ANALYZING PRESENT AND FUTURE CONNECTIONS BETWEEN FARM MANAGEMENT AND PERFORMANCE: COMPARATIVE STUDY USING STATISTICAL METHODS AND SUPERVISED LEARNING

Victor--Dumitrel TIȚA[1], Daniel NIJLOVEANU[1], Doru Anastasiu POPESCU[2,]
Nicolae BOLD[1]

[1]University of Agronomic Sciences and Veterinary Medicine of Bucharest, Faculty of Management and Rural Development, Slatina Branch, 150, Strehareți Street, Slatina City, Olt County, Romania, E-mails: victortita@yahoo.com, nijloveanu_daniel@yahoo.com, bold_nicolae@yahoo.com
[2]University of Pitești, Department of Mathematics and Informatics, 1, Târgul din Vale Street, Pitești City, Arges County, Romania, E-mail: dopopan@yahoo.com

Corresponding author: bold_nicolae@yahoo.com

Abstract

The determination of the influences of the management on the economic performance of an agricultural holding can be an important process for the farm manager, as a source of information which can consist in a valuable component needed in the decisional process. In this matter, the analysis is useful during a specific period of time, which also comprises future predictions. In this paper, we will present a study of the connection between the farm management approach, represented by several characteristics of the farm and the farmer, and the farm economic performance, represented by the financial result of the farm activity. This study is presented as a comparative analysis of two methods that establish the existence and intensity of the mentioned connection, the first one being based on statistical methods and instruments and the second one being based on machine-learning based tools, specifically supervised learning. This study aims to find alternative means of studying causal implications of the management type on the economic activity within a farm, based on digital-based tools. The obtained results for the mentioned research showed that the methods based on supervised learning can be an important tool of analysis, being complementary with the traditional statistical methods regarding the analysis of the microeconomic agricultural environment and performance, providing supplementary key data regarding the economic indicators.

Key words: *farm, manager, supervised learning, regression*

INTRODUCTION

The development of the digital-based methods used for the analysis of certain phenomena had a great implication on the possibility of the performance improvement in a great deal of domains.

This trend is supported by the extent of the digital instruments that are used in various context, including the economic ones. In this matter, the determination of causal implications of certain indicators on the economic performance can be made and improved using methods that take into account a great deal of variables, generated as a form of Big Data from the surrounding environment.

This gain of capability in computing a great deal of variables and parameters lead to a better understanding of the phenomena and, as a direct effect, to a better decisional system within an economic environment.

In the economic field, a special attention must be given to the micro-level of a national economy, namely at the enterprise level. In agriculture, one of the most important causal relationships that affects the economic performance is made with the managerial decisions within an agricultural holding.

The managerial decisions are based on an entire set of previous experiences, known as educational background and work experience. One of the most known modalities to achieve a good educational background is the initial education, but other forms of formal education, such as training, have also a great deal of impact on the educational background of the manager of the agricultural holding.

Thus, we can imply that a direct connection between the educational training type of the farm manager and the economic performance of the farm can be established. Our approach was based on the general classification of the educational agricultural training types, namely elementary training, exclusively practical training and complete agricultural training, and the association of the economic performance of the farm with the profit indicator.

Regarding the short description of the training types, we can consider (Eurostat, 1996) [2], (NIS, 2020) [8]:

-the elementary agricultural training is considered for those who have completed any training cycle in a basic agricultural education school and/or in a training center that is oriented towards certain disciplines (horticulture, viticulture, forestry, pisciculture, veterinary science, technological agriculture and related disciplines);

-the exclusively practical training refers to the experience gained through practical work on a farm;

-the complete agricultural training refers to the completion of courses specific to agriculture, lasting at least two years after the completion of compulsory education completed in an agricultural school, college or university.

This paper presents a study related to the existence and the intensity of a relationship between the two variables presented previously.

