IRSTI 55.01.01

Diana Bairamova[1]
[1]Suleyman Demirel University, Kaskelen, Kazakhstan

DATA COLLECTION TO IDENTIFY STUDENTS AT RISK OF NOT COMPLETING A COURSE USING MACHINE LEARNING

Abstract. One of the most important methods in the study of various subjects is the understanding at an early stage of the learning process on the part of both the teacher and the student that the student is in a risk group that will not complete the course successfully. Identifying this group of students at an early stage of learning increases the level of motivation of students to start studying well in time and can help the teacher individually determine which student needs help. Before identifying a group of students at risk of not completing the course successfully, an important part is to collect and prepare the necessary data (predictors) for teaching machine learning algorithms. Currently, this is necessary for both online and offline education. In the presented method of determining a group of students, various types of algorithms were used, where one of the best results of determining a group of students with risk and without risk was shown by Logistic Regression with a high AUC =0.8003. The SMOTE method was used in the work, which coped well with the problem of data imbalance of the "Pass" and "Not Pass" classes, while increasing the accuracy of the forecast for the minority class "Not Pass" by 11%. Using certain predictors of student performance, it is possible to derive additional information such as the level of interest in the lesson, the determination of the final score for the lesson, a certain category (A, B, C, D) of students with different characteristics and other indicators that contribute to the involvement of students in the lesson at the earliest stage of learning.

Keywords: Machine learning, student's "at risk" prediction, significant predictors, Academic Performance Categories, SDV.

***

Аңдатпа. Әртүрлі пәндерді оқудағы ең маңызды әдістердің бірі- оқытушы тарапынан да, студент тарапынан да оқу процесінің бастапқы кезеңіде студенттің курсты сәтті аяқтай алмау мүмкіндіңгінің бар екенін түсіну. Оқушылардың бұл тобын алдын-ала анықтау оқытру процессіне деген ынтасын жоғарлатуға және оқутышыға қай оқушының көмекке мұқтаж екенін жеке анықтауға көмектеседі. Курсты сәтсіз аяқтау қаупі бар студенттер тобын анықтамас бұрын, маңызды бөлігі Машиналық оқыту алгоритмдерін оқыту үшін қажетті деректерді (болжаушыларды) жинау және дайындау болып табылады. Қазіргі уақытта бұл онлайн және офлайн білім беру үшін қажет. Ұсынылған студенттер тобын анықтау әдісі алгоритмдердің әртүрлі түрлерін қолданды, мұнда тәуекелі бар және тәуекелсіз студенттер тобын анықтаудың ең жақсы нәтижелерінің бірі жоғары AUC=0,8003 Логистикалық Регрессия әдісін көрсетті. Жұмыста SMOTE әдісі қолданып "Pass" және "Not pass" топтарының деректер теңгерімсіздігі мәселесін жақсы шешіп, "Not pass" азшылық тобы үшін болжам дәлдігін 11% арттырды. Оқушылардың үлгерімінің белгілі бір болжаушыларын қолдана отырып, сабаққа деген қызығушылық деңгейі, сабақтың қорытынды бағасын анықтау, әртүрлі сипаттамалары бар оқушылардың белгілі бір санаты (A, B, C, D) және оқушылардың сабаққа қатысуына ықпал ететін басқа көрсеткіштер сияқты қосымша ақпараттар алуға болады.

Түйін сөздер: Машиналық оқыту, "тәуекел" тобындағы студенттердің болжамы, Академиялық үлгерім санаттары, SDV.

***

Аннотация. Одним из наиболее важных методов при изучении различных предметов является понимание на ранней стадии процесса обучения как со стороны преподавателя, так и со стороны студента того, что студент находится в группе риска, которая не завершит курс успешно. Выявление этой группы учащихся на ранней стадии обучения повышает уровень мотивации учащихся к тому, чтобы вовремя начать хорошо учиться, и может помочь учителю индивидуально определить, какой ученик нуждается в помощи. Прежде чем определить группу студентов, подверженных риску неудачного завершения курса, важной частью является сбор и подготовка необходимых данных (предикторов) для обучения алгоритмов машинного обучения. В настоящее время это необходимо как для онлайн, так и для офлайн образования. В представленном способе определения группы студентов использовались различные типы алгоритмов, где один из наилучших результатов определения группы студентов с риском и без риска показал метод Логистической Регрессии с высоким AUC =0,8003. В работе был использован метод SMOTE, который хорошо справился с проблемой дисбаланса данных классов "Pass" и "Not Pass", увеличив при этом точность прогноза для класса меньшинства на 11%. Используя определенные предикторы успеваемости учащихся, можно получить дополнительную информацию, такую как уровень интереса к уроку, определение итоговой оценки за урок, определенная категория (A, B, C, D) учащихся с различными характеристиками и другие показатели, способствующие вовлечению учащихся на уроке на самом раннем этапе обучения.

