points predicted to be class 1 belong to class 4. This observation clearly explains that the classes are getting replaced. That is, class 2 and class 7 are getting replaced while predicting similarly class 1 and class 4 are also highly replaced while predicting.

In the Recall matrix, as shown in figure 4, the row sum is 1, and the diagonal needs to be 1 for correct classification, considering fewer columns and rows, for class 7, of all points that originally belonged to class 7, 90% of points are predicted as class 7 which is good value. While for class 8 and class 1, of all the points belonging to class 8, 66% are predicted as class 1, which could be acceptable as class 8 has fewer data points. For class 2 and class 7, of all the originally belonged to class 2, 56% of the points are classified as class 7. While for class 4 and class 1, of all the points that are originally belonged to class 4, 20% of points are classified as class 1, thus similar to the precision matrix, the target classes 2, 7 and class 4 and 1 are getting replaced while predicting which is still a disadvantage for classification leading to increased misclassification points. Thus now trying a tree-based model for classification.



	1	2	3	4	5	6	7	8	9
1	58.000	0.000	0.000	18.000	8.000	1.000	6.000	0.000	0.000
2	1.000	26.000	0.000	3.000	0.000	1.000	41.000	0.000	0.000
3	1.000	0.000	2.000	3.000	3.000	0.000	5.000	0.000	0.000
4	23.000	2.000	1.000	64.000	5.000	2.000	13.000	0.000	0.000
5	10.000	2.000	0.000	1.000	13.000	5.000	8.000	0.000	0.000
6	5.000	0.000	0.000	2.000	1.000	28.000	8.000	0.000	0.000
7	2.000	8.000	2.000	2.000	0.000	0.000	139.000	0.000	0.000
8	2.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
9	2.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4.000

Figure2: Confusion Matrix For Naïve Bayes.



	1	2	3	4	5	6	7	8	9
1	0.558	0.000	0.000	0.194	0.267	0.027	0.027		0.000
2	0.010	0.684	0.000	0.032	0.000	0.027	0.186		0.000
3	0.010	0.000	0.400	0.032	0.100	0.000	0.023		0.000
4	0.221	0.053	0.200	0.688	0.167	0.054	0.059		0.000
5	0.096	0.053	0.000	0.011	0.433	0.135	0.036		0.000
6	0.048	0.000	0.000	0.022	0.033	0.757	0.036		0.000
7	0.019	0.211	0.400	0.022	0.000	0.000	0.620		0.000
8	0.019	0.000	0.000	0.000	0.000	0.000	0.005		0.000
9	0.019	0.000	0.000	0.000	0.000	0.000	0.000		1.000

Figure3: Precision Matrix For Naïve Bayes.

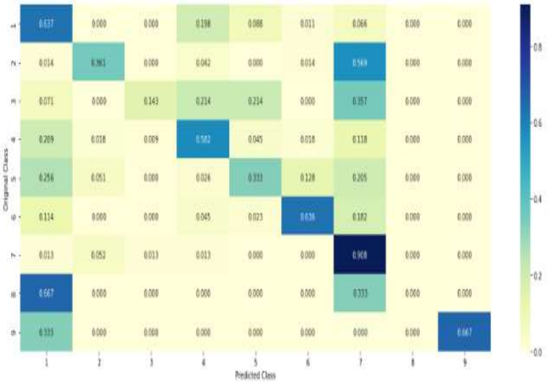


Figure4: Recall Matrix For Naïve Bayes.

B. Multivariate analysis with K Nearest Neighbour Classification

The K Nearest Neighbour classification works poorly for high dimensional data; thus, the data model built using response coding is considered for classification. The KNN at first finds the best k value to fit the model. The data model is finely tuned with several k values like [9, 11, 15, 21, 31, 41, 51, …99] as shown in figure 5, from it the best k is 15 that obtained train log loss as 0.6557, CV log loss as 1.0450 and accuracy as 86.59%, precision as 0.869, recall as 0.848, and F1 measure as 0.8584. The CV log loss, accuracy score, precision and recall values for KNN is better compared to Naïve Bayes CV log loss. Thus, KNN that implemented with response coding works better than Naïve Bayes.

