Clustering of Student data using KNN-FCM techniques

[1]M. Satya Srinivas, [2]S Siva Rama Raja, [3]MVVAL Sunitha, [4]PVVS Eswar Rao

Sasi Institute of Technology & Engineering, Andhra Pradesh, India

[1]maddipativas@gmail.com, [2]sivaramaraja@sasi.ac.in, [3]mvvlsuntha@sasi.ac.in, [4]eswar@sasi.ac.in

Abstract

Even though it is crucial for planning and administering educational pedagogy, the study of students' academic performance is a major concern for universities and schools of higher learning. Students must be classified based on their performance in the system. It's also simple to develop and control the teaching pedagogy based on the students' comprehension levels. Many machine learning algorithms for education have been suggested during the last decade. Problems arise when trying to enhance student achievement with data that is skewed. As a result of this investigation, this article gives a comprehensive review of the subject topic. Students may be clustered using machine learning algorithms based on an unbalanced dataset. As a first step, use the K-Nearest Neighbors (KNN) technique to fill in any blanks in the characteristics of the dataset. Fuzzy C-Means can be used to group students based on their performance in the entrance exam. The values of the missing characteristics in the dataset are filled using the KNN method. Fuzzy C-means method is then used to group the data. This technique is capable of accurately predicting student performance in the early stages of the year. It's possible for students to understand about their own learning and comprehending abilities so they may develop. Teachers are able to quickly determine the grasping capabilities of their pupils. Planning and managing the instructional strategy depending on students' comprehension levels is a snap. Academically, they may improve their standing by increasing their grades.

Keywords: *Teaching Pedagogy, Student Performance, K-Nearest Neighbor, and Fuzzy C-Means*

1. Introduction

Data mining methods have been utilised by universities to investigate educational records, such as student performance, instructor ratings, gender inequality, and so on, in educational institutions. Pupils' enrollment and dropout rates, early identification of failing students, and correct resource allocation might all be improved with the use of data mining tools.

Student accomplishment is significant in educational institutions as it is commonly used as a criteria for judging the institution's performance. Universities are strengthening their efforts to quantify student accomplishment by tracking student engagement and involvement in online learning environments [6]. Early detection of at-risk kids, as well as preventive actions, may dramatically enhance their chances of success. Machine learning methods have been frequently employed for prediction in recent years. Machine learning systems have a lot of promise for supporting teachers in recognizing poor student performance by giving system capable of detecting risk before it becomes a serious problem. Henceforth, teachers may commit extra time to such challenging kids in order to prepare them for summative exams.

Cluster analysis organizes observations into clusters to seek for patterns in data collecting. The objective is to produce an ideal clustering in which the observations within one cluster are similar but different from those within the other.. The cluster analysis involves two elements: first, a measure that indicates the degree of similarity between items; and second, a mechanism for grouping or clustering things (methods) [5]. Clustering is crucial in data analysis and data mining applications. To put it another way, the purpose of clustering is to group similar objects together so that they may be more readily recognised (clusters) (clusters). Clustering may be performed using a number of techniques, including hierarchical, dividing, and grid approaches.

In the education industry, predicting student performance and classifying data are crucial. The major purpose of this initiative is to cluster the pupils as per their performance in the entrance exams by using the KNN-FCM techniques. Based on their learning capability, cluster them into different categories or groups. So, we can improve the planning and management of teaching pedagogy by assigning the appropriate faculty to the students. As the faculty can teach according to the level of the students' capability and this leads to an increase in their performance. Thus, it increases the overall performance of the institution.

2. Related Work

There are concerns in the academic community about monitoring and analysing student academic performance, according to James Manoharan et al. (2014) [1]. The author explains how to calculate the similarity distance in simple terms using the Euclidean distance and a predictive clustering algorithm. The author uses K-means clustering and a deterministic combination model to apply the technique to a data set of computer science students' results from six courses. k-means' appeal is based on its scalability, simplicity, and speed in dealing with sparse data. A starting point for the k-means method is the original centroid coordinates and the coordinates are adjusted to meet the Equation's objective function until they do. The smallest local value can be

found using K-means clustering. When the initial cluster points are chosen, they determine which minimum is found. Keep adjusting centroids until the k-means method finds the local minimum. In the K-means algorithm, each cluster's distance and centroids are computed using l-iterations, where n is the number of iterations needed to discover the clusters The number of iterations l is determined by the initial cluster points. O(nVl) is the total number of iterations for k-means clustering, which means the total time complexity of the algorithm is O(nVl).

Individual performance results were evaluated numerically among 65 students. Students at higher education institutions can be monitored more effectively thanks to k-mean clustering, which allows institutions more accurate data in a shorter period of time. According to YuniYamasari et al. (2020) [2], traits that are unimportant can be deleted from clusters to improve them. Students can be assigned to groups based on their performance clusters with a high degree of accuracy. Non-dominant data included in the clustering method may necessitate some refinement of the results. Unsupervised evaluation has traditionally used only one metric to assess the quality of clusters. The Gini Index feature selection method has been employed in this study. A clustering strategy called K-means is used in this work for mining purposes. Using silhouette coefficient, ANOVA, and t-test as metrics for evaluating the study's findings, the paper concludes. The Gini index measures the degree to which a feature is isolated from other features within a given category. When points from one cluster are compared to those from another cluster, the silhouette value is calculated. Clustering techniques such as K-means and ANOVA were also utilised. It was determined if the clusters formed by each method differed significantly based on the value of the silhouette. In the end, the t-test was used to compare the cluster quality between the two methods based on the silhouette value.

