IJEEC

International Journal of Electrical Engineering and Computing
Vol. 5, No. 2 (2021)

Case Study
UDC 330.341.1:[004.738.5:316.323.6
DOI 10.7251/IJEEC2102094B

Predicting Study Programme Selection with Data Mining Classification Technique

Rade Božić

Faculty of Business Economics, University of East Sarajevo, Republic of Srpska, Bosnia and Herzegovina

E-mail address: rade.bozic@fpe.ues.rs.ba

Abstract— **The application of data mining in the field of education (Educational Data Mining - EDM) is becoming more and more popular. Predicting final grades during studies, measuring student and lecturer performance, targeting students, curriculum improvement, are just some of the examples that can support the development of this area. The focus of this article is on the prediction of the study programme that students will select during their higher education at the Faculty of Business Economics in Bijeljina. The analysis was conducted on the data of the faculty wherein the first two years students attend the same courses, while at the beginning of the third year they select a specific study programme. The aim of this paper is to use classification methods to predict the selected study programme based on the final grades achieved on courses during the first two years of study. The highest accuracy was obtained using random forest algorithm (59,94%). Model evaluation results show that choice of study programme does not depend only on the success achieved in all courses during the first two years of study. The analysis was performed using open-source WEKA mining tool, and the obtained results were presented and interpreted.**

Keywords: data mining, classification, prediction, study programme

I. INTRODUCTION

In modern society, a large amount of data is stored in various information systems. In order to use their maximum potential, the goal is to extract usable information that can help in decision-making process. Data Mining is the science of collecting, preparing, processing, analyzing, and extracting usable information from data sets [1]. It can be applied in various areas such as medicine, biology, transport, telecommunications, meteorology, engineering, art, etc.

The application of data mining in the field of education (EDM) stands out as a special domain. It is defined as the process of transforming raw data from information systems in education into usable information, which can be used by lecturers for corrective activities or providing answers to research questions [2]. It provides support in education development through the analysis of study programmes and curriculum, evaluation of the performance of lecturers and students, predicting success during studies, targeting students, and a large number of other examples of application, especially in higher education institutions.

Data mining can also be used to predict the choice of the study programme. This allows educational institutions to see the advantages and disadvantages of the current curriculum, as well as changes that are required in the current development strategy. It can also provide assistance in planning the way that teaching is conducted, the teaching staff that needs to be hired, the work profiles that will join the labor market, the regulation of the enrollment quota, etc. In this paper, classification methods of data mining will be applied to predict the choice of study programme at the Faculty of Business Economics in Bijeljina. The first two years of study at a given faculty are the same for all students, while at the beginning of the third they select a specific field of study: 1. finance, banking and insurance (FBI), 2. foreign trade, taxes and customs (FTTC) and 3. business informatics (BI). The data set used in this paper includes the final grades of students in courses during the first two years of study, which also represents the attributes of the set. The paper is organized as follows: 2. Classification technique in data mining 3. Educational Data Mining, 4. Related work, 5. Data set, 6. Analysis process, 7. Results, 8. Results comparison 9. Conclusion.

II. CLASSIFICATION TECHNIQUE IN DATA MINING

Classification is a supervised learning data mining technique. It can be applied in a lot of different areas such as economy and business, medicine, biology, transport, meteorology, art, education, etc. The idea is to put objects into different classes based on their characteristics. Data classification has two steps in its execution, learning (model constructing) and classification.

Due to this, the data set is split into training and testing parts. The training part is used for the learning process where the algorithm builds a classifier by analyzing training data [3].

Classifier recognizes patterns among data and classifies them into different categories or classes (labels). The next step is to use the test part of the data set to classify unseen data. After this, an evaluation of the model is done and it shows the performance of the classifier [4]. The most used classification methods are Naïve Bayes, decision trees, neural networks, k-nearest neighbors (k-NN), support vector machines (SVM), linear classifiers, etc.

III. EDUCATIONAL DATA MINING – EDM

Data mining in education is becoming a popular research topic. Some authors believe that it is the science of learning, as well as a wide area suitable for the application of data mining. Reason for this is the growth and availability of data in education. It enables data-based decision-making to improve current educational practice and teaching materials [5].

