Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 December; 6 10s(2): 1583-1591

Drop-Out Prediction in Higher Education Using Imbalanced Multiclass Dataset

Juan Antonio Contreras Montes[1], María Claudia Bonfante Rodríguez[2]*, María Andrea Chamorro[1]

[1] Department of Research and Development, Zabud Technologies S.A.S, Cartagena, Colombia
[2] Faculty of Engineering, Universidad del Sinú, Cartagena, Colombia

***Corresponding Author:** - María Claudia Bonfante Rodríguez
*Email: maria.bonfante@unisinu.edu.co

Received: 4-Decembe-2023
Revised: 4-Decembe-2023
Accepted: 4-Decembe-2023
Online First: 4-December-2023

Abstract
Introduction: High quality education has the potential to drive social change, promote equity and alleviate poverty. The prosperity of nations is closely linked to the calibre of their education systems. However, student attrition at the tertiary level poses a major obstacle to mitigating social disparities. While many factors contribute to this phenomenon, harnessing machine learning and data analytics to identify influencing variables and predict potential student dropouts is an effective approach to address this problem.

Objectives: To analyses risk-factors for attrition (drop out) of students at Higher Education Institutions and machine learning algorithms for early detection of such students that could benefit all the stakeholders.

Methods: The study used an unbalanced dataset from a higher education institution to build a classification model to predict academic dropout. The dataset was balanced using oversampling technique and tested using three machine learning algorithms: Random Forest (RF), Support Vector Machines (SVM) and Multinomial Logistic Regression (LR).

Results: The best result was achieved with RF model, with high values of recall, specificity, F1 and balanced accuracy for each of classes: Dropout, Enrolled and Graduate.

Conclusions: A total of 23 features were selected. With 80% of the balanced data, the training of three machine learning models was carried out. For the validation process, the remaining 20% of the data from the original (unbalanced) dataset was used. The results showed a high accuracy in two of the trained models: RM and SVM, with an overall accuracy higher than 0.93.

Keywords: student dropout; student attrition; higher education; significant differences; features selection; machine learning.

1. Introduction

It has been demonstrated that there is a positive correlation between higher education and poverty reduction because higher education expands skills and abilities, creates employment opportunities (Bhandari, 2021), increases lifespan and is a key role in human development (Guo et al, 2022). High quality education could enable the social transformation of countries, increase the equity and reduce the poverty. The social and economic success of the states is directly determined by the quality of their education systems. However, attrition student in higher education is a great obstacle to reduce the social inequality, even more if the dropout is motivated by a financial problem of the student (Ceglédi et al, 2022).
Attrition rate is the unit of measure used to determine the proportion of students leaving the higher education system after their first year. Attrition student can be a result of financial problems (Müller and Klein, 2022), social and personal difficulties, low student-teacher ratio, de-motivating institute environment (orthodox teaching methods, outdated syllabus, lack of student-engagement activity, etc.), poor administrative management, lack of student support, etc., but these factors could differ from one case to another (Ahmad-Tarmizi et al, 2019; Guzman et al, 2021). However, it is necessary to identify between academic failure and withdrawal due to transfer or a temporary leave.
Student attrition and student retention have emerged as a significant and expensive challenge for higher education institutes all around the world. In the last three decades, there has been a significant growth of research published in: 1) identifying the most relevant dropout motives; 2) predicting student attrition. It is very important the early identification (prediction) of those students who are considered to have a higher probability of failing academically or dropping out of

an academic programmer, but the first step is to know the variables (features) influencing student dropout (Nurmalitasari et al, 2023).

2. Related Word

According to Behr et al (2020), the most relevant dropout motives are a lack of interest in the field of study and wrong expectations. It could be due to lack of information about different study programmers, study requirements, organization requirements, job opportunities in specific fields and study alternatives. Morison and Cowley (2017) reviewed studies on the causes of EP (University-based enabling programs) attrition at Australian universities and concluded that the most crucial factor in attrition of students in EPs are: 1) Personal time pressures (employment change, work hours, relationship stress, illness, etc.); 2) An unexpected life event or change in personal circumstances; 3) Lack of use of student support services; 4) Low student engagement.

