Publisher homepage: www.universepg.com, ISSN: 2663-7804 (Online) & 2663-7790 (Print)

https://doi.org/10.34104/ajeit.022.065077

Australian Journal of Engineering and Innovative Technology

Journal homepage: www.universepg.com/journal/ajeit

Students Performance Measurement and Prediction Based on Academic Features Through the Machine Learning

Akash Kumar Pal[1], Shahrin Sumona[1], Md. Atikur Rashid[1], Mirza A.F.M. Rashidul Hasan[2], Mithun Kumar[1], and Md. Anwar Hossain[3]*

[1]Dept. of Computer Science and Engineering, Bangladesh Army University of Engineering and Technology (BAUET), Natore-6431, Bangladesh; [2]Dept. of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh; and [3]Dept. of Information and Communication Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

*Correspondence: manwar.ice@gmail.com (Md. Anwar Hossain, Associate Professor, Department of Information and Communication Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh).

ABSTRACT

The education sector plays a vital role in achieving long-term economic and national development. During the last decades, due to the availability of resources and technology, efficient and higher standards in education have become easier to attain. The amount of data in educational institutions is growing rapidly. Through the data mining and machine learning methodologies, it has become easier to look into data from a different perspective and extract various information from the data. In our research, we use various algorithms to find out the correlation between features and predict students' performance using their academic records. We try to find out the factors which affect and influence the performance of students. We implemented both unsupervised and supervised learning algorithms in our research. K-mean clustering has been used as an unsupervised learning method to group the students and find out the dependencies between the features. For prediction purposes, we use classification techniques like KNN and Linear Regression to predict students' performance. Our research not only aims into finding useful information but also provides insight into students' preferred teaching methods, potentiality, and performance. This information can guide the students for their future and guide them to their preferred fields according to their skill sets.

Keywords: Clustering, Classification, KNN, Prediction, Linear regression, and Machine learning.

INTRODUCTION:

We are living in the age of data. Like all modern organizations, educational institutions are collecting a large volume of data involved around their students. All educational institutions are working in a strongly competitive environment. Most of this data is usually used in producing simple queries and making traditional reports. A significant amount of data is kept unused. Data mining is a process of discovering useful and meaningful patterns as well as information from a large amount of data. Data mining provides us with tools and technology that can help us in the knowledge discovery process. It is a step-by-step process from integrating data from different sources to mining useful patterns and information and representing it in a way that is easy for human interpretation.

In this research, with the help of data mining and machine learning methodologies, we are trying to find out useful patterns and information from available data about students. We will try to find out the factors that are affecting students' results and overall performance. In the field of academics, data mining is very effective in discovering valuable information which can be used for profiling students based on their academic records (Suhem *et al.,* 2012). Predictive information provides valuable time in which we can take measures like counseling and additional coaching to improve the performance of degrading students. Data mining, generally defined as the process of discovering meaningful patterns in large quantities of data, offers a great variety of techniques, methods, and tools for a thorough analysis of available data in various fields (Dorina K., 2012). Through our research, we can get insight into students' preferred teaching methods and fields. It can help us to improve education quality and tech-nical skills through the findings. The research aims to find students' performance for developing his/her career and also track the factors affecting their performance. It can also give us an idea of the dependencies and patterns among the factors.

The motivations behind our research are based on the huge possibilities and goals that we can achieve through it. The education sector plays a key role in the development of an individual as well as a nation. Hence continuous improvement and development of this sector is a must. Gathering important information about students and the factors that affect their results and performance might come in handy. It can help us in finding interesting patterns and how to use them in the advancement of student performance. Various tools, technology and methods are available due to the advancement of science. The availability of these tools and technology can monitor the regular performance of the students. The students can aware his/her present and future academic condition based on this type of research hence we are motivated to do the research work in this field.

Many institutions still manage students' information manually where significant data are kept unused. They gather various information about students but do not use it for further analysis or any kind of prediction. We have to use this huge amount of data in some productive way to our advantage. Manual management of student data limits the usability of this data. Every student has a preferred and suitable subject or sector to study. But all cannot choose their preferable subjects or sectors for many issues. Some students have to forcefully study non-favorable subjects or fields which limits their potential. Data analysis can help in understanding their skillsets and potentiality which can eventually guide them. All students are not in the same categories. There is always a gap between good, medium, and bad students. For the continuation of the gap between students, they do not cope with the boundaries. Lack of guidance remains between them as some are to be handled differently from others. For this reason, the good student usually always performs well and bad students remain lagging. We pressurized them to good in study and gain good results in exams. If we focus on their potential, they might have a better future. Students can achieve their highest potential if their talents and skillsets are discovered in the early stages and are properly guided. Objectives describe what we expect to achieve from the research. The main objectives of the research are listed here.

Students' performance measurement
In this work, we will collect all the information from the student's track record and analysis those data to see how a student's performance is increasing or decreasing.

Finding students' potentiality
By analyzing students' track records and all informative datawe can know about a student's potentiality.

Achieve steps of learning approach
In this work, it will be helpful for theteachers to achieve steps of the learning approach from the findings. They can give the right effort to a student.

Improving education quality and technical skills
To find out the student's potentiality wecan improve educational quality and technical skills through the approach. That will help a student to make a better future.

Factors affecting students' performance
Finding the factors affecting students' performance through our analysis and taking necessary steps according to findings. It can help both teacher and a student to utilize the factors that affect those students. And also find out dependencies and patterns among those factors. Therefore the research will not only help us in improving the students' performance but also bring out their full potential. It will help to develop new teaching methods accor-ding to student types and provide insight into student sentiment. That may decrease the number of student dropouts. Following the introduction above, we conclude the paper with a satisfactory performance of the research. The remainder of this paper is organized as

follows. In section 2, we discuss the student's sentiments. Sections 3 and 4 consist of methodology and result analysis of the research; conclusions are discussed in section 5.

Students'sentiment analysis

Students' sentiment analysis is an analysis process that measures students' emotions, attitudes, or opinions based on students' data. In this study, we evaluate students' responses to their academic studies like how good they are at their class tests and final exam. These responses give them the outcome in their result. We compare their present result with the past result to evaluate their progress or downfall in the study. We use their numeric information which is attendance, performance marks, class-test marks, and exam mark to predict their result and the status of the result. We can evaluate the student's sentiments from their academic activities. Therefore if the students do not perform according to their expectations, we can consult with them to know if they are facing any problems or difficulties. Students are a crucial asset for the long-term development of any nation. If we give them proper guidelines and show them the best paths, they will surely succeed in bringing prosperity to both their society and the nation.

Datasheet attributes

Initially, to analyze students' sentiment, we need to have data consisting of their various features. For this purpose, we used an xlsx file to store data. We prepared 2 datasheets for two years of 1000 students in four subjects. We entered Students ID, Name, per subjects Class Attended, Attendance_ percentage, Attendance, Class Test (CT_1, CT_2, CT_3), Exam, Grade, and Status as an attribute. Then we converted the excel file into a CSV file. It is a plain text file that contains a list of data and is used for exchanging data between different methods.We stored students' information digitally and preprocessed their information for better quality analysis. If any student's information was missing, we filled the data with the algorithm of filling the missing value. We can fill the missing data with mean or median values. We have normalized the data for ranging all the students' data for better use and to reduce complexity. Then we classified the data and visualized those data to founding the desired outcome.

