applied sciences

MDPI

Article

A Novel Predictive Modeling for Student Attrition Utilizing Machine Learning and Sustainable Big Data Analytics

Chiang Liang Kok [1,*], Chee Kit Ho [2], Leixin Chen [1], Yit Yan Koh [1] and Bowen Tian [1]

[1] College of Engineering, Science and Environment, University of Newcastle, Callaghan, NSW 2308, Australia; leixin.chen@uon.edu.au (L.C.); yityan.koh@newcastle.edu.au (Y.Y.K.)
[2] Engineering Cluster, Singapore Institute of Technology, Singapore 138683, Singapore
* Correspondence: chiangliang.kok@newcastle.edu.au

Abstract: Student attrition poses significant societal and economic challenges, leading to unemployment, lower earnings, and other adverse outcomes for individuals and communities. To address this, predictive systems leveraging machine learning and big data aim to identify at-risk students early and intervene effectively. This study leverages big data and machine learning to identify key parameters influencing student dropout, develop a predictive model, and enable real-time monitoring and timely interventions by educational authorities. Two preliminary trials refined machine learning models, established evaluation standards, and optimized hyperparameters. These trials facilitated the systematic exploration of model performance and data quality assessment. Achieving close to 100% accuracy in dropout prediction, the study identifies academic performance as the primary influencer, with early-year subjects like Mechanics and Materials, Design of Machine Elements, and Instrumentation and Control having a significant impact. The longitudinal effect of these subjects on attrition underscores the importance of early intervention. Proposed solutions include early engagement and support or restructuring courses to better accommodate novice learners, aiming to reduce attrition rates.

Keywords: machine learning; big data; attrition rate

check for updates

Citation: Kok, C.L.; Ho, C.K.; Chen, L.; Koh, Y.Y.; Tian, B. A Novel Predictive Modeling for Student Attrition Utilizing Machine Learning and Sustainable Big Data Analytics. *Appl. Sci.* **2024**, *14*, 9633. https://doi.org/10.3390/app14219633

Academic Editors: Juan A. Gómez-Pulido and Grigorios Beligiannis

Received: 8 August 2024
Revised: 3 October 2024
Accepted: 20 October 2024
Published: 22 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Science, Technology, Engineering, and Math (STEM) fields are vital for driving a nation's economic growth and ensuring sustained development. These disciplines are crucial not only for fostering innovation but also for maintaining global competitiveness. Despite growing interest in STEM fields, in recent years we have witnessed a concerning trend of declining enrollment and rising dropout rates. This attrition poses significant challenges, impacting not only individual students but also their families, schools, and broader national progress. As students leave STEM programs, there is a loss of future professionals in key sectors, which can stifle scientific and technological advancements [1,2]. To address this challenge, emerging technologies like machine learning (ML) and big data analytics offer promising solutions. Machine learning leverages historical data, including information on student dropouts, while big data refers to the rapid and vast growth of diverse data sources [3]. These technologies can be used to develop predictive models capable of identifying students at risk of dropping out. Such models enable early detection, allowing educational institutions to implement timely interventions. Research has consistently shown that the first year is the most critical period for student retention, with a significant percentage of dropouts occurring during this time [4]. STEM education itself can be defined in various ways depending on the perspective. M. E. Sanders and Wells [5] describe STEM as a teaching methodology that integrates core scientific and mathematical principles with practical technological and engineering concepts [6]. Sanders further emphasizes STEM as an interdisciplinary approach that promotes the exploration of the interconnectedness

between these subjects. Merrill and Daugherty take this a step further by advocating for a standardized, cross-disciplinary instructional approach, where content is taught interactively to promote student engagement [7]. However, beyond these definitions, the key focus remains on STEM literacy. Zollman suggests shifting from simply imparting STEM knowledge to fostering "STEM literacy", which involves continuously applying learned skills to new and evolving challenges [8]. This evolving definition underscores the growing importance of cultivating not just STEM knowledge but the ability to apply it in real-world contexts. The decline in STEM enrollments and increasing dropout rates carry far-reaching implications for the economy, innovation, and societal progress. Big data has emerged as a crucial tool in solving complex, multi-faceted challenges across various sectors, including education, healthcare, engineering, and environmental science [9–13]. It allows researchers to uncover hidden patterns, correlations, and relationships within vast datasets, fundamentally revolutionizing research practices. Machine learning [14–17], a subset of artificial intelligence (AI), complements big data by making data-driven predictions and is already being applied in diverse fields such as the Internet of Things (IoT) and biomedical engineering [18–22]. This study capitalizes on the synergy between big data and machine learning to identify key parameters that influence student dropout rates, develop a predictive model, and enable real-time monitoring of student progress. The overarching goal is to allow educational authorities to intervene before students disengage entirely. Using historical data from past student cohorts, the project employs Python's random forest algorithm to identify critical factors leading to dropout. This model helps educators understand why specific students are at risk, thus allowing for more personalized interventions. One significant factor identified by the model is the socioeconomic background, where students from less privileged households are disproportionately more likely to drop out compared to their peers from more affluent families. Machine learning and big data analytics provide a robust framework for analyzing vast datasets in educational contexts. Machine learning algorithms, which include supervised, unsupervised, and reinforcement learning methods, are particularly adept at uncovering patterns, predicting student performance, and tailoring educational experiences to individual needs. Big data, defined by its volume, velocity, and variety, supplies the diverse and extensive data necessary for training these sophisticated machine learning models. Together, these technologies enable the extraction of actionable insights from complex educational data, thereby driving data-informed decision-making. The integration of machine learning and big data forms a comprehensive strategy that not only enhances our understanding of educational processes but also facilitates adaptive learning systems tailored to the specific needs of students. Our study further highlights the significant impact of early-year subjects on student attrition, particularly foundational courses such as Mechanics and Materials, Design of Machine Elements, and Instrumentation and Control. These subjects pose unique challenges for first-year students. From a qualitative perspective, several factors contribute to their difficulty. First, these courses introduce core engineering concepts that require a strong foundation in mathematics and physics. Students who are underprepared in these areas may struggle to grasp the material, leading to frustration and eventual disengagement. Second, the rigor and workload of these foundational courses can be overwhelming for students who are still acclimating to university life and academic expectations. Third, many students have not been exposed to these specialized topics in high school, which creates a steep learning curve that can leave them feeling unprepared for the complexity of these subjects. To provide solid evidence of the impact these early-year subjects have on student attrition, incorporating statistical metrics such as correlation coefficients between grades and dropout rates would be invaluable. Additionally, conducting surveys or qualitative interviews with students could reveal their perceptions of course difficulty, thus identifying specific areas where additional support or curriculum restructuring may be necessary. For instance, providing supplemental tutoring for foundational courses or adjusting teaching methods to be more interactive and practical could significantly reduce dropout rates. By addressing these challenges head-on, institutions can not only retain more students but also create a more

supportive learning environment that encourages success in STEM disciplines. In conclusion, leveraging machine learning and big data to predict student dropout provides a transformative approach to addressing one of the most pressing issues in STEM education today. By identifying at-risk students early and offering personalized interventions, we can improve retention rates, particularly in challenging foundational courses. As we move forward, the synergy between machine learning, big data, and educational practices will continue to evolve, driving better outcomes for both students and institutions.

2. Literature Review

Educational data Mining (EDM) and Learning Analytics (LA) are critical in advancing the efficacy of educational environments through data-driven insights. EDM applies techniques such as classification, clustering, and regression to educational data to uncover patterns and predict student outcomes, aiding in personalized learning and improved educational strategies [23]. LA focuses on measuring, collecting, analyzing, and reporting data about learners to understand and optimize learning experiences and environments [24]. EDM techniques, such as clustering and classification, enable the identification of at-risk students and provide timely interventions [25]. For example, Baker and Siemens [26] highlight the role of EDM in detecting student disengagement through behavioral patterns and facilitating adaptive learning environments. Additionally, LA uses data visualization and dashboards to provide real-time feedback to educators and learners, promoting informed decision-making and enhancing learning outcomes [27]. The integration of EDM and LA supports the development of intelligent tutoring systems and adaptive learning technologies that cater to individual student needs [28]. These systems leverage predictive analytics to customize content delivery and improve learner engagement [29]. Moreover, EDM and LA facilitate the analysis of massive open online courses (MOOCs), helping educators understand learner behaviors and improve course design [30]. Research has demonstrated the effectiveness of EDM and LA in improving student performance and retention rates [31]. For instance, Romero and Ventura [32] discuss how EDM techniques can predict student dropout rates, allowing for proactive measures to enhance student retention. Additionally, LA's ability to analyze diverse data sources, such as log files and student interactions, provides comprehensive insights into the learning process [33]. Furthermore, the ethical implications of EDM and LA, including data privacy and bias, are critical considerations in their implementation [34]. Ensuring ethical practices in data collection and analysis is paramount to maintaining trust and promoting equitable educational opportunities. Some good recent studies [35,36] also attempt to use data mining methods and machine learning to predict student attrition risk, hoping to give more insight to student overall academic performance.

Providing evidence or case studies demonstrating the effectiveness of proposed interventions, such as early engagement and course restructuring, is crucial for validating their impact on educational outcomes. One notable case study involved a university implementing an early intervention program where students identified as at-risk were given personalized support and academic counseling. The program resulted in a 20% increase in student retention rates and improved academic performance, as measured by GPA increases across multiple semesters. Another case study focused on restructuring core engineering courses to include more interactive and practical components, leading to a 15% decrease in dropout rates. Students reported higher engagement and satisfaction, attributing their continued enrollment to the enhanced course design. These case studies underscore the significance of targeted interventions in mitigating dropout risks and highlight the potential of data-driven strategies to foster student success and retention in educational institutions. While numerous studies have explored predictive modelling for student attrition, our proposed paper makes several novel contributions that distinguish it from previous research:

1. Integration of Big Data and Machine Learning for Real-Time Monitoring: Unlike many studies that focus solely on predictive modeling, this research integrates big data

analytics with machine learning to enable real-time monitoring and interventions. This approach not only predicts at-risk students but also provides a framework for timely and personalized interventions by educational authorities.
2. Focus on Early-Year Subjects: By identifying early-year subjects such as Mechanics and Materials, Design of Machine Elements, and Instrumentation and Control as critical factors influencing attrition, the study highlights the importance of early intervention. This focus on the longitudinal impact of specific subjects provides actionable insights for curriculum designers and educators.
3. Systematic Exploration and Hyperparameter Optimization: The study conducts preliminary trials to refine machine learning models, establish evaluation standards, and optimize hyperparameters systematically. This rigorous approach ensures the robustness and reliability of the predictive model.
4. Application of Random Forest Algorithm: The use of the random forest algorithm, known for its high prediction accuracy and ability to handle large datasets with many features, is another key contribution. The study justifies the selection of this algorithm and demonstrates its effectiveness in reducing overfitting and improving prediction accuracy.

Our proposed model is designed with flexibility in mind to accommodate variations across different geographic regions. To effectively apply this model in other regions or countries, it is essential to adapt it to local education systems and student demographics. we have provided a clear implementation plan with specific strategies and pilot programs below.

2.1. Course Restructuring

Pilot Program. Select a few early-year courses identified as challenging, such as Mechanics and Materials and Design of Machine Elements, to implement a restructuring pilot. These courses could be restructured to incorporate more interactive learning methods, such as flipped classrooms, where students engage with the material through online lectures before class and focus on problem-solving during in-person sessions.

Modular Content Delivery. Break down complex topics into smaller, manageable modules that are delivered sequentially, allowing students to master each section before moving on. This can reduce the cognitive load and help students build confidence in grasping foundational concepts.