Raisibe-Mathye (2020) employed various algorithms of machine learning for detecting students into four risk profiles: highest risk profile (19%); high risk profile (with the largest proportion: 34%); medium risk profile (26%); lowest risk profile (21%). He used a dataset with 41 variables and 50.000 observations which included three kinds of attributes: Biographical information; Academic information; and Pre-Schooling attributes. The number of variables were reduced down to 24 after Information Gain Ranking (IGR) Feature Selection Algorithm was applied. The eleven most relevant variables were: English First Lang, English First Additional, Plan Description, Number Of Years for Degree, Race Description, Year Started, Progress outcome, Quintile, Qualified, Plan Code; Home province. Accuracy, Recall, Precision and F1-score were employed as metrics to assess the performance of the algorithms of machine learning: Random Forest; Decision Tree; Support Vector Machines; Multinomial Logistic Regression; and Bayesian Network. The best model was Random Forest and the poor model was Multinomial Logistic Regression.

Raisibe-Mathye (2020), adapting Tinto's framework, defined the following factors: 1) Background of family (include: gender, marital status, age, race description, family background, area and qualifications, living place, language, spoken home language, birth place, home province, race description); 2) Individual attributes (medical conditions, academic literacy, quantitative literacy, mathematical literacy, admittance exam characteristics, absence rates, plan code, tardiness and majors); 3) Pre-college or schooling characteristics (cumulative grade point average, field of study, final marks, university score, course enrollment)

By other hand, early prediction of student at risk (predicting student dropouts) is one of the main major concerns for any academic institution.

Chai and Gibson (2015) developed a student attrition model for predicting which first year students are most susceptible of dropping-out at different points in time during their first semester of study using a dataset of 23.291 students (observations) who started their first semester at Curtin University, Australia, between 2011 - 2013. Four models were evaluated using features that are available at four different time periods in the semester. These models are the pre-enrolment model (17 features: Gender, Birth Country=Australia, Socio economic status, High school / tertiary entrance score, Highest education qualification, etc.), enrolment model (5 features: Age at enrolment, Course credit value, Attendance mode, Field of study, Study load type), In-semester model (2 features: LMS logins, Portal logins) and End of semester model (4 features: Course average, Surveys completed, Units completed, Units withdrawn). However, a data discretization process is included portioning feature values into ranked intervals. Then, each interval is treated as a categorical value. After data discretization, cleansing and preparation, the dataset was limited to 23290 observations, 19222 (83%) retained and 4068 (17%) attrition students, with 148 encoded features. The authors evaluated the following supervised learning methods; logistic regression, decision trees and random forests. Logistic regression and random forests achieved the best precision performance. Finally, logistic regression model was selected due to its relative ease in extracting insights.

Aulck et al (2016) used data gathered from University of Washington (UW) system between 1998 and 2006 with a graduation rate about 76.5%. In this case, student dropout is defined as those students who did not complete at least one undergraduate degree during 6 calendar years of first enrollment. The dataset contains more than 32500 observations. Information about the exact number of features is not showed but the data set include demographic information (Gender, Previous Schooling, Race/Ethnicity, Birth Date, Resident Status, and identification as Hispanic), pre-college entry information (SAT and ACT scores), and complete transcript records (classes taken, time at which they were taken, GPA, and majors declared). The most significant features were: GPA in math, English, chemistry, and psychology courses, year of enrollment and birth year. They reported results from the following machine learning models, based on Accuracy and AUC (Area Under Curve): regularized logistic regression, k-nearest neighbors and random forests. The best results were obtained using regularized logistic regression.

Cherian et al (2020) show the results of a study focused only on the quantitative relationship between entry grades and drop-out rates in UK higher education. They analyzed a data set which cover full-time and part-time undergraduate students using correlation analysis and concluded that here is a relationship between entry grades and the non-continuation of full-time and part-time, young and mature undergraduate entrants.

Sani et al (2020) compared three machine learning algorithms to predict attrition among B40 students in bachelor's degree programs in Malaysia's public universities. B40 corresponds to the 40 percent of Malaysia's people with the lower income. After data pre-processing, 10 features were selected: age, gender, marital status, institution, programmers, study mode (part time, fulltime), sponsorship, qualification (Student's qualification for bachelor's degree admission), cgpa, and class (C=dropout, G=students who manage to graduate). The machine learning algorithms employed were Decision Tree, Random Forest and Artificial Neural Network algorithm, and the best model in predicting student attrition was Random Forest.