Ключевые слова: Машинное обучение, прогноз студентов в группе “риска”, значимые предикторы, категории академической успеваемости, SDV.

I. Introduction

Currently, there is an assumption among students of current generations that at the beginning of their studies they may not put enough effort into a good and successful study of subjects and that in any case they will be able to study by the end of the semester, raise their grades by the end of studying the subject, while not allocating enough time to study the subject from the beginning of the learning process. Most of these students do not understand how they are at risk of not finishing their subject properly, after which the student either has to drop out due to lack of motivation in their achievements at the learning process, or start studying the subject anew while spending valuable resources like time and other equally important resources.

By using Neural Networks, scientist Cameron I. Cooper (January 2022), identified groups of students at risk of not completing the course [1]. Earlier, from 2007 to 2014, via meta-analysis [2], Watson, Li, Christopher and Frederick W. B. (2014) determined that 1st-year students entering the specialty Information Systems or Computer Science have very low academic performance in the subject “CSS101 - Introduction to Computer Programming”. Their studies show that only 67% of students complete this course successfully.

Giving a student the opportunity to see information about whether they are at risk of students who do not complete the course successfully from the very beginning of the subject and tracking their progress throughout the course can be a good motivation to start studying in time and not by the end of the course, while students can try to improve their academic performance before the end of the course. This opportunity will also be a hint for the teacher which students most do not understand the subject or do not show sufficient activity. On the other hand, it will be economically advantageous [3] from the point of view, since the authors of Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock and Jevin West (January 20, 2016) of one study note how important it is to understand whether a student belongs to a risk group, since a large number of students drop out in the first years of study, and this also effects on economic expenditures of the state budget.

Nowadays, when the consequences of the pandemic have had a great impact on the change in the format of education (from offline education to the online), this problem in determining students at risk of not completing the course successfully concerns not only teaching subjects offline at a university or college, but also lessons translated into an online format, also for online courses it is much more difficult for a teacher to understand which students need help in the learning.

To identify students from the risk group, real data on the academic performance of North American University students collected by scientists A. Mubayed, M. Injadat, A. Shami and H. Lutfiya (March 2020) [4] were used as an initial data. The data contains various types of information about the academic performance of both online and offline formats of education. In addition to real data, thanks to the use of one of the well-known methods of data augmentation - SDV (Synthetic Data Vault) [5], it was possible to increase the sample (dataset) to train machine learning models from 486 rows of data on student performance to 5486 with a fairly high 93% score of augmented data.

To determine the group of students with risk and without risk, such well-known algorithms as Logistic Regression, Neural Networks, Decision Tree, Bagging Classifier, k nearest neighbors Method (KNN), Support Vector Method, Naive Bayesian Classifier with cross-validation and hyperparameter tuning were used.

In addition to all available predictors, the following characteristics were identified, such as the category of academic performance (A,B,C,D) and the level of student activity (low and high) in their studies, which have a positive impact on understanding the overall picture of student performance and in addition to the fact that it is possible to identify groups of students with risk and without risk, thanks to the derived additional parameters, it is possible to fully describe the involvement of students in various subjects.

II. Literature review

Earlier, in a study of detecting students who may not complete the course successfully, scientist Cameron I. Cooper (January 2022), uses Neural Networks, discovered whether a student is included in a group of students at risk of failing or not failing the course [1]. However, the scientist took in their research only one subject - "Introduction to Computer Programming", since only 67% of students successfully completed the course in this subject in the period from 2007 to 2014 at colleges and universities. This was revealed using meta-analysis by scientists Watson, Li, Christopher and Frederick W. B. (2014). The scientists in the study propose an alert system that can improve students' academic progress [2]. Researchers collect the data across 7 years of study only the subject "Introduction to Computer Programming" and they collect about 592 rows of data [2] to train and test data.

In the study researcher try different 25 types of Neural Networks and choose the PNN (Probabilistic Neural Network) with the higher accuracy and after that he grow the accuracy by using backward elimination and choose most important inputs [1]. By using Sensitivity Analysis researcher found the most important periods [1] which is help to instructors in time correct the situation and help to improve academic achievements of students.

