A Comparative Analysis of Feature Selection Algorithms in Order to Improve Students Success

Marko Bursać
School of Railroad Transport of Applied Studies
Academy of Technical and Art Applied Studies Belgrade
Belgrade, Serbia
marko.bursac@vzs.edu.rs

Zoran Ćirović, Jelena Mitić
School of Electrical and Computer Engineering of Applied Studies
Academy of Technical and Art Applied Studies Belgrade
Belgrade, Serbia
zoran.cirovic@viser.edu.rs,
jelena.mitic@viser.edu.rs

Marija Blagojević
Faculty of technical sciences Čačak
University of Kragujevac,
Čačak, Serbia
marija.blagojevic@ftn.kg.ac.rs

Abstract—**Educational Data mining and artificial intelligence are among the most common areas when it comes to discovering the pattern of education. This paper presents the research of feature selection algorithms as well as a comparative analysis of their influence on accuracy when determining the success of students. In this paper we use four most common feature selection algorithms in combination with a multilayer perceptron classification algorithm. Research has shown that the ReliefAttributeEval algorithm gives the best results.**

Keywords- educational data mining; feature selection; students success; prediction

I. INTRODUCTION

Student success is an important aspect of the success of an educational institution, in addition to improve the quality of the teaching process and hiring adequate teaching staff. By applying data mining technologies in education, it is possible to improve the education process, predict the success of students at the level of courses, years of study, exams, etc.

One of the definitions of Data Mining (DM) or Knowledge Discovery in Data-bases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections [1]. Educational Data Mining (EDM) is relatively new disciplines of DM. EDM is an interdisciplinary area of research that aims to improve the educational process by using methods of statistics, machine learning, data analysis. One of the important steps in data preprocessing in DM (EDM) is feature selection. Feature selection has a significant impact on students' performance.

Feature selection (FS) or attribute selection is process of selecting relevant attributes for predictive model. It can be used to identify and eliminate unnecessary, irrelevant and redundant attributes from data that do not contribute to the predictive model's accuracy or may even reduce model accuracy [2]. FS enable to reduce time of computation, improve prediction performance and allow us to better understand data. After data preprocessing steps (data cleaning, statistical analysis, data transformation) FS is last step in data preprocessing [3] and also main task before applying data mining techniques [1].

This paper represent analysis of performance diferent FS algorithms in order to improve accuracy of student performance prediction using the artificial neural network method. The research was conducted on a data set, collected from the student database School of Electrical and Computer Engineering Applied Studies (Academy of Technical and Art Applied Studies Belgrade).

The main goal of the research is to identify the most suitable FS and which in combination with the multilayer perceptron method gives the best results. One of the goals of the paper is to identify subjects that affect success in the Visual Programming Techniques course (fourth semester).

For his purposes we use WEKA an open-source machine learning software written in Java, which is the most used software in EDM. One of the reasons to use WEKA is many pre-build tools for data preprocessing, classification, association rules, regression and visualization [3]. In this research we use four FS algorithm (CorrelationAttributeEval, GainRadioAttributeEval, InfoGainAttributeEval, ReliefAttributeEval) which are implemented in the WEKA software and Multilayer perception classification algorithm.

II. RELATED WORK

Authors in the paper [4] provide a comparative analysis of FS algorithms in determining student performance, as well as the selection of the most suitable classification algorithm. Research done using WEKA software provides a comparative view of 15 classification algorithms for six FS algorithms. The results of the research indicate that ReliefAttributeEval, ChiSquaredAttributeEval and CfsSubsetEval are one of the most important FS algorithms in determining students' performance. By applying the mentioned FS algorithms and classification algorithms, the accuracy in determining the performance of subjects can be increased by 10 to 20 percent.

The most important factors when predicting student success are shown through the research presented in the paper [5]. The authors used a data set composed of demographic, socio-economic and academic data. During data preprocessing, GainRadioAttributeEval and InfoGainAttributeEval algorithms were used, which determined 11 and 9 attributes with

significance greater than 0.1. By applying the classification algorithms, it was found that the highest accuracy was obtained using the J48 and Random Forest algorithms.

