European Research Studies Journal
Volume XXVII, Issue 4, 2024

pp. 248-272

The Application of Selected Supervised Machine Learning Methods in the Classification of Family Businesses in the Context of Cluster Formation

Submitted 13/09/24, 1st revision 23/09/24, 2nd revision 14/10/24, accepted 30/10/24

Daria Wotzka[1], Paweł Frącz[2], Jolanta Staszewska[3], Joachim Foltys[4], Małgorzata Smolarek[5], Krzysztof Orzechowski[6]

Abstract:

Purpose: *The article focuses on the application of selected supervised machine learning methods for the classification of family businesses in the context of cluster formation. The research aim was to evaluate various learning algorithms to develop a tool for classifying entrepreneurs, intended for use in an online application.*

Design/Methodology/Approach: *Through a comprehensive survey, 448 responses were gathered, addressing various aspects of clusters and related experiences. Based on the collected data, classification methods for respondents were developed in the context of cluster formation. The classifier categorizes entrepreneurs based on their under-standing of cluster concepts, managers' perceptions of clusters, companies' experiences with clusters, the operational status of clusters, and experience in business networks. The article conducts a comparative analysis of the classification outcomes derived from the application of decision trees and neural networks across diverse configurations. This analysis, based on distinct evaluation metrics, culminates in the identification of the most optimal algorithm suited for the task at hand.*

Findings: *As a result of the conducted research, a supervised machine learning algorithm in the form of an ensemble bagged tree was selected. This algorithm achieves an average effectiveness of 82%, measured as the arithmetic mean of accuracy, specificity, precision, sensitivity, F1 score, and the Matthews correlation coefficient. The median value was 96%.*

Practical Implications: *The presented results have been implemented in the form of a computer application that allows for the simulation and classification of entrepreneurs based*

[1] *Opole University of Technology, Faculty of Electrical Engineering Automatic Control and Informatics, ORCID 0000-0002-8861-7974, email: d.wotzka@po.edu.pl;*

[2] *Opole University, Faculty of Economics, ORCID 0000-0003-1677-6084, email: pawel.fracz@uni.opole.pl;*

[3] *Humanitas University, Institute of Management and Quality Sciences, ORCID 0000-0001-914-212, email: jolanta.staszewska@humanitas.edu.pl;*

[4] *Humanitas University, Institute of Management and Quality Sciences, ORCID 0000-0003-4836-3161, email: joachim.foltys@humanitas.edu.pl;*

[5] *Humanitas University, Institute of Management and Quality Sciences, ORCID 0000-0002-3766-8843, email: malgorzata.smolarek@humanitas.edu.pl;*

[6] *Humanitas University, Institute of Management and Quality Sciences, email: krzysztof.orzechowski@humanitas.edu.pl;*

on their business experiences. The developed tool is being deployed as a web-based application, serving as a platform to showcase the numerous possibilities and benefits of cluster formation.
Originality/Value: *This study represents a novel approach, as there are no available articles specifically applying machine learning techniques to classify entrepreneurs, particularly family-owned businesses, in the context of cluster formation.*

Keywords: *Family businesses classification; machine learning, cluster formation.*

JEL Classification: *C45, C52, D22.*

Paper Type: *Research article.*

Funding: *Publication was funded by the state budget under the program of the Polish Ministry of Education and Science called "Science for Society" project number NdS/545753/2022/2022, amount of funding 678 000,00 PLN total value of the project 754 000,00 PLN (Poland)*

1. Introduction

This study represents a novel approach, as there are no available articles specifically applying machine learning techniques to the classification of family businesses or entrepreneurs in the context of cluster formation. The existing literature primarily focuses on traditional statistical methods or qualitative analyses when examining cluster formation and entrepreneurial activities.

By leveraging advanced machine learning algorithms such as decision trees and neural networks, this research bridges a significant gap in the current body of knowledge. It introduces a data-driven methodology that enhances the accuracy and efficiency of classifying entrepreneurs based on their experiences and involvement in clusters.

This pioneering effort not only contributes to the theoretical framework but also offers practical implications for policymakers and business practitioners aiming to foster cluster development. The innovative application of these techniques provides deeper insights and a more nuanced understanding of the dynamics at play, thereby setting a new standard for future research in this domain.

The objective of this study was to utilize survey results for the classification of family-owned businesses in the context of their experiences with cluster formation. Given the pioneering nature of this research, the following hypothesis was formulated: The application of machine learning methods enables effective classification of family-owned businesses in the context of clusters, identifying key characteristics and patterns specific to these enterprises. Machine learning methods are sufficiently advanced to effectively process and analyze data on family-owned businesses, allowing for their precise classification (Tyagi *et al.,* 2023).

This approach enhances the understanding of the diversity among family-owned businesses and their approach to cluster formation. Furthermore, machine learning algorithms have the capability to identify patterns and characteristics unique to family businesses, which are challenging to discern using traditional methods. These patterns may include specific management strategies, organizational structures, and models of collaboration within clusters.

Effective classification can serve as a foundation for further research and analysis aimed at developing more efficient support and development strategies for family businesses. The findings can also be valuable for policymakers and business practitioners in formulating policies and programs that support the creation and growth of clusters.

This article consists of several sections, as follows. In the Research and Methodology section, we present the survey structure and respondent characteristics, the stated research hypotheses, and the assumptions for the classification process. In the Results and Discussion section, we present the comparative results regarding the effectiveness of the classification algorithms considered in the study and select the best classification method. Finally, we conclude and propose further research possibilities.

2. Literature Review

Family businesses constitute a significant part of the economy in many countries. They are one of the pillars of these economies, contributing to job creation, income generation, and economic stability (Birdthistle and Hales, 2023). Their importance lies, among other things, in their long-term approach to business. Unlike companies focused on quick profits, family businesses often concentrate on sustainable and stable development, taking care of relationships with employees, customers, and local communities (Norena-Chavez and Thalassinos, 2023; Liapis *et al.,* 2013).

This long-term perspective makes them more inclined to reinvest profits into the development of the company, which fosters innovation and sustainable growth. Economic clusters can bring many benefits to family businesses, such as increased competitiveness, better access to resources and technology, and greater opportunities for collaboration. Family businesses often have strong ties with local communities. Supporting their development through clusters can contribute to the growth of local economies and job creation (Fitzgerald and Muske, 2016).

Creating clusters within business operations constitutes a significant factor supporting sustainable development of both businesses and local communities.

Clusters, being concentrations of enterprises operating within a specific region, bring forth a multitude of benefits that translate into economic, environmental, and social aspects. Analyzing cluster aspects in the context of sustainable business activities allows us to recognize their significant contribution to various areas. Resource efficiency stands as a pivotal aspect of clusters. Through collaboration and sharing of both resources and infrastructure, companies within clusters can achieve higher resource utilization efficiency.

This, in turn, enables the reduction of energy, water, and other resource consumption, thereby promoting sustainable resource management. Clusters also contribute to the development of local communities. By bringing together companies operating within a defined region, they support the local economy by increasing employment and creating new job opportunities (Zoltan J. Acs and Laszlo Szerb, 2006).

This local collaborative dynamic translates into social and economic growth. In the context of sustainable innovations, clusters create conducive conditions for collaboration in developing innovative solutions. These actions encompass the development of renewable technologies, recycling strategies, or emission reduction initiatives, which are crucial for achieving sustainable development. Clusters also bolster the competitiveness of firms (Mills, Reynolds, and Reamer, 2008).

Through knowledge exchange, experience sharing, and collaborative project work, companies within clusters can enhance their operational efficiency, leading to cost reduction and increased sustainable profitability. Corporate Social Responsibility also finds its place in cluster activities. Establishing clusters enables companies to actively engage in initiatives that positively impact the local community and environment, constituting a significant element of sustainable business practices. Clusters also support product lifecycle management (Corallo, Del Vecchio, Lezzi, and Luperto, 2022).

