Factors affecting Bangladesh undergraduate students' performance

Sumayya Akhter[1], Fahmid Abrar Samin[1], and Nafi Bin Noor[1]

[1]Department of Electrical and Computer Engineering, North South University

Abstract - As a country with a significant number of undergraduate students pursuing higher education, Bangladesh places immense importance on the academic performance of its undergraduate population. Understanding the factors that influence the academic success of these students is crucial for improving the quality of education and ensuring the future prospects of the nation. This paper aims to investigate and analyze the various factors affecting the performance of under- graduate students in Bangladesh with machine learning. Through a comprehensive research study, this paper examines potential factors, including but not limited to demographics, family support, study habits, class attendance and extracurricular involvement. Dataset was collected from a sample of undergraduate students in Bangladesh through surveys specifically designed for this research project. Statistical analysis was conducted on this novel dataset to identify correlations and patterns. The findings of this study provide valuable insights into the multifaceted dynamics of undergraduate student performance in Bangladesh. By shedding light on the most influential factors, educational institutions, policymakers, and students themselves can make informed decisions to enhance academic achievement.

Index Terms - Academic performance, undergraduate students, dataset, machine learning, statistical analysis.

1. Introduction

Bangladesh's education system is experiencing remarkable growth and transformation with increasing access to education at all levels, the country is witnessing a surge in enrollment and improvements in the quality of learning as the functional literacy rate is 60.77% in 2023 among the people aged 15 years and above in the country [1]. And to extend the education to the undergraduate section, 164 universities are conducted among which 53 are public universities [2]. Partnerships with international organizations is also playing a pivotal role in this educational evolution. The number of female undergraduate students are rising which is 37.58% in 2023 reflects the growth of education in Bangladesh [2].

However, Higher education in Bangladesh is a critical component of the nation's human capital development and future prosperity. With 1034320 students pursuing undergraduate degrees in 2022 [2], the academic performance of these students is a matter of paramount concern. This paper delves into the multifaceted domain of factors influencing the academic performance of undergraduate students in Bangladesh. The overarching issue we address in this study is the performance of undergraduate students, specifically, the factors affecting their academic success. While this is a part of the broader landscape of education in Bangladesh, it is essential to focus on this specific area to gain a deeper understanding of the challenges and opportunities within the education system. With the stakes higher than ever, there remains a need to address the disparities in academic performance that persist among this diverse student population. To do so, we must explore the various factors that play a role in shaping the academic outcomes of undergraduate students. This involves investigating a range of variables, including demographics, family support, study habits, study routine, class attendance, class interaction, stress level, extracurricular involvement, and part time jobs.

The problem we aim to tackle is twofold: first, to identify the key factors that contribute to academic success among undergraduate students in Bangladesh, and second, to under- stand how these factors interplay to create an environment conducive to learning and achievement. By shedding light on these factors, we hope to provide insights that can inform policies and practices aimed at enhancing the educational experience and performance of students in the country. To address this problem, we have conducted a comprehensive research study among the students from different universities that explores the diverse dimensions of undergraduate student performance in Bangladesh. By delving into this specific do- main, we can better understand the unique challenges and opportunities that exist within the higher education landscape. Through the analysis of a broad range of data, we aim to offer a nuanced perspective on the complex interplay of factors affecting undergraduate student performance. This research endeavor is crucial for the future of Bangladesh's education system and its contribution to the overall development of the nation. By clarifying the domain and problem statement, we set the stage for a rigorous investigation into the factors influencing the academic journey of undergraduate students in Bangladesh.