The authors in the paper [6] optimize the artificial neural network (ANN) model for predicting student success using the FS algorithm (ReliefAttributeEval). Applying the algorithm resulted in 11 out of 21 attributes ranked with more than 0.5, which were used in the creation of a new ANN model. The results indicate an increased accuracy of the model by 25%.

The research in the paper [7] provides a comparative view of the application of different FS algorithms (CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttribute, GainRadio AttributeEval, PrincipalComponents and Relief AttributeEval) and 15 different classification algorithms available in Weka software. The results shown through the F-measure, precision, and recall values indicate that the best results were achieved by applying the principal components FS algorithm and the Decision Tree (DT).

Similar research was conducted in the paper [8], where the authors use nine classification algorithms and two FS algorithms (CorrelationAttributeEval and Wrapper-Based algorithm). According to the obtained results, SMO and J48 have the highest accuracy with CorrelationAttributeEval, while Naïve Bayes has the highest accuracy with the wrapper subset feature selection algorithm when determining students' grades.

Abeje Orsango Enaro and Sudeshna Chakraborty [9] use four different FS algorithms and six classification algorithms in their research. The results of applying FS algorithms and classification algorithms indicate that the CfsSubsetEval algorithm with Random Forest algorithm gives the best results (Correctly Classified Instances 77.29%) compared to other algorithms.

III. METHODOLOGY

The research methodology includes data collection, initial data preparation, data preprocessing and applying data mining model (Figure 1).

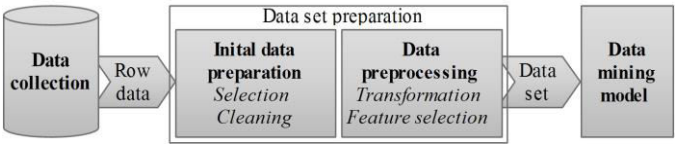

Figure 1. Research methodology

A. Data collection

In this research, we use the data set, collected from the student database of School of Electrical and Computer Engineering Applied Studies. The data set contains previous education and exams data for two generations (2018/2019 and 2019/2020) of students on the New Computer Technologies study program.

The data set include 131 instances with 23 input attributes (previous education data), points on the eighteen exams from the first three semesters. The following table shows the attributes used in this research (Table I).

TABLE I. ATTRIBUTE LIST

No	*Attribute*	*Values*
1	Average - I year of high school	0,0 - 5,0
2	Average - II year of high school	0,0 - 5,0
3	Average - III year of high school	0,0 - 5,0
4	Average - IV year of high school	0,0 - 5,0
5	Entrance Exam	0 - 60
6	Engineering Mathematics	0 - 100
7	German Language	0 - 100
8	English Language	0 - 100
9	Application Software	0 - 100
10	Computer Architecture and Organization I	0 - 100
11	Information Technology Fundamentals	0 - 100
12	Computer Architecture and Organization II	0 - 100
13	Digital Multimedia	0 - 100
14	Discrete Mathematics	0 - 100
15	Programming Fundamentals	0 - 100
16	Computer Graphics	0 - 100
17	Introduction to Object Programming	0 - 100
18	Database	0 - 100
19	WEB Design	0 - 100
20	Probability and Statistics	0 - 100
21	Programming Languages	0 - 100
22	Computers and Peripherals	0 - 100
23	Introduction to Internet Technology	0 - 100

B. Data set preparation

We use vertical data selection to remove attributes that are not relevant and may negatively affect the results, and horizontal data selection to remove instances that have no previous education data, or have data errors. Also, to clean data we delete all records who has abnormal distance from other values of variables.

