Machine Learning Based Model for Predicting Student Outcomes

Dijana Oreški
Associate Professor of Data Mining
Faculty of Organization and Informatics
University of Zagreb
Varazdin, Croatia
dijana.oreski@foi.unizg.hr

Dora Zamuda
Student of Software Engineering
Faculty of Organization and Informatics
University of Zagreb
Varazdin, Croatia
dora.zamuda@foi.unizg.hr

Abstract

Machine learning provides various algorithms for application in different domains. In the educational field, huge volume of students' data is generated and machine learning algorithms serve as valuable tool for pattern identification in students' behavior. In this paper, CRISP DM standard for data mining is applied in the research with decision tree algorithm used for modelling on Croatian dataset to develop predictive models for students' outcomes prediction. Data set consisted of 264 students of largest Croatian university collected by online survey. The results prove that decision tree modelling achieves superior results in terms of high accuracy and reliability together with interpretability of tree structure and obtained rules in prediction of students' academic performance. This approach shows promise to be used in student success prediction in the universities in an automatic manner. Such model can be used to: (i) improve students' learning and develop personalized recommender systems for optimal learning paths, (ii) emphasize to professors most important determinants of students' academic success (iii) help management of higher education institutions to facilitate the provision of detailed student learning and adjust institutions strategies, (iv) automate adaptation of the course modules and faculty programs.

Keywords
Machine Learning, Decision Tree, CRISP DM, Academic Performance.

1. Introduction
Machine learning algorithms have great potential in mining of students' learning data. Sector of higher education have discovered potential positive impact of machine learning algorithms on the learning process and outcomes in order to move towards a university of the new era. Machine learning algorithms serve as a useful tool to educational policy makers for enhancing the quality of learning and teaching process. Having this in mind, application of machine learning algorithms could lead to a positive change and can have a positive impact to higher education institutions to find out solutions for their problems (Van Barneveld et.al., 2012). Thus, machine learning based predictive models can serve as valuable tool for the decision-making process (Peña-Ayala, 2014). A higher education institution takes students' academic performance as one of the criteria for students' evaluation. Often, their focus is on the students tracking to identify students with better academic performances (Soni et.al., 2018). This is one of the fields of machine learning application in educational domain. Student performance is critical factor for measuring learning results and defining learning activities. However, previous papers did not provide enough evidence of which factors affects students' performance and how students could make progress to perform better. To solve those challenges, we will apply machine learning algorithm decision tree on extensive dataset of students' characteristics in order to develop accurate and reliable models of student performance.

Within this paper we provide three contributions: (i)evaluation of decision tree as machine learning algorithm for development of student academic performance predictive models, (ii) a case study using a the trained model at Croatian universities, (iii) discussion of related issues, focusing on student motivation.
This paper is structured as follows. Second section gives review of related work in the given topic. Section 3 explains data and machine learning algorithm used in the empirical research. Section 4 provides main research results and findings, whereas section 5 concludes the paper and gives guidelines and recommendations for further research.

2. Literature Review

This section provides review of the previous scientific papers' usage of different machine learning algorithms on various types of educational data in predicting student academic performance. There are a number of relevant research papers in the domain of student academic performance prediction published so far. These studies mainly examined student's exam performance analysis in educational institutions. Their results have shown that one of the most successful strategies for students' academic success is identification of relevant success factors and their optimization for students.

Different authors have used various methods for predictive models' development. Kolo et.al. (2015) used a decision tree machine learning algorithm for prediction of academic performance. Authors used student's financial status, gender and motivation as predictors of academic performance. Their results indicated gender differences between students. Hamsa et.al. (2016) also used decision tree algorithm for predictive model development. They developed various models for each year of study. Their target group where computer science students. Raut and Nichat (2017) tried to identify how to motivate students of low socio demographic characteristics. They also used decision tree algorithm for model development. Olaniyi et.al. (2017) predicted students' performance by using decision tree algorithms. Decision tree served as tool for identification of dropouts and students of special attention. Hasan et. al (2018) investigated students' outcomes by means of decision tree machine learning algorithm. They used activity of students and academic data of students as input parameters to estimate the performance algorithm for discovery of students' performance. Hasan et. al. (2018) used grades as students' performance indicators. Sharma and Kumar (2016) used decision tree as machine learning algorithm for predictive model development based on observations about an item to inference about student's outcome.

Some of the other research papers used artificial neural networks as machine learning algorithm for predictive model development. Wankhede (2014) used multilayer neural network model. Several other authors used similar approach: e.g. (Binh and Duy, 2017), (Okubo et.al., 2017) and (Bendangnuksung Prabu, 2018). Bayesian approach was also employed in some of the previous research papers. Authors such as Olaniyi et. al (2017), Razaque et. al. (2019), Divyabharathi and Someswari (2021) used Naive Bayes approach for student's classification. Support vector machines algorithm was used in research papers of Raihana and Farah Nabilah (2021), Asogobon et.al. (2016), Pratiyush and Manu (2016), Alamri et.al. (2020), Elmannai et. al. (2018).

Different research papers considered different variables for modelling students' academic achievement. So far, there has not been an in-depth analysis aimed at identifying optimal input dataset considering different types of student related features (e.g. past student performance, engagement and demographic data) for achieving the most precise predictions (Tomasevic et.al., 2020). For instance, Sudani and Palaniappan (2019) used a combination of institutional, academic, demographic, psychological and economic factors to predict students' performances using a multi-layered artificial neural network (NN) to classify students' degrees into either a good or basic degree class. Attendance at timetabled sessions shown to be a good predictor of student achievements in the research of Fike and Fike (2008).

3. Methods

Research methodology applied here is based on the CRISP DM standard for data mining which imposes six steps of data analysis: domain understanding, data understanding, data preparation, modelling, evaluation and deployment. In domain understanding phase literature review was performed and results are presented in section 2. Data understanding and data preparation are presented through data description in the next section, whereas modelling phase is presented here. Evaluation of the model is explained through section 5.

Having in mind the goal of this paper, representative state of the art of machine learning algorithms is described in Section 2. Several algorithms were considered for implementation in the case of student academic performance prediction. At the end, decision tree algorithm is applied due to their simplicity and interpretability.

Decision tree provides a hierarchical way of presenting knowledge. Decision tree is a semantic tree who's each node is associated with a set of possible answers. Decision tree consists of nodes and branches. Each branch connects a parent node with a child. The root node is the name for the node that is at the top of the tree and there is no parent node for it, while the leaf is the last node, i.e. the one for which there is no child node. The leaves of the tree represent all possible solutions to a given problem, and are also called response nodes. All other nodes located in the tree are called decision nodes. Each decision node displays a question, and by answering it (with yes or no, or choosing among the values offered), the branch to be followed is selected.

Decision tree is classification technique which works at divide and conquer principle (Quinlan, 1986). Divide and conquer principle indicates dataset is divided progressively in smaller subsets. These subsets are based on the values of an input variable. Each input variable is select according to criterion of selection. This criterion tries to identify the attribute that gives the best separation of a data set into individual classes. For each subset a child node is created and the subset data is included in it. These steps are repeated on the data of the child nodes, until a criterion of determination is met. Decision tree model look like a tree in which each node is either a leaf (indicating the value of the class) or a decision node (indicating some test which needs to be performed on input variable). Each branch shows an outcome of the test. In this research, the goal is to create a predictive model of student success. With an aim to create a decision tree model, the configuration of the supervised learning model is performed and explained in the next section.

4. Data Collection

The data used in this research was collected through an online survey questionnaire. The survey was conducted in the academic year 2020/2021. This survey was completed by students who studied and enrolled for the first time in the first year of college not before 2010. Number of 264 respondents filled out the questionnaire. These characteristics indicate a specific target population for conducting the survey. The data contain characteristics of respondents (students'), their socio-demographic characteristics, background, habits of learning and motivation. The questions were personal questions about students, as well as those related to the choice of studies, students' habits, whether private or in college, questions related to the environment in which they are located. First, sociodemographic characteristics were asked, such as gender, age, name of the faculty, year of enrollment, reason for faculty selection and distance from the faculty. Following group of the questions was related to the study in which the students are, and this includes the current status of the student, whether student receives a scholarship, whether student ever dropped a year or one of the subjects, whether student used instructions, the amount of time devoted to learning and the average grade so far. Students were also asked about their living environment: place of residence, life of parents, siblings, partner and children. The last group of questions was related to student activities such as employment, playing sports, using mobile phones, playing games, watching movies / series, going out and consuming alcohol.

Collected data set consists of 26 attributes of which two are related to students` personal information, 12 to study, five to the environment and 7 to activities. Attributes are initially divided into two tables according to whether they are categorical or numeric. Description of the questions included in the questionnaire can be founded in table 1 (description of the categorical variables) and table 2 (description of the numerical variables). There are 24 categorical and 2 numerical attributes.

Table 1: Description of the categorical variables

Variable name	Description	Values	Mode
Gender	Students` gender	M, F	F
Faculty	Name of the faculty	Various values	Faculty of Organization and Informatics
Reason_of enrollment	Reason of enrollment	proximity, recommendation, reputation, desire, other	desire
Faculty_distance	Distance of the faculty from the place of residence	less than or equal to 1 hour drive, less than or equal to 5 hours drive,	less than or equal to 1 hour drive

		less than or equal to 10 hours drive, more than 10 hours drive	
Student_status	Status of the student	Full time student, part time student,	Full time student
Scholarship	Does student take scholarship	Yes, no	Yes
Dropout_year	Did student had dropout of certain academic year	Yes, no	No
Course_repetition	Did student have repetition on one of the courses	Yes, no	No
Instructions	Did student pay for the instructions	Yes, no	No
Paying for help	Paying for help at online exams (cheating)	Yes, no	No
Learning	Time spent on learning	Each day, once a week, before exam, other	before exam
GPA	Grade point average	1,0 – 1,49 1,5 – 2,49 2,5 – 3,49 3,5 – 4,49 4,5 – 5,0	3,5 – 4,49
Place of residence	Place of residence	Town, countryside	countryside
Parents	How do parents live	In marriage, separated, single parenting, other	In marriage
Sibilings	Number of brothers or sisters	More than 3, 3 or less, No sibilings	3 or less
Partner	Does the student have partner	married, engaged, in the relationship, not in a relationship	not in a relationship
Children	Does the student have children	Yes, no	No
Employment	Is student employed	Yes, full time Yes, part time student job No	Yes, part time student job
Sport	Whether a person engages in sports activities	Regulary, sometimes, never	sometimes
Cell_phone	Usage of cell phone on daily basis	Less than 1 hour, From 1 till 3 hours, From 3 till 5 hours, More than 5 hours no	From 1 till 3 hours
Computer_games	How often does student play computer games	often, sometimes, never	never
Movies/series	Does the student watches movies or series and how often	A lot, average, no	average
Going_out	How often does student going out	often, few times a week, once a week, no	once a week

Alcohol	Whether a student consumes alcohol	often, at weekends, sometimes, no	sometimes

Categorical variables are represented by mode as average value, whereas for numerical variables there are range of the values, minimum and maximum value, median, mean and standard deviation. See Table 2, where those descriptive data are presented.

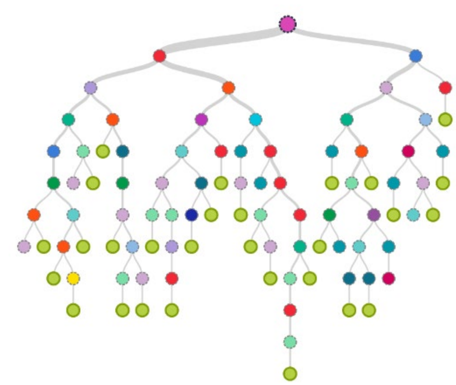
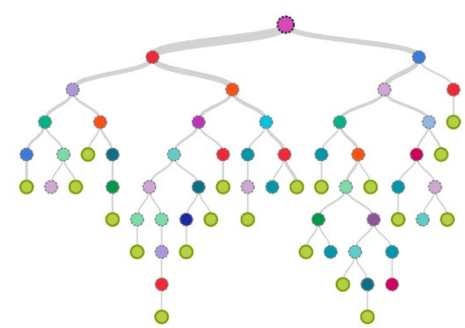
Table 2: Description of the numerical variables

Variable name	Description	Range of the values	Minimum	Maximum	Mean	Median	Standard deviation
Age	Student age	0-99	18	28	21,64	21	1,77
Year of enrollment	Year of enrollment at the faculty	2010-2021	2013	2021	2018,01	2018	1,54

5. Results and Discussion

The tool BigML is used in this paper to implement machine learning approach decision tree. This tool is cloud open source collection of many machine learning algorithms, covering whole process of knowledge extracion from data, including data preparation and selection, various classification algorithms and clustering algorithms. Hereinafter, we present results of decision tree modelling on our dataset. Results of the modelling are demonstrated through table 3. Process tree punning is performed wih an aim to developed optimal tree structure. Table 3 gives comparison of three different types of prunning applied for the development of desicion tree model. Columns 2 and 3 of table 3 provide measures for model quality. At the the end, column 4 gives visualization of each model.

Table 3. Decision tree models

Type of prunning	Reliabiliy	Accuracy of the model	Visualization of the model
Smart pruning	45,40%	100%	
Active statistical pruning	45,40%	96,7%	

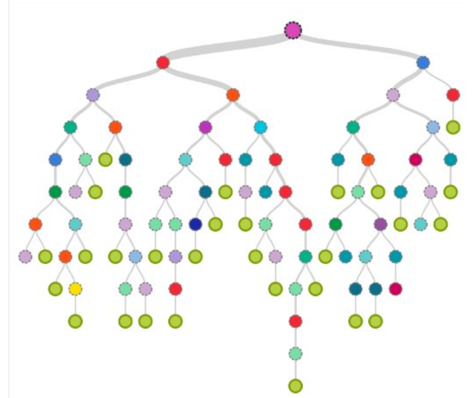
No statistical pruning	45,40%	100%	

Table 3 shows three decision tree models developed by using different forms of pruning during model configuration: *smart pruning*, *active statistical pruning, no statistical pruning.* For further data analysis and a more detailed description of the model, we opted for *active statistical pruning* primarily because of the shape of the tree. Tree structure is the smallest, and reliability of the model is the same as for two other pruning options. Furthermore, this model has the highest prediction accuracy. Reliability of the selected model is 45.40%, and the error is between 2 % and 4%, which would mean that the model is quite accurate. High accuracy is also visible in the graphical representation of the actual and predicted distribution. As seen in Figure 1. the bars are almost the same.

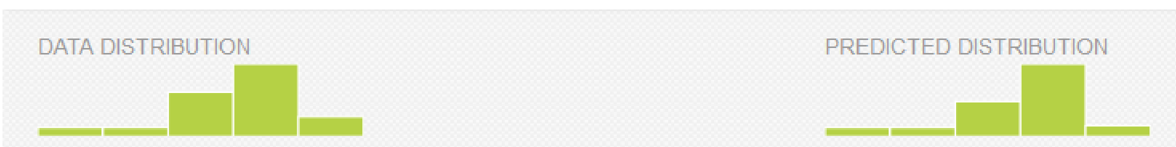
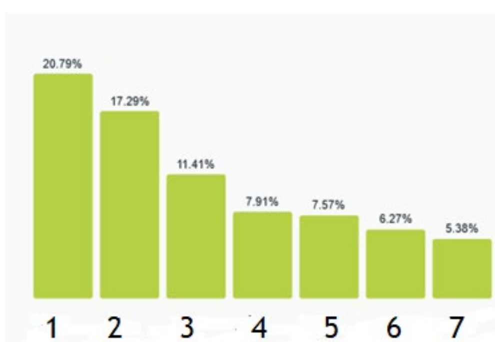


Figure 1. Model accuracy for *active statistical pruning*

Sensitivity analysis is performed according to guidelines of Kamiński et. al. (2017). Figure 2 shows results of sensitivity analysis on the form of variable importance in the decision tree model. The most important variables are: *course_repetition* with 20.79%, *enrollment_reason* with 17.29%, *enrollment_year* with 11.41%, *family* with 7.91%, *learning* with 7.57%, then *series/movies* with 6.27% and *place of residence* with 5.38 %. Other variables have a significance of less than five percent.



1= course_repetition, 2= enrollment_reason, 3= enrollment_year, 4= family, 5= learning, 6= series/movies, 7= place of residence.

Figure 2. Variable importance in decision tree model

The more significant attributes are closer to the root of decision tree. In our case, repetition of the course, followed by the reason for enrollment and the family are first nodes in decision tree model. Decision tree can be extracted into series of IF - THEN rules. Root-to-leaf routes represent those rules (McGonagle, 2021). Hereinafter, we will emphasize few rules for demonstration. First rule, listed at figure 3, predicts path when GPA is in range from 3,5 till 4,49. Reliability of such rule is 64,57%. Decision nodes here are *course_repetition, family* and *enrollment_reason*.

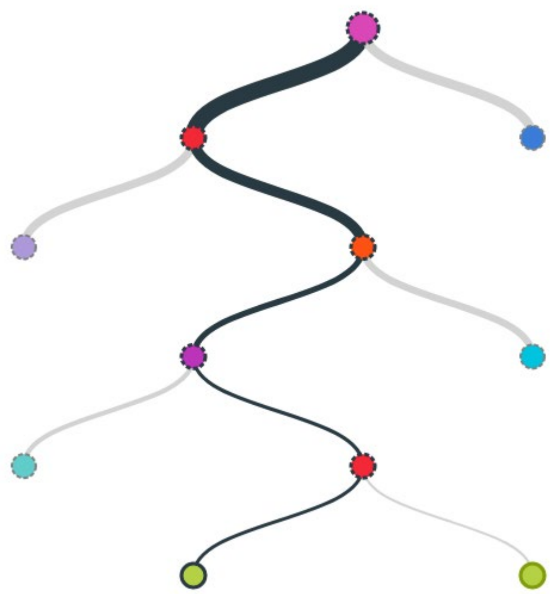
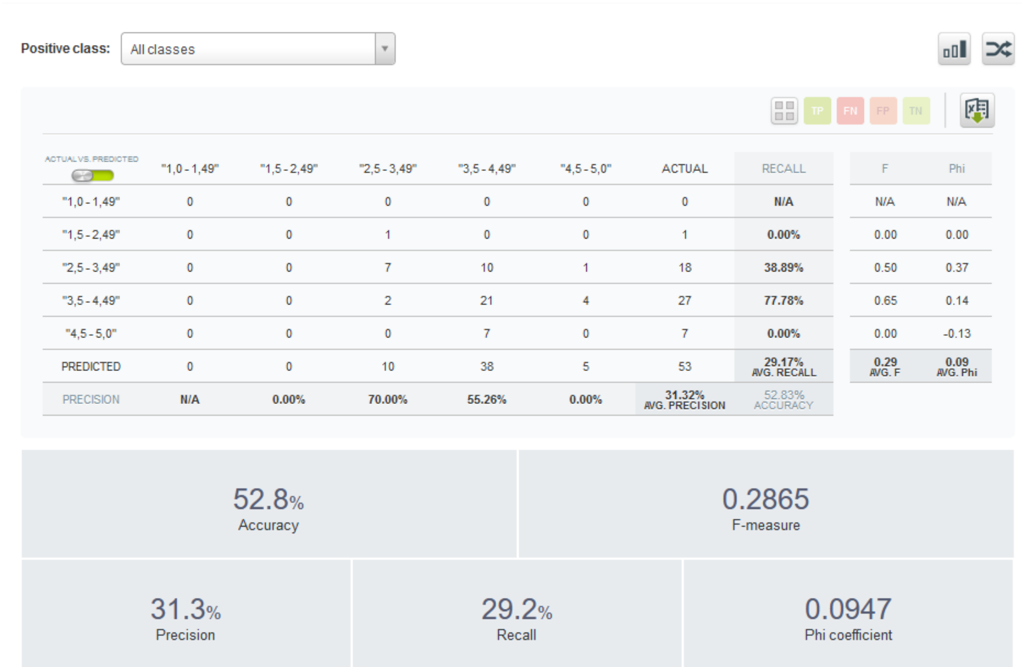


Figure 3. Rule for prediction of success "3,5-4,49"

```
IF (course_repetition = YES && family != less than 3 && enrollment_reason
!=NOT (reputation OR wish)
then GPA = „3,5 – 4,49‟
```

Second rule, listed below, achieves reliability of 74,12%. Used attributes are: *course_repetition, enrollment_reason, year of enrollment,* and *instruction.* In *if-then* rule looks like this:

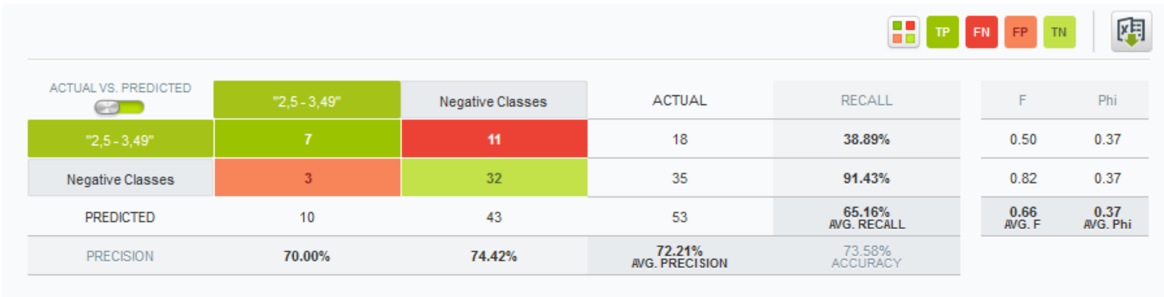
```
if (course_repetition = no && enrollment_reason!= wish or recommendation &&
year of enrollment > 2018 && instructions = yes)
then GPA = „2,5 – 3,49‟.
```


Positive class: All classes

ACTUAL VS. PREDICTED	"1,0 - 1,49"	"1,5 - 2,49"	"2,5 - 3,49"	"3,5 - 4,49"	"4,5 - 5,0"	ACTUAL	RECALL	F	Phi
"1,0 - 1,49"	0	0	0	0	0	0	N/A	N/A	N/A
"1,5 - 2,49"	0	0	1	0	0	1	0.00%	0.00	0.00
"2,5 - 3,49"	0	0	7	10	1	18	38.89%	0.50	0.37
"3,5 - 4,49"	0	0	2	21	4	27	77.78%	0.65	0.14
"4,5 - 5,0"	0	0	0	7	0	7	0.00%	0.00	-0.13
PREDICTED	0	0	10	38	5	53	29.17% AVG. RECALL	0.29 AVG. F	0.09 AVG. Phi
PRECISION	N/A	0.00%	70.00%	55.26%	0.00%	31.32% AVG. PRECISION	52.83% ACCURACY		

Figure 4. Metrics of decision tree model quality

After evaluating the decision tree model using the remaining 20% of the dataset allocated for evaluation phase, accuracy of 52.8% for all classes is achieved (figure 4). If we take a deeper look at the details for each interval separately, there is no data for the average "1.0 - 1.49", this most likely occurred because there is only one such data in the whole data set, and in this case it ended up in the evaluation set. Furthermore, when we observe the next interval of "1.5 - 2.49", the accuracy is high 98.1%, but for "2.5 - 3.49" it is slightly lower, 73.6%, however still very good. The interval "3.5 - 4.49" has an accuracy of 56.6%. The last interval has an accuracy of 77.4%, which is slightly higher than the previous one.



ACTUAL VS. PREDICTED	"2,5 - 3,49"	Negative Classes	ACTUAL	RECALL	F	Phi
"2,5 - 3,49"	7	11	18	38.89%	0.50	0.37
Negative Classes	3	32	35	91.43%	0.82	0.37
PREDICTED	10	43	53	65.16% AVG. RECALL	0.66 AVG. F	0.37 AVG. Phi
PRECISION	70.00%	74.42%	72.21% AVG. PRECISION	73.58% ACCURACY		

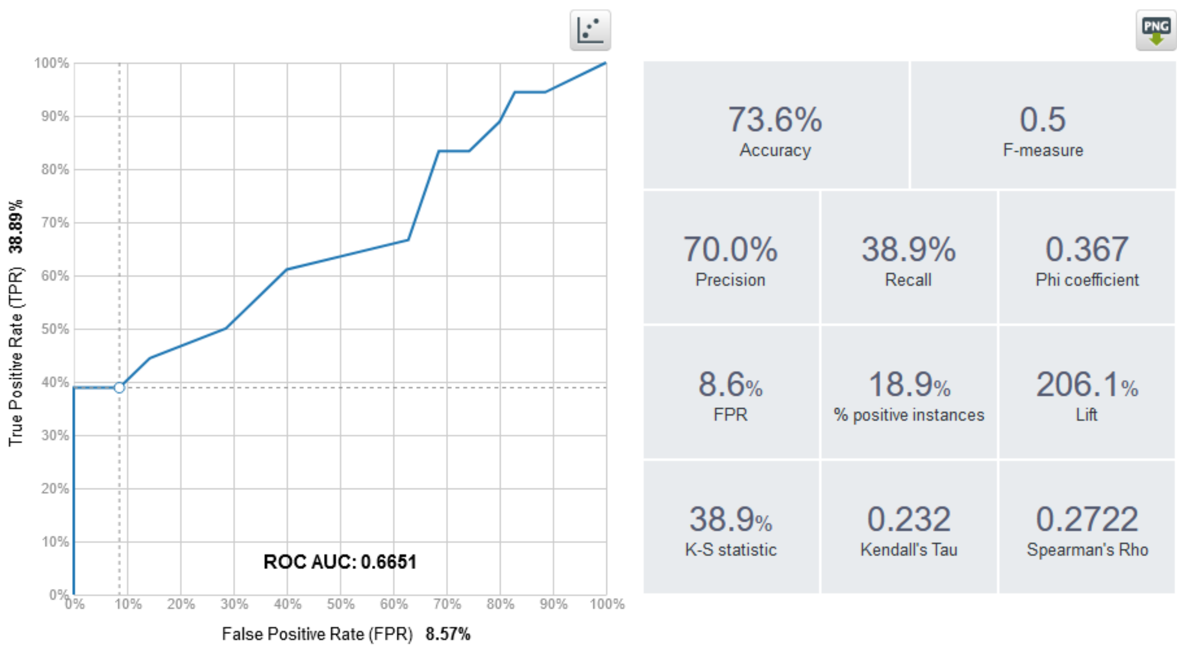


Figure 5. ROC curve for "2,5 - 3,49" part of decision tree model

We will take a closer look at figure 5, and explain the interval "2.5 - 3.49". The shape of the curve itself tells us that for better results we need more data, that is more students include into sample. Furthermore, the table shows that the model accurately predicted seven positive values and 32 negative values, and incorrectly predicted three positive values and 11 negative ones. So, in total, there are more accurately classified than negatively classified. Final accuracy of the model is 73.6%.

6. Conclusion

In this paper we have used decision tree algorithm for prediction of the students' grade point average. The results show the students' performance and its prediction seems to be accurate. This analysis has shown that the proper machine learning application on student's performance data can be efficiently used for vital extraction of valuable hidden knowledge from the vast amount of data generated on daily basis.

Hereinafter, we provide various guidelines and directions for future research. To exploit the full potential of the student academic performance prediction,further data integration is needed. Besides students demographic data, their learning behaviours, habits and familiy background, nowadays student interaction with the learning environment is a prerequisite to ensure sufficient amount of data for analysis. Having in mind the recent increase in the availability of learning management system data and popularity of such learning environments, data from various sources should be integrated.
Future studies should evaluate other machine learning algorithms, such as: artificial neural networks, k-nearest neighbors, Bayesian approaches, support vector machines and ensemble methods to assess whether performance is improved.

References

Alamri, L.H., Almuslim, R.S., Alotibi, M. S., Alkadi, D. K., Ullah Khan, I. and Aslam, N. "Predicting Student Academic Performance using Support Vector Machine and Random Forest," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, vol. PartF168981, pp. 100–107, doi: 10.1145/3446590.3446607, 2020.

Asogbon, M. G., Samuel, O. W., Omisore, M. O., & Ojokoh, B. A. (2016). A multi-class support vector machine approach for students academic performance prediction. Int. J. Multidiscip. Curr. Res, 4, 210-215. G. Pratiyush and S. Manu, "Classifying Educational Data Using Support Vector Machines:A Supervised Data Mining Technique," *Indian J. Sci. Technol.*, vol. 9, no. 34, doi: 10.17485/ijst/2016/v9i34/100206, 2016.

Bendangnuksung, A. and Prabu, D. "Students' Performance Prediction Using Deep Neural Network," *Int. J. Appl. Eng. Res.*, vol. 13, no. 2, pp. 1171–1176, Accessed: Nov. 01, 2021. [Online]. Available: http://www.ripublication.com, 2018.

Binh H.T. and Duy, B.T. "Predicting Students' performance based on learning style by using artificial neural networks," *Proc. - 2017 9th Int. Conf. Knowl. Syst. Eng. KSE 2017*, vol. 2017-January, pp. 48–53, doi: 10.1109/KSE.2017.8119433, 2017.

David Kolo, K., Adepoju, S. A. and Kolo Alhassan, J., "A Decision Tree Approach for Predicting Students Academic Performance," *Int. J. Educ. Manag. Eng.*, vol. 5, no. 5, pp. 12–19, doi: 10.5815/IJEME.2015.05.02, 2015.

Divyabharathi, Y., and Someswari P., "JOURNAL OF ADVANCEMENT IN ENGINEERING AND TECHNOLOGY Journal homepage: www.scienceq.org/JAET.php A Framework For Student Academic Performance Using Naive Bayes Classification Technique," Accessed: Nov. 01, 2021. [Online]. Available: www.scienceq.org/JAET.php, 2021.

Elmannai, H. Anis, M. and Saber, M. "Support Vector Machine for Remote Sensing image classification., 2018.

Al-Sudani, S. and Palaniappan, R. "Predicting students' final degree classification using an extended profile," *Educ. Inf. Technol. 2019 244*, vol. 24, no. 4, pp. 2357–2369, doi: 10.1007/S10639-019-09873-8, 2019.

Fike, D.S. and Fike, R. "Predictors of First-Year Student Retention in the Community College:," *http://dx.doi.org/10.1177/0091552108320222*, vol. 36, no. 2, pp. 68–88, doi: 10.1177/0091552108320222, 2008.

Hamsa, H., Indiradevi, S. and Kizhakkethottam, J.J. "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technol.*, vol. 25, pp. 326–332, doi: 10.1016/J.PROTCY.2016.08.114, 2016.

Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S. and Sarker, K.U. "Student Academic Performance Prediction by using Decision Tree Algorithm," *2018 4th Int. Conf. Comput. Inf. Sci. Revolutionising Digit. Landsc. Sustain. Smart Soc. ICCOINS 2018 - Proc.*, doi: 10.1109/ICCOINS.2018.8510600, 2018.

Kamiński, B.,Jakubczyk, M. and Szufel, P. "A framework for sensitivity analysis of decision trees," *Cent. Eur. J. Oper. Res. 2017 261*, vol. 26, no. 1, pp. 135–159, doi: 10.1007/S10100-017-0479-6, 2017.

McGonagle, C., George, J., Hsu, S.,Khim, A. and Williams, J. "Backpropagation | Brilliant Math & Science Wiki." https://brilliant.org/wiki/backpropagation/ (accessed Nov. 03, 2021).

Okubo, F., Shimada, A., Yamashita, T. and Ogata, H. "A neural network approach for students' performance prediction," *ACM Int. Conf. Proceeding Ser.*, pp. 598–599, doi: 10.1145/3027385.3029479, 2017.

Olaniyi, A., Kayode, A.S., Abiola, S. Y., , Tosin, H. M. and Babatunde, "Students performance analysis using decision tree algorithms," *Ann. Comput. Sci. Ser.*, vol. 15, no. 1, 2017.

Peña-Ayala, A., "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, doi: 10.1016/J.ESWA.2013.08.042, 2014.

Razaque, F. *et al.*, "Using Naïve Bayes Algorithm to Students' bachelor Academic Performances Analysis."

Raut A.B. and Nichat A.A. "Students Performance Prediction Using Decision Tree Technique," *Int. J. Comput. Intell. Res.*, vol. 13, no. 7, pp. 1735–1741, Accessed: Nov. 01, 2021. [Online]. Available: http://www.ripublication.com, 2017.

Raihana, Z., Farah Nabilah A.M "CLASSIFICATION OF STUDENTS BASED ON QUALITY OF LIFE AND ACADEMIC PERFORMANCE BY USING SUPPORT VECTOR MACHINE | Journal of Academia." https://myjms.mohe.gov.my/index.php/joa/article/view/8185 (accessed Nov. 01, 2021), 2021.

Quinlan, J.R. "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, doi: 10.1007/bf00116251, 1986.

Sharma, S. and Kumar, H. "A Survey on Decision Tree Algorithms of Classification in Data Mining," *Int. J. Sci. Res.*, vol. 5, no. 4, pp. 2094–2097, doi: 10.21275/V5I4.NOV162954, 2016.

Soni,A., Kumar, V., Kaur, R. and Hemavathi, D. "PREDICTING STUDENT PERFORMANCE USING DATA MINING TECHNIQUES," 2018.

Tomasevic, N., Gvozdenovic, N. and Vranes, S. "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, p. 103676, doi: 10.1016/J.COMPEDU.2019.103676, 2020.

Van Barneveld, J. P., Arnold, A., Campbell, K.E., "Analytics in Higher Education: Establishing a Common Language | EDUCAUSE," *EDUCAUSE Learning Initiative*, 2012. https://library.educause.edu/resources/2012/1/analytics-in-higher-education-establishing-a-common-language (accessed Nov. 01, 2021).

Wankhede, S.B. "Analytical Study of Neural Network Techniques: SOM, MLP and Classifier-A Survey," *IOSR J. Comput. Eng.*, vol. 16, no. 3, pp. 86–92, doi: 10.9790/0661-16378692, 2014.

Biographies

Dijana Oreski works as associate professor at the University of Zagreb, Faculty of Organization and Informatics. She received her phd degree in data science field developing new feature selection technique. She teaches various courses related to data mining, machine learning and artificial intelligence. Focus of her scientific work is in application of data science and machine learning approaches in social sciences. She was involved in numerous international projects as project leader or collaborator. She is editorial board member of various scientific journals and scientific committee member of various international scientific conferences. She received awards for her scientific engagement.

Dora Zamuda is a first-year graduate student in Software Engineering at the Faculty of Organization and Informatics in Varaždin, University of Zagreb, where she has already completed her undergraduate studies in Information Systems. There were many different subjects and areas in the undergraduate study, but some of them were more interesting than others, which is why she chose the topic of her final work related to artificial intelligence, machine learning and big data. As she is currently in graduate school and majoring in software engineering, she has found herself in the field of studies in which she is much more involved in the application and software development and her current area of interest have moved more to the above. She would like to merge those two branches (software and mobile development with artificial intelligence and machine learning) into one and work on them simultaneously. Currently she is working on developing one mobile application and one software for real life company with other students and a few professors outside her studies so she could expand her area of expertise.