Towards a Machine Learning Failure Prediction System Applied to a Smart Manufacturing Process

Tainá Da Rocha, Arthur Beltrame Canciglieri, Anderson Luis Szejka, Leandro Dos Santos Coelho, Osiris Canciglieri Junior

Towards a machine learning failure prediction system applied to a smart manufacturing process

Tainá da Rocha[1], Arthur Beltrame Canciglieri[1], Anderson Luis Szejka[1][0000-0001-8977-1351]*, Leandro dos Santos Coelho[1,2][0000-0001-5728-943X] and Osiris Canciglieri Junior[1][0000-0002-8503-9275]

[1]Industrial and Systems Engineering Graduate Program, Pontifical Catholic University of Parana, Curitiba, Brazil
[2]Department of Electrical Engineering (PPGEE), Federal University of Parana (UFPR), Polytechnic Center, Curitiba, Brazil
*anderson.szejka@pucpr.br , osiris.canciglieri@pucpr.br , leandro.coelho@ufpr.br

Abstract. At a time when the competitive market is operating rapidly, manufacturing industries need to stay connected, have interchangeability and interoperability in their factories, ensuring that there is heterogeneous communication between sectors, people, machines and the client, challenging the manufacturing industry to discover new ways to bring new products or improve their manufacturing process. Precisely because of the need to adjust to these new market demands, factories pursue complex and quick decision-making systems. This work aims to propose applications of Machine Learning techniques to develop a decision-making platform applied to a manufacturing line reducing scrap. This goal will be achieved through a literature review in the fields of Artificial Intelligence (AI) and Machine Learning to identify core concepts for the development of a failure prediction system. This research has demonstrated the problems and challenges faced by manufacturing daily, and how, through the application of AI techniques, it is possible to contribute to assist in these problems by improving quality, performance, scrap rates and rework, through connectivity and integration of data and processes. This paper contributes to evaluate the performance of machine learning ensembles applied in a real smart manufacturing scenario of failure prediction.

Keywords: Machine Learning, Interoperability, Industry 4.0, Artificial Intelligence, Integration Manufacturing.

1 Introduction

Performance gains have been discussed lately, not only in the manufacturing perimeter, but also in different areas, especially in the business strategy, which has shown prominence [1]. This theme is increasingly correlated with the Industry 4.0 paradigm, which is supported by nine pillars: Collaborative Robotics, Simulation, Systems Integration, Industrial Internet of Things, Cyber Security, Cloud Computing, Additive Manufacturing, Augmented Reality and Big Data and Analytics. Its goal is to make

processes faster, more flexible and more efficient, promoting the union and/or representation of physical resources with digital ones, connecting machines, systems and assets to produce higher quality, more profitable, lower-cost items resulting in a better performant process [2].

In this context, it is necessary to take as support a combination of these pillars to achieve this result. Currently, these pillars form a strategy for implementing techniques from Industry 4.0, bringing benefits such as manufacturing flexibility, autonomy and intelligent make-decision through the data processing of an integrated manufacturing system and providing data quickly and reliably to assist in faster decision making [3]. Factories must change their physical and procedural structures, integrating horizontally and vertically, concepts and paradigms, respecting the addendums of the third industrial revolution: Automation. At this stage, the manufacturing industries already encounter several problems due to obsolete machinery and systems that need to be automated and connected to a network when possible.

The motivation for this work came from the need to predict and reduce parts failure, optimize performance indicators and reduce scrap costs through a regression approach. In this way, this paper presents an intelligent test system applied to a metal-mechanic assembly line through performance metrics such as Mean Absolute Error (MAE). The main research contribution is the use of artificial intelligence techniques to promote interoperability between machines and systems, reduce the information gap, make improvements in production lines and failure prediction between operations to optimize performance indicators, reduce costs in rework, optimize workflow and increase quality.

The remainder of this work is as follows: Section 2 illustrates the problem statement addressed in this work. Section 3 presents the Technological Background with concepts based in the literature. In session 4 exposes the conceptual proposal applied in the case study. Section 5 shows the preliminary results of the application in this case study. Finally, session 6 presents the conclusion and future works.

