Comparative Analysis of Machine Learning Algorithms on Data Sets of Different Characteristics for Digital Transformation

D. Oreški*, I. Pihir* and D. Višnjić*
* University of Zagreb / Faculty of Organization and Informatics, Varaždin, Croatia
{dijana.oreski, igor.pihir, dunja.visnjic}@foi.unizg.hr

Abstract **- The application scenarios for machine learning algorithms are getting more complicated as machine learning and real-world situations converge more and more. All fields of study have adopted and benefit from diverse machine learning algorithms implementation. The challenge is to determine which algorithm is best suited to solve a given problem. This problem is especially challenging in social sciences. To tackle that issue, this paper explores a group of machine learning algorithms used for predictive models' development in social science domains of business and education. Several machine learning algorithms are applied here (algorithms of artificial neural networks, k-nearest neighbors, decision tree) along with characteristics of datasets measured by meta-features. In the empirical part of the research, algorithms are compared on the data sets using standard predictive model evaluation metrics. Data sets are extracted from the education and business domain. Research results provide insights into machine learning algorithms' performance depending on their meta-features. Meta-features are significant predictors of machine learning algorithms' performance in both education and business domain. Machine learning-based predictive models developed in this paper are a step forward to the acceleration of digital transformation in the education and business sector.**

Keywords **- comparative analysis; machine learning algorithms; meta-features; digital transformation.**

I. INTRODUCTION

Real-world situations today include digital technology usage on daily basis. In such scenarios, huge amounts of data are generated. Various machine learning algorithms are developed in order to extract knowledge from the data [1]. Every academic discipline, including social sciences, has embraced machine learning and benefited from its use. However, the application of machine learning algorithms is complex and time-consuming. The difficult part is choosing the algorithm that will solve a given problem [2]. The social sciences are particularly challenged by this issue since machine learning for quick and on-demand data analysis, could be a problem or an opportunity to produce new possibilities for new business value and possible digital transformation.

Digital Transformation (DT) is a contemporary paradigm, still growing and emerging in a large number of research papers, used and practiced in business development as well as in research with great interest worldwide [3]. Governments and private companies go for DT to improve business processes, digitally transform their operations, redefine business models and enhance products and services. DT is strategy-oriented and customer-centric, it is based on the introduction of new informational and communicational technology and organizational changes in business models, business processes, and/or products/services [4]. Technological and business changes, as a result of digital transformation, are introduced in new business models. Available data has become the core of interactions in digital business due to its unlimited amount. The digital world has shaped a culture of interaction based on data, and digital transformation leaders are expected to create innovative products and services, as well as increase data readiness to create and share information and deliver seamless digital services [5]. Digital transformation use/reuse technology-related concepts [6] such as Big data, Data analytics, Machine Learning, Artificial Intelligence, Data-Based Decision Making, and Knowledge Management [7], [8], [9] to open possibilities and accelerate digital transformation in business as well as in education and academia.

This paper is structured as follows. Section II overviews related work in the field. Section III describes two datasets used in the empirical research and gives a brief overview of two machine learning algorithms. Section IV provides insights into the characteristics of data measured by meta-features and compares the results of predictive models developed by different machine learning algorithms. Section V concludes the paper by providing guidelines for machine learning algorithms selection in the social science domain and results interpretation in the context of education and business digital transformation.

II. RELATED WORK

Knowledge of artificial intelligence, especially machine learning, is required to process the increasing amount of data that is becoming more accessible in today's digital

environment. A large number of algorithms have been developed in this area that can be used to develop predictive and descriptive models in different domains.

However, since the developed algorithms have different effects on different data, it is necessary to determine which algorithm is best to use in a particular situation, domain, and data set. According to the "no free lunch" theorem there is no best algorithm for all data [10]. Choosing the best algorithm from a large set of available algorithms is a challenging and time-consuming task known in the literature as the algorithm selection problem [11]. Conventional approaches such as trial and error, theoretical analysis, or expert knowledge, have several drawbacks, including high computational costs, difficulty in finding experts, and experts' tendency for personal prejudices and preferences [12], [13]. Considering the shortcomings of traditional approaches to algorithm selection, recent research has focused on automatic algorithm selection [14]. One prominent approach to automated algorithm selection is meta-learning-based algorithm recommendation.

