International Journal on Applications in Information and Communication Engineering
Volume 7 : Issue 2 : May 2021 , pp 31 –37 www.aetsjournal.com ISSN (Online) : 2394 - 6237

Using a robust clustering method to analyze student performance

Suganya Mahalakshmi A, Kanagaprabha S, Priyanka, Priyadharashini A
Department of Computer Science and Engineering,
Rathinam Technical Campus,
Coimbatore, Tamilnadu,India

Abstract- Using techniques such as statistical analysis, machine learning, artificial intelligence, and database management systems, data mining is the process of extracting useful information from vast datasets. Using data mining techniques, educational institutions with a higher level of education seek to discover a student's academic achievement. Various data mining tasks can be utilized to evaluate pupils' overall performance. The objective of classification is to generate data that may be used to measure student achievement. In addition, other strategies are employed to support the classification of data by the decision tree method. Using decision trees allows very exact student performance forecasts. Prior research developed an SVM Prediction method for predicting a student's Grade Point Average (GPA) upon completion of first-, second-, and third-year computer education and instructional technology courses. The SVM prediction data mining technique consists of three independent processes: data preparation, formulation of the prediction model, and SVM linear regression evaluation. The SVM prediction model is able to perform classification and regression tasks by applying linear mathematics and integrating based on linear data. However, the SVM algorithm does not provide any suggestions for enhancing the student's grade point average. There is no room for subjectivity in the grading procedure, therefore this method cannot determine the exact recall levels. In the proposed work, the Spectral Cluster-based Classification Tree Machine Learning Technique for Analyzing Academic Achievement in Higher Education Institutions is provided as a remedy for these problems. In an effort to promote genuine positive values, the purpose of this research is to develop a Cluster-based Decision Tree technique that can accurately predict student GPAs. We need your help to develop solutions based on decision trees that are suggestive and aid students with low GPAs in raising their grades. Several metrics, including Prediction Accuracy, True Positive, False Positive, and Number of Decision Rules, are utilized to assess the efficacy of the proposed technique. This is also true when applying cluster analysis approaches, where these obstacles may result in inferior clustering results. This occurrence is common. Robust Clustering Methods were designed to circumvent these undesirable impacts.

Key words -Student Performance, Data Mining, and Grade Point Average True Positive instead of False Positive

I. INTRODUCTION

Data mining is a technique for obtaining useful information from vast and intricate datasets. Data analysis and identification can be facilitated by the application of data mining techniques. Classification and prediction are two different approaches to data analysis that are used to examine and develop specific class models. Numerous classification strategies, such as the Decision tree approach, the Bayesian network, the neural network, and the Genetic algorithm, can be used to develop a classification model. Based on past trends, this form of categorization can help forecast future preferences. Data mining, also known as database knowledge discovery, is the process of extracting knowledge from massive amounts of data. Data mining is used in education to enhance students' comprehension of the learning process by analyzing, extracting, and estimating learning process-related attributes.

The vast amount of data already saved in academic databases contains information that can be used to generate precise estimates of students' academic progress. Institutions of higher education exist primarily to educate students with a superior education so they can make more responsible decisions. By judging student performance based on the discovery of new knowledge, the higher education system achieves the highest attainable standard of excellence. Using data provided by the student management system, a student's performance was determined at the end of the semester. This data includes the number of current students and their analytical, course, and assignment grades.

II. LITERATURE SURVEY

The author of this work [1] suggests utilizing Latent Semantic Analysis to reveal the underlying structure of the information. This is achieved by classifying data objects according to their degree of similarity. This sort of learning is also known as unsupervised learning, because class labels do not require an a priori classification of the objects being learned. Before beginning the learning process, data objects are paired with known classes in supervised learning. This article [2] explores the potential application of cluster analysis to the investigation of data structures. Because density structures are designed to implement class breakdown, they can substantially improve the performance of decision tree classifiers. It is possible to decompose Classes and estimate cluster mergers using decision tree classifiers by employing cluster analysis. The C4.5 and CART classifiers are then utilized to generate a statistic for class breakdown. This paper[3] covers the difficulties of detecting speakers in densely populated areas with a high background noise level. The bulk of speaker recognition techniques are founded on Mel-Frequency-Cepstral

Coefficients (MFCC), the Gaussian Mixture Model (GMM), and the Universal Background Model (UBM) (UBM). It is common knowledge that these approaches excel at recognizing small populations in environments with little background noise.

