A REVIEW ON PERFORMANCE OF STUDENT'S PREDICTION USING DATA MINING TECHNIQUES

Mr. Rahul Sharma
Research Scholar
Computer Science and Engineering
Rabindranath Tagore University, Bhopal

Dr. Shiv Shakti Shrivastava
Professor
Computer Science and Engineering
Rabindranath Tagore University, Bhopal

Abstract

The success of an academic institution can be measured in terms of the quality of education provides to its students. In the education system, the highest level of quality is achieved by exploring the data relating to redirection about students' performance. These days the lack of an existing system to analyze and judge the student's performance and progress is not being addressed. There are two reasons why this is often happening. First, the present system isn't accurate to predict student's performance. Second, because of shortage of consideration of some vital factors that are affecting student's performance. Predicting student's performance is a more challenging task as a result of a large amount of information in the academic database. This proposed system can help to predict student's performance more accurately. For these suitable data mining, approaches will be applied. In this approach, the pre-processing step will be applied to the raw data set so that the mining algorithm will be applied properly. The prediction about a student's performance can help him/her to enhance the performance.

Key Words: *Education, student, performance, data mining, pre-processing, database, prediction.*

1. INTRODUCTION

Improving student's academic performance isn't an easy task for the academic community of higher learning. The academic performances of the students during their first year at university is a turning point in their educational path, usually encroaches on their General Point Average (GPA) in a decisive manner. The student's evaluation factors like class quizzes mid and final examination assignment lab -work is studied. It is recommended that all this correlated information should be conveyed to the class teacher before the conduction of the final exam. This study will help the teacher to reduce the drop-out ratio to a significant level and improve the performance of students. This paper, we present hybrid procedures based on the Decision Tree of Data mining methods and Data Clustering enables academicians to predict student's GPAs (SGPA & CGPA), and based on that instructors can take the necessary steps to improve student academic performance.

Graded Point Average (GPA) is a commonly used indicator of academic performance. Many universities set a minimum GPA that should be maintained. Therefore, GPA still remains the most common factor used by academic planners to evaluate progression in an academic environment. Many factors could act as barriers to students attaining and maintaining a high GPA that reflects their overall academic performance, during their tenure in university. These factors could be targeted by the faculty members in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance. With the help of clustering algorithm and decision tree of data mining technique, it is possible to discover the key characteristics for future prediction. Data clustering is a process of extracting previously unknown, valid, positional useful, and hidden patterns from large data sets.

The amount of data stored in educational databases is increasing rapidly. The clustering techniques are the most widely used technique for future prediction. The main goals of clustering are to partition students into homogeneous groups according to their characteristics and abilities. These applications can help instructors and students to enhance education quality. This study makes use of clusters analysis to segment students into groups according to their characteristics. Decision tree analysis is a popular data mining technique that can be used to explain different variables like attendance grade ratio and ratio. Clustering is the technique often used

in analyzing data sets. This study makes use of clusters analysis to segment students into groups according to their characteristics, uses a decision tree for making a meaningful decision for the student.

2. PREVIOUS WORK

Data mining (sometimes known as knowledge or information discovery) is the method of analyzing information from totally different views and summarizing it into useful information. Information that may be used to increase revenue, cuts costs, or both data mining software system is one of the varieties of analytical tools for analyzing information. It permits users to analyze the information identity. Technically, data mining/data processing is the process of finding correlations or patterns among dozens of fields in massive relational databases. Following are the survey papers being studied:

- Paris et. al.(1), compared the data mining methods accuracy to classifying students in order to predict the category grade of a student. These predictions are more helpful for identifying the weak students and helping the administration to take remedial measures at the initial stages to produce excellent graduates which will graduate at least with the upper category [1].
- Rathee and Mathur applied ID3, C4.5, and CART decision tree algorithms on educational information for predicting student performance in the examination. All the algorithms are applied to the internal assessment information of students to predict their academic performance in the final examination. The efficiency of various decision tree algorithms will be analyzed based on their accuracy and the time was taken to derive the tree. The prediction obtained from the system has helped the class teacher to identify the weak students and improve their performance. C4.5 is the best algorithm among all three because it provides higher accuracy and efficiency than the other algorithms [3].
- Kortemeyer and Punch applied data mining classifiers as a means of comparing and analyzing students' use and performance who have taken a technical course via the web. The results show that combining multiple classifiers leads to a significant accuracy improvement in a given data set. The prediction performance of combining classifiers is often better than a single classifier because the decision is relying on the combined output of several models [3].

