Application of Logistic Regression Technique for Predicting Student Dropout

Berat Ujkani, Daniela Minkovska, Lyudmila Stoyanova
Department of Programming and Computer Technologies,
Faculty of Computer Systems and Technologies, Technical University of Sofia
8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria
{bujkani, daniela, lstoyanova}@tu-sofia.bg

Abstract – **Student dropout in higher education is a complex issue and as a process it includes many factors which may affect each other. This paper explores the use and application of a probabilistic supervised machine learning technique for predicting university student dropout to obtain insights on the students at risk and prevent them from dropping out the studies. Data from a public university in the Republic of Kosovo were obtained and examined. The dataset comprises instances of students' dropouts for the past six academic years along with their demographics, grades, enrollment details etc. Logistic Regression, as one of the most widely used Machine Learning and Artificial Intelligence algorithms, was used to build the model and produce the predictions. First, a statistical analysis was conducted and after the data preprocessing, logistic regression classifier was implemented. The results show that a high prediction accuracy was reached, with a percentage of 90%, and a F1 score of 0.85, indicating that the model is performing great and the predictions results are reliable.**

Keywords – **student dropout, higher education, machine learning, logistic regression, numpy, scikit-learn**

I. Introduction

Student dropout is a significant issue for many societies and universities around the globe and as a phenomenon has been continuously addressed by the scientific community. Successful studies are also crucial for students themselves. The data show that there is a large number of dropouts in almost every country, while there are many reasons why students may not graduate. For instance, only in the first weeks of studies, almost 75% of all dropout rates occur [1].

The dropout rate among university students is a problem that causes negative consequences to the individual, the institution, and society at large. Furthermore, the high dropout rate is a signal that the quality of education needs improvement [1]. While there are a plethora of other consequences and reasons for student dropout, one of the ways researchers have to study the phenomenon of dropout among students is by trying to predict the dropout rate so appropriate preventive measures and strategies can be developed and applied in time.

The analysis of the dropout rate dates back almost fifty years ago to the model developed by [2]. Since then, many models and techniques have been developed and gained much attention in trying to figure out the factors and relevant reasons why students dropout [3].

Nowadays, Artificial Intelligence powered techniques are being used more and more in education and contributing to transforming many aspects of higher education. A lot of universities are leveraging Artificial Intelligence and trying to benefit from unlocking various Machine Learning techniques, particularly in helping them to forecast enrollment, predict educational outcomes as well as identify students that might dropout, thus contributing to improving the quality and developing prevention strategies that help in student retention [4].

The importance of pursuing a study in analyzing the dropout rate remains significant in many countries. No studies have been conducted in Kosovo about the dropout rates, the reasons, and the factors that make students leave the university. This study is the first of its kind in the context of Kosovo, which analyses the dropout rates and presents a simple but useful model for predicting the dropout in the future using a popular machine learning method. The paper uses as a case study the data from a public university in Kosovo, but the results might be useful for all higher education institutions and other relevant policymakers.

II. Related Work

The mission of a university is to ensure qualitative education that is intrinsically linked to the development of a country or society. Moreover, universities must ensure equal opportunities of access and completion of studies for all enrolled students [5]. However, the dropout phenomenon at university levels is concerning for all actors involved in the education sector and causes negative consequences in terms of time, money, or other unexploited resources [6].

The high percentage of students dropping tertiary education may signify to the labor market that highly qualified individuals will not be easily found in the next few decades. On the other hand, high dropout rates signify that university teaching is not providing quality education and from the student's perspective it is considered a personal failure [7].

Identifying the students at risk and predicting dropout rate in higher education has proven to be an attractive study topic for researchers and thus, various studies have been conducted throughout the years to analyze and predict student success and dropout phenomenon.

Machine learning is an excellent approach to tackle this problem, and it has been explored in several papers to date. L. Kemper et al. in [8] created machine learning models to predict dropouts at the Karlsruhe Institute of Technology using Decision Trees and Logistic Regression. By feeding the models data from the transcript of records, they were able to calculate the dropout probability on a case-by-case basis with an accuracy of up to 95% after three semesters or over 83% after only one semester. One of the key features in the

978-1-6654-9878-4/22/$31.00 ©2022 IEEE

2022 XXXI International Scientific Conference Electronics (ET) | 978-1-6654-9878-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ET55967.2022.9920280

dataset for dropout prediction was the combination of the average and total number of completed and failed exams or average grades.

