UDK 004.85:378
DOI 10.7251/ZRSNG2101001M

Explainable machine learning methods as a tool for higher education improvement

Vladislav A. Miškovic, Sinergija University

***Abstract* – Group work and communication problems that occurred during the pandemic have significantly increased the use of new technologies, especially in higher education, where methods and systems of artificial intelligence, primarily machine learning, are increasingly used in the automation of various aspects of the teaching process. An application of these methods introduces completely new ethical and legal problems, related to the nature and manner of use of automatically created knowledge in various social processes. Confidence in automated generated knowledge and the issue of responsibility for the results of decisions that are made on the basis of that knowledge are trying to be solved by using the so-called Explainable Machine Learning methods. The paper discusses the application of these methods in the automation of various aspects of higher education, and demonstrates practical examples of their use in predicting student performance and teacher evaluation. In the practical examples, only publicly available data on the realization of parts of the teaching process were used, as well as open source software tools from the *Weka* data mining system and several libraries of explainable machine learning methods for the *R* and *Python* programming languages.**

***Keywords* – Higher education; Automation; Prediction; Machine Learning; Explainable metods**

***Apstrakt* — Problemi grupnog rada i komunikacije do kojih je došlo u uslovima pandemije značajno su povećali upotrebu novih tehnologija, posebno u visokoškolskom obrazovanju, gde se u automatizaciji različitih aspekata nastavnog procesa sve više koriste metodi i sistemi veštačke inteligencije, pre svega mašinskog učenja. Primena ovih metoda uvodi sasvim nove etičke i pravne probleme, vezane za prirodu i način upotrebe automatizovano stvorenog znanja u različitim društvenim procesima. Poverenje u automatizovano generisano znanje i pitanje odgovornosti za rezultate odluka koje se na osnovu tog znanja donose pokušavaju se rešiti korišćenjem tzv. objašnjivih metoda mašinskog učenja (*Explainable Machine Learnig*). U radu se razmatra primena ovih metoda u automatizaciji različitih aspekata visokoškolskog obrazovanja i ilustruje praktičnim primerima njihove upotrebe u predviđanju uspešnosti studenata i ocenjivanju nastavnika. U praktičnim primerima korišćeni su samo javno dostupni podaci o realizaciji delova nastavnog procesa i programski alati otvorenog koda iz sistema za istraživanje podataka *Weka* i više biblioteka objašnjivih metoda mašinskog učenja za jezike *R* i *Python*.**

***Ključne reči* – visoko školstvo; automatizacija; predviđanje; mašinsko učenje; objašnjivi metodi**

I. INTRODUCTION

In the current pandemic conditions, the use of new technologies at all levels of social organization has significantly increased, especially in higher education. New technology has changed the forms of learning, the way of preparation and realization of teaching and set new, higher requirements for teachers and higher education institutions [1].

High requirements for the quality of teaching at this level of education can be achieved by better adaptation to the possibilities and needs of students, as well as the institutions themselves, which is very difficult to realize without the automation of important aspects of the learning process, e.g. checking the level of prior knowledge and later monitoring the progress and achieved learning outcomes [2],[3].

Experimenting with the learning process with new technologies does not always give positive results [3]. In practice, higher education institutions are somewhat more careful in introducing innovations than economic entities, because their mission is significantly different, as well as the type of responsibility for the final outcome of such changes. However, the experience shows that many technological innovations have proved useful both for outcomes of the education process, and the students themselves, especially those related to the application of artificial intelligence methods [3], [4].

But the use of new automation methods, especially methods of machine learning, introduces new ethical and legal problems related to the nature and manner of use of automated knowledge in various social processes [5], [6], [7]. In order for humans to have confidence in automated learned knowledge, it must be created in such a form and scope that humans can successfully interpret and use it. Nowadays, the "Right to Explanation" is promoted as a new civil right [5], and such an automated system must provide an explanation for their output. One branch in the field of Artificial Intelligence deals with machine learning methods and practical applications of explainable empirical models (Explainable Machine Learning, XML).

The classification and short review of these methods and their practical usage in the improvement of higher education processes are discussed in this paper.

II. MACHINE LEARNING METHODS

Machine learning can be defined as the process of estimating unknown dependencies or structures in a system using a limited number of observations [8]. Methods of machine learning can be classified by basic strategy as rote learning, learning by being told, learning by analogy, and inductive learning, which includes learning by examples and learning by experimentation and discovery [8], [9]. Inductive learning methods are especially important for the discovery of new, previously unknown knowledge.

Inductive learning can be seen as the process of estimating an unknown function, dependency or structure of a system *S* using a limited number of examples *x*. The finite set of empirical data is commonly called a *training set* or a *data set*. A model of learning instances is a basic form of available background knowledge, and is usually specified by a set of attributes or features x_i, i=1..*n*.

Most important inductive learning tasks are classification, regression, and clustering. Classification and regression methods are forms of supervised machine learning, where the system generalizes examples of solved problems to create models which can be used to solve new, previously unseen problems. There are numerous types of machine learning models. Most popular classification models are Bayesian models, decision trees, decision rules, Support Vector Machines (SVM), Artificial Neural Networks (ANN) and multiple models (ensembles) [8], [10], [11]. Decision trees and rules are models that are generally understandable to humans and can be directly translated into the natural language. Machine learning models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), especially Deep Neural Networks and ensembles produce models that are not understandable to humans at all [11], [12].

For problems where dependent variables (attributes) are numeric, there are regression learning methods and models, such as linear regression and regression trees. Linear regression [10], as a weighting sum of attribute values, is a simple regression model which is usually understandable for humans. Regression trees [13] and model trees [11] are nonlinear prediction models, which are understandable for humans in many cases. But many popular high-performance models, such as Support Vector Regression (SVR) and Neural Network Regression are not understandable.

Separate explanation methods are needed to explain predictions for majority of usually superior, high-performance classification and regression models.

III. EXPLAINABLE MACHINE LEARNING METHODS

Explainable Machine Learning (XML) is an emerging field of artificial intelligence (AI) whose aim is to help decision makers to understand and trust underlying machine learning methods and tools [5], [14].

Models and methods of XML can be classified according to its understandability to human as *transparent* (white-box, glass-box) and *non-transparent* (black-box, opaque) [5], [15]. Machine learning models can be explainable by design or need post-hoc explanation, which can be textual or visual.

Explanation methods can explain empirical models in general (*global*, dataset level explanations) or only for some areas of interest or representative instances (*local*, instance level explanations). Some explanation methods are *model-specific*, but there are many *model-agnostic* methods [5], [15].

General model-agnostic methods are applicable to all machine learning models and can explain models of any type, no matter how complex they are. Many model-agnostic explanation methods are graphical, because visualization is model-independent, and enable the compression of large amounts of information into small, observable space.

IV. APPLICATION OF EXPLAINABLE MACHINE LEARNING METHODS IN EDUCATION

Educational institutions collect a large amount of data about their students and the teaching process. Aggregated data can be used to find interesting patterns in that data and use it to improve a teaching process. Nowadays, such analyses can be automated using Data Science, especially Machine Learning methods and tools. Machine learning is typically used to create models for an evaluation of the teaching process, to predict the performances of individual students and teachers, and to provide insights into the impact of various factors of the teaching process and its outcomes.

The main application of *explainable* machine learning methods is as *model generation tools* in educational analytics and prediction. Also, there are other applications of these methods [16], [17], [18], such as Intelligent Tutoring Systems (ITS), for providing support to learners through adapting to their needs by some kind of automated recommender system [19], or components of Learning Management Systems (LMS).

In this section we will describe the use of explainable machine learning methods to create models that solve two practical problems of predictive analytics in education: the prediction of both student and teaching assistant performance.

A. Example 1: Student Performance Prediction

The problem *student-math* [20], [21] contains results of a Mathematics course for 395 high school students. Each instance (student) is described by a vector of 32 attribute values and one target numeric value of attribute G3, which represents the final grade of students at the end of the school year in the range from 0 to 20, Fig. 1.

No	1: school	2: sex	3: age	4: address	5: famsize	6: Pstatus	7: Medu	8: Fedu	9: Mjob	...	30: absences	31: G1	32: G2	33: G3
	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal	Numeric	Numeric	Nominal		Numeric	Numeric	Numeric	Numeric
1	GP	F	18.0	U	GT3	A	4.0	4.0	at_home		6.0	5.0	6.0	6.0
2	GP	F	17.0	U	GT3	T	1.0	1.0	at_home		4.0	5.0	5.0	6.0
3	GP	F	15.0	U	LE3	T	1.0	1.0	at_home		10.0	7.0	8.0	10.0
4	GP	F	15.0	U	GT3	T	4.0	2.0	health		2.0	15.0	14.0	15.0
5	GP	F	16.0	U	GT3	T	3.0	3.0	other		4.0	6.0	10.0	10.0
6	GP	M	16.0	U	LE3	T	4.0	3.0	services		10.0	15.0	15.0	15.0
7	GP	M	16.0	U	LE3	T	2.0	2.0	other		0.0	12.0	12.0	11.0
8	GP	F	17.0	U	GT3	A	4.0	4.0	other		6.0	6.0	5.0	6.0
9	GP	M	15.0	U	LE3	A	3.0	2.0	services		0.0	16.0	18.0	19.0
10	GP	M	15.0	U	GT3	T	3.0	4.0	other		0.0	14.0	15.0	15.0

Fig. 1. Dataset student-math, first 10 instances (of 395)

1) Explanation by using transparent model

The method for predicting a student's grade in the future can be obtained by machine learning regression models based on these historical data. Machine learning will be illustrated by linear regression and the support vector regression method.

The linear regression model belongs to *transparent* models, which are self-explanatory. The model is learned by the *LinearRegression* method from the *Weka* system [11] and predict students' grade with mean absolute error of 1.3059. It has the following form:

```
G3 =
    -0.5303 * school=GP +
    -0.2568 * age +
    -0.4192 * Fjob=services,health,teacher +
     0.5391 * Fjob=health,teacher +
    -0.2845 * activities=yes +
     0.3167 * romantic=no +
     0.4022 * famrel +
     0.1355 * Walc +
     0.0474 * absences +
     0.1687 * G1 +
     0.9718 * G2 +
     0.7893
```

According to this linear model the students' success is equally influenced by 10 out of 32 attributes: *school*, *age*, *Fjob*, *activities*, *romantic*, *famrel*, *Walc*, *absences*, *G1* and *G2*.

2) General explanations of black-box models

The model learned by the *Support Vector Regression* (*SMOreg*) method from the *Weka* system [11] predicts a student's grade with the mean absolute error of 0.9785. Such a model is represented as a (possible) large matrix of mapping coefficients, and it has no intelligible textual or visual explanation. Because such a regression model cannot be directly understood, it is not clear how the students' assessments are made, and in order to provide an explanation, suitable global methods for explaining non-transparent models must be used.

One of the traditional ways for explaining non-transparent models is to calculate and display the importance of individual attributes (variable importance) in predicting the value of the target attribute [5], [10], [22]. Nowadays, instead of a simple feature importances list, we can use a more informative explanation method, *SHAP Summary Plot* from the *Python* package *shap* [23], which contain information not only about feature importance, but also about their values, Fig. 2.

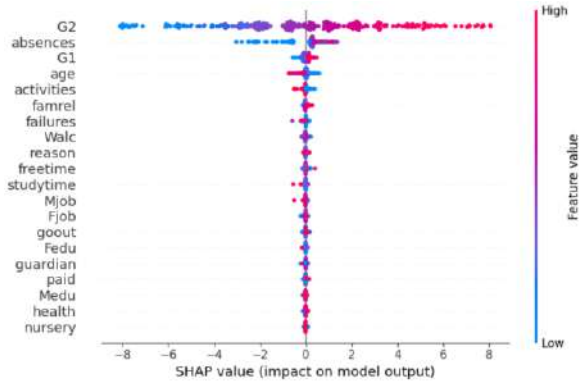

Fig. 2. SHAP Summary of feature importances and their values for explanation of student-math problem

Feature importance scores are calculated on the basis of game theory as a fair distribution of the prediction (so-called payout) among the features, i.e. Shapley values [5], [23]. The method is usually more consistent with our knowledge than the traditional feature importance methods like the permutation importance based on the Random Forest [22], which depends on the performance of a specific model.

The plot shows a relationship between feature values and its impact on predictions. Every point on the *SHAP Summary plot* is a Shapley value for a feature and an instance [5]. Features are ordered by their importance. The point color represents the feature value for instance, from blue-colored low values to high values marked in red.

It is clear that in the general case, the prediction of success is mostly influenced by the results in the second half of the year (G2), followed by the number of absences. All other attributes have much less effect on the prediction of the success of an arbitrary student.

Another graphical method of explainable learning is the *Funnel plot* method from the *R* package DALEX [24], which is used to compare the explanations of predictions of several different machine learning models. One or more models (*Challengers*) are compared to the selected base model (*Champion*) [5].

The dataset on which these models are built is divided by creating categories according to the quantiles of the columns in the data. For each category, the difference is calculated using the appropriate metric (specified by the user, or RMSE for regression, 1-AUC or cross-entropy for classification). A positive value of this difference means that the basic model has a better performance in the specified category, while a negative value means that one of the other models is better.

Fig. 3 shows a comparison of the regression tree forest model (RF) on the left with the basic linear regression model (LM) on the right. It is evident that in all attributes, the RF model is the best for predicting students' success.

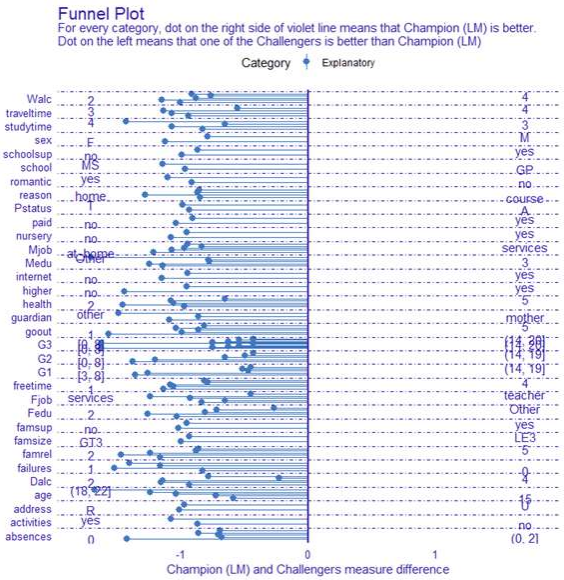

Fig. 3. The comparison of several model explanations

3) Explanations of specific decisions of black-box models

A precise explanation of the assessment of a *specific* student is especially important if it produces consequences, i.e.

influences some decisions. Local explanation methods are used for this type of explanation. Shown in Fig. 4 is the data for student number 394:

No.	1: school	2: sex	3: age	4: address	5: famsize	6: Pstatus	7: Medu	8: Fedu	9: Mjob	...	30: absences	31: G1	32: G2	33: G3
	Nominal	Nominal	Numeric	Nominal	Nominal	Nominal	Numeric	Numeric	Nominal		Numeric	Numeric	Numeric	Numeric
394	MS	M	18.0	R	LE3	T	3.0	2.0	services		0.0	11.0	12.0	10.0

Fig. 4. Feature importances for student-math problem

The prediction model for student 394 is learned by regression tree method and the explanation of the prediction using the local graphical method from the *R* package ExplainPrediction [25] is shown in Fig. 5.

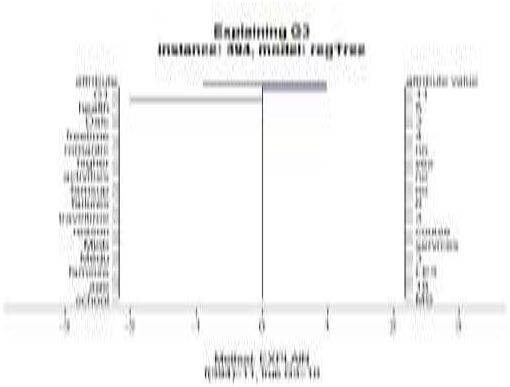

Fig. 5. Model and instance level explanation

It can be seen from this explanation that the model for student 394 predicted the grade 11 (out of 20), while that student really achieved a grade 10. It can also be seen that the prediction in the final exam (G3) according to this model is crucially influenced only by the value of the attribute G1, i.e. success in the first semester, while the influence of other attributes can be ignored for this student.

