Unlocking Student Dropout Patterns: A Machine Learning-Based Data Analysis

Anikinee Deb[1]
Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, SMU, Majitar, Sikkim, India

***Abstract*—In higher education, student dropout is still a major problem that has an impact on both the student outcome as well as institutional performance. This study uses a multi-dimensional dataset that includes academic, demographic, and economic features to examine how supervised machine learning algorithms can be used for predicting student dropout. These features were used to train six classifiers - Random Forest, Support Vector Classifier (SVC), Logistic Regression, K-Nearest Neighbours (KNN), Decision Tree, and Naive Bayes and evaluated using the performance metrics – Accuracy and ROC-AUC. With ROC-AUC scores of 0.956, 0.951, and 0.948, respectively, and accuracies of up to 0.91, Logistic Regression, Random Forest, and SVM outperformed the other classifiers in terms of predictive performance. To evaluate each feature group's relative contribution, a methodical feature ablation analysis was performed. The Logistic Regression and Random Forest models' ROC-AUC scores dropped from 0.956 to 0.810 and 0.951 to 0.816, respectively, when academic features were removed, resulting in the biggest performance drop of any model. On the other hand, the performance of the model was only marginally affected by the elimination of demographic or economic characteristics. The findings show that the most important predictors of student dropout are academic indicators, which are followed by economic and demographic characteristics.**

***Index Terms*—Educational Data Mining, Random Forest, KNN, SVC, Logistic Regression, Decision Tree, Naïve Bayes, Machine Learning, ROC-AUC, Dropout Prediction, Data Science**

I. INTRODUCTION

One of the biggest issues facing higher education globally is student dropout, which affects both the performance of the institution and the individual outcomes of the students. University dropouts have long-term economic disadvantages, fewer job opportunities, and social inequality for individuals, in addition to the disruption of their academic lives [1], [2]. According to the institutional viewpoint, dropout results in reduced graduation rates, financing problems, and resource inefficiencies, all of which have an impact on stakeholder confidence and national education rankings [2], [3].

Academic achievement, socioeconomic status, mental health, institutional involvement, and external macroeconomic circumstances are some of the many interconnected elements that contribute to the complexity of dropout behaviour [4], [5]. Although it has been widely recognised that academic indicators such as GPA or credit completion rate are powerful predictors, current research highlights the increasing significance of non-academic characteristics, such socioeconomic and demographic factors, in predicting dropout risk [6], [7]. However, the predictive ability of these variables varies between datasets and learning contexts and is context-sensitive [8].

Modern educational research is increasingly using data-driven methods, especially machine learning (ML), to address this complex issue and enable the early identification of students who are at danger. When analysing diverse educational data, machine learning (ML)-based predictive algorithms have shown an impressive level of accuracy and flexibility, revealing detailed patterns and connections that conventional statistical techniques tend to overlook [9], [10]. By offering timely insights into student behaviour and performance, these systems help educational institutions reduce the risk of dropout and improve retention rates by facilitating the development of proactive intervention methods.

Despite significant advancements, there are still gaps in our knowledge of the relative significance of the academic, demographic, economic data domains in predicting dropout outcomes. By comparing the performance of several machine learning classifiers, such as Random Forest, Support Vector Machines, K-

Nearest Neighbors, Logistic Regression, Decision Trees, and Naive Bayes, this work fills this gap. Additionally, using ROC-AUC as the main assessment metric, we conduct feature group ablation research to assess the impact of excluding particular data types on model performance.

This research aims to accomplish two main goals. Primarily, it seeks to assess how well various machine learning (ML) algorithms predict student dropout in an academic environment. Secondly, it aims to measure the relative influence of different feature categories on model performance, including academic, economic, and demographic factors. This study provides valuable insight by identifying the most important data areas, which can help institutions of higher education create focused intervention plans, allocate resources as efficiently as possible, and eventually enhance student retention results.

II. LITERATURE SURVEY

Understanding the reasons behind students' early academic departures is a problem that educational institutions and instructors around the world are working to address. Recent developments in artificial intelligence and machine learning have created new avenues for identifying hidden trends and important variables influencing student dropout. Cutting-edge research that uses these technologies to not only anticipate dropout risks but also to enable proactive, data-driven interventions is highlighted in the curated review that follows. This review sheds light on the qualities that are most important in keeping students interested and on track.

