Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

Advancing Student Success: A Machine Learning Based Approach for Early Identification of Drop Out in Schools

Yassine Benlachmi
ISIC-TEAM, L2ISEI-Laboratory-ESTM
Moulay Ismail University
Meknes, Morocco

Moulay Lahcen Hasnaoui
CTE-TEAM, EDS- Laboratory-FSE,
Mohammed V University
Rabat, Morocco

Hatim Derrouz
Laboratory LaRi, Faculty of science Kenitra Ibn Tofail University
Kenitra, Morocco

Abdallah Rhathoy
ISIC-TEAM, L2ISEI-Laboratory-ESTM
Moulay Ismail University
Meknes, Morocco

***Abstract*— The achievement of students is crucial in educational institutions, and early identification of potential challenges can significantly improve their chances of success. Machine learning techniques have gained widespread use for predictive purposes but are primarily accessible to educators with expertise in artificial intelligence. Effective application of data mining methods involves defining student success, prioritizing student attributes, and selecting the most suitable method for a problem. This research aims to bridge the knowledge gap between educators and machine learning techniques for predictive analytics in educational institutions. The study synthesizes recent research and existing literature into a methodical procedure, defining student achievement, ranking relevant qualities, and choosing the right techniques for given problems. It explores important decisions and parameters for effective implementation and provides justifications and arguments in favor of these decisions. The research suggests a machine learning-based approach to forecast school dropout, a crucial component of student achievement forecasting. The Random Forest classifier was the most successful, achieving an unmatched prediction accuracy of 100%. The Naive Bayes classifier also showed impressive performance, with an accuracy of 98.8%. These findings demonstrate the method's effectiveness in outperforming advanced techniques and highlight its potential for real-world use. Future research may focus on improving the suggested approach and exploring new features.**

***Keywords*— Dropout Risk Factors; Educational Data Mining; AI in Educational Assessment; Machine Learning in Education; Schools Dropout Prediction**

I. INTRODUCTION

School dropout rate in many regions of the world is a complex issue that extends beyond academic institutions and impacts societal and economic well-being. This issue is not solely due to the end of classes, but rather due to pupils falling between educational system gaps, causing a progressively emptier classroom. In Morocco, the formal primary school admission age is six, and the academic year starts from September till June. The length of the elementary education cycle is six years, followed by three years for lower secondary and three years for upper secondary in the current system. There are 7,058,000 students in primary and secondary school in Morocco. Of these students, 4,211,000 (about 60%) are enrolled in elementary school. Figure 3 displays the greatest level of education attained by Moroccan youth aged 15 to 24. While some young people in this age range may still be enrolled in school and pursuing their educational objectives, it is noteworthy that over 26% of young people has never had any formal education and 23% have at most received an incomplete basic education [1]. Morocco's primary education drop-out rate in 2014 was 11.2%, a decrease from 1995 to 2014. Key statistics from this period include a gross enrolment ratio of 114.8%, a net enrolment rate of 98.4%, and a repeater rate of 9.3%. Private primary education had 17.3% students, with females making up 47.4% of the population. Out-of-school rates were low at 0.4%, and the adjusted net intake rate was 92.5%. The survival rate to the last grade was 94.3%, and the completion rate was high at 97.1% [2]. In the 2020-2021 school year, around 331,000 Moroccan children were denied their fundamental human right to education [3]. HCP research shows 1.5 million Moroccans are not enrolled in education, vocational training, or employment, with women comprising 73% of NEETs, 68% having degrees, and 41% married The drop-out rate in rural Morocco's primary schools was 5.7% in 2017 [4]. In 2005, only 54% of Morocco's rural population had access to an all-weather road, indicating a significant gap in opportunities, markets, and essential social services [5]. The study employs deep embedding clustering to analyze Morocco's PISA dataset, uncovering intricate educational trends and emphasizing the need for region-specific strategies to enhance outcomes. The widespread use of computers in the past 30 years has enabled the collection of vast

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

and diverse data sets, providing unprecedented opportunities to identify patterns and trends [6]. Data mining analytical techniques can be categorized into traditional statistical techniques [7], machine learning [8], and artificial intelligence [9], Deep Learning [10], which are used in various domains for purposes like disease detection [11]–[13], behavior prediction [14], [15], and pattern extraction. Many educational databases have been created as a result of the development of educational database management systems. The use of data mining techniques to extract useful information from these databases has been made possible by this breakthrough. As a result, Education Data Mining (EDM) has emerged as a separate academic discipline [6]. In a study authors [16] presented the challenges faced during the research for students dropout prediction.

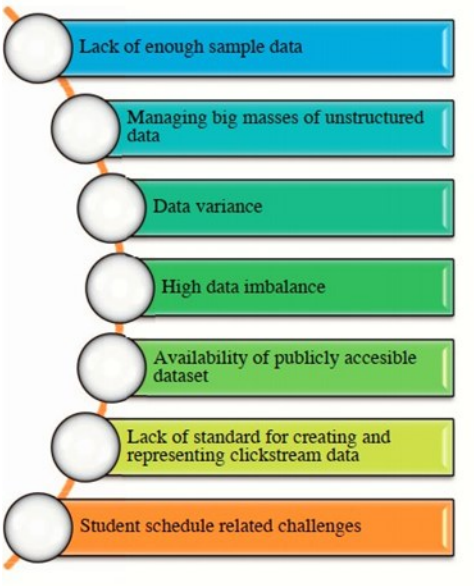

Fig. 1 Dropout Prediction Challenges using ML [16]

Pasina et al. [17] and Lailiyah et al. [18] developed prediction models for evaluating student performance using academic and demographic information. They used clustering algorithms and decision trees to forecast high school student's academic achievements, aiming to provide educational institutions with a risk assessment tool for identifying at-risk students and implementing intervention plans.

University dropout is a complex and negative event that impacts students and schools, causing significant life and economic impacts on both students and universities. It presents difficulties for both students and schools [19]. Dropouts can indicate organizational issues or course quality concerns. Machine learning techniques can predict dropouts by evaluating student's academic performance in education [20], [21]. In the literature, there are many research studies where researchers have proposed methods for predicting higher education dropouts, buthere in this research work, our focus will be on primary school dropouts.