Problem finding and necessary measures

Through our analysis using various methods, we find out the problems behind students' bad results, depression and inattentiveness. After analyzing the data, we can get various insights into the factors that are affecting their performance. This research explores as well the possibility of identifying the key indicators in the small dataset, which will be utilized in creating the prediction model, using visualization and clustering algorithms (Lubna M. A. Z., 2019). With our analysis, we can identify the students who are performing unexpectedly. With the information, we can cooperate with that student to get an idea of his/her problems or issues. Some students fail or gain bad results because they do not like that subject, some do not understand that subject, some students have not interested in education study, some have family issues or some need guidelines. Throughout the work, we can find out what problems they are facing and give them suggestions. If the problems are solely base on their academics we can provide them with extra classes, tutoring, individual tasks, etc. If they have family issues, we can monitor them so that we can give them extra care. If they feel depressed, we can give them counseling. If the financial state of the parents is in bad condition, we can help them to get motivated and provide them with a weaver or scholarship opportunities. Our findings can also help to understand their sentiment towards their studies. This can help them to bring out their potential and individual skillsets. Students will also be cheerful-minded as they are doing and working on what they love and admire.

METHODOLOGY:

The objective of our proposed methodology includes finding useful patterns and valuable information. To accomplish this,we have to go through a step-by-step process from datapreparation to knowledge represent ation. The existing methodologies of students' information management and performance measurement are as follows.

Manual Evaluation

In this world, the amount of data is increasing day by day. Therefore nowadays it is hard to handle this data in manual ways. It is not possible that anyone can record this data in notes or other handwriting ways. This process limits the usability of this data.

Excel Datasheet

We know excel is a good software to record data. But excel cannot handle a vast amount of data properly. And it is quite hard to evaluate some important data from an excel sheet. Nowadays it is the world of big data and we need to extract information that can help us predict something for further inquiry. Excel also has limited visualization and knowledge representation methods.

Limited Visualization

We need to increase our visualization and utilization of the data. It can help us in decision making for the future. Limited visualization limits human interpretation and a better understanding of the data.

Descriptive Information

In thisresearch, we are trying to find out the potentiality of data from the findings. Unused data can also help us if we can properly use it and extract necessary information from it. Currently, used methods can give us descriptive information for particular findings. They do not provide us with predictive information which can be very beneficial to our goal. Our overall methodology can be perceived through the diagram given in **Fig. 1**. It represents the steps from which we can receive our expected outcome. At first, we go through the data integration phase then we preprocess the data by implementing various techniques. After that, we implement different machine learning algorithms to get knowledge and finally implement knowledge representation.

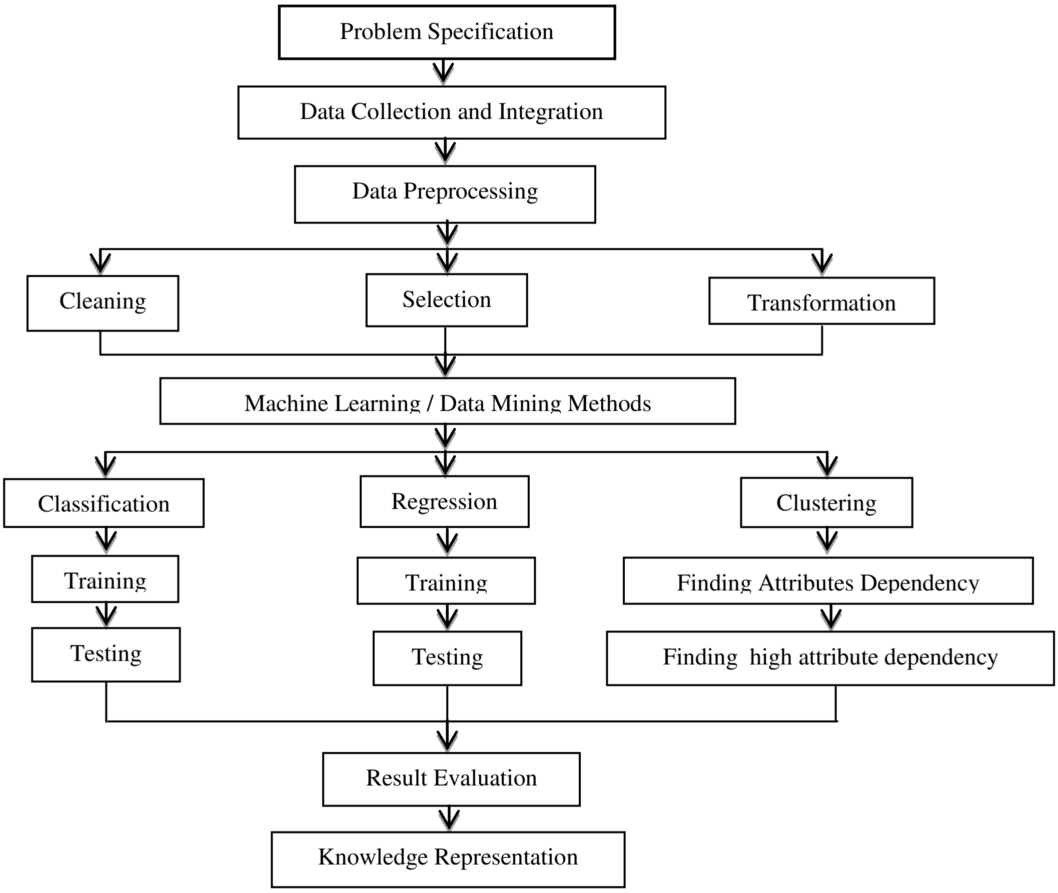
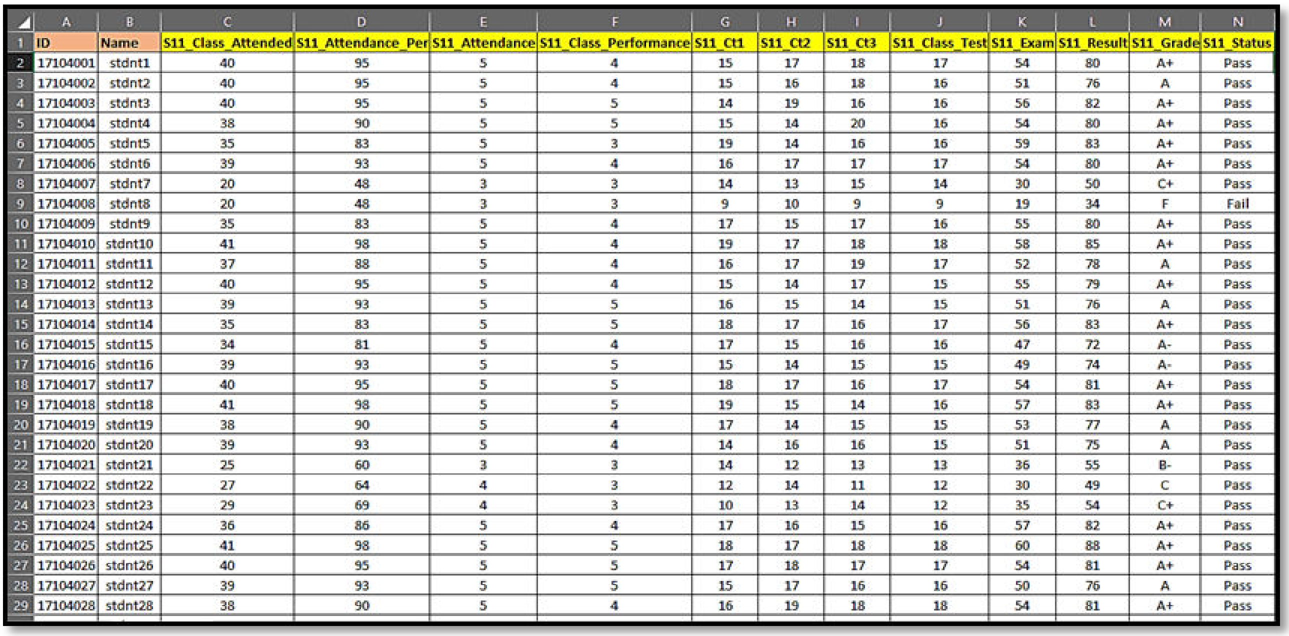


Fig. 1: Diagram of Research Workflow.

