OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 01 - Issue 12 (2024)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

Exploring the Predictive Potential of Educational Data Mining: A Study on University Students' Academic Performance

Faizan Sattar
Department of Computer Science, University of Sargodha, Punjab, Pakistan

Assad Latif
School of management and engineering North China University of water resources and electric power Zhengzhou Henan, China

Saima Ali Batool
Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.

Ahmad Murad
Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.

Meiraj Aslam
Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.

Abdul Manan Razzaq
Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.

Corresponding Author: Faizan Sattar. Email: faizan.sattar.khan@gmail.com

DOI: *https://doi.org/10.71146/kjmr163*

Article Info

Abstract

Predicting student's performance become more tough due to a huge volume of data in educational database. Even though data mining has been efficiently implemented in the business world. Use of data mining is relatively new in higher education, i.e., it used for recognition and extraction of potentially and advanced precious knowledge from huge amount of data using the concept of data mining. Model should be developed which can determine the conclusion from student's academic success. In this study, the concept of educational data mining have been used. It is a computational process of discovering patterns in a large dataset and transform it into understandable structure for further use. Student data collected from different institutes. The scope of this research is to identify the factors influencing the performance of students in universities. To predict the performance of university student, for clustering purposes K mean algorithm and two algorithms of classification Decision tree (J48) and Naive Bays have been used, also compare the results of Decision tree J48 and Naive Bay's algorithms which gives the better results according to their accuracy . For this purpose, Weka tool was used to implement the algorithms. This study helps the educational institutions to identify such factors which affect the student performance and to eliminate these causes to enhance their performance.

Keywords: *Classification, Student Performance Analysis, Machine Learning, Large Volume Analytics*

Introduction

Data mining commonly called knowledge discovery that is the way toward dissecting information from alternate points of view and condensing it into helpful data - data that can be utilized to build income, cuts costs, or both. The software of data mining is one of various diagnostic apparatuses for investigating information. It enables clients to break down information from a wide range of measurements or edges, classify it, and compress the relationships distinguished. In fact, data mining is the way toward discovering connections or arrangements among many fields in extensive comparative (relative) databases.

Despite the fact that data mining is a generally new term, the innovation is definitely not. Organizations have utilized effective PCs to filter through volumes of store scanner information and examine statistical surveying reports for a considerable length of time. Be that as it may, nonstop developments in PC handling power, storage of hard disk, and software of statistical are drastically expanding the exactness of investigation while dropping down the cost. Data mining is to store the knowledge involved by individuals from a large amount of data, and this learning is the essential data however demonstrative and earlier unbeknown. The strategies of data mining are obtained to fulfil application in many fields, and utilization of methods of data mining in university can quicken the advancement and improvement of the educational framework. The innovation of data mining can discover valuable learning from a huge amount of data, and this learning gives essential establishment to enhance administration arrangements of university. The classification and other is clustering is a basic task of data mining which widely used for different purposes like data categorization and investigation in both industry and educational institutions. Clustering is the way toward arranging objects which are unlabeled into gatherings which individuals are comparative somehow. (Larose 2006) Classification exists in algorithm of supervised learning. The concept of data mining is used for different purposes like business, government and scientific research. There are two types of data mining learning methods .one is supervised and other is unsupervised.

Mining of data or extraction of data, additionally prominently known as Knowledge Discovery in Databases (KDD), alludes to the nontrivial extraction of verifiable, already obscure and possibly valuable Information from information in databases. While information mining and learning disclosure in Databases (or KDD) are oftentimes regarded as equivalent words, information mining is entirely of the learning revelation process. With the advancement of information disclosure procedures, the estimation of the information is fundamentally improved.

In this area there are some hurdles that has a greater effect on application of Education data mining about their performance

- Comprehension: Concentration on ideas of organization, translation and analysis of information.
- Knowledge: concentration on memorization, acknowledgement and what's more, review of data.
- Application: concentrates on critical thinking, utilization of particulars, and standards.
- Analysis: concentrates on finding the hidden association, and the division of an entire into segments.
- Synthesis: concentrates on a blend of thoughts to make something new, making something one of a kind whether physical or verbal.
- Evaluation: concentrates on making judgments about issues, settling differences or contradictions.

1. Related work

Barata et al., stated that in past work we have recognized students of four different types, portrayed by gaming qualities, behavior, particular performance and engagement levels.

In this paper we displayed a novel examination where we contemplated how we can exploit what separates diverse sorts of students, in a gamified setting, to anticipate their behavior and efficiency by midterm. To this impact we broke down information of student covering both performance estimations and gaming inclinations, from one occurrence of the course, and utilized it to recognize relevant properties and prepare classifiers to test with information from another term. Our review demonstrates that the understudies' sort can be anticipated with up to 79 percent exactness by midterm, utilizing only data of performance. In any case, information involving both performance measurements and the understudies' player characterization as per the Brain Hex model was more precise in before focuses in time, giving 66 percent exactness after seven weeks and 47 percent even following five weeks of class. Different types of algorithms used in this study that's are given below J48, Bayes Net, SMO, IB1 and simple logistic model of regression, for experimental purposes WEKA tool used. From this review we took in a profitable lesson. Obviously, in the specific instance of our analysis, where different students were processed in light of execution accumulation, we anticipated that the best indicators would be metrics of performance. In any case, we discovered that prior in the course, performance of students gives off an impression of being less discriminative and its prescient power can be enhanced by combine it together with information that can be measured already, for example, their type of players. We trust our review lays vital preparation for the advancement of robust gamified learning situations. These oughts' to draw on achievement and gaming information to distinguish profiles of different types of students in close real time, which could be utilized to expeditiously adjust substance to fit the needs of students and would be an imperative tool to evaluate the progress of both the instructors and students.

Sorour et al stated that most of the predictive models are difficult for teachers to interact. New method is described to build an interpretable performance prediction model for the students which based on comment data mining. Comment data has been used to understand the behavior of student and situation .In current study the main six attributes denote the relationship between different set of interrelated variables of the student .For this purposes white box method (Decision tree and Random Forest implemented through WEKA open source learning tool) used .All algorithms are than implemented individually and compared their results to optimize these algorithms by using F-measure .overall prediction result of RF is better as compared to DT These findings helps the teachers to prepare their lesson and focus on the attributes of each group . The proposed strategy has two restrictions: in the first place, 'black box' and 'white box' models are not antipodes. On the off chance that we can join both methodologies in a general procedure, the benefits of both ideal models will be consolidated, and the disadvantages are disposed of. Second, we fabricated the glossary of traits manually which separated the mutual variables from every one of the understudies and the words identified with each trait. Our heading for future work is to quantify which show delivers higher forecast results and pick up the best arrangement of standards. New experiments required for these purposes. Like, link the Neural Network and Genetic Algorithm. For the second, we will recommend a strategy for building a glossary that reflects the remarks of students. To this end, we will gather new remarks to concentrate qualities and explore the contrasts between the current and the new remark information. This trial will help educators to concentrate on the most vital traits and give programmed criticism to understudies after every lesson to enhance their execution.

Singh et al proposed that universities and many technical organizations are facing the problem in term of analyzing their performance. The real difficulties are in placements of students, their admissions and curriculum. The two most imperative process over which information's are gathered and investigated are student's placement and their admissions. The

Ranking of universities depend on the placement of students and performance in term of study. Data set collected through the student Performa. Aside from academic achievements there are different variables which help in understanding the student's general achievements. To understand the student's behavior clustering analysis is used. open-source software Rapid Minor Studio used for that analysis also used for business analysis, machine learning, and data mining and predictive analysis..

Badr et al stated that Data mining in educational sector is a developing field that uses the information acquired from instructive data frameworks to find learning and discover answers to inquiries and issues concerning the education framework. High dropout rates and poor scholarly execution among understudies are cases of the most widely recognized issues that influence the notoriety of educational institutes. The academic record of students can be breaking down to investigate the elements behind these marvels. They presented a model of knowledge to predict the student's performance in programming courses that depend on the previous courses of different subjects. The main objective of this model is to enhance the quality of educational institutions by using classification method that told those students who are at risk of dropout .For this purposes teachers' pay an important role by adjusting their teaching method .This model helpful for the student's to predict their performance in programming subject before choose that subject .For the experiment purpose CBA algorithm used that generate rules .Result for experiment used of mathematics and English of the graduated students from king Saud university session 2008 to 2014 .Firstly the grades which used in two English courses and two mathematics generate four rules with 62.75% accuracy rate .secondly grades that used in two English courses also generate four rules with 67.33% accuracy rate .English courses have greater effect to predict the performance of programming course .

