Students’ Performance Analyses Using Machine Learning Algorithms in WEKA

Veselina Nedeva¹, Tanya Pehlivanova¹
¹Trakia University of Stara Zagora, Faculty of Technics and Technologies of Yambol, Bulgaria

e-mail: veselina.nedeva@trakia-uni.bg

Abstract. Predicting student performance is important for universities. Thus, they can identify students who need support and take measures to improve educational outcomes. The paper presents an analysis of the results of a study conducted at the Trakia University of Stara Zagora. The study aims to identify the most significant features that affect student performance and to select the most efficient machine-learning algorithm to predict their performance. The efficiency of four classification algorithms was compared - BayesNet (BN), Multilayer Perceptron (MLP), Sequential minimal optimization (SMO) and Decision tree (J48). For comparison, the indicators TP Rate, Precision, F-Measure, Accuracy and error measures - MAE, RMSE, RAE, and RRSE were used. The processing is done with Weka open-source software. The obtained results show that the MLP algorithm is the best for the used data. The obtained accuracy is sufficient to create an effective forecast model. 12 attributes have been identified that have the greatest impact on student performance.

1. Introduction
The main goal of every university is to provide quality knowledge and skills to students so that they are competitive in the labour market. One way to achieve this goal is to timely predict student performance. Thus, it is easy to identify students who need support and take measures to improve educational outcomes. This would help teachers to provide an effective approach to teaching.

The analysis of students' progress during their studies provides the university management with information about the probability of success of each student. Traditionally, this analysis is done by the lecturers who use their interactions with students in classroom activities and mid-term assessments to identify those "at risk" of dropping out and take timely action. In the modern system of higher education, the time for communication between lecturers and students is constantly decreasing and finding endangered students is becoming increasingly difficult. The reason for this is the growing number of working students and access to online learning materials for students [1]. More accurate predictions of student success can be made using modern techniques for data mining and machine learning analyses.

In the age of mass computerization, a lot of data is collected in educational institutions, but it often remains unused. Powerful tools are needed to reap the benefits of this data. Such tools are machine learning algorithms (MLA).

The paper aims to identify the key variables that affect educational success and to select the most effective ML algorithm to predict student achievement.
2. Related work

There are many studies in the literature related to the students’ performance using machine learning. Most research in this area seeks to identify the most appropriate algorithms by which predictions can be made and to identify the features that can be used for forecasting.

Scientific articles are known, the purpose of which is to make an in-depth review of the literature related to the prediction of students results [2, 3].

In [2], different techniques and different types of data used are analyzed. “Based on the data collected in this review, the most widely used technique for predicting students’ behaviour was supervised learning, as it provides accurate and reliable results. In particular, the SVM algorithm was the most used by the authors and provided the most accurate predictions. In addition to SVM, DT, NB and RF have also been well-studied algorithmic proposals that generated good results. As for the neural networks, they are a less used technique, but they obtain great precision in predicting the students’ performance.

The characteristics used are Demographic characteristics of students and their grades, grades obtained in other courses, high school exams, behavioural data, students’ interaction with certain Moodle modules, students' characteristics and their academic performance.

According to [3], there is a clear increase in the number of published studies in the field in recent years.

„The methodologies that are used can be split into Classification (supervised learning, e.g., Naive Bayes, Decision Trees), Clustering (unsupervised learning, e.g., partitioning data), Statistical (e.g., correlation, regression), Data mining (identifying features and trends) and other methods.“ [3]

„The features that have been used to predict student performance can be broadly split into five categories: demographic (e.g., age, gender), personality (e.g., self-efficacy, self-regulation), academic (e.g., high-school performance, course performance), behavioural (e.g., log data) and institutional (e.g., high-school quality, teaching approach). The majority of the articles used academic data for prediction (e.g., predicting course performance based on high-school performance). The use of data describing student behaviour in a course (log data), while becoming more popular within the computing education research domain, is still relatively rare.“ [3]

In Bulgarian universities, the topic of predicting student performance is poorly developed. There are few studies of Bulgarian authors related to this issue.

