Understanding Student Performance in Foundation Year: Insights from Logistic Regression, Naïve Bayes, and Random Forest Models

Abdallah Bashir Musa

Department of Basic Sciences, Deanship of Preparatory Year and Supporting Studies,
Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
Email: abhamad@iau.edu.sa (A.B.M.)

Manuscript received August 20, 2024; revised September 9, 2024; accepted November 4, 2024; published December 13, 2024

***Abstract*—Foundation programs enhance students' essential skills, equip them for degree programs, and impact academic performance, retention, and intrinsic motivation. Previous studies focused mostly on demographic factors and statistics. Limited literature has focused on students' performance in the foundation year. This study uses machine learning techniques to investigate the factors influencing foundation year students' performance. The study assesses 22 predictor factors, including demographics, secondary school achievement, language proficiency, and university experiences, using Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF) algorithms. The study's findings revealed that gender, school type, secondary school scores, desired college major, and English and math proficiency levels were the significant determinants of students' performance in their foundation year. Random Forest (RF) showed higher accuracy than both Naïve Bayes (NB) and Logistic Regression (LR). The study indicated that identifying performance factors can improve support services by maximizing learning and results via data-driven methodologies. In conclusion, this study revealed the potential of machine learning in evaluating student performance determinants, supporting targeted interventions, and individualized training through advanced machine learning algorithms and longitudinal data. Moreover, the study helps predict students' performance in the second semester. Consequently, it projects the enrollment figures for each college along with the anticipated dropout rates.**

***Keywords*—foundation year, Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF)**

I. Introduction

The academic standing of college students is the most essential consideration in determining a university's ranking. Student accomplishment is influenced by several variables, including demography, socioeconomic status, and elements associated with high school. Assessment of student performance during the foundation year acts as a fundamental standard by which colleges assess and choose new students, as well as track the efficacy of their instructional strategies. In today's fiercely competitive educational environment, many colleges struggle to draw in prospective students. Every university offers foundation year programs globally to strengthen the students' concepts and fundamental understanding of their respective subjects and skills, which they will need throughout their degree and in their practical lives. Imam Abdulrahman Bin Faisal University, established in 1975 with two colleges, has grown into a leading research university with 21 colleges in the Eastern Province and a student population of over 45,000 [1].

The Education Foundations program at Imam Abdulrahman Bin Faisal University aims to achieve excellence in education to meet global developments and community demands [2]. Their mission is to prepare scientific professionals, conduct academic studies, and provide quality consultations. Their goals include providing students with educational skills, developing research skills, providing educational consultations, conducting original research, training students to diagnose and address educational issues, and instilling Islamic values and ethics [3]. They also promote leadership, transparency, responsibility, quality, respect, honesty, and objectivity [1]. In addition, a major purpose of the foundation year program is to enroll students in a specific department, providing them with foundational courses for their chosen fields [4]. These students are divided into tracks like health, engineering, science, and humanities, each with its own subject requirements. English language proficiency is a common subject for all tracks. Students spend a year in this department, where their performance is evaluated, creating a competitive environment [5]. After completing this year, they are allocated to colleges associated with their chosen tracks. This study focuses on analyzing academic performance in science and engineering tracks, where English language and mathematics are core subjects. Further, understanding the factors affecting student performance is crucial for improving higher education quality and student development. This helps in refining teaching methods and providing tailored support to students facing difficulties. Factors such as socioeconomic background and past academic performance were found to be major in this regard [6]. Previous research has mainly focused on predicting students' performance using simple statistical techniques [7]. Research showed that foundation year programs emphasize the importance of foundation programs in higher education, highlighting the value of accommodation, financial support, and fostering a sense of belonging [8].

This study analyzed the academic performance of incoming students in a preparatory year, who come from diverse backgrounds and aspire to enroll in different colleges. The students struggle during the foundation year as they compete to enroll at their desired college. For instance, a student may be enrolled in the College of Business Administration while he/she is interested in studying at the College of Computer Science and Information Technology. Moreover, students who perform poorly will drop out of university. Identifying the factors that affect the student's performance helps the deanship administration take necessary supportive initiatives for poorly performing

students and decrease the dropout rate. Additionally, this study integrated the application of machine learning methods, which is a relatively unconventional approach in the realm of educational research. The growing adoption of machine learning in data science applications, especially in classifying and examining complex connections, presents an intriguing opportunity for categorizing students' educational data [9]. The research compared three popular machine learning algorithms: Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF) in the task of classifying students' performance. LR and NB are traditional statistical methods that have been widely used for classification and prediction in both statistics and machine learning.

On the other hand, random forest is a newer algorithm that has shown promising accuracy in classification and regression tasks [10]. As a whole, utilizing the machine learning method, this study provided valuable cultural insights, assessing the academic performance of students in their foundation year at Imam Abdulrahman Bin Faisal University. The study compared logistic regression, naïve Bayes, and random forest algorithms performance for student data. The study findings revealed that primary performance predictors include gender, school type, secondary school scores, desired college major, and English and math skill levels. Random Forest (RF) showed higher accuracy than both Naïve Bayes (NB) and Logistic Regression (LR). In essence, the study provided valuable insights for refining teaching and support services to optimize learning experiences and outcomes across diverse learners. The research promotes equity and excellence in higher education through continued investigation.