Khan et al (2021) used Decision Tree model to predict students at the risk of ending up with unsatisfactory result in a course. The dataset was obtained from students of a course taught at Buraimi University College (Sultanate of Oman) with 151 instances and 11 features: 10 input features and one output feature (prediction class). The input features are: Gender; Attendance; Major (of the student); Year (Year of study); Session (morning or evening); Grade-1_Cont (Grades in exam 1); CGPA (Cumulative GPA); Sponsorship; Dorm (Whether resides in hostel); and PreReq_Grades (Grades in the prerequisite subject). The output feature is Final_Grade (Low, High). The Gain Ratio Attribute Evaluator Filter with Ranker search methods was used as feature selection algorithm and the most significant features selected are: CGPA, PreReq_Grades, Grade-1_Cont and Attendance. They applied different machine learning algorithms: k-NN, Decision Tree, Artificial Neural Networks, and Naïve Bayes, and the best results were obtained with Decision Tree algorithm.

Opazo et al (2021) applied machine learning models to predict first-year engineering student dropout over enrolled students from data collected from two Chilean universities: Adolfo Ibáñez University (UAI) and Talca University (UT). The data collected from UAI (considering only engineering students) contains 3750 observations (instances) and 14 features: gender, place of residence, region of origin, high school data, average high school grades, student ranking according to their school, university application, year (year where the student entered the university), the preference ranking, university admission test scores (include scores for mathematics, language, science or history), average among all tests, and the output feature (Dropout). The considered data from (UT) contains 2201 observations and 17 features. 14 of these features are shared with dataset from UAI and the other three corresponding to: engineering degree that the student enroll to, education of the father, and family income. The results showed that it is better to apply a model per each data set (UAI, UT) instead of combining the datasets into a single dataset. They applied the following machine learning algorithms: K-Nearest Neighbors KNN, Support Vector Machine SVM, Decision Tree, Random Forest, Gradient-Boosting Decision Tree, Naive Bayes, Logistic Regression, and a Neural Network, and Gradient-Boosting Decision Trees reported the best model.

Cuizon (2021) created a model to predict student dropout using data of students admitted in two programs: Bachelor of Science in Information Technology [BSIT] and Bachelor of Science in Computer Systems [BSCS], of the University of San Jose-Recoletos, Cebu City, Philippines. The dataset contains 1100 instances with 8 demographic and academic attributes (average_grade, highschool_type, terms_enrolled, age, proximity, gender, religion, marital_status) and 7 predictors' personality traits (rationality, originality, extroversion, independence, resilience, creativity and assertiveness) gathered through a questionnaire. Author developed a predictive ensemble model from three classification models: support vector machine [SVM], random forest [RF], and k-Nearest Neighbor [k-NN] through plurality voting.

AlHashemi (2021) used different classifications models to predict students' attrition such as Logistic Regression + PCA, Random Forest + PCA, Random Forest without PCA, XGBoost with PCA, and Decision Tree with PCA. The dataset includes 56 attributes with information about: demographics (Age, Gender, etc.), type of the student (Student's High School GPA score, High School from where the student graduated, etc.), courses (Core course # opted in the First semester, Core course # opted in the Second semester, etc.), grades (Grade in Core course # opted in the First semester, Grade in Core course # opted in the Second semester, etc.), parents' information (Father's educational status, Mother's educational status, etc.), and the financial needs of students (Financial need of Student, Course Fees, etc). The Random Forest without PCA model has been reported as the best model.

Niyogisubizo et al (2022) developed a two-layer ensemble machine learning approach to predict student dropout. The first layer include the following models: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB). Then, a Feed-forward Neural Networks (FNN) is fed by the predictions generated in the first layer to obtain the final prediction of student dropout. Cross-validation is used in this second layer. The original dataset used contains 261 samples and 12 features: access, tests, tests_grade, exam, project, project_grade, assignments, result_points, result_grade, graduate, year and acad_year. The dependent variable has two values: 1 (non-dropout) and 0 (dropped out of the university course). The following metrics were employed: accuracy (ACC), Precision (PR), Recall (Rec), the Area Under the Curve (AUC) and F1-Score (F1).