This lesson's [4] purpose is to promote comprehension of these issues and their repercussions. This talk focuses on the fundamental properties of the various graphs created by Laplacians. Spectrum clustering algorithms must be derived from scratch using a number of techniques before they can be made public. This paper's author[5] provides a simple Matlab-based method for spectrum grouping. The technique is constructed using devices based on matrix perturbation theory, and it operates superbly. The author gives experimental findings for a variety of challenging clustering scenarios. The author discovers innovative spectral clustering cost functions in the mentioned paper [6]. Using the difference between a certain partition and a solution to the spectrum normalised cut problem, a computation is conducted. The purpose of these computations is to determine it. Utilizing spectral clustering algorithms will reduce partition cost functions. During one of the spectrogram segmentation steps, the learning technique is used to address both the blind one-microphone speech separation problem and the resulting cost issue. This study [7]explains one ID3 system and provides a method for synthesizing the decision trees employed by several additional systems. A basic algorithm is broken down into its component elements in this study. This essay [8] can be found at: (CRESPAR). Funding from the Institute of Education Sciences has enabled the formation of a national research and development center (IES). Also included is the expression of substance or opinions. In this study [9], the classification problem is utilized to evaluate student performance, and the decision tree approach is one of the classification methodologies employed. The knowledge of students is evaluated based on their performance in the semester's final examination. It is useful for identifying students who have dropped out of school as well as those who require specialized instruction and extra attention. This article[10] introduces a novel cost function for spectral clustering that evaluates the error between a group of divisions. A novel method for spectral clustering is created because to the decreased cost functions that come from partitioning. The author develops a tractable approximation and applies it to the cost function created by the power method of eigenvector computation.

The author provides an introduction to multiclass spectral clustering in this paper [11]. The formulation of the problem of continuous optimization is based on discrete clustering and Eigen decomposition. Using ortho-normal transforms, the author analyzes the origin of all optimal solutions as a function of eigenvectors. The author of this study [12] mixes temporal smoothness with evolutionary spectrum clustering in two frameworks. Both of these architectural blueprints are available on this website. The author builds both of these frameworks by applying lessons learnt while resolving the well-known k-means clustering problem. Then, similar cost functions are constructed and applied to evolutionary spectrum clustering problems. This paper's author[13] made heavy use of the spectral clustering algorithm, a data clustering technique that organizes data based on the eigenvectors of a similarity/affinity matrix. How to automatically determine the number of clusters and cluster noisy and sparse data successfully. After analyzing Eigen space properties, it is determined that (a) each and every eigenvector of a data similarity matrix is not informative and relevant for clustering; (b) eigenvector selection is different as a result of the use of uninformative and irrelevant eigenvectors, resulting in suboptimal clustering results; and (c) equivalent eigenvalues are not utilized for relevant eigenvector selection given a realistic data set. Using spectral techniques, the author of this paper [14] develops an effective clustering algorithm for large-scale graph data. Repeatedly, a limited number of "supernodes" are linked to normal network nodes. To do this, it is necessary to turn the initial graph into a sparse bipartite graph. Utilizing spectrum approaches to cluster the bipartite network enhances clustering efficacy, but substantially decreases clustering efficiency. This article [15] demonstrates how maximization of the Q function may be framed as a spectrum relaxation problem and presents two novel Q-maximizing spectral clustering techniques. These two contributions have been included in the publication. Experiments suggest that the newly created algorithms are effective and efficient at identifying acceptable clusterings and the optimal number of clusters across numerous real-world network data sets.

