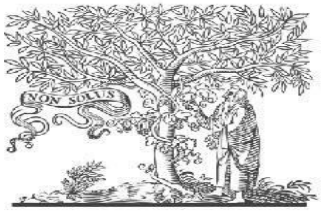
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 31st Mar 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 03

10.48047/IJIEMR/V12/ISSUE 03/113

Title **Student academic performance using machine learning algorithms**

Volume 12, ISSUE 03, Pages: 813-818

Paper Authors

DHANA LAKSHMI GALLA,PHANIDHRA RAJU KAGITALA ,BYULA MALLELA,

JAYADEEP THOTA

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Student academic performance using machine learning algorithms

[1]DHANA LAKSHMI GALLA,[2]PHANIDHRA RAJU KAGITALA ,[3] BYULA MALLELA,

[4]JAYADEEP THOTA

[1,2,3,4]Computer Science and Engineering, Kallam Haranadhareddy institute of technology

Guntur, India

208x5a0501@khitguntur.ac.in 198x1a05h7@khitguntur.ac.in
208x5a0503@khitguntur.ac.in

198x1a05e9@khitguntur.ac.in

Abstract

Developing tools to support students and learning in a traditional or online setting is a significant task in today's educational environment. The initial steps towards enabling such technologies using machine learning techniques focused on predicting the student's performance in terms of the achieved grades. The disadvantage of these approaches is that they do not perform as well in predicting poor-performing students. The objective of our work is two-fold. First, in order to overcome this limitation, we explore if poorly performing students can be more accurately predicted by formulating the problem as binary classification. Second, in order to gain insights as to which are the factors that can lead to poor performance, we engineered several human interpretable features that quantify these factors. These features were derived from the students' grades from the University of Minnesota, an undergraduate public institution. Based on these features, we perform a study to identify different student groups of interest, while at the same time, identify their importance.

Keywords—Svm algorithm, Decision tree algorithm, Random Forest algorithm, Gradient Boosting algorithm.

I. INTRODUCTION

Higher educational institutions constantly try to improve the retention and success of their enrolled students. According to the US National Center for Education Statistics [8], 60% of undergraduate students on four-year degrees will not graduate at the same institution where they started within the rest six years. At the same time, 30% of college fresh- men drop out after their rest year of college. As a result, colleges look for ways to serve students more efficiently and actively. This is where data mining is introduced to pro-vide some solutions to these problems. Educational data mining and learning analytics have been developed to pro- vide

tools for supporting the learning process, like monitor and measure student progress, but also, predict success or guide intervention strategies

Most of the existing approaches focus on identifying students at risk who could be net from further assistance in order to successfully complete a course or activity. A fundamental task in this process is to predict the Stu dent's performance in terms of grades. While reasonable prediction accuracy has been achieved [14, 10], there is a significant weakness of the models proposed to identify the poor-performing students [18]. Usually, these models tend to be over-optimistic for the performance of students, as most of the students do well, or have satisfactory enough performance.

We essentially identify two complementary groups of students, the ones that are likely to successfully complete a course or activity, and the ones that seem to struggle. After identifying the latter group, we can provide additional resources and support to enhance their likelihood of success. However, "success" and "failure" can be relative or not. For example, a B− grade might be considered a bad grade for an excellent student, while being a good grade for a very weak student. We investigated different ways to define groups of students taking a course: failing students, students dropping the class, students performing worse than expected and students performing worse than expected, while taking into consideration the difficulty of a course. In order to gain more insight into the learning process and its most important characteristics, we have created features that capture possible factors that influence the grades at the end of the semester. Using these features, we present a comprehensive study to answer the following questions: which features are good indicators of a student's performance? which features are the most important? The findings are interesting, as different features are the most important for different classification tasks.

II. Literature Survey

As we are interested in estimating next-term student performance, we will review the related work in this area of research. The binary classification has been used in various educational problems, like predicting if a student will drop out from high school [6] or to predict if a student will pass a module in a distance learning setting [7]. Multi-label classification has been applied to provide a qualitative measure of students' performance. In [17], decision tree and naive Bayes classifiers are used with data from a survey. Attributes collected by a learning management system have been employed to estimate the outcome as Fail, Pass, Good and Excellent [16], or to classify students [12]. Some approaches [11, 9] test different ways to label the student performance, with two (pass or fail) or more labels. Most of the approaches are small-scale studies, that are applied to a limited number of courses. In recent years, influenced by advances in the recommender systems, big data approaches have been also utilized in the area of learning analytics. Initially, the term "next-term grade prediction" was introduced by Sweeney et al. [18] in the context of higher education, and it refers to