Starting from the design stage, through production, distribution, and disposal, companies within clusters strive to minimize their negative environmental impact, aligning with the pursuit of sustainable development. Community development constitutes another significant aspect of clusters. By promoting collaboration among companies, educational institutions, non-governmental organizations, and other entities, clusters support the resolution of local issues and aim for achieving sustainable social and economic development (Chen, Wang, Miao, Ji, and Pan, 2020; Derlukiewicz *et al.,* 2020). Creating regional innovation clusters can drive economic growth by fostering innovation, increasing productivity, and improving regional economic performance through dense knowledge flows and entrepreneurship support (Kerr and Robert-Nicoud, 2019; Muro and Katz, 2010).

In today's dynamic and competitive business environment, companies need to be flexible and able to quickly adapt to changes. Competency reconfiguration means the process by which a company transforms, modifies, or develops its skills, resources, and capabilities to better meet new challenges and opportunities in the market. For family businesses, which often have deeply rooted structures and traditions, the ability to effectively adapt after such changes is crucial for their survival and development (Smith, 2023; Velinov *et al.*, 2023; Noja *et al.*, 2021).

This means they must be able to: Identify and develop new skills by being open to learning and implementing new technologies, management methods, or innovative products and services; Be ready, if necessary, to make changes in their organizational structures to better support new competencies and strategies; Ensure that change management processes are effective to minimize resistance and increase employee engagement in new directions of development, as the ability to quickly respond to new market opportunities is key for the company to maintain its competitiveness.

Due to the fact that family businesses often have long-term perspectives and goals, their analysis and classification can help develop strategies that will support their long-term development and success across generations (Yilmaz, Raetze, Groote, and Kammerlander, 2024). It can also lead to more effective support and development strategies, contributing to their stability and growth. Through detailed data analysis and classification of family businesses, decision-makers and managers can make more informed and strategic decisions, which will contribute to the long-term success and sustainability of these enterprises.

Classifying family businesses using machine learning methods may allow for the identification of key characteristics and needs of these firms. This would enable better tailoring of assistance programs and support policies, leading to the optimization of public and private resources.

Furthermore, precise classification of family businesses could help create more efficient and cohesive clusters. It is also important to note that the application of machine learning methods in the classification of family businesses promotes innovation and the adaptation of modern technologies in business management. This can also contribute to increasing the competitiveness of family businesses in the market.

Designing and developing neural network models is becoming increasingly crucial in simulating various business scenarios (IDEAMOTIVE, 2024). A central aspect of this process is the creation of advanced neural network models capable of replicating and simulating complex interactions within the firm's structure. It is worth emphasizing that the unique skills of individual family members, their roles, and the dynamics of their collaboration must be considered in the design of these models.

The development of machine learning algorithms plays a crucial role in creating models capable of analyzing and mimicking complex patterns of behavior within the simulated competency structure (Dong, Hou, Zhang, and Zhang, 2020). This requires continuous adjustment and improvement of algorithms to better reflect changing conditions and the specificity of the family firm. Validation and testing of neural network models are essential to ensure their effectiveness and accuracy in simulating anticipated scenarios.

This process involves various tests that allow for the assessment of the model's reliability and the identification of areas for further optimization. Ultimately, model optimization is a key stage in the design process, aiming to ensure the best possible performance and accuracy of the simulation. Adjusting model parameters and tailoring it to the specific needs of the family firm are crucial for achieving desired results. As a result, the process of designing and developing advanced neural network models for simulating the operation of a family firm after competency reconfiguration requires a comprehensive approach that takes into account both the unique characteristics of the organization and the specifics of its members (Legaard *et al.*, 2022).

In the context of designing advanced simulation models for a family firm following competency reconfiguration, it is imperative to consider various approaches and techniques. In addition to neural networks, the effectiveness of decision trees and ensemble methods of trees has been investigated to evaluate their utility in simulating complex interactions within the firm's structure. Decision trees, due to their simplicity of interpretation and relatively low computational cost, often present an attractive alternative to neural networks.

Their ability to represent complex decision structures and ease of interpreting results make them a popular choice in business analysis. Ensemble methods of trees, such as random forests or gradient boosting, take it a step further by integrating multiple decision trees to improve model accuracy and stability (Costa and Pedreira, 2022). Through mechanisms such as bagging or boosting, these methods can address overfitting issues and enhance overall predictive performance.

Comparative studies of the effectiveness of these different approaches have observed that each has its own advantages and limitations. Neural networks often achieve the highest prediction accuracy, especially with large and complex datasets, but their interpretability may be limited. On the other hand, decision trees are easier to understand but may be less precise with highly complex data. Ensemble methods of trees serve as a compromise between these two approaches, offering a good balance between interpretability and predictive performance (Halawi, Clarke, and George, 2022).

3. Research Methodology

This work is based on the findings of a comprehensive survey, that was conducted in the form of questionnaires, examining the landscape of family-owned businesses within the Polish market, particularly in relation to their engagement with cluster formation (Staszewska, Smolarek, Foltys, Wotzka, and Fracz, 2024). The study resulted in the collection of 448 responses that delved into diverse facets of clusters and the experiences associated with them.

3.1 The Survey Structure and Respondent Characteristics

In the survey, companies meeting specific criteria were considered for inclusion. These criteria encompassed firms in which at least two shareholders have familial ties and collectively hold a minimum of 50% plus 1 share, while also maintaining uninterrupted operation for a minimum of 8 years. Additionally, the selected companies were required to meet the following conditions: demonstrate a net revenue of no less than 10 million PLN in 2020; exhibit consistently positive financial results in recent years; not be undergoing bankruptcy proceedings or liquidation; maintain good financial standing; and possess an unblemished reputation.

Within the scope of family-owned companies listed on the Warsaw Stock Exchange (GPW), qualifying entities included those in which the individual who founded or acquired the company, along with their relatives and descendants, collectively held at least 25% of the voting rights at the General Meeting of Shareholders of said company.

The surveyed companies were characterized by various features. The majority of them were headquartered mainly in the Silesian (32.4%), Lower Silesian (29.5%), Lesser Poland (24.8%), and Opole (13.4%) voivodeships. The main sectors of activity for the companies included trade (38.4%), services (50.0%), and production (28.1%).

The legal forms of the companies were also diverse, with sole proprietorships being the most common (50.4%), followed by individual entrepreneurs (15.2%), civil partnerships (16.3%), capital companies (11.8%), and partnerships (7.4%). Most of the companies were microenterprises (79.7%), followed by small enterprises (13.4%), medium-sized enterprises (5.4%), and large enterprises (1.3%). Regarding the market of operation, the majority of companies operated mainly in the local market (57.6%), as well as in regional (14.3%) and national (13.4%) markets. The presence of companies in foreign (9.2%) and global (5.1%) markets was smaller.

The survey covered a group of respondents of various genders, with a clear majority being men (65.0%), while women constituted 35.0% of the surveyed population. In terms of education, the dominant group consisted of respondents with master's degrees (41.1%) and those with secondary education (39.3%). The remaining categories, namely bachelor's degrees, vocational education, and primary education, accounted for a smaller percentage in the surveyed group. Regarding the roles performed in the company, the highest number of respondents were owners (46.9%) and shareholders (19.9%).

Meanwhile, top-level executives accounted for only 2.7% of the respondents. As for length of employment, the majority of respondents (61.6%) had been working in their company for over 10 years, while those working from 2 to 5 years and from 7 to 10 years accounted for 11.6% and 13.8% of the surveyed population respectively. "Hard to say" responses were noted in only 0.2% of cases. Among the surveyed companies, the majority were those where the first generation of the founder currently manages the business (50%).