In [4] a study is presented, which aims to verify whether there are models available for the students’ performance at the university in the data on their personal and pre-university characteristics. The characteristics in the available data that most strongly influence the success of the students are also sought. Machine learning was used to achieve the goal. After comparing the results of the classifiers J48, Naive Bayes, Bayes Net, k-NN, Jrip, OneR, it was found that J48 performs best, then Jrip and k-NN.

The good performance of the models depends on the type of data used for forecasting. This explains the fact that in different studies different classifiers are offered as the most appropriate.

In [1] was discovered the potential for predicting the performance of students in small student groups with very limited attributes - attendance at lectures, access to a virtual learning environment and intermediate grades. From the compared machine learning techniques, K-Nearest Neighbors (KNN) and Random Forest (RF) give the most accurate forecasting results.

A different approach to data used to predict student performance is applied in [5 England, external data]. There, models have been developed for predicting student grades using internal institutional databases and external open data sources (the results of the National Student Survey). It’s proved that models based on internal and external data sources provide better efficiency and are more accurate than models based only on internal institutional data sources.

The literature review reveals that as classifiers with the highest accuracy for the problem under consideration are listed:

Support Vector Machine (SVM) [6, 7], Random Forest (RF) [1, 8], Decision tree (DT) [9], k-NN and Bayes classifiers [1], Artificial Neural Networks (ANN) [6], Naive Bayes (NB) [10].
To make an adequate prognosis, the reasons that determine the success of students must be known. The most commonly used attributes for predicting student achievement are: personal and pre-university characteristics [4], personal characteristics and academic achievements [11] demographic factors [10], grades obtained up to a certain stage of training [10, 7, 12], Moodle activity [8], use of online resources [13, 14] and others.

3. Materials and methods
As a result of the analysis of the literature sources, the initial data from which the forecast models will be created have been determined. A survey was compiled, which included 32 questions. The survey was provided to 115 students studying in 2 engineering programs at the Trakia University - Stara Zagora. The data cover three groups of characteristics - personal characteristics, academic environment and social factors. The survey was conducted in the period 10.10.2019 to 01.02.2020.

3.1. Preprocessing
From the data obtained through the survey, 27 questions are used, each of which forms an attribute (property) that gives the feature of the studied data set. The final data set used to execute the project contains 115 instances, each described with 27 attributes in WEKA. 25 of them are with nominal variables, and 2 are with numeric values.

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

The attributes, their description and values are presented in Table 1.

| Attributes | Description and Values |
|------------|-----------------------|
| 1.AGE | Age |
|            | {19-20, 21-23, 24-26, 27-29, 30-32} |
| 3B.COURSE | Course/Year of Education |
|            | {'First Year', 'Second Year', 'Third Year', 'Fourth Year'} |
| 4.MAR_STAT | Marital status of students |
|            | {Single, Married, Other} |
| 5.CHILD | Children of the student |
|            | {'Have not children', 'Have children'} |
| 6.JOB | Does the student work in the specialty |
|            | {'yes, in the other specialty', 'Yes, in the specialty', 'I do not work'} |
| 7.SATISF | Job satisfaction |
|            | {'neither yes nor no ', 'rather no ', 'yes, 'rather yes', no, 'no opinion'} |
| 8.INCOME | Family incomes |
|            | {medium, 'low ', 'high, 'very low ', 'very high'} |
| 9.PLACE | Place of residence |
|            | {'small town', 'big city', Sofia, village} |
| 10.EDU_PARENT | Parental education |
|            | {'higher education', 'higher education - secondary education', 'secondary education', 'secondary education-primary education', 'primary education', 'higher education-primary education '} |
| 12.PERS_STRESS | Personal stress |
|            | {'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 types', 'yes - change in personal purpose and problems with colleagues', 'yes - financial yes - learning problems  yes - problems with teacher', 'yes - illness yes -
3.2. Attribute selection

