ZKG INTERNATIONAL

ISSN: 2366-1313

PREDICTING STUDENT PERFORMANCE WITH DATA MINING

#1THUPAKULA SRIKANTH,
#2POLE SRIJA,
#3K.CHANDRASENA CHARY, *Associate Professor,*
Department of Computer Science and Engineering,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

ABSTRACT: The quality of instruction provided to students is what determines an educational school's performance. Real-world study on redirection and student performance can aid in determining the optimal level of quality for the educational system. Right now, we're not discussing how there is no mechanism in place to measure and study students' development and achievement. There are two reasons why this occurs frequently. To begin, the existing system is not particularly good at predicting how well children would perform in school. Furthermore, the students' academic achievement is jeopardized since critical aspects are being overlooked. There is a lot of data in the school database, making it difficult to predict how well students will do. There is a probability that the proposed method will improve accuracy in predicting how well youngsters will perform in school. We'll employ the proper methods to collect information for these. A step will be made to prepare the raw data so that the mining program can function properly in this procedure. The prediction of how well the student will do may help them perform better.

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

1. INTRODUCTION

Higher education institutions are confronted with a challenging endeavor when it concerns aiding students in enhancing their academic performance. Typically, the initial year of college represents a turning point in the academic careers of science and engineering majors, exerting a substantial influence on their grade point average. The criteria for assessments that students are required to complete, such as midterm and final exams, homework assignments, laboratory work, and quizzes, are examined.

It is imperative that the class instructor be apprised of all of this information before the final examination. By reducing the number of students who drop out and increasing student success, this endeavor will benefit educators. A composite method integrating data clustering and the Decision Tree Data Mining technique is introduced in this article. This approach enables educators to forecast the GPAs (SGPA, CGPA) of their pupils and assist them in improving their academic performance. The grade point average (GPA) is frequently employed as a metric for assessing academic performance. A minimum GPA requirement is imposed by numerous colleges.

Maintaining that is necessary. Conversely, academic administrators continue to rely primarily on the GPA as a metric for evaluating student progress. A multitude of Multiple obstacles may impede a college student's ability to attain and sustain a satisfactory grade point average. The cumulative academic performance of a student is reflected in their GPA. Teachers can develop instructional strategies to support students' academic progress and achievement by monitoring the evolution of their students' grades. By utilizing the decision tree and clustering method of data mining, it is possible to ascertain the most critical attributes that are indispensable for future predictions. Clustering data is a method for extracting authentic, positionally valuable patterns from extremely massive data sets that have never been observed before.

ISSN: 2366-1313

New information is continually being accumulated in academic libraries. Clustering is the prevailing method utilized to forecast future events. Students are categorized into similar categories according to their characteristics and abilities. These applications can aid in the improvement of courses, which is advantageous for both instructors and learners. Cluster analysis is employed in this research endeavor to categorize students into distinct groups according to shared characteristics.

Utilizing data mining to comprehend variables such as the attendance ratio and the grade ratio is frequently accomplished via decision tree analysis. Clustering is among the rudimentary methods utilized to analyze data collections. Students in this research are categorized into groups according to the degree of similarity in particular attributes. A decision tree is then utilized to assist the students in determining what actions to take in the actual world.

2. BACKGROUND WORK

Information or knowledge finding is a subset of data mining. Information is arranged to make sense and reviewed from several perspectives. One instrument for information analysis is data mining software. Data is found to increase income, reduce expenses, or do both. That makes information identification clearer. Known by another name, data processing, data mining searches massive relational databases for patterns or connections in hundreds of places. Reviewing the survey articles listed below:

Paris and colleagues (1) evaluated data mining techniques for estimating category grades and classifying students. These projections assist in the identification of underachievers and facilitate prompt implementation of the required adjustments by the school to produce outstanding graduates who will at least place in the top two categories.

Rathee and Mathur predicted exam performance using ID3, C4.5, and CART decision tree algorithms together with educational data. Every algorithm predicts final exam performance using data from internal testing for each student. We shall contrast time complexity and accuracy of decision tree methods. Through the system's predictions, the teacher was able to determine which students required additional assistance and how to raise their performance. Out of the three algorithms, C4.5 runs the fastest and most precisely.

