Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques

Abdellatif HARIF, Moulay Abdellah KASSIMI
Laboratory of Science of Information Technology Data, Mathematics and Applications-
National School of Applied Sciences, IBN ZOHR University, Agadir, Morocco

***Abstract*—Predicting student performance has become a strategic challenge for universities, essential for increasing student success rates, retention, and tackling dropout rates. However, the large volume of educational data complicates this task. Therefore, many research projects have focused on using Machine Learning techniques to predict student success. This study aims to propose a performance prediction model for students at IBN ZOHR University in Morocco. We employ a combination of Random Forest and Recursive Feature Elimination with Cross-Validation (RFECV-RF) for optimal feature selection. Using these features, we build classification models with several Machine Learning algorithms, including AdaBoost, Logistic Regression (LR), k-Nearest Neighbors (k-NN), Naive Bayes (NB), Support Vector Machines (SVM), and Decision Trees (DT). Our results show that the SVM model, using the 8 features selected by RFECV-RF, outperforms the other classifiers with an accuracy of 87%. This demonstrates the effectiveness and efficiency of our feature selection method and the superiority of the SVM model in predicting student performance.**

Keywords*—Student performance prediction; Recursive Feature Elimination (RFE); cross-validation; Random Forest (RF); feature selection; IBN ZOHR University

I. Introduction

Education is the foundation of human development, providing individuals with the knowledge and skills necessary to navigate the world and achieve their goals. As digitization accelerates and data volumes increase in educational environments, it becomes crucial to understand how these tools can be used to measure and promote student well-being while supporting personalized learning experiences. Educational institutions have begun to explore the potential of big data technologies and Educational Data Mining (EDM) to more effectively support learning and education [1] [2].

In recent years, the use of Data Mining and Machine Learning in educational settings has significantly evolved [3]. These techniques enable understanding student behaviors and implementing targeted interventions. In this context, performance prediction is particularly relevant in Moroccan universities, where higher education institutions collect a multitude of data on their students. This information includes both qualitative and quantitative variables such as academic performance and socio-economic [4]. Experimental processes include data collection and preprocessing, the application of prediction models, as well as the evaluation of results. Additional techniques such as feature selection and cross-validation are also employed to enhance the quality of predictions [5].

This research employs an empirical framework to evaluate the accuracy and efficiency of different machine learning models in forecasting academic outcomes. By implementing a comprehensive framework, and finding the optimal features required, the main objective of this study is to develop a robust performance prediction model for students at the University of IBN ZOHR in Morocco. This will enable the accurate identification of final grades and provide insights that can guide educational interventions, ultimately enhancing the educational outcomes for students.

The remainder of this article is organized as follows: Section II is dedicated to related works, Section III presents our methodology, Section IV exposes the experimental results, Section V discusses the obtained results and analyzes the implications of our study for educational practice, and Section VI concludes by presenting perspectives for future work.

II. Related Works

Predicting student performance has become a critical area of research in educational data mining and learning analytics. The ability to accurately forecast academic outcomes not only aids in identifying students at risk of failing but also helps in tailoring educational interventions to enhance student success. Several studies have explored various techniques for predicting student performance using data mining techniques and machine-learning algorithms. In this section, we analyze the literature from 2009 to 2023 focusing on articles that demonstrate the effectiveness of Machine Learning methods in predicting student performance. It emphasizes the importance of feature selection, model optimization, and the incorporation of diverse data types, including demographic and behavioral information, to enhance the accuracy and reliability of predictive models in educational settings [6].

For instance, Asselman et al. [7] proposed a new approach that combines models such as Random Forest, AdaBoost, and XGBoost. Their experiments on three different datasets demonstrate that the XGBoost model significantly outperforms other models and the original Performance Factors Analysis (PFA) algorithm. The study concludes that ensemble learning

methods, particularly XGBoost, enhance prediction accuracy in educational settings.

Similarly, Ajibade et al. [8] introduced behavioral features alongside traditional academic and demographic features as new predictors. Various classifiers, including Naïve Bayes, Decision Tree, K-Nearest Neighbor, Discriminant Analysis, and Pairwise Coupling, were used. The study found that incorporating behavioral features improved prediction accuracy from 72.6% to 84.2%. Furthermore, applying ensemble methods like AdaBoost, Bagging, and RUSBoost enhanced accuracy to 94.1%, demonstrating the effectiveness of these techniques in predicting academic performance.

Building on this, Shahiri et al. [9] aimed to identify gaps in current prediction methods, determine key attributes influencing student performance, and evaluate various predictive algorithms. Important attributes highlighted include cumulative grade point average (CGPA) and internal assessments.

Helal et al. [10] focused on predicting academic performance by considering student heterogeneity. Using data from an Australian university, the research shows that models trained on specific student sub-populations outperform those trained on the entire dataset. The study combines enrolment and LMS activity data, finding that this improves the precision of identifying at-risk students. Both black-box and white-box classification methods were used, with white-box methods being particularly useful for designing effective student support strategies.

Widyahastuti and Tjhin [11] aimed to predict students' performance in final examinations using linear regression and multilayer perceptron. Data was collected from e-learning logs and attendance records of 50 undergraduate students. The research concluded that the multilayer perceptron model provides better prediction results compared to linear regression in terms of accuracy, performance, and error rate. The findings highlight the importance of using neural network models for more accurate predictions in the educational context.

Furthermore, Yang et al. [12] investigated predicting student academic performance using Multiple Linear Regression (MLR) and Principal Component Analysis (PCA). Data was collected from 58 university students enrolled in a blended calculus course. The study found that combining MLR with PCA improves the predictive accuracy of the model. Traditional evaluation measures like MSE and R^2 were supplemented with new measures like pMSE and pMAPC to better assess predictive performance. The results indicated that using PCA components significantly enhances the model's accuracy.