The focus of this type of mining is the development of methods for analysing unique data of the educational context. They are collected from various sources such as classic face-to-face lectures, educational software, online courses and various types of tests. The knowledge gained in this field does not only help lecturers and organizers of educational activities, but also students [6].

Nabila and Idriss [7] believe that data mining in education can be applied to:

- performance assessment and student guidance during the learning process,
- collecting feedback and adapting learning process based on student behaviour,
- improving the learning process,
- evaluation of teaching material,
- detection of problems and non-standard behaviours in learning,
- as well as for a better understanding of the phenomenon of education.

Data mining in education can be considered as a combination of three main areas: computer science, education, and statistics [8]. As an interdisciplinary field, it applies methods and techniques from statistics, machine learning, data mining, recommendation systems, data retrieval, psycho-pedagogy, cognitive psychology [7], etc.

Few authors [9] [10] proposed a model for applying EDM, similar to other data mining application processes. It starts with providing data with educational context. After that, chosen data is prepared and cleaned, often with help of different data mining techniques. Some authors suggest that an information system for collecting this data could help in skipping this step with built-in tools and methods. The next step is applying appropriate data mining techniques for EDM. The last step is the evaluation and interpretation of obtained results from the preview step (Fig. 1) [11].

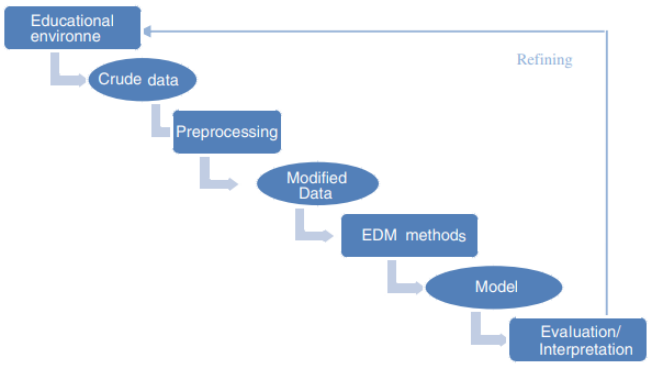

Figure 1. Proposed model for applying EDM [11]

IV. RELATED WORK

Prediction in EDM is one of the most common tasks in this research field. The subject of forecasting can be the performance of students and lecturers, grades in certain courses, success in taking the entrance exam, selection of the academic programme and field of study, duration of studies, probability of dropping out of studies, as well as many other cases.

The subject of research of this paper is the prediction of the study programme which students at the Faculty of Business Economics in Bijeljina will enroll. The teaching process is organized in such a way that students attend the same courses for the first two years. At the beginning of the third year, they choose one of three different study programmes. This kind of organization enables a unique analysis that refers to the prediction of the selection of study programme based on the final grades that the student has achieved in the courses during the first two years of study.

A large number of higher education institutions gives their students the opportunity to select a specific study field or programme at the beginning of their studies. For this reason, it is not possible to make predictions based on success in individual courses during higher education, which makes this analysis unique. This does not mean that it can be applied only at higher education institutions, but also at other levels of education.

A similar problem was described by Ghani, Cob, Drus and Sulaiman [12]. They were predicting study programme enrollment in higher education institutions in Malaysia. Their work was based on a comparison of three classification methods: logistic regression, Naïve Bayes method and decision tree. The attributes used in the analysis included data related to the student's nationality, parent employment status, location of the desired higher education institution, type of application, offered study area intended for the student, matching of the desired and offered area of study, department and enrolment status. Data on monthly income, previous education, as well as employment status were omitted from the analysis due to a large number of missing values. The aim was to predict students' orientation to study information technology, engineering or business management. The results obtained showed that the highest average accuracy was provided by the decision tree method (71%).

Ezz and Elshenawy [13] worked on the development of a system for recommending an educational path for students at Al Azhar University of Engineering. They developed a model

which suggests one engineering department for each student based on the performance achieved in the preparatory period (the period after high school, and before enrollment). Data for each study programme (engineering department) are stored in the system, and after that appropriate set of attributes and algorithm that provides the best results for each programme individually is selected. The data set included students who graduated in the period from 2012 to 2018. It consisted of two parts: the first part included grades from all courses that the student attended in the preparatory period, while the second part included final grades after graduation. They came to the conclusion that random forest provides the best results when recommending the Department of Mechanics, the k-nearest neighbor method is the best algorithm for the Department of Architecture, the linear regression algorithm achieves the best results at the Mining Department, etc.