Cameron I. Cooper show the results of their research where students have increased the success rate of the course by 23% using the use of alert systems [1].

In his study, Erkan Er (August 2012) about identification at-risk students only for online courses used only time-varying data that were variable over time, that is, when train his model, he used only such data as class attendance, midterm grades, and so on, but the researcher does not use such data as age, gender, since these data do not determine which risk group the student belongs to [6].

The author [6] mentions in his research that he used 3 stages for training data as 3 different stages of semester training. In the first stage, the author takes attendance for 4 lessons and an assessment for the first task. The second stage includes attendance for 8 weeks and evaluation for the first and second tasks. In the last stage, the author uses the attendance of 10 weeks and three grades for assignments as well as the overall grade of the midterm exam [6]. The author does not take all the attendance grades for the initial data, because he is sure that the teacher may not conduct lessons after 10 weeks of the learning.

After the author has divided his dataset into 3 stages [6], he then uses various types of algorithms and trains the model at all stages separately. In addition, to improve his results, the author uses a special technique proven earlier by scientists I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Lumos [7] which show that using the combined result of different algorithms gives better results. Thus, the author uses three schemes. In the first scheme, a student belongs to a risk group even if only one algorithm assigns him to a risk group. In the second scheme, a student is assigned to a risk group if at least two algorithms show that this student belongs to this group. And the last scheme 3 says that if all three algorithms determine a student as from a risk group [7], only in this case he will belong to this risk group, otherwise this student belongs to the successful group.

In his results, the author shows that the use of scheme 2 [6], that is, more than one algorithm leads to an improvement in the learning outcomes of the model.

Researchers Yujing Chan, Aditya Johri, Huzefa Rangwala use data on student learning for the period from 2009 to 2013 for analysis [8], their results show that the average GPA score and when the student was enrolled in the learning is an important factor for learning the model, in addition, the age of students shows that the older generation is aimed at finishing their studies while younger people can easily drop out [8].

III. Method and Materials

1.1 Data Collection

Based on the collected data on the academic performance of North American University students of the course in natural sciences by researchers [4] compiled a list of predictors (18) containing 486 lines, which contains information in addition to the scores of quizzes, assignments and midterm results, but also data such as the number of visits to the education department of the system, assignments delay, deadlines for completing tasks and others (see Table 1).

Table 1. Description of the collected data (predictors)

№	**Predictor Name**	**Description**
1	Student ID	Student identifier (will not be used for train models)
2	Visits	The number of student visits to the learning platform during the course
3	Study of materials	The number of times the course material has been studied on the platform
4	Reading notifications on the platform	How many times has a student read notifications on the platform from a teacher
5	Discussion Participation on the platform	How many times has a student written comments to notifications from a teacher on the platform
6	Revision of the quiz	The number of times the quiz has been sent on the platform
7	Quiz	Evaluation of the quiz
8	Assignment 1 delay	Shows 0 if the student missed the deadline for completing task 1 and 1 if he was not late
9	Assignment 1 completion	Hours of completion of the first task
10	Assignment 1	Evaluation of the assignment 1
11	Midterm	Evaluation of the midterm exam
12	Assignment 2 delay	Shows 0 if the student missed the deadline for completing task 2 and 1 if he was not late
13	Assignment 2 completion	Hours of completion of the second task
14	Assignment 2	Evaluation of the assignment 2
15	Assignment 3 delay	Shows 0 if the student missed the deadline for completing task 3 and 1 if he was not late
№	**Predictor Name**	**Description**
16	Assignment 3 completion	Hours of completion of the third task
17	Assignment 3	Evaluation of the assignment 3
18	Assignment assignments completion time	The average duration of execution in hours of all completed assignments

The above data in Table 1 were used in the study as predictors to identify groups of students at risk of completing and not completing a successfully taken course. As predicted data, in addition to the main purpose of determining the group of students with risk and without risk, the final assessment of the student, the student's level of interest in studying the subject, the student's academic performance category (A, B, C, D) shown in the (Figure 1), which determined from two additional derived features (see Table 2):

Table 2. Description of the target data

№	**Target Value Name**	**Description**
1	Pass/Not Pass	Predicting course completion or course failure
2	Final exam grade	Based on all previous predictors of academic performance, the prediction of the final exam score
3	Academic Performance Category	Prediction which of the **4** categories does the student belong to: 1) A **-** the student is strongly involved in the lesson and copes very well with the tasks 2) B **-** the student is involved in the lesson, performs all tasks but makes mistakes

