Educational Data Mining : A Review on Student Academic Performance Prediction Using Data Mining Methods

Kolinben Rakeshkumar Sukhadia

Depatrtment of Electrical and Computer Engineering, Lakehead University, 955 Oliver Road, ThunderBay,Ontario,Canada

ABSTRACT

Predicting the performance of student is very important in higher education management. The objective of this report is to accurately predict student performance by applying various data mining techniques such as classification, clustering, regression and association rule mining on student dataset. From last 10 years, there has been a fast development in education system which prompts huge amount of data. Because of this enormous amount of data, predicting student performance is more difficult. Analysts and researchers apply various data mining techniques namely classification, clustering, regression methods and association rule mining. By applying ensemble methodology on student dataset, it gives better classification accuracy and low errors. By tuning parameter of support vector machine such as kernel and penalty to proper value, it helps in increasing classification accuracy and reduce errors such as Root mean square error(RMSE), Mean Absolute Error(MAE), Relative Absolute Error(RAE). Type of school is not most influenced attribute to influence student performance. Other attribute such as parent’s occupation has huge impact on predicting student academic performance. By monitoring student’s performance semester by semester, it helps to improve student academic result prediction. IBM statistical package for social studies(SPSS) is used to apply chi-square automatic interaction detection(CHAID) in producing decision tree structure. Algorithm produce more accurate result if higher version of SPSS is used. Ensemble method is used to improve student performance classification accuracy.

Keywords: Educational data mining, student database, classification, clustering, association rule mining, Evaluation metrics

I.INTRODUCTION

The objective of this study is to to accurately predict student performance by applying various data mining techniques such as classification, clustering, regression and association rule mining on student dataset. Data mining is most powerful methodology to analyze useful information from the data warehouse. In data mining, there is prediction of hidden information by doing extraction.data mining is multidisciplinary field which includes various areas such as making learning from dataset, statistics, information technology, visualization of data and artificial intelligence. Educational data mining uses different data mining tools and methods for predicting student performance. Application of data mining in education is called educational data mining. There is extraction and interpretation of the raw data from education system which have huge impact on student retention rate, university success rate, and student performance. To make institutions provide best education to students, Educational data mining helps in predicting the future patterns. Data can be collected from many sources namely from online learning management system, and database of academic institutes. In data mining, there are different methodology to process the information such as classification, clustering and regression methods. For predicting student performance, there are various attributes such as student's academic year, student's previous semester GPA, father's education, mother's education, free time after school, quality time spent with family, weekly study time, time spent going out with friends. Research objectives, such as gaining a deeper understanding of the teaching and learning phenomena,identifying weaker students at an early stage, recommending them extra module or course depending on student's performance, identifying the factors affected to student success the most and analysing student behaviour in e-learning system. The major objectives of educational institution is to increase the university success rate and student success rate.

Learners, educators, researchers, and administrators are users in educational data mining.learners. Education data mining goals are student modelling, predicting students performance and learning outcomes, generating recommendations, analyzing learner's behaviour and communicating to stakeholders.

EDM ENVIRONMENT

two types of educational environments , traditional education and Computer based education. Before prediction, data can be pre-processed in proper ways .

offline data: From real time situation and settings, offline data is produced. It is also produced through traditional education, where knowledge transfer to students is based on face to face contact. Teacher-student interactions, student-to-student interactions, traditional classroom tests, participation in various activities by students attendance, data derived from various courses.

Online data: From weblogs, online data is derived. Emails, Elearning, Learning management system, intelligent tutoring system, adaptive educational hypermedia system and publication databases.

EDUCATIONAL DATA MINING METHODS

1.1 Classification and regression
1.2 Clustering

1.3Relationship mining

1.1 Classification and regression

Regression predicts continuous variables. It is different from classification technique. Linear regression and neural network are regression techniques. Various parameters nemely age, gender attendance, family income, occupation, qualification can be used as predictors.

Classification is supervised learning technique. Dataset is divided into training dataset and testing dataset. In this method, classifying data is based on training set and utilizes that patterns to introduce new dataset.

1.2 Clustering

There is grouping of similar records. It is unsupervised method which focus on high-dimensional data. For clustering, K-means is the most used method. Based on student’s learning style such as visual, aural and kinaesthetic, clustering can be used to group students.

1.3 Relationship mining

Relationship mining is there to discover relationship between various variables in dataset. Association rule mining is the most commonly used educational data mining method.Association rules in educational data mining are used to determine strong association rules from educational databases using support and confidence as the predefined measures. Apriori method is the most widely used methods in association mining.

Components of Educational Data Mining(EDM) : The major components of educational data mining are educational data, educational environment, stakeholders of the education system and data mining tools and techniques. Learners, faculties, parents, course researchers and educational developers and administrators are components of educational data mining. Figure 1 shows process of data mining in education data mining

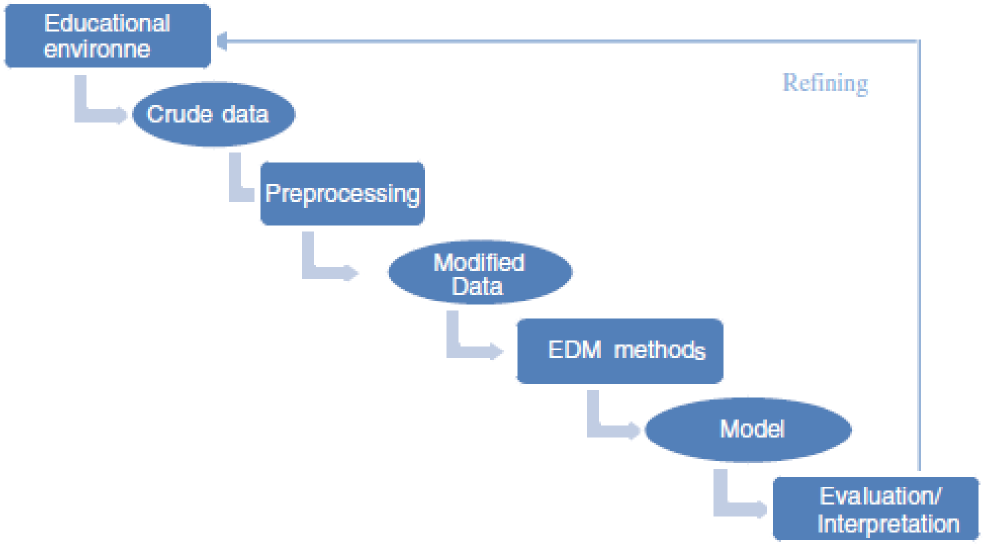


Figure 1. methodology of data mining in education data mining(Oyelade et al., 2010)

The process starts with collecting data related to educational environment. After that, data preprocessing technique is applied. In data preprocessing, there are heterogeneous data fusion, treatment of missing and incorrect values, converting the data to an appropriate form and feature selection. After that, proper Educational Data Mining(EDM) method such as classification, clustering and regression method is applied.

Objective of this report is to do survey of various data mining methods on educational student database. The main objective is to increase accuracy of student performance prediction and reducing errors.

In this report, I reviewed 30 journal papers on student academic performance prediction using data mining. I have focused more on predicting accuracy better and reducing errors such as Root Mean Square Error(RMSE), Relative absolute Error(RAE), Mean Absolute Error. For better prediction it is necessary to have good student performance prediction accuracy.

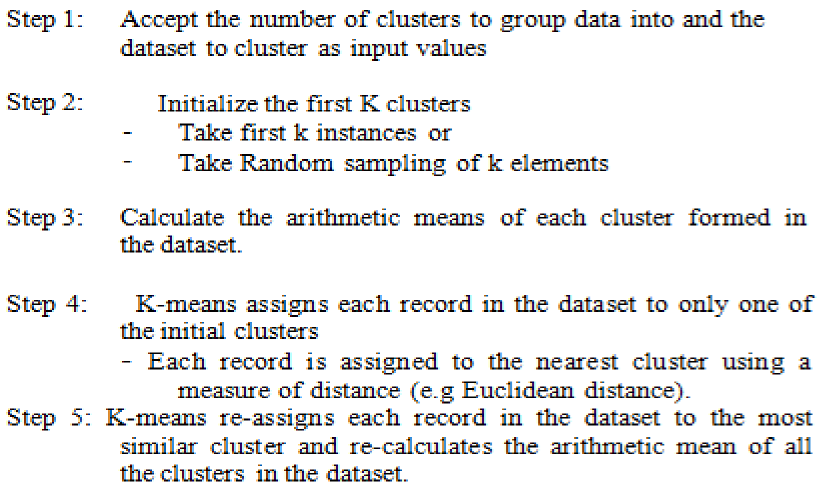
In next section, in the literature review, I have done survey on student academic performance using data mining method. This is educational data mining so various classifications, clustering and regression techniques are applied on student dataset. In technical discussion section, the performance of the various data mining algorithms will be compared and research gap is found. In the conclusion section, the summary of review would be there. There also merits and demerits of various data mining techniques would be there.

II.LITERATURE REVIEW

Data mining is most powerful methodology to analyze useful information from the data warehouse. In data mining, there is prediction of hidden information by doing extraction.data mining is multidisciplinary field which includes various areas such as making learning from dataset, statistics, information technology, visualization of data and artificial intelligence. Educational data mining uses different data mining tools and methods for predicting student performance. Application of data mining in education is called educational data mining. There is extraction and interpretation of the raw data from education system which have huge impact on student retention rate, university success rate, and student performance. To make institutions provide best education to students, Educational data mining helps in predicting the future patterns. The process starts with collecting data related to educational environment. After that, data preprocessing technique is applied. In data preprocessing, there are heterogeneous data fusion, treatment of missing and incorrect values, converting the data to an appropriate form and feature selection. After that, proper Educational Data Mining(EDM) method such as classification, clustering and regression method is applied.

[1] Oyelade et al.(2010) described K-means clustering algorithms for prediction of stufdent academic performance. Distance and centroid computations are performed before the k-means algorithm converges, and loops are executed a number of times, say l, where l is the number of k-means iterations. Even within the same dataset, the precise value of l changes depending on the initial beginning cluster centroids. As a result, the algorithm's computational time complexity is O(nkl), where n is the total number of items in the dataset, k is the required number of clusters, and l is the number of iterations, kn, ln.

Figure 2 shows pseudocode for k-means algorithm.



```
Step 1:   Accept the number of clusters to group data into and the
          dataset to cluster as input values

Step 2:   Initialize the first K clusters
          -   Take first k instances or
          -   Take Random sampling of k elements

Step 3:   Calculate the arithmetic means of each cluster formed in
          the dataset.

Step 4:   K-means assigns each record in the dataset to only one of
          the initial clusters
          - Each record is assigned to the nearest cluster using a
              measure of distance (e.g Euclidean distance).
Step 5: K-means re-assigns each record in the dataset to the most
          similar cluster and re-calculates the arithmetic mean of all
          the clusters in the dataset.
```

Figure 2. Psuedocode for k-means algorithm(Oyelade et al., 2010)

Students are classified in Excellent, very good, good, very fair, fair and poor categories. Table 1 shows studenyt category and its range. If student score is 70 and above, it is categorized as excellent. Students were classified as very good, good, very fair, fair and poor if their grade ranges from 60-69, 50-59, 45-49, 40-45 and below 45 respectively.

Table1. Performance index(Oyelade et al.,2010)

70 and above	Excellent
60-69	Very good
50-59	Good
45-49	Very fair
40-45	Fair
Below 45	Poor

Table 2 shows result for k=3. There are different cluster sizes and according to that, overall performance are different. Overall performance for cluster size 25 is 62.22. Overall performance for cluster size 15 and 29 are 45.73 and 53.03 respectively.

Table 2 . cluster size and overall performance for k=30(Oyelade et al., 2010).

For k=3,

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

Table 3 shows result for k=4. There are different cluster sizes and according to that, overall performance are different. Overall performance for cluster size 24 is 50.08. Overall performance for cluster size 16,30 and 9 are 65.00, 58.89 and 43.65 respectively.

Table 3. cluster size and overall performance for k=4(Oyelade et al., 2010).

For k=4,

Cluster	Cluster size	Overall performance
1	24	50.08
2	16	65.00
3	30	58.89

This paper(Oyelade et al., 2010) uses dataset of 79 students and produces numrerical interpretation of results for performance evaluation and by tuning number of clusters and cluster size, authors tried to get better overall performance ;however, paper(Md. Hedayetul et al., 2012), same method k-means is used but there is no tuning of hyperparameters such as number of clusters and cluster size. By taking internal assessment and prevoious exam grade, performance prediction is done.

[2] Md. Hedayetul et al.(2012) proposed prediction of student academic performance by using k-means clustering algorithm. Data clustering is used to deevelop relationship among variables of large dataset. It is unsupervised and statistical data analysis method. It is used to operate on a huge dataset to uncover hidden patterns and relationships, allowing quick and efficient decision-making. Student’s university academic performance is measured by internal and external assessment. Class test marks, lab performance, assignment, attendance and quiz are considered as internal assessment, while final semester grade and previous semester grade are taken as

external assess,ment. By using b Student's evaluation factors are Class quizzes, mid and final exam assignment. The main objective of clustering method is to partition students into groups according to their abilities and characteristics. Many factors considered while evaluating student's academic performance. Proposed method is k-mean clustering to predict student academic performance. Here is description of proposed methodology. By taking internal assessment and prevoious exam grade and by using data clustering technique, there is prediction of final grade of student. If quiz=good, assignment=complete, lab-performance=good, class-=test=good, attendance=regular and prev-grade=high, then student's final grade will be good. If quiz=good, assignment=incomplete, lab-performance=good, attendance=regular, prev-grade=average and class-test=average then final grade is average. Student's final grade is low, if parameters value prev-grade=low, quiz=average, assignment=incomplete, lab-performance=poor, mid-term=low and attendance=irregular. k-means method works as follows:

- Number of clusters to group data
- Initialize the first k clusters, take first k instances or take random sampling of k elements.
- After that, there is calculation of arithmetic means of each cluster formed in dataset.
- K-means assigns each record in data to only one of initial clusters.
- By using euclidean distance, each record is assigned to nearest cluster using measure of distance.

Proposed K-means re-assigns each record in data to the most similar cluster and recalculates the arithmetic mean of all the clusters in he dataset. There are two class labels passed and failed. If grade is above 2.20, then students are categoqized into passed and if grade is less than or equal to 2.20, then students are categorized into failed. There is cluster of student among their GPA. There are 8.33% student whose GPA is 2.00-2.20. There are 16.67% student whode GPA is 2.20-3.00. There are 28.33% student whose GPA is from 3.00-3.32. From 3.32-3.56 GPA, percentage is 25%. The student percentage is 21.67% between GPA 3.56-4.00.

Figure 3 represents graph of GPA and the percentage of student(Md. Hedayetul Islam Shovon and Mahfuza Haque, 2012)

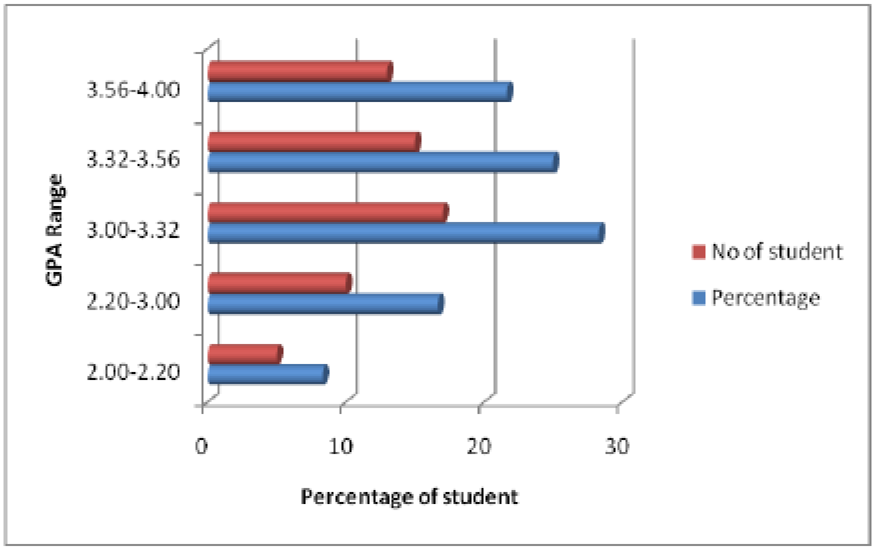


Figure 3. Number and percentage of students regarding to GPA(Md. Hedayetul Islam Shovon and Mahfuza Haque, 2012)

Table 4 represents different classes such as high, medium and low. According to those, students are categorized.

Table 4: Number of student percentage and their GPA based on class categories(Md Md. Hedayetul Islam Shovon and Mahfuza Haque, 2012)

Class	GPA	No. of student	Percentage
High	>=3.50	28	46.67
Medium	2.20<=GPA<3.5	27	45
Low	<=2.20	5	8.33

Figure 4 is graphical representation of percentage of students getting high, medium and low GPA. 46.47% students getting GPA>=3.50, 45% students getting GPA between 2.20 to 3.5 and 8.33% student is having GPA<=2.20.