Wanjau, Okeyo, and Rimiru [14] proposed a model which predicts the choice of academic study field at higher education institutions. The data used in the research were collected by a survey at Dedan Kimathi University of Technology. Attributes were arranged according to the following categories: interests and motivations, academic qualifications, educational context, and socio-demographic indicators. The authors determined that the final grade from high school, teacher inspiration, career flexibility, student orientation and grade in mathematics are the attributes that have the most influence on the choice of the study programme. The goal was to predict which academic study field students will choose: science, technology, engineering or mathematics. The best results were obtained by a classifier based on the decision tree with an accuracy rate of 85.2%.

Al-Radaideh, Ananbeh and Al-Kabi [15] proposed a model which helps high school graduate students to choose a suitable study programme. Data used in this paper was obtained from six schools in Mafraq in Jordan and consists of 248 instances. Selected attributes are the average grades from 10^{th}, 9^{th} and 8^{th} class and minimum grade acceptable for each study path. The study path represents the label which has one of four values: Science, Management, Academic and Profession. The method used in analysis process was C4.5 decision tree algorithm. Overall accuracy of the proposed model was 87.9%. Authors concluded that this model can help school management to determine the suitable study path for graduate students.

Jirapanthong [16] proposed a classification model which can help students to select undergraduate programme in private universities in Thailand. Data set consists of 7778 records with attributes like student's gender, major programme and GPA in high school, student's education type and label which represents faculty that student has enrolled. The decision tree algorithm was used in analysis process. After two test cases, the percentage of correctly classified instances was 87,3%. Focus of this research project was on the study of influencing factors on academic success of undergraduate students.

Mulugeta and Borena [17] suggested model which determines number of students' enrolment at different university departments like Medicine, Chemistry, Management etc. Data set included historical data from different government and private universities. Variables used in study are population count of the city, the strength of the economy and students' demographic information like age, gender, financial situation. Three methods were compared (Neural Network, Bayesian Classifier and Decision tree) and the best results were obtained by Multilayer Perceptron Neural Network.

V. DATA SET

The analysis described in this paper was conducted on a data set of the Faculty of Business Economics in Bijeljina. Before the analysis process, it is necessary to collect, prepare and check the data. The data set used for the analysis was formed from three different tables from the original database. The courses with their ID number were extracted from the first database table (Table 1). The student id number was extracted from the second database table together with the enrolled study program, while the third database table contained course grades. There were no instances with missing or inconsistent attribute values. The data set was integrated into the final table (CSV format) using Excel. After that, WEKA open-source mining tool was used for analysis process.

TABLE I. DATA SET ATTRIBUTES (COURSES IN FIRST TWO YEARS OF STUDY)

Course name	Course ID
Basic Economics	585
Sociology	584
Accounting	586
Business Informatics	587
Mathematics for Economists	588
Business Statistics	589
Enterprise Economics	590
English Language 1	591
Microeconomics	592
Monetary Economics	593
Financial Mathematics	594
Business Law	595
Management	596
Marketing	597
Business Finance	598
English Language 2	599

The final data set has 16 attributes which values represent the grades of graduates in the courses taken during the first two years of study, and a label that represents the selected study programme. It has 654 instances, i.e., graduated students who enrolled in the third year of study in the period from 2010 to 2018. All of the attributes were used in the analysis procedure, i.e., no method was applied to select the attributes that have the greatest impact on class prediction.

TABLE II. SAMPLE FROM THE DATA SET

592	593	594	595	596	597	598	599	Study programme
8	10	9	6	9	9	10	10	FBI
8	8	7	7	6	9	9	10	FBI
6	7	7	6	8	9	8	9	BI
6	6	7	7	6	7	7	6	FTTC
6	6	6	7	6	7	10	6	FTTC
7	6	8	6	7	7	7	9	BI
6	6	6	8	6	7	9	6	BI

Students' grades are expressed numerically, in the interval from 6 to 10. The column in which the study programme is recorded is expressed as a nominal variable with three possible values that represent the abbreviated name of the programme

(FBI, FTTC, BI). The student id number is not included in the set, because it does not provide any significance in the analysis. An example of instances from the set over which the analysis was performed is given in table 2.