2. Literature Review

Academic performance is a crucial aspect of a student's educational journey, and understanding the factors that influence it is of great importance. This literature review aims to pro- vide insights into the factors affecting Bangladeshi university students' academic performance, with a focus on recent re- search employing machine learning approaches. [6] investigates the use of machine learning techniques to predict the academic performance of undergraduate students. The study employs various features such as attendance, assignment scores, and exam results to develop predictive models. Key findings from this paper include the potential of machine learning models to predict student performance accurately and the importance of features like attendance in predictive accuracy. They have used the Information Gain (InfoGain) algorithm to select the most effective features and ensemble methods to compare the accuracy with more robust algorithms, including Logit Boost, Vote, and Bagging. The algorithms were evaluated based on the performance evaluation metrics such as accuracy, precision, recall, F-measure, and ROC curve, and then validated using 10- folds cross-validation. Furthermore, the paper underscores the significance of timely interventions for students at risk of poor academic performance. In the study [7], the authors focus on developing a machine learning-based model to predict academic performance in a Bangladeshi context. The research considers multiple factors, including students' prior academic records, socioeconomic background, and class attendance. The paper highlights the potential of machine learning in providing early warnings about students who may face challenges in their academic journey. Most of the analyses were predominantly performed using the basic logistic regression (LR) model. As an alternative, they used the advanced machine learning (ML) approaches for detecting significant risk factors and to predict the prevalence of stress among Bangladeshi university students. It emphasizes the need for personalized interventions and support systems to improve academic outcomes.

We have reviewed a paper which was not specifically focusing on machine learning, this paper [8] provides valuable insights into factors affecting students' motivation and, consequently, their academic performance. It explores the perceptions of Bangladeshi EFL (English as a Foreign Language) students regarding teaching factors that impact their motivation. The findings highlight the importance of pedagogical approaches, teacher-student relationships, and classroom dynamics in influencing students' motivation and, subsequently, their academic performance.

Overall, the reviewed papers suggest that machine learning approaches can be effective tools for predicting and improving the academic performance of Bangladeshi university students. Key factors influencing academic performance include attendance, socioeconomic background, prior academic records, and teaching quality. These findings emphasize the need for early intervention and personalized support systems to address challenges and enhance academic outcomes. Additionally, pedagogical approaches and motivation play significant roles in shaping students' performance and should be considered in educational strategies.

This review provides a foundation for understanding the factors affecting academic performance in the

context of Bangladeshi university students and the potential of machine learning techniques in addressing these challenges. Further research in this area can lead to more tailored interventions.

2. Methodology

We have done different experiments to solve our problem.
They are: KNN, Naïve Bayes, Logistic Regression, SVM, Decision Tree

2.1 KNN

K-Nearest Neighbors (K-NN), also called lazy learner [1] is a simple but effective supervised algorithm where an object is assigned to the most common class based on the neighbors' votes, as depicted in Fig. 1. It uses the distance to predict the class, and once the data is normalized, better accuracy is gained. It has many advantages: it can be implemented quickly; it supports nonparametric analysis and the time required to build the model is based on k. However, if the data has lots of outliers or missing data then this algorithm is ineffective [2].

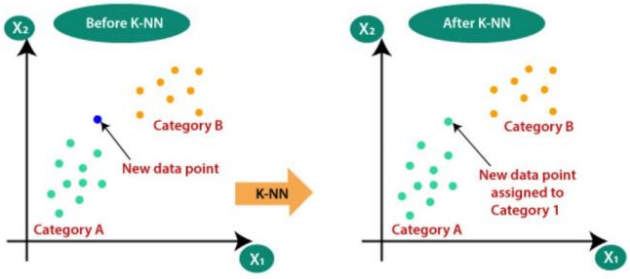

Fig-1: Classifying New Data Using K-NN Algorithm

```
y_pred = knn_model.predict(X_test)
```

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
1	0.16	0.27	0.20	15
2	0.20	0.27	0.23	15
3	0.25	0.19	0.22	26
4	0.10	0.06	0.08	16
5	0.33	0.25	0.29	12
accuracy			0.20	84
macro avg	0.21	0.21	0.20	84
weighted avg	0.21	0.20	0.20	84

Fig-2: Test result of KNN model

In the project, we employed the K-Nearest Neighbors (KNN) algorithm to predict students' academic performance in a machine learning context. Our methodology encompasses data collection, feature selection, and engineering, as well as the application of the KNN algorithm. We carefully split the data into training and testing sets, and evaluate the model's performance using relevant metrics.

