Student Performance AnalysisUsing Machine Learning

[1] Kotagiri Sanjana, [2]Bommidi Madhan Sainath Reddy, [3]T Sri Pranith Reddy, [4]Raheem Unnisa
B.Tech Student, Department of Computer Science and Engineering,CMR Technical Campus, Medchal, Hyderabad, Telangana, India
Assistant Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad Telangana, India

Abstract- Performance analysis of outcomes based on learning is a system that strives for excellence at different levels and diverse dimensions in the field of students' interests. This paper proposes a complete EDM framework in the form of a rule-based recommender system that is designed not only to analyze and predict the performance of students, but also to present the reasons behind it. Does the proposed framework analyze the students? To get all the necessary information about students, teachers, and parents, we collect demographic information, study-related characteristics, and psychological characteristics. Using powerful data mining techniques, to predict academic performance with the highest accuracy possible. The framework succeeds to highlight the student's weak points and provide appropriate recommendations. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones. Student Performance Analysis using Machine Learning.

INTRODUCTION

There are numerous enhancements and significant interests in using data mining in the educational sector. This ultramodern arising sector, called educational data mining, is concerned with better styles that prize knowledge from educational data. Data mining is a methodof sorting data that identifies patterns in huge databases. There are many different fields where these generalities and styles can be applied, including marketing, drugs, realestate, client relationships, engineering, and web mining. Educational data mining is one of the arising or advancedfashion of data mining that can be applied to data related to education. The data can be collected from history usingdata and functional data that live in educational institute databases. Scholars' data can be a variety of information or academic performance. Additionally, it can be accomplished by learning database systems, which have atremendous amount of data and information.

Several data mining generalities are implemented properly using it, such as Naive Bayes and K Nearest Neighbor. Using association rules, brackets, and clustering, different kinds of knowledge can be discovered. By utilizing this method, we prize knowledge that describes a scholar's performance in examination andall their detail information. First, we need to sort out these enormous amounts of data, then we use cluster analysis to classify the raw data. Clustering is a bunch of physical or abstract objects, as per the degree of similarity between them. It is divided into several groups, and makes the same data objects within a group of high similarity and different groups of data objects which aren't analogous. As we know in the moment's terrain, there's a lack of quality education,and also as well as the amount of competition is growing day by day. So there's a need for quality way to ameliorate the standard of the scholars and education as well.

For this several proponents give time to time suggestions and norms for performance enhancement. Still, the systems were not as good as they used to be. Consequently, the experimenters concluded that technology can be an important component of analyzing the excrences present in the current system. And also the use of technology makes the decision-making processeasy, as it can induce reports and graphs for analysis purpose. Education could be an important issue for achieving fiscal progress. The Students scholastic performance focuses on different aspects, creating analysis little bit delicate. In forthcoming times, there has been a rise within the chance .in rate of interest and concern overindividualists within the use of data mining for assaying academic rates. Data processing depicts growing and forthcoming

areas of inquiries in education and it has separate separate requirements that some fields Warrant. During this design, the performance analyses of scholar's are mentioned. The thing at of this design is furnishing scholars' performance using given strategies through different algorithms.

A lot of studies in this area examine how machine literacy can be applied to educational fields. It focuses on relating high-threat scholars and their performance. Ultramodern literacy institutions operate in a largely competitive and complex environment. Therefore, assaying performance, furnishing high-quality education, formulating strategies for assessing the scholars' performance, and relating unborn requirements are some challenges faced by the utmost universities at the moment.

LITERATURE SURVEY

The study conducted by Kotsiantis et al is one of the original studies which delved into operation of machine literacy methods in distance literacy for powerhouse vaticination. The most significant donation bythis study was that it was a colonist and sculpted the path for several similar studies. While machine literacy algorithms had been preliminarily implemented in several settings, this was maybe the first time they were applied toan academic terrain.

