J. Electrical Systems 20-10s (2024):3741-3750

[1] **Omiros Iatrellis**
[2] **Nicholas Samaras**
[3] **Konstantinos Kokkinos**
[4] **Costas Chaikalis**

# Predictive Analytics in Education: Evaluating Machine Learning Methods for Student Dropout Prediction

***Abstract: -*** Student dropout remains a pressing concern with significant socio-economic implications. This study utilizes supervised machine learning to forecast potential dropouts by analyzing a diverse array of factors including academic achievements, class attendance, socio-economic backgrounds, and behavioral patterns. These factors are integrated into a comprehensive predictive model that enhances our understanding of student retention and informs the design of targeted interventions. Through a comparative analysis of two prominent algorithms, K-Nearest Neighbors and Naive Bayes, our research assesses the effectiveness of these methods using a detailed dataset. The findings reveal that the Naive Bayes algorithm outperforms K-Nearest Neighbors in predicting student dropouts, offering valuable data for educational practitioners focused on data-driven strategies to enhance student retention. The study advances the application of machine learning in educational settings and contributes practical insights for the development of policies and interventions aimed at reducing dropout rates, thereby enriching the academic discourse and improving educational outcomes.

***Keywords:*** student dropout prediction, educational data analytics, machine learning, algorithmic performance comparison.

## I. INTRODUCTION

Educational attainment is a cornerstone of individual success and societal progress. However, the journey through education is fraught with challenges that lead some students to disengage and eventually drop out, resulting in significant personal and societal costs [1]. Addressing this issue requires innovative tools that can predict and mitigate the risk of dropout early in a student's academic career [2]. Artificial Intelligence (AI) offers a powerful framework for addressing complex issues by uncovering patterns within large datasets. Within AI, the subcategory of machine learning is particularly promising for tackling educational challenges such as predicting student dropout risks [3]. Supervised machine learning, where models are trained on labeled datasets to predict outcomes for new data, effectively identifies students at risk of dropping out based on a range of predictive indicators. This approach enables educators to design tailored interventions that help keep students on track, optimizing educational outcomes through data-driven insights.

In response to these challenges, the INVEST University alliance has developed the EDUC8EU platform as part of the Erasmus+ European Universities initiative [4]. The alliance aims to foster collaboration among seven European universities, developing innovative educational and research programs that embrace varied learning strategies such as multilingual learning, blended and work-based learning, and European mobilities [5]. The EDUC8EU platform, equipped with an intelligent academic advising system, streamlines the advising process, reduces inconsistencies, and optimizes resource utilization [6].

This study utilizes the EDUC8EU infrastructure, integrating it with advanced machine learning techniques to demonstrate the potential of this platform in the educational domain [7]. We aim to develop and compare two machine learning models—the K-Nearest Neighbors (KNN) and Naive Bayes algorithms—to predict student dropout risks. These models were chosen for their proven capabilities in classification tasks across various domains and are applied here to show how supervised learning can leverage the comprehensive capabilities of EDUC8EU. By employing these models on data collected through the EDUC8EU system, this study seeks not only to determine which algorithm better predicts student dropout but also to contribute to the broader field of educational research. This research demonstrates the practical application of machine learning in real-world educational settings,

[1] Dept. of Digital Systems, University of Thessaly, Larissa, Greece. iatrellis@uth.gr
[2] Dept. of Digital Systems, University of Thessaly, Larissa, Greece. nsamaras@utth.gr
[3] Dept. of Digital Systems, University of Thessaly, Larissa, Greece. kokkinos@uth.gr
[4] Dept. of Digital Systems, University of Thessaly, Larissa, Greece. kchaikalis@uth.gr

potentially paving the way for more personalized, data-driven approaches to student retention and success. Through this effort, EDUC8EU, supported by the resources and collaboration of the INVEST alliance, aims to function as a valuable asset for supporting students across all academic levels—undergraduate, graduate, and doctoral.

The structure of this paper includes Section II, which reviews related literature, Section III, describing our methodology and data collection, Section IV, presenting the comparative results of the K-Nearest Neighbors and Naive Bayes algorithms, and Section V, which concludes with the implications of our findings and suggestions for future research.