After initial data preparation we use data transformation method (min-max normalization) to improve the accuracy of data mining algorithms. Normalization represents process of rescaling attributes to the range from 0 to 1. The normalization process is not necessary for all data sets when the values are uniform. It is preferred and used when the variable values are very different. The last step before applying data mining techniques in data preprocessing is selection of FS algorithms. The most common type of classification FS algorithms are filters, wrappers, embedded and hybrid methods [10]. In this research, we use four FS algorithms built in WEKA software:

- CorrelationAttributeEval (CAE) - is known as Pearson's correlation coefficient and is widely used in statistics. Used to measure the correlation between every attribute and the target attribute 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 [11].
- GainRadioAttributeEval (GRAE) - measures the significance of attributes with respect to target class on the basis of gain ratio [12].

- InfoGainAttributeEval (IGAE) - is a feature evaluation method based on the entropy method and is widely used in machine learning. It is calculated by how much the term can be used to classify information in order to measure the importance of the lexical units of the classification [12].
- ReliefAttributeEval (RAE) - This method randomly selects an instance and its value and compares it with the nearest neighbors to find a relevance score for each attribute. The algorithm tries to create a list of attributes that can differentiate between instances from the class labels [13]. Relief calculates a proxy statistic for each feature that can be used to estimate feature 'quality' or 'relevance' to the target concept, i.e. feature weights [14].

The threshold for attribute selection after applying FS algorithms according to previous research was between 0.01 [5] and 0.5 [6]. In this research, a threshold of 0.05 was applied and all attributes with a lower threshold were considered irrelevant (Table II).

C. Data mining model

Data Mining model was created in WEKA software. We use multilayer perception (artificial neural network algorithm) to predict student success on Visual Programming Techniques subject. Multilayer perception is one of the most used techniques in educational data mining, which can produce a more accurate classification because it has better weight characteristics than other modeling [15]. Multilayer perception has an input layer to receive the input and the output layer to make the decision about the input. The computational engine of multilayer perception consists of hidden layers which are capable of approximating any type of continuous function [16].

Multilayer perception selected parameters are: Learning rate: 0.3; Momentum: 0.2.

IV. Results and Discussion

After applying the selected FS algorithms (CorrelationAttributeEval, GainRadioAttributeEval, InfoGainAttributeEval, ReliefAttributeEval), it is possible to identify subjects (from the first three semesters) that are most influential in determining the success of students in the Visual Programming Techniques course (fourth semester).

According to Table II, the largest number of the most influential subjects (19) were identified using the CorrelationAttributeEval algorithm. When using ReliefAttributeEval, the 9 most influential subjects with a threshold greater than 0.05 were identified. When using the InfoGainAttributeEval algorithm, the 8 most influential items with a threshold greater than 0.05 were identified. The lowest number of subjects with a threshold greater than 0.05 was identified using the GainRadioAttributeEval algorithm, 7 in total.

TABLE II. Attribute rank after applying FS algorithms (treshold greater than 0.05)

No	*Attribute*	*CAE*	*GRAE*	*IGAE*	*RAE*
1	Average - I year of high school	0.4	0.00	0.00	0.02
2	Average - II year of high school	**0.16**	0.00	0.00	0.03
3	Average - III year of high school	**0.21**	0.01	0.01	**0.05**
4	Average - IV year of high school	**0.16**	0.03	0.03	0.03
5	Entrance Exam	**0.18**	0.00	0.00	0.03
6	Engineering Mathematics	**0.26**	0.03	0.02	**0.34**
7	German Language	**0.11**	0.00	0.00	0.01
8	English Language	**0.13**	0.00	0.00	0.02
9	Application Software	**0.07**	0.00	0.00	0.02
10	Computer Architecture and Organization I	**0.09**	**0.23**	**0.22**	**0.07**
11	Information Technology Fundamentals	**0.14**	**0.36**	**0.16**	0.02
12	Computer Architecture and Organization II	**0.09**	**0.22**	**0.17**	**0.05**
13	Digital Multimedia	**0.12**	0.00	0.00	**0.06**
14	Discrete Mathematics	**0.23**	0.00	0.01	**0.07**
15	Programming Fundamentals	**0.40**	**0.32**	**0.30**	**0.16**
16	Computer Graphics	0.02	0.04	0.03	0.04
17	Introduction to Object Programming	**0.38**	**0.44**	**0.38**	**0.06**
18	Database	**0.29**	**0.28**	**0.28**	0.04
19	WEB Design	0.04	0.00	0.00	0.00
20	Probability and Statistics	**0.16**	0.00	0.00	0.03
21	Programming Languages	**0.29**	**0.45**	**0.34**	**0.12**
22	Computers and Peripherals	0.04	0.02	0.01	0.03
23	Introduction to Internet Technology	**0.21**	0.00	**0.20**	0.04

