Predicting Student Failure in University Examination using Machine Learning Algorithms

Article *in* International Journal of Innovative Technology and Exploring Engineering · March 2020

DOI: 10.35940/ijitee.E2643.039520

2 authors:

Vivek Raj
Indian Institute of Technology (Banaras Hindu University) Varanasi
1 PUBLICATION 3 CITATIONS
SEE PROFILE

Mani Vannan
SRM Institute of Science and Technology
12 PUBLICATIONS 8 CITATIONS
SEE PROFILE

Predicting Student Failure in University Examination using Machine Learning Algorithms

Vivek Raj S. N., S. K. Manivannan

Abstract: ***Student Performance Management is one of the key pillars of the higher education institutions since it directly impacts the student's career prospects and college rankings. This paper follows the path of learning analytics and educational data mining by applying machine learning techniques in student data for identifying students who are at the more likely to fail in the university examinations and thus providing needed interventions for improved student performance. The Paper uses data mining approach with 10 fold cross validation to classify students based on predictors which are demographic and social characteristics of the students. This paper compares five popular machine learning algorithms Rep Tree, Jrip, Random Forest, Random Tree, Naive Bayes algorithms based on overall classifier accuracy as well as other class specific indicators i.e. precision, recall, f-measure. Results proved that Rep tree algorithm outperformed other machine learning algorithms in classifying students who are at more likely to fail in the examinations.***

Keywords: Student Performance, Learning Analytics, Classification, Precision, Recall-Measure.

I. INTRODUCTION:

Student progression is one of the key indicators of health of higher educational institutions and also one of the major problems plaguing the education sector is increase in percentage of students who are not able to complete the degree. This is a matter of grave concern for all the stake holders involved in education sector, especially the educators who play a key role in mentoring the students to make them achieve their targets. The key challenge facing the teachers is identifying the students who are more likely to fail in the exam and provide personalized support for the students such that they can be given a better chance for passing the examinations. In this article we propose the use machine learning approach for identifying and classifying students who are the more likely to fail in the university examination. We have deployed and compared well known supervised machine learning classifiers to identify the target class and in turn to find out the best classification algorithm based various metrics like precision, recall, score and f-measure. The study is unique in the following ways as the majority of the previous studies carried out have been comparing the accuracy of the overall classifiers to choose the best classifiers but this article also includes class specific metrics to give a correct indication of classifier accuracy.

Main intention of this study is to correctly predict the students who have high likelihood of failing in the university exam rather than predicting student who are not in likelihood of failing.

In this scenario overall accuracy of the model is misleading and may not be useful for selecting a good classifier. The study uses various socio-economic and demographic characteristics of the students to predict the likelihood of failing in university examinations using Rep Tree, Jrip, Random Forest, Random Tree, Naive Bayes algorithms and also chooses the best classifier among the above using metrics like precision, recall, f score.

II. LITERATURE REVIEW

Since the arrival of learning analytics i.e. applying machine learning algorithms in education environment to improve student outcomes, there have been many prominent research works to predict student outcomes as follows.

(Delen, 2010) have compared individual classification algorithms with those of the ensemble learning algorithms for identifying students who are having more likelihood of attrition and found that balanced dataset provided better predictions than unbalanced data set .(Banumathi & Pethalakshmi, 2012) have used unsupervised machine learning algorithm namely UCAM (unique clustering with affinity measures) clustering algorithm and proved that it better than traditional clustering algorithms since the former avoids the problems of fixing initial cluster size and seeds . (Alam et al., 2018) proved that multi-layered neural network based approach performed better than other machine learning algorithms in terms of accuracy without variable reduction to classify the students in to high, medium and low categories. (Manrique, Nunes, Marino, Casanova, & Nurmikko-Fuller, 2019) have compared classification algorithms for predicting student dropouts in higher educational institutes using three different student representations and they have found that random forest and gradient boost ensemble algorithms performed better while Naive bayes performed least because of its strong interdependence assumption. (Puarungroj, Boonsirisumpun, Pongpatrakant, & Phromkhot, 2018) predicted student success in English exit exam using by using a c4.5 algorithm (j48) and found that English placement test result was a key predictor for student success in English exit exam success. (Supianto, Julisar Dwitama, & Hafis, 2018) compared the classification accuracy of various decision trees algorithms random tree, reptree, and c4.5 decision tree in predicting whether student will graduate in time or not.

There are other notable works which have compared the classifiers based on metrics other than overall classification accuracy such as (Márquez-Vera et al., 2016) proposed ICRM2 algorithm for predicting student dropout early and they proved that ICRM2 algorithm outperforms other classification algorithms in terms of True Negative rate and Geometric Mean of True Negative and True Positive rate. (Ratnaningsih & Sitanggang, 2016) have compared the three classification algorithms,

Revised Manuscript Received on March 02, 2020.