To achieve the goal of the research, the attributes are selected before the classifications are made. Four attribute selections are made using the following evaluation algorithms for selection: GainRatioAttributeEval, InfoGainAttributeEval, CorrelationAttributeEval and CfsSubsetEval. The Ranker search method is applied to the first three; and the GreedyStepwise search method is applied to CfsSubsetEval. The selection is made according to the coefficients of significance. The results for 4 of

| Attributes   | Description and Values                                                                 |
|--------------|----------------------------------------------------------------------------------------|
| 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 |
| {'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 |
| {'excellent', 'very good', 'good', 'satisfactory'} | 17.SPEC_2 At what rank request is admitted student in the specialty in which he is studying |
| {no, yes} | 18.ASSES_UNI The average success of the student from studying at the university |
| {'excellent', 'very good', 'good', 'satisfactory', 'fail'} | 19A.EXAM_FAIL Does the student have failed exams |
| Numeric | 19B.EX_FAIL_NUM Number of failed exams |
| {'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'} | 20.SATISF_TRAIN Satisfaction with the level of education (overall) |
| {'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'} | 21.SATISF_CURRI Satisfaction with the subjects in the curriculum of the specialty |
| {'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 |
| {'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'} | 23.SATISF_ADMIN Satisfaction with the administrative service of students |
| {'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'} | 24.SAT_REL_PROF Satisfaction with communication between teachers and students |
| {'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'} | 25.SAT_REL_STUD Satisfaction with student relationships |
| {'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 |
| {'yes', 'rather yes', 'neither yes nor no', 'rather no', 'no'} | 27.SATISF_QUALIF Satisfaction with opportunities for professional development |
| {'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 |
| {active, dropout} | 32.EDU_STATUS Student status |
the selections with the best coefficients of significance are compared. 10 of the attributes are at the beginning of each of the four selections. To them are added the attributes that occur in at least two selections. Thus, 14 attributes were selected.

A brief description of the selected 4 attribute selection techniques is given in Table 2.

### Table 2. Evaluators and their actions.

| Evaluator                  | Action                                                                 |
|----------------------------|------------------------------------------------------------------------|
| GainRatioAttributeEval;    | Evaluates the worth of an attribute by measuring the gain ratio with    |
|                            | respect to the class.                                                   |
|                            | GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) /           |
|                            | H(Attribute).                                                          |
| InfoGainAttributeEval      | Evaluates the worth of an attribute by measuring the information gain  |
|                            | with respect to the class.                                             |
|                            | InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute).              |
| CfsSubsetEval;             | Evaluates the worth of a subset of attributes by considering the      |
|                            | individual predictive ability of each feature along with the degree of |
|                            | redundancy between them. Subsets of features that are highly          |
|                            | correlated with the class while having low intercorrelation are        |
|                            | preferred.                                                             |
| CorrelationAttributeEval;  | Evaluates the worth of an attribute by measuring the correlation       |
|                            | (Pearson's) between it and the class. Nominal attributes are           |
|                            | considered on a value by value basis by treating each value as an      |
|                            | indicator. An overall correlation for a nominal attribute is arrived   |
|                            | at via a weighted average.                                             |

### 3.3. Compared parameters

The performance parameters are derived from the confusion matrix for each classification. The possible results of the classification (from the students' performance) are 5: a - excellent, b - very good, c - good, d - satisfactory, e - fail. Therefore, each confusion matrix is formed based on 25 classification results. An example of a confusion matrix for classification with the J48 algorithm is given in Table 3. The next parameters for each class can be read from it: true positive (TP) i.e. correct positive prediction, true negative (TN) i.e. correct negative prediction, false positive (FP) i.e. incorrect positive prediction and false negative (FN) i.e. incorrect negative prediction.

