Advances in Parallel Computing Technologies and Applications
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC210137

Student Performance Prediction Using Machine Learning

Priya S [a,1], Ankit T [b] and Divyansh D [b]
[a] *Assistant professor, Department of Computeer Science SRM Institute of Science and Technology*
[b] *Student, Department of Computeer Science SRM Institute of Science and Technology*

Abstract. Performance analysis of learning outcomes is a system that will aim for excellence at all levels and dimensions of the student's field of interest. This paper suggests a comprehensive EDM framework in the form of a rule-based recommender system that not only analyses and predicts student achievement, but also demonstrates the reasons for it. The suggested framework examines students' demographic information, study-related characteristics, and psychological factors to gather as much information as possible from classmates, teachers, and parents. School reports and queries used to collect the world's latest data (e.g., student marks, population, social and school-related factors. Using a set of potent data mining methods, aiming for the greatest possible precision in academic performance prediction. The framework is effective in identifying the student's weaknesses and making relevant recommendations. In contrast to current frameworks, the proposed framework outperforms them in a practical case study involving 200 individuals.

Keywords. Data Mining, Student Performance Prediction, Classification

1. Introduction

Recently, online educational systems have grown in popularity, and student information stored has grown to be large in number. This allows for the development of rules and forecasts. Data mining tools used to process student teaching data. A variety of information about the student's social status, learning environment, or study notes can be used to make predictions about his or her achievements or failures. The purpose of this study is to predict the final level of students to assist teachers in taking steps to protect vulnerable children. To improve the prediction model's accuracy, a number of data preprocessing procedures were used. The level of accuracy of the three popular data mining algorithms (decision tree, random forest, and inexperienced Bayes) is then compared. In addition, the effects of the two separate phase ranges on data mining are investigated in this study: five-level grade categorization and binary grade categorization. Research in this field. The methods used in this study are briefly described to provide thorough understanding of the concepts. Experimental experiments with dataset descriptions, data pre-processing, and experimental result subtitles are summarized in Section 4. Section 5 concludes with a conclusion and recommendations for future studies.

[1] Priya S, Department of Computeer Science, SRM Institute of Science and Technology, India.
E-mail: spriyasrmist@gmail.com.

2. Related Works

Predicting students' educational performance is one among the most topics of educational data mining. With the upcoming of new technology, technology investment in the education sector has increased, in combination with technological advances, e-Learning platforms such as the web on-line learning and multimedia technologies have evolved, and each learning prices have reduced, and time and area limitations are eliminated, the rise of on-line courses and therefore the hike of online transactions and school-based transactions led to an increase in digital knowledge during the field [1, 2]. Costa stressed the info concerning the failing rate of the students; the teachers were involved and raised vital concerns about the failure prediction. Estimating student's performances becomes harder as a result of the massive volume of information in coaching databases. Descriptive statistical analysis is effectively accustomed offer the fundamental descriptive data of a given information. However, this isn't always sufficient. to inform faculty and students ahead of time, students may also be ready to determine ahead of time, using calculable modelling strategies, there is an advantage to defining university students in terms of how they can work in education in order to increase achievement levels and manage resources effectively [3–6]. Universities' huge expansion of electronic knowledge ends up in a rise within the ought to acquire significant data from these large amounts of information. By utilizing data processing techniques on educational data, it's possible to enhance the standard of the education [7–9]. To date, data processing algorithms have been used in various fields of education such as engineering education, physical education, and English language education. Some fields are oriented toward high school students, many of whom are highly interested in higher education. While some data analysis studies focus on predicting overall student outcomes, some studies focus on teacher outcomes as a whole. The authors of [10–12] made a project to learn the patterns of university student retention . Since reading the literature, it is clear that the present state of the art has a lot of space for change. Improvements are possible, as seen in the article, if we examine different learning styles; choose features carefully; examine the application of the studied hypothesis and not its moderate application, but also by the variation in that performance; and investigate the delta of student variables between those who remain and those who are retained. The following characteristics were considered to be informative using these methods for determining whether students would stay for the first three years of an undergraduate degree: family history and family's social-economic standing, high school GPA, and test scores. The authors of [13–15] researched the classification techniques used in data mining. There are three main components of data mining. Clustering, association laws and sequence analysis are all techniques for evaluating results. Process classification / collection process is the process of analyzing data and creating a set of collection rules that can be used to label a potential data label. The process of extracting data from a data set and converting it into an understandable structure is known as data mining. It is a mathematical method that incorporates methods from artificial intelligence, machine learning, analytics, and database systems to detect integration in big data sets [16–18]. The real data mining activity entails the automated or semi-automated processing of vast amounts of data in order to uncover previously identified trends [19, 20]. There are six different types of data mining activities. Anomaly identification, association rule learning, clustering, classification, regression, and summarization are some of the techniques used to diagnose anomalies. Classification is a popular data mining technique that is used

in a number of fields. Classification is a method of digging data to predict the membership of a data model group. The basic classification techniques are discussed in this article. The aim here is to provide a systematic overview of various classification strategies in data mining, including decision tree inference, Bayesian networks, and the k-nearest neighbor classifier.