Vivek Raj S.N., Research Scholar, School of Management, SRM Institute of Science & Technology, Kattankulathur, Tamilnadu, India.

Dr. S. K. Manivannan, Associate Professor, School of Management, SRM Institute of Science & Technology, Kattankulathur, Tamilnadu, India.

Naive Bayes, Bagging, and C4.5in determining non active students in higher education setup at Indonesia Open University based on cross-validation, confusion matrix, ROC curve, recall, precision, and F measure.

This study follows the guidelines of the previous prominent previous research works fulfills the research gap by comparing the classification algorithms based on other metrics like precision, recall and f score as put forward by (Hossin & Sulaiman 2015).

III. RESEARCH METHODOLOGY

The study is based on data collected from second and third year studying bachelor of business administration degree in a private university in Chennai, India. Purposive sampling technique is used since the data is collected only from the second and third year students of the college as they are more exposed to the learning environment than the first year students. A total of 127 students responded the survey among 220 students. Data collected include student's demographic, economic, academic details.

We have followed data mining approach with cross validation for this research following the footsteps of prominent research conducted by (Delen, 2010). We have used to k- fold cross validation method for assessing the model performance. We have arbitrarily set the value of k =10 based on the recommendations of (Kohavi, 2001). (Berrar, 2019) In 10 fold cross validation dataset is split in to disjoint subsamples and the algorithm is trained and validated as follows.

The Steps in 10 fold cross validation are as follows.

1. Divide the Data set in to 10 disjoint subsample
2. Each subsample can be termed as $s_{1,}$ s_2, s_3... s_{10}.
3. For n=1 to 10

a. s_n is used as a validation dataset for remaining 9 subsamples

b. Remaining subsamples to be used for training the algorithm.

4. Compute the accuracy for all the ten runs and the mean value of results is used for prediction.

The study predicts the students who have the likelihood of failing in the university examination and thus needed intervention can be provided in teaching learning process. Dependent variable in the study is likelihood of student failing in the university examination. Dependent variable has two cases i.e. yes or no. The predictor variables used in the study are given in Table-I. Supervised learning algorithms have been deployed to predict student success. Algorithms are then compared based on metrics to select the best classifier.

Table –I: List of Independent Variables.

Sl.no	Independent variables	Type
1	Gender	Binary-nominal
2	Higher secondary school board	Multinomial
3	Higher secondary percentage	Ordinal
4	Mode of travel to college	Multinomial
5	Mothers educational qualification	Multinomial
6	Fathers educational qualification	Multinomial
7	Higher secondary course	Multinomial
8	English language competency [reading]	Ordinal
9	English language	Ordinal
10	English language competency [writing]	Ordinal
11	Accommodation	Multinomial
12	Distance from home to college	Ordinal
13	Mothers occupation	Multinomial
14	Family business	Binary-nominal
15	Whether first graduate in family	Binary-nominal
16	Year of study in college	Multinomial
17	Age	Number
18	Fathers occupation	Multinomial
19	Number of siblings	Number

IV. DATA ANALYSIS

The objective of the study is to develop a model to predict the student who is at the likelihood of failing in the university examination. We have used the demographic characteristics of students as independent or predictor variables to classify the students using various classification algorithms and choose the best algorithm of the given.

For the preliminary analysis the overall classifier accuracy of the various classification algorithms have been compared.

Table –II: Comparisons of Classification Algorithms Based on Overall Accuracy

Sl. no	Classifier	Correctly classified instances	Overall accuracy (%)	Kappa statistic	Relative absolute error (%)
1	Rep tree	92	72.44	0.381	73.47
2	Jrip	76	59.84	0.13	90.99
3	Random forest	84	66.14	0.24	88.26
4	Random tree	73	57.48	0.08	89.78
5	Naive Bayes	79	62.20	0.15	79.51

Table-II compares the various classification algorithms based on over all accuracy, kappa statistic and relative absolute error. Metrics in the table give us the lead that the reduced error pruning tree algorithm performs much better than the other algorithms since the algorithm correctly classifies 92 out of 127 instances followed by random forest algorithm and Naive Bayes, as random forest correctly classifies 84 out of 127 instances and Naive Bayes correctly classifies 79 out of 127 instances. Kappa statistic value also suggests that rep tree and random forest algorithms perform much better compared to other classification algorithms. But these measures provide us only the tip of the iceberg indicators because of the following constraints. Priority of this study is to correctly identify the students who have failed in the university examinations rather than identifying the students who have passed the university examinations.