Another work with a similar approach has been published by [9]. Gradient Boosting, Random Forest, Support Vector Machine and Ensemble models were used for predicting dropouts at different stages of university studies – before enrollment and at the end of each semester for the first two years. In addition to academic data, the models also took into consideration students' personal data, their family and environment, and their interactivity with online learning tools. According to the results, it's possible to detect dropouts with an accuracy of 82.91% by the end of the first semester. The precision increases to 91.5% by the end of the fourth semester.

A comparable dropout prediction accuracy between 77 and 93% was also achieved by [10]. The dataset contained data from four academic years and was used to train classification models such as Logistic Regression, Decision Tree, Random Forest, Naive Bayes and SVM. The five features of the dataset included number of course views, assignment scores, tests, examinations and projects. As a result, the models could be trusted to predict the dropout probability by the end of only one semester with a satisfactory precision.

J. Niyogisubizo et al. in [11] published a paper on the student dropout topic using the ensemble machine learning approach. Similarly to the previous papers, the dataset consisted mainly of access, examination, assignment, project, and test results. The combination of Random Forest, Extreme Gradient Boosting, Gradient Boosting, and Feed-forward Neural Networks resulted in a training accuracy range from 86.67% to 96.67% and a testing accuracy range from 76.67% to 92.18%. Such results could be helpful for the university to identify and prevent potential students from dropping out.

III. Methodology

A common methodology in building machine learning models is composed of two phases: first being concerned with data gathering, statistical analysis and data pre-processing, while the second phase dealing with the real implementation of logistic regression technique to build the model and predict the results. The same exact methodology was decided to be followed. One important thing to note here is that we refer to dropout as leaving the higher education system entirely without being awarded with a degree as defined by [1] as the term can be understood in different ways. For instance, a student can change the field of study, the type of university, or can leave the university entirely.

A. Data

The study uses data extracted from the student's management system of an academic unit of a public university in the Republic of Kosovo. The dataset includes data from six academic years, from the academic year 2015/2016 up to 2020/2021 with attributes related to students, starting from the demographic information, gender, age, enrollment date, etc. The total number of students data gathered was 4,818. 121 records were removed from the dataset when the data cleaning process was employed. The final dataset included 4,697 records.

Table 1 shows the final dataset, containing nine attributes, which are gender, age, city, enrollment date, level of study, number of exams to be taken, exam passed, average grade, current status - whether student is active, inactive, graduated or has dropped out. Since we chose to apply a binary classifier, an extra field named class was added with two possible values, 0 for graduate and 1 for dropout status. This field was also set as the feature or predictable attribute.

TABLE 1. DATASET ATTRIBUTES, TYPES AND POSSIBLE VALUES

Attribute	*Type*	*Value*
ID	Numeric	Any number
City	Nominal	Any string
Department	Nominal	0 = Computer Science and Engineering 1 = Production Engineering 2 = Economics Engineering
Enrollment	Date	1/10/2015-30/9/2021
Level of study	Nominal	0 = Bachelor 1 = Master
Age	Numeric	18-35
Gender	Nominal	0 = Male 1 = Female
Exams Number	Numeric	15 - 30
Passed Exams	Numeric	1 - 30
Average Grade	Numeric	6 - 10
Status	Nominal	0 = Graduate 1 = Active 2 = Inactive
Class	Nominal	0 = Dropout 1 = No dropout

Before developing the machine learning model, training and testing the data, a statistical analysis based on extracted data was conducted. It is of great importance to figure out the key statistics in university dropout, in particular understand for instance dropout based on gender, dropout in relation with the number of exams passed, etc. In the figure below are represented the dropouts over the academic years 2015-2021 divided by gender.

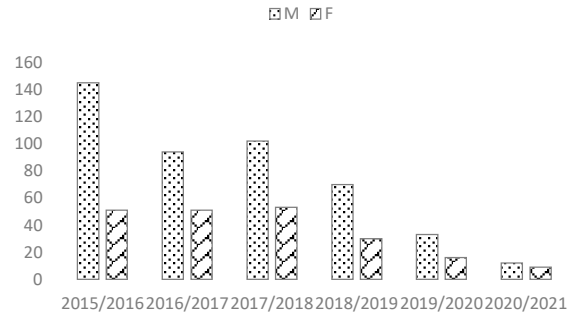

Fig. 1. Retention and dropout rates by cohort

The figure shows that the number of female students that dropped out during the 2017/2018 academic year was the highest, while the greatest difference in dropout and gender can be noticed in the academic year 2015/2016. Similarly, in the table below are represented the data related to dropout for the same academic years in conjunction with the number of successful exams passed.