III. RIGID COMPOSITION

The robustclustering algorithm was designed to receive a number of diverse cluster sets as input and then build a set of robust clusters from these sets. A resilient cluster can only consist of objects that are grouped together in each and every input cluster formed by the different clustering techniques. Formation of an initial n x n upper triangular agreement matrix. In each cell of this matrix, represented by row and column indices, the number of category agreements between the two variables is indicated (see Figure 1). This matrix is then employed to cluster variables based on their existing clusteragreement (as found in the matrix).

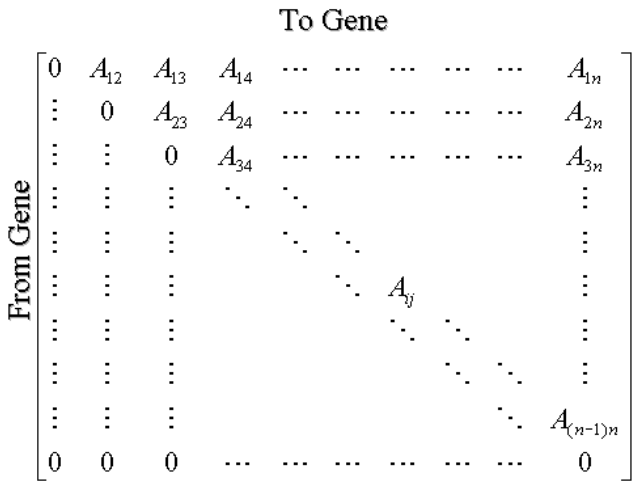

Figure-1: Confusion Matrix

The agreement matrix serves as an input to the algorithm, which produces a list called List. This list contains all pairs for which the corresponding cell in the agreement matrix contains a value equal to the total number of techniques being combined (i.e. full agreement). Beginning with a collection of empty robust clusters represented by RC, where RCi represents the ith robust cluster, the first cluster is constructed. The contents of the first pair of elements in List are then stored in this first cluster. Then, the pairs in List are examined individually to determine if any of the persons comprising the current pair are members of any of the groups contained in RC. If one of the members of the current pair is located and the other is not, the missing element will be added to the cluster that contains the found element. If neither component of the pair is present in RC, a new cluster containing both components of the pair is added to RC. The collection of robust clusters, often known as RC, is obtained once the list has been exhausted.The results of applying robust clustering are summarized in Table7 and the algorithm follows.

Input: Agreement Matrix (n×n), A; The Number of Clustering Methods
1) Set List = all pairs (x, y) in A, with agreement = number of methods
2) Set RC to be an empty list of clusters
3) Create a cluster and insert the first pair (x, y) in List into it
4) For i = 2 to size of List
5) For j = 1 to number of Clusters in RC
6) If x or y of $List_i$ is found within RC_j
7) If the other member of the pair $List_i$is not found in RC_j
8) Add the other member to RC_j
9) End If
10) Else ifthe other member of the pair $List_i$is not found in RC_j
11) Add a new cluster to RC containing x and y
12) End If
13) End For
14) End For

Output: Set of Robust Clusters, RC

The proposed technique is divided into three phases are:

a) Higher Educational Student Data Storage
b) Spectral Cluster for Student Data Organizing
c) Decision Tree on Student Activities and Prediction

a) Data Storage for Students Who Are Attending Postsecondary Institutions

b)One approach to measure the institution's overall degree of development is by looking at the percentage of an institution of higher education's alumni who go on to achieve success in their respective fields after graduating from that institution. Research is carried out on the scholastic achievements of students who are currently enrolled in institutes of higher education. The purpose of this study is to develop hypotheses and projections regarding the academic performance of pupils in the years to come. This is a necessary condition that must be met in order to raise the standard of education overall. The processes of data mining combine the most important function in data analysis with the actions of institutions, which collect data to be disseminated from a wide variety of data sources stored at a wide range of student activities. A more precise analysis of the data is produced as a consequence of this.

