IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

Predicting Student Performance using Data Mining

Mabel Christina

Dept. Of Information Science and Engineering, MVJ College of Engineering, VTU, Bangalore, INDIA

*Corresponding Author: mabelchristina@mvjce.edu.in, Tel.: +91-8553635345

Available online at: www.ijcseonline.org

Accepted: 19/Oct/2018, Published: 31/Oct/2018

***Abstract*—** Data mining focuses on collection information from knowledge bases or data warehouses and therefore the info collected that had never been famous before, it's valid and operational. today instructional data processing is associate rising discipline, involved with varied Approaches like Predicting student performance, Analysis and visual image of information, Providing feedback for supporting instructors, Recommendations for college students, Social network analysis and then thereon mechanically extracts that means from giant repositories of information generated by or associated with people's learning activities in instructional setting. One of the most important challenges is to enhance the standard of the academic processes therefore on enhance student's performance. Thus, it's crucial to line new ways and plans for an improved management of the present processes. This model helps to predict student's future learning outcomes mistreatment knowledge sets of senior students.

***Keywords*—**Data Mining, Educational Data Mining

I. INTRODUCTION

Educational information has become an important resource during this era, contribute abundant to the welfare of the society. [12] Educational establishments are getting additional competitive due to the quantity of establishments growing apace. To remain afloat, these establishments area unit focusing additional on up numerous aspects and one vital issue among them is quality learning. For providing quality education and to face new challenges, the institutions have to understand their potentials that area unit expressly seen and that area unit hidden. The truths behind today's instructional establishment's area unit a considerable quantity of data are hidden. To be competitive, the establishments ought to establish their own potentials hidden and implement a way to bring it out. In recent years, instructional data processing has placed on a mammoth recognition among the analysis realm because it has become an important would like for the tutorial establishments to enhance the standard of education.

The higher education establishments has potential data like tutorial performance of scholars, body accounts, potential data of the school, demographic details of the scholars and lots of different info in a very hidden kind. The technique behind the extraction of the hidden data is data Discovery method. Recently data processing is wide used on instructional dataset. Educational data mining (EDM) has become a awfully helpful analysis space [1]. Data mining helps to extract the data from accessible dataset and will be created as data intelligence for the good thing about the establishment. Teaching will categories the scholars be their tutorial performance. Several factors influence the tutorial performance of the scholar. The model is especially targeted on exploring numerous indicators that have an impact on the tutorial performance of the scholars. The extracted info that describes student performance will be hold on as intelligent data for deciding to enhance the standard of education in establishments. The data stored is employed for predicting the student's performance ahead.

II. RELATED WORK

Data mining in higher training is a current studies subject and this location of studies is gaining popularity because of its potentials to academic institutes. Data Mining may be used in educational area to enhance our knowledge of learning process to attention on identifying, extracting and comparing variables related to the learning process of pupil. Mining in educational environment is known as Educational Data Mining.

Educational information mining is emerging as a research region with a collection of computational and mental methods and research processes for knowledge how college students analyze [11].

[1] T he paper depicts the users, components in addition to the diverse processes in EDM. [2] In this paper a method to improve the pupil's performance is noted via mapping the scholar's document using K-mean clustering set of rules and grouping data sets into cluster however there is no destiny performance prediction.

[5]The paper offer a prediction of Applying records mining method to pick out whether students' on-line learning reviews may be assessed based totally on their log documents however It is restricted to the available records in on line database while factors such as college students' position inside the collaborative organization and Structure of the collaborative responsibilities is not taken into consideration .

[7]The reference introduces a CHAID prediction model to discover the elements influencing the overall performance of students in final examinations and predicting the grade of students the use of .NET framework. Decision Tree, and Multilayer Perception however attributes aside from grades of pupil aren't considered.

[9]The reference paper depicts an Efficient grouping of on line Students records with similar Characteristics using CRISP – DM Methodology information in Moodle Database furthermore No real time information collection is executed.
[6] The reference presents a technique for Improving Quality of Educational Process via the use of clustering, correlation analysis and association rules on the opposite part decision making system for reinforcing the first-rate of educational sports isn't cited.

The paper have an on the spot result on quality assurance in e-learning and on the advance of the teaching method through the difference of content by predicting behavior patterns mistreatment OLAP Analysis, Moodle, LMS but his study offers improvement in terms of the report system within the field of e-learning. however not think about development and implementation of latest modules, likewise as user authentication.

[3][12]The study specialize in predicting students performance and distinctive the slow learners among students employing a comparative study of Classification algorithms like Multilayer Perception, J48 and REP Tree however not targeted on Integration of knowledge mining techniques with database management system and E- learning [14].