2 Problem Statement

The lack of communication between manufacturing operations causes many impacts, among them the loss of productivity [4][5]. In this work, the focus is the integration of data communication of two specific operations in an industry of the metal-mechanical sector, having as objective the improvement of the communication between stations, anticipating through a forecast, under the possible failures in products, consequently reducing the waste of the operation. For reasons of industrial manufacturing secrecy, the data cannot be disclosed in this work and it was uncharacterized both the characteristics and the name of the operation.

A schematic drawing is presented in Fig. 1 to better illustrate the adjustment performed. "X" represents the measurement of the internal dimension of the cylinder. "Y" represents the measurement of the other cylinder's height, which is fitted to the first cylinder mentioned. "Z" represents the compensation/adjustment washer.

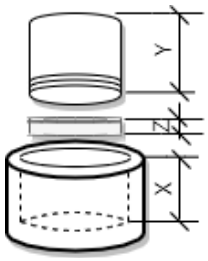

Fig 1. Product Illustration

In this production process, one operation (station D) is responsible for making a mechanical adjustment and another operation ahead (station J) for checking this previously adjusted measurement. The problem between them is that according to process influences, oscillation and deviations as product geometry, machining influences, and so on, it is necessary to perform an adjustment at station "D" according to measurements from station "J". This adjustment is performed manually, using a measurement compensation washer, and has a 15 pieces delay (Fig. 2). Therefore, the purpose of this article is to optimize this adjustment so that it is anticipated by forecasting using Machine Learning techniques.

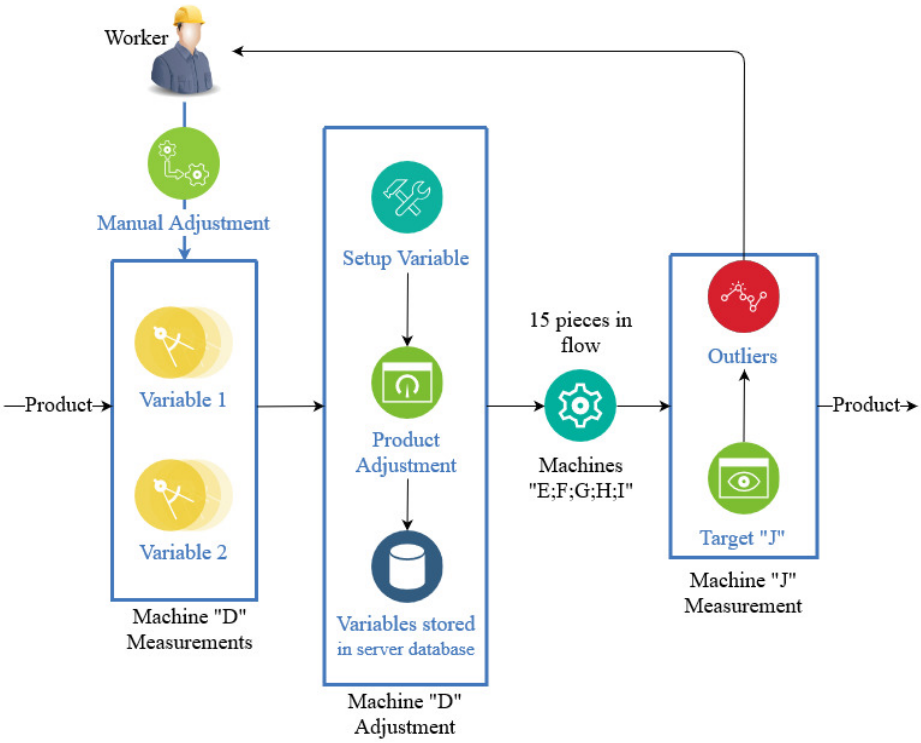

Fig 2. Conceptual proposal architecture.

3 Technological Background

Nowadays, the information is dynamic, and its complete understanding is essential for quality control of the resulting product. It is necessary to involve a series of sectors from the suppliers to the delivery for a customer, where it goes through several stages consisting of requirements such as quality, personalization, cost, time, budget, for example, that need to be respected, and this is when there are hidden risks due to misinformation, misunderstanding, divergence and indirect information [6]. These risks, also called impacts, arise from heterogeneity in the sector (domain) and its peers semantically.