El Aissaoui et al. [13] proposed a multiple linear regression-based approach to predict student outcomes, utilizing multivariate adaptive regression splines to select the most relevant variables, thereby improving the model's performance. Their methodology demonstrated that variables selected through Multivariate Adaptive Regression Splines (MARS) led to more accurate predictive models compared to other variable selection methods.

The approach used by Alshanqiti and Namoun [14] combined collaborative filtering, fuzzy set rules, and Lasso linear regression to optimize prediction accuracy. It also utilizes an optimized self-organizing map for multi-label classification to identify various factors affecting student performance. The method was tested on seven datasets, demonstrating significant improvements over baseline models, highlighting the importance of combining supervised and unsupervised learning for accurate predictions and explanatory insights into student performance.

The study by Turabieh et al. [15], proposed an enhanced version of the Harris Hawks Optimization (HHO) algorithm to improve feature selection for predicting student performance. By controlling population diversity using k-nearest neighbors (kNN) clustering, the modified HHO algorithm aims to overcome premature convergence and prevent trapping in local optima. The study employs various machine learning classifiers, such as kNN, Layered Recurrent Neural Network (LRNN), Naïve Bayes, and Artificial Neural Network, to evaluate the prediction system using a dataset from the UCI machine learning repository. Results indicate that the combination of the enhanced HHO and LRNN achieves the highest accuracy of 92%, highlighting the importance of early prediction to avoid student failure and improve educational outcomes.

Shivaji et al. [16] proposed a feature selection technique to enhance the performance of machine learning classifiers in predicting bugs in software changes. By applying the Gain Ratio for feature selection, the study aims to reduce the number of features, thereby improving classifier accuracy and scalability. The technique was evaluated using Naïve Bayes and Support Vector Machine (SVM) classifiers across multiple open-source projects. Results indicate that feature selection significantly improves bug prediction performance, achieving high precision and reducing false positives.

Another study by Zaffar et al. [17] investigated the effectiveness of various feature selection algorithms in predicting student academic performance. The research evaluates six filter-based feature selection algorithms (CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttribute Eval, GainRatioAttributeEval, Principal Components, and ReliefAttributeEval) using two different datasets with varying numbers of features. The study finds that there is a significant performance difference based on the number of features, with a 10-20% accuracy variation.

Adejo and Connolly [18] investigated and compared the efficiency of multiple data sources, different classifiers, and ensemble techniques in predicting student academic performance. Using data from the University of the West of Scotland, the study employs Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM) classifiers, as well as their ensembles. Results indicate that combining multiple data sources with heterogeneous ensemble techniques significantly improves prediction accuracy and helps identify at-risk students early. The proposed hybrid model, which integrates various classifiers and data sources, achieves higher accuracy and efficiency compared to individual base classifiers.

Imran et al. [19] proposed a model to predict student performance using supervised learning algorithms. The research addresses common issues such as data high dimensionality, class imbalance, and classification errors. Using data from the UCI Machine Learning Repository, the study evaluates three classifiers: J48, NNge, and MLP, with J48 achieving the highest accuracy of 95.78%. The study demonstrates the importance of data preprocessing and the use of ensemble methods to improve prediction accuracy. This model is designed to help educational institutions make early interventions to support at-risk students.

Similarly, Razaque and Alajlan [20] evaluated six machine learning models (Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, AdaBoost, and Stochastic Gradient Descent) to predict student performance. The dataset includes academic and demographic data from the UCI Machine Learning Repository. The models are assessed based on accuracy, precision, sensitivity, and F-measure. The results indicate that Stochastic Gradient Descent outperforms other models, achieving the highest accuracy. The study underscores the importance of preprocessing and proper model selection to enhance prediction accuracy, aiming to identify at-risk students early and support their academic success.

The study by Ghorbani and Ghousi [21] and Alija et al. [22], explored the application of various supervised machine learning algorithms for predicting student performance, with a particular emphasis on managing imbalanced datasets. The authors utilize the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset. Multiple algorithms are evaluated in the research. These findings underscore the necessity of addressing imbalanced data to enhance the accuracy and reliability of predictive models in educational data mining.

H. Alamri et al. [23] explored the use of SVM and Random Forest algorithms to predict academic performance based on various influencing factors such as prior grades and social conditions. The study utilized two types of datasets focused on mathematics and Portuguese language courses, applying both binary classification and regression techniques. The results show that SVM and RF models achieve high accuracy levels, with RF performing slightly better in binary classification scenarios.

Moreover, the application of Recursive Feature Elimination (RFE) has shown considerable promise in various domains, particularly in enhancing the accuracy and efficiency of predictive models [24], [25]. In the context of student performance prediction, several studies have leveraged RFE to identify the most critical features influencing academic outcomes. Syed Mustapha [26] employed RFE with Random Forest to refine feature selection for predicting student grades. This approach evaluated the effectiveness of different models such as the Boruta algorithm and Lasso regression for regression tasks, and Recursive Feature Elimination (RFE) and Random Forest Importance (RFI) for classification tasks. Key findings included the superior performance of Gradient Boost in regression tasks and the effectiveness of Random Forest in classification tasks. The study emphasized the importance of proper feature selection to improve the accuracy and efficacy of predictive models.

III. Our Methodology

In this section, we detail the methodology adopted to conduct our study on student performance prediction. Our approach, illustrated in Fig. 1, is based on a series of structured processes, including data collection and preprocessing, feature selection, and model construction and evaluation. To build an efficient prediction model, we have integrated a method that optimizes feature selection. This approach aims to identify the most relevant attributes that directly influence the performance of the model.

