Educational Administration: Theory and Practice
2023, 29(3), 381-388
ISSN: 2148-2403
https://kuey.net/

Research Article

Comparison of Logistic Regression, Naive Bayes and Random Forest Classifier Methods for Drug Review

Priyanka Masih[1*], Sunita Kushwaha[2]

[1*]Research Scholar, MATS School of Information Technology, MATS UNIVERSITY, Raipur (C.G.), India, Email: priyanka.masih123@gmail.com
[2]Associate Professor, MATS School of Information Technology, MATS UNIVERSITY, Raipur (C.G.), India,

Citation: Priyanka Masih (2023) Comparison Of Logistic Regression, Naive Bayes And Random Forest Classifier Methods For Drug Review, *Educational Administration: Theory and Practice*, *29* (3), 381 - 388
Doi: 10.53555/kuey.v29i3.4667

ARTICLE INFO

Received: 05-07-2023
Accepted: 08-08- 2023

ABSTRACT

Machine Learning techniques are popularly used in a wide range of applications. However, it is not yet clear which classifier is best suited for which data. Moreover, the proposed work comparing how Nave Bayes, Random Forest and Logistics Regression differ from each other based on a given Drug review dataset. Drug review analysis has become very useful in present times as classifying medicines based on their effectiveness through analyzing online reviews from users can assist future consumers in collecting knowledge and making better decisions about a particular drug. Here, we are collected drug review dataset and processed for analysis. For analytical study, R Programming is used. This dataset provides patient reviews on specific drugs along with related conditions, and the reviews are analyzing by patient rating, which reflects overall patient satisfaction. The objective of this proposed research is to measure the effectiveness level of a particular drug. This paper is comparing classifiers by evaluating their classification accuracy; precision, recall, F1-score, and area under the ROC curve are compared in terms of performance factor.

Keywords: Machine Learning, Logistic Regression, Naive Bayes, Random Forest, R Programming.

1. INTRODUCTION

In opinion mining, most of the researchers have worked on general domains such as electronic products, movies, and restaurant reviews, but not much on health and medical domains. Patients using drugs are often looking for stories from patients like them on the internet which they cannot always find among their friends and family. Online health forums are becoming an increasingly popular platform for people to search for health-related information. The opinion mining method employed in this work focuses on predicting the drug satisfaction level among the other patients who have already experienced the effects of a drug.

Analyzing drug reviews presents a unique challenge due to the unstructured nature of text data and the inherent complexity of human language. Traditional methods of manually categorizing and interpreting these reviews are time-consuming. Machine learning techniques help to automate the process of sentiment analysis and categorization of drug reviews. In this context, classification algorithms such as Logistic Regression, Naive Bayes, and Random Forest have emerged as popular choices for analyzing text data and predicting sentiment labels. Drug reviews, enabling automated classification into categories such as positive and negative sentiments. The objective of this study is to conduct a comparative analysis of these classification algorithms Logistic Regression, Naive Bayes, and Random Forest in the context of drug review analysis. By evaluating their performance on a real-world dataset of drug reviews, Finding the optimal structures and their values to implement the algorithms, understanding and predict the data to support predicting and knowledge gathering process, i.e. to classify unclassified drug-review data that will help the patients to decide whether to use or not. The outcome can be beneficial for both consumers and manufacturers to understand the effectiveness of drugs as well as whether a particular drug has any significant side-effects or not [1, 4].

