iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34222

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Student's Performance Prediction

Tummala Sri Ranga Sai Krishna[1], Jaya Sai Hrithik Padavala[2], Vinay Reddy Poreddy[3], K Bhagya Laxmi[4]
[1, 2, 3, 4]*Computer Science and Engineering, Matrusri Engineering College*

Abstract: ***The advent of technology the world currently experiencing is an outcome of the adapting pedagogy being implemented by the educational institutions around the world. One trait of such efficient pedagogy is to ensure the underprivileged students of the class receive extra support to make up to their fellow classmates. Identification of such students plays a significant role. In this paper, we discuss two machine learning techniques that build a classifier which then predicts the performance of students from a dataset provided by the machine learning repository of University of California Irvine. The machine learning techniques include Decision Tree and Logistic Regression. ROC index performance measure and the classification accuracy are used as holistic measures for comparison of the discussed machine learning techniques. In addition, we were able to identify five significant factors that influence the performance of students from the given dataset.***
Keywords: Prediction of student's performance, Machine Learning, Decision Tree, Logistic Regression, Supervised Learning

I. INTRODUCTION

Making better schooling inexpensive has a considerable impact on making sure the nations' financial prosperity and a vital cognizance of the authorities while making educational policies. Yet, the financial burden the higher education is imposing on the students and their families is observing an exponential increase since the past few decades, importantly in the developed countries such as the United States of America. The student loan debt in NewYork alone crossed a trillion mark which is more than the combined credit and auto loan debts of the entire America[1]. Few countries such as Iraq, India provide partial or full funding to the higher education students.Yet, the delay in the graduation of students is resulting in an extra burden for the governments and educational institutions that are providing funding. Machine learning techniques can be employed to predict the student's performance and an appropriate action can be taken based on the prediction.Choosing the attributes to be used as input for the machine learning technique play an important role. The attributes can be classified into grades, educational background, psychological evaluation and demographics[2]. Two machine learning techniques Decision Tree and Logistic Regression are used to build the machine learning model. We used ROC index and classification accuracy for comparing the discussed techniques. The dataset used for the implementation of techniques is retrieved for the machine learning repository of University of California Irvine. The dataset has information of 396 students.

II. LITERATURE REVIEW

The use of machine learning for the prediction of potential students at risk is not as novel as it seems to be. There has been notable research conducted in this field. To predict the efficacy of a student his/her grade is an inevitable measure. But it cannot be taken as a holistic characteristic to determine a student's efficiency. There are several other characteristics such as age, family background etc which can be classified into demographics that play a significant role[2]. This research also took in an additional characteristic of usage of the internet. Along with the Decision Tree and Logistic Regression there are several other techniques such as Naives Baye, Artificial Neural Network, Decision List,etc. Table 1 shows a summary of research papers that relate to the study[3].

Table I
Summary Of Research Papers That Relate To Study

Paper	Feature	Dataset size	Machine Learning Algorithm
Meier et al, 2015 [4]	Grades	700	KNN, Logistic Regression, SVM
Guleria et al, 2014[5]	Class Performance, Attendance, Assignment, Lab Work	120	Decision Tree

Xu et al, 2017[1]	Grades, Backgrounds	1169	Linear Regression, Logistic Regression, RF, KNN
Altabrawee et al, 2019 [3]	Grades, Background, Demographics	161	Naive Bayes, ANN, Logistic Regression, Decision Tree
Huang et al, 2011[6]	GPA and Grades	239	Linear Regression, ANN, SVM

III. METHODOLOGY

In this section, machine learning techniques that are to be compared in this paper are introduced.

A. Decision Tree

A decision tree model represents a tree structure congruent to that of a flowchart. In this, every internal node represents a test on a dataset attribute while the test outcomes are represented by each tree branch. In addition, each leaf node represents a target feature label and the upper first node in the tree represents the root node. Decision trees can be a binary or a non-binary trees. Decision trees are popular classification techniques because using them does not need prior knowledge of the problem domain or a complicated setting of the classification parameters. In addition, they can be converted to classification rules easily and they can be understood easily.