[10] The paper depicts a comparative analysis for management of student retention mistreatment CRISP – DM, ten fold Cross validation Approach mistreatment prediction model which will be used as call aid in predicting students retention however Students outcome prediction in real time is loophole.

III. PROPOSED FRAMEWORK

Educational data processing has large quantity of information that has got to be organized during a consistent manner .To org anise, analyse and classify students details K-mean cluster algorithmic rule is been used supported tutorial records. Thereby forming 3 clusters supported students record

- Low performance Student
- Average Student
- Smart Student

But it merely specifies this situations whereas no future prediction is offered and variables used for analysis square measure solely supported demographic
and educational records.

Therefore to reinforce the prevailing system the planned model is meant by collection Students Personal and educational knowledge from the senior students of the establishment and Thereby Grouping the student's performance supported bound conditions as

- best
- good
- average
- Poor

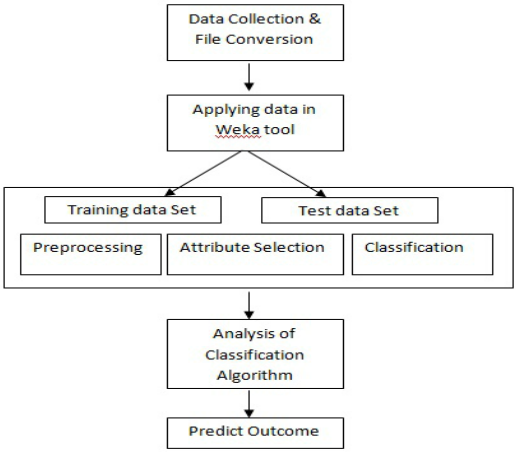

Figure 1. Student performance Framework

Student model is designed for the prediction of the outcome of the student based on the framework given below in Figure 1.This system provides an efficient analysis on student performance by data collection and result prediction.

There is a piece methodology that governs a series of stages. The methodology starts from the matter definition, then knowledge assortment from form and Students info. Attribute choice, Nominal conversion, file conversion and weka tool implementation. Comparative analysis of economical classification rule is completed to predict student's performance by creation of student model

4.1 Data Collection and Preparation

The datasets of around 350 understudies are gathered from I , II, III, IV year B.E CSE) In this procedure, a poll shape is utilized to gather the genuine information from the understudies that portray the connection between learning conduct and their scholastic execution. The factors for making a decision about the learning and scholastic conduct of understudies utilized in the poll are understudy statistic points of interest, School subtle elements, Attendance, CGPA and Final review in last semester. These information's are along these lines recorded in exceed expectations sheets for investigation.. Informational collections around 300 understudies were gathered by 3weeks.Among the dataset around 250 are been utilized as preparing dataset and 50 datasets as test information to plan understudy display.

4.2 Data Selection and Transformation

In this stage just the information required for information mining are chosen. A couple of determined factors were chosen. From the accessible database, a portion of the data for the factors is gathered. The information gathered from Feedback structures and database .at first trait determination is finished. In this progression just those fields were chosen which were required for information mining. A couple of inferred factors were chosen. While a portion of the data for the factors was extricated from the database. The procedure of characteristic choice manages choosing the most suitable traits for characterizing the informational collections. By the investigation among the 24 traits, characteristics of higher positioning are utilized for arranging the preparation dataset.

The attributes are

- CGPA
- Arrears
- Attendance
- 12marks
- Engineering Cut-off
- Medium of Education
- Type of Board

All the indicator and reaction factors which were gotten from the database are given in Table 1.On trait determination the investigation is improved the situation school and school dataset independently in view of specific conditions as given underneath in figure lastly breaking down the execution utilizing last condition on both the records of school and school in light of conditions.

For school points of interest the conditions are connected and assembled as Best, Good, Average and Poor in view of CGPA, ATD and ARR .yet for school subtle elements the gathering depends on MOE, TOB, TWM and ECUT. Data gathered from understudies as input and from database. The profile of understudies is characterized in light of the scholastic and statistic points of interest of understudies.

The understudies' scholastic foundation is estimated utilizing the section necessities to be satisfied to get passage into the college/school. In this stage the main the information required for information mining are chosen.

A few derived variables were selected. From the available database, some of the information for the variables is collected. The data collected from Feedback forms and database are entered in excel sheets and converted to ARFF format for further processing in WEKA tool.