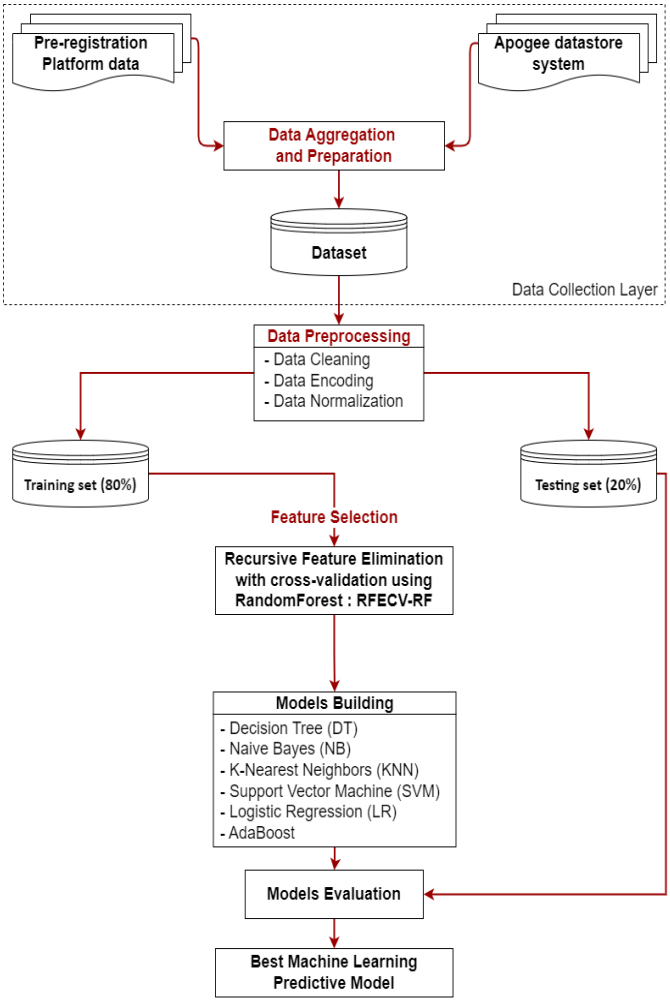

Fig. 1. Architecture of the proposed model.

A. Data Description

The data utilized in this research project belongs to IBN ZOHR University of Morocco and pertains to students enrolled at open-access establishments. It is sourced from two main systems:

- Pre-registration platform: Before enrolling at the faculty, students are required to fill out a form, providing a range of information.

- APOGEE datastore System: This system centralizes academic data, capturing student results throughout their academic journey.

Our dataset comprises 174,135 records and 21 attributes, captured during the period from 2016 to 2020. The following Fig. 2 displays the distribution of data during this period, and Table I lists the considered variables.

B. Data Preprocessing

Data preprocessing is an essential step in machine learning, ensuring the integrity and dependability of the data used for analysis. It involves cleaning, encoding, and normalizing data, reducing biases, and enhancing the precision of predictive models [27]. This section outlines the techniques used in the preprocessing stage, designed to effectively prepare the data for thorough analysis.

1) Data aggregation: The initial stage of our analysis involves aggregating data from the Pre-Registration Platform and the APOGEE Data Storage System. In this process, we merge these data sources utilizing SQL-style joins, which facilitate a precise combination of the data, ensuring thorough synchronization. This technique guarantees the effective integration of each student's information from both sources, minimizing data redundancy and enhancing the overall consistency of the dataset.

Fig. 2. Distribution of students by registration year.

TABLE I. DESCRIPTION OF CONSIDERED STUDENT ATTRIBUTES

Attributes	Values	Description
GENDER	{Female,Male}	Gender
LIV_ENV	{Urban, Rural}	Type of Environment in which the student lives
AGE_ENR	{(20-22) G1,(23-25) G2,(26-30) G3,(31 and above) G4}	Age at the time of Enrollment: G1, G2, G3 refer to Group 1, Group 2, and Group 3, respectively.
DISCIPLINE	{Literary,Scientist,Technical}	Discipline chosen
TEACH_LANG	{Mother tongue,Foreign Language}	Language of Studying at university
FAM_STAT	{Single,Married,Divorced}	Student's Family Status
REG_RES	{Sous Massa Region,Southern Regions,Rest of the Country}	Region of Residence
DISABLED	{No,Yes}	Indicates if the student is Disabled
STD_PRF	{Student,Professional Activity}	Student's Profession
FA_PRF	{Deceased,Unemployed/At Home,Public-Service/Army/police,Retirement,Low-income jobs,Middle-income jobs,Good-income jobs}	Father's Profession
FA_EDU_LIV	{None,Elementary,Intermediate,High}	Father's Education Level
MO_PRF	{Deceased,Unemployed/At Home,Public-Service/Army/police,Retirement,Low-income jobs,Middle-income jobs,Good-income jobs}	Mother's Profession
MO_EDU_LIV	{None,Elementary,Intermediate,High}	Mother's Education Level
PAR_REL	{Married,Divorced}	Parents Relationship
BAC_TYPE	{Literary,Scientist,Technical}	Baccalaureate Type
BAC_DEG	{Pass,Satisfactory,Good,Very Good}	Baccalaureate Degree
BAC_ACAD	{Agadir Sous Massa Region,Southern regions,Rest of the Country,Foreign Academy}	Baccalaureate Academy
MTANGUE_GRD	[3,19]	Grade in Mother Tongue
1F_LANG_GRD	[3,19]	Grade in First Foreign Language
2F_LANG_GRD	[3,19]	Grade in Second Foreign Language
F_GRADE	{<6 (Bad),[6,10[(Poor),[10,13] (Medium), >=13 (Good)}	Target: Final Grade

2) Data cleaning and encoding: During the initial preprocessing stage, crucial measures were taken to ensure data integrity and suitability for further analysis. The dataset initially consisted of 177,193 entries across 38 variables, which, after removing 3,058 redundant records and eliminating irrelevant attributes, was reduced to 174,135 entries and 21 variables. Following this, attention was directed towards data encoding, where, for example, the 'GENDER' attribute was encoded using label encoding, and 'AGE_ENR' as well as 'F_GRADE' were handled using ordinal encoding to facilitate computational processing and maintain the natural order of values. Fig. 3 illustrates the distribution of the target 'F_GRADE' for visualization.