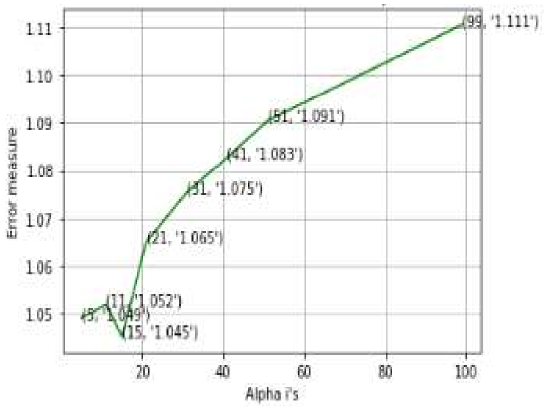


Figure5: Cross-Validation Error For Each Alpha

Testing the model with best hyperparameters

From testing the best alpha, that is, several neighbours k points, the log loss the model obtained is 1.0450, and the misclassification rate is 38% which is almost closer with Naïve Bayes. At the same time, test results are far different from NB because the misclassified points are calculated in the form when a predicted y hat is not equivalent to the original target y, then it is counted as a misclassification point. However, where log loss of CV is measured from actual probabilities, this Percentage of misclassification point doesn't depend on the log of CV and test, and it doesn't change much while log loss of test and CV change much

. From Figure 6 and the confusion matrix of KNN, classes 1, 2, 4, 7 dominate the data. Figure 7, precision matrix still holds the same confusion between class 7 and class 2 and in class 4 and 1. But it worked better for class 8, getting 100%, while for class 9, also 100% that is KNN is good at handling the less distributed target classes, where NB doesn't get good results for class 8. In the recall matrix, as shown in figure 8, target classes 8 and 9 are considered in the classification, but still, the problem arises among classes 2, 7 and class 4 and 1.