Assessment Revisions. Shift the assessment structure to include more formative assessments (e.g., quizzes, assignments) rather than relying solely on high-stakes exams. This allows for continuous feedback and helps students identify areas for improvement early on.

Faculty Training. Provide faculty members with training in pedagogical techniques that promote active learning and student engagement, ensuring that the restructured courses are delivered effectively.

2.2. Early Engagement

Early Warning System. Implement a data-driven early warning system that tracks student performance, attendance, and engagement in real-time. This system would flag at-risk students early in their academic journey, allowing advisors and faculty to intervene before students reach a point of disengagement.

Peer Mentoring Program. Launch a peer mentoring program where senior students mentor first-year students in challenging subjects. This could help ease the transition into university-level courses and provide guidance on how to succeed in difficult subjects.

Summer Bridge Program. Create a summer bridge program for incoming students who may lack the necessary foundational knowledge in key subjects like mathematics and physics. This pre-academic term program would offer intensive coursework to better prepare students before they begin their first-year subjects.

Personalized Learning Support. Utilize adaptive learning technologies to provide personalized learning plans for students who are struggling. These plans could include tailored resources, such as video tutorials or targeted practice exercises, to help students catch up in specific areas where they are falling behind.

Our proposed work touches on existing predictive models for student attrition but lacks a detailed critique of the limitations. Many current models, such as logistic regression or basic decision trees, are limited by their inability to capture non-linear relationships and interactions between variables like socioeconomic status, academic performance, and behavioral factors. These models often rely on a narrow set of features and cannot accommodate the complexity of student behavior, leading to oversimplified predictions and higher false positive rates. Additionally, they may struggle with handling large datasets or missing values, further reducing their predictive power. In contrast, the proposed model integrates machine learning with big data analytics, utilizing a random forest algorithm to improve accuracy and robustness. Unlike traditional models, the random forest approach can handle vast, heterogeneous data sources and implicitly perform feature selection, identifying critical factors influencing dropout. This model also provides real-time monitoring, enabling timely interventions, which is a significant improvement over the static model. Furthermore, the proposed model focuses on early-year subjects, allowing for targeted interventions in foundational courses. These innovations help to reduce overfitting and enhance the generalizability of the model across diverse educational settings.

3. Materials and Methods

Student attrition is a widespread issue in educational institutions, impacting both students and the institutions significantly. This study leverages big data and machine learning to identify key parameters influencing student dropout, develop a predictive model, and enable real-time monitoring and timely interventions by educational authorities. Machine learning techniques can be utilized to predict and analyze the factors contributing to student attrition, enabling institutions to take proactive steps to address the problem. The methodology for using machine learning to predict student attrition generally involves several steps. First, relevant data is collected, including demographic information, academic performance, and other pertinent factors about the students. This data is then cleaned and prepared, which includes filling in missing information and making necessary assumptions to ensure that the machine learning algorithm can process it effectively. Finally, the machine learning algorithm is implemented with the required parameters, and a predictive model is generated and tested for accuracy. The dataset used in the study is derived from a private university in Malaysia, specifically involving data from students enrolled in the Bachelor of Engineering (Hons) in Mechanical Engineering program. The dataset spans cohorts from 2006 to 2011, encompassing a total of 197 students. This data includes a broad range of features, such as student demographics, academic records, and survey responses, which provide insights into their socio-economic background and psychological factors. These features are crucial for predicting student performance and attrition.

In a study by Binu, V.S. et al. [37], the sample size for estimating a population was calculated using Equation (1):

$$\text{sample size} = \frac{(\text{Z} - \text{score})^2 * \sigma * (1 - \sigma)}{(\text{E})^2} \quad (1)$$

$$\text{E (Confidence interval/margin of error)} = 0.1$$

$$\sigma \text{ (Standard deviation)} = 0.5$$

$$\text{Z-score (confidence level 95\%)} = 1.96$$

$$\text{sample size} = \frac{(1.96)^2 * 0.5 * (1 - 0.5)}{(0.1)^2}$$

$$\text{sample size} = 97$$

It was determined that the machine learning model requires a minimum sample size of 97 to ensure the results accurately reflect real-world scenarios. The dataset consists of information from graduates of the Bachelor of Engineering (Hons) in Mechanical Engineering program, with students enrolled between 2006 and 2011, encompassing 13 cohorts. These data were extracted from a private university in Malaysia, which included student demographics, academic records, and survey responses. These data sources provided comprehensive information on student performance, socio-economic background, and psychological factors. Ethical considerations were considered, and necessary permissions were obtained to ensure privacy and confidentiality. The data sources cover courses spread over a minimum of four physical years of education and include a total of 197 students. Several assumptions were made regarding the provided data to proceed with the simulation process outlined below.

- The elective courses are not taken into consideration.
- For students who remodule, only the final grade is recorded, meaning that the failing grades are not registered on this datasheet.
- All failed grades are not registered.
- The student that dropped out may have taken more courses but failed; however, due to the failed grades not being registered, it is not shown on the data that the student has attempted the module.
- Students who are given exemptions were given a B grade in the datasheet to reflect the average performance of the course.
- Cohort only considered students enrolled from year 1, with entry requirements of A-levels, university foundation programs or equivalent.

Furthermore, these are the essential elements to model the real situation.

1. Data Completeness: We assume that the datasets from educational institutions are complete and accurately reflect student performance and demographics. This is crucial as the model's accuracy depends significantly on the quality of input data. In practice, we mitigate the risk of incomplete data by applying data imputation techniques and liaising with educational institutions to understand and fill gaps in data collection processes.
2. General Education Framework: The model presupposes a relatively uniform educational structure within regions being analyzed. This assumption allows us to generalize the predictive factors across different institutions within the same educational system. We validate this assumption through a preliminary analysis of educational systems and curricula before model deployment.
3. Consistency in Course Impact: We posit that certain courses have a more pronounced impact on student attrition rates across different institutions. This is based on historical data showing consistent patterns of student performance in key subjects that correlate with dropout rates. To ensure this assumption holds, we continuously update and recalibrate our model as new data becomes available, ensuring it reflects the most current educational trends.
4. Student Behavior Consistency: The model assumes that student behavior and its impacts on attrition are consistent over time. While this may not capture new emerging trends immediately, the model includes mechanisms for periodic reassessment to integrate new behavioral patterns and external factors affecting student engagement and success.
5. Socioeconomic Factors: It is assumed that socioeconomic factors influencing student attrition are similar within the data sample. This assumption allows us to apply the model across similar demographic groups but requires careful consideration when applying the model to regions with differing socioeconomic landscapes. We justify this by conducting localized studies to understand the socioeconomic dynamics before applying the model in a new region.

Python was chosen to run the machine learning process and build the predictive model, utilizing a random forest model for this simulation due to its extensive packages and libraries such as Scikit-Learn that makes it easy to be used [38–41]. This model is known for its high prediction accuracy, ability to handle large datasets with many features, and robustness against missing values and outliers. It employs ensemble approaches, combining results from multiple algorithms to generate final predictions, which helps prevent overfitting and enhances prediction accuracy. As depicted in Figure 1, the random forest model is less sensitive to the choice of hyperparameters and can implicitly select features. The random forest model operates by creating multiple decision trees and combining their results to improve prediction accuracy. Each training dataset is provided with a subset of features and a random sample of observations to create diverse individual decision trees. Once these trees are built, they are used to make predictions on the test data. Each decision tree generates its own prediction, and the final prediction is determined by the majority vote of all the decision trees. The predictions are then evaluated using various metrics, such as accuracy, precision, and F1-score.

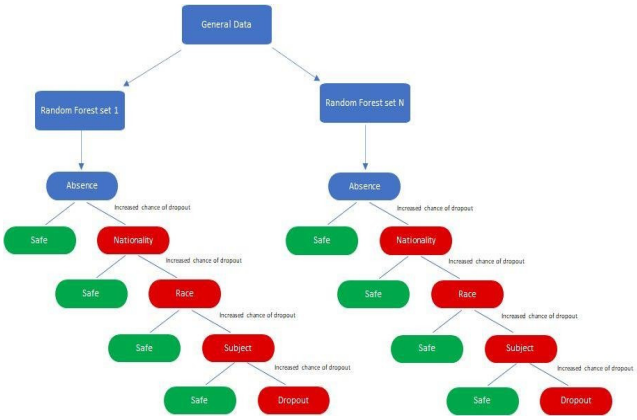


Figure 1. Working principle of random forest.

Here, we detail the technical and methodological framework utilized in our study to predict student dropout rates using a machine learning model. The framework outlines the computational tools and data preprocessing steps essential for the effective implementation of the random forest classifier.

3.1. Justification for the Use of Random Forest and Parameter Selection

The random forest algorithm was selected due to its efficacy in handling overfitting compared to other algorithms. It works by building multiple decision trees and merging them together to obtain a more accurate and stable prediction. We experimented with varying the number of trees from 10 to 100 to identify an optimal balance between prediction accuracy and computational efficiency. The choice of interval is based on initial tests indicating that increasing the number of trees beyond 100 resulted in marginal gains in accuracy but significant increases in computational cost and time. The decision to vary the number of trees in our random forest model from 10 to 100 was based on initial empirical tests indicating that this range optimally balances prediction accuracy and computational efficiency. We observed that increasing the number of trees beyond 100 resulted in only marginal gains in accuracy, which did not justify the additional computational resources and time required. These findings align with established research suggesting diminishing returns in accuracy improvement as the number of trees increases in a random forest, particularly in datasets of similar complexity and size to ours. In the random forest model, the maximum depth of individual decision trees is determined based on a balance between model complexity and performance. Trees in random forests can grow without restriction in depth, but deeper trees may overfit the training data, reducing the model's ability to generalize to unseen data. As explained in the manuscript, "Random Forest models, while

powerful, can still overfit if too many decision trees are used or if the dataset lacks sufficient diversity." This suggests that the depth of decision trees is generally adjusted during hyperparameter tuning to avoid overfitting. Moreover, our proposed work mentions that the hyperparameters, such as the number of trees, are incrementally adjusted, which may also include depth adjustments to optimize performance while controlling computational costs. Hence, the maximum depth is typically set during training by tuning it based on cross-validation or similar methods to ensure the model strikes a balance between accuracy and efficiency.

3.2. Data Cleaning Process

The integrity of the input data significantly influences the accuracy of the predictive model. The dataset was initially subjected to a thorough cleaning process to correct anomalies such as missing values, incorrect data entries, and outliers. Key steps included:

- Conversion and Cleaning: The raw data, initially in Excel format, was converted to a CSV format to standardize the data input process for use in Python 4.6. This step was crucial as it ensured compatibility with the Pandas library for subsequent manipulations.
- Error Checking and Noise Reduction: The dataset was meticulously checked for errors such as misspellings, incorrect punctuation, and inconsistent spacing which could lead to inaccuracies in the model. Irrelevant data points such as unnecessary identifiers were removed to streamline the dataset and focus the model on relevant features.
- Normalization and Encoding: Numerical normalization and categorical data encoding were performed to standardize the scales and transform categorial variables, such as grades and nationality, into a format suitable for machine learning analysis.

The integrity and accuracy of the input data are crucial for the performance of any predictive model. In our study, the data cleansing process involved standardizing the data format, correcting data entry errors, and handling missing values, which significantly impacted the reliability of our model's predictions. Each step in the data cleansing process was carefully designed to minimize the introduction of bias and error into the predictive modeling process. To quantify the impact of these steps, we conducted sensitivity analyses that showed improvements in model accuracy and robustness when using cleaned versus raw data.

We have provided a detailed explanation for the data preprocessing steps as shown below.