Most students enrolled in the FTTC study programme (342 or 52.29%), followed by the FBI (241 or 36.85%) and finally the BI (71 or 10.86%) (Fig. 2).

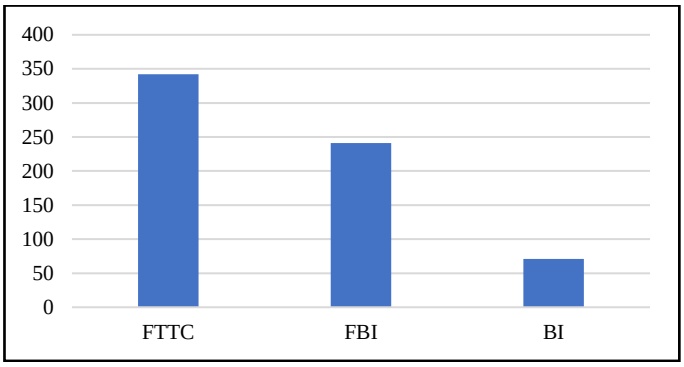

Figure 2. Distribution of classes in data set

VI. ANALYSIS PROCESS

This chapter describes the tools used for the analysis process, the method of dividing the set into a part for training and testing, model evaluation and used classification methods.

A. WEKA

The software open-source tool WEKA (Waikato Environment for Knowledge Analysis) was used in the analysis process. It is a set of machine learning algorithms for data mining and it also contains tools for data preparation, classification, regression, clustering, association rules and visualization. It was developed by the University of Waikato in New Zealand. The used version is 3.8.5. Model evaluation data were collected from this software [18].

B. K-fold Cross validation

The k-fold cross-validation method was used to divide the set into training and testing parts. If a data set consists of N instances, they are divided into k equal parts (subsets or folds). If N is not divisible by the number k, then the last subset contains fewer instances than the other subsets. Then the training and testing process is performed k times. Each individual subset is used for testing, while the other k-1 subsets are used for training. The total number of correctly classified instances (through overall execution) is divided by the total number of instances N, in order to obtain a general level of prediction accuracy [19]. This method is applied in order to reduce the bias and the possibility of model overfitting. It has been empirically proven that when k has a value of 5 or 10, an estimated error rate is not influenced by high bias or large variance [20]. According to the recommendation of the WEKA tool, the number of folds (k) in this paper is 10, while the testing process is performed 11 times (10 for each individual fold and another test on the whole data set).

C. Model evaluation

Four measures were used to assess model evaluation: accuracy, precision, recall, and F1-score. The calculation of these indicators is based on a confusion matrix. In binary classification, i.e., classification in which there are only two classes (positive and negative), the confusion matrix consists of (Table 3):

- True Positives (TP) - the number of positive instances that model has classified as positive.
- True Negatives (TN) - the number of negative instances that model has classified as negative.
- False Positives (FP) - the number of negative instances that model has classified as positive.
- False Negatives (FN) - the number of positive instances that model has classified as negative.

TABLE III. BINARY CLASSIFICATION CONFUSION MATRIX

Correct classification	Classified as	
	+	-
+	true positives (TP)	false negatives (FN)
-	false positives (FP)	true negatives (TN)

Equations for model evaluation metrics:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F\ score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

There are three labels in this example. When calculating indicators based on given formulas, one class becomes positive, while the other classes together form a negative. Accuracy is not a reliable indicator of model evaluation in data sets in which there is an uneven distribution of classes. For that reason, precision, model recall and F1-score are used. Precision is the participation of exact positives in a set of positively classified instances. The recall (True Positive Rate - TPR) of the model shows the percentage of positives that are correctly classified. The F1-score combines the precision and recall of the model in one measure, i.e., it represents the harmonious mean of these two measures.

D. Classification methods

In order to predict the study programme enrolment, the following classification methods were used in the analysis process: Naïve Bayes, decision tree, nearest neighbour method and random forest. These methods are often used in EDM.

