Will the Student Get an A Grade? Machine Learning-based Student Performance Prediction in Smart Campus

Ali Alnoman
Department of Computer Science and Engineering
American University of Ras Al Khaimah, UAE

2023 Advances in Science and Engineering Technology International Conferences (ASET) | 978-1-6654-5474-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASET56582.2023.10180545

***Abstract*—As an important component in the structure of smart campus, student performance prediction can help to observe the academic progress of students and make timely decisions towards improving the overall learning process. In order to achieve accurate performance prediction, it is required to extract useful data such as the past performance of students from the student information system and employ the extracted data to train an appropriate machine learning model. In this paper, different machine learning models are utilized to predict whether or not a student will get an A grade (i.e, 90 or higher). The work utilizes a real dataset that is extracted from three different courses that require computer and programming skills. The dataset includes three features about the student past performance, namely, high school grade, course midterm grade, and absence percentage. The dataset is then used to train different machine learning models, specifically, linear discriminant, logistic regression, Naive Bayes, support vector machine, decision tree, K-nearest neighbors, and bagged trees. In order to highlight the effectiveness of these classifiers, different metrics were used to evaluate the classification performance such as accuracy, precision, recall, and F1-score. Besides, these models are tested considering one, two, and three features from the dataset to evaluate the significance of each feature in the classification process. After comparing the performance of these machine learning models, it is shown that predicting student performance is indeed applicable with an accuracy that reached 99%. It is also shown that the bagged trees and K-nearest neighbors succeeded to achieve the highest classification accuracy compared to the other models.**

***Index Terms*—Machine learning, student performance prediction, classifier, smart campus.**

I. Introduction

The concept of smart campus emerged to promote student learning through a wide variety of advanced technologies such as data mining and artificial intelligence. These technologies are technically enabled with the help of the underlying smart campus infrastructure which is categorized into different layers such as the sensing layer, network layer, and computing layer.

One of the effective technologies in smart campus is the Internet of Things (IoT) which forms the sensing layer and involves a wide range of applications in different aspects related to smart buildings, smart furniture, and smart wearable sensors [1]. The sensing layer enables the computing layer which is in charge of making decisions based on the various needs of learners that can be identified by processing data related to student attendance, participation, performance, and health [2]. Here, it is worth to mention that the goal of implementing such technologies in the context of smart campus is not only to enhance the technical capabilities, but more importantly to fuse these technologies with the educational process in order to establish an efficient student-centered learning environment [3]. Early prediction of students performance helps instructors to identify the learning weaknesses at early stages and to make proactive steps that help to engage students and enhance the learning process.

Besides, the integration of machine learning and IoT in the information system of smart campus can assist in predicting student attendance which in turn helps to quantify the occupancy of classrooms and hence to optimize classroom utilization, air conditioning performance, and energy efficiency [4].

Large volumes of educational data are continuously generated by the information systems in universities and educational institutions. It is possible to achieve high accuracy predictions about student performance by adopting effective preprocessing techniques and appropriate machine learning models. The goal of predicting student performance is to help instructors and educational authorities to make effective and timely decisions toward improving the learning process. Among the popular machine learning models used for predicting student performance are decision trees, support vector machines (SVM), and neural networks (NN) [5]. The contribution of this work is as follows:

- Predict student performance as whether the student will get a final grade of A (i.e., 90 or higher).
- Use a real dataset to train and test the student performance prediction system. The utilized dataset includes three features, namely, high school grade, midterm grade, and attendance.
- Compare the classification performance of different machine learning models, specifically, decision trees, bagged trees, neural networks, K-nearest neighbors, support vector machine, linear discriminant, and KNN.
- Train and test the machine learning models using one, two, and three features from the dataset to observe the effectiveness of each feature (or the combination of different features) on the prediction accuracy.