Meta-learning, in the context of machine learning, is the process of learning from previous experience gained by applying different learning algorithms to data sets with different characteristics [15]. There have been numerous successful applications of the meta-learning process in algorithm selection. For instance, Sivakumar et al. [16] compare algorithms in the medical domain for the early detection of cancer. They conduct a comparative analysis of classification algorithms, and the research results in the proposal of a classification algorithm for the medical domain. In the medical domain, meta-learning has also been used to predict covid-19 status based on chest CT scans [17], and [18] proposes MetaPred, a meta-learning framework for predicting clinical risk (e.g. hospital mortality, re-admission to the hospital, etc.). Furthermore, Garcia-Saiz and Zorrilla [19] compare regression algorithms in the educational domain, with the goal of predicting student performance in e-learning courses. Similarly, Romero et al. [20] use meta-learning to recommend a subset of classification algorithms based on Moodle data.

Characteristics extracted from data sets, known as meta-features, play a significant role in the successful application of meta-learning since they can influence the performance of the considered algorithms [21]. They describe different types of data properties that can be used to predict the performance of machine learning algorithms [22], [23]. Prior studies in the domain of classification methods suggest that the characteristics of data sets have a significant effect on the performance of algorithms and demonstrate that the selection of the "best" algorithm depends on the properties of a given data set [24], [25[, [26]. Recent research has not earned much attention to these issues. Kiang suggests that the characteristics of the data set have a significant impact on the performance of classification algorithms [27]. Similarly, Smith's research on the problem of choosing a classification algorithm confirms that it is necessary to understand the characteristics of the data set and that they should be related to the classification algorithm's performance [28].

Many empirical studies [21], [22], [29], [30], [31] proposed different sets of meta-features, but there is no ultimate list that uniformly describes, organizes, and calculates them. Rivolli et al. [30] recently systematized and standardized meta-features into six groups: 1) general; 2) statistical; 3) information-theoretical; 4) model-based measures; 5) landmark measures; and 6) others. This is the most comprehensive list to date.

In this research, we are investigating general and information-theoretical meta-features on datasets from the social sciences domain. Some examples of such usage of different machine learning algorithms and large or open data are explored in various business domains such as agriculture [32], [33], possible use at large and open data in traffic [34], law and government data [35], and used in an education setting [36], [37].

III. DISCUSSION

This section describes two datasets used in the research and introduces two machine learning algorithms applied to the data.

A. Data description and data understanding

Datasets are downloaded from the publicly available repository Kaggle.

The first dataset contains information about research that was conducted in 2008 in two Portuguese schools. The research was conducted for two courses: mathematics and Portuguese language. In this paper, the selected research was on mathematics courses. The number of male and female respondents is approximately equal with 53% female and 47% male. This dataset consists of 395 instances and 33 variables. This dataset is from the education domain and refers to student grade prediction [38].

The second dataset is from the business domain and refers to house sales prediction [39]. The second dataset contains 21 attributes and 21613 instances. Attributes are mostly of numerical type with two categorical ones - date and waterfront. Distributions are of different types, from exponential to multimodal. The most common data distribution is exponential. There is no missing data in the set.

Data preparation consisted of detecting missing values and/or incorrect data. Data understanding included variables description and correlation analysis of numerical variables.

Correlations are performed in the data understanding phase. Results of correlation analysis for the first dataset are presented in the table below. Relations between dependant variable and independent variables are presented.