Our research objectives are as follow:

1) Feature Analysis and Selection:

a) Analyze dataset features to identify key variables influencing student dropout.
b) Utilize statistical methods for feature distribution analysis.
c) Select relevant features with strong predictive potential for model development.

2) Model Training and Evaluation:

a) Implement machine learning classifiers (Random Forest, SVM, k-NN, Naive Bayes, MNB, CNB).
b) Train models on preprocessed data, considering encoding and missing value handling.
c) Evaluate model performance using ac- curacy, precision, recall, F1-score, and ROC AUC.

3) Comparative Analysis of Models:

a) Conduct a comparative analysis of model performance, considering inter- pretability and efficiency.
b) Assess strengths and weaknesses of each model for informed decision- making.

4) Interpretation of Results:

a) Interpret machine learning model re- sults to identify significant predictors.
b) Examine confusion matrices and ROC curves for insights into classification abilities.

5) Optimization and Fine-Tuning:

a) Explore model optimization through hyperparameter tuning and feature en- gineering.
b) Refine models iteratively for improved accuracy.

6) Practical Implications and Recommendations:

a) Translate research findings into practical implications for educational stakeholders.
b) Provide actionable recommendations for early identification of at-risk students and targeted interventions.

Through addressing these objectives, the study aims to contribute insights into student dropout prediction, advance educational analytics, and offer practical strategies for improving student retention.

II. RELATED WORK

The global issue of children and young people not attending school persists, with around 258 million out of school in 2018 [22]. The UN General Assembly adopted 17 Sustainable Development Goals (SDGs) in 2015 [23], with 169 targets aiming for 2030. The fourth SDG aims to ensure all girls and boys complete free, equitable, and quality primary and secondary education. However, three years later, no progress has been made in reducing the global number of out-of-school

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

children, adolescents, and youth [22]. The study [24], examines the reasons behind student dropouts in rural Pakistan, focusing on fathers whose sons have also dropped out of secondary education. The research employs a qualitative approach and conducts in-depth interviews with 14 fathers to understand the factors contributing to student dropouts. The fathers found family poverty, poor academic performance, and issues related to teachers' engagement with teaching as the primary factors. The study highlights the complexities of the tension in mainstream teachers' educational roles and the urgent need for increased investment in secondary education, especially in remote rural areas. The research contributes to the ongoing discourse on educational challenges in developing regions and sheds light on the multifaceted factors influencing secondary school dropout through the lens of fathers in rural Pakistan.

The study [25] focuses on social, economic, political, and environmental aspects, and provides insights into reasons for dropout, social and economic factors, government initiatives, school environment, geographical considerations, social norms, and expert opinions. The study highlights the interconnected nature of dropout and economic development, highlighting potential long-term consequences for the nation. Key causes of dropout include chronic poverty, parental unwillingness, financial constraints, poor school infrastructure, biased social practices, inadequate education quality, and geographical isolation, unequal access to education, and security concerns for girls. School dropout is a significant issue, impacting various aspects of society, including social, economic, political, and academic domains. Francisco and al [26] introduces an IoT framework to predict dropout using machine learning methods like Decision Tree, Logistic Regression, support Vector Machine, K-nearest neighbors, Multilayer perceptron, and Deep Learning. The system uses socioeconomic data from preregistration to identify students at risk of dropout. The automation of the prediction process allows for more precise and efficient predictions, enhancing management and service-related functions. The system's effectiveness was validated, with the Decision Tree method achieving 99.34% accuracy, F1 score, 100% recall, and 98.69% precision. This system offers a promising solution for universities to predict and address potential student dropouts effectively. The systematic literature review [27] examines the challenges and opportunities faced by the urban poor in accessing education. It reveals that despite attending school, the urban poor struggle to fully realize their right to education, with deficiencies in learning outcomes and educational trajectory. The review emphasizes the need for comprehensive public policies and a multisectoral approach to break the cycle between urban poverty and education, highlighting the complexity of the challenge and the need for a multi-sectorial approach.

The study [28] compares machine learning models for the early prediction of student performance using a dataset from Ubon Ratchathani University. The results show Random Forest's superiority with an 86.87% F1-measure. Key predictors like national test scores and entry types were identified. The study provides practical insights for educational institutions to enhance student performance, plan strategies, and make informed decisions, offering a robust foundation for improving student success in diverse educational contexts.

Recent studies have focused on predicting student dropout rates in traditional face-to-face courses. Researchers in Brazil used school data from 2011 to 2016 to identify students at risk [29], [30]. They used predictive factors like age, gender, and demographic region. The Decision Tree classifier achieved a 69% [29] accuracy rate, while logistic regression algorithm is utilized by the authors Calixto et al. [30] and achieved an 87% accuracy rate. The re- search yielded significant findings, highlighting the importance of predicting student dropout rates in education. Despite efforts to promote inclusivity in the Indian education system, 35-

60 million children aged 6-14 years are not enrolled, exacerbated by gender, regional, and caste-based inequalities. Key challenges include high dropout rates, insufficient learning outcomes, inadequate school infrastructure, teacher absenteeism, numerous teacher vacancies, sub- par education quality, and insufficient funding. Vulnerable groups, such as orphans, child la- borers, street children, and victims of riots and natural disasters, also face barriers to accessing education [31]. Enrolling all eligible school-age children is crucial for achieving 100% literacy in India. Low retention can be attributed to subpar education, particularly in slums due to factors like low parental literacy rates, socioeconomic status, high juvenile delinquency rates, and low status of female children.