Moreover, this paper also aims to present a comparative analysis between the traditional methods used for the determination of the mentioned relationship, using statistical methods, and an empirical perspective on the usage of information technology-based methods, namely from supervised learning (Imandoust & Bolandraftar, 2013) [4] area, that can lead to innovative methods for the analysis of the mentioned relationship.

In order to accomplish this objective of the paper, a study on this relationship has been designed using statistical instruments (the questionnaire) and implemented for a determined number of farm managers during September – December 2020. The results of the implementation of the questionnaire were then processed and analyzed using the association test and also analyzed using supervised learning methods (Khan, și alții, 2022) [6], (Shin, Hou, Park, Park, & Choi, 2013) [11], namely regression and k-Nearest Neighbor (kNN) algorithm, which measures the Euclidean distance between the points in order to establish a correlation between them.

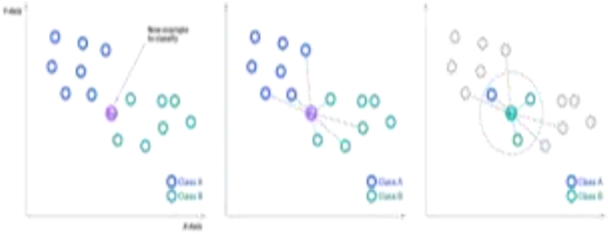


Fig. 1. Description of the kNN algorithm
Source: IBM (2023) [3].

kNN method is used in various domains and the list of domains includes the agriculture, in this matter the applications being used especially for predictions (Samhith, Rajinikanth, & Kavya, 2022) [10], crop performance (Reddy & Kumaran, 2023) [9], (Kaur, Gulati, & Kundra, 2014) [5] or farm machinery (Waleed, Um, Kamal, & Usman, 2021) [13].

MATERIALS AND METHODS \

The study is made based on the analysis and prevision of specific data obtained based on a questionnaire. The questionnaire, invented by Sir Francis Galton, is the most used method for obtaining primary data (Capotă, Popa, & Ghinescu, 2006) [1] used in social and economic studies. In the problems of the socio-economic domains, the association test is applied, after the compilation of contingency tables in which the data are classified according to one, two or more segmentation variables (Mihăiţă, 2021) [7]. The questionnaire must have an orderly format, with clear questions that are easy to complete and analyse (Şandor) [12].

For this purpose, specific methods related to statistical and ML-based techniques used for regression and clustering were selected.KNN

In this matter, the statistical part of the comparison is comprised of the usage of the association test (Chi, Hi or theoretical $\chi 2$) for the obtained data of the questionnaire. Data analysis through the association test, introduced by Karl Pearson (1857-1936) in 1900 involves the verification of the hypothesis of association between: the answers obtained in a questionnaire to the alternatives of a question and the verification of a particular set of data which may follow a known statistical distribution. This test allows highlighting the existence/non-existence of an association link between the subgroups created by the studied segmentation variables.

The field survey was conducted in the year 2020 on a sample of 49 respondents, aged between 25 and 70 years old from Olt county, with elementary, exclusively practical or complete agricultural training (Table 1).

The structure of the respondents is as follows:

-34 respondents fall into the age category of 25-44 years, of which 6 have elementary training, 5 exclusively practical training and 23 complete agricultural training;

-12 respondents belong to the age group of 45-65 years, of which 4 have elementary training, 1 exclusively practical training and 7 complete agricultural training;

-3 respondents belong to over 65 category, all 3 from complete agricultural training.

Table 1. The structure of the respondents based on age

Professional training	**MU**	**Age**			**Total**	**Percentage**
		25 – 44	**45 – 65**	**>65**	No.	%
Elementary training	no	6	4	0	10	20.40
Exclusively practical training	no	5	1	0	6	12.24
Complete agricultural training	no	23	7	3	33	67.34
Total	no	34	12	3	49	100.00
	%	69.38	24.49	6.13	49	100.00

Source: Field survey, 2020.