Firms managed by the second generation (counting from the founder) accounted for 46%. Currently, in the surveyed companies (counting from the founder), the most common are the second-generation members (elders and successors), constituting 60%. Only elders currently manage 36% of the surveyed firms. Relatively few (<4%) of the companies surveyed were those where three or more generations are currently employed.

3.2 Research Hypothesis and Assumptions for Classification Process

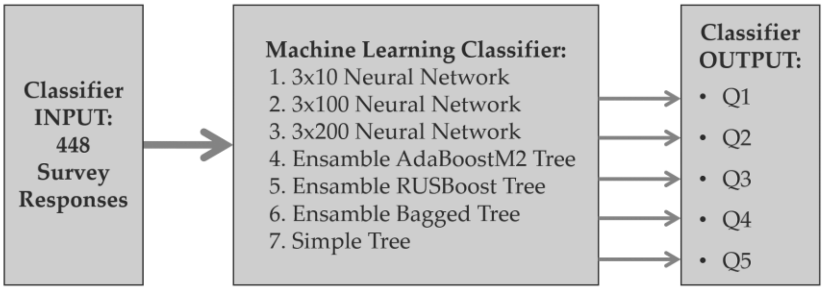
In the research it is hypothesized that an entrepreneur can be accurately classified based on their responses to selected questions about their business activities using machine learning methods.

H1: *It is possible to classify an entrepreneur based on their responses to selected questions about their business activities using machine learning methods.*

In order to validate the stated research hypothesis, a classification model was developed, which structure is presented in Figure 1. The input consists of responses to the survey questions, while the output, as mentioned earlier, comprises questions/answers Q1 to Q5. Various types of algorithms were evaluated as classifiers, including three-layer neural networks with 10, 100, and 200 neurons in the hidden layer, a simple decision tree, and ensembles of trees utilizing three distinct aggregation methods: bagging, RusBoost, and AdaBoost.

Survey questions were presented to entrepreneurs, with all responses subsequently digitized. This process yielded a matrix consisting of 436 columns, each representing an individual response, and 448 rows, corresponding to individual respondents. Among the myriad features, 5 questions specifically pertaining to cluster formation were selected. The labels, descriptions, and response types for these questions are detailed in Table 1. These questions were designated as the output variables for the classifier model in the analysis.

Figure 1. *The structure of the classification model*



***Source**: Authors' own elaboration.*

Table 1. *Designations and description of the question/answer considered as classifier output*

Question label	Description
Q1	Full: "Is the concept of a cluster as a form of organizing business activities known to Managers?" Short: “Understanding cluster concepts” Responses: 1 = YES, 2 = NO, 3 = I DON'T KNOW
Q2	Full: "How do Managers perceive clusters?" Short: “Managers' cluster perception” Responses: 1 = POSITIVE, 2 = NEGATIVE, 3 = NEUTRAL, 0 - REFUSAL TO ANSWER
Q3	Full: "Does your company have any experience with clusters?" Short: “Company's cluster experience” Responses: 1 = YES, 2 = NO, 3 = I DON'T KNOW, 0 - REFUSAL TO ANSWER
Q4	Full: "Is your company already operating within a cluster?" Short: “Cluster operation status” Responses: 1 = YES, 2 = NO, 3 = I DON'T KNOW, 0 - REFUSAL TO ANSWER
Q5	Full: "Does your company have experience participating in business networks?" Short: “Experience in business networks” Responses: 1 = YES, 2 = NO, 3 = I DON'T KNOW, 0 - REFUSAL TO ANSWER

***Source:** Authors' own elaboration.*

Firstly, the feasibility of employing a fully connected feedforward neural network as a classifier for entrepreneur clustering was investigated (Bishop, 1996). The term "feedforward" pertains to the flow of data through the network in a unidirectional manner, from the input layer through the hidden layers to the output layer, without any feedback loops.

The fundamental unit in such networks is the neuron, which receives inputs, processes them using activation functions, and generates outputs. Neurons are organized into layers, and a network may have one or more hidden layers between the input and output layers. The input layer receives input data, corresponding to the predictors (survey results), while the output layer produces final results (predicted class labels Q1-Q5) that are interpreted by the user.

During the training of the feedforward network, a limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) (Najafabadi, Khoshgoftaar, Villanustre, and Holt, 2017; Reza Godaz *et al.*, 2021) was utilized as the technique for minimizing its loss function, wherein the software minimizes the cross-entropy loss. The LBFGS solver employs a standard line-search method with an approximation to the Hessian matrix. The Rectified Linear Unit (ReLU) (Abien Fred M. Agarap, 2019; Xu, Wang, Chen, and Li, 2015) activation function was employed in the hidden layers, which performs a threshold operation on each element of the input, setting any value less than zero to zero. The activation function for the final fully connected layer is the softmax function.

In the course of research into selecting a classification model, consideration was given to a single decision tree classifier (Simple Tree) (Shalev-Shwartz and Ben-David, 2014). This model represents the structure of a tree graphically, where each node represents a test on one of the features, and each branch emanating from a node represents a possible outcome of that test. The tree's leaves correspond to target classes or classification decisions.

During the training of a decision tree classifier, the algorithm recursively splits the data based on features to best separate observations of different classes at each node. The Gini's diversity index (Ultsch and Lötsch, 2017) was employed as the splitting criterion, serving to measure the diversity of classes at a given node. The primary objective of this criterion is to minimize impurity, leading to improved classification quality.

Decision trees are relatively robust to the presence of irrelevant features in the training data and can handle data with nonlinear structures. As part of the procedure for selecting a decision tree-based classification model, a maximum number of splits in the tree was set at 403. It is assumed that the larger this value, the more complex the tree can be adapted to the training data. Additionally, it was assumed that all variables are considered at each split, which may contribute to obtaining more stable decision trees.

In the subsequent phase of the research, the constructed decision tree was integrated into an ensemble of classification models, which were trained over 30 cycles (the number of base trees in the ensemble), utilizing three different merging methods: AdaBoost Multi-class (Ens. AdaBoost.M2 Tree), Random Under-Sampling Boosting (Ens. RUSBoost Tree), and Bagging (Ens. Bootstrap Aggregating).

AdaBoost.M2 Tree (Hastie, Rosset, Zhu, and Zou, 2009) represents a generalized version of the popular AdaBoost method, applied in multiclass classification tasks. It is based on the concept of sequentially training weak classifiers and assigning greater weight to misclassifications to focus on challenging examples. It extends the AdaBoost algorithm to handle multiple classes by employing "one against all" or "one against one" strategies. In each iteration of the AdaBoost.M2 algorithm, a new classifier is added, attempting to improve the classification of difficult cases. Then, the weights of examples are updated to focus on in-stances that were misclassified by previous classifiers. In this manner, AdaBoost.M2 constructs a strong classifier by combining multiple weak classifiers.

RUSBoost Tree (Dwiyanti, Ardiyanti, and Ardiyanti, 2017; Seiffert, Khoshgoftaar, Van Hulse, and Napolitano, 2010) is a machine learning ensemble method that combines boosting techniques with methods for reducing sample imbalances within classes. Un-like traditional boosting, where all training samples are used in each iteration, RUSBoost employs random undersampling instead.

This entails randomly selecting a sample from the minority class (the class with fewer examples) in each iteration and using it along with selected examples from the majority class to train a weak classifier. Thus, RUSBoost focuses on challenging cases while simultaneously reducing the dominance of the majority class through random sample selection. This is particularly useful for imbalanced data sets, where one class significantly outnumbers the other.

Bagging (Jason Brownlee, 2020; Tanha, Abdi, Samadi, Razzaghi, and Asadpour, 2020), on the other hand, is a machine learning ensemble technique that involves training multiple models on different subsets of the training data and then combining the results of these models to obtain the final classification.