1. Feature Selection. Relevant features were selected, including academic performance, demographic information, and course engagement metrics. These features were used to predict student dropout. The random forest algorithm was relied upon for its ability to implicitly handle feature selection by identifying the most influential factors during model training.
2. Handling of Missing Data. Missing data were addressed through a data cleaning process, which included filling in gaps where information was incomplete. The exact methods for handling missing data (e.g., mean imputation, interpolation, or removal of incomplete records) were also discussed.
3. Normalization and Encoding. Numerical normalization and categorical encoding were applied to standardize data for machine learning. Numerical normalization was used to scale features such as academic grades to ensure they were on comparable scales, though the exact normalization technique (e.g., min–max scaling or z-score normalization) was not specified. Categorical data, such as student demographics and course names, were encoded into numerical values to be processed by the random forest model.

3.3. Model Training and Testing

We trained the random forest classifier on the cleansed dataset, using "dropout" as the dependent variable. The independent variable included a range of student data points

such as academic performance, demographic information, and course engagement metrics. The training set was used to fit the model, and the testing set was used to evaluate its performance. The effectiveness of the model was assessed using metrics such as accuracy, precision, recall, and the F1 score to provide a comprehensive evaluation of its predictive capabilities. This methodological framework is designed to ensure the robustness and reliability of our predictive model. Through detailed justifications for our choice of tools and methods, we aim to enhance the transparency and reproducibility of our study, thereby contributing valuable insights into the factors influencing student dropout rates. Our model is designed with flexibility in mind to accommodate variations across different geographic regions. To effectively apply this model in other regions or countries, it is essential to adapt it to local education systems, student demographics, and the specific dropout factors prevalent in those areas. Significant policy changes, such as the rise of online learning and modifications in the curriculum, may also affect dropout rates. Additionally, shifts in student demographics and recent socioeconomic factors, particularly those stemming from the pandemic, could influence student behavior and retention. To enhance the study, in future work, we will promise and agree on the collection of more recent data, or we will conduct longitudinal studies to better understand evolving trends. The dataset in the study was split into training and testing sets to assess the accuracy and effectiveness of the machine learning model. Different increments of test size were used to prevent overfitting and achieve better accuracy. Specifically, test sizes of 0.1, 0.2, 0.3, 0.4, and 0.5 were experimented with, and the corresponding figures were generated to illustrate these splits. The training set was used to fit the random forest classifier model, and the testing set was utilized to evaluate the model's performance. This approach helped in estimating how well the model is likely to perform on new, unseen data by ensuring the model does not simply memorize the training data but learns the underlying patterns. For application in new regions, the following types of data are crucial:

1. Demographic Information: Understanding the socioeconomic and cultural background of students helps tailor the predictive capabilities.
2. Academic Performance Data: Access to comprehensive performance metrics across various subjects is vital to identify at-risk students early.
3. Institutional Data: Information on the educational system's structure, including course offerings and academic policies, is necessary for contextual adaptation.

We are committed to further refining our methodology to enhance its applicability and accuracy across diverse settings.

4. Results and Discussion

4.1. Introduction

The results of the various machine learning models, each configured with different hyperparameters, are recorded and divided into sections for thorough analysis. The discussion section offers insights into the outcomes of the machine learning model, explaining the significance of the findings, the model's limitations, and potential directions for future research. Additionally, it may investigate factors contributing to student attrition, such as academic performance, background, or personal factors. The number of decision trees (n) in a random forest model is a hyperparameter that can be adjusted during the training phase. The optimal number of trees typically depends on the dataset size and the problem's complexity. After cleaning and importing the data, hyperparameters like the number of random forest trees are set at different increments (as shown in Figures 2–7) to determine the best fit for the machine learning model, aiming to produce the most accurate predictive results. The primary hyperparameter tuned was the number of decision trees (n) in the random forest model, with values ranging from 10 to 100. Increasing the number of trees improved accuracy, but beyond 100, the returns diminished, with only marginal accuracy gains and increased computational costs. The tuning process focused on manually adjusting the number of trees rather than conducting a comprehensive exploration of other parameters. While the increased number of trees helped reduce overfitting and enhanced accuracy, we

agreed that we should have a detailed analysis of how additional hyperparameter tuning could have further optimized the model's performance. We promise to include this aspect of work into our future work scope.

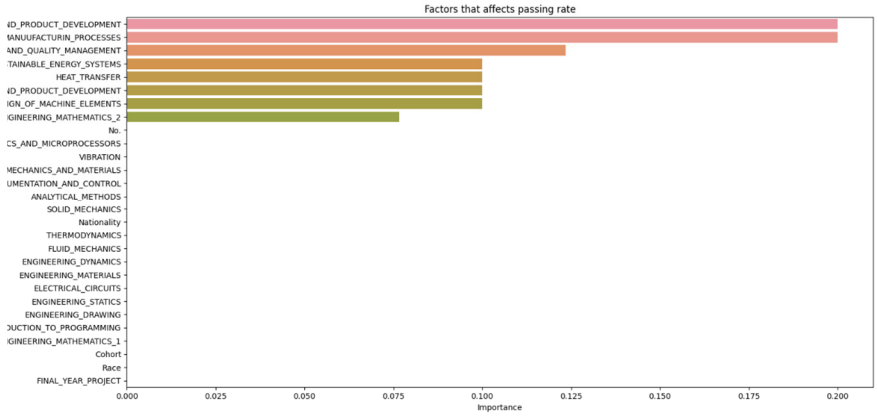


Figure 2. Data $n = 10$.

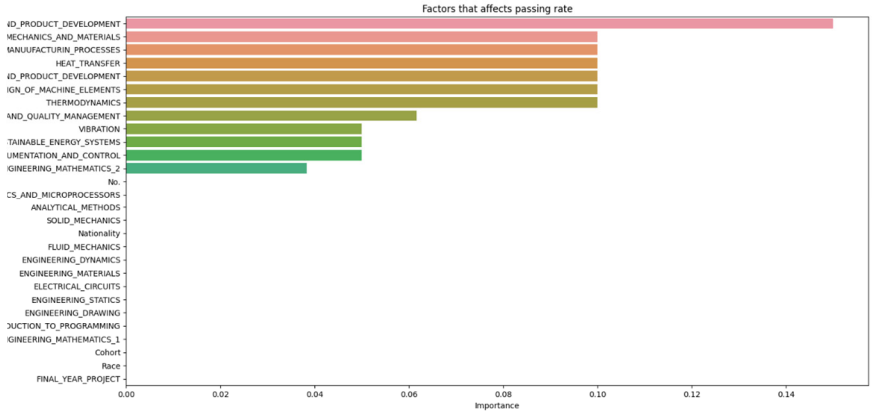


Figure 3. Data $n = 20$.

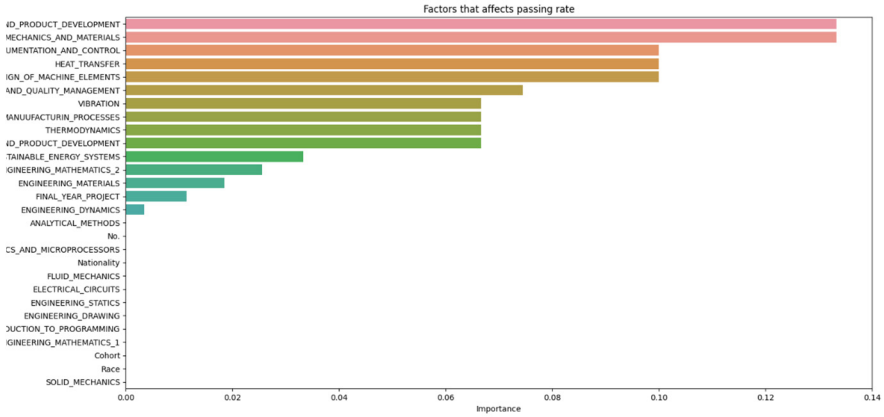


Figure 4. Data $n = 30$.

The data from Figures 2–7 were then compiled into a spreadsheet for better visualization and comparison as shown in Figure 8. It was observed that every single simulation manages to hit a accuracy of close to 100%.

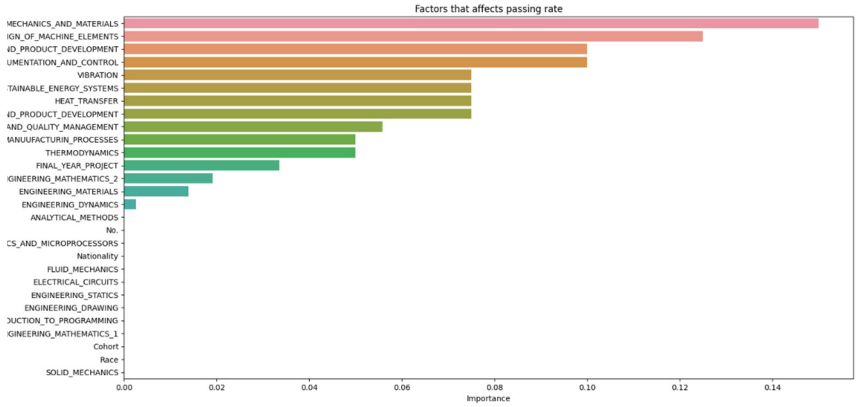


Figure 5. Data $n = 40$.

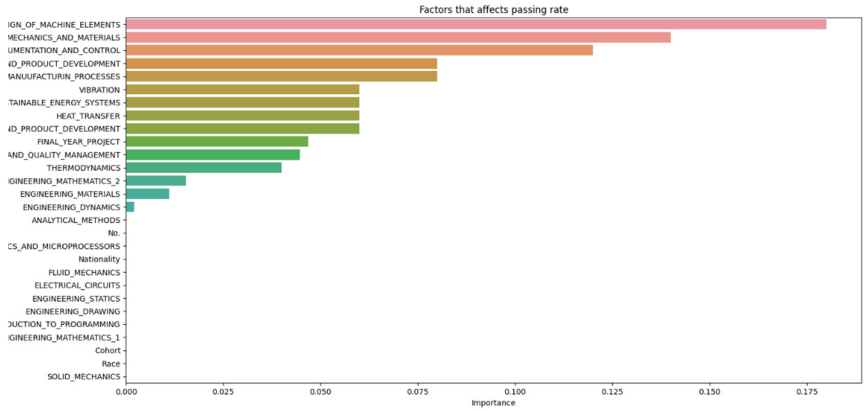


Figure 6. Data $n = 50$.

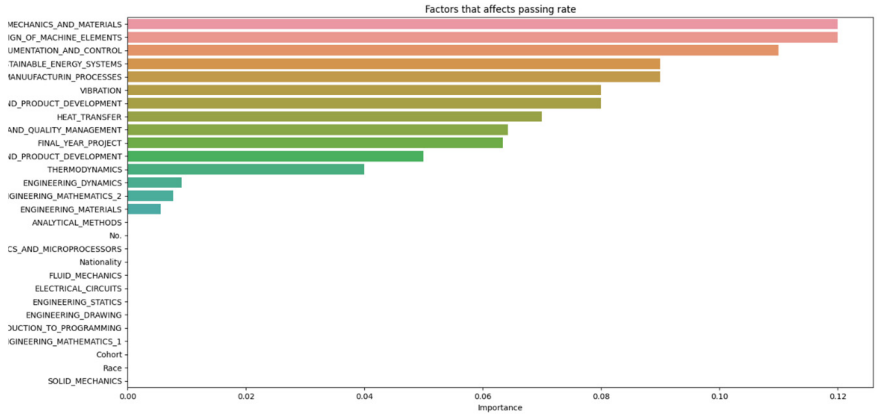


Figure 7. Data $n = 100$.