Naïve Bayes classifier is based on Bayes' conditional probability theorem. The algorithm combines a priori and conditional probabilities in one formula that can be used to

calculate the probabilities of each possible class. It is often written as:

$$P(c_i) \times \prod_{j=1}^{n} P\,(a_j = v_j |\, class = c_i) \qquad (5)$$

where $P(c_i)$ represents prior probability of the class c_i and $\prod_{j=1}^{n} P\,(a_j = v_j |\, class = c_i)$ indicates the product obtained by multiplying together the n values of $P(a_1 = v_1 |\, c_i), P(a_2 = v_2 |\, c_i)$ etc., where $a_1, a_2 \ldots a_n$ represents attributes and $v_1, v_2 \ldots v_n$ instance values of those attributes. After that, the class with the highest value is selected. Although it shows weaknesses in theory, this method provides good results in practice [19].

Decision tree is a method in which the classification process is modeled by a set of hierarchical decisions based on attributes, forming a tree-like structure. It consists of root node, internal nodes and leaf nodes. There are two phases in decision tree construction: building phase and pruning phase. In building phase, different attribute selection measures are used (Gini index, information gain, gain ratio, etc.). A condition on a particular tree node is a division criterion based on one or more attributes in the part of the set related to training, dividing it into two or more parts. Pruning phase eliminates subtrees in order to achieve better accuracy. Decision trees are human readable and easy to understand [1]. In this paper J48 algorithm will be used, which represents improved version of C4.5 algorithm.

K-nearest method starts from the idea that the class of an unseen instance is determined by the class of the corresponding instance or instances that are closest to it. The distance itself needs to be calculated in an appropriate way, and it is most often calculated using the Euclidean distance. The formula for Euclidian distance between points $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ in n-dimensional space is:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \qquad (6)$$

The algorithm finds the nearest instances and takes the class that occurs most frequently among them, and assigns it to an unseen instance [19].

Random forest can be used for classification and regression problems. This method consists of a large number of individual decision trees that function as an ensemble. Randomness is added in building process of every tree. Each individual tree in a random forest predicts the class, and the class that has the most votes, that is, the most times predicted by the trees, is taken as the model prediction. This algorithm is a modification of the bagging method that involves training each individual tree on a different subset of data [21].

VII. RESULTS

Classification methods were applied in the analyzing process using WEKA software tool. The obtained results are presented in this chapter. Confusion matrix, precision, recall and F1-score of selected methods are shown in a table form.

A. Naïve Bayes

The confusion matrix obtained using the Naïve Bayes method is shown in Table 4. Table 5 shows the precision, recall, and F1-score of the model by individual classes.

TABLE IV. *NAÏVE BAYES* CONFUSION MATRIX

Correct classification	Classified as		
	FTTC	**FBI**	**BI**
FTTC	**280**	62	0
FBI	130	**108**	3
BI	46	22	**3**

TABLE V. *NAÏVE BAYES* PRECISION, RECALL AND F1-SCORE

Label	Precision	Recall	F1-score
FTTC	**0,614**	**0,819**	**0,702**
FBI	0,563	0,448	0,499
BI	0,500	0,042	0,078

The accuracy of *Naïve Bayes* method is **59.79%**. The highest precision (61.4%) had the FTTC class, while the same class had the highest percentage of recall (81.9%) and F1-score (70.2%). The BI class had the lowest precision (50%), recall percentage (4.2%) and F1-score (7.8%).

B. Decision tree

As in the previous example, decision tree confusion matrix is presented in Table 6, while evaluation of the model is shown in Table 7.

TABLE VI. DECISION TREE CONFUSION MATRIX

Correct classification	Classified as		
	FTTC	**FBI**	**BI**
FTTC	**237**	89	16
FBI	125	**100**	16
BI	39	23	**9**

TABLE VII. DECISION TREE PRECISION, RECALL AND F1-SCORE

Label	Precision	Recall	F1-score
FTTC	**0,591**	**0,693**	**0,638**
FBI	0,472	0,415	0,442
BI	0,220	0,127	0,161

The accuracy of decision tree method is **52.91%**. The highest rate of precision (59.1%), recall (69.3%) and F1-score (63.8%) had the FTTC class, while the lowest precision (22%), recall (12.7%) and F1-score (16.1%) had the BI class. For the root of the tree, the algorithm chose financial mathematics, while it had a total number of 72 leaves. The information gain measure was used for attribute splitting.