### Table 3. Confusion matrix of J48.

| Classified as /Predicted as | N=115 | FN  |
|-----------------------------|-------|-----|
| a d e                        |       |     |
| 4 1 0 0 0                    | a – excellent | 1  |
| 0 32 5 0 0                   | b – very good | 5  |
| 1 6 46 0 1                   | c – good | 8  |
| 0 3 2 11 0                   | d – satisfactory | 5  |
| 0 0 0 3                      | e – fail | 0  |
| 1 10 7 0 1                   | FP    | 19  |

The parameters we compare are calculated as follows [15]:

#### 3.3.1. TP Rate

Rate of true positives (instances correctly classified as a given class)
3.3.2. *Precision* \((P)\)

Precision is the number of correct positive classifications divided by the total number of positive classifications. So,

\[
P = \frac{TP}{TP + FP}
\]  

(2)

3.3.3. *F-measure* \((Fm)\)

F-measure is a harmonic mean of precision \((P)\) and recall \((R)\). So,

\[
Fm = \frac{2PR}{P + R}
\]

(3)

where \(R = \frac{TP}{(TP + FN)}\).

3.3.4. *Accuracy*

Accuracy is the number of all correct classifications divided by the total numbers of cases. So,

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

(4)

The different error measures used for classification methods are as following.

3.3.5. *Mean Absolute Error* \((MAE)\)

MAE estimates how far the predictions or forecasts differ from the actual values.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |p_i - a_i|
\]

(5)

where \(n\) is the number of errors, \(|p_i - x_i|\) are the absolute errors.

3.3.6. *Root Mean Square Error* \((RMSE)\)

RMSE is an evaluator of the differences between the predictor values and the actual observed values.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}}
\]

(6)

where \(p_i\) are predicted values and \(a_i\) are actual values at time/place \(i\).

3.3.7. *Relative Absolute Error* \((RAE)\)

RAE is defined as the ratio of absolute error by the magnitude of the actual value. It is represented as below,

\[
RAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |a_i - \bar{a}|}
\]

(7)
where \( p_i \) is the forecast value, \( a_i \) is the actual value and \( \bar{a} \) is the average of actual values \( \bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i \).

### 3.3.8. Root Relative Squared Error (RRSE)

It is denoted as a mean absolute error (MAE) divided by the classification model error. It can be represented as below,

\[
RRSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i=1}^{n} (a_i - \bar{a})^2}}
\]

where \( \bar{a} \) is the average of actual values.

### 3.4. Classifications

The following ML algorithms are applied in the paper:

- **BayesNet**

  Bayes Net model represents probabilistic relationships among a set of random variables graphically. It models the quantitative strength of the connections between variables, allowing probabilistic beliefs about them to be updated automatically as new information that becomes available. It is a directed acyclic graph (DAG) that encodes a joint probability distribution, where the nodes of the graph represent a random variable and are representative of the correlation between variables. [15]

- **MultilayerPerceptron**

  MLP – Multi-Layer Perceptron method is considered as one of the most common extensively employed methods in a general supervised learning artificial neural network algorithm. Multilayer Perceptron is a feed-forward artificial neural network model that maps collections of input data in many satisfactory outputs. MLP has multiple layers of neurons where every layer is entirely related to the next one that can classify many data values in the non-linear matter. These layers can learn how to map and direct input data values into a required form of responses. [16]

- **SMO**

  Sequential minimal optimization (SMO) is an iterative algorithm for solving the quadratic programming (QP) problem using SVMs. It implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels.

- **J48**

  The C4.5 algorithm for building decision trees is implemented in Weka as a classifier called J48. It is an extension of the ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 is called a statistical classifier. J48 uses a measure called “information gain” to choose the attribute at each stage.