the problem of predicting the grades for each student in the courses that he/she will take during the next semester. Models based on SVD and factorization machines (FM) were tested. In another approach [15], the previous performance of students controls the grade estimation in two different ways while building latent models. In [19], some additional state-of the-art methods were used, as well as, a hybrid of FM and random forests (RF). The data used are the historical grades and additional content features, representing student, course and instructor characteristics. At the same setting, [14] and [10] developed course-specific methods to perform next-term grade prediction based on linear regression and matrix factorization. All these methods assign a specific numerical grade to each student's attempt to take a course. A limitation identified in these approaches was that the developed models perform poorly for failing students. In [5], failing students have been completely removed from the dataset. As this is the subpopulation of students that needs additional support the most, it is very important for a model to be able to accurately identify these students at risk. This work is a more general study of the factors that influence the student performance, in a very large scale. The only observed data that we have available are the students' grades at the end of the semester. In our approach, we formulated this problem as a binary classification task, in order to detect the different group of students. In other words, we keep the classification methodology, but apply it on the context of big data.

support to extract the knowledge which is hidden in the data. The aim of our work is to support the education system. As education plays a vital role in enhancing the overall of the human being. It is the most important part of life and plays a vital role in improving one being. But the level of education is decreasing because of certain reasons and those should be eliminated. It motivated us to work on student dataset. The data collection, categorization, and classification are being performed manually. The main aim of this work is to improvise the student performance in studies based on some important factors. The main disadvantage of this process is the delay in results; remedial measures are not taken properly due to a late analysis of student performance. There will be a delay in the results announcements which leads to the poor performance of the students in the next examination due to lack of planning in their preparation. When the counting of students increases, the analysis of the performance of a student becomes difficult. To overcome this difficulty we now introduce an educational datamining tool. When institutes store their student's details in the cloud, it will be difficult to analyse large data often called big data. By applying data mining on the data stored, we can easily categorize and analyse the results of a student in a short time without anydifficulties.2 Literature Review In this paper [1], an approach for classification of eye disease using the Random Forest algorithm is proposed

III. Problem Statement

All these methods assign a specific numerical grade to each student's attempt

to take a course. A limitation identified in these approaches was that the developed models perform poorly for failing students. In, failing students have been completely removed from the dataset. As this is the subpopulation of students that needs additional support the most, it is very important for a model to be able to accurately identify these students at risk**.**

IV. Existing System

In accordance with the impacts of learning-teaching on the sustainability of intermediate & secondary education and tech-boosted learning [5], We must carefully describe the basic information technology requirements that will serve us instead of being a hitch in learning and teaching. For instance, the preparation of teaching and managerial personnel's for the production of predictive analytic skills as it is crucial for measuring the latent outcomes of the computer-aided framework usage [7]. In to discussed technologies earlier, that is executed with a greater impression in the academic set-ups to produce a large volume of data and save it in means that it could be efficiently presented ubiquitously [8]. The data size can exceed the processing volume sometimes, storing and evaluating it with conventional methods. New technologies should be considered in order to perform data analysis such as data mining, intelligent systems, association rules mining, optimization based data mining [9] and big data. The bunch of these novel technologies will enable simple and effective analysis of educational-data, and can be utilized to transform the educational-data in a new shape which could be more beneficial [10,11,12].

With deep learning mining of educational data is a growing field for research that enable us to analyze and process the educational information collected from different roots [13]. For analyzing educational data, several statistical methods, data-mining, visualization, and ML gears are utilized. The study analytics generated from academic-data intends to investigate obtained data from the institutional databases. Learning-management frameworks interprets the information, improve learning procedures and atmosphere in which the data befalls .

V. Proposed System

Having as input the historical grading data, we derived different features to capture possible factors for a student's poor performance. The features can be separated into three distinct categories: the student-specific (independent from course c), course-specific features (independent from student s) and student- and course-specific features (they are a function of both s and c). All extracted features are described in Tables 1, 2, where related features are grouped together into eight different subcategories. The keywords on bold are used to indicate the corresponding group of features later. Note that for each {s, t, c}, where student s took course c in semester t, we generate a different set of features. Every set of features characterize a student's attempt to take course c at the specific point of his/her studies. These features are either numerical, categorical or indicator variables. For indicator features, we use the values of 0 or 1. The categorical features are encoded via a numerical value. For example, the feature about the current semester is categorical,

