Entrepreneurial Business and Economics Review

2025, Vol. 13, No. 2

10.15678/EBER.2025.130206

Exploring entrepreneurial phases with machine learning models: Evidence from Hungary

Aron Szennay, Judit Csákné Filep, Melinda Krankovits

ABSTRACT

Objective: The article aims to explore the potential differences between the two phases of entrepreneurship, i.e., total early-stage entrepreneurial activity and established business, as defined by the Global Entrepreneurship Monitor (GEM). The study aimed to classify entrepreneurs using various machine learning models and to evaluate their classification performance comparatively.

Research Design & Methods: Using the Hungarian GEM datasets from 2021 to 2023, we analysed a subsample of 964 entrepreneurs. Due to inconsistent results from traditional analyses (e.g., correlations, regressions, principal component analyses), we employed machine learning approaches (supervised learning classification methods) to uncover latent relationships between variables.

Findings: The study utilized seven machine learning classification methods to examine the feasibility of grouping companies within the sample using Hungarian GEM data. Findings indicate that machine learning techniques are particularly effective for classifying businesses, although the performance of each method varies significantly.

Implications & Recommendations: These results provide valuable insights for researchers in selecting methodologies to identify various business phases. Moreover, they offer practical benefits for market research professionals, suggesting that machine learning techniques can enhance the classification and understanding of entrepreneurial phases.

Contribution & Value Added: The study adds to the existing body of knowledge by demonstrating the effectiveness of machine learning methods in classifying business phases. It highlights the variability in performance across different machine learning techniques, thereby guiding future research and practical applications in market research and entrepreneurship studies.

Article type:	research article
Keywords:	entrepreneurship; responsibility; Global Entrepreneurship Monitor; GEM; machine learning
JEL codes:	L26, C38

Received: 12 July 2024 | Revised: 1 February 2025 | Accepted: 25 March 2025

Suggested citation:
Szennay, A., Csákné Filep, J., & Krankovits, M. (2025). Exploring entrepreneurial phases with machine learning models: Evidence from Hungary. *Entrepreneurial Business and Economics Review,* 13(2), 101-122. https://doi.org/10.15678/EBER.2025.130206

INTRODUCTION

Entrepreneurship is a fundamental driver of economic growth, but its role varies across countries at different stages of economic development (Stel *et al*., 2005). Although just a small but significant share of new business ventures as innovators, who contribute to the diffusion of new products, services, and even processes into the economy. Moreover, since the sheer number of new ventures is large, entrepreneurship fosters both innovation and competition in economies, contributing it its continuous restructuration (Sternberg & Wennekers, 2005). Furthermore, the role of enterprises is also crucial in achieving sustainability. Agenda 2030, the framework for sustainable development of the United Nations, explicitly calls for all businesses to apply their creativity and innovation to solving sustainable development challenges (United Nations, 2015).

We aimed to elucidate the differences between early-stage and established enterprises by examining a comprehensive set of attributes (*e.g.*, demography, entrepreneurial motivations, market scope, attitudes towards responsibility, *etc*.) using the Global Entrepreneurship Monitor data for Hungary. Entrepreneurship demographics is a well-studied research area (see Wach & Głodowska, 2021). However, quantitative methods often fall short when investigating determinants of particular activities (for example in the case of responsible behaviour (Krankovits *et al*., 2023) or even differences between phases of entrepreneurship. Thus, we employed machine learning techniques to determine whether these variables can accurately determine the phase of entrepreneurship. This research is motivated by the increasing trend of utilizing machine learning in social science research, which offers a robust alternative to traditional analytical methods that often fall short of uncovering complex patterns (Celbiş, 2021; Chung, 2023; Razaghzadeh Bidgoli *et al*., 2024). Furthermore, by focusing on entrepreneurs in Hungary, we sought to provide localized insights that can contribute to both regional policymaking and the broader theoretical understanding of entrepreneurial dynamics. Our findings aim to bridge the gap in the existing literature by demonstrating the efficacy of machine learning in identifying nuanced differences in entrepreneurial phases, thereby offering a novel methodological approach that one can replicate in other contexts. Therefore, we posed two research questions:

RQ1: Is it feasible to determine the phase of entrepreneurship with sufficient accuracy using a variable set available in the Global Entrepreneurship Monitor?

RQ2: Are machine learning techniques adequate methods to classify entrepreneurs based on their attributes?

The article is structured as follows. The next chapter will summarize both (1) the conceptual framework applied and the variables analysed on the base of this, and (2) the background of the machine learning approach applied. Then, we will describe the dataset and the methodologies used. The article will conclude with the results, discussion, research limitations, and suggestions for further empirical research.

LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

Conceptual Framework and Explanation of Variables Used

The article uses the Hungarian data of the Global Entrepreneurship Monitor (GEM) and thus, the revised GEM conceptual framework (Kelley *et al*., 2016). The GEM is the world's foremost study of entrepreneurship, collecting data directly from entrepreneurs (GEM, 2024). GEM is a consortium of national teams to collect and analyse data on entrepreneurship and entrepreneurship ecosystems, representing countries with almost half of the global population and two-thirds of GDP in 2021, 2022, and 2023 (GEM, 2022, 2023, 2024). Sternberg and Wennekers (2005) summarise the main objectives of GEM in four points: (1) to empirically examine variations in entrepreneurial activity between countries over time, (2) to identify reasons behind differing levels of entrepreneurship, (3) to explore policies that may boost entrepreneurial activity, and (4) to understand the link between entrepreneurship and economic growth.

The GEM methodology distinguishes four phases of entrepreneurship considering the three typical entrepreneurial barriers (Reynolds *et al*., 2005). The first one is when the startup of a business including any self-employment or selling any goods or services is expected in the next three years. Nascent entrepreneurs have an existing enterprise which did not pay wages or salaries for three months, while baby businesses paid wages or salaries between three and 42 months. These latter two together are called total early-stage entrepreneurial activity (TEA). The fourth phase is the established business (EB), for which salaries or wages have been paid for more than 42 months (GEM, 2023).

According to the revised GEM conceptual framework (Kelley *et al*., 2016), both (1) social values about entrepreneurship and (2) individual attributes, including demographic characteristics (*e.g.*, gender, age), self-perceptions and motivations moderate the relationship between entrepreneurial activity and social, cultural, political, economic context. A variable set was chosen for the analysis to reflect this conceptual framework. However, the final model comprises only variables with significant determining power.

The literature on the demographic attributes of entrepreneurship is rather rich. Age is a principal determining factor of entrepreneurial activity. Younger age cohorts tend to be entrepreneurially more active (see for example Csákné Filep *et al*. (2023) or Lafuente and Vaillant (2013) which has a striking consequence on the ageing societies (Lévesque & Minniti, 2011). However, Kautonen *et al*. (2014) highlight that the entrepreneurial activity of owner-managers shows an inverted U-shaped curve with a threshold age of 40 years, while in the case of self-employers, it increases almost linearly with age. Similarly, educational attainment generally correlates with entrepreneurial activity (Csákné Filep *et al*., 2023), and the performance of SMEs (Filser & Eggers, 2014), while the entrepreneur's managerial knowledge, expertise and skills positively affect the firm's early internationalisation (Wach & Głodowska, 2021). Formal education has a positive causal effect on any type of self-employment for women, while it contributes to the shift from shrinking industry self-employment to high-growth one in the case of men (Ahn & Winters, 2023). However, Kurczewska *et al*. (2020) found that both education and professional experience are necessary for entrepreneurial success. Furthermore, the authors' model implies that the gender of the entrepreneur also contributes to the success, as businesses run by men are more likely to survive.

The GEM methodology investigates the role of four entrepreneurial motives: (1) to make a difference in the world, (2) to build great wealth or a very high income, (3) to continue a family tradition, and (4) to earn a living because jobs are scarce. Weber (1982) mentions the motive of building great wealth or a very high income as an entrepreneurial goal. He considers it the foundation of capitalism. However, empirical research does not support the notion that wealth accumulation is the sole or primary motivation for starting a business (*e.g.*, Amit *et al*., 2001). Continuing family tradition can also motivate entrepreneurial activity, whether it involves establishing a new enterprise or taking over and continuing an existing family business. However, Gorgievski *et al*. (2011) suggest that measuring an entrepreneur's performance requires more than just considering business criteria (*e.g.*, growth, profit, etc.). Factors that may have a trade-off relationship, such as achieving social impact or work-life balance, also require examination. The examination of entrepreneurial motivations widely acknowledges the dichotomy between necessity-driven and opportunity-driven entrepreneurship – this is explored by the fourth entrepreneurial motivation of the GEM, which investigates livelihood motives. Noteworthy, while opportunity-driven entrepreneurship is characteristic of developed countries, necessity-driven entrepreneurship is more typical of developing nations (Acs, 2006; Szerb, 2004).

Filser and Eggers (2014) suggest that while the manager's risk-taking and innovativeness affect the performance of Rhine Valley (Austria, Liechtenstein, Switzerland) SMEs, their proactiveness does not. Furthermore, Ključnikov *et al*. (2019) found that the risk-taking and competitive aggressiveness of Czech and Turkish SME managers differ by gender, while their innovativeness, proactiveness, and autonomy are similar.

According to the GEM methodology, market scope captures the regionally farthest group of consumers. As Wach and Głodowska (2021) found, both age and education increase the pace of internationalisation in the case of Polish entrepreneurs.

