# USING MACHINE LEARNING TO ANALYZE UNIVERSITY STUDENTS' DROPOUT RATE – A CASE STUDY

*Veselina Nedeva, Tanya Pehlivanova*

Trakia University, Faculty of Technics and Technologies - Yambol
e-mails: veselina.nedeva@trakia-uni.bg; tanya.pehlivanova@trakia-uni.bg
Bulgaria

**Abstract:** The article presents case study from Trakia University – Stara Zagora. A study was conducted with students from the professional field "Electrical Engineering, Electronics and Automation". Socio-demographic indicators, educational factors, attitude towards the university, etc. were studied. The survey data were analyzed using machine learning. The aim of the article is to select appropriate machine learning classification algorithms for analyzing the reasons for dropping out of university.

**Key words:** Machine learning, BayesNet, NaiveBayes, OneR, JRip, WEKA, dropping out of university

## 1. INTRODUCTION

With the continuous development of technology, the need for highly educated specialists with higher education is increasing. One of Europe 2020 strategy's targets is that at least 40% of 30-34-year-olds in the EU should have completed tertiary education by 2020. According to Eurostat data in 2019 in the EU as a whole, this target has been achieved. However, not all Member States have achieved their national targets. One of these countries is Bulgaria. One way to increase the number of people who have successfully completed higher education is to reduce the dropout rate of students.

The study success is an important topic on the European policy agenda. The HEDOCE study found that study success is regarded as important in three quarters of the 35 European countries surveyed. In almost half of the countries it is high or very high on the policy agenda [1]. Bulgaria is among the countries where the solution to this problem is little relevance. There are few studies of Bulgarian authors related to this issue.

Due to the demographic crisis and the steadily increasing admission in Bulgarian universities, there are students in the universities who are not always sufficiently motivated and with sufficient prior preparation. This is even more true

for "unattractive" technical majors. The stress for first-year students is very high. At Faculty Technics and Technologies (FTT), Trakia University – Stara Zagora, Bulgaria, about 44% of the admitted students complete their studies within four years for a bachelor degree. For some majors and forms of training, this percentage is lower. Timely predicting the risk of students dropping out makes it possible to take measures to retain them.

The aim of the article is to select and to apply appropriate machine learning classification algorithms to analyze the reasons for students dropping out of university. This is not only a personal problem, but also causes social impact in society and the country's economy. We plan to continue the study for all majors in Faculty Technics and Technologies - Yambol.

This article is organized as follows: Section 1 is Introduction. A Literature review is provided in Section 2. Section 3 Methods includes Data collection, Data preprocessing, Classification algorithms. Section 4 Results and discussion includes Confusion matrix, Accuracy percentage, True Positive and Precision, Feature selection. The article ends with conclusions and plans for future research.

## 2. RELATED WORK

There are many studies in the literature related to the success of the students and predicting students dropping out using machine learning. The most commonly used classifiers are Decision tree (DT), Random Forest (RF), Support Vector Machine (SVM), k-NN and Bayes classifiers.

Various student information is used - academic achievement, personal and pre-university characteristics, social and demographic factors, attitude towards the university and others.

There is only one study available for a Bulgarian university [2], which aims to verify whether students' performance at the university can be predicted on the basis of their personal and pre-university characteristics. The characteristics, in the available data, which have the greatest impact on student success, are identified. The results obtained by two rule learners, a decision tree classifier, two popular Bayes classifiers and a Nearest Neighbor classifier (J48, Naive Bayes, Bayes Net, k-NN, Jrip, OneR) are compared. The results show that the decision tree classifier (J48) performs best, followed by the rule learner (JRip) and the k-NN classifier. The Bayes classifiers are less accurate than the others.

Student dropout prediction on Time Series Data was made in [3, 4]. Data for the subjects studied, the time of passing the exam, the number of attempts, the grades obtained and more are used. The classifiers used are SVM and k-NN that uses the Dynamic Time Warping Distance DTWD.