		3) C - the student is not involved in the lesson, but performs tasks 4) D - the student is not involved in the lesson, does not perform all tasks and makes mistakes
4	Activity level	Level of interest in the course: 1) Low level 2) High level

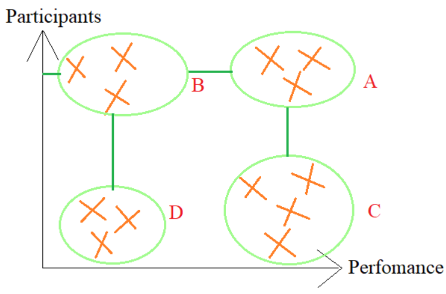

Figure 1. Academic Performance Category Detection

Each category of academic performance (Figure 1) was determined from two features as Performance (quiz scores and all assignments score) and Participants (determined from the predictors such as Visits, Study of materials, Reading notifications, Participation in the discussion). In addition to the category of academic performance, a feature was derived as "Activity level". This feature can also be predicted by having information on the number of visits to lessons, reading all posts, participating in discussions, studying materials (see Table 1). If this overall indicator is lower than the average according to these metrics, then the student's activity is defined as "Low" and if the score is higher than the weighted average, then the students are considered as active - "High".

1.2 Data Augmentation

After collecting all the necessary indicators for a complete presentation of information about the progress of students, the method of artificial data augmentation based on real data was applied.

In this research, the method of generating synthetic data was used - SDV (Synthetic Data Vault). This technique is one of their most effective methods of generating real data that can be replaced with generated data [5]. As a result, from 486 rows of data on student performance, 5486 results were obtained with higher accuracy of the generated data - 93%, which is an excellent indicator for the using data in the study. The difference (Figure 2) between the actually compiled data and those generated data, where we see that the generated data almost completely describe the real data.

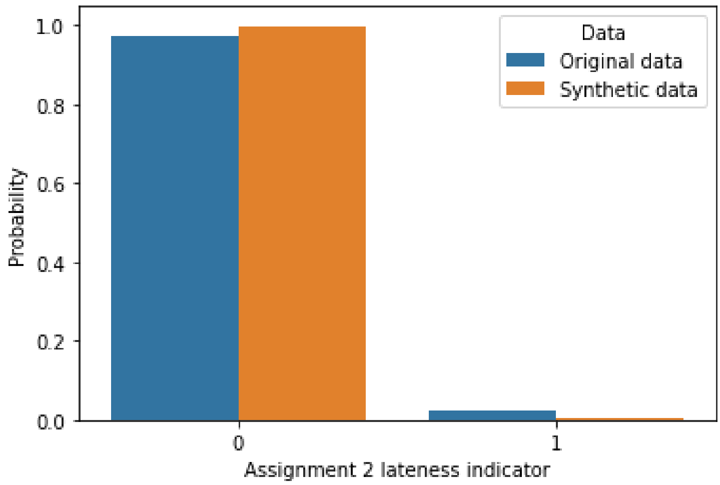

Figure 2. Original and Synthetic data compare of Assignment 1 lateness indicator

1.3 Data Description and Analyzing

Before conducting experiments with various algorithms and after collecting all the necessary data, as well as after adding the generated data, all the collected data were analyzed, where we can see the distributions of already existing data. From the (Figure 3) see the distributions of the "Pass and Not Pass" students. In general, in the collected list of data, almost 79% of students successfully completed the course, while the remaining 21% are those who could not successfully finished class.

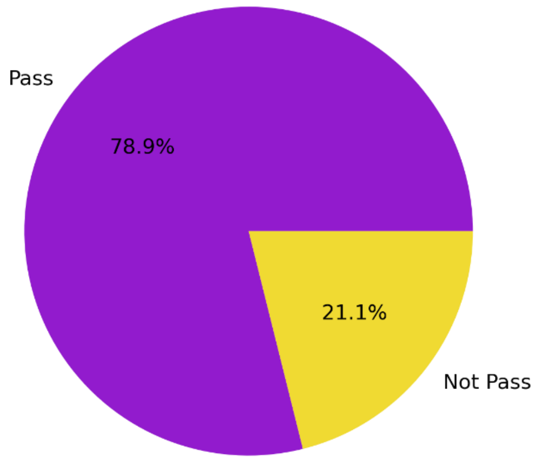

Figure 3. Distributions of the "Pass/ Not Pass" students.