B. Example 2: Teaching Assistant Evaluation

The dataset has 151 examples of assessments for assistants from the Statistics Department of the University of Wisconsin-Madison during several semesters of teaching [20]. The examples are described by a vector of values of five attributes and a score in a discrete classification attribute with values of *Low*, *Medium* and *High*. The model of training examples and data for the first ten teacher assessments are shown in Fig. 6.

No.	1: TA_native_English_speaker	2: Course_instructor	3: Course	4: Summer_or_regular_semester	5: Class_size	6: Class
	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal
1	English_speaker	23	3	Summer	19.0	High
2	non-English_speaker	15	3	Summer	17.0	High
3	English_speaker	23	3	Regular	49.0	High
4	English_speaker	8	2	Regular	33.0	High
5	non-English_speaker	7	11	Regular	55.0	High
6	non-English_speaker	23	3	Summer	20.0	High
7	non-English_speaker	9	5	Regular	19.0	High
8	non-English_speaker	10	3	Regular	27.0	High
9	English_speaker	22	3	Summer	58.0	High
10	non-English_speaker	16	3	Summer	20.0	High

Fig. 6. Part of data for the first 10 teaching assistants (of 151)

1) Explanation by using transparent model

One of the transparent models for predicting the success of assistants is the model of 16 decision rules [11], whose prediction accuracy was estimated at 59.6% by the cross-validation method:

Summer_or_regular_semester = Summer AND
Class_size > 15 AND
Course <= 13: High (16.0/2.0)

TA_native_English_speaker = English_speaker AND
Summer_or_regular_semester = Regular AND
Course <= 5 AND
Class_size > 18: High (11.0/1.0)

Course <= 10 AND
Course_instructor > 21 AND
Course_instructor > 24: Medium (4.0)

Course <= 10 AND
TA_native_English_speaker = English_speaker AND
Course_instructor <= 20: Low (3.0)

Summer_or_regular_semester = Summer AND
Course_instructor <= 7: Medium (2.0)

Course > 16 AND
Class_size > 39: Medium (7.0)

Course > 16 AND
Course > 22 AND
Class_size <= 23: Medium (3.0)

Course > 16 AND
Course_instructor > 8 AND
TA_native_English_speaker = non-English_speaker AND
Course > 20: High (7.0)

Summer_or_regular_semester = Summer AND
Class_size <= 10: Medium (2.0)

Summer_or_regular_semester = Regular AND
TA_native_English_speaker = English_speaker AND
Course <= 16 AND
Class_size <= 29: Medium (3.0)

Summer_or_regular_semester = Regular AND
TA_native_English_speaker = non-English_speaker AND
Course <= 4 AND
Class_size <= 25 AND
Course_instructor > 9 AND
Course > 1 AND
Class_size <= 24: Low (10.0/1.0)

Summer_or_regular_semester = Regular AND
Class_size <= 18 AND
Course <= 17: Medium (12.0/2.0)

Summer_or_regular_semester = Regular AND
TA_native_English_speaker = non-English_speaker: Low (66.0/31.0)

Class_size <= 31: High (3.0)

Course_instructor <= 10: Medium (1.0)

: Low (1.0)

2) General explanations of black-box models

The ensemble model Random Forest [22] predicts assessments with the accuracy of 64.9%. Because it is represented as a committee of at least 15 different decision trees, it is not transparent by itself and requires a separate explanation. As is done in Example 1, one of the post-hoc methods of explanation can be used.

The influence of individual attributes to the predictions of teaching assistants' assessment is shown in Fig. 7 using the new stacked SHAP feature importance plot method provided by the *Python* package *shap*, based on the SHAP method [23], [24].