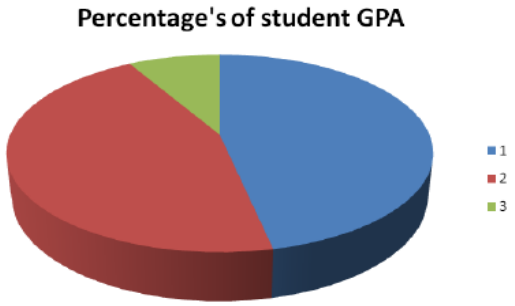
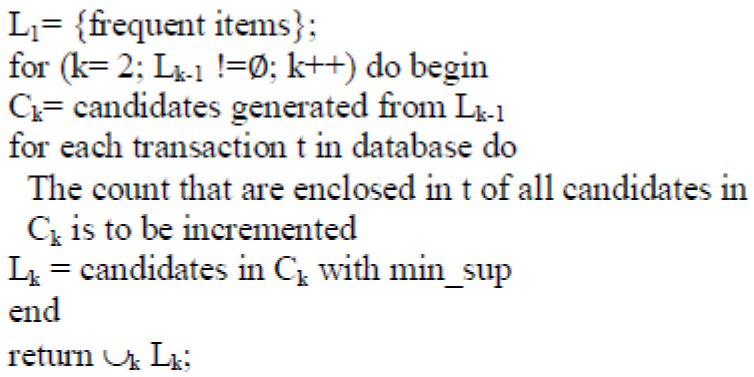


Figure 4. Graphical representation of percentage of students(Md. Hedayetul Islam Shovon and Mahfuza Haque, 2012)

In this research(Md. Hedayetul et al., 2012), k-mean clustering data mining method is used to predict student academic performance. In this method, there is no rules generation for student performance prediction; however, paper(D. Magdalene Delighta Angeline, 2013) studied student performance analysis using Association rule generation for student performance analysis using apriori algorithm. There were rule generation in apriori method and based on that student performance is evaluated.

[3]D. Magdalene Delighta Angeline(2013) studied student performance analysis using Association rule generation for student performance analysis using apriori algorithm. This paper uses association rules to extract the student performance with apriori methodology. In first pass, there is determination of item occurences to obtain large 1- itemsets. Next pass K contains two phases. In first phase, large itemsets found in the (k-1)th pass are used to generate candidate itemsets $C_{K.}$ After that, it is important to determine the candidates in C_{K} .Dataset used in this study was taken from department of computer science, Dr. G.u .Pope College of Engineering from year 2011 to 2012. Attributes of dataset are programme, duration, register number, programme code, Gender, Address, City1, City2, Percentage, assignment mark, correct response, self confidence, assignment submission, financial lack, parental education. Target variable is analysis report. Good, Average and Poor are three categories of target variable analysis report.

Figure 5 represents student academic performance prediction using apriori algorithm.



```
$L_1$= {frequent items};
for ($k$= 2; $L_{k-1}$ !=$\emptyset$; k++) do begin
$C_k$= candidates generated from $L_{k-1}$
for each transaction t in database do
  The count that are enclosed in t of all candidates in
  $C_k$ is to be incremented
$L_k$ = candidates in $C_k$ with min_sup
end
return $\cup_k L_k$;
```

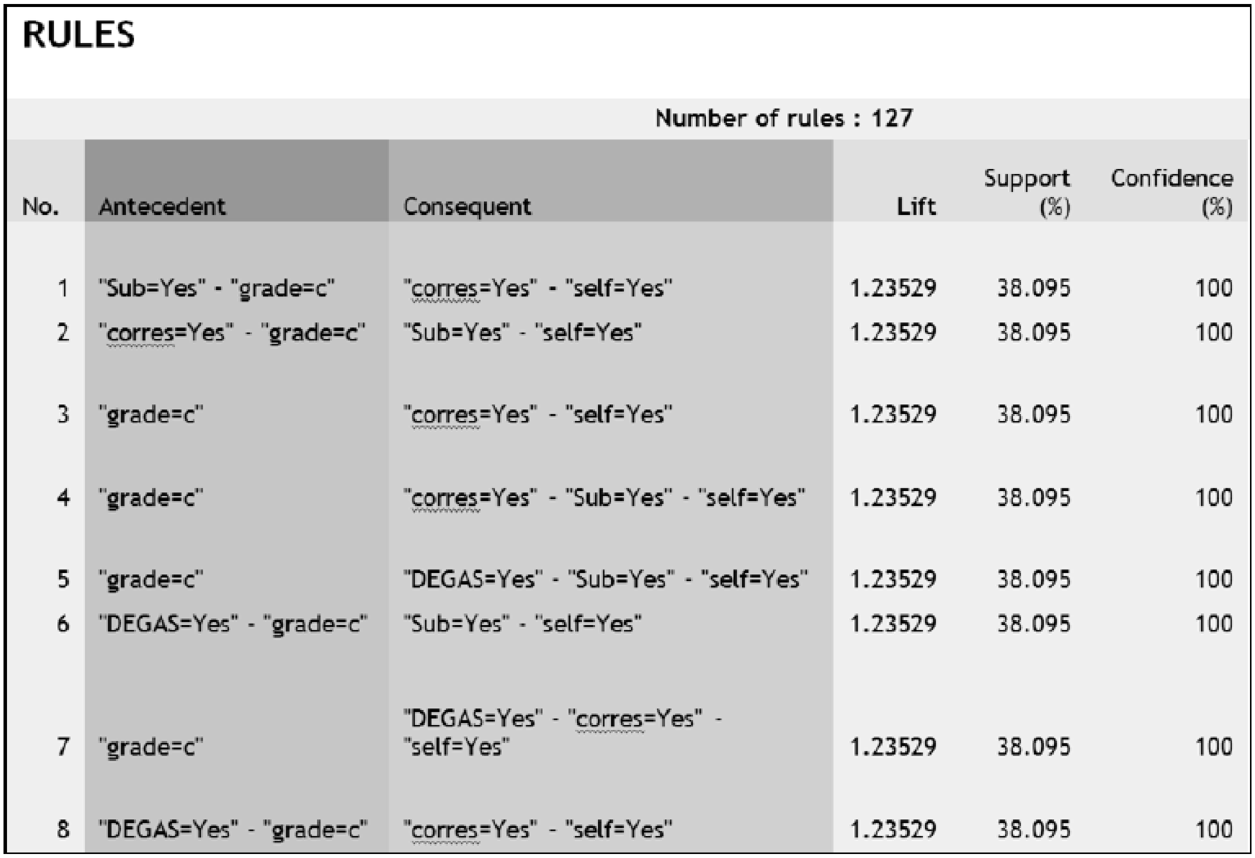
Figure 5. apriori algorithm(D. Magdalene Delighta Angeline, 2013)

Table 5 shows apriori parameters used in methodology such as support minimum, confidence minimum, max rule length and lift filtering.

Table 5. apriori parameters used in methodology(D. Magdalene Delighta Angeline, 2013)

Support minimum	0.33
Confidence minimum	0.75
Max rule length	4
Lift Filtering	1.1

Figure 6 shows rules generated using apriori methodology.



RULES

Number of rules : 127

No.	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"Sub=Yes" - "grade=c"	"corres=Yes" - "self=Yes"	1.23529	38.095	100
2	"corres=Yes" - "grade=c"	"Sub=Yes" - "self=Yes"	1.23529	38.095	100
3	"grade=c"	"corres=Yes" - "self=Yes"	1.23529	38.095	100
4	"grade=c"	"corres=Yes" - "Sub=Yes" - "self=Yes"	1.23529	38.095	100
5	"grade=c"	"DEGAS=Yes" - "Sub=Yes" - "self=Yes"	1.23529	38.095	100
6	"DEGAS=Yes" - "grade=c"	"Sub=Yes" - "self=Yes"	1.23529	38.095	100
7	"grade=c"	"DEGAS=Yes" - "corres=Yes" - "self=Yes"	1.23529	38.095	100
8	"DEGAS=Yes" - "grade=c"	"corres=Yes" - "self=Yes"	1.23529	38.095	100

Figure 6. Rules generated using Apriori method(D. Magdalene Delighta Angeline, 2013)

[4] V.Ramesh et al.(2013) described which parameter from student database has huge influence on student performance. For this study, dataset was collected from higher secondary students. The district is kancheepuram. There were 900 samples of data. There are various attributes such as beloew in table

Table 6 shows values of student database attributes

Table 6 . student database attributes(V.Ramesh et al.,2013)

PS	Parental status. It indicates parental status of students
Mtts	Mode of transportation to school
Comm	Community. Values are like scheduled cast, scheduled tribres, most backwardv classes, others

PTUI-SEC	Private tution at secondary level
X-GRA	Marks obtained at secondary level
GRP-HSC	Types of group study
TOS	Type of school
PTUI-HSEC	Private tution at higher secondry level
HSCGRADE	Marks obtained at higher secondary level

Table 7 shows various classification algorithm performance, where naïve bayes gives 49.5% accuracy, multilayer perception gives 72.38% accuracy, SMO gives 57.25% accuracy, and J48 gives 64.88% accuracy.

Table 7. comparison of classification performance(V.Ramesh et al.,2013)

Naïve bayes	Multilayer perception	SMO	J48	REPTREE
49.5%	72.38%	57.25%	64.88%	60.14%

Table 8 shows testing hypothesis and its results.

Table 8. testing hypothesis(V.Ramesh et al.,2013)

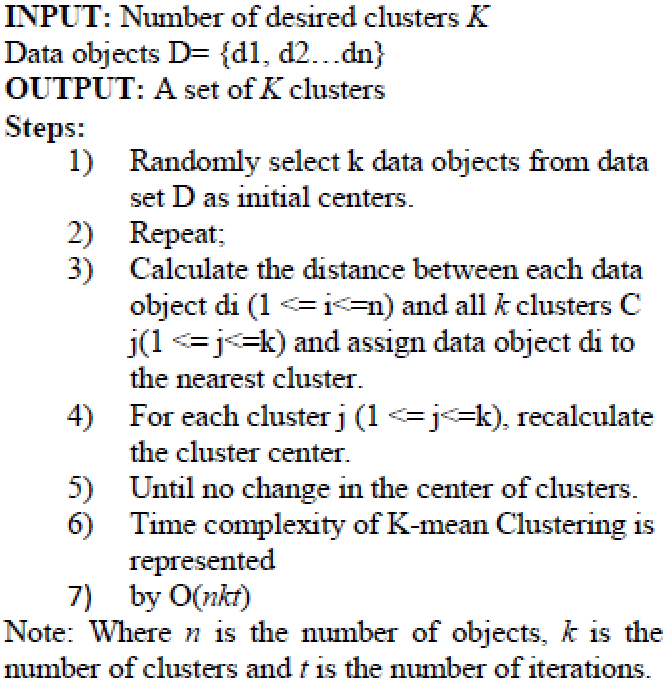
In testing hypothesis,

Hypothesis	Result
Type of school is not influencing grade obtained	Accepted
Private tution is not influencing grade obtained	Rejected
Study at home is not influencing grade obtained	Rejected
Parent education is not influencing grade obtained	Rejected

This paper(V.Ramesh et al., 2013) has done work on feature importance. It finds parameters from student dataset which has more influence on performance prediction; however, research(M.Durairaj and C. Vijitha, 2014) done on student dataset did not focus on feature importance.

[5] M.Durairaj and C. Vijitha(2014) described clustering algorithm for prediction of student performance.The dataset contains students detail of different subject marks semester wise. Clustering is a method which divide data into group of similar objects. Clustering plays crucial role in scientific data exploration, information retrieval, text mining, spatial database applications, web analysis marketing. Data mining methodologies namely k-mean clustering and naïve bayes clustering is proposed in this research to predict student academic performance.

Figure 7 depicts steps of k-mean clustering algorithm. In k-mean clustering methodology, k data elements are choosen as initial clusters. Distances of all data elements are calculated by euclidian distance formula. The data components that are closer to the centroids are transferred to the relevant clusters.

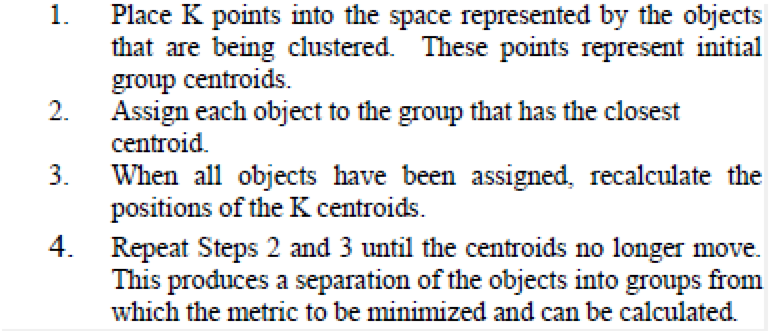


```
INPUT: Number of desired clusters K
Data objects D= {d1, d2…dn}
OUTPUT: A set of K clusters
Steps:
    1)  Randomly select k data objects from data
        set D as initial centers.
    2)  Repeat;
    3)  Calculate the distance between each data
        object di (1 <= i<=n) and all k clusters C
        j(1 <= j<=k) and assign data object di to
        the nearest cluster.
    4)  For each cluster j (1 <= j<=k), recalculate
        the cluster center.
    5)  Until no change in the center of clusters.
    6)  Time complexity of K-mean Clustering is
        represented
    7)  by O(nkt)
Note: Where n is the number of objects, k is the
number of clusters and t is the number of iterations.
```

Figure 7. proposed method k-mean clustering steps(M.Durairaj and C., 2014)

The design of a system necessitates a thorough understanding of the problem domain. In an IT IN industry, knowledge engineering is used to determine the datasets and input qualities.

Another method such as naïve bayes clustering is used for student academic performance prediction. Benefit of naïve bayes is that it requires a small amount of training data to estimate parameters. The value of a feature is assumed to be unrelated to the presence or absence of any other feature by a naïve bayes classifier. The naive bayes classifier considers each of these features to contribute independently of the other features' existence or absence. Figure 8 represents proposed data mining naïve bayes method steps. In naïve bayes clustering methodology, for each class, variances of variables need to be determined.



1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized and can be calculated.

Figure 8. proposed methodology naïve bayes method steps(M. Durairaj and C. Vijitha, 2014)

The tool weka 3.7.9 is used to implement data mining approach. It is collection of open source of data mining algorithms, preprocessing on data, classification, clustering and association rule extraction. It is java based open source tool.

Figure 9 represents weka open source tool which is java based open source tool.

Waikato Environment for Knowledge Analysis
Version 3.7.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure 9. Weka open source implementation tool(M. Durairaj and C. Vijitha, 2014)

The probability model for a classifier is a conditional model

P(C|F1,……Fn) There is condition on several feature variables F1 through Fn. There is a dependent class variable C. There is calculation of cluster instances by using K-mean clustering algorithm. Performance measures such as sensitivity, specificity and precision is used to predict student's academic performance.

Sensitivity = $\frac{t-pos}{pos}$

Specificity =$\frac{t-neg}{neg}$

Precision = $\frac{t-pos}{t-pos_{f}-pos}$

Where t-pos is number of true positives tuples that were correctly classified, pos is the number of positive tuples, t-neg is the number of true negative tuples which are correctly classified. Neg is number of negative tuples and f-pos is the number of false positive tuples. Another evaluation metric namely accuracy is also defined

Accuracy = Sensitivity$\frac{pos}{pos+neg}$+Specificity$\frac{neg}{pos+neg}$

Precision is calculation of positive predicted values, which is the fraction of retrieved documents which are relevant. It is formula is given as:

Precision = $\frac{|\{relevant\ documents\}^\wedge\{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$

Precision = $\frac{tp}{tp+fp}$

Recall is the fraction of the documents which are relevant to the query. Recall formula is given as below.

Recall = $\frac{|\{relavant\ documents\}^\wedge\{retrieved\ documents\}|}{relavant\ documents}$

Recall = $\frac{tp}{tp+fn}$

F Measure = It is harmonic mean of precision and recall. It is calculated using following formula.

F-Measure = $2\,\frac{precision*Recall}{precision+Recall}$

Table 9 demonstrates the comparison of weighted average for decision tree and naïve bayes. As a result, naïve bayes method gives more accurate result than decision tree.

Table 9. Comparison of Weighted Saverage for various techniques(M. Durairaj and C. Vijitha, 2014)

Technique	TP rate	FP rate	Precision	Recall	F-Measure	MCC
Decision tree	0.947	0.474	0.922	0.947	0.934	0.660
Naïve Bayes	1.000	0.000	1.000	1.000	1.000	1.000

It is concluded that, TP rate for decision tree is 0.947 and naïve bayes is 1.000. So, naïve bayes performs better than decision tree on student dataset.Clustering methodolologies such as k-mean clustering and naïve bayes clustering are applied on student dataset(M.Durairaj and C. Vijitha, 2014); however, author have done great experiment by applying classification method decision tree on student dataset and classify students based on four cases(Abeer et al.,2014).

[6] Abeer et al.(2014) describes educational organizations are one of the important parts of society. Author proposed classification methods for student performance prediction. Dataset was

taken from information system department of educational institution. It was from year 2005 to 2010. The size of data are 1548 recpords.In this proposed methodology, decision tree is used for prediction. Decision tree structure is like tree structure. In this methodology, internal node is depicted by rectangles and by ovals leaf nodes are represented. In ID3 method, there is construction of decision tree by using top-down, greedy search. It is important to select most influencing and most useful attributes. It is selected by using metric-information gain. There is another concept called information gain in decision tree method. Splitting criteria is used for splitting of tree nodes. Splitting criteria used for splitting of nodes is called information gain.

Table 10 represents dataset values, its description and its possible values. Dataset values are department of students, high school degree of students, lab test grade, assignment, seminar performance, attendance, homework and final grade marks. Knowledge represented by decision tree can be explained by rules.

Table 10. Student related variables(Abeer et al., 2014)

Variable	Description	Possible Values
Dep	Department of students	Literary, Scientific, Mathematics, Secondary, Industrial Technical, Commercial
HSD	High school degree of students	Good, Acceptable
Midterm	Midterm marks	Excellent >= 85%, Very good >=75 & <85%, good >= 65 & <75% Acceptable >=50 & <65% Fail <50%
LG	Lab Test Grade	Poor, average, good
SEM	Seminar Performance	Poor, average, good
ASS	Assignment	Yes, no
SP	Measure of student participate	Yes, no
ATT	Attendance	Poor, average, good

HW	Homework	Yes, no
FG	Final grade marks	

Case 1 - If Midterm Mark = Excellent, Lab Test Grade = Good, Student Participate = No, Homework = No, Seminar Performance = Good, Department = Scientific Mathematics then Final Grade = Very Good.