## II. Literature Review

### *2.1 Advancements in Machine Learning in Education*

The burgeoning field of educational technology has seen significant advancements in the use of machine learning to enhance academic advising systems. This literature review explores various innovative approaches to academic advising, highlighting the integration of machine learning techniques and the effectiveness of these systems in improving student guidance and retention:

The integration of fuzzy logic and semantic web technologies in academic advising systems, such as the EDUC8EU platform, has demonstrated substantial effectiveness in providing personalized learning pathways [6]. This approach utilizes a hybrid software infrastructure that combines expert systems with fuzzy reasoning and ontological tools to offer tailored recommendations to students. Such systems not only address the inherent fuzziness in learner models but also enhance decision-making by incorporating dynamic, context-aware advice based on a comprehensive analysis of individual student needs.

Personalized learning object recommender systems leverage user data to recommend educational content that aligns with individual learning preferences and needs. These systems employ machine learning algorithms to analyze user interactions and performance data, thereby facilitating a more customized learning experience. The effectiveness of these systems is evident in their ability to adapt recommendations in real-time based on continuous learner feedback [8].

Adaptive learning systems that utilize machine learning to adjust the educational content based on the learner's pace and understanding have shown promising results in improving student engagement and outcomes. These systems analyze the performance data to identify patterns and predict future learning paths that might be most effective for the student, thus personalizing the learning experience at scale [9].

Tools that integrate machine learning to provide career pathway advice based on past performance, interests, and market trends are increasingly being used. These systems utilize predictive analytics to forecast career trajectories and recommend educational paths that align with market demands and individual student profiles, thereby enhancing the relevance and timeliness of academic advice [10].

Recent advancements include the development of decision support systems that incorporate multiple criteria, such as academic performance, psychological factors, and personal preferences. These systems use complex algorithms to weigh various factors and provide a holistic view of the student's options, enhancing the quality and specificity of the guidance provided [11].

### *2.2 Multifaceted Approaches to Predicting Student Dropout*

Predictive modeling in education has significantly evolved, incorporating a wide array of data ranging from academic records to socio-economic and engagement metrics. This literature review examines several seminal studies that illustrate the breadth of methodologies and the importance of a multifaceted approach in dropout prediction.

Gershenfeld et al. highlighted the critical role of cumulative GPA and course completion rates across multiple institutions [12]. Their extensive analysis confirmed that students with lower GPAs are consistently at higher risk of dropping out, suggesting that academic performance is a strong predictor of student retention. However, they also noted the need for integrating support systems early in the academic journey to assist at-risk students.

Aina et al. explored how socio-economic factors such as financial stress and family income influence students' decisions to leave school [13]. Their findings indicated that students experiencing financial difficulties are significantly more likely to drop out, underscoring the need for financial aid and counseling services within educational institutions to help mitigate these challenges.

Mahuteau et al. focused on the impact of social integration on student persistence by analyzing peer interaction data [14]. The study concluded that a robust social network significantly enhances student retention, pointing to the importance of fostering a supportive community environment within educational settings.

Kuhn et al. investigated how students' perceptions of their curriculum's relevance to their career goals influenced their motivation and educational outcomes. Their research showed that students who perceive a high alignment between their studies and prospective careers exhibit higher persistence rates, highlighting the value of career-oriented educational planning [15].

Cuseo provided insights into the effects of proactive academic advising on retention [16]. By focusing on advisor-student engagement, their study supported the notion that regular and meaningful interactions with academic advisors can play a crucial role in reducing dropout rates, particularly for students identified as at-risk [17].

Nurmalitasari et al. quantified the impact of rising tuition fees on student dropout rates [18]. Their analysis provided compelling evidence that increases in educational costs are directly correlated with higher dropout rates, emphasizing the importance of economic considerations in dropout predictions.

Behr et al. delved into the demographics of dropout, focusing on age as a significant factor. Their work suggested that non-traditional, older students face unique challenges that can affect their educational persistence, advocating for tailored support structures for different age groups within academic institutions [19].

Each of these studies contributes to our understanding of the complex factors influencing student dropout. Our research builds upon these foundations by introducing a novel integration of diverse attributes, including academic performance, socio-economic status, and student engagement metrics. This literature review confirms the necessity of our approach, which combines multiple dimensions into a single predictive model, setting it apart from more narrowly focused studies. Additionally, our implementation of the EDUC8EU platform, now enhanced with advanced supervised learning techniques, provides educational institutions with a robust tool for early identification of at-risk students. This integration of advanced data analytics and practical technological solutions promises to improve intervention strategies and ultimately reduce dropout rates, fulfilling a crucial need in educational research and practice.

