Evaluating Machine Learning Models for Predicting Graduation Timelines in Moroccan Universities

Azeddine Sadqui[1], Merouane Ertel[2], Hicham Sadiki[3], Said Amali[4]
Informatics and Applications Laboratory (IA), Faculty of Sciences, Moulay Ismail University, Meknes, Morocco[1, 2, 3]
Informatics and Applications Laboratory (IA), FSJES, Moulay Ismail University, Meknes, Morocco[4]

***Abstract*—The escalating student numbers in Moroccan universities have intensified the complexities of managing on-time graduation. In this context, Machine learning methodologies were utilized to analyze the patterns and predict on-time graduation rates in a comprehensive manner. Our dataset comprised information from 5236 bachelor students who graduated in the years 2020 and 2021 from the Faculty of Law, Economic, and Social Sciences at Moulay Ismail University. The dataset incorporated a diverse range of student attributes including age, marital status, gender, nationality, socio-economic category of parents, profession, disability status, province of residence, high school diploma attainment, and academic honors, all contributing to a comprehensive understanding of the factors influencing graduation outcomes. Implementation and evaluation of the performance of five different machine learning models: Support Vector Machines, Decision Tree, Naive Bayes, Logistic Regression, and Random Forest, were carried out. These models were assessed based on their classification reports, confusion matrices, and Receiver Operating Characteristic (ROC) curves. From the findings, the Random Forest model emerged as the most accurate in predicting on-time graduation, showcasing the highest accuracy and ROC AUC score. Despite these promising results, it is believed that performance enhancements can be achieved through further tuning and preprocessing of the dataset. Insights from this study could enable Moroccan universities, among others, to better comprehend the factors influencing on-time graduation and implement appropriate measures to improve academic outcomes.**

Keywords*—*Machine learning; logistic regression; classification reports; on time graduation; Moroccan universities

I. Introduction

The integration of technological advancements within higher education has stimulated a shift towards data-driven strategies to manage burgeoning student enrollments and optimize institutional systems. Particularly, on-time graduation, a significant performance metric, is becoming increasingly challenging to predict and manage [1]. This predicament isn't confined to a single region, as institutions worldwide are contending with it. The case of Morocco, with its unique socio-economic contexts, is even more compelling [2].

This research intends to tackle this critical issue by applying machine learning methodologies to forecast on-time graduation rates at the Faculty of Law, Economic, and Social Sciences at Moulay Ismail University. A comprehensive dataset of 5236 bachelor students who graduated in 2020 and 2021 was utilized. The dataset includes numerous student characteristics such as age, marital status, gender, nationality, socio-economic category of parents, profession, disability status, province of residence, high school diploma attainment, and academic honors. It is posited that the detailed analysis of these variables could reveal valuable insights into the determinants of on-time graduation [3].

To identify the most effective method for predicting on-time graduation rates, a comparative analysis of five machine learning models - Support Vector Machines (SVM), Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR), and Random Forest (RF) - was conducted. Each model's performance was evaluated based on several statistical measures, including classification reports, confusion matrices, and Receiver Operating Characteristic (ROC) curves.

The implications of this research go beyond academic discourse. The findings could provide actionable insights for higher education institutions, particularly in Morocco, facilitating the formulation of effective data-driven strategies for improving on-time graduation management.

The paper is organized as follows: Post the introduction, the methodology, including data collection and analysis procedures, is detailed. Subsequent sections present a comprehensive discussion of the results, interpreting and comparing the performance of the different machine learning models. Potential implications of the findings are then outlined, offering a practical perspective. Finally, the paper concludes with a summary of the research findings and suggests directions for future research.

II. Related Work

Several studies have addressed the question of predicting student success and on-time graduation using machine learning techniques. These studies, while varying in scope and methodology, provide valuable insights into the potential of machine learning in education.

