PREDICTION OF STUDENT LEARNING DIFFICULTIES AND PERFORMANCE USING REGRESSION IN MACHINE LEARNING

K. Bhuvaneswari, Assistant Professor
Guru Nanak College, Chennai.
bhukrish06@gmail.com

G. Kanimozhi, Research Scholar
University of Madras.
kanimozhi.may2004@gmail.com

V.Shanmuganeethi, Professor
NITTTR Chennai
shanneethi@nitttrc.ac.in

ABSTRACT

Predicting student’s learning difficulties and student’s performance is a significant research area. Because it can help teachers stop students from quitting before end-of-semester exams, predicting students' learning challenges and performance is an important research subject. The goal of this study is to forecast the learning challenges that students will face in current and upcoming courses. Some kids with learning difficulties could have trouble paying attention in class, reading, writing, or doing math. Universities are increasingly predicting student success using machine learning and big data analytics. Based on cognitive and academic data, researchers have categorized students and forecasted their future outcomes using cutting-edge statistical approaches and machine learning algorithms. In addition to classification tasks, machine learning algorithms like SVM (Support Vector Machine) can be used to forecast learning problems. It can be used to anticipate learning issues, which was done in this study to anticipate the difficulties that particular students would face. Students' learning challenges can also be predicted using the VARK analysis. These results suggest that the SVM algorithm can be an effective tool for identifying students' learning issues. However, the data's quality, feature choices, and applied settings all affect how accurate the performance is. Hence, before making any forecasts, the data must be thoroughly gathered, pre-processed, and analyzed. In order to improve student learning, academic instructors can take proactive steps by using prediction of student academic results to gain a better idea of how actively or poorly students fared in their classes. In order to predict students' final academic performance, this study used undergraduate computer science students for categorization using the SVM method and prediction using linear regression. Some of the key causes of learning difficulties were discovered.

Keywords: Educational data mining (EDM), machine learning - Learning difficulties - Support vector machine (SVM) Performance of the class as measured by a linear regression model.

Introduction

Everyone has struggles while studying, and getting through these obstacles is an important aspect of learning, particularly for students who have a lot of courses to complete. These problems can range from inadequacies at the start to persistent lack of drive and decreased productivity. They will be able to conquer these challenges with their focused mindset and a lot of determination. In this essay, we covered a few persistent causes of learning difficulties and study issues that can impact students at any stage of their academic careers. Children with learning disabilities may struggle with reading, writing, math, or maintaining focus in class. They could also display indicators of poor social and emotional health, such as disengagement.

Using data from educational databases to apply data mining and machine learning techniques is known as "educational data mining," and it is one way to predict students' success. The goal of the educational data mining sector is to use machine learning algorithms and data mining techniques on data that can be retrieved from educational institutions in order to assess the behaviors of the students and improve their learning process.

A machine learning algorithm called SVM (Support Vector Machines) can be utilized for categorization tasks, including foretelling learning challenges. Gather information from a range of variables, such as demographics, past academic performance, classroom conduct, medical history, etc., that may be associated with learning challenges. After processing the data, any unnecessary or incomplete information will be removed. Also, normalize the data to make sure that every feature is scaled equally. to choose the preprocessed data's most pertinent attributes. This can be accomplished by investigating the relationship between various attributes and the desired outcome (learning difficulties). SVM looks for the best hyper plane (also known as a decision boundary) that can distinguish between various classes (e.g., students with learning difficulties and those without).

One of the effective methods used in supervised machine learning to predict student performance and outcomes is regression. A training dataset is mapped with a test dataset, just like classification, to produce predictions. Data was carefully analyzed, and this method can predict outcomes more accurately than the alternative. The output variable in classification differs from the regression variable in that it is categorical data rather than numerical data. It uses datasets with more than two or more variables as input.

Review of Literature

The review of literature is provided as follows after studying the review papers in order to comprehend the work done in previous years related to the supervised machine learning algorithm (SVM) and to specify the thesis as a well-structured idea about how the linear regression approach is used in predicting students' academic performance:

Albreiki & et.al (2021) used machine learning to analyze data logged by the TEL system known as the Digital Electronics Education and Design Suite (DEEDS). To predict learning difficulties with an accuracy of 80%, the author used machine learning methods from the SVM classifier. A DEED interacts with ANN algorithm to study how students learn in order to enhance performance. Shehri (2017) proved that Comparison of various students' performance in using ML techniques SVM – SMOTE is more efficient and Random Forest classifier achieved best result.