### *2.3 Our Contributions*

This study makes several significant contributions to the field of educational research. Through meticulous methodology and analytical techniques, our research offers novel insights and practical tools for educational institutions. The key contributions of this study are as follows:

- *Comprehensive model development for understanding student dropout*: Our predictive model incorporates a wide range of factors, from socio-economic backgrounds to academic performance and engagement metrics. This holistic approach broadens the understanding of why students drop out and what can be done to prevent it, providing valuable insights that go beyond traditional single-factor analyses. By integrating these diverse predictors, our model offers a more nuanced view of the complexities surrounding student retention, enabling targeted interventions that are tailored to the specific needs of at-risk students.

- *Advanced Analysis with Machine Learning*: This study rigorously tested two supervised learning algorithms to demonstrate their effectiveness in predicting student dropouts. The comparative analysis of the abovementioned algorithms under various configurations not only highlights the strengths and limitations of each approach but also advances the application of machine learning in educational settings. This methodological rigor provides a solid foundation for educational institutions looking to implement predictive analytics.

- *Technological implementation and practical tools for institutions*: Our study provides educational institutions with the technological means to adopt and perform the predictive approach. We have implemented the EDUC8EU platform, which is now enhanced with supervised learning techniques. This integration equips institutions with a robust tool for early identification of at-risk students, enabling proactive interventions and support.

## III. Research Methodology

### *3.1 Integration of the EDUC8EU Framework in Dropout Prediction*

The predictive model for student dropout has been designed around a core framework adapted from EDUC8EU, an intelligent academic advising system known for its comprehensive assessment of student needs across various dimensions. The EDUC8EU approach is a sophisticated platform originally designed for academic advising, integrating advanced technologies such as artificial intelligence, fuzzy logic, and semantic web tools [2], [20]. For our study, EDUC8EU's diverse components were instrumental in gathering comprehensive data about students' academic experiences, personal motivations, skills, and personality types. An example output of EDUC8EU is depicted in Fig. 1. These data were crucial for feeding into our advanced machine learning analysis. EDUC8EU comprises the following main components:

1) *Interest and Motivation Detection*: Utilizing fuzzy logic, this component of EDUC8EU assessed students' engagement levels by collecting data on their academic performance, perceived difficulty of their courses, and their career prospects. This data is crucial for understanding students' commitment and satisfaction with their educational pathways.

2) *Skill Gap Analysis*: Through integration with the ESCO and O*NET frameworks, EDUC8EU evaluated the skills of students against those required by their programs. This assessment was pivotal in identifying specific areas where students lacked the necessary competencies, pinpointing risks that could potentially lead to dropout.

3) *Personality Matching Using RIASEC*: By assessing students' personalities and matching them with appropriate academic environments, EDUC8EU ensured that each student's educational journey was aligned with their inherent strengths and preferences [21]. Data collected from these assessments informed our model about the likelihood of student persistence based on personality-environment fit.

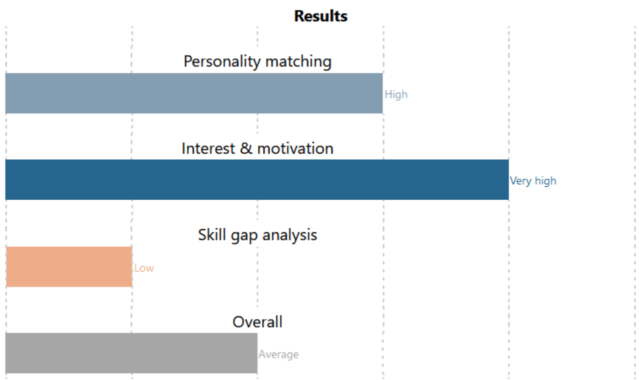


**Fig 1.** Sample output of EDUC8EU components

Once comprehensive data were collected through the EDUC8EU platform, we created a predictive model containing the following factors based on the EDUC8EU model presented in [6] that impact student retention:

1) *Age* (Numeric): Age is a fundamental demographic factor that can influence educational engagement and outcomes [19].

2) *Perceived Difficulty of the Learning Pathway/Object* (Ordinal): This factor is included in the model to capture the perceived difficulty of academic programs, which is directly linked to student stress and anxiety [12]

3) *Friends and Peers in the Learning Pathway* (Numeric): Our model incorporates the number of friends and peers within the learning pathway, as peer networks provide social integration that can enhance student persistence [14].