The study [32] evaluates artificial neural net- work (ANN)-based educational data mining (EDM) in higher education, focusing on modeling, learning procedures, and cost function optimization. It identifies hardware, training, theoretical, and data quality challenges. The findings provide a roadmap for future research and guide future scholars in addressing existing is- sues. The study aims to enhance the visibility and relevance of ANN-based EDM, contributing to sustainable development in higher education. The findings will help guide future scholars to- wards areas where their contributions can make a meaningful impact. The study by [33] used an elementary school database from 2008 in North Carolina to study predictive modeling for student outcomes. They used the Support Vector Machines (SVM) classifier, which had a remarkable accuracy rate of 90.80%. This high accuracy indicates the SVM classifier's effectiveness in predicting outcomes, providing valuable insights for educators, administrators, and policymakers to improve interventions and support systems to reduce dropout rates and foster academic success. Aulck et al. [34] study uses longitudinal student records from the North Carolina Department of Public Instruction to evaluate prediction techniques for high school graduation and dropout risk. They use tree-based classification methods and support vector ma- chines, incorporating 74 predictors from Grades 3-8. The findings show a shift toward lower dropout rates, particularly among male students, and suggest machine learning can identify at- risk students for targeted intervention programs. Jacqueline Knapke et al. [35] analyzed learning analytics in higher education, highlighting its potential for improving learning environments and outcomes. They highlighted benefits like student retention, progression, and cost savings from

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

educational data mining (EDM). However, the study also emphasized existing challenges and aimed to guide stakeholders in efficient application of EDM, ultimately enhancing teaching and learning experiences. The article did not explore the use of artificial neural networks (ANNs) in implementing EDM in higher education. The study [36] aims to improve dropout prediction in Massive Open Online Courses (MOOCs) by incorporating behavior features and adopting a multi-view semi-supervised learning approach. The authors aim to overcome limitations in current methods, which often use general features without considering diverse student learning behaviors. They also address the issue of insufficient labeled data for training models. They propose a novel multi-view semi-supervised learning model, specifically designed for MOOC dropout prediction, leveraging various types of learning behaviors. This innovative approach acknowledges the diversity in student learning behaviors and leverages a large pool of unlabeled data to enhance prediction performance. The study's experiments on the KDD Cup 2015 dataset show that the proposed method outperforms state-of-the-art approaches, demonstrating significant advancements in dropout prediction.

This study [37] uses machine learning algorithms to predict student performance, prevent academic failure, and identify factors contributing to dropout. It uses data from 15,825 undergraduate students at Budapest University of Technology and Economics, spanning 2010 to 2017. The models are trained using Decision Tree-based algorithms, Naive Bayes, k-NN, Linear Models, and Deep Learning. The best models, Gradient Boosted Trees and Deep Learning, achieved AUC values of 0.808 and 0.811, respectively. The findings highlight the potential of machine learning in predicting student outcomes and dropout risk, providing valuable insights for educational stakeholders. As the global economy evolves, it is widely acknowledged that advanced education plays a pivotal role in fostering competitiveness and success for both individuals and countries. This recognition has cemented the pursuit of a master's degree as a crucial part of postgraduate education. Such degrees not only enhance existing abilities but also equip students with new skills pertinent to specific careers. Despite the growing global interest in master's programs, there is a noticeable trend of high dropout or failure rates. This issue can be observed from two perspectives: the university, which experiences the loss of a student, and the student, who discontinues their academic journey. University dropouts are seen as an academic setback, and the urgency to address this issue stems from four key reasons: economic, social, individual, and pedagogical [38]. In response, various countries have implemented initiatives aimed at increasing the number of highly educated individuals needed to meet the demands of a knowledge-driven society and economy [39].

The study [40] aims to tackle high-school dropout in Denmark by using machine learning techniques for prediction. The research uses a large-scale study, focusing on predicting dropout within three months for students already six months into their education. The study integrates data from the MaCom Lectio study ad- ministration system and public online sources to improve prediction accuracy. The study is the first of its kind, involving a larger sample size of 36,299 pupils for both training and testing phases. The results show that a random forest classifier achieved remarkable accuracy of 93.47%, providing valuable insights for educational institutions seeking effective preventive measures. This study demonstrates the potential of machine learning for dropout prediction on a broader scale.

Machine learning [41], is a data analysis methodology that uses supervised learning to identify patterns and make judgments. It uses labeled training data to infer functions, enabling forecasts and informed decision-making. This framework reveals hidden patterns and complex data links, making it crucial for delicate subjects like university student retention. There exist several research studies in which authors have utilized advanced machine learning algorithms for the dropout prediction from schools and universities. Numerous studies have been conducted to explore the factors influencing school dropout in Morocco, whether at the university level or within the context of MOOCs. However, none of these studies has focused on primary education dropout. It is important to emphasize that the dropout rate remains particularly high in this category in Morocco, especially in rural regions, despite the efforts deployed by the Minister of National Education, Preschool, and Sports, as well as more broadly by the Moroccan state. The research study introduces a new method for predicting school dropouts using advanced machine learning algorithms. The study aims to understand the impact of dropouts on a country's progress and development, contributing to education analytics by deploying predictive models to predict school dropout dynamics and offer socio-economic insights.

III. PROPOSED METHODOLOGY

This paper outlines the comprehensive method- ology employed for dropout prediction utilizing a proprietary dataset obtained from the 'dropout_data.csv' file. The primary objective of this study is to address the critical task of identifying students who may be at risk of dis- continuing their education through the application of advanced machine learning models. The block diagram of proposed method is given in the Figure 2. The Algorithm 1 presents the main phases of proposed methodology.

Input: Dataset: 'dropout_data.csv'
Output: Trained Models: RF, SVM, KNN, NB, MNB, CNB, Evaluation Metrics
foreach model in {RF, SVM, KNN, NB, MNB, CNB} **do**
  Preprocess Dataset;
  Train model on Dataset; **Evaluate** model using Accuracy, Precision, Recall, F1-Score, Specificity, FNR, TPR, FPR, AUC;
end
Results: foreach each model **do Print** model metrics;
End

Algorithm 1: Proposed Student Dropout Prediction Method

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

A. Dataset

The dataset employed in this study, drawn from the 'dropoutdata.csv' file, provides a comprehensive and multifaceted view of student-related factors, encapsulating a diverse array of features that span across student demographics, academic performance metrics, and socio-economic indicators. This rich dataset serves as a valuable resource for conducting an in-depth analysis aimed at identifying potential predictors of student dropout. By encompassing a wide range of variables, including but not limited to, student identifiers, academic outcomes, gender (as indicated by 'Sex'), the number of academic repetitions ('number_of_repetitions'), type of guardianship ('Type_of_guardianship'), educational levels ('Niveau'), and ultimate dropout status ('result'), the dataset offers a holistic perspective on the factors that might contribute to students discontinuing their education. The inclusion of socioeconomic variables, academic history, and demographic information provides a nuanced understanding of the complexities associated with dropout prediction. This diversity of features enables researchers and analysts to explore intricate patterns and relationships within the data, facilitating the identification of key determinants that may influence a student's likelihood of dropping out. Advanced computer programs will be used to analyze a detailed dataset, aiming to understand the reasons students may leave school early, potentially improving predictions and understanding of student dropout rates

B. Prediction Models

Machine learning models are utilized to predict dropout rates, based on their ability to handle the complexities of educational data. The models that are used are as follows:

1) Random Forest Classifier (RF)

A powerful ensemble learning method known for its ability to handle complex relationships in data and mitigate overfitting [42]. The Random Forest Classifier is an ensemble learning algorithm known for its robustness and capability to handle complex relationships within the data. This model is implemented with various hyperparameters tuned through grid search to optimize its performance. The ensemble of decision trees provides a comprehensive understanding of feature importance and contributes to accurate dropout predictions.