Data integration

Data integration is a process where data is combined from different heterogeneous sources into a single source and provides a unified view of the total data. The data integration process is a significant process for a variety of situations like data analysis, data processing, etc. Our dataset consists of academic records of various attributes of 1000 students. Each subject consists of various data of that subject like attendance, class performance, class test, exam, result, grade, etc. These are the commonly used attributes in universities for students' academic records. Our research explores as well the possibility of identifying the key indicators in the small dataset, which will be utilized in creating the prediction model, using visualization and clustering algorithms (Lubna M. A. Z., 2019). In our dataset, we integrated per subject attendance from different sources like data from different course teachers. From the attendance, we calculated the per subjects' attendance marks and percentages. Based on the student's performance in their classroom teacher gives them performance marks. We integrated their performance mark into our dataset. Each subject has three class-test marks and the average marks are gathered into the dataset. Final exam marks are also gathered into the dataset. We calculated the final result with all these marks. Based on the final result, we gave the student grade marks and the outcome pass or fail. In this manner, we combined 4 subjects' data into our dataset for 1st year and 2nd year.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Name	S11_Class_Attended	S11_Attendance_Per	S11_Attendance	S11_Class_Performance	S11_Ct1	S11_Ct2	S11_Ct3	S11_Class_Test	S11_Exam	S11_Result	S11_Grade	S11_Status
2	17104001	stdnt1	40	95	5	4	15	17	18	17	54	80	A+	Pass
3	17104002	stdnt2	40	95	5	4	15	16	18	16	51	76	A	Pass
4	17104003	stdnt3	40	95	5	5	14	19	16	16	56	82	A+	Pass
5	17104004	stdnt4	38	90	5	5	15	14	20	16	54	80	A+	Pass
6	17104005	stdnt5	35	83	5	3	19	14	16	16	59	83	A+	Pass
7	17104006	stdnt6	39	93	5	4	16	17	17	17	54	80	A+	Pass
8	17104007	stdnt7	20	48	3	3	14	13	15	14	30	50	C+	Pass
9	17104008	stdnt8	20	48	3	3	9	10	9	9	19	34	F	Fail
10	17104009	stdnt9	35	83	5	4	17	15	17	16	55	80	A+	Pass
11	17104010	stdnt10	41	98	5	4	19	17	18	18	58	85	A+	Pass
12	17104011	stdnt11	37	88	5	4	16	17	19	17	52	78	A	Pass
13	17104012	stdnt12	40	95	5	4	15	14	17	15	55	79	A+	Pass
14	17104013	stdnt13	39	93	5	5	16	15	14	15	51	76	A	Pass
15	17104014	stdnt14	35	83	5	5	18	17	16	17	56	83	A+	Pass
16	17104015	stdnt15	34	81	5	4	17	15	16	16	47	72	A-	Pass
17	17104016	stdnt16	39	93	5	5	15	14	15	15	49	74	A-	Pass
18	17104017	stdnt17	40	95	5	5	18	17	16	17	54	81	A+	Pass
19	17104018	stdnt18	41	98	5	5	19	15	14	16	57	83	A+	Pass
20	17104019	stdnt19	38	90	5	4	17	14	15	15	53	77	A	Pass
21	17104020	stdnt20	39	93	5	4	14	16	16	15	51	75	A	Pass
22	17104021	stdnt21	25	60	3	3	14	12	13	13	36	55	B-	Pass
23	17104022	stdnt22	27	64	4	3	12	14	11	12	30	49	C	Pass
24	17104023	stdnt23	29	69	4	3	10	13	14	12	35	54	C+	Pass
25	17104024	stdnt24	36	86	5	4	17	16	15	16	57	82	A+	Pass
26	17104025	stdnt25	41	98	5	5	18	17	18	18	60	88	A+	Pass
27	17104026	stdnt26	40	95	5	5	17	18	17	17	54	81	A+	Pass
28	17104027	stdnt27	39	93	5	5	15	17	16	16	50	76	A	Pass
29	17104028	stdnt28	38	90	5	4	16	19	18	18	54	81	A+	Pass

Fig. 2: First year data of subject one.

The above **Fig. 2** shows year one' one subjects' data among the total four of our main datasets. In our dataset, we integrated per subjects' attendance from different sources like data from different course teachers. From the attendance, we calculated the per subjects' attendance marks and percentages. Based on the student's performance in their classroom teacher gives them performance marks. We integrated their performance mark into our dataset. Each subject has three class-test marks and the average marks are gathered into the dataset. Final exam marks are also gathered into the dataset. We calculated the final result with the help of attendance marks, performance marks, class-test average marks and final exam marks.

Data preprocessing

Data preprocessing is a process of preparing the raw dataand making it suitable for data mining and the machinelearning process. It is the first and crucial step while creatinga machine learning model. If the dataset consists of muchredundant or irrelevant data then the knowledge discoveryprocess from the dataset becomes difficult and inefficient. Data preprocessing is a step-by-step process. The steps are -

1) To gather the dataset
2) To import the necessary libraries
3) To find the missing values and handle them
4) Encoding the categorical data
5) Removal of noise or outliers
6) Scale the values

While working on a large amount of data we often face situations like missing values. It can also occur during a data transaction. In the dataset, we handle the missing values with median values. Missing values have a significant effect on the observation. Therefore, the missing values need to be handled otherwise we may face unwanted results. Missing values can also be handled manually but it consumes valuable time. The empty values can also be filled with mean values. We must remember the fact that the values filled by this process will not be accurate. Then we normalized the value in our dataset between the 0-1 range which is used to reduce the data redundancy and improve the data integrity. Min-max normalization is the simplest method and consists in rescaling the range of features to scale them in a predefined range. The equation through which min-max normalization is calculated is given below.

$$v' = \frac{(v-min_a)}{(max_a-min_a)}(new\_max_a - new\_min_a) + new\_min_a \qquad (1)$$

Where,
v'= Normalized value, v= Value to be calculated
min_a= Minimum value of that attribute
max_a= Maximum value of that attribute
new_min_a= New minimum value of that attribute
new_max_a= New maximum value of that attribute

Another data preprocessing method is noise or outlier analysis and their removal. The solution can be found through clustering. The cluster can show us if there is any noise or outliers present in the dataset.

Algorithms

After data preprocessing, we get consistent data in our hands. Now we can implement various kinds of algorithms using specific attributes according to the needs of a specific pattern. In data mining and machine learning methods, there are generally two types of algorithms. One of them is unsupervised learning and another is supervised learning. An unsupervised learning algorithm is an algorithm that learns patterns and information from unlabeled data. This method is appropriate to use when there is a need of discovering hidden patterns or groups from a dataset without the need for human intervention. In our research, we implemented the K-means clustering method as our unsupervised learning algorithm. K-means clustering is the process of grouping a set of data points into several groups such that objects in the same group are more similar than the objects in other groups. In K-means clustering the K value defines the number of clusters. K-means tries to make the intra-cluster data points as similar as possible and at the same time keep the clusters as different as possible. This algorithm assigns data points to a group or cluster such that the sum of the squared distance between the data point value and the cluster centroid value is at the minimum. The process of how the k-mean algorithm works are given below -

1) Specify the number of clusters
2) Initialize centroids by first shuffling the dataset and randomly selecting k data points for the centroids without replacement
3) Continue the iterations until no changes to the centroids are found
4) Compute the sum of the squared distance between data points and all centroids
5) Assign each data point to the nearest cluster
6) Compute the centroids for the clusters by taking the average of the data points that belong to each cluster