3) Data normalization: Normalization is an essential preprocessing technique in machine learning, ensuring that the range of independent variables is uniform across the dataset. This process helps in achieving faster convergence during training, reduces the complexity of the model, and often leads to better overall performance [28].

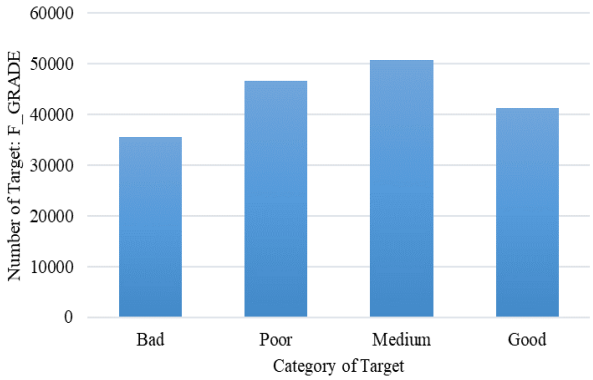

Fig. 3. Distribution of the F_GRADE target.

In our study, which analyzes student performance data with a wide variety of scales and distributions, Z-score normalization is identified as the most appropriate technique. It's a statistical technique that transforms features to have a mean of zero and a standard deviation of one. This normalization is especially useful in machine learning when features vary in scale because it ensures that each feature contributes equally to the analysis and helps in improving the convergence of many algorithms. The formula for the Z-score given by:

$$Z = \frac{(X-\mu)}{\sigma} \quad (1)$$

where X is the original data point, μ is the mean of the data, and σ is the standard deviation.

4) Outliers' analyses: The Z-score is utilized not only for normalization but also for the effective identification of outliers by highlighting data points that substantially diverge from the mean. Once these outliers are detected, several strategies can be employed for their management, such as removal, capping, transformation, or imputation [29]. For our case study, we have chosen to manage outliers through logarithmic transformation. This technique mitigates the influence of outliers by compressing extreme values closer to the median, which helps in reducing skewness and improving the uniformity of the data distribution.

C. Features Selection

Feature selection is a critical step in machine learning, aimed at identifying the most relevant features for use in predictive models. This process is vital as it improves model performance by reducing overfitting, enhancing accuracy, and decreasing training time [30]. In our study, we adopted a systematic approach to optimize feature selection. Specifically, we employed Recursive Feature Elimination combined with Cross-Validation using a Random Forest as an estimator (RFECV-RF).

1) Random Forest (RF): Random Forest is a sophisticated machine learning technique extensively used for classification and regression tasks [31]. As an ensemble learning method, it constructs numerous decision trees during training and combines their outputs to boost classification accuracy and mitigate overfitting. By combining the predictions from multiple trees, RF lowers model variance and enhances generalization capabilities. This robust technique is particularly effective in managing large datasets with intricate feature interactions. Additionally, it provides estimates of feature importance, allowing for feature selection, because it can capture complex relationships and interactions between features, resulting in more reliable and accurate predictive performance. This makes it an excellent choice for predicting student performance [32].

2) Recursive Feature Elimination with Cross-Validation using RF (RFECV-RF): RFE is a wrapper feature selection method that iteratively removes the least important features based on model performance to identify and rank the most significant predictors. To determine the optimal number of features that maximize the performance of machine learning models, we combined RFE with cross-validation [33],[34]. We employed RF as the classification model within the RFECV framework to evaluate and iteratively eliminate features that did not improve classification accuracy. As shown in Algorithm 1, RFECV-RF initializes with the complete feature set S and an empty elimination list R. The algorithm sets a predefined number of features to eliminate in each iteration, known as step_size which is set to 1 in our case, then employs 5-fold cross-validation to robustly evaluate the RF classifier's performance on S. During each iteration, the classifier is trained, and the performance of the feature set is evaluated through cross-validation. Feature importance scores are calculated, and the least significant features, determined by the step_size, are moved from S to R and then removed from S. This process continues until S is empty, ensuring all features are assessed. The refined set S is then re-evaluated with 5-fold cross-validation to validate its effectiveness, and the algorithm outputs R, listing the eliminated features, and S, the curated set of key features for precise prediction.

Algorithm 1: RFECV-RF algorithm for feature selection

Input: Training sample set
1: Initialize the full feature set S = {1, 2, ..., N} where N is the total number of features.
2: Initialize the feature ranking list R = []
3: Determine the set of features to eliminate in each step, termed as step_size.
4: Specify n_folds = 5 for the cross-validation process.
5: **While** len(S) $\neq$ 0 do
6: **For** (each subset of features) do
7: Train the Random Forest classifier on the training data using the features in S.
8: Perform 5-fold cross-validation to estimate the model's performance for each subset of features.
9: Calculate the importance score for each feature in the current feature set S using the Random Forest.
10: Rank the features based on their importance scores.
11: **If** (condition) then
12: Identify the least important features equal to step_size
13: Add these to R and remove them from S.
14: **End If**
15: **End For**
16: If S becomes empty, break the loop.
17: **End While**
18: The final set of features in S is used to perform a final round of training and 5-fold cross-validation.
19: Output the set R as the eliminated features and S as the selected feature subset.

