The 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
22-25 September, 2021, Cracow, Poland

Application of Machine Learning Methods in Carbon Footprint Optimization

Piotr Milczarski,
Faculty of Physics and Applied Informatics, University of Lodz, Pomorska str. 149/153, Lodz, Poland
piotr.milczarski@uni.lodz.pl

Abstract—**In the paper, the study of the carbon footprint optimization process is shown in order to receive low-carbon products. Three clusterization methods are used to the frozen vegetables production line in order to derive the models to assess the carbon footprint for the onion and spinach production in the factory. Three results of the clusterization by three chosen methods are assessed by five classification methods: k-Nearest Neighbors, Multilayer Perceptron, C4.5, Random Forrest and Support Vector Machines with a radial basis kernel function. In the chosen model with five clusters, the best clusterization methods are k-means followed by Canopy.**

Keywords—Carbon Footprint; clusterization; Canopy, k-means, Expectation-Maximization; k-Nearest Neighbors; Multilayer Perceptron; C4.5; Random Forrest; Support Vector Machines;

I. INTRODUCTION

Greenhouse gas emissions from human activities have been a major contributor to global warming since the mid-twentieth century. Agriculture and land-use change contributed to 17% of global anthropogenic greenhouse gas emissions in 2010 [1]. By 2050 the population will be 9 billion people [2] to ensure supplyment of food, agricultural production should be increased by 60%. Climate change can affect food availability; for example, an increase in temperature, a change in the structure of rainfall or extreme weather events may result in a reduction in agricultural productivity [3, 4]. Therefore, its main challenge has become to mitigate the threats that climate change poses to food security.

In the case of the CFOOD project, it is focused a focus on the optimization of the frozen food production process, so we consider a segment of the product life cycle from the moment of raw material delivery to the shipment of the finished frozen food to the recipient.

The methods of calculating the carbon footprint are most often based on well-known standards. Among them, the most used are: ISO14040: 2006 [5] – Environmental management-life cycle assessment: principles and framework, ISO14064-1: 2018 [6] – Greenhouse gases - Part 1: Specification with guidance at the organization level for quantification and reporting of greenhouse gas emissions and removals, ISO/TS 14067:2018 [7] – Greenhouse gases - Carbon footprint of products - Requirements and guidelines for quantification, PAS2050 [8] – Specification for the assessment of the life cycle greenhouse gas emissions of goods and services.

II. CFOOD PROJECT AND PRODUCTION ASSESSMENT

In the CFOOD projects, the production process is divided into several smaller stages e.g. for the onion production:

- S1 – initial cooling of the raw materials before the processing;
- S2 – the raw material preparation for the production;
- S3 – raw material pre-processing on the production line;
- S4 – product freezing in the cold tunnel;
- S5 – product preparation to coldstore.

Each of the process stages is connected to electric meter units. Each production stage has also a preparation phase that is measured separately, e.g. S1 has a preparation phase that is denoted pS1, etc.

The stages S1 and S4 have the biggest impact on energy utilization because they are connected with freezing processes. In Tab 1 there are presented the results in kWh/t for some stages and their preprocessing ones.

During the monitoring of the processes, the authors measured the parameters more thoroughly at the stages S1-S5. We have checked the whole production process that usually lasts 24-36 hours and its output is around 20-100 tons of production. During each supervision of the production line in Unifreeze, we have gathered not only average values e.g. input mass of raw materials and output production assets (see Fig. 1), the temperatures of the materials, etc. As we can see the stage S1 energy consumption is very chaotic, but S3 and S4 have smooth power values.

In the research section, we have tested several of clusterization methods. and choose three: Canopy, k-Means (KM) and Expectation-Maximization (EM) [9][10]. We have tested several options with the cluster numbers and chosen five clusters for each method that should represent according to our experience some real-time situations that occur during the production and their accounting systems:

- Optimal production – the product has the temperature -25°C ± 2 °C at the end of the line and the production goes without obstacles

- Close to optimal – during the high season energy consumption should be lower so as to make through-output higher. That is why the product temperature is allowed to be between -6°C and -18°C.
- Wrong accounting of some parameters. Sometimes, operators of the system can make mistakes. That will result in too high or too low results e.g. the through-output.
- Malfunction of the energy meters. It is a different situation from the above one and might result in random results.

The clusterization model with five clusters should have at least 80-100 processes. After a year of the process measurement, till June 2021, we have collected 152 results only for the frozen onion production and 75 for the spinach. The other vegetables have less than 50 cases..