Using data mining, Kortemeyer and Punch compared and examined how students in online technical courses utilized and excelled in the course material. Accuracy of data is increased in several ways.

The choice is based on several models, hence combining classifiers frequently produces better predictions than using only one.

STEPS OF DATA MINING

Data mining is the procedure of extracting numerous models, derived values, and summaries from a given amount of data. Engineers, scientists, and other professions that use traditional processes to extract conclusions from data must see the challenge of identifying or estimating dependencies from information, or acquiring new information, as merely one component of a comprehensive experimental strategy. Generally speaking, the following steps are taken regularly in order to recognize and interpret models and patterns in data:

1Establish the application domain, the end-user goals, and relevant prior information (formulate the hypothesis).

2. Data Collection: Determine the data to use for modeling and how to extract it. We need initially to find out which data sources are available. Different data can be included in spreadsheets, files, and hard copy (paper) lists.

3. Data integration: joining of many files, databases, or data cubes. A key component of the integration process is creating a data map that outlines how every data element in every data set must be prepared for expression in a common format and record structure.

4. Data selection: Following its collection and combination from several sources, only relevant

ISSN: 2366-1313

data is chosen for data mining. We select just relevant information.

5. Pre-processing: Known as data at times.

Cleansing. It deals with the finding and removing of errors from information to improve its quality. Two instances of information cleaning include finding or removing outliers and filling in missing numbers.

Data Transformation: Operations are further procedures in data preprocessing that improve the results of data mining and support the mining process. Data converting techniques include normalization, differences and ratios, and smoothing.

Data Reduction: Large datasets often require an interim data reduction phase before using data mining methods. Better mining results may come from larger datasets, but this is not a guarantee. Data reduction keeps analytical results stable while reducing the quantity of datasets.

6. Building the model: In this stage, the appropriate data mining technique, data mining task (such association rules, serial pattern identification, classification, regression, and clustering), and data processing algorithm or algorithms are selected and implemented to build the model.

7. Interpretation of the discovered knowledge (model /patterns): It is evident from the interpretation of the recognized pattern or model whether the patterns are intriguing. Often called "Model Validation/Verification," the aim of this phase is to properly characterize the result so that it may be thoroughly examined.

8. Decisions / Use of Discovered Knowledge: It helps to make better decisions when one has the necessary information.

3. THE PROPOSED APPROACH

The goal of this project is to change the way higher education is done now and find out what factors might help students do well. Colleges and universities want to learn a lot about students who do well, so they reward those students. One way to do this is to help students learn how to manage and process knowledge well.

Classification

The classification algorithm is a type of data mining method that helps us organize data into groups that we already know about. This method of supervised learning needs training data that has already been sorted so that rules can be made for putting test data into groups that have already been set up. The process has two steps. In the first step, called the "learning phase," training data is looked at and rules for grouping are made. In the second step, called the classification step, the test data is put into groups that have already been set up according to the rules that have already been set up. Because classification algorithms need to be able to put things into groups based on the values of an information component, we gave each student a "performance" component that could be "Good" or "Bad."

4. TOOLS AND METHODOLOGY

A. Models

Four different kinds of classification models are used to teach the algorithm how to make good guesses. The models are used to look at how the tests turned out. Their choice was based on how often they show up in new works of writing. Here is a list of the techniques:

Decision Tree

There are several options in a decision tree, and each branch node represents a different choice. Each leaf branch will represent a choice. A decision tree is often used to help people gather knowledge that will help them make a choice. At the decision tree's root node, people can decide what to do. The decision tree learning technique gives us a way to split each node from the root node into its own separate nodes. The end result is a decision tree, where each branch shows a possible situation where the choice could be made and what would happen.

Naive Bayes

The Naive Bayes algorithm is a simple way to put things into groups. The Bayesian theory, which is

ISSN: 2366-1313

based on the idea of chance, makes it possible. It's called "naive" because it solves problems based on two main assumptions: that all variables are conditionally independent and have the same classification, and that there are no secret variables that could change the way it looks at things. As an algorithm, this classifier is useful because it sorts data into groups and offers a way to find new information.