Case 2 – If Midterm Marks = Excellent, Lab Test Grade =Good, Student Participate = No, Attendance = Good, Homework = No, Department = Secondary Technical Commercial then Final Grade = Very Good.

Case 3 – If Midterm Marks = Excellent, Lab Test Grade =Good, Student Participate = No, Attendance = Good, Homework = No, Department = Secondary Technical Commercial then Final Grade = Very Good.

Case 4.- – If Midterm Marks = Excellent, Lab Test Grade =Good, Student Participate = No, Attendance = Good, Homework = No, Department = Secondary Technical Commercial then Final Grade = Very Good.

This paper(Abeer et al., 2014) used classification method decision tree; however, paper(Samy S. Abu-Naser et al., 2015) used artificial neural network method for student performance prediction and it performs better than proposed methodology in paper(Abeer et al., 2014).

[7] Samy S. Abu-Naser et. al(2015) described how to predict sophomore student performance using artificial neural network. Objective of this study are to identify suitable factors that affect a student performance,to convert these factors into appropriate forms for an adaptive system coding, and to predict student performance by applying an artificial neural network. A neural network is made up of a group of artificial neurons that work together to interpret data. ANN is an adaptive system that alters its structure in response to external or internal data. Neural networks represent complex interactions between inputs and outputs to find data patterns. Layers are collection of neurons with similar functionality. Input layer, hidden layer and output layers are there in ANN. Layer of neurons that receives user programme input is called input layer. The output layer is the layer of neurons that sends data to the user programme. Hidden layers exist between the input and output layers. In a neural network, every neuron has the ability to influence processing. Any layer of the neural network can be used for processing.Input variables are obtained from student file and registrar system. High school score, results in math 1 , results

in math 2 in the student freshman year, results in electrical circuits and electronics 1 in the student freshman year, number of credits passed, CGPA of the freshman year, type of high school(public or private), high school location and student's gender are input variables. Output variable is cumulative grade point average.output variable is classified in excellent, very good, good and poor. Performance of student is represented by output variable. In design of neural network, multilayer perceptron neural network is used using linear sigmoid activation function.

Figure 10 shows artificial neural system architecture. There are input layer, hidden layer and output layer. Moreover, there are input signals and error signals are there

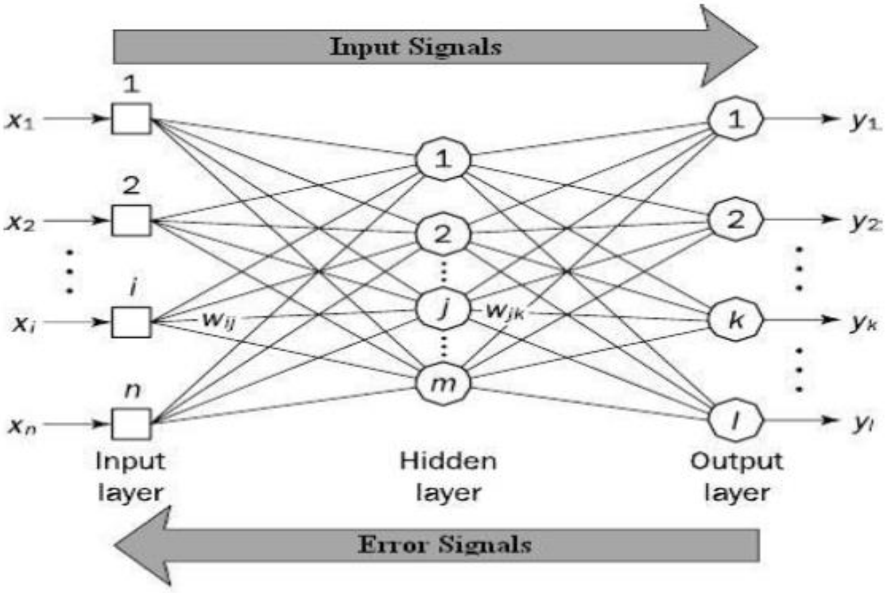


Figure 10. Architecture of Artificial neural network(Samy S. Abu-Naser et. al,2015)

Table 11 depicts input data transformation. Various input variables are high school score, math I, Math II, Electrical Circuit I, CGPA of freshman year

Table 11. Input data transformation.

Input Variable	Domain
High school score	Above 80% 75-79 70-74

MathI	Above 85% 75-84 60-74
Math II	Above 85% 75-84 60-74
Electrical Circuit I	Above 85% 75-84 60-74
Electronics I	Above 85% 75-84 60-74
Number of credits passed	32 hours 26-31 24-25
CGPA of freshman year	Above 90%, 80-89,70-79,65-69
Zone of high school attended	Palestine, outside of palestine
Type of high school	Public private
Gender	Male, female

There is evaluation to predict sophomore student performance in the faculty of engineering and information technology. In this paper, there are 10 inputs(input layer) , 6 inputs(hidden layer) and 4 outputs(single output layer). As training set, 60 % of total data are used, and for testing 30 % data used. For cross validation, 10% data is used. 11 out of 13 for the excellent data are accurately predicted . There is accurate prediction of 10 out of 12 for the very good data.

Figure 11 shows Performance of Neural network model. Students are classified into poor, good, very good, and excellent based on their performance.

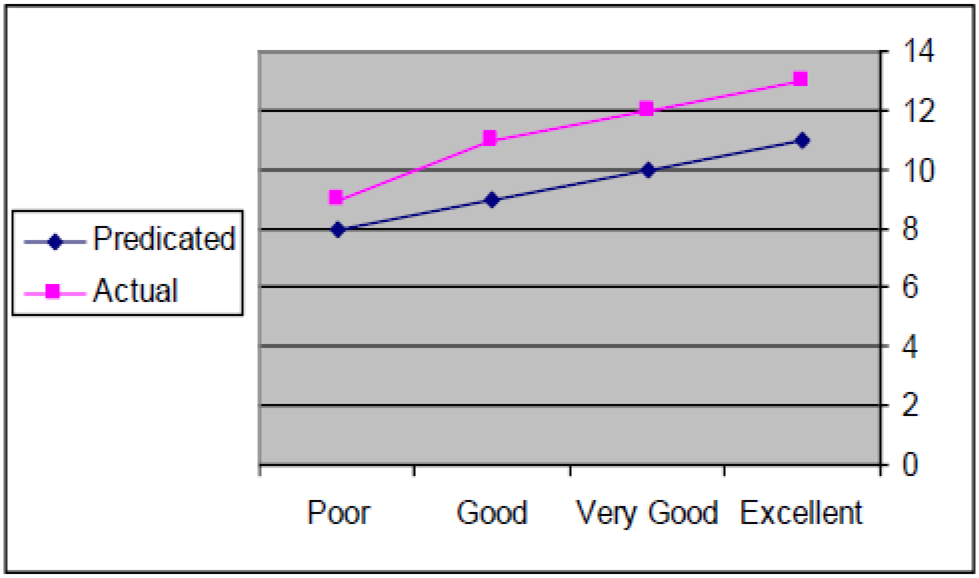


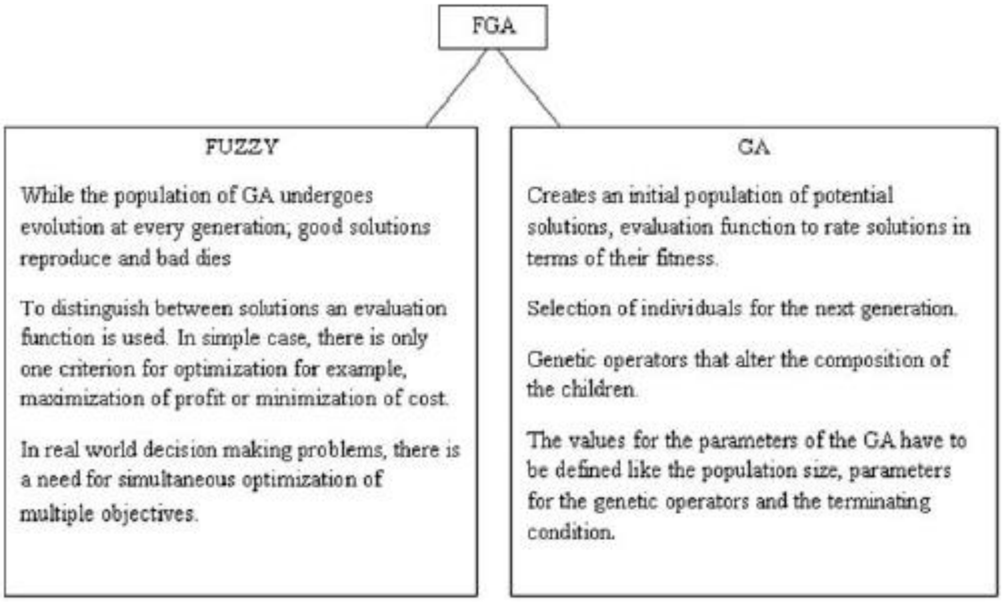
Figure 11: performance of neural network methodology on student dataset(S. Abu-Naser et. al,2015)

[8]. Sayana et al.(2015) proposed kmeans clustering method to predict student academic performance. K mean divide students into clusters according to their performance. Performance is depend on various attributes like percentage of marks, attendance and class test.kmean is simple and qualitative method . Euclidian distance is taken as similarity distance. To predict the pass percentage and fail percentage of the students, kmean is resourceful way of predicting student success. Kmean clustering has advantages such as it is robust, very fast, and easier to understand. Moreover, when dataset are distinct, it gives good result; however, it has demerits such as number of cluster must be predefined. Number of clusters assigned automatically so it is efficient to apply kmean algorithm.

[9] Hashmia Hamsa et al., (2016) described predction of student performance by decision tree algorithm. Dataset is collected by schedules namely, assignment submission time, daily attendance, admission time. This includes 120 1nd 48 students from bachelor and master program respectively. There are two batches of 30 students from computer science program and other two batches of 30 students from elctronic and communication department. Among 48 students data, there are two batches of 12 students from master computer science program and

two batches of 12 students from electronic and communication department. Admission score, internal marks and average for two sessional marks are attributes in dataset. Decision tree is working as follows : decision tree is same as tree structure. It classifies instances from root node to leaf node. A single node serves as the root of the tree. The path returns a leaf node if the tuples in D are of the same class C. Path produces a D majority class leaf node, if the attribute list is empty.To determine the splitting criterion, the algorithm uses the attribute selection method. As splitting attribute, attribute with highest informational gain is selected. Splitting attribute is SpA. One branch is grown for value of SpA, if SpA is discrete valued. Two branches are grown, if SpA is continuous valued.

Figure 12 shows how fuzzy genetic algorithm is working for student academic performance prediction Fuzzy genetic algorithm has two methods working simuktaneously, genetic algorithm and Fuzzy Fitness Finder.



While the population of GA undergoes evolution at every generation, good solutions reproduce and bad dies

To distinguish between solutions an evaluation function is used. In simple case, there is only one criterion for optimization for example, maximization of profit or minimization of cost.

In real world decision making problems, there is a need for simultaneous optimization of multiple objectives.

Creates an initial population of potential solutions, evaluation function to rate solutions in terms of their fitness.

Selection of individuals for the next generation.

Genetic operators that alter the composition of the children.

The values for the parameters of the GA have to be defined like the population size, parameters for the genetic operators and the terminating condition.

Figure 12. Working of FGA algorithm(Hashmia Hamsa et al., 2016)

Figure 13 shows results got after applying proposed methodology for mathematics subject for bachelor and master students. It also depicts how many students are at risk and determine how many students safe. Based on this, it predicts student academic performance.

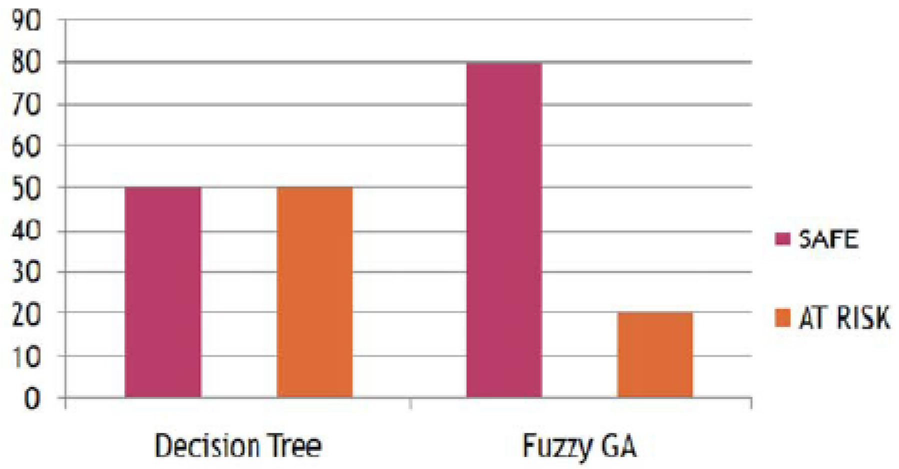


Figure 13. Prediction of mathematics result(Hashmia Hamsa et al., 2016)

In this paper(Hashmia Hamsa et al., 2016), authors have done comparison between decision tree and fuzzy genetic algorithm. Between these two methods, fuzzy genetic method performs better. Moreover, there is no feature visualization in this paper; however, in following paper(Elaf Abu et. al, 2016), different feature visualizations were there. Additionally, there were comparison among classification methods, bagging and boosting method.

[10] Elaf Abu et. al(2016) described student academic performance prediction using ensemble techniques. There are student features categories such as demographical, academic background features, parents participation on learning process features and behavioral features. Nationality, gender, place of birth, and parent responsible for student are demographical features. Educational stages, grade levels, section id, semester, topic and student absence days are academic background features. Parent answering survey, parent school satisfaction are there in parent participation on learning process feature category. Behavioral feature category includes features such as discussion groups, visited resources, raised hand on class, viewing announcements. Data preprocessing technique is applied for improving dataset quality. Data visualization is preprocessing task. To understand complex data, there is graphical representation. Figure 14 shows gender feature visualization. In this research, weka tool is used. The dataset is visualized based on gender feature into 305 males and 175 females

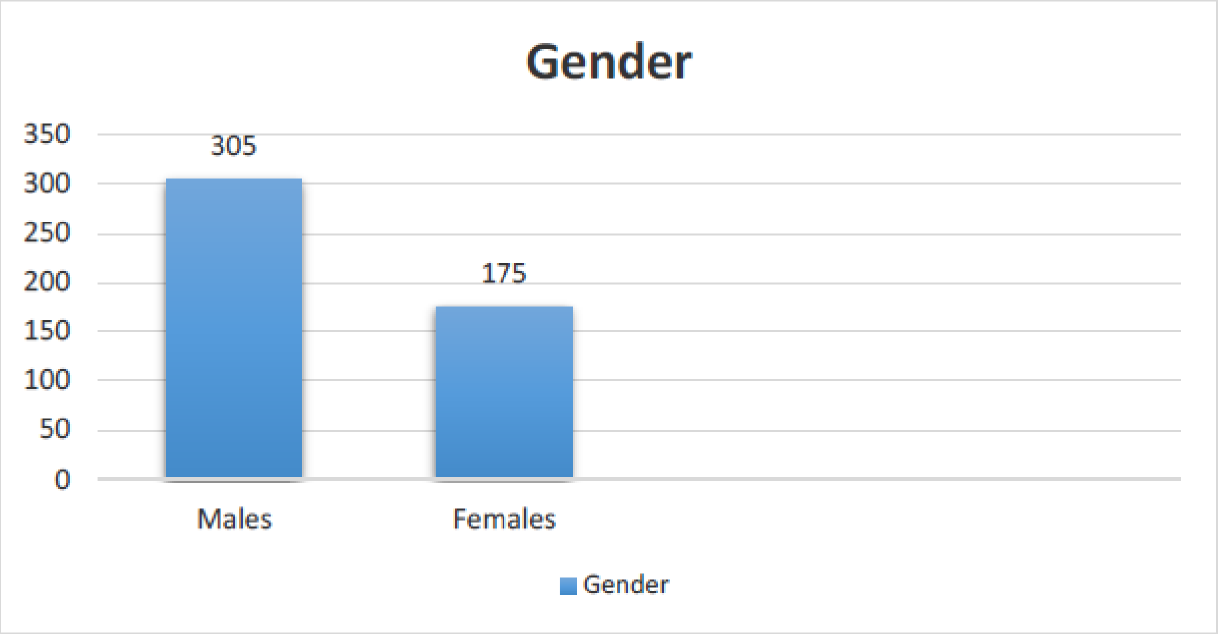


Figure 14. Gender Feature Visualization(Elaf Abu et. al, 2016)

Figure 15 shows educational stages visualization. Students are partitioned into educational stages such as lower level stage, middle level stage and high level stage. There are 199 students in lower level, 248 students in middle level, and 33 students in high level.