Aissaoui and Ouafae (2020) the author discussed how various educational data sets can be predicted or classified by machine learning algorithms. The author also discussed how the results could be improved by taking into account various types of data selection and machine learning algorithms.

Rasheed and Wahid (2022) provides a necessary framework for pedagogical support in order to facilitate decision-making processes in higher education towards sustainable education. It uses a deep artificial neural network to predict at-risk students and offer early intervention measures based on a set of distinctive handcrafted attributes extracted from clickstream data from virtual learning environments. This model's classification accuracy ranges from 84% to 93%.

Felder and Silverman (1988) has made an effort to pinpoint characteristics that can be used to identify multiple intelligences as well as characteristics associated with the Felder-Silverman model of learning performance. Data was gathered, and algorithms including decision trees, Support Vector Machines, K-Nearest Neighbor, Naive Bayes, Linear Discriminant Analysis, Random Forest, and Logistic Regression were used to test the accuracy of classification algorithms. The SVM algorithm had the best accuracy for identifying the learner's dominant intelligence.

Waheed & et.al (2019) compares various resampling methods to address the issue of imbalanced data while projecting student performance using two different datasets, including Borderline SMOTE, Random over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek. Although this paper used the Friedman test to determine the best resampling technique, it appears that classifiers perform better on data that has been balanced by the SVM-SMOTE method. The test's results demonstrate that SVM-SMOTE outperforms other resampling techniques in terms of performance. Additionally, when using the SVM-SMOTE resampling technique, the Random Forest model outperformed all other classifiers in terms of results.

Khan (2021) demonstrated that naive bayes and ripper are the best algorithms for predicting students' performance, with naive bayes having some advantages over RIPPER. In order to predict the student's performance, Raheela Asif. used Decision Trees, K-Nearest Algorithms, Rule Induction, Naive Bayes, and Artificial Neural Networks. On the student data throughout the experiment, they applied SVM regression and IBK classifiers.

Hussain & et.al (2018) has put a lot of effort into predicting student performance in order to achieve a variety of goals, including: identifying at-risk students, ensuring student retention, allocating resources and courses, and many others. The outcome was then determined by comparing the calculated accuracy rates of the Linear Discriminant Analysis (LDA) and SVM classification algorithms, which were the highest and were deemed to be highly significant in producing a reasonable accurate prediction rate of student performance.

Boddeti and Bala (2020) examines the various applications of various educational data mining and machine learning techniques to forecast at-risk students' academic performance in educational settings, identify and forecast students' dropout from ongoing courses, and assess students' performance based on dynamic and static data. Finally, this review succeeded in improving student performance by identifying at-risk students and dropouts, demonstrating the value of using both static and dynamic data.

Yağcı (2022) suggests a brand-new machine learning-based model to forecast undergraduate students' final exam grades. In order to predict the students' performance, the effectiveness of the machine learning algorithms random forests, nearest neighbour, support vector machines, logistic regression, Naive Bayes, and k-nearest neighbour was calculated and compared. Waheed. (2020), on the other hand, discovered that the SVM algorithm outperformed the LR algorithm. SVM, the NN algorithm, and the decision tree are the algorithms with the highest and lowest performances, respectively, according to Xu. (2019). This finding suggests that RF, NN, and SVM algorithms perform more effectively.

Zohair (2019) focuses on supervised learning, more specifically on predictive analysis because it can alert teachers to students who may drop out of a course and can offer extra support to students who need to improve their academic performance. In this research paper, the author looks into a variety of methods for quantifying variables with the intention of choosing one that can identify the most crucial elements that go into creating a model for predicting student performance. The authors used R software to apply data mining techniques and machine learning algorithms that are used to predict outcomes. They also tested a regression model on the dataset to see if there are any factors that affect student performance.