Each model in bagging is trained independently on a random subset of the training data generated using bootstrap, which involves randomly selecting samples with replacement. Subsequently, during classification, the results from all models are aggregated using a majority voting scheme. Bagging helps reduce the variance of the model, which can lead to better generalization performance, especially for complex models such as decision trees, which may tend to overfit on training data.

Additionally, in the training of boosting algorithms, a learning rate of 0.1 was defined, which determines the magnitude of contribution of each new model to the final decision of the ensemble.

In the context of developing classification models for the study, the input feature set was partitioned into a training set and a testing set using an 85/15% split ratio. To address the uneven distribution of examples across classes, it was essential to apply appropriate stratification techniques to ensure that each class was adequately represented in both the training and testing sets.

The study was conducted using the Monte Carlo method (Shonkwiler and Mendivil, 2024), wherein the training and testing procedure for each algorithm was performed 100 times, each time randomly selecting training and testing data sets. Additionally, during the algorithm training phase, a five-fold cross-validation (CV) approach was employed, further enhancing the robustness and reliability of the model evaluation process.

In the following section, we present the comparative results regarding the effectiveness of classification algorithms as discussed in the this section.

4. Research Results and Discussion

4.1 Comparison of Classification Methods Effectiveness

Based on the results obtained, various evaluation metrics were computed, which will be discussed below. In Figure 2 a) - c) sample visualizations of confusion matrices for three types of machine learning algorithms applied to classify respondents based on their responses to the question Q1: "Understanding cluster concepts” are presented.

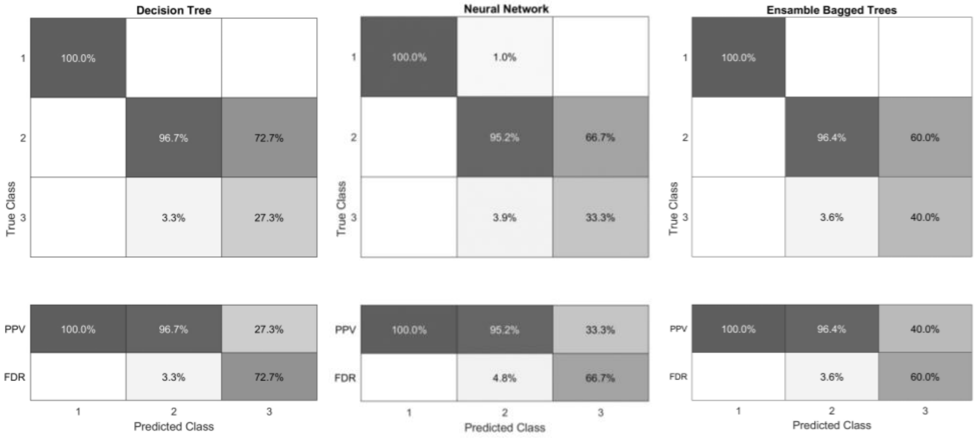
The confusion matrix in classification is a Table utilized to evaluate the performance of a classification model. It illustrates the number of correct and incorrect classifications for each class by the model. True Positive (TP) refers to the instances where the model correctly predicted a sample belonging to the positive class. False Positive (FP) denotes the instances where the model incorrectly predicted a sample belonging to the positive class when it actually belongs to the negative class (Type I error).

True Negative (TN) represents the instances where the model correctly predicted a sample belonging to the negative class. False Negative (FN) signifies the instances where the model incorrectly predicted a sample belonging to the negative class when it actually belongs to the positive class (Type II error). This table facilitates a precise assessment of the model's performance, including its accuracy, precision, sensitivity (recall, true positive rate), specificity, Matthews correlation coefficient (MCC), and F1 score.

As indicated in Table 1, respondents could provide three answers to question Q1: 1 = YES, 2 = NO, 3 = I DON'T KNOW. Heatmap-style graphs, presented in Figure 2, demonstrate variability in classification results depending on the group, yet they appear similar regardless of the type of machine learning algorithm used. The color blue corresponds to correctly classified respondents (TP, TN), while red indicates incorrectly classified respondents (FP, FN). Below the confusion matrix, there is a matrix of Positive Predictive Value (PPV) and False Discovery Rate (FDR) values.

PPV values should converge to 100%, while FDR should approach 0%. Based on observations from the heatmaps, a pattern emerges where class 1 (YES) is classified correctly 100% of the time, while class 2 (NO) is classified correctly over 95% of the time. However, there is a challenge with the classification of class 3 (I DON'T KNOW), which is only correctly classified around 27% to 40% of the time.

Figure 2. *Example confusion matrices for: (a) Decision tree; (b) Neural network; (c) Ensamble Bagged Tree. The models were developed based on the survey results for question Q1*



(a) **(b)** **(c)**

Source: *Authors' calculations.*

In further research confusion matrices were computed for seven considered machine learning methods and for all five analysed questions Q1-Q5, based on which other quality assessment metrics were determined (Tables 2-7). In Table 2, a summary of results (median value) for individual questions and machine learning algorithms is provided. The 5-fold CV Accuracy represents the model accuracy calculated through 5-fold cross-validation. A higher value indicates better performance in classifying data.

Testing Accuracy denotes the model's accuracy on a test set, i.e., data not used during model training. This metric is crucial as it demonstrates how well the model generalizes to new, unseen data. Training Time indicates the time required to train each model, a significant metric, especially concerning large datasets or limited computational resources. For the "Understanding cluster concepts" (Q1), three neural network models (3x10, 3x100, 3x200) achieve high accuracy values in both cross-validation (0.95) and on the test set (0.96), with the largest network (3x200) attaining the highest accuracy while maintaining an acceptable training time (3.08).

However, ensemble models (AdaBoostM2 Tree, RUSBoost Tree) exhibit lower cross-validation accuracy, although their test set results surpass those of cross-validation. Also high measures, exceeding 0.95, were achieved for the Bagged Tree and Simple Tree algorithms. Training time for neural networks in-creases with their size, while ensemble models demonstrate relatively shorter training times.

For “Managers' cluster perception“ (Q2), results are similar, albeit with slightly lower accuracy across all models compared to Q1. Classification results of “Company's cluster experience“ (Q3) and “Cluster operation status” (Q4), results are similar to those obtained in Q1 and Q2, although accuracies for all models appear slightly higher. For “Experience in business networks” (Q5), results are diverse. Ensemble AdaBoostM2 Tree and RUSBoost Tree models exhibit significantly lower cross-validation and test set accuracy compared to other models.

However, their training times are relatively shorter. Overall, neural network models achieve high accuracy, especially when adequately sized, albeit requiring longer training times. The decision tree learns the fastest while achieving high levels of accuracy simultaneously. The ensemble method of Bagged Trees demonstrates effectiveness comparable to neural networks, with significantly shorter learning times.

Table 2. *The aggregation of median values derived from 100 cycles of training/testing accuracy and time for Q1-Q5 classification using various machine learning algorithms*

Question	Metric name	Metric value (median over 100 trials) for Machine Learning Algorithm applied						
		3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoostM2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
Q1	5-fold CV Accuracy	0.93	0.94	0.95	0.23	0.80	0.96	0.95
	Testing Accuracy	0.94	0.94	0.96	0.73	0.79	0.96	0.96
	Training time	1.20	1.73	3.08	0.37	1.26	1.65	0.12
Q2	5-fold CV Accuracy	0.85	0.87	0.87	0.78	0.86	0.91	0.88
	Testing Accuracy	0.87	0.88	0.88	0.79	0.87	0.91	0.89
	Training time	7.78	2.55	4.50	0.23	2.30	2.91	0.27
Q3	5-fold CV Accuracy	0.90	0.93	0.93	0.77	0.90	0.95	0.94
	Testing	0.90	0.93	0.93	0.76	0.91	0.96	0.94

	Accuracy							
	Training time	2.78	1.79	3.17	0.36	1.11	1.48	0.13
Q4	5-fold CV Accuracy	0.94	0.97	0.97	0.77	0.53	0.99	0.99
	Testing Accuracy	0.94	0.96	0.97	0.76	0.58	0.99	0.99
	Training time	3.66	1.65	3.48	0.32	1.70	1.74	0.16
Q5	5-fold CV Accuracy	0.87	0.90	0.90	0.01	0.65	0.94	0.94
	Testing Accuracy	0.88	0.91	0.91	0.88	0.69	0.95	0.94
	Training time	5.06	1.86	3.17	0.36	1.45	1.31	0.09

Source: *Authors' calculations.*

In Table 3, the results (median over 100 iterations) of Matthews Correlation Coefficient (MCC) calculations for various machine learning algorithms are presented. Higher MCC values indicate better classification quality by the respective model. MCC ranges from -1 to +1, where +1 signifies perfect classification, 0 indicates random classification, and -1 signifies perfect inverse classification.