Overall accuracy and kappa statistic compares the algorithms based on identifying both true positives and true negative classes. Our aim is correctly identifying the quadrant one in Table III i.e. true positive rate which gives the power of the algorithm in predicting the students who have arrears in university examination. In this scenario true indicator of power of algorithm is precision and recall and weighted harmonic mean of precision and recall which leads to f – measure.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = TP/(TP + FN)$$

Table – III: Confusion Matrix

		Actual Class	
		Failure	No Failures
Predicted Class	Failure	I. True Positives	II. False Positives
	No Failure	III. False Negatives	IV. True Negatives

Table –IV: Class Wise Comparison of Classifiers

Sl.No	Classifier	Precision	Recall	F Score
1	Rep Tree	0.68	0.5	0.578
2	Jrip	0.465	0.417	0.44
3	Random Forest	0.571	0.417	0.482
4	Random Tree	0.432	0.396	0.413
5	Naive Bayes	0.5	0.375	0.429

The Table-IV Compares the classification algorithms based on the precision, recall and f score values. Values in the tables give us the different interpretation of results compared to overall classification accuracy. Of all the classifiers only rep tree has a significant value of precision and f score. Random forest algorithm which had an overall accuracy of 66.14 % has an f score of 0.482 which is less than 0.5.Naive Bayes algorithm's over all accuracy is 62.20 % but it has a f score of only 0.429. If we had selected the classifier based on overall classifier accuracy then the results would be highly misleading. All the other algorithms other than rep tree failed significantly in identifying true positive classes since the f score values are less than 0.5.

V. FUTURE WORK AND RECOMMENDATIONS

The research carried out compared only few machine learning classification algorithms i.e. rep tree, Jrip, random forest, random tree, Naive Bayes algorithms. Further research can be carried out with more exhaustive list of algorithms to get more insights and judgements. Since our dataset is imbalanced ensemble machine learning algorithms can be used and compared with other traditional machine learning algorithms. All the algorithms in this research paper used classification cut-off value as 0.5, we propose to use Youden index in the future research to identify the optimal cut-off value.

VI. CONCLUSION

Student pass percentage and percentage of students earning degree have always been among the top key performance indicators for measuring the performance of higher educational institutions. In this article we have explored the application of machine learning algorithms on student data to classify the student having arrear and not having arrear in university examinations. We compared five machine learning algorithms Rep Tree, Jrip, Random Forest, Random Tree and Naive Bayes based on overall classifier accuracy, precision, recall and f-measure. Results revealed that it is meaningful and highly important to compare algorithms not only based on overall classifier accuracy but also with class specific metrics. Results also showed that rep tree algorithm outperformed other algorithms in all the metrics.

REFERENCES

1. Alam, M. M., Mohiuddin, K., Das, A. K., Islam, Md. K., Kaonain, Md. S., & Ali, Md. H. (2018). A Reduced feature based neural network approach to classify the category of students. *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence - ICIAI '18*, 28–32. https://doi.org/10.1145/3194206.3194218
2. Banumathi, A., & Pethalakshmi, A. (2012). A novel approach for upgrading Indian education by using data mining techniques. *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)*, 1–5. https://doi.org/10.1109/ICTEE.2012.6208603
3. Berrar, D. (2019). Cross-Validation. *In Encyclopedia of Bioinformatics and Computational Biology* (pp. 542–545). https://doi.org/10.1016/B978-0-12-809633-8.20349-X
4. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003
5. Hossin, M., Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201
6. Manrique, R., Nunes, B. P., Marino, O., Casanova, M. A., & Nurmikko-Fuller, T. (2019). An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, 401–410. https://doi.org/10.1145/3303772.3303800
7. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135
8. Puarungroj, W., Boonsirisumpun, N., Pongpatrakant, P., & Phromkhot, S. (2018). Application of Data Mining Techniques for Predicting Student Success in English Exit Exam. *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication - IMCOM '18*, 1–6. https://doi.org/10.1145/3164541.3164638
9. R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, *The Proceedings of the 14th International Conference on AI (IJCAI), Morgan Kaufmann, San Mateo, CA, 1995*, pp. 1137–1145.
10. Ratnaningsih, D. J., & Sitanggang, I. S. (2016). Comparative analysis of classification methods in determining non-active student characteristics in Indonesia Open University. *Journal of Applied Statistics*, *43*(1), 87–97. https://doi.org/10.1080/02664763.2015.1077940
11. Supianto, A. A., Julisar Dwitama, A., & Hafis, M. (2018). Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University. *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, 308–311. https://doi.org/10.1109/SIET.2018.8693158

AUTHORS PROFILE

Vivek Raj S.N., Research Scholar, School of Management, SRM Institute of Science & Technology, Kattankulathur, Tamilnadu.

Dr. S. K. Manivannan, Associate Professor, School of Management, SRM Institute of Science & Technology, Kattankulathur, Tamilnadu.

Retrieval Number: E2643039520/2020©BEIESP
DOI: 10.35940/ijitee.E2643.039520

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication