IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Student Performance Prediction System using Data Mining Approach

Kalpesh P. Chaudhari[1], **Riya A. Sharma**[2], **Shreya S. Jha**[3], **Rajeshwari J. Bari**[4]

Department of Computer Engg, SSBT COET, Jalgaon[1,2,3,4]

Abstract: The success of an academic institution can be measured in terms of quality of education provides to its students. In the education system, highest level of quality is achieved by exploring the data relating to redirection about students performance. These days the lack of existing system to analyse and judge the students performance and progress isn't being addressed. There are 2 reasons why this is often happening. First, the present system is not accurate to predict students' performance. Second, because of shortage of consideration of some vital factor those are affecting students' performance. Predicting students' performance is more challenging task as a result of large amount of information in academic database. This proposed system can help to predict students' performance more accurately. For these suitable data mining approach will be applied. In this approach, preprocessing step will be applied to raw dataset so that the mining algorithm will be applied properly. The prediction about students' performance can help him/her to enhance the performance.

Keywords: Education, student, performance, data mining, pre-processing, database, prediction.

I. INTRODUCTION

Improving student's academic performance is not an easy task for the academic community of higher learning. The academic performance of engineering and science students during their first year at university is a turning point in their educational path and usually encroaches on their General Point Average (GPA) in a decisive manner. The students evaluation factors like class quizzes mid and final exam assignment lab -work are studied. It is recommended that all these correlated information should be conveyed to the class teacher before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. In this paper, we present a hybrid procedure based on Decision Tree of Data mining method and Data Clustering that enables academicians to predict student's GPA (SGPA, CGPA) and based on that instructor can take necessary step 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 the academic planners to evaluate progression in an academic environment. Many factors could act as barriers to student 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. Clustering technique is most widely used technique for future prediction. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities. These applications can help both instructor and student to enhance the education quality. This study makes use of cluster 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 ratio and grade ratio. Clustering is one of the basic techniques often used in analysing data sets. This study makes use of cluster analysis to segment students in to groups according to their characteristics and use decision tree for making meaningful decision for the student's.

II. LITERATURE SURVEY

Data mining (sometimes known as knowledge or information discovery) is the method of analysing 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 analysing information. It permits users to analyse 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 data mining methods accuracy to classifying students in order to predicting category grade of a student. These predictions are more

DOI10.17148/IJARCCE.2017.63195

helpful for identifying the weak students and helping administration to take remedial measures at initial stages to produce excellent graduates which will graduate at least with the upper second category [1].

- Rathee and Mathur applied ID3, C4.5 and CART decision tree algorithm on educational information for predicting a student performance in the examination. All the algorithms are applied on the internal assessment information of student to predict their academic performance in the final examination. The efficiency of various decision tree algorithms will be analysed based on their accuracy and time 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 the 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. 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].