Salam et al. used ELM technique for the diagnosis of heart diseases. The diagnosis of heart disease, which affects millions of people, is one of the most important aspects of the use of machine learning technology. Patients with heart disease have a number of independent variables in general, such as age, sex, serum cholesterol, blood sugar, and so on, that can be used to diagnose them effectively. To model these variables, an Extreme Learning Machine (ELM) algorithm is used in this article. The proposed system would be used to supplement an expensive routine checkup with a warning system that alerts patients to the possibility of heart failure. The method is based on actual data obtained by the Cleveland Clinic Foundation, which included information on about 300 patients. According to simulation data, this architecture is around 80% accurate in detecting heart disease.

Shadab Adam Pattekari and Asma Parveen developed a predictive system for heart disease using the Naïve Bayes algorithm. The key goal of this study is to create an Intelligent System using the Naive Bayes data mining simulation methodology. It's a web-based programmed in which the user asks pre-determined questions. It pulls secret data from a database and compares user values to a trained data set. It will address with complex questions about heart disease treatement, allowing healthcare workers to make correct treatment choices than old decision systems. It also aims to lower healthcare costs by delivering successful care.

Oliver and Mangas made a health gear, a wearable machine to monitor and analyze physiological signals. Health Gear is a real-time wearable device that monitors, visualizes, and analyses physiological signals in real time. Health Gear is made up of a series of noninvasive physiological sensors that are attached to a mobile phone through Bluetooth and store, transfer, and interpret physiological data before displaying it to the user in an understandable manner. The emphasis of this paper is on a Health Gear implementation that uses a blood oximeter to track the user's blood oxygen level and pulse as they sleep. We further discuss two separate algorithms for predicting sleep apnea episodes automatically, as well as the overall system's success in a sleep trial of 20 volunteers.

3. Existing System

Previous predictive models relied solely on demographic data from students, such as gender, age, marital status, family income, and qualifications. Additionally, study-related characteristics such as homework and study hours, as well as past successes and grades, are included. This previous studies were restricted to predicting academic performance or loss without elaborating on the causes for this prediction. The majority of recent studies aimed to collect more than 40 attributes in their data collection in order to estimate a student's academic success. These characteristics came from the same type of data group, whether demographic, study-related, or both, resulting in a lack of variety in predicting laws. As a result, the information for the reasons for the student's dropout was not completely extracted by these created laws. Aside from the work described above, previous

predictive research models from the viewpoint of educational psychology performed a couple of experiments to analyze the relationship between mental health and academic success. The guidelines were too brief, and they failed to demonstrate how to implement them.

4. Proposed System

The suggested paradigm starts by integrating demographic and study-related attributes with educational psychology areas, by applying psychological features to the historically used data collection (i.e., students' demographic and study-related data). We selected the most important attributes based on their justification and association with academic success after surveying the previously used variables for predicting the student's academic performance. The proposal's goal is to look at a student's longitudinal statistics, study-related information, and psychological attributes in terms of their final state and see whether they are on target, struggling, or even failing. In addition, we conducted a thorough analysis of our proposed model with previous similar models.

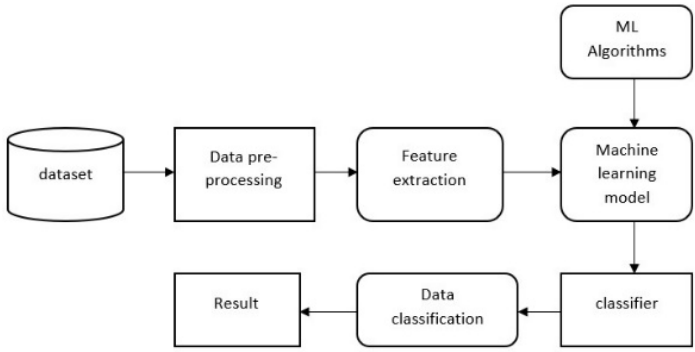

Figure 1. System Architecture

5. Method

As a result of the surge in digitalization, we now have an abundance of data in every field. When the right information and tools to use it are known, having more data becomes valuable. Using a variety of machine learning methods, data mining aims to extract data from data. It is possible to build links to data and make accurate predictions for the future using the data mine. Education is one of the applications for data mining. In education it is a field that encourages us to predict future predictions by analyzing historical data in the field of education and applying electronic learning methods to it. Data mining is divided into three categories: classification, collection, and organizational mine control. The classification task is the focus of this research. Depending on the analysis area and

the sort of data we have, different data mining techniques may be used. On educational datasets, four well-known Data Mining Methods, Applications, and Systems classification algorithms were used to predict students' final grades.