The test statistic χ^2 equals 3.0272, which is in the 95% region of acceptance: [-∞ : 9.4877]. The requested test was calculated, however, this may not be the right of test for the hypothesis. The priori power is low (0.3514).

Table 2. χ^2 test parameters

Indicator	**Value**	**Explanation**
k	3 x 3 = 9	Number of categories
n	49	Sample size
χ^2	3.0271984551396316	Chi square test statistic
DF	4	(Rows-1)*(Columns-1) = (3-1)*(4-1) = 6
Phi effect (Φ)	0.248555	$\Phi=\sqrt{(\chi 2/n)}$
DFmin	2	Min(Rows-1,Columns-1) = Min(3-1,4-1) = 2
Cramer's V effect	0.175755	$V = \Phi/\sqrt{DFmin}$

Source: Own results based on the data from Field Survey, 2020.

The farmers' distribution by age is presented in Fig. 2.

Also, the experience of the farmers is an extremely important indicator of their work and performance and will be taken into consideration as a parameter in the analysis of the correlation.

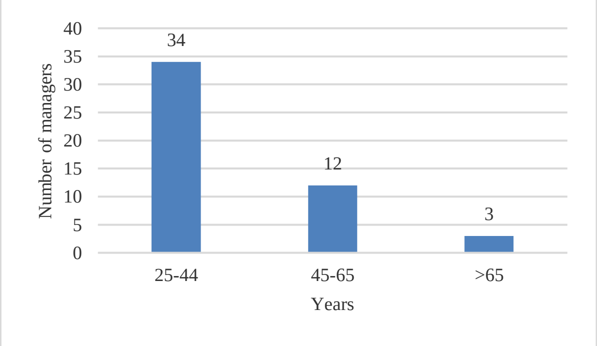


Fig. 2. Farmers' distribution by age (%)

Source: Own results based on the data from Field Survey, 2020.

Figure 3 presents the distribution of the farmers based on their experience in terms of years of work as farm manager.

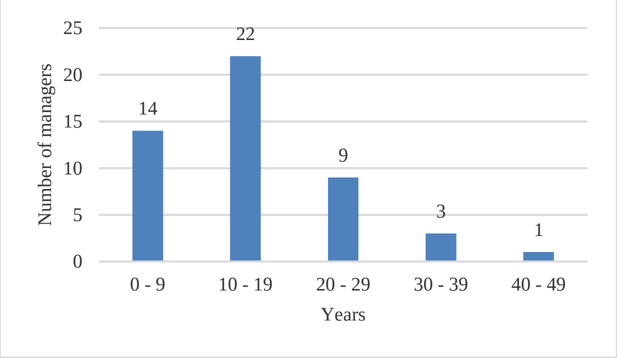


Fig. 3. Farmers' distribution based on their years of experience as farm manager
Source: Own results based on the data from Field Survey, 2020.

The studied farms, numbering 49, belong to three categories of legal forms: Individual Enterprise (IE), Limited Liability Company (LLC), Commercial Company. One of them is a Commercial Company, which is managed by people with elementary training. In the case of LLC type farms, there are 21 companies, from which 2 managers have basic training, 3 exclusively practical and 16 complete agricultural training. In case IE of the 26 farms, 6 administrators have basic training, 3 have exclusively practical and the rest of 17 complete agricultural training.

Figure 4 presents the farmers' distribution according to the profit level category obtained by their company.

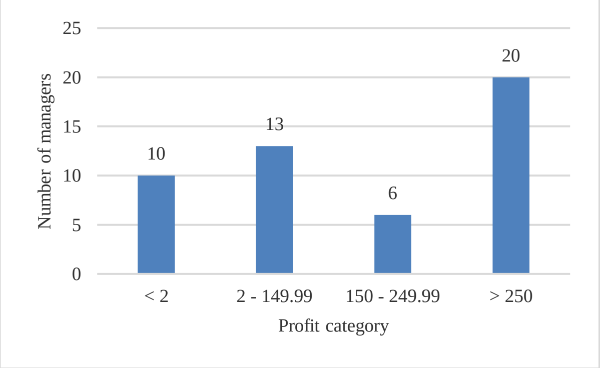


Fig. 4. Farmers' distribution based on the profit category of the company (%)
Source: Own results based on the data from Field Survey, 2020.

