International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

P-Value Feature Selection Technique for Prediction of Student Performance

Ritu Aggrawal[a], *Saurabh Pal*[b]*

[a]*Research Scholar,VBS PU Jaunpur,U.P., India.*
[b]*Head Dept. of MCA, VBS PU Jaunpur,U.P., India.*

ABSTRACT

Applications of Machine learning algorithms increased the growth in various fields like disease prediction, student's performance prediction, and crop productions prediction and in various other fields. Early prediction of students' performance can help high-risk students to take attention. We propose a student performance prediction model to predict high risk students which take special attention in study.We have used logistic regression (LR), K-nearest neighbor (KNN), Naïve Bayes (NB), decision tree (DT), random forest (RF) and gradient boosting classifier (GBC) on student dataset. The dataset is taken from VBS Purvanchal University, Jaunpur. P-value feature selection method is used to select important features which play major role in prediction. A subset of the original dataset is obtained after selecting important features to compare the results of used six machine learning techniques and ensemble approach as on whole dataset. We calculated the accuracy of the classifiers, the confusion matrix; ROC curve and AUC value are for the model verification to show the performance of the proposed model.

Keywords: Logistic Regression (LR), K-nearest neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF) and Gradient Boosting Classifier (GBC)

1. Introduction

Machine learning algorithms help to identify and understand huge data sets. These learning algorithms classify large problems into class-level sub-problems in the target variable. Each class in the target variable provides information [130-135].

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis. The p-value is actually the probability of getting a sample IF the null hypothesis is true. So, we assume the null hypothesis is true and then determine how "strange" our sample really is. If it is not that strange (a large p-value) then we don't change our mind about the null hypothesis. As the p-value gets smaller, we start wondering if the null really is true and well may be we should change our minds (and reject the null hypothesis) [1]. Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output.

1.1 Steps of Backward Elimination:

Below are some main steps which are used to apply backward elimination process:
Step-1: Firstly, We need to select a significance level to stay in the model. (SL=0.05)
Step-2: Fit the complete model with all possible predictors/independent variables.
Step-3: Choose the predictor which has the highest P-value, such that.
If P-value >SL, go to step 4.
Else Finish and Our model is ready.
Step-4: Remove that predictor.

* *Corresponding author.*
E-mail address: drsaurabhpal@yahoo.co.in

Step-5: Rebuild and fit the model with the remaining variables.

- **Iteration Log**

This is a listing of the log likelihoods at each iteration. (Remember that logistic regression uses maximum likelihood, which is an iterative procedure.) The first iteration (called iteration 0) is the log likelihood of the "null" or "empty" model; that is, a model with no predictors. At the next iteration, the predictor(s) are included in the model. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have "converged", the iterating is stopped and the results are displayed [2].

- **Log Likelihood**

This is the log likelihood of the final model. The value has no meaning in and of itself; rather, this number can be used to help compare nested models. [3].

- **Number of Observation**

This is the number of observations that were used in the analysis. This number may be smaller than the total number of observations in your data set if you have missing values for any of the variables used in the logistic regression. Statistics uses a list wise deletion by default, which means that if there is a missing value for any variable in the logistic regression, the entire case will be excluded from the analysis[4-5].

- **LR Chi2(3)**

This is the likelihood ratio (LR) chi-square test. The likelihood chi-square test statistic can be calculated. This is minus two (i.e., -2) times the difference between the starting and ending log likelihood. The number in the parenthesis indicates the number of degrees of freedom. In this model, there are three predictors, so there are three degrees of freedom [6-8].

- **Prob > Chi2**

This is the probability of obtaining the chi-square statistic given that the null hypothesis is true. In other words, this is the probability of obtaining this chi-square statistic if there is in fact no effect of the independent variables, taken together, on the dependent variable. This is, of course, the p-value, which is compared to a critical value, perhaps .05 or .01 to determine if the overall model is statistically significant. In this case, the model is statistically significant because the p-value is less than .000 [9-11].

- **Pseudo R-squared**

This is the pseudo R-squared. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in OLS regression (the proportion of variance explained by the predictors), suggested to interpreting this statistic with great caution [12].

- **Honcomp**

This is the dependent variable in our logistic regression. The variables listed below it are the independent variables[13.