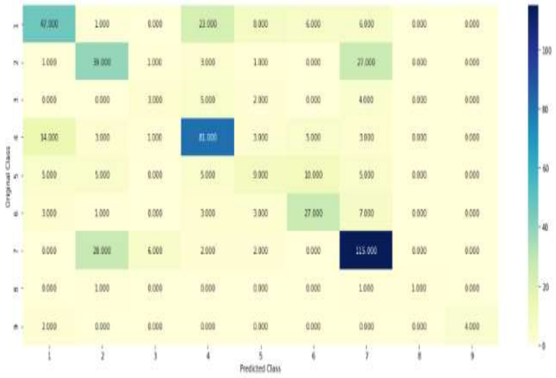


Figure6: Confusion Matrix For Knn Classification

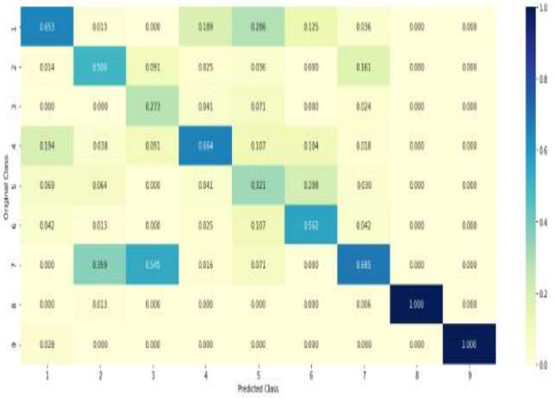


Figure7: Precision Matrix For Knn Classification

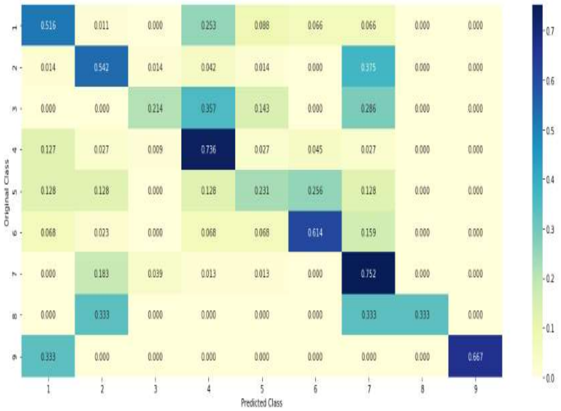


Figure8: Recall matrix for KNN

C. Multivariate Analysis Logistic Regression with class balancing:

The results of the model's NB and KNN are seen. Now logistic regression is performed with one-hot encoding where Logistic Regression (LR) can easily handle high dimensional data, which separate various classes effectively. In LR, the paper considers performing two models: LR with balancing the class and LR with not balancing the class as we did in NB and KNN. Considering the first model LR with class balancing and performing hyperparameter tuning for several alphas values with L2 regularization, got CV log loss as 1.1024 with alpha = 0.001, and train log loss as 0.5242 figure 9.

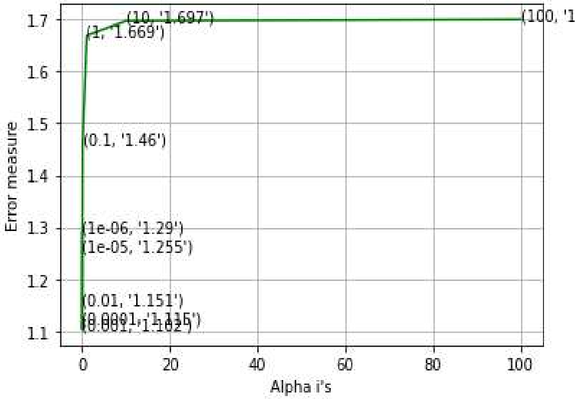


Figure 9: Cross-Validation Error For Each Alpha

Testing the model with best hyperparameters:

Fitting the model with best fit alpha = 0.001, the results gave test log loss as 1.1024, accuracy as 87.09%, precision as 0.882, recall as 0.868, and F1 measure as 0.8744 misclassification rate as 33%, which are good results compared to NB and KNN. From figure 10, as usual, class 7, class 4, class 1 performed well compared to others. The precision matrix, as shown in figure 11, and the recall matrix, as shown in figure 12, gave betters all the classes like classes 3, 6, 5, 8, 9 also performed class balancing by oversampling technique thus all the classes equally gets recognized.