Comparative analysis of available methods, Naive Bayes, Random forest, Gradient Boosting Tree, Support Vector Machines, k-NN with Dynamic Warping Distance, based on a proposed classification of strategies for predicting dropout,

and an experiment based on a dataset of student information acquired through an automated management system from a university in Brazil was made in [5].

In [6] was presented an analysis of different machine learning techniques, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and Random Forest (RF), applied to the task of dropout occurrences prediction for university students. The analysis has been conducted using data available at the moment of the enrolment unlike most approaches that use data on University careers of students. Accuracy (ACC), specificity (SPEC), and sensitivity (SENS) were used to compare the classifiers.

An early student dropout predicting system was developed in [7] using administrative data for students. The AdaBoost Algorithm to combine regression analysis, neural networks, and decision trees — instead of only one specific method was used. Performance of the AdaBoost was evaluated with the indicators accuracy, recall, correctly and incorrectly predicted cases.

Article [8] analyzes the correlation between demographic indicators and academic performance to predict student dropout using three single classifiers, K-Nearest Neighbor (k-NN), Naïve Bayes (NB) and Decision Tree (DT). Much better results are obtained when combining algorithms with Ensemble Classifier Methods using Gradient Boosting as a meta-classifier. Accuracy is best at 10-fold cross-validation.

In [9], three machine learning techniques are applied and compared - Decision Tree (DT), k-Nearest Neighbor (k-NN), and Random Forest (RF) to analyze and predict student performance. The DT, k-NN, and RF methods are chosen because they perform well with incomplete data. The most accurate prediction results are obtained with Random Forest and k-Nearest Neighbors machine learning techniques.

After the analysis of the literature, we decided that the machine learning (ML) classification algorithms, BayesNet, NaiveBayes, OneR and Jrip are most suitable for our study. The processing of the results is done with WEKA open source software.

## 3. METHODS

### 3.1. Data collection

The data were collected through a survey of students from FTT Yambol at Trakia University - Stara Zagora. The questionnaire contains 32 questions. It was developed after conducting interviews with current and graduate students. The experience of the authors, which they have acquired over more than 25 years of teaching activity, is also reflected. The preparation of the questionnaire takes into account the results of the literature reviewed which studies and analyzes the reasons for students dropping out of university education. The questionnaire covers three groups of questions (personal characteristics, academic environment, and

social conditions) which, according to the authors, will allow us to find the answer to the questions and to solve the problems related to students dropping out of education.

The study covers students from two majors in the professional field Electrical Engineering, Electronics and Automation – "Automation and Computer Systems" and "Electrical Engineering". The general population includes 280 students, and the statistical sample covers the number of 115 students surveyed, ie. 41,07 %. The survey was conducted with 91 students still doing their studies and 24 students who had stopped their studies for various reasons. It was conducted in the period from 10.10.2019 to 01.20.2020. The choice of students is random. The questionnaire is offered to students from I to IV course.

### 3.2. Data Pre-processing

The data collected cannot be used until they have been properly prepared for processing and analysis by Weka. "The raw data obtained does not give any information in the form it appears. The raw data stored could contain errors due to multiple reasons like, missing data, inconsistencies that arise due to merging data, incorrect data entry procedures, and so on" [10]. "Deriving meaningful information from the raw data requires preprocessing of data that converts real-time data into computer-readable format." [11]

Data pre-processing involved several procedures that can be represented as follows: Data accumulation; Data cleaning; Data transformation.

The final data set used to execute the project contains 115 instances, each described with 23 attributes with nominal variables. The attributes, their description and the values used in the analysis are presented in Table 1.