3. Objectives

This study aims to analyse the risk factors for student dropout (dropout) in higher education institutions, making use of machine learning algorithms to generate trained models that can predict cases of student dropout and influence early and targeted decision making to intervene or provide support to students at risk.

4. Methods

Three stages in the solution of the problem were considered:
1. Dataset collection and Data Description
2. Features selection
3. Building the Model

4.1 Dataset collection and Data Description

The dataset employed in this study was created from a higher education institution (Polytechnic University of Portalegre, in Portugal) and built from several disjoint databases. The dataset contains 4424 records of students enrolled in various undergraduate degrees between academic years 2008/2009– to 2018/2019 and includes information at time of student enrollment, such as academic path, socioeconomic, macroeconomic, and demographic factors, and academic performance at the end of the first semester and at the end of the second semester (Martins et al, 2021). The workflow designed to collect data and create the dataset is described by Realinho et al (2022). Also, basic statistics information about the attributes, analysis of multicollinearity, features with correlation coefficient greater than 0.7 between them and the most importance features (obtained using different methods) are showed in Realinho et al (2022) too.
The dependent variable is called “Target” and has tree classes: Graduate (the majority class) represents 49.93% (2209 rows) of the records, Dropout represents 32.12% (1421) of the records and Enrolled (the minority class) represents 17.95% (794) of total records, as showed in figure 1. There is an uneven distribution of observations. Data has been downloaded from: https://www.kaggle.com/datasets/ankanhore545/dropout-or-academic-success.

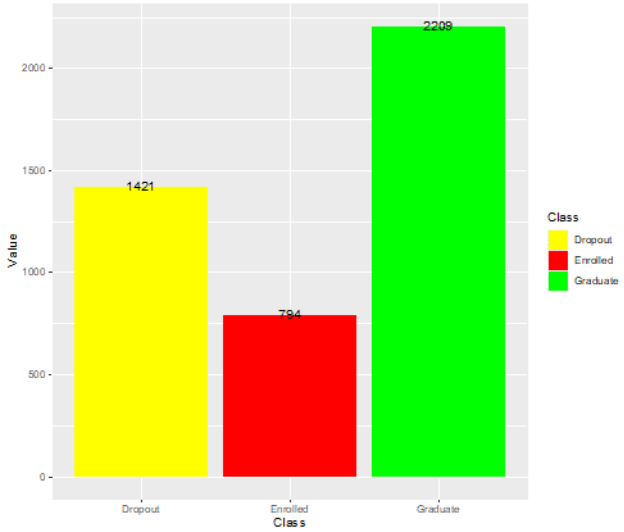

Figure 1. Imbalance between classes

In figure 1, an imbalance is evident among each of the classes. This imbalance affects the accuracy of classification models, as high accuracy is obtained in the majority class, but poor accuracy in the minority class. Due that the classes are unbalanced, in this study the number of samples across the classes were balanced using oversampling methods. The R software, version 4.2.1, and the "upSample" function were used to balance the classes, with the results shown in Figure 2.

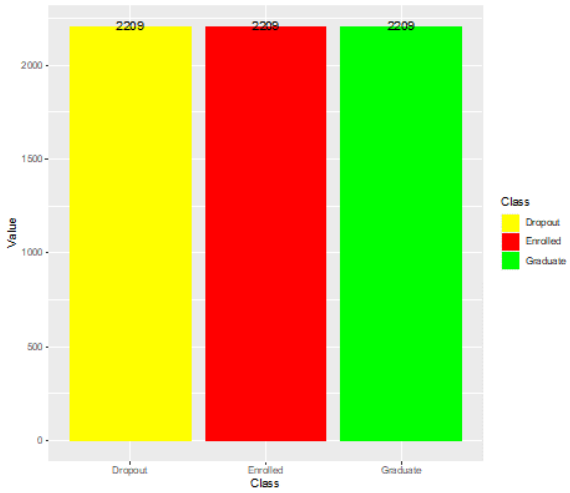

Figure 2. Balanced dataset

4.2 Features selection

Identifying significant differences among classes is important because it allows for a better understanding of the data and how the different features relate to the target variable (i.e., the classes). By identifying the most important features that are significantly different among the classes, it is possible to develop a more accurate model that can accurately predict the class labels for new data. To select features, the procedure consisted of identifying those characteristics that presented significant differences among each of the classes. Previously, tests were performed to determine whether the features had a normal distribution. The Kolmogorov-Smirnov test has been applied and it was observed that features had not a normal distribution. The homogeneity of variance test method used was the Levene test. As a result, 23 features with $p < 0.05$ were extracted from the original 36 features. These features are shown in Table 1.