TABLE 1. CORRELATION BETWEEN INDEPENDENT VARIABLES AND DEPENDENT VARIABLE IN THE EDUCATIONAL DATASET

Variable	by Variable	Correlation	Significance
age	G3	-0.161	p < 0.05
medu	G3	0.217	p < 0.05
fedu	G3	0.152	p < 0.05
Travel time	G3	-0.117	p < 0.05
Study time	G3	0.097	p > 0.05
failures	G3	-0.364	p < 0.05
famrel	G3	0.051	p > 0.05
Free time	G3	0.011	p > 0.05
gocout	G3	-0.132	p < 0.05
Dalc	G3	-0.054	p > 0.05
Walc	G3	-0.051	p > 0.05
health	G3	-0.061	p > 0.05
abscences	G3	0.034	p > 0.05

Most of the correlations are statistically significant. However, there are no high correlations. Table 2 demonstrates a correlation between independent variables and the dependent variable, price, in the case of the business dataset.

TABLE 2. CORRELATION BETWEEN INDEPENDENT VARIABLES AND DEPENDANT VARIABLE IN BUSINESS DATASET

Variable	by Variable	Correlation	Significance
bathrooms	price	0.525134	p > 0.05
bedrooms	price	0.308338	p > 0.05
condition	price	0.036392	p < 0.05
floors	price	0.256786	p > 0.05
grade	price	0.667463	p > 0.05
lat	price	0.306919	p > 0.05
long	price	0.021571	p < 0.05
sqft_above	price	0.605566	p > 0.05
sqft_basement	price	0.323837	p > 0.05
sqft_living	price	0.702044	p > 0.05
sqft_living15	price	0.585374	p > 0.05
sqft_lot	price	0.089655	p < 0.05
sqft_lot15	price	0.082456	p < 0.05
view	price	0.397346	p > 0.05
waterfront	price	0.266331	p > 0.05
yr_built	price	0.053982	p < 0.05
yr_renovated	price	0.126442	p < 0.05
zipcode	price	-0.05317	p < 0.05

Correlation analysis results can determine how the variables are related and that an increase in the construction quality index can cause a price increase, i.e. higher-quality built and designed properties have a higher price than properties of lower construction quality.
In the case of the business dataset, most of the correlations are not statistically significant. Differences in correlation analysis result potentially indicate differences in data characteristics between the two domains of social sciences.

Both datasets are further characterized by meta-features. Simple and information-theoretic meta-features were extracted from the data. Table 3. provides an overview of meta-feature values for those two categories.

TABLE 3. OVERVIEW OF META-FEATURE VALUES

Meta-feature	Educational dataset	Business dataset
General meta-features		
nr_attr	31	19
nr_bin	0	1
nr_inst	395	21613
nr_cat	1	1
nr_num	30	18
attr_to_inst	0.078	0.0008
inst_to_attr	12.742	1137.526
cat_to_num	0.033	0.055
num_to_cat	30	18
nr_outliers	395	386
Information theoretic		
attr_conc.mean	0.456	0.028
attr_conc.sd	0.138	0.056
attr_ent.mean	0.278	3.356
attr_ent.sd	1.549	2.122

The educational dataset has higher dimensionality than the business dataset. The business data set has a higher proportion of numerical attributes in the data. The number of instances is significantly higher in the business dataset. The ratio of outliers is higher in the educational dataset.
Average values of information-theoretic measures show that the educational dataset is more homogenous (measured by attribute concentration), whereas the business dataset has the higher entropy.

B. Machine learning algorithms

The first machine learning algorithm method used in this paper is the decision tree algorithm. A decision tree is a series of nodes connected by branches that extend downward from the root node until they end in a leaf node. The root node is located at the very top of the decision tree graph, and after it, the attributes are tested according to which branches are created by finding new results. Each branch leads to a new decision node or ends with a leaf node.

Decision tree characteristics are:

(i) The decision tree algorithm falls under supervised learning and as such requires the class of the variable being predicted. A training dataset must be applied that provides the algorithm with the values of the variable to be predicted,

(ii) The training data set must have a lot of variation in the result and have a large number of instances, which provides the decision tree algorithm with a large number of branches to classify the data that is added later. A decision tree learns based on examples, if missing examples for a data set, classification, and prediction will be poor or impossible for that data set,

(iii) The attribute to be predicted must be precise, which means that the values of that attribute must be precisely defined so that it is easy to recognize whether an instance belongs to a certain set or not.