5.1. Artificial Neural Network

Artificial Neural Network is a collection of input and output units connected by a heavy connection. ANN learns by adjusting link weights in such a way that it can predict the appropriate label of such input data sets. The retrieval algorithm is one of the most well-known learning methods for ANN training. ANN has many advantages, including high tolerance to noisy data sets and good results in the differentiation of untrained patterns, so it is helpful in situations where the relationship between class markers and negative data properties is understandable. Image and handwriting recognition, voice recognition, diagnostic medicine, and disease are just a real global launch of the ANNs. ANNs can be categorized according to their structure and style. ANN for fully integrated feed supply, with input layer, one or more hidden layers, and output output, is one example. In addition, connectors facing the input unit or output unit in the previous layer do not go back. In addition, each L unit layer provides details on each L + 1 unit layer. In this study, the three-layer ANN is fully integrated. The input layer, the two hidden layers, and the output layer form a network. Twenty input cells, or neurons, form an input layer, while six inverted connections form the first hidden layer. There are three other secret secrets in the second hidden layer. The output layer, which has a single output unit, is the fourth layer. The function of secret operating units has been implemented using the Rectifier Linear Unit.

5.2. Logistic Regression

The arithmetic simulation model defines the relationship between multiple independent variables, X1 ... XK, and the dependent variables, D. The preparation model uses a mathematical form known as the logistic equation, with a scale of 0 to 1 for all inputs provided. The order model can be used to define the probability of an event, which is usually a number between 0 and 1. The order model is represented by the formula below.

$$P(D = 1|X1, X2, \ldots, Xk) = \frac{1}{1 + e(a + \sum_k b_i x_i)} \tag{1}$$

Model parameters, a and b, can be read in a series of scenarios named in the training database. During the training process, the Gradient Descent Algorithm used for finding best values for model.

5.3. Support Vector Machine

It's a promising approach for both linear and non-linear data classification [21]. It transforms the initial training data into higher dimension using non-linear projection. We've described each since then. Each student has several variables, and each of them is referred to as a multidimensional entity. It looks for the linear optimal separating within this new dimension. A hyperplane is a "decision boundary" that separates students from one class from those of another. A hyperplane will often be used to distinguish data from

two groups (H1 and H2). Support vectors (”critical” planning information samples) and edges (Large edge and Small edge, which are characterised by aid vectors) are used by the SVM to find this hyperplane. Many data analysts claim that this approach is slow during the training process, but it has a high precision, particularly for small numbers of support vectors that are independent of the object's higher dimensions. As a result, we can conclude that SVM is an excellent tool for classifying a small number of training samples with various parameters.

6. Experimental Result

The study collected a single set of data containing details and details of students from two secondary schools in Portugal. Database features include student marks, social and demographic data and school-related features. This is all based on records and inquiries. The database contains student achievement data from the Mathematics section, and the second contains student data for the Portuguese language course. The database has 33 attributes in total.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Figure 2. Attribute Dataset

As in several European countries, the range of results in data ranges from 0-20. The data had to be converted into groups since the final grade of students is in the form of whole numbers, and the result should be in the form of category numbers. We used and compared two different classification schemes in the study: the inclusion of five levels and the binary grade. We divided the last measure into five sections first. The 0-9 scale applies to grade F, which is the lowest and points to ”failure.” Remaining category marks (D (adequate), C (satisfactory), B (good), and A (good / excellent) corresponding to D (sufficient), C (satisfactory), B (good), and A (good) excellent / excellent). We also divided the final grade into two categories: ”pass” and ”failed.” This helped us to

compare marks. Finally using the dataset we dropped the attributes that are unique to the student and shows the lowest relationship with the target attribute. We use the train and test split to measure the accuracy of our model we split the data set into two training datasets 80% for training and the rest remaining 20% is utilized for the testing purpose. Using the various machine learning techniques mentioned above we test our data and find the accuracy of the model, at last we find the algorithm with the highest accuracy for predicting the student grades.