For instance, Marbouti [4] conducted a study using machine learning to predict the success of first-year engineering students based on high school academic performance. They used logistic regression and decision tree models, highlighting the significant role of high school mathematics grades in predicting success.

Similarly, Delen [5] used decision tree, neural network, and logistic regression models to predict students' graduation status based on demographic and academic data. They found that the

neural network model performed best in predicting student graduation status.

In the context of Moroccan higher education, however, the application of machine learning for predicting on-time graduation remains relatively unexplored. This study contributes to filling this research gap by applying machine learning techniques to a dataset from the Faculty of Law, Economic, and Social Sciences at Moulay Ismail University.

Notably, this work extends beyond the previous studies by comparing the performance of five different machine learning models: Support Vector Machines, Decision Tree, Naive Bayes, Logistic Regression, and Random Forest, in predicting on-time graduation. Moreover, a diverse range of student attributes is incorporated, aiming to create a more comprehensive prediction model.

Through this approach, the aim is to further the understanding of the factors influencing on-time graduation and contribute to the development of more effective strategies for academic success in Moroccan universities.

III. Materials and Methods

This section outlines the data collection process, the variables incorporated in the study, the preprocessing steps undertaken, and the machine learning models employed for analysis.

A. Data Source

The dataset was derived from the student records of the Faculty of Law, Economic, and Social Sciences at Moulay Ismail University, encompassing 5236 bachelor students who graduated in the years 2020 and 2021. The dataset was collected and anonymized in strict compliance with data privacy regulations.

B. Dataset Features

The dataset incorporated a variety of student characteristics (Fig. 1) such as age, marital status, gender, nationality, socio-economic category of parents, profession, disability status, province of residence, high school diploma attainment, and academic honors. The target variable was 'Graduate on Time,' a binary variable indicating whether the student graduated within the standard duration of the program. (Table I) summarizes the main variables of this study.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5233 entries, 0 to 5232
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   5233 non-null   object
 1   Marital_status        5233 non-null   object
 2   Gender                5233 non-null   object
 3   Nationality           5233 non-null   object
 4   Parents_Categ_Soscio  5233 non-null   object
 5   Profesion             5233 non-null   object
 6   Disability            5233 non-null   object
 7   Province              5233 non-null   object
 8   Hight_school_diploma  5233 non-null   object
 9   Academic_honor        5233 non-null   object
 10  Graduate_on_time      5233 non-null   int64
dtypes: int64(1), object(10)
memory usage: 449.8+ KB
```

Fig. 1. Dataframe information.

TABLE I. Student Characteristics

Variable	Definition
Age	Age of the student in years
Marital_status	Marital status of the student
Gender	Gender of the student
Nationality	Nationality of the student
Parents_Categ_soscio	Socio-economic category of the student's parents
Profession	Profession of the student
Disability	Indicates whether the student has a disability
Province	Province of the student's residence
High_School_Diploma	Indicates the type of high school diploma the student has.
Academic_honor	Academic honors achieved by the student
Graduate_on_time	Indicates whether the student graduated on time

C. Label Encoding

The dataset under consideration encompasses a range of features that capture the student's demographic and academic characteristics. The 'Age' feature represents the age of the student in years. 'Marital_status' indicates the student's marital status, represented by numeric values where 1 signifies being Single, 2 stands for Married, 3 denotes Divorced, and 4 implies a Widower. 'Gender' signifies the student's gender, with 0 indicating Female and 1 representing Male. 'Nationality', 'Parents_Categ_soscio', 'Profession', and 'Province' are represented by IDs corresponding to different nationalities, socio-economic categories of parents, professions, and provinces respectively. 'Disability' is a binary indicator that highlights whether a student has a disability, where 0 denotes No and 1 stands for Yes. 'High_School_Diploma' points to the type of high school diploma the student possesses, represented by different IDs for each type of diploma. Academic_honor' delineates the academic honors a student has achieved, ranging from 1 (Passing), 2 (Good), 3 (Very Good), to 4 (Outstanding). Lastly, 'Graduate_on_time' is a binary variable that indicates whether the student graduated within the standard duration of the program, represented by 0 (No) and 1 (Yes). These features collectively provide a comprehensive profile of the students' demographic and academic landscape (Table II, Fig. 2).

D. Correlation Features

The features used in this study have differing levels of correlation with the target variable, 'Graduate on Time'. The correlation values signify the strength and direction of the relationship between each feature and the target [6]. These values were computed and visualized through a correlation matrix. A positive correlation indicates that as the feature value increases, the likelihood of on-time graduation also increases, and vice versa. Conversely, a negative correlation means that as the feature value increases, the likelihood of on-time graduation decreases.

TABLE II. METHOD FOR ENCODING OF VARIABLES

Variable	Definition	Possible Values
Age	Age of the student in years	Numeric values
Marital_status	Marital status of the student	1(Single), 2(Married), 3(Divorced), 4 (Widower)
Gender	Gender of the student	0(Female), 1 (Male)
Nationality	Nationality of the student	IDs corresponding to different nationalities
Parents_Categ_soscio	Socio-economic category of the student's parents	IDs corresponding to different socio-economic categories
Profession	Profession of the student	0 (No), 1 (Yes)
Disability	Indicates whether the student has a disability	0 (No), 1 (Yes)
Province	Province of the student's residence	IDs corresponding to different provinces
High_School_Diploma	Indicates the type of high school diploma the student has.	IDs corresponding to different high school diploma
Academic_honor	Academic honors achieved by the student	1 (Passing), 2 (Good), 3 (Very Good), 4 (Outstanding)
Graduate_on_time	Indicates whether the student graduated on time	0 (No), 1 (Yes)