Comparing results for different algorithms across the same questions allows us to observe that there is no single algorithm that dominates in all cases. It is noteworthy that results vary depending on the question. Some algorithms may be more effective in classifying certain types of responses (e.g., "YES" vs. "NO") than others.

For example, in question Q1 - Understanding cluster concepts, the "Simple Tree" algorithm achieved an MCC of 1.00 for the "YES" response, suggesting perfect classification, while for other questions, the same algorithm may have lower MCC values. Ensemble algorithms such as AdaBoostM2 Tree, RUSBoost Tree, and Bagged Tree often exhibit stability and good performance compared to individual models.

However, in some cases, individual models may achieve similar or even better performance. It is worth noting that for some cases, MCC values are marked as NaN, indicating that the MCC could not be calculated for these cases. This may be due to lack of data or other technical factors preventing model fitting. The table includes mean and median values for each algorithm, calculated from all questions/answers.

The highest values are highlighted in gray. The Ensemble Bagged Tree and Simple Tree algorithms achieved the best average MCC scores, reaching 0.73 and 0.69 respectively and also best median values both exceeding 0.81.

Table 3. *The aggregation of median Matthews correlation coefficient values derived from 100 cycles of testing for Q1-Q5 classification using various machine learning algorithms*

Question and Responses		Matthews Correlation Coefficient (median over 100 trials) for Machine Learning algorithm applied						
		3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoost M2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
Q1	1 (YES)	0.96	0.96	0.96	NaN	1.00	1.00	1.00
	2 (NO)	0.85	0.85	0.89	NaN	0.61	0.88	0.88
	3 (I DON'T KNOW)	0.28	0.21	0.39	NaN	0.17	0.04	0.28
Q2	1 (POSITIV)	0.87	0.87	0.87	NaN	0.96	0.96	0.92
	2 (NEGATIVE)	0.30	0.31	0.25	NaN	0.55	0.48	0.44
	3 (NEUTRAL)	0.03	0.03	0.03	NaN	0.28	0.02	0.30
	0 (REFUSAL)	0.54	0.56	0.55	NaN	0.45	0.72	0.61
Q3	1 (YES)	0.92	0.98	0.96	NaN	1.00	1.00	1.00
	2 (NO)	0.04	0.30	0.04	NaN	0.33	0.57	0.39
	3 (I DON'T KNOW)	0.76	0.80	0.80	NaN	0.73	0.87	0.81
	0 (REFUSAL)	0.03	0.02	0.02	NaN	0.03	1.00	0.49
Q4	1 (YES)	0.96	0.96	0.96	NaN	0.17	1.00	1.00
	2 (NO)	0.02	0.02	1.00	NaN	0.27	1.00	1.00
	3 (I DON'T KNOW)	0.87	0.87	0.92	NaN	0.10	0.96	0.96
	0 (REFUSAL)	0.02	0.02	0.02	NaN	0.05	0.51	0.02
Q5	1 (YES)	0.70	1.00	1.00	NaN	0.23	1.00	1.00
	2 (NO)	0.58	0.70	0.62	NaN	0.18	1.00	1.00
	3 (I DON'T KNOW)	0.59	0.58	0.53	NaN	0.28	0.75	0.76
	0 (REFUSAL)	0.04	0.05	0.03	NaN	0.07	0.03	0.28
Mean		0.49	0.53	0.57	NaN	0.39	0.73	0.69
Median		0.58	0.58	0.62	NaN	0.28	0.88	0.81

Source: *Authors' calculations.*

In Table 4, the results for sensitivity (median over 100 iterations) of machine learning algorithms are compiled in the context of different questions. Sensitivity is a measure of the model's ability to correctly detect positive cases. Higher sensitivity indicates better performance in identifying positive cases. Sensitivity values range between 0 and 1, where 1 signifies perfect ability to detect positive cases, and 0 indicates an inability to detect them.

Interpreting the relative effectiveness of algorithms in different contexts, in some cases, such as question about the understanding cluster concepts (Q1) for the "YES" response, algorithms like Simple Tree, Ens. AdaBoostM2 Tree, Ens. RUSBoost Tree, and Ens. Bagged Tree achieve high sensitivity values (1.00), suggesting excellent ability to detect positive cases.

However, for other questions and responses, different algorithms exhibit varying effectiveness in detecting positive cases. For example, in question about the experience in business networks (Q5), the "YES" response, the 3x100 Neural Network algorithm achieves a sensitivity of 0.50, while the other models have lower values. The table presents mean and median values for each algorithm, derived from all questions and answers.

The highest values are shaded in gray, making it evident that the Ens. Bagged Tree algorithm achieves the highest values, with an average of 0.66 and a median of 0.98. Similar to MCC, Ensemble algorithms such as AdaBoostM2 Tree, RUSBoost Tree, and Bagged Tree demonstrate stability and good performance compared to individual models in some cases, especially in question about the understanding cluster concepts (Q1).

Similarly to MCC, some sensitivity values are zero, indicating that for these cases, the model did not detect any positive cases, as in Q2 - “Managers' cluster perception” response “NEUTRAL” and refusal to answer Q3 - “Company's cluster experience”, Q4 – “Cluster operation status” or Q5 – “Experience in business networks”. The RUSBoost Tree and Simple Tree algorithms achieved the best average sensitivity scores, both reaching 0.62 while the best median value, equal to 0.62, was achieved by the Ensemble Bagged Tree.

Table 4. *The aggregation of median sensitivity values derived from 100 cycles of testing for Q1-Q5 classification using various machine learning algorithms*

Questions and Responses		Sensitivity (median over 100 trials) for Machine Learning algorithm applied						
		3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoostM2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
Q1	1 (YES)	0.93	0.94	0.94	0.00	1.00	1.00	1.00
	2 (NO)	0.98	0.98	0.98	1.00	0.73	0.98	0.96
	3 (I DON’T KNOW)	0.33	0.17	0.33	0.00	0.50	0.00	0.33
Q2	1 (POSITIV)	0,96	0,98	0,98	1,00	0,98	0,98	0,98
	2 (NEGATIVE)	0,33	0,33	0,00	0,00	0,67	0,33	0,33
	3 (NEUTRAL)	0,00	0,00	0,00	0,00	0,50	0,00	0,50
	0 (REFUSAL)	0,60	0,63	0,63	0,00	0,33	0,90	0,67
Q3	1 (YES)	0.97	1.00	1.00	1.00	1.00	1.00	1.00
	2 (NO)	0.00	0.33	0.00	0.00	0.50	0.33	0.50
	3 (I DON’T KNOW)	0.83	0.92	0.85	0.00	0.67	1.00	0.83
	0 (REFUSAL)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q4	1 (YES)	0.98	1.00	1.00	1.00	0.67	1.00	1.00
	2 (NO)	0.00	0.00	0.00	0.00	1.00	1.00	1.00
	3 (I DON’T KNOW)	0.93	0.93	1.00	0.00	0.14	1.00	1.00
	0 (REFUSAL)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q5	1 (YES)	0.00	0.50	0.00	0.00	0.50	1.00	0.00
	2 (NO)	0.60	0.50	0.50	0.00	0.40	1.00	1.00

	3 (I DON'T KNOW)	0.95	0.97	0.98	1.00	0.73	0.98	0.97
	0 (REFUSAL)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean		0.49	0.54	0.48	0.26	0.54	0.66	0.64
Median		0.6	0.5	0.5	0	0.5	0.98	0.83

Source: *Authors' calculations.*

In Table 5, the results for specificity (median over 100 iterations) of machine learning algorithms are compiled in the context of questions Q1-Q5. Specificity is a measure of the model's ability to correctly identify negative cases. Higher specificity indicates better performance in identifying negative cases. Specificity values also range between 0 and 1, where 1 signifies perfect ability to detect negative cases, and 0 indicates an inability to detect them.