- **Coef.**

These are the values for the logistic regression equation for predicting the dependent variable from the independent variable. They are in log-odds units. Similar to OLS regression, the prediction equation is[14-15]

$$logit(p) = log\left(\frac{p}{1} - p\right)$$

$$= b_0 + b_1 * Sexmale + b_2 * age + b_3 * cigsPerDay + b_4 * totChol + b_5 * sysBP + b_6 * glucose \qquad (1)$$

where p is the probability of being in honors composition [16]. Expressed in terms of the variables used in this example, the logistic regression equation is

$$log(p/1 - p) = -9.1264 + 0.5815 * Sexmale + 0.0655 * age + 0.0197 * cigsPerDay + 0.0023 * totChol + 0.0174 * sysBP + 0.0076 * glucose \qquad (2)$$

- **Std. Err.**

These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from

0; by dividing the parameter estimate by the standard error you obtain a z-value (see the column with z-values and p-values). The standard errors can also be used to form a confidence interval for the parameter, as shown in the last two columns of this table [17].

- **z and P>|z| Values**

These columns provide the z-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0. If we use a 2-tailed test, then compare each p-value to preselected value of alpha. Coefficients having p-values less than alpha are statistically significant. For example, if chose alpha to be 0.05, coefficients having a p-value of 0.05 or less would be statistically significant. If we use a 1-tailed test then we can divide the p-value by 2 before comparing it to preselected alpha level. With a 2-tailed test and alpha of 0.05, we may reject the null hypothesis that the coefficient for female is equal to 0 [18].

- **Odds ratio (OR) and Logistic Regression (LR)**

Odds are determined from probabilities and range between 0 and infinity. Odds are defined as the ratio of the probability of success and the probability of failure.

$$OR = \frac{p}{1-p} \quad (3)$$

At the point when a LR is determined, the LR coefficient (b) is the assessed increment in the log chances of the result per unit increment in the estimation of the presentation [19].

- **Confidence Intervals (CI)**

A confidence interval is a bound on the estimate of a population variable. It is an interval statistic used to quantify the uncertainty on an estimate. [20]. A confidence interval is different from a tolerance interval that describes the bounds of data sampled from the distribution. It is also different from a prediction interval that describes the bounds on a single observation. Instead, the confidence interval provides bounds on a population parameter, such as a mean, standard deviation, or similar.

$$Upper\ 95\%\ CI = e^\wedge[ln(OR) + 1.96\sqrt{{}^{1}/_{a} + {}^{1}/_{b} + {}^{1}/_{c} + {}^{1}/_{d}}] \quad (4)$$

$$Lower\ 95\%\ CI = e^\wedge[ln(OR) - 1.96\sqrt{{}^{1}/_{a} + {}^{1}/_{b} + {}^{1}/_{c} + {}^{1}/_{d}}] \quad (5)$$

- **Model Validation**

Model validation is the set of processes and activities intended to verify that models are performing as expected. The developed model is evaluated on the basis of different performance metrics and the model is validated according to these performance values. The testing of the model to validate the model is fundamental of testing [21-22]].

2. Methodology

The student data set is collected from the BCA students of VBS Purvanchal University, Jaunpur. The goal of this paper is to predict whether the student is at risk of under performance. The data set provides student data. It contains 1000 records and 22 attributes.

Logistic regression classifier is implemented using python language to calculate the results of prediction. Standard error, z-value, p-value, and confidence interval (25-95%) are evaluated and only those features were selected whose p value is less than 0.5. Six machine learning algorithms (such as logistic regression, random forest, decision tree, naive Bayes, k-nearest neighbor, and gradient boosting) were applied to obtain performance of proposed model in terms of accuracy. The model is developed using the following steps as shown in Fig. 1.