Research Methodology

Using student datasets to predict student performance is one use case for educational data analytics. It is regarded as the most established and well-liked use of data mining in the field of education. In order to test the impact of various factors on student performance, a model is created in this study to predict student performance using linear regression. This study uses a machine learning algorithm called linear regression to forecast how well students will perform in various courses. Getting the data set needed for the research work is the first step in the implementation process. A dataset containing the data of the students is subjected to the methodology. Figure 1 illustrates the system's architecture. Student datasets, which are independent variables, should be used as the study's input data. This dataset, which contains information about students, should be viewed as a tabular format (that is age, gender, academics record, personal information).

The dataset was first gathered, and it was then divided into training and testing datasets. A training dataset is used in a dataset to build a model, and a testing dataset is used to verify the model.

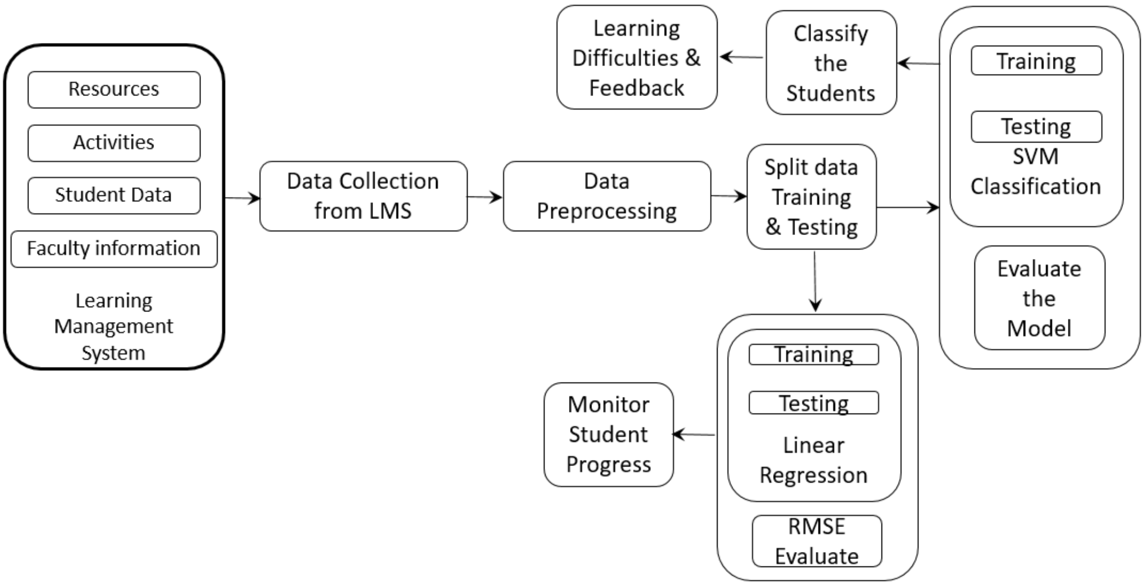


Figure -1: Architecture of prediction of Student performance and Learning Difficulties

Results have been generated based on the prediction model, and they demonstrate that only study time and absences can have an impact on students' performance. Predictive modelling is frequently used in educational data mining techniques to forecast student performance. Classification, regression, and categorization are some of the tasks used to construct the predictive modelling. Regression analysis focuses on attributes and prediction methods as the two main predictors of student performance. The dataset description for the classification and regression model is shown in Table 1.

The sample data were the ones used in this study. 100 students make up the sample dataset. We deal with 150 instances and 8 attributes in this study. The following figure includes information on all the independent and dependent variables.

S. No.	Variable	Description	Data Type
1	Gender	Student Gender	Categorical
2	Age	Student Age	Numerical
3	Study Time	Study Time duration of the student	Numerical
4	Absent Days	No. of days student is absent	Numerical
5	Parent Education	Education of the parents	Categorical
6	Travel Time	Time Taken by the student to travel	Numerical
7	Test Preparation	Whether the student's preparation for the test is completed or not	Categorical
8	Academic Score	Academic grade of the student	Categorical

Table 1: Data Description

This study focuses on the application of linear regression to student academic performance taking into account the student dataset.