*Table 1. Attributes, description and values*

| Attributes | Description and Values |
|---|---|
| 1.AGE | Age<br>*{19-20, 21-23, 24-26, 27-29, 30-32}* |
| 3B.COURSE | Course/Year of Education<br>*{'First Year', 'Second Year', 'Third Year', 'Fourth Year'}* |
| 4.MAR_STAT | Marital status of students<br>*{Single,Married,Other}* |
| 5.CHILD | Children of the student<br>*{'Have not children','Have children'}* |
| 6.JOB | Does the student work in the specialty<br>*{'yes, in the other specialty','Yes, in the specialty','I do not work'}* |
| 7.SATISF | Job satisfaction<br>*{'neither yes nor no ','rather no ',yes,'rather yes',no,'no opinion'}* |

| Attributes | Description and Values |
|---|---|
| 8.INCOME | Family incomes<br>*{medium,'low ',high,'very low ','very high'}* |
| 9.PLACE | Place of residence<br>*{'small town', 'big city', Sofia, village}* |
| 10.EDU_PARENT | Parental education<br>*{'higher education','higher education - secondary education', 'secondary education', 'secondary education-primary education' ,'primary education' ,'higher education-primary education '}* |
| 12.PERS_STRESS | Personal stress<br>*{'yes - financial' ,'have not stress', 'problems with colleagues and learning problems', 'yes - illness', 'yes - problems with teacher', 'yes - learning problems', 'yes - change in personal purpose', 'yes - other type', 'yes - change in personal purpose and problems with colleagues', 'yes - financial yes - learning problems yes - problems with teacher', 'yes - illness yes - learning problems', 'yes - change in personal purpose yes - problems with teacher', 'yes - learning problems yes - change in personal purpose', 'yes - learning problems yes - problems with teacher'}* |
| 14.HIGH_SCH | Profile of completed secondary education<br>*{'profiled high school (language mathematics or other)', 'vocational high school with different specialty', 'vocational high school in the specialty', 'non profiled high school'}* |
| 15.ASSES | Average success from high school<br>*{excellent, 'very good', 'good', 'middle'}* |
| 20.SATISF_TRAIN | Satisfaction with the level of education (overall)<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 21.SATISF_CURRI | Satisfaction with the subjects in the curriculum of the specialty<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 22.SATISF_INFRA | Satisfaction with the educational infrastructure (laboratories, dormitory, office, etc.) at the university<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 23.SATISF_ADMIN | Satisfaction with the administrative service of students<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 24.SAT_REL_PROF | Satisfaction with communication between teachers and students<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |

| Attributes | Description and Values |
|---|---|
| 25.SAT_REL_STUD | Satisfaction with student relationships<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 26.SATISF_DIFFIC | Satisfaction with the difficulty and volume of the content of the curriculum in the subjects of the specialty<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 27.SATISF_QUALIF | Satisfaction with opportunities for professional development<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 28.SATISF_FUTU | Graduation from university is a prerequisite for professional success in the future<br>*{'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'}* |
| 32.EDU_STATUS | Student status<br>*{active, dropout}* |

### 3.3. Classification Algorithms

The article applies machine learning (ML) algorithms to analyze the results of the survey and to identify the significant attributes of the data collected to suggest a hypothesis that will continue the study.

There are three different learning styles in machine learning algorithms: Supervised Learning, Unsupervised Learning, and Semi-Supervised Learning. The style that is applied depends on the type and the content of the data. The Machine learning algorithms we are testing refer to Supervised Learning. Supervised learning deals with labeled data, which are contained in the prepared data set.

To analyze the data, we use well-known data mining algorithms, including two rule learners, a decision tree classifier, two popular Bayes classifiers. The WEKA software is used for the study implementation since it is freely available to the public and is widely used for research purposes in the data mining field.

The algorithms listed in Table 2 are applied in the study. Each algorithm is executed on the full set of data (full training set). Then each of the algorithms is executed again, but with the following test options: Stratified cross-validation Fold - 10 (The algorithm is applying 10 times. The data is divided into 10 folds. Each time 9 of the folds are used for training and 1 fold is used for testing) and Percentage Split 66% (Training is on 2/3 of the data. Testing is on 1/3 of the data.In our case, these are 39 instances.)