C. K-nearest neighbor method

The confusion matrix of the 5-NN method is presented in Table 8, while evaluation metrics are shown in Table 9.

TABLE VIII. 5-NN CONFUSION MATRIX

Correct classification	Classified as		
	FTTC	FBI	BI
FTTC	**286**	52	4
FBI	156	**82**	3
BI	46	22	**3**

TABLE IX. 5-NN PRECISION, RECALL AND F1-SCORE

Label	Precision	Recall	F1-score
FTTC	**0,586**	**0,836**	**0,689**
FBI	0,526	0,340	0,413
BI	0,300	0,042	0,074

In this method *k*=5. From all tested values, this one provided the best model evaluation results. The accuracy achieved by applying the k-nearest neighbor (5-NN) method is **56.73%.** Euclidean distance was used in the calculation. As in the previous two cases, again the FTTC class had the highest accuracy (58.6%), recall (83.6%) and F1-score (68.95%), while the BI class had the lowest accuracy (30%), recall (4.2%) and F1-score (7.4%).

D. Random forest

Confusion matrix of the random forest method is shown in Table 10, while evaluation metrics are shown in Table 11.

TABLE X. RANDOM FOREST CONFUSION MATRIX

Correct classification	Classified as		
	FTTC	FBI	BI
FTTC	**274**	65	3
FBI	127	**111**	3
BI	41	23	**7**

TABLE XI. RANDOM FOREST PRECISION, RECALL AND F1-SCORE

Label	Precision	Recall	F1-score
FTTC	**0,620**	**0,801**	**0,699**
FBI	0,558	0,461	0,505
BI	0,538	0,099	0,167

The accuracy of the model that used the random forest method in the analysis procedure is **59.94%**. FTTC class had the highest precision (62%), recall (80.1%) and F1-score (69.9%), while BI had the lowest precision (53.8%), recall (9.9%) and F1-score (16.7%).

VIII. RESULTS COMPARISON

This chapter summarizes and compares all obtained performance results from the analysis process. Each measure of model evaluation is individually presented in a table form. Precision, recall and F1-score are grouped by labels and classification methods.

A. Accuracy

Table 12 and Fig. 4 show the accuracy of the methods used in the analysis process.

TABLE XII. ACCURACY COMPARISON OF DIFFERENT METHODS

Method	Accuracy (%)
Naïve Bayes	59,79
Decision tree	52,91
5-NN	56,73
Random forest	**59,94**

The accuracy of the methods according to the selected attributes ranged from 50% to 60%. Random forest method had the highest accuracy of **59.94%.** Decision tree had the lowest accuracy of 52.91%. Naïve Bayes and random forest methods provided similar accuracy results (59-60%) (Fig. 3). As expected, the random forest method provided better results than the decision tree algorithm. Although random forest algorithm provides the best results, 40,06% of students were not correctly classified, what makes almost 2/5 of test set. These results show that all models should be improved. Because of the uneven distribution of classes, accuracy measure is not reliable, so it is necessary to calculate other evaluation measures.

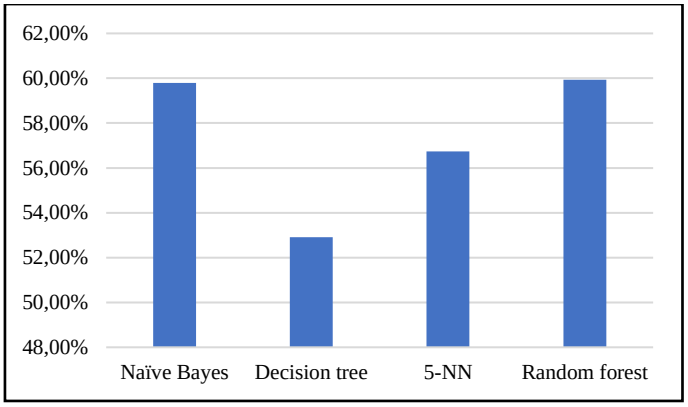

Figure 3. Methods accuracy comparison

B. Precision

Table 13 shows the summary of precision results.