```
<class 'pandas.core.frame.DataFrame'>
Index: 5233 entries, 0 to 5233
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   5233 non-null   int64
 1   Marital status        5233 non-null   int64
 2   Gender                5233 non-null   int64
 3   Nationality           5233 non-null   int64
 4   Parents Categ Soscio  5233 non-null   float64
 5   Profesion             5233 non-null   int64
 6   Disability            5233 non-null   int64
 7   Province              5233 non-null   int64
 8   Hight School Diploma  5233 non-null   int64
 9   Academic Honor        5233 non-null   int64
 10  Graduate on Time      5233 non-null   int64
dtypes: float64(1), int64(10)
memory usage: 490.6 KB
```

Fig. 2. Dataframe after encoding.

In the dataset, the 'Academic Honor', 'High School Diploma', and 'Province' are the features that show the highest correlation with on-time graduation as depicted in the correlation matrix (Fig. 3). The 'Academic Honor' feature, representing the academic performance of students, shows a positive correlation, suggesting that students with higher academic honors are more likely to graduate on time. Similarly, the 'High School Diploma' and 'Province' features also have a positive correlation with on-time graduation, indicating that the type of high school diploma and the province of residence can have an influence on graduation times (Fig. 4).

These highly correlated features are particularly beneficial in predictive modeling, as they provide significant insight into the factors that influence on-time graduation. By focusing on these variables in the machine learning models, it becomes possible to make more accurate predictions and gain a deeper understanding of the factors contributing to graduation times.

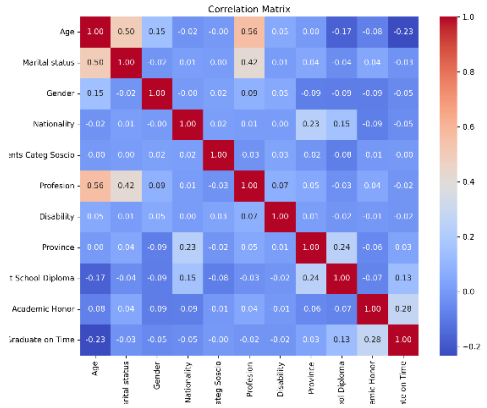

Fig. 3. Heat map for checking correlated columns for graduate on time.