II. Literature Review

When examining the literature on the study of academic performance, it is evident that most studies primarily rely on three types of information regarding students. These include (1) demographics and socio-economic factors, (2) high school-related information, and (3) college enrollment data [11]. The demographic and socio-economic information commonly used in these studies encompasses the student's gender, place of residence, parental status (both parents alive, one parent alive, or both parents deceased), parents' occupations and educational levels, whether the student is employed or not, and the number of family members [12]. The high school-related information utilized in this study consists of the student's high school GPA, the type of high school attended (public or private), the student's proficiency level in English (advanced, intermediate, or beginner), and their proficiency level in mathematics (advanced or not advanced). Further, before starting their classes, students must undergo placement tests in English and mathematics. These tests evaluate their proficiency in these subjects, and based on their scores, students are assigned to the appropriate English or mathematics proficiency level. Additionally, several variables may impact students' performance, such as the number of hours they study per day (≤2, 2 - 4, or >4), their chosen academic track (engineering or science), and their desired college major (engineering, design, computer science, or business administration) [13]. Students in the science track are allocated to either the College of Computer Science and Information Technologies or the College of Business Administration, while students in the engineering track are assigned to engineering or design colleges based on their Grade Point Average (GPA) at the end of the second semester. Furthermore, an important factor to consider is whether the student utilizes the services of the Deanship Learning Resources System Center (LRSC). The primary objective of this center is to provide academic support to students, with one instructor available for each subject. It has been observed that students who consistently visit the center experience significant improvements in their academic performance [14]. The evaluation of the student's performance is based on their GPA for the first semester, which is assessed on a scale of 5; the grades for each subject range from A+ to F. Given that the preparatory year is crucial in determining a student's future career, it is common for students to strive for a high GPA to secure admission into their desired college. It is rare for students to have a GPA lower than 3.0 out of 5.0, as this would result in dropping them out of the university [15]. Consequently, the students' performance is categorized into two classes based on their GPA: less than 4 or greater than or equal to 4. Students who obtain a GPA lower than 4 are unable to enroll in colleges and typically transfer to another university.

A. Machine Learning Systems for Educational Data

Machine learning algorithms have become essential in analyzing complex educational data sets because they can model non-linear relationships more efficiently without imposing explicit programming [16]. In educational data mining, hybrid techniques including decision trees, clustering, artificial neural networks, and Naïve Bayes are more effective at predicting student achievement. Applying machine learning and artificial intelligence in education has recently gained popularity. Several educational applications exist for artificial intelligence, machine learning, and data analysis. These applications provide personalized training, administrative help, and data-driven decision-making [17]. Without previous grades, machine learning algorithms can predict future grades, identify poor students, and enhance educational institutions [18]. Additionally, they can identify students who are likely to drop out, improving instructional strategies and lowering dropout rates [19]. For instance, Huynh-Cam applied the decision tree and random forest to predict freshmen students' performance at an earlier time. CART is identified as the top classifier. The significant factors were the mother's occupation, department, father's occupation, the main source of living expenses, and admission status [20]. Masangu studied the factors influencing students' performance in different grades. Various machine learning algorithms are applied, including support vector machines, logistic regression, and random forest classifiers. The support vector machine was the most effective classifier, followed by random forest and logistic regression. The number of days a student is absent is the key factor impacting academic performance [21]. Sixhaxa investigated the academic and behavioral factors that affect students' performance. Five machine learning techniques were used, including random forest and logistic regression. Random forest performed better than logistic regression. The most important factors that affect students' performance are visited resources, raised hands, and student absent days

[22]. Lavidas also investigates the factors influencing students' use and intention to use artificial intelligence technology for academic purposes. This study applied the Unified Theory of Acceptance and Use of Technology (UTAUT2) model. The expected performance, habit, and enjoyment of artificial intelligence applications are the key factors influencing teachers' intentions to use them. Additionally, the actual use of AI applications is shaped by behavioral intention, habits, and enabling situations [23]. Most of the pre-mentioned studies investigated the factors that affect students' performance using demographic features. However, a few researchers, like Sixhaxa (2022), expanded their analysis by incorporating academic and behavioral factors. This study's novelty is that it combines demographic features with secondary school performance, language proficiency, mathematics skill level, and university experiences. It specifically examines the factors affecting student performance within the Deanship of Preparatory Year and Supporting Studies at Imam Abdulrahman Bin Faisal University, making it very useful in providing valuable insights for the Deanship members to formulate effective educational policies. This study applied logistic regression, naïve Bayes, and random forest. Logistic regression uses the maximum likelihood estimation technique by building a logistic function. It is widely used for classification tasks in educational research due to its interpretability and mathematical simplicity [24]. Logistic regression was found to have an accuracy of 0.8 in predicting drop-out students based on demographic and performance measures [25]. Additionally, Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem and supposes the independence of predictors. It has shown comparable performance to more complex techniques in classification text with a difference of only 1-5% and accurate prediction in the areas of college enrolment using the attributes of the students (85-90% accuracy) [26]. Because of its strength in uncertain situations, Naïve Bayes is good at representing student knowledge. Random forest is an ensemble method that utilizes a collection of decision trees, created from bootstrap samples, to form a strong classifier and effective feature selector, abstaining from overfitting [27]. This method has shown impressive performance, often rivaling or surpassing other algorithms. Specifically, the Random Forest technique achieved around 90% accuracy in predicting changes in college majors based on academic and demographic factors [28]. The algorithm also delivered commendable results in classifying text from Discussion Forums, achieving accuracy rates between 84% to 89%. It has also shown promise in modeling intelligence within intelligent tutoring systems [29]. Naïve Bayes outperformed logistic regression and decision trees during those studies comparing multiple methods that investigated predicting the exam scores, while the random forest was among the top performers for identifying students at risk based on individual assessments [30]. While each of these algorithms has proven to be valuable on its own, a direct comparison is essential to determine the most suitable approach based on specific factors, including the complexity and size of the dataset [31]. Machine learning can exhibit strong potential to complement educational research with advanced modeling techniques applied to big educational data of great complexity.