A. Elbouknify et al. (2025) – AI-Based Identification and Support of At-Risk Students: A Case Study of the Moroccan Education System

This study uses data from the Ministry of National Education and sophisticated machine learning algorithms to forecast student dropout rates in Morocco. By outlining which characteristics affect forecasts, SHAP (Shapley Additive Explanations) offers transparency and aids educators in comprehending the reasons behind dropouts. With an AUC of 87% and an accuracy and recall of 88%, the model demonstrated successful real-world performance. This strategy provides insightful information for focused interventions aimed at lowering dropout rates. [4]

B. Kim et al. (2023) – University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques

In this study, academic, demographic, socioeconomic, and macroeconomic data types were used to predict university dropout rates. To predict whether students would graduate or drop out, four binary classifiers were trained. The average ROC-AUC score for the classifiers' overall effectiveness in predicting dropout status was 0.935. Academic data had the greatest impact on model performance; the average ROC-AUC score dropped to 0.811 when academic-related elements were removed. According to the study's findings, academic information greatly improves dropout prediction accuracy. [11]

C. Andrade-Girón et al. (2023) – Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review

In order to predict student dropout, this thorough evaluation examined 23 papers that used machine learning (ML) and deep learning (DL) methods. According to the results, Random Forest was the most widely utilised algorithm, sometimes reaching 99% accuracy. In assessing the efficacy of the model, the study underlined the significance of performance indicators like accuracy, sensitivity, specificity, and AUC. It also emphasised the growing importance of deep learning models in identifying intricate patterns in educational data and emphasised issues that have been mentioned frequently in the literature, such as feature selection and data imbalance. The review's conclusion raised the possibility that hybrid strategies that combine ML and DL could enhance prediction accuracy and interpretability even further. [12]

D. K. Sood, A. L. Jimenez Martinez, and R. Mahto (2023) – Early Detection of At-Risk Students Using Machine Learning

The purpose of this study is to identify students who are at risk of dropping out in the early years of their academic careers by using machine learning techniques. The authors tested a number of algorithms using educational datasets that included behavioural indicators and academic performance measurements, such as Random Forest, Support Vector Machines

(SVM), Logistic Regression, and Gradient Boosting. With an accuracy of up to 92%, the Random Forest classifier outperformed the others; Gradient Boosting and SVM models came in second and third, respectively. The study emphasises how crucial early identification is in allowing teachers to take action before kids completely lose interest. The paper also addresses issues with feature selection and data quality, stressing the importance of thorough preprocessing and pertinent feature engineering in creating trustworthy predictive models. [13]

III. DATA PREPARATION AND ANALYSIS

A. Data Collection

This study makes use of the "Predict Students' Dropout and Academic Success" [14] dataset, which provides a comprehensive knowledge of undergraduate students from a Portuguese higher education institution. An extensive range of academic, demographic, and socioeconomic data are captured by the dataset, which has 4,424 instances and 35 attributes. It has several disjoint tables divided into three primary domains:

Demographic Factors: Factors associated with academic perseverance include age, gender, marital status, and nationality.

Socioeconomic Factors: Regional economic factors like GDP, inflation, and unemployment rate; parental education; work status; and scholarship assistance.

Academic and Institutional Data: Course type, application method, and academic performance indicators (enrolled credits, approved credits, evaluated credits, and semester grades).

By integrating academic, personal, and economic factors, this varied dataset makes it possible to identify dropout predictors.

B. Data Preprocessing

A crucial first step in guaranteeing high-quality input for machine learning models is data preprocessing.

- Missing Data: To avoid data distortion and guarantee a clean dataset, rows with missing values were eliminated.
- Categorical Encoding: To make nominal features appropriate for model input, they were converted using one-hot encoding, whereas ordinal features were encoded using label encoding.
- Feature Scaling: To ensure that every feature contributes equally to model training, we used StandardScaler to standardise numerical values by eliminating the mean and scaling to unit variance.

C. Exploratory Data Analysis

Initially, 4,424 student records were gathered where 1,421 of the students had dropped out, 2,209 had graduated, and 794 were still enrolled as shown in Figure 1. A final dataset of 3,630 students, consisting solely of graduates and dropouts, was obtained by excluding the 794 enrolled students because of their unclear academic status in order to concentrate on a binary classification assignment. 1,249 of these were men and 2,381 were women. According to Figure 2's gender-specific graduate and dropout distribution, female students were more likely to graduate, indicating greater academic performance than male students.