Analytic strategy for e-learning data analysis was described by Alana M. de Morais et al. (2015) [3]. Responses from students will be used to form groups. The paper's two most important objectives are to lead the next learning activities and to identify which criteria are most relevant to tutor support for each group. It is possible for teachers to make better decisions by using the VLE's helpful responses. If a pedagogical resource didn't have the desired effect, find out why. Expertise in virtual classrooms is essential to resolving issues and establishing new learning methods for students. For instance, some children learn better by doing physical activities while others learn better by comprehending the concepts they're studying. With visual cues, some are better than others at following directions (pictures, graphs, or diagrams). A variety of teaching methods can be used to teach the same subject matter. Stepwise Backward Regression, a semi-automated technique that relies on the projected coefficients of the variables to add or remove them from consideration in the t-test, was used by the researchers to build a model. A separate study found that "Incorrect" and "Correct First Attempts" appear in three regression models generated by the approach. This study's findings shed light on two different aspects of VLE students' answer profiles. In the beginning, multivariate analysis methods are used to characterize the behavioral characteristics of answers using Open Learning Data. The findings here also add to the current understanding of how

student performance impacts instructor decisions in virtual learning environments (VLEs).

Luz The research carried out by Stella Robles Pedrozo et al. (2013) [4] With the help of an adaptive technique, it is possible to classify and group students with similar conceptual flaws in order to facilitate feedback between teachers and students. Student performance evaluation has always been important in the learning process because it provides information about the student's level of knowledge and progress. It's difficult to pinpoint a single issue because students are judged on their abilities. Building an assessment with items covering all the areas to be evaluated and questions of varying complexity can be one of the most difficult challenges in conducting assessments and trying to estimate ability levels of those being assessed. How to arrive at a realistic estimate of the parameters that will be used to evaluate a person's level of expertise is explained in detail in this article.

A 10-item exam was given to a group of 141 students in tenth grade who were taking a Mathematics subject. Given that the entire feedback method was done manually before this experience, the test proved to be extremely beneficial. There are just one correct answer for each question on the exam. Each question offers four response options (one of which is accurate; the other three are incorrect). In order to evaluate their students' abilities in the topic of randomness, a group of math teachers designed this test "k" is an array of "data-frame" variables, which contains all the student responses (in zeros and ones).

Sonali Shankar et al. (2016) [5] purpose is to analyse the students' performance in respect to their nation based on numerous variables. Data from Harvard University's online students for the academic year 2013-2014 was made available on May 14th, 2014. The study's purpose is to employ K-mean clustering to examine the student dataset. Clustering is the strategy applied in this article, which splits the dataset into two or more groups for simpler analysis. Distance is the basic parameter for cluster formation. Different clustering approaches, including as DB scan, partitioning, hierarchical, grid, and model-based clustering, may be used to compute the distance between two locations using the Euclidean formula. K-mean clustering's optimal value is defined by the silhouette index (SI) (SI). Cluster Sum of Squares is a measure of the compactness of the clusters' points (WSS) (WSS). International students' average performance is judged by looking at factors like organised activities, chapters studied, and how many days they spend engaged with the course. Consequently, these attributes are compared to averages across countries, and it is concluded that grades are not the lone predictor of high course understanding.

Ravneil Nand et al. (2021) [6] employed Clustering to investigate college student's online presence. The purpose of this article is to look into how Covid-19 has affected the online presence of students in two courses of undergraduate mathematics. The students' actions in these two classes were tracked using activity logs. Each student's click history is saved in the activity log. Lecture capture records are kept separate on the Opencast system. There are two steps to students' online presence in undergraduate courses. In Step 1, students were clustered using a Self-Organizing Map (SOM) to determine how effective their interaction was post-Covid-19 for the two courses. The data were

examined with Matlab software, which resulted in the clustering of online student presence. Step 2 was to perform a Prediction on Semester data to assess how well the qualities matched the data. Through clustering and prediction, the analysis of student performance in two undergraduate mathematics courses was explored in this study. It can be seen that clustering can result in the formation of diverse online presence clusters. The levels of online presence are matched to student performance in these areas. The data was used to make the prediction, and the regression diagram showed that there is a linear relationship. A major advantage of clustering is that it is a simple approach to implement can divide students into groups. The drawback of this methodology is that it takes longer to implement because it requires a large number of samples.

Nova Rijati et al. (2020) [7] presented a Students' academic conduct may be utilised as a technique of identifying students who have the potential to be successful entrepreneurs. Knowledge discovery in databases depends significantly on the quality of data, which consists of a set of characteristics that define the data's properties. Optimal data mining performance needs the suitable attribute selection method. Mapped aspects impacting student entrepreneurship employing cognitive science theory and features from Indonesia's Higher Education Database were utilised to train machine learning models in this scenario. Dataset qualities may be examined using four distinct methods: correlation, information gain, one R, and relief F. Clustering data using the Simple K-Means approach leads in a 17 percent decrease in Sum of Squared Errors when compared to the other three clustering methods. The instances cluster profile created vary based on the most essential feature variances in each attribute selection approach.

