Turkish Journal of Computer and Mathematics Education Vol.13 No. 01(2022), 149-163
Research Article

An Extensive Analysis on Computing Students' Academic Performance in Online Environment using Decision Tree

Nyme Ahmed[a], Rifat-Ibn-Alam[b], Md. Golam Ahsan Akib[c], Syed Nafiul Shefat[d], Dr. Dip Nandi[e]

[a, b, c, d, e] Department of Computer Science, American International University- Bangladesh

Abstract: Maintaining the continuation of study is a vital element as it holds students' concentration to achieve what the external world left them to explore. COVID-19 acts as some kind of a barrier in front of this continuation of study. Online education lifts this barrier and gives the students a free open road to roam around. But to be sure that students are maintaining the pace and continuing their sturdy approach to achieve their goals, there has to be some monitoring. Educational Data Mining (EDM) is a new discipline that arose from applying data mining techniques to educational data. EDM can be used to understand students and their learning environments better, improve teaching support, and make decisions in educational systems. The main objective of this paper is to analyze the factors that have a profound impact on students' academic performance while conducting EDM applications, more specifically using decision trees. Four distinct datasets are derived from X University students' academic marks in four different undergraduate program courses during an online semester. The decision trees' knowledge reveals critical factors in analyzing students' performance. The findings of this paper will help educators develop new strategies to cope with various challenges and ultimately the betterment of education.
Keywords: Data Mining, Decision Tree, Performance Analysis, Online Education, COVID-19.

1. Introduction

COVID-19 struck the world unexpectedly, putting human life in quarantine. It compelled the traditional education model to close its doors, and it would have remained closed had online education not been introduced. The COVID-19 has a detrimental effect on all fields of global education. It has enforced a global lockdown, with a devastating influence on the kids' lives. At first, teachers and students were perplexed and unsure how to deal with the unexpected scenario that resulted in the suspension of educational activities. It laid the groundwork for educational institutions to design and adopt virtual learning [1]. However, online education was a novel experience for the majority of pupils. Adapting to a new environment takes time. However, online education has several advantages. E-learning eliminates the need for paper documentation, removes the need for transportation, and is less expensive, resulting in significant energy savings.

Analyzing the students' academic performance is essential as the pandemic forced the education sector to confine all its activities virtually. Therefore, to maintain the quality of education, it is crucial to analyze academic performance and take necessary actions to better education. This paper is a part of a larger research, where a comparison is made on students' online and on-campus academic performance. This research aims to aid in identifying the aspects that contribute significantly to students' academic performance in virtual classes. It can be used as a guide while making major educational decisions, particularly for computing students. Four separate programming courses were chosen for the study, covering students from their first year through their final year of the undergraduate degree.

The paper is distributed into five sections, which are arranged in the following order: the introduction is in section 1, followed by a related works in section 2, data collection and analysis in section 3, section 4 contains the results & discussions of this paper, and lastly, section 5 consists of the conclusion.

2. Related Works

Analyzing student academic performance is becoming an essential factor in improving academic instruction, assisting students as they study, and giving tutors more options when preparing their students. In recent years, numerous works on this topic have been conducted. Several literature reviews are discovered in this section that analyzed student academic performance from a variety of perspectives.

Data mining techniques are frequently used in educational settings to investigate and evaluate student performance. Educational data mining is used to assess educational or academic behavior to implement essential changes to improve

the quality of education. Students' success is contingent upon various factors, including their personal, economic, social, and environmental actions [2]. Numerous educational institutions use the experiment results to establish a pattern of students' behavior and devise a new method for overcoming academic performance barriers and enhancing educational quality [3]. Educational Data Mining (EDM) lends a helpful hand in adapting courses to match students' needs and faculty capabilities. Students are zealous about passing tests and completing assigned tasks [4]. Many EDM approaches are available for classifying students according to their overall performance, including Naive Bayes, K-Nearest Neighbors, Decision Trees, and the Apriori algorithm [5].

In the virtual classes [6], a downfall was noticed in the interaction between students and teachers. Surprisingly, the interaction between the classmates was much more stable. Thus, students might lack a particular topic, which the teacher could quickly solve, but it was not the case, as the students did not ask the question in the first place. This inconsistency in the students' participation is an influential factor in students' academic performance. Students face various problems in virtual classes. Some of them include inappropriate study environment, insufficient study material, internet connection problems, load shedding, and so on. These issues deeply affect the students' academic performance. On top of that, they face several physical and mental hazards such as weak eyesight, overweight, sleep deprivation, psychological issues, and many more [7]. [8] used many classification algorithms such as C5.0, J48, CART, KNN, SVM, Naive Bayes, and Random Forest on three distinct datasets from high school, college, and virtual classrooms where C5.0 and Random Forest outperformed the rest of the techniques in all three datasets. Three distinct classification techniques were used by [9] to identify the various factors that affect student performance: Multi-layer Perceptron (MLP), Boosting algorithm and Bagging. The dataset was obtained from the UCI online repository, which contains secondary school students' performance. The features are classified into three categories: academic background, personal attributes, and economic background to analyze performance. The findings indicate that only students' economic background affects their performance. MLP classifiers achieved 72% accuracy with economic background attributes, Bagging classifiers achieved 88% accuracy, and MultiBoost classifiers achieved 86% accuracy. The authors of [10] performed a classification technique on a dataset obtained from the University of California, Irvine website using Naive Bayes and K-means. Moreover, clustering was used to analyze students' academic achievement, and 98.866 percent accuracy was achieved in forecasting students' academic performance.