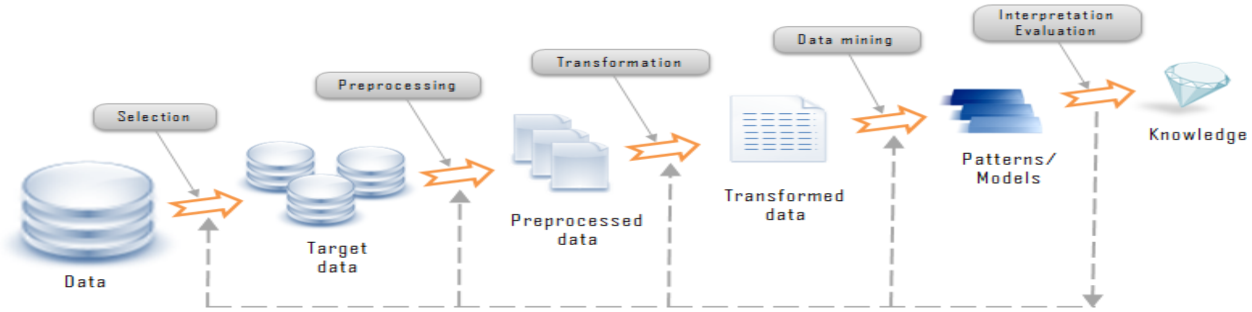
III. STEPS OF DATA MINING

Data mining is the method of discovering numerous models, derived values and summaries from a given collection of information. It’s necessary that the problem of discovering or estimating dependencies from information or discovering new information is simply one part of the overall experimental procedure utilized by engineers, scientists and others who apply standard steps to draw conclusions from information. The overall method of finding and decoding 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 goals of the end-user (formulate the hypothesis).
2. **Data Collection**: Determining how to find and extract the right data for modeling. First, we need to identify the different data sources are available. Data may be scattered in different spreadsheets, files, and hard-copy (paper) lists.
3. **Data integration**: Integration of multiple data cubes, databases or files. A big part of the integration activity is 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, and we select only the data which useful for data mining. Only relevant information is selected.
5. **Pre-processing**: The Major Tasks in Data Pre-processing are: Cleaning, Transformation and Reduction.

- **Data cleaning**: Additionally known as data cleansing. It deals with errors detection and removing from information so as to improve the quality of information. Information cleaning sometimes includes fill in missing values and identify or remove outliers.
- **Data Transformation**: Data transformation operations are additional procedures of data pre-processing that would contribute toward the success of the mining process and improve data-mining results. Some of Data transformation techniques are Normalization, Differences and ratios and Smoothing.
- **Data Reduction**: For large datasets there’s an increased probability that an intermediate, data reduction step should be performed before applying data mining techniques. While massive datasets have potential for higher mining results, there’s no guarantee that they’ll produce better knowledge than small datasets. Data Reduction obtains a reduced dataset representation that’s much smaller in volume, however produces constant analytical results.

6. **Building the model**: in this step we elect and implement the appropriate data mining task (ex. association rules, serial pattern discovery, classification, regression, clustering, etc.), the data mining technique and also the data processing algorithm(s) to create the model.
7. **Interpretation of the discovered knowledge** (model /patterns): 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/ 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 take better decisions [7].



DOI10.17148/IJARCCE.2017.63195

- **DATA MINING PROCESSING**

Today's academic system, a student's performance is decided by the internal assessment and end semester examination. The interior assessment is administered by the teacher based mostly upon student's performance in academic activities like class test, seminar, assignments, general proficiency, attendance and practical work. The end semester examination is one that's scored by the student in semester examination. Every student must get minimum marks to pass a semester in internal similarly as end semester examination[6].

A. DATA PREPARATION :

The data set used in this study is obtained from SSBT College of Engineering and Technology, Jalgaon. The result of previous students of different branches is collected from the college database and their behaviours are collected from their respective faculties and students. It is stored in other database and it is used to predict the performance of the present students in their next semesters[4].

B. DATA SELECTION AND TRANSFORMATION

In this step only those were chosen that were needed for data mining. Some derived variables were chosen. Whereas some of the data for the variables was extracted from the database. All the predictor and response variables that were derived from the database are given in following table for reference.

In this paper 17 different behaviours of the students are considered which is used to predict the performance of the present students in their upcoming semester. The complete description of the different behaviours with their possible values and variables are as shown in Table.

C. DATA CLUSTERING:-

data clustering is unsupervised and statistical information analysis technique. It is used to classify identical information into a homogenous cluster. It is used to operate an oversized data-set to find hidden pattern and relationship helps to make decision quickly and with efficiency. In a word, cluster analysis is employed to segment an oversized set of information into subsets referred to as clusters. Each cluster may be a collection of information objects that are similar to each other are placed inside an equivalent cluster but are dissimilar to objects in other clusters[5].

- **Implementation of k-means clustering algorithm:-**

k-means is one amongst the simplest unsupervised learning algorithms used for clustering. K-means partitions "n" observations in to k clusters during which every observation belongs to the cluster with the nearest mean.

