Journal of Advances and Scholarly Researches in Allied Education
Vol. 22, Issue No. 01, January-2025, ISSN 2230-7540

DOI: https://doi.org/10.29070/4sfq6836

Effectiveness of a machine learning classifier model

Sonal Sakalle [1] * **, Sakshi Rai** [2]

1. Research Scholar, LNCT University, Bhopal, M.P. , India
sonalsakalle@gmail.com ,
2. Professor, LNCT University, Bhopal, M.P., India

Abstract: As a statistical tool, predictive modelling may foretell how something will act in the future. In order to foretell how people will act in the future, machine learning has become a popular tool. Determining which of the many accessible algorithms is best suited to the data at hand is an intriguing challenge. The field of study that finds the greatest value in predictive modelling is educational data mining. Accurately predicting undergraduate students' grades has several benefits for both students and teachers. Students might be more motivated to choose their future endeavours with the support of early prediction. Using data gathered from undergraduate studies, this study displays the outcomes of many machine learning methods. It uses data obtained from undergraduate studies to assess the efficacy of several machine learning methods. Choosing the proper characteristics and the right prediction algorithm are two big problems that are addressed.

Keywords: Machine Learning, Predictive Analytics, Algorithms

X

INTRODUCTION

Organisations go through a series of stages known as the recruiting process to find, assess, and ultimately choose new personnel. Job advertisements, screening of resumes, first interview, testing, background checks, final interview, offer, and acceptance are the usual steps. Finding the right person for the job and making sure they have an easy transition into the firm are the main goals of the process, which might change based on the size, culture, and nature of the organisation.

In order to produce data-driven, impartial hiring judgements, machine learning (ML) is seeing increased application in the recruiting process. This method is gaining traction in the recruiting industry as a result of its many advantages, including the possibility of more accuracy, less prejudice, more efficiency, lower costs, and more consistency. Machine learning algorithms are able to sift through mountains of data in search of correlations and patterns that humans may miss. As a result, the quality of hiring may be enhanced and forecasts about applicant success can be made with better precision.

It is possible to increase efficiency and decrease the likelihood of bias and prejudice in the recruiting process by automating some of the manual labour involved.

This article's focus is on using decision tree algorithms to test out various strategies for making hiring and rejection decisions. Evaluation metrics should be used to assess each algorithm's efficacy. To make the best hiring or rejection judgements, use the algorithm with the maximum efficiency.ML, data mining being the most important.

It is possible for humans to make errors while attempting to build correlations between various variables or

when conducting analysis. This makes it hard for them to figure out how to fix specific issues.In many cases, these issues may be effectively addressed by using machine learning, which in turn improves system efficiency and machine design.The feature set employed by machine learning algorithms is consistent across all datasets.

The characteristics might be real, hypothetical, or both. Learning is said to be supervised when examples are provided with known labels, or the matching accurate outputs as opposed to unsupervised learning, which occurs when instances are not. Researchers want to find new, potentially valuable classes of things by using these unsupervised (clustering) techniques (Jain et al., 1999).

LITERATURE REVIEW

Acharjya, D. P. (2020), To correctly diagnose thyroid illness, the CS method employs a number of ensemble classifiers. Sensitivity, specificity, and accuracy are used to assess the performance of various ensemble classifiers, including those based on bagging, stacking, boosting, and voting, all of which are considered in this study. The best classification method is the stacking ensemble.

Narayan, S., & Sathiyamoorthy, E. (2019), In order to identify the important characteristics of Ischemic Stroke and the most efficient methods of intervention and therapy, a DSS has been established. The writers of this piece take into account the global stroke database. The Shapiro-Wilk method and Pearson correlations are used to determine whether traits are significant. In the realm of prediction, several methods are used, including VC, MLP, RF, and AdaBoost.