TABLE 2. DROPOUT NUMBER PER COHORT

	Exams Passed		
Academic Year	*<5*	*5-10*	*10+*
2015/2016	86	70	40
2016/2017	88	37	20
2017/2018	122	22	11
2018/2019	33	3	13
2019/2020	10	7	4

From the figure it can be seen that the largest number of student dropout happened to those that passed less than five exams, with the academic year 2017/2018 being the period where students have dropped out the most. The overall calculated average percentage of dropout is 23,74%.

B. Model

The next step in conducting the research was to perform the prediction using a machine learning method based on the available dataset. We recall that the goal of the prediction is to create a model based on the class attribute that attempts to predict dropout or successful completion of studies in the future. This is a typical classification problem where a binary classifier should be implemented in order to predict whether a student is expected to dropout or not. While there are many different machine learning classification methods which can be used in order to build a suitable predictive model, for the purpose of this study where the dataset has a categorical output, Logistic Regression technique was chosen among other available supervised machine learning techniques.

Logistic Regression is a supervised machine learning method that allows classification of data points according to two categories and predicting probability distributions as opposed to discrete values in linear regression. Basically, logistic regression is an extension to linear regression.

The task is to determine whether a student will dropout or not based on the data related to the number of exams passed and the average grade. We decided to implement Logistic Regression technique using Python and its popular libraries: NumPy and Scikit-Learn. Firstly, a set of steps were taken to load the data and split them so that we could train and evaluate properly. The dataset was divided into two parts: 70% for training and 30% for model testing, to better understand the model performance. Labels were encoded into integers so 0 to represent students who have dropped out and 1 to represent students that have not dropped out. Students who have active or inactive status were out of the scope for the purpose of this study. All these were easily achieved by using scikit-learn built-in functions. The goal is to learn a logistic model $\hat{y}$ that models y given X.

$$\hat{y} = \frac{e^{XW_y}}{\sum_j e^{XW}} \tag{1}$$

where $\hat{y}$ are the predictions, X the inputs and W the weights.

The results in the form of prediction gained need to be compared with the target values using cross-entropy loss objective function.

$$J(\theta) = -\sum_i \ln(\hat{y}_i) = -\sum_i \ln\left(\frac{e^{X_i W_y}}{\sum_j e^{X_i W}}\right) \tag{2}$$

Next, gradient of loss $J(\theta)$ is calculated with regards to the model weights W.

$$\frac{\partial J}{\partial W_j} = \frac{\partial J}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial W_j} = -\frac{1}{\hat{y}}\frac{\partial \hat{y}}{\partial W_j} =$$

$$= -\frac{1}{\frac{e^{XW_y}}{\sum_j e^{XW}}}\frac{\sum_j e^{XW} e^{XW_y} 0 - e^{XW_y} e^{XW_j} X}{\left(\sum_j e^{XW}\right)^2} = \frac{X e^{XW_j}}{\sum_j e^{XW}} = X\hat{y} \tag{3}$$

$$\frac{\partial J}{\partial W_y} = \frac{\partial J}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial W_y} = -\frac{1}{\hat{y}}\frac{\partial \hat{y}}{\partial W_y} =$$

$$= -\frac{1}{\frac{e^{XW_y}}{\sum_j e^{XW}}}\frac{\sum_j e^{XW} e^{XW_y} X - e^{W_y X} e^{XW_y} X}{\left(\sum_j e^{XW}\right)^2} = \frac{1}{\hat{y}}(X\hat{y} - X\hat{y}^2) =$$

$$X(\hat{y} - 1) \tag{4}$$

And finally, the weights W are updated so they set a higher probability for the correct class prediction (y) and discourage the probabilities for incorrect classes (i).

C. Evaluation

An important step when running a prediction in machine learning is the evaluation process and checking whether the used technique is performing in a proper way [12]. To measure the efficiency and performance of the model, we decided to use the following metrics: confusion matrix, accuracy, recall, precision and F_1 metric. A short explanation of each of the metrics as in [12] is given in the lines below.

