Clustering and Classification of Student Performance Using K-Medoids, Fuzzy-C-Means, and Ensemble Methods

Tasnia Haque Kheya*
23kheya@gmail.com(Id:0424052084)
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Shaikh Prothomaa Binte Minhaz*
prothomaaminhaz@gmail.com(Id: 0424052087)
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Md. Mahadi Hasan Ankon*
mahadi.ankon@gmail.com(Id: 0424058012)
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh

Abstract

The prediction of student performance is a crucial task in modern education systems, allowing educators and institutions to identify struggling students early and provide targeted interventions. Traditional approaches to monitoring academic progress are often inefficient, subjective, and time-consuming. With the rise of Educational Data Mining (EDM) and Machine Learning (ML), data-driven approaches have gained traction in accurately predicting student outcomes. This study proposes a robust predictive model that utilizes ensemble learning techniques to enhance the accuracy of student performance predictions. Data preprocessing steps, such as one-hot encoding, clustering, feature selection, and handling of missing values, were applied to ensure data integrity and improve model efficiency. Multiple machine learning algorithms were implemented to achieve optimal performance, including Decision Trees, Random Forest, Naïve Bayes, Support Vector Machines (SVM), AdaBoost, and Gradient Boosting. These models were combined using an ensemble architecture with soft voting, leading to significant improvements in prediction accuracy. The final model achieved an accuracy of 97.77%, demonstrating the effectiveness of ensemble learning in student performance prediction. The results of this study provide meaningful insights into the factors affecting student success and highlight the potential of ML-driven solutions in educational settings. By classifying students into performance categories (Excellent, Moderate, and At-Risk), this research enables data-driven decision-making for teachers, institutions, and policymakers.

Keywords

Data mining, Ensemble, K-means, DBSCAN, Classification, Clustering

*All authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Tasnia Haque Kheya, Shaikh Prothomaa Binte Minhaz, and Md. Mahadi Hasan Ankon. 2025. Clustering and Classification of Student Performance Using K-Medoids, Fuzzy-C-Means, and Ensemble Methods. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The prediction of student performance has become an essential area of research in the field of educational data mining (EDM) and artificial intelligence (AI), as it enables institutions to proactively address academic challenges, identify at-risk students, and optimize learning experiences. Traditional approaches to evaluating student performance, such as standardized testing, teacher assessments, and periodic examinations, often suffer from inefficiencies, subjectivity, and delayed feedback, limiting their ability to provide timely interventions. Moreover, these conventional methods tend to focus primarily on academic scores while overlooking other crucial factors, such as student engagement, attendance, socioeconomic background, psychological well-being, and extracurricular activities, which significantly impact learning outcomes. In response to these challenges, machine learning (ML) techniques have emerged as powerful tools for analyzing complex educational datasets and predicting student success with higher accuracy. By leveraging ML algorithms, institutions can automate the prediction process, allowing for real-time monitoring of academic progress, early identification of students in need of support, and the implementation of personalized learning strategies tailored to individual needs. In particular, ensemble learning—a method that combines multiple models to improve prediction accuracy—has proven to be a highly effective approach in various domains, yet its application in student performance prediction remains relatively underexplored. This study aims to bridge this gap by developing an ensemble learning framework that integrates multiple ML classifiers, including Decision Trees, Random Forest, Support Vector Machines (SVM), Naïve Bayes, and AdaBoost, to enhance prediction accuracy and robustness. The dataset used in this research, "Student Performance BD," comprises 8,612 student records with 24 features encompassing academic, demographic, and behavioral factors, making it a comprehensive dataset for evaluating different machine learning approaches. To ensure data quality and improve model efficiency, various preprocessing techniques, such as missing value imputation, feature selection, clustering, and outlier detection, are applied before training the predictive models. Furthermore, the study explores the impact of different clustering techniques, including K-Means,

K-Medoids, and DBSCAN, to identify meaningful student performance categories—Excellent, Moderate, and At-Risk. The results of the study demonstrate that ensemble learning significantly improves prediction accuracy, with the proposed model achieving an impressive accuracy of 97.77% which outperforms traditional single-model approaches. By providing a more precise and data-driven method for predicting student performance, this research offers valuable insights for educators, policymakers, and academic institutions to develop targeted interventions, improve teaching methodologies, and enhance overall student success. Ultimately, the adoption of machine learning in educational settings represents a paradigm shift in how student performance is analyzed, monitored, and improved, paving the way for more intelligent and adaptive learning systems that cater to the diverse needs of students.

2 Related works

Feng et al. [5] proposed a deep learning-based approach to predict student performance using educational data mining techniques. The study utilized three datasets from a university, where students within the same dataset were enrolled in the same courses. The researchers employed K-Means clustering (k = 4) to group students based on their academic performance and a five-layer convolutional neural network (CNN) with ReLU activation to construct a predictive model. The model achieved accuracy levels of 94.59%, 94.29%, and 93.29% for three different datasets. However, a key limitation of this study was the random selection of initial clustering centers in K-Means, which can lead to inconsistent results. Additionally, the authors did not compare the effectiveness of K-Means with alternative clustering methods, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or hierarchical clustering. Furthermore, the study lacked advanced evaluation techniques like Receiver Operating Characteristic - Area Under Curve (ROC-AUC) analysis, which could provide a more comprehensive assessment of model performance.

Sarker et al. [7] explored student performance prediction using a college dataset and a synthetic dataset for algorithm development and validation. The study applied Decision Trees, K-Nearest Neighbors (KNN), Naïve Bayes, Neural Networks, and Random Forest for classification and analysis. The researchers proposed two novel GPA calculation approaches (GPA-1 and GPA-2) based on internal examination marks and compared them with official board GPA scores. The results highlighted how Decision Trees effectively identified subjects that significantly correlated with student performance, thereby offering deeper insights into academic outcomes. However, the study faced several limitations, including small and imbalanced datasets, which may have affected the generalizability of the findings. Moreover, the authors did not explore advanced feature selection techniques, such as Least Absolute Shrinkage and Selection Operator (LASSO), which could enhance prediction accuracy by eliminating irrelevant features. Additionally, the study did not benchmark Decision Tree results against more modern ensemble learning techniques or deep learning architectures, leaving scope for further improvements in predictive modeling.

Chen et al. [4] developed a student performance prediction approach based on clustering and hybrid neural networks. The study applied one-hot encoding and data binning for preprocessing and used the Louvain algorithm to cluster students based on modularity metrics. The core predictive model, a hybrid neural network (RMHNN), achieved an impressive accuracy of 93.1% in forecasting student grades. Despite its high accuracy, the study had some notable limitations. The authors did not assess the model's robustness to noise or missing data, which is a crucial factor in real-world educational datasets. Additionally, the study focused primarily on the Louvain clustering algorithm without benchmarking it against alternative clustering techniques, such as hierarchical clustering or Gaussian Mixture Models (GMM), which could have provided a comparative perspective on performance clustering methods.

Satyanarayana et al. [8] employed ensemble classification and K-Means clustering, achieving accuracies of 91-95%. Despite high accuracy, the study's small dataset and failure to address missing or imbalanced data limited the generalizability of the results. More sophisticated ensemble methods like bagging or boosting were not explored.

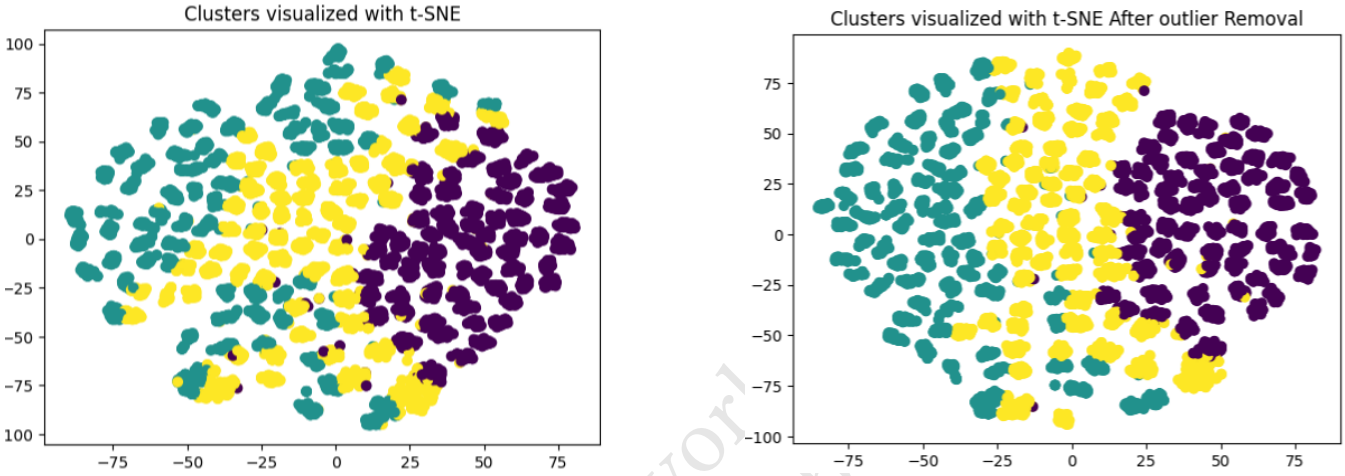
Alhazmi et al. [3] conducted a study on early student performance prediction in higher education using a private dataset. The authors applied feature extraction using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm and compared multiple ML models, including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and XGBoost. Among these, XGBoost demonstrated the highest predictive accuracy, achieving 66.37%, which was significantly lower than other state-of-the-art approaches. One of the key limitations of this study was the low accuracy achieved, indicating the need for improved feature engineering and model tuning. Additionally, the study focused on data from a single institution, reducing the diversity and robustness of the findings. Moreover, the research did not incorporate self-improvement mechanisms, where students could receive automated feedback based on predictions, a feature that could enhance practical applications in educational settings.

Ahmad et al. [2] introduced a multi-model ensemble approach to predict student performance using data from a Pakistani provincial university. The study implemented Naïve Bayes (NB), J48 Decision Trees, and a Multilayer Perceptron (MLP) within an ensemble framework that included bagging, boosting, stacking, and voting methods. Among these techniques, stacking-based ensemble classification achieved the highest accuracy of 95%, demonstrating the effectiveness of ensemble learning in educational data mining. However, the study faced several challenges, including the lack of outlier handling during preprocessing, which could impact classification accuracy. Additionally, the model was developed for binary classification (Pass or Fail), limiting its applicability for more detailed academic performance predictions. The dataset was also sourced from a single university, restricting the generalizability of the model to diverse student populations.

Guerrero et al. [6] conducted a review of ML approaches, concluding that supervised learning techniques typically outperform unsupervised ones in predicting student success. The review identified research gaps in integrating deep learning and ensemble models for enhanced accuracy.

ID	Full Name	Age	Gender	Location	Family Size	Mother Edu	Father Edu	Mother Job	Father Job	Guardian	Parental Inv.	Internet	Study Time	Tutoring	School Type	Attendance	Extra Activities	English	Math	Science	Soc. Science	Art/Culture	Student Group
2	Avi Biswas	16	Male	Urban	6	SSC	HSC	No	No	Father	Yes	Yes	8	Yes	Private	95	Yes	95	98	92	94	98	Science
3	Taslima Sultana	18	Female	Rural	6	SSC	HSC	No	Yes	Father	Yes	No	4	No	Semi_govt	92	No	65	71	40	78	80	Commerce
4	Md Adilur Rahman	15	Male	Rural	4	SSC	SSC	Yes	Yes	Father	Yes	Yes	5	Yes	Govt	81	Yes	64	78	58	86	74	Commerce
5	Saleh Ahmed	16	Male	Rural	6	SSC	SSC	Yes	Yes	Father	Yes	Yes	7	Yes	Private	90	Yes	84	90	85	86	88	Science

Table 1: Dataset - "Student Performance BD"



(a) abc **(b) def**

Figure 1: Visualization of the clusters without and with outlier removal respectively)

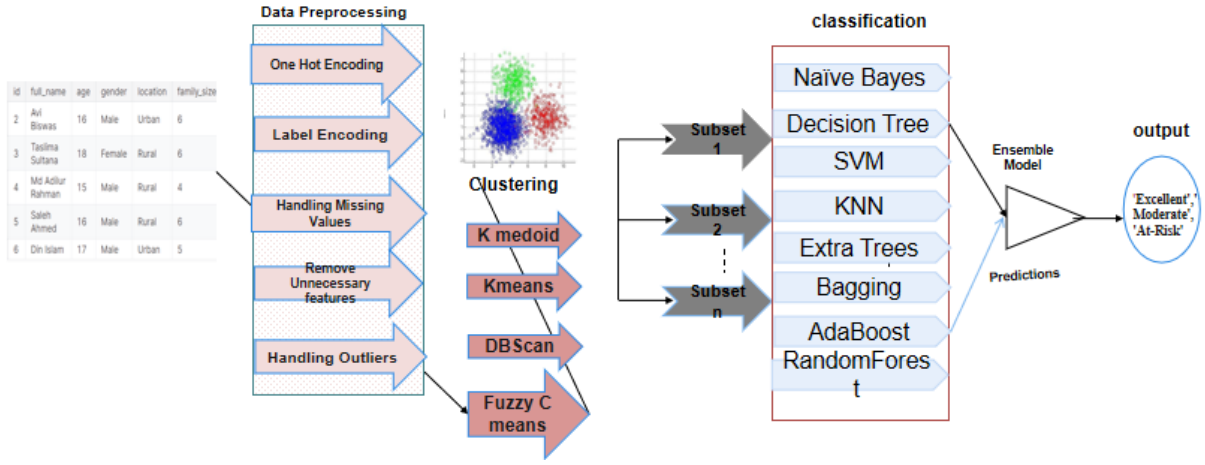


Figure 2: Proposed Main flow

3 Dataset

This study utilizes a dataset "Student Performance BD", collected from Kaggle [1], It is publicly available and authorized for research and academic use. This dataset contains information on the academic performance of students from various regions in Bangladesh. It contains about 9000 data of different students and features 24 attributes related to demographics, academic metrics, extracurricular activities, and other relevant factors.The authors of this dataset collect about 80% of the features data on Bangladeshi students from online platforms and the rest 20% was collected by them from physical surveys of 3 different schools.

4 Preprocessing

The preprocessing phase aims to standardize and improve the quality of the input data. The data set is examined for missing values and appropriate imputation techniques are applied where necessary. Several categorical attributes, such as gender, location, family background, and subject choices, are encoded using label encoding to convert them into numerical values for processing.Continuous numerical attributes, such as grades in various subjects and attendance rates, are standardized using the StandardScaler to ensure uniformity across different feature ranges.Principal Component Analysis (PCA) and t-SNE are used for dimensionality reduction to identify the most significant features for clustering.Upon completion of the preprocessing phase, the dataset is refined and optimized for clustering and classification, ensuring that all attributes are in a suitable format for subsequent analysis.

5 Proposed Methodology

Following preprocessing, clustering techniques are employed to group students based on their academic performance and other relevant attributes. Multiple clustering algorithms, including K-Medoids, K-Means, DBSCAN, and Fuzzy C-Means, are utilized to explore different data structures and improve cluster interpretability. Empty clusters and overlapped clusters were handled using Fuzzy C-Means clustering technique. The Fuzzy C-Means algorithm in this study handles overlapping clusters by assigning membership probabilities to each data point, allowing for soft clustering rather than strict categorization. The fuzziness parameter (m=2) controls cluster diffusion, ensuring smooth transitions between performance categories. To address empty clusters, a reallocation strategy is employed, where low-certainty points are reassigned, and cluster centroids are reinitialized if necessary. This approach ensures robust clustering, effectively capturing variations in student performance.All the clustering results not only help in identifying distinct performance groups but also play a crucial role in detecting potential outliers that may skew the classification results.

In the final phase, the clustered data is used for classification, where various machine learning models are employed to predict student performance categories. The classification framework includes algorithms such as Naïve Bayes, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Extra Trees, Bagging, AdaBoost,and Random Forest. These classifiers operate on different subsets of the clustered data, ensuring robust and diverse predictions. To enhance the overall accuracy and reliability of the system, an ensemble model is constructed by aggregating the predictions from multiple classifiers. The ensemble approach effectively combines the strengths of individual models using hard voting technique which reduces overfitting and improving generalization.The final output categorizes students into three performance groups: "Excellent," "Moderate," and "At-Risk". The integration of clustering for outlier detection, combined with ensemble-based classification, ensures a highly adaptive, interpretable, and accurate student performance analysis.

6 Results

Several clustering algorithms were evaluated to segment the dataset into three clusters: Excellent, Moderate, and At-risk. These clusters were intended to serve as labels for further classification tasks. The clustering methods tested included K-Means, DBSCAN, and K-Medoids, with the resulting clusters subsequently classified using SVM, KNN, and Decision Tree algorithms. From the results, DBSCAN did not effectively separate the clusters, as the majority of data points were grouped into a single cluster, making it unsuitable for further classification. In contrast, both K-Means and K-Medoids provided a well-distributed clustering. However, a comparison of classification metrics reveals that K-Medoids outperformed the others, achieving the highest accuracy across all classifiers. Given these findings, K-Medoids was selected for further study, as it provided a more balanced and effective clustering structure for classification.

The performance of different clustering algorithms are shown in table 2

Table 3 demonstrates that filtering the data using the mean and a carefully selected threshold, which is determined through trial and error with the Random Forest classifier, significantly improved classification performance. This filtering step refined the dataset, reducing noise and enhancing class separability, leading to better model generalization. SVM saw the highest accuracy boost (0.95 to 0.97), while Random Forest itself improved from 0.89 to 0.94, highlighting the impact of the optimized threshold. Naïve Bayes and Decision Tree also benefited, with accuracy increasing by 5% and 3%, respectively. However, KNN's performance remained unchanged (0.74 to 0.73), likely due to its sensitivity to local variations rather than global trends. These results confirm that filtering through statistical means and an optimized threshold is a crucial step in enhancing classifier performance.

To enhance classification performance, we selected the best-performing models, such as SVM, Decision Tree, and Random Forest—and applied an ensemble approach. By leveraging the strengths of each classifier, the ensemble effectively reduced individual model biases and improved overall generalization. The results show a notable improvement across all evaluation metrics, with an accuracy of 0.9777, precision of 0.9777, recall of 0.9777, and an F1 score of 0.9777. These near-perfect and balanced metrics indicate that the ensemble approach achieved a robust and well-calibrated classification, minimizing false positives and false negatives. Compared to individual models, the ensemble demonstrates superior stability and reliability, confirming its effectiveness in handling the dataset more efficiently than any single classifier alone. To further enhance performance and address overlapping data points, we applied Fuzzy C-Means (FCM) clustering. This helped in refining the dataset and improving classification accuracy. The results show a significant improvement, with an accuracy of 0.9912, precision of 0.9914, recall of 0.9912, and an F1 score of 0.9913. These near-optimal metrics indicate that FCM effectively reduced misclassifications by handling overlapping regions more flexibly, further enhancing model robustness and generalization.

The confusion matrix and ROC curve for the ensemble model with Fuzzy C-Means are shown in Figure 3 and Figure 4, respectively. These visualizations confirm the superior classification accuracy and the model's ability to distinguish between classes effectively.

Clustering Algorithm	Excellent	Moderate	At-risk	SVM accuracy	KNN accuracy	Decision Tree accuracy
Kmeans	3156	2823	2620	0.86	0.73	0.85
DBSCAN	7825	343	429	-	-	-
K-Medoids	3097	2749	2753	0.89	0.74	0.87

Table 2: Performance on different clustering algorithms.

Model	Accuracy (Before)	Precision (Before)	Recall (Before)	Accuracy (After)	Precision (After)	Recall (After)
SVM	0.95	0.95	0.95	0.97	0.97	0.97
Decision Tree	0.88	0.88	0.88	0.91	0.92	0.91
KNN	0.74	0.74	0.74	0.73	0.73	0.73
Naive Bayes	0.87	0.87	0.87	0.92	0.92	0.92
Random Forest	0.89	0.90	0.89	0.94	0.94	0.94

Table 3: Performance of various classifiers for the data labeled by the K-Medoids algorithm.

Model	Accuracy	Precision	Recall	F1 Score
Ensemble (SVM + Decision Tree + Random Forest)	0.9777	0.9777	0.9777	0.9777
Ensemble + Fuzzy C-Means (FCM)	**0.9912**	**0.9914**	**0.9912**	**0.9913**

Table 4: Performance comparison of ensemble classification and ensemble with Fuzzy C-Means clustering.

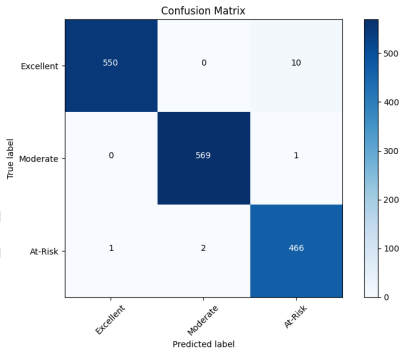


Figure 3: Confusion Matrix for the ensemble model with Fuzzy C-Means.

7 Discussion

In this study, we aimed to cluster a dataset containing demographic, academic, behavioral, and environmental features into three distinct categories: excellent, moderate, and at-risk students. The clustering process was initially performed using K-Medoids, which outperformed other algorithms, including K-Means and DBSCAN, in terms of class separation and accuracy. To further refine the dataset, outliers were handled using filtering techniques based on the mean and an optimized threshold, which was carefully selected through trial and error, significantly improving model performance, especially with Random Forest. In order to address the challenge of overlapping data points, we employed Fuzzy C-Means clustering, which allowed for a more nuanced classification by assigning data points to multiple clusters with varying membership degrees, thereby improving the handling of ambiguous cases. Subsequently, an ensemble approach combining SVM, Decision Tree, and Random Forest was utilized to leverage the strengths of each classifier, resulting in a substantial performance improvement. The final results demonstrated a high level of classification accuracy, with the ensemble model achieving an accuracy of 0.9777 and Fuzzy C-Means boosting it further to 0.9912, underscoring the effectiveness of these methods in accurately categorizing students based on their multi-faceted characteristics.

8 Conclusion

This study successfully applied a series of advanced machine learning techniques to classify students into three distinct categories: excellent, moderate, and at-risk. By leveraging K-Medoids clustering and handling outliers through filtering and optimized thresholds, we improved the overall performance of the classification models.