This research focuses on a comparative analysis of four FS algorithms using an educational institution data set (131 instances, 23 input attributes).

Based on the obtained results, the performance of the observed four FS algorithms implemented in the WEKA software is presented. The following table (Table III) shows the accuracy results for different FS algorithms after applying multilayer perception.

TABLE III. Accuracy for diferent FS algorithms

	CAE	*GRAE*	*IGAE*	*RAE*
Correctly Classified Instances (%)	50.3817	55.7252	57.2519	58.7786
Incorrectly Classified Instances (%)	49.6183	44.2748	42.7481	41.2214
Kappa statistic	0.2139	0.2474	0.2493	0.3033
Mean absolute error	0.1388	0.1457	0.1372	0.1332
Root mean squared error	0.3262	0.3111	0.3041	0.3032

The following table shows the average values for TP Rate, FP Rate, Precision, Recall, F-measure for different FS algorithms (table IV).

TABLE IV. ACCURACY FOR DIFERENT FS ALGORITHMS

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*
CorrelationAttributeEval (CAE)	0.504	0.276	0.509	0.504	0.504
GainRadioAttributeEval (GRAE)	0.557	0.301	0.480	0.557	0.514
InfoGainAttributeEval (IGEA)	0.573	0.315	0.492	0.573	0.534
ReliefAttributeEval (RAE)	0.588	0.271	0.559	0.588	0.561

The obtained results indicate that ReliefAttributeEval is the most suitable for predicting the success of students in combination with the multilayer perceptron method.

The following figure (Figure 2) shows the comparison of four FS algorithms CorrelationAttributeEval (CAE), GainRadioAttributeEval (GRAE), InfoGainAttributeEval (IGEA), ReliefAttributeEval (RAE) with respect average TP Rate, FP Rate, Precision, Recall, F-measure.

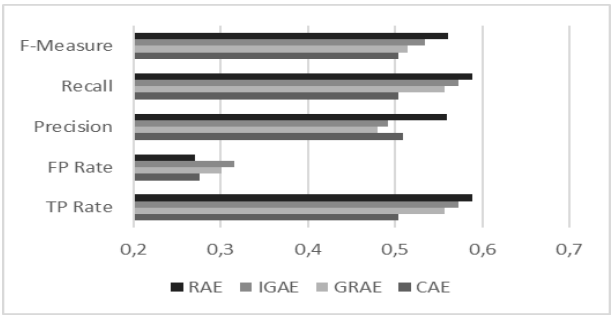

Figure 2. Feature selection algorithms matrics

Confirmation of the obtained results using the ReliefAttributeEval algorithm was done by changing the thresholds and changing the number of the most significant attributes. After applying different thresholds during the selection of appropriate attributes, the most correctly classified instances were obtained when using the 0.05 threshold, which is in accordance with the presented methodology.

V. CONCLUSION

This paper presents a survey of four different FS algorithms implemented in WEKA software. By applying FS algorithms, the initial data set (row set) was preprocessed, which led to a reduction in the number of attributes that were not relevant in determining student success.

The best results for student success were obtained using the ReliefAttributeEval algorithm in combination with the multilayer perception method. The obtained results confirmed the presented methodology. By choosing the appropriate FS algorithm, the most favorable result was determined, which was the main goal of the work. The comparison of the results of correctly classified instances with other researches is not satisfactory, which indicates a small number of attributes in the data set. As a weakness of the model, it is possible to identify a small number of attributes and instances, which caused low accuracy during classification. Also, attribute values (Table I) below 51 are not relevant, which also caused an error, which in further work should be defined through evaluation or such instances removed from the data set.