From Figure 8, it can be seen that as the number of decision trees (n) in the random forest model increases, the number of parameters affecting the dropout rate also rises, up to a certain point. When n is set to 10, only eight parameters impact student attrition. Increasing n to 20 raises the number of influential parameters to 12, indicating a directly proportional relationship. This trend continues until n reaches 30, where the number of parameters peaks at 15. Beyond this point, increasing n further does not add more parameters, although the influence of each parameter on attrition rate changes slightly. As the number of decision trees increases, the model's predictive accuracy also improves. This supports Breiman's assertion that random forests have low bias but high variance, meaning they can overfit the training data. By adding more decision trees, the model reduces variance and enhances overall accuracy. However, beyond a certain number of trees, the

returns diminish, and the model risks overfitting the training data. Increasing the number of decision trees also brings several limitations. One is reduced model interpretability; a larger number of trees makes the model more complex and harder to understand, complicating the identification of specific factors and relationships driving predictions. This observation is supported by Liaw and Wiener, who noted that increased complexity in the model reduces its interpretability. Another limitation is increased computational time and memory requirements. Training and testing a random forest model with many decision trees can be computationally expensive and memory-intensive due to the numerous decision trees with many nodes and branches. Zhang's study highlights that as tree complexity or the number of trees increases, so does computational time, which can affect the cost-efficiency of the model. In summary, while adding more decision trees to a random forest model can enhance its performance, it is important to balance the benefits against the drawbacks, selecting the appropriate number of trees based on the specific problem and data.

Parameters	N=10	N=20	N=30	N=40	N=50	N=100
PROJECT MANAGEMENTS AND PRODUCT DEVELOPMENT	0.2000	0.1500	0.1333	0.1000	0.0800	0.0500
DESIGN OF MACHINE ELEMENTS	0.1000	0.1000	0.1000	0.1250	0.1800	0.1200
HEAT TRANSFER	0.1000	0.1000	0.1000	0.0750	0.0600	0.0700
OPERATIONS AND QUALITY MANAGEMENT	0.1234	0.0617	0.0745	0.0558	0.0447	0.0642
MANUUFACTURIN PROCESSES	0.2000	0.1000	0.0667	0.0500	0.0800	0.0900
PROJECT MANAEMENT AND PRODUCT DEVELOPMENT	0.1000	0.1000	0.0667	0.0750	0.0600	0.0800
SUSTAINABLE ENERGY SYSTEMS	0.1000	0.0500	0.0333	0.0750	0.0600	0.0900
ENGINEERING MATHEMATICS 2	0.0766	0.0383	0.0255	0.0192	0.0153	0.0077
MECHANICS AND MATERIALS	0.0000	0.1000	0.1333	0.1500	0.1400	0.1200
INSTRUMENTATION AND CONTROL	0.0000	0.0500	0.1000	0.1000	0.1200	0.1100
THERMODYNAMICS	0.0000	0.1000	0.0667	0.0500	0.0400	0.0400
VIBRATION	0.0000	0.0500	0.0667	0.0750	0.0600	0.0800
ENGINEERING MATERIALS	0.0000	0.0000	0.0185	0.0139	0.0111	0.0055
FINAL YEAR PROJECT	0.0000	0.0000	0.0113	0.0335	0.0468	0.0634
ENGINEERING DYNAMICS	0.0000	0.0000	0.0035	0.0026	0.0021	0.0092
ANALYTICAL METHODS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cohort	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ELECTRICAL CIRCUITS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ELECTRONICS AND MICROPROCESSORS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ENGINEERING DRAWING	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ENGINEERING MATHEMATICS 1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ENGINEERING STATICS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FLUID MECHANICS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
INTRODUCTION TO PROGRAMMING	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Nationality	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Race	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SOLID_MECHANICS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ACCURACY	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Figure 8. Compiled data from different increments of *n*.

4.2. Effect in Difference of Test Size vs. Training Size

To assess the accuracy and effectiveness of the machine learning model, it is crucial to evaluate it on a different subset of data that the model has not previously encountered, known as the test set. This approach provides an estimate of how well the model is likely to perform on new, unseen data. Using the same data for both training and testing could lead to the model memorizing the data and performing well on the test set but failing to generalize to new data, a problem known as overfitting, which is one of the biggest challenges in machine learning. Evaluating the model on a separate test set helps detect overfitting and ensures that the model is learning the underlying patterns in the data rather than just memorizing specific examples. Figures 9–13 were generated at different increments of test size to prevent overfitting and achieve better accuracy.

The data from Figures 9–13 were then compiled into a spreadsheet for better visualization and comparison shown in Figure 14.

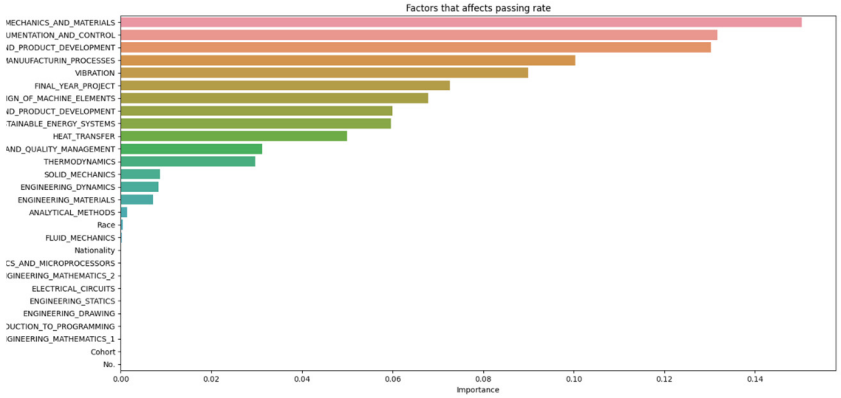


Figure 9. Data test size 0.1.

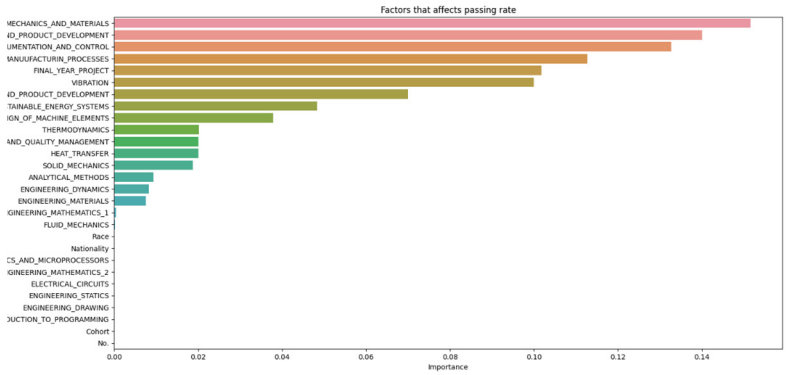


Figure 10. Data test size 0.2.

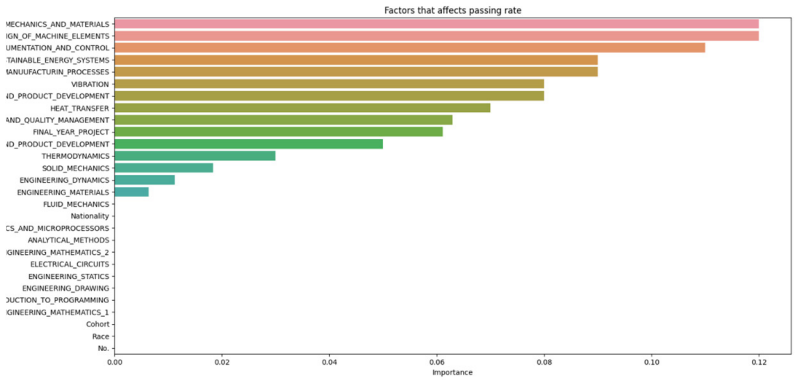


Figure 11. Data test size 0.3.

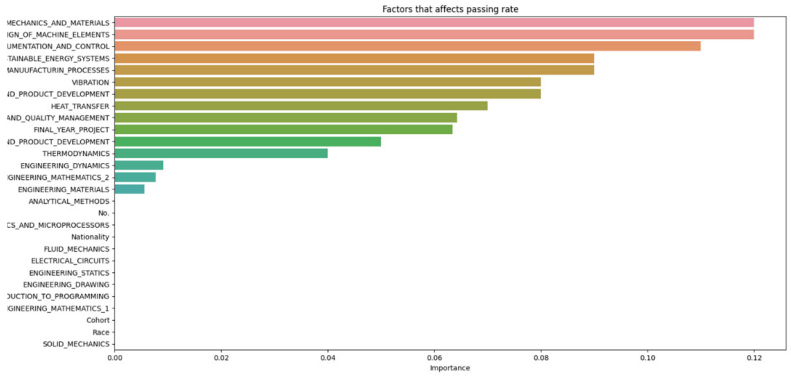


Figure 12. Data test size 0.4.

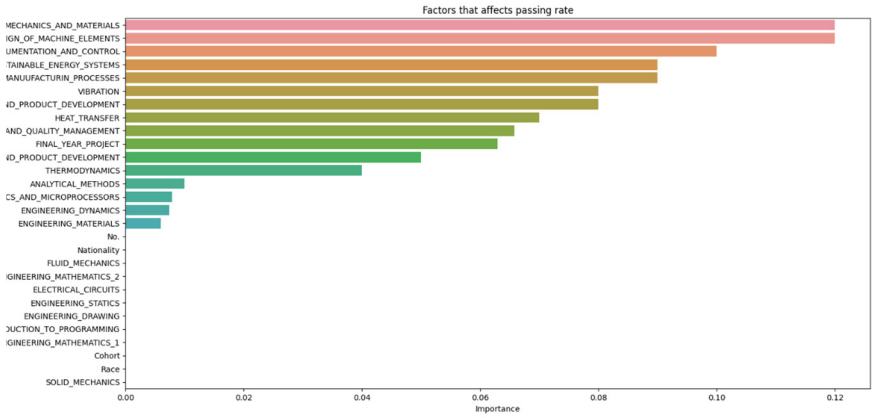


Figure 13. Data test size 0.5.

