OAIJSE

OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

DATA MINING ANALYSIS ON STUDENT'S ACADEMIC PERFORMANCE THROUGH EXPLORATION OF STUDENT'S BACKGROUND AND SOCIAL ACTIVITIES

MISS. SAINDANE AISHWARYA[1] PROF. HARISH K. BARAPATRE[2]

PG Students, Dept. of Computer, YTIET, Maharashtra, India[1]

Assistant Professor, Dept. of Computer , YTIET, Maharashtra India[2]

Abstract: Application of data mining and machine learning techniques in the field of education is gaining prominence these days as it would help Identify weak areas of the student and predict student performance. This would help both the student and the educational institution. Research till date has only predicted the final grade of the student. In this paper, we attempted to analyze the various factors that contributed to the student performance like the background of the student, involvement in social activities, student performance in the coursework etc. data mining techniques such Decision Tree J48, Naïve Bayesian and KNN algorithms are used to predict the student performance in mathematics at secondary school. We find that the background of the student and involvement in the social activities have played an important role in forecasting student performance. This model could be adopted for the early prediction of student achievements and assist in improving on the weak areas of the student.

Keywords: *Educational Data Mining, Classification*

INTRODUCTION

With rapid growth of artificial intelligence and machine learning, these two techniques are being applied in almost every domain to gain insights into the data and predict outcomes. Data mining and machine learning techniques are being applied in the field of education also to understand and predict student performance. Applying data mining and machine learning techniques in the educational domain would help to take decisions beforehand to understand and predict weak areas of a student and improve student performance. Thus, this benefits all the stakeholders domain in the educational domain – The student, educational institutions, parents and teachers. As it is a win-win situation for everyone, more and more educational institutions are being interested in applying the data mining techniques and predict student performance. The trainers would have enough time to improve on the weak areas of the student based on the discovered insights and the recommendations from the data mining system would help the students improve their performance. Application of data mining techniques in the educational domain also helps in categorizing academic records based on student’s learning patterns, activities etc.

Current literature has been focusing on student performance in the academics, teaching quality, learning methodologies as the prime elements in predicting a student’s overall performance. But there are other factors which can impact the student performance such as the background of the student, attendance of school, involvement in social activities, study habits etc. Considering the impact of the above-mentioned factors would help to improve the student performance in a particular subject as early as possible. In this paper, we try to analyze and recognize the impact of the background characteristics of a student in evaluating and predicting student performance. The performance forecasting model is built using the supervised data mining techniques such as Decision Tree J48, Naïve Bayesian and KNN algorithms. It has been observed that student background and social activities have a significant impact on student performance and their impact can be graphically visualized from the decision tree structure generated by the models.

II LITERATURE SURVEY

Performance Analysis and Prediction in Educational Data Mining – A Research Travelogue

Nowadays due to digitalization and computerization many changes have occurred in the field of education. Large amount of data is residing in the educational databases, but it is being utilized effectively. Latest Technologies and powerful tools like machine learning and Big Data have to be used to analyze the data in the databases and gain insights from it. This paper gives an insight into the research made in the field of Educational Data Mining. Based on this paper, we have selected the algorithms used to predict the student performance. They are: Decision Tree J48, Naïve Bayesian algorithms and KNN algorithms.

Application of Big Data in Education Data Mining And earning Analytics

With the advancement in mobile technologies and ubiquitous usage of smart phones, they have become a part of student's life to access the online content. Online activities of the students generate enormous amounts of data. This data is remaining unused and getting wasted as conventional analytical systems do not have the capacity to process them. This paper gives insights into how Big Data and Machine learning can be used in the field of Educational data mining. Some of the applications of Big Data in EDM mentioned in this paper are: Student Performance Prediction, Student Drop-outs - Attrition Risk Detection, Data Visualization to identify trends, Development of Intelligent feedback systems, Course Recommendation System, Skill Estimation of Student etc.

Role of Data warehousing in Educational Data Analysis

Most of the educational institutions have lot of data but there are no efficient tools to analyze the high volume of data and translate it into information or knowledge which could be useful. This paper gives insights into the data warehousing strategies that must be used by educational institutions to build a good datawarehouse that can be used to analyze the data present with institution over a period. It discusses about how to collect the data from various source systems, clean it, load it and integrate it in the data warehouse to gain insights into the data. This would lead to understand the elements that impact the success of the student, optimizing the resources and increase productivity.