A study was conducted [11] on the impact of detachment from the campus on students' mental health during the global pandemic. A survey of young students was conducted to gauge their feelings about their mental health in the aftermath of the Covid-19. Students have suffered significant emotional trauma because of their absence from regular class. [12] analyzed numerous factors to determine the most relevant variables affecting students' academic achievement in online classrooms, specifically MOOCs. When examining the students' performance, the authors discovered the best factors connected to the MOOC's exercises. [13] used three different decision tree algorithms (ID3, C4.5, and CART). The CART algorithm outperformed C4.5 and ID3 in terms of accuracy. Furthermore, this study showed that students' academic performance was influenced by various qualitative factors such as students' parents' qualifications, living location, economic status, friends & relatives' support, resource accessibility, attendance, and academic results. [14] generated a set of rules and identified the dominant factor to analyze students' performance on online platforms using the Decision Tree (J48) classification technique. From "X-University" and Microsoft Teams software, the authors extracted 589 instances of seven courses. The decision tree revealed that 'Final term' and 'Mid-term' were the most significant attributes for analyzing students' performance. At the same time, other factors such as quiz, gender, and attendance had a negligible impact. Decision trees and clustering algorithms like K-Means clustering were used by [15] to analyze student performance. The dataset was compiled from various undergraduate courses at a single university in Pakistan. Three different types of decision trees were used with an accuracy range of 60.58-69.23 percent. [16] analyzed students' learning behavior in virtual learning to evaluate the students' performance. They gathered students' high-dimensional behavioral characteristic data and performed correlation analysis. Then they developed the performance evaluation model using a decision tree and achieved an accuracy of 88%. According to them, completing the video task point is the most important factor in the students' performance.

Student academic data and campus behavior data were used by [17] to analyze students' patterns and correlations for the improvement in the teaching activities and teaching management. They used the global search advantage of the genetic algorithm to develop a GABP hybrid prediction model. The data validation results showed that Recall reaches 95%, F1 is around 86 percent, and the accuracy of the algorithm prediction results is significantly improved. [18] developed a model to predict the students' results and identify their shortcomings to improve the quality of education. They gathered the data from the university's dataset and surveyed it filled out by students. These include ECA,

programming skills, assessment marks, assignment marks, attendance, GPA, and so on. They used the WEKA tool to develop three decision tree models (J48, REPTree, and Hoeffding Trees), where J48 outperformed the other two algorithms. The authors of [19] used Decision Tree to improve students' performance, and at that point, they were able to predict the appropriated scholarly achievement in each major. They analyzed the data from 1200 students to come up with their conclusions. Accuracy and error rate are two of the metrics used to evaluate the framework. This method yields a 95.55% accuracy rate and a 4.55% error rate, respectively.

Decision Trees are the simplest to express and understand. Because most people are familiar with hierarchical trees, a straightforward illustration will assist in conveying the findings. If any dataset needs to be classified, a decision tree is generally the best place to begin. It will provide an excellent overview and will aid in the comprehension of the classification. It will provide instructors with a comprehensive understanding of the aspects that influence students' performance. Therefore, this research will be conducted using the decision tree classification algorithm.

3. Data Collection and Analysis

The datasets are compiled and analyzed briefly in this section. Four distinct datasets are obtained from X University students' academic marks of four different courses of the Bachelor's program in the department of Computer Science and Engineering in one semester conducted virtually. The courses are Introduction to Programming (IP), Object Oriented Programming 1 (OOP1), Object Oriented Programming 2 (OOP2), and Web Technologies (WT), all of which are taught throughout the four years undergraduate program. These four courses were chosen primarily to provide an overview of all students' performance from the first to the final years. Freshmen students are taught the Introduced to Programming course. After completing this course, they can enroll in the second-year programming course, Object Oriented Programming 1 (OOP1). Third-year students take Object Oriented Programming 2 (OOP2) before completing the first two programming courses. Finally, Web Technologies (WT) is available to the final year students to meet the preceding programming courses. A total of 273 instances are available in these four datasets. The overall marks gained by the students is consist of the marks from both mid and final term where both of them contribute equally. The students are classified into three categories (Good, Average, Below Average) considering their marks in the courses stated in the following table.

Table 1. Category of Students

Overall Marks	80-100	60-79	40-59
Category	Good	Average	Below Average

A. Dataset 1

This dataset is gathered from the students' marks from the Introduction to Programming (IP) course. This dataset contains 16 factors with a population size of 40, which are stated briefly in the following table.

Table 2. Description of Dataset 1

Course: Introduction to Programming (IP)					
Attribute	**Type**	**Summary**			
		Mean	**Standard Deviation**	**Max**	**Min**
MidAttendance(10%)	Numeric	10	0	10	10
MidAssignment(10%)	Numeric	9.75	1.58	10	0
MidPerformance(10%)	Numeric	8.8	0.65	10	8
MidQuiz(30%)	Numeric	21.5	3.19	29	11
MidAssessment(40%)	Numeric	28.58	4.23	37	20
MidMarks(100%)	Numeric	78.63	6.9	94	65
FinalAttendance(10%)	Numeric	9.88	0.52	10	7
FinalAssignment(10%)	Numeric	9.93	0.47	10	7
FinalPerformance(10%)	Numeric	8.35	1.33	10	4
FinalQuiz(25%)	Numeric	19.68	2.76	25	15
FinalAssessment(20%)	Numeric	11.05	3.97	18	4
FinalViva(25%)	Numeric	18.10	5.25	25	0
FinalMarks(100%)	Numeric	76.75	9.8	96	48
OverallMarks(100%)	Numeric	77.68	7.34	95	60

StudentLevel	Nominal	Good (80 <= OverallMarks <= 100) = 17 Average (60 <= OverallMarks < 80) = 23
Gender	Nominal	Male (M) = 30, Female (F) = 10

B. Dataset 2

This dataset is compiled from the marks of students enrolled in the Object Oriented Programming 1 (OOP1) course. This dataset has 10 characteristics with a population size of 78, summarized in the table below.

Table 3. Description of Dataset 2

Course: Object Oriented Programming 1 (OOP1)					
Attribute	**Type**	**Summary**			
		Mean	**Standard Deviation**	**Max**	**Min**
MidPerformance(10%)	Numeric	8.9	2.47	10	0
MidQuiz(20%)	Numeric	15.06	2.45	19	3
MidViva(20%)	Numeric	11.95	3.83	20	0
MidMarks(50%)	Numeric	35.89	5.34	47	20
FinalProject(20%)	Numeric	11.86	6.35	20	0
FinalViva(30%)	Numeric	17.82	7.1	30	10
FinalMarks(50%)	Numeric	29.68	11.96	50	10
OverallMarks(100%)	Numeric	65.56	14.9	95	42
StudentLevel	Nominal	Good (80<= OverallMarks <= 100) = 19 Average (60 <= OverallMarks < 80) = 23 Below Average (40 <= OverallMarks <60) = 36			
Gender	Nominal	Male (M) = 61, Female (F) = 17			

C. Dataset 3

This dataset is comprised of student marks for the Object Oriented Programming 2 (OOP2) course. This dataset has 15 characteristics and population size of 111 individuals, and the table below summarizes these characteristics.