  It is known that academic performance depends primarily on the results that students receive in the exams at the end of each semester. Therefore, the output attribute is 18.ASSES_UNI. For the present analysis, 4 classification algorithms have been selected for which the accuracy is the highest. A classification is made with all 26 attributes. Then classifications are made with the already selected attributes in the following order:

  - The algorithm is implemented with the first 10 attributes;
  - One attribute is added at each subsequent iteration and the classification is implemented again;
  - Classifications are implemented until the 14 selected attributes are reached;
  - The results obtained from the implementation of ML algorithms are compared according to the already presented parameters.

  The obtained results from the classifications are compared by TP Rate, Precision, F-Measure, Accuracy. In this analysis, it is necessary to include a comparison according to different error measures used for classification methods - MAE, RMSE, RAE, and RRSE. The reason for this is that for some of the classifications the results are equal or close and the values for measuring the errors must be monitored.
4. Results and Discussion

4.1. Attributes selection
The selection of the attributes has been made according to the methodology described in item 3.2. Four attribute selection techniques are used with the following options:

- Sel 1 Evaluator: weka.attributeSelection.GainRatioAttributeEval
  Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

- Sel 2 Evaluator: weka.attributeSelection.InfoGainAttributeEval
  Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

- Sel 3 Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
  Search: weka.attributeSelection.GreedyStepwise -R -T -1.7976931348623157E308 -N -1 -num-slots 1

- Sel 4 Evaluator: weka.attributeSelection.CorrelationAttributeEval
  Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

The results obtained for all selections are shown in Table 4.

Table 4. Results from the selection of attributes.

| Sel 1                | Sel 2                | Sel 3                | Sel 4                |
|----------------------|----------------------|----------------------|----------------------|
| 0.3427 19B.EX_FAIL_NUM 0.2684 12.PERS_STRESS | 0.2199 19A.EXAM_FAIL 0.2555 19B.EX_FAIL_NUM | 0.1965 32.EDU_STATUS 0.2092 1.AGE | 0.1976 19B.EX_FAIL_NUM 0.2693 12.PERS_STRESS |
| 0.1396 12.PERS_STRESS 0.1871 14.HIGH_SCH | 0.1116 1.AGE 0.1549 19A.EXAM_FAIL | 0.1109 12.PERS_STRESS 0.1667 14.HIGH_SCH | 0.2464 19A.EXAM_FAIL 0.1236 12.PERS_STRESS |
| 0.1098 14.HIGH_SCH 0.1540 7.JOB_SATISF | 0.0867 15.ASSES 0.1426 3B.COURSE | 0.0871 24.SAT_REL_PROF 0.1253 32.EDU_STATUS | 0.0878 19B.EX_FAIL_NUM 0.2520 12.PERS_STRESS |
| 0.0769 24.SAT_REL_PROF 0.1253 32.EDU_STATUS | 0.0762 3B.COURSE 0.1245 15.ASSES | 0.0762 3B.COURSE 0.1245 15.ASSES | 0.0768 27.SATISF_QUALIF 0.1133 21.SATISF_CURRI |
| 0.0708 2.GENDER 0.1239 27.SATISF_QUALIF | 0.0697 4.MAR_STAT 0.1165 20.SATISF_TRAIN | 0.0671 20.SATISF_TRAIN 0.1113 24.SAT_REL_PROF | 0.0672 27.SATISF_QUALIF 0.1133 21.SATISF_CURRI |
| 0.0624 7.JOB_SATISF 0.1100 22.SATISF_INFRA | 0.0624 7.JOB_SATISF 0.1100 22.SATISF_INFRA | 0.0624 7.JOB_SATISF 0.1100 22.SATISF_INFRA | 0.0624 7.JOB_SATISF 0.1100 22.SATISF_INFRA |

The obtained results show that a part of the attributes participate among the top ten in each of the selections, regardless of the algorithm by which they are evaluated and the method by which they are searched. Four more attributes have been added to them, which are ranked among the top fourteen in at
least two of the selections. In this way, the following fourteen attributes are selected and ranked: 1. AGE; 2. GENDER; 3B. COURSE; 12. PERS_STRESS; 14. HIGH_SCH; 15. ASSES; 19A. EXAM_FAIL; 19B. EX_FAIL_NUM; 27. SATISF_QUALIF; 32. EDU_STATUS; 7. JOB_SATISF; 4. MAR_STAT; 24. SAT_REL_PROF; 20. SATISF_TRAIN. The selected attributes will participate in the analysis with MLA.