Bhardwaj and Pal conducted a study in India, Faizabad to determine factors that most heavily affected pupil performance. They used Bayesian Bracket for their study

The study by Erkan Er was grounded upon Kotsiantis'as well as other analogous studies. It concludedthat Naive Bayes was more accurate than any other machine learning algorithm. Still, the pivotal donation of this study was that time-steady features may be mischievous to the machine literacy process, and hence are more left out of the study entirely. According to him, rather than using demographic characteristics of scholars, original attendance and schoolwork grades are more effective at the earlier stages of the process.

Bhardwaj and Pal conducted a study on the pupil performance grounded by opting 300 scholars from 5 different degree council conducting BCA (Bachelorette of Computer Operation) course of Dr.R.M.L. Awadh University, Faizabad, India. By means of Bayesian bracketsystem on 17 attributes, it was plant that the factors like scholars' grade in elderly secondary test, living position, medium of tutoring, mama's qualification, Among scholars, other habits, family income, and the pupil's family status were largely related to the student's academic performance. In the present study, those variables whose probability values were lesser than0.70 were given due considerations and the largely impacting variables with high probability values have been shown in Table 1.

These features were used for vaticination model construction. For both variable selection and vaticination model construction, publishers use MATLAB. From the table, it can be seen that the alternate high implicit variable for scholars' performance is their living position. The third high implicit variable for scholars' performance is the medium of tutoring. In Uttar Pradesh the mama lingo language of scholars is Hindi. Hence, scholars tend to be more comfortable in Hindi and other languages, than in theEnglish language.

The study conducted by Erkan Er proved precious in attesting the oneness of the proposed operation. Based on his research, he concluded that powerhouse rates for a distance literacy program were predicted by all current operations of machine literacy in an academic setting. Possibly, there is no operation that attempts to forecast the absolute performance of the student. However, it has not been published yet, If one does live. Kotsiantis et al compared five algorithms, viz. Decision Trees (C4.5), the Naive Bayes algorithm (Bayesian networks), 3-NN (kNN), RIPPER (Rule Literacy) and WINNOW (Perceptron grounded neural networks). This study consisted of two experimental stages, training and testing. During these stages, the number of attributes was increased step-by-step. For illustration, while only demographic data was included in the first step, performance attributes were added in the coming step. Five algorithms were tested for each these posterior way and also they were compared. This relative study helped in narrowing down campaigners for our own operation.

Data Mining can be used in educational field to enhance our understanding of literacy process to concentrate on relating, rooting and assessing variables related to the literacy process of scholars as described by Alaa elHalees. Mining in

educational terrain is called Educational Data Mining. Han and Kamber describe data mining software that allows the druggies to dissect data from different confines, classify it and visualize the connections linked during the mining process.

In their study, Pandey and Pal selected 600 students from various colleges at Dr.R.M.L. Awadh University, Faizabad, India in order to gauge pupils' performance. Bymeans of Bayes Bracket on order, language and background qualification, it was plant that whether new adventurer scholars will performer or not. Hijazi and Naqvi conducted as study on the pupil performance by opting a sample of 300 scholars from a group of sodalitiescombined to Punjab university of Pakistan. The thesis was stated as "Student's station towards attendance in class, hours spent in study on diurnal basis after council, scholars' family income, scholars' mama's age and mama'seducation are significantly related to pupil performance" was framed. By means of simple direct retrogression analysis, it was found that factors such as mother's education and pupil's family income were largely associated with the pupil's academic performance. Khan conducted a performance study on 400 scholars comprising 200 boys and 200 girls from the senior secondary academy of Aligarh Muslim University, Aligarh, a university located in Aligarh, India with a main ideal to establish the prognostic value of different measures of cognition, personality and demographic variables for success at advanced secondary position in wisdom sluice. The selection was grounded on cluster slice fashion in which the entire population of interest wasdivided into groups, or clusters, and a arbitrary sample of these clusters was named for farther analyses. It was found that girls with high socioeconomic status had fairly advanced academic achievement in their studies and boys with low socioeconomic status had fairly advanced academic achievement in general. Galit gave a case studythat uses scholars' data to dissect their literacy geste to prognosticate the results and to advise scholars at risk before their final examinations. A Review on Data Miningways and factors used in Educational Data Mining to prognosticate pupil amelioration.