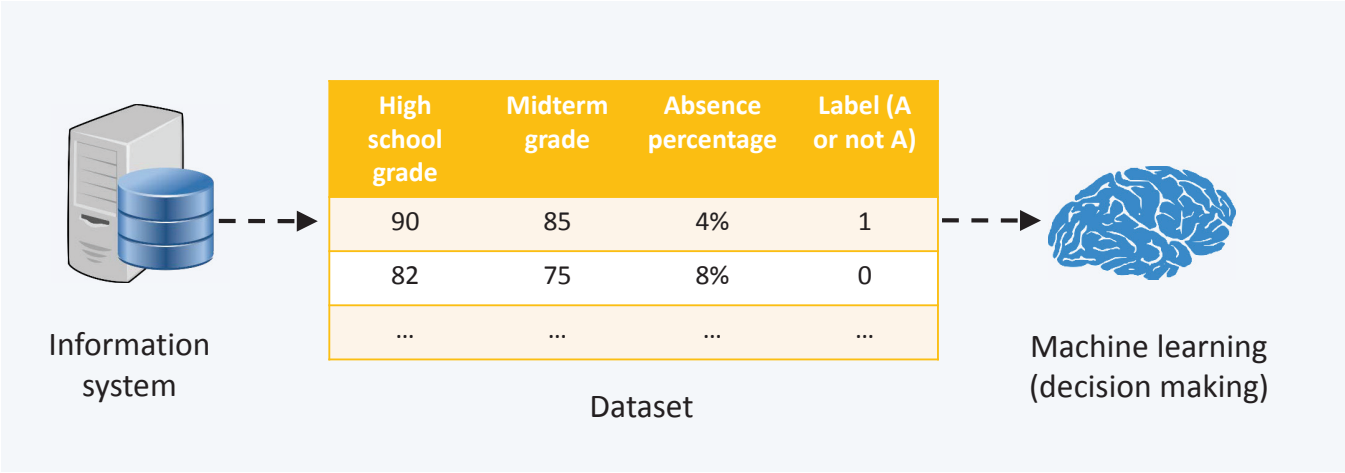


Fig. 1: Proposed system model.

The rest of the paper is organized as follows. Section I introduces the related research work. In Section III, the utilized machine learning models and implemented dataset are demonstrated. Section IV presents simulation results and discussions. Finally, section V provides the concluding remarks.

II. Related Work

Machine learning models comprise a set of algorithms and computations that can enable decision making systems in smart campus to adapt to the dynamic behavior of students and instructors. It has been shown that high-accuracy machine learning models can perform human-like tasks by learning from past experiences. Moreover, data mining techniques, which makes it possible to extract hidden features from educational data, have gained a growing attention in education due to the availability of educational data and the prosperity of artificial intelligence. In general, the goal of implementing machine learning in smart campus is to enable intelligent applications such as student performance prediction, context awareness, content customization, emotion recognition, and virtual tutoring [6].

Motivated by the increasing attention towards realizing smart campus, different machine learning approaches have been conducted to achieve different aspects of student performance prediction. Here, it is worth to mention that the accuracy of machine learning that is based on non-academic attributes such as gender and background information have exhibited lower accuracy (e.g. 60.9% [7]) compared to academic attributes such as historical data of student grades (e.g., 93.6% [8], 90.7% [9]). Furthermore, several approaches have been used to count the number of attendees in indoor environment such as WiFi-based counting, camera-based, and thermal imaging [4]. Each of these methods has its own advantages and disadvantages in terms of privacy, cost, accuracy, and ease of use. Nevertheless, the attendance data used in this work is based on the real data obtained from the student information system. Below is a summary of the different machine learning aspects that are considered to achieve student performance prediction:

- Different machine learning models such as decision trees, random forest, neural networks, support vector machine, K-nearest neighbors, Naive Bayes, and regression.
- Different classifier types (e.g., two-class classifier such as pass/fail or multi-class classifier [10] such as predicting the score or the high/average/low grades).
- Different features (e.g., grades, attendance, demographics, gender).
- Different datasets (e.g, data from one course or multiple courses, one semester or multiple semesters, one level or multiple levels, one institution or multiple institutions, does the course require programming/mathematics [11] skills or not).