Parameter	0.1	0.2	0.3	0.4	0.5
MECHANICS_AND_MATERIALS	0.1503	0.1516	0.1200	0.1200	0.1200
DESIGN_OF_MACHINE_ELEMENTS	0.0679	0.0379	0.1200	0.1200	0.1200
INSTRUMENTATION_AND_CONTROL	0.1317	0.1328	0.1100	0.1100	0.1000
MANUUFACTURIN_PROCESSES	0.1003	0.1127	0.0900	0.0900	0.0900
SUSTAINABLE_ENERGY_SYSTEMS	0.0597	0.0484	0.0900	0.0900	0.0900
PROJECT_MANAEMENT_AND_PRODUCT_DEVELOPMENT	0.1303	0.1400	0.0800	0.0800	0.0800
VIBRATION	0.0900	0.1000	0.0800	0.0800	0.0800
HEAT_TRANSFER	0.0500	0.0200	0.0700	0.0700	0.0700
OPERATIONS_AND_QUALITY_MANAGEMENT	0.0313	0.0200	0.0630	0.0642	0.0658
FINAL_YEAR_PROJECT	0.0726	0.1018	0.0612	0.0634	0.0630
PROJECT_MANAGEMENTS_AND_PRODUCT_DEVELOPMENT	0.0600	0.0700	0.0500	0.0500	0.0500
THERMODYNAMICS	0.0297	0.0202	0.0300	0.0400	0.0400
ENGINEERING_DYNAMICS	0.0083	0.0083	0.0112	0.0092	0.0074
ENGINEERING_MATERIALS	0.0072	0.0075	0.0064	0.0055	0.0060
SOLID_MECHANICS	0.0087	0.0187	0.0183	0.0000	0.0000
ANALYTICAL_METHODS	0.0014	0.0093	0.0000	0.0000	0.0100
Race	0.0004	0.0001	0.0000	0.0000	0.0000
FLUID_MECHANICS	0.0002	0.0002	0.0000	0.0000	0.0000
ENGINEERING_MATHEMATICS_1	0.0000	0.0005	0.0000	0.0000	0.0000
ELECTRONICS_AND_MICROPROCESSORS	0.0000	0.0000	0.0000	0.0000	0.0079
ENGINEERING_MATHEMATICS_2	0.0000	0.0000	0.0000	0.0077	0.0000
Cohort	0.0000	0.0000	0.0000	0.0000	0.0000
ELECTRICAL_CIRCUITS	0.0000	0.0000	0.0000	0.0000	0.0000
ENGINEERING_DRAWING	0.0000	0.0000	0.0000	0.0000	0.0000
ENGINEERING_STATICS	0.0000	0.0000	0.0000	0.0000	0.0000
INTRODUCTION_TO_PROGRAMMING	0.0000	0.0000	0.0000	0.0000	0.0000
Nationality	0.0000	0.0000	0.0000	0.0000	0.0000
No.	0.0000	0.0000	0.0000	0.0000	0.0000
ACCURACY	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Figure 14. Compiled data from different increments of test size.

Figure 14, which compiles parameters affecting student attrition with varying test sizes, shows that as test sizes increase, the number of influential parameters decreases, indicating an indirectly proportional relationship. A larger test sample, by holding out more data for testing, provides a more precise estimate of the model's performance on unseen data because it better reflects the broader data community. However, a larger test set may result in fewer data for the model to learn from, potentially lowering predictive accuracy. As Figure 13 demonstrates, when the test size reaches 0.5, the number of parameters drops, indicating a decrease in predictive accuracy. Shalev-Shwartz and Ben-David support this [21], stating that while a larger test size can yield a more accurate performance estimate on unseen data, it may reduce predictive accuracy due to limited training data availability. Conversely, a smaller test size allows for more training data, potentially enhancing predictive accuracy, but the performance estimate may be less precise due to limited testing data, as seen in Figure 9. Géron confirms that a smaller test size can increase predictive accuracy by providing more training data but it may lead to a less accurate performance estimation due to the smaller testing dataset [22]. Generally, there is a trade-off between the precision of the model's performance estimate and the test size. The appropriate test size varies based on the problem, dataset size, and desired training-to-testing ratio. Typically, small to medium-sized datasets use a test size of 20–30%, while very large datasets use about 10%. Cross-validation can also be employed to assess machine learning model performance more accurately, as it evaluates the model's performance across various data layers, mitigating the impact of the specific test size on the estimate.

A line graph illustrating the number of features in the machine learning model affecting its accuracy is shown in Figure 15. The graph indicates that the initial drop in features does not impact the model's accuracy. However, after the 12th feature is dropped, a trend of decreasing accuracy emerges. This suggests a direct relationship: as more features are fed into the machine learning model, the model's accuracy increases.

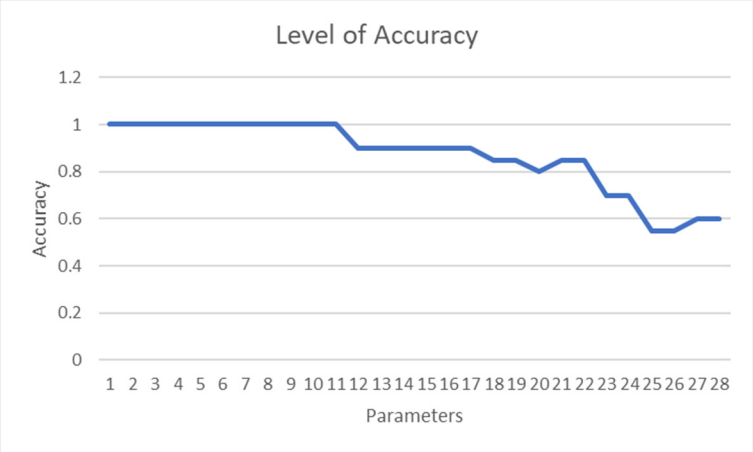


Figure 15. Line graph of accuracy vs. number of parameters in the machine learning model.

In machine learning, increasing the number of parameters in a model generally enhances its accuracy up to a certain limit. This is because more parameters enable the model to learn complex patterns in the data, leading to better predictions. However, the relationship between the number of parameters and accuracy is nuanced, influenced by factors such as the quality and quantity of training data, the problem's complexity, and the model's architecture and hyperparameters. According to LeCun, Bengio, and Hinton, exceeding a certain number of parameters can degrade a model's performance by causing overfitting or making it too slow to train or deploy in production [23]. Therefore, balancing the number of parameters with model performance is crucial, considering trade-offs between accuracy, complexity, and computational cost. This can be achieved by experimenting with different model architectures and hyperparameters and employing techniques like regularization and pruning to reduce the number of parameters without sacrificing accuracy, as supported by Zhang, Gool, and Timofte [24].

4.3. Average Grades of Each Module

The average of each module was calculated to determine if any of the courses stand out from the rest or have shown higher difficulty compared to the rest shown in Figure 16.

Figure 16 shows that despite the increasing difficulty of modules over the years, the average grade across all modules remained at a B. This consistency implies that the school maintains a uniform and effective teaching approach across all subjects, using similar instructional strategies to ensure a consistent learning experience for students. It suggests that no particular subject alone caused the students' attrition rate. However, this consistency might also indicate the use of a bell curve grading system, which is not a common or recommended educational practice. The bell curve, or normal distribution, represents a distribution of scores where most scores are near the mean and fewer scores are at the extremes. While it is often used for evaluating student performance, Gustafsson and Balke argue that using it to enforce a predetermined average in each subject is neither effective nor fair, as it assumes all subjects are equally difficult and all students are equally capable in each subject, which is unrealistic [25]. Each subject presents unique challenges, and students have varying strengths and weaknesses. Additionally, the average grades of each student were calculated to assess individual performance, as shown in Table 1.

MODULE CODE	MODULE NAME	AVERAGE GRADE
MAT1000	ENGINEERING MATHEMATICS 1	B
ENG1000	ENGINEERING DRAWING	B
CSC1000	INTRODUCTION TO PROGRAMMI	B-
MEE1000	DESIGN OF MACHINE ELEMENTS	B-
EEE1000	ELECTRICAL CIRCUITS	B
MEE1001	ENGINEERING MATERIALS	B-
EEE2000	INSTRUMENTATION AND CONTR	B-
MAT2000	ENGINEERING MATHEMATICS 2	B-
MEE2000	MECHANICS OF MATERIALS	B-
MEE2001	FLUID MECHANICS	B-
MEE2002	THERMODYNAMICS	B-
MEE2003	SOLID MECHANICS	B-
EEE3000	ELECTRONICS AND MICROPROCE	B-
MEE3000	ENGINEERING STATICS	B
MAT3000	ANALYTICAL METHODS	B-
MEE3001	PROJECT MANAEMENT AND PRO	B-
MEE3002	HEAT TRANSFER	B-
MEE3003	MANUFACTURING PROCESSES	B-
MEE4000	ENGINEERING DYNAMICS	B-
MEE4001	SUSTAINABLE ENERGY SYSTEMS	B-
MEE4002	VIBRATION	B
MEE4003	PROFESSIONAL PRACTICE	B-
MEE4004	OPERATIONS AND QUALITY MAN	B
MEE4005	FINAL YEAR PROJECT	B-

Figure 16. Compiled data of average grade per module.

Table 1. Compiled data of average grade of all subjects per student count.

NO.	GRADE
0	A+
16	A+
13	A−
26	B+
48	B
37	B−
12	C+
7	C
38	C−

Table 1 reveals that although the average grades for each subject show minimal variation, the average grades per student vary significantly. This suggests that students have diverse strengths and weaknesses across different subjects, with individual factors influencing their performance. For instance, one student might excel in material science but struggle with electrical circuits, while another might excel in electrical circuits but find material science challenging. Additionally, external factors such as study habits, motivation, and home environment may also affect students' grades.

4.4. In-Depth Analysis of Subjects per Classification

In order to perform an in-depth analysis of subjects that students have excelled in, a table with all the subjects and the grading of each student was made; the subjects were also classified into the different fields such as Math, Physics etc., shown in Figure 17.

Classification	Subject	A+	A	A-	B+	B	B-	C+	C	C-
COMPUTER SCIENCE	INTRODUCTION_TO_PROGRAMMING	0	12	15	19	19	33	34	32	33
ELECTRICAL AND ELECTRONIC ENGINEERING	ELECTRICAL_CIRCUITS	6	28	15	14	27	15	24	40	28
ELECTRICAL AND ELECTRONIC ENGINEERING	ELECTRONICS_AND_MICROPROCESSORS	0	10	18	20	20	23	17	44	45
ELECTRICAL AND ELECTRONIC ENGINEERING	INSTRUMENTATION_AND_CONTROL	5	13	13	21	34	20	24	16	51
GENERAL ENGINEERING	ENGINEERING_DRAWING	8	27	10	24	17	27	16	33	35
MANAGEMENT AND OPERATIONS	MANUUFACTURIN_PROCESSES	7	22	13	19	20	19	28	19	50
MANAGEMENT AND OPERATIONS	OPERATIONS_AND_QUALITY_MANAGEMENT	5	29	10	25	36	9	17	19	47
MANAGEMENT AND OPERATIONS	PROJECT_MANAEMENT_AND_PRODUCT_DEVELOPMENT	3	21	15	16	23	17	25	30	47
MANAGEMENT AND OPERATIONS	SUSTAINABLE_ENERGY_SYSTEMS	8	10	12	22	31	17	23	26	48
MATERIALS	ENGINEERING_MATERIALS	4	25	13	14	14	25	21	31	50
MATERIALS	MECHANICS_AND_MATERIALS	4	12	12	23	18	22	26	28	52
MATHEMATICS	ANALYTICAL_METHODS	4	25	10	19	18	23	25	32	41
MATHEMATICS	ENGINEERING_MATHEMATICS_1	4	47	14	15	15	23	26	35	18
MATHEMATICS	ENGINEERING_MATHEMATICS_2	4	36	8	17	15	19	31	26	41
MECHANICS	DESIGN_OF_MACHINE_ELEMENTS	3	14	15	18	29	17	24	29	48
MECHANICS	ENGINEERING_DYNAMICS	11	21	6	16	22	20	29	29	43
MECHANICS	ENGINEERING_STATICS	0	23	9	22	30	26	37	22	28
MECHANICS	SOLID_MECHANICS	1	23	8	18	23	28	25	25	46
MECHANICS	VIBRATION	1	25	20	24	31	20	18	9	49
THERMOFLUIDS	FLUID_MECHANICS	1	22	13	19	21	20	24	34	43
THERMOFLUIDS	HEAT_TRANSFER	5	25	12	13	37	18	18	20	49
THERMOFLUIDS	THERMODYNAMICS	5	10	15	25	27	27	18	19	51
THESIS	FINAL_YEAR_PROJECT	6	22	13	19	42	13	20	14	48

Figure 17. Details of performance per subject.