Table 4. Description of Dataset 3

Course: Object Oriented Programming 2 (OOP2)					
Attribute	**Type**	**Summary**			
		Mean	**Standard Deviation**	**Max**	**Min**
MidAttendance(10%)	Numeric	8.94	1.53	10	3
MidLabPerformance(20%)	Numeric	13.78	3.93	19	0
MidLabExam(20%)	Numeric	14.6	3.55	19	0
MidViva(20%)	Numeric	12.31	5.39	20	0
MidQuiz(30%)	Numeric	22.15	2.51	29	15
MidMarks(100%)	Numeric	71.78	12.58	95	38
FinalLabPerformance(10%)	Numeric	5.15	3.47	10	0
FinalLabExam(20%)	Numeric	9.88	6.87	20	0
FinalQuiz(20%)	Numeric	17.03	1.9	20	8
FinalViva(20%)	Numeric	12.86	5.63	20	0
FinalProject(30%)	Numeric	20.02	6.89	29	0
FinalMarks(100%)	Numeric	64.95	20.23	96	16
OverallMarks(100%)	Numeric	68.72	15.18	93	41
StudentLevel	Nominal	Good (80<= OverallMarks <= 100) = 31 Average (60 <= OverallMarks < 80) = 47 Below Average (40 <= OverallMarks <60) = 33			
Gender	Nominal	Male (M) = 86, Female (F) = 25			

D. Dataset 4

This dataset contains the marks earned by students in the Web Technologies (WT) course. This dataset has 15 factors and a population of 44 participants, and these attributes are summarized in the table below.

Table 5. Description of Dataset 4

Course: Web Technologies (WT)					
Attribute	**Type**	**Summary**			
		Mean	**Standard Deviation**	**Max**	**Min**
MidLabPerformance(10%)	Numeric	7.39	2.1	10	0
MidQuiz(30%)	Numeric	16.55	3.9	24	6
MidProject(30%)	Numeric	20.02	5.56	30	5
MidViva(20%)	Numeric	8.5	3.23	14	0
MidReport(10%)	Numeric	6.23	2.72	9	0
MidMarks(100%)	Numeric	58.68	10.42	82	40
FinalLabPerformance(20%)	Numeric	12.32	4.18	20	0
FinalQuiz(20%)	Numeric	13.73	2.58	18	5
FinalProject(30%)	Numeric	17.45	7.93	30	0
FinalViva(20%)	Numeric	12.32	4.13	19	0
FinalReport(10%)	Numeric	7.11	2.36	9	0
FinalMarks(100%)	Numeric	62.95	16.34	94	25
OverallMarks(100%)	Numeric	61.36	11.59	88	40
StudentLevel	Nominal	Good (80<= OverallMarks <= 100) = 2 Average (60 <= OverallMarks < 80) = 22 Below Average (40 <= OverallMarks <60) = 20			
Gender	Nominal	Male (M) = 34, Female (F) = 10			

4. Result and Discussion

In this section, four different Decision Trees are generated from these four datasets using the CART algorithm in Jupyter Notebook. As mentioned earlier in table 1 that students are classified into three categories based on their overall marks. Therefore, we are considering 'StudentLevel' as the class attribute. Hence, some attributes [MidMarks(100%), FinalMarks(100%), OverallMarks(100%)] are reduced while analyzing further as the gist of these attributes are already present in the 'StudentLevel' attribute. Moreover, the nominal attribute 'Gender' attribute does not have any influence on the students, so it is not taken into consideration. The trees and their corresponding rules are illustrated in detail.

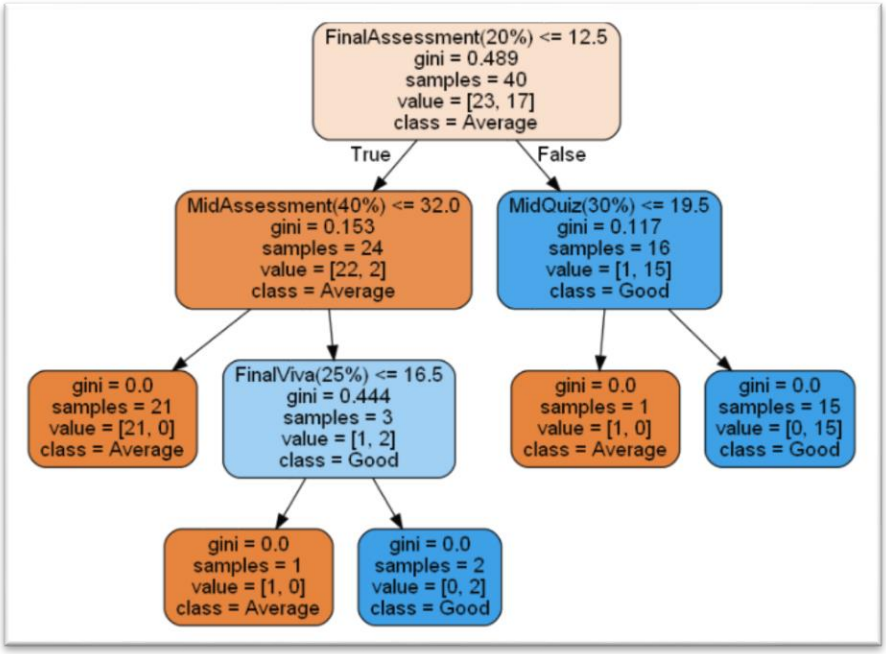

Fig.1. Decision Tree derived from Dataset 1 (IP)

Table 6. Derived set of rules from Dataset 1 (IP)

Rules	Description	Samples
R1	If [FinalAssessment(20%) <= 12.5 && MidAssessment(40%) <= 32] then StudentLevel = 'Average'	21
R2	If [FinalAssessment(20%) <= 12.5 && MidAssessment(40%) > 32 && FinalViva(25%) <= 16.5] then StudentLevel = 'Average'	1
R3	If [FinalAssessment(20%) <= 12.5 && MidAssessment(40%) > 32 && FinalViva(25%) > 16.5] then StudentLevel = 'Good	2
R4	If [FinalAssessment(20%) > 12.5 && MidQuiz(30%) <= 19.5] then StudentLevel = 'Average'	1
R5	If [FinalAssessment(20%) > 12.5 && MidQuiz(30%) > 19.5] then StudentLevel = 'Good'	15
Total Samples		40