The study also consisted of a series of questions regarding respondents' participation in various vocational training courses, hypothesizing that agricultural experience does not influence decisions to attend these courses.

In order to study the correlation between the parameters, we will use the test of hypotheses for the statistical method, based on the difference between the p-value of the association test and the significance level α. In this matter, two hypotheses will be taken into consideration:

- H_0: p-value > α;
- H_1: p-value ≤ α.

The null hypothesis (H_0) will be confirmed if the value of the significance level will be lower than the p-value of the association test.

Related to the kNN algorithm, the steps taken into account are:

S1.The data was is split into train data and test data using specific methods. The training data is used to fit the model, while the test data is used to evaluate the performance of the model.

S2.The Euclidean distances between the points are calculated and normalised. The normalisation consists in the creation of a fair comparison between the calculated distances of the values of the parameters.

S3.A number of neighbours (k) is optimally chosen or determined. defines how many neighbours will be checked to determine the classification of a specific query point.

S4.An instance of the kNN model is created and the train data is fitted.

S5.The model is tested using test data by obtaining the accuracy parameter, which is basically the proportion of the correct answers given by the model compared to the test data.

RESULTS AND DISCUSSIONS

The profit obtained in most farms, namely 20, exceeds 250,000 euros, in 6 farms it is between 150,000 and 250,000 euros, in 13 farms it is between 2,000-149,999 euros, and less than 2,000 euros were obtained by 10 farms.

The profit in the studied farms is not influenced by the vocational training of the respondents, as indicated by the calculation of the Chi-square test in Table 3.

Table 3. The relationship between training type and farm profit

Professional training	MU	Farm profit (thousand euro)				Total	
		<2	2 – 149.99	150 – 250	>250	no	%
Elementary training	no	2	3	1	4	10	20.40
Exclusively practical training	no	0	1	0	5	6	12.24
Complete agricultural training	no	8	9	5	11	33	67.36
Total	no	10	13	6	20	49	100.00
	%	20.40	26.53	12.24	40.83	100.00	X

Source: Own results based on the data from Field Survey, 2020.

Expressed as a percentage, the profit obtained in the farms represents more than 30% for 10 farms, between 10-30% for 20 farms, between 2-10% for 8 farms and less than 2% obtained by 11 of the farms (Table 3).

It is found from the type of elementary vocational training, that 6 respondents have profit from 10% to 30% and 4 respondents have profit >30%.

Table 4. χ^2 test parameters

Indicator	Value	Explanation
k	4 x 3 = 12	Number of categories
n	49	Sample size
χ^2	5.850493395493395	Chi square test statistic
DF	6	(Rows-1)*(Columns-1) = (3-1)*(4-1) = 6
Phi effect (Φ)	0.34554	$\Phi=\sqrt{(\chi2/n)}$
DFmin	2	Min(Rows-1,Columns-1) = Min(3-1,4-1) = 2
Cramer's V effect	0.244334	$V = \Phi/\sqrt{}$DFmin

Source: Own results based on the data from Field Survey, 2020.

Since p-value > α, H_0 is accepted. The statistical model fits the observations, but there is not enough evidence to suggest an association between the professional training type and the profit category. The p-value equals 0.4401, ($p(x\leq\chi^2) = 0.5599$). It means that the chance of type I error, rejecting a correct H_0, is too high: 0.4401 (44.01%). The larger the p-value the more it supports H_0. Also, The test statistic χ^2 equals 5.8505, which is in the 95% region of acceptance: [-∞ : 12.5916]. The observed effect size phi is medium, 0.35. Cramer's V effect size is 0.24. This indicates that the magnitude of the difference between the observed data and the expected data is medium.