Table 1. features obtained from Levene test

Feature	F Value	Pr(>F)
Curricular.units.1st.sem..enrolled.	589.272032	1.354031e-227
Curricular.units.2nd.sem..enrolled.	534.878495	9.248452e-209
Curricular.units.1st.sem..without.evaluations.	194.800603	8.344917e-82
Age.at.enrollment	122.989006	1.046073e-52
Mother.s.occupation	117.322425	2.257639e-50
Curricular.units.1st.sem..credited.	93.182434	2.296740e-40
Curricular.units.2nd.sem..credited.	35.489045	5.125291e-16
Curricular.units.2nd.sem..evaluations.	20.185531	1.876375e-09
Application.order	19.880368	2.538983e-09
Marital.status	19.8323	2.663e-09
Application.mode	16.3518	8.407e-08
Curricular.units.1st.sem..evaluations.	11.437319	1.110815e-05
Curricular.units.1st.sem..approved.	9.974542	4.762728e-05
Nacionality	9.736228	6.038020e-05
Mother.s.qualification	8.898423	1.390662e-04
Curricular.units.1st.sem..grade.	8.895548	1.394650e-04
Father.s.occupation	7.979355	3.474158e-04
Previous.qualification	7.004526	9.178726e-04
Father.s.qualification	6.740963	1.193691e-03
Curricular.units.2nd.sem..approved.	6.546757	1.448710e-03
Admission.grade	6.113784	2.230911e-03
Course	4.871673	7.701715e-03
Previous.qualification..grade.	4.378643	1.259681e-02

It is important to mention that in the application of the Levene test, binary type features were not considered.

The tables 2-4 contain statistical measures of the academic attributes that were selected, indicated for each of the three classes (Dropout, Enrolled and Graduate). These measures are mean, median, standard deviation, and maximum and minimum values.

Table 2. Basic statistics information about academic attributes at the enrollment

Feature	Class	Mean	Median	Sd	Min	Max
Admission.grade	Dropout	124.669	123.000	15.124	95	190
	Enrolled	125.208	124.000	13.814	95	190
	Graduate	128.429	127.000	14.077	95	190
Previous.qualification..grade.	Dropout	131.091	133.000	12.869	95	190
	Enrolled	131.200	130.000	12.871	96	190
	Graduate	134.074	133.000	13.342	97	184
Previous.qualification	Dropout	5.311	1.000	10.310	1	43
	Enrolled	4.786	1.000	11.058	1	42
	Graduate	4.031	1.000	9.806	1	43
Application.order	Dropout	1.593	1.000	1.216	1	6
	Enrolled	1.625	1.000	1.214	1	9
	Graduate	1.851	1.000	1.396	0	6

Table 3. Basic statistics information about academic attributes at the end of the first semester

Feature	Class	Mean	Median	Sd	Min	Max
Curricular.units.1st.sem..credited.	Dropout	0.609	0.000	2.105	0	18
	Enrolled	0.508	0.000	1.715	0	14
	Graduate	0.847	0.000	2.686	0	20
Curricular.units.1st.sem..enrolled.	Dropout	5.821	6.000	2.326	0	21
	Enrolled	5.964	6.000	1.988	0	17
	Graduate	6.670	6.000	2.664	0	26
Curricular.units.1st.sem..evaluations.	Dropout	7.752	8.000	4.922	0	31
	Enrolled	9.341	9.000	3.463	0	24
	Graduate	8.277	8.000	3.810	0	45
Curricular.units.1st.sem..approved.	Dropout	2.552	2.000	2.858	0	21
	Enrolled	4.319	5.000	2.289	0	15
	Graduate	6.232	6.000	2.583	0	26
Curricular.units.1st.sem..grade	Dropout	7.052	10.000	5.867	0	18
	Enrolled	10.770	12.000	3.578	0	17
	Graduate	12.213	13.000	2.638	0	18
Curricular.units.1st.sem..without.evaluations.	Dropout	0.192	0.000	0.794	0	8
	Enrolled	0.178	0.000	0.741	0	8
	Graduate	0.088	0.000	0.589	0	12