D. Classification of Machine Learning Models and Evaluation Metrics

After performing feature selection to identify the most relevant predictors, we developed classification models to predict student performance using various supervised machine learning techniques, including Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and AdaBoost. Evaluation measures such as accuracy, precision, recall, and F1 score were employed to assess the performance and robustness of each model. These measures provide a comprehensive understanding of the model's effectiveness in accurately identifying patterns, minimizing false positives and negatives, and managing imbalanced classes. Table II details the hyperparameters used for each classification algorithm, while Table III outlines the formulas for each evaluation metric.

TABLE II. HYPERPARAMETERS OF CLASSIFICATION MODELS

Classifier	Hyperparameters
Decision Tree (DT)	Criterion=entropy, max_depth = 10, splitter = best, min_samples_split = 2
Naive Bayes (NB)	Gaussian Naive Bayes doesn't require parameter tuning
K-Nearest Neighbors (KNN)	n_neighbors=5, weights=uniform , metric=minkowski
Support Vector Machine (SVM)	C=1, gamma=scale, kernel=rbf
Logistic Regression (LR)	C=100, penalty=l2, solver= newton-cg
AdaBoost	max_depth=3, n_estimators=200, learning_rate=0.1, algorithm=' SAMME.R'

TABLE III. EVALUATION METRICS AND THEIR DEFINITIONS

Metric	Formula	Description
Accuracy	$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$ (2)	-TP (True Positives): samples correctly classified as positive. -TN (True Negatives): samples correctly classified as negative. -FP (False Positives): samples incorrectly classified as positive. -FN (False Negatives): Instances incorrectly classified as negative.
Precision	$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ (3)	
Recall	$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ (4)	
F1 score	$\text{F1} = 2 * \frac{\text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}}$ (5)	

IV. RESULTS AND EXPERIMENTS

A. Hardware Used

The Experiments were run on a desktop computer using the Ubuntu 20.04 LTS Operating System. The system's technical specifications include 32GB of RAM, an Intel Core i7-12700F processor operating at a clock speed of 2.10 GHz with 12 cores, and an NVIDIA GeForce RTX 3060 graphics card.

B. Experimental Results

Our study aimed to evaluate the importance of feature selection and determine the most effective classifier for predicting the “Final Grade” target. We conducted several experiments using various supervised learning methods on our dataset. Initially, we used RFECV-RF for feature selection, as shown in Table IV. This method helped us identify the optimal subset of features that strongly predict our target. The selected features were then used to train and test several classifiers including DT, NB, KNN, SVM, LR, and AdaBoost to comprehensively evaluate each model's performance.

TABLE IV. USED CONFIGURATION FOR RFE ALGORITHM

Configuration	Value
Model	Random Forest (RF)
Cross-Validation	5-fold
Steps(step_size)	1

Fig. 4 depicts the relationship between the number of features and the classification scores across five cross-validation folds using RFECV-RF. This analysis ranked each feature based on its contribution to enhancing the prediction model's accuracy. The graph illustrated a notable increase in classification scores as the number of features increased from 1 to 6, highlighting the significance of these initial features. The classification score leveled off around eight features, indicating that this number optimally captured the essential information required for effective classification, stabilizing at approximately 87%. The eight highest-ranked features were identified as follows: “BAC_DEG”, “LIV_ENV”, “MTANGUE_GRD”, “FA_PRF”, “FA_EDU_LIV”, “AGE_ENR”, “1F_LANG_GRD”, and “2F_LANG_GRD”. The findings were consistent across different cross-validation folds, demonstrating the robustness and reliability of the approach. Furthermore, beyond 12 features, a slight decrease in the classification score was observed, suggesting that adding more features introduced noise or redundant information.

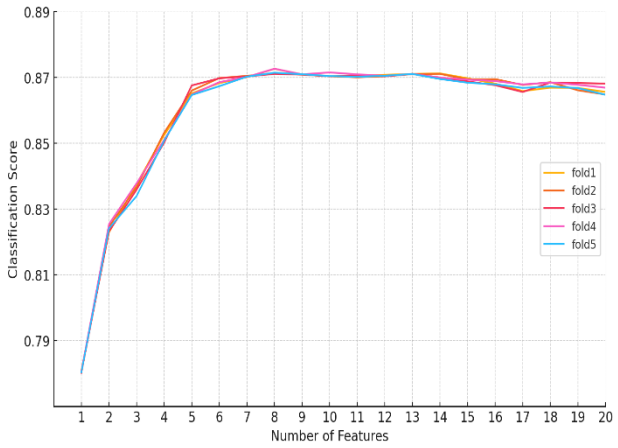

Fig. 4. Classification score vs. Number of selected features using RFECV-RF.

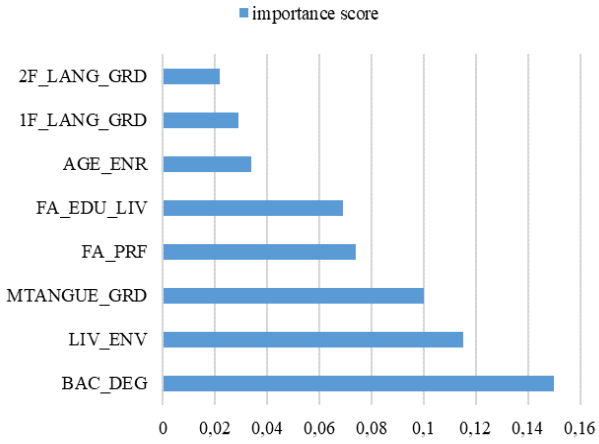

Fig. 5. Feature importance using RFECV-RF algorithm.