Table 1: Dataset Description

Variable	Description	Possible values
TWM	Twelfth total marks	{best,good,average,poor}
MOE	Medium of Education	{English, Tamil}
TOB	Type of Board	{State board, Matriculation,CBSE, Diploma}
ECUT	Engineering Cut- off	{best,good,average,poor}
CGPA	Current CGPA	{best,good, average, poor}
ARR	No. of arrears	{no,average,poor,very poor}
ATD	Attendance percentage	{best,good,average,poor}

IV. IMPLEMENTATION OF MODEL

The principle objective is to investigate on the off chance that it is conceivable to anticipate the execution of the understudy (yield) in view of the different logical (input) factors which are held in the model. The characterization demonstrate was manufactured utilizing a few unique calculations and every one of them utilizing distinctive order systems. The WEKA Explorer application is utilized at this stage.

The usage of the dataset is finished utilizing an information

mining apparatus WEKA. WEKA is open source programming that actualizes a substantial accumulation of machine inclining calculations and is generally utilized in information mining applications. WEKA remains for Waikato Environment for Knowledge Analysis.

A. Applying Training Data in WEKA Tool

WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA stands for Waikato Environment for Knowledge Analysis. From the above data, student.arff file is created, and then this file is loaded into WEKA explorer for processing.

- Choose "WEKA 3.7.x" from Programs. The first interface that appears looks like the one given below.
- Explorer: An environment for exploring data. It supports data preprocessing, attribute selection, learning and visualization
- Get to the WEKA Explorer environment and load the training file using the Preprocess mode.
- Get to the Classify mode (by clicking on the Classify tab) as shown below:
- Now, you can specify the test options (by checking the corresponding button):
- Use training set means that you use the training set (the file you loaded in Preprocess) for testing.

B. Comparative identification of efficient Classification algorithm

- Initially the datasets are filtered with attributes of higher rank for classification based on select attribute option.
- the datasets are thereby tested with various classification algorithms
- Classifiers simulated are :
  - Naïve bayes
  - Multilayer Perception
  - SMO
  - J48
  - REP Tree

The usage of the dataset is finished utilizing an information mining apparatus WEKA. WEKA is open source programming that actualizes a substantial accumulation of machine inclining calculations and is generally utilized in information mining applications. WEKA remains for Waikato Environment for Knowledge Analysis.

C. Applying Training Data in WEKA Tool

WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA stands for Waikato Environment for Knowledge Analysis. From the above data, student.arff file is created, and then this file is loaded into WEKA explorer for processing.

- Choose "WEKA 3.7.x" from Programs. The first interface that appears looks like the one given below.
- Explorer: An environment for exploring data. It supports data preprocessing, attribute selection, learning and visualization
- Get to the WEKA Explorer environment and load the training file using the Preprocess mode.
- Get to the Classify mode (by clicking on the Classify tab) as shown below:
- Now, you can specify the test options (by checking the corresponding button):
- Use training set means that you use the training set (the file you loaded in Preprocess) for testing.

D. Comparative identification of efficient Classification algorithm

- Initially the datasets are filtered with attributes of higher rank for classification based on select attribute option.
- the datasets are thereby tested with various classification algorithms
- Classifiers simulated are :
  - Naïve bayes
  - Multilayer Perception
  - SMO
  - J48
  - REP Tree

V. DESIGN OF MODEL

The planning of understudy show J48 calculation give a most extreme precision in arranging the occasions in a proficient way. The understudy display is made in Net Beans utilizing java coding. The J48 calculation recognized by similar investigation of order calculation is taken for outlining.

The student model is meant to look at models like
1. Students
2. Student detail record
3. Student outcome analysis

J48 may be a tree primarily based learning approach, supported Iterative dichotomous (ID3) algorithmic program. It uses divide-and-conquer algorithmic program to separate a root node into a set of 2 partitions until leaf node (target node) occur in tree.
Given a collection T of total instances the subsequent steps square measure wont to construct the tree structure.

Step 1: If all the instances in T belong to constant cluster category or T has fewer instances, than the tree is leaf tagged with the foremost frequent category in T. Step 2: If step one doesn't occur then choose a check supported one attribute with a minimum of 2 or bigger potential outcomes. Then take into account this check as a root node of the tree with one branch of every outcome of the check, partition T into corresponding T1, T2, T3 ....., in line with the result for every various cases, and also the same is also applied in algorithmic thanks to every sub node. Step 3: data gain and default gain magnitude relation square measure graded victimization 2 heuristic criteria by algorithmic program J48.

The design of type includes a login type as shown in Figure five representational process username and positive identification for access. Student details square measure gathered victimization the shape as in Figure six.students performance square measure analyses as per the prediction with suggestions as in Figure seven. the end result of student's square measure analyses as if Best, Good, Average or Poor from the scholar model generated and thereby providing suggestions for upliftment of scholars performance.