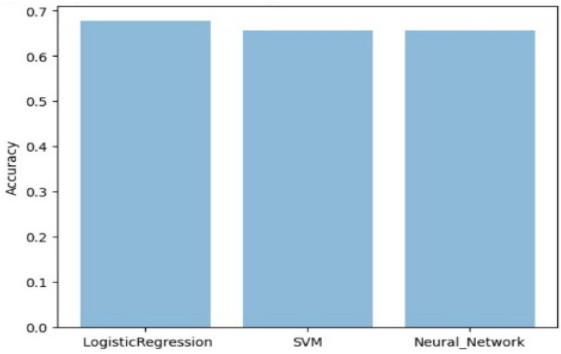

Figure 3. Result Graph

7. Conclusion and Future Work

Finally, student success monitoring is a big issue. It's vital that they're dealt with. The study presented in this thesis demonstrates the use of machine learning methods in conjunction with supervised learning algorithms to better understand the efficiency of the algorithm with respect to student records, where we analysed student performance and classified it into three categories: high, medium, and poor, with a 79% precision We will have some technological solutions in the future to increase the quality of student success. The user interface model may be developed to automatically include student records and to send staff warning messages about students who are doing poorly. We may use a Neural Network to construct the prediction and anticipate improved performance. Non-academic qualities should be combined with academic attributes.

References

[1] Sai Ramesh LS, Ganapathy S, Bhuvaneshwari R, Kulothungan K, Pandiyaraju V, Kannan A. Prediction of user interests for providing relevant information using relevance feedback and re-ranking. International Journal of Intelligent Information Technologies (IJIIT). 2015 Oct 1;11(4):55-71.

[2] Selvakumar K, Sai Ramesh L, Kannan A. Enhanced K-means clustering algorithm for evolving user groups. Indian Journal of Science and Technology. 2015 Sep 1;8(24):1-8.

[3] Pardo A, Gašević D, Jovanovic J, Dawson S, Mirriahi N. Exploring student interactions with preparation activities in a flipped classroom experience. IEEE Transactions on Learning Technologies. 2018 Jul 23;12(3):333-46.

[4] Shaleena KP, Paul S. Data mining techniques for predicting student performance. In 2015 IEEE international conference on engineering and technology (ICETECH) 2015 Mar 20 (pp. 1-3). IEEE.

[5] Shahiri AM, Husain W. A review on predicting student's performance using data mining techniques. Procedia Computer Science. 2015 Jan 1;72:414-22.
[6] Meier Y, Xu J, Atan O, Van der Schaar M. Predicting grades. IEEE Transactions on Signal Processing. 2015 Oct 30;64(4):959-72.
[7] Arsad PM, Buniyamin N. A neural network students' performance prediction model (NNSPPM). In 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) 2013 Nov 25 (pp. 1-5). IEEE.
[8] Gray G, McGuinness C, Owende P. An application of classification models to predict learner progression in tertiary education. In2014 IEEE International Advance Computing Conference (IACC) 2014 Feb 21 (pp. 549-554). IEEE.
[9] Buniyamin N, bin Mat U, Arshad PM. Educational data mining for prediction and classification of engineering students achievement. In2015 IEEE 7th International Conference on Engineering Education (ICEED) 2015 Nov 17 (pp. 49-53). IEEE.
[10] Kleinbaum DG, Klein M. Logistic Regression A Self-Learning Text. 3rd ed. New York: Springer-Verlag New York. 2010.
[11] Kotsiantis S, Pierrakeas C, Pintelas P. Predicting Students ' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence. 2004 May 1;18(5):411-26.
[12] Muñoz-Bullón F, Sanchez-Bueno MJ, Vos-Saz A. The influence of sports participation on academic performance among students in higher education. Sport Management Review. 2017 Aug 1;20(4):365-78.
[13] Hamsa H, Indiradevi S, Kizhakkethottam JJ. Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Procedia Technology. 2016 Jan 1;25:326-32.
[14] Sarker F, Tiropanis T, Davis HC. Linked data, data mining and external open data for better prediction of at-risk students. In2014 International Conference on Control, Decision and Information Technologies (CoDIT) 2014 Nov 3 (pp. 652-657). IEEE.
[15] Huang S, Fang N. Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques. In2012 Frontiers in Education Conference Proceedings 2012 Oct 3 (pp. 1-2). IEEE.
[16] Anand J, Srinath D, Janarthanan R, Uthayakumar C. Efficient Security for Desktop Data Grid using Fault Resilient Content Distribution. International Journal of Engineering Research and Industrial Applications. 2009;2(8):301-13.
[17] Vaishnavi R, Anand J, Janarthanan R. Efficient security for Desktop Data Grid using cryptographic protocol. In2009 International Conference on Control, Automation, Communication and Energy Conservation 2009 Jun 4 (pp. 1-6). IEEE.
[18] Anand J, Janarthanan R, Kannan P, Konar A. Efficient Data Storage in Desktop Data-Grid Computing using Real-Time parameters. International Journal of Computer Science and Technology. 2011 Sep;2(3):392-7.
[19] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3th ed. Morgan Kaufmann publications. 2012.
[20] Kelleher JD, Mac Namee B, D'arcy A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press; 2020 Oct 20.
[21] Sabena S, Kalaiselvi S, Anusha B, Ramesh LS. An Multi-Label Classification with Label Correlation. Asian Journal of Research in Social Sciences and Humanities. 2016;6(9):373-86.