Educational Data Mining (EDM) is an innovative interdisciplinary area that handles the development of approaches to explore data arising in educational fields. Computational approaches used by EDM is to examine educational data in order to study educational questions. As a result, it provides natural knowledge of tutoring and literacy process for effective education planning. This paper conducts a comprehensive study on the recent and applicable studies put through in this field to date. The study focuses on styles of analysing educational data to develop models for perfecting academic performances andperfecting institutional effectiveness. Literature is accumulated and relegated, consequential work is identified, and it is mediated to calculating preceptors and professional bodies in this paper. We identify exploration that gives well- fortified advise to amend edifying and amp the further impuissant member scholars in the institution. The results of these studies give sapience into ways for upgrading pedagogical processes, prognosticating pupil performance, compare the perfection of data mining algorithms, and demonstrate the maturity of open source tools. Data Mining Approach For Predicting Student Performance. This work proposes an innovative approach- substantiated soothsaying-to take into account the successional effect in prognosticating pupil performance (PSP). Rather than using all literal data as other styles in PSP, the proposed styles only use the information of the individual scholars for optimizing his/ her own performance. Also, these styles also render the "pupileffect" (e.g. how good/ clever a pupil is, in performing thetasks) and" task effect" (e.g. how delicate/ easy the task is) into the models. Experimental results show that the proposed styles perform nicely and much faster than the other state-of-the- art styles in PSP. A new approach for upgrading Indian education by using data mining ways. Education is the backbone of all developing countries. Elevation of the education system, upgrades the country tothe world top ranking position. One of the major problemsthat the education system facing is prognosticating the geste of scholars from large database. This paper focus on upgrading Indian education system by using one of the ways in Data.

PROBLEM STATEMENT

The main objective of this design is to extemporize pupil performance in studies grounded on some critical factors. Education is essential for a country's

betterment and progress. It enables the people of a country cultivated and well mannered. Nowadays developing new styles to discover knowledge from educational database inorder to assay pupil's trends and behaviours towards education. To assay the data from different confines classify it and to epitomize the connections. It motivated us to work on pupil dataset analysation. The data collection, categorization and bracket is being performed manually.

EXISTING SYSTEM

As of now, being system take only performanceinto consideration which isn't sufficient for having system,which can help us to estimate performance of a pupil. Wearen't having a system which would help us to integrate theperformance and undesirable into consideration. Disadvantages of Existing System, Being system miss theundesirable data for the scholars And It may not check thesocial data for the pupil.

DISADVANTAGE

Consequently, these generated rules did not fully extract the reasons for the reasons behind the student's dropout. Apart from the previously mentioned work, there were previous statistical analysis models fromthe perspective of educational psychology that conducted a couple of studies to examine the correlation between mental health and academic performance. The recommendations were too brief, they missed illustrating the methodologies to apply them.

PROPOSED SYSTEM

The work aims to develop a trust model using data mining ways, which mines needed information, so that the present education system may borrow this as a strategic operation tool. The proposed system use educational data mining ways to estimate performance andidentify undesirable geste. In the educational sector, Data mining is used for wide variety of operations including asperformance of the scholars like mark, attendance, staff opinion, adulterous conditioning, Ragging and stress. The data mining techniques used for relating the pupil's performance using K- means and KNN algorithms. Advantages of Proposed System, Educational database contain the useful information for Evaluating Students. The data mining techniques are more helpful in classifyingeducational database and help us in evaluating the performance and undesirable behaviour of a student.