III. Research Methodology

In the previous work [11][12] to assess the production processes we have prepared the set of verified data and to assess the trustworthiness of the production data we have compared the results of processes classification using 5 classifiers: k-Nearest Neighbors, Multilayer Perceptron [13], C4.5, Random Forrest and Support Vector Machines with a radial basis kernel function [9].

In the current paper, we have focused on unsupervised methods i.e. clusterization [9] into the onion and spinach processes..

In Tables 2-4 and 6-8 there are classification results of the production processes using the following clusterization methods with five clusters as it was explained in Sec. III:

- Canopy: max-candidates = 100; periodic-pruning = 10000 ; min-density = 2.0; T2 radius = 0.804 and T1 radius = 1.005
- k-Means (KM) with Euclidean distance, max-candidates = 100, periodic-pruning = 10000, min-density = 2.0, T1 = -1.25 and T2 = -1.0.
- Expectation–Maximization (EM) with max-candidates = 100, "minimum improvement in log likelihood" = 1E-5, "minimum improvement in cross-validated log likelihood" = 1E-6, and "minimum allowable standard deviation" = 1E-6 .

IV. Results for the Onion Production

Corresponding clusters for the three chosen methods are very similar:

TABLE I. THE CLUSTERIZATION OF THE ONION PRODUCTION RESULTS USING CANOPY, K-MEANS (KM) AND EXPECTATION-MAXIMALIZATION(EM) CLUSTERIZATION.

ID	Mass[kg]	pS1	S1	S3	pS4	S4	pt	Canopy	KM	EM
87	32904	0,067	1,711	1,168	0,920	26,642	2,256	cluster0	cluster2	cluster3
98	37040	0,305	6,646	1,021	0,763	28,547	2,327	cluster0	cluster3	cluster3
101	46660	0,146	4,109	0,955	0,728	25,151	2,679	cluster3	cluster3	cluster3
102	34020	0,369	11,363	1,311	1,105	38,498	1,552	cluster2	cluster2	cluster0
175	80408	0,037	9,030	1,232	0,497	33,598	1,973	cluster4	cluster3	cluster3
237	18600	27,824	53,249	14,191	30,384	32,829	1,860	cluster2	cluster0	cluster1
311	25860	0,025	1,823	0,149	0,151	4,082	2,111	cluster1	cluster4	cluster4
370	30820	0,101	1,919	0,312	0,802	14,716	5,779	cluster3	cluster1	cluster2
392	64740	0,052	3,670	1,271	0,642	34,832	2,002	cluster4	cluster3	cluster3
393	50060	0,079	2,096	0,967	0,808	28,956	2,356	cluster0	cluster3	cluster3
394	36270	0,037	1,855	1,534	0,263	9,552	2,053	cluster1	cluster4	cluster4
1370	24750	0,064	1,577	1,302	1,513	33,441	2,020	cluster4	cluster3	cluster3
2387	20530	0,087	1,917	1,166	1,822	31,935	2,106	cluster4	cluster3	cluster0
2426	41352	11,032	3,171	0,648	59,214	17,813	1,894	cluster3	cluster1	cluster2