```
correlation['Graduate on Time'].sort_values(ascending=False)
```

```
Graduate on Time        1.000000
Academic Honor          0.277324
Hight School Diploma    0.129158
Province                0.033249
Parents Categ Soscio   -0.003765
Disability             -0.016831
Profesion              -0.020155
Marital status         -0.026202
Nationality            -0.045326
Gender                 -0.052751
Age                    -0.233321
Name: Graduate on Time, dtype: float64
```

Fig. 4. Ranking of correlations with graduate on time.

E. Modeling

In this study, the modeling procedure takes place after the data preprocessing stage. This procedure entails training machine learning algorithms to predict whether a student will graduate on time based on their academic and demographic characteristics. Several well-regarded machine learning techniques were employed, including Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). These classification models are popular and efficient for dealing with such a binary classification task. The models were trained based on attributes such as Academic Honor, High School Diploma, and Province, aiming to classify students into two categories: those who are likely to graduate on time and those who are not. The Python scikit-learn library was used for data analysis and model implementation. The models were evaluated using a split-test method, partitioning the original dataset into a training set (80%) to train the model, and a test set (20%) to evaluate it. This technique is commonly used in machine learning to assess the effectiveness and efficiency of predictive models. By comparing the performance of the different models, the aim is to identify the one that provides the most accurate predictions for on-time graduation among students at the Faculty of Law, Economic, and Social Sciences at Moulay Ismail University. The entire procedure of the experiment is depicted in (Fig. 5).

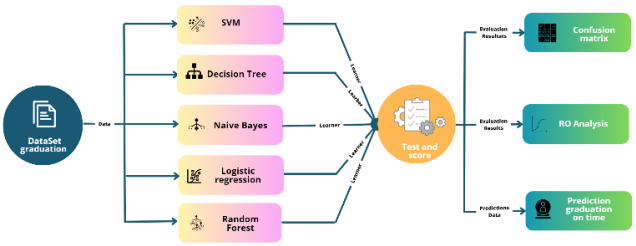

Fig. 5. Model machine learning use.

1) Machine learning algorithms: In this study, we utilized five different machine learning algorithms, each with its strengths and applicable use cases. These include:

a) Support Vector Machine (SVM): SVM is a widely used classification algorithm that finds the hyperplane in an N-dimensional space that distinctly classifies the data points. It is especially useful in high dimensional spaces and situations where the number of dimensions exceeds the number of samples [7].

b) Naive Bayes (NB): The Naive Bayes classifier is a simple and efficient machine learning algorithm often used in text classification, spam filtering, recommendation systems, etc. It is based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of a feature [8].

c) Decision Tree (DT): Decision Trees are a type of flowchart-like structure where each internal node represents a feature (or attribute), each branch represents a decision rule, and each leaf node represents an outcome. They are widely used due to their interpretability and simplicity [9].

d) Logistic Regression (LR): Logistic regression is a type of regression analysis used for predicting the probability of a binary outcome. It's a statistical model that uses a logistic function to model a binary dependent variable [10].

e) Random Forest (RF): Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the majority vote of individual trees for classification problems or average prediction for regression problems. It's a powerful algorithm known for its robustness and simplicity [11].

F. Performance Indicators

In this study, the performance of the various machine learning models was evaluated using several widely recognized performance indicators:

1) Confusion matrix: The Confusion Matrix is another key performance indicator used in this study. It is a specific table layout that visualizes the performance of an algorithm, typically a supervised learning one. The matrix contrasts the actual and predicted classifications of the instances in a dataset to measure the quality of the output of the classifier. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class (Fig. 6).

In general, the confusion matrix provides four types of classification results with respect to a classification target k [12].

- True positive (TP): correct prediction of the positive class ($c_{k,k}$)
- True negative (TN): correct prediction of the negative class $\sum_{i,j\in N\setminus\{k\}} c_{ij}$
- False positive (FP): incorrect prediction of the positive class $\sum_{i\in N\setminus\{k\}} c_{ik}$
- False negative (FN): incorrect prediction of the negative class $\sum_{i\in N\setminus\{k\}} c_{ki}$

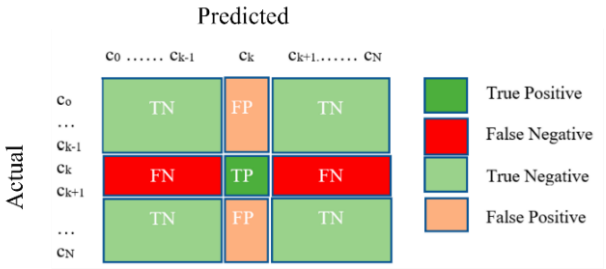

Fig. 6. Confusion matrix for multi-class.

The Confusion Matrix allows us to compute various other classification metrics, including precision, recall, F1-score, and support. By providing a more detailed view of how the classification model is performing, the Confusion Matrix plays a crucial role in understanding the behavior of the model beyond simple accuracy [13].

2) Classification report: A classification report offers a comprehensive synopsis of how well a classification model has performed. It consolidates key performance metrics such as accuracy, precision, recall, F1-score, and support.

a) Accuracy: Defined as the ratio of correctly predicted observations to total observations, accuracy is the most straightforward performance measure [14]. The accuracy of the model is defined as:

$$OverallAccuracy = \frac{(\sum_{i=1}^{N} c_{(i,i)})}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{(i,i)}} \quad \# \tag{1}$$

b) Precision: Precision, calculated as the ratio of correctly predicted positive observations to total predicted positive observations, indicates the model's ability to correctly identify only the relevant instances [15]. It's defined as:

$$Precision_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \quad \# \tag{2}$$

c) Recall (Sensitivity): Also known as sensitivity, recall measures the model's ability to identify all relevant instances, defined as the ratio of correctly predicted positive observations to all actual positives [16], it's defined by equation (3).

$$Recall_{class} = \frac{TP_{class}}{TP_{class} + FN_{class}} \quad \# \tag{3}$$

d) F1-Score: The F1 score combines precision and recall into a single metric by taking their harmonic mean, effectively balancing the trade-off between the two measures [17]. It's defined by equation (4).

$$F\ 1 - Score = \frac{2 * TP_{class}}{2 * TP_{class} + FN_{class} + FP_{class}}\ \# \quad (4)$$

e) Area Under Curve: The Receiver Operating Characteristic (ROC) curve, plotting the true positive rate against the false positive rate, indicates the model's discriminative power. The Area Under the Curve (AUC) offers a single measure summarizing the overall quality of the classifier [18].

G. Training and Validation

In this study, the modeling procedure commenced with partitioning the dataset into a training and validation set. As per the widely accepted practices in machine learning research, 80% of the total data was allocated for training the algorithms, while the remaining 20% was set aside for validation [19]. The objective was to ensure a realistic estimate of the models' performance on unseen data, providing a valuable measure of their generalizability.