The two most used algorithms for decision tree development are the classification and regression trees algorithm (CART) and C4.5 algorithm. Decision trees created employing CART algorithm are binary, which means that each decision has only two branches. The C4.5 algorithm has an advantage over the CART algorithm because it is not limited to binary branches. This algorithm creates decision trees with different branching layouts [40].

Deep learning has become one of the most used machine learning approaches. Artificial neural networks are first representative of the deep learning approach. Neural networks work on the principle of learning from distributed data. The simplest neural network consists of: an input layer, one hidden layer and an output layer.

The process of training a neural network involves the following steps:

(i) defining the structure or architecture of the neural network. This is a very important step, if we create a very extensive network with a large number of neurons then our model will not generalize the data very well.
(ii) select a nonlinear transformation to apply to each link. This transformation controls the efficiency of each neuron in the network.
(iii) decide on the loss function we will use for the output layer. This is true if we have a problem that uses supervised learning,
(iv) to learn the parameters of the neural network, that is, determine the weight values of each connection. The weight values are determined by optimizing the loss function. [41]

IV. RESEARCH RESULTS

The dataset is split into two train and test data, where 70% of the data goes under variables for training, and 30% under variables for testing. In the data preparation phase, outlier detection was performed. In order to perform a neural network model, data are normalized so that only values between 0 and 1 are found in the variables.

A. Predictive models on a business dataset

Optimization of hyperparameters was performed for both machine learning algorithms. Artificial neural network parameters are presented in table 4.

TABLE 4. ARTIFICIAL NEURAL NETWORK PARAMETERS

Artificial neural network parameter	Value
Number of hidden layers	One
Number of hidden nodes	10
Activation function	Sigmoid
Learning rate	0.1

Pruning was performed in decision tree model development to optimize the mode. The smart pruning approach yielded the best model. Models' performance results in terms of error rate are presented in table 5.

TABLE 5. MODELS PERFORMANCE ON THE BUSINESS DATASET

Machine learning algorithm	Error rate
Decision tree	7,55 %
Artificial neural networks	1,6 %

A comparison of decision tree and neural network models indicates the following factors: construction quality index and square footage have proven to be the factors that have the highest impact on price formation according to both models. Furthermore, according to the decision tree model, the third most important factor includes latitude, whereas the year of construction stands out in the neural network model. By adjusting the values of the attributes, significant changes in the price are observed, and thus we can conclude that there is a positive connection between the price of the property, the square footage of the interior space and the quality of construction. A positive correlation between the price of real estate and the number of rooms was also confirmed. Predictive models have shown that the number of rooms has much less influence on the price than expected.

B. Predictive models on an educational dataset

Optimization of hyperparameters was performed for the educational dataset.

TABLE 6. ARTIFICIAL NEURAL NETWORK PARAMETERS

Artificial neural network parameter	Value
Number of hidden layers	One
Number of hidden nodes	15
Activation function	Sigmoid
Learning rate	0.1

Pruning was performed in decision tree model development to optimize the mode. The smart pruning approach yielded the best model. Models' performance results in terms of error rate are presented in the table 7.

TABLE 7. MODELS PERFORMANCE ON EDUCATIONAL DATASET

Machine learning algorithm	Error rate
Decision tree	5,62 %
Artificial neural networks	7,34 %

Based on the model evaluation, we can conclude that there are differences in the performances of different machine learning algorithms on datasets from different domains. In order to establish a connection between algorithms performances and data characteristics, meta-features are examined.

C. Meta-features based meta-model

According to our sample, using data of similar characteristics, an artificial neural network as a machine algorithm should be recommended to do data analysis on a dataset of business domains and a decision tree algorithm in the case of the educational dataset. The educational dataset has higher values for information-

centered meta-feature of attribute concentration. Such measure come from the information-theory field and try to capture the amount of information in the data. The same approach in learning has a decision tree algorithm, which is an information-based machine learning algorithm. Since decision tree-based predictive model yielded more accurate predictions in the case of the educational dataset, inherited information in data provided good basis.