Uday Kumar et al. (2020) [8] utilising machine learning approaches to forecast the performance of a student at a university Predicting the result of given inputs may be done using machine learning approaches.. In nature, hierarchical clustering algorithms are split into two types: aggregative and divisive. The three agglomerative hierarchical clustering approaches are single linkage clustering, average linkage clustering, and full linkage clustering. three unique forms of hierarchical clustering are available: single linkage clustering, the average linkage clustering, and full linkage clustering. They employed the complete linkage clustering technique. The leaf node represents the label, the branch reflects the outcome, and the internal nodes reflect the qualities. In addition to being a prediction and classification approach, Naive Bayes is also both. The Bayesian theorem is utilised to develop the Naive Bayes classifier. A dataset of K L University students was submitted to clustering and classification in order to evaluate their performance and forecast whether or not they would pass a technical test administered as part of the recruitment process. Accurate results have been reached.

Ishwank Singh et al. (2016) [9] assessed students' performance using a clustering technique. Data mining was utilised in this research to examine student performance and put students into numerous categories. students must constantly grow in order to compete in today's world. The practise of collecting data from a collection of data is known as data mining. Extracting information from data is termed data mining, from a piece of data and works with a broad variety of patterns. Clustering requires arranging comparable things in one cluster and dissimilar ones in another. The K-means clustering technique was utilised to group the data in this investigation. In the K-Means method,

each observation is allocated to the cluster with the closest mean among the 'K' groups. Clustering analysis was offered by the author as an acceptable standard/benchmark to examine whether students' performance increased consistently over time in this assignment. The analysis is beneficial in the admissions and placement procedures. Clustering data is done using K-means as it is simple to implement and has strong computational efficiency.

Shiwani Rana et al. (2016) [10] examined student performance at an institution using a Hierarchical Clustering Algorithm. Important classification and grouping strategies are discussed in this research, It may be utilised to assess student success in the third semester of the BE (Information Technology) curriculum. Students at the University Institute of Engineering and Technology (UIET) at Panjab University apply the Hierarchical Clustering Algorithm to estimate their digital electronics proficiency (PU) (PU). For the aim of grouping the data, the K Means approach is applied. If the class or labels for any data are unknown, then the clusters must be found without this information. This is a sort of unsupervised learning There is a short comparison of a few machine learning algorithms in this publication. Students' performance is measured using a range of classification and clustering methods. students of UIET, PU, Chandigarh's third-semester BE (Information Technology) degree is grouped using the Hierarchical Clustering Algorithm. Using the Hierarchical Clustering Algorithm, theoretical analysis is done on the top five students' grades.

Sujith Jayaprakash et al. (2020) [11] suggested a “Predicting Students Academic Performance using an Enhanced Random Forest Classifier” utilising the improved random forest method. The major purpose of this study project is to explore the components that impact children' academic achievement and to help in the identification of youngsters who are at risk. Other algorithms, like Naive Bayes, Bagging, Boosting, and Random, fade in light. The improved random forest approach expects to boost categorization and prediction accuracy.. The data illustrate how different categorization methods including the Nave Bayes Algorithm, Bagging, Logit Boosting, Random Forest, and Enhanced Iterative Random Forest compare. With a 93 percent accuracy rate, the improvised random forest beat the other classifiers, as evidenced in the data below.. In addition to analysing student performance, this research examined the efficacy of many models developed utilising classification algorithms. The Random Forest algorithm beat the other ensemble techniques including the Nave Bayes algorithm with greater iteration and bag size.

Dagim Solomon et al. (2018) [12] offered a “Predicting Performance and Potential Difficulties of University Student utilising Classification: Survey Paper” using knowledge discovery techniques (ID3, random forest, artificial neural network, and logistic regression) (ID3, random forest, artificial neural network, and logistic regression). The major purpose is to anticipate the performance of the student utilising the existing data set based on the dominating characteristic. With the aid of these educational data mining methods like ID3, random forest, artificial neural network, and logistic regression, and academic performance of the student and diverse factor effect that lead the student to the failure are successfully detected. Classification, grouping, and prediction are all strategies for assessing a student's performance. The benefit of

assessing student performance is to assist the student pay attention to who is at danger point or not. Another advantage is the benefit of the institution to take various activities to right track the student. To study the student data for constructing models, data mining methods are employed, notably classification algorithms. The categorisation helps prediction.

Yupei Zhang et al. (2020) [13] proposed a "Graphs Regularized Robust Matrix Factorization and Its Application on Student Grade Prediction" using the Regularized robust matrix factorization (GRMF) using graphs. Many investigations make use of the method of matrix factorization (MF). This research uses GRMF to examine two publicly available data sets for rating prediction and picture recovery in order to validate our approach. Our university has 1325 students in 832 classes, and this article uses GRMF to analyse this data. GRMF's testing findings clearly demonstrate that it is able to handle a wide range of data problems and produces more effective characteristics than other approaches. On our educational data set, GRMF also outperforms other approaches in terms of prediction accuracy. In higher education, this strategy can help with tailored teaching and learning.