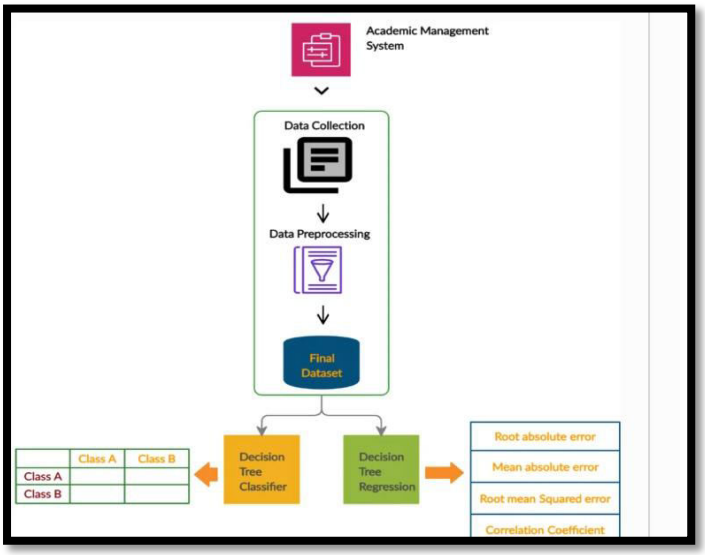
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

and the values {fall, spring, summer} are transformed to {0,1,2}, respectively.

Our motivation was to identify groups of students that need further assistance and guidance in order to successfully complete a course. These students could benefit from informed interventions. We consider this to be a binary classification problem, where these students form one of the classes and the remaining students form the other class.

VI. Block Diagram:

The proposed approach is based on the identification of important features which are responsible for any task. The features are important for the objective, because they provide the support for the right decision like we have find out that the absence of the student is one of the important features for taking care of the grade of the students. In the results, every classification method proves this feature is more important. It shows better results in terms of precision, recall and f1-score. We



have analyzed the different variables to identify the importance of features. All the classification methods are providing the accuracy not better than the Random Forest. Because, first identify the attributes which does not have any effect on the marks of students then check the accuracy of the model. There are two types of the attributes - categorical and numerical and we have find the correlation of these attributes with the target attribute and if there is no correlation between them then we eliminate that attribute. Hence the efficiency of model that is based on Random Forest increases. In the future we can estimates the classifier with over fitting and under fitting.

CONCLUSION

The purpose of this paper is too accurate identify students that are at risk. These students might fail the class, dropout, or perform worst than they usually do. We extracted features from historical grading data, in order to test different simple and sophisticated classification methods based on big data approaches. The best performing methods are the Gradient Boosting and Random Forest classifiers, based on AUC and F1 score metrics.

We also got interesting findings that can explain the student performance.

Future Enhancement

An approach is proposed in this article for observing and forecasting the students marks and grades in an automated way. This research study aims to gain better accuracy for the classification and low root means square error. This study also led us to make groups of students who have same

education historic record, for instance, students have taken the same subjects in the same academic session. This job is not simple and easy, fact that intermediate & secondary grade students do not have the same conduct while studying in the same group. Thus, to attain reliable forecasting outcomes it is essential to choose students of the same academic section and group. The student marks and grade were analyzed in this study by knowledge areas. It can be justified that a grade from one subject can be utilized to predict from the grade of a student who took the exam in the previous academic session.

References

[1] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshan. Classification and regression trees. CRC press, 1984.

[3] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):273–297, 1995.

[4] J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

[5] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran. Machine learning based student grade prediction: A case study. arXiv preprint arXiv:1708.08744, 2017.

[6] J. E. Knowles. Of needles and haystacks: Building an accurate state-wide dropout early warning system in Wisconsin. Journal of Educational Data Mining, 7(3):18–67, 2015.

0using machine learning techniques. Applied Artificial Intelligence, 18(5):411–426, 2004

. [8] J. McFarland, B. Hussar, C. de Brey, T. Snyder, X. Wang, S. Wilkinson-Flicker, S. Gebrekristos, J. Zhang, A. Rathbun, A. Barmer, et al. Undergraduate retention and graduation rates. In The Condition of Education 2017. NCES 2017-144. ERIC, 2017.

[9] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In Frontiers in education, 2003. FIE 2003 33rd annual, volume 1, pages T2A–13. IEEE, 2003.

[10] S. Morsy and G. Karypis. Cumulative knowledge-based regression models for next-term grade prediction. In Proceedings of the 2017 SIAM International Conference on Data Mining, pages 552–560. SIAM, 2017.\

[11] E. Osmanbegovi´c and M. Sulji´c. Data mining approach for predicting student performance. Economic Review, 10(1):3–12, 2012.