III. System Model

The performance prediction process begins by extracting the relevant information from the student information system to create a dataset that is used to train the machine learning models. Specifically, the dataset contains 103 samples from different courses and levels about the high school grade, course midterm grade, and the absence percentage of students. Moreover, the dataset is extracted from courses in the computer science discipline where computing and programming skills are essential. Despite the fact that student attendance is dependent on different factors such as the lecture topic, teaching quality, and student motivation, attendance data can give useful information about the performance of students. Different machine learning models are utilized to achieve the intended classification objective. The purpose of classification is to predict whether or not a student will get an A grade (i.e., 90 or higher). Fig. 1 depicts the proposed system model, and Fig. 2 shows portion of the utilized

TABLE I: Predictions of the proposed machine learning models.

	Linear Discriminant	Logistic Regression	Naïve Bayes	SVM	Decision Tree	NN	KNN	Bagged Trees
Accuracy (%)	93.2	92.2	91.3	94.2	95.1	96.1	98.1	99
True Positive	20	20	19	20	20	23	22	22
True Negative	76	75	75	77	78	76	79	80
False Positive	4	5	5	3	2	4	1	0
False Negative	3	3	4	3	3	0	1	1

High School Grade	Absence	Midterm Grade	Label
87	0.03	48	0
94	0	76	1
82	0	72	0
88	0.07	60	0
94	0	52	0
:	:	:	:
83	0.07	76	0
91	0.07	76	0
76	0.11	60	0
90	0.1	72	0
93	0.11	84	1

Fig. 2: Portion of the utilized dataset.

dataset. It is worth to mention that in Fig. 2 an absence value of 0.03 means 3%.

Since the data extracted from the information system include different courses and grading scales, data preprocessing is carried out before feeding the dataset into the machine learning models. Here, different preprocessing steps are implemented:

- All grades are scaled to 100% since the maximum grade of midterm exams from different courses can be different.
- Grades that are categorized as A (i.e., grades of 90 or higher) are labeled as 1 while all other grades are labeled as 0.
- Samples that include missing values are removed.

IV. Simulation Results

To evaluate the system performance, simulations were conducted using the classification learner application in MATLAB. Machine learning models that were used in simulations are linear discriminant, logistic regression, Naive Bayes, SVM, KNN, neural network, decision tree, and bagged trees. After training the aforementioned machine learning models, the performance is evaluated by carrying out a 10-fold cross-validation which helps to protect against overfitting. Specifically, a 10-fold cross-validation will partition the dataset into 10 disjoint randomly chosen subsets (folds) such that in each iteration one fold is used to validate the model while the other folds are used to train the model. The app then calculates the average validation error over all folds. Fig. 3 shows the accuracy of the utilized machine learning models. It can be observed that the bagged trees model outperformed the other models with an accuracy of 99% followed by KNN with an accuracy of 98.1%. The bagged trees model is technically similar to the Random Forest model in the sense that it is an ensemble machine learning model that applies a number of decision tress on subsets of the dataset in order to achieve the highest decision accuracy. Accordingly, the results shown coincide with several research works such as [10] [11] [12] [13] which revealed that the Random Forest model outperformed other machine learning models in predicting student performance. Fig. 3 also shows that using all the existing features helps to attain higher accuracy compared to using only one or two features. Moreover, using only the high school grade or only the absence information as a feature achieved poor performance.