This distribution of data (Figure 3) suggests that the algorithm models will work well with the dominant (majority) group, in this case it is a group of students who have completed the course well. For this reason, in this study, such methods as SMOTE and Near Miss [9] were used in the training of various algorithms (Logistic Regression, Neural Networks, KNN and others). These two methods work well with the problem of unbalance in the data.

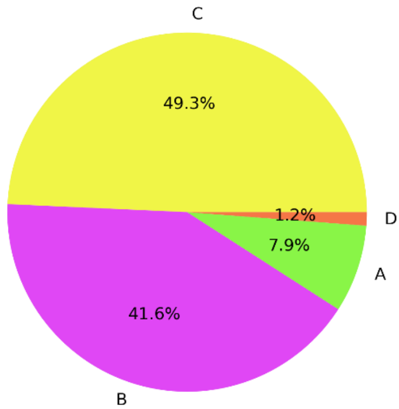

Figure 4. Data distributions of the “Activity Performance Category”.

The largest portion of students is occupied by students who do not participate in the lesson, but perform tasks - group C (49%), as well as those students who are involved in the lesson but make mistakes when performing tasks - this is group B with a portion (42%) and the smallest part are those students who do nothing at all and are not involved in the lesson - category D (1%).

In the context of the activity level indicator, students who actively behave in relation to the study of the subject very well receive higher grades (>= 80) more than those who do not show interest in learning process (see Figure 5). Basically, those who show no interest in learning get the most score of 60 on the final exam (see Figure 5).

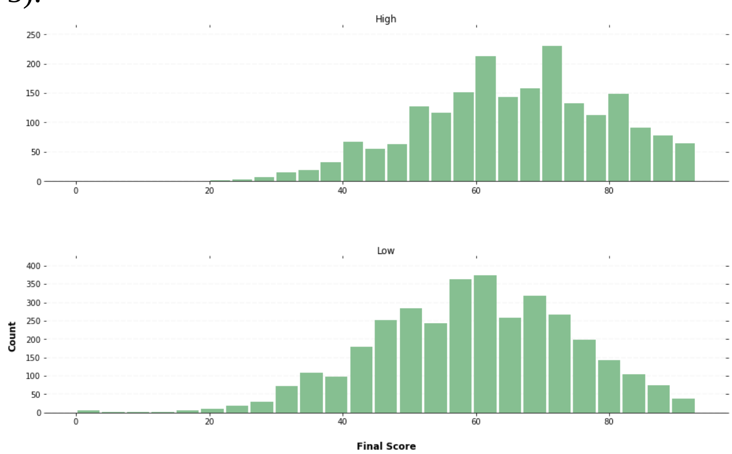

Figure 5. Final score estimates by “Activity level” indicator

1.4 Data Preprocessing

One of the most important points before starting the training of various algorithms is the method of preprocessing and data preparation. This allows you to improve the results of training models for the highest accuracy of the forecast. In this study, some of the main methods of data preprocessing were used. Firstly, all missing values were replaced with the averages of these predictors, in addition, some unnecessary columns, such as the Student ID, were removed, and after removing the unnecessary column, the Min-Max Scaler method was used, which was necessary for data normalization.

The method is used to scale the data, and using the min-max scaling method, the data has been normalized. Also, as far as preprocessing is concerned, the data can have highly correlated characteristics, which means that if one function (predictors) increases or decreases, the other function also increases or decreases. This can be illustrated using a correlation matrix (Figure 6), the characteristics of which can be correlated.

The graph illustrates the correlation between all available predictors (Visits, Study of materials, Reading notifications on the platform, Discussion Participation on the platform, Revision of the quiz, Quiz, Assignment 1 delay, Assignment 1 completion, Assignment, Midterm, Assignment 2 delay, Assignment 2 completion, Assignment 2, Assignment 3 delay, Assignment 3 completion, Assignment 3, Assignment assignments completion time).

From the figure 6, predictors such as 'Assignment_1_completion hours', 'Assignment_2_ completion hours', 'Assignment_3_completion_hours' correlate very strongly with each other, with a predictor like 'Average_assignments_completion_time'. This suggests that it is enough to leave one of the most significant predictors and it describes all these previous listed predictors.

After determining the most correlated data in the study, the correlation method in the library (matplotlib - .corr() function) was used. The method revealed the 3 most correlated predictors. This method made it possible to remove unnecessary predictors. As a result, after using correlation method (.corr), only one predictor out of three predictors about the performance of the assignment delay.