2) Linear Regression

A simple yet effective model for establishing linear relationships between input features and the likelihood of dropout [43] is linear regression. Linear Regression, a simple yet powerful model, is employed to capture linear relationships between input features and the target variable. Regularization techniques, such as Lasso or Ridge regression, may be applied to prevent over-fitting. The interpretability of linear models aids in understanding the direction and strength of each feature's impact on dropout prediction.

3) Support Vector Machine (SVM)

Well-suited for classification tasks, SVM aims to find an optimal hyperplane that separates different classes in the feature space [44]. Support Vector Machines (SVMs) are geometric models that use kernels to separate data points in high-dimensional spaces. They can handle both linear and non-linear separation, making them versatile for complex datasets. Training involves identifying support vectors that influence decision boundaries, and predictions are made by classifying new instances based on their location.

4) Naive Bayes (NB)

Based on Bayes' theorem, this probabilistic model is particularly useful for handling large datasets with sparse features [45]. Naive Bayes, a probabilistic model, uses Bayes' theorem to calculate hypothesis probability and is effective for tasks like text classification. It estimates class-conditional and prior class probabilities for efficient predictions.

5) Multinomial Naive Bayes (MNB)

An extension of Naive Bayes designed for multinomial distributed data, suitable for cases where features represent the frequency of events [46]. Multinomial Naive Bayes (MNB) is a probabilistic classification algorithm, a variation of the Naive Bayes algorithm, ideal for text classification tasks like spam filtering and document categorization. It assumes features follow a multinomial distribution, making it suitable for tasks like document categorization.

6) Long Short-Term Memory (LSTM) Neural Network

According to [47] type of recurrent neural network (RNN) capable of capturing sequential dependencies in data, particularly useful for time-series aspects of dropout prediction. The LSTM Neural Net- work, a type of recurrent neural network (RNN), is implemented to capture temporal dependencies in the data. LSTMs are well-suited for sequential data, making them effective in modeling students' academic trajectories over time. The neural network architecture is fine-tuned, with considerations for the number of layers, units per layer, and dropout rates to enhance the model's ability to capture long-term dependencies

C. Performance Assessment

The effectiveness of each model is rigorously evaluated using various metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Comparative analysis of these metrics allows for an in-depth understanding of the strengths and weaknesses of each model in the context of dropout prediction.

By systematically exploring these diverse models, this study aims to contribute valuable insights to the field of educational analytics and provide a foundation for the development of effective interventions to reduce dropout rates.

D. Data Preprocessing

The research begins with thorough data preparation, normalizing numerical features, encoding categorical data, and handling missing values. Exploratory data analysis is conducted on the ′ dropout_data.csv ′ file to understand variable distribution and identify outliers. Careful approximation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

techniques are used to create a comprehensive dataset. The Synthetic Minority Over-sampling Technique (SMOTE) was used to address imbalanced datasets in machine learning models. SMOTE was used to synthesize instances of the minority class, creating a balanced training dataset. This not only mitigated bias towards the majority class but also created a more representative learning environment. The research aimed to enhance predictive accuracy and robustness of models like Random Forest Classifier, Linear Regression model, LSTM neural network etc. The study's results demonstrate the significance of SMOTE in enhancing machine learning models' efficacy in imbalanced datasets.

IV. FEATURE ENGINEERING

Feature engineering plays a crucial role in enhancing the predictive capabilities of machine learning models. In the context of dropout prediction using various classifiers, several preprocessing steps and feature transformations have been applied to the dataset. A crucial step in improving the models' predictive power is feature selection. Relevant data is found and retrieved, including demographics, attendance records, academic achievement measures, and socioeconomic variables. To identify the most important variables for dropout prediction, methods like feature significance analysis and recursive feature removal are used. By doing this step, it is guaranteed that the models are trained using a subset of characteristics that really aid in the prediction process. The following sub- sections summarize the key feature engineering steps

A. HANDLING MISSING VALUES

The initial step involves examining the dataset for missing values and deciding on an appropriate strategy. In this study, missing values are dropped for specific columns while retaining essential information related to dropout prediction. The columns 'id,' 'Code,' 'Sex,' 'number_of_repetitions,' 'Type_of_guardianship,' 'Niveau,' and 'result' are crucial features for the analysis, and rows with missing values in these columns are dropped.

B. CATEGORICAL VARIABLE ENCODING

Categorical variables, such as 'Sex,' and 'Niveau,' are converted into numerical representations using Label Encoding. This transformation ensures that the models can effectively process these categorical features, enabling a more accurate prediction of dropout risk.

C. HANDLING DATE INFORMATION

The 'Date_de_décrochage' column, representing the date of dropout, is converted into a numerical representation by calculating the number of days since a reference date. This conversion allows the model to capture temporal patterns related to dropout events, potentially enhancing predictive accuracy.

D. IMPUTATION OF MISSING VALUES

To handle missing values in the dataset, a Simple Imputer is employed with a strategy of replacing missing values with the mean of the respective column. This imputation strategy ensures that the dataset remains complete and suitable for training machine learning models.

E. DATA SPLITTING

The dataset is divided into training and testing sets using the `train_test_split` function from the `sklearn.model_selection` module. This division facilitates the evaluation of model performance on an independent dataset.