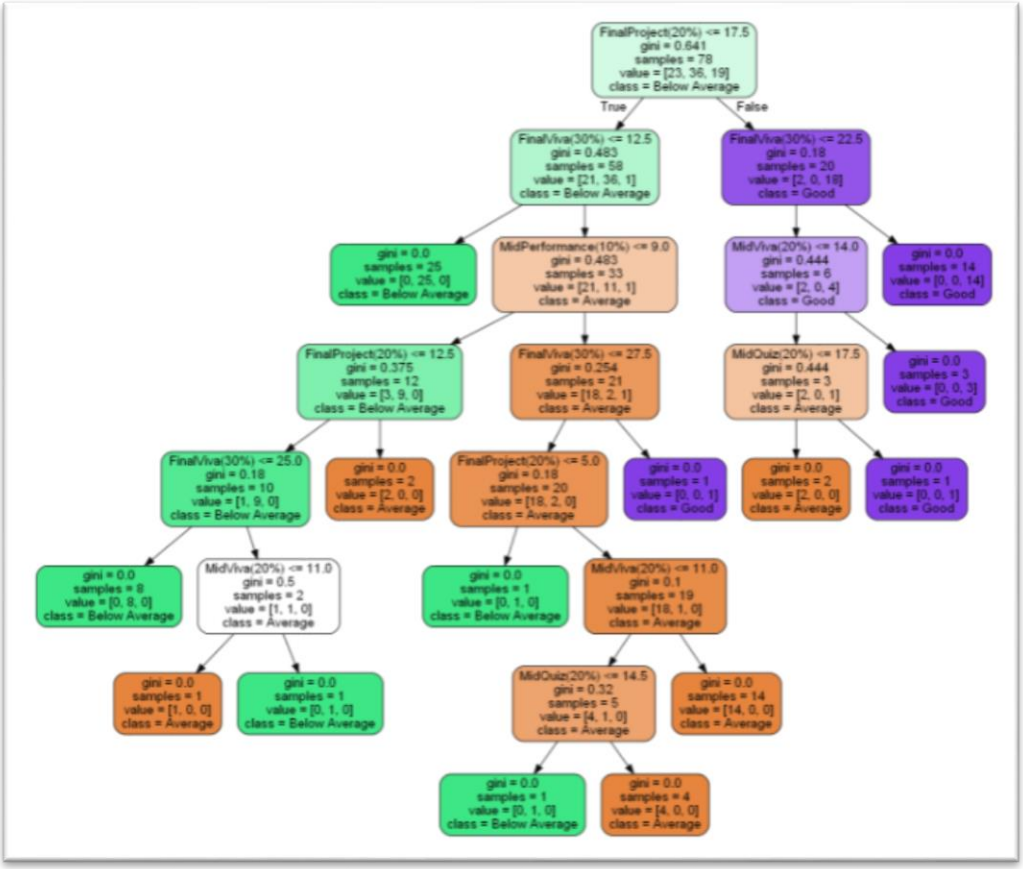

Fig.2. Decision Tree derived from Dataset 2 (OOP1)

Table 7. Derived set of rules from Dataset 2 (OOP1)

Rules	Description	Samples
R1	If [FinalProject(20%) <= 17.5 && FinalViva(30%) <= 12.5] then StudentLevel = 'Below Average'	25
R2	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) <= 9 && FinalProject(20%) <= 12.5 && FinalViva(30%) <= 25] then StudentLevel = 'Below Average'	8
R3	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) <= 9 && FinalProject(20%) <= 12.5 &&	1

	FinalViva(30%) > 25 && MidViva(20%) <= 11] then StudentLevel = 'Average'	
R4	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) <= 9 && FinalProject(20%) <= 12.5 && FinalViva(30%) > 25 && MidViva(20%) > 11] then StudentLevel = 'Below Average'	1
R5	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) <= 9 && FinalProject(20%) > 12.5] then StudentLevel = 'Average'	2
R6	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) > 9 && FinalViva(30%) <= 27.5 && FinalProject(20%) <= 5] then StudentLevel = 'Below Average'	1
R7	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) > 9 && FinalViva(30%) <= 27.5 && FinalProject(20%) > 5 && MidViva(20%) <= 11 && MidQuiz(20%) <= 14.5] then StudentLevel = 'Below Average'	1
R8	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) > 9 && FinalViva(30%) <= 27.5 && FinalProject(20%) > 5 && MidViva(20%) <= 11 && MidQuiz(20%) > 14.5] then StudentLevel = 'Average'	4
R9	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) > 9 && FinalViva(30%) <= 27.5 && FinalProject(20%) > 5 && MidViva(20%) > 11] then StudentLevel = 'Average'	14
R10	If [FinalProject(20%) <= 17.5 && FinalViva(30%) > 12.5 && MidPerformance(10%) > 9 && FinalViva(30%) > 27.5] then StudentLevel = 'Good'	1
R11	If [FinalProject(20%) >17.5 && FinalViva(30%) <= 22.5 && MidViva(20%) <= 14 && MidQuiz(20%) <= 17.5] then StudentLevel = 'Average'	2
R12	If [FinalProject(20%) >17.5 && FinalViva(30%) <= 22.5 && MidViva(20%) <= 14 && MidQuiz(20%) > 17.5] then StudentLevel = 'Good'	1
R13	If [FinalProject(20%) >17.5 && FinalViva(30%) <= 22.5 && MidViva(20%) > 14] then StudentLevel = 'Good'	3
R14	If [FinalProject(20%) >17.5 && FinalViva(30%) > 22.5] then StudentLevel = 'Good'	14
Total Samples		78

Fig.3. Decision Tree derived from Dataset 3 (OOP2)

Table 8. Derived set of rules from Dataset 3 (OOP2)