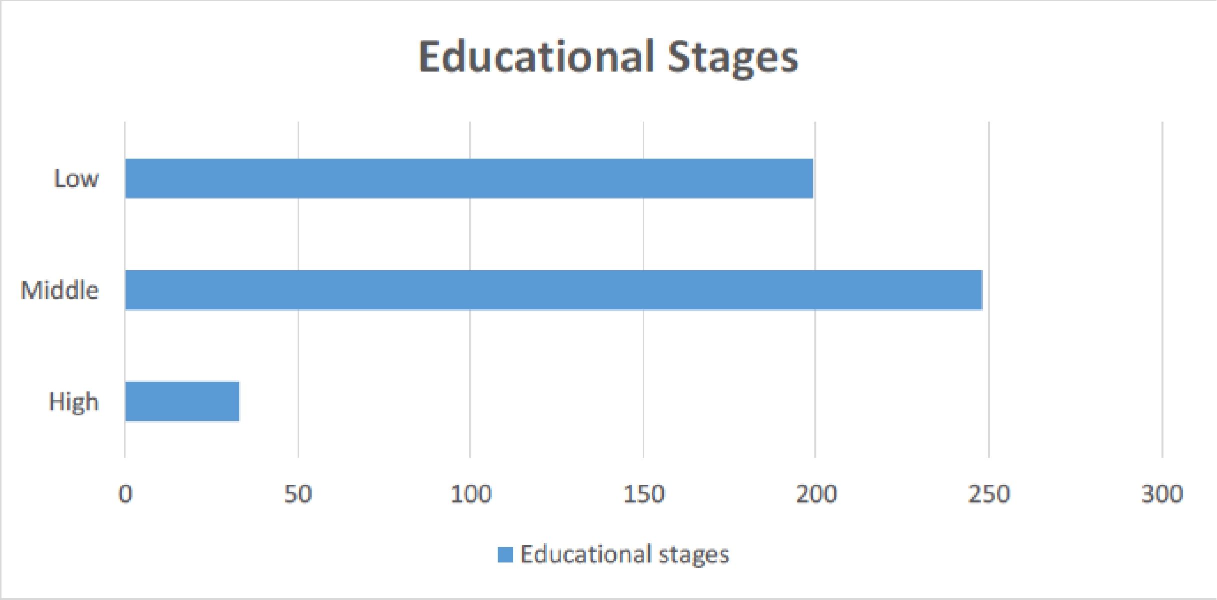


Figure 15. Educational stages visualization(Elaf Abu et al., 2016)

Bagging is an ensemble method. This methodology helps to increase accuracy of classifier. Bagging method starts with resampling original data into training datasets. Each bootstrap sample size is equal to the size of the original training set. Using different classifiers, bootstrap samples will trained. Individual classifiers results combine by majority vote process.

Each classifier is influenced by the performance of previous classifier in boosting method. Instances are choosen with a probability that is propotional to their weight. Bagging works best with high variance model. It produces variance generalization behavior with small changes to training data. In random forest, there is integration of randomized feature selection. Random forest utilizes random decision trees to select random subset of features.

Table 12 represents accuracy, recall, precision and f-measure evaluation metrics. There are methods namely, naïve bayes, decision tree and artificial neural network. Decision tree, artificial neural network and naïve bayes method accuray are 75.8, 79.1 and 67.7 respectively. By using decision tree, ANN and naïve bayes as base learners, Boosting method accuracies are 77.7, 79.1 and 72.2 respectively. Bagging method accuracies are 75.6, 78.9 and 67.2 respectively.

Table 12. Classification method results(Elaf Abu et al., 2016)

Evaluation measure	Classification methods			Bagging			Boosting		
Classifiers type	DT	ANN	NB	DT	ANN	NB	DT	ANN	NB
Accuracy	75.8	79.1	67.7	75.6	78.9	77.7	79.1	79.1	72.2
Recall	75.8	79.2	67.7	75.6	79.0	67.3	77.7	79.2	72.3
Precision	76.0	79.1	67.5	75.7	78.9	67.1	77.8	79.1	72.4
F-measure	75.9	79.1	67.1	75.6	78.9	66.7	77.7	79.1	71.8

Ensemble classification methods(Elaf Abu et. al, 2016) were used to predict student performance and it provides good accuracy and low errors than traditional classification methods; however, paper(Amjad, 2016) used traditional classification methods namely naïve bayes and decision tree.

[11] Amjad(2016) described ID3, CHAID decision tree method and naïve bayes method. ID3 generates decision tree from dataset. To create decision tree, setting are used.Splitting criterion, minimal size of split, minimal leaf size, minimal gain are used. After running the ID3 decision tree algorithm with the 10- fold cross validation on the dataset, the following confusion matrix was generated. Table 13 performance measure using ID3 method. Excellent, very good, good and pass are student categories.

Table 13. performance measure using ID3 method(Amjad, 2016)

		Actual				Class Precision(%)
		Excellent	Very Good	Good	Pass	
Prediction	Excellent	20	12	7	6	44.44
	Very Good	25	39	35	34	29.32
	Good	9	11	18	8	39.13
	Pass	6	19	8	13	28.26
Class Recall (%)		33.33	48.15	26.47	21.31	

C4.5 decision tree method is successor of the ID3 method.It uses pruning in the generation of decision tree. Moreover, settings such as Splitting criterion = information gain ratio, Minimal size of split = 4, Minimal leaf size = 1, Minimal gain = 0.1, Maximal depth = 20, Confidence = 0.5 are used in this method. After running the C4.5 decision tree algorithm with the 10-fold cross validation on dataset, the following confusion matrix was generated.

Table 14. performance measure using C4.5 method(Amjad, 2016)

		Actual				Class Precision(%)
		Excellent	Very Good	Good	Pass	
Prediction	Excellent	23	12	8	6	46.94
	Very Good	20	40	30	29	33.61
	Good	11	10	18	12	35.29
	Pass	6	19	12	14	27.45
Class Recall (%)		38.33	49.38	26.47	22.95	

CART decision tree method is classification and regresasion tree is decision tree method. It uses minimal cost-complexity pruning.

Following are the settings used with the CART operator toproduce the decision tree:

- Minimal leaf size = 1
- Number of folds used in minimal cost-complexitypruning = 5

Table 14. performance measure using CART method(Amjad, 2016)

		Actual				Class Precision(%)
		Excellent	Very Good	Good	Pass	
Prediction	Excellent	43	16	10	6	57.33
	Very Good	12	40	38	26	34.48
	Good	4	10	2	6	9.09
	Pass	1	15	18	23	40.35
Class Recall		71.67	49.38	2.94	37.70	

(%)					

[12] S.A.Oloruntoba et al.(2017) studied student academic performance prediction from student’s registration file from department of computer science of federal polytechnique in Nigeria for 2015-2016 for graduate students[new1]. This research focused on predictiong student performance based on ‘O’ level results. In this paper, there are also comparison of support vector machine with other data mining methods. In preprocessing stage, data is transformed into required form. There are many steps namely, data fetching, data cleanning, data filtering and data transformation. In data cleaning, the data with errors and irrelevant data are discarded. Data filtering reduces large data . new attributes are derived in data transformation stage. For small dataset, Support vector machine is most suitable. There is an optimal hyperplane to maximize the margin. In the next stage, this method is also applied to non-linear separable problems. After that, there is mapping of the data to high dimensional space. For implementation of proposed method, there is usage of python sklearn. Support vector regression is compared with other algorithms namely Linear Regression, Decision tree, LASSO(LASSO regression), elastic net and K nearest neighbour.Mean square error(MSE) for proposed method support vector regression is -0.170131. which is lowest compared to other methods. The parameters of support vector regression are kernel functions and penalty parameters. Kernel has values namely linear, poly and RBF. In this research, authors measure accuracy of various kernels with c values are 10,100 and 1000. There are accuracy prediction results for various kernels.

Figure 16 represents proposed methodology diagram for student performance prediction using support vector machine.

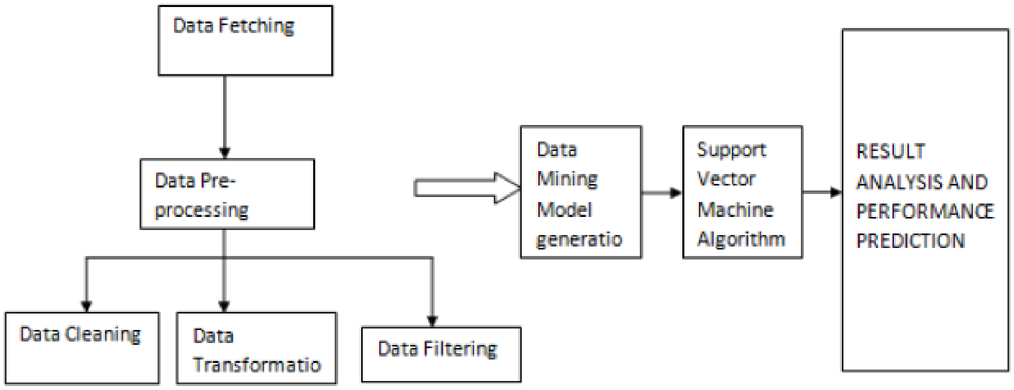

Data Fetching
Data Pre-processing
Data Cleaning
Data Transformatio
Data Filtering
Data Mining Model generatio
Support Vector Machine Algorithm
RESULT ANALYSIS AND PERFORMANCE PREDICTION

Figure 16. Proposed methodology of support vector machine on student dataset(.A.Oloruntoba et al., 2017)

Table 15 depicts training as well as testing accuracy after tuning kernel and penalty parameter values.

Table 15. Training accuracy and testing accuracy for linear, Poly and RBF kernel(A.Oloruntoba et al., 2017)

Kernel	Training accuracy	Testing accuracy
Linear(C=10)	77	79
Linear(C=100)	78	75
Linear(C=1000)	45	49
Poly(C=10)	94	65
Poly(C=100)	94	51
Poly(C=1000)	94	51
RBF(C=10)	94	97
RBF(C=100)	94	98
RBF(C=1000)	94	96

Support vector machine classification method is applied on student dataset in this paper(S.A.Oloruntoba et al., 2017). Moreover, there is tuning of kernel and penalty parameter values; however, in this paper(Snehal Bhogan et al., 2017), unsupervised machine learning enhanced K-strange point clustering method is applied on student dataset.

[13] Snehal Bhogan et al.(2017) described that Kmean is unsupervised machine learning method. There are clusters with similar group of similar data.In Kmean clustering algorithm , the data points of the dataset is partitioned into clusters. When there is unlabeled data, this technique is used.

Enhanced K-strange point clustering algorithm and naïve bayes classification algorithm is proposed to predict student performance. It overcomes the disadvantage of k-mean clustering algorithm and decision tree. In K-mean algorithm, clusters or centroid does not converge in which it go into infinite iteration. The limitation of decision tree is that, it does not consider all the attributes of the student academic performance dataset. So to overcome this in proposed methodology, there are enhanced k-strange clustering algorithm and naïve bayes methodology.

In this research, authors used student database of batch 2012-2016 and batch 2013-2017 of Agnel Institute of technology and design. Authors used database of batch 2012-2016 as training dataset and 2013-2017 as testing data. The proposed method of enhanced k- strange point clustering algorithm finds minimum of dataset. It is known as Kmin. After that, in the next stage, it finds maximum distance from Kmin. In the next stage Next, from Kmax and Kmin from the dataset, the algorithm computes maximally separated third point. Using Euclidean distance formula, computation of separation is calculated. Results indicated that in k-mean clustering technique, some of medium and low class tuple was assigned to high and medium class; however in enhanced k-strange clustering, this is not happened.

[14]Alaa Khalaf Hamoud et al.(2017) described student success prediction using bayesian methods of machine learning. In this paper, authors collected data in two ways. The first way is to questionnaire to collect answers from computer science and information technology department of college. The another way is to collect answers online by google forms. Dataset file is comma seperated value files and there are total 161 data in dataset. The tool used for this research is weka tool. There are questionarrie based on various parameters namely Gender,Address, Status, Work, LiveWithParent, ParentAlive, FatherWork, MotherWork, Failed courses per semester, number of days of absence per semester, Grade Point Average,

YearsOfStudy, notes writing during lectures. Model construction diagram to predict student academic performance is as follows :

Figure 16 shows construction diagram for predicting student academic performance. There are stages such as data preprocessing, attribute selecttion, applying algorithms and results evaluation.

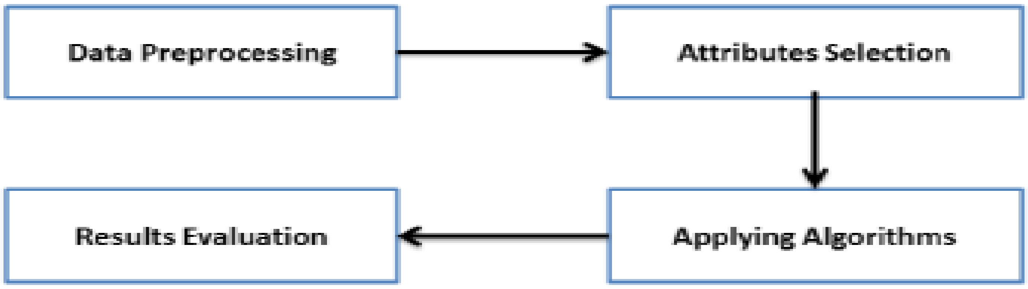


Figure 16. diagram for predicting student academic performance(Alaa Khalaf Hamoud et al.,2017)

In this research, To evaluate the correlation between the class and other attributes, filter correlation attribute evaluation is used. Table shows correlation average of questions.

In this research, by student's success, success of educational institutions can be measured.two bayesian algorithms are applied in this model . the questionnaire have 62 questions. Questions are related to health, social activity, relationships and academic performance. Weka 3.8 tool is used to build this model. There are two stages in which this model is working. In the first stage, there are most correlated questions to the final class. In the second stage, algorithms are applied on dataset and then find the optimal algorithm. There are two bayesian algorithms applied on student academic performance datasets. In this research proposed method, there are naïve bayes and bayesnet. Naïve bayes classification working as follows. T be training dataset of samples, each sample is n-dimensional vector , X={x1, x2,….xn}. When given sample x, Bayesian network is graphical depiction of relationships among variables. By round nodes, variables are represented and by arrow, dependence between one variable and another is determined. When there is sample X, the classifier predict x which belongs to the class which has the highest posteriori probability. So there is class that maximizes probability of Ci given x.

Table 16 shows correlation average of questions. For example Q26 has average of 0.088, and Q10 has correlation average of 0.174.

Table 16 : Correlation Average of Questions(Alaa Khalaf Hamoud et al.,2017)

Sequence	Question	Average
1	Q26	0.088
2	Q10	0.174
3	Q51	0.21
4	Q27	0.168
5	Q45	0.173
6	Q4	0.062
7	Q40	0.145
8	Q32	0.039
9	Q34	0.131

There are various evaluation metrics namely True Positive Rate, False Positive Rate, Precision and recall used for prediction of student academic performance.True positive rate, False positive rate, Precision and recall for bayesnet algorithms are 0.655, 0.432, 0.643, 0.655 respectively. For naïve bayes true positive rate, false positive rate, precision and recall are 0.667, 0.297,0.706 and 0.667 respectively. Naïve bayes algorithm has higher true positive rate so it is best methodology to predict student academic performance.

Figure 17 shows comparison of evaluation metrics of naïve bayes and bayesnet.

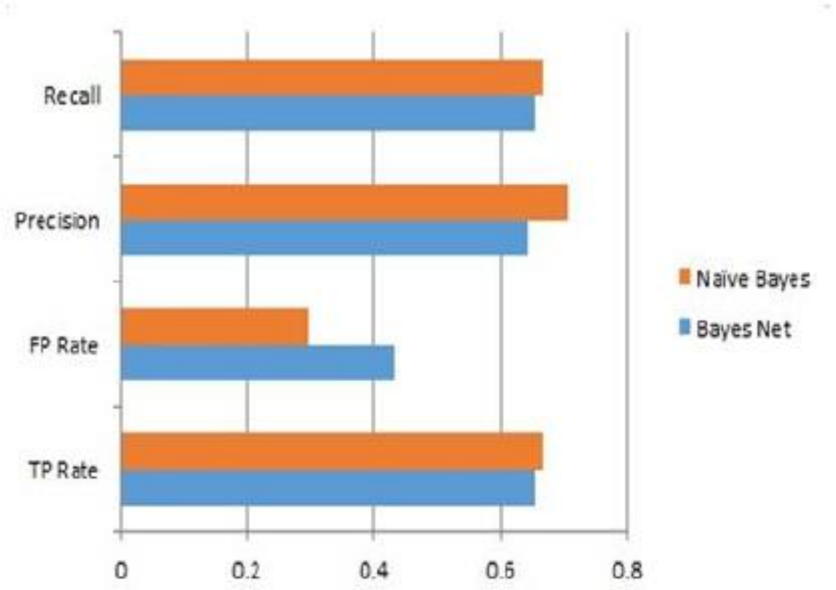


Figure 17. Evaluation metrics bayesian algorithms(Alaa Khalaf Hamoud et al.,2017)

In this paper(Alaa Khalaf Hamoud et al., 2017), authors collected data in two ways. The first way is to questionnaire to collect answers from computer science and information technology department of college. The second way is to collect answers online by google forms; however, in paper(Devine Grace Doble Function et al., 2017), there are not two ways to collect data. Dataset contains students enrolled in computer organization course for 2016-2017 and 2017-2018.

[15] Devine Grace Doble Function et al.,(2017) proposed decision tree J48 algorithm for prediction of student academic performance. In this research, dataset consists of 2nd year BSIT students admitted in computer organization course for 2016-2017 and 2017-2018. There are six variables in data such as, midterm, final exam, projects, remarks, recitation and quizzes. Possible values for quizzes is 3 grade. Recitation values are from 0 to 5. The value of project is from 1 to 10. Possible values of midterm examination are 60-95 grades and final examination values are from 60-95 grade. Software used for this research is weka tool. The programming language used in weka is java. Process consists of data collection and data preprocessing and data visualizations. For prediction, there are five phases such as training, pattern, testing, result evaluation and knowledge representation. Decision trees methodology is supervised learning method. It is most popular classification method. It has several merits such as these methods can work for attributes which has different types such as nominal and numerical. The other advantage of decision tree is that these can classify new examples fast. Its simplicity makes easy

to understand. Working of j48 is as follows: tree is leaf labeled with the most frequent type in T, if all T belongs to the same group of instances. In step 2, There is partition of T into T1, T2 and T3 and in recursive way, same thing applied. If step 1 occurs an error, selection of outcome in single and more test-based attributed then take this test as a root node of the tree with one branch of each issue of the trial. Using two heuristic criteria, data gain and default gain proportion are positioned. Cross validation method is used to validate the predicted model. There is division of training data into some folds or partitions by CV test. 10-fold and bootstrap are two types of cross validation.