Decision tree classification technique has been used in many real word applications such as financial analysis, medicine, molecular biology, manufacturing production, and astronomy. During building the decision tree, the algorithm uses an attribute or feature selection measure which is used in selecting the attribute or the feature that best divides the dataset instances into distinct target classes. Such measures include the Information Gain, Gain Ratio, and Gini Index. Popular decision trees algorithms include ID3, CART, and C4.5[7][3].

B. Logistic Regression

Logistic Regression represents a mathematical modeling technique which describes the relationship between several independent variables, X1...XK, and a dependent variable, D. The logistic model uses the logistic function as a mathematical form which has the range between 0 and 1 for any given input. The logistic model can describe a probability of an event which is always a value between 0 and 1. The following formula represents the logistic model.

$$P(D = 1 | X1, X2 \ldots XK) = 1/1 + e^{-a + \sum_1^k \beta i + xi} \quad (1)$$

Where and are the model's parameters that can be learned from a set of labeled instances in the training dataset. Gradient Descent Algorithm can be used to find the best values of the model's parameters during the training phase [8]

IV. THE EXPERIMENT

A. Dataset and Data Sources

The dataset used in this research is collected from the machine learning repository handled by the University of California Irvine. Two data sources have been used, surveys collected from the students and the students' grades data records. The dataset contains 382 student records.

The dataset contains twenty attributes. The attributes can be divided into five categories which are personal and lifestyle, studying style, family related, educational environment satisfaction, and student's grades. Table2 shows the attributes used in order to construct the dataset. Each student has been labeled as Weak or Good based on his/her final grade. The weak student is the student who has a final grade less than sixty out of hundred. On the other hand, the Good student is the student who has a final grade equal or greater than sixty. Identifying the weak status students is more important than identifying the good status students, therefore the weak status is considered a positive value of the target attribute.

Table II
The Dataset Attributes

Attribute	Attribute Definition
School	Student's School
Sex	binary: "F" - female or "M"-Male
Address	binary: "U"- urban or "R"-rural
famsize	binary: "LE3"- less than 3 or "GT3"- greater than 3
pstatus	binary: "T"- living with parents or "A"- living alone
Medu	mothers education numeric:-0-4
Fedu	fathers education numeric: 0-4
Mjob	mothers job
Fjob	fathers job
reason	reason to choose this school nominal
guardian	student's guardian nominal: "mother", " father", "other"
traveltime	numeric
studytime	numeric
failures	numeric
schoolsup	extra support binary: "yes" or "no"
famsup	educational support from family binary: "yes" or "no"
paid	paid tutoring binary: "yes" or "no"
activities	extra curricular activities binary" "yes" or "no"
internet	binary: "yes" or "no"

B. Accuracy and Performance measures

In this experiment, a three folds cross validation method has been used. In this method, the dataset is divided into three equal size sets. The learning and testing are executed three times. At each fold or execution, the machine learning algorithm selects one set to be the test set and the remaining two sets as the training sets. The accuracy and the performance measures are aggregated over all the folds in order to calculate the final performance and the final accuracy of the model. The ROC index and the performance measure, has been used to evaluate the performance of the classification models. This measure is a well-known measure that is relying on the ROC curve and it is calculated by using the prediction scores. Equation 2 is used to calculate the ROC index [9]. In addition to the ROC index, many important measures have been used such as the accuracy, the classification error, and the F Measure. Equation 4 is used to calculate the F Measure. The F Measure is a useful alternative to the misclassification rate measure. [9] [3]

$$ROC\ index\ =\ \sum_{i=2}^{|T|}\ (\ FPR(T[i])\ -FPR(T[i-1])\ )\ *\ (\ FPR(T[i])\ -FPR(T[i-1]))\ /\ 2 \quad (2)$$