TABLE XIII. PRECISION COMPARISON OF DIFFERENT METHODS

Label	Naïve Bayes	Decision tree	5-NN	Random forest
FTTC	0,614	0,591	0,586	**0,620**
FBI	**0,563**	0,472	0,526	0,558
BI	0,500	0,220	0,300	**0,538**

The highest level of precision for the FTTC (62%) and BI (53.8%) class was provided by the random forest method, while for the FBI class the highest precision was provided by the Naïve Bayes method (56.3%). The lowest precision in the FBI (47.2%) and BI (22%) classes was provided by the decision tree, while in the FTTC class the 5-NN method had the lowest precision (58.6%). The lowest precision of all models was in the BI class, while the highest precision was achieved in the FTTC class. It should be noted that with the original data set,

the situation is identical with the class distribution. While analysing the best results from this evaluation measure, only 62% of total predicted FTTC students were correctly classified. In the case of another two classes, percentage level is even lower. The random forest method again had the most success in this measure, but its level is not significant. Like accuracy, this validation measure shows that models didn't provide good results.

C. Recall

Table 14 shows the recall of the models according to the methods.

TABLE XIV. RECALL COMPARISON OF DIFFERENT METHODS

Label	Naïve Bayes	Decision tree	5-NN	Random forest
FTTC	0,819	0,693	**0,836**	0,801
FBI	**0,448**	0,415	0,340	0,461
BI	0,042	**0,127**	0,042	0,099

The highest percentage of model recall for the FTTC class was provided by the 5-NN method (83.6%), for the FBI Naïve Bayes (44.8%), while for the BI it was the decision tree (12.7%). The lowest percentage of recall for the FTTC class was provided by the decision tree (69.3%), for the FBI the 5-NN method (34%), while for the BI class the same percentage was achieved by Naïve Bayes and 5-NN (4.2%). Again, all models had the lowest recall in the BI class, while in the FTTC they had the highest. Analysing the best results, 16,4% of FTTC, 55,2% of FBI and 87,3% of BI students were incorrectly classified. Although FTTC class had good result with 5-NN algorithm, other two classes, especially BI, had very low recall percentage with the same method. This time random forest algorithm didn't provide the highest percentage in any class.

D. F1-score

Table 15 shows the F1-score of different methods. For the FTTC class, the best F1-score was provided by the Naïve Bayes method (70.2%). For FBI (50.5%) as well as BI (16.7%) random forest provided the best results. The lowest F1-score in the FTTC class was provided by the decision tree (63.8%). In the FBI (41.3%) and BI (7.4%) classes, the 5-NN method had the lowest F1-score. BI had the smallest F1-score, while the FTTC class had the highest. The random forest algorithm had the best results with two classes (BI, FBI), but very low percentage level on both of them. Low level of F1-score indicates that analysed models need improvement.

TABLE XV. F1-SCORE COMPARISON OF DIFFERENT METHODS

Label	Naïve Bayes	Decision tree	5-NN	Random forest
FTTC	**0,702**	0,638	0,689	0,699
FBI	0,499	0,442	0,413	**0,505**
BI	0,078	0,161	0,074	**0,167**

E. Summarizing evaluation measures results

The accuracy, precision, recall and F1-score of the used methods are not high. Although model recall in the FTTC class shows a high percentage, in the other two classes (especially BI) this is not the case. The random forest method had the best results in accuracy, precision and F1-measure, but it was on the low percentage level. This indicates that all models need to be improved. For this reason, it can be concluded that total course grades are not the only factor influencing the choice of a specific study programme. In addition to this, there may be influence by other factors like high school of students, the area of their interest, time needed to complete course, information on the renewal of the school year, etc. Also, certain combinations of attributes could provide better results when it comes to model evaluation, which can be done in future work.