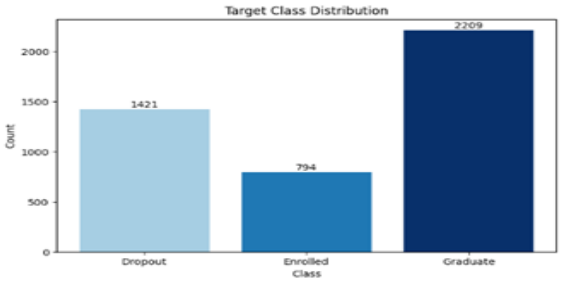

Fig 1: Target Class Distribution.

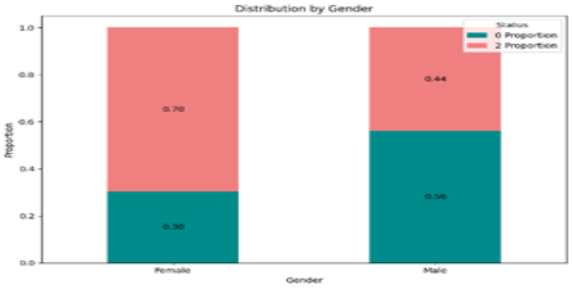

Fig 2: Distribution of target class by Gender.

Further analysis of student outcomes across a number of categorical factors, including marital status, attendance time (day versus evening), special education needs, displacement status, debtor status, tuition price payment status, scholarship holding, and international student status, is provided in Figures 3, 4, 5 and 6. Strong associations between socioeconomic characteristics and dropout risk were found; around

94% of students with past-due tuition dropped out, and 76% of debtors did not finish their education as shown in Figure 3. 86% of students who received scholarships, on the other hand, graduated successfully, which may indicate that they were more motivated to study and supports earlier research on the benefits of financial aid. Results were also impacted by marital status; students who were officially separated had greater dropout rates, whereas single students were more likely to graduate as shown in Figure 4.

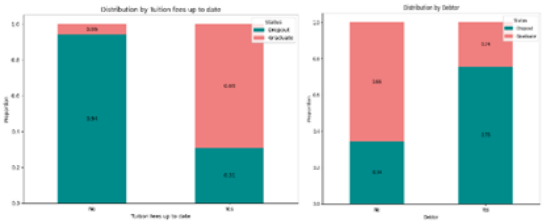

Fig 3: Distribution of target class by Tuition fees up to date and Debtor.

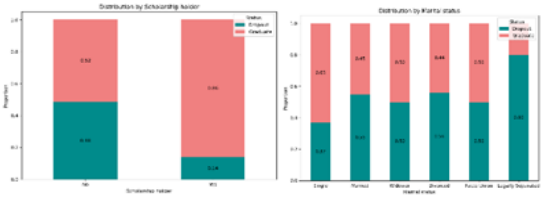

Fig 4: Distribution of target class by Scholarship holder and Marital status.

Academic outcomes were linked to the kind of attendance, as students who attended classes in the evening had a greater dropout rate (51%) than those who attended classes during the day (38%). Performance was impacted by displacement status; students who were displaced had a reduced graduation rate and a 34% dropout rate, most likely as a result of instability or outside pressures as shown in Figure 5. Likewise, the dropout rate for students with education special needs was marginally higher (42%) than that of their counterparts (39%), suggesting that they faced more academic difficulties. With graduation rates of 63% and 61%, respectively, international students showed an almost comparable dropout distribution to local students, indicating that dropout likelihood may not be greatly impacted by international status in this sample as shown in Figure 6.

To investigate the connections between different academic, demographic, social, and macroeconomic characteristics, a correlation matrix visualisation was created as shown in Figure 7.

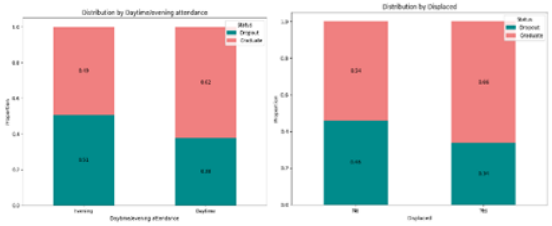

Fig 5: Distribution of target class by Daytime/evening attendance and Displaced.