- The main idea is to define k centroids, one for every cluster. Associate every point belonging to a given data set and to the nearest centroid. in any case the points in the data set are over, the first step is completed and an early grouping is finished. Re-calculate k new centroids as barycentre of the clusters ensuing from the previous step. when k new centroids has been calculated, a new binding has got to be done between a similar data set points and also the nearest new centroid. A loop has been generated. As results of this loop the k centroids modification their location step by step till no additional changes is done. This algorithm aims at minimizing an objective function, during this case a squared error function. The objective function is given as[7]

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i(j) - c_j||2$$

Algorithmic steps for k-means clustering:

Let X = {x1,x2,x3,……..,xn } be the set of data points and V = {v1,v2,…….,vc } be the set of centers.

1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data points and cluster center.
3. Assign the data points to the cluster center whose distance from the cluster center is least of all the cluster centers.
4. Recalculate the new cluster center using following formula:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_i$$

where, 'ci' represents the number of data points in ith cluster.

5. Again calculate the distance between each data points and new obtained cluster center.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

D. Classification:

Classification is a data processing task that predicts group membership for information in a given dataset. As for the classification, there are totally different academic objectives for using classification, such as: To group students who are hint-driven or failure-driven and find common misconceptions that students possess. A decision tree may be a flowchart-like tree structure, where every internal node, non leaf node, denotes a test on an attribute, every branch represents an out- come of the test, and every leaf node (or terminal node) holds a category label. Decision tree is thus popular and are used for classification in many application areas. It became therefore popular for many reasons: The development of decision tree classifiers does not require any domain information or parameter setting, decision trees will handle high dimensional information, easy to assimilate by humans, and therefore the learning and classification steps of decision tree induction are simple and quick. On the other angle decision tree classifiers have excellent accuracy [7].

- **Naive Bayes Algorithm:**

The performance of the student is predicted using data mining technique referred to as as classification rules. The Naïve Bayes classification algorithm is employed by the administrator to predict the performance of the student within the approaching semester based on their previous semester result and their behaviour. A Naïve Bayes classifier is a easy probabilistic classifier based on applying Bayes theorem from bayesian statistics with strong (naive) independence assumptions. A lot of descriptive term for the underlying probability model would be "independent feature model".

In easy terms, a Naïve Bayes classifier assumes that the presence or absence of a specific feature of a class is unrelated to the presence (or absence) of the other feature. For instance, a fruit could also be considered to be an apple if it's red, round, and about 4" in diameter. Even if these features depend upon one another or upon the existence of the opposite features, a Naïve Bayes classifier considers all of those properties to severally contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, Naïve Bayes classifiers can be trained very efficiently during a supervised learning setting. In several practical applications, parameter estimation for Naïve Bayes models uses the method of maximum likelihood; in different words, one will work with the Naïve Bayes model without believing in bayesian probability or using any bayesian strategies[7].

- **Steps of Naïve Bayes:**

Step 1: Scan the dataset (storage servers)

Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]

Step 3: Apply the formulae

$$P\left(\frac{ai}{vj}\right) = \frac{n\_c + m * p}{n + m}$$

Where:

- n = the number of training examples for which v = vj
- n_c = number of examples for which v = vj and a = ai
- p = 1/number of subject values
- m = the equivalent sample size [number of attributes]

Step 4: Multiply the probabilities by p

Step 5: Compare the values and classify the attribute values to one of the predefined set of class[7].

E. Decision Tree Algorithm:

Decision tree in data mining is one in all the best and easiest methods that are most often used by the researchers on their work. The basis node of the decision tree may be a top node resembles easy question also referred to as as a posture that bears multiple branches known as sub nodes with answers for the basis node question. Successively every answer associated with a set of queries or conditions that help us to predict the information, on that the final decision is made. ID3 and C4.5 are referred to as as induction algorithm of decision tree developed by the scientist known as Ross Quinlan.

Each algorithm supports greedy technique, top-down recursive in divide-and-conquer manner and that they don't support backtracking. C4.5 is additionally known as superset of ID3.

The advantages of this technique are, it doesn't need elaborated information, it deals with complicated information, these are easy to understand, and information Classification becomes easier, makes learning easier, it produces very accurate conclusion [8].

- **C4.5 Algorithm:**

The algorithm is given as follows [8]**:**

Step 1: Read trained data instances.