Prediction of Student Performance through Modeling Small Dataset Size was proposed by Abu Zohair et al (2019) [14] [15]. Project objectives are to demonstrate that it is possible for a small dataset, and to demonstrate that an accurate prediction model can be built. For this study, we also investigate how to identify significant Prediction model building indicators in a small dataset by employing visualisation and clustering techniques The best indicators were examined using a variety of machine learning approaches in order to discover the most accurate model possible. For the strategies we examined, the clustering algorithm was the most effective at identifying relevant indicators even in small datasets. SVM and discriminant analysis algorithms were found to be effective in training small datasets and achieving acceptable accuracy and reliability rates, according to the primary findings.

Nave Bayes and Rule-Based classification methods were used by Fadhilah Ahmad et al. (2015) [15] to predict students' academic performance using classification data mining techniques [15].]. Classification algorithms are most accurate when used in the following situations, according to their findings: There was a large amount of data that had to be mined, often thousands. Only a few data points in the dataset for analysis are distorted or missing. WEKA is also used in this study to compare three categorization methods: DT, NB, and RB. The experimental results show that the RB has a better classification accuracy than the RB-based system in comparison to the NB and DT. Teachers will be able to intervene earlier to help students with average or below-average grades. A limitation of this study is the lack of data, which is due to incomplete and missing information. More data from different years or universities will be added to this study in the future to improve the prediction's accuracy.

According to a model proposed in BratislavPredic et al. (2018) [16], a student's final grade is influenced by their performance in various educational settings. A comparison of classifier performance was carried out to find the best classifier for a multiclass feature dataset. The majority vote strategy is then used to build an ensemble of Nave Bayes, Hidden Nave Bayes, J48 decision trees, and Random Forest. Using Naive Bayes, Hidden

Naive Bayes, J48 decision trees, and Random Forest classifiers, we constructed an ensemble using the majority vote strategy. Experimental results show that the proposed model can now more accurately predict students' grades in a mixed learning environment. This paper makes a significant contribution by developing an effective multi-class prediction model that can be used in the context of the study reported in this paper.

A model based on artificial intelligence was suggested by Khalfan Al Mayahi et al. (2019) [19]. This model seeks to predict whether or not a student will succeed in a current course by assessing his former grades and test results. If you're making a prediction like this, you should do it at least two to four months before the exam date. Because of this, students and teachers alike will be able to devote more time to learning this field. Classification is one of the most common supervised learning tasks in machine learning. Several machine learning methods, including linear regression, support vector machines (SVM), and the naïve Bayesian classifier, were investigated. They found that SVC and Elastic Net were accurate models. We used a model that predicted whether or not the student would pass the test. ' Accuracy was found to be 87 percent.

Suhas S Athani et al. (2018) [20] proposed a method to automatically classify students into five categories using Multiclass SVM. This method takes into account failure rates in terms of basic subjects like mathematics. Portugal's school used student reports and a questionnaire to collect the project's actual data. Students will be grouped according to the grades they received in the range of grades they were allotted. "A" is the highest grade, with the next four grades ranging from 'B' to 'D,' with 'F' denoting failure. Many machine learning algorithms are able to do this kind of analysis. Weka contrasted neural networks with multiclass support vector machines. It is because of this that the result is very predictable. K-fold cross-validation is used to assess accuracy, and the Support Vector Machine Classifier is shown to have an accuracy rate of 89%.

The study of Hina Gull et al. focused on predicting the grades of students who would be taking the same course again in the future. This article uses logistic regression, linear discriminant analysis, K-nearest neighbours, classification and regression trees, Gaussian Naive Bayes, and support vector machines to analyse historical data on student grades from one of the undergraduate courses. Student grades (A, B, C, D, E, or F) may be predicted ahead of time by this project, which seeks to identify poor learners and help them overcome hurdles they experienced throughout the learning process. With early grade prediction, instructors are able to identify pupils in need of more assistance in the classroom, and students may strive to improve their grades. To better serve their pupils, teachers may utilise the data to improve their teaching techniques. In this study, researchers looked at students' previous performance in a course to see how accurate their forecasts of future success were. According to the author, linear discrimination analysis was the most accurate way of classification. It was also possible to predict the final test with a 90.74 percent accuracy rate using a model based on LDA.

Several machine learning approaches were utilised to examine the Open University dataset to predict student achievement by Fatema Alnassar et al. (2021) [22]. To improve curriculum and teaching techniques, educational institutions have found that using a range of relevant data to predict student performance is an effective methodology. When it comes to automated educational data processing, Machine Learning (ML) and

Artificial Intelligence (AI) are hot issues in academia. Student performance prediction on the Open University (OU) dataset is solved using machine learning techniques such as Support Vector Classifier, k-Nearest Neighbor (k-NN), and Artificial Neural Network (ANN). Demographics, involvement, and performance are the three most important metrics to look at when analysing educational statistics. The experimental study found that the k-NN technique was the best for OU experiments when compared to the available literature. Changes in how missing values are handled and data standardisation techniques have been attributed with enhancing the outcomes. The study had Data preparation, ML implementation, and critical analysis are the three main steps in the process. According to the experiments, K-NN outperformed SVM and ANN algorithms for a variety of feature permutations. According to student data, the kind of test and the number of prior attempts have a substantial association.