4.2. Classifications with selected attributes

According to the methodology described in item 3.4, classifications were made using the following four algorithms with the options indicated to them:

- **Class 1 Scheme**: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5 (BN)

- **Class 2 Scheme**: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a – TP Rate 0.991 (MLP)

- **Class 3 Scheme**: weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4" – TP Rate 0.957

- **Class 4 Scheme**: weka.classifiers.trees.J48 -C 0.25 -M 2 – TP Rate 0.835

The results of the classifications using all 26 attributes are shown in Table 5. The lowest values of the indicators are obtained in Class 1 (BN). Therefore, this algorithm is excluded from subsequent studies. The best indicators are obtained with the algorithm MLP (Class 2) - TP Rate 0.991; Precision 0.991 and F-Measure 0.911.

| Classification | TP Rate | Precision | F-Measure | Accuracy % | MAE  | RMAE | RAE % | RRSE % |
|----------------|--------|-----------|-----------|------------|------|------|-------|--------|
| Class 1        | 0.748  | 0.754     | 0.749     | 74.78      | 0.1306 | 0.2795 | 49.4476 | 77.2573 |
| Class 2        | 0.991  | 0.991     | 0.991     | 99.13      | 0.0088 | 0.0596 | 3.3406  | 16.4682 |
| Class 3        | 0.957  | 0.957     | 0.956     | 95.65      | 0.2417 | 0.3188 | 91.5477 | 88.1276 |
| Class 4        | 0.835  | 0.846     | 0.835     | 83.48      | 0.0912 | 0.2135 | 34.5225 | 59.0169 |

It is known that with an appropriate selection of attributes it is possible to increase the accuracy of classification. In our study, this does not happen. The described comparison indicators for the selected three classifiers applied for a different number of attributes were calculated. The results show that in all cases the indicators of the classifications with a reduced number of attributes have slightly lower values. The reduction, however, is not significant. The best results are from the application of the MLP algorithm with 12 attributes.

Although it does not improve the accuracy of the classification, the reduction of attributes is applicable. The need to choose the most appropriate attributes arises from the fact that often the information collected by students is sensitive. It concerns data on marital status, salary, education, grades, family stress, personal stress, etc., which respondents are not always willing to share.

| Classification | TP Rate | Precision | F-Measure | Accuracy % | MAE  | RMAE | RAE % | RRSE % |
|----------------|--------|-----------|-----------|------------|------|------|-------|--------|
| MLP Class 2-10 | 0.939  | 0.943     | 0.939     | 93.91      | 0.0382 | 0.1222 | 14.4762 | 33.7875 |
| MLP Class 2-11 | 0.954  | 0.958     | 0.956     | 95.65      | 0.0279 | 0.0952 | 10.5831 | 26.3256 |
| MLP Class 2-12 | 0.974  | 0.976     | 0.974     | 97.39      | 0.0190 | 0.0802 | 7.1875  | 22.1619 |
| MLP Class 2-13 | 0.974  | 0.975     | 0.973     | 97.39      | 0.0175 | 0.0929 | 6.6098  | 25.6934 |
Table 6. Results of classifications with 10 to 14 attributes.