c)Even if information pertaining to students is stored in a wide variety of various knowledge sources, it is still possible to mine information from historical as well as modern data sets. The databases are updated with the scores of the students who were successful in completing the graduation exam. These students were successful in their attempts to graduate. The score that an individual student obtained on an examination is calculated by applying the conventional scoring method to the subject matter that was being tested. The majority of the time, information about students is stored in educational databases. These databases contain a wide variety of information provided from a number of different points of view, and they are where most of the information about students is held. Data mining is a technique that examines the aspects that influence student performance at higher education institutions and identifies the correct views of data for specific student activities. This method can also be described as "student performance analysis." Students from these types of higher education institutions are recruited for this method.

d) The database of an educational facility will store many data points in order to create accurate estimates regarding the number of students who will enroll in a particular class and the number of students who will be unable to successfully

complete it. This is done so that any odd values that might appear on the student grade sheets can be found and identified. e)Spectral Cluster for Student Data Organizing. The student data are grouped using spectral clustering in order to test the efficacy of an analysis that successfully accounts for the shifting parameters of student activity in the institution. This evaluation is done in order to determine whether or not the analysis is effective. This evaluation is carried out so that a determination can be made regarding the efficacy of the analysis that is in question. The spectral cluster approach is one that is substantially more effective when it comes to categorizing students' grade point averages according to a number of indications that span the spectrum. This is because the spectral cluster method takes into account the entire spectrum. The procedure of producing a spectral cluster begins with the application of a standard threshold to the same spectrum range as the previous phase. After this stage, the students' test scores that fall within the same range as the previous one are grouped into a variety of distinct spectra values. This step is the final step in the process. When there is a change in the level of student engagement, the instantaneous threshold is used to calculate the varying spectral similarity range value for a number of different measures. In order to accurately calculate both the true positive and the false positive values of the examination score, it is necessary to first regenerate the spectral clusters based on the instantaneous threshold and then use a wide range of distinct spectral cluster objects.

.

b) A Determination Tree Concerning Student Activities and Their Estimates

The purpose of the decision tree model is to identify the elements that contribute to current performance, as well as to estimate future success on upcoming examinations by examining data regarding previous performance. In order to generate it, a spectral cluster object that has a threshold for its instantaneous range is used. The decision rules provide guidance on the parameters that should be controlled in order to enhance the overall performance of the students as a group.

The fact that Spectral Clusters and run gobjects are rather similar to one another can be used as a point of reference when developing the essential components of decision rules. A decision tree model is used in the construction of the tree in order to determine the numerous factors that contribute to the overall performance of the pupils. This is done in order to construct the tree accurately. Significant criteria that were derived from the spectral cluster object are taken into consideration throughout the process of sorting pupils into groups according to the traits that they bring with them (grades).

In addition, the efficacy of the decision tree algorithm is validated by applying the cross validation and percentage split approaches in the process. The administration of the higher education institution makes use of the decision tree and the rules that it creates in order to take concrete steps toward improving the institution's overall success rate for its student population as a whole.

IV. PERFORMANCE EVALUATION

In this section of the paper, we will analyze the efficacy of the Decision Tree Data Mining Technique for analyzing student performance in educational institutions of higher learning. These institutions include universities and colleges. Because of this work, we now have the capability to reliably predict how well children will do in higher education, which is one of the most significant things that has been accomplished as a result of this work. A decision tree that is constructed on the basis of the spectral cluster object similarity range threshold has been developed for the purpose of recommending a strategy for enhancing one's grade point average after graduation. Before attempting to determine the true positive and false positive values, it is required to first evaluate the accuracy and recall of the cluster item. Some examples of performance measures that can be applied to the parameters include the Student Strength metric, the prediction Accuracy metric, the True Positive rate metric, the False Positive rate metric, the number of decision rule metric, and the Prediction Rate metric.