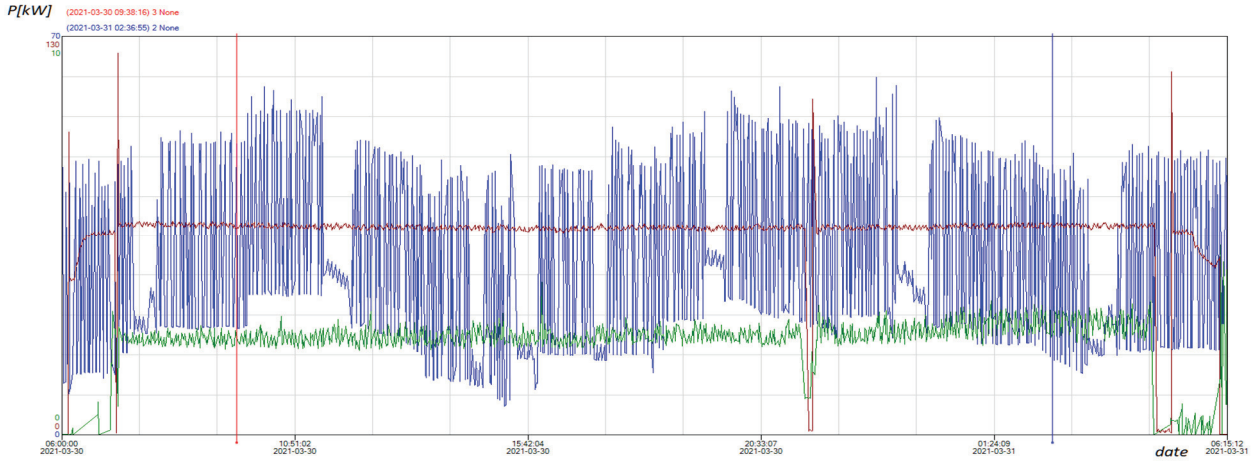

Figure 1. Energy consumption in the stage S1(blue), S3 (green) and S4(brown), during the process 2426. The time axis shows the current date.

- Canopy C0, KM C3, EM C3
- Canopy C1, KM C4, EM C4.
- Canopy C2, KM C1, EM C2

In the discussion below and their corresponding tables, we have highlighted the optimal clusters. All values for the stages and their preprocessing phase are in kWh/ton, the production through output (pt) in [ton/h]. Canopy seems to provide the best assessment of the processes because it's the best cluster that has the lowest energy consumption from the three optimal clusters for each clusterization To assess and to choose the clusterization method we have used five machine learning methods as in our previous work [11][12]. All the clusterization results were assessed by the classification methods with the same parameters. In Tab. 5 there are classification results of the production processes using the following classifiers:

- 3NN (kNN) 3-Nearest Neighbors;
- Multilayer Perceptron (MLP) with a hidden layer with 16 nodes for the onion and 15 nodes for the spinach production with a learning rate equal to 0.79 and momentum equal to 0.39 [13];
- binary tree C4.5 with a confidence factor equal to 0.25, with a minimum number of instances per leaf equal 2;
- Random Forrest (RF) with the bag size percent equal to 100, with maximum depth unlimited, number of execution slots equal to 1 and 100 iterations;
- Support Vector Machine (SVM) with a radial basis function (RBF) given by the Eq. (1):

$$K(x,y) = exp(-0.05*(x-y)^2) \quad (1)$$

Canopy clusterization

Cluster 0 has a lower output of 2.29 ton/h and the lowest preprocess values, but it is the first choice as the optimal process. Cluster 1 has too low S4 stage energy. Possible cause: incorrect accounting in the system by the operator. Cluster 2 too high S1and S4, lower output. Cluster 3 has the highest through-output, lower energy consumption, but the processes: 201, 2426, 2388 have too high energy consumption at the preprocessing stages pS1 and pS4. Cluster 3 occurs during the high season. Cluster 4 has a high value of S4, but it would be the 2nd choice for the optimized process cluster.

TABLE II. THE CANOPY CLUSTERIZATION RESULTS OF THE CENTROIDS PARAMETERS OF THE CHOSEN STAGES.

Cluste	Division	pS1	S1	S3	pS4	S4	pt
0	**62**	**0.18**	**4.43**	**0.98**	**0.98**	**25.7**	**2.29**
1	28	0.06	2.81	0.83	0.05	0.18	2.34
2	5	0.36	11.3	1.31	1.11	38.5	1.55
3	20	0.12	2.40	0.83	2.42	22.8	2.93
4	37	0.05	2.93	1.25	1.16	34.1	2.02

K-means clusterization

Cluster 0 – has very high energy consumption. Possible cause – wrong data. Cluster 1 has high preprocess values pS1 and pS4. Cluster 2 would be the 2nd choice as the optimal process. Cluster 3 is 1st choice for optimal process cluster because it has fair freezing stages energy values and low energy consumption in other stages. Cluster 4 has too low the S4 stage value.

TABLE III. THE K-MEANS CLUSTERIZATION RESULTS OF THE CENTROIDS PARAMETERS OF THE CHOSEN STAGES.

Cluster	Division	pS1	S1	S3	pS4	S4	pt
0	1	27.82	53.25	14.20	30.38	32.83	1.86
1	3	6.23	3.43	0.47	25.60	15.17	3.45
2	20	0.14	7.16	1.05	0.89	30.56	2.18
3	**97**	**0.15**	**4.79**	**1.05**	**1.01**	**28.95**	**2.34**
4	31	0.06	2.72	0.67	1.16	34.11	2.02

EM clusterization

Cluster 0 has fair values and it would be 2nd choice for the optimal processes. Cluster 1 – too high values due to a wrong accounting. Cluster 2 –has other preprocess values high. Cluster 3 is 1st choice, and its preprocess values are rather low. Cluster 4 has too small S4 values.