K-means can give us an idea of the data we are dealing with and also provide insight into the dependencies between the attributes. It can also show us if there are any outliers and noise present in the dataset. So, it is a useful process for outlier analysis as well as noise removal. There is one more thing we have to keep in mind while doing K-means clustering. The value of k determines how many clusters we will get. So, we need to choose the optimal k value which is hard for humans to select. That's why we used Elbow Method for this purpose which is used to select the k value. This method can provide us an idea of what the optimal number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters centroid. K is chosen at the spot where SSE starts to flatten out and form an elbow. Using this method, we can hope to get the best k value according to our dataset (Alisa B. Z., 2020). The supervised learning algorithm is the process of learning from a labeled dataset or when correct examples are given. Our objective is to predict students' performance based on several criteria such as their evaluation in previous grades (Feras *et al*., 2017). In our research, we used two types of supervised learning. One is for categorical target variables and another for the continuous target variable. For the categorical target variable, we used the KNN classification algorithm. This algorithm provides us with discrete values. Such as; yes-no, 0-1, pass-fail. So based on the problem, we can use KNN to predict if a student will pass or fail the exam. The KNN algorithm assumes that similar data exist near each other. KNN captures the idea of similarity by calculating the distance between two data points. Like all classification algorithms, KNN also has two phases; training and testing. We can also use other features to predict result status if those features provide enough influence on the result. For continuous target variables, we used Multiple Linear Regression that can easily handle continuous values. In multiple linear regression, there are multiple independent variables and one dependent variable. The value of the dependent variables changes according to the independent variables. The equationused in multiple linear regression is given below –

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \varepsilon \qquad (2)$$

Where,
Y= The predicted value of the dependent variable
β0= The Y intercept (value of y when all other parameters are set to 0)
β1X1= The regression coefficient (β1) of the independent variable (X1) (the effect that increases the value of the independent variable has on the predictted Y value)
βnXn= The regression coefficient of the last independent variable
ε= model error (how much variation there is on our estimate of y)

In multiple linear regression, we can use the values of attendance, class performance, class test, etc. as the independent variables and predict examination marks or results as the dependent variable to get the prediction of how many marks a student is likely to get according to the given values. The coefficient values can give us insight into which independent variable or feature influences the performance of a student (Ouafae *et al.,* 2020). The greater the coefficient value the more impact it has on students' performance.

Knowledge representation and visualization

Knowledge representation is a technique for human-interpretation through the experience of past problems. Itcan be viewed from different perspectives. The usefulnessof the knowledge and inform-ation depends on how they arerepresented. It is not just transferring information but alsoattempts to ensure students receiving the visual information understand greater insights and perspectives surrounding agiven subject.

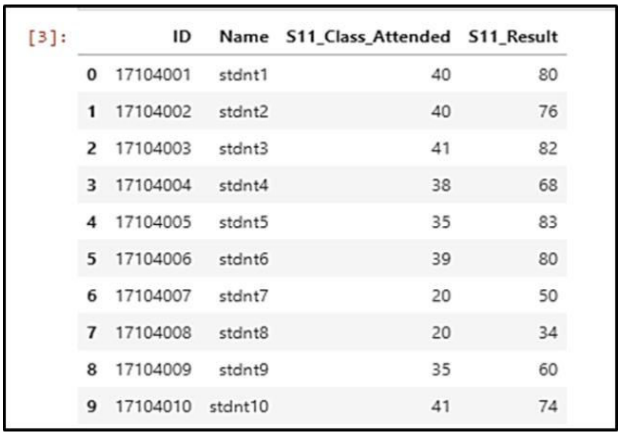
To make data more understandable for humansand pull insights from it, effective visualization is a must. In this research, we used a scatter plot as it was suitable forour needs. A scatter plot is a visualization method that displaysthe values of two different variables as points. The data foreach point is represented by its horizontal (x) and vertical (y) position on the visualization. We used this plot-ting techniqueto find out the dependency among the attributes through thex and the y axis. It can give us insight into how the featuresare related to one another and also show the groupings amongthe data and outliers.

Necessary measures

After implementing the various algorithms, we can hope to find out the features that are directly linked to students' performance. Clustering methods can help to find out groupings of the students' performance or features as well as to get knowledge about the dependencies among the attributes. Hence, this can help us to pinpoint the features that are influencing students' performance. From the findings, the authority can focus on those features and take necessary measures to improve them which will eventually provide a better outcome for the students. Through the classification and the regression method, we can predict if a student is likely to pass or fail. It is possible to provide timely warning and support to low-achieving students and advise high-performing students (Raheela *et al.,* 2017). We may even predict how many marks or points will they get according to their academic records with a notable amount of accuracy which is very crucial for the students. With this information, we can provide various measures to the students who have the pos-sibility of failing or performing badly. The meas-ures can be additional classes, individual tutoring, implementing more effective teaching methods, counseling to gain focus, etc.

RESULTS:

In the result part, we concentrate on the outcome of the research and know the findings after implementing the algorithms. We compared the actual result with our expected result by implementing the K-means clustering, K-nearest classification and multiple linear regression algorithms.



[3]:

	ID	Name	S11_Class_Attended	S11_Result
0	17104001	stdnt1	40	80
1	17104002	stdnt2	40	76
2	17104003	stdnt3	41	82
3	17104004	stdnt4	38	68
4	17104005	stdnt5	35	83
5	17104006	stdnt6	39	80
6	17104007	stdnt7	20	50
7	17104008	stdnt8	20	34
8	17104009	stdnt9	35	60
9	17104010	stdnt10	41	74

Fig. 3: Selected data for clustering.

Here, we have selected class attended (S11_Class_Attended) and result (S11_Result) attribute of 1st year 1st subject from our main dataset to implement clustering among them. The updated dataset is given in **Fig. 3**. The value of k determines how many clusters we will get. Therefore, we need to choose the optimum k value which is difficult for a human to select. Hence, we used the Elbow method and chose k=3 as it is on the elbow point of the graph as shown in **Fig. 4**. Which is used to select the k value. We choose the Elbow point to determine the k values.

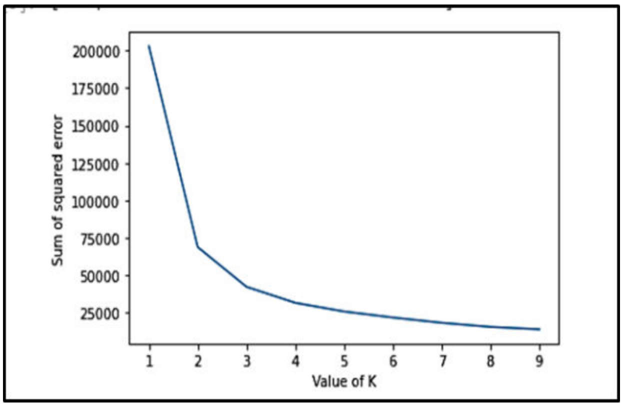


Fig. 4: Elbow plot for K selection.

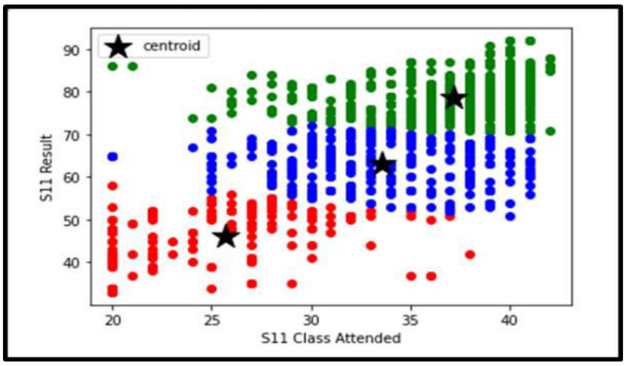


Fig. 5: The cluster of S11 class attended and result.