III. Method and Instruments

A. Research Design

The research employed a quantitative methodology to explore the factors that affect students' academic performance. Data was collected via surveys focusing on prior education, university experiences, and socioeconomic backgrounds from 427 foundation-year students enrolled in science and engineering tracks at Imam Abdulrahman Bin Faisal University (IAU). By applying binary logistic regression, six key predictors were identified, and the relationships among various factors and performance levels were analyzed using chi-squared and t-tests. The dataset was divided, assigning 30% for testing and 70% for model training. Techniques like bootstrap sampling, data splitting, and randomization played a significant role in enhancing the study [32]. The classification of the students' data was performed using random forest, Naïve Bayes, and logistic regression approaches To ensure a thorough assessment, we calculated several machine learning metrics. This allowed us to make clearer comparisons between the three techniques and understand their significance in educational contexts.

B. Classification Methods

The database classification task is performed using supervised machine learning techniques such as logistic regression, naïve Bayes, and random forest, as explained below:

1) Logistic regression

Logistic Regression (LR) [33] is a widely used statistical method for classifying binary data. By considering the training data set $X \in R^{nxk}$, where n represents the data size while k represents the number of features, and let y_i be the binary outcome $y_i \in \{1,0\}$, y=1 for the positive class with probability π and 0 for the negative one with probability $1-\pi$. The goal is to classify the instance x_i as positive or negative. By assuming the independence of the training features, the logistic model expresses the conditional probabilities associated with the instance x_i as follows [34]:

$$p(y_i = 1)) = \pi_i = \frac{exp\left(\beta_0+\beta_1 X_1+\beta_2 X_{2+\cdots+\beta_k X_k}\right)}{1+exp\left(\beta_0+\beta_1 X_1+\beta_2 X_{2+\cdots+\beta_k X_k}\right)} \quad i = 1,2,\ldots\ldots,n \tag{1}$$

where $x_1, x_2, \ldots\ldots, x_k$ are the model predictors and $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are regression parameters as β_0 is the regression intercept.

The odd ratio (OR) of $y_i = 1$ is defined as

$$odd = \frac{p(y_i=1)}{1-p\,(y_i=1)} = \frac{\pi_i}{1-\pi_i} \tag{2}$$

The logistic (logit) transformation is the logarithm of the odds is defined as

$$logit\,(y_i = 1) = ln(odd) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x + \cdots. + \beta_k x_k \tag{3}$$

The transformation in Eq. (3) has lots of the desirable properties of the linear regression model. The logit is linear in the regression parameters. These parameters are often

estimated with the maximum likelihood (ML) function. This function is defined as:

$$L(\beta) = \prod_{i=1}^{n}(\pi_i)^{y_i}(1-\pi_i)^{1-y_i} = \prod_{i=1}^{n}\left[\frac{exp\left(\beta_0+\beta_1X_1+\beta_2X_{2+\cdots+\beta_k}x_k\right)}{1+exp\left(\beta_0+\beta_1X_1+\beta_2X_{2+\cdots+\beta_k}x_k\right)}\right]^{y_i}\left[\frac{1}{1+exp\left(\beta_0+\beta_1X_1+\beta_2X_{2+\cdots+\beta_k}x_k\right)}\right]^{1-y_i} \quad (4)$$

and the log-likelihood is:

$$lnL(\beta) = \sum_{i=1}^{n}[y_i\, ln(\frac{exp\left(\beta_0+\beta_1X_1+\beta_2X_{2+..+\beta_k}x_k\right)}{1+exp\left(\beta_0+\beta_1X_1+\beta_2X_{2+..+\beta_k}x_k\right)}][(1-y_i)ln(\frac{1}{1+exp\left(\beta_0+\beta_1X_1+\beta_2X_{2+..+\beta_k}x_k\right)}] \quad (5)$$

The maximum likelihood estimates (MLE) for logistic regression are obtained through numerical optimization techniques. The predicted class for the logistic regression model will be as shown below:

$$if\ (\hat{\pi}_1 \geq 0.5)\ the\ instance\ is\ classified\ to\ the\ class\ y = 1 \quad (6)$$

$$if\ (\hat{\pi}_1 < 0.5)\ the\ instance\ is\ classified\ to\ the\ class\ y = 0$$

For testing the significance of each parameter, the Wald statistic is used and computed for each parameter as:

$$w_j = \frac{\beta_j^{\,2}}{SE\,\beta_j^{\,2}} \quad (7)$$

The Wald statistic has a chi-square distribution with 1 degree of freedom, it compares with a critical value of chi-square.