Similar to sensitivity, some models achieve high specificity, especially for certain questions and responses. For instance, for "Understanding cluster concept" (Q1), and "Experience in business networks" (Q5), response "YES", all models achieve full specificity (1.00). However, for other questions and responses, specificity values may vary depending on the model and context. Like sensitivity, Ensemble algorithms and individual models exhibit differing effectiveness in detecting negative cases in various scenarios.

In some cases, such as question about Managers' cluster perception (Q2) for response "POSITIVE", individual models achieve higher specificity compared to Ensemble algorithms. The table displays also the mean and median values for each algorithm, calculated from all questions and answers. The highest values, shaded in gray, approach unity, indicating exceptional performance. It is worth noting, however, that nearly all values are remarkably high, underscoring the overall effectiveness of the algorithms. All models except Ensemble AdaboostM2 achieved an average specificity exceeding 0.94.

Table 5. *The aggregation of median specificity values derived from 100 cycles of testing for Q1-Q5 classification using various machine learning algorithms*

Question and Responses		Specificity (median over 100 trials) for Machine Learning algorithm applied						
		3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoostM2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
Q1	1 (YES)	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2 (NO)	0.88	0.89	0.89	0.00	0.94	0.89	0.89
	3 (I DON'T KNOW)	0.97	0.97	0.98	1.00	0.80	0.98	0.97
Q2	1 (POSITIV)	0.93	0.93	0.86	0.00	1.00	1.00	0.97
	2 (NEGATIVE)	0.97	0.98	0.98	1.00	0.95	0.98	0.97
	3 (NEUTRAL)	0.97	0.98	0.98	1.00	0.94	1.00	0.98
	0 (REFUSAL)	0.93	0.93	0.93	1.00	0.98	0.91	0.95
Q3	1 (YES)	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	2 (NO)	0.98	0.98	0.97	1.00	0.95	1.00	0.97
	3 (I DON'T	0.95	0.96	0.96	1.00	0.98	0.96	0.96

	KNOW)							
	0 (REFUSAL)	0.98	1.00	1.00	1.00	0.97	1.00	1.00
Q4	1 (YES)	1.00	0.94	1.00	0.00	0.50	1.00	1.00
	2 (NO)	0.98	1.00	1.00	1.00	0.91	1.00	1.00
	3 (I DON’T KNOW)	0.96	0.96	0.96	1.00	0.90	0.98	0.98
	0 (REFUSAL)	0.99	1.00	1.00	1.00	0.89	1.00	1.00
Q5	1 (YES)	1.00	1.00	1.00	1.00	0.89	1.00	1.00
	2 (NO)	0.97	0.98	0.98	1.00	0.89	1.00	1.00
	3 (I DON’T KNOW)	0.63	0.50	0.46	0.00	0.71	0.63	0.75
	0 (REFUSAL)	0.97	0.97	0.98	1.00	0.93	0.98	0.97
Mean		0.95	0.95	0.95	0.74	0.9	0.96	0.97
Median		0.97	0.98	0.98	1	0.94	1	0.98

Source: *Authors‘ calculations.*

In Table 6, the results for precision (median over 100 iterations) of machine learning algorithms are presented in the context of different questions. Precision is a measure of the proportion of correctly predicted positive cases out of all predicted positive cases by the model. Higher precision indicates fewer false positive cases. Precision values also range between 0 and 1, where 1 signifies perfect ability to predict positive cases without false alarms, and 0 indicates an inability to predict them.

Similar to sensitivity and specificity, some models achieve high precision, especially for certain questions and responses. For instance, for question about “Understanding cluster concepts”, response "YES", all models achieve perfect precision (1.00). However, for other questions and responses, precision values vary depending on the model and context. Like sensitivity and specificity, ensemble algorithms and individual models exhibit differing effectiveness in predicting positive cases in various scenarios.

In some cases, such as question “Managers' cluster perception” for response "POSITIVE", individual models achieve higher precision compared to ensemble algorithms. Similarly to previous cases, some precision values are zero or could not be calculated (NaN), suggesting that for these cases, the model did not predict any positive cases. The highest values are shaded in gray, clearly indicating that the bagged tree algorithm attains the top performance, with an average of 0.75 and a median of 0.96.

Table 6. *The aggregation of median precision values derived from 100 cycles of testing for Q1-Q5 classification using various machine learning algorithms*

Questions and Responses		Precision (median over 100 trials) for Machine Learning algorithm applied						
		3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoostM2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
Q1	1 (YES)	1.00	1.00	1.00	NaN	1.00	1.00	1.00
	2 (NO)	0.96	0.96	0.96	0.73	0.97	0.96	0.96
	3 (I DON’T KNOW)	0.25	0.17	0.33	NaN	0.09	0.00	0.20

Q2	1 (POSITIV)	0.98	0.98	0.96	0.79	1.00	1.00	0.99
	2 (NEGATIVE)	0.33	0.33	0.25	NaN	0.41	0.50	0.50
	3 (NEUTRAL)	0.00	0.00	0.00	NaN	0.20	0.00	0.20
	0 (REFUSAL)	0.63	0.62	0.60	NaN	0.75	0.64	0.67
Q3	1 (YES)	1.00	1.00	1.00	0.76	1.00	1.00	1.00
	2 (NO)	NaN	0.25	NaN	NaN	0.25	1.00	0.33
	3 (I DON'T KNOW)	0.77	0.83	0.83	NaN	0.85	0.85	0.85
	0 (REFUSAL)	NaN	NaN	NaN	NaN	NaN	1.00	0.50
Q4	1 (YES)	1.00	0.98	1.00	0.76	0.82	1.00	1.00
	2 (NO)	0.00	0.00	1.00	NaN	0.09	1.00	1.00
	3 (I DON'T KNOW)	0.87	0.87	0.88	NaN	0.28	0.93	0.93
	0 (REFUSAL)	0.00	0.00	0.00	NaN	0.00	0.50	0.00
Q5	1 (YES)	0.50	1.00	1.00	NaN	0.09	1.00	1.00
	2 (NO)	0.57	0.78	0.80	NaN	0.17	1.00	1.00
	3 (I DON'T KNOW)	0.95	0.94	0.93	0.88	0.94	0.95	0.97
	0 (REFUSAL)	0.00	0.00	0.00	NaN	0.00	0.00	0.20
Mean		0.52	0.56	0.61	0.79	0.47	0.75	0.7
Median		0.57	0.78	0.83	0.76	0.28	0.96	0.93

Source: *Authors' calculations.*

In Table 7, results encompassing F1 score values (median over 100 iterations) for the considered machine learning algorithms across the investigated question-answer scenarios are presented. The F1 measure combines precision and recall, providing a balanced classification metric. Values close to 1 indicate a balance between precision and recall.

The interpretation of the obtained results is as follows: Question and answers about the "Understanding cluster concepts" (Q1): F1 values for all models are high when the answer is "YES", indicating proficient classification of correct responses. For "NO", F1 values are varied, but Ensemble Ada-BoostM2 Tree model achieve lower or none (NaN) results than neural networks, other ensembles or the simple tree. Results for "I DON'T KNOW" are significantly lower.

Question and answers about the "Man-agers' cluster perception" (Q2): F1 values are high for "POSITIVE" responses. both for neural networks and some ensemble models. For "NEGATIVE", "NEUTRAL", and "REFUSAL" responses, results are much wors, not exceeding 0.75 for the Bagging Tree. Question and answers about the "Company's cluster experience" (Q3): Similar to Q1, models attain high F1 values for "YES" responses. For "NO" results are varied, with ensemble models tending to achieve lower F1 values.

Results for "I DON'T KNOW" are diverse, but ensemble models frequently achieve lower F1 values. Question and answers about the "Cluster operation status" (Q4): F1 values are high for "YES" responses across all models. For "NO", ensemble models (Ada-BoostM2 and RUSBoost) often achieve lower F1 values. Similar for "I

DON'T KNOW" and "REFUSAL" responses. Question and answers about the “Experience in business networks” (Q5): F1 values are high for "YES" responses for most models. For "NO", ensemble models often achieve lower F1 values. Results for "I DON'T KNOW" are diverse, but ensemble models tend to achieve lower F1 values.

For "REFUSAL" responses, results are diverse, but ensemble models achieve lower F1 values compared to neural networks. Overall, neural network models frequently achieve higher F1 values for various response categories, particularly for positive responses. Ensemble models yield diverse results, but often exhibit lower F1 values compared to neural networks.

When considering the F1-score metric, it becomes evident that algorithms Ens. Bagged Tree and Simple Tree outperformed others, exhibiting the highest median values, exceeding 0.97. Notably, the algorithm Ens. Bagged Tree achieved the most favorable mean score among all algorithms assessed, reaching 0.86, with a median of 0.98.

Table 7. *The aggregation of median F1 score values derived from 100 cycles of testing for Q1-Q5 classification using various machine learning algorithms*

Questions and Responses		F1 Score value (median over 100 trials) for Machine Learning algorithm applied						
		3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoostM2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
Q1	1 (YES)	0.97	0.97	0.97	NaN	1.00	1.00	1.00
	2 (NO)	0.96	0.96	0.97	0.84	0.84	0.97	0.97
	3 (I DON'T KNOW)	0.38	0.40	0.50	NaN	0.17	0.40	0.40
Q2	1 (POSITIV)	0.97	0.97	0.97	0.88	0.99	0.99	0.98
	2 (NEGATIVE)	0.40	0.45	0.40	NaN	0.55	0.50	0.50
	3 (NEUTRAL)	0.40	0.40	0.50	NaN	0.33	NaN	0.50
	0 (REFUSAL)	0.60	0.62	0.62	NaN	0.46	0.75	0.67
Q3	1 (YES)	0.98	1.00	0.99	0.86	1.00	1.00	1.00
	2 (NO)	0.40	0.50	0.40	NaN	0.40	0.67	0.50
	3 (I DON'T KNOW)	0.80	0.83	0.83	NaN	0.76	0.89	0.85
	0 (REFUSAL)	0.50	0.67	0.67	NaN	0.31	1.00	0.67
Q4	1 (YES)	0.99	0.99	0.99	0.86	0.74	1.00	1.00
	2 (NO)	1.00	1.00	1.00	NaN	0.25	1.00	1.00
	3 (I DON'T KNOW)	0.90	0.90	0.93	NaN	0.22	0.97	0.97
	0 (REFUSAL)	0.75	NaN	NaN	NaN	0.19	1.00	0.45
Q5	1 (YES)	1.00	1.00	1.00	NaN	0.20	1.00	1.00
	2 (NO)	0.60	0.67	0.60	NaN	0.25	1.00	1.00
	3 (I DON'T KNOW)	0.95	0.96	0.95	0.94	0.82	0.97	0.97
	0 (REFUSAL)	0.33	0.40	0.50	NaN	0.29	0.40	0.37
Mean		0.73	0.76	0.77	0.88	0.51	0.86	0.78
Median		0.8	0.86	0.88	0.86	0.4	0.98	0.97

Source: *Authors‘ calculations.*

4.2 Selection of the Best Classification Method

In order to select the most suitable algorithm for implementation in an expert system, Table 8 presents the mean and median values of all metrics. In particular, table 8 presents the aggregation of median and mean values (excluding NaN), calculated as averages for all Q1-Q5 classifications using various machine learning algorithms. The table includes average values for the following performance metrics: F1 score, Matthews correlation coefficient (MCC), accuracy, sensitivity, specificity, precision and the overall mean over all metrics.

Based on the data, the Ens. Bagged Tree algorithm emerged as the most appropriate choice, followed by the Simple Tree algorithm. Neural networks, however, did not perform as effectively. Although the Ens. AdaBoostM2 Tree algorithm produced reasonably good results, it frequently had metrics with NaN values, indicating that the algorithm did not effectively classify all classes.

Table 8. *The aggregation of median and mean values (excluded Nan) derived from all Q1-Q5 classification using various machine learning algorithms*

Metric name	3x10 Neural Network	3x100 Neural Network	3x200 Neural Network	Ens. AdaBoostM2 Tree	Ens. RUSBoost Tree	Ens. Bagged Tree	Simple Tree
	Median value						
F1 score	0.8	0.86	0.88	0.86	0.4	0.98	0.97
MCC	0.58	0.58	0.62	NaN	0.28	0.88	0.81
Accuracy	0.9	0.93	0.93	0.76	0.79	0.96	0.94
Sensitivity	0.6	0.5	0.5	0	0.5	0.98	0.83
Specificity	0.97	0.98	0.98	1	0.94	1	0.98
Precision	0.57	0.78	0.83	0.76	0.28	0.96	0.93
Overall Mean	0.74	0.77	0.79	0.68	0.53	0.96	0.91
	Mean value						
F1 score	0.73	0.76	0.77	0.88	0.51	0.86	0.78
MCC	0.49	0.53	0.57	NaN	0.39	0.73	0.69
Accuracy	0.91	0.92	0.93	0.79	0.77	0.95	0.94
Sensitivity	0.49	0.54	0.48	0.26	0.54	0.66	0.64
Specificity	0.95	0.95	0.95	0.74	0.9	0.96	0.97
Precision	0.52	0.56	0.61	0.79	0.47	0.75	0.7
Overall Mean	0.68	0.71	0.72	0.69	0.60	0.82	0.79

Source: *Authors‘ calculations.*

5. Conclusions

This study focuses on the classification of entrepreneurs through five key questions: understanding cluster concepts, managers' perceptions of clusters, companies' experiences with clusters, the operational status of clusters, and experience in business networks.

An extensive survey conducted through questionnaires gathered 448 responses, providing comprehensive insights into these aspects within the Polish market of family-owned businesses.

The research hypothesizes that entrepreneurs can be accurately classified by applying machine learning methods to their responses to selected questions about their business activities. The article details the classification procedure using simple decision trees, ensembles of trees with various aggregation methods, and neural networks. It provides the necessary information to replicate the experiment.

The classification results from different algorithms are compared using quality measures such as accuracy, sensitivity, specificity, precision, Matthews correlation coefficient, and the F1 score. The presented results confirm the feasibility of accurately classifying entrepreneurs based on their knowledge of cluster formation with high effectiveness and thereby also validate the research hypothesis.

The chosen algorithm was implemented in an online classifier for users within a tool for to simulate the functioning of a family business within the framework of cluster and outsourcing concepts. The tool allows for real-time evaluation and support for entrepreneurs, leveraging the algorithm's superior classification performance to provide accurate and reliable assessments.

The presented tool is not fully accurate, achieving an average overall effectiveness of 82% (arithmetic mean of several quality measures), which is due to a limited set of training data. Additionally, the results are based on a limited set of survey data from four regions in Poland.