Decision tree has correctly classified 106 instances as PASS and 13 cases as conditional leading to misclassification.Moreover, this method has correctly classified 13 instances as FAILED and three instances as conditional leading to misclassification. Decision tree has incorrectly classified 6 PASS and five failed to lead to misclassification.

Table 17. proposed method output(Devine Grace Doble Function et al., 2017)

Correctly classified instances	124	82.1192
Incorrectly classified instances	27	17.8808
Kappa statistic	0.5393	
K & B relative info score	7935.1148	
K & B information score	87.3548	0.5785
Class complexity	163.5452	1.0831
Class complexity	8655.735	57.3227
Complexity improvement	-8492.19	-56.2397
Mean absolute error	0.1277	
Root mean squared error	0.3094	
Relative absolute error	45.3953	
Root relative squared error	83.0006	
Total number of instances	151	

Table 18. output evaluation metrics(Devine Grace Doble Function et al., 2017)

	True positive rate	False positive rate	precision	Recall	f-measure	MCC	Class
	0.946	0.333	0.891	0.946	0.918	0.657	Pass
	0.722	0.023	0.813	0.722	0.765	0.737	Failed
	0.238	0.085	0.313	0.238	0.270	0.173	Conditional
Weighted average	0.821	0.262	0.801	0.821	0.809	0.599	

Although authors have got better accuracy of 82 % by applying proposed classification method j48 method in this paper(Devine Grace Doble Function et al., 2017), paper(Stephen J.H Yang et al., 2017) studied prediction of student academic performance using multiple linear regression and principal component analysis.

[16] Stephen J.H Yang et al.(2017) studied prediction of student academic performance using multiple linear regression and principal component analysis. In this research, for improving prediction accuracy, multiple linear regression is used with principal component analysis(PCA) . There are limitations of traditional multiple linear regression are that the coefficient of determination and mean square error can not evaluate the accuracy of multiple linear regression. In this paper, predictive MSE(pMSE) and predictive mean absolute percentage correction(pMAPC) measures for calculating accuracy of regression model. Evaluation measures such as R^2 and mean squared error(MSE) evaluate goodness of fit of regression models. In this proposed methodology, goodness of fit of regression model is measured using traditional evaluation measures. But traditional evaluation metrics can not evaluate accurate regression model performance. So there are additional evaluation measures for predicting regression methodology performance. So this helps teachers to evaluate predictive accuracy. The shuffling concept overcome the issue of higher residual errors influenced by outlier data. R^2 and MSE only measure goodness of fit of regression model. It does not help in predicting accuracy. To reduce risk of wasting resources , teachers need to know accuracy. There is comparison of Pmapc and Pmse values for multiple linear regression with principal component analysis.The

Pmse and pmapc values for multiple linear regression with principal component analysis are 198.62 and 0.81 respectively. For RQ 2, t-test is performed to examine the predictive performance measures between multiple linear regression with and without principal component analysis. MLR performance result can be improved by using six components of principal component analysis. There is strong correlations among independent variables. So that performance can be improved using PCS, and best predictive performance is predicted by six components.There are optimal Pmse and pmapc values of regression algorithms by six components, this is due to seventh to twenty-seventh components are influenced by outliers.

In this research(Stephen J.H Yang et al., 2017), dataset attributes are of numerical type. Additionally, student performance prediction is done using regression method; however, in paper(Stephen J.H Yang et al., 2017), dataset attributes types are numerical and categorial.

[17] Hafez Mousa et al., (2017) proposed school student performance prediction using data mining classification method. Data mining includes three classification methods such as naïve bayes, decision tree and k-nearest neighbour classification technique.

Naïve bayes method : It is a popoular methodology. It is used with independent attribute datasets.It requires less computation than other methodologies.It has less accuracy and it is faster technique. For training, it require small amount of data. Decision tree method is known as ID3, then it is improved to C4.5. In K-nearest neighbour method, there is comparison of test cases with training cases. It is easiest machine learning algorithm in data mining method. Figure 18 shows proposed methodology phases. There are five phases in methodology steps. Phase 1 is business understanding. Phase 2 is data collection. Phase 3 is data preprocessing. In phase 4, data mining classification algorithms are applied on dataset. Result evaluation is there in phase 5.

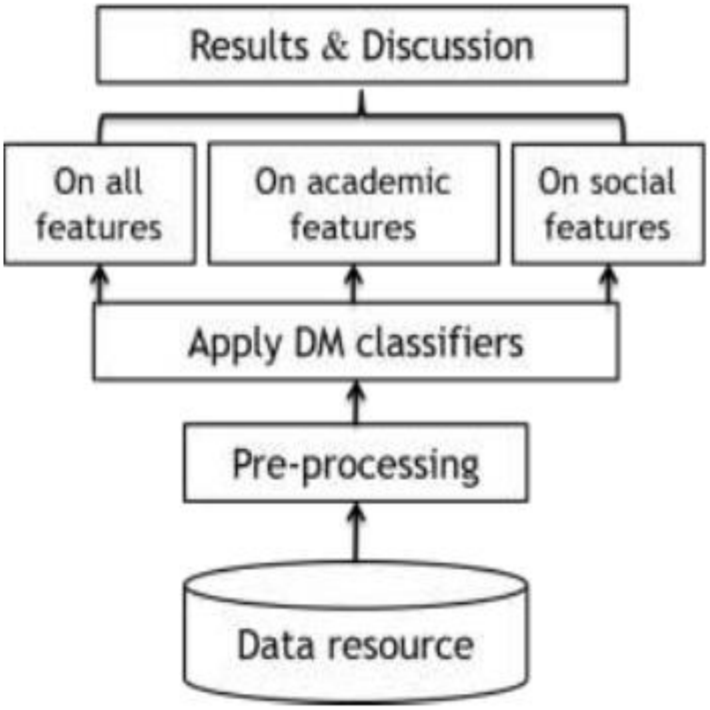


Figure 18. Proposed method steps(Hafez Mousa et al., 2017)

There is data of 1036 students . It is collected from preparatory school(7th, 8th, 9th grade) in year 2014-2015. It is of UNRWA schools. Dataset attributes are level, orphan, SHC, BirthYear, Camp, FatherWork, FailYears, PrevYear, FirstTerm, FinalResult. Table 19 represents student dataset attributes, its type and attribute description. Level is of categorial type, orphan, SHC, camp, fatherwork, prevyears, firsttem and finalresult are of categorial type.

Table 19. Dataset Attributes(Hafez Mousa et al., 2017)

Attrinbutes	Type	Description
Level	Categorial	7th, 8th, 9th
Orphan	Categorial	Father, mother, both
SHC	Categorial	Yes, no
Birthyear	Number	
Camp	Categorial	Yes, no
Fatherwork	Categorial	Yes, no

Failyears	Number	
Prevyears	Categorial	Success, fail
Firstterm	Categorial	Excellent, very good, good, fair, poor, very poor
Finalresult	Categorial	Success, fail

In this research, there is classification results by using all attributes of dataset is measured.44

Table 20. Classification results by using all attributes of student dataset(Hafez Mousa et al., 2017)

	Naïve bayes	Decision tree	k-nn
Accuracy	91.79	92.96	88.86
Recall	63	97.83	63.04
Precision	72.50	66.18	58.00
f-measure	67.44	78.95	60.42

This table 21 represents classification method results by using academic attributes such as FailYears, PrevYear and FirstTerm.Classification methods are naïve bayes, decision tree and k-nn.

Table 21. Classification results by using academic attributes of student dataset(Hafez Mousa et al., 2017)

	Naïve bayes	Decision tree	k-nn
Accuracy	91.79	92.96	88.86
Recall	63	97.83	63.04
Precision	72.50	66.18	58.00

f-measure	67.44	78.95	60.42

This table 22 represents classification method results by using social attributes such as orphan, SHC, Birhyear, Camp, Fatherwork and failyears. Classification methods are naïve bayes, decision tree and k-nn.

Table 22.Classification results by using social attributes of student dataset(Hafez Mousa et al., 2017)

	Naïve bayes	Decision tree	k-nn
Accuracy	86.80	87.10	82.99
Recall	17.39	13.04	34.78
Precision	53.33	60.00	36.36
f-measure	26.23	21.43	35.36

Authors(Stephen J.H Yang et al., 2017) studied data mining classification methods such as naïve bayes, decision tree and k-nn. Accuracy is low in this research(Stephen J.H Yang et al., 2017) ; however, ensemble methods such as boosting and bagging are applied in paper(Olugbenga et al., 2017).

[18] Olugbenga et al.(2017) described ensemble approach to predict student academic performance. Objective of this study is to compare different classifiers and ensembles of classifiers. In this paper, there are comparison of various data mining techniques such as Artificial neural network, decision tree, bagging, boosting and stacking. Bagging is a short form for bootstrap aggregation, which is an ensemble learning process.individual model is taken in bagging and equal weight is assigned to each individual model. Boosting is different ensemble learning method It is used to improve accuracy of classifier and also lowering the error of weak classifier. It focuses on training each new model instance from the error or misclassification of previous stage.

Figure 19 shows ensemble model using stacking approach. Stacking method is known as stacked generalization. It is widely used ensemble method than bagging and boosting. It mimnimize generalization error. It also improve prediction classification accuracy.

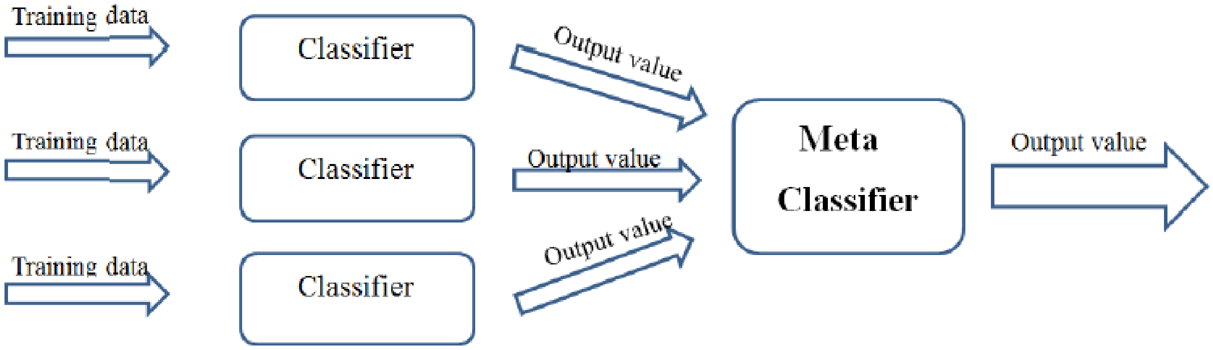


Figure 19. Ensemble model using stacking approach(Olugbenga et al., 2017)

Table 23 represents comparison and evaluation of the performance pof different methods. There are seven models in this research. Model1 uses banner variables as input factors, model2 uses moodle variables, model 3 uses survey variables. In model 4, banner and moodle models are used as input factors. Banner and survey models are used as input factors in model 5, In model 6, moodle and survey models are used as input factors. Banner, moodle and survey model are used as input factors in model 7.

Table 23. Comparison and Evaluation of the performance different Algorithms(Olugbenga et al., 2017)

Model	Algorithms	Performance measurement					
		PAP	RMSE	Precision	Recall	F-Score	Classification Error
Model 1	DT	34.81	0.751	10.78	13.84	12.12	65.19
	ANN	39.62	0.753	12.17	15.34	13.57	60.38
	SVM	49.62	0.710	7.09	14.29	9.48	50.38
	DT	50.33	0.685	25.76	25.36	24.60	49.67

Model 2	ANN	50.19	0.697	21.42	23.39	22.36	49.81
	SVM	52.48	0.667	15.50	19.18	17.14	47.52
Model 3	DT	78.05	0.464	13.66	15.67	14.16	21.95
	ANN	73.10	0.504	15.48	16.22	15.84	26.90
	SVM	82.95	0.411	13.83	16.67	15.11	17.05
Model 4	(Ensemble Model)	81.62	0.421	77.74	74.52	76.10	18.38
Model 5	(Ensemble Model)	73.10	0.503	66.37	60.17	63.12	26.90
Model 6	(Ensemble Model)	80.81	0.421	70.00	71.94	70.96	19.19
Model 7	(Hybrid Model)	81.67	0.396	79.62	75.86	77.69	18.33

This paper(Olugbenga et al., 2017) studied various ensemble methode and applied it on dataset; however, G. Sujatha et al.(2018) described predicting student performance using personalized analytics on educational databases.

[19] G. Sujatha et al.(2018) described predicting student performance using personalized analytics on educational databases. The main objective of this study are to find the best regression method on student data. Regression is used to predict numeric or continuous value. For prediction of the student’s success or failure. The other objective of this research is to predict the final grade. In this study, various attributes are marks in c programming, marks in maths, marks in english, higher secondary school board of study, medium of study, group of study, computer programming experience or computer programming studied, internal marks of c programming. Possible values of marks in c programming, marks in maths, and marks in english are from 0 to 100. Possible values of higher secondary school board of study are state, ICSE and CBSE. Local and english are medium of study. Possible values of group of study include biology and maths.Yes and no are possible values of student computer programming experience. Possible values of internal marks of c programming are 0 to 50. In this research, dataset os collected from quba college of engineering. The data is about computer science students programming marks. It consists of 3000 data of 3 years. The dataset is divided inn train and test data. 70% data is considered as train and 30% data is considered as test dataset. Student performance is classified as fail, weak programmer, good programmer, and excellent programmer. Grade between 0 to 49

is classified as fail. Grade between 50 to 64 is classified as weak programmer. Good programmer value ranges are from 65 to 79. Grade between 80 to 100 is considered as excellent programmer. Evaluation metric is Root Mean Square Error for measuring algorithm's error. In proposed methodology, regression method is used to predict student performance. On training set, There is multi linear regression for programming course and evaluated the prediction results by applying different algorithms such as support vector machine, random forest and decision tree. By applying multilinear regression model on trining dataset, formula lm(formula = c ~ m + e + s + l + bck +cs + int, data = trai). Then error rate is calculated by applying algorithms namely multi linear regression, support vector machine and random forest. Error rate for multilinear regression, support vector machine, decision tree and random forest are 4.855, 4.38, 5.278, and 5.176 respectively. In support vector regression technique, idea is to find the best fit line. The best fit line is hyperplane that has maximum number of points. Regression models for predicting student academic performance in an engineering dynam student performance. Multiple linear regression also known as multiple regression. It is a statistical technique that uses expllanatory variables to predict the outcome of a response variable. ,Multiple regression is an extension of linear regression that uses one explanatory variable. Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets.

This paper(G. Sujatha et al., 2018) used regression method to predict student performance; however, Mr. shubham Agrawal and Mr. Kapil Sahu(2018) proposed different data mining method to apply on student academic performace dataset.kmean algorithm is the best to predict student academic performance.

[20] Mr. shubham Agrawal and Mr. Kapil Sahu(2018) proposed kmean method to apply on student academic performace dataset.kmean algorithm is the best to predict student academic performance. They also compared kmean algorithm prediction result with decision tree algorithm prediction and concluded that proposed method kmrean is better in terms of accuracy and has less errors. In the first stage, To group data, it accepts the number of cluster . after that, in the next stage, there is initialization of first k cluster. In the next stage, for each cluster, there is calculation of the arithmetic mean. After that, there is assignment of each record in the dataset to initial cluster. In the next stage, there is reassignment of each record to most similar cluster. And there is recalculation of the mean of the cluster in dataset. This technique is known as cluster

analysis. Clustering analysis is used in many areas such as prediction, market research, data analysis and image processing. There is partoytion of k clusters from n classifications. There are various merits of kmean algorithm. The first merit is that, when centroids are recomputed, an instance can change cluster. This method helps to produce higher clusters. Additionally, it is very easy to implement. There are k clusters from n clarifications. Cluster centroid is defined as mean value of objects in cluster. There are k centers . For each cluster, there is k center. In the next step, the better way to place middles far from each other .different locations give different results so middles should be placed at accurate location. Then, there is recalculation of k new centroids. After that, a new procedure is applied on same dataset with nearest new center. In the next stage, there is generation of loop. It indicates that k centers change their location until there is no more change. In this research, student dataset parameters are student id, semester,subject-marks, and sem, result. There are five subjects so sub1, sub2, sub3, sub4 and sub5 are attributes of dataset.

Proposed methodology of kmean in this paper(Mr. shubham Agrawal and Mr. Kapil Sahu, 2018) used weka tool for student performance prediction. Study by A.s Arunachalam et al.(2018) used matlab tool for implementation of method on student dataset.

[21] A.s Arunachalam et al.(2018) described analyzing student performance using evolutionary artificial neural network. Proposed methodology is based on genetic algorithm and artificial neural network. Due to the capability of assigning proper weights to each node under the hidden layer and poor modeling structure, traditional artificial neural network lacks predicting student performance. The dataset attributes are overall semester mark, age, day scholar, board studied, attendence, co-curricular activities such as paper presentations, seminar attendance. In this paper, evolutionary artificoial neural network is proposed as methodology. It follow working principle of genetic algorithm. Proposed methodology genetic algorithm have three main operations first is process of selection, process of cross over among parents, process of mutation among parents. In first Stage of crossover, Crossover is a process in which there are two steps. Figure 20. shows working of genetic algorithm.There are phases such as fitness, selection, mating and mutation and final selection.