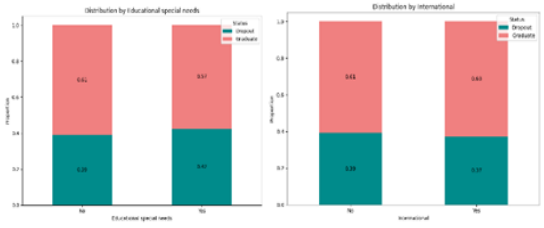

Fig 6: Distribution of target class by Educational special needs and International.

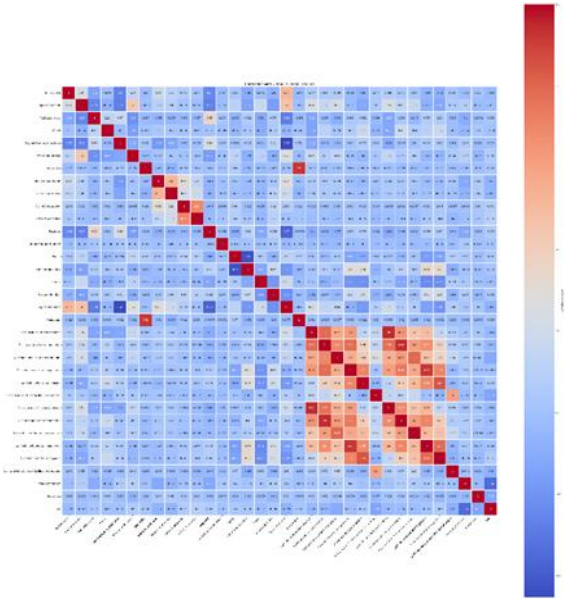

Fig 7: Correlation Matrix for all features

Significant relationships between "Father's qualification" and “Mother's qualification," as well as between "Father's occupation" and "Mother's occupation," are evident in Figure 6. Academic performance measures like "Curricular units 1st sem (enrolled)" and "Curricular units 2nd sem (enrolled)" also show substantial relationships, and "approved" and "grade" metrics show comparable trends between semesters. In terms of the consistency of family

background traits and academic performance, these results are in line with predictions.

IV. EXPERIMENTS

A. Experimental Design

Several essential steps make up the experimental workflow as shown in figure 8, which guarantees reliable model construction and assessment. Raw data first goes through preprocessing, which includes feature engineering, data cleaning, and exploration. After the dataset has been cleaned, it is divided into training (80%) and testing (20%) sets. The training set is used to train several classification algorithms, such as Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest Neighbours, and Naïve Bayes. Based on performance on the test set, the top-performing model is chosen, and hyperparameter tuning is used to further optimise it. The effectiveness and generalisability of the optimised model are next assessed using the test data, and the outcomes are examined.

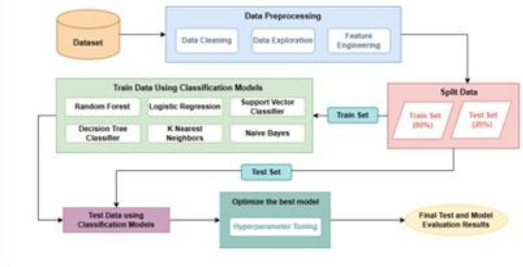

Fig 8: Experimental Workflow

B. Methodology

The purpose of this study is to assess how various feature groups affect model performance and to predict student dropout using machine learning approaches. The dataset comprises 4424 student records that are classified as either enrolled, graduate, or dropout. Enrolled students' records were eliminated to guarantee a binary classification problem because they had not yet arrived at a definitive conclusion. As a result, the dataset was condensed to 3630 instances with the distinct labels of Graduate and Dropout. In order to maintain class proportions, the data was randomly divided into training (80%) and testing (20%) sets.

To improve the quality and suitability of the dataset for machine learning techniques, data preprocessing was done. Missing value instances were eliminated. For categorical data, one-hot encoding was used for nominal variables and label encoding for ordinal properties. Additionally, the StandardScaler was used to standardise all continuous numerical features. This was an important consideration for algorithms that are sensitive to the size of input data, such as K-Nearest Neighbours, Support Vector Machines, and Logistic Regression. Six supervised machine learning classifiers were used: Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN). Performance measures like accuracy, precision, recall, and F1-score were used to assess these models on the test set after they had been trained on the training data.