*Table 2. Classification machine learning algorithms used in the study*

| ***Classifiers/ Options*** | ***Description*** |
|---|---|
| *NaiveBayes* | *This supervised learning algorithm is a probabilistic classifier and uses a statistical method for each classification.* |
| *BayesNet* | *The BayesNet classifier creates simple graphics that include all the first-level input attributes.* |
| *Rules.JRip* | *It implements a propositional rule learner called as "Repeated Incremental Pruning to Produce Error Reduction (RIPPER)" and uses sequential covering algorithms for creating ordered rule lists. The algorithm goes through 4 stages: Growing a rule, Pruning, Optimization and Selection [12].* |
| *Rules.OneR* | *A simple classification that produces one rule for each predictor in the data and then the rule with the smallest total error is selected [13].* |
| *Stratified cross-validation Fold – 10 – for every algorithm* | *Cross-validation (using 10 folds and applying the algorithm 10 times – each time 9 of the folds are used for training and 1 fold is used for testing) [2].* |
| *Percentage Split 66% - for every algorithm* | *Each classifier is applied for percentage split (2/3 of the dataset are used for training and 1/3 – for testing) [2].* |

## 4. RESULTS AND DISCUSSION

After the collection, control, purification and conversion of the data, a classification was made for the extraction of data, taking into account the nature of the data and the proposed machine learning algorithms. The analyzes are based on models with outputs for 32.EDU_STATUS with participation of the other 22 input attributes. The purpose of Classification is to classify instances of the dataset into different classes based on some constraints.

### 4.1. Confusion matrix

The confusion matrix visualizes the performance of the classification algorithm. It contains information about the actual and predicted classes.

Confusion matrices of different classifications have been found. The accuracy indicators are calculated from them: TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area and PRC Area. For the purposes of the article, the focus is on Accuracy, TP Rate and Precision.

With this study, we aim to determine whether a student is active "A" or dropout "D". Table 3 shows the confusion matrices for the four selected algorithms

(BayesNet, NaiveBayes, Rules.OneR and Rules.JRip) for each of the three testing options Full training set, Cross validation Folds 10, Percentage Split 66%.

*Table 3. Confusion matrix for all ML algorithms*

| ***BayesNet Predicted class:*** *A = active; D = dropout* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Actual class* | *Full training set* | | *Total* | *Cross validation Folds 10* | | *Total* | *Percentage Split 66%* | | *Total* |
| | *A* | *D* | | *A* | *D* | | *A* | *D* | |
| *A* | *87* | *4* | *91* | *82* | *9* | *91* | *30* | *0* | *30* |
| *D* | *7* | *17* | *24* | *17* | *7* | *25* | *7* | *2* | *9* |
| *Total* | *94* | *21* | *115* | *99* | *16* | *115* | *37* | *2* | *39* |
| ***NaïveBayes Predicted class:*** *A = active; D = dropout* | | | | | | | | | |
| *Actual class* | *Full training set* | | *Total* | *Cross validation Folds 10* | | *Total* | *Percentage Split 66%* | | *Total* |
| | *A* | *D* | | *A* | *D* | | *A* | *D* | |
| *A* | *87* | *4* | *91* | *82* | *9* | *91* | *30* | *0* | *30* |
| *D* | *9* | *15* | *24* | *18* | *6* | *24* | *7* | *2* | *9* |
| *Total* | *96* | *19* | *115* | *100* | *15* | *115* | *37* | *2* | *39* |
| ***Rules.OneR Predicted class:*** *A = active; D = dropout* | | | | | | | | | |
| *Actual class* | *Full training set* | | *Total* | *Cross validation Folds 10* | | *Total* | *Percentage Split 66%* | | *Total* |
| | *A* | *D* | | *A* | *D* | | *A* | *D* | |
| *A* | *89* | *2* | *91* | *89* | *2* | *91* | *28* | *2* | *30* |
| *D* | *18* | *6* | *24* | *20* | *4* | *24* | *8* | *1* | *9* |
| *Total* | *107* | *8* | *115* | *109* | *6* | *115* | *36* | *3* | *39* |
| ***Rules.JRip Predicted class:*** *A = active; D = dropout* | | | | | | | | | |
| *Actual class* | *Full training set* | | *Total* | *Cross validation Folds 10* | | *Total* | *Percentage Split 66%* | | *Total* |
| | *A* | *D* | | *A* | *D* | | *A* | *D* | |
| *A* | *89* | *2* | *91* | *86* | *5* | *91* | *28* | *2* | *30* |
| *D* | *19* | *5* | *24* | *22* | *2* | *24* | *8* | *1* | *9* |
| *Total* | *108* | *7* | *115* | *108* | *7* | *115* | *36* | *3* | *39* |