Table 4. Basic statistics information about academic attributes at the end of the second semester

Feature	Class	Mean	Median	Sd	Min	Max
Curricular.units.2nd.sem..credited.	Dropout	0.450	0.000	1.680	0	16
	Enrolled	0.359	0.000	1.329	0	12
	Graduate	0.667	0.000	2.212	0	19
Curricular.units.2nd.sem..enrolled.	Dropout	5.780	6.000	2.108	0	18
	Enrolled	5.938	6.000	1.831	0	17
	Graduate	6.628	6.000	2.297	0	23
Curricular.units.2nd.sem..evaluations.	Dropout	7.174	7.000	4.817	0	25
	Enrolled	9.436	9.000	3.567	0	28
	Graduate	8.142	8.000	3.246	0	33
Curricular.units.2nd.sem..approved.	Dropout	1.940	0.000	2.574	0	16
	Enrolled	4.058	4.000	2.180	0	12
	Graduate	6.177	6.000	2.269	0	20
Curricular.units.2nd.sem..grade	Dropout	5.740	0.000	5.954	0	17
	Enrolled	10.772	12.000	3.503	0	17
	Graduate	12.280	13.000	2.624	0	18
Curricular.units.2nd.sem..without.evaluations.	Dropout	0.238	0.000	0.994	0	12
	Enrolled	0.188	0.000	0.780	0	8
	Graduate	0.081	0.000	0.523	0	12

In the figure 3 has been plotted a boxplot showing the data distribution of the feature "Curricular.units.1st.sem..grade" in each group: Dropout, Enrolled and Graduate.

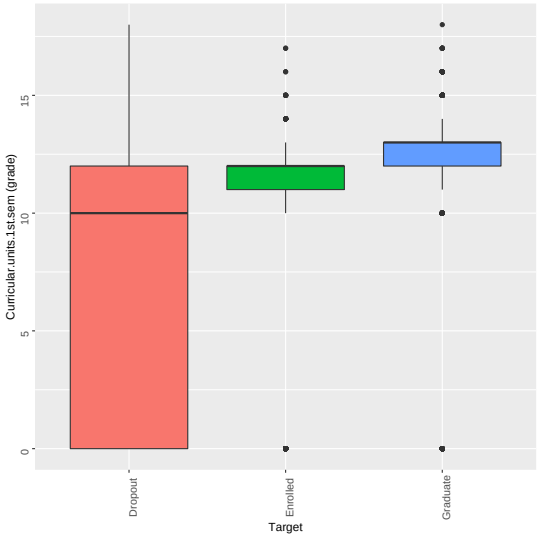

Figure. 3. Boxplot of the feature "Curricular.units.1st.sem..grade".

3. Building the Model

The selected features were used as input variables, or predictors, in the training of the following machine learning algorithms: RM (Random Forest), SVM (Support Vector Machine), LR (Multinomial Logistic Regression). For the training process, 80% of the data was used, after being balanced, using only the selected features. For validation, the remaining 20% of the data was used, with the original class distribution (without balancing).

5. Results

The results obtained for each model are presented in table 5.

Table 5. Evaluation metrics used for evaluating classifier performance.