Students' social involvement during the Covid-19 pandemic was researched using machine learning approaches by Jigna B Prajapati et al. (2021) [23]. Nave Bayes, J48 tree, REPTree and random forest approaches are tested on a structured dataset with more than 1200 occurrences to see how they compare. In this article, the research provides and evaluates the most popular social applications and platforms. Additionally, it compares them to other machine learning techniques. Social engagement among students will be impacted significantly by one of the most widely used platform during the Covid-19 epidemic, which is the topic of this research. To summarise the comparative outcomes, this research focuses on accuracy, F-measure, and time. The results and dynamic analysis of the research suggest that preprocessing a student social interaction dataset using a machine learning/deep learning approach may increase accuracy and other aspects. The results may be predicted by comparing the performance of the ML algorithm to the most popular social networking site.

BoddetiSravani et al. (2020) [26] implement innovative teaching and learning with the use of machine learning that takes into account the students' backgrounds, prior academic success, and other variables. This study examines how machine learning applications affect teaching and learning in higher education, as well as how to improve the learning environment. Websites such as Course Era and Udemy were increasingly crucial to students as they got more interested in online courses. A higher dropout rate could result from the difficulty of supporting each and every open learning student in such large class sizes. Students' academic progress was predicted using a machine learning approach called linear regression. A machine learning algorithm and the amount of data it consumes are critical to its success in the field of educational technology. An algorithm for predicting student performance must be chosen carefully. Correctness is determined using a machine learning algorithm. An method known as Linear Regression was used in this investigation to support the claim. Research shows that a student's ability to succeed in school is affected by a variety of factors, including his or her family background. According to a number of studies, a student's history and other traits play a significant role on their academic achievement. Machine learning has been increasingly important in recent years across many industries, and it may also be applied efficiently in academia.

3. Proposed Methodology

The proposed methodology has three main steps which are collecting the dataset, filling the missing values using KNN, clustering by fuzzy c-means. First, the students entrance exam scores are acquired from the Vietnam dataset. The dataset obtained contains missing values so it is said to be incomplete.so as to make the dataset complete the method called KNN is used. As a result, the missing values in the dataset are filled. Finally, the clusters are generated using the FCM approach. Figure 1 shows architecture of proposed methodology.

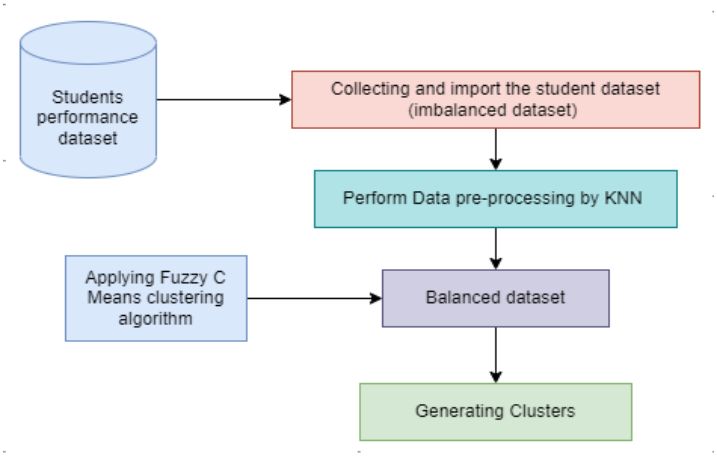


Figure 1. Architecture diagram

A. Dataset

The dataset which has been used is the college entrance exam scores in the Vietnam dataset. This is a dataset about test scores that candidates achieved in the college entrance examination for the school year 2021 in Vietnam. The college entrance exam scores in Vietnam data contain 10 fields namely, id, foreign language, literature, mathematics, biology, chemistry, physics, civic education, geography, and history.13701 missing values are present in the dataset. Data is acquired from a variety of sources in today's world and used for a variety of purposes, including analysis, insight generation, theory validation, and so on. Some information may be missing from the information gathered from various sources. This could be the result of a human error during the data collecting or extraction procedure. As a result, Preparation of the data is complicated if you don't take into account how to deal with missing numbers. To have a positive impact on one's career, the method of imputation must be chosen carefully. Data preprocessing is essential before its actual use.

A dataset with missing values is a hornet's nest for any data scientist. With variables that have no values, it can be difficult to deal with them. To put it another way, there is no easy answer. The lack of values in real-world datasets is a concern. It's possible that a variable's observations will be missing values due to a variety of factors, such as a problem with the equipment or machinery, a lack of participants, a mistake in the researcher's data entry, a participant's amnesia, an accounting error, etc. The KNN algorithm was employed to resolve this issue.

B. KNN Algorithm

The missing values are filled in using this algorithm. Impute missing data values using k-Nearest Neighbors (kNN). It is possible to estimate missing values by comparing the

completed values of nearby observations using the method k-Nearest Neighbors (k - nearest neighbours). Next-neighbour algorithms are designed to find the 'k' nearest-neighbor samples in a dataset. Data points that are missing can be estimated using the 'k' samples that have been collected. The dataset's mean value of the 'k's-neighbors is used to impute the missing values of each sample. The steps for filling in the missing values are outlined below.

Step 1: Neighbors with the number K are the ones you should choose.

Step 2: Neighbors K in the Euclidean distance should be calculated as shown in figure 2.

