A COMPREHENSIVE ANALYSIS ON EDUCATIONAL DATA MINING FOR STUDENTS

MADDU NAGA SAI PRASAD SANTHOSH, Mr. MD. SHAKEEL AHMED

1MCA STUDENT, DEPARTMENT OF MCA, VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, NAMBUR, GUNTUR, ANDHRA PRADESH.

2 Associate Professor, DEPARTMENT OF MCA, VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, NAMBUR, GUNTUR, ANDHRA PRADESH.

ABSTRACT

Since student performance and pass rates in school reflect teaching level of the school and even all instruction systems, it is basic to improve student pass rates and lessen dropout rates. Decision Tree (DT) calculation and Support Vector Machine (SVM) calculation in data mining, have been utilized by analysts to discover significant student includes and foresee the student pass rates, be that as it may, they didn't think about the coefficient of introduction, and whether there is a reliance between student highlights. Consequently, right now, propose another idea: highlights conditions, and utilize the network search calculation to advance DT and SVM, so as to improve the exactness of the calculation. Besides, we added 10-overlay cross-approval to DT and SVM algorithms. The outcomes show the test can accomplish better outcomes right now. The reason for this examination is to give help to students who have more prominent challenges in their investigations, and students who are in danger of graduating through data mining techniques.

Keywords: Data mining, Higher education, Clustering, Decision tree, neural network, classification, prediction, association rule analysis.

I. INTRODUCTION

In a country's life education plays a vital role to ensure the survival of the state and the nation[1]. In today's scenario educational technologies aide the process of learning and teaching (TL) as they are being used in educational domains including the traditional form of classrooms where it’s all about face to face and even the learning platforms available online. Educational actors (students, teachers and administrators) have been benefitted as they are provided with the relevant information in which they have to act upon and thereby end up in promoting the quality based innovations in this domain[2]. These days universities are run in a very powerful and dynamically viable manner. A large amount of data is gathered in the form of marks, records,
documents, files, performance et cetetra all related to student performa[3]. Educational data mining act as a bridge between the two one is the education and the other is computer science. The subfields of computer science, Data mining and Machine learning are used. Data mining is used to uncover the hidden patterns in the unstructured data. It is devoted to discover the knowledge and then generate the relevant information. Due to the data mining advancements it has become possible to mine educational data to get the useful data. This relevant information serves to benefit its handlers[4]. The rapid growth in educational domain brought the fact of distilling the massive amounts data which let to the emergence of educational data mining(EDM)[5]. Educational data mining is a specific field of data mining that helps in discovering invisible patterns in the data to take decisions for students, teachers and administrators. EDM make predictions which are further characterize the learner behavior, domain content knowledge, educational functionalities, assessment outcomes, applications and the applications. The outcomes can guide the students in learning process, tutors in teaching process to enhance the educational practices and the administrators in managing process[6].

II. Related work

The advancements in the EDM are progressively evolving the data mining techniques to inform the educational domains. Ultimately, the objective is to gather relevant information about the learning and teaching for the pedagogical enhancements. Accordingly, EDM has been divided into four phases. The first phase, discovers the relationships between data using data mining techniques such as classification, clustering, regression, sequential pattern mining and association rule mining. The second phase, validates whether the selected relationships are theoretically validated or not. If the relations are validate further processing is done otherwise not. The third phase, predictions are made according to the relationships that are theoretically validated for the future aspects in educational learning and teaching context. The fourth phase, supports the predictions made in third phase and make policy level decisions which would help in pedagogical improvement in the educational era[6].