4) *Perceived Reputation of the Learning Object* (Ordinal): The perceived reputation of educational programs and institutions is a key factor in the model, contributing to student satisfaction and retention [15].

5) *Cost of the Learning Pathway* (Numeric): The financial strain of the learning pathway is included in our model to account for its direct influence on a student's ability to continue their studies[18].

6) *GPA Average* (Numeric): Academic performance is a critical predictor in our model, strongly indicating student retention [12].

7) *Average Study Hours per Week* (Numeric): This metric is incorporated into the model as it closely links to academic success and engagement [12].

8) *Perceived Career Opportunities* (Ordinal): The alignment of educational programs with career aspirations, measured by perceived career opportunities, is included in the model to reflect its significant impact on student motivation and persistence [15].

*3.2 Machine Learning Algorithms*

In this study, we utilized the EDUC8EU infrastructure, which encompasses the eight attributes previously described, to apply and compare the effectiveness of two well-established machine learning algorithms - Naive Bayes and KNN. These algorithms were specifically selected for their proven effectiveness in classification tasks across various domains and their unique capabilities to efficiently handle diverse types of data, making them particularly suited to the multifaceted nature of educational datasets.

The Naive Bayes algorithm is favored for its simplicity and effectiveness in classification scenarios, operating under the assumption of independence between predictors [22]. This simplicity is deceptive, as it allows for high efficiency and good performance, especially in scenarios where the dimensions of the input features are high and the data are abundant. Naive Bayes excels in handling categorical data and has been successfully applied in various educational applications, from predicting student performance to analyzing learning behaviors [23].

In the context of dropout prediction, Naive Bayes offers the following advantages:

- *Probabilistic understanding*: It provides a probabilistic output for each prediction, offering a clear understanding of the likelihood of dropout, which is invaluable for making informed interventions.
- *Scalability*: The algorithm's efficiency and scalability make it well-suited for handling large educational datasets, ensuring quick processing and timely insights.
- *Performance with incomplete data*: Naive Bayes can perform well even with incomplete data profiles, common in educational settings, by making use of probabilistic inference.

Contrastingly, the KNN algorithm classifies based on the closest training examples in the feature space, making it a non-parametric method that doesn't assume a specific distribution for the data [24] This characteristic is particularly beneficial for educational data, which often contains irregular and atypical patterns not easily captured by parametric methods [25].

KNN is chosen for its:

- *Flexibility and simplicity*: KNN is inherently simple yet powerful in capturing complex relationships in data through its distance-based approach.
- *Adaptability*: It adapts well to changes in the input data, making it suitable for dynamic educational environments where student data evolve over time.
- *Interpretability*: The rationale behind each classification is transparent and easy to communicate to educators, as it is based on similarity measures with known cases.

Both algorithms have demonstrated substantial success in various predictive modeling tasks beyond education, which validates their robustness and adaptability. By employing these models on data collected through the EDUC8EU system, this study not only seeks to determine which algorithm better predicts student dropout but also aims to showcase how traditional machine learning techniques can be innovatively applied in modern educational contexts. The comparative analysis under various configurations highlights the strengths and limitations of each approach, providing a solid foundation for educational institutions looking to implement predictive analytics.

*3.3 Model Testing and Evaluation Model Testing and Evaluation*

In this study, the dataset underwent a split-sample approach, divided into varying training data ratios from 80% to 10%. This approach enabled preliminary assessments of each machine learning model under different conditions. Using the KNN and Naive Bayes algorithms, several configurations were tested to explore the stability and performance of each setup. The models are tested using several key performance metrics:

- *Accuracy*: This measures the overall correctness of the model, defined as the ratio of true predictions to total cases.
- *Precision*: This evaluates the model's accuracy in predicting positive labels.
- *Recall*: This assesses the model's ability to detect all actual positives.

By assessing each configuration, we could determine which conditions favor specific algorithms, providing valuable insights into their practical applications and effectiveness within the educational context. This structured approach ensures a comprehensive evaluation of the predictive capabilities of each algorithm, contributing significantly to our understanding of their application in real-world educational settings.