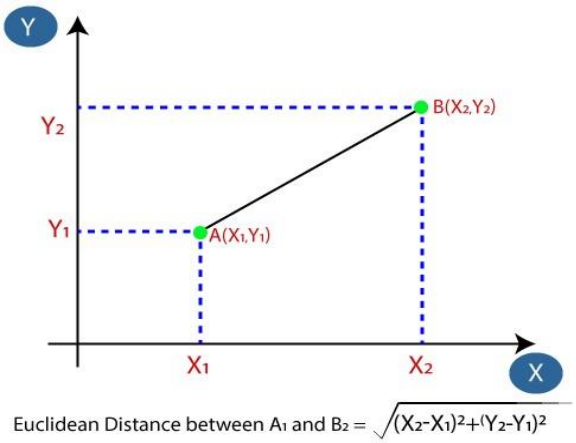


Figure 2: Finding distance between points

Step 3: Choose the K closest neighbors based on Euclidean distance calculations.

Step 4: Data points in each category are counted among the k-next nearest neighbors.

Step 5: In order to properly classify the new data, it should be placed in the category with the mostclose neighbors.

Euclidean distance can be used to identify the closest neighbors. After learning to use feature points, machine learning classifies sample data. Data generalization is necessary for improved performance. After missing values are filled the next step is to perform an unsupervised machine learning technique which is fuzzy c-means on the complete dataset. By applying this technique the students are classified into clusters according to their scores in the entrance exam. Based on the clusters formed the grasping levels of the students can be analyzed. This methodology is mainly beneficial for educational institutions as the institutions can understand the student’s grasping levels and act accordingly.

C. Fuzzy C-Means Algorithm

The attributes of the dataset are then clustered after the missing values have been filled in. Using fuzzy logic, the data clustering method Fuzzy C-Means An unsupervised clustering algorithm can be used to create a fuzzy partition from data. By assigning each point a percentage of membership in each cluster centre ranging from zero to 100 percent, fuzzy logic can be used to cluster multidimensional data. There are many advantages to this method over traditional clustering, which requires that every cluster be identified by an individual label for each point. This algorithm assigns each cluster center's data point a membership based on the distance between the cluster centre and the data point. The closer a piece of data is to the cluster centre, the more likely it is to be part of a cluster. Obviously, all of the data points

should have the same number of members. Following is the algorithm for calculating fuzzy c-means scores:

Let X be the collection of data points (x1, x2, x3..., xn) and V be the set of centres (v1, v2, v3..., vc) (v1, v2, v3..., vc).

Step 1: Randomly assign the locations of the cluster centers 'c'.

Step 2: The fuzzy membership 'μij' can be calculated using:

$$\mu_{ij} = 1/\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\left(\frac{2}{m}\right)}}$$

Step 3: Calculate the 'vj' fuzzy centers using:

$$v_{ij} = \sum_{i=1}^{n} {\mu_{ij}}^{m} x_i / \sum_{i=1}^{n} {\mu_{ij}}^{m}$$

Step 4: Repeat steps 2 and 3 until the 'J' value achieves the minimum necessary for the experiment or ||U(k+1) - U(k)|| < β.

Where ‘k’ represents the iteration step.

‘β’ is the condition for ending the range [0, 1].

‘U = (μij)n*c’ represents the fuzzy membership matrix.

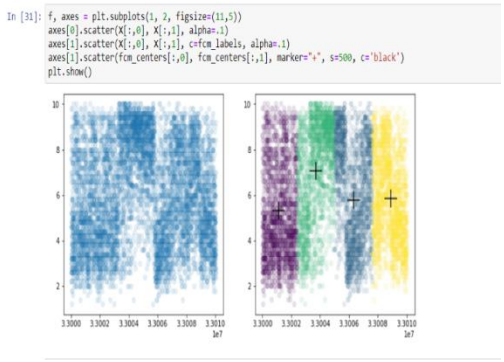
‘J’ stands for the objective function.

4. Results & Discussion

For clustering students based on their performance in entrance examination here imported an entrance exam scores dataset. The dataset obtained consists of 4500 records. It consists of 10 fields namely, id, foreign language, literature, mathematics, biology, chemistry, physics, civic education, geography, history.13701 missing values are present in the dataset as shown in figure 3&4. Missing values in datasets can pose issues for many machine learning techniques. As a consequence, it's a good idea to discover and replace missing values in each column of the input data before modelling the prediction job. Missing data imputation, or simply imputing, is the name for this. A prominent technique for missing data imputation is the use of a model to anticipate missing values. This includes constructing a model for each input variable that contains missing values. Although any model may be used to estimate missing values, the k-nearest neighbour (KNN) approach, generally known as "nearest neighbour imputation," has been found to be usually successful.

To fill these missing values and make the dataset balanced here used KNN Imputer for fitting KNN.