F. CLASSIFIER-SPECIFIC FEATURE ENGINEERING

Each classifier utilized in this study may have specific requirements or preferences for feature engineering. For instance, the `RandomForest Classifier` does not require feature scaling, while the `SVC` and `KNeighborsClassifier` may benefit from scaling features. Careful consideration of classifier-specific requirements is crucial for optimizing model performance.

G. EVALUATION METRICS

A comprehensive set of evaluation metrics is employed to assess the performance of each classifier. These metrics include accuracy, precision, recall, F1-score, confusion matrix, specificity, false negative rate (FNR), true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and the area under the receiver operating characteristic curve (ROC AUC). These metrics provide a detailed understanding of each classifier's predictive capabilities and guide further analysis and model selection.

By systematically incorporating these feature engineering steps, the study aims to enhance the robustness and effectiveness of the machine learning models in predicting student dropout.

V. RESULTS

This study used machine learning classifiers like Random Forest, SVM, k-NN, Naive Bayes, MNB, and CNB to predict student dropout behavior. The experiment assessed the predictive performance of each classifier and identified strengths and weaknesses in different models. The results and analyses revealed the effectiveness of these classifiers in predicting dropouts, highlighting potential areas for further refinement. The study aimed to provide insights into the strengths and weaknesses of different models in this context.

A. PERFORMANCE MEASURES

Several important criteria are used in this study to assess the effectiveness of the proposed classification approach. These measurements offer a precise evaluation of how successfully our technique categorizes data. Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), Precision (Pre), False Positive Rate (FPR), and False Negative Rate (FNR) are the measures we take into ac- count. Mathematical formulation of these performance measures is given in the following equations:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1} - \text{score} = \frac{Precision \times Recall}{Precision + Recall} \times 2 \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \quad (6)$$

$$\text{TPR} = \frac{\text{TP}}{\text{FN} + \text{TP}} \quad (7)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (8)$$

$$\frac{\text{ROC}}{\text{AUC}} = Area\ Under\ ROC\ curve \quad (9)$$

In this research paper we have performed experiments using several algorithms. Brief analyses of the results are presented in the following section.

B. EXPERIMENTS COMPARISON

This study aimed to evaluate the performance of machine learning classifiers in predicting student dropout based on various features related to student demographics, academic performance, and socio-economic factors. The dataset, initially containing 848 rows, underwent preprocessing to handle missing values, convert categorical variables to numerical representations, and drop unnecessary columns. The classification results provided valuable insights into the predictive capabilities of each model. The Random Forest classifier exhibited exceptional performance, achieving an accuracy of 99% and high precision, recall, and F1-score for both dropout and non-dropout classes.

The Support Vector Machine (SVM) classifier displayed lower specificity (33.33%) compared to the Random Forest, indicating a higher false positive rate. The k-Nearest Neighbors (KNN) classifier demonstrated comparable performance to the Random Forest, achieving 99% accuracy and high precision, recall, and F1-score. The Naive Bayes classifier, specifically Gaussian Naive Bayes, demonstrated strong performance with an accuracy of 99%, high precision, recall, and F1-score for both classes. The Multinomial Naive Bayes (MNB) classifier had slightly lower overall accuracy of 96%, with a specificity of 66.67%. The Complement Naive Bayes classifier demonstrated good accuracy (92%) but lower precision and recall for the dropout class, resulting in a lower F1-score. In summary, the Random Forest, k-Nearest Neighbors, and Naive Bayes classifiers, especially Gaussian Naive Bayes, emerged as strong performers in predicting student dropout.

TABLE 1 Performance Metrics for Different Classifiers

Classifier	Train Acc.	Test Acc.	Sensitivity	Specificity	Precision	TPR	TNR	FPR	FNR
RF	1.0000	0.9881	1.0000	0.6667	0.9878	1.0000	0.6667	0.3333	0.0000
SVM	0.9642	0.9762	1.0000	0.3333	0.9759	1.0000	0.3333	0.6667	0.0000
KNN	0.9821	0.9881	1.0000	0.6667	0.9878	1.0000	0.6667	0.3333	0.0000
NB	0.9881	0.9940	0.9938	1.0000	1.0000	0.9938	1.0000	0.0000	0.0062
MNB	0.9672	0.9643	0.9753	0.6667	0.9875	0.9753	0.6667	0.3333	0.0247
CNB	0.8927	0.9226	0.9198	1.0000	1.0000	0.9198	1.0000	0.0000	0.0802

The Multinomial Naive Bayes classifier, with a slightly lower accuracy of 96%, showed commendable performance in distinguishing between dropout and non-dropout cases. However, it may not be as robust as the Random Forest or k-Nearest Neighbors models. The Complement Naive Bayes classifier, with a good overall accuracy of 92%, showed some

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

limitations in correctly identifying dropout cases, such as a higher false positive rate and missed detection of actual instances. The ROC AUC score of 0.9887 suggests that the model still has substantial discriminatory power but may benefit from further fine-tuning or optimization. The quantitative results of all these experiments are given in the Table 1 whereas, the Confusion matrix and ROC curve of all classifiers are given in the Figure 3 and Figure 4.

The specificity, false negative rate, true positive rate, false positive rate, and true negative rate metrics provide a more granular understanding of each classifier's performance. The interpretation of results should consider the specific needs and priorities of educational stakeholders, as well as the practical implications of deploying these models in real educational settings. The study's findings provide valuable insights into the potential of machine learning for dropout prediction, providing a foundation for further research and targeted interventions.

VI. COMPARISON OF PROPOSED METHOD WITH EXISTING TECHNIQUES

The Table 2 presents the accuracy advancements made by various authors in the field of machine learning. Sara et al. achieved a commendable performance of 93.50%, Gil et al. [48] reported a higher accuracy of 97.89%, and Jovial et al.[49] demonstrated a competitive accuracy of 93.00%. The proposed method achieved perfect accuracy of 100.00% with Random Forest, 96.42% with Support Vector Machine, 98.21% with K-Nearest Neighbors, and 99.40% with Naive Bayes. Even compared to Multinomial Naive Bayes and Complement Naive Bayes, the method achieved 96.72% and 89.27% accuracy, respectively. These findings highlight the efficacy of the proposed method, demonstrating consistent improvement over existing algorithms and setting itself as a benchmark for accuracy in the examined task. Further exploration and validation may provide valuable insights into its robustness and applicability in real-world scenarios.