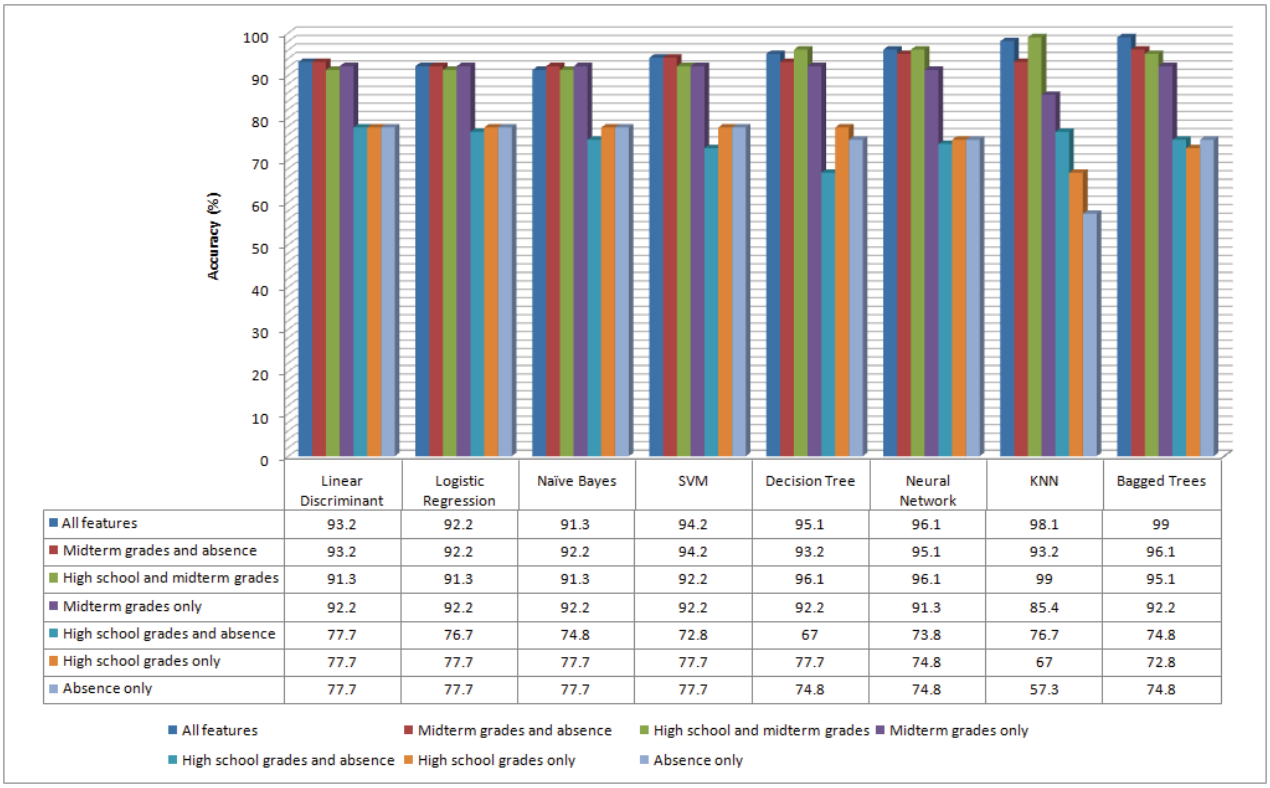
It is worth to mention that using the accuracy as the sole metric for evaluating the performance of machine learning models is insufficient. Therefore, other measurements such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are used as shown in Table I which shows these measurements for the utilized machine learning models considering all features. In Table I, the number of samples that are labeled 1 is 23 out of 103 samples whereas the number of samples labeled 0 is 80 out of 103. To give more details, true positive here means that the model predicted an A grade (label 1) and the student indeed obtained an A grade. The true positive, true negative, false positive, and false negative values help to calculate more important metrics, namely, the accuracy, precision, recall, and F1-score which combines the precision and recall in one metric. These values are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