III SYSTEM ANALYSIS

Existing system:

Data mining techniques are being applied in the field of education to understand and predict student performance and this helps in focusing on the weak areas of a student as early as possible. In the current literature, the factors considered for predicting student performance are student performance in academics, teaching quality and learning methodologies. The existing system does not consider the other important elements that impact distorting performance such as the background of the student, students involvement in social activities, the study habits of the student, the learning patterns of the student, Attendance in school etc.

Disadvantages:

1. The accuracy of the model using few attributes like teaching quality, student performance and learning methodology is not satisfactory.
2. Desired results are not achieved.

IV PROBLEM DEFINITION

Traditional data analysis approaches are not capable of extracting information from the educational systems and predict the student performance well ahead based on existing situations and attributes associated with the student. Due to exponential growth in data associated with students on various online platforms data mining techniques and related analytical approaches have to be followed to make use of the existing multidimensional data in the educational systems and gain insights. Hence the data associated with the educational systems is BigData and

V PRPOPSED SYSTEM

In this paper we try to analyze the impact of other important attributes that affect student performance like the background of the student, students involvement in social activities, study habits, attendance in school etc along with student performance in coursework on a 2-level classification and 5-level classification using the data mining algorithms such as Decision Tree J48 Naïve Bayesian algorithms and KNN algorithms.

Advantages:

1. Accuracy of proposed system is good
2. Early prediction is possible
3. Desired results are achieved.

Decision Trees – J48 Algorithm:

Decision Trees – J48 algorithm is a statistical classifier and can work effectively on student datasets. It is one of the best data mining techniques that is capable of classifying information based on input features in an effective way. Information Entropy is the basis of this algorithm.

KNN Algorithm

KNN or K-Nearest Neighbour algorithm is a statistical algorithm which does not use any parameters. The features that are used for prediction and training the model are analyzed and the classification of other records in the dataset is done based on identifying its nearest neighbor and grouping them together. It is a form of example based learning.

Naives – Bayesian algorithm:

Bayes theorem in statistics is the base for the Naives Bayesian algorithm. This is one of the preferred algorithms for classification due to its high accuracy and it is probabilistic in nature.

VI SYSTEM DESIGN

System Architecture:

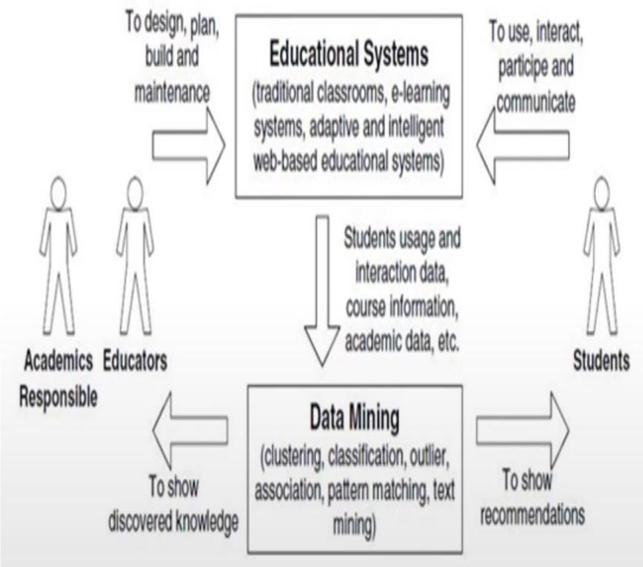

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, particpe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 1 System Architecture

The Figure1 shows the system architecture. The data collected from various educational systems is fed to the data mining system. The recommendations from the Data Mining system are taken by the educators to plan the courses and address the difficulties faced by the students to perform better. The students can understand their weak areas and prepare ahead.

2-Level Classification and 5-Level Classification

This experimental result also has shown that student coursework results are significant attributes in predicting student performance in mathematic final grade as it has highest precision accuracy 0.924 in 2-level classification and 0.791 in 5-level classification. In overall, accuracy of algorithms in 2-level classification are out performed models in 5-level classification. The models accuracy are> 0.5, this indicated that student background and student social activities are viable to be used to perform early analysis and prediction of at-risk student to determine whether it pass or fail the subject.

2. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

3. Data visualization

Data visualization is an important preprocessing task, which used graphical representation to simplify and understand complex data. Visualization techniques have been recently used to visualize online learning aspects. Instructors can utilize the graphical representations to understand their learners better and become aware of what is occurring in the distance classes.

Data Flow Diagram:

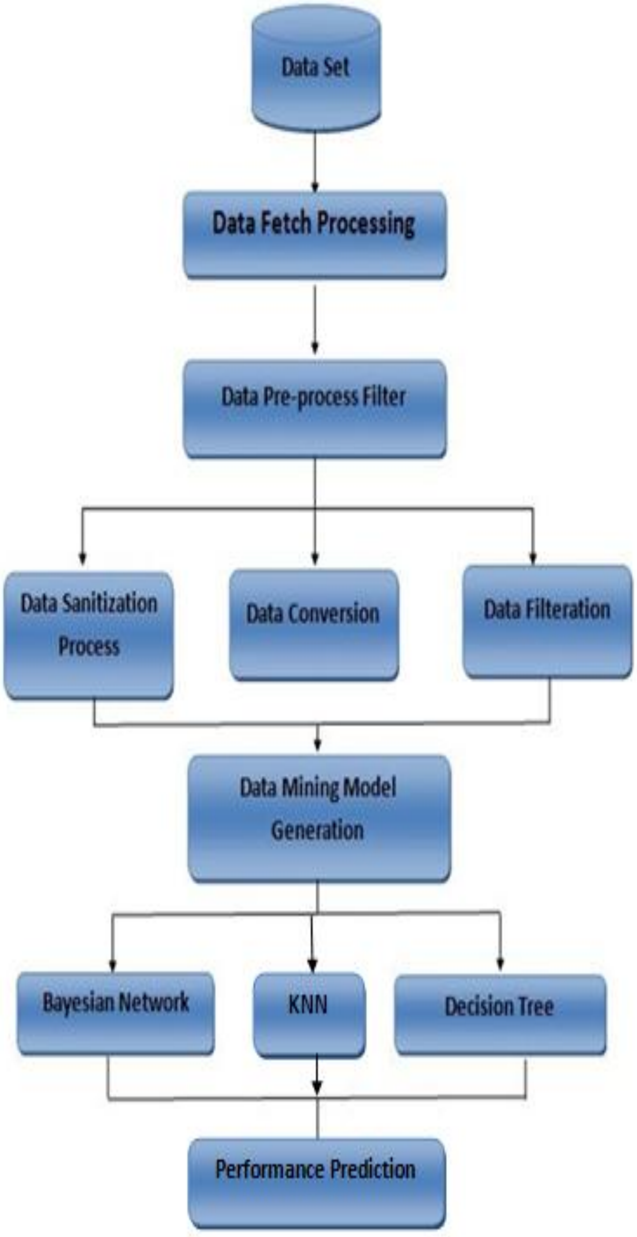


Figure 2 Data Flow Diagram

The above diagram represents the flow of data within the system. The data is collected from different systems and integrated into a single dataset. As the dataset is used to train our model, it has processed to remove any data anomalies, normalize data and filter out any unnecessary data. This cleaned and normalized data is used to generate a model and apply the data mining algorithms like Naïve Bayesian, KNN and Decision Tree – J48. The results predicted by each of these algorithms are analyzed individually for accuracy.

VII IMPLEMENTATION

This project has the below Five modules:

1. **Initial Data Processing Technique**
2. **Crucial Data Mining Process**
3. **Data Clustering Methodology**
4. **Data Classification Process**
5. **Data Facts Representation**

1. Initial Data Processing Technique:

Collection of data is been given as an input data set in order to perform analytical evaluation of students performance. This data processing section raw data from a specific data set is been taken into consideration and being evaluated and converted into proper segmented dependent variable classification in order to improve the quality of the data. This process of data free processing mechanism has to been taken over effectively in order to predict the secure information enabling data access policies for the data mining models. This initial stage of prior processing of data is been driven effectively and efficiently in order to eliminate semi finished values so that we could achieve out-rated retrieval of data from dataset. The above said methodology is demanded in order to maintain equality of data processing prior to the data mining approach.