Fig. 5, generated through our algorithm, illustrates the relative importance of various features in the predictive model. The "BAC_DEG" feature emerged as the most impactful with an importance score of 0.15, highlighting the significant influence of the baccalaureate degree on our target prediction. The living environment “LIV_ENV” and proficiency in the mother tongue “MTANGUE_GRD” also featured prominently, indicating their critical roles within socio-economic and linguistic contexts. Other attributes such as the father's profession “FA_PRF” and his educational level “FA_EDU_LIV” exhibited considerable significance.

Fig. 6 illustrates the impact of feature selection on the accuracy of various machine learning models by comparing their performance with all features versus the top 8 selected features. The DT model shows an improvement in accuracy from 81% to 83%, indicating that feature reduction can help mitigate overfitting while maintaining the model's ability to make accurate predictions. The NB model, which has the lowest accuracy among the models, sees a marginal increase from 69% to 70%, suggesting that while feature selection provides some benefit, the model might still not be optimal for our dataset. The accuracy of the KNN model increases from 80% to 82%, likely benefiting from the dimensionality reduction. The SVM model, which already performed well with all features at 84% accuracy, further improves to 87% with eight features selected, highlighting the effectiveness of choosing the most relevant features for this model. LR sees a slight increase from 75% to 76%, suggesting that it remains relatively stable relative to the number of features. Finally, AdaBoost's improvement in accuracy from 82% to 84% with selected features indicates a positive response to feature selection, likely due to a reduction in variance and noise in the data.

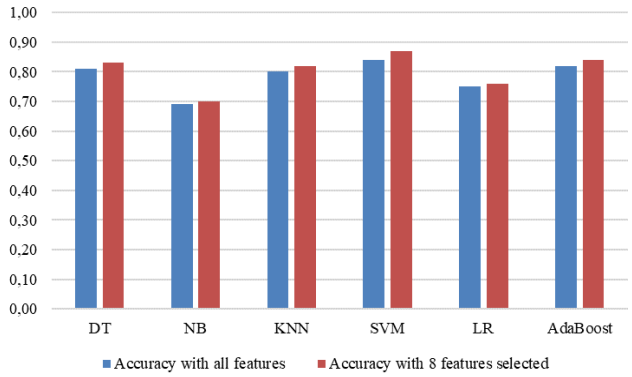

Fig. 6. Accuracy of models using all features vs. Top 8 selected features.

The comparison of various classifiers using full and top eight selected feature sets, as illustrated in Table V and Fig. 7, showcases how feature selection impacts the performance of machine learning models across several metrics such as precision, recall, and F1 score. Notably: the SVM model showed an impressive increase in precision from 76.33% to 82.09%, and in recall from 81.29% to 84.43%, with a corresponding improvement in the F1 score from 79.21% to 81.10%. These results indicate that the SVM model, with its ability to maximize the margin between classes, benefits from reducing complexity and noise by eliminating irrelevant features. Similarly, the AdaBoost model, demonstrated significant progress in terms of precision, increasing from 74.42% to 76.31%, and in recall from 80.14% to 82.71%, with an enhancement of its F1 score from 77.70% to 78.36%. This improvement shows that precise feature selection can indeed optimize AdaBoost’s capability. KNN also displayed notable improvements in all performance metrics with effective feature selection. Precision increased from 73.47% to 75.65%, recall from 75.86% to 76.29%, and the F1 score from 74.87% to 75.75%.

The performance of other models on datasets with all features and with the eight selected features also shows interesting results, though less dramatic than for SVM and AdaBoost. The DT model observed a slight improvement after feature selection. Precision increased from 72.29% to 73.10%, recall from 77.29% to 78.43%, and the F1 score from 76.29% to 76.53%. This modest improvement suggests that even for a relatively simple model like Decision Trees, which is less prone to overfitting, removing non-essential features can help clarify classification decisions.

TABLE V. EVALUATION METRICS OF VARIOUS CLASSIFIERS WITH ALL FEATURES VS. TOP EIGHT SELECTED FEATURES

Models	Dataset with all features (20 features)				Dataset with top 8 features selected			
	Accuracy	*Precision*	*Recall*	*F1 Score*	*Accuracy*	*Precision*	*Recall*	*F1 Score*
DT	0,81	0,722857143	0,772857143	0,762857143	0,83	0,730967742	0,784285714	0,765254237
NB	0,69	0,702857143	0,632857143	0,602857143	0,70	0,710967742	0,653654875	0,601254237
KNN	0,80	0,734705882	0,758571429	0,748709677	0,82	0,756451613	0,762857143	0,757457627
SVM	**0,84**	**0,763333333**	**0,812857143**	**0,792068966**	**0,87**	**0,820897652**	**0,844285714**	**0,811034483**
LR	0,75	0,701428571	0,616428571	0,607142857	0,76	0,711828571	0,616885714	0,609142857
AdaBoost	0,82	0,744193548	0,801428571	0,777017544	0,84	0,763103448	0,827142857	0,783559322