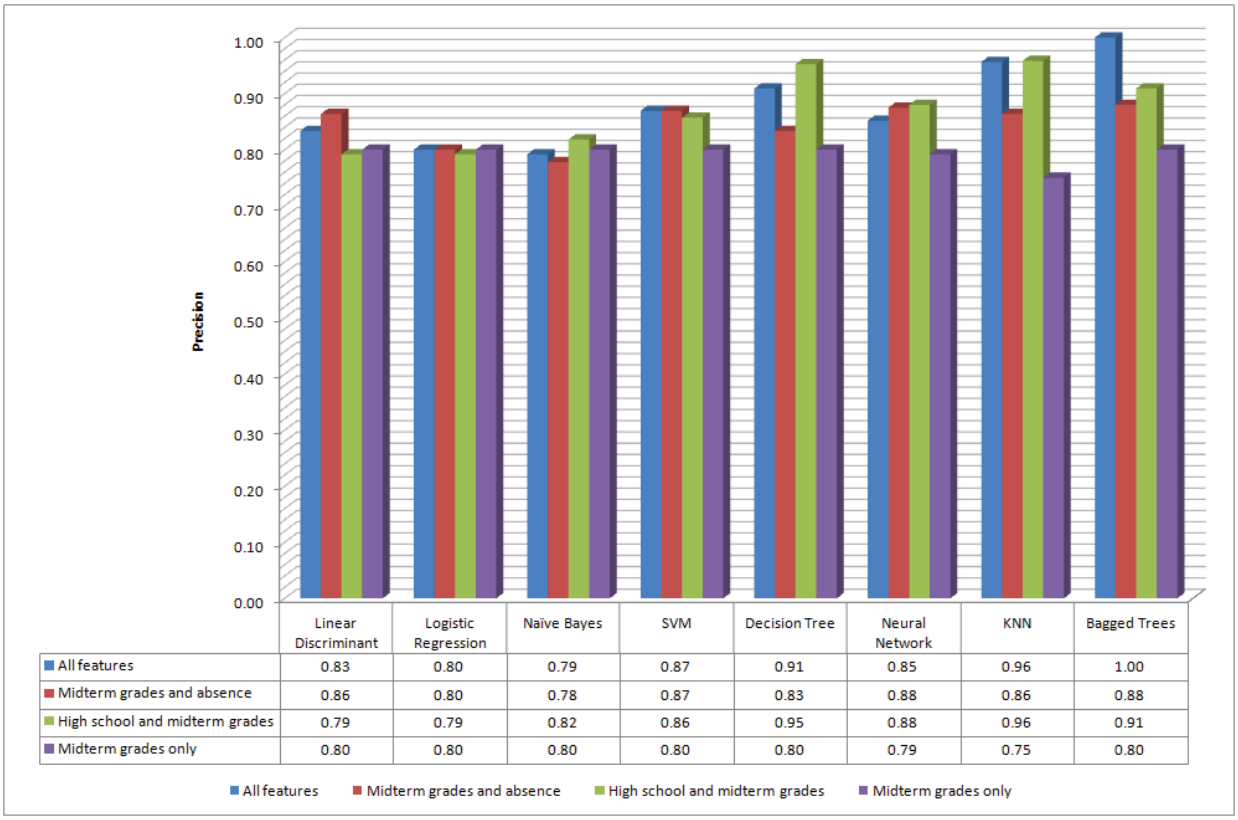
$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$