A. PERFORMANCE ANALYSIS

By the simulation of data set with the various classifiers the accuracy of correctly classified instances is as below in Table 2.

Classification Algorithm	Accuracy (In %)
Naive Bayes	85.92
MultilayerPerception	94.94
SMO	94.34
Decision table	96.10
J48	97.27
REP Tree	95.33

As J48 algorithm classifies the instance with maximum accuracy ,it is used in designing the student model to predict students performance by analyzing training data and test data .thereby predicting students performance as Best, Good, Average or Poor.

VI. RESULTS AND DISCUSSION

Education Data mining fundamental centre is to examine the training framework. The model spotlights on dissecting the expectation exactness of the understudy's execution .The dataset that contains of all scholarly and individual elements of the understudies. This model can be helpful in the instructive framework like Universities and Colleges. By this model we can know the scholarly status of the understudies ahead of time and can focus on understudies to enhance their scholastic outcomes and arrangements. In this way enhance their guidelines and proprieties. Subsequently the nature of training can be progressed. The consequences of the information digging calculations for the characterization of the understudies in light of the properties chose uncovers that the forecast rates are not uniform among the calculations. The scope of expectation differs from (80-98%).Thereby by near examination of order calculations, (for example, Naïve bayes, MLP ,SMO ,Decision Table, REP tree, J48) utilizing WEKA device, it is demonstrated that the traits looked over the first dataset have high impact utilizing J48 with an exactness of 97% under investigation and utilized for foreseeing test informational collection for future result as best, good, average or poor.

REFERENCES

[1] Crist´obal Romero, *Member, IEEE, and Sebasti´an Ventura, Senior Member, IEEE,* "Educational Data Mining: A Review of the State of *the* Art " VOL. 40, NO. 6, NOVEMBER 2010.

[2] Parneet Kaura, Manpreet Singhb ,Gurpreet Singh Josanc "Classification and Prediction based DataMining Algorithms to Predict Slow Learners in Education Sector" Science Direct Procedia Computer Science 57 (2015) 500 – 508 2015 (ICRTC-2015).

[3] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri, "DataMining Models for Student Careers", Science Direct - Expert Systems with Applications 42 (2015) 5508–5521.

[4] Harwatia, Ardita Permata Alfiania, Febriana Ayu Wulandari," Mapping Student's Performance Based on Data Mining Approach", Science Direct Agriculture and Agricultural Science Procedia3 (2015) 173 – 177.

[5] Manolis Chalaris, Stefanos Gritzalis, Manolis Maragoudakis, Cleo Sgouropoulou and Anastasios Tsolakidis," Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques", Science Direct - Procedia - Social and Behavioral Sciences 147 (2014) 390 – 397.

[6] V.Ramesh Assistant Professor Department of CSA, SCSVMV University Kanchipuram India "Predicting Student Performance: A Statistical and Data Mining Approach", International Journal of Computer Applications (0975 – 8887) Volume 63– No.8, February 2013 35.

[7] Krpan, Slavomir Stankov, "Educational Data Mining for Grouping Students in E-learning System", Proceedings of the ITI 2012 34th Int. Conf. on Information Technology Interfaces, June 25 – 28, 2012, Cavtat, Croatia.

[8] Dursun Delen, "A Comparative Analysis of Machine Learning Techniques for Student Retention Management", Science Direct - Decision Support Systems 49 (2010) 498–506.

[9] Ali Buldua, Kerem Üçgüna, "DataMining Application on Students' Data", Science Direct - Procedia Social and Behavioral Sciences 2 (2010) 5251–5259.

[10] Jiawei Han , Micheline Kamber , Jian Pei ,"Data Mining: Concepts and Techniques", Third Edition (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition.

[11] http://www.intechopen.com/books/theory-and-applications-for-

advanced-text-mining

[12] Nurbiha A Shukora , Zaidatun Tasira, Henny Vander Meijden, "An Examination of Online Learning Effectiveness using DataMining",Science Direct - Procedia - Social and Behavioral Sciences 172(2015) 555 –562.

Authors Profile

Ms. Mabel Christina Pursued her Bachelor of Engineering from Visvesveraya Technological University in 2010 and Master of Engineering from Visvesveraya Technological University in year 2016. She is currently working as Assistant Professor in Department of Information Sciences, Visvesveraya Technological University since 2018. She has published a research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and it's also available online. His main research work focuses on Data Mining Big Data Analytics, IoT and Computational Intelligence based education. He has 6 months of teaching experience and 6 months research Experience.