Where |T| represents the number of thresholds that are used, FPR(T[i]) represents the false positive rate at the threshold i, and TPR(T[i]) represents the true positive rate at the threshold i. A larger ROC index indicates a better classification model. A model with ROC index above 0.7 considered a strong model while a model with ROC index below 0.6 considered a weak model.[9]

$$F\ measure\ =\ 2\ *\ \frac{(Precision\ *\ Recall)}{(Precision\ +\ Recall)} \quad (3)$$

$$Precision\ =\ \frac{TP}{TP\ +FP} \quad (4)$$

$$Recall\ =\ \frac{TP}{TP\ +FN} \quad (5)$$

TP, True Positives, is the number of data rows in the test set which had a positive target and that were predicted to have a positive target. TN, True Negatives, is the number of data rows in the test set that had a negative target and that were predicted to have a negative target. FP, False Positives, is the number of data rows in the test set which had a negative target but that were predicted to have a positive target. FN, False Negative, is the number of data rows in the test set that had a positive target but that were predicted to have a negative target [9][3].

C. Implementation

All the models have been implemented by the Rpubs by RStudio software. A Cross Validation operator has been used in order to execute the three folds validation operations during the training and the testing phases. The operator is used for sampling the property set to linear sampling. In order to find the best set of the models' parameters, the Optimize Parameters (Grid) operator has been used. The Optimize Parameters operator has been set to find the best value of the learning rate and the L2 regularization. For the learning rate and the L2 regularization, the configuration is set to use 100 steps on a linear scale from 0 to 1[3]. For building the DT model, the Optimize Parameters operator has been set to find the best value of the splitting criterion, and the minimal size for split properties. Also, apply pruning property has been set by the optimization operator. All the other parameters have been set to the default values. The Logistic Regression operator has been set to use regularization and the optimization operator set to find the best value for the solver method and the lambda. The lambda search property is set to use sixty steps on a linear scale starting from 0 to 1.787. All the other parameters have been set to the default values.

D. The Results

Two classification models have been created and tested using Two machine learning techniques, Logistic Regression, and Decision Tree. Table3 shows the accuracy and the performance measures for each model as well as the confusion matrices.

Table III
The Accuracy And Performance Measures For The Models

Model	TP	FP	TN	FN	Precision	Recall	Accuracy	ROC index
Decision Tree	70	10	26	12	87.50	85.36	81.82	0.765
Logistic Regression	73	7	30	8	91.25	90.12	87.29	0.786

V. CONCLUSIONS

To solve the problem of identifying the students who have a poor academic performance, two classification models have been built to predict the performance of the students. Two machine learning techniques, Decision Tree, and Logistic Regression, have been used. The models have been compared to one another using the ROC index performance measure and the classification accuracy. Logistic Regression model has the highest ROC index that equals to 0.786 and accuracy of 87.29. In addition, the decision tree model showed that not all the attributes are involved in the classification process. We find that the variables which actually impact the prediction of final grades are Absences, Fathers' job and Grades in Exam 1 and Exam 2 as found in decision trees algorithm.

REFERENCES

[1] J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 201

[2] K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.

[3] Altabrawee, Hussein & Ali, Osama & Qaisar, Samir. (2019). Predicting Students' Performance Using Machine Learning Techniques. JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences. 27. 194-205. 10.29196/jubpas.v27i1.2108..

[4] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, "Predicting grades," IEEE Trans. Signal Process., vol. 64, no. 4, pp. 959–972, 2016. .