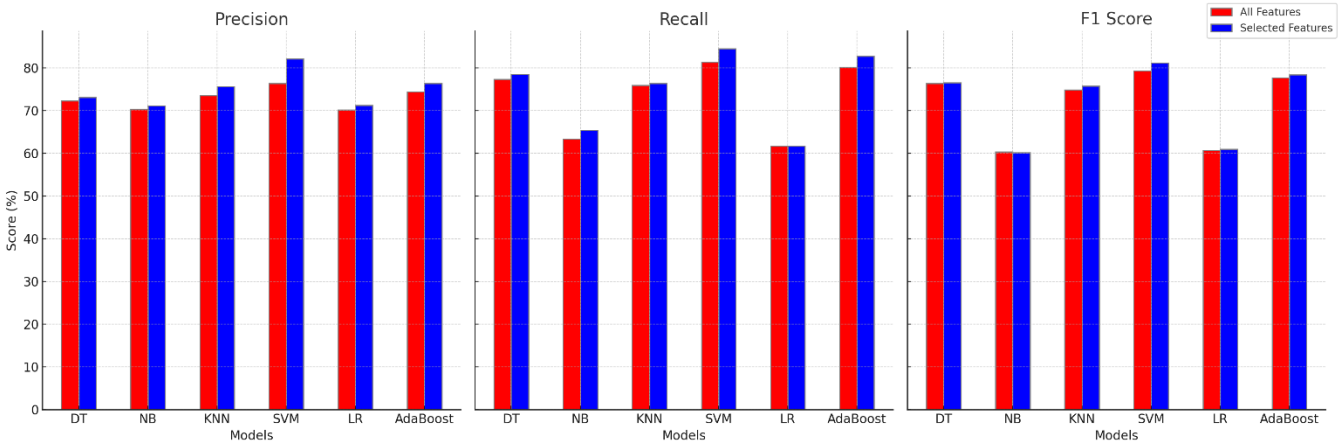

Fig. 7. Classifier metrics comparison with all features vs. selected features.

The NB model saw a minimal increase in precision, from 70.29% to 71.10%, and in recall, from 63.29% to 65.37%, but a slight decrease in the F1 score from 60.29% to 60.13%. These results indicate that while feature selection improved precision and recall, the overall impact on the harmony between these metrics was minimal. LR showed a slight increase in precision from 70.14% to 71.18% and recall from 61.64% to 61.69%, with a similar rise in the F1 score from 60.71% to 60.91%.

The analysis demonstrates that reducing the number of features from 20 to 8 generally enhances performance across most classifiers, albeit by varying degrees. These observations provide valuable insight into how different models respond to feature reduction and can guide modeling and preprocessing choices in future studies.

V. DISCUSSION

These results collectively highlight that feature selection can generally enhance the performance of the machine learning models in terms of accuracy, precision, recall, and the F1 score. Most models show an improvement across all metrics, especially notable in SVM and AdaBoost, which suggests that reducing the number of features to the most relevant ones can significantly enhance model performance. The NB model, while showing improvement in recall, does not show a proportional increase in the F1 score, this pattern may stem from the inherent probabilistic characteristics of this classifier. Also, the marginal improvements observed in the LR model suggest that not all models uniformly benefit from feature reduction, possibly due to the specific nature of the data.

The effectiveness of the SVM in student prediction applications is demonstrated by its higher performance when combined with feature selection utilizing Random Forest (RF) and Recursive Feature Elimination with Cross-Validation (RFECV). This method works especially well when dimensionality reduction is essential to increasing the interpretability and efficiency of the model, and when there is a complex relationship and interactions between attributes. The strategic feature selection approach complements SVM's resilience in determining the best hyperplane for classification and its ability to handle high-dimensional spaces.

However, it is imperative to recognize the possible disadvantages and limitations linked to this methodology. If the procedure is not appropriately fine-tuned, one drawback is the possibility of overfitting. It's also possible that the computational requirements will rise dramatically. Furthermore, the particular hyperparameters selected may significantly impact the method's effectiveness. This underscores the necessity for tailored feature selection strategies that align with the strengths of each model to optimize performance.

By focusing heavily on the predictive models of machine learning, we risk neglecting particular cases that could represent unique educational paths or specific challenges encountered by certain groups of students. Additionally, an overreliance on predictive analytics could lead decision-makers

to prioritize adherence to the model, possibly sidelining broad educational goals. It is important to consider these factors when analyzing the effectiveness and suitability of strategies for predicting student's performances.

VI. Conclusion and Future Work

In this study, we developed a robust model for predicting student performance at IBN ZOHR University by employing a combination of Random Forest and Recursive Feature Elimination with Cross-Validation (RFECV-RF) for optimal feature selection. Our dataset consists of 174,135 records and 21 attributes, collected over the period from 2016 to 2020.

Our experiments demonstrated that the SVM classifier, utilizing the top 8 features selected through RFECV-RF, outperformed other models, achieving an impressive accuracy of 87%. This underscores the efficacy of our feature selection approach and the SVM model's robustness in accurately predicting student performance. Other classifiers, such as AdaBoost, Decision Tree (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Logistic Regression (LR) also showed varying degrees of improvement with feature selection, but none matched the performance of SVM.

Regarding future work, we look forward to addressing class imbalance within our dataset. Our current dataset shows variations in the representation of some classes. To address this variation, we plan to explore several rebalancing techniques. Additionally, we plan to explore additional feature selection techniques such as genetic algorithms, which have the potential to refine the selection of relevant features further and enhance the model's predictive accuracy. Furthermore, we plan to test our model on datasets from other universities to validate the generalizability of our approach. We aim to ensure the model's robustness and applicability across different student populations and academic environments by applying it to diverse educational contexts.

References

[1] C. Dziuban, C. R. Graham, P. D. Moskal, A. Norberg, and N. Sicilia, "Blended learning: the new normal and emerging technologies," International journal of educational technology in Higher education, vol. 15, pp. 1–16, 2018.

[2] C. Romero and S. Ventura, "Data mining in education," Wiley Interdisciplinary Reviews: Data mining and knowledge discovery, vol. 3, no. 1, pp. 12–27, 2013.

[3] R. Baker and P. Inventado, "Educational Data Mining and Learning Analytics: Learning analytics Springer," New York, NY, 2014.