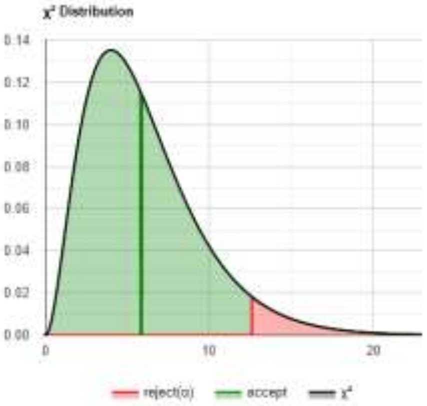
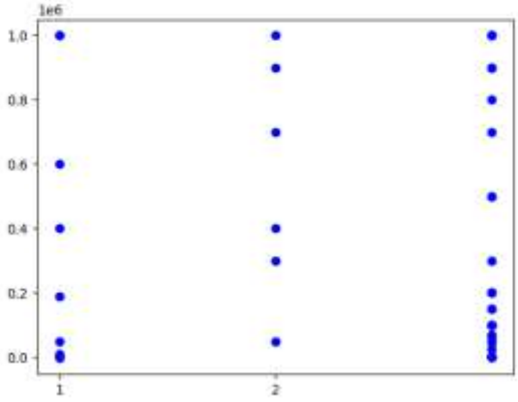


Fig. 5. χ^2 distribution
Source: Own results from Field Survey, 2020.

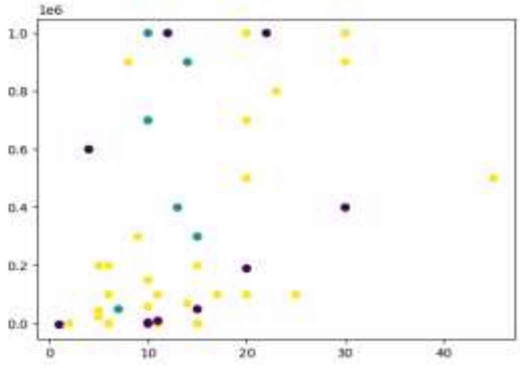
Regarding the ML-based model implementation, a regression model was built and a model based on kNN algorithm was made. The kNN model was built also for the specific nature of the values, which group data into categories, which are conceptualized for the kNN algorithm and specifically used for these types of data processing models.

For the regression model, the results were similar to the ones related to the statistical approach. Thus, a Pearson coefficient between the training category and the profit category equal to 0.01071065174315855 was obtained and a Pearson coefficient between the training category and profit values equal to 0.003941743741843418 was obtained, which shows a low connection between the two pairs of parameters.

For the kNN-based model, there was also taken into account the years of experience as a manager, due to the nature of the used tool, which uses raw data as train and test data and the categorisation is made using a different approach. In this matter, the data was set and the distributions presented in Figure 6a and 6b were resulted.



(a)



(b)

Fig. 6. (a) Farmers distribution based on the profit type; (b) Farm distribution based on profit and years of experience.

Source: Own design based on the obtained results.

Figure 6 (a) shows a specific distribution of the profit of the farmers based on the professional training category. 1, 2 and 3 on the horizontal axis are codifications for the three professional training categories, elementary training, exclusively practical training and complete agricultural training, respectively.

While for the category 2 the profit is not concentrated, for the other two categories the profit has a tendency of concentration for a level below 200 thousand euros.

In Figure 6 (b), the blue markers represent the farmers with exclusively practical training, the yellow ones the farmers with complete agricultural training and the purple markers represent the farmers with elementary training. As we can observe, the distribution based on profit related to the experience tends to be lower for the farmers with complete agricultural training and fewer years of experience.

Following the methodology presented in the previous section, the values presented in Table 5 were obtained, for the given values. The first instance of the implementation took into account the determination of the training category based on the profit and the experience.