Approaches in EDM Prediction

The aim is to infer the single label of the data that has been collected, by adjoining the other aspects of the data collection. The name itself indicates, it will predict the future aspects on the basis of the past data. In educational domain it can be used to predict the student’s behavior and their educational success[7]. Classification Classification is a supervised learning technique. It
maps the different unknown samples into the given samples[8]. This technique is only used when the predicted variable is none other than categorical. It is employed on to the predefined classes in test situations, training and mapping data. The decision tree, K-nearest neighbor, support vector machine and neural network algorithms are classified under the classification technique[7]. Regression This technique is used only when the predicted variable is numerical[8]. The regression technique is a technique to predict the relationship among dependent variables and independent variable. The independent variable is already known and the response variables are to be predicted. Regression analysis is further divided into linear regression, non linear regression and multiple regression[8]. Clustering This technique groups the similar data together into clusters. In the same clusters each object is very similar and as compared to other clusters they are less similar. Clustering and classification differ because clustering does not follow a specific criteria, instead they find the logic between the data collected. Beforehand no one knows in how many groups or in how many different kind of groups the data is to be divided in clustering. Clustering is used for preprocessing the data[8]. It is used to map the preferencesess of the student to explore the different types of educational contents and to look for the different learning patterns. In educational data mining clustering is used for figuring out the typical behavior of the students that can be grouped together in different groups. The methods that are classified under the clustering technique are K-means, Hierarchical clustering et cetera[9]. Relationship Mining: It is done to find out the relationships among the variables and to further use them as rules. We try to identify those variables are strongly connected to the other variable of the same kind. In educational data mining the relationship mining helps in discovering the student’s behavioural patterns and difficulties they face during the learning and even the mistakes are detected in the learning process that would help the teachers[7]. Association rule mining This technique is based on the “if then” rules. Here the relationships among the different variables is done as “If the student is willing to improve his or her performance level he or she should get the usage of the available help”[8]. Sequential pattern mining Sequential pattern mining is the technique that helps in discovering the temporary associations among the different labels or events. In educational data mining it can be used to infer the student’s request regarding the aide they require over time to improve their educational status[7]. Correlation mining In correlation mining we identify the linear correlation between the different variables as being positive or negative. This provides aide to EDM by making the
relationships among the students’s attitude for a particular activity as if that is positive it means the students tried to finish it and if negative it means the student did not attempt the activity[7]. Casual data mining Casual data mining is used to discover the relationship causes among the different variables. It identifies if the particular event is caused/originated by another. Here the predictions are made about the factors affecting the students’s performance related to the activity they were given[7]. Discovery within models The aim is to identify the structure according to the relationships without the predefinition of the idea to the fact that what is to be found. Discovery within models opposes the predicting methods as it does not allow the access to the previous correlations between the variables. In educational data mining it is used to find the factors that are responsible for the student’s competencies development[7]. Distillation of data for human judgement It provide the aide for understanding the analysis of results. It makes the complex data easily more understandable by humans to leverage their judgement[9].

III. LITERATURE SURVEY

Alom, et al., (2018) [6] trackes the students of Australia beginning from their primary school year 1 to completion of their high school successfully and further more it too tracks their admissions into the higher educational universities or institutions. The criteria on which they are calculating the successive rates of the students is the gender of the students. In this paper data mining software Wilson calculator is used which is a practical meta-analysis effect size calculator and Orange so as to analyze data. Orange for given data sets provide the predictive modeling and the visualizations solutions.

Tavares, R., et al., (2017) [7] The prime focus of this paper is to enhance the digital educational resources in primary school for science education by implementing on the data mining techniques. There is a lot of impact on the students for the self regulated learning after adopting the learning approach. The analyses of the students behaviour is done after getting the particular help and the recommendations. Zhang, W., et al., (2018) [8] This paper provides the guidance over the techniques of data mining and the related applications in the online educational scenario. The knowledge extraction is done with the help of educational data mining from raw data.

Asif, R., et al., (2017) [9] Studies the performance related to the education of the students. The data taken of the students is focused on two aspects. Firstly, achievement of the student is predicted at the completion of the four year study programme. Then predictions are combined
with the progressions. The outcomes is the generation of two groups of students, low and high grade achieving students. By this teachers get to support the students at the low level by giving them activities and the task and to the high level students more opportunities are given.

Bakhshinategh, B., et al., (2018) [10] In this paper different types of existing Educational data mining tasks and applications have been listed with their categorization on the basis of their purpose. A comparative study is done over the existing surveys related to educational data mining and all the task are reported in a taxonomy.

Manjula, M. (2018) [11] Implements K-means clustering algorithm and uses the weka tool for analyzing the educational data set. With weka the accuracy level improves and the results obtained are the perfect graphs. The data taken has preprocessed to clear the null values, remove the unwanted memory space and the unwanted data. Data mining two methods have been used such as classification and clustering.