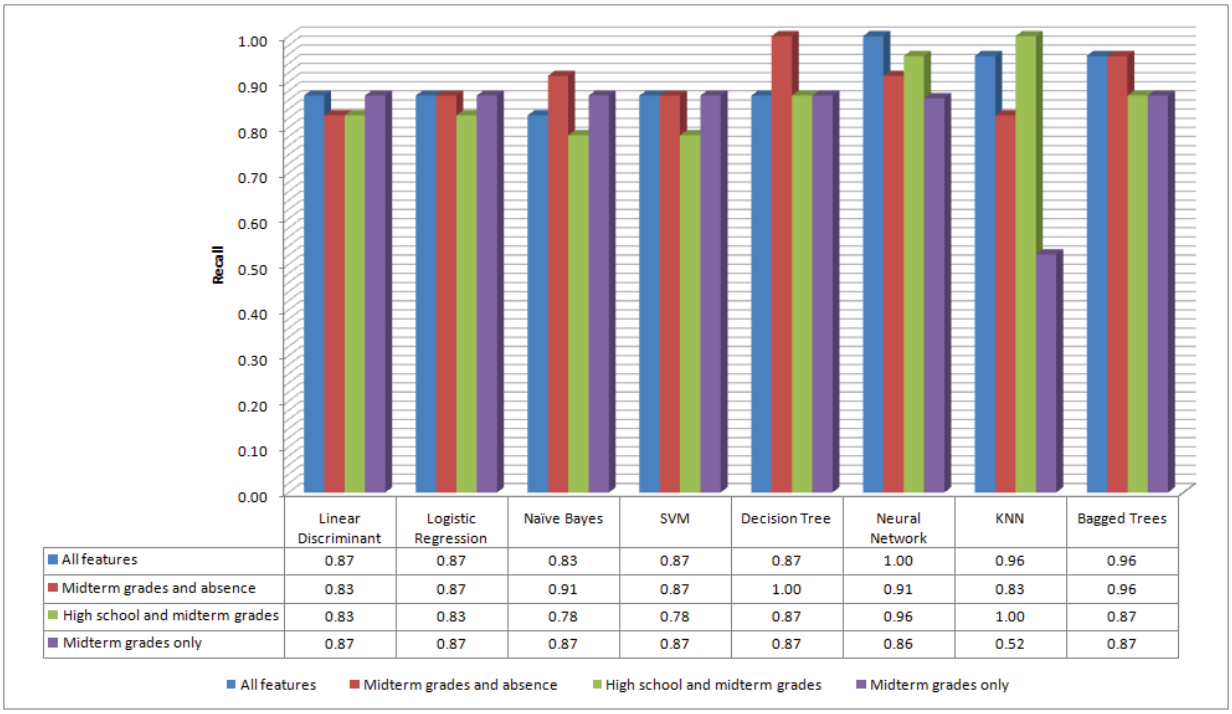
	Linear Discriminant	Logistic Regression	Naïve Bayes	SVM	Decision Tree	Neural Network	KNN	Bagged Trees
All features	93.2	92.2	91.3	94.2	95.1	96.1	98.1	99
Midterm grades and absence	93.2	92.2	92.2	94.2	93.2	95.1	93.2	96.1
High school and midterm grades	91.3	91.3	91.3	92.2	96.1	96.1	99	95.1
Midterm grades only	92.2	92.2	92.2	92.2	92.2	91.3	85.4	92.2
High school grades and absence	77.7	76.7	74.8	72.8	67	73.8	76.7	74.8
High school grades only	77.7	77.7	77.7	77.7	77.7	74.8	67	72.8
Absence only	77.7	77.7	77.7	77.7	74.8	74.8	57.3	74.8

Fig. 3: Accuracy of the implemented machine learning models.