Sorour et al. stated that today, the utilization of learning of analytics is becoming more vital in the learning condition with the end goal of comprehension and upgrading student's learning circumstances. The goal of this research is to inspect the effects of Teacher Interventions (TIs) on achievements and attitude of students convoluted with the lesson by dissecting their freestyle remark information after each lesson. The present review proposes another strategy for building an open expectation model of prediction, which speaks to student's exercises, circumstances and perspectives; the strategy classifies words in the remarks of student into six characteristic sorts and shows the most critical sorts that influence the forecast comes about. Promote, the predicted results are contrasted and the topic-based method of statistical that utilize Latent Dirichlet Allocation and model of Support Vector Machine. The outcomes demonstrated that there were sure relationships amongst TIs and the nature of composing remarks that effect on enhancing the result of prediction. The results of those experiments represent AM method gives better results of prediction. This research manages on the bases of computer application course and also applied to another course.

2. Method and materials

2.1. Data Mining

Generally, Data mining is of about preparing information and recognizing examples and patterns in that data so you can choose or judge. The standards of data mining have been around for a long time, however, with the appearance of enormous information, it is much more common. A lot of information caused a blast in the utilization of more broad data mining methods, halfway on the grounds that the extent of the data is considerably bigger and in light of the fact that the data has a tendency to be more changed and broader in its exceptionally nature and substance. Data mining (DM) is a very important tool and helps the educational institutes to discover the performance of students and predicted the future academic performance. This research, a new predictive method will be proposed for predictive based analysis of student's data of educational institutes. The data that will be used for predictive based analysis will be retrieved from

different educational institutes. Important attributes will be extracted from the collected data from different educational institutes, converting it into digital format to form dataset and stored by using any DBMS for analyzing and prediction purpose. The comparative information will be assembled into various classes and diminishes the comparability of interclass. There is no predefined classes are accessible for doling out information. The concept of Statistical or measurable ideas is utilized to relate distinctive objects consequently. The assignment of collection is not recently just to execute the calculations in which class mark of each example not knows and distinctive classes are preparing. The choosing criteria of gatherings rather the other resembles arrangement, association and substance extraction is that every one of the investigations of understudy's information, the procedure of clustering give very noteworthy outcomes.

- Clustering (Bunching) is a strategy for unsupervised learning, and a typical procedure for measurable information investigation utilized as a part of many fields, encompasses the machine learning, information mining, design acknowledgment, analysis of image, data recovery, and Bio-informatics. To identify the task of grouping the records which are similar to each other but must be different from others and identifying those variables which will provides efficient clustering. These clusters will helpful to discover the students committed by one or same types of individuals (Chen *et al.*, 2011).
- The algorithm of clustering K-Means will be utilized to look specifically of n articles without utilizing various leveled(hierarchical) affiliation. This algorithm really purposed by J. Macintosh Queen in 1967. K-Means begins from choosing K components through which the underlying seeds is shaped. There are many ways for selecting K elements like selecting the k-preliminaries, randomly select the k observation and also select the k observation have distinct values. The principal reason of choosing the algorithm of k-mean is that in numerous pasts explores, the algorithm of k-mean is turned out to be an outcome situated algorithm. In addition, this calculation is very effective to deal with the information of educational institute students with huge volume and low computational cost.

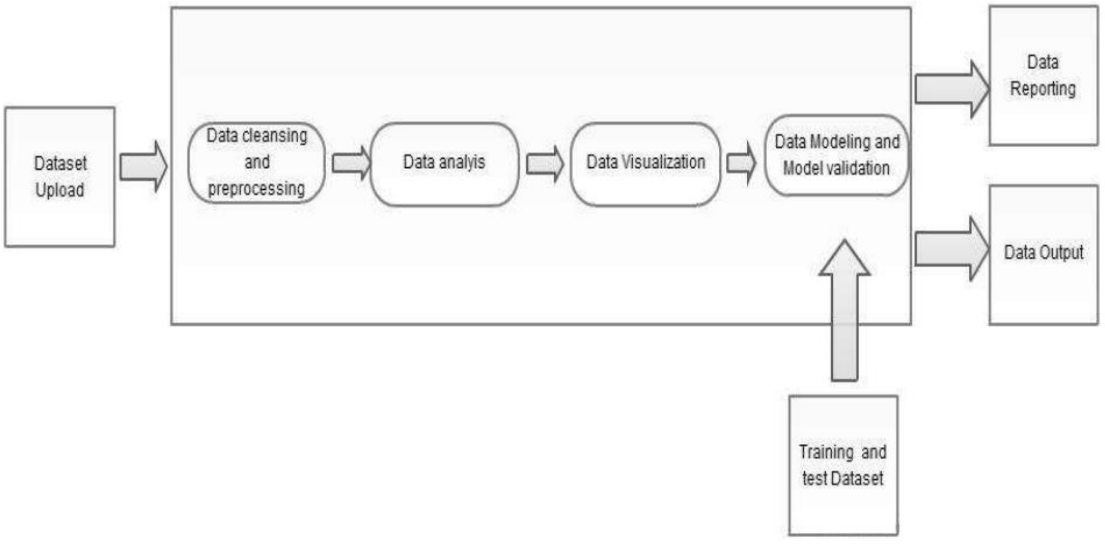

Figure 1: K-mean Clustering

2.2. Multi-Layer Perceptron

Neural networks are utilized as a part of the bland territories of expectation, anticipating, order and issues of classification. The yield from these methods is a certainty level in the vicinity of zero and one, the more like one means that how great the classification procedure has been. A wide assortment of down to earth application zones utilizes this method some of which are: - educational institutes, medical, utilities, recreation, design acknowledgment, mechanical autonomy, agriculture, legal, manufacturing, condition monitoring, speech and text processing, and chemical analysis. One type of neural network is the Multi-Layer Perceptron (MLP), which is an administered grouping system being able to shape classes in view of gaining from cases inside the information. Utilizing known information which are isolated into a preparation set and a testing set, the system is prepared on the previous and tried on the last which it has not already observed(Krikke & Alfonsi 2006). One specific example where MLP's are successfully used, which may be considered similar to the work undertaken within this study, is in the area of fraud detection. Many of the High Street banks and credit card companies use such technologies to reduce fraud. Within excess of 400,000 transactions every day, a devaluation of only 2.5% of fraudulent transactions translates into a saving of $1 million per year and 80% correctly identified fraud transactions is considered a good result for this type of problem. Difficulties can be experienced when using MLPs. To create such a network there will be an input and output layer together with one or more hidden layers and each of which will contain a huge number of nodes. Establishing the correct number of layers and nodes can be problematic; however, software tools can assist in this area.

3.2.1 Self Organizing Map

Self-organizing outlines are utilized to group the data and are utilized in a variety of practical arenas. Examples include space exploration, signal processing/robotics, language processing, utilities/manufacturing, medical, finance, document classification, commerce and chemical analysis. Clustering techniques have been used in student's performance analysis of educational institutes to find links between the performances of students whose are near about to dropout and other students are found by using the dataset that collected from different educational institutes. A Self Organizing Map (SOM) is a sustain forward neural system that uses an unsupervised preparing calculation, and through a procedure called self-association, designs the yield units into a topological portrayal of the first information. It utilizes the numerous factors in the informational index (multi measurements) and perceives examples to frame a 2-dimensional framework of cells. This strategy has likenesses to the measurable approach of Multi-Dimensional Scaling.

2.3. Dataset Collections

The data employed in this study was collected from the students of different institutes. Manual information is gathered and modernized it according to the prerequisite of this study. Table 1 shows the attributes extracted from the manual data.