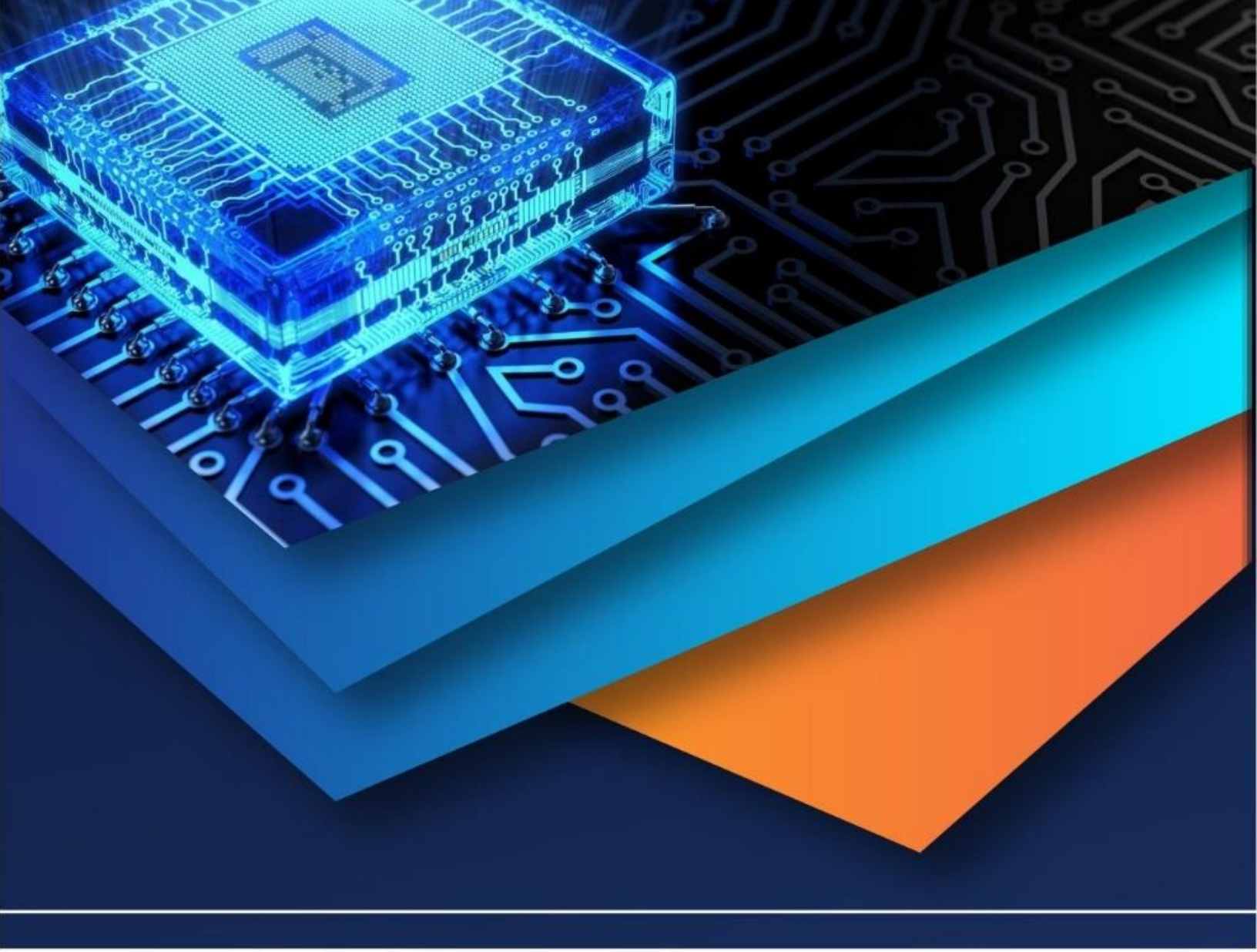
[5] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126–129, 2015.

[6] S. Huang and N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques," Proc. - Front. Educ. Conf. FIE, vol. 1, pp. 3–4, 2012.

[7] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann publications, 2012.

[8] D. G. Kleinbaum and M. Klein, Logistic Regression A Self-Learning Text, 3rd ed. New York: Springer-Verlag New York, 2010.

[9] *J. D. Kelleher, B. Mac Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies. 2015.*



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY