**Previous academic analysis**

**ABSTRACT:**

Educational Data Mining (EDM) and predictive analytics have emerged as crucial tools in higher education for analyzing and improving student performance. EDM focuses on extracting meaningful patterns from educational data to enhance curriculum design, teaching strategies, and assessment methods. Similarly, predictive analytics leverages advanced machine learning techniques to forecast academic outcomes and identify key performance indicators like student grades. This study combines insights from both domains to analyze the performance of students, addressing challenges such as imbalanced datasets and multi-class prediction accuracy. By examining higher secondary students in Bangladesh, the research explores four critical aspects: predicting final academic performance based on internal examinations, assessing the impact of individual subjects, monitoring performance progression, and identifying consistent trends in internal exams. Complementarily, six machine learning models, including Decision Tree, Support Vector Machine, and Random Forest, were evaluated for grade prediction accuracy on a real dataset of 1,282 students. The integration of Synthetic Minority Oversampling Technique (SMOTE) and feature selection with Random Forest achieved a remarkable 99.5% f-measure, significantly improving prediction accuracy for imbalanced datasets. These findings underscore the potential of EDM and machine learning to empower educators in identifying at-risk students, fostering consistent performers, and enhancing overall academic outcomes.

**Introduction**

In higher education institutions (HEIs), evaluating and improving student academic performance is critical for fostering educational success and producing skilled professionals. HEIs maintain comprehensive academic management systems to record student data, including grades and examination results, which are traditionally used for performance reporting and course evaluation. However, the vast repositories of student data offer opportunities for uncovering meaningful patterns and insights through Educational Data Mining (EDM) and predictive analytics. These approaches enable institutions to analyze trends, predict outcomes, and implement strategies for enhancing academic achievement ([Han et al., 2022](https://www.sciencedirect.com/science/article/abs/pii/B9780123814791000010); [Baker, 2010](https://www.sciencedirect.com/science/article/pii/B9780080448947017497)).

Recent studies highlight that socio economic factors, demographics, and learning activities significantly influence student performance (Kono et al., 2018). Predictive analytics, employing advanced machine learning techniques, has demonstrated considerable success in addressing challenges like dropout prediction, academic early warnings, and course selection ([Solomon et al., 2020](https://link.springer.com/article/10.1007/s10462-019-09708-0)). However, challenges persist, particularly in handling imbalanced datasets for multi-class student grade prediction. Addressing these issues, this study combines EDM and predictive analytics to analyze student performance from multiple perspectives, including subject-wise impact, consistent performance trends, and imbalanced data handling using oversampling techniques like Synthetic Minority Oversampling Technique (SMOTE) and feature selection (Maxwell, 2012; Alo, 2020).

By leveraging real-world datasets, including higher secondary students in Bangladesh and diploma students in Malaysia, this study aims to identify effective predictive models, evaluate the influence of internal examination results on final grades, and propose actionable insights for educators. These findings not only aid in reducing academic disparities but also contribute to shaping data-driven strategies for empowering students and institutions to excel in a competitive educational landscape.

**Research Objectives**

This study seeks to enhance the understanding of academic performance among students enrolled in a two-year Higher Secondary Certificate (HSC) humanities program in Bangladesh. By focusing on analyzing internal examination performance and its correlation with final outcomes, the research aims to provide actionable insights to educators and institutions. Specifically, the study aims to predict final academic performance using classification models based solely on internal examination marks, bypassing demographic and socioeconomic variables, to ensure a focused and unbiased approach ([Han et al., 2022](https://www.sciencedirect.com/science/article/abs/pii/B9780123814791000010)).

A key objective is to identify subjects that act as strong predictors of student performance and classify them as booster, degrader, or at-risk subjects, thereby enabling targeted interventions. The study emphasizes the use of interpretable models such as Decision Trees (DT) to ensure results are easily comprehensible by non-technical stakeholders, including teachers and administrators ([Raileanu & Stoffel, 2004](https://ieeexplore.ieee.org/document/5666579)). Moreover, by examining performance progression across the two-year period, the study strives to uncover consistent patterns and trends, allowing early detection of struggling or at-risk students and offering opportunities for timely support ([Asif et al., 2017](https://link.springer.com/article/10.1007/s11423-017-9506-7)).

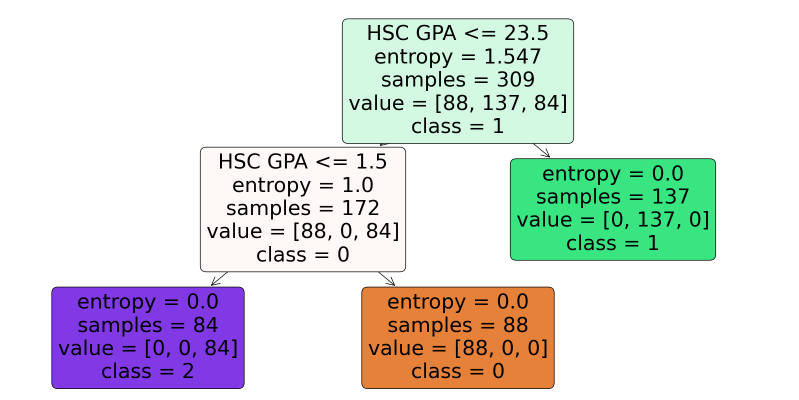
The research further evaluates the effectiveness of proposed GPA calculation methods compared to board GPA for predicting outcomes and analyzes the impact of feature selection techniques in improving classification accuracy. By employing advanced machine learning techniques such as SMOTE and Decision Trees, the study addresses challenges posed by imbalanced datasets and ensures robust predictive performance ([Al-Barrak, 2016](https://www.sciencedirect.com/science/article/pii/S1877050917301677)). These objectives collectively aim to bridge the gap between academic analytics and practical interventions, ensuring a holistic improvement in students’ academic trajectories.

**DATA MINING TECHNIQUES:**

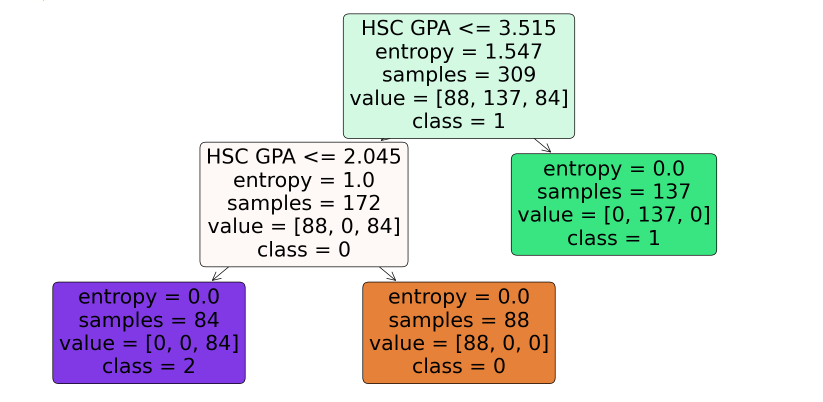
**DECISION TREES:**

Decision Tree techniques play a vital role in student performance prediction by analyzing internal examination scores and final academic outcomes. These models effectively classify students into performance categories by identifying key subject-wise trends and their impact on grades. The Decision Tree classifier, when combined with feature selection techniques like Information Gain and Gini Index, enhances interpretability and accuracy in predicting student success (Asif et al., 2017)​. Additionally, integrating Synthetic Minority Oversampling Technique (SMOTE) addresses dataset imbalances, improving classification performance in multi-class prediction scenarios (Bujang et al., 2021)​. By refining attribute selection and reducing bias (Raileanu & Stoffel, 2004; Han et al., 2022)​, Decision Tree models provide valuable insights for educators to identify at-risk students, monitor academic progression, and implement targeted interventions, ultimately fostering improved learning outcomes.

Decision Tree techniques play a vital role in student performance prediction by analyzing internal examination scores and final academic outcomes. These models effectively classify students into performance categories by identifying key subject-wise trends and their impact on grades. The Decision Tree classifier, particularly J48, is widely used in multi-class classification as it can handle missing values and high-dimensional data while delivering optimal accuracy with a minimal number of features​. When combined with feature selection techniques like Information Gain and Gini Index, Decision Trees enhance interpretability and accuracy in predicting student success (Asif et al., 2017)​. Additionally, integrating Synthetic Minority Oversampling Technique (SMOTE) addresses dataset imbalances, improving classification performance in multi-class prediction scenarios (Bujang et al., 2021)​. By refining attribute selection and reducing bias (Raileanu & Stoffel, 2004; Han et al., 2022)​, Decision Tree models provide valuable insights for educators to identify at-risk students, monitor academic progression, and implement targeted interventions, ultimately fostering improved learning outcomes.



Decision tree for hsc gpa



(proposed gpa)

**KNN**

K-Nearest Neighbors (K-NN) is a simple yet powerful classification algorithm used for student performance prediction, particularly in multi-class scenarios. It operates by identifying the ‘K’ closest training data points in the dataset and assigning the majority class label to the new instance based on these neighbors. The Euclidean distance is commonly used to measure proximity between data points, ensuring accurate classification by minimizing the distance between similar instances (Kramer & Kramer, 2013)​. While K=1 is often used in binary classification, multi-class problems require a larger K value to improve prediction stability and reduce variance, with this study setting K=5 for optimal performance (Yuan et al., 2008)​. However, K-NN can be computationally expensive, particularly for large datasets, as it must calculate distances for each new data point against all existing training instances (Sun & Huang, 2010)​. Despite this limitation, K-NN remains a valuable method for student grade prediction, offering high accuracy in cases where well-distributed datasets and appropriate feature selection techniques are applied.

**SVM**

Support Vector Machine (SVM) is a powerful classification algorithm that constructs optimal hyperplanes to distinguish between different classes in a dataset. It is particularly effective in multi-class classification tasks, such as predicting student academic performance, where it identifies complex decision boundaries between grade categories. In student performance prediction, SVM has demonstrated strong accuracy in classifying students based on internal examination scores and final academic outcomes. However, its performance is significantly affected by imbalanced datasets, which can lead to misclassification of minority classes. To address this challenge, oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE) have been applied, enhancing the model’s ability to handle skewed class distributions and improving overall predictive accuracy (Bujang et al., 2021)​. Furthermore, in educational data mining, SVM has been utilized to analyze patterns in student learning behaviors and academic trends, reinforcing its effectiveness in predictive analytics (Sarker et al., 2024)​. By integrating feature selection methods and optimizing hyperparameters, SVM models can deliver high-precision predictions, assisting educators in identifying at-risk students and developing targeted interventions to enhance learning outcomes.

**NAIVE BAYES:**

Naïve Bayes is a probabilistic machine learning algorithm widely used for classification tasks, particularly in educational data mining for predicting student performance. Based on Bayes’ theorem, it assumes feature independence and calculates the posterior probability of a class by considering the likelihood of given features and their prior probabilities (Friedman et al., 1997). The algorithm efficiently classifies students into different academic performance categories by analyzing their internal examination scores and other learning attributes (Sarker et al., 2024)​. Despite its simplicity and fast computation, Naïve Bayes can struggle with imbalanced datasets, often leading to misclassification of minority class instances. To mitigate this, techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been employed, enhancing its accuracy in multi-class student grade prediction (Bujang et al., 2021)​. In comparative analyses with other classifiers like Decision Tree and Support Vector Machine, Naïve Bayes has shown competitive performance, particularly when integrated with feature selection methods that refine attribute importance and reduce bias (Bujang et al., 2021)​. By leveraging prior probabilities and optimizing classification strategies, Naïve Bayes remains an effective tool for academic performance prediction, assisting educators in identifying at-risk students and implementing targeted interventions to enhance learning outcomes.

**RANDOM FOREST**

Random Forest is a powerful ensemble learning algorithm widely used for classification tasks, including student performance prediction. By constructing multiple decision trees and aggregating their outputs, Random Forest improves predictive accuracy and reduces the risk of overfitting, making it highly effective for handling complex educational datasets (Sarker et al., 2024)​. In student grade prediction, Random Forest has demonstrated superior performance compared to other classifiers like Support Vector Machine and Naïve Bayes, particularly when combined with techniques such as the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance issues (Bujang et al., 2021)​. Among different attribute evaluation methods, the Gini Index within Random Forest has achieved the highest accuracy, indicating its effectiveness in selecting the most relevant features for classification tasks (Bujang et al., 2021)​. Additionally, Random Forest excels in educational data mining by analyzing patterns in student learning behaviors and performance trends, helping institutions make data-driven decisions to enhance student success (Sarker et al., 2024)​. By leveraging its ability to handle high-dimensional data and improve interpretability through feature importance analysis, Random Forest remains one of the most effective machine learning models for academic performance prediction, providing educators with actionable insights to optimize learning outcomes and support at-risk students.

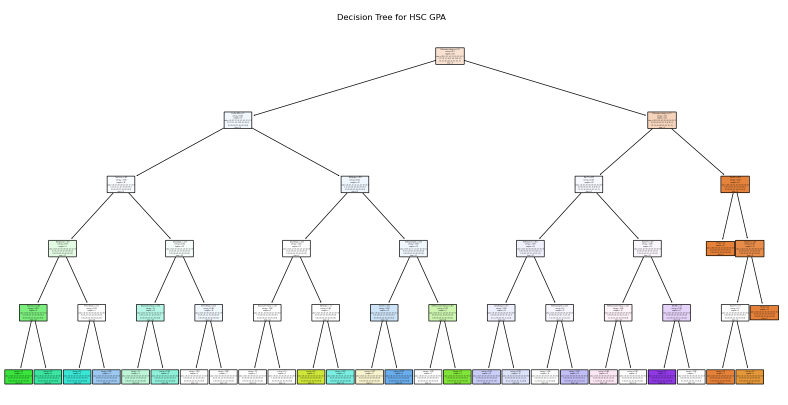
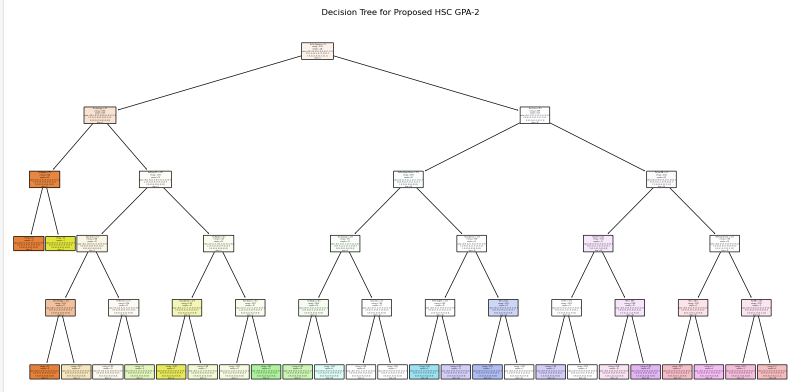
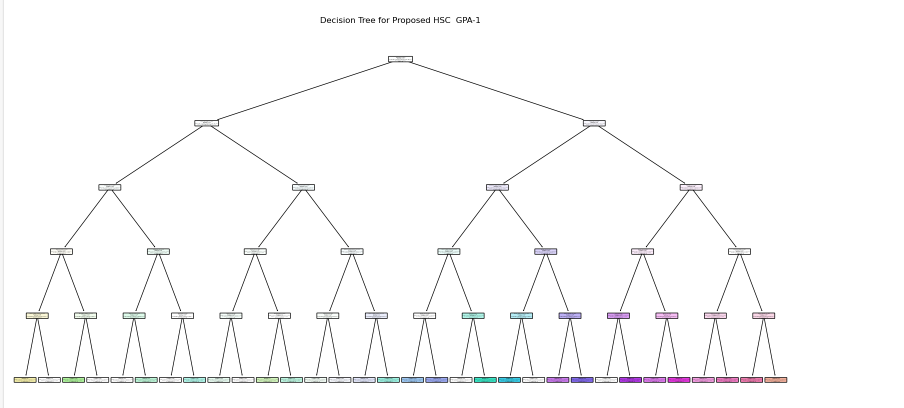
**J48:**

J48, a widely used Decision Tree algorithm, is highly effective in student performance prediction due to its ability to handle missing values and high-dimensional data while maintaining optimal accuracy. It constructs decision trees based on attribute selection measures like Information Gain and Gini Index, making it particularly suitable for multi-class classification tasks in educational data mining (Sarker et al., 2024)​. In a comparative study on student grade prediction, J48 achieved high precision, with a score of 0.989, ranking among the top-performing classifiers alongside Random Forest (Bujang et al., 2021)​. However, its performance was impacted by imbalanced datasets, leading to misclassification of minority class instances. To mitigate this, oversampling techniques like SMOTE were applied, improving the classification accuracy and overall f-measure of J48 to 99.1% (Bujang et al., 2021)​. Additionally, J48 has been instrumental in identifying key subject-wise performance trends, assisting educators in early detection of at-risk students and optimizing intervention strategies for improved learning outcomes.

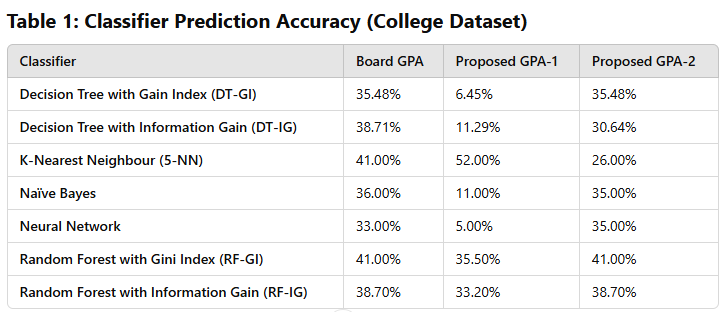
DATA SET::

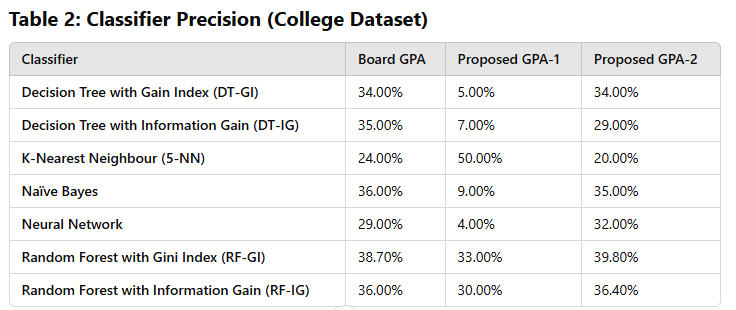
This study utilizes two primary datasets: a **college dataset** and a **synthetic dataset**, both designed to analyze student academic performance. The **college dataset** comprises records of 309 students enrolled in a two-year higher secondary humanities program in Bangladesh, capturing marks from four internal examinations—Half-Yearly, Yearly, PreTest, and Test—across seven subjects. The performance of students is categorized into three levels: **Good (60%-100%)**, **Average (50%-59%)**, and **Poor (0%-49%)**, enabling a structured classification approach (Sarker et al., 2024)​. Additionally, detailed analysis was conducted on **641 students** undertaking **Computer System Architecture (CSA) and Introduction to Computer Systems (ICS)**, where ICS students exhibited **higher mean grades and fewer failures**, whereas CSA posed greater challenges, particularly for students weak in mathematics (Bujang et al., 2021)​.

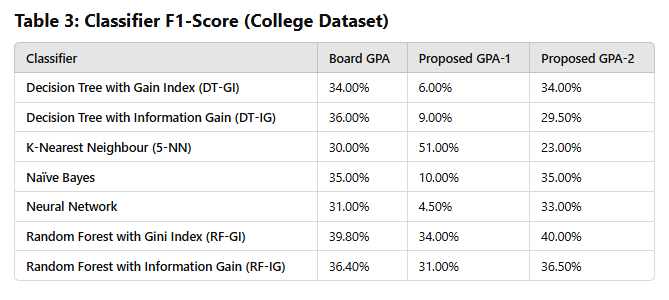
To improve class distribution and enhance predictive analysis, a **synthetic dataset** containing **1000 instances** was generated with integer values ranging from **30 to 90**, ensuring a more balanced multi-class dataset. The synthetic dataset facilitates robust evaluation of machine learning models, reducing biases associated with real-world imbalanced datasets. All data preprocessing and feature extraction were performed using **RapidMiner and Google Colaboratory**, ensuring optimal accuracy in classification, feature selection, and predictive modeling.

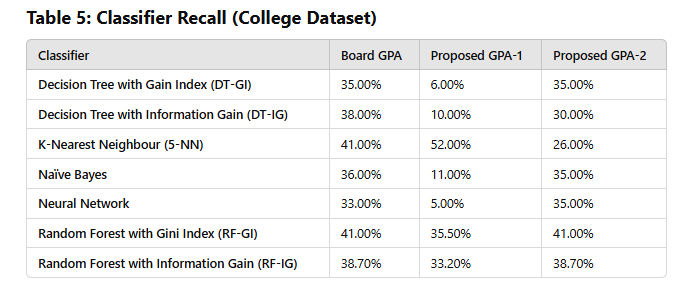
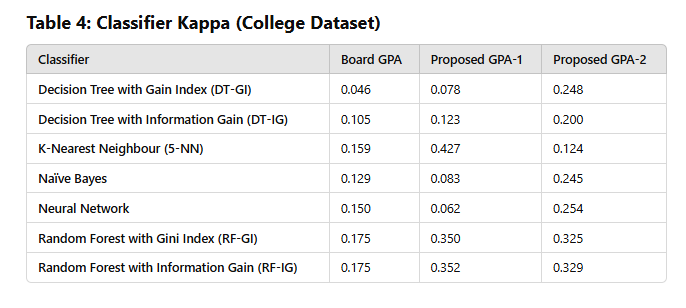


Results:

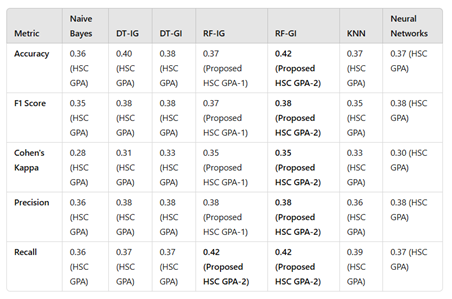








result:



REFERENCES:

1))

Alo, P. (2020). Hsc results of 2020 in January after issuance of ordinance: Dipu moni. *Prothom* *Alo*. Retrieved from [https://en .prothomalo .com /youth /education /hsc -results of -2020 -in -january -after -issuance -of -ordinance -dipu -moni](https://en.prothomalo.com/youth/education/hsc-results-of-2020-in-january-after-issuance-of-ordinance-dipu-moni).