Table 1: Attributes

Serial No.	Attribute	Description
1	school	Student's institute
2	sex	Student's sex (binary: 'F' - female or 'M' - male)
3	age	Student's age (numeric: from 15 to 22)
4	address	Student's home address type (binary: 'U' - urban or 'R' - rural)
5	famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	Parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7	Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€ “5th to 9th grade, 3 â€ “secondary education or 4 â€ “higher education)
8	Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€ “5th to 9th grade, 3 â€ “secondary education or 4 â€ “higher education)
9	Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
10	Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
11	Guardian	student's guardian (nominal: 'mother', 'father' or 'other')
12	Travel time	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
13	Study time	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
14	schools up	Extra educational support (binary: yes or no)
15	Famsup	Family educational support (binary: yes or no)
16	Paid	Extra paid classes within the course subject (binary: yes or no)
17	Activities	Extra-curricular activities (binary: yes or no)
18	Higher	wants to take higher education (binary: yes or no)
19	Internet	Internet access at home (binary: yes or no)
20	Famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
21	Free time	Free time after school (numeric: from 1 - very low to 5 - very high)
22	Go out	Going out with friends (numeric: from 1 - very low to 5 - very high)
23	Health	Current health status (numeric: from 1 - very bad to 5 - very good)
24	Absences	Number of school absences (numeric: from 0 to 93)

More than 2400 records have recorded; however, still there are a number of records stored manually awaiting to be entered to the automated system. Although theoretically large volume of data is more important to train data mining models, due to time constraint to preprocess the data, a sample data containing 1100 records are taken for this study. The records of 651 students have been selected for analysis so that an effective model can be build and tested.

3.3.1 Data Pre-Processing

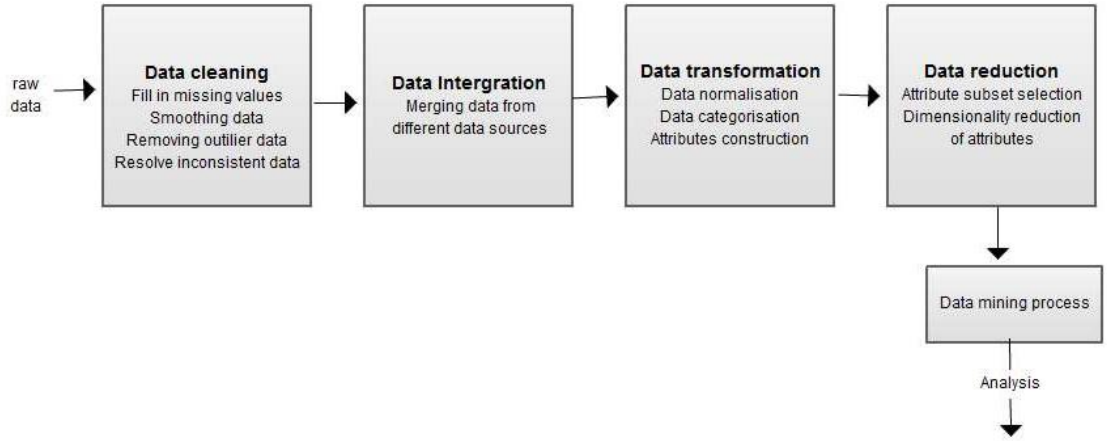

Figure 2: Data Preprocessing

The information of understudies, which is utilized with the end goal of this review, experiences various constraints. These Truth be told, as expressed by Witten and Frank (2000), one of the basic issues in building the model of data mining is restrictions in the information itself. In this manner, an ideal model could be developed once complete, perfect and robotized information is all around arranged.

3.3.2 Data cleaning

The dataset of students has 32 factors which are appeared below and information sorts are incorporated. Information was changed over as csv and furthermore an exceed excel document. The goal of my study did not require every one of the factors recorded henceforth there was requirement for data arrangement, decrease and pre - handling. Information lessening is performed by choosing the most instructive properties in a dataset, while endeavoring to lose no critical data for order. There was requirement for expulsion of the factors which I didn't require. From the 32 factors just 24 were useful for the investigation. The selection of attributes depends on human comprehension of informational collection. When managing an extensive number of attributes to build a decision by using human knowledge as a practical and furthermore considered that exclusive those attributes are picked which don't contain any missing esteems.

3.3.2 Integration of data

incorporate missing values, anomalies and encoding irregularity in different trait values.

For the purposes of analysis, it requires 2 datasets the understudy dataset and the dataset of weather thus there was requirement for integration of data. This was attainable by having the capacity to recognize factors which coordinate in both informational indexes and incorporating the information basing on those factors. For this situation the reason for my incorporating was the month and year where these factors were the same. The understudy informational collection demonstrated the year and month where a specific low achievement of understudy. The weather informational collection additionally had the month and year when certain climate designs happened.

3.3.3 Transformation of data

The transformation of data incorporates standardization and accumulation. In the understudy informational dataset there was have to normalize the information to transfer into Weka. I likewise ordered the time into PM and AM in dataset. These changes enabled to load information into Weka for the purpose of analysis. The principal phase of my study was to group the factors into identifiable groups. This procedure changed the information into an arrf document and which utilized further in my research for arrangements. The specimen of the arrf document is shown below:

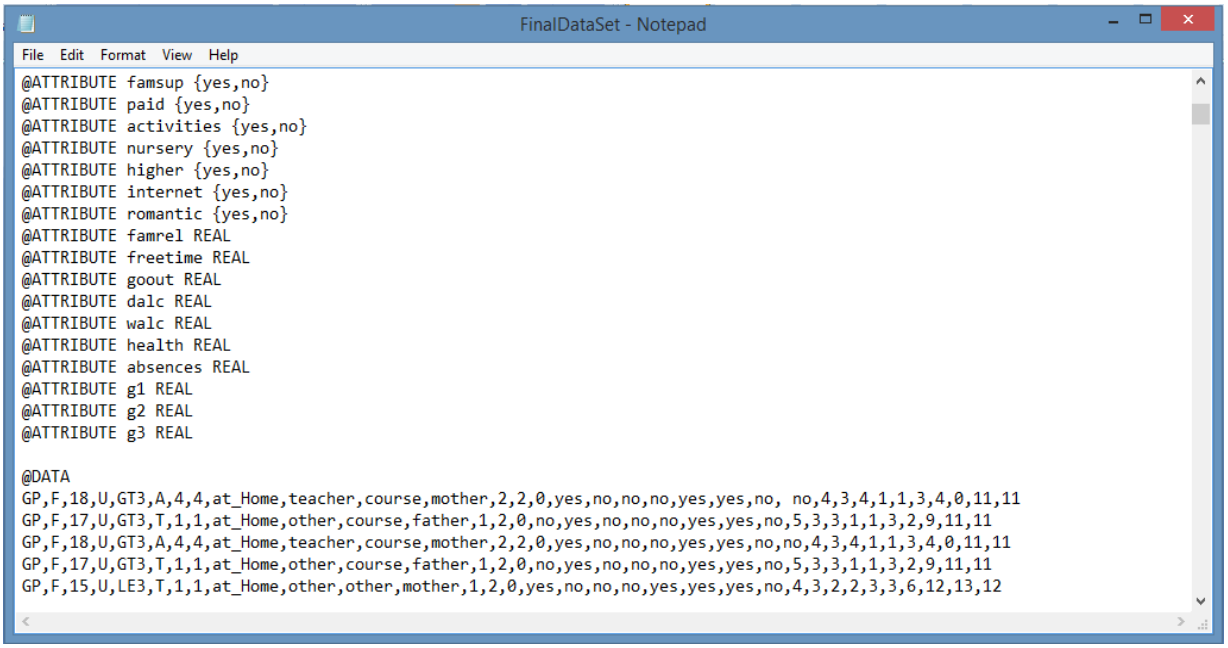

Figure 3: Arrf file

3.4 Main Algorithms Used

The major algorithm utilized as a part of this paper is the algorithm of clustering K-mean, DT J48 and NB. The recommended model for the purposes of analysis incorporates clustering the information at first in order to assemble comparable qualities together. The grouped information will add another class to the other attribute clusters. After the grouping we can then characterize the clustered information. Figure 4 clarifies all the procedure. The working of algorithm is explained in following section given below.

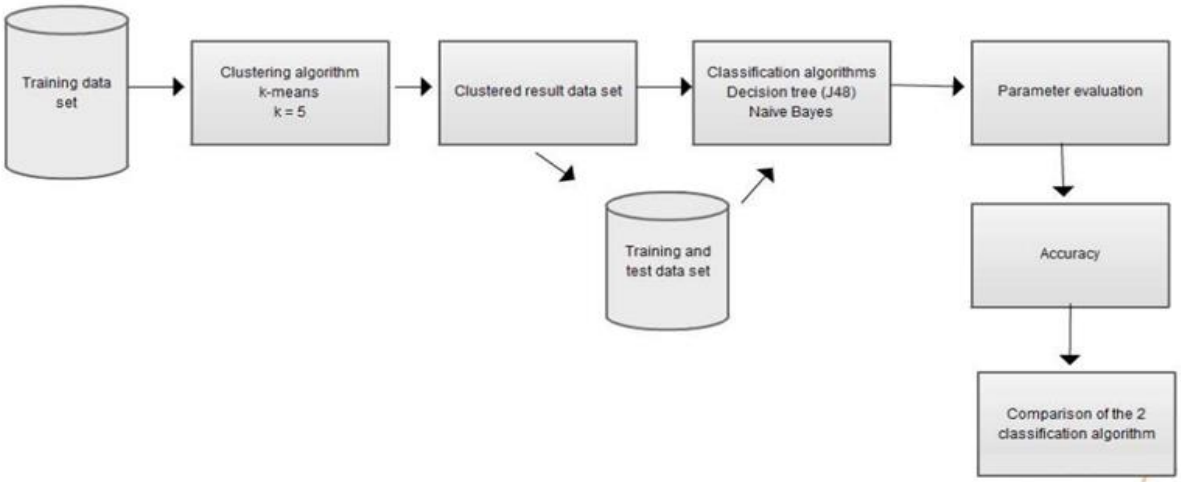

Figure 4: Recommended model of classification

3.4.1 Clustering algorithm K-Means

Clustering is a strategy for unsupervised learning, and a typical procedure for statistical analysis of data utilized as a part of many fields, including data mining, design acknowledgment, data recovery, and analysis of image, machine learning and Bio-informatics. To identify the task

of grouping the records which are similar to each other but must be different from others and identifying those variables which will provides efficient clustering. These clusters will helpful to discover the students committed by one or same group of individuals (Rohe et al., 2011). The algorithm of clustering K-Means will be utilized to seek straightforwardly of n substances without utilizing various leveled association. Macintosh Queen in 1967 proposed this algorithm. K-Means begins from choosing K components through which the underlying seeds are build. There are many ways for selecting K elements like selecting the k-preliminaries, randomly select the k observation and also select the k observation have distinct values. The principal reason of choosing the algorithm of k-mean is that in numerous past researches, the algorithm of k mean is turned out to be an outcome-oriented algorithm. In addition, this calculation is very productive to deal with prediction of student's performance with huge volume and low computational cost (Wang nad Lin, 2011).

3. Implementation and results

The fundamental reason for this study is to investigate the utilization of data mining applications in education to predict and analysis the performance of the students. To accomplish this, sample data of 1100 student were taken arbitrarily and 651 records have been selected for analysis so that an effective model can be build and tested. Following image show the pre-process phase of Weka tool.

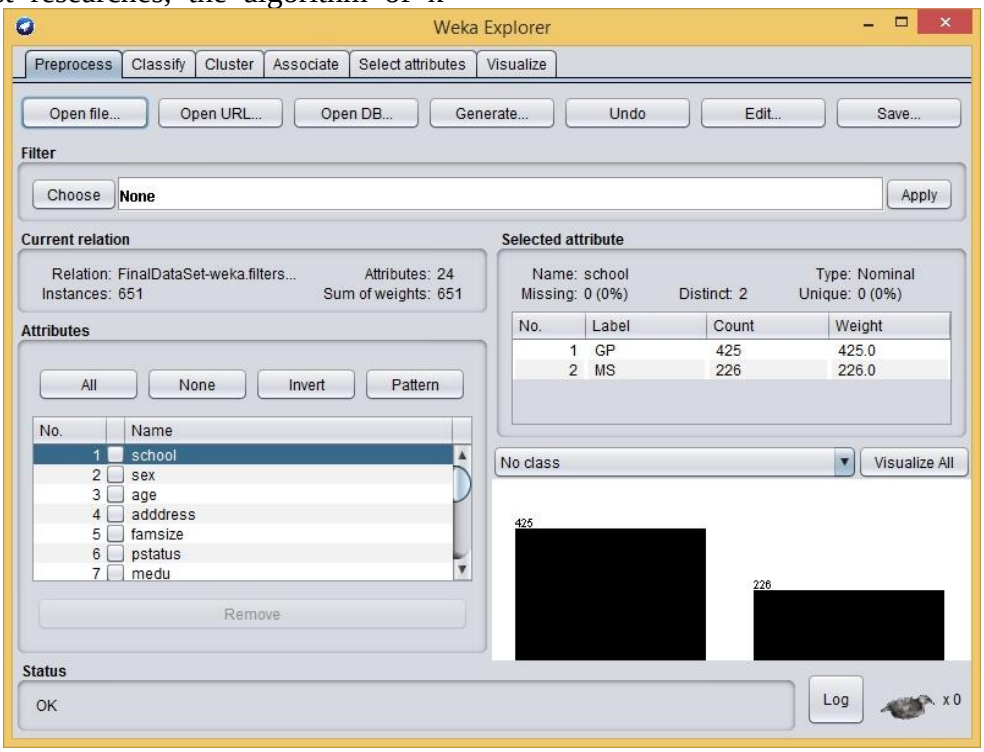

Figure 5: Preprocess phase of WEKA tool

Figure 6: Graphically representation of attributes

In this study, more classification techniques are used in the method of data mining to forecast the student's achivements. This approach of classification are utilized because a wider look and understanding of final and exit results,and hat will lead to a comparative degree of study results.

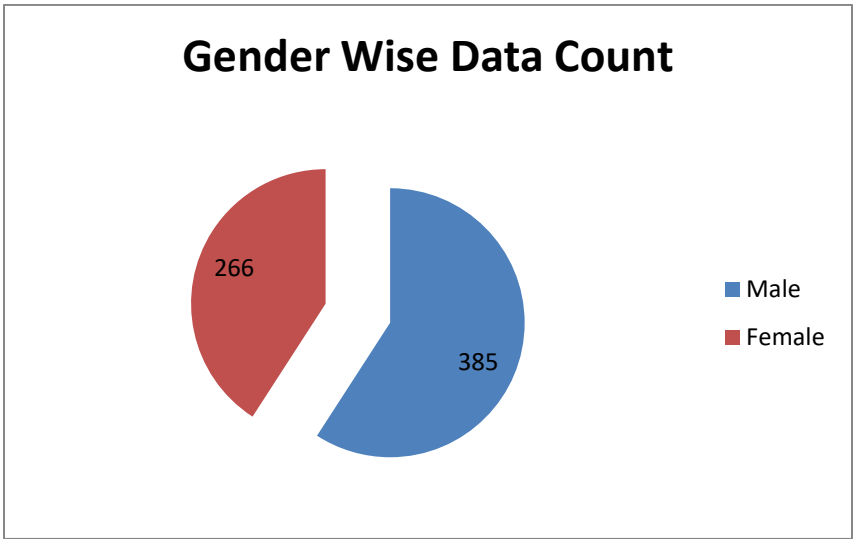

Figure 7: Gender wise data count

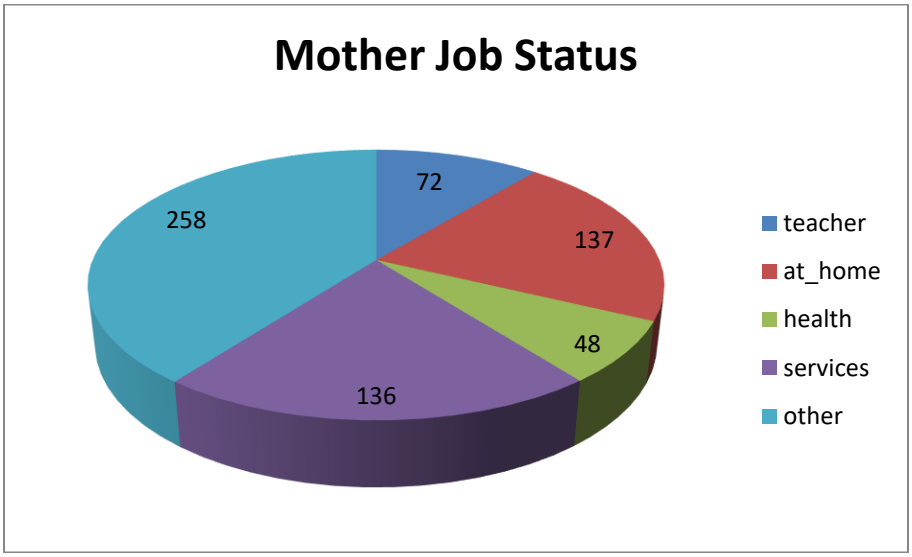

Figure 8: Status of mother job

The above said chart show that the mothers of 78 students are teachers, 137 are house wifes, 48 are health professionals, 136 are belongs to services and others are 258.