Rules	Description	Samples
R1	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) <= 7.5 && FinalProject(30%) <= 24.5] then StudentLevel = 'Below Average'	23
R2	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) <= 7.5 && FinalProject(30%) > 24.5] then StudentLevel = 'Average'	3
R3	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) <= 5] then StudentLevel = 'Below Average'	4
R4	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) <= 1.5 && MidLabExam(20%) <= 14.5] then StudentLevel = 'Below Average'	4
R5	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) <= 1.5 && MidLabExam(20%) > 14.5] then StudentLevel = 'Average'	3
R6	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) > 1.5 && FinalProject(30%) <= 26 && MidLabPerformance(20%) <= 16.5 && MidViva(20%) <= 9 && FinalProject(30%) <= 18.5] then StudentLevel = 'Below Average'	1
R7	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) > 1.5 &&	4

	FinalProject(30%) <= 26 && MidLabPerformance(20%) <= 16.5 && MidViva(20%) <= 9 && FinalProject(30%) > 18.5] then StudentLevel = 'Average'	
R8	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) > 1.5 && FinalProject(30%) <= 26 && MidLabPerformance(20%) <= 16.5 && MidViva(20%) > 9] then StudentLevel = 'Average'	31
R9	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) > 1.5 && FinalProject(30%) <= 26 && MidLabPerformance(20%) > 16.5 && FinalLabPerformance(10%) <= 6.5] then StudentLevel = 'Average'	2
R10	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) > 1.5 && FinalProject(30%) <= 26 && MidLabPerformance(20%) > 16.5 && FinalLabPerformance(10%) > 6.5] then StudentLevel = 'Good'	1
R11	If [FinalLabExam(20%) <= 15.5 && MidViva(20%) > 7.5 && FinalViva(20%) > 5 && FinalLabPerformance(10%) > 1.5 && FinalProject(30%) > 26] then StudentLevel = 'Good'	1
R12	If [FinalLabExam(20%) > 15.5 && FinalProject(30%) <= 22.5 && MidLabExam(20%) <= 18 && FinalQuiz(20%) <= 14] then StudentLevel = 'Below Average'	1
R13	If [FinalLabExam(20%) > 15.5 && FinalProject(30%) <= 22.5 && MidLabExam(20%) <= 18 && FinalQuiz(20%) > 14] then StudenLevel = 'Average'	4
R14	If [FinalLabExam(20%) > 15.5 && FinalProject(30%) <= 22.5 && MidLabExam(20%) > 18] then StudenLevel = 'Good	1
R15	If [FinalLabExam(20%) > 15.5 && FinalProject(30%) > 22.5] then StudentLevel = 'Good'	28
Total Samples		111

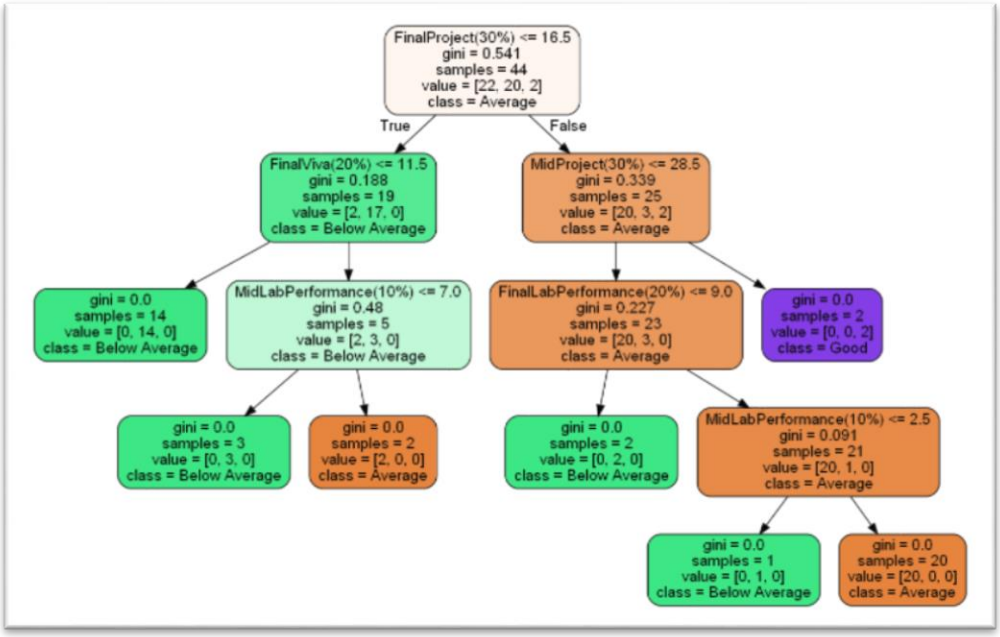

Fig.4. Decision Tree derived from Dataset 4 (WT)

Table 9. Derived set of rules from Dataset 4 (WT)

Rules	Description	Samples
R1	If [FinalProject(30%) <= 16.5 && FinalViva(20%) <= 11.5] then StudentLevel = 'Below Average'	14
R2	If [FinalProject(30%) <= 16.5 && FinalViva(20%) > 11.5 && MidLabPerformance(10%) <= 7] then StudentLevel = 'Below Average'	3
R3	If [FinalProject(30%) <= 16.5 && FinalViva(20%) > 11.5 && MidLabPerformance(10%) > 7] then StudentLevel = 'Average'	2
R4	If [FinalProject(30%) > 16.5 && MidProject(30%) <= 28.5 && FinalLabPerformance(20%) <= 9] then StudentLevel = 'Below Average'	2
R5	If [FinalProject(30%) > 16.5 && MidProject(30%) <= 28.5 && FinalLabPerformance(20%) > 9 && MidLabPerformance(10%) <= 2.5] then StudentLevel = 'Below Average'	1
R6	If [FinalProject(30%) > 16.5 && MidProject(30%) <= 28.5 && FinalLabPerformance(20%) > 9 && MidLabPerformance(10%) > 2.5] then StudentLevel = 'Average'	20
R7	If [FinalProject(30%) > 16.5 && MidProject(30%) > 28.5] then StudentLevel = 'Good'	2
Total Samples		44

Tables 6-9 demonstrate that all the Decision Trees could transform the total instances to rules with the perfect distribution. These rules will assist in identifying the factors that have the maximum impact on the students' academic performance. The following table shows the most impactful factors in all the datasets-

Table 10. List of impactful factors from the datasets

Dataset	Most Impactful Factors from Decision Tree
Dataset 1 (IP)	MidQuiz(30%), MidAssessment(40%), FinalAssessment(20%), FinalViva(25%)
Dataset 2 (OOP1)	MidPerformance(10%), MidQuiz(20%), MidViva(20%), FinalViva(30%), FinalProject(20%)
Dataset 3 (OOP2)	MidLabPerformance(20%), MidLabExam(20%), MidViva(20%), FinalLabPerformance(10%), FinalLabExam(20%), FinalViva(20%), FinalQuiz(20%), FinalProject(30%)
Dataset 4 (WT)	MidLabPerformance(10%), MidProject(30%), FinalLabPerformance(20%), FinalProject(30%), FinalViva(20%)

The following graphs show the range of obtained marks for all the three categories of students in the factors from table 10.