The dataset was separated into three primary groups—demographic, academic, and economic features—in order to examine the function of various feature categories. "Marital status," "Application mode," "Application order," "Course," "Daytime/evening attendance," "Previous qualification," "Nationality," parental professions and qualifications, "Displaced," "Educational special needs," "Debtor," "Tuition fees up to date," "Gender," "Scholarship holder," "Age at enrolment," and "International" were among the demographic characteristics. Enrolment, evaluation, approval, grades, and credits for both semesters were among the academic aspects that included information about students' achievement in the course. National indicators, specifically the "Unemployment rate," "Inflation rate," and "GDP," were reflected in economic aspects.

Each feature group's contribution was evaluated using a feature exclusion analysis. One of the three experiments' categories—economic, academic, or demographic—was taken out of the training set, but the other two were kept. The updated dataset was used to retrain and test each classifier. A baseline model trained with all features was used to compare the predicted performance in each of these scenarios. This made it possible to determine which feature group had the biggest impact on precise dropout prediction. Lastly, hyperparameter tuning was done so that the model's parameters were set as best they could be for robust performance on unknown data and generalisability.

V. RESULTS AND DISCUSSIONS

A. Model Performance

Using six classifiers—Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, K-Nearest Neighbours (KNN), and Decision Tree—the efficacy of various machine learning methods in predicting student status (Dropout-0 vs. Graduate-2) was evaluated. There were four primary evaluation metrics used to evaluate performance:

- Accuracy: The percentage of all predictions that are accurate.
- Precision: The proportion of correctly predicted positive observations to total predicted positives.
- Recall: The proportion of correctly predicted positive observations to all actual positives.
- F1-Score: The harmonic means of recall and precision, which evenly distributes the two.

Of all the models, Logistic Regression performed the best, with 91.4% accuracy. It showed good capacity to identify graduates and dropouts, with a macro-averaged F1-score of 0.91. With a macro F1-score of 0.89 and an accuracy of 89.9%, SVM also demonstrated competitive performance. Random Forest came in third with a macro F1-score of 0.88 and an accuracy of 89.3%. The performance of KNN, Decision Tree, and Naïve Bayes was moderate. Despite its 85.5% accuracy, Naïve Bayes' precision and recall varied little between courses. Although KNN's accuracy was 85.4%, its recall for the dropout class was lower. With an accuracy of 84.8%, Decision Tree's prediction performance was balanced but relatively poorer as shown in table 1.

The table below provides a summary of each model's specific metrics:

Model	Accuracy	Precision (0/2)		Recall (0/2)	F1-Score (0/2)	Macro F1-Score
Logistic Regression	91.4%	0.92/0.91		0.85/0.96	0.88/0.93	0.91
SVM	89.9%	0.94/0.88		0.79/0.97	0.86/0.92	0.89
Random Forest	89.3%	0.90/0.89		0.81/0.95	0.85/0.92	0.88
Naïve Bayes	85.5%	0.85/0.86		0.75/0.92	0.80/0.89	0.84
KNN	85.4%	0.91/0.83		0.69/0.96	0.78/0.89	0.84
Decision Tree	84.4%	0.80/0.88		0.81/0.87	0.80/0.88	0.84

Table 1: Model Evaluation Metrics

These results show that more probabilistic and instance-based classifiers are not as effective as linear and ensemble-based models, especially Logistic Regression, SVM, and Random Forest, when it comes to the binary categorisation of student outcomes.

Each machine learning classifier's ability to predict student outcomes (Dropout or Graduate) was assessed by plotting Receiver Operating Characteristic (ROC) curves and computing ROC-AUC (Area Under the Curve) scores. AUC values were used to assess the performance of six machine learning classifiers: Random Forest, SVM, K-Nearest Neighbours, Naïve Bayes, Decision Tree, and Logistic Regression. Logistic Regression scored the highest (0.956), with Random Forest (0.951) and SVM (0.948) following closely after. Table 2 shows that Decision Tree received the lowest score (0.840), while KNN and Naïve Bayes also did well.

Model	LR	RF	SVM	KNN	NB	DT
ROC-AUC	0.956	0.951	0.948	0.896	0.885	0.840

Table 2: ROC-AUC Scores with All Features

Models were retrained after academic, demographic, and economic features were removed in order to evaluate the contribution of each feature category. Their crucial importance was confirmed when academic characteristics were removed, since this resulted in the largest performance decline (average fall of 0.145). Random Forest fell to 0.816 and Logistic Regression to 0.810. While economic data had no influence and models maintained constant AUCs, excluding demographic data had a small impact (average reduction of 0.069) (Table 3).