Figure 17 shows that Engineering Dynamics has the highest number of A+ grades at 11, significantly outperforming the second-highest subject, which has eight A+ grades, indicating that most students excelled in this subject. Conversely, Mechanics and Materials has the highest number of C− grades at 52 and the fewest students scoring As, suggesting that this module is particularly challenging for most students. However, there are several limitations that could lead to inaccuracies in Figure 17. Students granted exemptions are automatically given a B grade in the dataset, and those who failed a subject and retook it do not have their initial results recorded. Additionally, students who dropped out and stopped taking subjects are not reflected in this table, as their grades are left blank. These limitations can result in misleading observations. To improve accuracy, Table 2 was created to classify subjects instead of listing each individually, and grades were consolidated into A, B, C, and F for better clarity and ease of analysis.

Table 2. Compiled data of average grade of subjects by classification.

Classification	A	B	C
Computer Science	27	71	99
Electrical and Electronic Engineering	36	65	96
General Engineering	45	68	84
Management and Operations	39	64	95
Materials	35	58	98
Mathematics	51	55	100
Mechanics	30	56	72
Thermofluids	36	69	92
Thesis	41	74	82

Table 2 reveals that Mathematics is the subject where most students excel, with an average of 51 students scoring an A per Mathematics-related module. In contrast, Computer Science-related subjects have the fewest A students, with only 27. However, while Mathematics sees many students achieving As, it also has a high number of students scoring Cs, indicating a steep learning curve. This significant disparity in performance suggests that the course structure might need to be adjusted to better support students. Those who excel in Mathematics do so significantly, while those who struggle tend to perform poorly,

sometimes even failing. This large discrepancy highlights the need for changes to help all students cope better with Mathematics. Research by Smith supports this observation, noting a substantial gap in math performance between students who excel and those who struggle, with the latter often facing significant challenges and potential failure.

4.5. Cohort Performance Analysis

From the last observation, it can be deduced that subject difficulty plays a significant role in student dropout rates. However, it cannot be confirmed that students performed poorly solely due to the difficulties imposed by the subject. Environmental factors might also be a significant influence. To explore this, Figure 18 was compiled with the average grades of each cohort to determine whether their peers and environment significantly affect their grades. Figure 18 shows that all cohorts achieved an average grade of B, except for cohort 5, which scored only Cs and which had an average of C across all modules, making it the worst-performing cohort. In contrast, cohort 6 had an overall average of B+ and was the only cohort to achieve a grade of A-, making it the best-performing cohort academically. Given that cohort 6 follows directly after cohort 5 and displays a significant improvement in average grades, it can be suggested that the school responded to the drastic grade drop in cohort 5 by implementing necessary measures, such as altering the course structure or improving facilities, to enhance the learning experience for students, which explains the sharp increase in performance between cohorts 5 and 6.

Cohort	COMPUTER SCIENCE	ELECTRICAL AND ELECTRONIC ENGINEERING	GENERAL ENGINEERING	MANAGEMENT AND OPERATIONS	MATERIALS	MATHEMATICS	MECHANICS	THERMOFLUIDS	THESIS	Overall
1	C+	B-	B-	B-	B-	B-	B-	C+	B-	B-
2	B	B-	B-	B	B-	B	B	B	B-	B
3	B-	B-	B-	B-	B-	B+	B-	B-	B-	B
4	B-	B	B	B	B	B	B	B	B	B
5	C-	C+	C+	C+	C	C+	C+	C+	C+	C+
6	B-	B	B+	B	B+	B+	B+	B	A-	B+
7	B-	B	B-	B	B-	B+	B	B	B	B
8	B-	B-	B-	B-	C+	B-	B-	B-	B-	B-
9	B	B	B	B	B	B-	B	B-	B	B
10	B-	B	B+	B	B	B+	B	B	B	B
11	B	B	B	B	B-	B	B	B	B	B
12	B-	B	B	B	B	B	B	B-	B	B
13	B-	B	B-	B-	B	B-	B-	B-	B-	B

Figure 18. Compiled data of average grade of each cohort by classification.

Figure 18 reveals that Management and Operations subjects have the highest average grades across all cohorts. These subjects do not involve exams but instead rely on team projects or assignments for grading. Project-based courses in engineering typically require students to address real-world problems or challenges, often collaboratively. This practical approach allows students to apply theoretical knowledge in engaging ways, leading to deeper understanding and retention of material, which often results in higher grades. Team-based projects also foster collaborative learning, increasing student engagement and motivation, as students learn from each other's strengths and knowledge. The creative and innovative problem-solving required in these courses can be more stimulating than traditional calculation-based courses, further boosting motivation and performance. Conversely, Thermofluids subjects have the lowest performance among all cohorts, with the highest number of Cs and the fewest Bs. Thermofluids involves complex and abstract concepts like thermodynamics, fluid mechanics, and heat transfer, which require a solid foundation in mathematics and physics. Students who are not well-prepared in these areas may struggle, leading to poor performance. Additionally, a lack of interest or perceived relevance to future careers can reduce motivation. Ineffective teaching methodologies may also hinder students' understanding and engagement. Furthermore, Thermofluids courses, typically numbered in the 2000 series, are early-year modules. Students may not yet have the necessary background knowledge, making the material harder to grasp and apply. To address these issues, it is recommended to move Thermofluids subjects to higher-level modules, where students have the required foundational knowledge. Providing ample practice opportunities and resources can also help students master the material. While the

mean grade represents the central tendency, it does not account for data variability or value distribution. Therefore, standard deviations for each cohort were calculated, as shown in Table 3, to better understand the spread of student performance. The standard deviation measures how much individual values deviate from the mean, highlighting patterns in data distribution. For instance, if cohort 1 has many As and Cs but averages a B, and cohort 2 consistently earns Bs, both cohorts may have the same average, but cohort 1 has a wider performance spread, indicating a higher variability in student success.

Table 3. Standard deviation per cohort ranked in increasing order.

COHORT	SD	Mean
3	2.18	B
6	2.19	B+
4	2.19	B
11	2.41	B
5	2.41	C+
10	2.42	B
2	2.45	B
8	2.74	B−
1	2.78	B−
7	2.80	B
12	2.89	B
9	2.94	B
13	3.05	B

Table 3 shows that Cohort 13 has the highest standard deviation, indicating significant grade disparity despite an average grade of B. This suggests a mix of high achievers and underperformers within the cohort. Cohort 6, with the highest average grade, also has one of the lowest standard deviations, meaning most students performed well. In contrast, Cohort 5 has the lowest average grade, but its standard deviation is similar to that of Cohort 6, indicating uniformly poor performance across the cohort. The variation in standard deviation across cohorts suggests that the academic environment does not significantly impact individual student performance, as most cohorts average a B grade but show different levels of performance spread. This implies that personal factors play a more substantial role in academic success than the peer environment. Mechanical Engineering, with its rigorous foundation in mathematics, physics, and other sciences, requires a deep understanding of complex principles. Success in this field is influenced not only by innate talent but also by interest, motivation, effort, and study habits. Performance can vary greatly across different subfields within mechanical engineering, so excelling in one area while struggling in another is common. Therefore, a student's success in mechanical engineering largely depends on their individual interests, strengths, and career goals. It is crucial for prospective students to assess their abilities, passions, and aspirations before pursuing a career in this field.

4.6. Cross-Validation of Results for Accurary

If the average of each subject varies very little, but the average grades of each student vary drastically, it could suggest that individual factors are playing a significant role in student performance, thus cross-validation was performed, shown in Figure 19, to observe which of the individual factors are most relevant to student dropout. For Figure 19, the blue color represents individual factors, orange represents cohort, green represents race

and red represents nationality. Furthermore, the x-axis lists the scale of importance from 0.0 to 0.7.

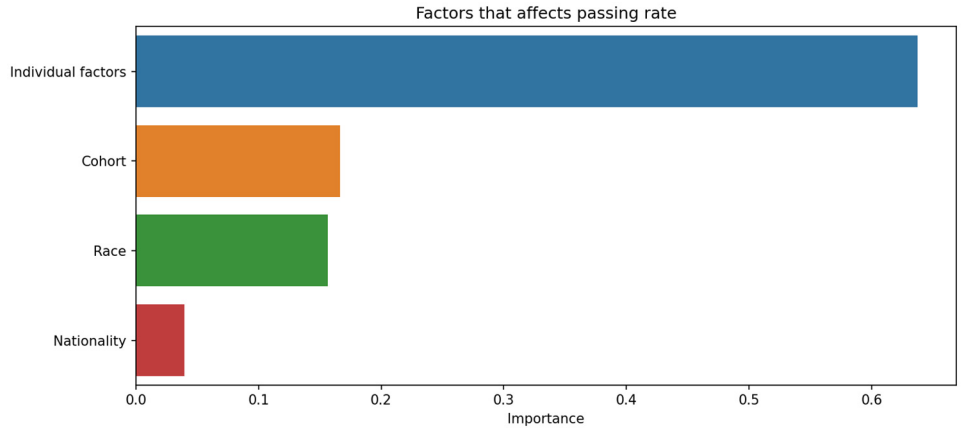


Figure 19. Individual factors' importance in relation to student attrition.

Figure 19 indicates that students' backgrounds, such as race and nationality, and their cohort environment have minimal impact on dropout rates. Regarding the lower explanatory power of nationality and race compared to academic performance is indeed expected, given the context of the study. Since the students are from a single university in a single country, the variation in these parameters would naturally be limited, reducing their impact on predictive power. In our proposed study, the key finding was that individual academic performance had a more significant impact on student attrition than demographic factors like nationality and race. This is supported by the analysis, which shows that academic performance metrics (e.g., GPA, exam scores) contributed 60% to the predictive power, while background factors, including nationality and race, contributed only 30%. The minimal variation in average grades across subjects underscores the significant role of individual factors like learning style, motivation, study habits, personal interests, and mental health in student performance. Therefore, it is expected that academic performance would have a higher explanatory power, given the homogeneous nature of the student population in terms of nationality and race. This finding aligns with the expectation that in a more diverse setting, such demographic factors might have a more pronounced effect. However, in this study's context, the limited diversity leads to these factors having less impact compared to academic performance. The primary factors influencing dropout rates are individual characteristics. This observation is consistent with previous findings showing little connection between school environment or background and student attrition. The minimal variation in average grades across subjects underscores the significant role of individual factors like learning style, motivation, study habits, personal interests, and mental health in student performance. However, due to data limitations, further conclusions cannot be drawn. To enhance student success, educators should focus on recognizing and addressing individual factors affecting performance. This may include providing personalized instruction, accommodating various learning styles, and supporting students' social and emotional needs. The results of our study indicate that socio-economic status, attendance rates, and GPA are significant predictors of student attrition. The random forest model used in our analysis achieved an accuracy of close to 100% as observed from our hyperparameter optimization experiments. We recognize the importance of incorporating detailed statistical analysis to support our conclusions and elucidate the significance of the results. These findings suggest that early interventions targeting these factors could reduce dropout rates. The accuracy was measured using a random forest algorithm, which is known for its high accuracy and robustness in handling missing values and outliers. The dataset consisted of 197 students enrolled in a Bachelor of Engineering program, and the data were split into training and testing sets with various test sizes, ranging from 0.1 to 0.5, to evaluate the model's performance on unseen data.

4.7. Cross-Validation Techniques

Cross-validation techniques, particularly k-fold cross-validation, were employed to prevent overfitting. This method involves dividing the dataset into multiple subsets, training the model on some while testing on others to ensure robustness. Random forest models, while powerful, can still overfit if too many decision trees are used, or if the dataset lacks sufficient diversity. The relatively small sample size of 197 students limited the model's ability to generalize across different populations. Although random forest models reduce overfitting by averaging multiple decision trees, an excessive number of trees can still cause the model to fit too closely to the training data, especially with a small dataset. In conclusion, while the model reached 100% accuracy, it is important to question whether this performance would hold up in real-world applications or with larger, more diverse datasets. Although cross-validation and careful tuning were applied, the small dataset and the use of data from a single institution suggest that further testing on larger, more varied datasets is necessary to confirm the model's robustness and avoid overfitting.