For the balanced dataset obtained here apply Fuzzy C means technique. But, before applying FCM, have to pick the clusters. Figure 3 &4 represents the clusters in visualization and texual form respectively



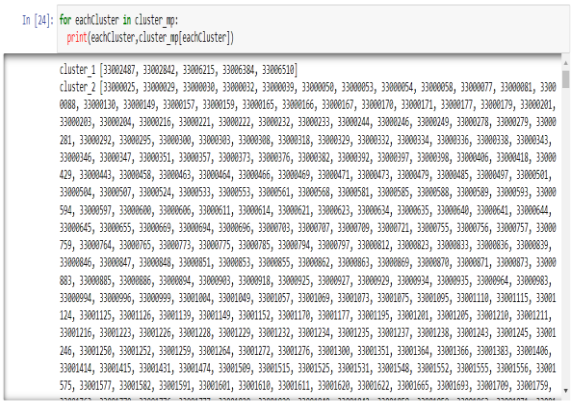


Figure 3: Formation of Clusters Figure 4: The students in different clusters

For measuring the performance of clustering algorithms, SSE values have to be used in lieu of accuracy. In prior studies, the K-Means clustering method is employed however in our project, we conducted fuzzy-c means clustering with an SSE value of 2, while k-means has an SSE value of 3.5 suggesting poor performance as shown in figure 14. This project increased the performance with an SSE of 2. Here we dropped the SSE value by roughly 1.5 and boosted the performance of the project. For these clustering projects if the SSE value is low then the performance is high and vice versa. Whereas, for accuracy, it is the reverse as the value grows its performance improves, and vice versa. Figure 5 shows SSE values for K-means & Fuzzy C Means

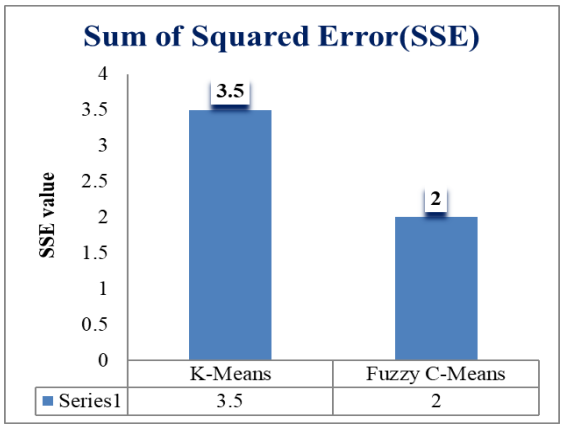


Figure 5: SSE values for K-Means and Fuzzy C-Means

5. Conclusion

Clustering is the most commonly deployed approach for predicting the future. Clustering's principal objective is to classify pupils into homogenous groups based on their traits and skills. These apps may aid both professors and students in increasing the quality of their education. To carry out this project, apply one of the clustering algorithms, specifically the Fuzzy-c-means clustering algorithm. This approach splits the data into clusters. To benchmark, the FCM, K-means clustering is employed. Fuzzy c-means clustering surpasses the k-Means approach. The fuzzy c-means approach, unlike

the k-Means algorithm, lets data points to belong to several clusters with a probability. Fuzzy c-means clustering gives improved results for overlapping data sets. This strategy has the ability to increase educational quality by taking the correct measures at the right time to boost student performance and minimize the chance of failure.

In this initiative, only students mark the entrance exams. In the future, it wants to enhance this approach to provide more useful and accurate outputs by considering the family history, mental states of the students when writing the test which would be beneficial for the instructors to improve student learning outcomes. This may also be constructed and turned as an app or online application.

References

[1] *J. Jamesmanoharan; S. Hari Ganesh; M. LovelinPonnFelciah; A. K. Shaheen Banu, "Discovering Students' Academic Performance Based on GPA using K-Means Clustering Algorithm", 2014 World Congress on Computing and Communication Technologies.*

[2] *YuniYamasari; Anita Qoiriah; Hapsari P. A. Tjahyaningtijas; Ricky E. Putra; AgusPrihanto; Asmunin "Improving the Quality of the Clustering Process on Students' Performance using Feature Selection", IEEE access, 2020*

[3] *Alana M. de Morais; Joseana M. F. R. Araújo; Evandro B. Costa "Monitoring Student Performance Using Data Clustering and Predictive Modelling", IEEE access, 2015.*

[4] *Luz Stella Robles Pedrozo; Miguel Rodríguez-Artacho "A cluster-based analysis to diagnose students' learning achievements", IEEE access, 2013.*

[5] *Sonali Shankar; Bishal Dey Sarkar; Sai Sabitha; Deepti Mehrotra "Performance Analysis of Student Learning Metric using K-Mean Clustering Approach", IEEE access, 2016.*

[6] *Ravneil Nand; Ashneel Chand; Mohammed Naseem "Analyzing students online presence in undergraduate courses using Clustering", IEEE access, 2021.*

[7] *Nova Riyanti; Surya Sumpeno; MauridhiHery Purnomo "Attribute Selection Techniques to Clustering the Entrepreneurial Potential of Student based on Academic Behavior", IEEE access, 2020.*

[8] *V. Uday Kumar; Azmira Krishna; P. Neelakanteswara; CMAK Zeelan Basha "Advanced Prediction of Performance of A Student in An University using Machine Learning Techniques", IEEE access, 2020.*

[9] *Ishwank Singh; A Sai Sabitha; Abhay Bansal "Student performance analysis using clustering algorithm", IEEE access, 2016.*

[10] *Shiwani Rana; Roopali Garg "Application of Hierarchical Clustering Algorithm to Evaluate Students Performance of an Institute", IEEE access, 2016.*