TABLE IV. THE EM CLUSTERIZATION RESULTS OF THE MEAN PARAMETERS OF THE CHOSEN STAGES.

Cluster	Division	pS1	S1	S3	pS4	S4	pt
0	21	0.25	8.09	1.10	1.37	29.90	2.24
1	1	27.82	53.25	14.19	30.38	32.83	1.86
2	5	4.24	4.08	0.48	18.18	14.94	2.94
3	**95**	**0.97**	**4.55**	**1.05**	**0.78**	**29.19**	**2.33**
4	30	0.05	2.56	0.68	0.16	6.53	2.14

TABLE V. THE ASSESSMENT OF THE CLUSTERIZATION METHODS BY FIVE MACHINE LEARNING METHODS.

Classifier	Classification results [%]		
	Canopy	KM	EM
3NN	98,7	**98.0**	87.5
C4.5	98.0	**99.3**	98.7
MLP	92.1	**97.4**	89.5
RF	100	**100**	100
SVM	96.1	**96.7**	88.8

The best clusterization method is k-means followed by Canopy.

V. RESULTS FOR SPINACH PRODUCTION

Spinach has already 75 full processes in the CFOOD database. In the discussion below and their corresponding tables, we have highlighted the optimal clusters. All values for the stages and their preprocessing phase are in kWh/ton, the production through output (pt) in [ton/h] and an average energy consumption in one hour Et in [kWh/h]. Canopy seems to provide the best assessment of the processes because it's the best cluster that has the lowest energy consumption from the three optimal clusters for each clusterization. Figures 2-4 show the clusters for the corresponding Tables 6-8.

Canopy clusterization

Cluster 3 has the best results and it would be the 1st choice for the optimized process cluster.

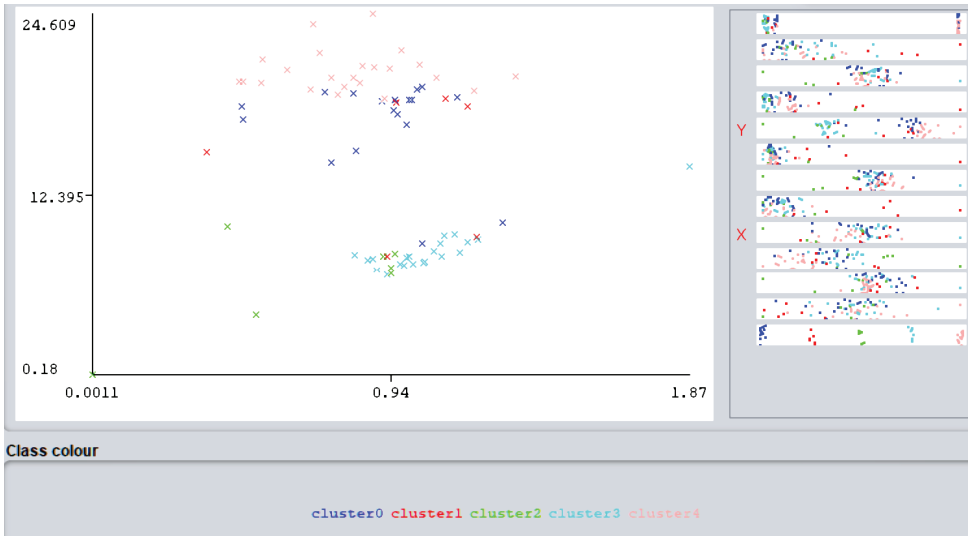

Figure 2. The Canopy cluster division in the space (S4, S2): where the stage S4(x-axis), S2 (y-axis).