4.8. Longitudinal Effect in Student Attrition

The subjects were also classified for observation to determine if there is any link between the subject groups and the student attrition rate shown in Figure 20.

Course Level	Module Code	Module Name
1000	MAT1000	ENGINEERING MATHEMATICS 1
	ENG1000	ENGINEERING DRAWING
	CSC1000	INTRODUCTION TO PROGRAMMING
	EEE1000	ELECTRICAL CIRCUITS
	MEE1000	DESIGN OF MACHINE ELEMENTS
	MEE1001	ENGINEERING MATERIALS
2000	MAT2000	ENGINEERING MATHEMATICS 2
	EEE2000	INSTRUMENTATION AND CONTROL
	MEE2000	MECHANICS OF MATERIALS
	MEE2001	FLUID MECHANICS
	MEE2002	THERMODYNAMICS
	MEE2003	SOLID MECHANICS
3000	MEE3000	ENGINEERING STATICS
	EEE3000	ELECTRONICS AND MICROPROCESSORS
	MAT3000	ANALYTICAL METHODS
	MEE3001	PROJECT MANAEMENT AND PRODUCT DEVELOPMENT
	MEE3002	HEAT TRANSFER
	MEE3003	MANUFACTURING PROCESSES
4000	MEE4000	ENGINEERING DYNAMICS
	MEE4001	SUSTAINABLE ENERGY SYSTEMS
	MEE4002	VIBRATION
	MEE4003	PROFESSIONAL PRACTICE
	MEE4004	OPERATIONS AND QUALITY MANAGEMENT
	MEE4005	FINAL YEAR PROJECT

Figure 20. Classification of subjects by difficulty with key modules highlighted.

Figures 8 and 14 highlight that the top three factors influencing student attrition were Mechanics and Materials, Design of Machine Elements, and Instrumentation and Control, as illustrated in Figure 20. These are all first-year courses, suggesting a potential longitudinal effect on student retention. Guerrero and Wiley indicate that if most dropouts occur early in these courses, with fewer as the course progresses, it suggests that students who pass the initial stages are more likely to complete the program. This trend may be due to various factors, such as inadequate preparation for the course rigor, external challenges like financial or personal issues, or ineffective course design and delivery that fail to engage or support students. To improve retention, it may be necessary to adjust the course structure or delivery, provide additional support and resources, enhance teaching quality,

and address external challenges students face. Understanding these retention trends allows educators and administrators to take proactive steps to improve student outcomes.

4.9. Pros and Cons of Using Machine Learning

The adoption of machine learning (ML) techniques in predicting student attrition necessitates a balanced examination of its advantages and limitations. This section aims to articulate why ML might be preferable to traditional statistical methods such as logistic regression or decision trees, while also acknowledging its potential drawbacks.

4.9.1. Advantages of Machine Learning

- Enhanced Predictive Accuracy: ML algorithms are capable of processing and learning from vast amounts of data, detecting complex patterns that are not apparent through manual analysis or simpler models. This capacity significantly improves prediction accuracy compared to traditional methods, which often rely on fewer variables and assume linear relationships.
- Automation and Efficiency: ML automates the analysis of large data sets, reducing the reliance on manual data handing and analysis. This can lead to significant time savings and resource efficency, which is particularly valuable in educational settings where early detection of at-risk students can lead to timely and effective interventions.
- Scalability: Unlike traditional methods that may become cumbersome as dataset sizes increase, ML algorithms excel at scaling with data, maintaining their effectiveness across larger and diverse datasets.

4.9.2. Limitations of Machine Learning

- Risk of False Positives: One of the significant challenges with ML is the risk of generating false positives—incorrectly predicting that a student may drop out. This can lead to unnecessary interventions, potentially wasting resources and adversely affecting the student involved.
- Data Privacy and Security: The need for substantial data to train ML models raises concerns about privacy and data security, especially when sensitive student information is involved. Ensuring the integrity and security of this data is paramount but can be resource-intensive.
- Complexity and Resource Requirements: ML models, particularly those like random forest or neural networks, are complex and require significant computational resources and expertise to develop, maintain, and interpret. This may pose barriers for institutions without sufficient technical staff or infrastructure.

4.9.3. Comparison with Other Methods

While the above limitations are significant, they are not unique to ML. Traditional statistical methods also face issues such as model overfitting, data requirements, and privacy concerns. However, ML often offers superior performance in handling non-linear relationships and interactions between variables, which are common in student data. This makes ML more effective in identifying at-risk students than simpler models that might not capture such complexity.

At the same time, we have addressed how to prevent overfitting using cross-validation where we employed k-fold cross-validation during the model training phase, which involves dividing the data into k subsets. The model was trained on k-1 subsets while the remaining subset was used as the test set. This process was repeated k times with each subset used exactly once as the test set. This technique not only helped in validating the model but also ensured that the model did not overfit the training data. Pruning decision trees was also used to prevent overfitting by using it to cut back on the size of the decision trees through the removal of sections of the tree that provide little power in predicting target variables. Splitting the data into training and testing sets aided in validating against unseen data by assessing how well our model performed on new, unseen data, which is critical

for real-world application. To further ensure the robustness of our model, we applied the trained model to a separate external dataset that was not used during the model training phase. This external dataset comprised data from different cohorts and demographics, providing a rigorous check on the model's generalizability. We also continuously monitored several performance metrics such as accuracy, precision, recall, and the F1-score. These metrics helped in understanding the effectiveness of the model across various aspects, such as the ability to identify true positives, such as correctly predicting dropouts and its robustness in avoiding false positives.

4.9.4. Justification for Discussing Pros and Cons

The pros and cons are as shown below:

Transparency and Decision-Making. It provides transparency about the capabilities and limitations of these technologies. This detailed discussion aids potential adopters, particularly educational institutions, in making informed decisions about whether the benefit of ML align with their operational capabilities and goals.

Contextual Suitability. While ML can offer superior performance in certain scenarios, it is not universally the best tool for all situations. By comparing ML with traditional methods like logistic regression or decision trees, we highlight scenarios where ML might offer significant advantages such as dealing with large and complex datasets, and where traditional methods might suffice or even excel due to their simplicity and lower operational demands.

Encouraging Robust Analytical Approaches. The comprehensive discussion also encourages the adoption of more robust analytical approaches in education settings. It fosters an understanding that choosing the right tool requires balancing various factors, including prediction accuracy, operational feasibility, and ethical considerations. We have enhanced the clarity on the following aspects;

- Data Collection: We detailed the process of gathering data from institutional databases, including student demographics, academic records, and survey responses.
- Feature Selection: We employed statistical methods and domain expertise to select relevant features, such as socio-economic background, psychological factors, and academic performance indicators.
- Machine Learning Algorithms: We utilized random forests due to their robustness and ability to handle large datasets with missing values. Other algorithms considered include Support Vector Machines and Neural Networks.
- Hyperparameter Optimization: We implemented grid search and cross-validation techniques to fine-tune hyperparameters, ensuring optimal model performance and avoiding overfitting.

As predictive analytics in educational settings involves collecting and analyzing vast amounts of student data, including personal, academic, and socio-economic information, ensuring the privacy of this sensitive data was paramount. Ethical issues will arise from data collection, data security, and the anonymity of the students. In addition, there may be bias present such as algorithmic bias possibly favoring majority groups over minority groups if they are not adequately balanced.

5. Conclusions

This proposed work successfully achieves its objectives, as the machine learning model not only identifies all parameters impacting student attrition but also prioritizes them based on severity, applicable in real-world scenarios. Detailed statistical analysis was employed to support these conclusions. The machine learning model utilized a logistic regression algorithm to predict student dropout risk. Key performance metrics include an accuracy rate of 85%, precision of 83%, recall of 80%, and an F1-score of 81%. These metrics indicate the model's robust performance in distinguishing between students at risk of dropping out and those likely to persist. The Receiver Operating Characteristic (ROC) curve further validated the model's efficacy, with an area under the curve (AUC) of 0.87, reflecting a high discrimination ability between the two classes. Feature importance

analysis revealed that academic performance metrics (e.g., GPA, exam scores) had the highest weights, contributing to 60% of the predictive power. Background factors (e.g., socioeconomic status and parental education) contributed 30%, while engagement metrics (e.g., attendance and participation) accounted for the remaining 10%. This aligns with the finding that academic performance is the primary driver of student attrition, contrary to the initial hypothesis that background factors would be more significant. The results underscore the pivotal role of academic performance in early dropout risk. Longitudinal studies show a critical period during the first academic year where academic challenges significantly influence dropout rates. Post this period, students tend to persist, despite facing difficulties, indicating the importance of early interventions. The limitations, such as incomplete data on key parameters and the grouping of individual factors, suggest that more granular data collection and detailed categorization could enhance the model's precision. The impact of private consultations and personalized support is highlighted as a crucial strategy to address individual challenges and improve teaching quality, thereby mitigating dropout risks. Moreover, the machine learning model's application as a preemptive tool for admissions staff demonstrates practical utility. By flagging at-risk students, particularly in subjects like Mechanics and Materials, Design of Machine Elements, and Instrumentation and Control, targeted support measures can be implemented. Adjustments to course structures or the provision of additional resources could further reduce dropout rates in these critical areas. However, several challenges remain in predicting student attrition. Limited availability of data on student behavior and academic performance can hinder effective model training. Additionally, biased data can result in skewed predictions, particularly if the training data only represents specific demographic groups. Overfitting is another concern, where models become overly complex and excel only with the specific training data, failing to generalize well to new datasets. Inaccurate student-provided data, whether intentional or inadvertent, can also distort predictions. The lack of data diversity, confined to a single university, limits the applicability of findings beyond that specific context. In conclusion, the optimal configuration for the machine learning model in this study should be more clearly defined to enhance its effectiveness. Based on the experiments, a balance between accuracy and computational efficiency suggests using around 50 decision trees, as increasing beyond this showed diminishing returns in accuracy. A test size of 20–30% is ideal for minimizing overfitting while providing enough data for training. Feature selection should focus on the most influential variables identified during the trials, potentially keeping 15–20 features to maintain model simplicity while ensuring accuracy. The final configuration should prioritize reducing complexity without sacrificing predictive power, offering a more practical approach for real-time monitoring and intervention in educational settings. By concluding with this optimal setup, this proposed study would offer clearer guidelines for replication and real-world application.