The KNN model has been trained with the five nearest neighbor nodes. The accuracy of the KNN model is 20% which is lower than the expectation. One of the reasons that the accuracy is low because the dataset was collected from different university of Bangladesh and each university have different grading policy. Results reveal the effectiveness of KNN in predicting student performance.

2.2 Naïve Bayes

Naïve Bayes is one of the simplest probabilistic supervised algorithms based on the Bayes' theorem [3]. This classifier predicts each feature's probability and assigns it to its belonging class [4]. Simplicity, scalability, and applicability to real-world problems are common characteristics of the Naïve Bayes algorithm: see Figs. 3 and 4.

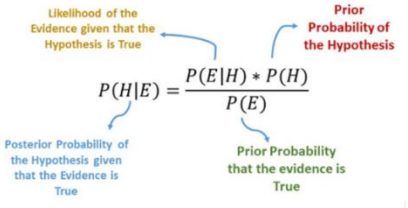

Fig-3: Naïve Bayes Equation

```
y_pred = nb_model.predict(X_test)
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
1	0.17	0.13	0.15	15
2	0.22	0.13	0.17	15
3	0.39	0.35	0.37	26
4	0.33	0.12	0.18	16
5	0.26	0.75	0.39	12
accuracy			0.29	84
macro avg	0.28	0.30	0.25	84
weighted avg	0.29	0.29	0.26	84

2.3 Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. In statistic, the logistic model is a model that models the probability of an event taking place by having the log odds for the event be a linear combination of one or more independent variables.

```
y_pred = lg_model.predict(X_test)
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
1	0.06	0.07	0.06	15
2	0.10	0.07	0.08	15
3	0.38	0.35	0.36	26
4	0.00	0.00	0.00	16
5	0.17	0.42	0.24	12
accuracy			0.19	84
macro avg	0.14	0.18	0.15	84
weighted avg	0.17	0.19	0.17	84

2.4 SVM

Support Vector Machine (SVM), depicted in Fig. 4, is a supervised learning method that classifies the features based on their categories by dividing them as wide as possible from consequences [5]. It can be used in small databases and less time is required to build the model than other classification algorithms. SVM can be used in linear and non-linear classification by using the “kernel trick” to map the inputs into high dimensional space. The accuracy of SVM is 19% which is comparatively low.

2.5 Decision Tree

In our problem statement methods, we got highest train set accuracy in Decision Tree Classifier in Fig-7&8

```
y_pred = svm_model.predict(X_test)
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
1	0.22	0.33	0.26	15
2	0.18	0.27	0.22	15
3	0.19	0.12	0.14	26
4	0.18	0.12	0.15	16
5	0.17	0.17	0.17	12
accuracy			0.19	84
macro avg	0.19	0.20	0.19	84
weighted avg	0.19	0.19	0.18	84

	precision	recall	f1-score	support
1	0.12	0.20	0.15	15
2	0.17	0.20	0.18	15
3	0.27	0.23	0.25	26
4	0.42	0.31	0.36	16
5	0.29	0.17	0.21	12
accuracy			0.23	84
macro avg	0.25	0.22	0.23	84
weighted avg	0.26	0.23	0.23	84

Fig-8: Decision Tree outcomes