A novel classifier was created by Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G. and Finn, D. P. (2019) that makes use of a decision tree and the greedy step-wise technique to increase precision. The classifiers here employ the greedy stepwise method to zero down on the most important features, while the decision tree approach is put to use in the prediction phase. Japanese stroke patients are used to test the effectiveness of the suggested classifier. A higher rate of accuracy is achieved by the aforementioned method.

Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., & Akbari, E. (2019), For the treatment of stroke, a smart ensemble approach is created. Authors identify the semantic and syntactic link among key aspects of stroke illness. In all, 507 patients are taken into account here. The patient's medical record provides the list of stroke symptoms. Information from medical records is mined using tagging and entropy methods. In addition, SVM, boosting, random forest, and bagging approaches are utilized to improve the accuracy of predicting stroke patients. Compared to SVM, boosting, bagging, and random forest, ANN produced much superior results.

Different data mining methods were used by Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., and Lyu, C. (2019) to forecast the occurrence of ischemic stroke. In addition to the health records of 80 stroke patients and 112 controls, this dataset also includes 16 characteristics for predicting ischemic stroke. SVM, SGB, and PLR are three machine learning classifiers used to differentiate between stroke patients and healthy individuals. The authors argue that SVM approaches provide the most encouraging outcomes.

RESEARCH METHODOLOGY

Data collection and pre-processing

Data is first retrieved from many healthcare facilities' online forums in Coimbatore, India. The data was collected by means of a Python script that was developed utilizing the Outscraper.com API. We started out collecting ratings in Excel, but we eventually switched to CSV. Every review, whether original or retweeted, has been carefully categorized as favorable, negative, or neutral. We have omitted independent reviews. Eighty percent of the data came from training, and twenty percent came from tests, at six different hospitals in and around Coimbatore.

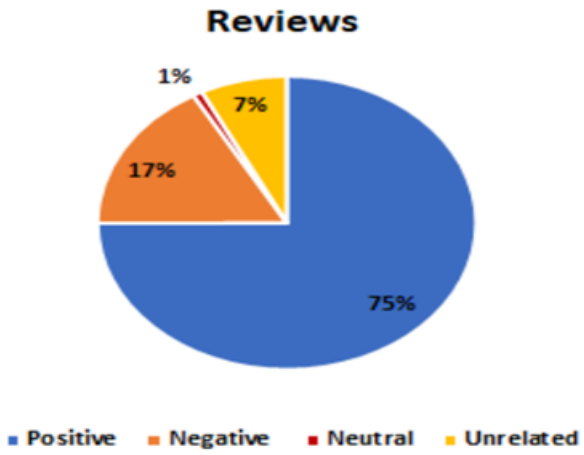

Figure 1: Class Distribution on Reviews

Proposed architecture

Each review is carefully classified as positive, negative, or neutral after removing irrelevant ones. Original and repeated reviews are then included. The opinions expressed by six medical facilities in the Coimbatore area were taken into account. Figure 3.3 shows the results of the preprocessing of 1,350 reviews, as mentioned before. There were 1250 reviews—positive, negative, and neutral—that made it to processing after pre-processing and duplicate removal.

Data Set on Breast Cancer

The "Breast Cancer Wisconsin (Diagnostic) Data Set" from Kaggle was used for the purpose of making predictions on the prevalence of breast cancer. There is one target feature and thirty-one medical predictor features in this dataset. Diagnostic, id, radius-mean, texture-mean, perimeter-mean, area-mean, smoothness-mean, compact-mean, concavity-mean, and concavity points-mean are a few of the essential qualities

ANALYSIS

Machine Learning classifier

Logistic Regression

The models' outputs are either 0 or 1, and the binary classification issue is solved using Logistic Regression. The selected feature successfully predicts patient satisfaction towards different HHCS, as shown by the model's 98.01 accuracy and 98% model performance recall score. The results are shown in Figure 2 and Table 1.