Performance can be evaluated based on a variety of factors, including the number of decision rules, the accuracy of predictions, the true positive rate, the false positive rate, and the prediction rate.

4.1 percent of questions with the correct response

When information is accurately classified as belonging to a certain class, this is what is indicated by the term "true positive rate," which refers to the rate at which the classification is accurate. In most instances, it is presented in the form of a percentage. You may find the mathematical formula that is used to determine the real positive rate further down in this article.

$$\mathbf{TPR} = \left(\frac{\textbf{informations correctly identified as belonging to a class}}{\mathbf{C}}\right)$$

Table: 4.1. No. of decision rulesVs True Positive rate (%)

No. of decision rules	True Positive rate(%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	13	15
20	16	18
30	19	23
40	22	27
50	25	31

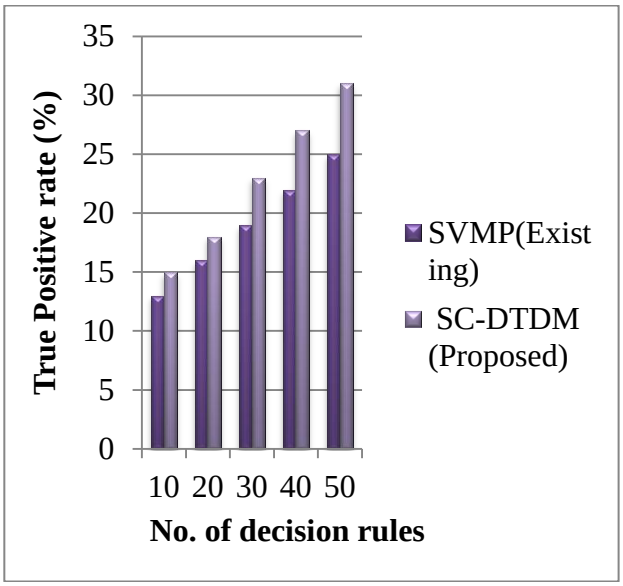

Figure: 4.1. No. of decision rulesVs True Positive rate (%)

Figure: 4.1Display the proportion of cases determined to be true positives.

The calculation of the True Positive rate makes use of both the SVM-based Prediction method (SVMP) Technique and the Spectral Cluster-based Decision Tree Data Mining (SC-DTDM) Technique that we suggested. SVMP stands for SVM-based Prediction method. SC-DTDM stands for Spectral Cluster-based Decision Tree Data Mining. Along the X-axis is a representation of the number of decision rules, and along the Y-axis is a representation of the degree to which forecasts are correct. The increase in the overall number of decision rules that were put into place was exactly proportionate to the rise in the percentage of "true positives." The rate of True Positives is exhibited by utilizing both the currently implemented SVMP as well as the suggested SC-DTDM methodology. Both of these methods are used in conjunction with one another. The significance of this subject was taken into consideration at some point. Figure 4.1 illustrates that the recently established SC-DTDM strategy is superior in terms of decision rules to both the previously utilized SVMP method and the proposed SC-DTDM approach. This is demonstrated by the fact that the newly developed SC-DTDM strategy was developed. The Spectral Cluster based Decision Tree Data Mining Technique has the potential to achieve a high performance variation of True Positive rate that is 15 percent greater than that of the system that is being used at the moment. This is because the technique is based on decision trees.