| Classification       | TP Rate | Precision | F-Measure | Accuracy % | MAE  | RMAE | RAE % | RRSE % |
|----------------------|---------|-----------|-----------|------------|------|------|-------|--------|
| MLP Class 2-14       | 0.965   | 0.966     | 0.965     | 96.52      | 0.0244 | 0.112 | 9.2449 | 30.9726 |
| SMO Class 3-10       | 0.765   | 0.770     | 0.765     | 76.52      | 0.2497 | 0.3311 | 94.5774 | 91.5303 |
| SMO Class 3-11       | 0.791   | 0.802     | 0.787     | 79.13      | 0.2487 | 0.3296 | 94.1822 | 91.1228 |
| SMO Class 3-12       | 0.809   | 0.818     | 0.805     | 80.87      | 0.2483 | 0.3291 | 94.0505 | 90.9769 |
| SMO Class 3-13       | 0.800   | 0.804     | 0.799     | 80.00      | 0.2487 | 0.3294 | 94.1822 | 91.0499 |
| SMO Class 3-14       | 0.835   | 0.841     | 0.833     | 83.48      | 0.2473 | 0.3274 | 93.6553 | 90.5082 |
| J48 Class 4-10       | 0.643   | 0.707     | 0.612     | 64.35      | 0.1877 | 0.3063 | 71.0692 | 84.6771 |
| J48 Class 4-11       | 0.748   | 0.792     | 0.738     | 74.78      | 0.1344 | 0.2593 | 50.9080 | 71.6668 |
| J48 Class 4-12       | 0.748   | 0.792     | 0.738     | 74.78      | 0.1344 | 0.2593 | 50.9080 | 71.6668 |
| J48 Class 4-13       | 0.748   | 0.781     | 0.734     | 74.78      | 0.1350 | 0.2598 | 51.1149 | 71.8124 |
| J48 Class 4-14       | 0.748   | 0.781     | 0.734     | 74.28      | 0.1350 | 0.2598 | 51.1149 | 71.8124 |

The twelve attributes for which the best indicators are obtained with the Multilayer Perceptron classification algorithm are the following: 1.AGE; 2.GENDER; 3.B.COURSE; 12.PERS_STRESS; 14.HIGH_SCH; 15.ASSES; 19A.EXAM_FAIL; 19B.EX_FAIL_NUM; 27.SATISF_QUALIF; 32.EDU_STATUS; 7.JOB_SATISF; 4.MAR_STAT.

4.3. Comparison of selected indicators for different MLA

The comparison of the selected indicators for the different algorithms, with a different number of attributes, is shown in Figures 1 to 4. The indicators TP Rate, Precision, F-Measure and Accuracy are the highest in the classification with MLP (Figures 1 and 2). As for the number of attributes, maximum values are obtained for 12 pieces.

Figure 1. Comparing F-Measure, Precision and TP Rate for MLP, SMO and J48 Classes.
Figure 2. Comparing Accuracy (%) for MLP, SMO and J48 Classes.

The different error measures - MAE, RMSE, RAE and RRSE are the lowest again for the MLP classifier and twelve attributes (Figures 3 and 4).

Figure 3. Comparing RMSE and MAE for MLP, SMO and J48 Classes.

The error values for the MPL algorithm using 12 and 13 attributes are very close. A classification with 12 attributes was chosen because in this case the RMSE and RRSE indicators are the lowest.
5. Conclusion
In this report, MLA is used to predict the performance of students at the university. Data from a study conducted among students from the Trakia University - Stara Zagora were used. The data cover three groups of characteristics - personal characteristics, academic environment and social factors. Based on the review of the literature, four MLA has been selected, which give good results in similar research - NB, MLP, SMO and J48. Attribute selection techniques are applied and 14 attributes are selected and ranked. The processing is done with Weka open source software.

The comparison of the indicators TP Rate, Precision, F-Measure, Accuracy and error measures - MAE, RMSE, RAE and RRSE for the different classification algorithms and a different number of attributes shows that the best results are obtained using the algorithm Multilayer Perceptron MLP, applied to 12 attributes. The values of the main indicators are: TP Rate = 0.974; Precision = 0.976, F-Measure = 0.974 and Accuracy = 97.39%. The values for the errors in this case are the lowest: MAE = 0.0190, RMAE = 0.0802, RAE = 7.1875%, RRSE = 22.1619%.