The above **Fig. 5**. shows cluster without the normalization process, therefore the clusters lookscrambled. After normalization, the clusters are more well-formed and easier to distinguish. From **Fig. 5** we can observe that the students with better class attendance have a better result. We can see that the S11 class attendance and S11 result have an almost linear relation among them. This information not only proves the importance of class attendance for a better result but also provides further analysis potential for this attribute. We can also see the groupings of the data points which show us that majority of the student have good and average class attendance thus having good and average results as well as performance. We implemented a similar process for the various subject of the same and different years to validate the above information about class attendance attributes or features.

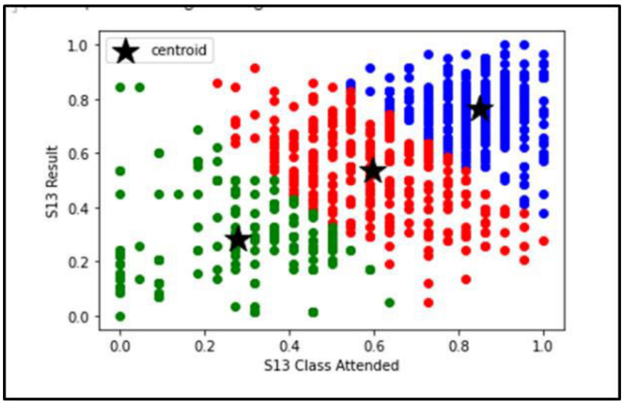


Fig. 6: Cluster of S13 class attended and result.

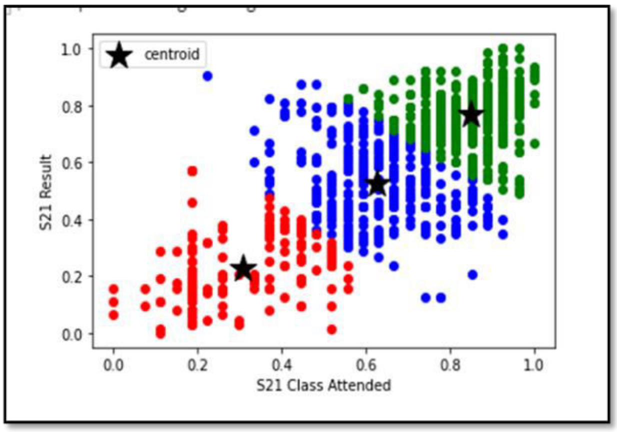


Fig. 7: Cluster of S21 class attended and result.

From the above **Fig. 6** and **Fig. 7**, we can observe that they show similar patterns and characteristics to **Fig. 5**. This validates all the information we gained from our cluster for the class attendance feature. Following the same procedures, we collected inform ation and insight into the other attributes. Every cluster can give us useful information which will be beneficial for our objective. They can also provide us with information on the factors that influence a student's performance.

From **Fig. 8**. we can observe that the student's class performance and results create scrambled clusters where the clusters are not distinguishable from one another. We can see that the S11 class performance and S11 result have nonlinear relations among them. This information proves the importance of class performance for better results is relatively less than other features. If we look deep into it, we can find the reason as very few students are responsive to their lecturers or teachers. Most often students feel shy and fear as some of them are introverts. We can also see the groupings of the data points which show us that majority of the student have good and average class performance marks including one outlier.

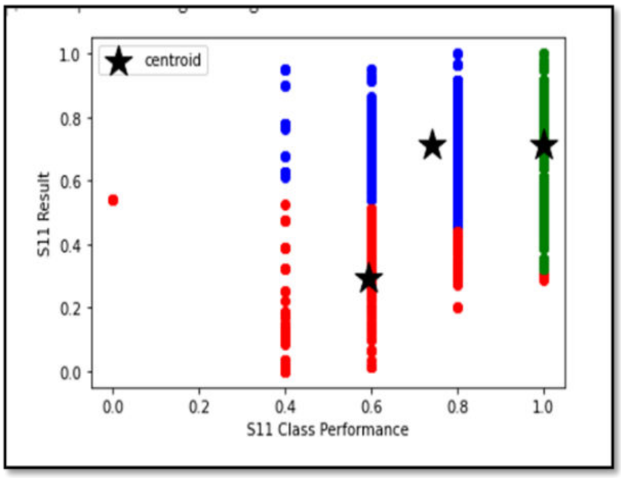


Fig. 8: Cluster of S11 class performance and result.

Now, we implemented a similar process for the various subject of the same and different years to validate the above information about class perform-ance attributes or features. For example, the cluster of S13 class performance and result is given in **Fig. 9**. It shows a similar pattern and result.

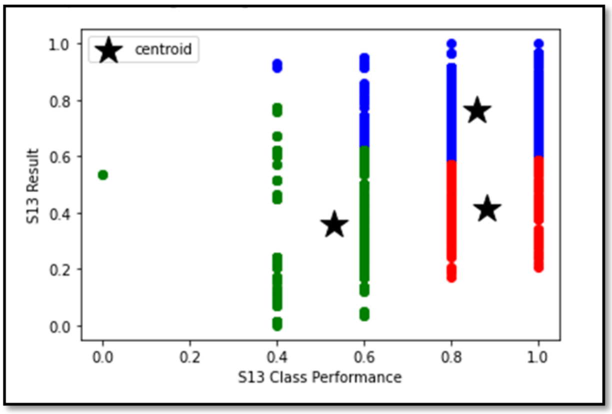


Fig. 9: Cluster of S13 class performance and result.

From the **Fig. 10**. we can observe that the students with better class test marks have a better result. We can see that the S11 class test and S11 result have an almost linear relation among them. This information not only proves the importance of class test marks for a better result but also provides further analysis potential for this attribute. We can also see the groupings of the data points which show us that majority of the student have good and average class test marks thus having good and average results.

We implemented a similar process for the various subject of the same and different years to validate the above information about class test attributes or features. For example, the cluster of S21 class test marks and results is shown in **Fig. 11**. It shows a similar pattern and result.

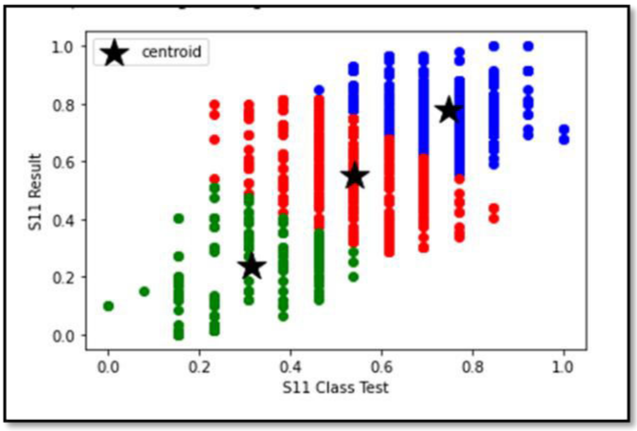


Fig. 10: Cluster of S11 class test and result.

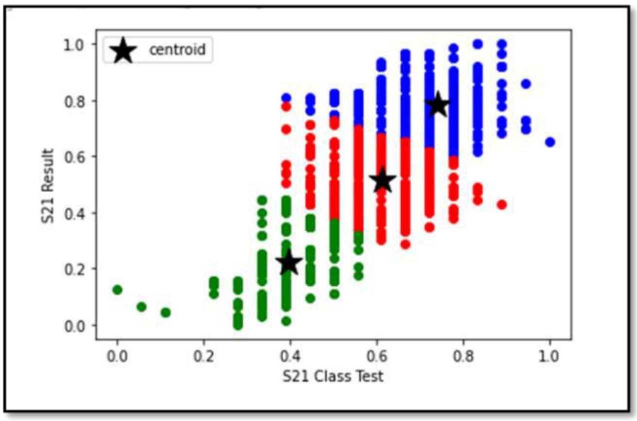


Fig. 11: Cluster of S21 class test and result.