Step 2: Calculate Overall entropy:

$$\mathrm{Entropy}(t) = -\sum_{j} p(j|t)\log_2 p(j|t)$$

Step 3: Calculate entropy of every attribute:

$$M_i = \mathrm{Entropy}(N_i) = -\sum_{j} p(j|N_i)\log_2 p(j|N_i)$$

Step 4: Calculate information gain of every attribute:

$$\mathrm{Gain}_{\mathrm{split}} = \mathrm{Entropy}_{\,(p)} - \left(\sum_{i=1}^{k} \frac{n_i}{n}\mathrm{Entropy}(i)\right)$$

Step 5: Build Tree.

Step 6: build prune trees.

IV. PROPOSED SYSTEM

The Architecture of system is given in following figure. For the project we are collecting student's information from SSBT's college of engineering and technology which comes under North Maharashtra University. We pre-process the information we collected for deletion of information that doesn't required. Based on the rule, Students year down and backlog is being predicted. In proposed work we will use k-means clustering, Naive Bayes and C4.5 algorithms to predict student's failure. Accuracy of this classification algorithm is compared in order to check best performance. Student ranking is done on the basis of student's internal assessment and other student's academic related activities. The ranking of students will be decided by average percentage calculated and by sorting average percentage in descending order.

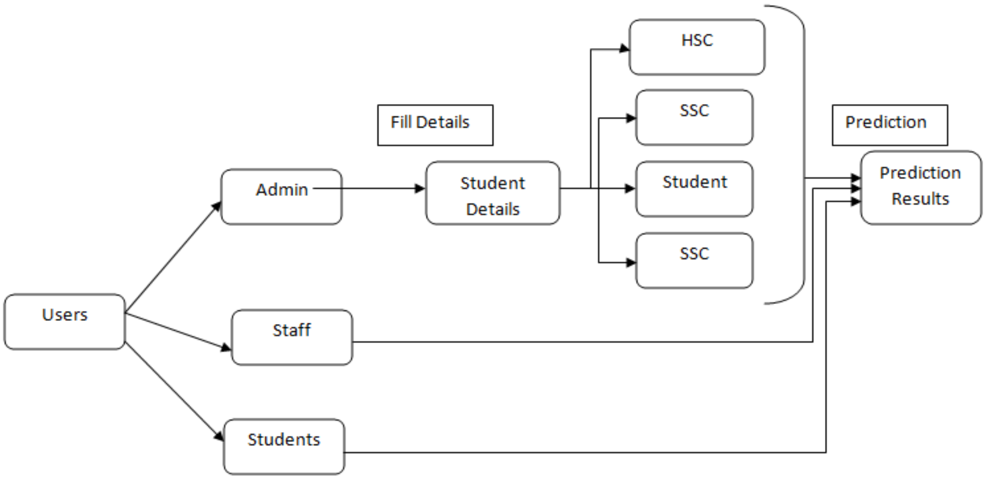


Fig. System Architecture

The data set used in this study is obtained from SSBT College of Engineering and Technology, Jalgaon. The result of previous students of different branches is collected from the college database and their behaviours are collected from their respective faculties and students. It is stored in other database and it is used to predict the performance of the present students in their next semesters. Following Table shows all the attributes of the students which are taken under observation in this system [8].

Variables	Description of Variables	Possible Values
Regular	Student is regular or not in class	Regular, Irregular
Hrs.	No of hrs spent on study	1, 2, 3, 4, >4
Interactive	Interactive in class or not	Yes, no
Study Materials	Types of books refer to study	Local books, Reference books
Library Visit	No of times student goes to library	2, 3 , >3
Book Type	Type of books issued	Technical, Novels, Sports
Grasping Ability	Ability to grasp knowledge	Poor, Good, Better, Best
Time Management	Ability of time management	Poor, Good, Better, Best
Decisions	Ability to make right decisions	Poor, Good, Better, Best
Sports	Participated in any sport	Yes, No (if yes then name of sport)
Audit points	Collect audit points on time	Yes, No
Extracurricular activities	Participate in extracurricular	Yes, No
Result Record	Previous semester results	A+-85-100 CGPA-10 A – 74-84 CGPA-9 B – 64-73 CGPA-8 C – 55-63 CGPA-7 D – 47-54 CGPA-6 E – 40-46 CGPA-5 F – Less than 40
SSC	Result of 10th	Distinction:85-100% First:60-84% Second:45-59% Third:35-44% Fail: Less than 35%
HSC	Result in 12th	Distinction:85-100% First:60-84% Second:45-59% Third:35-44% Fail: Less than 35%
Faculty Guidance	Take guidance from faculty	Good, Better, Best
Higher study	Interested in higher studies	Yes, No