Because the data contains personal information that students are not always willing to share, it has been studied how the reduction in the number of attributes affects the accuracy of the classification. It was found that when using twelve selected attributes, the accuracy remains close to the initial. These are: 1.AGE; 2.GENDER; 3B.COURSE; 12.PERS_STRESS; 14.HIGH_SCH; 15.ASSES; 19A.EXAM_FAIL; 19B.EX_FAIL_NUM; 27.SATISF_QUALIF; 32.EDU_STATUS; 7.JOB_SATISF; 4.MAR_STAT.

At the next stage, data for students from other technical programs at the Faculty of Technics and Technologies - Yambol will be included. The analyzes will be performed with the selected ML algorithms Multilayer Perceptron, which gives the best results.

References
[1] Wakelam E, Jefferies A, Davey N and Sun Y 2020 The potential for student performance prediction in small cohorts with minimal available attributes British Journal of Educational Technology Vol 51 No 2 pp 347–370
[2] Rastrillo-Guerrero J, Gómez-Pulido J and Durán-Domínguez A 2020 Analyzing and Predicting Students’ Performance by Means of Machine Learning: A Review. Applied Science 10, 1042
[3] Hellas A, Ihantola P, Petersen A, Ajanovski V, Gutica M, Hynninen T, Knutas A, Leinonen J, Messom C and Liao S 2018 Predicting Academic Performance: A Systematic Literature
[4] Kabakchieva D 2013 Predicting Student Performance by Using Data Mining Methods for Classification, Cybernetics and information technologies vol. 13 no 1 pp 61-72

[5] Sarker F, Tiropanis T and Davis H 2013 Students’ Performance Prediction by Using Institutional Internal and External Open Data Sources, Proc. Int. Conf. on Computer Supported Education (CSEDU 2013) (Aachen, Germany)

[6] Hussain M, Zhu W, Zhang W, Abidi S and Ali S 2018 Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review vol. 52 no 1 pp 1–27

[7] Huang S and Fang N 2013 Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education vol. 61 pp 133–145

[8] Nespereira C, Elhariri E, El-Bendary N, Vilas A and Redondo R 2015 Machine learning based classification approach for predicting students’ performance in blended learning Proc. Int. Conf. on Advanced Intelligent System and Informatics (AISI2015) (Beni Suef, Egypt) pp 47-56

[9] Sultana S, Khan S and Abbas M 2017 Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts International Journal of Electrical Engineering & Education vol. 54 pp 105–118

[10] Kotsiantis S, Pierrakes C and Pintelas P 2004 Predicting Students’ Performance in Distance Learning using Machine Learning Techniques Applied Artificial Intelligence vol. 18 no 5 pp 411–426

[11] Tan M and Shao P 2015 Prediction of student dropout in E-learning program through the use of machine learning method International Journal of emerging technologies in learning vol. 10 pp 11–17

[12] Bydžovská H 2015 Student performance prediction using collaborative filtering methods Proc. Int. Conf. on Artificial Intelligence in Education (AIED 2015) (Madrid, Spain) pp 550–553

[13] Sagar M, Gupta A and Kaushal R 2016 Performance prediction and behavioral analysis of student programming ability. Proc. Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI) (Jaipur, India) pp 1039–1045

[14] Saqr M, Fors U and Tedre M 2017 How learning analytics can early predict under-achieving students in a blended medical education course Journal Medical Teacher vol. 39, no 7 pp 757–767

[15] Witten I, Frank E and Hall M 2011 Data Mining: Practical machine learning tools and techniques 3rd. ed. (Morgan Kaufmann, San Francisco) pp 172

[16] Nazhat S and Maryam A 2019 Analysing Students’ Learning Style to Predict the Most Important Factors that Affects the Performance of e-Learning Platform for Iraqi Postgraduate Studies using MLP-ANN Algorithm and Electronic Questionnaire International Advanced Research Journal in Science, Engineering and Technology IARJSET Vol. 6, no 5 pp 93-107