IV. PROPOSED SYSTEM:

Right now, utilize the arrangement algorithmin data mining innovation, otherwise called directed learning DT calculation and SVM calculation. Especially, in light of the fact that the grid search calculation is a technique to upgrade the model performance by navigating a given blend of parameters, the calculation can be utilized to enhance the DT and SVM algorithms to improve the exactness of the calculation. Moreover, in light of the fact that distinctive student highlights have diverse effects on student pass rates and there is some reliance between the different highlights, we present the highlights dependences and master direction.

Preprocessing Student Data:

In reality, there is a lot of conceivably important data in the field of instruction. So the data must be institutionalized and preprocessed before we use data mining to burrow data. The datasets utilized right now be downloaded from the machine learning site. Right now, essentially expand on the best way to manage various kinds of data, for example, numeric, ostensible, parallel and so on., mostly managing ostensible sorts.

The ostensible sorts of data were changed into a 1-of-C encoding and all highlights were institutionalized to a zero mean and one standard deviation. This implies as opposed to normalizing somewhere in the range of 0 and 1 every classification is given its own opening. For instance, there are five sorts of work, instructors, specialists, government employees, (for example, authoritative or police), at home or other. On the off chance that the mother is an
instructor, you can utilize the parallel exhibit as [1, 0, 0, 0, 0], if the mother's work is a specialist, you can utilize the double cluster as [0, 1, 0, 0, 0]. Contrasted and the ostensible kind, we will legitimately utilize the numerical data and paired sort of data converted into 1-0.

Decision Tree technique:
In the field of training, the student pass rates speak to the teaching nature of the school to a limited degree and even the teaching level of the entire instructive foundation. Diverse student highlights affect student pass rates in various degrees, and it is critical to discover key highlights that influence student pass rates. The record used to quantify the level of polluting influence is Entropy. So the data entropy in the DT calculation is regularly used to ascertain the heaviness of student highlights. DT is a branch structure made out of rules. Leaf hubs speak to the consequences of characterization; the root hub speaks to the key highlights. The way toward developing a branch structure in a DT is called recursive apportioning. This fanning structure can be communicated as though THENform, effectively comprehended by a human. Since the data set contains highlights qualities and label traits, we can utilize classes to recognize.

**Support Vector Machine strategy:**
For this investigation, contrasted and different algorithms, support vector machines are profitable for paired grouping issues, so we utilize the SVM calculation to ascertain the student pass rates and the precision of the model. SVM is a nonlinear capacity of regulated learning. The greatest
bit of leeway of the SVM calculation is that when taking care of the issue, it is unimportant to the component of the example, and the little example can in any case keep up a solid versatility to the new example.

Figure 3: Support Vector Machine

PROPOSED ARCHITECTURE

Fig. 4. System Model.
In the training, data mining innovation is for the most part utilized for student data examination, assistant instructors and supervisors to make right decisions. Data mining incorporates numerous algorithms, such as DT, affiliation rules, SVM, bunching examination, etc. Notwithstanding, contrasted and other data mining algorithms, the DT is simpler to be comprehended and can be built a decision tree for data sets with numerous characteristics, and SVM can solve nonlinear issues and take care of high-dimensional issues. Online Education Teacher Evaluation Based on Big Data Analysis. Specifically, the general calculation flowchart is appeared in Figure 4.

CONCLUSION

In the training field, a considerable lot of the past specialists have demonstrated incredible enthusiasm for student pass rates or in any event, finding significant student highlights. Nonetheless, they frequently disregard the connections that exist between student highlights. Not the same as the past examination, right now, consider the reliance between student highlights, presenting highlights conditions and master direction to instate coefficient with the goal that the calculation can be combined quicker. Specifically, so as to additionally decide the underlying estimations of the coefficients, we propose the instatement coefficients rules. Besides, so as to improve the exactness of the model, we utilize the SVM calculation and the DT calculation upgraded by the matrix search calculation to foresee the student pass rates and utilize the data increase to discover the highlights that impact the student performance. Examinations show that the calculation has accomplished great outcomes, so we can utilize the model to distinguish the significant highlights of students and student pass rates.
REFERENCES