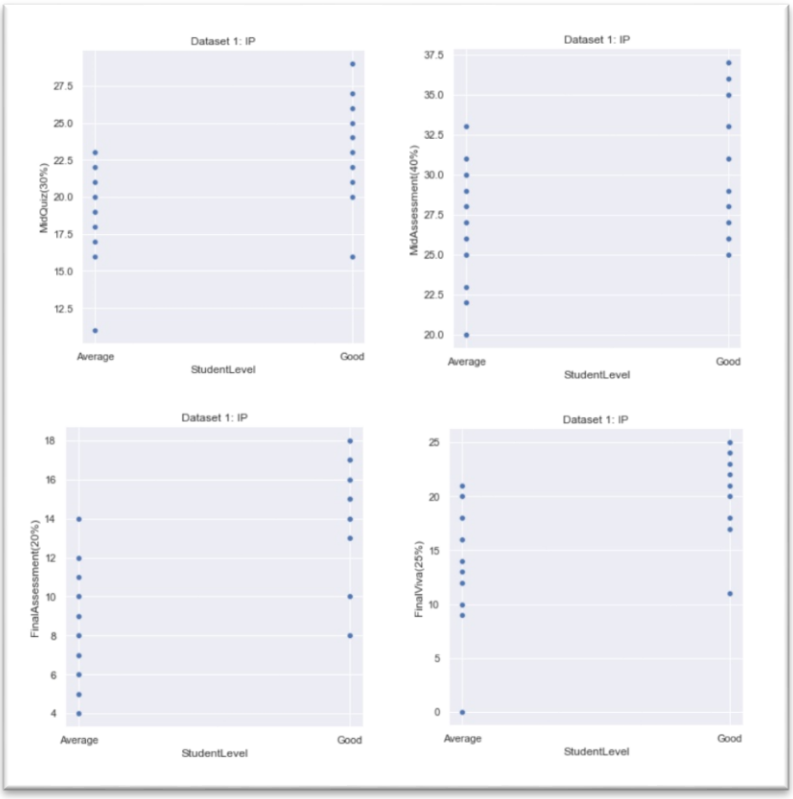

Fig.5. The impact of the factors corresponding to StudentLevel in Dataset 1 (IP)

Fig 5 depicts the impact of 'MidQuiz', 'MidAssessment', 'FinalAssessment' and 'FinalViva' in 'StudentLevel' for Dataset 1 (IP). 'Good' students obtained marks between 16 and 29 in 'MidQuiz(30%)', between 25 and 37 in 'MidAssessment(40%)', between 8 and 18 in 'FinalAssessment(20%)' and between 11 and 25 in 'FinalViva(25%)'. On the contrary, the students who are 'Average' gathered marks from 11 to 23 in 'MidQuiz(30%)', from 20 to 33 in 'MidAssessment(40%)', from 4 to 14 in 'FinalAssessment(20%)' and from 0 to 21 in 'FinalViva(25%)'.

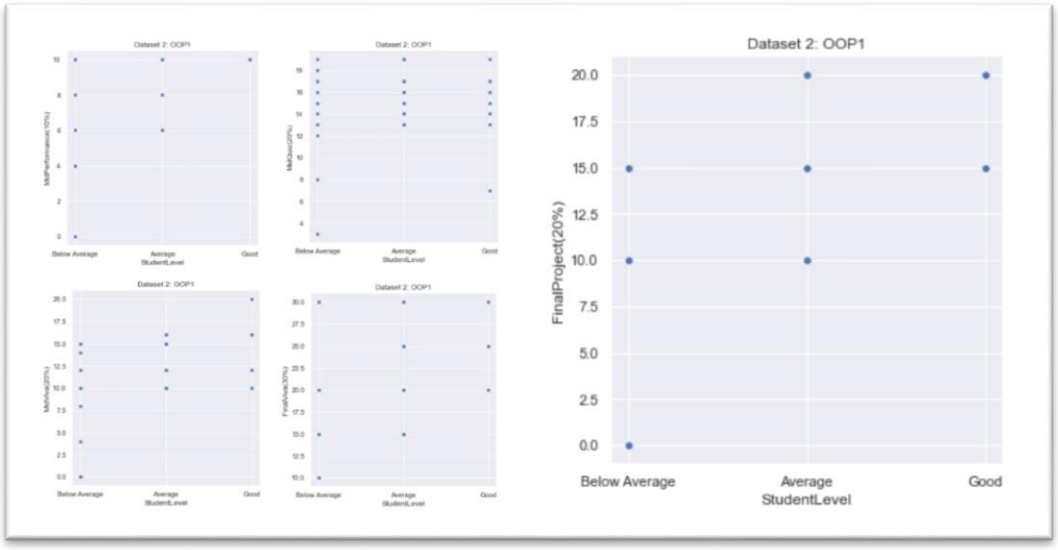

Fig.6. The impact of the factors corresponding to StudentLevel in Dataset 2 (OOP1)

Fig 6 implies the impact of all factors in 'StudentLevel' for Dataset 2 (OOP1). For instance, 'Good' students achieved full marks in 'MidPerformance(10%), between 8 and 19 in 'MidQuiz(20%)', between 10 and 20 in 'MidViva(20%)', between 20 and 30 in 'FinalViva(30%)' and between 15 and 20 in 'FinalProject(20%)'. These description patterns can be used to illustrate the rest of the categories of students ('Average' and 'Below Average').