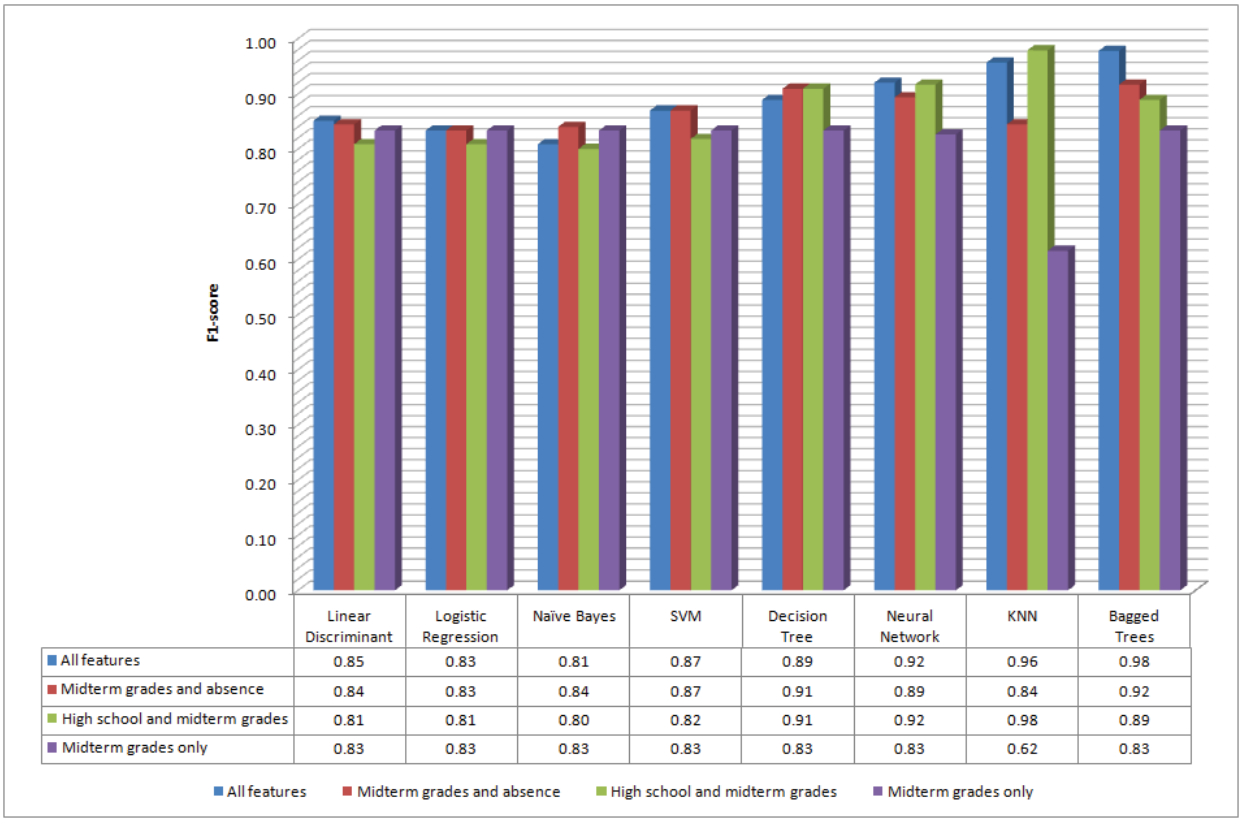
	Linear Discriminant	Logistic Regression	Naïve Bayes	SVM	Decision Tree	Neural Network	KNN	Bagged Trees
All features	0.83	0.80	0.79	0.87	0.91	0.85	0.96	1.00
Midterm grades and absence	0.86	0.80	0.78	0.87	0.83	0.88	0.86	0.88
High school and midterm grades	0.79	0.79	0.82	0.86	0.95	0.88	0.96	0.91
Midterm grades only	0.80	0.80	0.80	0.80	0.80	0.79	0.75	0.80

Fig. 4: Precision of the implemented machine learning models.



	Linear Discriminant	Logistic Regression	Naïve Bayes	SVM	Decision Tree	Neural Network	KNN	Bagged Trees
All features	0.87	0.87	0.83	0.87	0.87	1.00	0.96	0.96
Midterm grades and absence	0.83	0.87	0.91	0.87	1.00	0.91	0.83	0.96
High school and midterm grades	0.83	0.83	0.78	0.78	0.87	0.96	1.00	0.87
Midterm grades only	0.87	0.87	0.87	0.87	0.87	0.86	0.52	0.87

Fig. 5: Recall of the implemented machine learning models.