Explanation of the results serves as guidelines for machine learning algorithm selection based on the dataset characteristics measured by meta-features.

V. CONCLUSION

This research examined characteristics of data in education and business. Data characteristics were measured by meta-features. Two machine learning algorithms were applied on the datasets to develop predictive model. Model evaluation led to a conclusion regarding machine learning algorithm selection based on the meta-features. Explanation of the results serves as scientific contributions and guidelines for machine learning algorithm selection based on the dataset characteristics measured by meta-features in favor of the application and research of digital transformation in education and the business world. In the furture research, sample of datasets will be increased, as well as number of extracted meta-features and number of employed machine learning algorithms.

ACKNOWLEDGMENT

This work has been fully supported by Croatian Science Foundation under the project UIP-2020-02-6312.

REFERENCES

[1] Y. Peng, G. Wang, G. Kou, and Y. Shi, (2011). An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing, 11(2), 2906-2915.

[2] D. Oreski, (2022). Framework of Intelligent System for Machine Learning Algorithm Selection in Social Sciences. J. Softw., 17(1), 21-28.

[3] I. Pihir, K. Tomičić-Pupek, and M. T. Furjan, "Digital Transformation Insights and Trends," *29th CECIIS*, pp. 141–149, 2018.

[4] I. Pihir, "DIGITAL TRANSFORMATION AS UNIVERSITY COURSE - DEVELOPMENT AND IMPLEMENTATION," in *EDULEARN22 Proceedings*, 2022, pp. 7725–7729. doi: 10.21125/edulearn.2022.1799.

[5] L. Hrustek, M. T. Furjan, and I. Pihir, "Enabling Open Data Paradigm for Business Improvement," *56 th Int. Sci. Conf. Econ. Soc. Dev.*, no. May 2021, pp. 174–183, 2020.

[6] Tomičić Furjan, M., Tomičić-Pupek, K., Pihir, I. "Understanding Digital Transformation Initiatives: Case Studies Analysis", Business Systems Research, Vol. 11, No. 1, pp. 125-141, 2020, doi: 10.2478/bsrj-2020-0009

[7] Schwab, K. "The fourth industrial revolution", UK: Portfolio penguin, 2017.

[8] Roedder, N., Dauer, D., Laubis, K., Karaenke, P., Weinhardt, C. (2016), "The digital transformation and smart data analytics: An overview of enabling developments and application areas", IEEE International Conference on Big Data, IEEE, Washington, DC, USA, pp. 2795-2802, doi: 10.1109/BigData.2016.7840927

[9] Gartner inc., "Hype cycle research methodology," gartner.com, 2018, [Online]. Available: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle. [Accessed: Jan. 23, 2023]

[10] D. H. Wolpert, "The Supervised Learning No-Free-Lunch Theorems," in *Soft Computing and Industry*, London: Springer London, 2002, pp. 25–42. doi: 10.1007/978-1-4471-0123-9_3.

[11] R. Ali, A. M. Khatak, F. Chow, and S. Lee, "A case-based meta-learning and reasoning framework for classifiers selection," *ACM Int. Conf. Proceeding Ser.*, pp. 1–6, 2018, doi: 10.1145/3164541.3164601.

[12] G. Wang, Q. Song, and X. Zhu, "An improved data characterization method and its application in classification algorithm recommendation," *Applied Intelligence*, vol. 43, no. 4. pp. 892–912, 2015. doi: 10.1007/s10489-015-0689-3.

[13] R. Ali, S. Lee, and T. C. Chung, "Accurate multi-criteria decision making methodology for recommending machine learning algorithm," *Expert Syst. Appl.*, vol. 71, pp. 257–278, 2017, doi: 10.1016/j.eswa.2016.11.034.