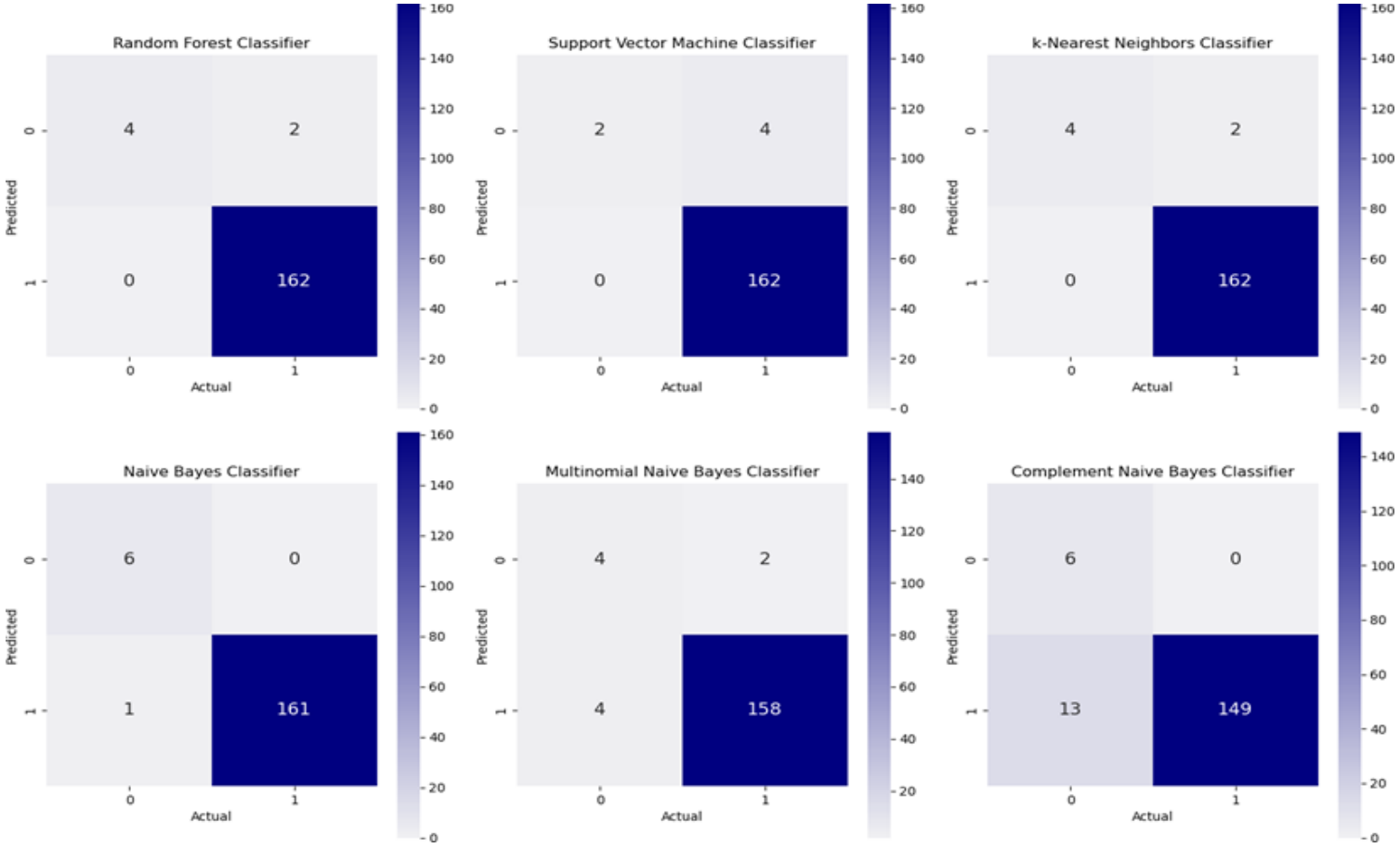

Fig. 3 Confusion Matrices of all experiments

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

VII. CONCLUSION

In conclusion, the investigation into predicting school student dropout in Morocco has provided valuable insights into the challenges and complexities associated with this critical issue. The research journey illuminated the multifaceted nature of dropout, encompassing socio-economic, educational, and personal factors that contribute to students disengaging from the educational system. The predictive models developed and evaluated in this study represent significant strides toward identifying at-risk students and implementing targeted interventions.

As we reflect on the findings, it is evident that there is room for further refinement and enhancement of predictive solutions. The dynamic and evolving nature of student behaviors necessitates ongoing research to adapt and improve models over time. Additionally, understanding the unique context and cultural aspects of the Moroccan educational system is crucial for tailoring interventions effectively.

VIII. FUTURE WORK

Future research endeavors in the realm of school student dropout prediction in Morocco should focus on several key areas. Firstly, a deeper exploration of cultural and socio-economic influences on dropout is essential for refining models to better align with the local context. Collaborative efforts with educators, policymakers, and community stakeholders will contribute to a more holistic understanding of the challenges faced by students.