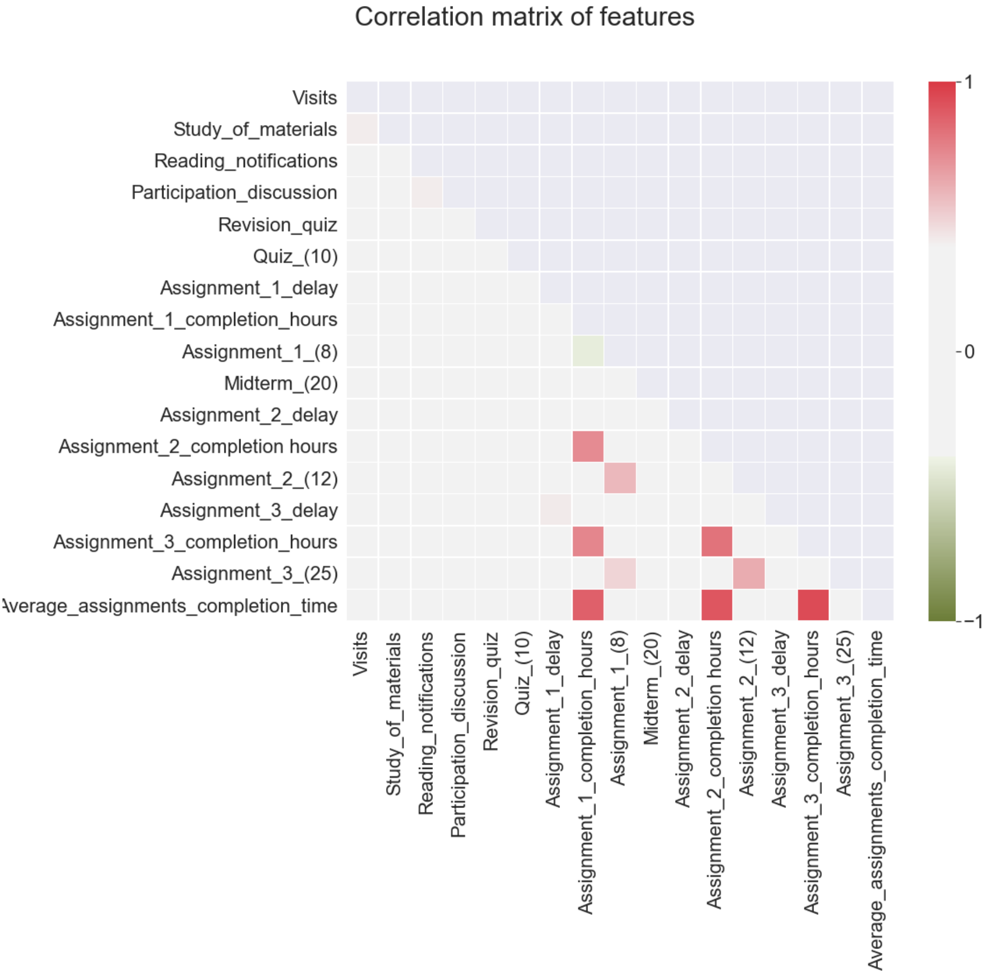

Figure 6. Correlation Matrix of the all features

1.5 Train data process

After normalization of the available data by dividing the data into training and test data using the method train_test_split by X (predictors values from the first column "Visit" up to the last "Assignment_3") and Y (target value) - "Pass/Not Pass".

In the study, various algorithms were used to determine the group of students with the risk of not completing the course successfully and without the risk of completing the course successfully (classification task), first of all, a Logistic model with classical settings was used, the accuracy of which was 82%. As a result, (see Figure 7), the Logistic model shows on the classification report that basically the model is well trained on the majority class - "Pass" with a forecast accuracy of 89%, while the forecast accuracy of the model on the "Not Pass" class is only 45%, which is a very low indicator.

	precision	recall	f1-score	support
Not Pass	0.65	0.31	0.42	290
Pass	0.84	0.96	0.89	1082
accuracy			0.82	1372
macro avg	0.74	0.63	0.65	1372
weighted avg	0.80	0.82	0.79	1372

Figure 7. Logistic Regression Classification Report results

This indicates the imbalance of the available data and this attracted the fact that it was necessary to distribute the data in such a way that any model that is used in the future trained well not only for the dominant group (Pass) but also for a smaller group ("Not Pass").