```
Classifier:  DecisionTreeClassifier()
Train Accuracy:  0.819
Test Accuracy:  0.179
```

Fig-7: Train & test Accuracy

3 Result

Pairplot: The bellow Fig-9 reflects the scatterplot distribution of the combination of each attribute. The diagonal kdeplot reflects the distribution of each attributes marking with respect to their CGPA.

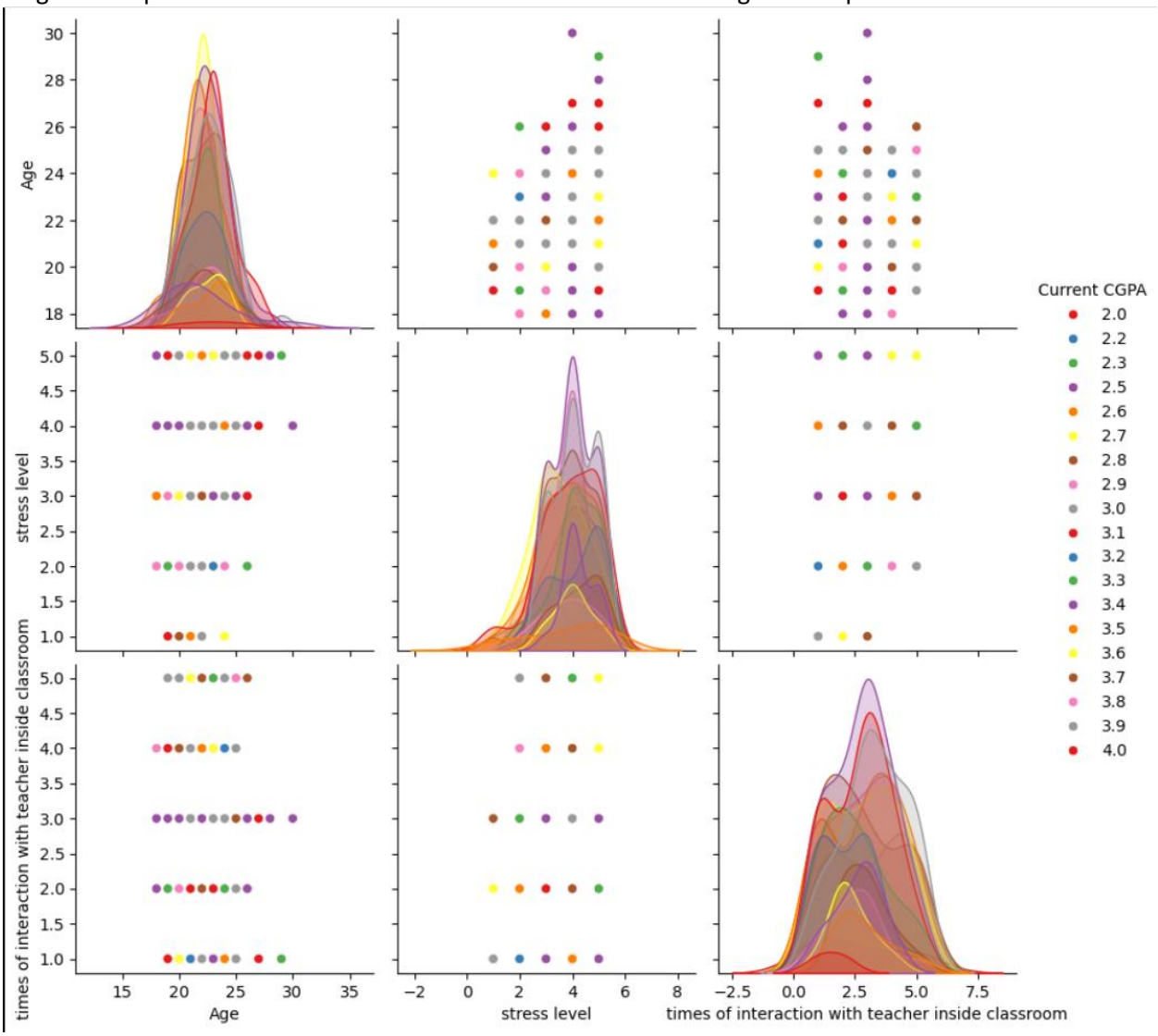

Fig-9: Students CGPA based on each features(characteristics)
The graph shows how students' behaviors and habits effect their result performance.

Histogram**:** In the below histogram in Fig-9, X-axis defines the frequency number of a student's activity on a particular attribute and Y-axis defines the count**.**

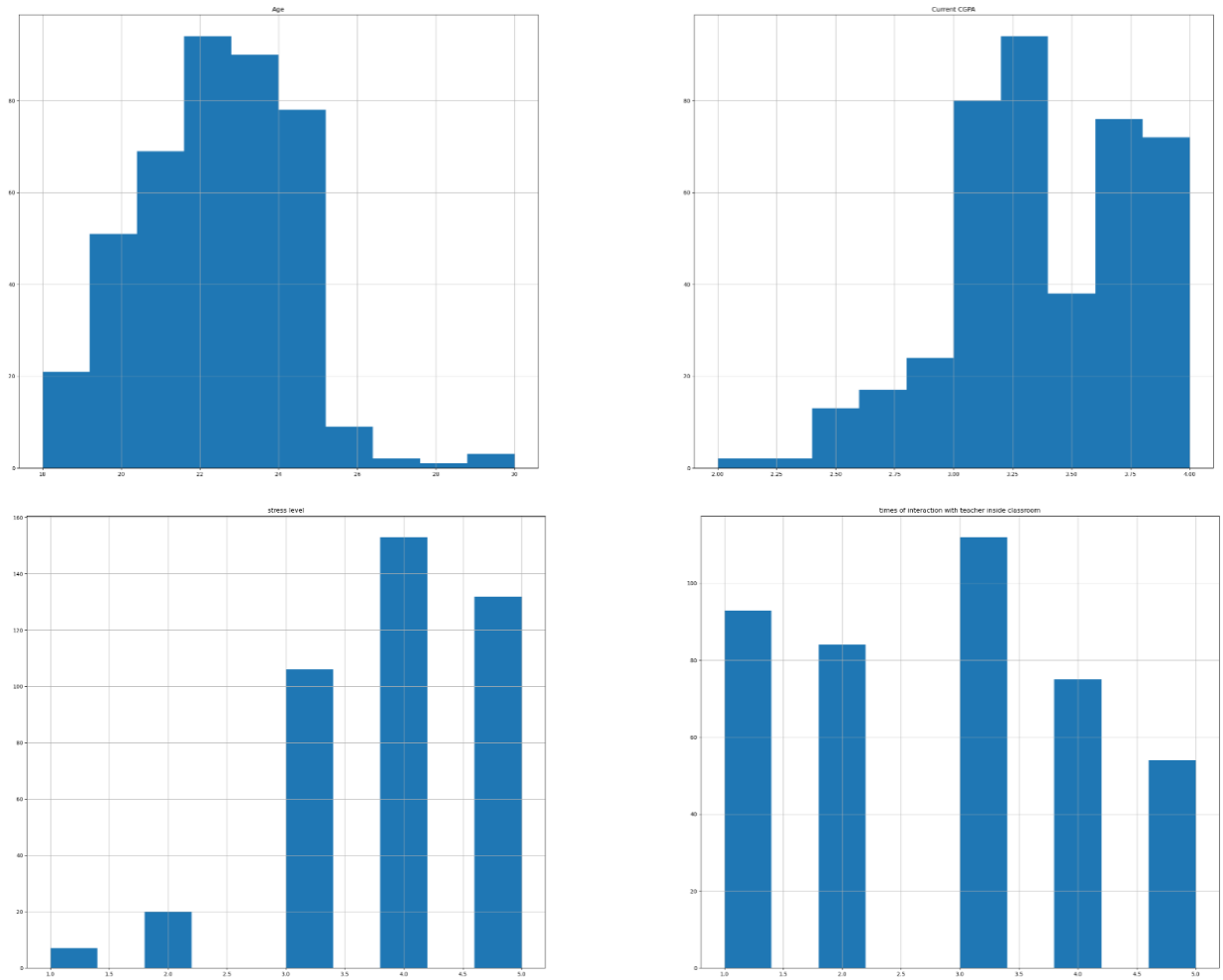

Fig-19: Histogram distribution in discreate format

The histogram shows that -*

- *Majority students stress level is very high.*
- *Majority of the students are less eager to interact with teacher during class.*

PairGrid plot: From the below pair plot in Fig-10, it is quite a lot clear to see that Hige and Low CGPA students are distinguishable in every combination of pair plot but Medium level students are scattered in almost everywhere and mixed up.

