Application of Student Performance Analysis Using Machine Learning Algorithms

Perusomula Manohar Reddy[1], Dr.C.Hemalatha,Ph.D[2], Gajjala Amruth Reddy[3]
[1,2,3]*Department of CSE, Sathyabama university, Chennai, India*

Abstract: This post presents a methodology to predict students' academic achievement in a higher education environment. Logistic regression is a machine learning technique that is decision tree-based. Also, the importance of a number of distinctive qualities, or "features," is considered in order to determine which of these are connected to student achievement. The results of an experiment showing how well machine learning works in this application are presented last.

1. INTRODUCTION

Together with statistical analysis and data mining, the machine learning methodology is one of the most often used approaches for analysing student performance or success. Academic performance is a difficult problem for postsecondary schools all across the world. Data analytics is a technology that is being used by numerous educational institutions to identify students who are having academic difficulties and to increase throughput. The project fits the definition of educational data mining (EDM). EDM is a branch of data mining that focuses on the creation, assessment, and use of a number of automated methods for the study of enormous amounts of data from academic contexts.

This investigation looks at how a student's activity in a learning management system relates to their academic growth. Participation is a key determinant of a student's success or failure in any learning environment. There are several options for how education is delivered, including traditional classroom settings, online learning, blended learning, and non-traditional methods. The learning management system is a tool that facilitates digital communication between instructors and students (LMS). LMS systems are utilised in learning on a regular basis for a number of reasons.

The ease of learning at the student's speed, increased cost-efficiency for the institutions, and comprehensive coverage of a sizeable number of students are a few factors that contributed to the acceptance of LMS. The phrase "hybrid learning," often known as "blended learning," describes a mix of traditional and modern educational environments. an effective teaching technique, physical interaction between students and instructors, the improvement of academic achievement, and learner autonomy are some examples of aims for blended learning. Yet, failure rates in blended learning undergraduate programs have recently sharply increased. The COVID-19 pandemic has encouraged the adoption of one kind of online learning or another, which has accelerated research into the elements that enhance student achievement in this mixed learning environment. In an effort to help students, reduce the number of students who drop out of blended learning courses, and improve academic success, this research provides machine learning models with the best performance in predicting undergraduate student accomplishment.

2.LITERATURE SURVEY

2.1 Classification-based Analysis of Educational Data Mining

Higher education institutions frequently give a lot of weight to overall student achievement. Thus, teachers must utilise a range of methodology, including physical testing, statistical tools, and the most popular data mining techniques, in order to forecast students' success. In the developing academic field of educational data mining, data mining techniques are applied. To assist the user in comprehending a student's academic achievement, study habits, and potential for improvement, it employs statistical methodologies and machine learning algorithms. The various data mining methods that may be used to assess students' performance levels will be evaluated in this study. Research in data mining and machine learning is currently of utmost importance, and educational institutions lay great focus on it. Those are some of the best study places. for finding relevant information gleaned from historical data stored in huge databases. Data mining for education, also known as Educational Data Mining, is the application of data mining techniques in educational situations (EDM). It is an important field of study that enhances our capacity to foresee useful information from educational databases in order to improve academic performance and get a deeper comprehension of how students learn. The best method for mining educational data may be a combination of learning science and data mining. Using

educational data mining to the development of a user perception, action, and trial model may be advantageous. Knowledge discovery, another name for data mining, is gaining popularity. because it is so good at analysing data from all angles and turning it into usable knowledge. K-nearest neighbour, neural networks, decision trees, support vector machines, naive bayes, and many more data mining methods are used in education. Using data mining technologies, there are numerous methods for swiftly analysing data. Examples of open-source software made for data analysis and to lay a solid basis for future use include Weka, Fast Miner, Orange, Knime, and SSDt. In this study, WEKA (Waika to Environment for Knowledge Analysis) is used since it is most suited for data analysis and building models that produce predicted results.

2.2CLASSIFICATION TECHNIQUES

A very beneficial and effective method for making judgements is data mining. One of the most popular and essential data mining techniques is classification. A working comprehension of training data is important to comprehend classification. Two steps make up the categorising process: the construction of a training model, analysis of test data, and model evaluation

Algorithms are used in a number of classification methods, including:

Algorithms based on statistics: Instead of only providing a simple classification, statistical processes usually use a core probability model that is accurate and offers probabilities of belonging to each class.

Correlation analysis, a statistical technique, is used to evaluate the strength of the relationship between two continuously recorded, numerical variables (such as age and weight). a little unique.

Bayesian Model : The Bayesian Model is a frequentist theory-based approach. The use of probability in data analysis is the cornerstone of frequentist methodology. Calculations using the Bayesian approach only consider the hypothesis' probability.

Any item mapped to a certain class may be seen as identical to other objects already existing in that class using distance-based approaches.

unlike those of other classes, the objects. There are two methods for classifying things based on their distance, i.e..

The basic technique makes the assumption that every class's centre is a representation of that class. A new item may join the class that has the highest possible similarity value.

nearest neighbours in K It is a non-parametric method that depends on measurements of distance.. When a new instance is entered, it may be classified using the distance function while still keeping all of the cases that are now open for access. Using decision trees as the base As stated by Using this strategy, a tree must be built to represent the categorization process. Two phases make up this procedure.

Method of classification:

a. Build a tree named with Decision Tree
b. Implementation of Decision tree to database

Neural network-based algorithms: With this technique, a model that offers a structure for data representation may be built. All of a tuple's relevant attributes are routed into a graph at the time of classification. Rule-based algorithms: This approach permits the application of if-then-else rules to data classification..

2.3CLASSIFICATION

2.3a.DATA SET DESCRIPTION

The learning management system-gathered Kalboard 360 dataset, which relates to education, is used in this study (LMS). This kind of technology always enables users to use up-to-date instructional resources by using a device and an Internet connection. The learner activity tracker tool is an application used to gather data. There are 480 student entries in this collection, along with 16 attributes. The top three feature subcategories are as follows:

- Characteristics of the population, such as gender and nationality
- Educational characteristics like grade level, section, and stage; psychological
- Characteristics like raised hand in class;

2.4 Prediction Graphs for Student Performance Using Markov Networks in the Classroom

Colleges and universities are becoming more and more interested in gauging student development as they monitor and work to improve their retention and graduation rates. Effective interventions should be devised based on early markers of student progress—or lack thereof—to enhance the possibility that students will succeed. Here, we offer a technique for calculating students' performance in the early stages of their academic careers utilising data mining and machine learning techniques, namely linear regression and a Markov network (MN).

The findings suggest that the suggested framework, when used to analyse performance over the course of

one semester, may accurately predict student development, or more specifically, student grade point average (GPA) within the planned major. Also, as performance increases, the predicted GPA for the upcoming semesters becomes more precise, enabling professors to discuss with students their anticipated academic achievement early on in their academic careers.

A few of the index terms include student performance, Markov networks, linear regression, and educational analytics.

2.5IMPLEMENTED TECHNIQUES

Markov Networks

Markov networks are probabilistic models that may be represented by undirected graphs and can contain any number of cycles, in contrast to directed graphical models. The probability distribution considers the largest cliques, or clusters, of fully linked nodes in the network. There is a function c that might be used to deliver a positive value for each maximum clique c for maximal clique c signifies the subset of random variables x(c). The potential functions c are not always associated with marginal distributions of subsets of nodes and do not always have a probabilistic interpretation. The joint distribution of an MN can be described by the partition function or normalisation constant that guarantees that p(x) integrates to a particular value. 1. the restricted The MN may be defined in terms of the independent properties of each random variable. Each node v is conditionally independent of every other node with regard to its close neighbours. The process of drawing conclusions about variable x from the observable or visible variable y is known as inference. Several variables (nodes) in the graphical model have values that are often seen in real-world applications. a unique, multiple, or combination marginal distribution of the relevant variables. For the purposes of our goals, the posterior distribution p(x|y) will be used to decide how all unknown variables, x, should be set in relation to observed y. Because exact inference, models are typically relatively conservative. Inference in graphics is not frequently used. difficult There are several methods for approximating inference. These types include graph-cuts, variational, sampling-based, (local) optimization, etc. It should be noted that a pairwise MN with maximum cliques linking only pairs of nodes was created for this experiment.

using our framework

Achievement in one area might be a good indicator of future success in other areas or a student's ability. To put it another way, a child's past academic performance may be a reliable predictor of future success. An "A" high school student, for instance, is frequently believed to perform better in college than a "C" student if all other conditions are equal.

Students who earned a "A" in calculus II should do better in calculus III than those who do so. A consequence. \ We can see that these requirements and the functionalities offered by the MN mesh flawlessly. A student's future performance may be successfully influenced by the MN based on their past academic achievement due to factors including age, gender, parental education level, emotional issues, etc.

We believe that providing new data will enhance the functionality of our system. By monitoring a child's growth, this app can help identify opportunities for early intervention and improve student outcomes.

3 SYSTEM DESIGN

3.1Existing system:

• Data mining techniques are creating a tonne of computerised information across a variety of fields.

• The creation of student accomplishment prediction models to predict students' performance in academic institutions is one of the key areas of the development of education data mining.

• Based on their grades from the 10th, 12th, and previous semesters, a prediction approach has been given.

The study's analysis makes use of biological regression, decision trees, entropy, and KNN classifiers.

3.1.a. Disadvantage:

The result is unclear.

The project accuracy is low, the approach is not suitable for the dataset, this system has certain characteristics based on mark, and it is based on the 10th and 12th grades.

System proposed in 3.

• The ultimate goal of any educational institution is to give students the best information and learning opportunities available.

Reaching this goal will entail identifying the students who need more assistance and taking the appropriate actions to raise their performance.

• Using four machine learning algorithms, a classifier that can predict students' study performance was created in this study.

• The student mark system, question-based learning, and emotional recognition form the foundation of our feature.

advantage:

• easier to use, comprehend, and very effective for training.
• understand the relationship between the dependent variable and one or more independent variables.
• Outstanding performance
• The accuracy is quite good.
• This system's best algorithm has the highest output efficiency.

Block diagram:

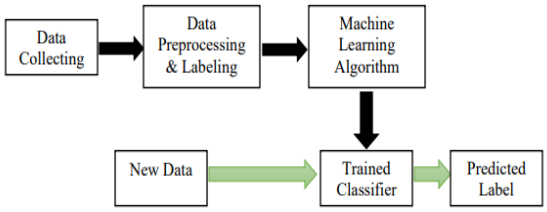

Fig 1 block diagram

Flow diagram:

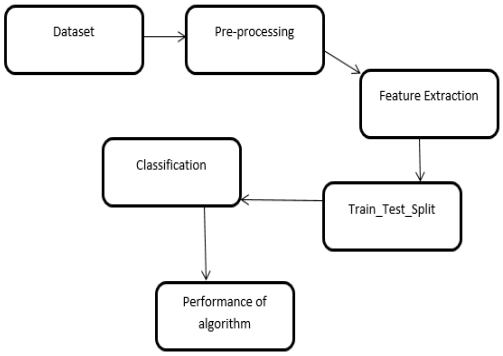

Fig 2 flow daigram

4.DATA PROCESSING AND AASSESMENT
4.1 Modules
Data Collection
Data Pre-processing
OneHot Encoding
Training and Testing
Performance Evaluation
Prediction

4. 1a. Dataset gathering
• By collecting data, you may maintain track of prior occurrences so that data analysis can be used to look for recurring patterns. You may build prediction models using machine learning techniques from these patterns that follow trends and anticipate future changes.
• As predictive models are only as good as the data on which they are based, effective data collection procedures are crucial for producing high-performing predictive models.

4.1b. Cleaning up data
• In this module, data cleaning is done to make the data ready for analysis by removing or changing any information that could be incorrect, missing, duplicate, or poorly organised.
Any machine learning effort must begin with data cleansing.

4.1c. Feature Extractio
• Training time is halved while accuracy is enhanced by lowering the amount of features in the dataset.
• The feature extraction method is used in machine learning, pattern recognition, and image processing to create derived values (features) that are meant to be useful and non-redundant from a set of measured data. This mechanism accelerates generalisation and learning, and under some circumstances, it improves human interpretations. Dimension reduction and feature extraction go hand in hand.

4.1d.Model training
A dataset used to train an ML algorithm is called a training model. Both individual instances of output data and important input data groupings that have an impact on the outcome are included. In order to compare the processed output to the sample output, the input data are sent through the algorithm using the training model. The model is modified using the correlation's findings. Model fitting refers to this iterative procedure. To produce an accurate model, the training dataset or validation dataset must be exact.
•Giving data to a machine learning algorithm is known as "model training," which enables it to recognise and learn the acceptable ranges for all linked attributes.
4.1e. Testing model : The machine learning model developed during this session is evaluated using the test dataset.
• In order to ensure that the software system meets the requirements, quality assurance must be used. Has the full potential of each feature been utilised? Is the software operating properly? A list of all the standards you use to assess the application must be included in the technical specification document.
• Any mistakes and flaws connected to development may be found during software testing. You don't want your audience to come to you in tears if the application's issues are discovered after it has been made public. Finding problems involves using a variety of testing methods that are not apparent until runtime.

4.1.f.Performance Evaluation

• Using performance assessment metrics including F1 score, accuracy, and classification error, we test the efficacy of taught machine learning models in this session.

•If the model isn't functioning properly, we alter the machine learning algorithms to improve it.

• Efficiency Evaluation is a method that is carefully thought out and carried out to assess an employee's performance in relation to their duties at work. It is used to evaluate how much value an employee brings to the company in terms of improved revenue in comparison to industry norms and total employee return on investment (ROI).

4.1.g.Prediction

- When assessing the likelihood of a certain outcome, such as whether or not a customer would leave in 30 days, "prediction" refers to the output of an algorithm that has been trained on past data and applied to current data.
- The algorithm will provide likely values for each record in the new data, assisting the model builder in determining what the value of the unknown variable will most likely be.
- used a decision tree, an algorithm, and logistic regression
- The basic concept of logistic regression is the use of a logistic function to describe a binary dependent variable, however there are many more sophisticated expansions. Regression analysis, a type of binary data analysis, uses logistic regression, also known as logit regression, to compute the parameters of a logistic model.

By computing probabilities using a logistic regression equation, it is possible to utilise statistical software to comprehend the relationship between the dependent variable and one or more independent variables. With the use of analysis, you may gauge the probability of an action taking place or a decision being reached.

con.....

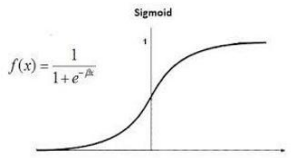

Fig 3 performance graph

It is common to use CART, C4.5, C5, and ID3 decision trees. An input variable (X) and a split on that variable are represented by a node if the variable is numerical.

• A decision tree frequently starts the process by selecting a root node, also known as the tree's terminal node, since it has an output variable (y) that is crucial for prediction. Calculate each node's information gain or entropy before the split. Determine which node has a larger information gain or less entropy. Split the node once again to repeat the procedure.

CHAPTER 5

CONCLUSION:

According to recent studies, a student's former accomplishment has a significant impact on their future academic success. Our research shows that a student's achievement is significantly influenced by their earlier performance. In addition, we demonstrated that neural network performance scales with dataset size. Since its inception, machine learning has come a long way, and it has the potential to be a useful tool for academics. Any academic institution may incorporate future applications like this one, along with any improvements made to them.

REFERENCE

[1] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Preventing student dropout in distance learning systems using machine learning techniques,” AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, pp. 3-5, September 2003.

[2] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[3] Erkan Er. "Identifying At-Risk Students Using Machine Learning Techniques", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp. August 2012.

[4] S. Kotsiantis, I.D. Zaharakis, and P. Pintelas, "Assessing Supervised Machine Learning Techniques for Predicting Student Learning Preferences ““Performance funding for higher education,” National Conference of State Legislatures, Feb. 2012. [Online]. Available: http://www.ncsl.org/