2) Naïve bayes

Naïve Bayes [35] is a Bayesian Network Classifier that is highly efficient for inductive learning in machine learning and data mining, utilizing data to make accurate predictions based on Bayes' rule. For the current study's data training data set $x \in R^n$ where x_i is the ith attribute in x, with an associated target variable $y \in \{-1,+1\}$ as the paper deals only with binary classification. According to Bayes rule the probability of x_i being classified to $y \in \{-1,+1\}$ is as follows [35]:

$$p(x_i/y) = \frac{p((x_i).p((y/x_i)}{p(y)} \quad (8)$$

Given the value of the class, it is assumed that all the attributes are independent, the probability of an instance $E_k = (x_1\ , x_2\ , \ldots, x_n\ )$ where k = 1,2, …, m is being classified to $y \in \{-1,+1\}$ is:

$$p((x_1\ ,x_2\ ,\ldots,x_n/y) = \prod_{i=1}^{n}\frac{p((x_i).p((y/x_i)}{p(y)} \quad (9)$$

Instance E is classified $y = +1$ if and only if

$$(E_k) = \frac{p(y=+1/(x_1\ ,x_2\ ,\ldots,x_n)}{p(y=-1/(x_1\ ,x_2\ ,\ldots,x_n)} \geq 1 \quad (10)$$

where $n_b(E_k)$ is called a Bayesian classifier.

Probability estimates are usually derived from the frequency counts using smoothing functions such as the Laplace estimate.

3) Random forest

The Random Forest [36]is a widely used machine-learning technique for regression and classification problems, utilizing decision trees and ensemble learning to tackle complex problems. It consists of multiple decision trees, each associated with bootstrap samples from the original data set. The nodes are divided based on the entropy associated with a subset of features. The algorithm then employs the bagging technique, bootstrap aggregating, to select the best trees. Bagging repeatedly selects a random sample of features with replacements from the training set and fits trees to these samples. This method leads to significant improvements in performance and accuracy compared to using individual classifiers. The steps for constructing a random forest are as follows [10]:

- Choose a random subset of features from the original training dataset using bootstrapping.
- Construct a decision tree using this bootstrapped sample.
- Determine the desired number of trees, N, to build.
- Make predictions for the class label, y, of each decision tree and assign the new data points to the category that receives the majority votes.
- Repeat the above steps.
- Combine all the predicted y values together.

These steps are shown in Fig. 1:

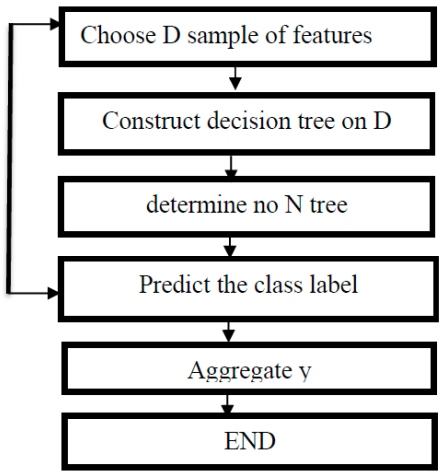

Fig. 1. Random forest' flow chart.

C. Data Sample

A structured questionnaire was administered to 427 out of 700 engineering and science students enrolled at Imam Abdulrahman Bin Faisal University's Deanship of Preparatory Year and Supporting Studies. This sample was selected using simple random sampling (SRS). Among the respondents, 311 were male, representing 73% of the total, while 116 were female, accounting for 27%. Within the cohort of students enrolled in the preparatory year during the 2021-2022 academic year, 283 were in the engineering track, making up 66%, and 144 were in the science tracks, which comprised 34%. The dataset, gathered after the first semester, includes 22 predictor variables, such as socioeconomic background, secondary school factors like entrance examination scores and school type, and university-related elements like proficiency in English and mathematics. Additionally, other variables could also influence students' academic performance.

D. Hypothesis Testing

The classification task involves a comprehensive examination of how students' socioeconomic, family, and educational backgrounds relate to their academic performances. This study utilizes statistical methods, including chi-square tests and two-sample t-tests, to explore the effects of secondary school entrance exam scores and the number of family members on student performance. Specifically, the chi-square test is used to assess whether factors such as gender, academic track (science or engineering), levels of English proficiency (advanced, intermediate, beginner), mathematics skills (advanced or not), type of school (public or private), geographic location, parents' jobs, and education, along with students' work status (yes or no), influence their performance in the first semester. For the analysis, we set the significance level at $\alpha = 0.05$ for the t-test and chi-square test.