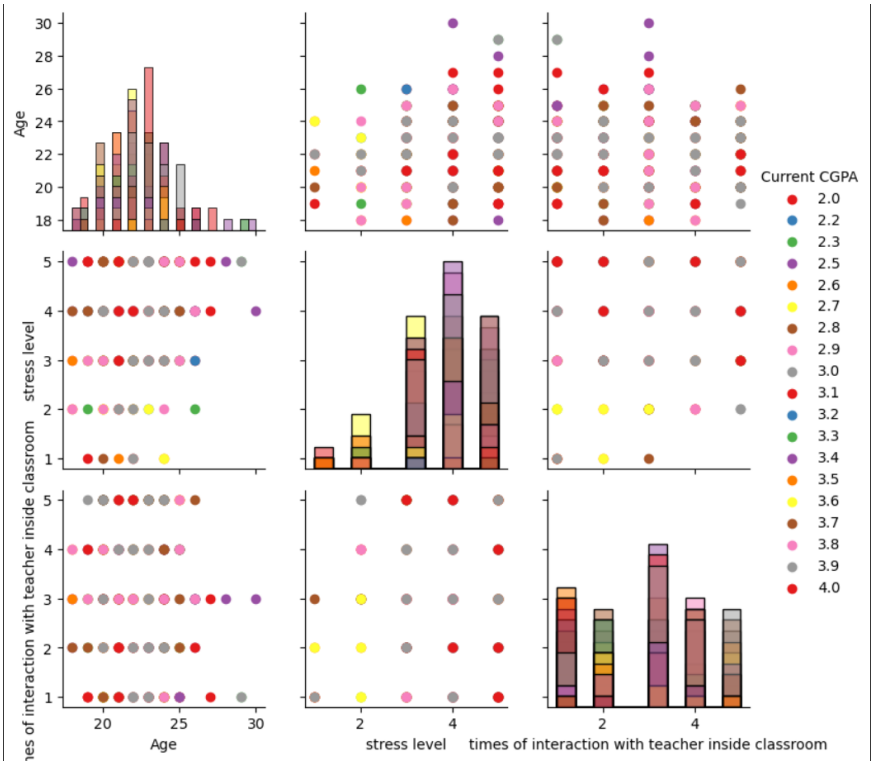

Fig-10: Each 2 Features and CGPA

Swarmplot:

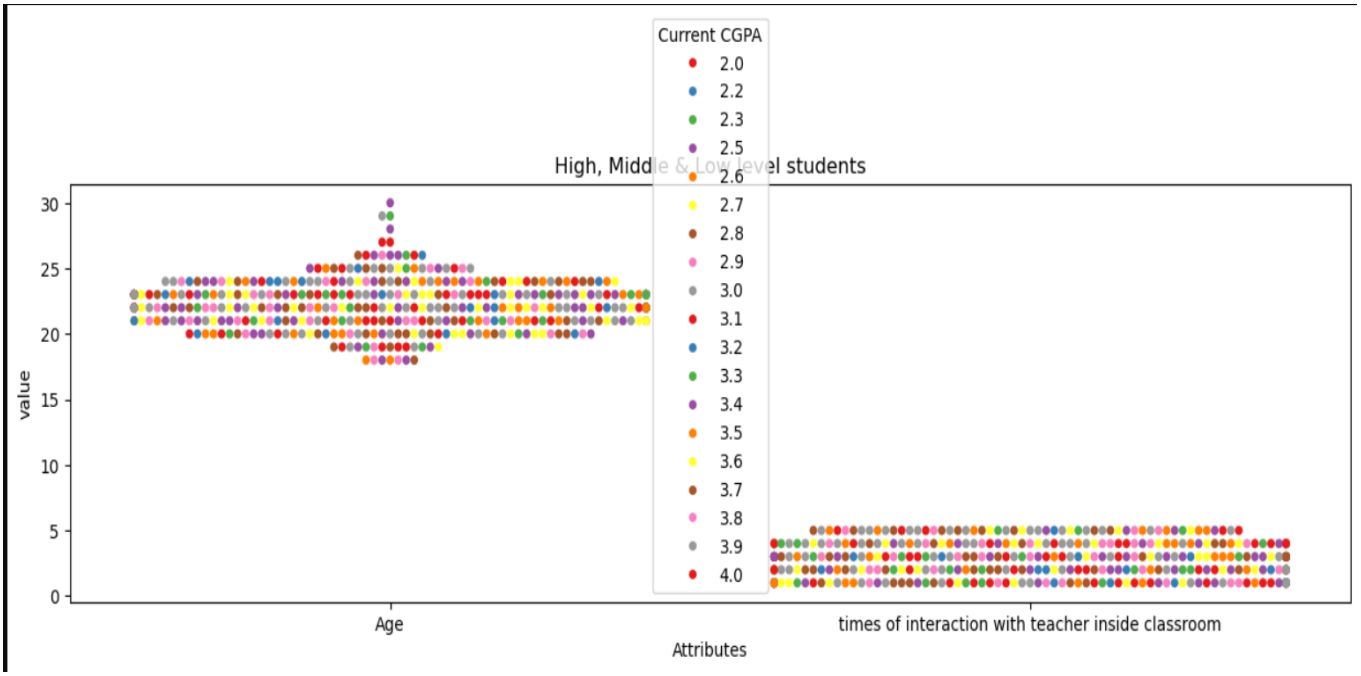

Fig-11: Swarmplot of age and times of interaction with teacher w.r.t CGPA

From the above swarm plot, it is noticeable that – how age and times of interaction with teacher affect students' performance on result.

Algorithm	KNN	Naïve Bayes	Log. Reg.	SVM	Decision Tree
Accuracy	20%	29%	19%	19%	23%

So far we have seen, for our problem, we get best result from Naïve Bayes model.

4 Discussion

We wanted to work on our project mainly to help students showing some statistical result that how students' performance is affected over various attributes, features and characteristics. Our problem is very uncommon that no so many works or papers have published by now. Furthermore, we added some unique levels in our dataset that distinguish from existing works so far.
The experimental results revealed that students who interact more with teachers have good CGPA and male students tends to get highest cgpa but most of the female students got average cgpa .

