Predicting Student Performance Using Clickstream Data and Machine Learning

M. Kowsalya[1], B. Akilan[2], E. Devendran[3], N. Kowsik Kumar[4]
Assistant Professor[1] ,Students [1 2 3] ,Department of Computer Science and Engineering, Erode Sengunthar Engineering college, Perundurai, Erode, Tamil Nadu, India.

ABSTRACT-Predictive analysis of student performance has grown in importance in education in recent years. It makes it possible to learn more about how to improve teaching and learning, identify students who are at risk, and comprehend how students learn. Learning The board Framework information have as of late been utilized by a great deal of scientists to foresee understudy execution. This project focuses on the use of clickstream data for this purpose. A lot of student databases are collected from reliable sources like KAGGLE. The prediction of student performance is the output, and the data serves as the input. The LINEAR REGRESSION algorithm serves as the foundation for the proposed method.

KEY WORDS : : Learning analytics, Educational data mining, Linear regression, Random forest,Virtual learning environm

I. INTRODUCTION

Learning Assessment (LA) is a creating investigation field. " The most common definition of LA is "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." As the definition proposes, LA is associated with instructive information accumulated from Virtual Learning Conditions (VLEs), like Learning The executives Frameworks (LMSs), and PC upheld learning conditions. Huge instructive informational indexes have become open because of the expanded utilization of LMSs throughout recent many years. The examination of these informational collections can give a more profound perception and understanding into the growing experiences and encounters of students on the off chance that the fitting LA and information mining procedures are utilized. In current LA-based research, descriptive, diagnostic, predictive, and prescriptive analytics are some of the methodologies examined. In training, prescient examination involves gathering mysterious future occasions or results connected with educating or learning. The effect of a specific informative system on students or the enlistment for a course, understudy maintenance, or other part of educating can be anticipated by certain undertakings. Foreseeing scholastic achievement, course grades, or expertise obtaining are instances of different assignments that put an accentuation on learning and the points of view of understudies. The prediction of student performance is one important area.

II.MACHINE LEARNING

The focus of the field of study known as machine learning (ML) is the study of methods that "learn," or methods that use data to improve performance on a particular set of tasks. It is believed to be a part of artificial intelligence. A model is built by machine learning algorithms using sample data, or training data, to make decisions or predictions without being explicitly programmed to do so. Machine learning algorithms are used in a wide range of applications, including computer vision, agriculture, speech recognition, email filtering, and medicine, when it is difficult or impossible to develop conventional algorithms that can perform the required tasks. Various approaches are utilized in the field of machine learning to instruct computers on how to complete tasks for which there is no fully satisfactory algorithm. One technique is to name a portion of the right responses as legitimate when there are a great deal of potential responses. This can then be used as training data by the computer to improve the algorithms it uses to find correct answers. For instance, the MNIST dataset of digits written by hand has frequently been used to train a system for digital character recognition. During the 1990s, the redesigned field of AI (ML) started to thrive. The goal of the field shifted from creating artificial intelligence to solving real-world problems. Rather than zeroing in on simulated intelligence's representative methodologies, it directed its concentration toward measurements' techniques and models.

III. DATA ANALYTICS

The most common way of inspecting, cleaning, changing, and demonstrating information is known as information examination. The objective of information investigation is to find valuable data, support navigation, and illuminate ends. Business, science, and social science are just a few of the areas where data analysis is utilized. It encompasses a variety of methods that go by a variety of names, and it has a variety of aspects and approaches. Information examination assists organizations with pursuing more logical choices and run all the more effectively in the present business world. While business insight envelops information investigation that is vigorously dependent on collection and principally centers around business data, information mining is a specific technique for information examination that puts an accentuation on factual displaying and information disclosure for prescient purposes as opposed to simply expressive ones. The three types of data analysis utilized in statistical applications are descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). CDA focuses on confirming or misrepresenting existing hypotheses, whereas EDA focuses on locating new informational highlights. Text investigation utilizes measurable, semantic, and underlying strategies to remove and group data from printed source a kind of unstructured information while prescient examination centers around the utilization of factual models for prescient guaging or characterization. There are numerous approaches to data analysis.

IV. PREDICTION

A prediction or forecast is a statement about data or an upcoming event. They are habitually, however not dependably, in view of information or experience. The precise distinction between the two terms is not universally agreed upon; Meanings are assigned in different ways by authors and disciplines. It is impossible to guarantee accurate information regarding the future due to the inherent uncertainty of future events. Anticipating potential advancements can profit from forecasts.

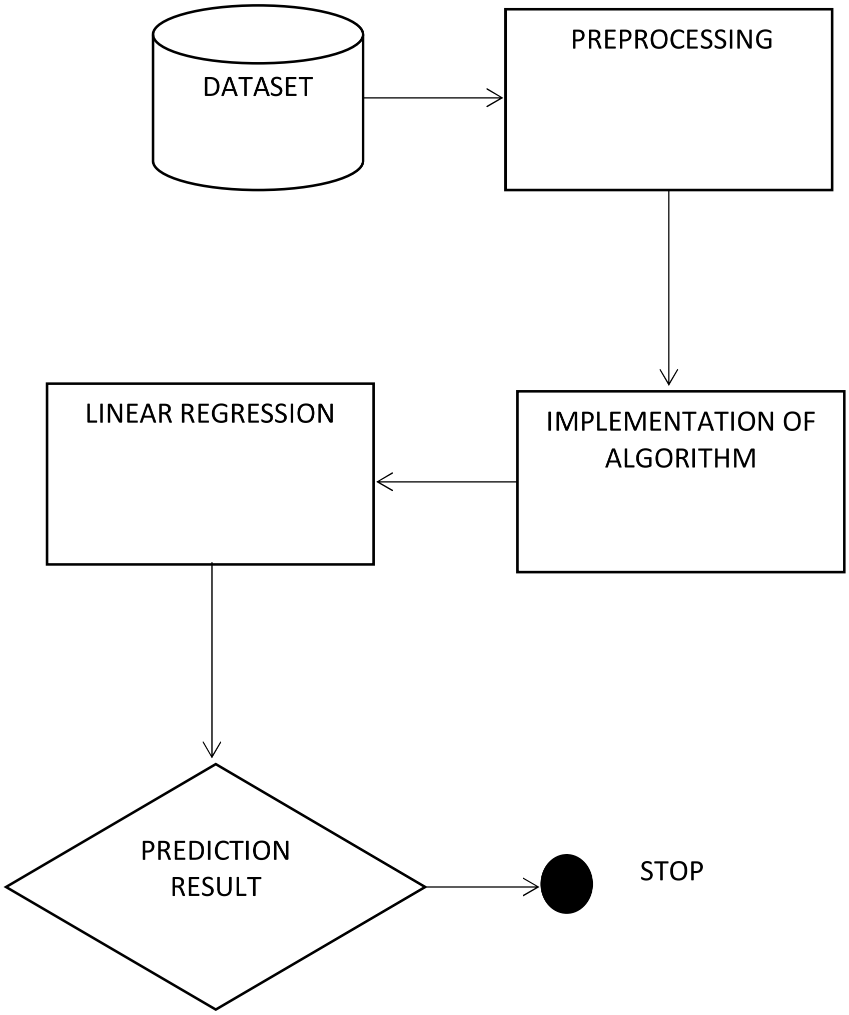
V. PROPOSED SYSTEM

The prepossessing model will go through a free processing step model after the input data are provided. The data set will be divided into two groups of 80 and 20, with measurements being taken on 80% of the train data set and 20% of the tested data set. The input must be provided as a data set in accordance with the proposed methodology. Each student's specifics will be listed in trained rows and columns that make up the data set. In the wake of preprocessing, the Direct Relapse Calculation will be carried out. This calculation is utilized to foresee an understudy's exhibition in light of the contribution with a more elevated level of exactness than the past model. Accordingly, we can expect precision more prominent than 90% over the current lstm model. Machine learning-based prediction gained a lot of traction. We execute the new AI applications in educating and picking up considering the understudies' experiences, past scholarly scores, and different attributes. Due to the large class sizes, which can increase the dropout rate at the end of open learning courses, it would be difficult to assist each individual student. In this paper, we employ the linear regression machine learning algorithm to predict a student's academic performance. The most significant advantage of linear regression models is linearity: It improves on the assessment cycle and, above all, these separately deciphered direct conditions are straightforward. The direct relapse calculation could be utilized to accomplish an elevated degree of exactness

VI. EXISTING SYSTEM

This research trained multiple predictive models using machine learning methods and clickstream data, achieving up to 90.25% (89.25% + 0.97%) accuracy. The results provide insights into effective ways to extract features, train and evaluate predictive models in student performance prediction tasks using students' clickstream data. From a data science perspective, this research contributes three major findings through answering its three Research Objectives: (1) Feature extraction, (2) Feature selection, and (3) Model evaluation. In addition, although this research was conducted based on a data science approach rather than with a pedagogical focus, the identified important features from the best model can inform future course design and teaching interventions. The 90 % of the acuuracy is less when compared with the existing machine learning models such as LSTM this algorithm is .

VII. FLOW DIAGRAM



VIII. INPUT DATASET

The information in an informational collection is taken from a bundle or an information module. The information is provided as the contribution for the AI cycle because the recreation is conveyed in yield, and it contains the column and coulum in the subsequent or csv document information.

8.1. PREPROCESSING

Information can be utilized by organizations to assist them with pursuing better choices and develop their business from practically any source, including inner information, connections with clients, and information from everywhere the web.

On the other hand, programs for machine learning and analytics cannot be immediately run on raw data. For your information to it be effectively "read," or comprehended by machines, you should first preprocess. A step known as data preprocessing transforms raw data into a format that computers and machine learning can understand and analyze during the process of data mining and analysis.

Information from this present reality, like message, pictures, video, etc, is tangled. It may have irregularities and blunders, yet it might likewise be inadequate and not have a reliable plan.

At the point when they read information, machines like to deal with it in a slick and clean manner. Hence, it is easy to compute organized information like rates and entire numbers. Be that as it may, prior to leading investigation, unstructured information like text and pictures should initially be arranged and cleaned.

8.2.LINEAR REGRESSION IMPLEMENTATION

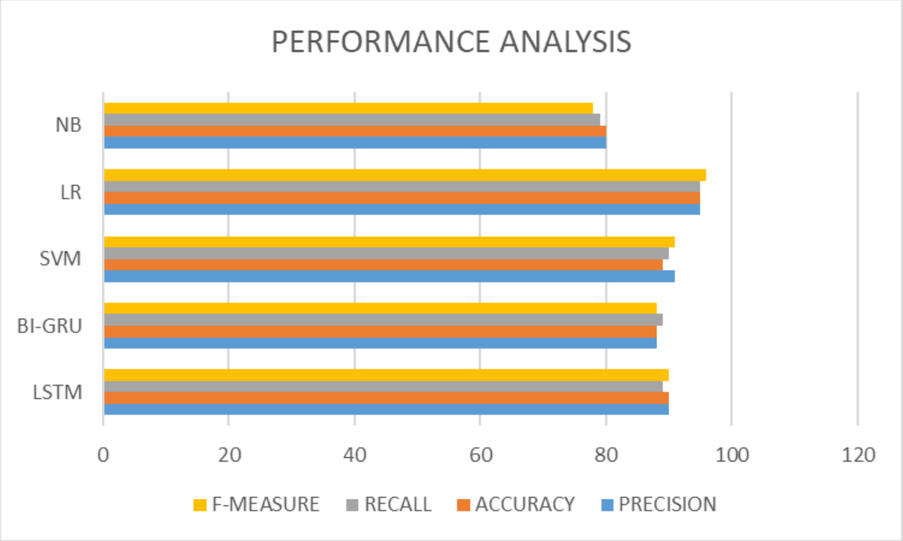
Linear regression analysis is used to predict a variable's value from another variable's value. The one you want to predict is the dependent variable. You can predict the value of the other variable by using the independent variable.

This kind of assessment evaluates the coefficients of the straight condition, including no less than one independent factors that best anticipate the value of the dependent variable. Direct relapse delivers a surface or straight line with the most minimal conceivable distinction among anticipated and genuine result values. For a bunch of matched information, there are direct straight relapse number crunchers that utilize the "least squares" strategy to decide the best-fit line. The independent variable's value, Y, is used to estimate the value of the dependent variable, X. Linear-regression models are straightforward and provide a mathematical formula for making predictions that are easy to understand. Straight backslide can be applied to various locales in business and academic audit.

Direct relapse is utilized in business as well as the organic, conduct, natural, and sociologies. The utilization of straight relapse models to precisely and deductively foresee what's to come is currently deeply grounded. Due to its long history in statistical analysis, linear-regression models have well-understood properties that can be quickly trained.

IX. RESULT

Students' performance can be predicted using performance prediction tasks by utilizing feature importance analysis. The understudies' snap ways of behaving on the course landing page and subpages are the most critical to the forecast errands, as indicated by this study's component significance examination. The second most important activity is accessing the course material via the dataset. The low significance of the remaining activity categories suggests that the patterns of click behavior on those websites have little impact on predicting academic outcomes.



From the above chart and table linea regression model is the highest level of accuracy and provides the better output than any existing algorithm. This performance analysis was tested based on the survey process in the previous papers.

X. CONCLUSION

The purpose of this study is to develop a model for predicting student performance using clickstream data. In the preliminaries, different perceptive models were ready and analyzed, using gathered click data (number of snaps in seven days by week and month to month premise), artificial intelligence estimations Straight Backslide computation is used as the proposed procedure to find the better accuracy exactly as expected the result was gotten. what's more, a technique for choosing highlights. This investigation revealed that the LR model and week-to-week based click include conglomeration as board information are the most effective methods for this understudy execution expectation case. In this occasion, highlight choice is discretionary. Also, this study found that the info dataset from kaggle is critical in anticipating understudy execution by examining significant highlights from the best model. Teachers can utilize the upside of understudies visiting the course's landing page and subpages to further develop the internet learning climate in view of these discoveries. Teaching intervention methods should be used to help students who are at risk.

REFERENCES

[1] Siemens, G., LAK 2011 General and Program Chair Message. In the LAK11 Proceedings: First International Conference on Learning Analytics and Knowledge, 27 February–1 March 2011, Banff, Alberta, Canada

[2] Norst, N.; Hernández-Garcíac, Á. What kinds of information are utilized in learning examination? a summary of six cases. Comput. Hum. Behav. 2018, 89, 335–338. [CrossRef]

[3] Society for Learning Analytics Research Online availability: https://www.solaresearch.org/about/what-is-learninganalytics (accessed August 30, 2022). Educ. Sci.

[4] Akçapnar, G.; 2023, 13, 17, 13 of 14 A. Altun; A skar, P. Developing an early warning system for students at risk through the use of learning analytics. Int. J. Educ. Technol. High. Educ. 2019, 16.

[5] Chen, F.; CrossRef] Utilizing Student Time Series Behavior in Learning Management Systems for Early Course Performance Prediction, Y. J. Study Anal. 2020, 7, 1–17. [CrossRef]

[6] M. Imran; S. Latif; D. Mehmood; Shah, M.S. Understudy Scholastic Execution Forecast utilizing Administered Learning Procedures. Int. Emerg. J. Technol. Learn. 2019, 14, 92–104.

[7] Yang, Y.; CrossRef] D. Hooshyar; Pedaste, M.; Wang, M.; Yang M. Huang; Lim, H., "Prediction of students' procrastination behavior in online learning using their submission behavior pattern." Ambient, J. Intell. Humaniz. Comput. 2020, 1–18. [CrossRef]

[8] C.G. Brinton; Chiang, M. Clickstream data and social learning networks are used to predict MOOC performance. In the 2015 IEEE Conference on Computer Communications (INFOCOM), which took place in Kowloon, Hong Kong, from April 26 to May 1, 2015; pp. 2299–2307.

[9] F. Marbouti; H.A. Diefes-Dux; K. Madhavan's models for using standards-based grading to identify students in a course who are at risk early on. Comput. Educ. 2016, 103, 1–15.

[10] Fischer; E. Potma; Warschauer, M. Supporting first-generation college students in an online chemistry course by utilizing clickstream data mining techniques. In the LAK21 Proceedings: Irvine, California, USA, 12-16 April 2021, 11th International Conference on Learning Analytics and Knowledge; pp. 313–322.