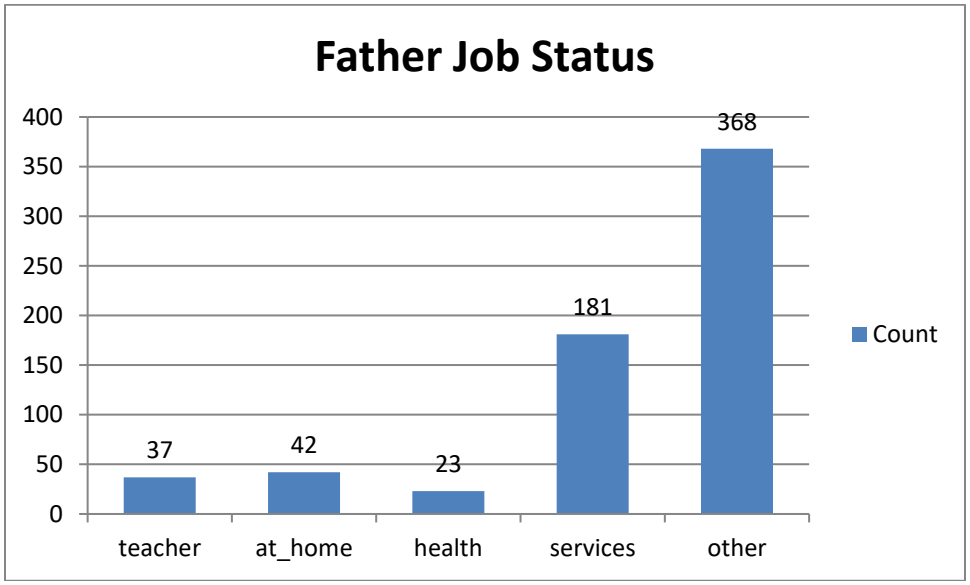

Figure 9: 5 Status of father job

The above said chart show that the father of 37 students are teachers, 42 are house wifes, 23 are health professionals, 181 are belongs to services and others are 368.

4.1 Clustering

For my assessment I at first grouped the data into different clusters for making utilization of abilities by method for data descriptive assessment. I utilized K-mean which is agreeing MacQueen (1967) most likely the least difficult unsupervised discovering algorithm that settles the great perceived clustering trouble. The strategy takes after a simple and convenient way to arrange a given informational index by method for a focused-on number of clusters (expect k groups) settled from the earlier. The essential motivation is to characterize k centroids, one for every last cluster. Regularly, in classification order you have a suite of pretend courses and wish to know which sort a fresh out of the box new object has a place with yet group tries to workforce a suite of articles and to discover whether there might be some connection between the items. Inside the setting of laptop studying, clustering and classification is basically

unsupervised learning and supervised learning. For the most part quite often it's no helpful to decide designs by means of straightforward questioning the data thus clustering framework making utilization of data mining is utilized to keep up monstrous measures of data. For my assessment I utilized the k = 2 in objective to put my data into 2 different clusters. The advantages of k-technique if factors are gigantic, then alright way lots of the events computationally quicker than progressive clustering, on the off chance that we safeguard affirm smalls and alright strategy deliver more tightly groups than various hierarchical grouping, particularly if the groups are globular. However, the fundamental disadvantages of k-approach it's confounded to foresee k-value and it works well. Likewise unique introductory parcels can bring about various last clusters. In spite of the difficulties of K-means I ran the calculations ordinarily to find the best number of groups.

We got following results after using the k mean clustering:

=== The Model of clustering (complete set of training) =kMeans
Total number of repetition: 3
Total sum of sqaure of error in that cluster is: 2858.56014061588
basic starting points (arbitrary):
Cluster 0:
GP,F,17,U,GT3,A,4,4,other,teacher,mother,2,2,yes,yes,no,no,yes,no,4,1,4,1,2
Cluster 1:
MS,F,17,U,GT3,T,2,2,other,at_Home,mother,1,3,no,no,no,yes,yes,no,3,4,3,3,8
Missing values globally replaced with mean/mode
Final cluster centroids:

Cluster#			
Attribute	Full Data	0	1
	(651.0)	(336.0)	(315.0)
school	GP	GP	GP
sex	F	F	F
age	16.7465	16.8125	16.6762
adddress	U	U	U
famsize	GT3	GT3	GT3
pstatus	T	T	T
medu	2.5146	2.3839	2.654
fedu	2.3072	2.2232	2.3968
mjob	other	other	other
fjob	other	other	other
guardian	mother	mother	mother
traveltime	1.5684	1.5923	1.5429
studytime	1.9309	1.875	1.9905
schoolsup	no	no	no
famsup	yes	yes	yes
paid	no	no	no
activities	no	no	yes
higher	yes	yes	yes
internet	yes	yes	yes
famrel	3.9324	3.881	3.9873
freetime	3.1797	3.0268	3.3429
goout	3.1859	3.0863	3.2921
health	3.5346	3.5149	3.5556
absences	3.6575	3.7232	3.5873

Time taken to build model (full training data) : 0.05 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 336 (52%)
1 315 (48%)

4.2 Naive Bayes' Algorithm

The algorithm of Naive Bayes' is a clear probabilistic classifier that figures an arrangement of chances by methods for numbering the recurrence and blends of qualities in a given informational collection. The algorithm utilizes the theorem of Bayes' hypothesis and accept all ascribes to be autonomous given the value of the classification variable. This contingent freedom presumption not frequently holds genuine in true purposes, thus the portrayal as Naive yet the calculation has a tendency to perform well and learn quickly in more than a couple supervised arrangement issues. Subsequent to applying the algorithm of Naive Bayes' the outcomes are introduced in work area eight. The legitimately named occasions is 75.42% which is a decent pointer and the kappa is 0.34 which is additionally an astounding index for this model.

Naive Bayes Classifier

Class Attribute	yes (0.77)	no (0.23)
Sex		
F	287.0	100.0
M	214.0	54.0
[total]	501.0	154.0

The probability of Female students to get minor grades are extreamly higher. Moving from higher to lower grades, it increases the probablility.