5. Limitations:

- Diverse grading policies across universities hinder model generalizability.
- Limited data availability and quality may affect model performance.
- Feature engineering and selection could be further optimized.

6. Conclusion:

This study explored the application of machine learning algorithms to predict the academic performance of Bangladeshi university students. Five different algorithms were experimented with: K-Nearest Neighbors (KNN), Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Decision Tree. Key findings include: KNN exhibited limited success with an accuracy of 20% due to diverse grading policies, Naïve Bayes yielded a higher accuracy of 29%, Logistic Regression provided moderate results, SVM achieved an accuracy of 19%, and Decision Tree demonstrated the highest accuracy. Future research directions include: exploring more relevant features, combining algorithms through ensemble techniques, standardizing data, implementing student interventions, and conducting longitudinal analysis.

References:

[1]Wang, X. (2011, July). A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality. In The 2011 International Joint Conference on Neural Networks (pp. 1293–1299). IEEE

[2]Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(2), e1289.

[3] Yang, S. (2019). An Introduction to Naïve Bayes Classifier: From theory to practice, learn underlying principles of Naïve Bayes. from https://towardsdatascience.com/ introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf

[4] Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 143, 1–15

[5]Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 143, 1–15.

[6] "Predicting Academic Performance of Undergraduate Students: A Machine Learning Approach" (https://link.springer.com/article/10.1007/s10639-023-11700-0)

[7] "A Machine Learning Approach for Predicting Academic Performance in Bangladesh" (https://link.springer.com/article/10.1186/s41043-021-00276-5)

[8] "Teaching Factors That Affect Students' Learning Motivation: Bangladeshi EFL Students' Perceptions" (https://www.researchgate.net/publication/356499487_TEACHING_FACTORS_THAT_AFFECT_STUDENTS'_LEARNING_MOTIVATION_BANGLADESHI_EFL_STUDENTS'_PERCEPTIONS)