For this situation, two methods of solving this problem have been tested - the SMOTE method. This method allows you to equate data in such a way that the class that is a minority will be synthesized from already existing data and thereby equated to the class with the majority [9]. The results of this method gave good results in relation to the minority class, the prediction accuracy increased by 11% in relation to the minority class, and the level of the Recall indicator for the minority class also increased from 31% to 73%.

The next method to solve the problem of data imbalance is the NearMiss method, which is the opposite of the SMOTE method and works the other way around, instead of increasing the set of the missing class, the NearMiss method reduces the size of the minority class and thereby equating this class with the majority class.

As a result, after using the NearMiss method, the accuracy of predicting the minority class "Not Pass" also increased by 11% and became equal to 53%, however, compared to the SMOTE method, the Recall indicator describing how well and accurately the algorithm works showed 71% accuracy. This NearMiss method is less by 2% of Recall indicator compared to the SMOTE method. As a result, it was the SMOTE method that was used in all subsequent machine learning algorithms to classify students with and without risk.

An important part for the successful prediction of groups of students is the setting of hyperparameters and the use of cross validation for accurate prediction. The study used K-fold cross validation with 30 folds, which showed 81% as the average accuracy of the forecast. To configure hyperparameters, one of the well-known and effective methods of configuring hyperparameters, known as - GridSearchCV, was used. This method allows automatically select the necessary settings for the best performance of the algorithm model [10].

IV. Results

In total, 7 machine learning algorithms (see Table 3) were used in the study, which showed different results in predicting the accuracy of students at risk of not completing the course and completing the course:

1. Logistic Regression
2. MLPClassifier
3. DecisionTreeClassifier
4. BaggingClassifier
5. KNN
6. SVM Classifier
7. Naïve Bayes

As a result, the most effective of the 7 taken algorithms for predicting a group with a risk of not completing the course successfully and completing successfully is the Logistic Regression algorithm, which, compared to other algorithms, works well on a smaller "Not Pass" class with 30 - fold cross validation and configured hypeparameters with a 53% accuracy of correct prediction, also for the majority class "Pass" with a prediction accuracy of 81% (see Table 3). Despite the fact that the Bagging classification method showed good results for the majority class with an accuracy of up to 84%, however, for the minority class, model predicts only 47% correctly, which is 6% less than the Logistic Regression model.

Table 3. Classification Algorithms Results

Class	**Not Pass**			**Pass**		
Name	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	42%	72%	53%	91%	73%	81%
MLPClassifier	37%	78%	51%	92%	65%	76%
DecisionTreeClassifier	34%	52%	41%	85%	73%	78%
BaggingClassifier	43%	53%	47%	87%	81%	84%
KNN	37%	74%	50%	85%	71%	75%
SVM Classifier	43%	53%	47%	90%	71%	79%
Naïve Bayes	42%	72%	50%	86%	71%	79%

The results of the Logistic Regression model showed the results of predicting students at risk to finish successfully and not successfully finish the course with an AUC score of 0.8003, which is a high prediction result (see Figure 8).

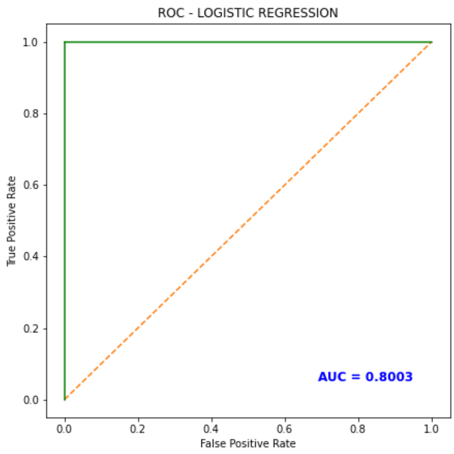

Figure 8. ROC AUC Curve for Logistic Regression model

V. Conclusion

This study is to determine at an earlier stage which group of students each of the students of a course participants belongs to - those who pass the course and those who do not finish the course successfully (Not Pass). In addition to the main task in determining the risk group to finish and not to finish successfully the course is to predict the final grade as well as the category - (A, B, C, D) derived from existing predictors, as well as determining the group with high and low activity for the lesson, but these tasks will be considered later.