Age		
mean	16.7535	16.7237
std. dev.	1.2314	1.1653
weight sum	499	152
precision	1	1
adddress		
U	371.0	85.0
R	130.0	69.0
[total]	501.0	154.0
Famsize		
GT3	351.0	110.0
LE3	150.0	44.0
[total]	501.0	154.0
Pstatus		
A	57.0	26.0
T	444.0	128.0
[total]	501.0	154.0
Medu		
mean	2.6774	1.9803
std. dev.	1.1102	1.048
weight sum	499	15
precision	1	1
fedu		
mean	2.4148	1.9539
std. dev.	1.1103	0.989
weight sum	499	152
precision	1	1
mjob		
teacher	69.0	5.0
at_Home	79.0	60.0
health	44.0	6.0
services	122.0	16.0
other	190.0	70.0
[total]	504.0	157.0
Fjob		
teacher	29.0	10.0
at_Home	26.0	18.0
health	19.0	6.0

	services	146.0	37.0
	other	284.0	86.0
	[total]	504.0	157.0
Guardian			
	mother	347.0	111.0
	father	121.0	35.0
	other	34.0	9.0
	[total]	502.0	155.0
Traveltime			
	mean	1.489	1.8289
	std. dev.	0.6858	0.8719
	weight sum	499	152
	precision	1	1
studytime			
	mean	1.9479	1.875
	std. dev.	0.8311	0.8136
	weight sum	499	152
	precision	1	1
schoolsup			
	yes	51.0	20.0
	no	450.0	134.0
	[total]	501.0	154.0
Famsup			
	yes	317.0	84.0
	no	184.0	70.0
	[total]	501.0	154.0
Paid			
	yes	33.0	8.0
	no	468.0	146.0
	[total]	501.0	154.0
Activities			
	yes	254.0	63.0
	no	247.0	91.0
	[total]	501.0	154.0
Higher			
	yes	453.0	131.0
	no	48.0	23.0
	[total]	501.0	154.0
Famrel			
	mean	3.976	3.7895
	std. dev.	0.8928	1.1217
	weight sum	499	152
	precision	1	1
freetime			
	mean	3.2164	3.0592
	std. dev.	1.0195	1.1311
weight sum		499	152

precision 1 1

goout
mean 3.2445 2.9934
std. dev. 1.1623 1.1893
weight sum 499 152
precision 1 1

health
mean 3.517 3.5921
std. dev. 1.4524 1.4112
weight sum 499 152
precision 1 1

absences
mean 3.7668 3.0114
std. dev. 4.9746 3.7067
weight sum 499 152
precision 1.3913 1.3913

=== Evaluation on training set ===

To test the model on data of training data set the time consumed is 0.34 seconds.

=== Summary ===
Instances of Correctly Classified 491 75.4224 %
Instances of Incorrectly Classified 160 24.5776 %
Statistic of Kappa 0.3491
Mean absolute error 0.2773
Root mean squared error 0.4143
Relative absolute error 77.3871 %
Root relative squared error 97.9215 %
Total Number of Instances 651
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.816	0.447	0.857	0.816	0.836	0.351	0.771	0.907	yes
0.553	0.184	0.477	0.553	0.512	0.351	0.771	0.525	no
Weighted Avg. 0.754	0.386	0.768	0.754	0.760	0.351	0.771	0.818	

=== Confusion Matrix ===a b <-- classified as
407 68 | a = yes
92 84 | b = no

The matrix of confusion consists of information about the values of predicted performance and actual performance of students which done by algorithm of classification. By using this matrix, the performance of students evaluated, total instance is 651.

Actual performance of students = 407+68 / 651 = 0.7296 *100 = 72.96%

Predicted performance of students = 407+92 /651 =0.7665*100= 76.65%

The predicted values show that the performance of students can be improved to follow the instance that defined in a dataset.

Table 2: Confusion Matrix of Naive Bay's Algorithm

	Predicted class		
Actual class		Class = Yes	Class = No
	Class = Yes	TP =407	FN = 68
	Class = No	FP = 92	TN = 84

In Naive Bay's classification algorithm, the correctly predicted observations are true negative and true positive. When the values of performance of predicted class and actual class is yes then it is called true positive. E.g., if the actual values of performance of students increase then predicted values also increase. The values of true positive is 407. when the values of student performance of predicted class are Yes and actual class is No then it is called False positive. E.g., if the actual values of performance of students decrease then predicted values increase. The values of False positive are 92. when the values of student's performance of predicted class is No and actual class is yes then it is called False Negative. E.g., if the actual values of performance of students increase then predicted values decrease. The values of False Negative are 68. when the values of student's performance of predicted class is No and actual class is No then it is called True Negative. E.g., if the actual values of performance of students decrease then predicted values also decrease. The values of True Negative are 84.

TP=407, FP = 92, FN = 68, TN = 84

Accuracy is the most intuitive student performance measure. It is obtained through ratio of the observations of correctly predicted values to the overall positive predicted values

Accuracy = TP + TN / TP+FP+FN+TN

= 407+84/407+92+68+84

=491/651

Accuracy = 0.75 *100 = 75%

The accuracy values of performance of students of Naive Bay's classification algorithm are 75%

4.3 Algorithm Decision trees (J48)

The tree of decision is strong and extensively utilized the method of data mining for characterization and forecast. For this examination using the accompanying characteristics offense kind, pm/am, area/segment, zone/beat, Cluster, month and temperature. The classifier of J48 is a straightforward C4.5 decision tree for arrangement. The assurance tree technique makes a binary tree that is most valuable in grouping problem. With this framework, a tree is produced to demonstrate the characterization approach. When the

tree is created, it is used to new data set. Subsequent to making utilization of the algorithm the outcomes are shown that given below. The occasions of conveniently categorized are 95% and the measurement (Statistics) of kappa is 0.68 which is close to one and these figures are a decent marker of a without flaw model. A measurement (Statistic) of kappa of 1 can affirm that agreement surpasses danger stages and the inverse is correct and that is said additional given in a below section. A matrix of confusion outlines the precision of the answer for an order concern. It contains data about genuine and expected characterizations finished by means of an arrangement procedure. Effectiveness of such frameworks is much of the time assessed utilizing the data inside the network and the grid for this information.

=== Assessment on training dataset ===

To test the model on data of training data set the time consumed is 0.4 seconds.

=== Summary ===

Instances of Correctly Classified	619	95.0845 %
Instances of Incorrectly Classified	32	4.9155 %
Statistic of Kappa	0.689	
Mean absolute error	0.0869	
Root mean squared error	0.2084	
Relative absolute error	45.621 %	
Root relative squared error	67.7121 %	
Total amount of Instances	651	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.995	0.420	0.952	0.995	0.973	0.712	0.902	0.981	yes
0.580	0.005	0.930	0.580	0.714	0.712	0.902	0.727	no
Weighted Avg. 0.951	0.376	0.950	0.951	0.946	0.712	0.902	0.954	

=== Confusion Matrix ===

a	b	<-- classified as
579	3	a = yes
29	40	b = no

The matrix of confusion consists of information about the values of predicted performance and actual performance of students which done by algorithm of classification. By using this matrix, the performance of students evaluated, total instance is 651.

Actual performance of students = 579+3 / 651 = 0.89 *100 = 89%

Predicted performance of students = 579+27 /651 =0.93*100= 93%

The predicted values show that the performance of students can be improved to follow the instance that defined in a dataset.

Table 3: Confusion Matrix of Decision Tree J48 Algorithm

	Predicted class		
Actual class		Class = Yes	Class = No
	Class = Yes	TP =579	FN = 3
	Class = No	FP = 29	TN = 40

In Decision tree j48 classification algorithm the correctly predicted observations are true negative and true positive. When the values of performance of predicted class and actual class is yes then it is called true positive. E.g., if the actual values of performance of students increase then predicted values also increase. The values of true positive is 579. when the values of student performance of predicted class are Yes and actual class is No then it is called False positive. E.g., if the actual values of performance of students decrease then predicted values increase. The values of False positive are 29. when the values of student's performance of predicted class is No and actual class is yes then it is called False Negative. E.g., if the actual values of performance of students increase then predicted values decrease. The values of False Negative are 03. when the values of student's performance of predicted class is No and actual class is No then it is called True Negative. E.g., if the actual values of performance of students decrease then predicted values also decrease. The values of True Negative are 40.