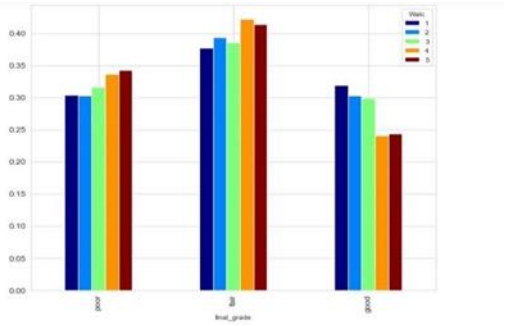

Fig.1

ADVANTAGE

The proposal aims to analyse students' demographic data, study-related details, and psychological characteristics in terms of final state to figure out whether the student is on the right track or struggling or even failing. Addition, we compare our proposed model to the existing related models. These recommendations are based on experiments that have been conducted to improve the student's academic performance. In addition to the above mentioned functionalities, the System will also alert all parties to the possible upcoming mental illnesses that the studentmight suffer from.

MODULE DESCRIPTION

The following modules are used in this project:

Data Collection: In this module, Student data's will be collected from the council. Student's data like mark, attendance, staff opinion, Social media, extracurricular activities, Ragging and stress.

Pre-processing: Data pre-processing is done to remove the deficient noisy and inconsistent data. Data must be preprocessed before using in point selection task.

Classification Module: The data mining method for relating the pupil's performance using Naïve Bayes andKNN algorithms. These two algorithm's identifies and analyses the performance of the pupil.

Prediction: Based on pupil marks, attendance, staff opinions, social media, extracurricular activities, and stress, we predict pupil performance in this module.

HARDWARE AND SOFTWARE REQUIREMENTS

Hardware:

Windows 7,8,10 64

bitRAM 4GB

Software:

Data Set

Python 2.7

Anaconda Navigator

PROJECT DESIGN ANDANALYSIS

1. ARCHITECTURE:

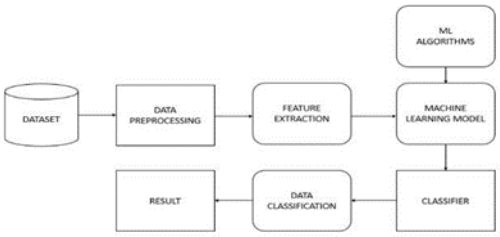

Fig.2

2. PROPOSED ALGORITHMS:

- K-nearest neighbor (knn) classification method
- Naive bayes algorithm

K-NEAREST NEIGHBOR (KNN) CLASSIFICATION METHOD

K-NN is a type of case- grounded literacy, or lazy literacy, where the function is only approached locally and all calculation is remitted until bracket. Thek-NN algorithm is among the simplest of all machine learningalgorithms. The neighbors are taken from a setof objects for which the class (for k-NN bracket) or theobject property value (for k-NN retrogression) is known.

STEP 1 BEGIN

STEP 2 Input D = (x1, c1),..., (xN, cN)

STEP 3 x = (x1... xn) new case to be classified

STEP 4 FOR each labelled case (xi, ci) calculate d(xi, x)

STEP 5 Order d (xi, x) from smallest to loftiest, (i =1... N)

STEP 6 select the K nearest cases to x Dkx

STEP 7 Assign to x the most frequent class in Dkx

STEP 8 END

NAIVE BAYES ALGORITHM

It's a bracket fashion grounded on Bayes 'Theorem with an supposition of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular point in a class is unconnected to the presence of any other point.For illustration, a fruit may be considered to be an apple if it's red, round, and about 3 elevation in periphery. Indeed if these features depend on each other or upon the actuality of the other features, all of these parcels singly contribute to the probability that this fruit is an apple and that's why it's known as' Naive'. Let's understand it using an illustration. BelowI've a training data set of rainfall and corresponding target variable' Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not grounded on rainfall condition. Let's follow the below way to perform it.

Step 1 Convert the data set into a frequency table.

Step 2 Produce Liability table by chancing the chances like Heavy probability = 0.29 and probabilityof playing is0.64.