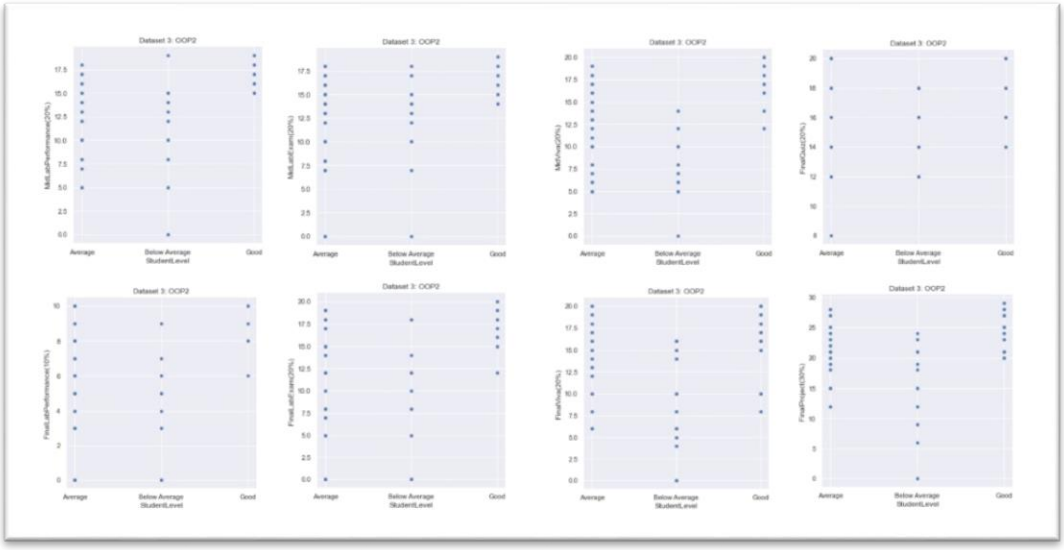

Fig.7. The impact of the factors corresponding to StudentLevel in Dataset 3 (OOP2)

Fig 7 indicates the impact of 'MidLabPerformance', 'MidLabExam', 'MidViva', 'FinalLabPerformance', 'FinalLabExam', 'FinalViva', 'FinalQuiz', 'FinalProject' in 'StudentLevel' for Dataset 3 (OOP2). For instance, 'Average' students got marks from 5 to 18 in 'MidLabPerformance(20%)', from 0 to 10 in 'FinalLabperformance(10%)', from 0 to 18 in 'MidLabExam(20%)', from 0 to 19 in 'FinalLabExam(20%)', from 5 to 19 in 'MidViva(20%)', from 6 to 20 in 'FinalViva(20%)', from 8 to 20 in 'FinalQuiz(20%)' and from 12 to 28 in 'FinalProject(30%)'. These description patterns are applicable for illustrating the rest of the categories of students ('Good' and 'Below Average').

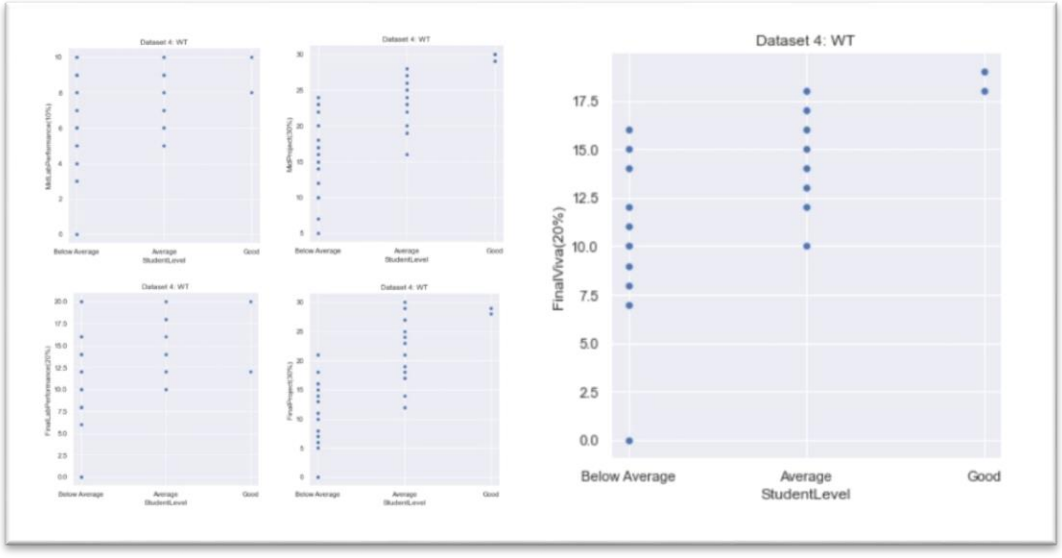

Fig.8. The impact of the factors corresponding to StudentLevel in Dataset 4 (WT)

Fig 8 point out the impact of 'MidLabPerformance', 'FinalLabPerformance', 'MidProject', 'FinalProject', 'FinalViva' in 'StudentLevel' for Dataset 4 (WT). For instance, the students who are 'Below Average' collected marks between 0 and

10 in 'MidLabPerformance(10%)', between 5 and 24 in 'MidProject(30%)', between 0 and 20 in 'FinalLabPerformance(20%)', between 0 and 21 in 'FinalProject(30%)' and between 0 and 16 in 'FinalViva(20%)'. Similarity has been observed in the functionality of the rest categories of students ('Good' and 'Average').

This analysis is based on four consecutive years of data from the Computer Science department's four programming courses (Introduction to Programming, Object Oriented Programming 1, Object Oriented Programming 2, and Web Technologies). The findings from these experiments indicate that a range of factors influences students' online performance. Students may join the courses remotely due to the online format of the courses. They were not presented with the hurdles that students often face when on-campus classes, which may explain why students' attendance' and lab performance rates are so high in online courses. Introduction to Programming (IP) introduces students to the fundamentals of programming. The majority of students already have some prior understanding of the contents of the course, which makes it simpler for them to grasp the concepts. That is why they earned a decent score on the assessment. Students performed well on their mid-viva in Object Oriented Programming-1 (OOP1) since they had some prior understanding of the subjects covered in Introduction to Programming. However, they were required to complete a final term project based on the knowledge obtained over the semester. Since this course was performed online, they may have certain knowledge deficits that they could not fill due to a lack of contact between group members and the course instructor. As this is their first project as undergraduate students, several students struggled with the project, which impacted their viva, dependent on the project they delivered. Students in Object Oriented Programming-2 (OOP2) earned good grades on the project and viva due to their prior expertise with projects from the earlier semester. In Web Technologies (WT), students earned an average grade on the project and viva since they were required to submit two distinct projects during the mid and final terms, and they were required to work with multiple languages. To generalize, first-year students outperformed seniors by a significant margin.