From the **Fig. 12**. we observed that the students with better exam marks have a better result. We can see that the S11 exam and S11 result have linear relations among them. We know that exams hold the highest mark in students' overall results. This information not only proves the importance of class test marks for better results but also provides further analysis potential for this attribute. We can also see the groupings of the data points which show us that majority of the student have good and average class test marks thus having good and average results.

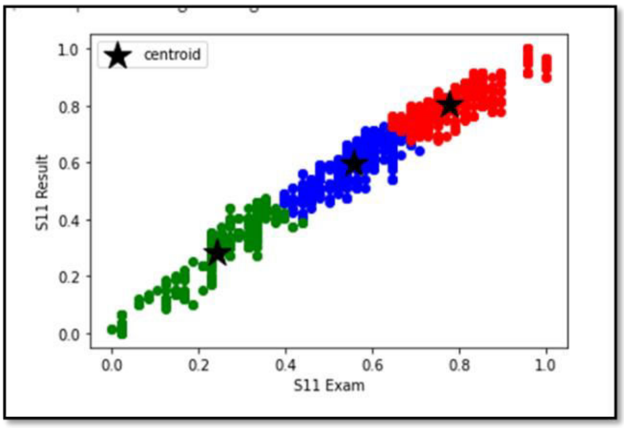


Fig. 12: Cluster of S11 exam and result.

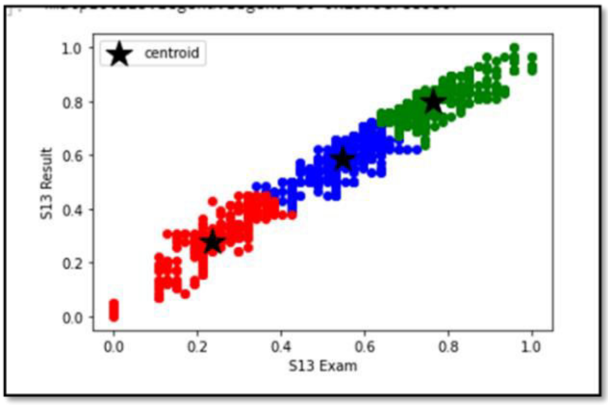


Fig. 13: Cluster of S13 exam and result.

We implemented a similar process for the various subject of the same and different years to validate the above information about exam attributes or features. For example, the cluster of S13 exam marks and results is shown in Figure 13. which shows a similar pattern and result. All the clusters we used were to identify and verify significant features of a subject and its effect on the result. We also found groupings of the students through those graphs. There is another way we can utilize the clusters to find more information and knowledge.

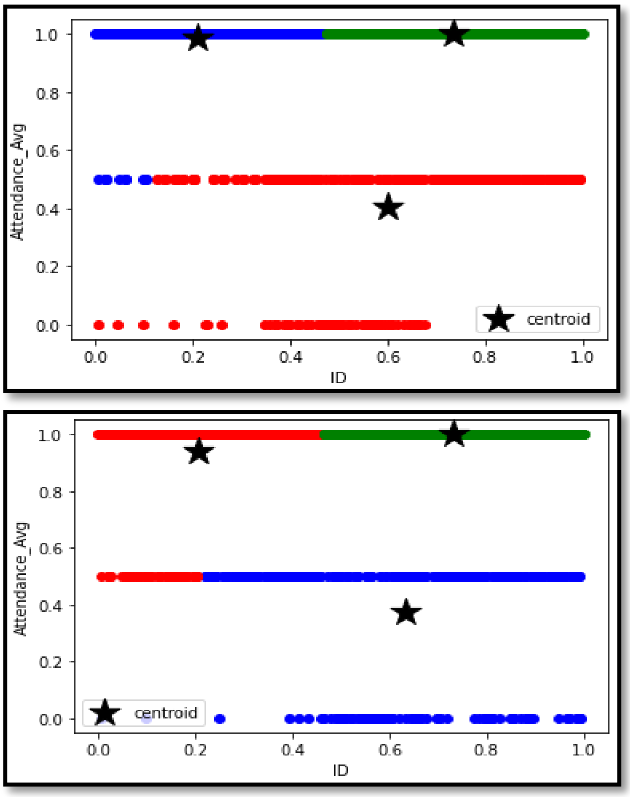


Fig. 14: Comparison between first year (above) and second year attendance.

We used this opportunity to find information about our cause. We implemented clusters to compare students' performance over two years which gave us insight into if students' performance is increasing or degrading. We also compared various subject features or attributes to get information on their status.

Here, from **Fig. 14** we can identify the difference between the attendance of the first year and the second year. We can look into the matter to find out the cause for why a student's attendance percentage has changed relative to the previous year.

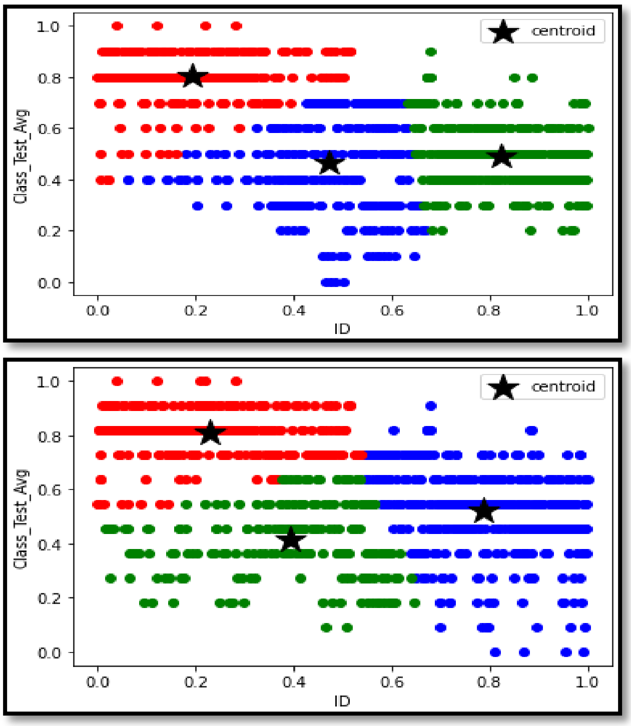


Fig. 15: Comparison between first year (above) and second year class test.

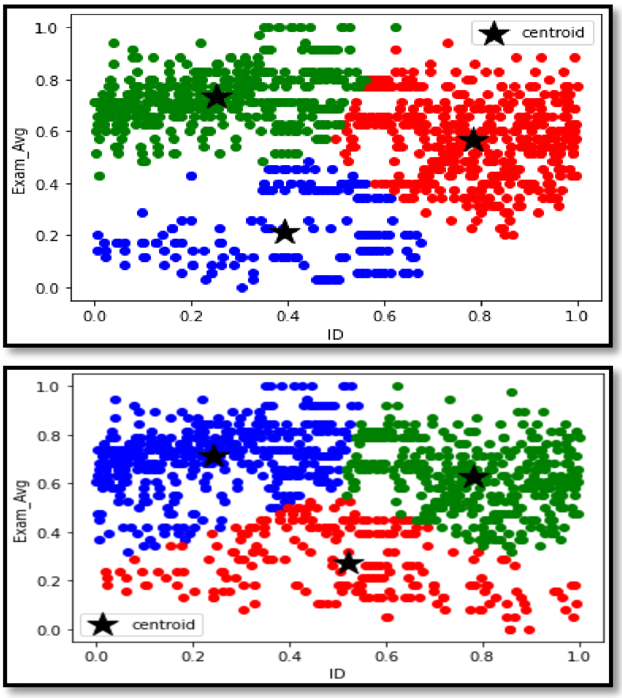


Fig. 16: Comparison between first year (above) and second year exam.