[14] B. M. Franklin, "Automatic Selection of MapReduce Machine Learning Algorithms : A Model Building Approach," p. 237, 2018, [Online]. Available: http://digitalcommons.mtu.edu/etdr/604/

[15] P. Brazdil, C. Giraud-Carrier, C. Soares, and R. Vilalta, *Metalearning*. Cognitive Technologies, 2009. doi: 10.1007/978-3-540-73263-1_3.

[16] S. Sivakumar, S. R. Nayak, S. Vidyanandini, J. A. Kumar, and G. Palai, "An empirical study of supervised learning methods for breast cancer diseases," *Optik (Stuttg).*, vol. 175, pp. 105–114, 2018, doi: 10.1016/j.ijleo.2018.08.112.

[17] E. Güldoğan, İs. O. Yildirim, S. Sevgi, and C. Çolak, "Artificial Intelligence-Assisted Prediction of Covid-19 Status Based on Thorax Ct Scans Using a Proposed Meta-Learning Strategy," *Acta Medica Mediterr.*, vol. 38, no. 3, pp. 1515–1521, 2022, doi: 10.19193/0393-6384_2022_3_228.

[18] X. S. Zhang, F. Tang, H. H. Dodge, J. Zhou, and F. Wang, "MetaPred," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2019, pp. 2487–2495. doi: 10.1145/3292500.3330779.

[19] D. García-Saiz and M. Zorrilla, "A meta-learning based framework for building algorithm recommenders: An application for educational arena," *J. Intell. Fuzzy Syst.*, vol. 32, no. 2, pp. 1449–1459, 2017, doi: 10.3233/JIFS-169141.

[20] C. Romero, J. L. Olmo, and S. Ventura, "A meta-learning

approach for recommending a subset of white-box classification algorithms for Moodle datasets," *Proc. 6th Int. Conf. Educ. Data Mining, EDM 2013*, 2013.

[21] B. Bilalli, A. Abelló, and T. Aluja-Banet, "On the predictive power of meta-features in OpenML," *Int. J. Appl. Math. Comput. Sci.*, vol. 27, no. 4, pp. 697–712, 2017, doi: 10.1515/amcs-2017-0048.

[22] C. Castiello, G. Castellano, and A. M. Fanelli, "Meta-data: Characterization of Input Features for Meta-learning," 2005, pp. 457–468. doi: 10.1007/11526018_45.

[23] I. Khan, X. Zhang, M. Rehman, and R. Ali, "A Literature Survey and Empirical Study of Meta-Learning for Classifier Selection," *IEEE Access*, vol. 8, pp. 10262–10281, 2020, doi: 10.1109/ACCESS.2020.2964726.

[24] N. Dessì and B. Pes, "Similarity of feature selection methods: An empirical study across data intensive classification tasks," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4632–4642, 2015, doi: 10.1016/j.eswa.2015.01.069.

[25] C. Chen and M. L. Shyu, "Clustering-based binary-class classification for imbalanced data sets," *Proc. 2011 IEEE Int. Conf. Inf. Reuse Integr. IRI 2011*, pp. 384–389, 2011, doi: 10.1109/IRI.2011.6009578.

[26] O. Kwon and J. M. Sim, "Effects of data set features on the performances of classification algorithms," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1847–1857, 2013, doi: 10.1016/j.eswa.2012.09.017.

[27] M. Y. Kiang, "A comparative assessment of classification methods," *Decis. Support Syst.*, vol. 35, no. 4, pp. 441–454, 2003, doi: 10.1016/S0167-9236(02)00110-0.

[28] S. Ali and K. A. Smith, "On learning algorithm selection for classification," *Appl. Soft Comput. J.*, vol. 6, no. 2, pp. 119–138, 2006, doi: 10.1016/j.asoc.2004.12.002.

[29] A. Filchenkov and A. Pendryak, "Datasets meta-feature description for recommending feature selection algorithm," *Proc. Artif. Intell. Nat. Lang. Inf. Extr. Soc. Media Web Search Fruct Conf. AINL-ISMW Fruct 2015*, no. November 2015, pp. 11–18, 2016, doi: 10.1109/AINL-ISMW-FRUCT.2015.7382962.