3. STEPS OF DATA MINING

Data mining is the method of discovered numerous models, derived values, and summaries from a given collection of information. The problem of discovering or estimating dependencies from the information or discovering new information must be simply one part of the overall experimental procedure utilized by engineers, scientists, and others who apply standard steps to conclude from the information. The overall method of finding and de-coding patterns and models from information involves the recurrent application of the subsequent steps [6]:

1. Understand the application domain, the relevant previous knowledge, and the goal of the end-user (formulate the hypothesis).
2. **Data Collection:** Determining how to find and extracts the right data for modeling. First, we need to identify the different data sources that are available. Data may be scattered in different spreadsheets, files, and hard-copy lists.
3. **Data integration:** Integration of multiple data cubes, files, databases. A big part of the integration activity to build a data map, which expresses how each data element in each data set must be prepared to express it in a common format and record structure.
4. **Data selection:** First of all the data are collected and integrated from all the various sources, we select only the data which useful for data mining. Only relevant information is selected.
5. **Pre-processing:** The Major Task in Data Preprocessing are: Cleaning, Transformation and Reduction.
   - **Data cleaning:** Additionally known as data cleansing. It deals with errors detection and removal from information to improve the quality of information. Information cleaning sometimes includes fill in missing values, identify or remove outliers.
   - **Data Transformation:** Data transformation operations are additional procedures of data preprocessing that would contributes toward the success of the mining process and improve data-mining results. Some of the Data transformation techniques are Normalization, Differences, and ratios, and Smoothing.
   - **Data Reduction:** For large dataset, there's an increased probability that an intermediate, data reduction step should be performed before applying data mining techniques. While massive datasets have the potential for higher mining results, there's no guarantee that they'll produce better knowledge than small datasets. Data Reduction obtains a reduced datasets representation that's much smaller in volume, however, produces constant analytical results.

6. **Building the model:** In this step we select and implement the appropriate data mining task (ex. association rules, serial pattern discovery, classification, clustering, and regression, etc.), the data mining technique, and also the data processing algorithms to create the model.
7. **Interpretation of the discovered knowledge (model /pattern):** The interpretation n of the detected pattern or model reveals whether or not the patterns are interesting. This step is additionally known as Model Validation or Verification and uses it to represent the result in an appropriate approach so it may be examined completely.
8. **Decisions / Use of Discovered Knowledge:** It helps to make use of the knowledge gained to make better decisions [7].

4. THE PROPOSED APPROACH

The aim of this project is to improve the current trend in the higher education systems and to find out which factors might help in creating successful students. It is really necessary to find the successful student's as it motivates higher education systems to know them well and one way to know this is by using valid management and processing of the student's database.

- **Classification**

  A classification algorithm is a data mining technique that helps us to map data into predefined categories. It is a supervised learning technique that needs categorized training data so it can be creating rules for categorizing test data into a pre-arranged category. [2] Its a 2 phase process. The first phase is the learning phase, where the classification rules are generated and training data is analyzed. The second phase is the classification phase, where test data is classified into predefined groups according to the generated rules. Since classification algorithms require predefined classes based on values of the information component, we had created a component "performance" for all students, for which they may have a value of either "Good" or "Bad".

4. TOOLS AND METHODOLOGY

A. Models

We are using four kinds of classification model so as to learn the predictive function which is required. The models are used for experimental analysis. They are selected on the basis of their frequent usage in the existing works of literature. The lists of methods are as follows:-

1) Decision Tree

A decision tree is a tree in which each branch node will represent a choice between several alternatives and each leaf node will represents a decision. A decision tree is commonly used for obtaining information so as to fulfill the purpose of decision making. The decision tree starts from a root node which is there for users to take action. From root node users split each and every node recursively into different nodes according to the decision tree learning algorithm. The final result is a decision tree where each branch represents a possible context of the decision and its outcomes.

2) Naive Bayes

The Naive Bayes algorithm is actually based on the probability theory, i.e. the Bayesian theorem [3], and is a simple classification method. It is named naive because it solves problems based on two critical assumptions: it assumes that there are zero hidden components that will affect the process of analyzing and it supposes that the prognostic component is conditionally independent with similar classification. This classifier provides an efficient algorithm for data classification and it represents the promising approaches to the discovery of knowledge.