2. Crucial Data Mining Process:

This process includes Data Selection & Data Transformation activities. About two processes are effectively driven in order to trigger variable data fields that are been identified in the process of data digging raw data mining in such that dependent variables are they selected are been mined from the desired data set. So this extraction of dependent variables from the desired data set helps us to evaluate the student performance by effectively implementing the data prediction model supervised approach in such that student background and social activities are being considered in predicting the student performance using an effective analytical approach.

3. Data Clustering Methodology:

This crucial stage of clustering mechanism is being driven with unsupervised data valuation policies along with arithmetic effective approaches so as to handle data

duplication methodologies over homogeneous data cluster environment. Data set that is been provided in data preprocessing stage is related with huge data information which has to get big with the proper data protection. Association rule mining and data decision evaluations are been carried out in a most effective and efficient way a using concealed data pattern extractions in a rational approach. The total data set that is been submitted in the initial stage of data pre processing has to get fragmented with the proper subsets of information in order to make a move analytical approach so that these data subsets are been categorized into various data instance clusters wherein each of that has a complete information over the different variable elements which has a correlation between 1 and other are been grouped effectively where as independent data items or elements are been isolated and dumped into eliminated clusters.

4. Data Classification Process:

The above said effective and efficient data clustering alone can't help us to address analytical approach in evaluating the performance of a student it which pen need to be carried out into data categorization for classification show that the prediction process will be effectively be driven. In our technical analytical approaches of Naïve Bayesian, Decision Tree J48 and K-Nearest neighbor algorithms in general demands data preprocessing, data clustering, data classification are to be driven effectively in such that data mining strategies are being implemented effectively in order to predict the student performance in an efficient and effective way. Data classification plays a crucial role in such that student data can be mined properly so that desired student group data set could be extracted with their behavioral, social, background attributes are considered and production process made simpler in a most effective way.

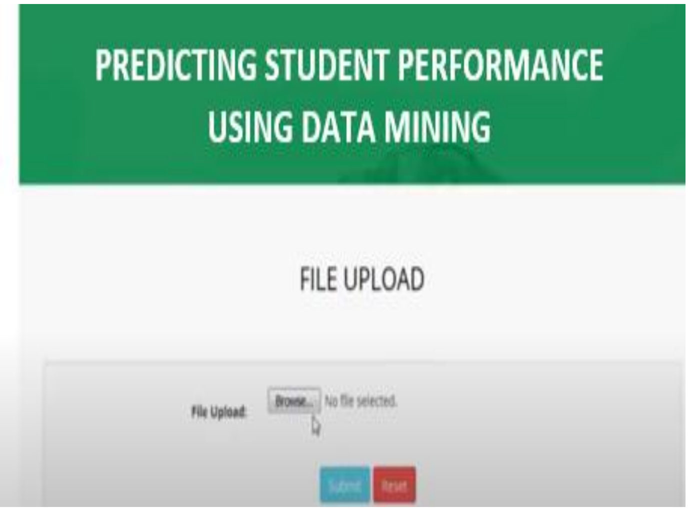
5. Data Facts Representation

In data knowledge segmentation huge data is been extracted from the desired academic Student data set that contained both dependent and independent wearable items which is been supplied to the above said linear classifier methodologies in such that proper segmentation or for clustering is been effectively done and inter-cluster comparison process will be driven effectively in order to predict the student performance evolution in a most effective way. Show the stages of data classification and clustering mechanism plays a crucial role in the approach of an analytical strategies in predicting the students performance any more efficient and effective ways so that the data representation or segmentation is been driven with a more precise optimal policies.