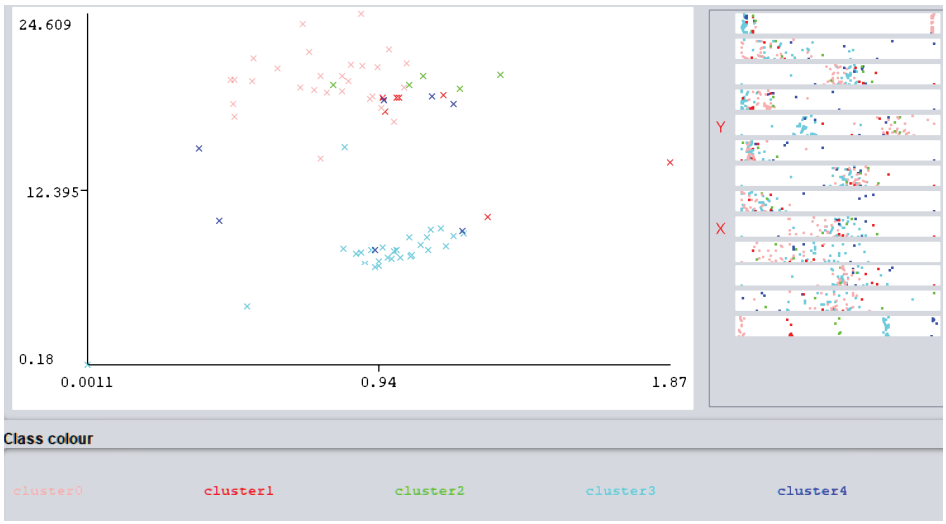

Figure 3. The k-means cluster division in the space (S4, S2): where the stage S4(x-axis), S2 (y-axis).

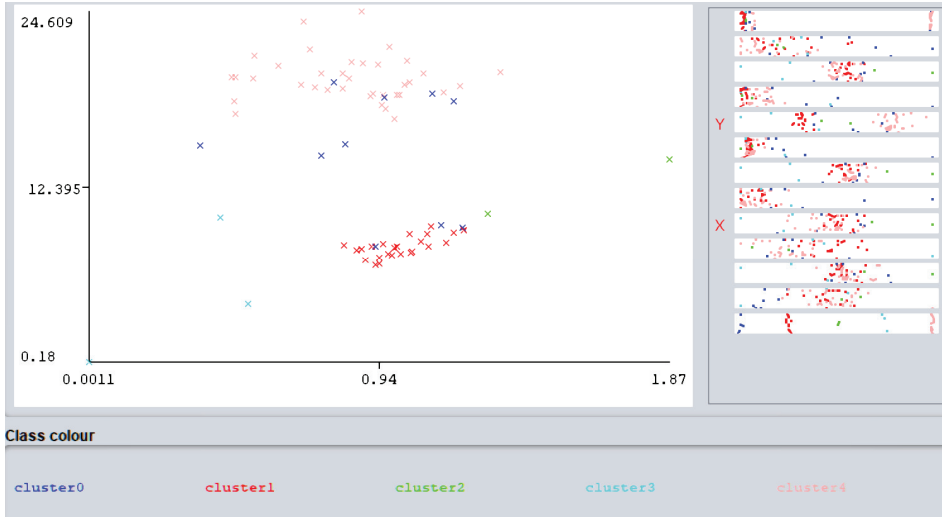

Figure 4. The Expectation–Maximization cluster division in the space (S4, S2): where the stage S4(x-axis), S2 (y-axis).

TABLE VI. THE CANOPY CLUSTERIZATION RESULTS OF THE CENTROIDS PARAMETERS OF THE CHOSEN STAGES AND PARAMETERS.

Cluster	Cases	S1	S2	S3	S4	pt	Et
0	18	31.83	15.49	16.79	0.91	1.70	147.7
1	6	34.70	13.54	16.83	1.06	1.76	166.7
2	7	22.29	7.20	11.01	0.70	2.19	114.4
3	**21**	**35.63**	**9.07**	**16.55**	**1.09**	**1.8**	**142.0**
4	23	27.99	21.02	18.24	0.85	1.44	135.4

K-means clusterization

Cluster 2 – has very low energy consumption. Possible cause – wrong data. Cluster 3 is 1st choice for optimal process cluster because it has fair freezing stages energy values and low energy consumption in other stages.

TABLE VII. THE K-MEANS CLUSTERIZATION RESULTS OF THE CENTROIDS PARAMETERS OF THE CHOSEN STAGES AND PARAMETERS.

Cluster	Cases	S1	S2	S3	S4	pt	Et
0	28	28.88	19.95	16.91	0.77	1.58	141.8
1	7	39.75	16.82	19.20	1.17	1.69	192.7
2	5	35.23	19.86	18.40	1.09	1.56	143.7
3	**28**	**31.3**	**8.05**	**15.06**	**0.95**	**1.88**	**137.7**
4	7	30.3	14.14	15.14	0.88	1.92	161.7

EM clusterization

Cluster 1 is 1st choice for optimal process cluster. Cluster 3 has too small S1-S4 and Et values.

TABLE VIII. THE EM CLUSTERIZATION RESULTS OF THE MEAN PARAMETERS OF THE CHOSEN STAGES.