The columns for each algorithm tell how the model classified samples - it's what the model predicted.

The top-left and bottom-right cell in the matrix show instances which our model classifies correctly.

The bottom-left and top-right cell of the matrix show instances which our model classifies incorrect.

The results show that for all classifiers, the performance is the best, with testing option Full training set

### 4.2. Accuracy percentage

The accuracy percentage can be calculated from Confusion matrix: (TP+TN)/(P+N). After the implementation of the ML Algorithms, it can be concluded that the Accuracy percentage for “dropout” class, when applying the BayesNet classification algorithm, has the highest number of correctly classified instances - 104, which is 90.43% of the total volume of data. The difference is small compared to NaiveBayes (102; 88.70%) and Rules.OneR (95 ; 82.61%), which are ranked second and third respectively. In fourth place is Rules.JRip with 94 cases, which is 81.74%. The other algorithms have smaller accuracy rate for “dropout” class.

The next studies were done for the first 4 algorithms that have the highest percentage of the accuracy - BayesNet, NaiveBayes, Rules.OneR, Rules.Jrip.

### 4.3. True Positive and Precision

#### 4.3.1. NaiveBayes and BayesNet

In this study, the comparison of the different ML classification algorithms is made by indicators True Positive Rate (TP Rate) and Precision. Where TP Rate is TP/(TP+FN) or the ratio of examples that were correctly classified as class x, among all examples that truly have class x. While Precision is TP/(TP+FP) or the ratio of examples that truly have class x among all those that were classified as class x. The obtained results for classifications with ML algorithms NaiveBayes and BayesNet are presented in Table 4.

The results reveal that True Positive Rate for NaiveBayes and BayesNet is equal for class 'active' - 0.956. The average True Positive Rate of the BayesNet classifier is slightly larger – it’s 0.904, and for the NaiveBayes classifier – 0.887. The BayesNet algorithm performs better in the 'dropout' class – 0.708, compared to 0.625 for NaiveBayes. The same is confirmed by the graphical presentation of these results (Figure 1.).

About Precision the results obtained show that, as expected, the ML algorithms by Bayes theorems, BayesNet performs better than the NaiveBayes algorithm. The average Precision is higher for BayesNet 90.1% compared to NaiveBayes - 88.2%. For class 'active' it is 92.6% compared to 90.6% for NaiveBayes and for class 'dropout' - 81% compared to 78.9% for NaiveBayes (Figure 2).

*Table 4. Results for Bayes classifications – Output 32.Edu_Status*

| Test options | ***Class*** | ***NaiveBayes*** | | ***BayesNet*** | |
|---|---|---|---|---|---|
| | | ***TP Rate*** | ***Precision*** | ***TP Rate*** | ***Precision*** |
| Full Training Set | *'active'* | 0.956 | 0.906 | 0.956 | 0.926 |
| | *'dropout'* | 0.625 | 0.789 | 0.708 | 0.810 |
| | *Weighted Avg.* | 0.887 | 0.882 | 0.904 | 0.901 |
| Cross validation Folds 10 | *'active'* | 0.901 | 0.820 | 0.901 | 0.828 |
| | *'dropout'* | 0.250 | 0.400 | 0.292 | 0.438 |
| | *Weighted Avg.* | 0.765 | 0.732 | 0.774 | 0.747 |
| Percentage Split 66% | *'active'* | 1.000 | 0.811 | 1.000 | 0.811 |
| | *'dropout'* | 0.222 | 1.000 | 0.222 | 1.000 |
| | *Weighted Avg.* | 0.821 | 0.854 | 0.821 | 0.854 |

### 4.3.2. Rule learners - Rules.JRip and Rules.OneR

Two algorithms for generating classification rules are considered. The results are presented in Table 5.