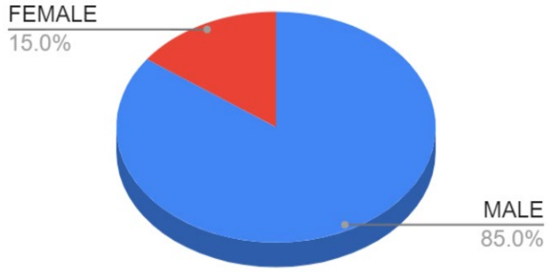


Figure 2: Student data on Gender

In this study, a questionnaire was used to determine the student's preferred learning style in order to create the ideal learning environment. The preferred method or approach that people take when learning new knowledge and skills is referred to as their learning style. You can optimise your learning process by being aware of your preferred learning style. There are four main learning styles, according to Felder-Silverman [14] and the VARK. Here are the specifics:

Visual learners: When information is presented visually, such as through diagrams, charts, or videos, visual learners are more likely to remember it. When studying, they might find it beneficial to use flashcards or other visual aids, and they might find it useful to make mind maps or flowcharts to arrange their ideas.

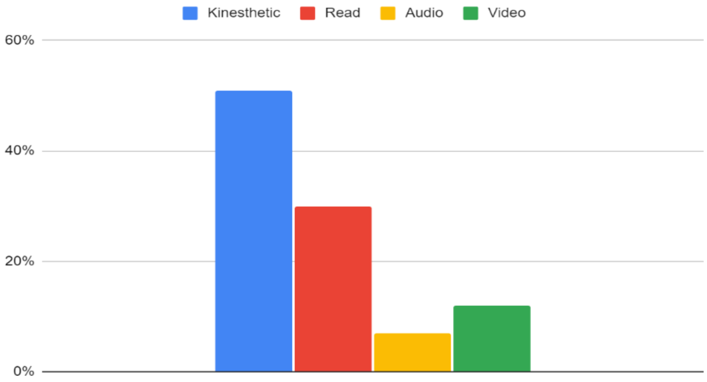
Auditory learners: Learning through listening is preferred by auditory learners, who may benefit from attending lectures, taking part in discussions, or listening to audio recordings. They might also find it helpful to read aloud to themselves or record their notes for later listening.

Kinesthetic learners: Active learning and hands-on experiences are the best teaching methods for kinesthetic learners. Taking breaks to stretch or move around during study sessions may be enjoyable for them, and using manipulatives or other tangible objects to aid in learning may be beneficial.

Reading/writing learners: Writing summaries, outlines, or taking notes can be useful for reading and writing learners who prefer to learn this way. In a journal or another writing format, they might also gain from expressing their thoughts or ideas.

While learning styles can be useful in determining how you learn best, it's important to remember that they are neither absolute nor prescriptive. There are many different ways to learn, and many people combine several different learning styles.

Figure 3: Students learning style based on VARK Analysis
This study discusses some of the most prevalent study issues and learning challenges that students may encounter at any point in their academic careers, as well as some solutions.
Through quizzes, assignments, assessment tests, and classroom activities, instructors can get a thorough understanding of how students are progressing throughout the course and predict how they will perform in various courses. Researchers have used a variety of machine learning and statistical techniques in educational data mining to accomplish this goal. The data were obtained from both conventional and online educational institutions.



It was interesting to look at different approaches to quantifying the importance of variables in this work, and the goal of this research is to choose one of those approaches that can help one identify the most crucial factors that go into creating a model for predicting student performance. A dataset called the Student Performance Dataset was taken from the UG students' Machine Learning Repository and used in this study.

Analysis

This study examined various issues that participants in learning difficulties experienced. The preferred method by which a person acquires, processes, and retains information is referred to as learning style. The most widely accepted model of learning styles identifies three main styles: visual, auditory, and kinesthetic. There are many theories about learning styles, though. Designing teaching strategies and materials that meet each student's unique needs can benefit from learning style prediction in education. The subject matter being learned, the learner's prior knowledge and experience, and the context in which learning occurs are just a few examples of the variables that can affect learning style preferences.