E. Experiment Setup

The entire data set is used to identify the significant factors. To perform the classification task, the original data set is split into two sets: 70% of the data is utilized for training the model, while the remaining 30% is reserved for testing the model. The logistic regression is conducted using SPSS_22, a widely used statistical software, to identify significant variables based on the complete data set. The backward-selecting procedure was used to select the significant variables. Hosmer and Lemeshow's [33] goodness of fit test is used to assess the model. 0.05 was used as a level of significance; only these significant variables were then employed to compute the classification measures on R version 4.1.2. The experiment involves a binary classification of the training data sets using logistic regression, naïve Bayes, and random forest. These methods are implemented following standard approaches. The metrics and ROC curve are calculated based on the test data.

IV. Results

A. The Significant Factors

The logistic regression results presented in Table 1 highlight the key factors that affect students' performance. Notably, gender, school type, secondary school score, and proficiency in English and math play significant roles. Specifically, the analysis indicates that gender is a considerable influencing factor (Wald = 6.961; sig = 0.008), with female students showing a 2.058 times greater likelihood of performing well compared to their male counterparts. Additionally, the type of school attended (Wald = 4.195; sig = 0.041) is also significant, as students from private schools tend to perform better, with an odds ratio of 0.586. Furthermore, students' secondary school scores, along with their proficiency levels in English and math have a positive effect on their performance, indicated by Exp(B) values greater than one. Interestingly, the desired college appears to be a protective factor for performance, reflected by an Exp(B) value of less than 1. In the table below, the results of the Hosmer and Lemeshow goodness of fit test show a p-value of 0.258, suggesting there's no evidence of a lack of fit in the model.

Table 1. Logistic regression' significant variables results

Variable	B	S.E.	Wald	Df	Sig.	Exp(B)
Gender	0.722	0.274	6.961	1	0.008	2.058
School type	−0.534	0.261	4.195	1	0.041	0.586
School score	0.117	0.031	13.864	1	0.000	1.124
English level	0.412	0.154	7.159	1	0.007	1.510
Math Level	0.572	0.223	6.596	1	0.010	1.772
college interest	−0.224	0.091	6.097	1	0.014	0.800
Constant	−11.543	2.846	16.445	1	0.000	0.000
Hosmer and Lemeshow Test						
Chi-square	10.108	**Df**	8	**Sig.**	0.258	

B. Relationship between Students' Performance and the Categorical Significance Variables

Table 2. Chi-square test result of the categorical factors and the student's performance levels

Variable	Chi-Square value	Df	p_value
Gender	10.571a	1	0.001
English_level	14.193a	2	0.001
School_type	2.228a	1	0.136
Math_level	13.691a	1	0.000
College_interest	20.295a	4	0.000

The chi-square test results presented in Table 2 indicate that gender, English proficiency, and math level have significant correlations with student performance. However, school type did not have a significant impact. The college interest suggests a strong influence on student performance

C. Relationship of Students Performance and Secondary School Scores

The t-test results shown in Table 3 compared secondary school scores and performance levels between the students' classes. Results showed that class one (students with GPA greater than or equal to 4) had a higher mean secondary school score (92.72) than class two (students with GPA less than 4), with a t-value of -5.758 and a p-value of 0.00 lower than the significance level of 0.05. This indicates that secondary school scores significantly impact students' performance levels, with class one having higher average scores (-2.087). This suggests that secondary school scores positively and significantly influence students' future academic success.

Table 3. Independent sample t-test of the secondary school score and the student's performance levels

Descriptive statistics of classes					
	Classes	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
School _score	1	160	90.6319	3.91989	0.30989
	2	267	92.7189	3.43755	0.21037

Independent sample _T test						
T_value	**Df**	**P_value**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the Difference**	
					Lower	Upper
−5.758	425	0.000	−2.08702514	0.36246724	−2.79947677	−1.37457351

D. Classifier Results by Performance Measures

The classification performance of the logistic regression, Naïve Bayes, and random forest algorithms is summarized in Table 4 using several machine learning metrics. The random forest algorithm stood out with the highest performance, achieving an accuracy of 78.13. Naïve Bayes followed closely with a respectable accuracy of 76.56. In contrast, logistic regression recorded the lowest values across all metrics, with an accuracy of 72.63. This indicates that random forest delivered the best classification performance among the three algorithms.

Table 4. The classification performance measurements results

Classifier	Accuracy	Specificity	Sensitivity	Precision	F_score	Kappa	AUC
Logistic Regression	0.7263	0.8293	0.5435	0.6410	0.5882	0.3710	0.717
Naïve Bayes	0.7656	0.8642	0.5957	0.7179	0.6512	0.4770	0.753
Random Forest	0.7813	0.8765	0.6170	0.7436	0.6744	0.5169	0.843

E. ROC Curve Analysis

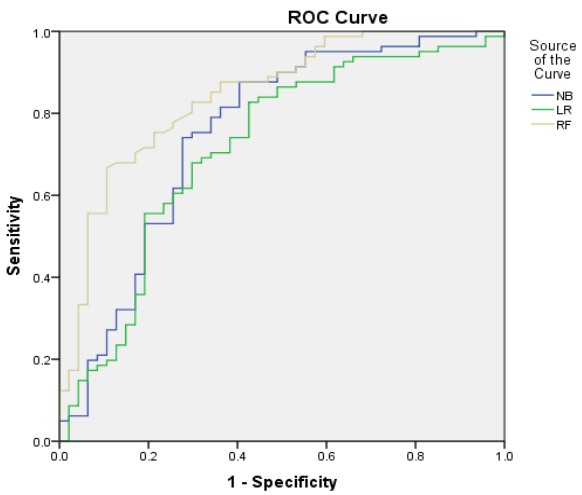

Fig. 2. The Roc curves of RF, NB, and LR.