2. LITERATUREREVIEW

Gopalakrishnan et al. studied that applying neural network based methods for opinion mining from social web in health care domain. We have extracted the reviews of two different drugs. Experimental analysis is done to analyze the performance of classification methods on reviews of two different drugs. The results demonstrate that neural network based opinion mining approach outperforms the support vector machine method in terms of precision, recall and *f*-score [1].
Gladence et al. proposed that various bayes classifiers like Bayes Network, Naive Bayes, Naïve Bayes Multinomial Text, and Naïve Bayes Updateable are working and how they differ with each other based on given data and these results are effectively compared with Logistics Regression. A result shows that Logistic Regression outperforms Bayesian Classification Methods in terms of various performance measures such as Precision, Recall, Mean Absolute error, Kappa, RMSE [2].
PRANCKEVIČIUS et al. studied that the paper is on comparing these classifiers by evaluating the classification accuracy, based on the size of training data sets, and the number of n-grams. In experiments, short texts for product-review data from Amazon were analyzed .The experimental results have shown that the Naïve Bayes classification method for product-review data achieves 1 – 2% higher average of classification accuracy than the Random Forest and Support Vector Machine method, but the difference is not statistically significant [3].
Nazim Uddin et al. proposed that the research, he have applied five machine learning algorithms. Unlike most of the similar researches in NLP when text mining is used for clustering the data, supervised learning methods have been implemented in this research to gain a better understanding of a drug by measuring its level of effectiveness. It is found that the Random Forest algorithm has generated the best accuracy among the four algorithms. In the case of Random Forest, higher precision, recall, and f1-score have been achieved for effective drugs compared to those measurements of ineffective drugs. The reason behind calculating the f-1 score is to get accuracy measurement from a different perspective as the f-1 score delivers the balance between precision and recall [4].
Garg proposed that uses patient reviews to predict the sentiment using various vectorization processes like Bow, TF-IDF, Word2Vec, and Manual Feature Analysis, which can help recommend the top drug for a given disease by different classification algorithms. The predicted sentiments were evaluated by precision, recall, f1score, accuracy, and AUC score. The results show that classifier Linear SVC using TF-IDF vectorization outperforms all other models with 93% Accuracy [5].
PARTHASARATHY et al. studied that Machine learning models plays a vital role in medical data analysis. This object deals with the comparison of two machine learning models for medical data. Based on the precision, recall, f-score we estimated model accuracy and identified best model among Logistic Regression and Naive Bayes. Model and NB models exhibited accuracy at 25% testing set respectively 92.3 and 86.53. This comparative study on the accuracy of the models in the analysis of medical data would let us conclude that the LR model is more accurate or the NB model [6].
Jacob et al. studied that for this work, twitter data was taken as the dataset. For training data, supervised machine learning algorithms like Naïve Bayes and Logistic Regression are used. For analyzing the data, Python programming was used. In Order to train it and then check its accuracy. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. The classification accuracies of Naive Bayes and Logistic Regression on the twitter data is compared and the result shows that Naïve Bayes classifier yielded more classification accuracy than Logistic Regression classifier. The accuracy obtained with Naïve Bayes classification technique on Bigdata is 73% whereas Logistic Regression produces 69% accuracy[7].
Rao et al. proposes a medicine recommendation system, which takes the patient review data and performs sentiment analysis on it to find the best medicine for a disease by using N-Gram model. In order to increase the accuracy, a Light gbm model is used to perform medication analysis. The paper also discusses the advantages, disadvantages and enhancements that can be incorporated to improve the accuracy [8].

3. EXPERIMENTALWORK

3.1 Problem Design

The following is the summary of our methodology for developing and validating the prediction models (Fig. 1).

- Data extraction: The main goal of this stage is to select only the required and related data fields to process the data and optimize memory usage. Only required fields are taken from the input dataset.
- Collected data stored in excel file and convert it into 'CSV' format for further processing in R.
- R-Tool: It is a tool for programming. Using this tool we can develop a program to predict the sentiment score of text.
- Importing libraries and packages which we need to apply in our model.

- Apply the following classification methods splitting the dataset into training and testing models in R.
  - Naive Byes Classification Method
  - Logistic Regression Method
  - Random Forest Method
- Predict the class (positive or negative) of each review in the train and test dataset.
- Compare the prediction results with actual values.
- Compute the models quality parameters and compare the prediction results and validate the models.