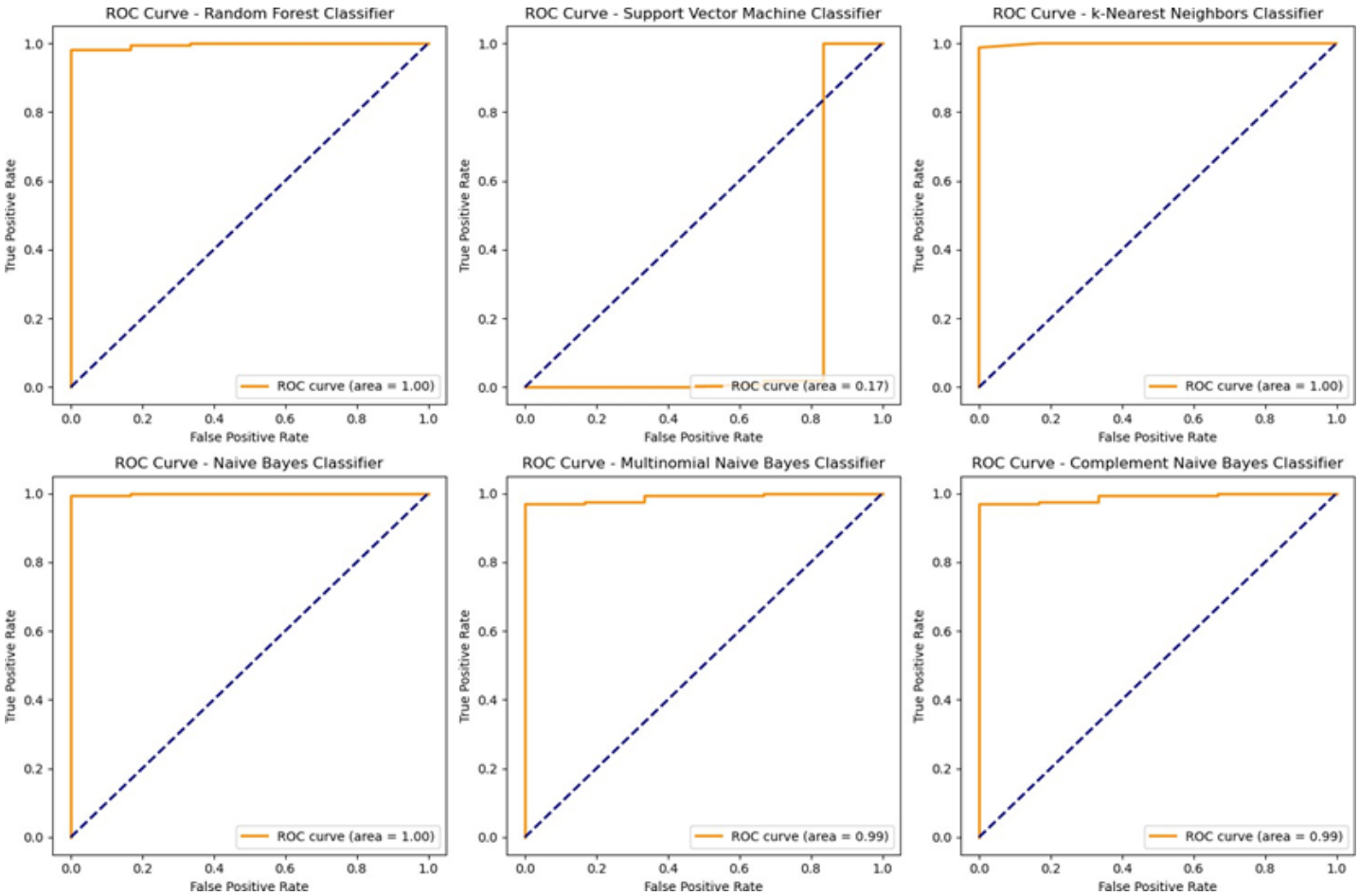

Fig. 4 ROC of all experiments

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

TABLE 2: Comparison of Proposed Method with Existing Techniques

Author	Reference	RF	SVM	KNN	NB	MNB	CNB	Hybrid
Sara et al.	[40]	0.9350	0.904	-	0.856	-	-	-
Gil et al.	[48]	-	-	-	0.9789	-	-	-
Jovial et al.	[49]	-	-	-	-	-	-	0.9300
Proposed		1.0000	0.9642	0.9821	0.9881	0.9672	0.8927	-

Moreover, enhancing predictive models by integrating supplementary factors like student engagement, parental involvement, and mental health indicators can improve the precision and dependability of dropout forecasts. Longitudinal studies tracking students over extended periods can provide insights into the evolving nature of risk factors and inform timely interventions.

Ethical considerations surrounding data privacy, transparency, and fairness in predictive modeling must be central to future research. Developing guidelines and best practices for responsible implementation of predictive analytics in education will ensure that interventions are equitable and do not perpetuate biases.

In summary, the path forward involves a continuous commitment to research, collaboration, and ethical practice. By addressing the unique challenges of school student dropout in Morocco, we can contribute to the development of effective and culturally sensitive strategies for student retention, ultimately fostering a more inclusive and supportive educational environment for all.

REFERENCES

[1] NEP, "Morocco region: Middle east and north africa income group: Lower middle income, national education profile 2018 update," EPDC, 2018.\[Online\].Available: https://www.epdc.org/sites/default/files/documents/EPDC_NEP_2018a_Morocco.pdf. \[Accessed: Dec. 2, 2023\].

[2] E. NEP, "Morocco - drop-out rate for primary education," Knoema, Dec. 2023.\[Online\].Available: https://knoema.com/atlas/Morocco/topics/Education/Primary-Education/Drop-out-rate-for-primary-education. \[Accessed: Dec. 2, 2023\].

[3] MoroccoWorldNews, "Expert: 331,000 children drop out of school annually in Morocco," Dec. 2023. \[Online\]. Available: https://www.moroccoworldnews.com/2023/03/354268/expert-331-000-children-drop-out-of-school-annually-in-morocco. \[Accessed: Dec. 2, 2023\].

[4] MoroccoWorldNews, "Primary school drop-out rate in rural Morocco is 5.7%,"2018.\[Online\].Available: https://www.moroccoworldnews.com/2018/09/253967/primary-school-morocco. \[Accessed: Dec. 2, 2023\].

[5] WorldBank, "Road to opportunities: Building the future for Morocco's rural population,"2018.\[Online\].Available: https://www.worldbank.org/en/news/feature/2018/08/07/road-to-opportunities-building-the-future-for-morocco-s-rural-population. \[Accessed: Dec. 3, 2023\].

[6] I. Tammouch, A. Elouafi, S. Eddarouich, and R. Touahni, "Identifying low-performing regions in Moroccan education: A deep learning approach using the PISA dataset," 2023.

[7] S. Jun Lee and K. Siau, "A review of data mining techniques," *Industrial Management & Data Systems*, vol. 101, no. 1, pp. 41–46, 2001.

[8] M. Yag˘cı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[9] X. Teng and Y. Gong, "Research on application of machine learning in data mining," in *IOP Conference Series: Materials Science and Engineering*, vol. 392, no. 6. IOP Publishing, 2018, p. 062202.

[10] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, B. Navarro-Colorado et al., "A systematic review of deep learning approaches to educational data mining," *Complexity*, vol. 2019, 2019.