Model	Without Academic	Without Demographic	Without Economic
LR	0.810	0.888	0.956
RF	0.816	0.884	0.951
SVM	0.793	0.884	0.948
KNN	0.751	0.853	0.896
NB	0.770	0.808	0.885
DT	0.665	0.744	0.840

Table 3: ROC-AUC Scores by Feature Group Exclusion

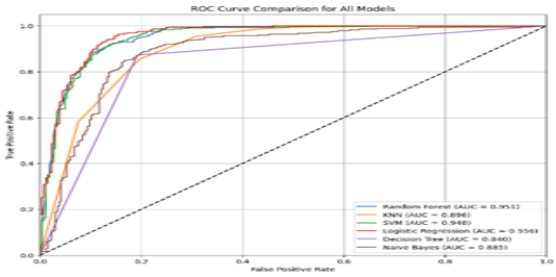

Fig 9: ROC curve graph for all features

These findings demonstrate that while demographic and economic information offer little added value, academic characteristics are the most reliable indicator of student performance.

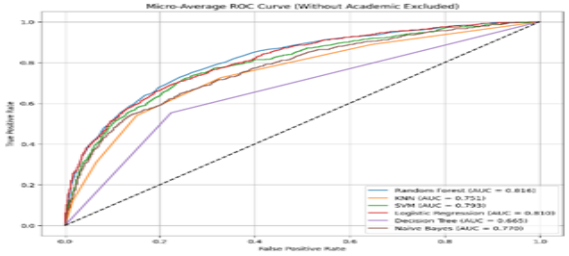

Fig10: ROC curve graph without academic features

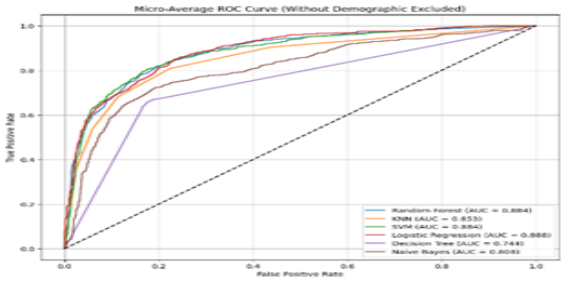

Fig11: ROC curve graph without demographic features

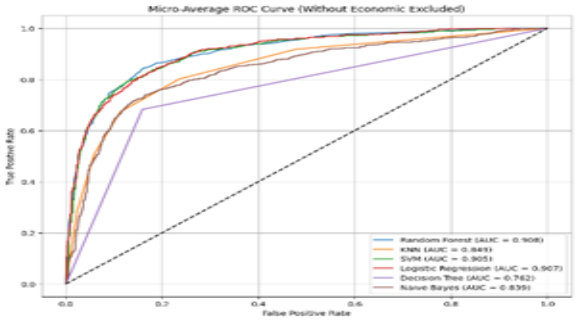

Fig12: ROC curve graph without economic features

B. Feature Importance

To determine which characteristics, have the most effects on student outcomes, feature importance calculations are crucial. We determined the key characteristics that influenced the Random Forest (RF) model, the best-performing classifier from our tests, in terms of its prediction ability. As shown in figure 10, the most significant characteristics have to do with the academic achievement of the pupils. "Curricular units 2nd sem (approved)", "Curricular units 2nd sem (grade)", and "Curricular units 1st sem (approved)" were the top three features, suggesting that student performance in the first and second semesters is a key factor in predicting dropout or graduation. Additional noteworthy aspects included "Curricular units 1st sem (grade)," "Curricular units 2nd sem (evaluations)," and "Age at enrolment," indicating the importance of both academic performance and student background traits. Macroeconomic factors (such as GDP and unemployment rate) and socioeconomic indicators (such as "tuition fees up to date," "father's and mother's occupation,"") also played a significant role in determining academic paths.