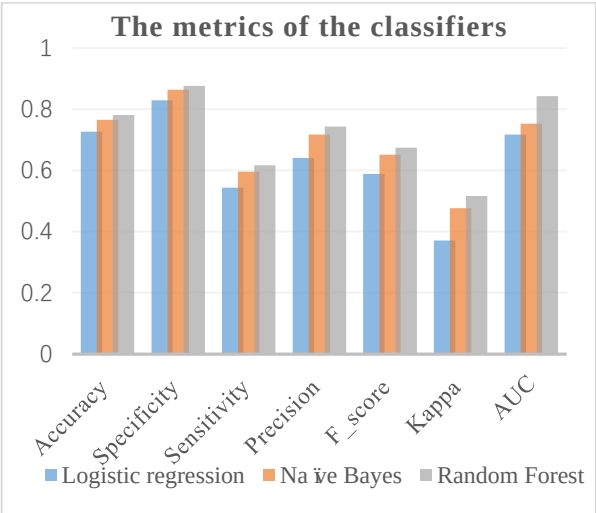

Fig. 3. The metrics for RF, NB, and LR.

The ROC curves displayed in Fig. 1 illustrate the performance of the naïve Bayes, logistic regression, and random Forest models. It's clear from the figure that the random Forest model consistently excels over naïve Bayes and logistic regression, achieving a higher area under the ROC curve (AUC) for the student's data. Consequently, random Forest stands out as the leading classifier based on AUC, with naïve Bayes following behind, while logistic regression performs the poorest. Fig. 2 presents the metrics for logistic regression, naïve Bayes, and random forest.

V. Discussion

The current study focused on exploring the factors that influence the performance of foundation year students. Previous research showed that foundation-year student success is influenced by academic achievement, social-emotional well-being, and critical thinking skills. Further, student retention is largely determined by academic performance, while academic adjustment is controlled by intrinsic motivation and degree program satisfaction [30, 37, 38]. By employing machine learning techniques, the study aimed to pinpoint various predictors of learning outcomes. It assessed the effectiveness of machine learning techniques such as logistic regression, Naïve Bayes, and random forest in classifying educational data, ultimately revealing valuable insights. Logistic regression was employed to identify the key factors affecting students' success in engineering and science programs at Imam Abdulrahman Bin Faisal's Deanship of Preparatory Year and Supporting Studies. The backward selection method was used to select the significant factors. The results are presented in Table 1. The table shows that the most significant factors are gender, school type, and scores in secondary school. English and mathematics proficiency and the desired college as their p-values are below the significance level of 0.05. Five categorical factors were identified as important, and the chi-square test was conducted to assess their relationship with student performance. The results of the chi-square test are shown in Table 2. The results are consistent with the logistic regression results, except for the school type factor, which has an insignificant impact on

student performance, as its P-value = 0.136, which is above the significance level of 0.5. For testing the significance of the quantitative factor of secondary school scores on student performance, the independent t-test was used; the results are shown in Table 3. The results of the independent t-test show the influence of secondary school grades on student achievement (t = -5.758, df = 425), with slight differences between the groups with higher (mean = 92.72, SD = 3.437) and lower achievement (mean = 90.63, SD = 3.919). The effect size of secondary school outcomes was moderate to large, which is consistent with previous literature. Research has shown that secondary education improves students' academic achievement in math and reading comprehension, but the effect size varies and decreases over time [39]. Elementary school test scores significantly predict secondary school outcomes, university enrollment, and hourly earnings [40].

Chi-square test and independent samples t-test results are reliable with logistic regression results. Consequently, the most important factors influencing student performance are gender, type of school, and score in secondary school. English and math proficiency and interest in college. These 6 factors are used for the classification task. Machine learning algorithms are applied to these important factors. Logistic Regression, Naïve Bayes, and Random Forest are applied as usual using the standard approach. The classification results of the three methods for the different machine learning measures are shown in Table 4. It can be seen from the table that Random Forest has the best overall classification performance compared to the other machine learning algorithms, indicating that it is well suited for predictive modeling applications in higher education.

The results of the study show that Random Forest has the best classification performance in evaluating machine learning algorithms with high accuracy (78.3%), specificity (87.65%), sensitivity (61.7%), precision (87.6%), F-score (67.44%), and AUC (8.043). In contrast, the Naïve Bayes algorithm outperformed logistic regression with an accuracy of 76.56%, a specificity of 86.42%, a sensitivity of 59.57%, a precision of 71.79%, an F-score of 65.12%, and an AUC of 0.753. However, some researchers believe that logistic regression is superior to Naïve Bayes or Random Forest in certain contexts. The random forest algorithm outperforms logistic regression in 69% of the data sets, emphasizing the need for careful selection and variant evaluation. It can accurately predict the failure of college examinations [41–43]. In addition, conflicting results have been reported regarding the effectiveness of different algorithms in predicting students' academic performance. In the existing literature, Bayesian algorithms, linear regression, logistic regression, k-nearest neighbor, and decision trees are reported as effective methods. The Naive Bayes algorithm has been shown to accurately predict student performance based on factors such as field of study, place of residence, relationships, occupation, and scholarships. However, the causal pathways and directionality remain unclear. Longitudinal studies and control of covariates could help to clarify these dynamics and deepen understanding [44, 45].