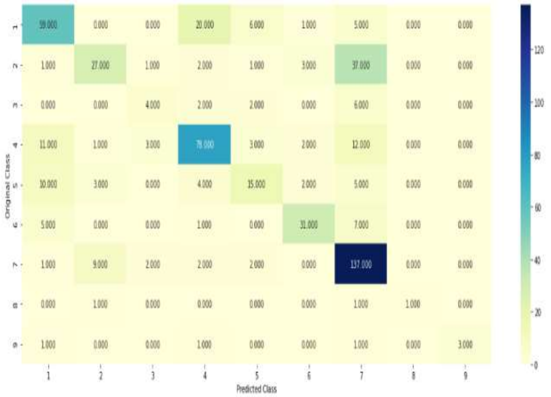


Figure10: Confusion Matrix for Logistic Regression with class balancing

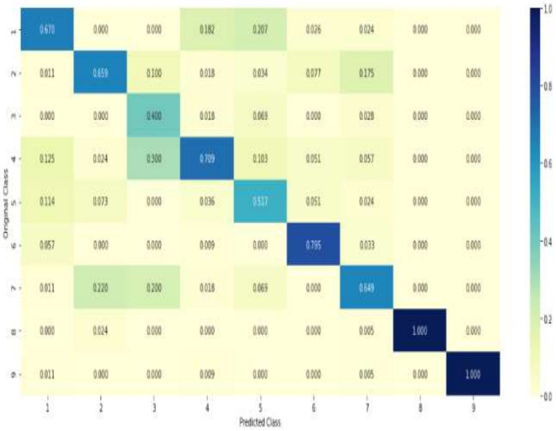


Figure11: Precision Matric For Logistic Regression With Class Balancing

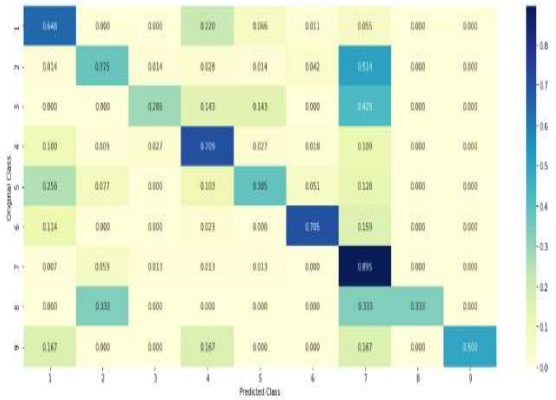


Figure12: Recall Matrix For Logistic Regression With Class Balancing

D. Multivariate analysis of Logistic Regression without Class balancing

The model is performed without class balancing technique as it gave the best fit in hyperparameter tuning, alpha = 0.001 it gave train log loss as 0.5206, while CV log loss as 1.1005 as shown in figure 13. The observation shows that the CV log loss results are almost closer to the previous LR model with class balancing.

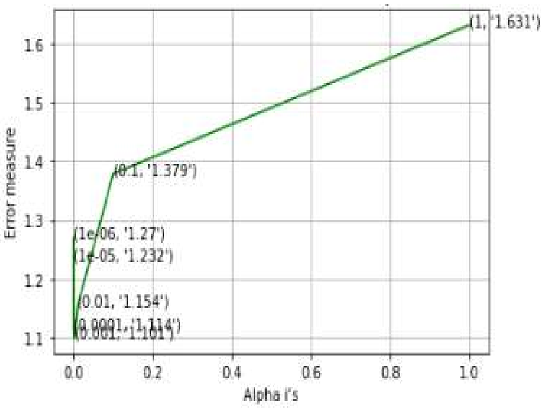


Figure13: Cross-Validation Error For Each Alpha

The testing model with best hyperparameters:

From testing the model with best-fit alpha = 0.001, the model gave test log loss as 1.1005, with accuracy as 85.14%, precision as 0.842, recall as 0.839, and F1 measure as 0.841 and misclassification point rate as 33%. As shown in figure 14, the confusion matrix gave some excellent results in addition to usual dominant classes 1, 4, 7, 2, and it gave class 5 a bit included in the analysis, as shown in figure 14. While precision shown in figure 15 and recall are shown in figure 16 matrixes, all the classes are considered equally and give good results. Still, the minor or less dominant types are seemed as less critical where it is needed to get 1 to be a good classifier. The class balancing model helped minor classes to get better values in the precision and recall matrix. There are only a few differences between LR with class balancing and LR without class balancing. Class balancing technique always sides more advantageous that minority classes or less frequently occurring classes tend to perform well which is highly useful in Logistic Regression and imbalanced type of datasets.