4.2 The percentage of tests that produce false positives

The false positive rate is also known as the false alarm rate and is defined as the ratio of the probability of improperly analyzed student performances to the total number of performances made. This ratio is also known simply as the false positive rate. This measurement is presented in the form of a percentage, and the mathematical derivation of this metric is as follows:

$$\mathbf{FPR} = \left(\frac{\textbf{Probability of wrongly analyzed as student performaces}}{\textbf{No.of performances made}}\right) * \mathbf{100}$$

Table: 4.2. No. of decision rulesVs False positive rate (%)

No. of decision rules	False positive rate (%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	62	56
20	66	59
30	70	63
40	73	67
50	77	69

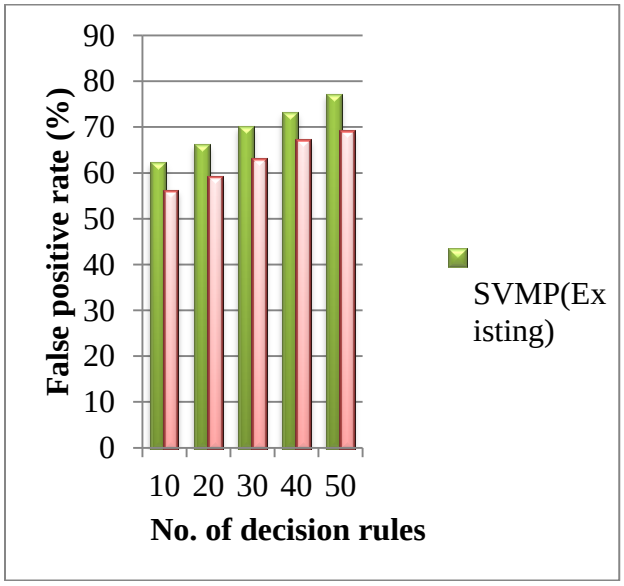

Figure: 4.2. No. of decision rulesVs False positive rate (%)

Figure: 4.2. Please display the false positive percentage.

When utilizing the SVM-based Prediction system (SVMP) Technique and the Spectral Cluster-based Decision Tree Data Mining (SC-DTDM) Methodology that we have supplied, the axis that is labeled "X" represents the number of decision rules, and the axis that is labeled "Y" indicates the number of false positives. Both of these techniques are referred to as the "Spectral Cluster-based Decision Tree Data Mining" (SC-DTDM) Technique. The acronyms "SVMP" and "SC-DTDM" are used to refer to these two different approaches,

respectively. Both of these strategies are shown in equal measure in the graph that is presented below. It is possible that the reduction in the number of false positives that have happened is attributable to the increase in the number of decision rules that have been implemented. Both the SVMP methodology, which is now being implemented, and the SC-DTDM method, which is projected to be utilized in the not-too-distant future, provide information concerning the rate of false positives in their respective methodologies. In terms of the decision rules, the recently developed SC-DTDM technique is superior to both the SVMP method, which had been utilized in the past, and even the proposed SC-DTDM approach. This is because the newly constructed SC-DTDM technique is superior to the SVMP method. Figure 4.2 illustrates this concept. The Spectral Cluster based Decision Tree Data Mining Technique has the potential to achieve a false positive rate that is ten percentage points lower than the method that is being used at the moment. When compared to the strategy that is being applied at the moment, this is an improvement.

The Degree of Accuracy That Could Be Obtained From the Predictions

One of the primary objectives of the data mining industry in the field of higher education is to determine the extent to which it is possible to accurately predict the academic success of students by focusing solely on the character traits that have the greatest impact on that achievement. This is one of the most important goals of the industry. The level of accuracy that may be achieved in the estimation of student score ranges by the utilization of a data collection that takes into consideration all of the students' academic, personal, and monetary qualities.