[30] A. Rivolli, L. P. F. Garcia, C. Soares, J. Vanschoren, and A. C. P. L. F. de Carvalho, "Meta-features for meta-learning," *Knowledge-Based Syst.*, vol. 240, p. 108101, 2022, doi: 10.1016/j.knosys.2021.108101.

[31] Y. Peng, P. A. Flach, C. Soares, and P. Brazdil, "Improved dataset characterisation for meta-learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2534, pp. 779–784, 2002, doi: 10.1007/3-540-36182-0_14.

[32] D. Oreški, I. Pihir, and K. Cajzek, "Smart Agriculture and Digital Transformation on Case of Intelligent System for Wine Quality Prediction", MIPRO 2021, 44th International Convention Proceedings, pp 1565-1570, 2021, doi: 10.23919/MIPRO52101.2021.9596979

[33] D. Oreški, I. Pihir, and N. Kadoić, N, "Smart Agriculture: Machine Learning in Modelling Wine Quality Based on Laboratory or IoT Sensory Analysis", Proceedings of 56th Croatian and 16th International symposium on Agriculture / Rozman, Vlatka ; Antunović, Zvonko - Osijek : Faculty of Agrobiotechnical Sciences Osijek, pp 717-722, 2021.

[34] M. Vujic, L. Dedic, M. Tomicic Furjan, and I. Pihir, "The Benefits of Open Data in Urban Traffic Network". In: Knapčíková, L., Peraković, D., Behúnová, A., Periša, M. (eds) 5th EAI International Conference on Management of Manufacturing Systems. EAI/Springer Innovations in Communication and Computing. Springer, Cham, pp 267–282, 2022, doi: 10.1007/978-3-030-67241-6_22

[35] C. Alexopoulos, I. Pihir, and M. Tomičić Furjan, "Automatic End-to-End Decomposition and Semantic Annotation of Laws Using High-Performance- Computing and Open Data as a Potential Driver for Digital Transformation", 33rd Central European Conference on Information and Intelligent Systems, pp 189-194, 2022, doi: http://archive.ceciis.foi.hr/app/public/conferences/2022/Proceedings/DTCC/DTCC4.pdf

[36] D. Oreški, I. Pihir, and M. Konecki, "CRISP-DM process model in educational setting", Li Yongqiang, Anica Hunjet, Ante Roncevic (ed.), Economic and Social Development, 20th International Scientific Conference on Economic and Social Development, Book of Proceedings. pp 19-28, 2017, doi: https://www.esd-conference.com/upload/book_of_proceedings/Book_of_Proceedings_esdPrague_2017_Online.pdf

[37] D. Oreški, M. Konecki, and I. Pihir, "Predictive Modelling of Academic Performance by Means of Bayesian Networks", Konecki, M., Kedmenec, I. & Kuruvilla, A. (ed.), 47th International Scientific Conference on Economic and Social Development, Book of Proceedings, pp 435-441, 2019, doi: https://www.esd-conference.com/upload/book_of_proceedings/Book_of_Proceedings_esdPrague2019_Online.pdf

[38] Student grade prediction, https://www.kaggle.com/dipam7/student-grade-prediction, downloaded 01/02/2023.

[39] House sales prediction https://www.kaggle.com/datasets/harlfoxem/housesalesprediction, downloaded 01/02/2023.

[40] D. T. Larose and C. D. Larose, *Data Mining and Predictive Analytics*, 2nd Editio. Wiley, 2015. doi: 10.1016/b978-0-12-800229-2.00003-1.

[41] D. Sarkar, R. Bali, and T. Sharma, *Practical Machine Learning with Python (A Problem-Solver's Guide to Building Real-World Intelligent Systems)*, 1st ed. ed. Apress, 2018. doi: 10.1007/978-1-4842-3207-1_11.