To further validate the stability and reliability of our models, we employ k-fold cross-validation. This technique involves dividing the dataset randomly into 'K' parts. In each fold of validation, one part is retained as the validation data for testing the model, and the remaining 'K-1' parts are used as training data. This process is repeated K times, with each of the K parts used exactly once as the validation data. Here's the mathematical representation often used to describe k-fold cross-validation:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} ModelEvaluation_i \qquad (1)$$

where $\text{ModelEvaluation}_i$ represents the evaluation metric (accuracy, precision, recall) calculated for each fold. This method not only helps in fine-tuning the models but also in assessing their effectiveness in a more generalized scenario beyond the specific split of training and testing data. The results from k-fold cross-validation provide insights into how each model might perform in real-world applications, ensuring that our conclusions about model superiority are robust and reliable.

## IV. Model Preparation and Data Processing

This section details the preparatory steps taken to ensure that our predictive models for student dropout are both reliable and accurate. It encompasses the collection, preprocessing, and cleaning of data sourced from the EDUC8EU platform, which is integral to the success of our analysis.

### *4.1 Data Collection*

Data collection was conducted through the EDUC8EU platform, developed as part of the INVEST European University alliance, during the first pilot phase of the project spanning the years 2021-2024. This platform has been instrumental in gathering a diverse and comprehensive dataset, comprising 790 student cases across various faculties within the alliance. The collected data encompasses multiple facets of students' academic lives, as detailed in section III. These data points are crucial for understanding the multi-dimensional nature of student dropout and provide a robust foundation for analyzing the factors that influence educational outcomes.

### *4.2 Data Pre-Processing*

The preprocessing stage involved several key steps to prepare the data for analysis:

- *Encoding*: All categorical data were transformed into a numerical format suitable for processing by machine learning algorithms. This included encoding scales for perceived difficulty, reputation, and other ordinal or nominal scales.
- *Normalization*: To ensure that each attribute contributes proportionately to the predictive models, we implemented normalization techniques. This was particularly important for features such as GPA, which vary significantly across different scales.

- *Handling Missing Data*: We addressed the issue of missing data with a combination of imputation for minor missing values and exclusion for cases where significant data points were missing. This approach helped maintain the integrity of our predictive models without introducing substantial bias.

*4.3 Data Cleaning*

The final step before model training involved rigorous data cleaning to enhance the quality of the dataset:

- *Error Checking*: We performed a systematic review of the dataset to identify and correct any anomalies or errors, such as out-of-range values for Age (e.g., below 18).

- *Removing Outliers*: Statistical methods were employed to identify outliers that could potentially skew the results of the analysis. These outliers were assessed and removed only if their exclusion was justified by statistical reasoning, ensuring that our dataset accurately reflected the student population.

- *Data Integration*: Although our dataset was primarily collected from the EDUC8EU platform, additional checks ensured that data formats and scales were consistent across different sources and variables.

*4.4 Results*

The KNN and Naive Bayes algorithms were effectively employed to predict student dropouts, analyzing a broad spectrum of educational and behavioral attributes with varying training data percentages from 80% to 10%. For the KNN algorithm, the optimal k value was determined through a series of validation tests that identified k=5 as providing the best balance between bias and variance, optimizing the accuracy while mitigating overfitting. This value was selected after testing different k values from 1 to 20, assessing their performance in preliminary runs across multiple data splits.

In contrast, the Naive Bayes algorithm was configured with a focus on enhancing predictive power by employing variable smoothing techniques. This approach adjusted the algorithm's sensitivity to feature variability, particularly beneficial given the diverse range of predictors involved in the study. The optimal parameter settings were identified through these experiments, with a specific focus on maximizing recall to ensure that as many at-risk students as possible were correctly identified by the model. This focus was crucial in an educational context where the cost of false negatives—failing to identify a student at risk of dropping out—could have significant implications for the student’s educational journey.

Classification accuracies for these algorithms are detailed in Table 1, covering eight distinct experiments. The table reveals that the highest accuracy was recorded in the first experiment with 80% training data, where KNN achieved an accuracy of 79.59% and Naive Bayes slightly outperformed it with an accuracy of 80.46%. The superior accuracy in the initial experiment prompted a series of further tests to compare these algorithms at different training data levels. Naive Bayes consistently outperformed KNN across all trials, maintaining an average accuracy of 77.47% compared to KNN’s 74.17%.

The charts (Fig. 2) and table above show that Naive Bayes generally provides higher accuracy, recall, and precision across most training percentages, indicating robust performance even with lesser training data. KNN shows reasonable consistency but slightly lower performance metrics, which could be attributed to its sensitivity to the k-value and the local data structure.