Table: 4.3. No. of decision rulesVs prediction accuracy (%)

No. of decision rules	Prediction accuracy (%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	36	48
20	45	56
30	54	67
40	68	79
50	71	84

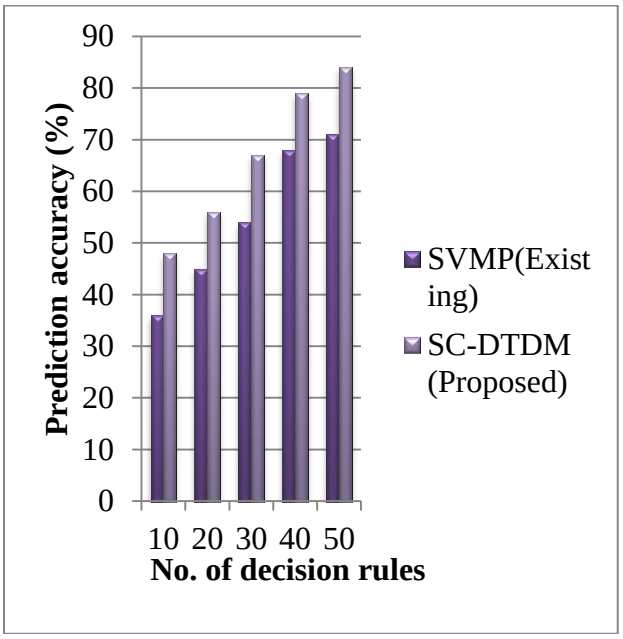

Figure: 4.3. No. of decision rulesVs prediction accuracy (%)

Figure 4.3: Please provide a concrete example that demonstrates how accurate the forecast was. This graph represents both the SVM-based Prediction technique (SVMP) Technique as well as our own proposed Spectral Cluster-based Decision Tree Information Retrieval (SC-DTDM) Technique, with the X axis trying to represent the number of majority voting and the Y axis representing the accuracy of each method's predictions. Both techniques are referred to as SVM-based Prediction techniques (SVMP). Both of these approaches are known collectively as SVM-based Prediction methods (SVMP). There was a correlation between the number of decision rules that were applied and the accuracy of the predictions that were made. This correlation was one-to-one. Both the SVMP, which is currently being put into use, and the SC-DTDM Technique, which is currently being suggested, are utilized in order to demonstrate that the prediction is accurate. Figure 4.3 shows that the suggested SC-DTDM technique is superior to both the existing SVMP and the recommended SC-DTDM method in terms of decision rules. This is illustrated by the fact that the suggested SC-DTDM methodology beats both of these methods. The method that is now being used has the potential to create an increase in prediction accuracy that is 16 percentage points lower than the potential increase that may be produced by the Spectral Cluster based Decision Tree Data Mining Technique.

The accuracy of the forecasting.

The proposed task performance can be evaluated using a prediction rate that is based on the grade point average of each individual student. By making use of the prediction rate, it is feasible to obtain an accurate prediction of the ranges of the students' similarity score. It is judged using standards established by the following: (percent).

Table: 4.4. No. of decision rulesVs Prediction rate (%)

No. of decision rules	Prediction rate (%)	
	SVMP(Existing)	SC-DTDM(Proposed)
10	40	45
20	45	49
30	49	53
40	53	58
50	59	65

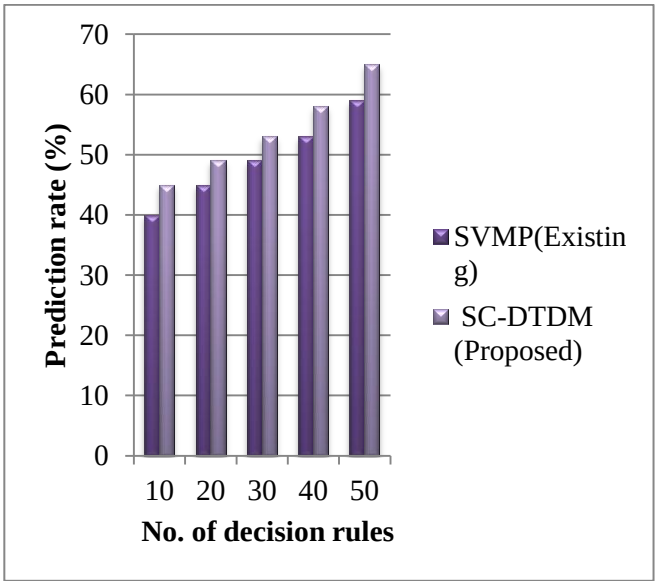

Figure: 4.4. No. of decision rulesVs Prediction rate (%)