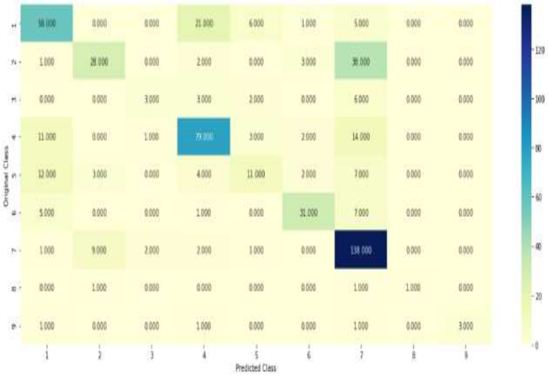


Figure14: Confusion Matrix For Logistic Regression Without Class Balancing

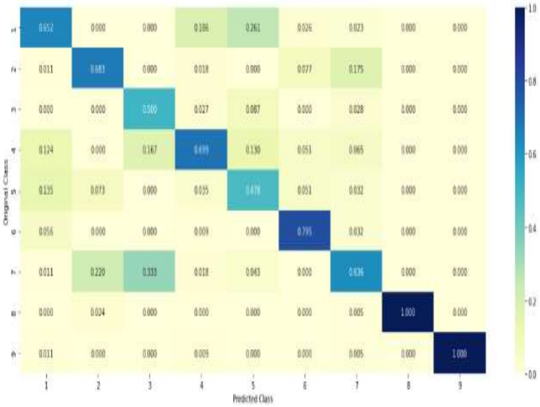


Figure15: Precision Matrix For Logistic Regression Without Class Balancing

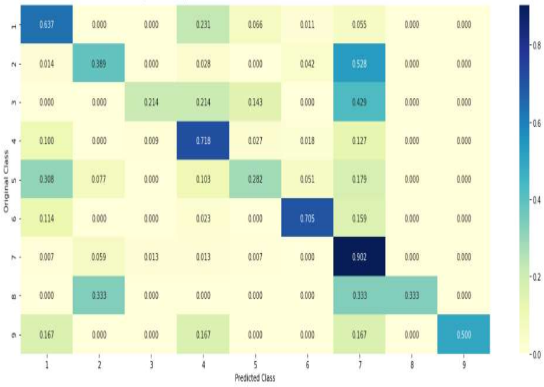


Figure16: Recall Matrix For Logistic Regression With Class Balancing

Table 1: Results Of All Models

	Parameter	Train [Log Loss]	CV [Log Loss]	Test [Log Loss]	Accuracy [Percentage]	Mis-classification points
Naïve Bayes	α=1	0.9258	1.2570	1.2567	83.83	37%
K-Nearest Neighbour	K=15	0.6557	1.0450	1.0196	86.59	38%
Logistic Regression [with class balancing]	α=0.001	0.5242	1.1024	1.0116	87.09	33%
Logistic Regression [without class balancing]	α=0.001	0.5206	1.1005	1.0208	85.14	33%

Table 1 describes that among all the models the paper incorporated, Logistic Regression with class balancing gave a least Log loss metric and more accurate results, but still, it is significantly closer to LR without class balancing model and KNN of test log loss and accuracy values. While considering misclassified points, LR in common with both class balancing and without class balancing gave the exact percentage value of 33%. Thus, we could say LR with class balancing is good from the mathematical proof compared to other models. But still, it gave only fewer differences in results, so in the future, as the continuation of the research work, we are planning to implement a hybrid model to analyse with the same dataset to check whether the hybrid model will give even lesser log loss and better accuracy than the used models.

6. CONCLUSION:

Thus, we could say LR with class balancing is good from the mathematical proof compared to other models, 1.0116 as test log loss metric, 87.09% as accuracy, precision as 0.882, recall as 0.868, F1 measure as 0.8744 and 33% of misclassification points in identification of cancer disease with gene mutation. But still, it gave very minute difference in results compared to other models, so in the future, as the continuation of the research work, we may intend to implement a hybrid model to analyse with the same dataset to check whether the hybrid model will give lesser or higher to the current log loss and accuracy metric.

REFERENCES:

[1] Kolachalama, V. B., & Garg, P. S. (2018). Machine learning and medical education. NPJ digital medicine, 1(1), 1-3.

[2] Richens, J. G., Lee, C. M., & Johri, S. (2020). We are improving the accuracy of medical diagnosis with causal machine learning. Nature communications, 11(1), 1-9.

[3] De Bruijne, M. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis.

[4] Kandula, Ashok & R, Sathya & S, Narayana. (2021). Performing Univariate Analysis on Cancer Gene Mutation Data Using SGD Optimized Logistic Regression. International Journal of Engineering Trends and Technology. 69. 59-67. 10.14445/22315381/IJETT-V69I2P209.

[5] Jagga, Z., & Gupta, D. (2015). Machine learning for biomarker identification in cancer research–developments toward its clinical application. Personalized medicine, 12(4), 371-387.