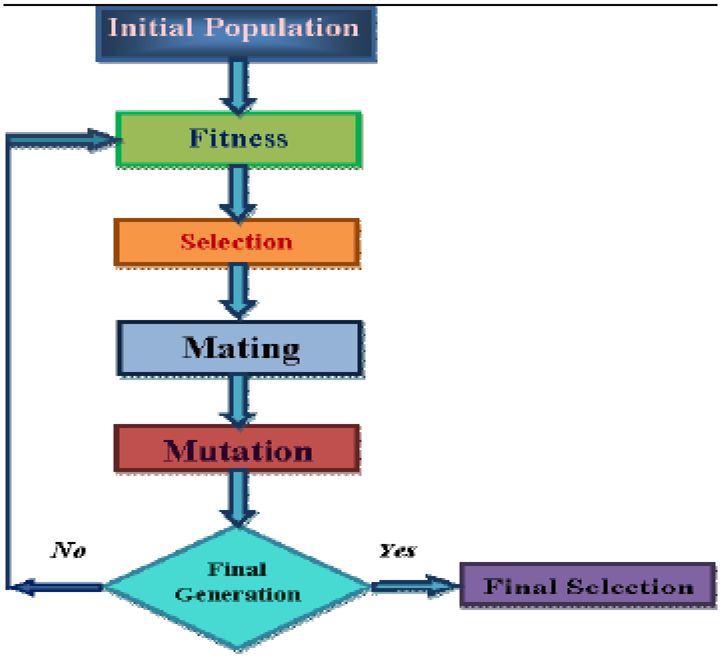


Figure 20. genetic algorithm process(A.s Arunachalam et al., 2018)

Figure 211 depicts workflow of proposed evolutionary artificial neural network. In the first step, there is initialization of weights and thresold. After that, there is calculation of fitness. In the next step, there is selection of genetic operator, crossover operator and mutation operator. After that, optimal and thresold is obtained. In the next stage, error is obtained. Finally, result is generated at the end.

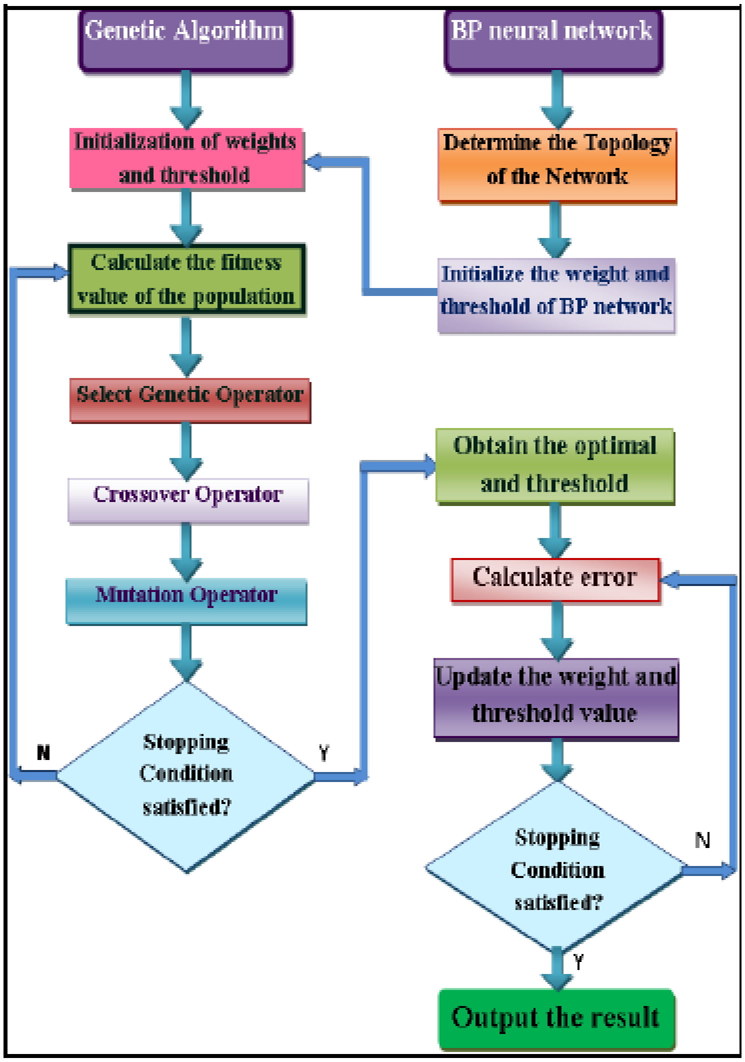

Genetic Algorithm
BP neural network
Initialization of weights and threshold
Determine the Topology of the Network
Calculate the fitness value of the population
Initialize the weight and threshold of BP network
Select Genetic Operator
Obtain the optimal and threshold
Crossover Operator
Calculate error
Mutation Operator
Update the weight and threshold value
N
Stopping Condition satisfied?
Y
Stopping Condition satisfied?
N
Y
Output the result

Figure 21: working of proposed methodology evolutionary artificial neural network(A.s Arunachalam et al., 2018)

In this method, for assigning and updating the hidden nodes weight, genetic algorithm is used. 1-m-n neural network pattern is there for valuation. 1 means there is one input node. N number of output nodes and m number of hidden nodes. The sigmoidal function is appled for hidden and output layer. The output of hidden layer Outh is considered as input of hidden layer Inph. The equation of out_h is like

N = (1 + n) *m. By this formula, amount of weight N is calculated.

L = N * d = (1+n)*m*d . By this formula, the distance of the chromosome L is measured. Dataset attributes are Gender, locality, paredu(parental education), Eco(Economic status), Attendence(Class attendence) and result. Possible values of gender are M and F. Possible values of locality are urban and rural. Values of paredu are Edu and unedu. Possible values pof Eco are high and low. Possible values of attendence are High and low. Attribute Result has values such as first, second, third and fail. The software tool used for implementation of proposed methodology is MATLAB tool. By using this tool, comparison of proposed methodology evolutionary artificial neural network is done with ANN and PNN.

[22] Raza Hasan et al.(2018) proposed decision tree method for student academic performance prediction. The data contains of student's activity and performance data from degree course at Middle East College. This study take data from Spring 2017 using a sample of 22 students who had registered in the said module from undergraduate level. The dataset comprises of variables related to two categories which are as follows:

- Student's Academic Information (Cumulative Grade Point Average (CGPA), High Risk(student having high failure rate in the same module), Term Exceed at Risk, At Risk (student failed 2 or more modules previously), Student Success Center (SSC), Coursework 1 (CW1), Coursework 2 (CW2), End Semester Examination (ESE) and Plagiarism Count).
- Students Activity (On Campus and Outside Campus access). This is the time spent by the student on Moodle in minutes from inside or outside the campus. The values were taken from MEC Moodle for E-Commerce Technologies in Fall 2017.

Figure 22 depicts proposed methodology of classification used in this research. In this research, data is gathered from student information system and number of minutes spend on the moodle.Various classification algorithms are decision tree, REPTree, RandomForest, LogisticMofdel tree, Hoeffding tree, decision stump, naïve bayes and SMO.

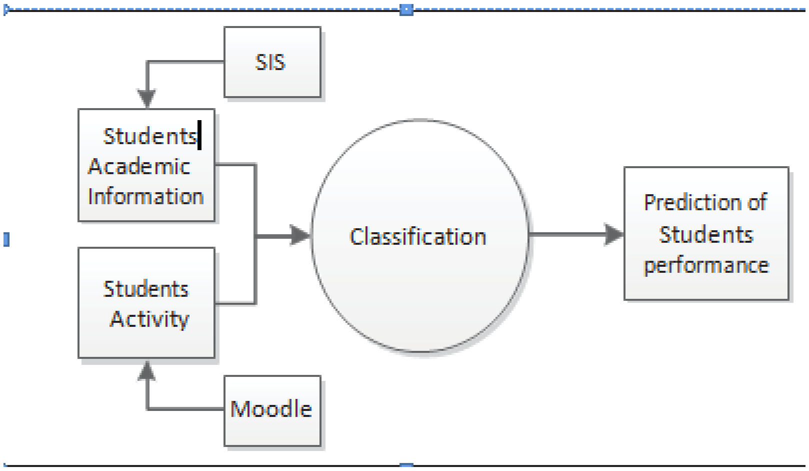

SIS
Students Academic Information
Students Activity
Moodle
Classification
Prediction of Students performance

Figure 22 .Proposed methodology(Raza Hasan et al, 2018)

As a result, different classifiers have different accuracy and kappa values. Decision tree, REP Tree, Random forest, logistic model tree, hoeffding Tree, decision stump classifier accuracies are 63.63%, 45%, 100%, 45.45445%, 45.4545% and 50% respectively. Among all these classifiers, random forest performs better on student dataset for predicting student academic performance. Kappa values of random forest and SMO are good. Its value is 1. Decision tree method has kappa value of 0.4172. 0.8616 is kappa value for naïve bayes classification method. 0.1103 and 0 are kappa values decision stump and REP Tree respectively.

In this paper(Raza Hasan et al., 2018), kappa value and accuracy are performance measures; however, root mean square error, mean absolute error are performance measures in Mehta Smruti and Ashish Adholiya(2019).

[23]Mehta Smruti and Ashish Adholiya(2019) studied this classification based on if percentage>80 then grade = o, if percentage>70 and percentage<=80 then grade=A, if percentage >60 and percentage<=70 then grade=b, if percentage>55 and percentage<=60 then grade=c, if percentage>50 and percentage<=55 then grade=D, if percentage<=50 then grade = E. In this research, authors used weka software data mining tool for implementation. weka software was developed at the universiTY of waikato in NewZealand. data pre-processing, classification, regression, clustering, visualization and association rules are implemented in weka software. Dataset consists of data of BCA students. Proposed method is decision tree algorithm. In this study, decision tree algorithm is used in statistical data mining and machine learning. As a result,

139 students fail, 69 students pass, first class received by 84 students, second class received by 68 students and 16 students got first class distinction. There are evaluation metrics are correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error, root mean squared error, relative absolute error and root relative squared error. 10 fold cross validation and training and test set are used for implementation. Accuracy for prediction after selecting attributes are 99.05%, incotrrectly classified instances are are 0.942%, kappa statistic is 0.985, mean absolute error 0.0048, root mean squared error 0.057, relative absolute error 2.2716 and root relative squared error is 17.6437. In case of training and test dataset, accuracy for correctly classified instances is 99.2467%, incorrectly classified instances 0.7533%. train test set technique is also applied on dataset. Various evaluation metrics are measured namely, correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error, root mean squared error, relative absolute error and root relative squared error. Classification method provided better results in paper(Mehta Smruti and Ashish Adholiya, 2019). In this paper(Farid Jauhari et al., 2019), boosting method is proposed.

[24] Farid Jauhari et al.,(2019) proposed boosting algorithm for prediction of student performance. The datasets consist of 2 subsets, Mathematics subject dataset and Portuguese subject dataset. There are 395 student in mathematics subject dataset and the Portuguese contain 649 student dataset. there are 33 attributes in dataset. The last attribute that named G3 is the final grade of the student in the subject. The range value of this attribute is between 0 and 20. The objective goals of the prediction are at this last attribute.

There are 3 general steps of research methodology in this research(Farid Jauhari et al., 2019), those are preprocessing, building model, and experiment as a model evaluation. This methodology is shown in Figure 23. Preprocessing employed to set the class of students based on their final grades (G3). Binary classification (pass/fail) and 5-levels classification (very good, good, satisfactory, sufficient, and fail) is there in this research.

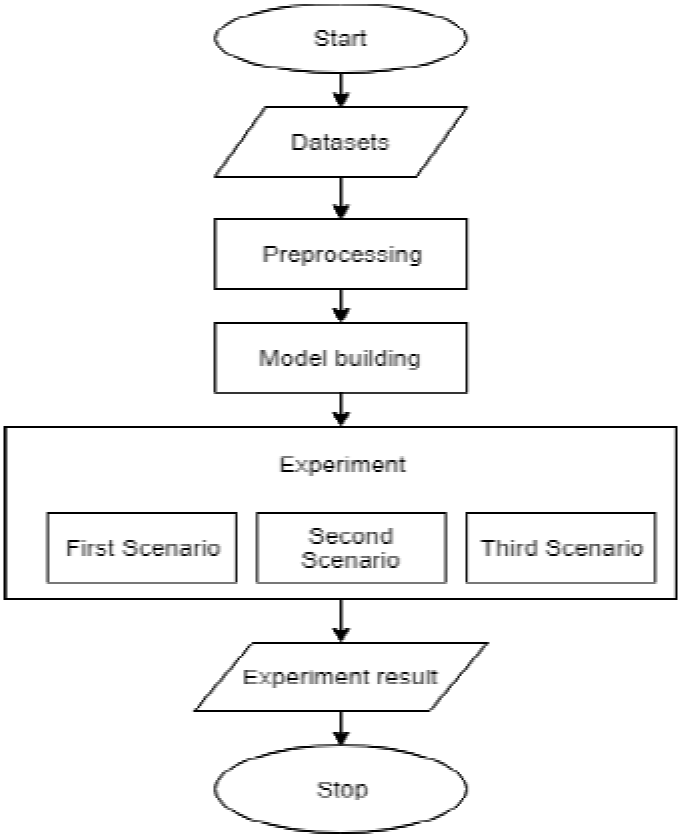


Figure 23. Proposed methodology(Farid Jauhari et al., 2019)

The accuracy of the prediction models was calculated by using Percentage Correct Classification (PCC). The PCC value near 100% means that the model has a high accuracy on classification. In this paper, there are 3 scenarios of evaluation. The first scenario compare those 3 boosting algorithms. In the second scenario, there is evaluation of 100% datasets as data training (it means data training = data test), 90% dataset as training data (it means randomly sampling of 90% datasets as training data and 10% datasets as test data), 80% - 10% dataset as training data.

Table 24 represents classification goals according to its grade. If grade is between 10-20, then students are classified as pass and if grade is between 1-9, then it is classified as fail.

Table 24. Grade in binary classification(Farid Jauhari et al., 2019)

Classoification gooals	Grade

Pass	10-20
Fail	1-9

In the first scenario, 10 fold cross validation is applied on student dataset. Highest PCC value is 93.197 by applying AdaBoost.SAMME on portuguese dataset and it is 91.377 by applying Adaboost.SAMME algorithm on student dataset.

Table 25. Result for 10-fold cross validation(Farid Jauhari et al., 2019)

			BINARY CLASSIFICATION			5-LEVEL CLASSIFICATION		
	RPART	C5.0	M1	SAMME	RPART	C5.0	M1	SAMME
Mathematics	90.392	88.96	90.893	91.377	77	60	74.567	72.434
Port	92.512	91.796	92.765	93.197	76	72.03	74.5436	71.978

In similar manner, second scenario and third scenario is applied on student datasets. In conclusion, Adaboost.M1 and Adaboost.SAMME models perform better than RPART on binary classification models: however, RPART outperforms boosting methods in 5-level classification.

In this paper(Farid Jauhari et al., 2019), proposed boosting ensemble method is used. In paper(P. ajay et al., 2020), ensemble method random forest was used.

[25] P. ajay et al.(2020) proposed random forest technique for predicting student performance on student dataset. In this proposed methodology,

Table 26 shows Dataset characteristics are as below.

Table 26 Dataset characteristics(P. ajay et al.,2020)

Dataset characteristics	Multivariate
AttributeCharacteristics	Multivariate
Associated task	Classification
Number of attributes	16
Number of instances	480

Weka tool is used for implementation of proposed methodology. Proposed methodology random forest working as follow There are two stages in random forest methodology one is random forest creation and other is form prediction from applying classifier.

In the first step of random forest, randomly choose k features from total m features.

After that, there is calculation of node d using best spplit point.

Then, using best split method, there is splitting of node into daughter nodes

Repeat steps until I number of nodes reached.

In the last stage of algorithm, apply random forest for n number of times for creation of n number of trees.

After applying proposed methodology, accuracy is determined. This proposed methodology of random forest is also compared with other methods such as naïve bayes and decision tree. Figure 24 shows evaluattion metrics values

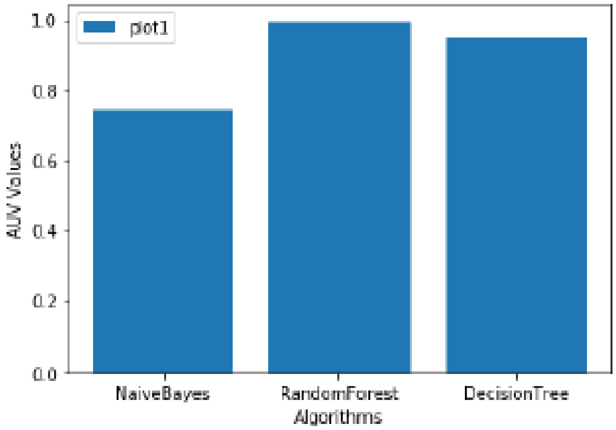


Figure 24. comparison of output between randomforest, naivebayes and decision tree(P. ajay et al., 2020).

In this paper(P. ajay et al., 2020), there was comparison among randomforest, naivebayes and decision tree. Whereas, paper(Abdullah Baz et al., 2020), used classification method naïve bayes.

[26] Abdullah Baz et al.(2020) studied naïve bayes method to predict student performance prediction. In this research, dataset contains 1398 records of students. Those students are graduated from Umm AI-Quara university database in year 2019. By using weka tool, classification method naïve bayes is used in this research. Weka contains collection of tools for classification, regression, association rules and clustering. In this research, dataset contains attributes namely, ID, gender, enhglish language grade, computer skill 1 grade, physics grade, mathematics(1) grade, maths(2) grade, computer programming skill, learning and study skills grade, technical english language grade, first semester GPA and final semester GPA. In data preperation, data cleaning is done to delete missing values. Students who does not contain all information, then extract tose data from dataset. Proposed methodology naïve bayes is used in this research. Bayes theorem is used in this method. It calculates probability of event based on conditions. There is also assumption that classification attributes are not dependent of value of class. It is very simple method. This method performs better with categorial data.
Equation is bayes theorem

$$P(A|B) = \frac{p(B|A)P(A)}{P(B)} \qquad (4)$$

Where p(A) = probability of A, P(B) = probability of B, P(B|A) = the probability of event B based on A condition, P(A|B) == the probability of event A based on B condition
Naibve bayes method classify student’s academic performance based on their grades in eight courses of first year.