VI. Conclusion

Assessment of student performance during the foundation year serves as a key baseline for colleges, assessing and selecting new students and tracking the effectiveness of their teaching initiatives. This study aims to discover the characteristics impacting student performance in engineering and science programs at Imam Abdulrahman Bin Faisal University's Deanship of Preparatory Year and Supporting Studies. The study analyzes 22 predictor factors, including demographics, secondary school performance, language proficiency, and university experiences. Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF) algorithms are used for the classification task. The findings indicated that performance was highly affected by gender, school type, English/math proficiency, college interest, and secondary school test results. All three methods successfully classified the data. The most effective classification method was random forest, which indicated the possibility for predictive modeling. However, the results are qualified by the limitations in the self-reported data and the absence of a causal analysis. By optimizing learning experiences and results through focused, data-driven methods, the findings highlight the potential to promote diversity and excellence in higher education and provide suggestions for improving support services to suit the requirements of diverse students. The study revealed preliminary insights into student performance drivers as well as the potential benefit of random forest. However, further research is needed to draw more robust conclusions. Furthermore, utilizing different designs and samples, future research should explore generalizability and causal processes; future work could be conducted for all university students, resulting in more accurate results. The findings of this study essentially have applications for improving student support services, as they may be used to customize tutoring and advice to meet the requirements of certain groups by identifying critical performance-influencing elements. Priority areas for growth can also be determined by considering the influence of abilities like math and English proficiency.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The author confirms sole responsibility for the following aspects: the study's conception and design, data collection, analysis and interpretation of the results, and manuscript preparation

Acknowledgment

I would like to express my heartfelt gratitude to the staff members and students at Imam Abdulrahman Bin Faisal University who assisted, motivated, and encouraged me in successfully finishing this study during the academic year 2021–2022.

References

[1] A. Yamini and K. S. Rekha, "Improved accuracy for identifying at-risk students at different percentage of course length using logistic regression compared with random forest predictive model," in *Proc. 2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, Dec. 2022.

[2] M. S. AL-Mekhlafi, "Quality evaluation of postgraduate programs from the perspective of students (Faculty of Education, Imam Abdulrahman Bin Faisal University (IAU), KSA)," *Journal of Education in Black Sea Region*, vol. 5, no. 2, pp. 76–95, 2020.

[3] W. Al-Abweeny, A. Al-Hamad, and J. Al-Qudah, "The growth of knowledge society in Saudi universities using electronic learning tools: Imam Abdulrahman Bin Faisal University as a model," *Multi-Knowledge Electronic Comprehensive Journal for Education & Science Publications (MECSJ)*, no. 25, 2019

[4] M. W. Meyer and D. Norman, "Changing design education for the 21st century," *She Ji: The Journal of Design, Economics, and Innovation*, vol. 6, no. 1, pp. 13–49, 2020.

[5] P. S. Rao, "The importance of speaking skills in English classrooms," *Alford Council of International English & Literature Journal (ACIELJ)*, vol. 2, no. 2, pp. 6–18, 2019.

[6] A. A. Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques," *Technology, Knowledge and Learning*, vol. 24, pp. 567–598, 2019.

[7] V. L. Miguéis, A. Freitas, P. J. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decision Support Systems*, vol. 115, pp. 36–51, 2018.

[8] T. L. Strayhorn, *College Students' Sense of Belonging: A Key to Educational Success for All Students*, Routledge, 2018.

[9] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-learning: Challenges and research opportunities using machine learning & data analytics," *IEEE Access*, vol. 6, pp. 39117–39138, 2018.

[10] A. B. Musa, "SFM: A sequential fitting method to address the overfitting problem of logistic regression," *International Journal of Advances in Soft Computing & Its Applications*, vol. 15, no. 3, 2023.

[11] A. E. Tatar and D. Düştegör, "Prediction of academic performance at undergraduate graduation: Course grades or grade point average?" *Applied Science*, vol. 10, no. 14, p. 4967, 2020.

[12] A. Gupta, D. Garg, and P. Kumar, "Analysis of students' ratings of teaching quality to understand the role of gender and socio-economic diversity in higher education," *IEEE Transactions on Education*, vol. 61, no. 4, pp. 319–327, 2018.

[13] E. Fernandes *et al*., "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *Journal of Business Research*, vol. 94, pp. 335–343, 2019.

[14] M. A. Almaiah and A. Almulhem, "A conceptual framework for determining the success factors of e-learning system implementation using Delphi technique," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 17, pp. 5962–5976, 2018.

[15] S. P. Springer, J. V. Morgan, N. Griesemer, and J. Reider, *Admission Matters: What Students and Parents Need to Know about Getting into College*, Hoboken, NJ: John Wiley & Sons, 2023.