*Table 5. Results for Rule learners – Output 32.Edu_Status*

| Test options | ***Class*** | ***Rules.JRip*** | | ***Rules.OneR*** | |
|---|---|---|---|---|---|
| | | ***TP Rate*** | ***Precision*** | ***TP Rate*** | ***Precision*** |
| Full Training Set | *'active'* | 0.978 | 0.824 | 0.978 | 0.832 |
| | *'dropout'* | 0.208 | 0,714 | 0.250 | 0.750 |
| | *Weighted Avg.* | 0.817 | 0.817 | 0.826 | 0.815 |
| Cross validation Folds 10 | *'active'* | 0.945 | 0.796 | 0.978 | 0.817 |
| | *'dropout'* | 0.083 | 0.286 | 0.167 | 0.667 |
| | *Weighted Avg.* | 0.765 | 0.690 | 0.809 | 0.785 |
| Percentage Split 66% | *'active'* | 0.933 | 0.778 | 0.933 | 0.778 |
| | *'dropout'* | 0.111 | 0.333 | 0.111 | 0.333 |
| | *Weighted Avg.* | 0.744 | 0.675 | 0.744 | 0.675 |

The results obtained show that the OneR rule algorithm performs better than the JRip ML algorithm. The overall TP Rate of the OneR classifier is 82.6% and for the JRip classifier, it is 81.7%. Both rule-based algorithms perform better for the 'active' class. For the 'dropout' class TP Rate is higher for OneR again. The results are also visible in the comparison chart for the Precision of the same ML algorithms (Figure 2.).

For the 'dropout' class Precision is big, but for the 'active' class is is slightly less for two classifiers – around 20-25% (Figure 2).