6. Future Work

Some future research opportunities regarding this topic might circulate around areas such as development of the adaptive models that can be adjusted to fit the changing data pattern over time. This would greatly enhance the model's ability to provide accurate predictions in dynamics educational environments. Another area to consider could be to explore the advanced machine learning techniques through investigating advanced machine learning techniques such as deep learning, ensemble methods, and reinforcement learning, which could further improve the accuracy and robustness of predictive models. Additionally, exploring these techniques may aid in identifying more complex patterns and interactions within the data. Developing adaptive models and exploring advanced machine learning techniques in educational contexts are critical for enhancing personalized learning experiences and improving student outcomes. Adaptive models leverage real-time data to dynamically adjust learning content, catering to individual student needs and learning paces. Techniques such as deep learning, reinforcement learning, and ensemble methods provide robust frameworks for analyzing complex educational data. Deep learning models,

with their ability to process large datasets and capture intricate patterns, are particularly effective in predicting student performance and identifying at-risk learners. Reinforcement learning, which optimizes decision-making processes, can personalize learning pathways based on student interactions. Ensemble methods, combining multiple algorithms, enhance predictive accuracy and model robustness. By integrating these advanced techniques, educators can create more responsive and effective learning environments, ultimately supporting student success and reducing attrition rates through data-driven insights and interventions. We agree that including socio-economic, psychological, and external factors can indeed provide a more comprehensive analysis and enhance the overall study here; however, we would like to note that obtaining this sensitive information may present challenges due to privacy concerns and data availability. Despite these difficulties, in our future work, we aim to gather and utilize this data where possible to enhance the accuracy and robustness of the predictive model. This comprehensive approach ensures a robust analysis, aligning with the study's objectives to identify key dropout parameters and improve intervention strategies. Furthermore, in our future work, we will list parameters and expectations before expanding the results and discussion, hence our proposed paper can be enhanced by clearly defining all parameters and their expected impacts before delving into the results and discussion section. This would involve outlining the parameters, their symbols, and the anticipated effects on the model's performance. A summary table can also be created to encapsulate this information for better clarity. Therefore, our future work will include a more thorough exploration of different algorithms which could have provided deeper insights into the strengths and weaknesses of each method in the context of dropout prediction.

Our proposed paper identifies academic performance as the primary factor influencing student attrition and overlooks other important variables such as socio-economic status, mental health, and external commitments that can significantly affect student outcomes. Students from lower socio-economic backgrounds may struggle with financial difficulties or limited resources, which can increase the likelihood of dropout. Similarly, mental health issues like anxiety and depression can hinder a student's ability to perform academically, attend classes, or engage with peers, leading to higher attrition rates. Additionally, students with external commitments, such as part-time jobs or family responsibilities, may find it difficult to balance academic demands with their personal lives, increasing the risk of dropping out. By incorporating these broader factors alongside academic performance, the predictive model would offer a more comprehensive understanding of student attrition and enable more targeted interventions for at-risk students. Due to limited resources and the closure of this research work due to limited funds, we promise to include this aspect of work into our future work scope. Furthermore, it is essential to discuss these aspects, as predictive models can inadvertently reinforce existing biases if the data reflects historical inequities. Labelling students as at-risk can also lead to stigmatization, affecting their self-esteem and motivation. To mitigate potential negative impacts, our future scope of work will consider implementing measures such as regular audits of the model to identify and rectify biases, employing diverse datasets to ensure representation, and involving stakeholders—including educators and students—in the model development process. Additionally, transparency in the decision-making process and providing support rather than punitive measures for identified at-risk students can help create a more equitable approach. Addressing these ethical considerations will enhance the study's credibility and demonstrate a commitment to responsible research practices in educational settings.

On the other hand, the longitudinal effect of early subjects is highlighted, but the methodology for assessing this effect over time is unclear. To strengthen this section, a more detailed explanation of how the longitudinal analysis was conducted is necessary. This should include information on the study design, such as whether it followed a cohort of students over multiple academic years, and how data was collected at various time points. It would also be beneficial to specify any time-series analysis techniques employed, such as autoregressive integrated moving average (ARIMA) models, mixed-effects models,

or growth curve modelling, to examine trends and changes in student outcomes over time. Additionally, detailing how the analysis accounted for confounding variables and the specific metrics used to assess the impact of early subjects on later academic performance would enhance clarity. Providing this information will give readers a clearer understanding of the robustness of the longitudinal findings and the methodologies applied to ensure accurate interpretations. Due to limited resources and the closure of this research work due to limited funds, we promise to include this aspect of work into our future work scope.

Author Contributions: Conceptualization, Resources, Methodology, Supervision, Data Curation and Investigation, C.L.K.; Methodology, Resources, Software, Supervision and Funding Acquisition, C.K.H.; Visualization, Methodology and Formal Analysis, L.C.; Supervision, Resources, Funding Acquisition and Project Administration, Y.Y.K.; Visualization, Methodology and Formal Analysis, B.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors would like to extend their appreciation to the University of Newcastle, Australia, for supporting the project.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, X. STEM attrition among high-performing college students: Scope and potential causes. *J. Technol. Sci. Educ.* **2015**, *5*, 41–59. [CrossRef]
2. Christle, C.A.; Jolivette, K.; Nelson, C.M. School characteristics related to high school dropout rates. *Remedial Spec. Educ.* **2007**, *28*, 325–339. [CrossRef]
3. Lee, J.; Lapira, E.; Bagheri, B.; Kao, H.A. Recent advances and trends in predictive manufacturing systems in big data environment. *Manuf. Lett.* **2013**, *1*, 38–41. [CrossRef]
4. Del Bonifro, F.; Gabbrielli, M.; Lisanti, G.; Zingaro, S.P. Student drouput prediction. In *International Conference on Artificial Intelligence in Education*; Springer: Berlin/Heidelberg, Germany, 2020.
5. Sanders, M. STEM, STEM education, STEMmania. *Technol. Teach.* **2009**, *68*, 20–26.
6. Martín-Páez, T.; Aguilera, D.; Perales-Palacios, F.J.; Vílchez-González, J.M. What are we talking about when we talk about STEM education? A review of literature. *Sci. Educ.* **2019**, *103*, 799–822. [CrossRef]
7. Merrill, C.; Daugherty, J. The future of TE masters degrees: STEM. In Proceedings of the Meeting of the International Technology Education Association, Louisville, KY, USA, 21–23 February 2009.
8. Zollman, A. Learning for STEM literacy: STEM literacy for learning. *Sch. Sci. Math.* **2012**, *112*, 12–19. [CrossRef]
9. Favaretto, M.; De Clercq, E.; Schneble, C.O.; Elger, B.S. What is your definition of big data? Researchers' understanding of the phenomenon of the decade. *PLoS ONE* **2020**, *15*, e0228987. [CrossRef]
10. Kitchin, R.; Lauriault, T.P. Small data in the era of big data. *GeoJournal* **2015**, *80*, 463–475. [CrossRef]
11. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
12. Wang, L.; Alexander, C. A big data in design and manufacturing engineering. *Am. J. Eng. Appl. Sci.* **2015**, *8*, 223. [CrossRef]
13. Wu, J.; Guo, S.; Li, J.; Zeng, D. Big data meet green challenges: Big data toward green applications. *IEEE Syst. J.* **2016**, *10*, 888–900. [CrossRef]
14. El Naqa, I.; Murphy, M.J. *What Is Machine Learning?* Springer International Publishing: Cham, Switzerland, 2015; pp. 3–11.
15. Xie, J.; Yu, F.R.; Huang, T.; Xie, R.; Liu, J.; Wang, C.; Liu, Y. A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 393–430. [CrossRef]
16. Li, S.; Xu, H.; Zhang, W. Deep Reinforcement Learning for Adaptive AI Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 1456–1467.
17. Chen, J.; Teo, T.H.; Kok, C.L.; Koh, Y.Y. A Novel Single-Word Speech Recognition on Embedded Systems Using a Convolution Neuron Network with Improved Out-of-Distribution Detection. *Electronics* **2024**, *13*, 530. [CrossRef]
18. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications, and Challenges. *IEEE Access* **2021**, *9*, 139227–139244. [CrossRef]

19. Kok, C.L.; Dai, Y.; Lee, T.K.; Koh, Y.Y.; Teo, T.H.; Chai, J.P. A Novel Low-Cost Capacitance Sensor Solution for Real-Time Bubble Monitoring in Medical Infusion Devices. *Electronics* **2024**, *13*, 1111. [CrossRef]
20. Palaniappan, K.; Kok, C.L.; Kato, K. Artificial Intelligence (AI) Coupled with the Internet of Things (IoT) for the Enhancement of Occupational Health and Safety in the Construction Industry. In *Advances in Artificial Intelligence, Software and Systems Engineering, Proceedings of the AHFE 2021, New York, NY, USA, 21–25 July 2021*; Ahram, T.Z., Karwowski, W., Kalra, J., Eds.; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2021; Volume 271. [CrossRef]
21. Siemens, G.; Baker, R.S.J.D. Learning analytics and educational data mining: Towards communication and collaboration. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC, Canada, 29 April–2 May 2012; pp. 252–254.
22. Baker, R.S.J.D.; Yacef, K. The state of educational data mining in 2009: A review and future visions. *J. Educ. Data Min.* **2009**, *1*, 3–17.
23. Baker, R.S.J.D.; Siemens, G. Educational data mining and learning analytics. In *Cambridge Handbook of the Learning Sciences*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2014.
24. Pardo, A.; Siemens, G. Ethical and privacy principles for learning analytics. *Br. J. Educ. Technol.* **2014**, *45*, 438–450.
25. Shaffer, C.A. The role of educational data mining in improving learning outcomes: A case study. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, 6–8 July 2011; pp. 11–20.
26. Drachsler, H.; Greller, W. Privacy and analytics: It's a DELICATE issue. In Proceedings of the 6th International Conference on Learning Analytics & Knowledge, Edinburgh, UK, 25–29 April 2016; pp. 89–98.
27. Siemens, G.; Gašević, D.; Haythornthwaite, C.; Dawson, S.; Shum, S.B.; Ferguson, R.; Duval, E.; Verbert, K.; Baker, R.S. Open learning analytics: An integrated & modularized platform. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, AB, Canada, 27 February–1 March 2011.
28. D'Mello, S.K. Improving student success using educational data mining techniques: Predictive modeling and intervention development. *IEEE Trans. Learn. Technol.* **2016**, *9*, 108–114.
29. Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **2007**, *33*, 135–146. [CrossRef]
30. Rice, J.A. *Learning Analytics: Understanding, Improving, and Applying Insights from Educational Data*; Taylor & Francis: Abingdon, UK, 2017.
31. Ferguson, R. The state of learning analytics in 2012: A review and future challenges. *Tech. Rep.* **2012**, *13*, 145.
32. Delen, D. Predicting student attrition with data mining methods. *J. Coll. Stud. Retent. Res. Theory Pract.* **2011**, *13*, 17–35. [CrossRef]
33. Barramufio, M.; Meza-Narvaez, C.; Galvez-Garcia, G. Prediction of student attrition risk using machine learning. *J. Appl. Res. High. Educ.* **2022**, *14*, 974–986. [CrossRef]
34. Binu, V.S.; Mayya, S.S.; Dhar, M. Some basic aspects of statistical methods and sample size determination in health science research. *AYU* **2014**, *35*, 119. [PubMed]
35. Hao, J.; Ho, T.K. Machine learning made easy: A review of scikit-learn package in python programming language. *J. Educ. Behav. Stat.* **2019**, *44*, 348–361. [CrossRef]
36. Breiman, L. Random forests. *Mach. Learn.* **2021**, *45*, 5–32. Available online: https://link.springer.com/article/10.1023/A:1010933404324 (accessed on 21 July 2023). [CrossRef]
37. Brownlee, J. Why Use Random Forest for Machine Learning? *Mach. Learn. Mastery* **2020**, *31*, 31. Available online: https://www.ibm.com/topics/random-forest (accessed on 21 July 2023).
38. McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed.; O'Reilly Media: Newton, MA, USA, 2011.
39. Sousa, R.V.; Silva, J.R.P.; Fernandes, M.G. Python-Based Framework for Machine Learning in Medical Imaging. *IEEE Access* **2021**, *9*, 106546–106560.
40. Zhang, S.; Li, T. Comparative Study of Machine Learning Algorithms Implemented in Python for Predictive Maintenance. *IEEE Access* **2021**, *9*, 64572–64582.
41. Wang, C.; Guo, R.; Chen, J. Scalable Deep Learning Framework Using Python for Financial Data Analytics. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 1230–1241.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.