Logistic Regression Accuracy:

98.01563493181

Table 1. Logistic regression classification report

	Precision	Recall	F1-Score
1	0.98	0.98	0.98
2	0.98	0.97	0.98
3	0.97	0.98	0.97
4	0.98	0.98	0.98
5	0.98	0.98	0.98

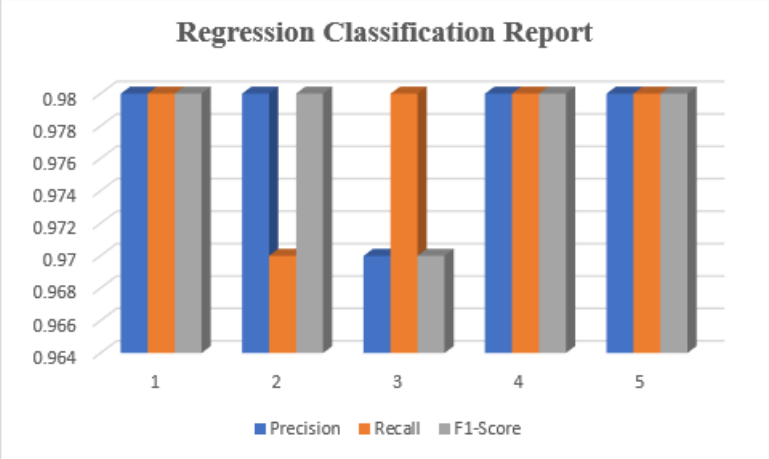

Figure 2: Classification report of Logistic regression

Random Forest

A feature-based hierarchical structure known as a decision tree is the theoretical basis of the RF classifier. A subset of characteristics associated with each node in the decision tree are used to categories them. The dataset samples were used to generate a set of decision trees. The chosen feature subset's Gini index is used to partition the nodes.

Random forest Accuracy:

99.86801583809944

Table 2 Random Forest

	Precision	Recall	F1-Score
1	1.00	1.00	1.00
2	1.00	0.99	1.00
3	0.99	1.00	0.99
4	1.00	1.00	1.00
5	1.00	1.00	1.00

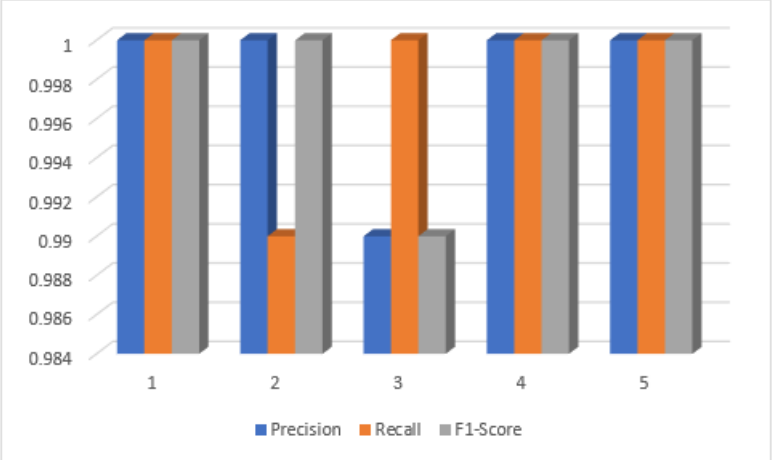

Figure 3: Classification report of Random Forest

Gradient Boost

As a machine learning classifier algorithm, it trains many underperforming learners to boost each other's performance. A new tree is developed at each level by fixing the mistakes made by earlier trees. The top healthcare-related machine learning algorithm produced the output.