3) Support Vector Machine

Support Vector Machine is used for classification which is also a supervised learning method. There are three research papers that have used the Support Vector Machine algorithm as their technique to analyze student's performance to review it thoroughly. Hamalainen et al. (2006) had chosen Support Vector Machine as their analyzing method because it suited well in small datasets. [4] Sembiring et al. (2011) demonstrates that the Support Vector Machine algorithm has a good ability to perform generalization and is actually found faster than other algorithms. [5] At the same time, the study done by Gray et al (2014) explained that the Support Vector Machine algorithm acquires the highest analyzing accuracy in identifying student's performance (Failing Risk). [6]

4) K-Nearest Neighbors

K-Nearest Neighbor is one of the simple Machine Learning algorithms based on the Supervised Learning technique. It is also called a lazy learner algorithm because it doesn't learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. K-NN algorithm stores all the available data and classifies a new data points based on the similarity. This means when new data appear then it can be easily classified into a good suite category by using K- NN algorithm.

B. Data Description

- Data source link: http://archive.ics.uci.edu/ml/datasets/ student performance
- Data format: Integer
- Size: 396 rows X 33 columns
- Number of Instances: 396
- Number of Attributes: 33

This data is of students' achievement in secondary education at a Portuguese school. The data attributes include student grade, demographic, social, and school-related features) and it was collected by using questionnaires and school reports. Dataset is provided regarding the performance in the subject: Mathematics. The target attribute G3 has a strong correlation with attributes G2, G1. This occurs because G3 is the final year grade, while G1, G2 correspond to the 1st and 2nd-period grades. During the data pre-processing set we found out that the data present in our dataset was clean, as a result, we did not have to perform the data cleaning methods. In our dataset, we had 33 attributes and as result we had to reduce some of the attributes which were not so important, to get better accuracy and a low-cost tree. In organizations these kinds of strategies are performed to reduce the data, so we also decided to do the same.

Fig – 1 Correlation Heatmap

Models	Decision Tree	KNN	Naïve Bayesian	SVM
Accuracy of test dataset	0.79	0.72	0.64	0.70
Error rate	0.21	0.27	0.35	0.29
Sensitivity	0.78	0.72	0.64	0.70
Specificity	0.21	0.27	0.35	0.29

Fig – 2 Comparison Table

Environment

We run the experiments on the 4 GB RAM PC, with 1.90 GHz of Intel i5 Processor. In evaluating these models, we used python Programming. We split the data into two parts, train data set containing 70% of the data and a test data set containing the remaining 30%.

5. CONCLUSION

In this paper, an effort is made to find the impact of our proposed features on student performances prediction with the help of classification models. A feature space is constructed by considered characteristics of family expenditure, family income, personal information, and family assets of students. The potential or dominant features selection is unavoidable as it provides us with a subset of features. By using the Decision Tree classification algorithms we found our analysis very effective for our proposed features of family expenditure and student personal information categories. It can be easily derived from the results we got that academic information, family details, and personal information have a very strong impact on the students' performance due to instinctive reasons provided in discussions. The meta-analysis on analyzing student's performance has encouraged us to carry out a further examination to be applied in our educational institutes. Hence, the educational system can take the help of this model to review the student's performance in a suitable manner.

6. REFERENCES

[1] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms", International Journal of Computer Science and Management Research, vol. 1, 2012.

[2] K. V. J.K. Jothi Kalpana, "Intellectual performance analysis of students by using data mining techniques", International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, 2014.

[3] V. Ramesh, "Predicting student performance: A statistical and data mining approach", International Journal of Computer Applications, vol. 63, no. 8, 2013.

[4] D. A. M. Dr. Abdullah AL-Malaise and M. Alkhozae, "Students performance prediction system using multi agent data mining technique", International Journal of Data Mining and Knowledge Management Process, vol. 4, no. 5, 2014.

[5] P. Kavipriya, "A Review on Predicting Students" Academic Performance Earlier, Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 12, December 2016.

[6] Shruthi P, Chaitra B P, "Student Performance Prediction in Education Sector Using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.

[7] Humera Shaziya, et.al. "Prediction of Students Performance in Semester Exams using a Naïve bayes Classifier", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 10, October 2015.

[8] Y. K. Salal, S. M. Abdullaev, Mukesh Kumar (April 2019) Educational Data Mining: Student Performance Prediction in Academic, ISSN: 2249-8958, Volume-8 Issue-4C edn, International Journal of Engineering and Advanced Technology (April 2019).

[9] Anoopkumar M , A. M. J. Md. Zubair Rahman , Model of Tuned J48 Classification and Analysis of Performance Prediction in Educational Data Mining, Volume 13 edn., : International Journal of Applied Engineering Research (2018).

[10] Eyman Alyahyan and Dilek Düştegör, Predicting Academic Success in Higher Education: Literature Review and Best Practices, International Journal of Educational Technology in Higher Education, Springer Open (2020).

[11] Juan L. Rastrollo-Guerrero, Juan A. Gómez-Pulido and Arturo Durán-Domínguez Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review, Applied Science (2020).