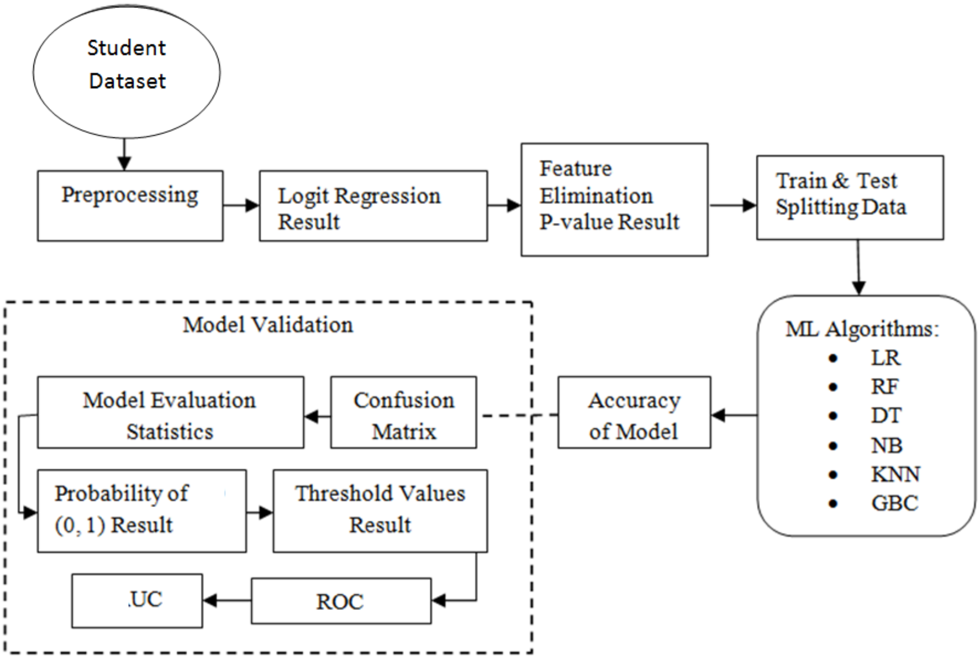


Fig. 1 Procedure for developing the model

Experimental Setup

The data for prediction of student performance is taken from VBS Purvanchal University, Jaunpur. The data set contains 1000 records and 22 attributes. Table 1 show the different attributes used in this study.

3. Results and Discussion

Six machine learning algorithms are applied for finding the predictions for student datasets using p-value back elimination feature selection techniques. In this experiment, we choose the student data subset by choosing only most important attributes. Logistic regression is used for calculating the p-values to choose the important attributes [23].

The calculated values of coef., std err, z P, are shown in Table 2.

Table: 2 Statistical Significance of Attributes

7	coef	std err	z	P>|z|	[0.025	0.975]
S1	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
S2	0.5742	0.107	5.345	0.000	0.364	0.785
S3	0.0641	0.007	9.799	0.060	0.051	0.077
S4	0.0739	0.155	0.478	0.633	-0.229	0.377
S5	0.0184	0.006	3.000	0.003	0.006	0.030
S6	0.1448	0.232	0.623	0.533	-0.310	0.600
S7	0.7193	0.489	1.471	0.141	-0.239	1.678
S8	0.2142	0.136	1.571	0.116	-0.053	0.481
S9	0.0022	0.312	0.007	0.994	-0.610	0.614
S10	0.0023	0.001	2.081	0.057	0.000	0.004
S11	0.0154	0.004	4.082	0.000	0.008	0.023
S12	-0.0040	0.006	-0.623	0.533	-0.016	0.009
S13	0.0103	0.013	0.827	0.408	-0.014	0.035
S14	-0.0023	0.004	-0.549	0.583	-0.010	0.006
S15	0.0076	0.002	3.409	0.001	0.003	0.012
S16	0.0021	0.002	2.087	0.054	0.001	0.004
S17	0.0155	0.003	4.082	0.000	0.009	0.023
S18	-0.0043	0.005	-0.623	0.537	-0.016	0.007
S19	0.0102	0.013	0.829	0.408	-0.014	0.038
S20	-0.0023	0.005	-0.549	0.584	-0.010	0.004
S21	0.0078	0.001	3.407	0.002	0.003	0.012

Table 3 shows the P value and the corresponding statistics after backward feature elimination techniques. Only important features with a value of P less than 0.05 are selected.

Table: 3 Selected Attributes Based on P-Value

	coef	std err	z	P>|z|	[0.025	0.975]
S1	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
S2	0.5742	0.107	5.345	0.000	0.364	0.785
S5	0.0184	0.006	3.000	0.003	0.006	0.030
S11	0.0154	0.004	4.082	0.000	0.008	0.023
S15	0.0076	0.002	3.409	0.001	0.003	0.012
S17	0.0155	0.003	4.082	0.000	0.009	0.023
S21	0.0078	0.001	3.407	0.002	0.003	0.012