TP=579, FP = 29, FN = 03, TN = 40

Accuracy is the most intuitive student performance measure. It is obtained through ratio of the observations of correctly predicted values to the overall positive predicted values

Accuracy = TP + TN / TP+FP+FN+TN

= 579+40/579+29+03+40

=619/651

Accuracy = 0.95 *100 = 95%

The accuracy values of performance of students of Decision Tree j48 classification algorithm are 95%. By comparing the result of both classification algorithms, we conclude that the accuracy values of Decision Tree J48 algorithm is high as compared to Naive Bay's Algorithm. So, the algorithm of classification Decision Tree J48 give better students' performance instead of Naive Bay's algorithms. Following is the decision tree which shows the whole data with their leaves

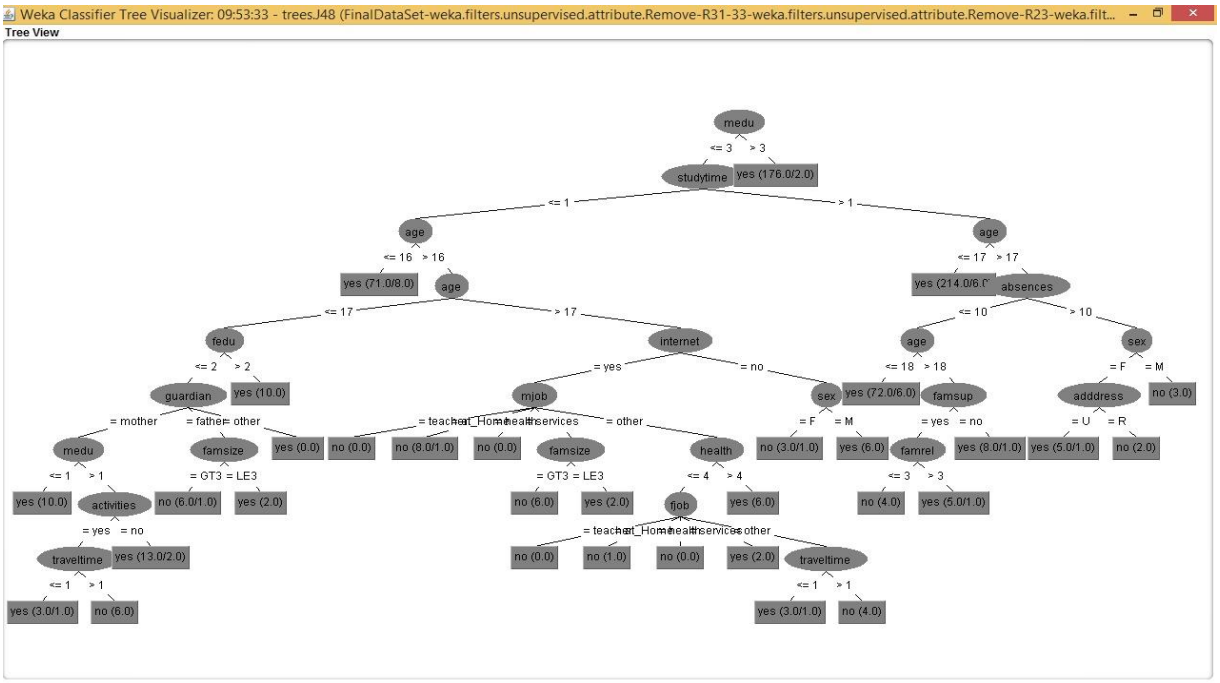

Figure 10 :Decision tree of attributes

Table 4: Factors affecting the performance of the students.

Sr. No:	Factor	Mean Value	Std. Dev
1	Age	16.7535	1.2314
2	Medu (Mother Education)	2.6774	1.1102
3	Fedu(Father Education)	2.4148	1.1103
4	studytime	1.9479	0.8311
5	Health	3.517	1.4524
6	Absensees	3.7668	4.9746

The following table shows that the job prefession of mohter greatly influence on the performance of the student as mothers at home give more attention towards the study of their students.

Table 5: Effect of mother job on performance

Sr. No	Job Type	Mean Value	Std. Dev
1	Teacher	69.0	5.0
2	At_home	79.0	60.0
3	Health	44.0	6.0
Total		192.0	71.0

The above said table shows that mothers at home give more attention towards the study of their students. Moreover, these results also shows the correctness of purposed model for educational data mining.

4. Conclusion

Predicting student's performance become more tough due to a huge volume of data in educational database. Even though data mining has been efficiently implemented in the business world. Use of data mining is relatively new in higher education, i.e., it used for recognition and extraction of potentially and advanced precious knowledge from huge amount of data using the concept of data mining. Data mining common called knowledge discovery that is the way toward dissecting information from alternate points of view and condensing it into helpful data - data that can be utilized to build income, cuts costs, or both. The software of data mining is one of various diagnostic apparatuses for investigating information. The innovation of data mining can discover valuable learning from a huge amount of data, and this learning gives essential establishment to enhance administration arrangements of university. The classification and other is clustering are a basic task of data mining which widely used for different purposes like data categorization and investigation in both industry and educational institutions. Clustering is the way toward arranging objects which are unlabeled into gatherings which individuals are comparative somehow. (Larose 2006) Classification exists in algorithm of supervised learning. The concept of data mining is used for different purposes like business, government and scientific research. There are two types of data mining learning methods .one is supervised and other is unsupervised Wang and Mitrovic (2002) proposed an intelligent agent of selection of problem which classifies the proper issue for different students in two phases. First of all, it predicts the quantity of errors the understudy will make an arrangement of issues, and afterward in stage of second determine on an appropriate issue for the scholars. The efficiency of achieved forecast is high, demonstrating that a neural network is fit for predictions. ANN used to predict the time that students will take to solve a problem. The student population model was firstly delivered utilizing every one of the information, with the prediction efficiency of 98.06%. This system was then used to anticipate the behavior of every individual pupil. In settings of academic, the methods of data mining were actualized in both, picking up information of and Administrative/scope situated troubles. (Baker and Yacef,2009) In concentrate, the technique might be cut up into learner and instructor arranged. Inside the underlying one, the point of convergence is on helping the understudy to concentrate more prominent viably through prescribed new substance; inside the last mentioned, the design is to offer the instructor an apparatus to permit him so learner guided with more prominent proficiently. (Romero,2010). In this study, student data will be collected from different institutes. The scope of this research is to identify the factors influencing the performance of students in universities and compare the results of both classification

algorithms which gives better results. To predict the performance of university student, for clustering purposes K-Mean Algorithm were used and for classification purposes Decision tree (J48) and Naive Bay's algorithms have been used. For this purpose, Weka tool will be used to implement the algorithms. This study will help the educational institutions to identify such factors which affect the student performance and to eliminate these causes to enhance their performance. After compiling results Decision tree j48 algorithm gives better results with accuracy of 95% as compared to Naive Bay's algorithm accuracy 75%

Reference:

1. Waqas, M., Ahmed, S. U., Tahir, M. A., Wu, J., & Qureshi, R. (2024). Exploring Multiple Instance Learning (MIL): A brief survey. Expert Systems with Applications, 123893.
2. Khan, S. R., Raza, A., Shahzad, I., & Ijaz, H. M. (2024). Deep transfer CNNs models performance evaluation using unbalanced histopathological breast cancer dataset. Lahore Garrison University Research Journal of Computer Science and Information Technology, 8(1).
3. Bilal, Omair, Asif Raza, and Ghazanfar Ali. "A Contemporary Secure Microservices Discovery Architecture with Service Tags for Smart City Infrastructures." VFAST Transactions on Software Engineering 12, no. 1 (2024): 79-92.
4. S. Abhari, S. R. N. Kalhori, M. Ebrahimi, and H. Hasannejadasl, "Artificial Intelligence Applications in Type 2 Diabetes Mellitus Care : Focus on Machine Learning Methods," vol. 25, no. 4, pp. 248–261, 2019.
5. Raza, A., Salahuddin, & Inzamam Shahzad. (2024). Residual Learning Model-Based Classification of COVID-19 Using Chest Radiographs. Spectrum of Engineering Sciences, 2(3), 367–396.
6. H. Kaur and V. Kumari, "Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach," *Appl. Comput. Informatics*, no. December, 2018, doi: 10.1016/j.aci.2018.12.004.
7. M. Wajid, M. K. Abid, A. Asif Raza, M. Haroon, and A. Q. Mudasar, "Flood Prediction System Using IOT & Artificial Neural Network", VFAST trans. softw. eng., vol. 12, no. 1, pp. 210–224, Mar. 2024.
8. I. M. Ibrahim and A. M. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," vol. 02, no. 01, pp. 10–19, 2021, doi: 10.38094/jastt20179.
9. Abbas, S. S., Shafique, P. H., Razzaq, A. M., & Ikhlaq, M. (2024). Enhancing Reliability and Sustainability of Green Communication in Next-Generation Wireless Systems through Energy Harvesting. Journal of Computing & Biomedical Informatics.
10. A. Mujumdar and V. Vaidehi, "ScienceDirect ScienceDirect ScienceDirect ScienceDirect Diabetes Prediction using Machine Learning Aishwarya Mujumdar Diabetes Prediction using Machine Learning Aishwarya Mujumdar Aishwarya," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.
11. M. Waqas, Z. Khan, S. U. Ahmed and A. Raza, "MIL-Mixer: A Robust Bag Encoding Strategy for Multiple Instance Learning (MIL) using MLP-Mixer," 2023 18th International Conference on Emerging Technologies (ICET), Peshawar, Pakistan, 2023, pp. 22-26.
12. S. Brian and R. R. B. Pharmd, "Prediction of Nephropathy in Type 2 Diabetes: An Analysis of the ACCORD Trial applying Machine Learning Techniques," no. 317, doi: 10.1111/cts.12647.
13. Faruque, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," *2019 Int. Conf. Electr. Comput. Commun. Eng.*, pp. 1–4, 2019.
14. Razzaq, A. M., Abbas, S. S., Ikhlaq, M., Shafique, P. H., & Shahzad, I. (2024). Development of OWL Structure for Recommending Database Management Systems (DBMS). Journal of Computing & Biomedical Informatics, 7(02).
15. Asif, S., Wenhui, Y., ur-Rehman, S., ul-ain, Q., Amjad, K., Yueyang, Y., ... & Awais, M. (2024). Advancements and Prospects of Machine Learning in Medical Diagnostics:

Unveiling the Future of Diagnostic Precision. Archives of Computational Methods in Engineering, 1-31.
16. Asif, S., Zhao, M., Li, Y., Tang, F., Ur Rehman Khan, S., & Zhu, Y. (2024). AI-Based Approaches for the Diagnosis of Mpox: Challenges and Future Prospects. Archives of Computational Methods in Engineering, 1-33.
17. J. Li *et al.*, "International Journal of Medical Informatics Establishment of noninvasive diabetes risk prediction model based on tongue features and machine learning techniques," *Int. J. Med. Inform.*, vol. 149, no. August 2020, p. 104429, 2021, doi: 10.1016/j.ijmedinf.2021.104429.
18. HUSSAIN, S., Raza, A., MEERAN, M. T., IJAZ, H. M., & JAMALI, S. (2020). Domain Ontology Based Similarity and Analysis in Higher Education. IEEEP New Horizons Journal, 102(1), 11-16.
19. S. G. Azevedo *et al.*, "System-Independent Characterization of Materials Using Dual-Energy Computed Tomography," *IEEE Trans. Nucl. Sci.*, vol. 63, no. 2, pp. 341–350, 2016, doi: 10.1109/TNS.2016.2514364.
20. Khan, S. U. R., Asif, S., Zhao, M., Zou, W., Li, Y., & Li, X. (2024). Optimized Deep Learning Model for Comprehensive Medical Image Analysis Across Multiple Modalities. Neurocomputing, 129182.
21. Shahzad, I., Khan, S. U. R., Waseem, A., Abideen, Z. U., & Liu, J. (2024). Enhancing ASD classification through hybrid attention-based learning of facial features. Signal, Image and Video Processing, 1-14.
22. Waqas, M., & Khan, M. A. (2018). JSOPT: A framework for optimization of JavaScript on web browsers. Mehran University Research Journal of Engineering & Technology, 37(1), 95-104.
23. Waqas, M., Tahir, M. A., & Qureshi, R. (2023). Deep Gaussian mixture model based instance relevance estimation for multiple instance learning applications. Applied intelligence, 53(9), 10310-10325.
24. Waqas, M., Tahir, M. A., & Khan, S. A. (2023). Robust bag classification approach for multi-instance learning via subspace fuzzy clustering. Expert Systems with Applications, 214, 119113.
25. B. G. Choi, S. Rha, S. W. Kim, J. H. Kang, J. Y. Park, and Y. Noh, "Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks," vol. 60, no. 2, pp. 191–199, 2019.
26. Khan, S.U.R.; Raza, A.;Waqas, M.; Zia, M.A.R. Efficient and Accurate Image Classification Via Spatial Pyramid Matching and SURF Sparse Coding. Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol. 2023, 7, 10–23.
27. Farooq, M.U.; Beg, M.O. Bigdata analysis of stack overflow for energy consumption of android framework. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019; pp. 1–9.
28. S. U. R. Khan, A. Raza, I. Shahzad and G. Ali, "Enhancing Concrete and Pavement Crack Prediction through Hierarchical Feature Integration with VGG16 and Triple Classifier Ensemble," 2024 Horizons of Information Technology and Engineering (HITE), Lahore, Pakistan, 2024, pp. 1-6.
29. F. Farrokhi *et al.*, "Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms," *World Neurosurg.*, 2019, doi: 10.1016/j.wneu.2019.10.063.
30. Waqas, M., Tahir, M. A., Al-Maadeed, S., Bouridane, A., & Wu, J. (2024). Simultaneous instance pooling and bag representation selection approach for multiple-instance learning (MIL) using vision transformer. Neural Computing and Applications, 36(12), 6659-6680.
31. Farooq, M. U., Khan, S. U. R., & Beg, M. O. (2019, November). Melta: A method level energy estimation technique for android development. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-10). IEEE.
32. Khan, S. U. R., & Asif, S. (2024). Oral cancer detection using feature-level fusion and novel

self-attention mechanisms. Biomedical Signal Processing and Control, 95, 106437.
33. Raza, A.; Meeran, M.T.; Bilhaj, U. Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers. VFAST Trans. Softw. Eng. 2023, 11, 80–92.
34. Dai, Q., Ishfaque, M., Khan, S. U. R., Luo, Y. L., Lei, Y., Zhang, B., & Zhou, W. (2024). Image classification for sub-surface crack identification in concrete dam based on borehole CCTV images using deep dense hybrid model. Stochastic Environmental Research and Risk Assessment, 1-18.
35. Khan, S.U.R.; Asif, S.; Bilal, O.; Ali, S. Deep hybrid model for Mpox disease diagnosis from skin lesion images. Int. J. Imaging Syst. Technol. 2024, 34, e23044.
36. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X.; Zhu, Y. GLNET: Global–local CNN's-based informed model for detection of breast cancer categories from histopathological slides. J. Supercomput. 2023, 80, 7316–7348.
37. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. Int. J. Imaging Syst. Technol. 2024, 34, e22975.
38. IoannisKavakiotis, Olga Tsave, Athanasios Salifoglou, and NicosMaglaveras, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, vol. 15, pp. 104– 116, 2017
39. Khan, U. S., & Khan, S. U. R. (2024). Boost diagnostic performance in retinal disease classification utilizing deep ensemble classifiers based on OCT. Multimedia Tools and Applications, 1-21.
40. Khan, M. A., Khan, S. U. R., Haider, S. Z. Q., Khan, S. A., & Bilal, O. (2024). Evolving knowledge representation learning with the dynamic asymmetric embedding model. Evolving Systems, 1-16.
41. Raza, A., & Meeran, M. T. (2019). Routine of encryption in cognitive radio network. Mehran University Research Journal of Engineering & Technology, 38(3), 609-618.
42. Al-Khasawneh, M. A., Raza, A., Khan, S. U. R., & Khan, Z. (2024). Stock Market Trend Prediction Using Deep Learning Approach. Computational Economics, 1-32.
43. Khan, U. S., Ishfaque, M., Khan, S. U. R., Xu, F., Chen, L., & Lei, Y. (2024). Comparative analysis of twelve transfer learning models for the prediction and crack detection in concrete dams, based on borehole images. Frontiers of Structural and Civil Engineering, 1-17.
44. Hekmat, A., Zhang, Z., Ur Rehman Khan, S., Shad, I., & Bilal, O. (2025). An attention-fused architecture for brain tumor diagnosis. Biomedical Signal Processing and Control, 101, 107221. https://doi.org/https://doi.org/10.1016/j.bspc.2024.107221