3.1 Industry 4.0

Industry 4.0, also called "Smart Factory", is the next revolution on the industry scene. According to [7], sensors, machines, workpieces and IT (Information Technology) systems will be connected along the value chain in addition to a single company. This cyber-physical system (CPS), is a set of transforming technologies that allow the connection of physical asset operations between computational resources. The CPS is controlled and monitored by computer-based algorithms and is fully integrated with its users (objects, humans, and machines) via the Internet [8], being able to interact with each other and analyze data to predict failures, configure and adapt changes. Sector 4.0, or I4.0, will make it possible to gather and analyze data between machines, enabling faster, more flexible, and more efficient processes to produce higher-quality goods at a reduced cost.

Also, automatic solutions will adopt versatile operations, consisting of operational components, devices, and analytics, such as "autonomous" manufacturing cells and adjustments that independently control and optimize multi-step manufacturing [9].

3.2 Artificial Intelligent

Artificial Intelligence (AI) enables systems to make decisions independently, supported by digitally established pattern logic, and can thus resemble situations, think about responses, make decisions, or act preventively. Thus, machine learning is a specific strand of AI that trains machines to learn from data, the closer to the data and real scenario, better.

Process management requires, at all levels, access, and display of the data necessary to oversee, diagnose and report the current status of the process [10]. But the existence of this type of technology does not inhibit human capacity because it is dependent on human training, teaching, creativity, and supervision, but it becomes a time optimization tool for repetitive and high-data decision-making activities. The creation of a data processing platform is the product of a group of information and techniques, which are: process data set to be applied, AI method, and, finally, the goal to be achieved. These methods from AI, belong to its subgroup called Machine Learning (ML).

Machine learning focuses on the question of how to build algorithms and computers

that improves through experience automatically. It is one of today's trends among data scientists, lying in the intersection of computer science and statistics, within the core of artificial intelligence and data science [11]. ML is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. Thus, instead of being programmed for specific actions only, machines use complex algorithms to make decisions and interpret data, automatically performing tasks. These programs can learn from the high power of data processing without human intervention. Waze, Netflix, Siri (Apple) are examples of use of the Machine Learning. Below follows a brief explanation of the ML methods used in this paper:

- K-Nearest-Neighbor (KNN)s: It is a classifier where learning is based on analogy [12]. The KNN algorithm is a supervised machine-learning algorithm used for classification and regression problems. Its operating principle is to separate individuals into groups (or classes) according to the existing similarity [13].
- Gradient Boosting Machine (GBM): It is a decision tree-based machine-learning algorithm that uses a gradient boosting framework, thus having several adjustment hyperparameters [14].
- Random Forest (RF): This technique generates several decision trees during training that can be randomly divided from a starting point [15]. This results in a "forest" of generated decision trees where the results are clustered by the algorithm [16]. The RF algorithm presents several advantages; it runs efficiently on large datasets, it is not sensitive to noise or over-fitting, it can handle thousands of input variables without variable deletion, and it has fewer parameters compared and adjustment with that of other machine-learning algorithms [17].
- Cross-Validation: Today's machine learning methods tend to overfit into the large datasets used for training and validating the algorithms. To solve this problem many data scientists have chosen to use cross-validation methods to attempt improving the robustness of the machine learning methods. In [18], indicate that the cross-validation method is one of the best techniques to evaluate the predictive performance of a model involving large samples. The process of Cross-validation is through randomly partition the sample into equally sized subsamples named folds. Finally, it is chosen the parameter with the best estimated average performance increasing the efficiency of the machine learning technic.

4 Discussion

Firstly, it was created a database with 35,634 rows of data to be used in the training and validation steps of the machine learning methods. As previously mentioned, for reasons of signed industrial confidentiality, the data cannot be shown in this work, so the name of the characteristics, as well as the name of the operation. The database includes two product variables, a setup variable, a product adjustment, and the target value of machine "J". The data was normalized between 0 and 1 and is shown in Tab. 1.

Tab. 1 Normalized data of test dataset

	Variable 1	*Variable 2*	*Setup Variable*	*Product Adjustment*	*Target "J"*
mean	0.2429	0.8656	0.0841	0.5027	0.2966
std	0.0090	0.0162	0.0488	0.1480	0.0662

With the dataset ready, a study of variable correlation between machine "D" and machine "J" was made to understand the linearity of the problem, since machine "D" makes a product adjustment in production and machine "J" measures this adjustment with higher precision, indicating that the variables from both machines should have in theory high linear correlation.