[16] R. K. Veluri *et al*., "Learning analytics using deep learning techniques for efficiently managing educational institutes," *Materials Today: Proceedings*, vol. 51, pp. 2317–2320, 2022.

[17] H. Abuhassna, F. Awae, M. A. Adnan, M. Daud, and A. S. Almheiri, "The information age for education via artificial intelligence and machine learning: A bibliometric and systematic literature analysis," *International Journal of Information and Education Technology*, vol. 14, no. 5, 2024.

[18] Р. К. Низамов, Д. М. Кордончик, А. В. Михеев, and Д. А. Сосунов, "Пленарное заседание," 2022.

[19] K. S. Rawat and I. Malhan, "A hybrid classification method based on machine learning classifiers to predict performance in educational data mining," in *Proc. 2nd International Conference on Communication, Computing and Networking: ICCCN 2018*, NITTTR Chandigarh, India, 2019.

[20] T. T. Huynh-Cam, L. S. Chen, and H. Le, "Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance," *Algorithms*, vol. 14, no. 11, p. 318, 2021.

[21] L. Masangu, A. Jadhav, and R. Ajoodha, "Predicting student academic performance using data mining techniques," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 1, pp. 153–163, 2021.

[22] K. Sixhaxa, A. Jadhav, and R. Ajoodha, "Predicting students' performance in exams using machine learning techniques," in *Proc. 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 635–640, Jan. 2022.

[23] K. Lavidas *et al*., "Determinants of humanities and social science students' intentions to use artificial intelligence applications for academic purposes," *Information*, vol. 15, no. 6, p. 314, 2024.

[24] R. Gomila, "Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis," *Journal of Experimental Psychology: General*, vol. 150, no. 4, p. 700, 2021.

[25] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," presented at IEEE Colombian Conference on Applications in Computational Intelligence, 2018.

[26] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," *Applied Science*, vol. 11, no. 7, p. 3130, 2021.

[27] K. Kargar, M. J. S. Safari, and K. Khosravi, "Weighted instances handler wrapper and rotation forest-based hybrid algorithms for sediment transport modeling," *Journal of Hydrology*, vol. 598, 126452, 2021.

[28] A. Al-Zawqari, D. Peumans, and G. Vandersteen, "A flexible feature selection approach for predicting students' academic performance in online courses," *Computers and Education: Artificial Intelligence*, vol. 3, 100103, 2022.

[29] G. Gorgun, S. N. Yildirim-Erbasli, and C. D. Epp, "Predicting cognitive engagement in online course discussion forums," *International Educational Data Mining Society*, 2022.

[30] P. A. Westrick *et al*., "The road to retention passes through first-year academic performance: A meta-analytic path analysis of academic performance and persistence," *Educational Assessment*, vol. 26, no. 1, pp. 35–51, 2021.

[31] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended isolation forest," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1479–1489, 2019.

[32] S. Jayaprakash, S. Krishnan, and V. Jaiganesh, "Predicting students' academic performance using an improved random forest classifier," presented at 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), 2020.

[33] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, Hoboken, NJ: John Wiley & Sons, 2013.

[34] A. B. Musa, "Comparative study on classification performance between support vector machine and logistic regression," *International Journal of Machine Learning and Cybernetics*, vol. 4, pp. 13–24, Feb. 2013.

[35] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems*, vol. 192, p. 105361, 2020.

[36] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *Journal of Petroleum Science and Engineering*, vol. 174, pp. 776–789, 2019.

[37] P. J. Van der Zanden, E. Denessen, A. H. Cillessen, and P. C. Meijer, "Domains and predictors of first-year student success: A systematic review," *Educational Research Review*, vol. 23, pp. 57–77, 2018.

[38] E. C. Rooij, E. P. Jansen, and W. J. Grift, "First-year university students' academic success: The importance of academic adjustment," *European Journal of Psychology of Education*, vol. 33, pp. 749–767, 2018.

[39] J. Dockx, B. Fraine, and M. Vandecandelaere, "Does the track matter? A comparison of students' achievement in different tracks," *Journal of Educational Psychology*, vol. 111, no. 5, p. 827, 2019.

[40] R. De Hoyos, R. Estrada, and M. J. Vargas, "What do test scores really capture? Evidence from a large-scale student assessment in Mexico," *World Development*, vol. 146, 105524, 2021.

[41] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, pp. 1–14, 2018.

[42] D. R. Nugroho *et al*., "Logistic regression and random forest comparison in predicting students' qualification based on students' half-semester performance," presented at 2023 11th International Conference on Information and Communication Technology (ICoICT), 2023.

[43] E. Heide *et al*., "Comparing regression, Naïve Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle," *Journal of Dairy Science*, vol. 102, no. 10, pp. 9409–9421, 2019.

[44] J. Kumari, R. Venkatesan, T. J. Jebaseeli, V. A. Felsit, K. S. Selvanayaki, and T. J. Sarah, "A comparison of machine learning techniques for the prediction of the student's academic performance," *Emerging Trends in Computing and Expert Technology*, 2020.

[45] N. Pandiangan, M. Lintang, and B. Priyudahari, "Naïve bayes for analysis of student learning achievement," *SHS Web of Conferences*, 2022.

Copyright © 2024 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).