Model	Class	Sensitivity/ Recall	Specificity	F1	Balanced Accuracy	Accuracy
RF	Dropout	0.9613	0.9850	0.9647	0.9731	0.9672
	Enrolled	0.9620	0.9834	0.9441	0.9727	95% CI :
	Graduate	0.9728	0.9819	0.9772	0.9773	(0.9532, 0.9779)
SVM	Dropout	0.9331	0.9666	0.9315	0.9499	0.9342
	Enrolled	0.9367	0.9738	0.9108	0.9553	95% CI :
	Graduate	0.9342	0.9570	0.9450	0.9456	(0.9159, 0.9497)
LR	Dropout	0.6655	0.9215	0.7269	0.7935	0.7078
	Enrolled	0.5949	0.8000	0.4736	0.6975	95% CI :
	Graduate	0.7755	0.8507	0.8057	0.8131	(0.6766, 0.7376)

The best result was achieved with the Random Forest method, with high values of recall, specificity, F1 and balanced accuracy for each of classes. The confusion matrix is shown in figure 4:

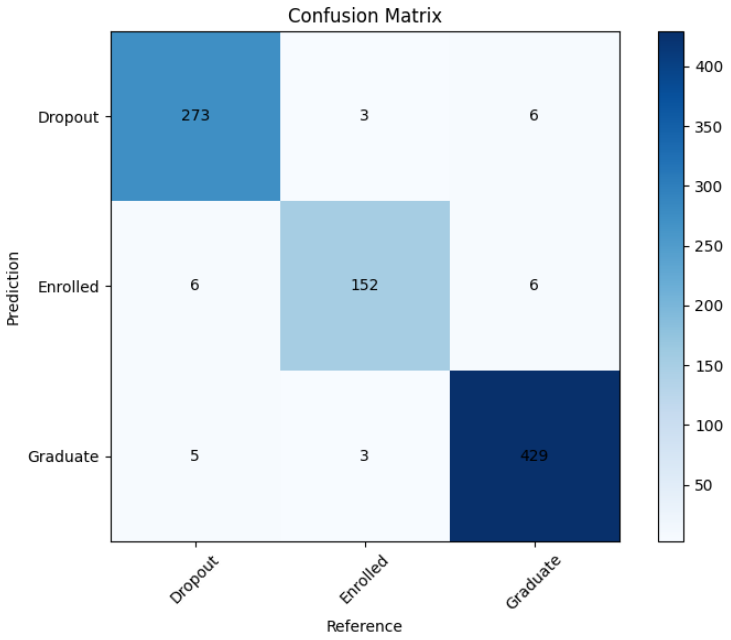

Figure 4. Confusion Matrix for Random Forest Model

6. Conclusion

There are many feature selection algorithms which help to select the most important features for training machine learning ML algorithm. However, the features that prove useful in one ML algorithm may turn out to be less useful in another one. In this study, the features that presented significant differences in each of the three classes using the Levene test were selected. The selected features were used as input variables, or predictors, in training models using supervised machine learning techniques.

Initially, classes were unbalanced: Graduate (the majority class): 2209 samples; Dropout: 1421 samples; and Enrolled: 794 samples. In order to obtain the same number of instances for all the classes, oversampling technique was employed to increase the number of instances in the minority classes by randomly replicating them. As result, all the classes have 2209 samples and 37 attributes (including output variable).

Later, the process of selecting characteristics that presented significant differences between each of the groups or classes was carried out. A total of 23 characteristics were selected. With 80% of the balanced data and 23 characteristics, the training of three machine learning models was carried out. For the validation process, the remaining 20% of the data from the original (imbalanced dataset) dataset was used. The results showed high accuracy for two of the trained models: RM (Random Forest) and SVM (Support Vector Machine), with an overall accuracy greater than 0.93.

References

1. Ahmad-Tarmizi, S.S., Mutalib, S., Abdul-Hamid, N.H., Abdul-Rahman, Sh. (2019). “A Review on Student Attrition in Higher Education Using Big Data Analytics and Data Mining Techniques”. International Journal of Modern Education and Computer Science. Vol. 8, pp. 1-14.
2. AlHashemi, Z. (2021). “Using Prediction ML algorithm for predicting early Student Attrition in Higher Education”. Master Thesis. Department of Graduate Programs & Research. Rochester Institute of Technology RIT, Dubai.
3. Aulck, L., Velagapudi, N., Blumenstock, J., West, J. (2016). “Predicting Student Dropout in Higher Education”. Proceedings of the 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications. New York, June 24, 2016.
4. Behr, A., Giese, M., Teguim, H.D., Theune, K. (2020). “Motives for dropping out from higher education – An analysis of bachelor’s degree students in Germany”, European Journal of Education, Vol. 56, pp. 325–343.
5. Bhandari, R. (2021). “Role of Higher Education in Poverty Reduction A Case Study of Tribhuvan University, Nepal”. Department of Education, Faculty of Educational Sciences, University of Oslo. Master Thesis.
6. Ceglédi, T., Fényes, H., Pusztai, G. (2022). “The Effect of Resilience and Gender on the Persistence of Higher Education Students”. Social Sciences 11: 93. https://doi.org/10.3390/socsci11030093