Table 5. kNN implementation results for the first instance

Farmer ID	Initial training category	Training category (k=10)	Training category (best k)
37	3	3	3
1	1	3	3
41	3	2	1
13	3	3	3
42	3	3	3
14	3	3	3
18	2	3	2
8	3	3	3
25	3	3	3
40	1	3	3
Accuracy	1.0	0.6	0.7

Source: Own results.

For the given results in Table 5, the values of the accuracy indicator are presented in Figure 7.

We can observe that the best accuracy of 0.7 is obtained for a number of 9 neighbours.

The accuracy values for different values of k (first implementation) are shown in Figure 7.

The second instance took into account the determination of the profit based on the category. The values are presented in Table 6.

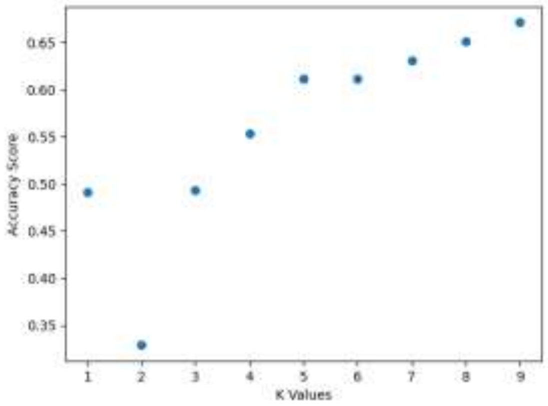


Fig. 7. Accuracy values for different values of k (first implementation)
Source: Own design based on the obtained results.

Table 6. kNN implementation results for the second instance

Farmer ID	Initial profit	Profit (k=10)	Profit (best k)
0	45	0.1	0.1
21	400	50.0	1,000.0
26	25	0.1	0.1
38	0	50.0	4.0
33	200	0.0	0.0
47	500	1,000.0	1,000.0
14	150	0.1	0.0
25	0	0.1	0.1
5	200	0.1	0.1
22	70	0.1	0.0
Accuracy	1.0	0.0	0.12

Source: Own results.

We can observe that the accuracy is quite low, which indicated the usage in the model of a categorisation of profit values instead of the profit value itself for all the instances that may be taken into account. In this matter, a categorisation of the profit was made and the results obtained for the third instance are presented in Table 7.

Also, the accuracy levels for the third instance of the implementation are presented in Figure 8.

The accuracy has a better value for a lower number of neighbours (k = 1). As we can observe, the model has given better results when the profit was categorised, in this way validating the nature of the kNN instrument and completing the comparison between the two methods for determining the relationship between the given parameters, which was not a consistent one, especially regarding the statistical approach.

Table 7. kNN implementation results for the third instance

Farmer ID	Initial profit category	Profit category (k=10)	Profit category (best k)
8	1	1	1
45	4	4	4
26	3	2	3
43	4	4	4
11	2	1	2
34	4	4	4
35	4	4	4
32	4	2	4
28	3	1	2
18	2	1	2
Accuracy	1.0	0.5	0.9

Source: Own results.

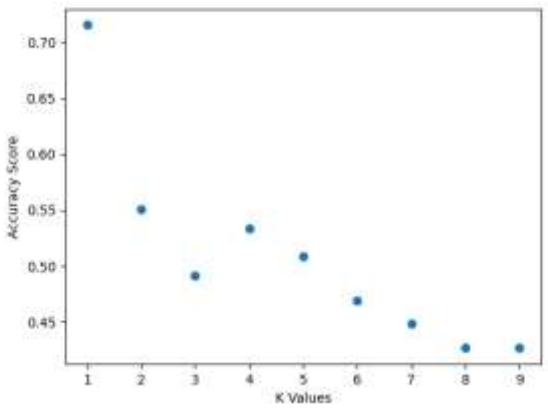


Fig. 8. Accuracy values for different values of k (third implementation)
Source: Own design based on the obtained results.