However, it is possible to expand the training data set by conducting further surveys in other parts of Poland or Europe and retraining the existing algorithm. Further work may therefore involve continuing the survey research and refining the developed classification tool.

References:

Abien Fred M. Agarap. 2019. Deep Learning using Rectified Linear Units (ReLU).

Birdthistle, N., Hales, R. 2023. The Meaning of a Family Business and Why They Are Important to Economies. Attaining the 2030 Sustainable Development Goal of Gender Equality, 13-24. DOI:10.1108/978-1-80455-832-420231002.

Bishop, C.M. 1996. Neural Networks for Pattern Recognition Advanced Texts in Econometrics. Oxford University Press.

Chen, X., Wang, E., Miao, C., Ji, L., Pan, S. 2020. Industrial Clusters as Drivers of Sustainable Regional Economic Development? An Analysis of an Automotive Cluster from the Perspective of Firms' Role. Sustainability, 12(7), 2848-2848. DOI:10.3390/SU12072848.

Corallo, A., Del Vecchio, V., Lezzi, M., Luperto, A. 2022. Model-Based Enterprise Approach in the Product Lifecycle Management: State-of-the-Art and Future Research Directions. Sustainability, 14(3), 1370. DOI:10.3390/SU14031370.

Costa, V. G., Pedreira, C.E. 2022. Recent advances in decision trees: an updated survey. Artificial Intelligence Rev, 56(5), 4765-4800. DOI:10.1007/S10462-022-10275-5.

Derlukiewicz, N., Mempel-Śniezyk, A., Mankowska, D., Dyjakon, A., Minta, S., Pilawka, T. 2020. How do Clusters Foster Sustainable Development? An Analysis of EU Policies. Sustainability, 12(4), 1297. DOI:10.3390/SU12041297.

Dong, Y., Hou, J., Zhang, N., Zhang, M. 2020. Research on How Human Intelligence, Consciousness, and Cognitive Computing Affect the Development of Artificial Intelligence. Complexity. DOI:10.1155/2020/1680845.

Dwiyanti, E., Ardiyanti, A., Ardiyanti, A. 2017. Handling Imbalanced Data in Churn Prediction Using RUSBoost and Feature Selection (Case Study: PT.Telekomunikasi Indonesia Regional 7). Advances in Intelligent Systems and Computing, 549 AISC, 376-385. DOI:10.1007/978-3-319-51281-5_38.

Fitzgerald, M.A., Muske, G. 2016. Family businesses and community development: the role of small business owners and entrepreneurs. Community Development, 47(4), 412-430. DOI:10.1080/15575330.2015.1133683.

Halawi, L., Clarke, A., George, K. 2022. Decision Trees and Ensemble. Harnessing the Power of Analytics, 61-81. DOI:10.1007/978-3-030-89712-3_5.

Hastie, T., Rosset, S., Zhu, J., Zou, H. 2009. Multi-class AdaBoost. Statistics and Its Interface, 2(3), 349-360. DOI:10.4310/SII.2009.V2.N3.A8.

IDEAMOTIVE. 2024. Implementing Artificial Intelligence in Your Business: All-In-One Guide. https://www.ideamotive.co/ai-developers/guide#benefits-of-artificial-intelligence-in-business.

Jason Brownlee. 2020. Bagging and Random Forest Ensemble Algorithms for Machine Learning. https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/.

Kerr, W.R., Robert-Nicoud, F. 2019. Tech Clusters (No. 20-063).

Liapis, K., Rovolis, A., Galanos, C., Thalassinos, E. 2013. The clusters of economic similarities between EU countries: a view under recent financial and debt crisis. European Research Studies, 16(1), 41-70.

Legaard, C.M., Schranz, T., Schweiger, G., Drgoňa, J., Falay, B., Gomes, C., Larsen, P.G. 2022. Constructing Neural Network-Based Models for Simulating Dynamical Systems. CSUR.

Mills, K.G., Reynolds, E.B., Reamer, A. 2008. Clusters and Competitiveness: A New Federal Role for Stimulating Regional Economies. Brookings.

Muro, M., Katz, B. 2010. The new „cluster moment": how regional innovation clusters can foster the next economy. Brookings.

Najafabadi, M.M., Khoshgoftaar, T.M., Villanustre, F., Holt, J. 2017. Large-scale distributed L-BFGS. Journal of Big Data, 4(1), 1-17. DOI:10.1186/S40537-017-0084-5/TABLES/4.

Noja, G.G., Cristea, M., Thalassinos, E., Kadłubek, M. 2021. Interlinkages between government resources management, environmental support, and good public governance. Advanced Insights from the European Union. Resources, 10(5), 41.

Norena-Chavez, D., Thalassinos, E. 2023. Impact of big data analytics in project success: Mediating role of intellectual capital and knowledge sharing. Journal of Infrastructure, Policy and Development, 7(3), 2583.

Reza Godaz, Benyamin Ghojogh, Reshad Hosseini, Reza Monsefi, Fakhri Karray, and Mark Crowley. 2021. Vector Transport Free Riemannian LBFGS for Optimization on Symmetric Positive Definite Matrix Manifolds. 13th Asian Conference on Machine Learning (ACML).

Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A. 2010. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and

Cybernetics Part A: Systems and Humans, 40(1), 185-197. DOI:10.1109/TSMCA.2009.2029559.

Shalev-Shwartz, S., Ben-David, S. 2014. Decision Trees. Understanding Machine Learning, 212-218. DOI:10.1017/CBO9781107298019.019.

Shonkwiler, R.W., Mendivil, F. 2024. Explorations in Monte Carlo Methods. Cham: Springer Nature Switzerland. DOI:10.1007/978-3-031-55964-8.

Smith, R.A. 2023. 10 Things That Help Family Businesses Preserve Their Legacy. Entrepreneur and Innovation Exchange. DOI:10.32617/979-652938799F9D6.

Staszewska, J., Smolarek, M., Foltys, J., Wotzka, D., Fracz, P. 2024. The Possibilities of Cooperation among Family Firms within a Cluster Environment. Eropean Research Studies Journal, XXVII(Issue 2), 132-154. DOI:10.35808/ERSJ/3375.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M. 2020. Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7(1), 1-47. DOI:10.1186/S40537-020-00349-Y/FIGURES/5.

Tyagi, P., Grima, S., Sood, K., Balamurugan, B., Özen, E., Thalassinos, E.I. (Eds.). 2023. Smart analytics, artificial intelligence and sustainable performance management in a global digitalised economy. Emerald Publishing Limited.

Ultsch, A., Lötsch, J. 2017. A data science based standardized Gini index as a Lorenz dominance preserving measure of the inequality of distributions. PLOS ONE, 12(8), e0181572. DOI:10.1371/JOURNAL.PONE.0181572.

Velinov, E., Kadłubek, M., Thalassinos, E., Grima, S., Maditinos, D. 2023. Digital Transformation and Data Governance: Top Management Teams Perspectives. In Digital Transformation, Strategic Resilience, Cyber Security and Risk Management (Vol. 111, pp. 147-158). Emerald Publishing Limited.

Xu, B., Wang, N., Chen, T., Li, M. 2015. Empirical Evaluation of Rectified Activations in Convolution Network. DOI:10.48550/arXiv.1505.00853.

Yilmaz, Y., Raetze, S., Groote, J.D., Kammerlander, N. 2024. Resilience in Family Businesses: A Systematic Literature Review. Family Business Review, 37(1), 60-88. DOI:10.1177/0894486523122337.

Zoltan J. Acs, Laszlo Szerb. 2006. Entrepreneurship, Economic Growth and Public Policy on JSTOR. Small Business Economics, 28(2/3), 109-122. DOI:10.1007/s11187-006-9012-3.