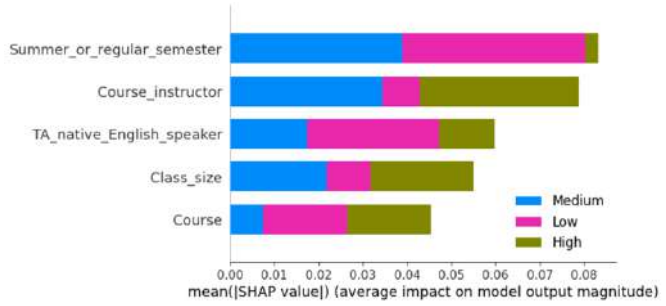

Fig. 7. SHAP Summary feature importance explanation of teaching assistants assesment using Random Forest model prediction

It can be seen that the rating of a teaching assistant in general is mostly influenced by the semester in which the classes are held (*Summer_or_regular_semester*) and teacher who teaches the subject (*Course_instructor*), except for Low rating assistants. Fluent English language speaking has a smaller effect, especially for High rating assistants. The class size (*Class_size*) is the next, but it is not too important for Low rated assistants, whose rating is more influenced by the subject (*Course*), the least important feature in general.

3) Explanations of specific decisions of black-box models

An explanation of the assessment of one *specific* teaching assistant can be obtained by some of the local methods of explanation.

The first example is method from the *R* package ExplainPrediction [25]. Fig. 8 shows the explanation of the rating prediction for the teaching assistant 151, obtained by the Random Forest model.

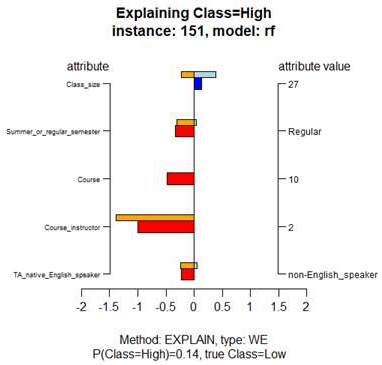

Fig. 8. ExplainPrediction explanation of Random Forest model prediction

It can be seen that the actual rating of this assistant is *Low*, although the model predicted *High*. The impact of individual feature values is explained by red-colored bars, and average positive and negative explanations of feature values are depicted by orange and blue bars. According to the graphical explanation, the prediction for teaching assistant 151 is mostly influenced by a course instructor is and the course in question. All other attributes are less important.

The second example is the prediction model based on the Support Vector Classification (SVC) for teaching assistant 2 using a similar graphical explanation method, the *Break-down plot* from the *Python* package DALEX [24] as shown in Fig. 9.

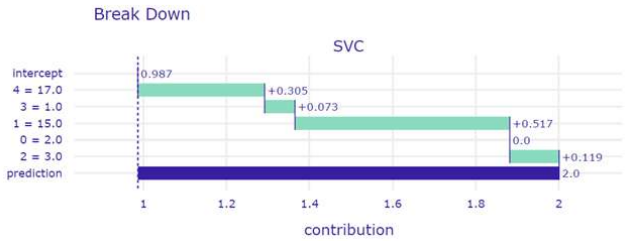

Fig. 9. Break-down explanation of Random Forest model prediction

The most influential attribute for the prediction of teaching assistant 2's assessment is attribute 1 (*Course_instructor*), followed by attribute 4 (*Class_size*).

V. CONCLUSION

In this paper a brief overview of the notion and application of explainable machine learning methods as a tool in higher education is provided, especially for model generation in educational analytics and prediction. Important methods and tools for automated knowledge generation are practically demonstrated. Predictions of generated models are explained (1) by using the structure of machine-learned transparent models for a complete understanding of the problem and (2) by methods of post-hoc explanation in general and for individual specific cases.

It is important to note that the explanation methods used are suitable for models with a small or moderate number of features. If there are thousands or millions of features, explanations methods which use a small subset of features are needed. The popular model-agnostic method is Local Interpretable Model-agnostic Explanations (LIME) [5], which locally approximates such huge non-transparent models by simpler transparent models.

All model-agnostic explanation methods used in this paper are graphical, because they are model-independent and more concise than other types of explanations. Also, there are model-specific methods, which generate explanations of non-transparent models in the form of rules (e. g. DeepRED [26], RxREN [27]) and decision trees (e.g.TREPAN [28]), which are already considered to be transparent models.