Future research would be reflected in the development and implementation of a new model for predicting student success that would include data related to the behavior of users on the electronic learning system, demographic data, interests, motivation, and learning styles. Also, future work would be based on improving the accuracy of determining success, as well as taking a larger number of instances, which would affect the accuracy.

REFERENCES

[1] C. Romero, J. R. Romero, S. Ventura, "A Survey on Pre-Processing Educational Data", Educational Data Mining. Studies in Computational Intelligence, vol 524, pp. 29-64, Springer, Cham, 2013.

[2] B. M. Olukoya, "Comparison of Feature Selection Techniques for Predicting Student's Academic Performance", International Journal of Research and Scientific Innovation (IJRSI), vol. 7, pp. 97-101, 2020.

[3] E. Alyahyan, D. Düştegör, "Predicting academic success in higher education: literature review and bestpractices", International Journal of Educational Technology in Higher Education , vol. 17, 2020.

[4] M. Zaffar, K. S. Savita, M. A. Hashmani, S. S. Hausain, "A Study of Feature Selection Algorithms for Predicting Students Academic Performance", International Journal of Advanced Computer Science and Applications, vol. 9, pp. 541-549, 2018.

[5] E. Osmanbegović, M. Suljić, H. Agić, "Determining Dominant Factor For Students Performance Prediction By Using Data Mining Classification Algorithms", Transition, vol. 17, pp. 147-158, 2014.

[6] N. Stanković, M. Blagojević, M. Papić, D. Karuović, "Artificial Neural Network Model for Prediction of Students' Success in Learning Programming", Journal of Scientific & Industrial Research, vol. 80, pp. 249-254, 2021.

[7] A. Triayudi, I. Fitri, "Comparison of the feature selection algorithm in educational data mining", TELKOMNIKA Telecommunication, Computing, Electronics and Control, vol. 19, pp. 1865-1871, 2021.

[8] C. Jalota, R. Agrawal, "Feature Selection Algorithms and Student Academic Performance: A Study", International Conference on Innovative Computing and Communications, Advances in Intelligent Systems and Computing, vol 1165. Springer, pp. 317-328, 2020.

[9] A. O. Enaro, S. Chakraborty, "Feature Selection Algorithms For Predicting Students Academic Performance Using Data Mining Techniques", International journal of scientific & technology research, vol. 9, pp. 3622-3626, 2020.

[10] A. Jović, K. Brkić, N. Bogunović, "A review of feature selection methods with applications", International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Croatia, pp. 1200-1205, 2015.

[11] http://infochim.u-strasbg.fr/cgi-bin/weka-3-9-1/doc/weka/attributeSelection/CorrelationAttributeEval.html (accessed: 15.01.2023.)

[12] S. Gnanambal, M. Thangaraj, V. T. Meenatchi, V. Gayathri, "Classification Algorithms with Attribute Selection:An Evaluation Study using WEKA", International Journal of Advanced Networking and Applications, vol. 9, pp. 3640-3646, 2018.

[13] C. ArunKumar, M. P. Soraj, S.Ramakrishnan, "A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets", International Conference on Advances in Computing & Communications, vol. 115, pp. 209-2017, 2017.

[14] Z. Cirović, N. Ćirović, "A Starcraft 2 Player Skill Modeling", Young Business and Industrial Statisticians Workshop on Recent Advances in Data Science and Business Analytics, pp. 121-128. 2019.

[15] A. Triayudi, W. O. Widyarto, "Educational Data Mining Analysis Using Classification Techniques", Journal of Physics: Conference Series, vol. 1933, 2021.

[16] G. Sinthia, M. Balamurugan, "Analyzing Student's Academic Performance Using Multilayer Perceptron Model", International Journal of Recent Technology and Engineering (IJRTE), vol. 7, pp. 156-160, 2019.