Figure represents final GPA of students. It is shown that, 51.44% of graduate students had an excellent GPA, 36.95% had very good final GPA, . 10.14% OF STUDENT WAS GOOD AND remaining students had an pass GPA.

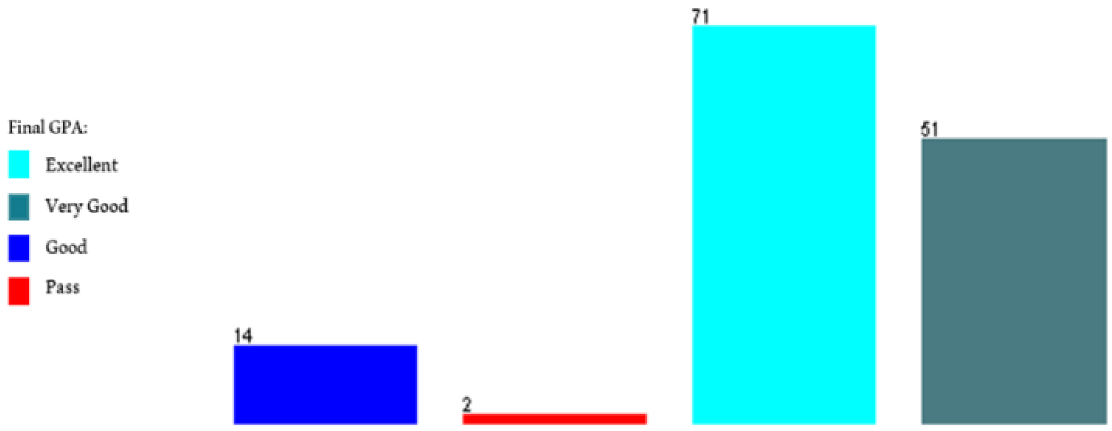


Figure 25. Final GPA of students(Abdullah Baz et al., 2020)

In this research, output measures are as below.

- True Positive (TP) Rate: The number of instances that aretruly classified for each class.
- False Positive (FP) Rate: The number of instances that arefalsely classified for each class.
- Precision: The number of truly classified instances divided by the total number of classified instances.
- Recall: The number of classified instances divided by the total number of instances for each class.
- F-Measure: The average between precision and recall.

In this research, correctly classified instances are 50 and accuracy in percentage is 72.4638%.

Incorrectly classified instances are 19 and 27.5362%.

This paper(Abdullah Baz et al., 2020) used bayesnet methodology; however, Randhir Singh and Surabh Pal et al.(2020) used data mining algorithms and ensemble methods to predict student performance.

[27] Randhir Singh and Surabh Pal et al.(2020) described machine learning algorithms and ensemble techniques to improve student performance prediction.The objective of this study is to finfd factors which affect low academic performance in BCA stream. In this paper, authors used three machine learning. Ensemble techniques such as bagging, and boosting are applied on

student dataset. Bagging classifier isn applied to decrease variance calculated by decision tree classifier. The purpose of this method is dividing dataset into subsets for training. Then, the average of the results obtained by each subset is considered. It performs better than single classifier. Boosting methodology is another ensemble method. It is applied to create group of classifiers. Performance measures are accuracy, recall, specificity and precision.

Table 27 represents bagging and boosting classifier accuracy, recall, specificity, precision and F-1 score. For bagging classifier, accuracy, recall, specificity, precision and F-1 score are 89.56, 70.94, 93.41, 73.41 and 68.47 respdectively. For boosting classifier, accuracy, recall, specificity, precision and f-1 score are 91.76, 76.33, 99.39, 80.55 and 65.77 respectively.

Table 27. Output of ensemble classifiers(Randhir Singh and Surabh Pal et al., 2020)

Classifier	Accuracy	Recall	specificity	precision	F-1 score
Bagging	89.56	70.94	93.41	73.41	68.47
Boosting	91.76	76.33	99.39	80.55	65.77

In this paper(Randhir Singh and Surabh Pal et al., 2020), bagging and boosting methods were used. Artificial neural network is used to predict student academic performance(Carlos Felipe Rodriguez-Hernandez etal., 2021).

[28] Carlos Felipe Rodriguez-Hernandez etal.(2021) studied artificial neural network in student performance prediction evaluation.In this research, artificial neural network outperforms other data mining methodologies. In this research, various stages of the educational data mining framework are initial preperation, statistical analysis, data preprocessing, model implementation and model evaluation. Variables in dataset are based on categories such as prior academic achievement, tution fees, student's socioeconomic status, student's home characteristics. In prior academic achievement, student;s score in seven subjects such as biology, physics, chemistry, mathematics, spanish, social sciences, and philosophy.In tution fees, how students pay for tution fees . They pay fees through one or more sources such as parents, educational loans, own resources and scholarships. Student's socioeconomic status includes parent's educational level, parent's occupation, and monthly family income. Student's home characteristics include number of rooms, it also depicts that student has computer and internet access at home or not. Other

parameters of dataset are student’s household status, student’s background information, high school characteristics, working status, university background and academic performance in higher education. In this research, tuning of various hyperparameters done . in step 1 the number of inputs, outputs, hidden layers, neurons in hidden layers and transfer function of the hidden layer is specified. In strep 2, the transfer functions of output layer is specified. Learning rate is also working as hyperparameters. Its values are 0.001, 0.0005, 0.0001, 0.00005 and 0.00001. prediction of accurate student performance, evaluation metrics are recall and F1 score. For implementing artificial neural network classification method, there are six stages such as data collection, initial preperation, statistical analysis, data preprocessing, model implementation and model evaluation. Multilayer perception has three layers namely input layer, hidden layer and output layer. Various activation functions can be selected when implementing multilayer perception for prediction. Backpropogation method occurs in two stages namely forward and second stage. The predictive weights of the multilayer perception is calculated in the forwarrd stage and input signal is transmitted to output through layers. Error signal is produced in second stage. Gradient descent optimization function is used in this paper to reduce error from mean squared error function. Learning rate and the momentum are two parameters of multilayer perception which can be tuned parameters. In this paper(Carlos Felipe Rodriguez-Hernandez etal., 2021), authors got better result by tuning parameters of neural network multilayer perception method. In this paper(Siti dianah abdul bujang et al., 2021), parameter tuning was not there. In this paper(Carlos Felipe Rodriguez-Hernandez etal., 2021), neural network multilayer perception was used. Whereas, paper(Siti dianah abdul bujang et al., 2021) used multiclass prediction method was used.

[29] Siti dianah abdul bujang et al.(2021) described student grade prediction using multiclass prediction model. In this proposed methodology, there is multiclass prediction model to reduce overfitting and misclassification results caused by imbalanced multi-classification based on oversampling synthetic minority oversampling technique with two feature selection techniques . following is the algorithm for multiclass prediction model

Input: the training dataset

output : the predicted student’s grtade label

- Import necessary library packages and select dataset
- Perform data preprocessing

    For oversampling, Select filters

    Set parameter of SMOTE(nearest neighbour, k=10)

    Select features with attribute evaluator & search method

    Select attribute selection mode.

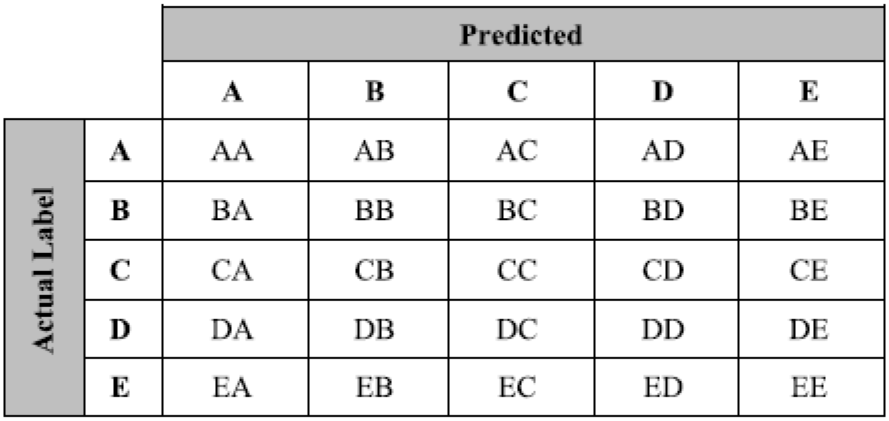
Use classification models to predict the results.

Splitting data into training and testing dataset using 10-fold cross validation

To predict grades using classification models such as j48, KNN, SVM, LR,NB and RF.

Evaluate the accuracy of classification models.

Figure 26 shows confusion matrix for student grade prediction.



		Predicted				
		A	**B**	**C**	**D**	**E**
Actual Label	**A**	AA	AB	AC	AD	AE
	B	BA	BB	BC	BD	BE
	C	CA	CB	CC	CD	CE
	D	DA	DB	DC	DD	DE
	E	EA	EB	EC	ED	EE

Figure 26: confusion matrix for student grade prediction(SITI DIANAH ABDUL BUJANG et al., 2021)

In this paper (Siti dianah abdul bujang et al., 2021), multiclass prediction with oversampling technique was there for student performance prediction. In Sourav Kumar Ghosh and Farhatul Janan, 2021), ensemble method random forest was used.

[30] Sourav Kumar Ghosh and Farhatul Janan(2021) studied student performance using ensemble method random forest classification technique. Dataset collected in this paper is from doing survey of 24 questions. Those factors are creating good notes, university facilities,

assignment submission, proper knowledge, exam strategy, group study, previous year questions. Moreover, it includes class test marks, preparation time, fear of examinations, hard questions, adaption to university. Figure 27 represents proposed methodology steps. Proposed methodology has following steps

- Identify factors through student survey
- Modification of factor rating by using fuzzy ANFIS.
- Classification of student;s performance by using random forest classifier.

After performing factor analysis based on survey, there was identification of important factors. After that rating of each attribute were merged with factor ratings.For that fuzzy ANFIS analysis is used.

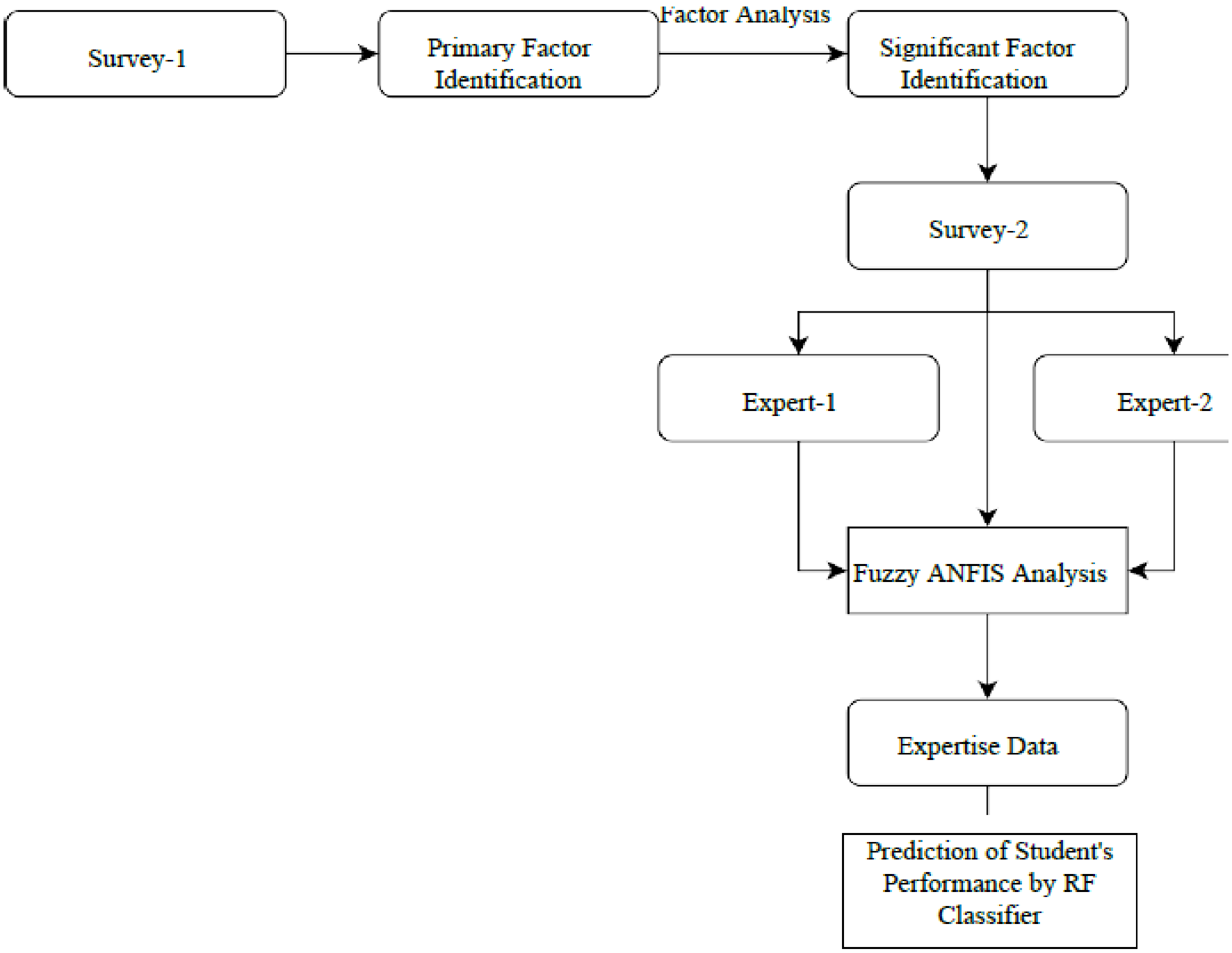

Survey-1
Primary Factor Identification
Factor Analysis
Significant Factor Identification
Survey-2
Expert-1
Expert-2
Fuzzy ANFIS Analysis
Expertise Data
Prediction of Student's Performance by RF Classifier

Figure 27. Process flowchart(Sourav Kumar Ghosh and Farhatul Janan, 2021)

In fuzzy ANFIS analysis, fuzzy logic imitates in a way that resembles human processing. Range of possibilities in fuzzy logic are certainly yes, can not say, possible yes, possible no, certainly no. After this analysis, random forest classification technique is applied.

Figure 28 represents four parts of fuzzy analysis. Those are ruule base, fuzzifier, defuzzifier and intelligence engine.

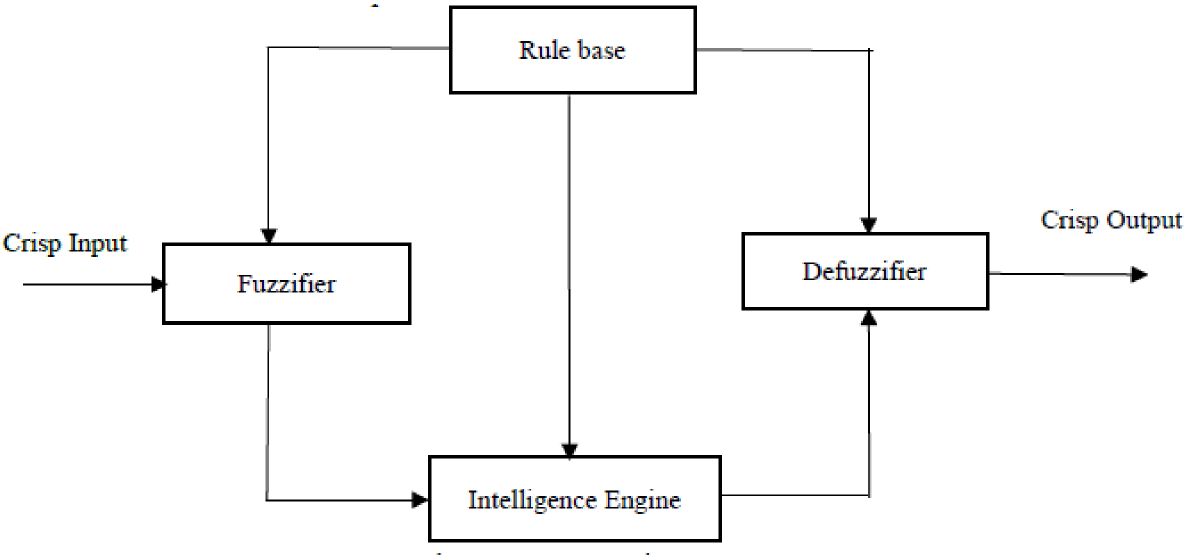


Figure 28. Fuzzy analysis structure(Sourav Kumar Ghosh and Farhatul Janan, 2021)

III. TECHNICAL DISCUSSION

3.1. Evoltionary artificial neural network(Arunachalam and Velmurugan et. al) performs better than existing artificial neural network and probabilistic artificial neural network on student performance dataset. Neural network with pattern 1-m-n is measured. 1 means one input node.N is number of output nodes. Hidden nodes value is m. Proposed methodology of evolutionary artificial neural network is based on genetic algorithm and artificial neural network. Magnitude of relative error(MMRE) values for evolutionary artificial neural network is 0.2158; however, MMRE value for Artificial neural network is 0.4384 and for probabilistic neural network, it is 0.4196. Pred(0.25) values for EVANN,ANN and PNN are 0.8012, 0.3921and 0.3693 respectiveltly. Pred(0.75) values for EVANN, ANN and PNN are 0.7843, 0.6832 and 0.6021 respectively. So Evolutionary artificial neural network performs better than Artificial neural network and probabilistic neural network due to its accurate modelling structure. In traditional artificial neural network , there os poor capability of assigning proper weights to each node under the hidden layer. So this is overcome by proposed evolutionary artificial neural network. This helps to increase prediction accuracy and reduce errors.