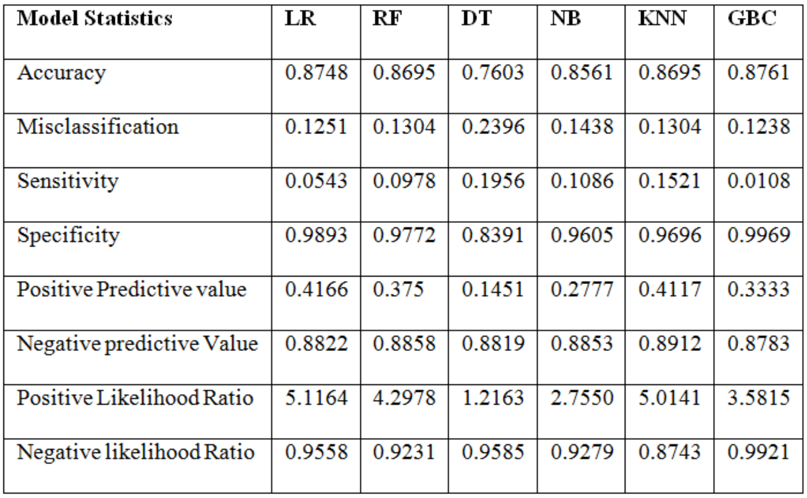
In Table 4 below, the odds ratio, confidence interval and P value are calculated.

Table: 4 Results of ODDS Ratio, Confidence Intervals and P-Values

	CI 95% (2.5%)	CI 95% (97.5%)	Odds Ratio	P-value
S1	0.000043	0.000272	1.788687	0.000
S2	1.455242	2.198536	0.000109	0.000
S5	1.054483	1.080969	1.017529	0.000
S11	1.011733	1.028128	1.019897	0.000
S15	1.000158	1.004394	1.002273	0.000
S17	1.013292	1.021784	1.067644	0.000
S21	1.004346	1.010898	1.007617	0.000

The values of the developed model shows that boys student (sex = 1) are 1.788687 more likely to be identified as having under performer than girls students (sex = 0) when all the other different characteristics remain the same. If we convert them in the percentage change, we will predict that boy's chances are 78.80% higher than girl's chances. If we take the grade in senior secondary school coefficient keeping all other attributes constant the predicted value is 1.067644, which will show that a 7% of increase in the chance of being increased in performance. In addition, those students who use laptop at home have a high chance of getting better performance by 2%, and for class calibration and category of students do not play an important role in the performance of students. Among the 22 features, we selected only 7 features for analysis through backward elimination based on P-value features [24]. A measurable investigation of the information was carried out, and an unmistakable measurement was solved for the segment and performance dominant factors. As shown in Table 5, the accuracy, misclassification, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio and negative likelihood ratio of the classifier are calculated. The accuracy of the GBC classifier is higher, so there are fewer classification errors.

Table: 5 Model Statistics



Model Statistics	**LR**	**RF**	**DT**	**NB**	**KNN**	**GBC**
Accuracy	0.8748	0.8695	0.7603	0.8561	0.8695	0.8761
Misclassification	0.1251	0.1304	0.2396	0.1438	0.1304	0.1238
Sensitivity	0.0543	0.0978	0.1956	0.1086	0.1521	0.0108
Specificity	0.9893	0.9772	0.8391	0.9605	0.9696	0.9969
Positive Predictive value	0.4166	0.375	0.1451	0.2777	0.4117	0.3333
Negative predictive Value	0.8822	0.8858	0.8819	0.8853	0.8912	0.8783
Positive Likelihood Ratio	5.1164	4.2978	1.2163	2.7550	5.0141	3.5815
Negative likelihood Ratio	0.9558	0.9231	0.9585	0.9279	0.8743	0.9921

The graphical representation of the values discussed in Table 5 is shown in Fig. 2. The highest accuracy achieved by gradient boosting classifier which is 87.61%. The other classifiers are achieving accuracy 87.48%, 86.95%, 86.95%, 85.61% and 76.03% in the decreasing order for logistic regression, random forest, k-nearest neighbor, naïve bayes and decision tree respectively.