It is necessary to periodically conduct assessment programs to determine the population's literacy proficiency; the results of these programs will serve as indicators of students' success and outcomes. Additionally, it can be useful for identifying each learner's specific challenges. This procedure gives feedback on how well students are able to comprehend and absorb the information being presented to them. A learner's competencies and skills are also identified by educational evaluation as they are acquired during the course of the learning process.

Since research has produced conflicting results, there is currently no way to predict someone's learning style. While some studies have found a relationship between demographic characteristics like age and gender and preferred learning styles, other research points to the complexity and context-dependence of individual learning style differences. As a result, rather than attempting to anticipate and accommodate each student's unique learning preferences, educators may find it more effective to adopt a multi-modal approach to teaching, incorporating a variety of teaching strategies and materials that are suited to different learning preferences.

Through the VARK Questionnaire, student academic characteristics and learning styles are gathered in order to evaluate the proposed research. Although information was fed into the regression model that shows students' learning progress, this analysis provided information on the motivational factors for the students.

The difference between the predicted and true values can be ascertained and minimized by using the regression line that fits the data the best. Two different types of linear regression exist. Simple Linear Regression, which employs a single independent variable, is one of them. Multiple Linear Regression, on the other hand, is the second kind of regression. Multiple independent variables are used in this research study's use of regression analysis of this kind.

Conclusion

Support Vector Machines (SVM), a machine learning classification algorithm, and a linear regression model were used to train the datasets for classification model evaluation and prediction, respectively.

It is crucial for educational institutions to develop strategies that improve their performance system because the caliber of teaching in educational institutions is regarded as one of the development keys of any nation. These plans can be made after evaluating student performance because early failure rate estimation enables educational institutions to take preventative measures to lower that rate. In the near future, efforts could be made to also look into how well the suggested techniques perform when applied to the problem of performance predictions' classification component. Looking forward to putting some of the great works that already exist into practice and putting more of an emphasis on the dynamic nature of student performance.

References

Albreiki, Balqis, Nazar Zaki, and Hany Alashwal. "A systematic literature review of student performance prediction using machine learning techniques." *Education Sciences* 11.9 (2021): 552.

Al-Shehri., "Student performance prediction using Support Vector Machine and K-Nearest Neighbor," *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Windsor, ON, Canada, 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946847.

Boran Sekeroglu, Kamil Dimililer," Student Performance Prediction and Classification Using Machine Learning Algorithms", ICEIT 2019, March 2019. http://doi.org/10.1145/3318396.331.

El Aissaoui, Ouafae, "A multiple linear regression-based approach to predict student performance." *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 1-Advanced Intelligent Systems for Education and Intelligent Learning System*. Cham: Springer International Publishing, 2020. 9-23.

Fareeha Rasheed, Abdul Wahid, (2022)"Learning style detection in E-learning systems using machine learning techniques" Elseveir

Felder, R. M., & Silverman, L. K. (1988). Learning styles and teaching styles in engineering education. Engineering Education, 78(7), 674–681.

Ghorbani, Ramin, and Rouzbeh Ghousi. "Comparing different resampling methods in predicting students' performance using machine learning techniques." *IEEE Access* 8 (2020): 67899-67911.

Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Raheel Nawaz, "Predicting Academic Performance of Students from VLE Big Data using Deep Learning Models, Computers in Human Behavior (2019). https://doi.org/10.1016/j.chb.2019.106189

Khan, Shakir. "Study Factors for Student Performance Applying Data Mining Regression Model Approach." *International Journal of Computer Science & Network Security* 21.2 (2021): 188-192.

Mushtaq Hussain, Wenhao Zhu, et.al" Using machine learning to predict student difficulties from learning session data", Springer, 2018. https://doi.org/10.1007/s10462-018-9620-8.

Ramin Ghorbani and Rouzbeh Ghousi," Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques: IEEE Access Digital Object Identifier 10.1109/ACCESS.2020.2986809.

Sravani, Boddeti, and Myneni Madhu Bala. "Prediction of student performance using linear regression." *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020.

Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **9**, 11 (2022).

Zohair, Abu, and Lubna Mahmoud. "Prediction of Student's performance by modelling small dataset size." *International Journal of Educational Technology in Higher Education* 16.1 (2019): 1-18.