Step 3 Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class withthe loftiest posterior probability is the outgrowth of vaticination.

UML DIAGRAMS

UML is simply anther graphical representationof a common semantic model. UML provides a comprehensive memorandum for the full lifecycle of object acquainted development. Advantages, To represent complete systems (rather of only the software portion) using object acquainted generalities. To establish an unequivocal coupling between generalities and executable law. To take into account the scaling factors that are essential to complex and critical systems. To creating a modeling language usable by both humans and machines.

UML defines several models for representing systems. The class model captures the stationary structure. The state model expresses the dynamic geste of objects. The use case model describes the conditions of the stoner. The commerce model represents the scripts and dispatches flows. The perpetration model shows the work units. The deployment model provides details that pertain to reuse allocation.

1.USE CASE DIAGRAM

Use case diagrams overview the operation demand for system. They're useful for donations to operation and/ or design stakeholders, but for factual development you'll find that use cases give significantly further value because they describe “the meant” of the factual conditions. A use case describesa sequence of action that provides commodity of measurable value to an action and is drawn as a vertical cirque.

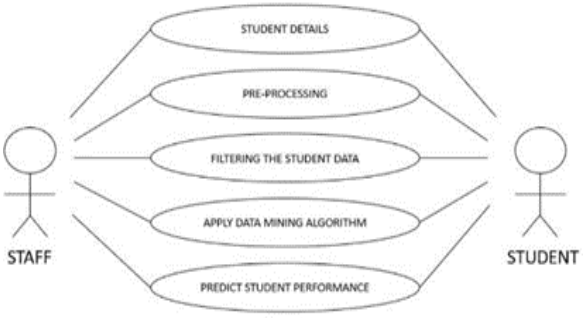

Fig.3

2.SEQUENCE DIAGRAM

Sequence Illustration model the inflow of sensewithin your system in a visual manner, enabling you both to validate and validate your sense, and generallyused for both analysis and design purpose. Sequence illustration are the most popular UML artifact for dynamic modeling, which focuses on relating the gestewithin your system.

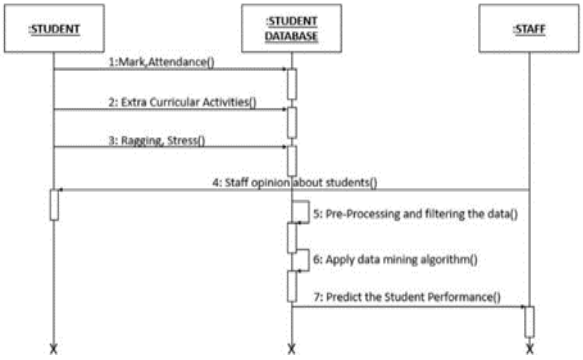

Fig.4

3.ACTIVITY DIAGRAM

Exertion illustration are graphical representations of workflows of accretive conditioningand conduct with support for choice, replication and concurrency. The exertion plates can be used to describe the business and functional step-by- step workflows of factors in a system. Exertion illustration correspond of Original knot, exertion final knot and conditioning in between.