Gradient Boost Accuracy: 99.8240211174659

Table 3: Gradient boost

	Precision	Recall	F1-Score
1	1.00	1.00	1.00
2	1.00	0.99	1.00
3	0.99	1.00	0.99
4	1.00	1.00	1.00

5	1.00	1.00	1.00

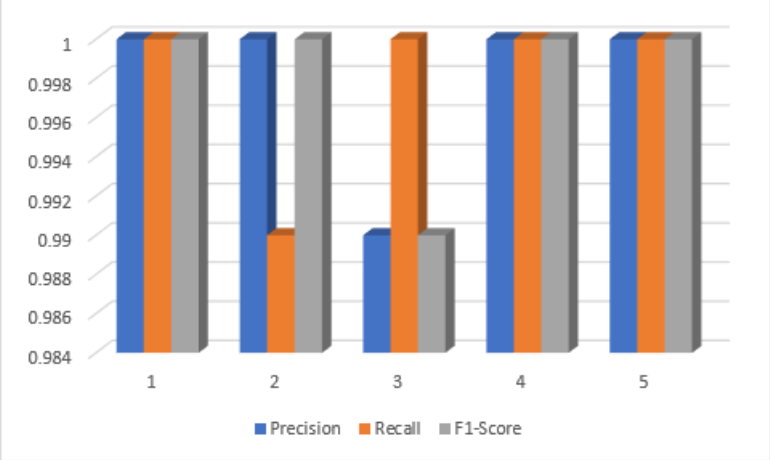

Figure 4: Classification report of Gradient Boost

Ada Boost

Among machine learning classifiers, AdaBoost is well-known. Together, the weak classifier and the majority voting system are used in conjunction with the boosting strategy. Training accuracy is directly proportional to the weights used in the final classifier's voting scheme. Among the classifiers that surpass many ML classifiers is the weights derived from the most recent research.

Ada Boost Accuracy: 86.8675758908931

Table 4. Ada Boost

	Precision	**Recall**	**F1-Score**
1	1.00	1.00	1.00
2	1.00	0.99	1.00
3	0.99	1.00	0.99
4	1.00	1.00	1.00
5	1.00	1.00	1.00

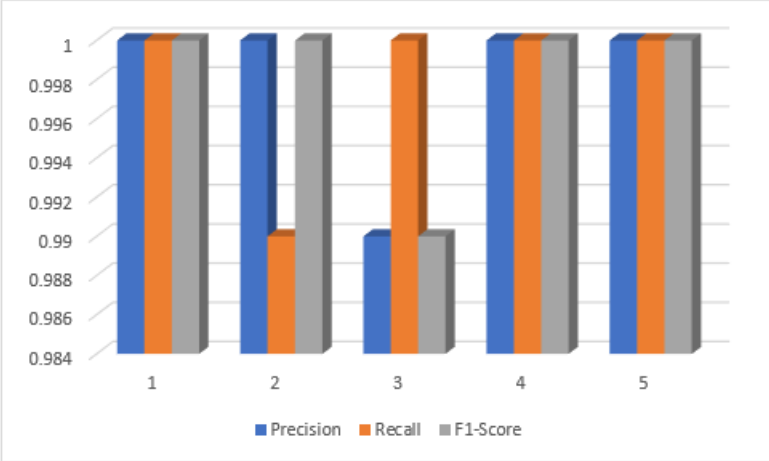

Figure 5: Classification report of Ada Boost

Model Validation

K-fold Cross validation

One way to train and validate an ML model is by cross validation, which involves comparing different models. In our study, we train and verify models like LR, RF, GB, and ADB using their accuracy score. In order to prevent over-fitting that happens while training with a limited data set, the model makes use of cross-validation. One way to validate a model is to divide samples into two groups: one to train the model and another to test it. We separated the data subsets into ten equal parts since we utilised 10-fold validation. A total of ten data subsets were used for the model's training and validation processes. The model was validated using the remaining data subset, out of a total of nine. Ten models, on average, are the output of this validation procedure. By using k=10 in K-fold cross validation, we find that GB and RF outperform LR (86.04) and ADB (86.8%), with an accuracy of 99.8%.