True positive (TP) rate represents the number of correctly predicted samples, while false positive (FP) rate represents the number of false positives, and they are defined as below.

$$TP = \frac{TP}{TP+FN} \tag{5}$$

$$FP = \frac{FP}{FP+TN} \tag{6}$$

Accuracy represents the overall precision of prediction and is calculated as the proportion between the correctly classified predictions and the total number of cases.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Recall focuses only on correctly classified predictions. It measures the probability that a student dropout is classified correctly out of all correctly positive and negative classifications.

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

The conditional probability that a student who is classified as a dropout is accurately classified is measured using the precision metric. This measure is also very useful to avoid incorrect predictions.

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

The last evaluation measure is F_1 score which finds the balanced measures of the performance of the model. It gives 0, when the model basically doesn't work up to 1, when the model works perfectly.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (10)$$

Observing the evaluation metrics in table 3, we can see the correctly classified values in the diagonal and the proposed model achieved an accuracy of 90%.

TABLE 3. EVALUATION METRICS

Dropout	*Yes*	*No*
Yes	1003	112
No	358	3224

	Accuracy	*Precision*	*Recall*	*F1*
%	0.90	0.89	0.81	0.85

IV. DISCUSSION

The research aimed at predicting the university dropout based on previous historical data of students who graduated respectively left the university entirely. Results show that a prediction accuracy has been reached with a value of 90%. and a recall value of 81 %. Looking at F1 metric, we can see that a score of 0.85 was achieved, which indicates that the model is performing good and the predicted results are reliable. In addition, comparing the model based on the number of exams passed, the results show that the prediction accuracy increases as more data are loaded.

The statistical results show that the overall percentage of dropout from the case study is 23%, which might not be a high dropout rate but since the size of the data is growing more features can be selected in the future, especially in considering the main reasons and motivation why students are leaving the studies and predicting the dropout rate.

V. CONCLUSION

This paper presented an application of logistic regression technique to predict whether a student will retain or dropout the studies based on data such gender, age, city, enrollment date, level of study, number of exams to be taken, exam passed, average grade, current status - whether student is active, inactive, graduated or has dropped out.

The data from six academic years were used to train and test the model built on the well-known supervised machine learning technique of Logistic Regression. The model was implemented in Python programming language using NumPy and Scikit-Learn libraries.

In the end, a high prediction accuracy was reached, with a percentage of 90%, and a F1 score of 0.85, indicating that the model is performing good and the predictions results are reliable.

ACKNOWLEDGMENT

The authors would like to thank the Research and Development Sector at the Technical University of Sofia for the financial support.

REFERENCES

[1] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A review and future research directions," IEEE Transactions on Learning Technologies, vol. 12, no. 3, pp. 384–401, Jul. 2019.

[2] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," Review of Educational Research, vol. 45, no. 1, pp. 89–125, 1975.

[3] R. Baker and G. Siemens, "Educational data mining and learning analytics," The Cambridge Handbook of the Learning Sciences, pp. 253–272, Sep. 2014.

[4] A. Behr, M. Giese, H. D. Teguim K, and K. Theune, "Early prediction of university dropouts – A random forest approach," Jahrbücher für Nationalökonomie und Statistik, vol. 240, no. 6, pp. 743–789, Feb. 2020.

[5] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting students drop out: a case study," Proceedings of the 2nd International Conference on Educational Data Mining (EDM2009), pp. 41–50, 2009.

[6] J. Luan, "Data mining and its applications in higher education," New Directions for Institutional Research, vol. 2002, no. 113, pp. 17–36, 2002.

[7] L. Shields, A. Newman, and D. Satz, "Equality of educational opportunity," plato.stanford.edu, May 2017, [Online].

[8] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," European Journal of Higher Education, vol. 10, no. 1, pp. 28–47, Jan. 2020, doi: 10.1080/21568235.2020.1718520.

[9] A. J. Fernandez-Garcia, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sanchez-Figueroa, "A real-life machine learning experience for predicting university dropout at different stages using academic data," IEEE Access, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/access.2021.3115851.

[10] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," Applied Sciences, vol. 11, no. 7, p. 3130, Apr. 2021, doi: 10.3390/app11073130.

[11] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," Computers and Education: Artificial Intelligence, vol. 3, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[12] B. Ujkani, D. Minkovska, and L. Stoyanova, "A machine learning approach for predicting student enrollment in the university," 2021 XXX International Scientific Conference Electronics (ET), Sep. 2021.