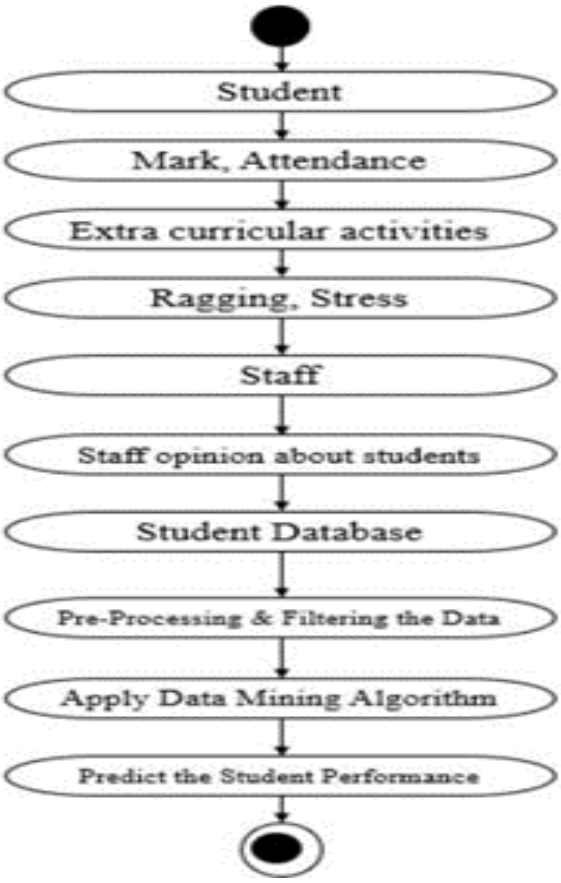

Fig.5

SYSTEM DESIGN

1. INPUT DESIGN

The input design is the link between the information system and the stoner. It comprises the developing specification and procedures for data medication and those way are necessary to put sale data in to a usable form for processing can be achieved by examining the computer to redate from a written or published document or it can do by having people conciliating the data directly into the system. The design of input focuses on controlling the quantum of input needed, controlling the crimes, avoiding detention, avoiding redundant way and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retainingthe sequestration. Input Design considered the following effects ,What data should be given as input? How the data should be arranged or enciphered? The dialog to guide the operating help in furnishing input. Styles for preparing input attestations and way to follow when error do.

2. OUTPUT DESIGN

A quality affair is one, which meets the conditions of the end stoner and presents the information easily. In any system results of processing are communicated to the druggies and to other system through labors. In affair design it's determined how the information is to be displaced for immediate need and also the hard dupe affair. It's the most important and direct source information to the stoner. Effective and intelligent affair design improves the system's relationship to help stoner decision- timber. The affairform of an information system should negotiate one or further of the following objects. Convey information about once conditioning, current status or protrusions of the Future. Signal important events, openings, problems, or warnings. Detector an action. Confirm anaction.

IMPLEMENTATION TESTING

A dataset of student data is collected and analyzed using data mining techniques throughout the implementation phase in order to produce rules for the analysis of student performance. Jupiter Notebook, an open-source software platform, is used to generate rules. Training and test sets are created from the dataset. The training set uses 25% of the dataset, with the test set using the remaining 75%.

The objective of this project is to analyse student performance. It is a machine learning model that takes student data as input and analyses using logistic regression and random forest. The below tabledisplays parameters used.

algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently.

After analysis race/ethnicity and parental level of education columns were excluded as they are not affecting the result of students. Previous exam scores were not considered and the remaining factors are considered for prediction which is done using LightGBM algorithm. LightGBM wasused as it requires less memory to execute and canhandle enormous amounts of data.

RESULTS

The below figure shows confusion matrix for Logistic Regression:

Fig.7

The confusion matrix shown above helps in analysing the accuracy of results when logistic regression is implemented. This test result's accuracy for testing data is 0.424.

The below figure shows confusion matrix for Random Forest:

Steps:

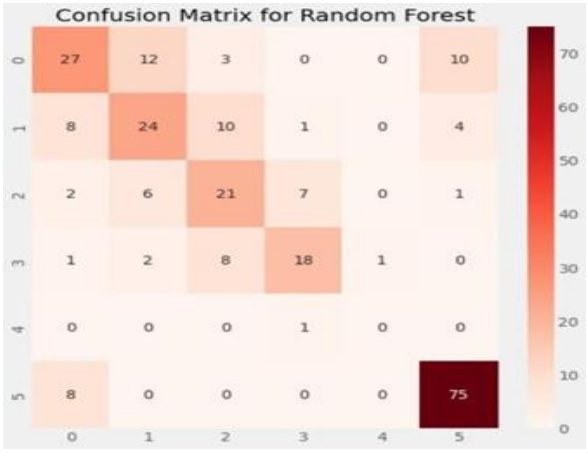

Fig.6

In first step, student's dataset is taken.