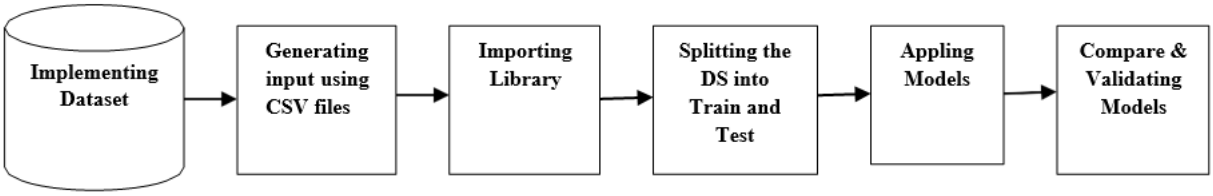

Fig 1: Problem Design

Table 1:Sample Drug Reviews Dataset

No.	Drug Name	Rating	Effectiveness	Side Effect	Condition	Benefits Review	Side Effects Review	Comments Review
1.	Biaxin	9	Considerably Effective	Mild Side Effects	sinus infection	The antibiotic may have destroyed bacteria causing my sinus infection.	Some back pain, some nausea.	Took the antibiotics for 14 days. Sinus infection was gone after the 6th day.
2.	lamictal	9	Highly Effective	Mild Side Effects	bipolar disorder	Lamictal stabilized my serious mood swings.	Drowsiness, a bit of mental numbness. If you take too much, you will feel sedated.	Severe mood swings between hypomania and depression with suicide ideation before Lamictal.
3.	depakene	4	Moderately Effective	Severe Side Effects	bipolar disorder	Initial benefits were comparable to the brand name version of this drug, Depakote.	Depakene has a very thin coating, which caused severe heart burn and stomach upset.	Depakote was prescribed to me by a Kaiser psychiatrist in Pleasant Hill, CA in 2006.

3.2 Dataset

The dataset that used in this experiment has been collected from UCI Machine Learning Repository [9] .where 3107 data is used. After collecting the dataset three machine learning algorithms have been applied to the dataset for binary classification. The classes for binary classification are class 0, and class 1, where 1 represents the effective drugs and class 0 indicates the less effective drugs. The algorithms used for binary classification are naïve Bayes classifier, Logistic Regression (LR) and Random Forest method applied to the dataset for classification. Table 2 represents the classes for effectiveness classification along with counts for each class. Class represents highly effective drugs with a count of 1330, whereas another class represents considerably effective drugs with a count of 928. Class represents moderately effective drugs with a count of 415. In addition, class represents marginally effective drugs with a count of 187. Besides, class represents ineffective drugs with a count of 247.

Table 2: Effectiveness Classification	
Effectiveness	**Count**
Highly Effective	1330
Considerably Effective	928
Moderately Effective	415
Marginally Effective	187
Ineffective	247
Total using data	3107

4. METHODS

This research work use the dataset with two columns first is rating and second is effectiveness. Here comparing dataset column name rating where less than 3 rating denoted by 0 which is less effective and grater then 3 rating denoted as 1 is effective there are total 1-10 rating available. After comparing adding new attribute review where binary digits 0 and 1 values are stored for further classification. Methods used in this work to develop the opinion classification system. The classification accuracy of the Nave byes method and the Logistic Regression method are evaluated by comparing.

4.1 Logistic Regression (LR)

Logistic regression is a Machine learning classification method that predicts the chances of specific classes based on some dependent variables. The logistic regression output is always between 0 and 1, which is appropriate for a binary classifier.
The standard formula for the theorem is [4]: Loggit=log (odds)

$$\text{odds} = \frac{P}{1-P}$$

$$P = \frac{\text{odds}}{1+\text{odds}}$$

$$\text{Log}\left[\frac{P}{1-P}\right] = \beta o_{+}\ \beta_{1+}\ldots\ldots\ldots\ldots\ \beta_q X_q$$

Exponential both Side:

$$\frac{P}{1-P} = e^y$$

$$\frac{1}{P} - 1 = \frac{1}{e^y}$$

$$\frac{1}{P} = \frac{1+e^y}{e^y}$$

$$P = \frac{e^y}{1+e^y}$$

Where P is probability

4.2 Naïve Byes Classifier (NB)

The Naive Bayes is a well-known machine learning algorithm for data classification. Based on its capability to work quickly and reject noise or redundant data, as well as its ease of implementation. The Bayes Theorem calculates the probability of a situation based on any foreknowledge or conditions that affect the event.