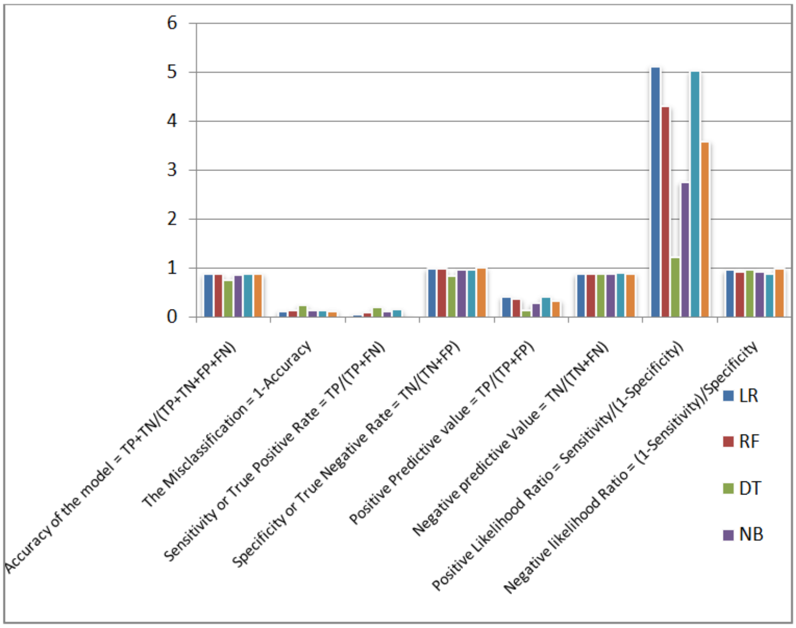


Fig. 2 Graphical Presentation of Model Statistics

Threshold Values Prediction (0.5), Since the model is forecasting students performance too many sort II errors isn't fitting. A False Negative is more dangerous than a False Positive for this situation. Consequently in order to expand the affectability, threshold can be brought down.The confusion matrix for Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbor and Gradient Boosting are shown in Fig. 3

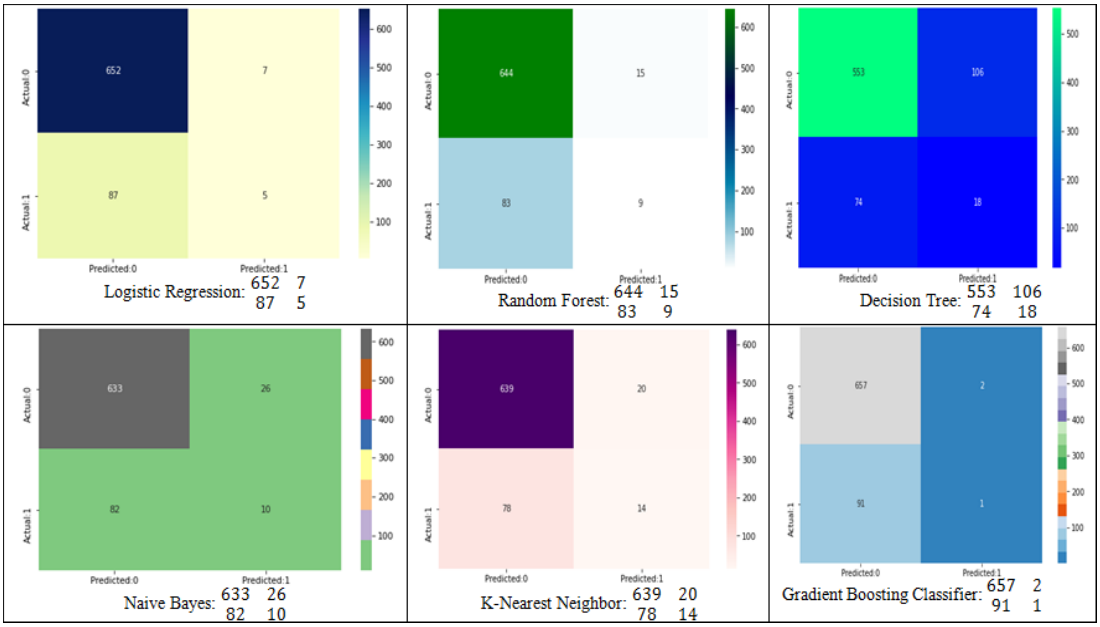


Fig. 3 Confusion Matrix of Classifiers

Various thresholds values are calculated using ROC and AUC curve. ROC is a plot of true positive rate with false positive rate on the two axis. In ideal situation ROC curve is towards the upper left corner of the figure where the true positive and false positive are at ideal levels. The threshold value is when increase in false positive is not seen.The area covered under the ROC curve calculates the precision of developed model, the more the area covered shows the difference between true positive and false negative.Fig. 4 show the ROC for different classifiers.