However, it was found that although theoretically there was a high correlation between the two variables (Variable 1 and Variable 2) from machine "D" and the measuring variable from machine "J", there was a correlation of less than 25%. The setup variable had a high correlation of almost 80% because it is a constant that changes only with a manual entry from the programmer and is adjusted within machine "J" parameters. All this information indicates that the automatic adjustment problem was more complex and could not be solved with linear regressions, indicating the need for Machine Learning methods along with improvements in the measurement process for good accuracy.

From this, it was necessary to establish a measurement basis for the accuracy of the methods that would be tested, so that it was possible to compare and validate the result. It was performed a study of the current state of the adjustment of machine "D" it was found that the worker made the adjustment based on the average of 9 measurements of machine "J" and observed the trend of the data and its outliers. Therefore, the worker would adjust downward if the 9 pieces were rising, and up if the 9 pieces were falling. However, this form of measurement is not periodically controlled, due to being checked visually by the operator, which is not exclusive for this task, and the machine operator "J" needed to indicate outliers to the machine operator "D" for an operating adjustment to be made.

Thus, it was seen that the current state of the adjustment process between machines "D" and "J" could be translated by a simple 9-measurement moving average method with 15-piece measurement delay, since there is a delay of 15 pieces between setting machine "D" and measuring it on machine "J".

After understanding the functionality of the process, its measurements and characteristics, the data from a manufacturing period were extracted and mined. At this stage, it was understood that these were non-linear variables. The time applied to understand the reality of production was of fundamental importance, as well as the knowledge of operators, experience of adjustments in "n" types of products, and the difference in how each product behaves in the same production line.

The data were extracted from historical databases, and were manually mined, excluding outliers, duplicate data, columns without information, thus preparing the data to be worked on in the methods to be chosen.

The ML methods used were chosen through trial and error. Method by method was

tested, its result, assertiveness, and stability evaluated. It was concluded that the best scenario found was a combination of methods.

The choice of the best model was carried out through the help of programming software for machine learning "H_2O", where combinations of models and ensembles were tested, using the Gradient Boosting Machine (GBM), Random Forest (RF) method as input, General Linear Models (GLM). Through this, it had a product, the best combination indicated for the application according to the problem presented.

Consequently, it was proposed three models of machine learning applied to the problem: Random Forest, Gradient Boosting Machine, and K-nearest neighbor. These models used only two variables of part measurement operation "D" and one set-up variable to obtain a prediction of the measurement result of operation "J" and thus perform the adjustment automatically. The conceptual proposal of horizontal integration of two manufacturing machines through Machine Learning tools can be seen in the figure below (Fig.3).

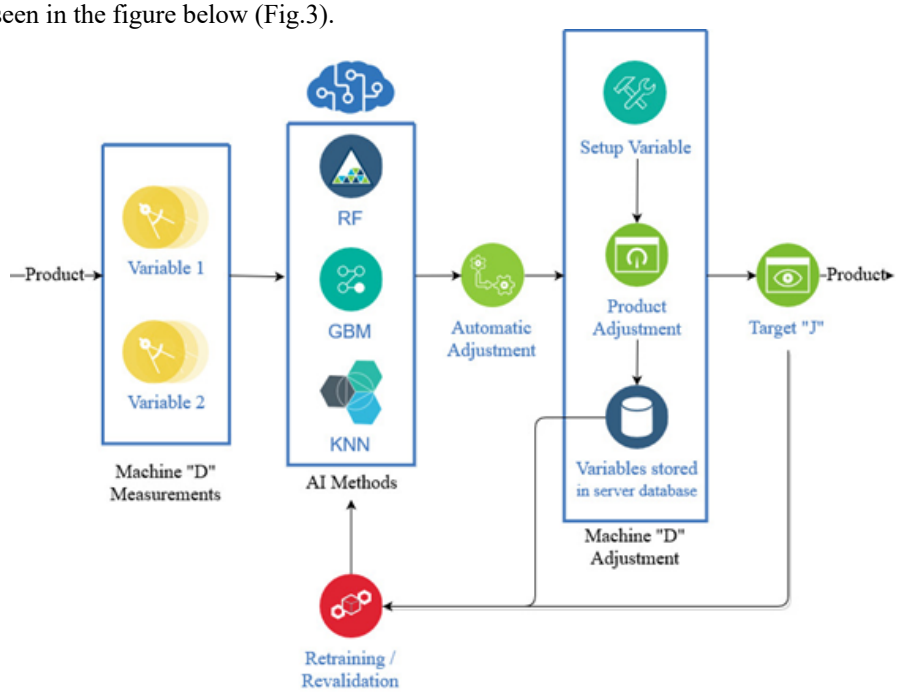

Fig.3 Proposed automatic adjustment architecture.

Fig. 3 shows the desired future scenario, in which machine "D" will perform product measurements (variable 1 and variable 2), then it will send the measurement data for processing in the selected AI method, which will send a signal with Box Magazine, according to the thickness washer that will best approach the chosen target to be reached in the "J" measurement operation. This Fig. 3 states, the moment when AI will be applied, as well as the replacement of the production operator, thus ceasing to be manual, becoming automatic. Another observation to make is that the AI analyzes the trend forecast for each product measurement and no longer as done in the scenario

presented (Fig. 2), which waited an average of 9-15 pieces to perform adjustments, if necessary. The problem is that many times this range of parts was outside the established limits, generating rework and scrap.

5 Preliminary Results

In this paper, we present a case study testing the proposed models (Random Forest, Gradient Boosting Machine, and K-nearest neighbor) in the previous item compared with the moving average model that represents the current state of the process. Each test was made using 35,634 test pieces and a cross-validation method of 5 folds to avoid overfitting of the models and that they could forecast with new data and parts without losing its accuracy, avoiding failures/rework, refuse, due to this problem of interoperability between workstations, thus optimizing the production process.

The result of the model validation is presented below in Tab. 2 with their square mean error, standard deviation of predictions, mean absolute error in microns, and their percentage improvement based on the current state. The ensembles were tested using weighted average from the model's prediction results of Gradient Boosting Machine (GBM), K-nearest neighbor (KNN), and Moving average (MA). The ensembles did not use the Random forest predictions since it had a worse performance than the moving average used as a baseline.

Tab. 2 Result of proposed ML methods and ensembles

Model	***Parameters***	***Square mean error (Microns^2)***	***Standard deviation of prediction (Microns)***	***Mean Absolute error (Microns)***	***Improvement (MAE)***
Moving average	n = 9 pieces delay = 15 pieces	16.7993	2.0847	**2.1814**	Base line
Random Forest	estimators = 1000	17.2245	1.9845	**2.3250**	**-6.585%**
K-Nearest Neighbours	neighbours = 100 weights = distance	14.3738	0.8026	**1.9304**	**11.50%**
Gradient Boosting Machine	estimators = 25 max_depth = 6 min_samples_split= 2 learning rate= 0.2 loss= Huber	14.3823	0.7503	**1.8212**	**16.51%**
Ensemble 1	0.5*GBM+0.5*KNN	14.1492	0.6470	**1.8437**	**15.48%**
Ensemble 2	0.9*GBM+0.1*KNN	14.2999	0.7103	**1.8197**	**16.58%**
Ensemble 3	0.4*GBM+0.3*KNN+ 0.3*MA	13.8360	0.7863	**1.8550**	**14.96%**
Ensemble 4	0.85*GBM+0.05*KN N+0.1*MA	14.1162	0.6939	**1.8135**	**16.86%**

As can be seen in Tab. 2, the best result was from model ensemble 4 with an improvement of 16.86% over the baseline model. This represents that the cloud computing with the machine learning models being ensemble could predict the result

in machine “J” almost 20% better than the operator could and then make automatically the adjustment for better control of the whole process of manufacture.

6 Conclusion

Throughout this case, we can conclude that the best standalone algorithm with the dataset presented was the Gradient Boosting Machine, and that the boosting technic helped the prediction of values more precisely. Also, the random forest method did not meet the expectation and presented a worse result than a simple moving average, not being able to adapt to the information presented. The ensembles presented were done only with a weighted average from the prediction values, needing more studies for stacked ensembles and more advanced methods of ensembling the machine learning algorithms.

It was also seen that the Industry 4.0 concepts have difficulties to be implemented, even with all its digital transformation over the years, regarding data collection and reliability, machine communication, consistency and stabilization of production processes, and people's resistance to the use of this technology. Throughout the implementation of this case, it was difficult to identify new features to be used, with some not being measurable, or even, stored in a database, by today’s process because they deal with external influences such as temperature and humidity, and others are not saved by the database.

We can conclude, after the application of AI, that because it is a nonlinear case, some Machine Learning techniques do not adapt, or do not have good accuracy and estimated error results, so it was used several techniques, from which the best result was almost 17% improvement, according to the current scenario. However, there are other techniques, an ensemble of techniques, of which still do not have much study on, but are demonstrating to be significant, insertion of new features and adjustment of the hyper-parameters that may contribute improving this result, thus being an input for future work (article).

Acknowledgments

The authors would like to thank Pontifical Catholic University of Parana (PUCPR) and Robert Bosch CtP for the financial support to the development of this research.

References

1. Neely, A., Kennerley, M.: Measuring performance in a changing business environment. *International Journal of Operations and Production Management*. **23**, 213-229 (2003).
2. Zdravkovic, M., Panetto, H.: The challenges of model-based systems engineering for the next generation enterprise information systems. *Information Systems and e-Business Management*. **15**(2) 225–227 (2017).
3. Ghobakhloo, M.: The future of manufacturing industry: a strategic roadmap toward Industry 4.0. *Journal of Manufacturing Technology Management*. **29**(2), 910–936 (2018).

4. Szejka, A. L., Aubry, A., Panetto, H., Canciglieri Júnior, O., Loures, E. R.: Towards a conceptual framework for requirements interoperability in complex systems engineering. Springer. *Lecture Notes in Computer Science*. **8842**, 229–240 (2014).
5. Szejka, A.L.; Canciglieri Junior, O.; Rocha Loures, E.; Aubry, A.; Panetto, H.: Requirements interoperability method to support integrated product development. In. *45th Conference on Computers and Industrial Engineering*, Metz, October 28–30. 1-10, (2015).
6. Karabegovic, I.: The role of industrial and service robots in the 4th Industrial Revolution - Industry 4.0. *ACTA Technica Corviniensis*. Tome XI, **2**, 11-16 (2018).
7. Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., & Harnisch, M.: Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston Consulting Group*. **9**, 54-89 (2015).
8. Adamczyk, B. S., Szejka, A. L. & Canciglieri, O. Knowledge-based expert system to support the semantic interoperability in smart manufacturing. *Computers in Industry.* **115**, 103161 (2020).
9. Lasi, H., Fettke, P., Kemper, H. G., Feld, T., & Hoffmann, M.: Industry 4.0. *Business & information systems engineering*. **6**(4), 239-242 (2014).
10. Reis, M.S., Kenett, R.: Assessing the value of information of data-centric activities in the chemical processing industry 4.0. AIChE. *AIChE Journal*. **64**(11), 3868–3881 (2018).
11. Jordan, M.I. and Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science, **349**(6245), 255-260 (2015).
12. Hadad, H.M., Mahmoud, H.A., Ali Mousa, F.: Bovines Muzzle Classification Based On Machine Learning Techniques. *International Conference on Communication, Management and Information Technology*. **65**, 864 – 871 (2015).
13. Zhu, X., Cheng, D., Zong, M., Li, X., Zhang, S.: Learning k for KNN classification. ACM Trans. *Intell. Syst. Technol*. **8**(3) 1-19 (2017).
14. Natekin, A., Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. **7**(21) 1-10 (2013).
15. Andronicus, A.A., Aderemi, O.A.: Classification of phishing email using random forest Machine Learning Technique. *Journal of Applied Mathematics*. **2014**(425731), 1-6 pages.
16. Lyncha, A.M., Abdollahib, B., Fuquac, J.D., de Carloc, A.R., Bartholomaic, J.A., Balgemannc, R.N., Berkeld, V.H., Frieboes, H.B.: Prediction of lung cancer patient sur- vival via supervised machine learning classification techniques. *Int J Med Inform*. **108**, 1–8. (2017)
17. Wang, L.A., Zhou, X., Zhu, X., Dong, Z. and Guo, W.: Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, **4**(3), 212-219 (2016)
18. Pinto, J.M. and Marçal, E.F.: Cross-validation based forecasting method: a machine learning approach. CEQEF. **49**, 1-18 (2019).