5. Conclusion

The primary goal of this research is to ascertain the critical parameters influencing students' performance in virtual classrooms. Jupyter Notebook is utilized to analyze students' performance during an online semester at 'X University.' Following that, various patterns for distinct factors are constructed using the Decision Tree classification model. Four separate decision trees are created from each of the courses' different datasets. The knowledge extracted from the decision trees indicates that viva, performance, and project are the most significant factors in analyzing students' performance, whereas quizzes, lab exams, and assessment had a lesser impact (table 10). These trees depicted crucial aspects affecting pupils' academic performance. As mentioned before, this paper is a subset of a bigger study. In another part, we did a comparison between online and on-campus students' performance. This study can be implemented on a bigger scale with much more data on different courses in the future, where a prediction will also be an option. If an epidemic should strike in the future, this study will aid educational authorities in confronting the challenges unleashed upon students' academic futures. Additionally, this paper can direct authorities to the critical factors that make a difference if they are obliged to transition the curriculum from offline to online or vice versa. Furthermore, future researchers will benefit from this research to undertake their ideas or studies.

References

[1] Jena, P. K. (2020). Impact of pandemic COVID-19 on education in India. International journal of current research (IJCR), 12.
[2] Mortada, L., Bolbol, J., & Kadry, S. (2018). Factors Affecting Students' Performance a Case of Private Colleges in Lebanon. J Math Stat Anal, 1, 105.
[3] Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418.
[4] Aliyeva, T., Rzayeva, U., & Khalilova, J. (2021). Problems and Prospects in the Applying Methods of Analysis Educational Data. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(6), 1255-1266.
[5] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
[6] He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior, 29(1), 90-102.
[7] Rastrollo-Guerrero, J. L., Gomez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. Applied sciences, 10(3), 1042.

[8] Sathe, M. T., & Adamuthe, A. C. (2021). Comparative Study of Supervised Algorithms for Prediction of Students' Performance. International Journal of Modern Education & Computer Science, 13(1).
[9] Malini, J. (2021). Analysis of factors affecting Student performance Evaluation using Education Dataminig Technique. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(7), 2413-2424.
[10] Ali, Z. M., Hassoon, N. H., Ahmed, W. S., & Abed, H. N. (2020). The Application of Data Mining for Predicting Academic Performance Using K-means Clustering and Naïve Bayes Classification. International Journal of Psychosocial Rehabilitation, 24(03).
[11] Nahar, A., Sohan, M. F. A. A., & Yasmin, S. (2021). Impact of Prolonged Isolation from the campus on the mental health of the students during Covid-19 pandemic. AIUB Journal of Science and Engineering (AJSE), 20(1), 59-64.
[12] Moreno-Marcos, P. M., Pong, T. C., Munoz-Merino, P. J., & Kloos, C. D. (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. IEEE Access, 8, 5264-5282.
[13] Lakshmi, T. M., Martin, A., Begum, R. M., & Venkatesan, V. P. (2013). An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data. International Journal of Modern Education & Computer Science, 5(5).
[14] Al Karim, M., Ara, M. Y., Masnad, M. M., Rasel, M., & Nandi, D. (2021). Student performance classification and prediction in fully online environment using Decision tree. AIUB Journal of Science and Engineering (AJSE), 20(3), 70-76.
[15] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.
[16] Yang, Y. (2021). The Evaluation of Online Education Course Performance Using Decision Tree Mining Algorithm. Complexity, 2021.
[17] Jiang, X. (2021). Online English Teaching Course Score Analysis Based on Decision Tree Mining Algorithm. Complexity, 2021.
[18] Hoque, M. I., Kalam Azad, A., Tuhin, M. A. H., & Salehin, Z. U. (2020). University students result analysis and prediction system by decision tree algorithm. Adv Sci Technol Eng Syst J, 5(3), 115-122.
[19] Sajja, V. R., Lakshmi, P. J., Naik, D. B., & Kalluri, H. K. (2021). Student Performance Monitoring System Using Decision Tree Classifier. In Machine Intelligence and Soft Computing (pp. 393-407). Springer, Singapore.

Authors' Profiles

Nyme Ahmed has earned a Bachelor of Science (BSc) in Computer Science and Engineering (CSE) from the American International University-Bangladesh (AIUB) in 2021. During his BSc, he was awarded the Summa Cum Laude and Dean's List Honors for academic excellence several times. He is interested in research areas including Data Mining, Data Analytics, Graph Theory, and a wide variety of Algorithms and Data Structures.

Rifat-Ibn-Alam has completed his Bachelor's in Computer Science & Engineering from American International University- Bangladesh (AIUB) in 2021. He was awarded the Summa Cum Laude award and Dean's List Honors multiple times for academic excellence during his undergraduate. His research interests include Data Mining, Data Analytics, and E-learning, and so on.

Md. Golam Ahsan Akib is pursuing his Bachelor's in Computer Science & Engineering from American International University- Bangladesh (AIUB) in 2021. He was awarded Dean's List Honors for academic excellence. His research interests include Data Analytics, Data Mining, and Machine Learning.

Syed Nafiul Shefat has received his Bachelor's in Computer Science & Engineering from American International University- Bangladesh (AIUB) in 2021. He was awarded the Summa Cum Laude and Dean's List Honors twice for academic excellence in his undergraduate period. His research interests are Data Mining, Data Analytics, Software Engineering, and Machine Learning.

Dr. Dip Nandi is currently working as a Professor and Director of Faculty of Science and Technology in American International University- Bangladesh. He has a vast range of research activities and contributions in various filed of Computer Science and Multidimensional researches. His research interest includes Data Mining, Software Engineering, Management Information Systems, E-learning etc. He can be contacted at dip.nandi@aiub.edu.