The results from the k-fold cross-validation showed that the Naive Bayes algorithm consistently outperformed the KNN algorithm across all folds. The average accuracy for Naive Bayes was 78.92%, with a standard deviation of 1.22%, indicating stable and reliable performance. In contrast, the KNN algorithm achieved an average accuracy of 75.14%, with a slightly higher standard deviation of 1.75%, suggesting more variability in its performance.

**Table 1:** Classification Accuracy from KNN and Naïve Bayes Model

| Attempt | Training Percentage | KNN Accuracy | KNN Recall | KNN Precision | Naïve Bayes Accuracy | Naïve Bayes Recall | Naïve Bayes Precision |
|---|---|---|---|---|---|---|---|
| 1 | 80% | 79.59% | 76.4% | 75.89% | 80.46% | 77.67% | 76.4% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 70% | 79.59% | 74.4% | 76.89% | 78.08% | 75.67% | 74.4% |
| 3 | 60% | 75.70% | 70.4% | 72.89% | 77.54% | 71.67% | 74.4% |
| 4 | 50% | 75.63% | 71.4% | 72.89% | 77.67% | 72.67% | 74.4% |
| 5 | 40% | 75.21% | 71.4% | 72.89% | 77.38% | 72.67% | 74.4% |
| 6 | 30% | 71.74% | 68.4% | 68.89% | 78.47% | 74.67% | 76.4% |
| 7 | 20% | 69.80% | 65.4% | 66.89% | 77.34% | 71.67% | 74.4% |
| 8 | 10% | 72.77% | 68.4% | 69.89% | 76.75% | 72.67% | 73.4% |

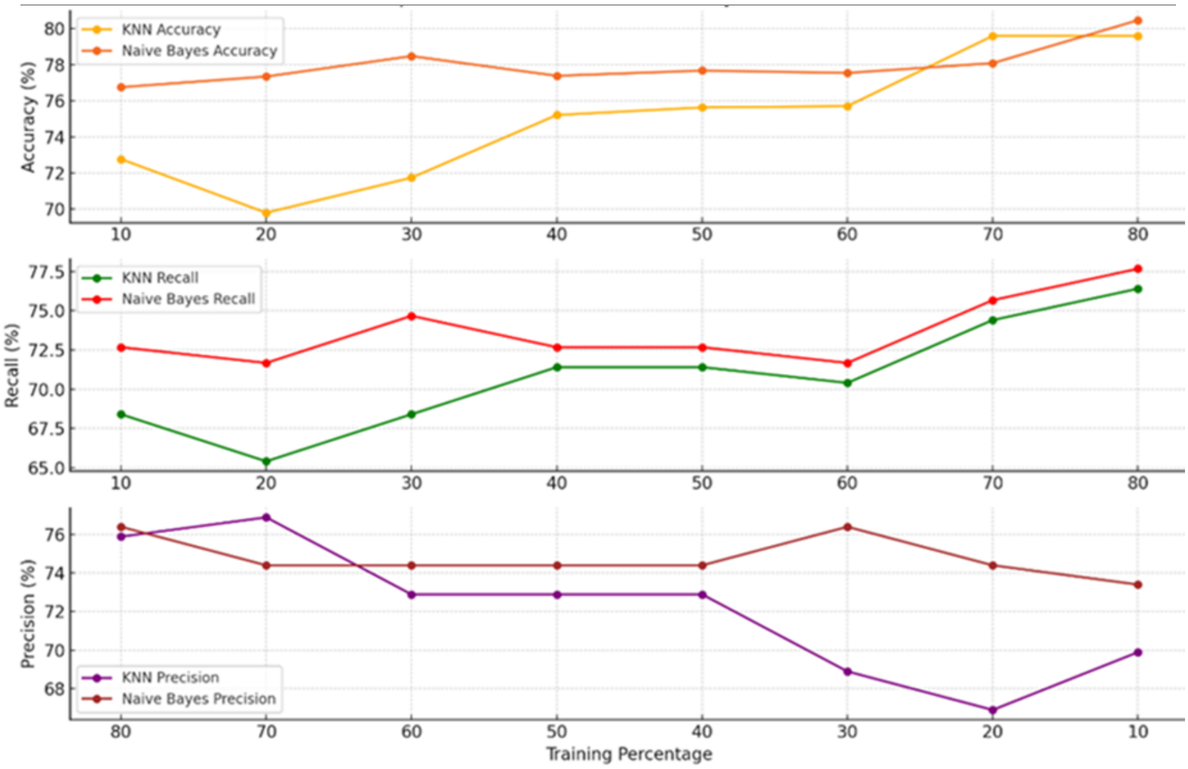


**Fig. 2.** Comparison of Naïve Bayes and KNN performance

While both algorithms demonstrated potential in predicting student dropouts, Naive Bayes consistently provided better performance. The consistency in performance metrics across both evaluation methods highlights the robustness of the Naive Bayes model and its effectiveness for early dropout prediction in educational settings.