The main objective of this study was to collect and prepare data for analysis and for further predictions of a group of students at risk of not completing the course and completing the course successfully. The data base was compiled earlier by scientists A. Mubayed, M. Injadat, A. Shami and H. Lutfiya (March 2020) from a North American university, where real data with online and offline lesson formats were located. In total, they collected 486 rows of data [4] were compiled about each student with various characteristics such as attendance at lessons, assessment of various tasks and quizzes, the number of readings of posts, participation in discussions, and so on. Further, an additional data augmentation method was used in the study – SDV (Synthetic Data Vault), which allowed generating an already existing list of data from 486 rows to 5486 rows of student data. The generated data list showed a good result of 93% accuracy of the generated data.

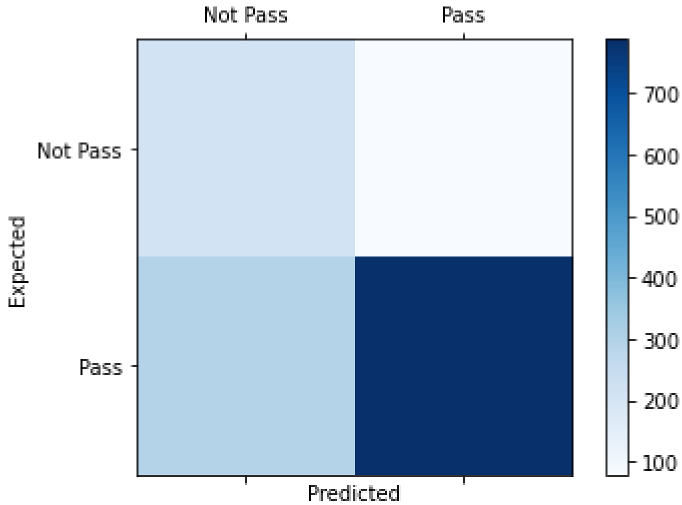

Figure 9. Confusion Matrix of Logistic Regression model

After full work on data collection and preparation, various types of algorithms for the prediction of the main task - determining the group of students with risk and without risk. In addition to working with algorithms and setting up hyperparameters using the GridSearchCV method, work was carried out on unbalanced data, since the majority class was students with the "Pass" group of 79%, the algorithms did not work well on the minority "Not Pass" group. To do this, the SMOTE and Near Miss methods [9] were used, which showed SMOTE works best, namely, it increased the prediction level of the minority group by 11%.

As a result of all the work done, the best and most efficient Logistic Regression algorithm (see Figure 9) was determined with a prediction accuracy of 73% and an AUC of 0.8003, which predicts the "Not Pass" group by 11% more accurately compared to other algorithms. In the future, the continuation of this study will be working with the accuracy of the algorithm, as well as with the prediction of additional signs such as student activity level of the lesson, the category of student performance and the prediction of the final grade for the course.

References

1 Cameron I. Cooper “Using Machine Learning to Identify At-risk Students in an Introductory Programming Course at a Two-year Public College“. Advances in Artificial Intelligence and Machine Learning; Research 2 (2) 407-421, (July 2022).

2 Watson, Christopher and Li, Frederick W. B. (2014) 'Failure rates in introductory programming revisited.', in Proceedings of the 2014 conference on Innovation & technology in computer science education

(ITiCSE '14). New York: Association for Computing Machinery (ACM), pp. 39-44.

3 Lovenoor Aulck, Nishant Velagapudi, JJoshua Blumenstock and Jevin West “Predicting Student Dropout in Higher Education” (20 Jan 2016).

4 A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, “Student Engagement Level in an e-Learning Environment: Clustering Using K-means”, American Journal of Distance Education, 34:2, pp. 137-156, Mar. 2020

5 Markus Endres , Asha Mannarapotta Venugopal , Tung Son Tran “Synthetic Data Generation: A Comparative Study” , pp. 94-102, August 2022

6 Erkan Er, “Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100”, International Journal of Machine Learning and Computing, Vol. 2, No. 4, (August 2012).

7 I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, 2009. “Dropout prediction in e-learning courses through the combination of machine learning techniques,” Computers & Education, vol. 53, no. 3, pp. 950-965, November 2009

8 Yujing Chan, Aditya Johri, Huzefa Rangwala “Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early”, (March 2018)

9 Ahmet Okan Arik, Çiğdem Çavdaroğlu “A Novel Method Based on Smote and Near Miss Methods for Intrusion Detection in Imbalanced Data Sets”, SSRN Electronic Journal, January 2022

10 Agus Minarno, Wahyu Kusuma “Human activity recognition utilizing SVM algorithm with Gridsearch”, ResearchGate, AIP Conference Proceedings, July 2022