Support Vector Machine

Support Vector Machine, a guided learning method, is used to group things together. The Support Vector Machine algorithm was used to look closely at how well students did in three different research projects. As a way to look at their data, Hamalainen et al. (2006) used Support Vector Machine.Because it did a great job with small samples. A study from 2011 by Sembiring et al. says that the Support Vector Machine method works better at generalization and gets results faster. According to Gray et al.'s (2014) study (Failing Risk), the Support Vector Machine method is the most accurate way to figure out how well a student is doing.

K-Nearest Neighbors

One more simple one is K-Nearest Neighbor.

Supervised learning is how machine learning systems are put together. The "lazy learner" algorithm gets its name from the fact that it can't learn right away from the training set. Instead, it saves the dataset and uses it for something when it's time to classify. The K-NN algorithm saves all the data that is available. It then groups new data points based on how similar they are to data points that have already been saved.

That is, the K-NN method makes it easy to quickly put new data into the right groups.

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

The performance of Portuguese secondary school students is represented by this data. Alongside grades, the dataset comprises social, demographic, and academic-related information. Surveys and school reports were utilized in its collection.

Datasets illustrating the mathematical performance of students are readily available. The objective attribute G3 is closely associated with the attributes G2 and G1. Because G1 and G2 represent the grades for the first and second periods, respectively, and G3 represents the final year, this is the case.

During the data pre-processing phase, we determined that the dataset contained clear data. This suggested that the implementation of data purification techniques was superfluous.

The dataset comprised a total of thirty-three attributes. A number of inconsequential characteristics had to be removed in order to produce a tree that was both more precise and cost-effective. These are the methods utilized by businesses to reduce data, so we chose to adopt them.

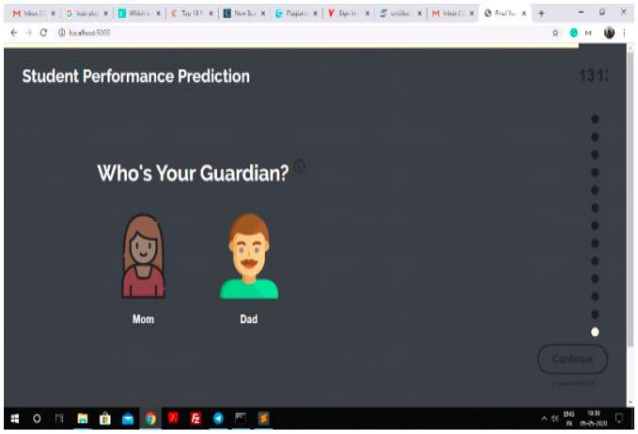
Fig Correlation Heatmap

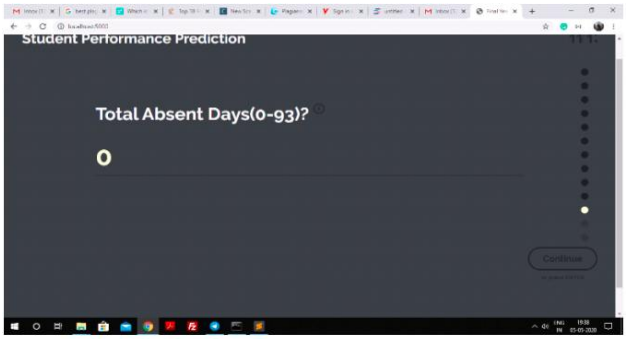
5. RESULTS

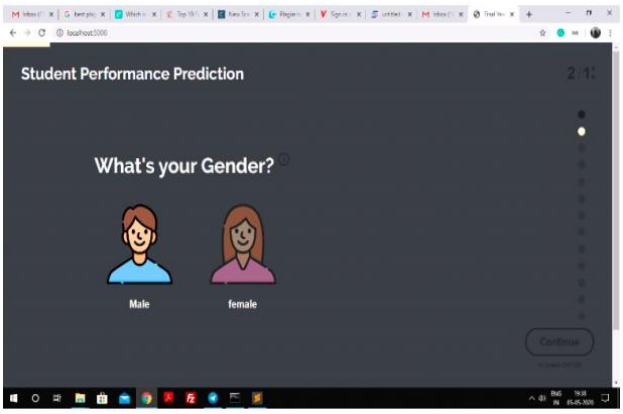
We finished our algorithmic implementation using Python. We built our categorization algorithms using Python modules and packages that came pre-installed. The following libraries and packages were used:

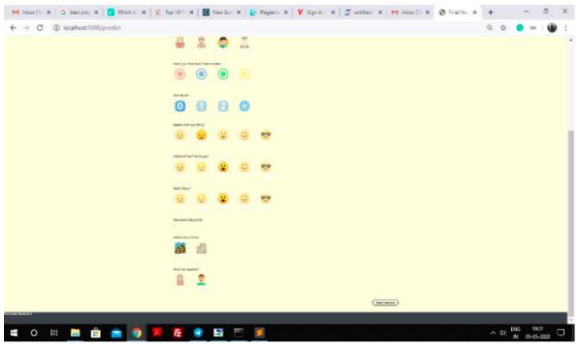
1. Numpy
2. Pandas
3. Scikit-learn
4. Matplotlib
5. Flask

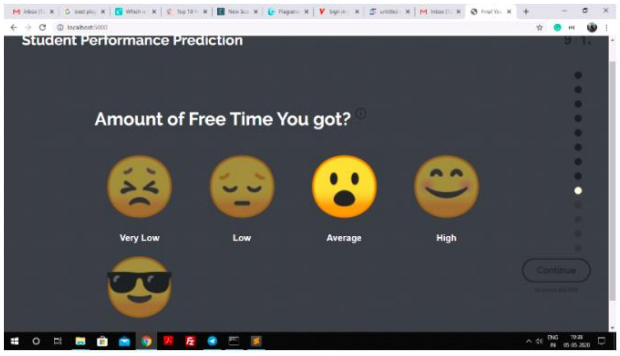
ISSN: 2366-1313

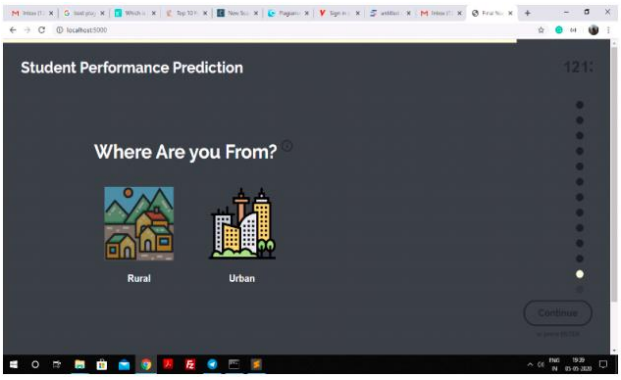


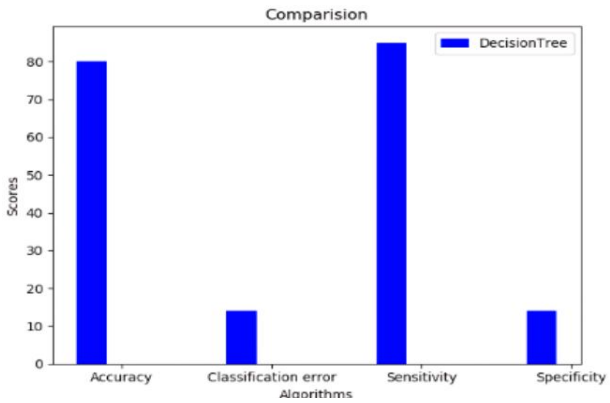






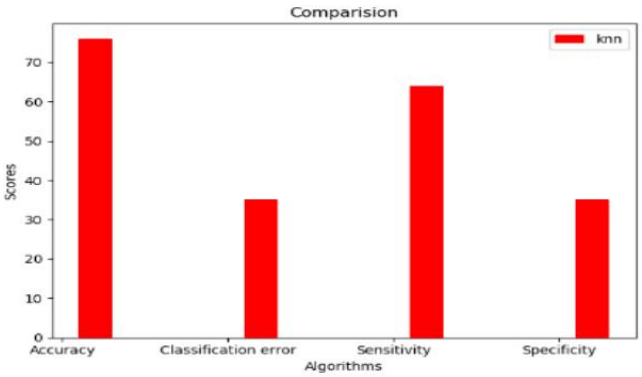




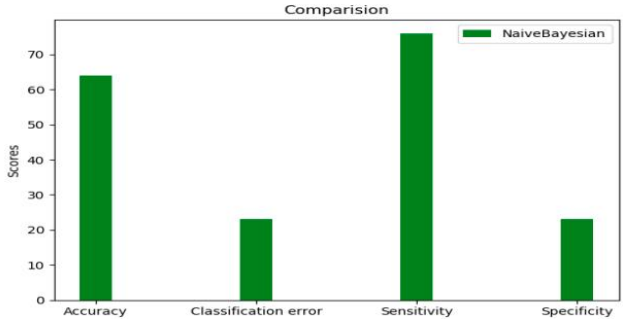


(a) Decision Tree

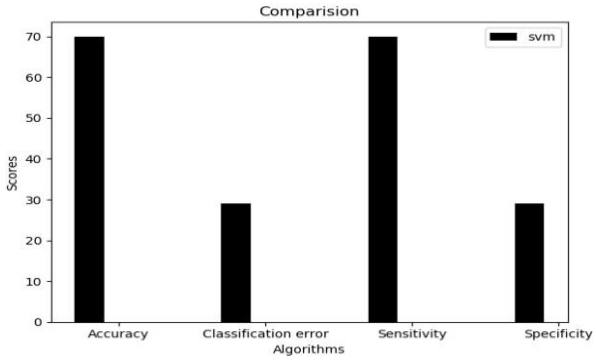
ISSN: 2366-1313



(b) K-Nearest Neighbors



(c) Naïve Bayesian



(d) Support Vector Machine

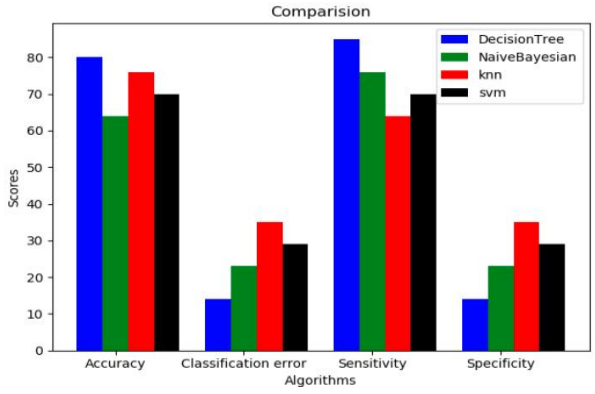


Fig Comparison of Models

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

As a result, all four algorithms were used efficiently, and their accuracy and performance were thoroughly documented. The analysis of the accuracy measures revealed that the Decision Tree strategy was the most appropriate for the provided dataset.

6. CONCLUSION

This study uses classification models to investigate how the proposed factors influence student performance prediction. A feature space for student families considers factors such as family income, spending, assets, and personal information. The inescapable process of selecting possible/dominant qualities leads to the creation of a subset of features. We discovered that by employing the Decision Tree classification technique, our examination of the proposed features of the categories of family expenditure and student personal information was quite effective. The findings of the conversations show that, for understandable reasons, academic, personal, and family information all have a significant and immediate impact on students' performance. The meta-analysis on student performance evaluation has inspired us to conduct additional research that will be useful in our educational institutions.

As a result, the educational system can use this paradigm to accurately measure students' performance.

REFERENCES

1. [K. V. J.K. Jothi Kalpana, "Intellectual performance analysis of students by using data mining techniques", International Journal of

ISSN: 2366-1313

Innovative Research in Science, Engineering and Technology, vol. 3, 2014.

2. V. Ramesh, “Predicting student performance: A statistical and data mining approach”, International Journal of Computer Applications, vol. 63, no. 8, 2013.
3. 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.
4. 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.
5. 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.
6. 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.