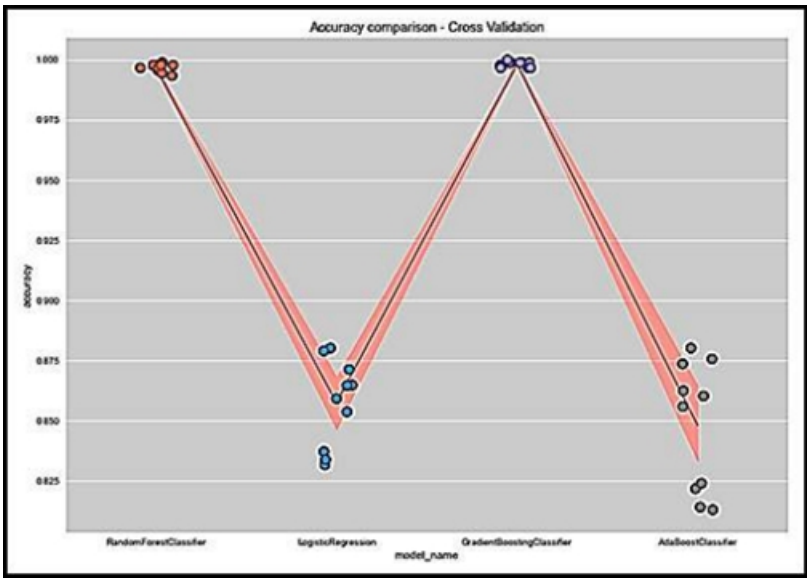

Figure 6 show the accuracy comparison using cross validation.

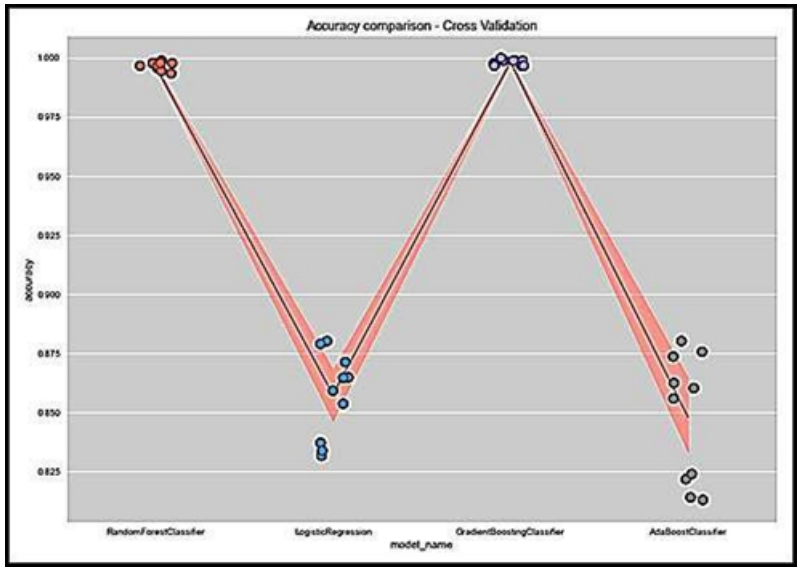

Figure 7: Model Comparison

Model Performance Measure

The accuracy of the predicted model's performance and performance indicators are used to assess ML classification algorithms. Several ML classification methods, including LR, RF, GB, and ADB, are used to forecast the model's performance. The performance measures used in this study endeavor are analysed using the ROC curve value. When it comes to patient satisfaction with HHCS, RF and GB stand head and shoulders above all other models (Table 4.11). The ROC value-based model performance is shown in Figure 4.11.