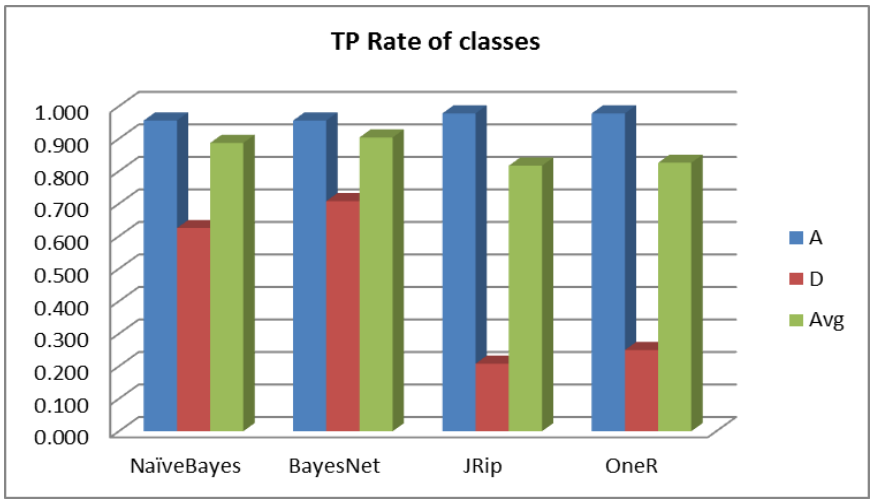


*Figure 1. Comparison for TP rate of classes*

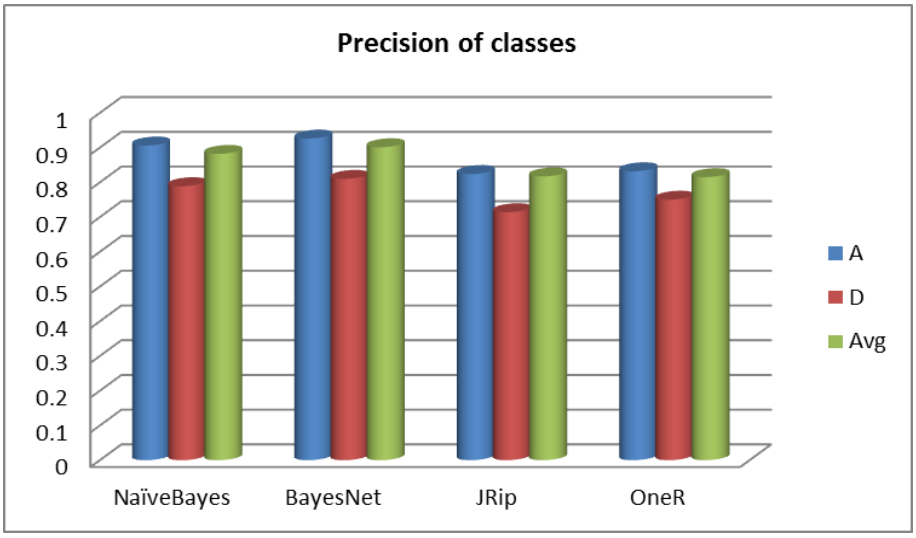


*Figure 2. Comparison for Precision of classes*

### 4.3.3. Comparison of TP rate and Precision

In order to select the appropriate ML algorithm for the purposes of the study of the following stages, it is necessary to compare the TP Rate results of the 'active' and 'dropout' classes from the four models selected at this stage from the applied algorithms: NaiveBayes, BayesNet, Rules.Jrip, Rules.OneR.

The summarized results from the comparison of TP rate and Accuracy are shown in Figure 1, and from the comparison of Precision - in Figure 2.

A comparison of the TP rates of the selected algorithms shows that BayesNet performs best. The value of the TP rate for NaiveBayes is almost the same. The difference is only 0.017. OneR and Jrip follow.

The comparison of Precision shows that BayesNet gives the best results for Weighted Average. The Precision value of NaiveBayes is lower, next is that of Jrip, and close to it is OneR.

The BayesNet classifier can be considered the most effective. It gives the best result for Precision and for TP rate. The average accuracy of the four selected classifiers is over 80%, which means that the results are with satisfactory reliability.

The comparison of the results for the two classes shows that the forecasts for the 'active' class are better.

### 4.4. Feature selection

Weka software offers various techniques for determining the most important attributes for the prediction: GainRatioAttributeEval; InfoGainAttributeEval; CfsSubsetEval et al. As a result of their application, the characteristics are ranked. In this study, different selection algorithms were applied and the optimal number of features for each of them was determined. To determine the most appropriate subset of features, classifications were made with Bayes Net selected as the most accurate classifier and the performances of the classifications were compared. The highest classification accuracy was found to be obtained with the subset containing 8 features determined by the InfoGainAttributeEval technique. The most important features, sorted by sugnificance are the following: **12.PERS_STRESS; 1.AGE; 3B.COURSE; 24.SAT_REL_PROF; 10.EDU_PARENT; 7.JOB_SATISF; 4.MAR_STAT; 25.SAT_REL_STUD.**

## 5. CONCLUSION AND FUTURE RESEARCH

This article analyzes the data from a survey among students from Faculty Technics and Technologies -Yambol, Trakia University - Stara Zagora in order to find the reasons for students dropping out of university. Based on the review of the literature sources and the experience of the authors, a questionnaire was developed, including 32 questions. A study involving 115 students was conducted.

The data is processed in the WEKA platform after going through the stages of accumulation, purification and transformation. The result is an ARFF file that contains 115 instances with 23 attributes.