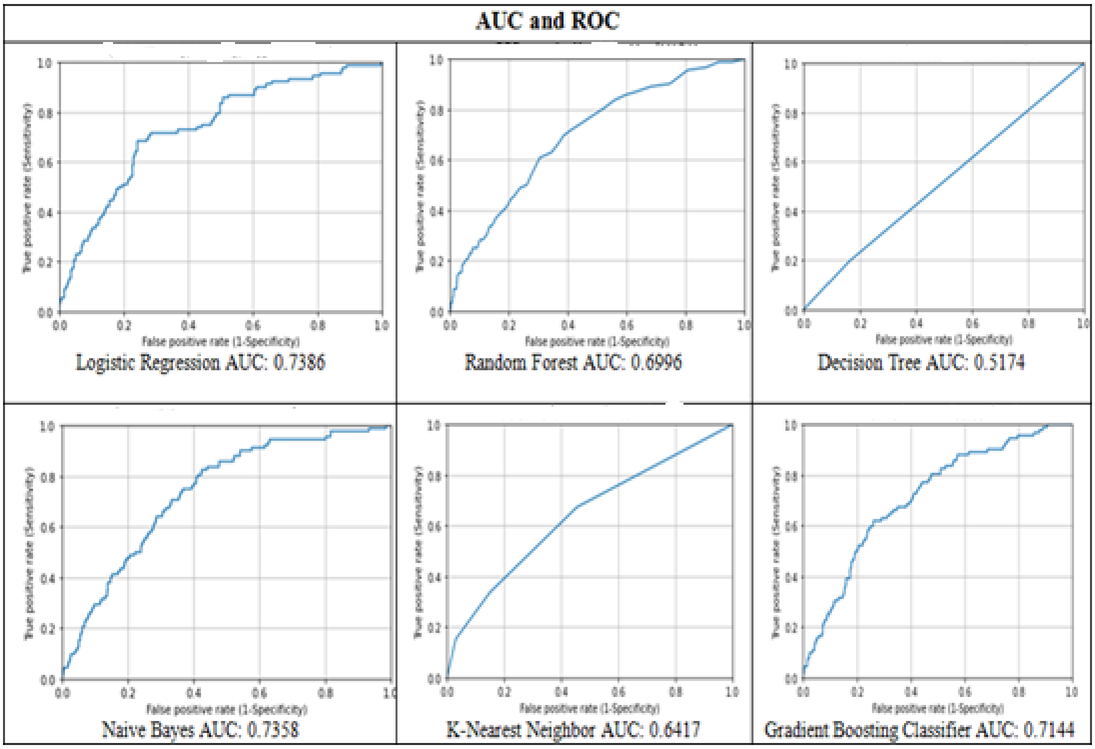


Fig. 4 ROC and AUC of the Classifier

4. Conclusion

We have used logistic regression (LR), K-nearest neighbor (KNN), Naïve Bayes (NB), decision tree (DT), random forest (RF) and gradient boosting classifier (GBC) on stuent dataset for the study. P-value feature selection method is used to select important features which play major role in prediction. A subset of the original dataset is obtained after selecting important features to compare the results of used six machine learning techniques and ensemble approach as on whole dataset. We calculated the accuracy of the classifiers, the confusion matrix, ROC curve and AUC value are for the model verification to show the performance of the proposed model.The developed model is based on p-value based statistical feature selection and six machine learning classifiers, such as logistic regression, random forest, decision tree, naive Bayes, k nearest neighbor and gradient boosting classifier. The highest accuracy achieved by gradient boosting classifier which is 87.61%. The other classifiers are achieving accuracy 87.48%, 86.95%, 86.95%, 85.61% and 76.03% in the decreasing order for logistic regression, random forest, k-nearest neighbor, naïve bayes and decision tree respectively.

The main findings of this paper are:

- Important features are chosen for P-Value less than 5%.
- The values of the developed model shows that boys student (sex = 1) are 1.788687 more likely to be identified as having under performer than girls students (sex = 0) when all the other different characteristics remain the same.
- If we take the grade in senior secondary school coefficient keeping all other attributes constant the predicted value is 1.067644, which will show that a 7% of increase in the chance of being increased in performance.
- The developed model achieved highest accuracy of 87.61% in the case of gradient boosting classifiers.
- The model is more specific than sensitive and the area covered by ROC is 73.86%.
- Those students who use laptop at home have a high chance of getting better performance by 2%, and for class calibration and category of students do not play an important role in the performance of students.