[6] Chaudhari, P., Agarwal, H., & Bhateja, V. (2019). Data augmentation for cancer classification in oncogenomics: an improved KNN based approach. Evolutionary Intelligence, 1-10.

[7] Cho, H. J., Lee, S., Ji, Y. G., & Lee, D. H. (2018). Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. PLoS One, 13(11), e0207204.

[8] Jung, W. S., Park, C. H., Hong, C. K., Suh, S. H., & Ahn, S. J. (2018). Diffusion-weighted imaging

of brain metastasis from lung cancer: Correlation of MRI parameters with the histologic type and gene mutation status. American Journal of Neuroradiology, 39(2), 273-279.

[9] Dorman, S. N., Baranova, K., Knoll, J. H., Urquhart, B. L., Mariani, G., Carcangiu, M. L., & Rogan, P. K. (2016). Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer-derived by machine learning. Molecular Oncology, 10(1), 85-100.

[10] Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proceedings of the National Academy of Sciences, 113(50), 14330-14335.

[11] Madhukar, N. S., Elemento, O., & Pandey, G. (2015). Prediction of genetic interactions using machine learning and network properties. Frontiers in bioengineering and biotechnology, 3, 172.

[12] Nguyen, L., Dang, C. C., & Ballester, P. J. (2016). Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. F1000Research, 5.

[13] Leung, M. K., Delong, A., Alipanahi, B., & Frey, B. J. (2015). Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE, 104(1), 176-197.

[14] A.Mallikarjuna Reddy, Vakulabharanam Venkata Krishna, Lingamgunta Sumalatha and Avuku Obulesh, “Age Classification Using Motif and Statistical Features Derived On Gradient Facial Images”, Recent Advances in Computer Science and Communications (2020) 13:65. https://doi.org/10.2174/221327 591266 6190417151247.

[15] Sharma, S., Aggarwal, A., & Choudhury, T. (2018, December). Breast cancer detection using machine learning algorithms. In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 114-118). IEEE.

[16] Maysanjaya, I. M. D., Pradnyana, I. M. A., & Putrama, I. M. (2018, June). Classification of breast cancer using Wrapper and Naïve Bayes algorithms. In Journal of Physics: Conference Series (Vol. 1040, No. 1, p. 012017). IOP Publishing.

[17] Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. International Journal of Innovative Science, Engineering & Technology, 2(9), 441-444.

[18] Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), 221.

[19] Krishnapuram, B., Carin, L., Figueiredo, M. A., & Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE transactions on pattern analysis and machine intelligence, 27(6), 957-968.

[20] Singh, V., Dwivedi, S. N., & Deo, S. V. S. (2020). The ordinal logistic regression model describes factors associated with nodal involvement in oral cancer patients and its prospective validation. BMC medical research methodology, 20, 1-8.

[21] Berrar, D. (2018). Bayes’ theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, 403-412.

[22] Ayaluri MR, K. SR, Konda SR, Chidirala SR. 2021. Efficient steganalysis using convolutional auto encoder network to ensure original image quality. PeerJ Computer Science 7:e356 https://doi.org/10.7717/peerj-cs.356.

[23] Gou, J., Qiu, W., Yi, Z., Xu, Y., Mao, Q., & Zhan, Y. (2019). A local mean representation-based K-nearest neighbor classifier. ACM Transactions on Intelligent Systems and Technology (TIST), 10(3), 1-25.

[24] Swarajya lakshmi v papineni, A.Mallikarjuna Reddy, Sudeepti yarlagadda , Snigdha Yarlagadda, Haritha Akkineni "An Extensive Analytical Approach on Human Resources using Random Forest Algorithm" International Journal of Engineering Trends and Technology 69.5(2021):119-127.

[25] Ilaiah Kavati, A. Mallikarjuna Reddy, E. Suresh Babu, K. Sudheer Reddy, Ramalinga Swamy Cheruku,Design of a fingerprint template protection scheme using elliptical structures,ICT Express,Volume 7, Issue 4,2021,Pages 497-500,ISSN 2405-9595,https://doi.org/10.1016/j.icte.2021.04.001