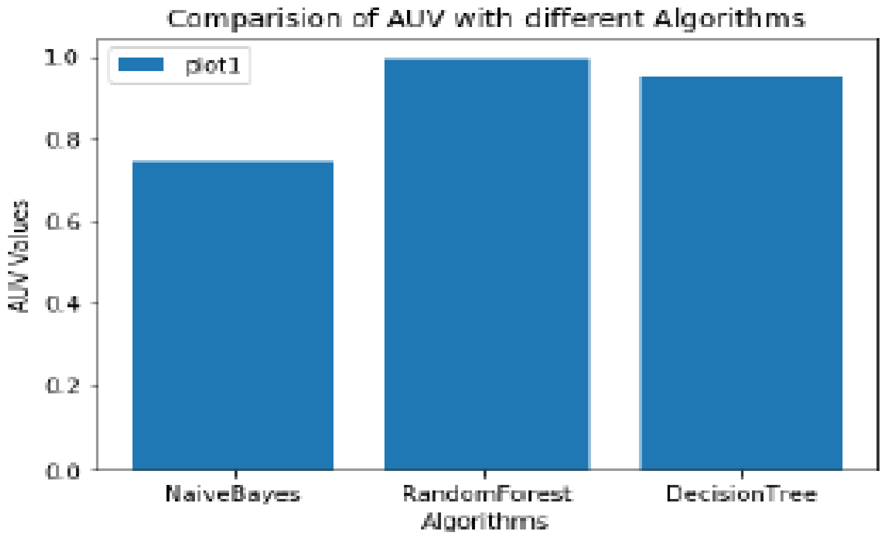
3.2. Decision tree methodology(Fergie Jonada and Reymon Rotikan,) on dataset consisting attributes namely gender, year, GPA, Fedu, Medu, Freetime, Goout, Famrel and studytime. It gives classification accuracy of 66.85% by applying 10-fold cross validation technique. Precision, recall and F-measure values are 0.622, 0.669 and 0.632 respectively. Proposed methodology of decision tree classification technique (David Kolo et. Al, 2015) on student dataset also give good accuracy and low errors which is better in student performance prediction.

3.3. Proposed random forest classification technique helps in improving classification accuracy of student performance prediction. It is an ensemble method. In this paper(P. Ajay et. al) , naïve bayes classifier and decision tree classifier is also applied on student dataset. Among all these classification methodologies, random forest got highest student performance prediction accuracy. Proposed method random forest got 99% accuracy: however, naïve bayes and decision tree classifier got 0.74 and 0.94 accuracy percentage. In paper(Kelly Spoon et. al) , there is also random forest ensemble method as proposed method. Proposed method is applied on student performance dataset which also achieved achieved highest accuracy than base classification techniques. A random forest with 10,000 regression or classification trees was used to create important variables for student performance prediction. In this, variable importance ranking is used by instructors or teachers or professors to to identify variables which are important in student success prediction. There are three methods by which there is evaluation of student performance.

- Using variable importance ranking
- Identifying variables most important to predict success of students
- Identify variables which is used in stepwise regression methods

Moreover, Mean Squared Error(MSE) is reduced by applying proposed random forest methodology on student dataset.

Figure 29 shows comparison of methodology random forest with naïve bayes and decision tree



methods.

Figure 29. Comparison of proposed methodology with naïve bayes and decision tree method(P. Ajay et al, 2017)

3.4. Proposed method of artificial neural network(ANN) (Carlos Felipe et al.,2021) has classification accuracy of 82%(high accuracy) or low accuracy 71% which is higher than other data mining methods in evaluation metrics such as the recall and F1 score. Paper(A.S Arunachalam and T.Velmurugan) has also low mean magnitude of relative error(MMRE) which is 0.2158 and highest accuracy. In artificial neural network method,

3.5. Proposed methodology of support vector machine(S.A. Oloruntoba and J.L.Akinode, 2017) has parameters namely, kernel and c . Possible values of kernels are linear, poly and RBF. By taking different c values and kernel values, training and testing accuracy vary.

Table 28 shows that When c value is 10 and according to different kernels , training and testing accuracy vary.

Table 28. Accuracy of different SVM kernel when c=10(S.A Oloruntoba et al.)

Kernel	Training	Testing
Linear	77	79
Poly	94	65
RBF	94	97

Table 29 shows that When c value is 100 and according to different kernels, training and testing accuracy vary.

Table 29. Accuracy of different SVM kernel when c=100(S.A Oloruntoba et al.)

Kernel	Training	Testing
Linear	78	75
Poly	94	51
RBF	94	98

So in this methodology, parameters of support vector machine are tuned to improve its accuracy. So by doing parameter tuning , there is increased in classification training as well as testing accuracy. SVM with parameter tuning gives better accuracy than SVM without parameter tuning. Parameters are kernel and c(penalty).

3.6. In this paper(Randhir Singh and Saurabh Pal, 2020), there is comparison of bagging and boosting methodologies. Out of these two ensemble methods, boosting performs better. Boosting provides good accuracy compare to bagging method.

Table 30 shows output of ensemble classifiers such as bagging and boosting. Performance evaluation measures are accuracy, recall, specificity, precision and F-1 score.

Table 30. Output of ensemble classifiers(Randhir Singh and Saurabh Pal, 2020).

Ensemble classifier	Accuracy	Recall	Specificity	Precision	F-1 Sore
Bagging	89.56	70.94	93.41	73.41	68.47
Boosting	91.76	76.33	99.39	80.55	65.77

In this paper(G.T. Prasanna et al., 2017), ensemble methods such as bagging and boosting are applied on student dataset to improve classification accuracy and decrease errors. Boosting algorithm also performs better. High classification accuracy prediction is occurred by boosting methodology.

Table 31 shows accuracy and error after applying ensemble methods to student dataset.

Table 31. Ensemble methods Accuracy and Error(G.T. Prasanna et al., 2017)

Ensemble classifier	Accuracy	Error
Bagging	35%	0.28
Boosting	69%	0.25

Proposed methodology boosting method(Farid Jauhari et al., 2019)

Research gap: classification accuracy on student dataset is low so it is improved by ensemble methods. In this , classification accuracy is increased by ensemble methodology such as bagging, boosting and random fores

3.7. Naïve bayes, decision tree and K-NN are applied on student dataset in proposed methodology(Hafez Mousa et al., 2017). Among all these classification methodologies, decision tree performs better. Its accuracy is highest among all three classifiers. Table 32 depicts performance measures for naïve bayes, decision tree and K-NN methods. Accuracy for naïve bayes, decision tree and k-nn classifiers are 86.80, 87.10 and 82.99 respectively. Recall values for naïve bayes, decision tree and k-nn are 17.39, 13.04 and 34.78 respectively. 53.33, 60.00 and 36.36 are precision values for naïve bayes, decision tree and k-nn respectively.

Table 32. Classification results(Hafez Mousa et al., 2017)

Evaluation Measure	Naïve Bayes	Decision Tree	K-NN
Accuracy	86.80	87.10	82.99
Recall	17.39	13.04	34.78
Precision	53.33	60.00	36.36
F-Measure	26.23	21.43	35.36

3.8. By tuning parameters such as number of clusters k and cluster size in k-mean methodology, overall performance vary(Oyelade et al., 2010). It helps to get more accurate performance by tuning parameters of k-mean algorithm. Table 33 shows overall performance value for cluster size 24 and 16. By choosing cluster size 16, there is better performance. Overall performance for cluster size 24 and 16 are 50.08 and 65.00 respectively.

Table 33. For number of cluster k=4, overall performance(Oyelade et al., 2010)

	Cluster size	Overall performance
1	24	50.08
2	16	65.00

Table 34 shows overall performance value for cluster size 19 and 17. By choosing cluster size 19, there is better performance. Overall performance for cluster size 19 and 17 are 49.85 and 60.97 respectively.

Table 34. For number of cluster k=5, overall performance(Oyelade et al., 2010)

	Cluster size	Overall performance
1	19	49.85
2	17	60.97

IV. CONCLUSION

In this report, I have reviewed 17 journal papers . the main objective of this study is to better prediction of student academic performance and reducing errors namely Root mean square, Relative absolute error and mean absolute error. Educational data mining affects various parts of education industry. It is advantageous in predicting student performance, predicting student profiling, planning and scheduling.This paper has reviewed 17 journal papers related to educational data mining to predict student academic performance.basic classification techniques namely decision tree, naïve bayes, K nearest neighbour, support vector machine are used for prediction. To enhance accuracy and reduce errors such as Root mean square error, relative absolute error, mean square error, further ensemble methodologies are applied on student performance datasets.

In decision tree methodology, IBM statistical package for social studies(SPSS) is used to apply chi-square automatic interaction detection(CHAID) in producing decision tree structure. Algorithm produce more accurate result if higher version of SPSS is used. Support vector regression methodology has parameters namely penalty parameters(C) and kernel functions. By tuning kernel parameters such as linear, poly and RBF there is improvement in prediction accuracy and reduction in errors. Due to poor modeling structure and the capability of assigning proper weights to each node in the hidden layer in traditional artificial neural network, Evolutionary artificial neural network performs better than existing artificial neural network and probabilistic artificial neural network. So this issue is solved by using genetic algorithm optimization approach. In artificial neural network, back-propagation training algorithm is used for training dataset. Two parameters of multilayer perception are adjusted during back-propagation algorithm, learning rate and the momentum. the value of weights are changed by learning rate.speed of the learning process is increased by momentum. This hyperparameters tuning helps to improve accuracy. By using softmax function as transfer function of the output layer , training duration is least. So it concludes that the softmax function performs faster towards the minimum error in comparison with two other transfer functions namely sigmoid and linear sigmoid.

It is concluded that type of school is not most influenced attribute to influence student performance. Other attribute namely parent’s occupation has huge impact on predicting student academic performance. By monitoring student’s performance semester by semester, it helps to improve academic result prediction. It also helps to increase prediction accuracy and reduce errors namely Root mean square error(RMSE), mean absolute error(MAE), and relative absolute error(RAE). In kmean algorithm, by varying k values and cluster sizes, there is improvement in prediction accuracy and reducing errors. Student performance prediction is enhanced by applying ensemble methods on educational student datasets. It helps with reduced errors and it helps early identification of student at risk of attrition. Ensemble classifier with three data sources(Banner + moodle+survey) is the best and then ensemble classifier with two data sources(banner + moodle models performance) are considered best in terms of accuracy. K mean clustering algorithm has disadvantage that clusters or centroid does not converge that it can go into infinite iteration; however, enhanced k-strange point clustering method is used until iteration depends on number of clusters. So limitation of kmean clustering is solved by enhanced k-strange clustering algorithm. It is concluded that in decision tree classification method, this technique is not considering all parameters and attributes of the dataset so on some student datasets, it does not perform good and it does not produce better result.

Decision tree is mostly robust to outliers. Due to overfitting, sampling error occurs. If sampled preperation data is different than evaluation data, then decision tree s not to generate good results. By applying bayesian algoritms on student performance dataset, there is improvement in prediction accuracy. Accuracy is affected by many factors. The number of attributes and cleaning dataset affect accuracy result.By removing incomplete answers, cleaning datasets and removing less cooelated questions, accurate result is obtaind. Random forset algorithm performs better than decision tree method. It provides better accuracy and low errors. Student database parameter school does not influence student performance, but parent’s occupation has huge influence on student performance.

Notation or nomenaclature :

EDM	Educational Data Mining
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
RAE	Relative Absolute Error
SPSS	statistical package for social studies
Fedu	Father’s Education
Medu	Mother’s Education
Goout	time spent going out with friends
Famrel	quality time spent with family
MMRE	Mean magnitude of Relative Error
EVANN	Evolutionary artificial neural network
ANN	Artificial Neural Network
PNN	Probabilistic artificial neural network
CHAID	chi-square automatic interaction detection

REFERENCES

[1]Oyelade, Oladipupo, Obagbuwa(2010). Application of k-Means Clustering algorithm for prediction of student's Academic Performance, International Journal of Computer Science and Information Security, 7(1), 292-295.

[2]Md. Hedayetul Islam Shovon, Mahfuza Haque(2012). Prediction of Student Academic Performance by an Application of K-means Clustering Algorithm, 2(7), 353-355.

[3] D. Magdalene, Delighta Angeline(2013). Association Rule Generation for Student Performance Analysis using Apriori Algorithm, Standard International Journals, 1(1), 12-16.

[4] V. Ramesh, P. Parkavi, K. Ramar(2013), Predicting Student Performance : A statistical and Data Mining Approach, International Jpournal of Computer Applications, 63(8), 35-39.

[5]M. Durairaj, C. Vijitha(2014). Educational Data mining for Prediction of Student Performance Using Clustering Algorithms, 5(4), 5987-5991.

[6]Abeer Badr EL Din Ahmed(2014), Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), 43-47.

[7]Samy Abu Naser, Ihab Zaqout, Mahmoud Abu Ghosh, Rasha Atallah, Eman Alajrami(2015). Predicting Student Prerformance Using Artificial Neural Network:in the Faculty of Engineering and Information Technology, International Journal of Hybrid Information Technology, 8(2), 221-228

[8]Sayana T S(2015), Prediction of Students Academic Performance using Data Mining, International Journal of Engineering Research & Technology, 3(30), 1-4.

[9] Hashmia Hamsa, Simi Indiradevi, Jubilant J. Kizhakkethottam(2016). Student academic performamnce prediction model using decision tree and fuzzy genetic algorithm, Elsvier, 25, 326-332.

[10]Elaf Abu Amrieh, Thair Hamtini, Ibrahim Aljarah(2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods, 9(8), 119-136.

[11]Amjad Abu Saa(2016). Educational Data Mining & Student's Performance Prediction, 7(5), 212-220.

[12] S.A. Oloruntoba, J.L. Akinode(2017) ,STUDENT ACADEMIC PERFORMANCE PREDICTION USING SUPPORT VECTOR MACHINE. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 6(12), 588-598.

[13] Snehal Bhogan, Kedar Sawant, Purva Naik, Rubana Shaikh, Odelia Diukar, Saylee Dessai(2017). PREDICTING STUDENT PERFORMANCE BASED ON CLUSTERING AND CLASSIFICATION, IOSR Journal of Computer Engineering, 19(3), 49-52.

[14] Alaa Khalaf Hamoud, Aqeel Majeed, Wid Akeel Awadh, Ali Salah Hashim(2017), "Student's Success Prediction based on Bayes Algorithms", International Journal of Computer Applications,178(7), 6-12.

[15] Devine Grace Doble Function(2017). Predicting Student Academic Performance in Computer Organization Course:Using J48 Algorithm, Indian Journal of Science and Technology, 11(47), 1-8.

[16] STEPHEN J.H. YANG, OWEN H.T. Lu, ANNA Y.Q. HUANG, JEEF C.H. HUANG, HIROAKI OGATA, ALBERT J.Q LIN(2017). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis, Journal of Information Processing, 26, 170-176.

[17]Hafez Mousa, Ashraf Maghari(2017), School Student's Performance Prediction Using Data Mining Classification, International Journal of Advanced Research in Computer and Communication Engineering, 6(8), 136-141.

[18]Olugbenga Wilson, Thomas Connolly(2018). Predicting student academic performance using multi-model heterogeneous ensemble approach, Journal of Applied Research in Higher Education, 10(1), 61-75.

[19] G. Sujatha, Sindhu.S, Savaridassan. P(2018). PREDICTING STUDENTS PERFORMANCE USING PERSONALIZED ANALYTICS, International Journal of Pure and Applied Mathematics, 119(12), 229-238.

[20] Mr. Shubham Agrawal, Mr. Kapil Sahu.(2018) Prediction of students Performance Using K-Means Algorithm, International Journal of Research in Electronics and computer engineering, 6(3), 1604-1607

[21] .S. Arunachalam, T.Velmurugan(2018). Analyzing student performance using evolutionary artificial neural network algorithm, International Journal of Engineering & Technology, 7, 67-73.

[22]Raza Hasan, Sellappan Palaniappan, Abdul Rafiez Abdul Raziff(2018). Student Academic Performance Prediction by using Decision Tree Algorithm, International Conference on Computer and Information Sciences, 1-5.

[23] Mehta Smruti Hemantkumar, Ashish Adholiya(2019). Predicting Students' Performance using J48 Decision Tree, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 4(4), 123-127.

[24] Farid Jauhari, Afif Supianto(2019). Building Student's Performance Decision Tree Classifier Using Boosting Algorithm, Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1298-1304.

[25] P.Ajay, M. Pranati, M. Ajay, P. Reena, T. BalaKrishna(2020), "PREDICTION OF STUDENT PERFORMANCE USING RANDOM FOREST CLASSIFICATION TECHNIQUE", International Research Journal of Engineering and Technology, 7(8), 405-408.

[26]Abdullah Baz, Fatima Alshareef, Ebtihal Alshareef, Hosam Alhakami, Tahani Alsubait(2020), Predicting Student's Academic Performance Using Naïve Bayes,International Journal of Computer Science and Network Security, 20(4), 182-190.

[27]Randhir Singh, Saurabh Pal(2020), Machine Learning Algorithms and Ensemble Techniques to Improve Prediction of Students Performance, International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 3970-3976.

[28] Carlos Felipe Rodriguez-Hernandez, Mariel Musso, Eva Kyndt, Edurado Cascallar(2021), “Artificial neural networks in academic performance prediction : Systematic implementation and predictor evaluation”, Computers and Education: Artificial Intelligence, 1-14.

[29] Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej krejcar, Enrique Herrera, Hamido Fujita, Nor Azura Md. Ghani(2021). Multiclass Prediction Model for Student Grade Prediction Using Machine Learning, 9, 95608-95621.

[30]Sourav Kumar Ghosh, Farhatul Janan(2021), Prediction of Student’s Performance Using Random Forest Classifier, International Conference on Industrial Engineering and Operations Management Singapore, 7089-7100.