## V. Conclusions

This study assessed the effectiveness of the supervised learning algorithms in predicting student dropout within the INVEST European University alliance, utilizing the comprehensive dataset provided by the EDUC8EU platform. Our analysis revealed that Naive Bayes consistently outperformed KNN across various training data proportions, achieving the highest accuracy of 80.46% with 80% training data compared to KNN's 79.59%. The average accuracy for Naive Bayes was 77.47%, surpassing KNN's 74.17%. These findings, validated through k-fold cross-validation, underscore the robustness and reliability of Naive Bayes as a critical tool for early dropout prediction. The EDUC8EU framework played a pivotal role in this research by facilitating the collection and integration of diverse educational data, which enabled a detailed and dynamic assessment of dropout risks. Its advanced data handling and analysis capabilities were instrumental in enhancing the predictive accuracy of our models.

Future research will focus on several areas to enhance the predictive capabilities and practical applications of the models. Firstly, expanding the dataset to include larger and more diverse samples from multiple educational institutions would improve the model's generalizability. Incorporating additional variables, such as psychological factors and extracurricular activities, could also enhance predictive accuracy by providing a more comprehensive understanding of student behavior and engagement. Moreover, exploring advanced machine learning algorithms, including Random Forests and deep learning models, could yield higher accuracy and offer comparative insights. Developing real-time prediction systems is another crucial area, enabling immediate feedback to educators and administrators based on live data, which would facilitate timely interventions.

## VI. Acknowledgement

The authors would like to thank the INVEST4EXCELLENCE project under the H2020-IBA-SwafS-Support-2-2020 program (Project No. 101035815, www.invest-alliance.eu) for providing support and thank the other project partners.

## References

[1] N. M. Alruwais, "Deep FM-Based Predictive Model for Student Dropout in Online Classes," *IEEE Access*, vol. 11, pp. 96954–96970, 2023, doi: 10.1109/ACCESS.2023.3312150.

[2] O. Iatrellis, A. Kameas, and P. Fitsilis, "EDUC8 : Self-evolving and Personalized Learning Pathways Utilizing Semantics," *2018 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp. 1–8, 2018.

[3] B. N. Anh, N. H. Giang, N. Q. Hai, T. N. Minh, N. T. Son, and B. D. Chien, "An University Student Dropout Detector Based on Academic Data," in *2023 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ISIEA58478.2023.10212223.

[4] "INVEST alliance." Accessed: Jul. 19, 2021. [Online]. Available: https://www.invest-alliance.eu/

[5] K. Kokkinos, N. Samaras, and O. Iatrellis, "European Universities Best Practices: The Case of INVEST EU Alliance," *European Scientific Journal ESJ*, vol. 25, Jan. 2024, doi: 10.19044/esipreprint.1.2024.p135.

[6] O. Iatrellis, E. Stamatiadis, N. Samaras, T. Panagiotakopoulos, and P. Fitsilis, "An intelligent expert system for academic advising utilizing fuzzy logic and semantic web technologies for smart cities education," *Journal of Computers in Education*, vol. 10, no. 2, pp. 293–323, Jun. 2023, doi: 10.1007/s40692-022-00232-0.

[7] O. Iatrellis, A. Kameas, and P. Fitsilis, "Personalized Learning Pathways using semantic web rules," in *ACM International Conference Proceeding Series*, 2017. doi: 10.1145/3139367.3139404.

[8] H. Imran, M. Belghis-Zadeh, T.-W. Chang, Kinshuk, and S. Graf, "PLORS: a personalized learning object recommender system," *Vietnam Journal of Computer Science*, vol. 3, no. 1, pp. 3–13, Feb. 2016, doi: 10.1007/s40595-015-0049-6.

[9] A. Essa, "A possible future for next generation adaptive learning systems," *Smart Learning Environments*, vol. 3, no. 1, Dec. 2016, doi: 10.1186/s40561-016-0038-y.