Table: Required Data.

- **K-means Algorithm:**

Table: Performance Index

70 or above	Distingtion
60-69	First class
50-59	H. second
45-49	Second class
40-45	Pass class
Bellow 45	Fail

Cluster	Cluster size	performance
1	25	62.22
2	15	45.73
3	29	53.03

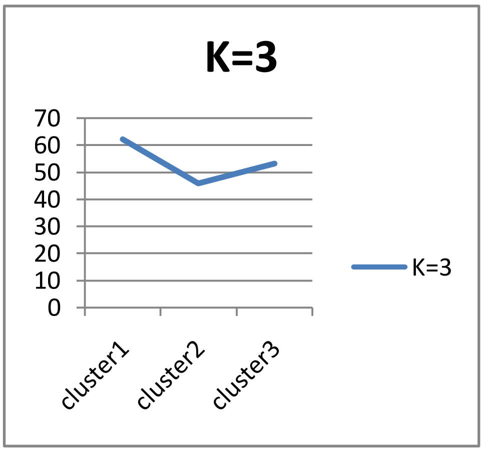


Table: K=4

Cluster	Cluster size	performance
1	24	50.08
2	16	65.00
3	30	58.89
4	9	43.65

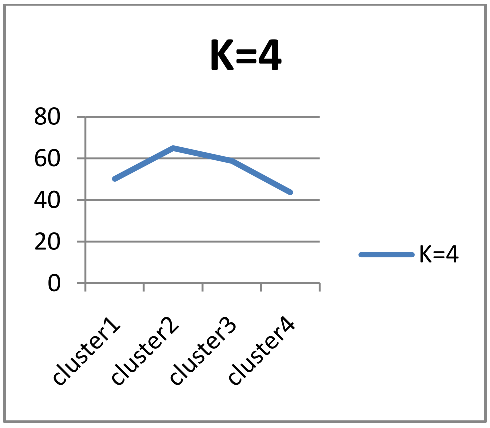
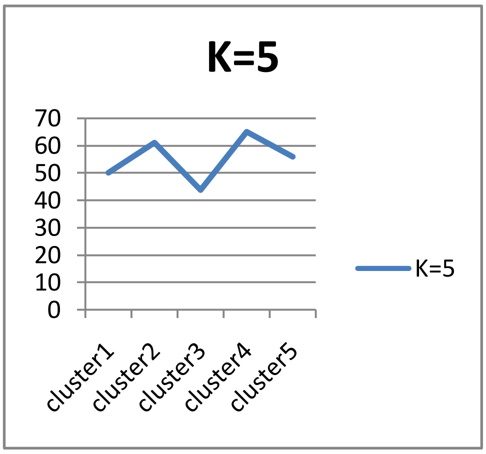


Table: K=5

Cluster	Cluster size	performance
1	19	49.85
2	17	60.97
3	9	43.65
4	14	64.93
5	20	55.79