REFERENCES

1. Hashemi, M., Azizinezhad, M., Najafi, V., and Nesari, A. J. (2011). What is mo-bile learning? challenges and capabilities.Procedia-Social and Behavioral Sciences,30:2477–2481.
2. Harichandan, S. (2009). Role of mobile technology in learning and teaching. InCollectedConference Papers and Abstracts September 2009, page 209.
3. Evjemo, B., Akselsen, S., Sletteme˚as, D., Munch-Ellingsen, A., Andersen, A., andKarlsen, R. (2014). I expect smart services!”: User feedback on nfc based servicesaddressing everyday routines.Mobility and Smart Cities, Mobility IoT.
4. Donner, J. (2008). Research approaches to mobile use in the developing world: A reviewof the literature.The information society, 24(3):140–159.
5. Wade, R. H. (2002). Bridging the digital divide: new route to development or new formof dependency?Global governance, pages 443–466.
6. Samuel, J., Shah, N., and Hadingham, W. (2005). Mobile communications in southafrica, tanzania, and egypt: Results from community and business surveys.Africa:the impact of mobile phones, 2:44–52.
7. Jensen, R. (2007). The digital provide: Information (technology), market performance,and welfare in the south indian fisheries sector.The quarterly journal of economics,pages 879–924.
8. Islam, M. S. and Gr¨onlund ,˚A. (2011). Bangladesh calling: farmers’ technology use prac-tices as a driver for development.Information Technology for Development, 17(2):95–111.
9. Chepken, C. (2012). Telecommuting in the developing world: a case of the day-labourmarket.
10. Buku, M. W. and Meredith, M. W. (2012). Safaricom and m-pesa in kenya: financialinclusion and financial integrity.Wash. Jl tech. & arts, 8:375.
11. Porteous, D. (2011). The enabling environment for mobile banking in africa, bankable-frontier.
12. Sife, A. S., Kiondo, E., and Lyimo-Macha, J. G. (2010). Contribution of mobile phones torural livelihoods and poverty reduction in morogoro region, tanzania.The ElectronicJournal of Information Systems in Developing Countries, 42.
13. Ssembatya, R. (2014). Designing an architecture for secure sharing of personal healthrecords: a case of developing countries.
14. Duncombe, R. (2012). Understanding mobile phone impact on livelihoods in developingcountries: A new research framework.
15. Donner, J. (2009). Mobile-based livelihood services in africa: pilots and early deploy-ments.Communication technologies in Latin America and Africa: A multidisciplinaryperspective, pages 37–58.
16. Gakuru, M., Winters, K., and Stepman, F. (2009). Innovative farmer advisory servicesusing ict.documento presentado en el taller de W3C “Africa perspective on the roleof movile technologies in fostering social development”, Maputo, 1.
17. Hughes, N. and Lonie, S. (2007). M-pesa: mobile money for the “unbanked” turningcellphones into 24-hour tellers in kenya.Innovations, 2(1-2):63–81.
18. Kafyulilo, A. (2014). Access, use and perceptions of teachers and students towards mo-bile phones as a tool for teaching and learning in tanzania.Education and InformationTechnologies, 19(1):115–127.
19. Traxler, J. and Kukulska-Julme, A. (2005). Mobile learning in developing countries. WMUTE’06. FourthIEEE International Workshop on, pages 98–102.
20. Collis, B. and Moonen, J. (2001).Flexible learning in a digital world: Experiences andexpectations. Psychology Press.
21. Pachler, N., Bachmair, B., & Cook, J. (2009). Mobile learning: Structures, agency, practices. Springer Science & Business Media.
22. Masters, K. (2005). Low-key m-learning: a realistic introduction of m-learning to devel-oping countries. InSixth Conference on Communications in the 21st Century: Seeing,Understanding, Learning in the Mobile Age, Budapest.
23. Ford, M. and Batchelor, J. (2007). From zero to hero–is the mobile phone a viablelearning tool for Africa.
24. Swaffield, S., Jull, S., and Ampah-Mensah, A. (2013). Using mobile phone texting tosupport the capacity of school leaders in ghana to practise leadership for learning.Procedia-Social and Behavioral Sciences, 103:1295–1302.