VIII. ANALYTICAL RESULT SCREENS

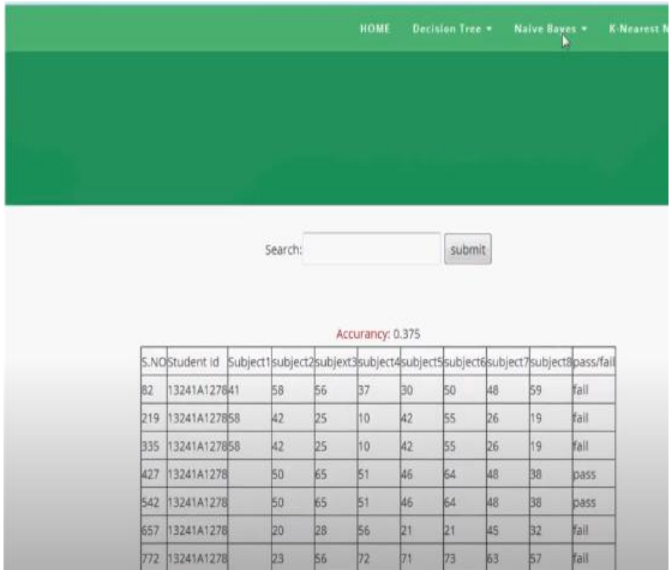
Upload Data:

This screen allows uploading the student data and predicting the student performance by applying different techniques like Decision Trees, Naïve Bayesian and K-Nearest Neighbour algorithms.



Performance Prediction using Decision Trees:

This screen applies the Naïve Bayesian algorithm to the dataset uploaded and shows the accuracy of the algorithm. A single student search is also provided where the predictions would be made for that student and the accuracy of that prediction is also shown



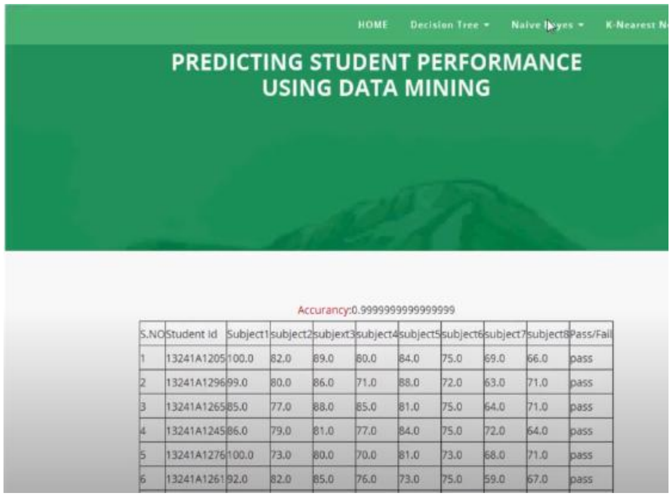
Search: submit

Accurancy: 0.375

S.NO	Student id	Subject1	subject2	subjext3	subject4	subject5	subject6	subject7	subject8	pass/fail
82	13241A1278	41	58	56	37	30	50	48	59	fail
219	13241A1278	58	42	25	10	42	55	26	19	fail
335	13241A1278	58	42	25	10	42	55	26	19	fail
427	13241A1278		50	65	51	46	64	48	38	pass
542	13241A1278		50	65	51	46	64	48	38	pass
657	13241A1278		20	28	56	21	21	45	32	fail
772	13241A1278		23	56	72	71	73	63	57	fail

Performance Prediction using K-Nearest Neighbor:

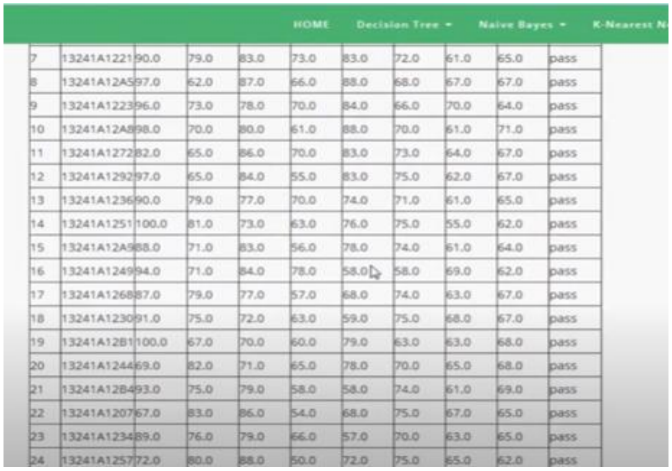
This screen applies the K-Nearest Neighbor algorithm to the dataset uploaded and shows the accuracy of the algorithm. A single student search is also provided where the predictions would be made for that student and the accuracy of that prediction is also shown