IX. CONCLUSION

In this paper, classification technique of data mining was used to predict the selection of specific study programme at the Faculty of Business Economics in Bijeljina. As the programme itself is selected at the beginning of the third year, the final grades from the courses completed in the first two years were attributes for an analysis process using four different methods: Naïve Bayes, decision tree, nearest neighbor method and random forest. Measures of the model evaluation show that the highest percentage of accuracy has the random forest method (59.94%). In addition to this measure, precision, recall and F1-score of models were also calculated, where the results for each class were presented separately. The random forest method had the best results in accuracy, precision and F1-measure, but it was on the low percentage level. This indicates that all models need to be improved. Every model has achieved the lowest precision, recall and F1-score in the business informatics programme. The reason for this is the lowest representation of a given class in the data set. Although the mentioned indicators were the highest in the programme of foreign trade, taxes and customs, their level is not significant because in other two classes are on the low level. It can be concluded that the choice of study programme doesn't depend only on the success achieved in all courses during the first two years of study. This creates an opportunity for future research that may include a number of other attributes, or be conducted with a particular combination of attributes used in this analysis. In addition, it is possible to apply other classification methods in order to achieve better results of model evaluation. This creates an opportunity for future research works.

REFERENCES

[1] C. C. Aggarwal, Data Mining - The text book, vol. 53, Springer International Publishing Switzerland 2015, 2015.

[2] K. Funatsu, New Fundamental Technologies in Data Mining, InTech, 2011.

[3] H. Hiawei, K. Micheline and P. Jian, Data Mining Concepts and Techniques, Elsevier Inc, 2011.

[4] C. C. Aggrawal, ata Mining - The text book, vol. 53, 2019.

[5] H. Cho, G. Gay, B. Davidson and A. Ingraffea, "Social networks, communication styles, and learning performance in a CSCL community," pp. 309-329, 2007.

[6] R. S. Cristóbal, M. Ventura and R. S. J. Pechenizkiy, Background. In Handobook of Educational Data Mining, Taylor & Francis Group, 2011.

[7] B. Nabila and B. Idriss, Which Contribution Does EDM Provide to Computer-Based Learning Environments? In Educational Data Mining Applications and Trends, Springer International Publishing Switzerland, 2014.

[8] S. Dawson, "Seeing the learning community: an exploration of the development of a resource for monitoring online student networking," *J. Educ. Technol.*, p. 41, 2010.

[9] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Trans. Syst. Man Cybern.*, p. 601–618, 2010.

[10] B. R. Sachin and S. M. Vijay, "A survey and future vision of data mining in educational field," *Proceedings of IEEE 2nd International Conference on Advanced Computing and Communication Technologies*, pp. 96-100, 2012.

[11] A. Peña-Ayala, Educational Data Mining: Applications and Trends, Studies in Computational Intelligence, 2014.

[12] N. L. Ab Ghani, Z. Che Cob, S. Mohd Drus and H. Sulaiman, "Student Enrolment Prediction Model in Higher Education Institution: A Data Mining Approach," *Lecture Notes in Electrical Engineering*, vol. 565, no. 1, 2019.

Rade Božić was born in Bijeljina, Bosnia and Herzegovina, in 1994. He received his B.Sc. and M.Sc. from the Faculty of Business Economics in Bijeljina, in 2013 and 2019. He is currently working toward the P.h.D. degree with the University of Novi Sad, Faculty of Economics in Subotica. He is currently a Teaching Assistant with the Faculty of Business Economics in Bijeljina. His research field is data mining.

[13] M. Ezz and A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student ' s best academic programme.," 2019.

[14] S. K. Wanjau, G. Okeyo and R. Rimiru, "Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions," 2016.

[15] Q. A. Al-Radaideh, A. Al Ananbeh and E. Al-Shawakfa, "A Classification Model for Predicting the Suitable Study Track For School Students," *IJRRAS*, 2011.

[16] W. Jirapanthong, "Classification Model for Selecting Undergraduate Programs," *Eighth International Symposium on Natural Language Processing*, 2009.

[17] M. Mulugeta and B. Borena, "Higher Education Students' Enrolment Forecasting System Using Data," 2015.

[18] "Weka 3: Machine Learning Software in Java," 2021. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/.

[19] M. Brammer, Principles of data mining, vol. 30, Springer-Verlag London, 2016.

[20] M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.

[21] L. Breiman, Random forests. Machine Learning, 2001, pp. 5-32.