The standard formula for the theorem is [4]:

$$P(X|Y) = \frac{P(X|Y)P(Y)}{P(X)}$$

In equation l,
P(XlY) = Posterior Probability
P(YlX) = Likelihood
P(Y) = Prior probability
P(X) = Marginal probability
Y = Class Variable
X is dependent feature vector (of size) where:
$X = (x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n)$

4.3 Random Forest

The Random Forest is ensemble classifier using many decision tree models. Random forest combines the idea of bootstrapping data from a learning dataset to form training data set and selecting parameters randomly to construct decision trees [4].In Random Forest classification, the algorithm aggregates predictions from multiple decision trees using a majority vote.
Let's say we have a Random Forest classifier consisting of K decision trees. Each decision tree Tk predicts the class label for a given in put sample x as $\hat{y}k$, where k=1,2,...,k=1,2,...,K.
The Random Forest prediction $\hat{y}_k$ for the input sample x is determined by the majority vote among the predictions of all decision trees:

$$\hat{y} = \text{mode}\ (\hat{y}1, \hat{y}2, \ldots\ldots\ldots, \hat{y}k)$$

Where,

- $\hat{y}_i$ is the predicted class label by the i^{th} decision tree.
- Mode represents the function that returns the most common label model among the predictions. Majority voting scheme helps to ensure robustness and improve the generalization performance of the Random Forest classifier.

5. RESULT

Many approaches are used to assess the quality of opinion classification methods. In this work, datasets were split into 75% of training data is 2195 and 25% of testing data is 912. The results obtained for the test dataset are evaluated using Naïve bias, Logistic regression and Random Forest models. Positive reviews total 2553 and negative 554 drug reviews are used. It important when making a decision on drug usage in this specific domain of drug reviews. As a result, the performance of the classifier of all positive and negative reviews must be measured independently. Thus, the predicted sentiment were measured using five metrics precision, recall, and f1-score ,accuracy and AUC on each specific class title (positive/negative) are evaluated. Describe the findings achieved as confusion matrices from all classification techniques used on each specific class label [1]. Binary predictions for each method used to assess the performance of the system, confusion matrix has been constructed. A binary prediction, which is one of the frequently used techniques for making predictions, is made up of the key elements of a ROC curve. [10].Table 3 shows the results using evaluation metrics.

Misclassification rate

Misclassification rate is a machine-learning metric that denotes the percentage of erroneous observations made by any classification system. Misclassification rate is defined as the ratio of number of wrongly classified reviews to the total number of reviews classified by the prediction method.

Misclassification Rate = (false positive + false negative) / (total predictions)

5.1 Performance measuring

- **Precision**

Precision is the percentage of correctly classified instances. The precision of positive comments is defined as the proportion of positive reviews correctly classified to the overall number of reviews. Classified as positive Low Precision means that a high percentage of the classes being classified as positive, which is not actually positive. Hence, Precision is expected to be high always [1].

$$\text{Precision} = \frac{\text{No. of correctly classified reviews}}{\text{Total no. of classified reviews}}$$

- **Recall**

Recall is defined as the ratio of number of positive reviews classified correctly to the total number of reviews [1].

$$\text{Recall} = \frac{\text{No. of correctly classified reviews}}{\text{Total no. of reviews}}$$

- **F1-Score**

F1-Score is a measure that combines precision and recall. It is the harmonic mean of precision and recall [1].

$$\text{F1-Score} = \frac{2 * (\text{Precision} * \text{recall})}{\text{Precision} + \text{recall}}$$