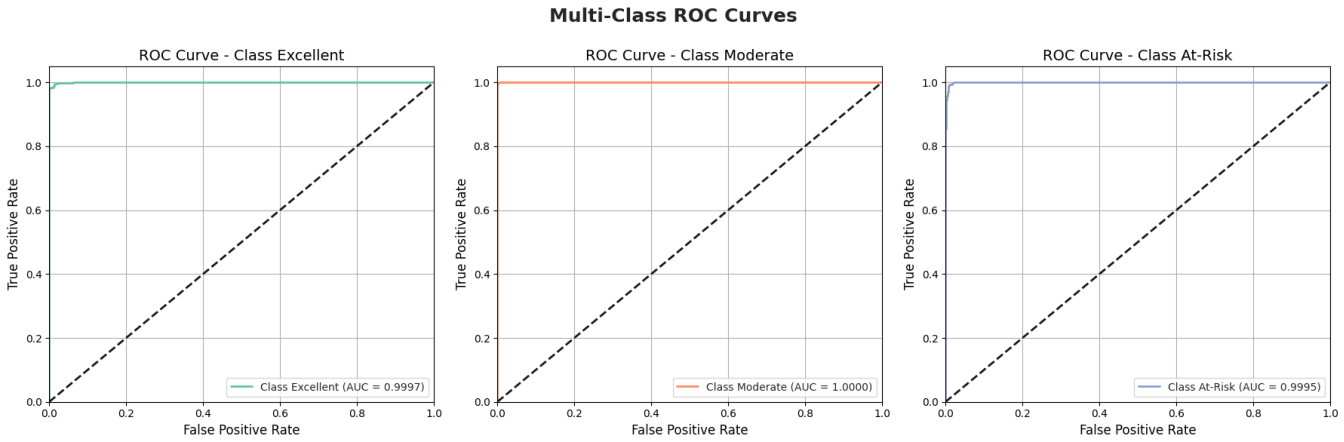


Figure 4: ROC Curve for the ensemble model with Fuzzy C-Means.

The use of Fuzzy C-Means addressed the issue of overlapping data points, enhancing the model's ability to handle ambiguous cases. The ensemble approach, combining SVM, Decision Tree, and Random Forest, resulted in a substantial improvement in classification accuracy, culminating in a final model that achieved impressive metrics across accuracy, precision, recall, and F1 score. These results demonstrate that combining clustering with ensemble methods and fuzzy techniques can yield highly effective solutions for complex classification tasks.

9 Future Work

In the future, we can expand the dataset to include additional features, such as psychological or socio-economic factors, to enhance the model's predictive capabilities. Exploring alternative clustering algorithms, such as Gaussian Mixture Models or Hierarchical Clustering, could offer insights into how different techniques handle complex data distributions. Moreover, we can investigate the use of deep learning-based approaches for feature extraction or classification, potentially improving performance further. Finally, applying this approach to real-world educational data will help validate the model's effectiveness and support the development of early intervention strategies for at-risk students.

References

[1] [n. d.]. Student Performance-BD — kaggle.com. https://www.kaggle.com/datasets/satayjit/student-performance-bd/data. [Accessed 09-02-2025].

[2] F. Ahmad, N.H. Ismail, and A.A. Aziz. 2015. The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences* 9 (01 2015), 6415–6426. doi:10.12988/ams.2015.53289

[3] Essa Alhazmi and Abdullah Sheneamer. 2023. Early Predicting of Students Performance in Higher Education. *IEEE Access* PP (01 2023), 1–1. doi:10.1109/ACCESS.2023.3250702

[4] Ziling Chen, Gang Cen, Ying Wei, and Zifei Li. 2023. Student Performance Prediction Approach Based on Educational Data Mining. *IEEE Access* 11 (2023), 131260–131272. doi:10.1109/ACCESS.2023.3335985

[5] Guiyun Feng, Muwei Fan, and Yu Chen. 2022. Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. *IEEE Access* 10 (2022), 19558–19571. doi:10.1109/ACCESS.2022.3151652

[6] Juan L. Rastrollo-Guerrero, Juan A. Gómez-Pulido, and Arturo Durán-Domínguez. 2020. Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences* 10, 3 (2020). doi:10.3390/app10031042

[7] Sazol Sarker, Mahit Kumar Paul, Sheikh Tasnimul Hasan Thasin, and Md. Al Mehedi Hasan. 2024. Analyzing students' academic performance using educational data mining. *Computers and Education: Artificial Intelligence* 7 (2024), 100263. doi:10.1016/j.caeai.2024.100263

[8] Ashwin Satyanarayana and Gayathri Ravichandran. 2016. Mining student data by ensemble classification and clustering for profiling and prediction of student academic performance. In *American Society for Engineering Education.*