Accurancy:0.9999999999999999

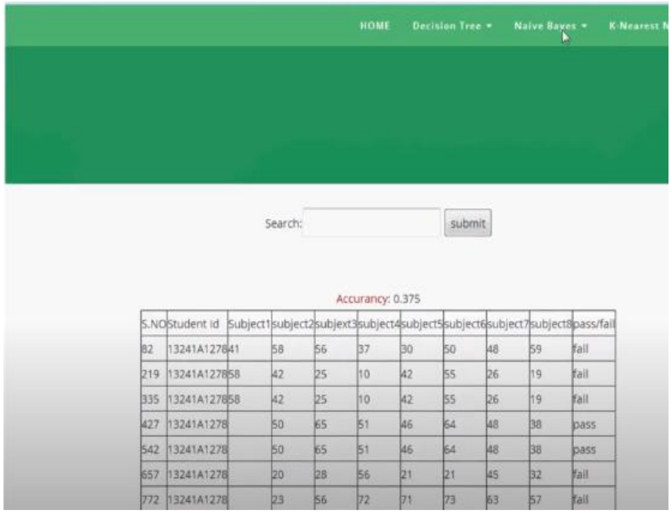
S.NO	Student id	Subject1	subject2	subjext3	subject4	subject5	subject6	subject7	subject8	Pass/Fail
1	13241A1205	100.0	82.0	89.0	80.0	84.0	75.0	69.0	66.0	pass
2	13241A1296	99.0	80.0	86.0	71.0	88.0	72.0	63.0	71.0	pass
3	13241A1265	85.0	77.0	88.0	85.0	81.0	75.0	64.0	71.0	pass
4	13241A1245	86.0	79.0	81.0	77.0	84.0	75.0	72.0	64.0	pass
5	13241A1276	100.0	73.0	80.0	70.0	81.0	73.0	68.0	71.0	pass
6	13241A1261	92.0	82.0	85.0	76.0	73.0	75.0	59.0	67.0	pass

This screen applies the Decision Trees algorithm to the dataset uploaded and shows the accuracy of the algorithm



7	13241A1221	90.0	79.0	83.0	73.0	83.0	72.0	61.0	65.0	pass
8	13241A12A5	97.0	62.0	87.0	66.0	88.0	68.0	67.0	67.0	pass
9	13241A1223	96.0	73.0	78.0	70.0	84.0	66.0	70.0	64.0	pass
10	13241A12A8	98.0	70.0	80.0	61.0	88.0	70.0	61.0	71.0	pass
11	13241A1272	82.0	65.0	86.0	70.0	83.0	73.0	64.0	67.0	pass
12	13241A1292	97.0	65.0	84.0	55.0	83.0	75.0	62.0	67.0	pass
13	13241A1236	90.0	79.0	77.0	70.0	74.0	71.0	61.0	65.0	pass
14	13241A1251	100.0	81.0	73.0	63.0	76.0	75.0	55.0	62.0	pass
15	13241A12A9	88.0	71.0	83.0	56.0	78.0	74.0	61.0	64.0	pass
16	13241A1249	94.0	71.0	84.0	78.0	58.0	58.0	69.0	62.0	pass
17	13241A1268	87.0	79.0	77.0	57.0	68.0	74.0	63.0	67.0	pass
18	13241A1230	91.0	75.0	72.0	63.0	59.0	75.0	68.0	67.0	pass
19	13241A12B1	100.0	67.0	70.0	60.0	79.0	63.0	63.0	68.0	pass
20	13241A1244	69.0	82.0	71.0	65.0	78.0	70.0	65.0	68.0	pass
21	13241A12B4	93.0	75.0	79.0	58.0	58.0	74.0	61.0	69.0	pass
22	13241A1207	67.0	83.0	86.0	54.0	68.0	75.0	67.0	65.0	pass
23	13241A1234	89.0	76.0	79.0	66.0	57.0	70.0	63.0	65.0	pass
24	13241A1257	72.0	80.0	88.0	50.0	72.0	75.0	65.0	62.0	pass

Performance Prediction using Naïve Bayesian Algorithm:



Search: submit

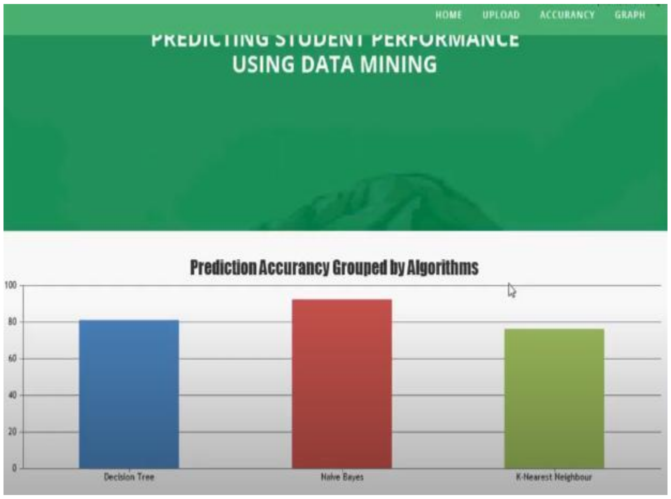
Accurancy: 0.375

S.NO	Student id	Subject1	subject2	subjext3	subject4	subject5	subject6	subject7	subject8	pass/fail
82	13241A1278	41	58	56	37	30	50	48	59	fail
219	13241A1278	58	42	25	10	42	55	26	19	fail
335	13241A1278	58	42	25	10	42	55	26	19	fail
427	13241A1278		50	65	51	46	64	48	38	pass
542	13241A1278		50	65	51	46	64	48	38	pass
657	13241A1278		20	28	56	21	21	45	32	fail
772	13241A1278		23	56	72	71	73	63	57	fail

Comparison of Algorithms for Student Performance Prediction:

The below screen shows a comparison of 3 different algorithms that are used to predict the student performance.

It can be observed that Naïve Bayesian algorithm can predict the student performance with 85% accuracy.



IX CONCLUSION

Educational data mining techniques are used to predict student performance at a very early stage due to the consideration of important factors such as student background and involvement of student in social activities. This helps the at-risk students at a very early stage and gives them scope to improve their performance. The teachers can also prepare early and come up with new techniques in teaching the topic.

In future, unsupervised machine learning algorithms and data mining techniques could be applied to discover the relationship and impact up the attributes in clusters. This attribute analysis and feature selection would help in building more accurate models to predict the student performance.

REFERENCES

[1] Performance analysis and prediction in educational data mining: a research travelogue, international journal of computer applications (0975 – 8887) volume 110 – no. 15, January 2015.

[2] Application of big data in education data mining and learning analytics – a literature review, ictact journal on soft computing: special issue on soft computing models for big data, July 2015, volume: 05, issue: 04

[3] The role of data warehousing in educational data analysis, shirin mirabedini department of computer engineering, payame noor university, Tehran, Iran

[4] Application of discriminant analysis to predict the class of degree for graduating students in a university system Erimafa J.T., Iduseri A. And Edokpa I.W.*

[5] Pool, Lorraine Dacre, Pamela Qualter, And Peter J. Sewell. "exploring the factor structure of the career edge employability development profile." education+ training 56.4 (2014): 303-313.

[6] SARANYA, S., R. AYYAPPAN, & N. KUMAR. "student progress analysis and educational institutional growth prognosis using data mining." international journal of engineering sciences & research technology, 2014

[7] Hicheur cairns, awatef, et al. "towards custom-designed professional training contents and curriculums through educational process mining." immm 2014, the fourth international conference on advances in information mining and management. 2014.

[8] Saptarshi Ray, “big data in education”, gravity, the great lakes magazine, pp. 8-10, 2013.

[9] Peggy a. ertmer and timothy j. newby, “behaviorism, cognitivism, constructivism: comparing critical features from an instructional design perspective”, performance improvement quarterly, vol. 6, no. 4, pp. 50-72, 1993.

[10] Wikipedia, “big data --- wikipedia, the free encyclopedia”, https://en.wikipedia.org/w/index.php?title=big_data&oldid=669888993. accessed 2015.

[11] Adebayo sb, jolayemi et (1998b). on the effect of rare outcome on some agreement/concordance indices. nig. j. pure and appl. sci.: 718-723.