[10] Y. , P. C. , Y. G. , & B. J. Chen, "Intelligent decision system for accessing academic performance of candidates for early admission to university," in *10th International Conference on Natural Computation (ICNC)*, 2014, pp. 687–692.

[11] R. Duarte, A. L. de Oliveira Pires, and Â. L. Nobre, "Mature Learners' Participation in Higher Education and Flexible Learning Pathways: Lessons Learned from an Exploratory Experimental Research," in *Contributions to Higher Engineering Education*, Springer Singapore, 2018, pp. 33–53. doi: 10.1007/978-981-10-8917-6_2.

[12] S. Gershenfeld, D. W. Hood, and M. Zhan, "The role of first-semester GPA in predicting graduation rates of underrepresented students," *J Coll Stud Ret*, vol. 17, no. 4, pp. 469–488, Feb. 2016, doi: 10.1177/1521025115579251.

[13] C. Aina, E. Baici, G. Casalone, and F. Pastore, "The determinants of university dropout: A review of the socio-economic literature," *Socio-Economic Planning Sciences*, vol. 79. Elsevier Ltd, Feb. 01, 2022. doi: 10.1016/j.seps.2021.101102.

[14] S. Mahuteau and K. Mavromaras, "Education Economics An analysis of the impact of socio-economic disadvantage and school quality on the probability of school dropout An analysis of the impact of socio-economic disadvantage and school quality on the probability of school dropout," 2014, doi: 10.1080/09645292.2014.918586.

[15] K.-A. L. Kuhn and S. R. Rundle-Thiele, "Curriculum Alignment: Exploring Student Perception of Learning Achievement Measures," *International Journal of Teaching and Learning in Higher Education*, vol. 21, no. 3, pp. 351–361, 2009, [Online]. Available: http://www.isetl.org/ijtlhe/

[16] J. Cuseo, "ACADEMIC ADVISEMENT AND STUDENT RETENTION: EMPIRICAL CONNECTIONS & SYSTEMIC INTERVENTIONS."

[17] O. Iatrellis, A. Kameas, and P. Fitsilis, "Academic Advising Systems: A Systematic Literature Review of Empirical Evidence," *Educ Sci (Basel)*, vol. 7, no. 4, p. 90, 2017, doi: 10.3390/educsci7040090.

[18] Nurmalitasari, Z. Awang Long, and M. Faizuddin Mohd Noor, "Factors Influencing Dropout Students in Higher Education," *Educ Res Int*, vol. 2023, 2023, doi: 10.1155/2023/7704142.

[19] A. Behr, M. Giese, H. D. Teguim Kamdjou, and K. Theune, "Dropping out of university: A literature review," *Review of Education*, vol. 8, no. 2, pp. 614–652, 2020, doi: 10.1002/rev3.3202.

[20] O. Iatrellis, A. Kameas, and P. Fitsilis, "EDUC8 ontology: semantic modeling of multi-facet learning pathways," *Educ Inf Technol (Dordr)*, vol. 24, no. 4, 2019, doi: 10.1007/s10639-019-09877-4.

[21] F. Hartmann, J.-H. Heine, and B. Ertl, "Concepts and Coefficients Based on John L. Holland's Theory of Vocational Choice—Examining the R Package holland," *Psych*, vol. 3, no. 4, pp. 728–750, Nov. 2021, doi: 10.3390/psych3040047.

[22] F. F. Balahadia and B. E. V. Comendador, "Adoption of Opinion Mining in the Faculty Performance Evaluation System by the Students Using Naïve Bayes Algorithm," *International Journal of Computer Theory and Engineering*, vol. 8, no. 3, pp. 255–259, Jun. 2016, doi: 10.7763/ijcte.2016.v8.1054.

[23] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowl Based Syst*, vol. 192, p. 105361, 2020, doi: 10.1016/j.knosys.2019.105361.

[24] Y. Li and L. Liu, "Improvement of Parallelization of KNN Algorithm Based on Sequential MapReduce," *International Journal of Computer Theory and Engineering*, vol. 13, no. 2, pp. 56–60, May 2021, doi: 10.7763/IJCTE.2021.V13.1290.

[25] S. Yang, H. Jian, Z. Ding, Z. Hongyuan, and C. L. Giles, "IKNN: Informative K-nearest neighbor pattern classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4702 LNAI, pp. 248–264, 2007, doi: 10.1007/978-3-540-74976-9_25.