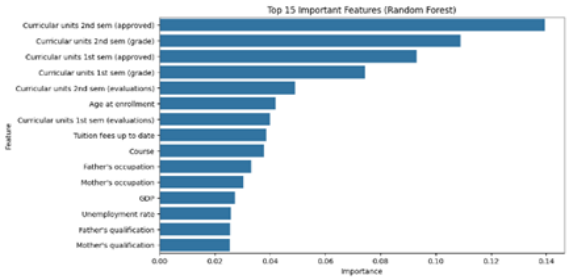

Fig 13: Feature Importance

C. Limitations

The unequal distribution of features across data sources presented challenges for this study. Of the 34 traits, 17 were found in academic data, compared to just 3 in economic data. Since removing academic characteristics resulted in the biggest decline in ROC-AUC scores, this probably distorted model performance. Its under-representation may be the reason for the low impact of eliminating economic data. To lessen prejudice, future research should strive for a more balanced feature collection.

VI. CONCLUSION

This study analysed how well academic, demographic, and economic characteristics predict a student's likelihood of graduating or dropping out. Using a

variety of machine learning classifiers, such as Random Forest, Support Vector Machine, K-Nearest Neighbours, Decision Tree, Logistic Regression, and Naïve Bayes, the study showed that Random Forest and Logistic Regression produced the best accurate predictions. The model's performance was consistently influenced by academic features more than by any other sort of data examined, suggesting that students' academic accomplishments and development are the best predictors of their chances of completing their education.

These results demonstrate the critical role that academic data plays in developing predictive models that effectively forecast student retention. By incorporating these models into institutional systems, at-risk students can be identified early on, enabling prompt interventions and focused support. Although economic and demographic factors also played a role, they had a less substantial effect, confirming that academic achievement was the best indicator of student outcomes. In order to improve forecast accuracy, future studies might address data imbalance and take into account more comprehensive elements like institutional policies and individual situations.

REFERENCES

[1] V. Tinto (1975) – "Dropout from Higher Education: A Theoretical Synthesis of Recent Research". Review of Educational Research, vol. 45, no. 1, pp. 89–125.

[2] R. W. Rumberger (2011) – "Why Students Drop Out of School and What Can Be Done". Cambridge, MA: Harvard University Press.

[3] C. R. Belfield and H. M. Levin (2007) – "The Economic Losses from High School Dropouts in California". Santa Barbara, CA: California Dropout Research Project.

[4] I. Elbouknify, I. Berrada, L. Mekouar, Y. Iraqi, E. H. Bergou, H. Belhabib, Y. Nail, and S. Wardi (2025) – "AI-Based Identification and Support of At-Risk Students: A Case Study of the Moroccan Education System". Benguerir, University.

[5] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West (2016) – "Predicting Student Dropout in Higher Education". arXiv preprint arXiv:1606.06364.

[6] S. Herzog (2005) – "Measuring Determinants of Student Return vs. Dropout/Stopout vs. Transfer: A First-to-Second Year Analysis of New Freshmen". Research in Higher Education, vol. 46, no. 8, pp. 883–928.

[7] E. Aguiar, G. Ambrose, B. Chawla, and D. Brockman (2015) – "Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time". Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (LAK '15), Poughkeepsie, NY, USA, pp. 93–102.

[8] G. Gray, C. McGuinness, and P. Owende (2014) – "An Application of Classification Models to Predict Learner Progression in Tertiary Education". Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, India, pp. 549–554.

[9] S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron (2014) – "Early alert of academically at-risk students: An open-source analytics initiative". J. Learn. Analytics, vol. 1, no. 1, pp. 6–47.

[10] J. W. You (2016) – "Identifying significant indicators using LMS data to predict course achievement in online learning". Internet High. Educ., vol. 29, pp. 23–30.

[11] S. Kim, E. Yoo, and S. Kim (2023) – "Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques". arXiv preprint arXiv:2310.10987.

[12] D. Andrade-Girón, J. Sandivar-Rosas, W. Marín-Rodriguez, E. Susanibar-Ramirez, E. Toro-Dextre, J. Ausejo-Sanchez, H. Villarreal-Torres, and J. Angeles-Morales (2023) – "Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review". IEEE Access, vol. 11, pp. 123456–123470.

[13] K. Sood, A. L. Jimenez Martinez, and R. Mahto (2023) – "Early Detection of At-Risk Students Using Machine Learning". Proceedings of the International Conference on Educational Data Mining, pp. 45–54.

[14] V. Realinho, J. Machado, L. Baptista, and M. V. Martins (2022) - "Predicting student dropout and academic success," Data, vol. 7, no. 11, p. 146.