- **Accuracy**

Measure of accuracy with sensitivity and specificity are statistical measure of the performance of a binary classification test [1]:

$$\text{Sensitivity} = \frac{TP}{TP\ (TP+FN)}$$

$$\text{Specificity} = \frac{TN}{TN/\ (TN+FP)}$$

$$\text{Accuracy} = \frac{(TP+TN)}{TP+TN+FP+FN}$$

Table 3: Results of evaluated matrices

Model	Classes	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	Negative positive	0.75 0.89	0.61 0.82	0.67 0.90	0.85	0.72
Naïve bias	Negative Positive	0.79 0.90	0.65 0.88	0.72 0.92	0.87	0.78
Random Forest	Negative Positive	0.75 0.92	0.72 0.96	0.83 0.94	0.90	0.84

All algorithms showed results ranging from 85% to 90% accuracy. Random Forest accomplished 84.3% AUC score. Even after achieving accuracy more prominent than logistic and Naïve Byes achieved only 72% and 78% AUC score.

Predicted in test data 147 controls test review of 0 is less than 765 cases review of 1 its Area under the curve: 84.33%. Predicted in train data 407 controls train review of 0 is less than 1788 cases train review of 1its Area under the curve: 86.61%

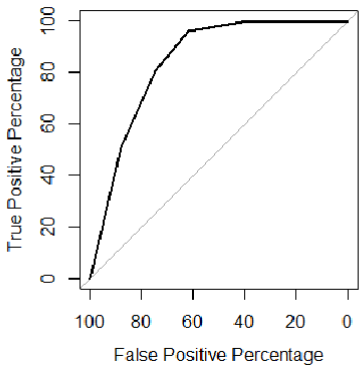

Fig: 2 ROC Curve for Predicted test data

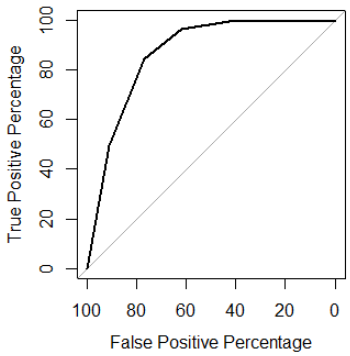

Fig: 3 ROC Curve for Predicted train data

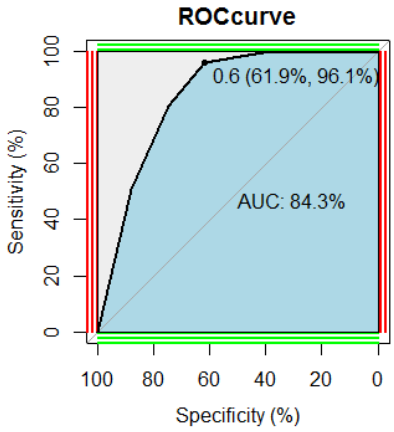

Fig: 2 ROC Curve with 0.6 threshold point

Here, Random Forest gives better result as compare to Logistic Regression and Naïve Byes Models. So Random Forest ROC Curve represent sensitivity with percentage 61.9% and Specificity with percentage 96.1% were AUC=84.3% .after comparing Accuracy of Random Forest Model Percentage shows that 0.6 threshold point give batter accuracy then 0.8.