Similarly, from **Fig. 15** we can identify the difference between the class test mark for the first year and second year. We can look into the matter to find out the cause for why a student's test mark has changed relative to the previous year. Then we can provide feedback according to the changes.

From **Fig. 16**. we can identify the difference between the exam mark of the first year and second year. We can look into the matter to find out the cause for why a student's exam percentage has changed relative to the previous year. We can take the necessary steps and provide feedback according to the findings to improve in future exams.

From **Fig. 17** we can identify the difference between the result of the first year and second year. We can look into the matter to find out the cause for why a student's result has changed relative to the previous year. We can take necessary steps and provide feedback according to the findings to improve future results and work on the issues that are causing the changes.

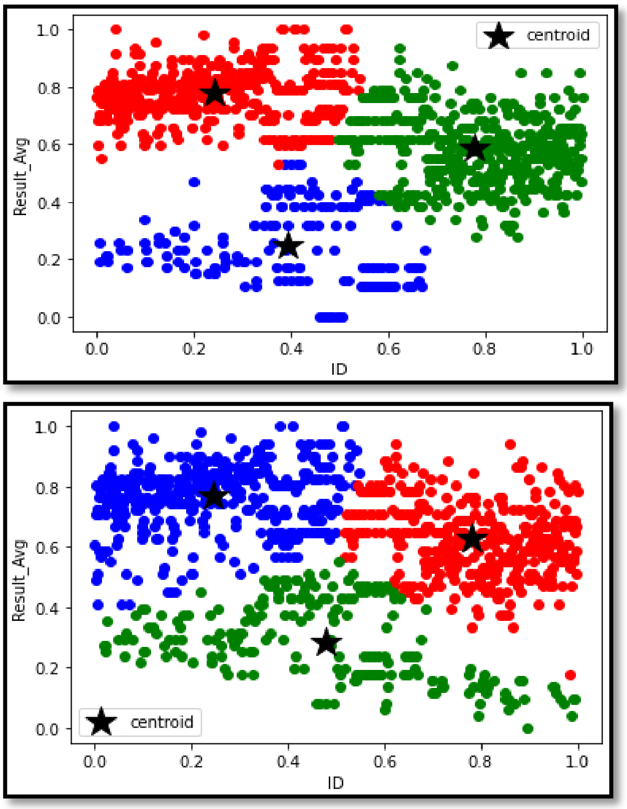
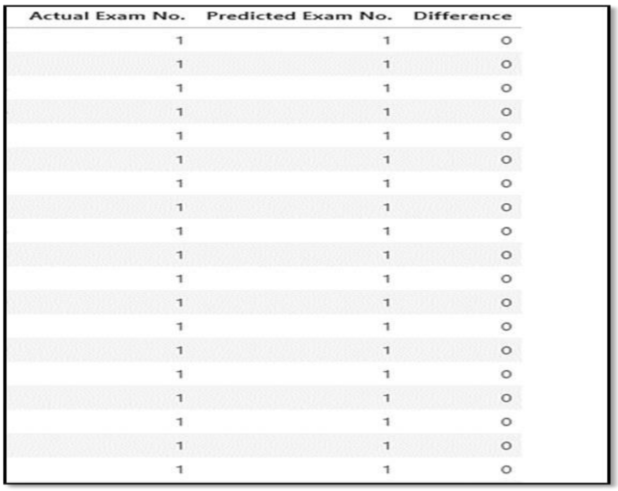


Fig. 17: Comparison between first year (above) and second year result.

Classification

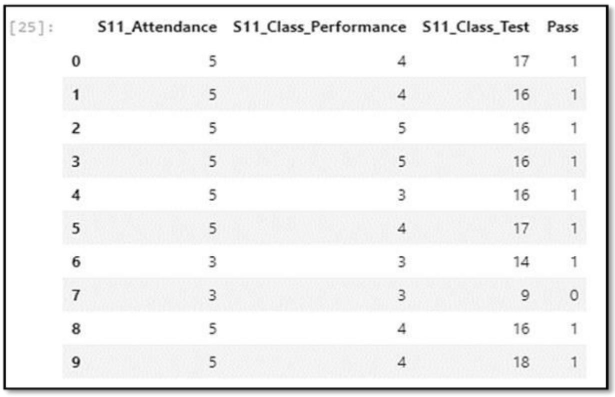
Classification techniques are used to build an educational model based on knowledge discovery in databases to predict learner behaviors. We implemented the K Nearest Neighbor classification method to predict students' results. As KNN can handle categorical data, we used various academic features to predict if a student is likely to pass or fail according to his or her available data.



Actual Exam No.	Predicted Exam No.	Difference
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0

Fig. 18: Selected attribute for KNN categorical data prediction.

In this part of the research, we predicted if a student will pass (1) or fail (0) according to his academic attributes such as attendance, class performance and class test. These three available attributes can be recorded before the final exam. This means there will be sufficient time to analyze these data before the exam and take necessary measures according to them. If the prediction given in **Fig. 19** shows that a student is likely to fail according to his or her current statistics various measures like extra classes, counseling, different teaching methods, etc. can be applied. Which will hopefully be beneficial to the students and the research objective. We can also find out the problems and the issues that the students might be facing and provide a solution to them. It will also give the authorities insight into how to cope with these problems or prevent them in the future.



[25]:

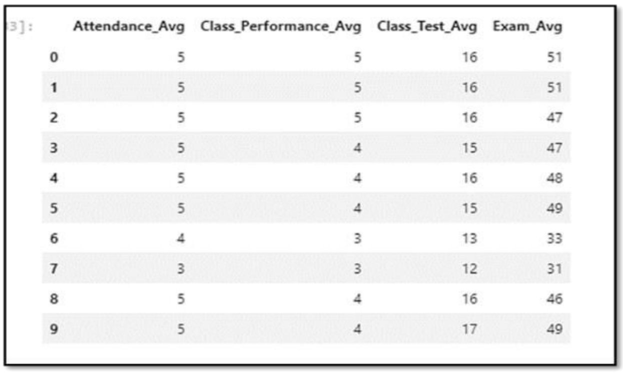
	S11_Attendance	S11_Class_Performance	S11_Class_Test	Pass
0	5	4	17	1
1	5	4	16	1
2	5	5	16	1
3	5	5	16	1
4	5	3	16	1
5	5	4	17	1
6	3	3	14	1
7	3	3	9	0
8	5	4	16	1
9	5	4	18	1

Fig. 19: Prediction through KNN categorical data.

Fig. 19 shows the actual exam status and predicted exam status. In our KNN classification, we got an accuracy rate of 99% as it predicts categorical data very efficiently.

Regression

As KNN could not handle continuous values we used multiple linear regression to predict continuous values like exam values. This method uses multiple independent variables to predict dependable variables. We have predicted students' exam numbers according to their various academic features. After predicting their exam values before the exam takes place, we can take necessary measures according to the findings. These measures might be very beneficial for their outcome and the findings can also help us to gather significant patterns and knowledge.



3]:

	Attendance_Avg	Class_Performance_Avg	Class_Test_Avg	Exam_Avg
0	5	5	16	51
1	5	5	16	51
2	5	5	16	47
3	5	4	15	47
4	5	4	16	48
5	5	4	15	49
6	4	3	13	33
7	3	3	12	31
8	5	4	16	46
9	5	4	17	49

Fig. 20: Selected attributes for the Regression algorithm.