The training phase of this research implemented five distinct machine learning models: Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). These models were selected due to their widespread adoption in predictive modeling tasks of a similar nature and their ability to effectively manage binary classification problems [20].

Python's scikit-learn library was utilized for the execution of these models [21]. The training of each model was conducted using the 'fit' function, while model parameters were optimized through the GridSearchCV function, an exhaustive search over a specified range of parameter values [22].

After the training, the models underwent validation using the validation set. Various performance metrics were employed to evaluate model performance, including Accuracy, Precision, Recall, and the F1-score. All of these metrics were featured in the Classification Report [15]. Additionally, the Receiver Operating Characteristic (ROC) curve was graphed for a visual assessment of the model's performance [23].

For this study, a 10-fold cross-validation technique was also incorporated during the training phase [19]. This technique subdivides the training set into 10 subsets, and the model is trained 10 times. In each training iteration, nine subsets are used for training, and one is used for validation. The ultimate model performance is computed as the average performance of the ten models. This method helps generate a more dependable performance estimate and mitigates the risk of overfitting [24].

IV. Results and Discussion

A. Analysis of Results

The performance of each model is summarized by highlighting key metrics and visual representations from the classification report and confusion matrix.

The SVM model achieved an overall accuracy of 69%. However, it demonstrated high recall (97%) for class 0 but had a low recall (11%) for class 1. This indicates that while the SVM model was able to identify the majority of class 0 instances correctly, it struggled to correctly classify instances from class 1 (Table III, Fig. 7).

TABLE III. Metric Report for SVM

Metric	Class 0	Class 1
Precision	0.69	0.64
Recall	0.97	0.11
F1-Score	0.81	0.18
Support	705	342

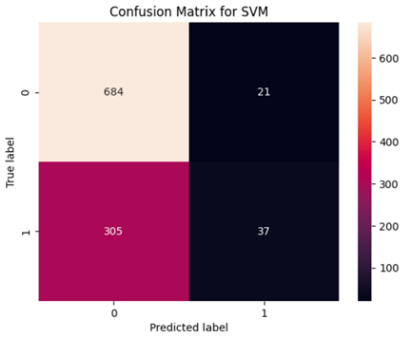

Fig. 7. Confusion matrix for SVM

The Decision Tree model performed better with an accuracy of 74%. It demonstrated a better balance in classifying both classes with a recall of 82% and 57% for class 0 and class 1 respectively (Table IV, Fig. 8).

TABLE IV. Metric Report for Decision Tree

Metric	Class 0	Class 1
Precision	0.80	0.60
Recall	0.82	0.57
F1-Score	0.81	0.59
Support	705	342

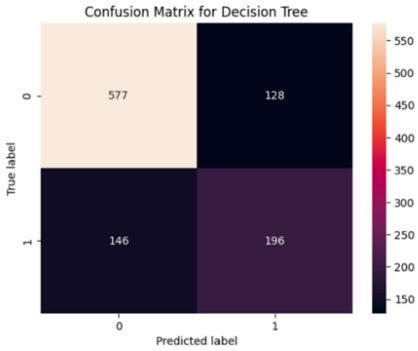

Fig. 8. Confusion matrix for decision tree.

The Naive Bayes model achieved an accuracy of 70%. While it showed good recall for class 0 (85%), it faced difficulties in classifying class 1 instances correctly, similar to the SVM model (Table V, Fig. 9).

TABLE V. METRIC REPORT FOR NAIVE BAYES

Metric	Class 0	Class 1
Precision	0.74	0.55
Recall	0.85	0.38
F1-Score	0.79	0.45
Support	705	342

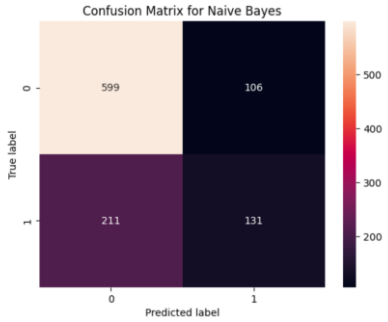

Fig. 9. Confusion matrix for naive bayes.

The Logistic Regression model performed with an accuracy of 75%. It showed a fairly balanced performance for both classes, with a recall of 89% for class 0 and 47% for class 1 (Table VI, Fig. 10).

TABLE VI. METRIC REPORT FOR LOGISTIC REGRESSION

Metric	Class 0	Class 1
Precision	0.78	0.67
Recall	0.89	0.47
F1-Score	0.83	0.55
Support	705	342

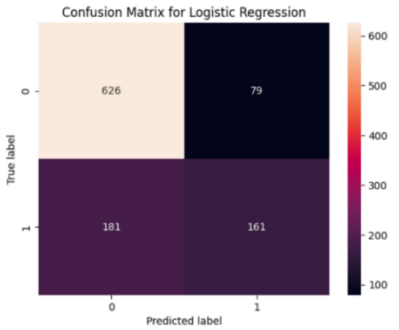

Fig. 10. Confusion matrix for logistic regression.

Random Forest was the top-performing model with an accuracy of 77%. It also demonstrated the most balanced performance with a recall of 85% and 62% for class 0 and class 1, respectively (Table VII, Fig. 11).

TABLE VII. METRIC REPORT FOR RANDOM FOREST

Metric	Class 0	Class 1
Precision	0.82	0.66
Recall	0.84	0.62