REFERENCES

[1] S. Verma, P. Tomar (eds), Impact of AI Technologies on Teaching, Learning, and Research in Higher Education, IGI Global, 2021

[2] T. van der Vorst, N. Jelicic, "Artificial Intelligence in Education: Can AI bring the full potential of personalized learning to education?", 30th European Conference of the International Telecommunications Society (ITS):Towards a Connected and Automated Society, Helsinki, Finland, 16th-19th June, 2019

[3] P. Thomas, Adaptive learning, AI in teaching and explainable AI: Is AI the new technology frontier for learning and teaching?, https://medium.com/haileyburyx/adaptive-learning-ai-in-teaching-and-explainable-ai-d1c6b5bd0802, retrieved 07.11.2021

[4] B. Ghai, Q. Vera Liao, Y. Zhang, R. Bellamy, K. Mueller, "Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers", Proceedings of the ACM on Human-Computer Interaction, Volume 4, Issue CSCW3, Article No. 235, pp 1–28, 2021

[5] P. Biecek, T. Burzykowski, Explanatory Model Analysis: Explore, Explain and Examine Predictive Models, Taylor&Francis Group, LLC, 2021

[6] P. Hall, N. Gill, An Introduction to Machine Learning Interpretability, 2nd Ed, O'Reilly Media, 2019

[7] C. Molnar, Interpretable machine learning: A Guide for Making Black Box Models Explainable, 2019

[8] Cherkassky V., Mulier F. M., Learning from Data: Concepts, Theory, and Methods, 2nd edition, John Wiley - IEEE Press, 2007

[9] R. Michalski, J. Carbonell, T. Mitchell (Eds.), Machine learning: An artificial intelligence approach (Vol. I), San Francisco, CA: Morgan Kaufmann, 1983.

[10] S. Raschka, V. Mirjalili, Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition, Packt Publishing, 2019

[11] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[12] W. Samek, T. Wiegand, K. R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", ITU Journal: ICT Discoveries, Special Issue No. 1, 13 Oct. 2017. Arxiv preprint arxiv:1708.08296.

[13] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, Belmont, 1984.

[14] D. Rothman, Hands-On Explainable AI (XAI) with Python, Packt Publishing, 2020

[15] Vaishak Belle, Ioannis Papantonis, "Principles and Practice of Explainable Machine Learning", Frontiers in Big Data, Volume 4, pp. 1-25, 2021 DOI: 10.3389/fdata.2021.688969

[16] J. Zhou, F. Chen (eds) Human and Machine Learning, Human–Computer Interaction Series, Springer, 2020 https://doi.org/10.1007/978-3-319-90403-0_2

[17] O. Biran, C. Cotton, "Explanation and justification in machine learning: A survey", in: Workshop on Explainable AI (XAI), 2017, pp. 8–13.

[18] J. M. Schoenborn, K. D. Althoff, "Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions", in: Case-Based Reasoning for the Explanation of intelligent systems (XCBR) Workshop, 2019.

[19] Behnoush Abdollahi, Olfa Nasraoui, "Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems", in J. Zhou, F. Chen (eds), Human and Machine Learning, Human–Computer Interaction Series, Springer, 2020 https://doi.org/10.1007/978-3-319-90403-0_2

[20] D. Dua, C. Graff, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[21] P. Cortez, A. Silva, "Using Data Mining to Predict Secondary School Student Performance", in A. Brito and J. Teixeira (eds), Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[22] L. Breiman, "Random Forests", Machine Learning, 45, pp. 5–32.,2001

[23] S. M. Lundberg, Su-In Lee, "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems, 2017

[24] Biecek P (2018). "DALEX: Explainers for Complex Predictive Models in R." Journal of Machine Learning Research, 19(84), 1-5.

[25] M. Robnik-Šikonja, I. Kononenko, "Explaining Classifications For Individual Instances", IEEE Transactions on Knowledge and Data Engineering, 20:589-600, 2008

[26] J. R. Zilke, E. L. Mencia, F. Janssen, "DeepRED - Rule Extraction from Deep Neural Networks", in T. Calders et al. (Eds.), Discovery Science: 19th International Conference, DS 2016, Bari, Italy, pp. 457–473, 2016

[27] S. K. Biswas, M. Chakraborty,B..t Purkayastha, P. Roy, D. M. Thounaojam, "Rule Extraction from Training Data Using Neural Network", International Journal on Artificial Intelligence Tools, Vol. 26, No. 03, 2017

[28] M. W. Craven, Extracting Comprehensible Models from Trained Neural Networks, PhD Thesis, Computer Science Department, University of Wisconsin, Madison, WI, 1996