[4] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," Educ Inf Technol, vol. 23, no. 1, pp. 537–553, Jan. 2018, doi: 10.1007/s10639-017-9616-z.

[5] H. Li, C. F. Lynch, and T. Barnes, "Early prediction of course grades: models and feature selection," arXiv preprint arXiv:1812.00843, 2018.

[6] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," Applied Sciences, vol. 10, no. 3, p. 1042, Feb. 2020, doi: 10.3390/app10031042.

[7] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interactive Learning Environments, vol. 31, no. 6, pp. 3360–3379, May 2021, doi: 10.1080/10494820.2021.1928235.

[8] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A Data Mining Approach to Predict Academic Performance of Students Using Ensemble Techniques," in Intelligent Systems Design and Applications, vol. 940, A. Abraham, A. K. Cherukuri, P. Melin, and N. Gandhi, Eds., in Advances in Intelligent Systems and Computing, vol. 940. , Cham: Springer International Publishing, 2020, pp. 749–760. doi: 10.1007/978-3-030-16657-1_70.

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Computer Science, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[10] S. Helal et al., "Predicting academic performance by considering student heterogeneity," Knowledge-Based Systems, vol. 161, pp. 134–146, Dec. 2018, doi: 10.1016/j.knosys.2018.07.042.

[11] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," in 2017 10th International Conference on Human System Interactions (HSI), Ulsan, South Korea: IEEE, Jul. 2017, pp. 188–192. doi: 10.1109/HSI.2017.8005026.

[12] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, "Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis," Journal of Information Processing, vol. 26, no. 0, pp. 170–176, 2018, doi: 10.2197/ipsjjip.26.170.

[13] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, "A Multiple Linear Regression-Based Approach to Predict Student Performance," in Advanced Intelligent Systems for Sustainable Development (AI2SD'2019), vol. 1102, M. Ezziyyani, Ed., in Advances in Intelligent Systems and Computing, vol. 1102. , Cham: Springer International Publishing, 2020, pp. 9–23. doi: 10.1007/978-3-030-36653-7_2.

[14] A. Alshanqiti and A. Namoun, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification," IEEE Access, vol. 8, pp. 203827–203844, 2020, doi: 10.1109/ACCESS.2020.3036572.

[15] H. Turabieh et al., "Enhanced Harris Hawks optimization as a feature selection for the prediction of student performance," Computing, vol. 103, no. 7, pp. 1417–1438, Jul. 2021, doi: 10.1007/s00607-020-00894-7.

[16] S. Shivaji, E. J. Whitehead, R. Akella, and S. Kim, "Reducing Features to Improve Bug Prediction," in 2009 IEEE/ACM International Conference on Automated Software Engineering, Auckland: IEEE, Nov. 2009, pp. 600–604. doi: 10.1109/ASE.2009.76.

[17] M. Zaffar, M. Ahmed, K. S. Savita, and S. Sajjad, "A Study of Feature Selection Algorithms for Predicting Students Academic Performance," ijacsa, vol. 9, no. 5, 2018, doi: 10.14569/IJACSA.2018.090569.

[18] O. W. Adejo and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," JARHE, vol. 10, no. 1, pp. 61–75, Feb. 2018, doi: 10.1108/JARHE-09-2017-0113.

[19] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student Academic Performance Prediction using Supervised Learning Techniques," Int. J. Emerg. Technol. Learn., vol. 14, no. 14, p. 92, Jul. 2019, doi: 10.3991/ijet.v14i14.10310.

[20] A. Razaque and A. M. Alajlan, "Supervised Machine Learning Model-Based Approach for Performance Prediction of Students," Journal of Computer Science, vol. 16, no. 8, pp. 1150–1162, Aug. 2020, doi: 10.3844/jcssp.2020.1150.1162.

[21] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[22] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection," IJCAI, vol. 47, no. 1, Mar. 2023, doi: 10.31449/inf.v47i1.4519.

[23] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ullah Khan, and N. Aslam, "Predicting Student Academic Performance using Support Vector Machine and Random Forest," in 2020 3rd International Conference on Education Technology Management, London United Kingdom: ACM, Dec. 2020, pp. 100–107. doi: 10.1145/3446590.3446607.

[24] W. Lian, G. Nie, B. Jia, D. Shi, Q. Fan, and Y. Liang, "An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning," Mathematical Problems in Engineering, vol. 2020, pp. 1–15, Nov. 2020, doi: 10.1155/2020/2835023.

[25] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," JSAN, vol. 12, no. 5, p. 67, Sep. 2023, doi: 10.3390/jsan12050067.

[26] S. M. F. D. Syed Mustapha, "Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods," ASI, vol. 6, no. 5, p. 86, Sep. 2023, doi: 10.3390/asi6050086.

[27] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Leaning," vol. 1, no. 1, 2006.

[28] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India: IEEE, Aug. 2020, pp. 729–735. doi: 10.1109/ICSSIT48917.2020.9214160.

[29] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using Python," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 12, pp. 4727–4735, 2019.

[30] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, "Educational data mining techniques for student performance prediction: method review and comparison analysis," Frontiers in psychology, vol. 12, p. 698490, 2021.

[31] Y. Manzali and M. Elfar, "Random forest pruning techniques: a recent review," in Operations research forum, Springer, 2023, p. 43.

[32] M. Sandri and P. Zuccolotto, "Variable selection using random forests," in Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005, Springer, 2006, pp. 263–270.

[33] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," Chemometrics and Intelligent Laboratory Systems, vol. 83, no. 2, pp. 83–90, Sep. 2006, doi: 10.1016/j.chemolab.2006.01.007.

[34] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-6849-3.