	Linear Discriminant	Logistic Regression	Naïve Bayes	SVM	Decision Tree	Neural Network	KNN	Bagged Trees
All features	0.85	0.83	0.81	0.87	0.89	0.92	0.96	0.98
Midterm grades and absence	0.84	0.83	0.84	0.87	0.91	0.89	0.84	0.92
High school and midterm grades	0.81	0.81	0.80	0.82	0.91	0.92	0.98	0.89
Midterm grades only	0.83	0.83	0.83	0.83	0.83	0.83	0.62	0.83

Fig. 6: F1-score of the implemented machine learning models.

Figs. 4, 5, and 6 further emphasize the superior performance of the bagged trees model especially when all features are used in the model training. However, the latter figures did not include the features "High school grades and absence", "High school grades only", and "Absence only" as they yield low accuracy as shown in Fig. 3. Following the bagged trees model, the KNN model with only two features (high school and midterm grades) showed a competing performance and was successful to achieve high-accuracy predictions.

V. Conclusion

In this paper, different machine learning models were used to predict student performance by indicating whether or not the student will get an A grade. The work utilized a real dataset extracted from courses that require computers knowledge and programming skills. The dataset includes three features about the student past performance, namely, high school grade, course midterm grade, and absence percentage. Afterwards, different machine learning models were trained and tested. These machine learning models are linear discriminant, logistic regression, Naive Bayes, support vector machine, decision tree, K-nearest neighbors, and bagged trees. To evaluate the performance of these classifiers, certain metrics such as accuracy, precision, recall, and F1-score where used while utilizing one, two, and three features from the dataset for training and testing. Results showed that student performance prediction is applicable with an accuracy that reached 99%. Moreover, the bagged trees and K-nearest neighbors models exhibited superior performance compared to other machine learning models.

References

[1] Y. Liang and Z. Chen, "Intelligent and Real-Time Data Acquisition for Medical Monitoring in Smart Campus," *IEEE Access*, vol. 6, pp. 74 836–74 846, Nov. 2018.

[2] X. Xu *et al.*, "Research on Key Technologies of Smart Campus Teaching Platform Based on 5G Network," *IEEE Access*, vol. 7, pp. 20 664–20 675, Jan. 2019.

[3] A. Alnoman, "A Framework for Technology-based Student-centered Learning in Smart Campus," in *IEEE Advances in Science and Engineering Technology International Conferences (ASET)*, Feb. 2022, pp. 1–4.

[4] T. Sutjarittham, H. Habibi Gharakheili, S. S. Kanhere, and V. Sivaraman, "Experiences With IoT and AI in a Smart Campus for Optimizing Classroom Usage," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7595–7607, 2019.

[5] S. Qu, K. Li, S. Zhang, and Y. Wang, "Predicting Achievement of Students in Smart Campus," *IEEE Access*, vol. 6, pp. 60 264–60 273, Oct. 2018.

[6] Z. Dong, Y. Zhang, C. Yip, S. Swift, and K. Beswick, "Smart campus: Definition, Framework, Technologies, and Services," *IET Smart Cities*, 2020.

[7] F. J. Kaunang and R. Rotikan, "Students' Academic Performance Prediction using Data Mining," in *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1–5.

[8] I. A. Abu Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 909–913.

[9] H. Gull, M. Saqib, S. Z. Iqbal, and S. Saeed, "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020, pp. 1–4.

[10] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access*, vol. 9, pp. 95 608–95 621, 2021.

[11] P. Sokkhey and T. Okazaki, "Comparative Study of Prediction Models on High School Student Performance in Mathematics," in *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2019, pp. 1–4.

[12] E. H. Yossy, Y. Heryadi, and Lukas, "Comparison of Data Mining Classification Algorithms for Student Performance," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, 2019, pp. 1–4.

[13] M. S. Ram, V. Srija, V. Bhargav, A. Madhavi, and G. S. Kumar, "Machine Learning Based Student Academic Performance Prediction," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 683–688.