Table 5: Model performance using ROC

Models	ROC
LR	0.758933127315834
RF	0.9960827310675624
GB	0.9972869097127411
AdaB	0.733487225430504

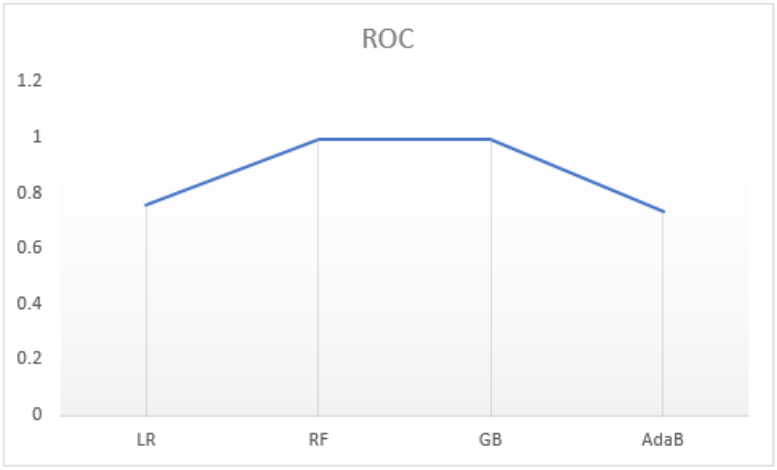

Figure 8: ROC

CONCLUSION

Opinion polling Two sides of the same coin: opinion mining. The process of extracting useful information from large datasets (e.g., reviews, surveys, comments, etc.) is known as data mining. This information is useful for understanding the sentiments expressed in the comments and for assessing the service's ability to meet the needs of its users. When it comes to determining the domain's market worth from the user's point of view, these online forum conversations now play a huge role. A lot of work goes into deciphering these viewpoints and drawing useful conclusions from the data they provide. According to their needs, the user is receiving the suggestion.

In recent years, recommender systems that incorporate sentiment analysis have grown in popularity as useful resources for consumers looking to find and assess products. These systems utilized a variety of methods to help people choose things that are perfect for them. Data mining approaches are being applied more and more in hybrid systems to enhance recommendations, even while popular CF-based algorithms still offer relevant, personalized answers in certain fields. Successful apps and standalone recommenders have successfully generated accurate suggestions in challenging domains in the past. In addition, apps' focus has shifted away from advising what to eat due to changes in machine learning classifier methods. Though they may have been nothing more than a fleeting trend at one point, recommender systems are clearly here to stay, and the algorithms behind them can and will play a significant role in making recommendations based on sentiment analysis.

The purpose of this study is to examine patient health care opinion systems using a new SCSP ensemble model. By using a machine learning strategy to forecast superior models in data analysis, the classification models categories patients' emotions as positive, negative, or neutral. Then, this study presents a new feature selection approach that uses patient satisfaction data to find the most important aspect of home health care services. Here, we took a look at the many parts that go into determining which health care services are better using various measures. The next section of this study proposal.

References

1. Acharjya, D. P. (2020). A Hybrid Scheme for Heart Disease Diagnosis Using Rough Set and Cuckoo Search Technique. Journal of Medical Systems, 44(1), 27.

2. Narayan, S., & Sathiyamoorthy, E. (2019). A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases. Neural Computing and Applications, 31(1), 93-102.

3. Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., ... & Finn, D. P. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. Expert systems with applications, 130, 157-171.

4. Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., & Akbari, E. (2019). A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. Journal of infection and public health, 12(1), 13-20.

5. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). A reliable method for colorectal cancer prediction based on feature selection and support vector machine. Medical & biological engineering & computing, 57(4), 901-912.

6. Vinodhini Gopalakrishnan et.al.,(2017),"Patient opinion mining to analyze drugs satisfaction using supervised learning",Journal of Applied Research and Technology, Vol.15, Iss 4,pp.311-319.

7. Wang, D., et.al.,(2018). A content-based recommender system for computer science publications. Knowledge-Based Systems, Vol.157, pp.1–9.

8. Zineb Nassr, et.al.,(2019), A comparative study of sentiment analysis approaches. ,SCA '19. Vol.91, pp.1–8.