The data were analyzed with ML algorithms offered in WEKA 3.9, namely: Bayes classifications (NaiveBayes, BayesNet) and Rule learners (Rules.JRip and Rules.OneR).

The most important results can be summarized as follows:

1. Supervised ML classification algorithms are appropriate for predicting university dropouts.

2. The application of ML algorithms to the data collected so far indicates that the highest degree of accuracy and the highest number of correctly classified instances are available for the BayesNet. Next are NaiveBayes, JRip and OneR classification algorithms.

The obtained results will serve as a basis for continuing the study of the reasons for dropping out of students from all specialties of Faculty of Technics and technologies - Yambol at Trakia University - Stara Zagora.

**REFERENCES**

[1] Vossensteyn, H. at al. *Dropout and Completion in Higher Education in Europe: Main Report*, Publications Office of the European Union, Luxembourg, 2015

[2] Kabakchieva, D. Predicting student performance by using data mining methods for classification, *Cybernetics and information technologies*, 1 (vol. 13), 2013, pp. 61-72.

[3] Askinadze, A., S. Conrad. Predicting Student Dropout in Higher Education Based on Previous Exam Results. *Proc. of the 12th Int'l Conf. on Educational Data Mining (EDM 2019)*, Montréal, Canada, 2019, pp. 500-503.

[4] Askinadze, A., S. Conrad. Application of the dynamic time warping distance for the student drop-out prediction on time series data. *Proc. of the 10th Int'l Conf. on Educational Data Mining (EDM 2017)*. Wuhan, Hubei, China, 2017, pp. 342-343.

[5] Manrique, R. at al. 2019. An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. *Proc. of the 9th Int'l Learning Analytics and Knowledge Conference (LAK-19)*. ACM, Tempe, Arizona, USA, March 2019, pp. 401-410.

[6] Del Bonifro, F., M. Gabbrielli, G. Lisanti, S. Zingaro. Student Dropout Prediction. *Proc. of the Int'l Conf. Artificial Intelligence in Education. AIED 2020*. Lecture Notes in Computer Science, vol 12163. Springer, Cham, 2020, pp. 129-140.

[7] Berens, J., K. Schneider, S. Gortz, S. Oster, J. Burghoff. Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, 3, (vol.11), 2019, pp. 1-41.

[8] Hutagaol, N., Suharjito, Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education, *Advances in Science, Technology and Engineering Systems Journal*, 4, (vol.4), 2019, pp. 206-211.

[9] Wakelam E., A. Jefferies, N. Davey, Y. Sun. 2019. The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology*, 2 (vol.51), 2019, pp. 347-370

[10] Garcia, S., J. Luengo, F. Herrera. *Data preprocessing in data mining*. Springer, Switzerland, 2015.

[11] Gupta, A., A. Mohammad, A. Syed, M. Halgamuge. A comparative study of classification algorithms using data mining: crime and accidents in Denver City the USA. *International Journal of Advanced Computer Science and Applications*, 7 (vol.7), 2016, pp. 374-381.

[12] Veeralakshmi, V., D. Ramyachitra. Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset. *International Journal of Computer Science Engineering*, 3, (vol. 1), 2015, pp. 79-85.

[13] Witten, I., E. Frank, M. Hall. *Data Mining: Practical machine learning tools and techniques* (3rd. ed.). Morgan Kaufmann, San Francisco, 2011.

***Information about the authors:***

**Veselina Nedeva** – Assoc. Professor of Informatics at Faculty of Technics and Technologies, Trakia University, Bulgaria. Areas of scientific research: Information Systems, Databases, Data Analysis, Machine Learning, Information Technologies in Education, E-learning, etc.

**Tanya Pehlivanova -** Assoc. Professor of Electrical Engineering at Faculty of Technics and Technologies, Trakia University, Bulgaria. Areas of scientific research: Electrical Engineering, Hybrid Solar-wind Energy Systems, E-learning, Machine Learning etc.

**Manuscript received on 8 July 2020**