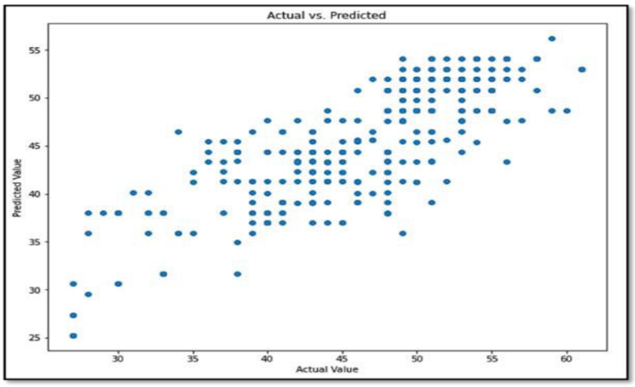
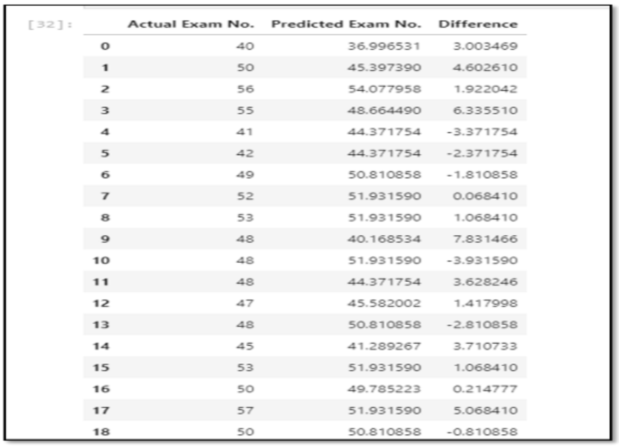


Fig. 21: Actual value vs predicted value for Linear Regression.

The **Fig. 20** shows the attributes that we worked on. Here, academic attributes like attendance, class performance, and class test average work as independent variables which are available before the exam. We used those to predict exam marks which are dependent here. From **Fig. 20** we observed that exam values are continuous so multiple linear regression was the best option for our objective. The attribute which has the highest coefficient value will have the most influence on the result.

Fig. 21 represents the actual values (x-axis) and the predicted values (y-axis) of the exam. This figure gives a certain idea of the regression model's accuracy. We observed that the actual values and the predicted values are almost linear. It means we can assume that the predicted values are almost similar to the actual values. Although there might be some outliers which is acceptable in this case as the values are continuous. Therefore, with a significant amount of efficiency, we can predict the exam values and take proper steps according to the findings. During clustering, we saw that the exam was most linear to the overall result. It has the most dependencies on the result. So, in this manner, we can get the idea of the overall result prediction through exam predicttion. Through this process, we can implement various other factors to identify their significance and use them to predict the result or see their influence on the result.



[32]:

	Actual Exam No.	Predicted Exam No.	Difference
0	40	36.996531	3.003469
1	50	45.397390	4.602610
2	56	54.077958	1.922042
3	55	48.664490	6.335510
4	41	44.371754	-3.371754
5	42	44.371754	-2.371754
6	49	50.810858	-1.810858
7	52	51.931590	0.068410
8	53	51.931590	1.068410
9	48	40.168534	7.831466
10	48	51.931590	-3.931590
11	48	44.371754	3.628246
12	47	45.582002	1.417998
13	48	50.810858	-2.810858
14	45	41.289267	3.710733
15	53	51.931590	1.068410
16	50	49.785223	0.214777
17	57	51.931590	5.068410
18	50	50.810858	-0.810858

Fig. 22: Actual and Predicted exam numbers.

Fig. 22 represents the actual exam numbers, predicted exam numbers, and their differences. As shown in **Fig. 22** the actual and predicted values are very close to each other despite being continuous values. We can see that most of the differences between the two values are very small although there are a few exceptions where the difference is relatively large. We can take various necessary measures for those students who are likely to get less or unexpected marks in the exam according to our findings. It can give us insight into the lacking's and clear the path to overcome them. We can use our findings or knowledge to develop more effective teaching methods and avoid mistakes in the upcoming future for a better outcome. It will hopefully help the students to perform better and keep on track which will bring out their full potential.

CONCLUSION:

The education sector is undoubtedly one of the most significant sectors for any nation. The development of this sector is essential for long-term economic and national prosperity. Like all other organizations, educational organizations also store large amounts of data that contain various features. Through analyzing these data, we can find interesting patterns and knowledge which might be beneficial to the development of the education sector. Our research is based on measurement and prediction methods that may influence the students' performance and sentiment. We can evaluate the students in various manners and look into their data from different perspectives. Through these methods, we can gather various knowledge and patterns and provide feedback or necessary measures according to the findings. Tak-ings the necessary steps or measures that help the students to stay on their track and bring out their best possible outcome. It can help us to identify the issues and problems that affect the performance of students and get an idea of how to overcome them in the upcoming future. There are huge potential appli-cations for our research. The future scope of our research is as follows:

1) Development of effective teaching methods
2) Maximization of educational institutes' efficiency and resource allocation
3) Improvement in the decision-making process for higher education and career
4) Potentiality to identify effects of various other attributes on students' performance.
5) Potentiality to track performance throughout a student's career and establish effective student profiling
6) Identification of undesirable behaviors & early dropouts of students to provide appropriate advising or counseling

ACKNOWLEDGEMENT:

This is to express our warm thanks to the dept. of Computer Science and Engineering at the Bangladesh Army University of Engineering and Technology (BAUET) for giving the opportunities and helping the research to reach its successful state.

CONFLICTS OF INTEREST:

The authors state that there is no potential conflict of interest in publishing this research article.

REFERENCES:

1) Alisa B. Z. (2020). Predicting Students' Academic Performance Based on Enrolment Data, *International Journal of Innovation and Economic Development*, **6**(4), 54-61. https://ideas.repec.org/a/mgs/ijoied/v6y2020i4p54-61
2) Dorina K. (2012). Student Performance Prediction by Using Data Mining Classification Algorithms, *International J. of Computer Science and Management Research*, **1**(4), 686-690.
3) Feras Al-O., Anna D. & Babar S. (2017). Analyzing students' performance using multi-criteria classification, *Cluster Computing*, **21**(1), 623-632. https://doi.org/10.1007/s10586-017-0967-4
4) Lubna M. A. Z. (2019). Prediction of Student's performance by modelling small dataset size, *International J. of Educational Technology in Higher Education*, **16**(27). https://doi.org/10.1186/s41239-019-0160-3
5) Ouafae E. A., Yasser E. A. E. M., Lahcen O. & Ahmed D. (2020). A Multiple Linear Regression-Based Approachto Predict Student Performance, Advanced Intelligent Systems for Sustainable Development (AI2SD' 2019), 9-23. http://dx.doi.org/10.1007/978-3-030-36653-7_2
6) Raheela A., Syed A. A. & Najmi G. H. (2017). Analyzing undergraduate students' performance using educational data mining, **113**, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007
7) Suhem P., Zain Z. & Fatima M. (2012). Application of data mining in educational databases for predicting academic trends and patterns, 2012, *IEEE*, *International Conference on Technology Enhanced Education* (ICTEE), 1-4. https://doi.org/10.1109/ICTEE.2012.6208617

Citation: Pal AK, Sumona S, Rashid MA, Hasan MAFMR, Kumar M and Hossain MA. (2022). Students performance measurement and prediction based on academic features through the machine learning. *Aust. J. Eng. Innov. Technol.*, **4**(4), 65-77. https://doi.org/10.34104/ajeit.022.065077