Figure: 4.4 It would be good to display the prediction's level of accuracy in order to better understand it. The SVM-based Prediction technique (SVMP) Methodology as well as the technique which we've developed, the Spatial frequency Cluster-based Decision Tree Data Mining (SC-DTDM) Technique, both have the X axis represent the number of decision rules, and the Y axis represent the prediction rate. The SVM-based Prediction technique (SVMP) Approach is also known as the Spectral Cluster-based Decision Tree Data Mining (SC-DTDM) Technique. This is due to the fact that the SC-DTDM Technique is based on a decision tree, but the SVMP Approach being based on a support vector machine (SVM). Both of these axes have already been drawn on a single graph that has been created. The number of prediction models that were put into place was found to have a positive link with the rise in the proportion of correct forecasts that they produced. In order to demonstrate how precise the prediction may be, both the SVMP method, that has already been deployed, and the SC-DTDM approach, which would be currently being investigated, are applied. In terms of decision rules, it is clear from looking at Figure 4.4 that the suggested SC-DTDM methodology outperforms both the currently. This is the case even though the latter methodology was the one that was advised. This argument is supported by the fact that the technique in question is superior, which acts as proof for the claim. The Spectral Cluster-based Classification Tree Analysis Technique with Modeling Level shown an improvement of 8 percent in overall accuracy when contrasted with the method that had been employed previously.

IV. CONCLUSION

In this piece of writing, a sophisticated data mining approach that is built on clustering has been established in order to analyze the overall performance of students. The goal of this approach is to cluster students into similar groups. It is feasible to obtain an accurate estimate of a student's GPA by using this method, which will lead to higher levels of true positive assessments. This will lead to increased rates of true positive evaluations. Because of this, the percentage of ratings that are authentically positive will increase.

REFERENCES

[1]. Saarela, M., & Kärkkäinen, T. (2015). Analysing student performance using sparse data of core bachelor courses. Journal of educational data mining, 7(1).

[2]. Prakash, K. P. (2021). An Intelligent Clustering Technique for Analysing the Performance of Students during Lockdown Period of Covid-19. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 2499-2512.

[3]. Yin, X. (2021). Construction of student information management system based on data mining and clustering algorithm. Complexity, 2021.

[4]. Scheidt, M., Godwin, A., Berger, E., Chen, J., Self, B. P., Widmann, J. M., & Gates, A. Q. (2021). Engineering students' noncognitive and affective factors: Group differences from cluster analysis. Journal of Engineering Education, 110(2), 343-370.

[5]. Cheng, W., & Shwe, T. (2019, October). Clustering Analysis of Student Learning Outcomes Based on Education Data. In 2019 IEEE Frontiers in Education Conference (FIE) (pp. 1-7). IEEE.

[6]. Asiah, M., Zulkarnaen, K. N., Safaai, D., Hafzan, M. Y. N. N., Saberi, M. M., & Syuhaida, S. S. (2019). A review on predictive modeling technique for student academic performance monitoring. In MATEC Web of Conferences (Vol. 255, p. 03004). EDP Sciences.

[7]. Xie, J., Girshick, R., & Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487). PMLR.

[8]. Rahman, M. M., Watanobe, Y., Kiran, R. U., Thang, T. C., & Paik, I. (2021). Impact of Practical Skills on Academic Performance: A Data-Driven Analysis. IEEE Access, 9, 139975-139993.

[9]. Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

[10]. Marrone, P., Gori, P., Asdrubali, F., Evangelisti, L., Calcagnini, L., & Grazieschi, G. (2018). Energy benchmarking in educational buildings through cluster analysis of energy retrofitting. Energies, 11(3), 649.