F1-Score	0.83	0.64
Support	705	342

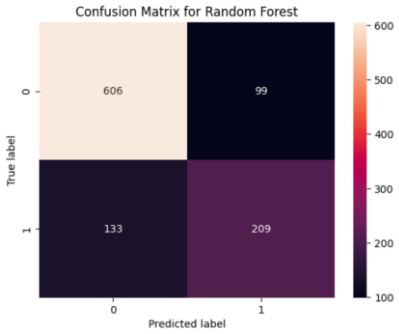

Fig. 11. Confusion matrix for random forest.

ROC and AUC curve

The Receiver Operating Characteristic (ROC) curve, in conjunction with the Area Under the Curve (AUC), offer insightful metrics for assessing the classification performance of a predictive model. In the context of this study, the ROC AUC scores reveal that the Random Forest model surpasses its counterparts, recording the highest score of 0.82. This indicates a superior capacity of the Random Forest model in distinguishing between students likely to graduate on time and those who aren't, across various thresholds.

Following closely, the Logistic Regression model achieved the second-highest ROC AUC score of 0.75. This suggests a commendable proficiency of this model in accurately classifying the students. The Decision Tree model was also noteworthy, with a score of 0.74. The SVM and Naive Bayes models demonstrated similar performance levels, with ROC AUC scores of 0.73 each. Although these scores are lower compared to the Random Forest and Logistic Regression models, they still denote reasonable classification capabilities (Fig. 12).

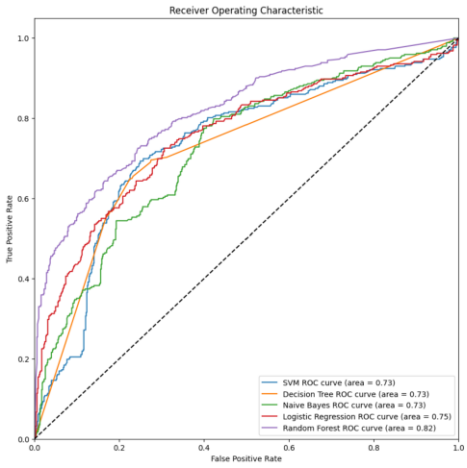

Fig. 12. ROC curve for SVM, DT, NB, LR, RF.

B. Discussion

In this study, the relationship between a variety of features and the ability to graduate on time was investigated. After a careful data transformation and correlation analysis, three features, 'Academic Honor', 'High School Diploma', and 'Province' were selected based on their strong correlation with the target variable.

Several popular machine learning models were used to construct and validate predictive models. These models included Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). Each of these models was trained using the train-test split method, and their performance was evaluated using a suite of metrics including accuracy, precision, recall, and the F1 score. Also, the ROC-AUC score was calculated to measure the performance of the models under different classification thresholds.

The results clearly demonstrate that the Random Forest model was superior to all other models, achieving top marks in accuracy, precision, recall, and F1 score for predicting timely graduation of students. In addition, it accomplished an ROC-AUC score of 0.82, a clear indicator of its excellent performance. Although the Decision Tree and Logistic Regression models exhibited commendable performance, they were unable to match the outstanding performance of the Random Forest model.

Future work in this area could include the incorporation of more features, utilization of diverse feature selection techniques, and experimentation with alternative machine learning models. Furthermore, additional data should be considered. Ultimately, the goal is to construct a reliable predictive model capable of accurately identifying students at risk of not graduating on time, thereby enabling early interventions to support these students in achieving success.

V. Conclusion

The application of various machine learning models was showcased in this study to predict students' ability to graduate on time. 'Academic Honor', 'High School Diploma', and 'Province' were used as predictors. The central goal was to identify the model that most accurately could assist educational institutions in pinpointing at-risk students and implementing interventions in a timely manner.

Among all the models tested, the Random Forest model stood out as the most effective, achieving the highest precision, recall, F1-score, and ROC-AUC score. The model's ability to manage high-dimensional spaces and generate an internal unbiased estimate of the generalization error distinguished it from the other tested models. Therefore, this study underscores the use of machine learning, and particularly ensemble methods like Random Forest, as potent tools in the educational sector.

As a part of future work, further fine-tuning of the Random Forest model is suggested, along with the exploration of other machine learning models and additional student features. A potential avenue for future research may also include the implementation of this model in other educational contexts, allowing for a more comprehensive understanding of its applicability and robustness.

In conclusion, this study accentuates the potential of machine learning in education to predict student outcomes and enable proactive measures. With the ongoing integration of technology into education, the importance of such predictive tools is projected to rise. Such tools equip educational institutions with better understanding of student needs, enabling them to tailor their support services more effectively and, ultimately, assist more students in achieving their educational goals.

References

[1] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, “Predicting time to graduation at a large enrollment American university,” PLOS ONE, vol. 15, no. 11, p. e0242334, Nov. 2020.

[2] V. Llorent-Bedmar, “Educational Reforms in Morocco: Evolution and Current Status,” Int. Educ. Stud., vol. 7, no. 12, p. p95, Nov. 2014.

[3] R. S. J. d. Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” Oct. 2009.

[4] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, “Models for early prediction of at-risk students in a course using standards-based grading,” vol. 103, pp. 1–15.

[5] D. Delen, “Predicting Student Attrition with Data Mining Methods,” vol. 13, no. 1, pp. 17–35.

[6] J. D. Febro, “Utilizing Feature Selection in Identifying Predicting Factors of Student Retention,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, 2019.

[7] C. Cortes and V. Vapnik, “Support-vector networks” vol. 20, no. 3, pp. 273–297.

[8] H. Zhang, “The Optimality of Naive Bayes”.

[9] J. R. Quinlan, “Induction of decision trees,” vol. 1, no. 1, pp. 81–106.

[10] J. S. Cramer, “The Origins of Logistic Regression”.

[11] L. Breiman, “Random Forests,” vol. 45, no. 1, pp. 5–32.

[12] S. Abraham, C. Huynh, and H. Vu, “Classification of Soils into Hydrologic Groups Using Machine Learning,” vol. 5, no. 1, p. 2.

[13] E. Merouane, A. Said, and E. F. Nour-eddine, “Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers,” International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, 2022.

[14] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”.

[15] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” vol. 45, no. 4, pp. 427–437.

[16] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in Proceedings of the 23rd international conference on Machine learning - ICML ’06, ACM Press, pp. 233–240. doi: 10.1145/1143844.1143874.

[17] H. B. Nembhard, “Statistical Process Adjustment Methods for Quality Control,” vol. 99, no. 466, pp. 567–568.

[18] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve.,” vol. 143, no. 1, pp. 29–36.

[19] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in International Joint Conference on Artificial Intelligence,

[20] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, and others, “Supervised machine learning: A review of classification techniques,” vol. 160, no. 1, pp. 3–24.

[21] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” no. arXiv:1201.0490. arXiv.

[22] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization.,” vol. 13, no. 2.

[23] T. Fawcett, “An introduction to ROC analysis,” vol. 27, no. 8, pp. 861–874.

[24] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection” vol. 4.