[11] *Sujith Jayaprakash; Sangeetha Krishnan; V. Jaiganesh "Predicting Students Academic Performance using an Improved Random Forest Classifier", IEEE access, 2020.*

[12] *D. Solomon, S. Patil, and P. Agrawal, ''Predicting performance and potential difficulties of university students using classification: Survey paper,'' Int. J. Pure Appl. Math, vol. 118, no. 18, pp. 2703–2707, 2018.*

[13] *Y. Zhang, Y. Yun, H. Dai, J. Cui, and X. Shang, ''Graphs regularized robust matrix factorization and its application on student grade prediction,'' Appl. Sci., vol. 10, p. 1755, Jan. 2020.*

[14] *L. M. Abu Zohair, ''Prediction of student's performance by modelling small dataset size,'' Int. J. Educ. Technol. Higher Educ., vol. 16, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s41239-019-0160-3.*

[15] *F. Ahmad, N. H. Ismail, and A. A. Aziz, ''The prediction of students' academic performance using classification data mining techniques,'' Appl. Math. Sci., vol. 9, pp. 6415–6426, Apr. 2015.*

[16] *B. Predić, G. Dimić, D. Rančić, P. Štrbac, N. Maček, and P. Spalević, ''Improving final grade prediction accuracy in blended learning environment using voting ensembles,'' Comput. Appl. Eng. Educ., vol. 26, no. 6, pp. 2294–2306, Nov. 2018.*

[17] *T. Alam, C. F. Ahmed, S. A. Zahin, M. A. H. Khan, and M. T. Islam, ''An effective recursive technique for multi-class classification and regression for imbalanced data,'' IEEE Access, vol. 7, pp. 127615–127630, 2019.*

[18] *Chew Li Sa; DayangHananibt.Abang Ibrahim; Emmy Dahliana Hossain; Mohammad bin Hossin "Student Performance Analysis System (SPAS)", IEEE access 2015.*

[19] *Khalfan Al Mayahi; Mahmood Al-Bahri "Machine Learning Based Predicting Student Academic Success", IEEE access 2020.*

[20] *Suhas S Athani; Sharath A Kodli; Mayur N Banavasi; P.G. Sunitha Hiremath "Student Performance Predictor using Multiclass Support Vector Classification Algorithm", IEEE access 2018.*

[21] *Hina Gull; Madeeha Saqib; Sardar Zafar Iqbal; Saqib Saeed "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning", IEEE access 2021.*

[22] *Fatema Alnassar; Tim Blackwell; ElahehHomayounvala; Matthew Yee-king "How Well a Student Performed? A Machine Learning Approach to Classify Students' Performance on Virtual Learning Environment", IEEE access, 2021.*

[23] *Jigna B Prajapati; Savan K Patel "Performance Comparison of Machine Learning Algorithms for Prediction of Students' Social Engagement", IEEE access, 2021.*

[24] *Pedro Manuel Moreno-Marcos; Ting-Chuen Pong; Pedro J. Muñoz-Merino; Carlos Delgado Kloos "Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics", IEEE access, 2020.*

[25] *Muhammad Adnan; Asad Habib; Jawad Ashraf; ShafaqMussadiq; Arsalan Ali Raza; Muhammad Abid; Maryam Bashir; Sana Ullah Khan "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models", IEEE access, 2021.*

[26] *BoddetiSravani; Myneni Madhu Bala "Prediction of Student Performance Using Linear Regression", IEEE access, 2020.*

[27] *Filipe Dwan Pereira; Elaine H. T. Oliveira; David Fernandes; Alexandra Cristea "Early performance prediction for CS1 course students using a combination of machine learning and an evolutionary algorithm", IEEE access, 2019.*

[28] *MeizarRaka Ramadhan; Sri Suning Kusumawardani; Paulus InsapSantosa; Maximilian Sheldy Ferdinand Erwianda "Predicting Student Academic Performance using Machine Learning and Time Management Skill Data", IEEE access, 2020.*

[29] *C. Jalota and R. Agrawal, Feature Selection Algorithms and Student Academic Performance: A Study, vol. 1165. Singapore: Springer, 2021.*

[30] *X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, ''Grade prediction of student academic performance with multiple classification models,'' in Proc. 14th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD), Jul. 2018, pp. 1086–1090.*

[31] *Hogo, Mofreh A. "Evaluation of e-learning systems based on fuzzy clustering models and statistical tools.". Expert systems with applications. Vol 37 (10), pp. 6891- 6903, 2010.*

[32] *Liang Zhao; Kun Chen; Jie Song; Xiaoliang Zhu; Jianwen Sun; Brian Caulfield; Brian Mac Namee "Academic Performance Prediction Based on Multisource, Multi Feature Behavioral Data", IEEE access, 2020.*

[33] *Haviluddin; NatanielDengen; Edy Budiman; MasnaWati; UmmulHairah "Student Academic evaluation using Naive Bayes Classifier algorithm", IEEE access, 2019.*

[34] *Buenaño-Fernández, D. Gil, and S. Luján-Mora, ''Application of machine learning in predicting performance for computer engineering students: A case study,'' Sustain., vol. 11, no. 10, pp. 1–18, 2019.*

[35] *M. A. Al-Barrak and M. Al-Razgan, ''Predicting students final GPA using decision trees: A case study,'' Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533, 2016.*