Cluster	Cases	S1	S2	S3	S4	pt	Et
0	10	32.5	14.8	16.27	0.93	1.77	159.7
1	**24**	**32.95**	**8.17**	**15.88**	**1.00**	**1.89**	**144.0**
2	2	48.96	4.84	25.24	1.58	1.50	214.5
3	3	9.93	4.84	5.68	0.31	2.10	65.99
4	36	30.79	19.91	17.12	0.85	1.61	144.4

TABLE IX. THE ASSESSMENT OF THE CLUSTERIZATION METHODS BY FIVE MACHINE LEARNING METHODS.

Classifier	Classification results [%]		
	Canopy	KM	EM
3NN	90,7	**94.7**	90,7
C4.5	93.3	**97.3**	98.7
MLP	96.0	**94.7**	97.3
RF	100	**100**	100
SVM	100	**98.7**	100

From Tab. 9 one can derive that the best clusterization method for the spinach production is also k-means followed by Expectation–Maximization, and then Canopy

VI. CONCLUSIONS

In the paper, we have shown three clusterization methods that allow us to assess the processes and their impact on energy consumption and hence, the carbon footprint. We have shown that all the clustering methods point out the processes that are proper from the manufacturing point of view. In the paper, the results for the onion and spinach production taking into account 152 and 75 processes respectively have been shown. Currently, we collect new processes for the other vegetable products. The will be analyzed using the clustering methods shown above

The k-means classifier is fast and simple, it has significant disadvantages because it is sensitive to emissions that distort the average value. Although, it gives with Canopy the best results in the assessment of the whole production it is planed to use k-SVD and fuzzy k-means methods in the future work.

ACKNOWLEDGMENT

The paper is co-financed by the National Center for Research and Development, grant CFOOD number BIOSTRATEG3/343817/17/NCBR/2018.

REFERENCES

[1] O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Kadner, A. Seyboth, I. Adler, S. Baum, G. Myhre, et al. “Climate Change 2014: Mitigation of Climate Change” *Working Group III Contribution to the IPCC Fifth Assessment Report*, Cambridge University Press: Cambridge, UK, 2015.

[2] *Food and Agriculture Organization of the United Nations (FAO). Regional Strategy for Sustainable Hybrid Rice Development in Asia*, Food and Agriculture Organization of the United Nations Regional Office for Asia and the Pacific: Bangkok, Thailand, 2014.

[3] D.B. Lobell, W. Schlenker, J. Costa-Roberts, “Climate trends and global crop production since 1980”, *Science 2011*, 333, 616–620. Energies 2019, 12, 138 11 of 13

[4] R.Y.M. Kangalawe, C.G. Mungongo, A.G. Mwakaje, E. Kalumanga, P.Z. Yanda, “Climate change and variability impacts on agricultural production and livelihood systems in Western Tanzania”. Clim. Dev. 2017, 9, 202–216.

[5] *ISO14040 (2006) Environmental management-life cycle assessment: principles and framework*. International Organization for Standardization, Geneva.

[6] *ISO14064-1 (2018) Greenhouse gases - Part 1: Specification with guidance at the organization level for quantification and reporting of greenhouse gas emissions and removals*. International Organization for Standardization, Geneva.

[7] *ISO/TS 14067 (2018) Greenhouse gases - Carbon footprint of products - Requirements and guidelines for quantification*. International Organization for Standardization, Geneva.

[8] *PAS 2050 (2011) “The Guide to PAS2050-2011, Specification for the assessment of the life cycle greenhouse gas emissions of goods and services*. British Standards Institution.

[9] P. Harrington, “Machine Learning in Action.” Manning Publ. 2012.

[10] A.P Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B*. 39 (1), 1977, 1–38.

[11] P. Milczarski, B. Zieliński, Z. Stawska, A. Hłobaż, P. Maślanka, P. Kosiński, "Machine Learning Application in Energy Consumption Calculation and Assessment in Food Processing Industry." *ICAISC (2)* (2020), Springer LNAI 12416, 369-379.

[12] Z. Stawska, P. Milczarski, et al., ”The carbon footprint methodology in CFOOD project.” *International Journal of Electronics and Telecommunications,* 2020, 66(4), 781–786.

[13] V. Golovko, Y. Savitsky, T. Laopoulos, A. Sachenko, L. Grandinetti, "Technique of learning rate estimation for efficient training of MLP", Proceedings of the International Joint Conference on Neural Networks 2000, pp. 323-328.