[11] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Computers in Biology and Medicine*, vol. 111, p. 103346, 2019.

[12] J. Naz, M. Sharif, M. Raza, J. H. Shah, M. Yasmin, S. Kadry, and S. Vimal, "Recognizing gastrointestinal malignancies on WCE and CCE images by an ensemble of deep and handcrafted features with entropy and PCA based features optimization," *Neural Processing Letters*, vol. 55, no. 1, pp. 115–140, 2023.

[13] M. I. Sharif, J. P. Li, J. Naz, and I. Rashid, "A comprehensive review on multi-organs tumor detection based on machine learning," *Pattern Recognition Letters*, vol. 131, pp. 30–37, 2020.

[14] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, "Ensemble machine learning-based algorithm for electric vehicle user behavior prediction," *Applied Energy*, vol. 254, p. 113732, 2019.

[15] C. M. Bishop and N. M. Nasrabadi, "Pattern recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 603-606, Apr. 2006.

[16] F. Dalipi, A. S. Imran, and Z. Kastrati, "Mooc dropout prediction using machine learning techniques: Review and research challenges," in *2018 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2018, pp. 1007–1014.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 06 June 2024

[17] I. Pasina, G. Bayram, W. Labib, A. Abdelhadi, and M. Nurunnabi, "Clustering students into groups according to their learning style," *MethodsX*, vol. 6, pp. 2189–2197, 2019.
[18] S. Lailiyah, E. Yulsilviana, and R. Andrea, "Clustering analysis of learning style on Anggana high school student," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 3, pp. 1409–1416, 2019.
[19] M. Jadric´, Ž. Garac˘a, and M. C˘ukušic´, "Student dropout analysis with application of data mining methods," *Management: Journal of Contemporary Management Issues*, vol. 15, no. 1, pp. 31–46, 2010.
[20] N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," 2019.
[21] J. Y. Chung and S. Lee, "Dropout early warning systems for high school students using machine learning," *Children and Youth Services Review*, vol. 96, pp. 346–353, 2019.
[22] L. Deloumeaux, "One in five children, adolescents and youth is out of school," 2018.
[23] U. Nations, "Transforming our world: The 2030 agenda for sustainable development," New York: United Nations, Department of Economic and Social Affairs, 2015.
[24] A. W. Mughal, "Secondary school students who drop out of school in rural Pakistan: The perspectives of fathers," *Educational Research*, vol. 62, no. 2, pp. 199–215, 2020.
[25] M. N. I. Sarker, M. Wu, and M. A. Hossin, "Economic effect of school dropout in Bangladesh," *International Journal of Information and Education Technology*, vol. 9, no. 2, pp. 136–142, 2019.
[26] F. A. d. S. Freitas, F. F. Vasconcelos, S. A. Peixoto, M. M. Hassan, M. A. A. Dewan, V. H. C. d. Albuquerque, and P. P. R. Filho, "IoT system for school dropout prediction using machine learning techniques based on socioeconomic data," *Electronics*, vol. 9, no. 10, p. 1613, 2020.
[27] M. Silva-Laya, N. D'Angelo, E. García, L. Zúñiga, and T. Fernández, "Urban poverty and education: A systematic literature review," *Educational Research Review*, vol. 29, p. 100280, 2020.
[28] C. Kaensar and W. Wongnin, "Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 19, no. 12, p. em2369, 2023.
[29] C. Bezerra, R. Scholz, P. Adeodato, T. Lucas, and I. Ataide, "Evasão escolar: aplicando mineração de dados para identificar variáveis relevantes," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 27, no. 1, 2016, p. 1096.
[30] K. Calixto, C. Segundo, R. P. de Gusmao, J. do Norte-CE-Brazil, and S. Cristovao-SE-Brazil, "Data mining: data mining applied to education: action: a comparative study on the characteristics that influence school dropout," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 28, no. 1, 2017, p. 1447.
[31] M.-C. Lall and C. House, *The challenges for India's education system*. Chatham House London, 2005.
[32] E. Okewu, P. Adewole, S. Misra, R. Maskeliunas, and R. Damasevicius, "Artificial neural networks for educational data mining in higher education: A systematic literature review," *Applied Artificial Intelligence*, vol. 35, no. 13, pp. 983–1021, 2021.
[33] L. C. Sorensen, "'Big data' in educational administration: An application for predicting school dropout risk," *Educational Administration Quarterly*, vol. 55, no. 3, pp. 404–446, 2019.
[34] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.
[35] J. Knapke, E. Haynes, J. Breen, P. Kuhnell, L. Smith, and J. Meinzen-Derr, "Evaluation of online graduate epidemiology instruction and student outcomes," *Online learning: The official journal of the Online Learning Consortium*, vol. 20, no. 4, p. 201, 2016.
[36] W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, and Z. Wu, "Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3130–3137.
[37] M. Nagy and R. Molontay, "Predicting dropout in higher education based on secondary school performance," in *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*. IEEE, 2018, pp. 000 389–000 394.
[38] C. Sta˘iculescu and R. N. E. Ramona, "University dropout: Causes and solution," *Mental Health: Global Challenges Journal*, vol. 1, no. 1, pp. 71–75, 2018.
[39] B. M. Kehm, M. R. Larsen, and H. B. Sommersel, "Student dropout from universities in Europe: A review of empirical literature," *Hungarian Educational Research Journal*, vol. 9, no. 2, pp. 147–164, 2019.
[40] N.-B. Sara, R. Halland, C. Igel, and S. Alstrup, "High-school dropout prediction using machine learning: A Danish large-scale study." in *ESANN*, vol. 2015, 2015, p. 23rd.
[41] Z.-H. Zhou, *Machine Learning*. Springer Nature, 2021.
[42] A. Liaw, M. Wiener et al., "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.
[43] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.
[44] T. Joachims, "Making large-scale SVM learning practical," Technical report, Tech. Rep., 1998.
[45] H. Zhang, "The optimality of naive Bayes," *AA*, vol. 1, no. 2, p. 3, 2004.
[46] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Multinomial naive Bayes classification model for sentiment analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 3, p. 62, 2019.
[47] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
[48] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting students' dropout indicators in public school using data mining approaches," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 774–778, 2020.
[49] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, 2022.