1. Next data cleaning is done where noisy data and incomplete data was removed.
2. In this step, data is divided into part for training and testing purposes. 25% of the data was kept for training the model and 75% for testing it.
3. Analysis of student's performance is done using logistic regression and random forest algorithms. Logistic regression was implemented as it is easier to interpret and very efficient to train. It takes less training time as compared to other

	Feature	Values	Description
1	Gender	Male/Female	Student's gender
2	Race/Ethnicity	Group: A/B/C	Group student belongs to
3	Parental level of Education	Bachelor's/ Master's/ other degree	Student parent's qualification
4	Sports	Y/N – 1/2	Student's participation in sports 1-yes 2-no
5	Test preparation course	Completed/ none	Student's testcompletion status
6	Math score	Integer:(0-100)	Student score in math subject
7	Reading score	Integer:(0-100)	Student score in reading
8	Writing score	Integer:(0-100)	Student score in writing
9	Attendance	Integer:(0-100)	The number of days the student is present

Fig.8

The confusion matrix shown above helps us in analysing the accuracy of results when random forest is implemented. This test result's accuracy for testing data is 0.648.

The dataset is split into training and test sets inboth techniques. The training set takes up 25% of the dataset, while the test set takes up the remaining 75%.

Test result is visualized as shown below:

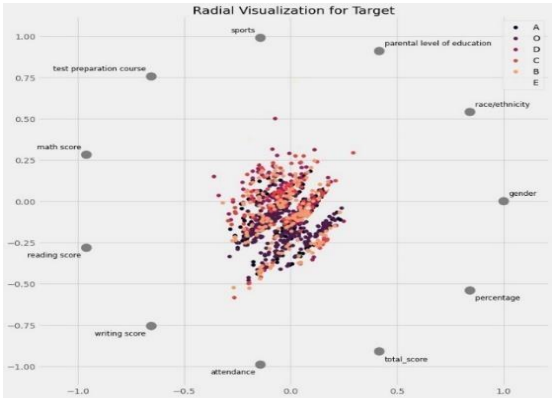

Fig.9

The student analysis is displayed using a scatterplot, where each colour corresponds to a certain grade. Students deviate more when their sports practice is prolonged or when absenteeism lowers the attendance rate.

CONCLUSION AND FUTURESCOPE

The research's main goal is to employ machine learning techniques to analyse how students' academic progress is developing. The analysis is done using random forest and logistic regression. This model can be utilized in variety of circumstances, including departmental level and at basic academic level for presenting a concise summary of the performance related to particular course. With a few modest alterations, any industrial organization or company can use this application to assess task participation and determine the perfect candidate based on productivity. This procedure can make it easier for the instructor to assess student performance and plan more effective academic improvement strategies. All other elements are taken into account for prediction besidespast exam results. Future updates to our dataset could include more features for improved accuracy and in-depth research.

ACKNOWLEDGEMENT

We thank CMR Technical Campus for supporting this paper titled "Student Performance Analysis Using Machine Learning", which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department Of Computer Scienceand Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work.

REFERENCES

[1.] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery databases.," AI Magazine, pp. 37-53, 1996.

[2.] "IndiceBrasscom de Convergencia Digital," 2015. [Online]. Available: www.brasscom.org.br. [Accessed:14- May-2016].

[3.] H. Witten, E. Frank and M. A. Hall, Data mining: Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann, 2011.

[4.] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," in ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18, Jun. 2009.

[5.] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance," in International Journal of Advanced Computer Science and Applications– IJACSA, vol. 2, no. 6, pp. 63-69,2011.

[6.] Poza-Lujan, Jose-Luis and Calafate, Carlos T. and PosadasYague. ”As- sessing the Impact of Continuous Evaluation Strategies: Tradeoff Between Student Performance and Instructor Effort”, IEEE Transactionson Edu- cation, vol.59, pp.17-23, Feb 2016.