- **Naive Bayes algorithm:**

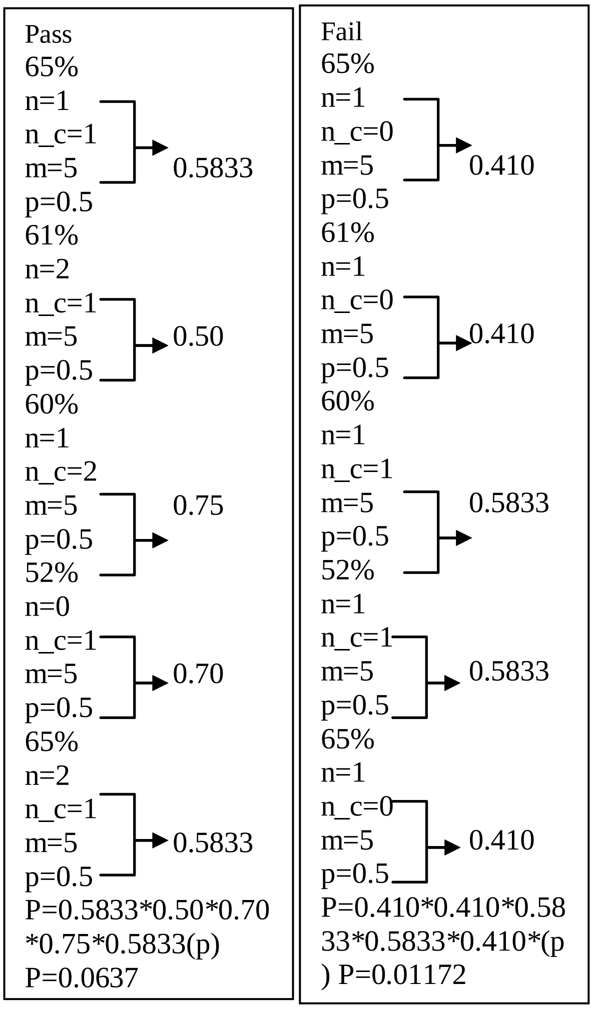
Attributes: 10th, 12th, FE, SE, TE. Result= pass or fail [p=1/2=0.5]
Training Data Set:-

Name	10th	12th	FE	SE	TE	result
Kalpesh	89%	61%	73%	72%	75%	Pass
Riya	82%	70%	74%	60%	72%	Pass
Shreya	60%	50%	51%	52%	39%	Fail
Rajeshwari	65%	55%	60%	70%	65%	Pass

New Student:- Mayur[65%,61%,60%,52%,65%]=?

$$P = \frac{n\_c + m * p}{n + m}$$

0.0637>0.01172
Hence the new student is predicted to be Pass.



Pass
65%
n=1
n_c=1
m=5 0.5833
p=0.5
61%
n=2
n_c=1
m=5 0.50
p=0.5
60%
n=1
n_c=2
m=5 0.75
p=0.5
52%
n=0
n_c=1
m=5 0.70
p=0.5
65%
n=2
n_c=1
m=5 0.5833
p=0.5
P=0.5833*0.50*0.70
*0.75*0.5833(p)
P=0.0637

Fail
65%
n=1
n_c=0
m=5 0.410
p=0.5
61%
n=1
n_c=0
m=5 0.410
p=0.5
60%
n=1
n_c=1
m=5 0.5833
p=0.5
52%
n=1
n_c=1
m=5 0.5833
p=0.5
65%
n=1
n_c=0
m=5 0.410
p=0.5
P=0.410*0.410*0.58
33*0.5833*0.410*(p
) P=0.01172

V. CONCLUSION

This paper, build a review on different students' performance supported data mining techniques under different circumstances. From this review, we will improve the speed of predicting the result. Within the previous study lot of your time taken in training the dataset i.e., over 60 minutes time in training and also the rest in testing. If we have a tendency to reduce the time in training, it greatly improves speed. The step taken to reduce the training dataset could also be by using clustering techniques. And additionally to enhance the standard of students' performance to create an early prediction with some classification techniques and ensemble clustering can even be used for the same. As an enhanced work of proposed system the important time information from any reputed colleges or from universities will be used for the betterment of effective result. The below mentioned table showed some of the actual accuracy rate predicted in previous study by using some classification techniques.

Algorithm	**Accuracy**
K-means clusturing	94%
Naive Bayes Algorithm	96%
C-means Algorithm	95%

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] R. R. Kabra, R. S. Bichkar, "Performance Prediction of Engineering Students using Decision Trees", International Journal of Computer Applications, Volume 36– No.11, December 2011.

DOI10.17148/IJARCCE.2017.63195