CONCLUSIONS

The determination of a relationship between farm indicators and, in our case, between the level of farmers' training and the performance of the farm has great levels of extent, due to the complex nature of the parameters.

In this matter, the statistical method is a well-founded one in the practice and the literature, giving specific accurate results for a given set of data.

On the other hand, the determination of this relationship brings a new type of approach and can lead to significant results of the analysis from new perspectives.

For example, this paper extended the analysis capability of the ML-based method, due to the fact that the presented ML model has given the possibility of training of the data and, therefore, the validation of new data input by the user.

Another conclusion would consist in the necessity of larger sets of data which can refine both the statistical and ML-based methods results.

As future work, the refinement would be made on the determination of a larger set of data.

Also, the comparison will be improved by modifying the ML model by introducing several new aspects, such as the categorisation of specific parameters.

REFERENCES

[1]Capotă, V., Popa, F., Ghinescu, C., 2006, Business Marketing (Marketingul afacerii). AkademosArt, Publishing House, București, pp. 46.

[2]Eurostat, 1996, Structures des exploitations: Méthodologie des enquêtes communautaire. https://ec.europa.eu/eurostat/documents/3859598/5826685/CA-98-96-493-FR.PDF.pdf/8b0bc3f8-3e6d-4f6e-b28d-fc120b702728?t=1414780269000, Accessed on the 15th of March 2023.

[3]IBM, 2023, IBM. https://www.ibm.com/topics/knn, Accessed on the 15th of March 2023

[4]Imandoust, S. B., Bolandraftar, M., 2013, . Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications, 3(5), 605-610.

[5]Kaur, M., Gulati, H., Kundra, H., 2014, Data mining in Agriculture on crop price prediction: Techniques and Applications. International Journal of Computer Applications, 99(12), 1-3.

[6]Khan, M. A., Abbas, K., Su’ud, M. M., Salameh, A. A., Alam, M. M., Aman, N., Aziz, R. C., 2022, Application of Machine Learning Algorithms for Sustainable Business Management Based on Macro-Economic Data: Supervised Learning Techniques Approach. Sustainability, 14(16), 9964.

[7]Mihăiţă, N. V., 2021, Strong, hidden, false and illusory statistical relationships (Relaţiile statistice puternice, ascunse, false și iluzorii). http://www.biblioteca-digitala.ase.ro/biblioteca/carte2.asp?id=388&idb=, Accessed on the 15th of March 2023.

[8]National Institute of Statistics, NIS, 2020, RGA Questionnaire RGA CAPI 2020 (Chestionar RGA CAPI RGA 2020) https://insse.ro/cms/files/RGA2020/aprilie2021/Chestionar-ghid-CAPI-RGA2020.pdf, Accessed on the 15th of March 2023.

[9]Reddy, B. B., Kumaran, J. C., 2023, A Comparison of the K Nearest Neighbor Algorithm and Naive Bayes for Predicting Agricultural Crop Yield. Journal of Survey in Fisheries Sciences, 10(1S), 1876-1883.

[10]Samhith, S. S., Rajinikanth, T. V., Kavya, B., Krishna, A. Y. S., 2022, Crop recommender system. International Journal of Engineering Applied Sciences and Technology, Vol. 7(10), 117-123. http://www.ijeast.com/Past-issue.php?title=Volume%207%20Issue%2010, position 15, Accessed on the 15th of March 2023.

[11]Shin, H., Hou, T., Park, K., Park, C. K., Choi, S., 2013, Prediction of movement direction in crude oil prices based on semi-supervised learning. Decision Support Systems, 55(1), 348-358.

[12]Şandor, S. D. (n.d.). Reserach methods and techniques in social sciences, A course support (Metode şi tehnici de cercetare în ştiinţele sociale, suport de curs.) Cluj – Napoca:Babeş – Bolyai University.

[13]Waleed, M., Um, T. W., Kamal, T., Usman, S. M., 2021, Classification of agriculture farm machinery using machine learning and internet of things. Symmetry, 13(3), 403.