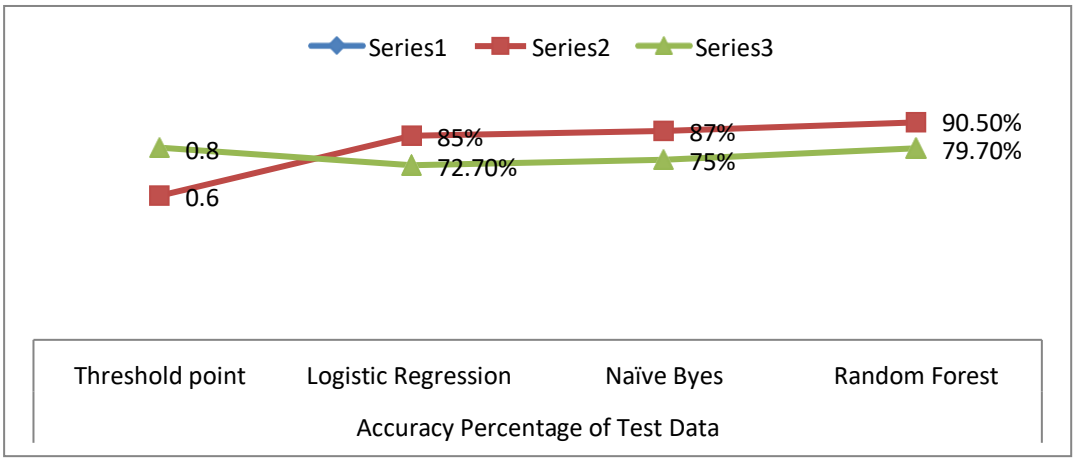

Table 9: Accuracy Percentage of Test Data			
Threshold point	**Logistic Regression**	**Naïve Byes**	**Random Forest**
0.6	85%	87%	90.5%
0.8	72.7%	75%	79.7%

Misclassification rate of Random Forest with threshold point 0.6 is 0.09.

6. CONCLUSION

This paper is focus on comparison of Logistic Regression, Naïve Byes and Random Forest algorithms for Drug review dataset. These algorithms also calculated two different Threshold points 0.6 and 0.8.For Comparison binary classification is used for drug review dataset. It focuses on analyzing the sentiments of the drug review and feeding the data to a machine learning model to training and testing the model. We evaluated them using five different metrics, precision, recall, f1score, accuracy, and AUC score. The findings indicate that the Random Forest classification method for drug-reviews has achieved the highest 90.5% classification accuracy in comparison with 87% Naïve Bayes,85% Logistic Regression classification accuracy with test data. In Future works could use Hybrid Algorithm to improve the performance of models by using real time drug review dataset.

REFRENCES

1) Vinodhini gopalakrishnan, chandrasekaranramaswamy (2 august 2017), patient opinion mining to analyze drugs satisfaction using supervised learning, journal of applied research and technology 15 (2017)311–319
2) l. mary gladence1, m. karthi and v. maria anu(august 2015), a statistical comparison of logistic regression and different bayes classification methods for machine learning, vol. 10, no. 14, august 2015
3) tomaspranckevičius, virginijusmarcinkevičius (2017), comparison of naïve bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification,

vol. 5 , no. 2, 221-232

4) mohammed nazim uddin, md.ferdous bin hafiz2 sohrabhossain shah mohammadmominulislam, drug sentiment analysis using machine learning classifiers, (ijacsa) international journal of advanced computer science and applications, vol. 13, no. 1,2022
5) satvikgarg(5 apr 2021), drug recommendation system based on sentiment analysis of drug reviews using machine learning, arxiv:2104.01113v2[cs.ir]
6) s. parthasarathy1 and v. madhu (2020), a comparative study on the logistic regression and naïve bayes models upon medical data through a machine learning approach, scientific journal 9, no.5, 2809– 2815, issn:1857-8365
7) sharonsusan jacob1, dr. r. vijayakumar(2019) , naïve bayes & logistic regression on big data: a performance analysis, international journal of applied engineering research issn 0973-4562 volume 14, number 5 , pp.1102-1105
8) t. venkatnarayanarao, anjumunnisa, kotha sreni(february 2020), medicine recommendation system based on patient reviews, international journal of scientific & technology research volume 9, issue 02, issn2277-8616
9) archive.ics.uci.edu. s. kallumadi and f. gräßer 2018. uci machine learning repository: drug review dataset (druglib.com)dataset.[Online]availableat :<https://archive.ics.uci.edu/ml/datasets/drug+review+dataset+%28druglib.com%29> [accessed 25 january2023].

10) f. kunneman, m. lambooij, a. wong, a. van den bosch, and l. mollema, "monitoring stance towards vaccination in twitter messages," bmc medical informatics and decision making, vol. 20, no. 1,feb 2020, art. no.33.