7. Chai, K.E., Gibson, D. (2015). “Predicting the Risk of Attrition for Undergraduate Students with Time Based Modelling”. Proceedings of the 12th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2015). Maynooth, Greater Dublin, Ireland. October 2015.
8. Cherian, J., Jacob, J., Qureshi, R., Gaikar, V. (2020). “Relationship between Entry Grades and Attrition
9. Trends in the Context of Higher Education: Implication for Open Innovation of Education Policy”. Journal of Open Innovation: Technology, Market, and Complexity. Vo. 6, No. 199; doi:10.3390/joitmc6040199.
10. Cuizon, J.C. (2021). “Ensemble Predictive Model for Academic Churn Risk Using Plurality Voting”. Mindanao Journal of Science and Technology. Vol. 19, No. 1, pp. 224-235.
11. Guo, T., Bai, X., Tian, X., Firmin, S., Xia, F.(2022). “Educational Anomaly Analytics: Features, Methods, and Challenges”. Frontiers in Big Data, Vol. 4, pp. 1-16.
12. Guzmán, A., Barragán, S., Vitery, F.C. (2021). “Dropout in Rural Higher Education: A Systematic Review”. Frontiers in Education. 6:727833, doi: 10.3389/feduc.2021.727833
13. Khan, I., Ahmad, A.R., Jabeur, N., Mahdi, M.N. (2021). “An artificial intelligence approach to monitor student performance and devise preventive measures”. Smart Learning Environments. Vol. 8, No. 17, pp.1-18.
14. Martins, M.V., Tolledo, D., Machado, J., Baptista, L. M.T., Realinho, V. (2021) "Early prediction of student’s performance in higher education: a case study" Trends and Applications in Information Systems and Technologies, vol.1, in Advances in Intelligent Systems and Computing series. Springer. DOI: 10.1007/978-3-030-72657-7_16 This dataset is supported by program SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal.
15. Morison, A., Cowley, K. (2017). “An exploration of factors associated with student attrition and success in enabling programs”. Issues in Educational Research, Vol. 27, No. 2, pp. 330-346.
16. Müller, L., Klein, D. (2022). “Social Inequality in Dropout from Higher Education in Germany. Towards Combining the Student Integration Model and Rational Choice Theory”. Research in Higher Education https://doi.org/10.1007/s11162-022-09703-w.
17. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., Nshimyumukiza, P.C. (2022). “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization”. Computers and Education: Artificial Intelligence. Vol. 3. Article 100066, pp. 1-12.
18. Nurmalitasari, Long, Z.A., Noor, M,F.M. (2023). “Factors Influencing Dropout Students in Higher Education”. Education Research International, https://doi.org/10.1155/2023/7704142.
19. Opazo, D., Moreno, S., Álvarez-Miranda, E., Pereira, J. (2021). “Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities”. Mathematics, Vol. 9, No. 20, pp. 1-27.
20. Raisibe-Mathye, M. (2020). “A Theoretical Model to Predict Undergraduate Attrition Based on Background And Enrollment Characteristics”. Master Thesis. School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg.
21. Realinho, V., Machado, J., Baptista, L., Martins, M.V. (2022). “Predicting Student Dropout and Academic Success”. Data, Vol. 7, No. 146. https://doi.org/10.3390/data7110146
22. Sani, N.S., Mohamed Nafuri, A.M., Othman, Z.A., Ahmad Nazri, M.Z., Mohamad, K.N. (2021). “Drop-Out Prediction in Higher Education Among B40 Students”. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, pp 550-559
23. Wan Yaacob, W.F., Sobri, N.M., Md Nasir, S.A., Norshahidi, N.D., Wan Husin, W.Z. (2020). “Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques”. Journal of Physics: Conference Series. 1496 012005