Thus, the first hypothesis of the study is as follows:

H1: Early-stage entrepreneurs and established businesses are different in their characteristics.

Machine Learning, as a Vehicle for Gaining a Deeper Understanding

If scholars cannot identify deeper correlations in statistical analyses, the question may arise whether deep learning or machine learning can help us. Both data mining methods are very popular, but there are significant differences in the focus of the method. While deep learning is an unsupervised method, *i.e.*, the data does not need to be labelled, machine learning classification methods are supervised learning, *i.e.*, there must be a test set and labels.

Deep learning presents significant advantages over traditional statistical analysis methods, enabling the analysis of complex data that may be challenging to analyse using conventional statistical approaches (Park & Hong, 2022). One of the key strengths of deep learning is its capability to handle

vast amounts of unlabelled and un-categorized data, making it particularly valuable in big data analytics scenarios (Najafabadi *et al*., 2015).

While statistical models like regression offer interpretability advantages over deep learning, the latter's ability to learn from data without the need for extensive hand-crafted feature engineering sets it apart (Staartjes *et al*., 2018). Moreover, deep learning models have been successful in tasks like medical image segmentation, object detection, and pollution forecasting, showcasing their versatility and effectiveness across diverse domains (Chen, 2023; Nath *et al*., 2021; Soria *et al*., 2020).

Data classification is a fundamental aspect of managing data in an entrepreneur's database. It involves organizing and categorizing information to facilitate decision-making processes and improve business operations. Through data classification, entrepreneurs can gain valuable insights into customer preferences, market trends, and operational efficiencies (Bhukya & Ramachandram, 2010). This structured approach enables entrepreneurs to identify patterns, trends, and relationships within their database, leading to informed strategic decisions and targeted marketing efforts (Wood & Salzberg, 2014).

Furthermore, data classification allows entrepreneurs to effectively segment their customer base, enabling personalized marketing campaigns and tailored product offerings (Stewart *et al*., 2019). By categorizing data into different classes based on common properties, entrepreneurs can better understand customer behaviour and preferences, ultimately enhancing customer satisfaction and retention (Bhukya & Ramachandram, 2010). Moreover, data classification supports risk assessment and fraud detection, helping entrepreneurs identify potential threats and take proactive measures to mitigate risks (Rezende *et al*., 2022).

In the context of relational databases, relational classification techniques offer advantages over propositional data mining approaches by directly classifying data involving multiple relations. This approach provides a more comprehensive understanding of interconnected data points, enhancing the entrepreneur's ability to extract meaningful insights from complex relational data structures and contributing to more accurate decision-making processes and business strategies (Vaghela *et al*., 2012).

In conclusion, data classification is indispensable for entrepreneurs to organize information, identify patterns, segment customers, assess risks, and make informed decisions. By leveraging data classification techniques, entrepreneurs can fully utilize their databases, leading to improved operational efficiency, targeted marketing strategies, and enhanced business performance.

Investigating the performance of machine learning techniques on entrepreneurship data, we formed a second hypothesis:

H2: Machine learning algorithms have reliable (above 90%) accuracy in distinguishing stages of entrepreneurial activity.

RESEARCH METHODOLOGY

We based the analysis on the GEM Adult Population Survey (APS) 2021, 2022 and 2023. Each APS dataset is representative of the 18-64-year-old adult population (n=2000), but we considered only the subsample of entrepreneurs in the analyses. The APS data collection was coordinated, supervised, and checked by the Global GEM team ensuring the consistency of responses in each GEM country. Thus, the resulting data were repeatedly checked before publication, so all variables and measures reflect the common GEM methodology (for example (GEM (Global Entrepreneurship Monitor), 2023; Reynolds *et al*., 2005)).

Hungarian GEM data are available only for 2021, 2022 and 2023 since the former national team terminated its membership in the international consortium in 2016. To have a larger sample of entrepreneurs, we merged these three years of data into one dataset database, assuming that the attitudes and behaviours of entrepreneurs do not change significantly over one year and there were no such new policies or other external circumstances which could significantly alter them. Thus, the sample comprised 964 entrepreneurs' answers (Table 1).

As the APS questionnaire has two similar question blocks with seven questions, each concerning responsibility in the case of nascent entrepreneurs and owner-managers, we merged answers to each pair of questions (see Table 2 and Table 3).

We elaborated relevant SDGs based on theUnited Nations (2015), where people, planet and profit were considered as social, environmental and economic pillars of sustainable development, respectively.

Table 1. Number of entrepreneurs in the sample

Year	Total early-stage entrepreneurial activity (TEA)	Established business owner (EB)	Total
2021	174	162	**336**
2022	186	138	**324**
2023	168	136	**304**
Total	**528**	**436**	**964**

Source: own study, based on GEM definitions.

Table 2. Variable descriptions

Variable label	Variable name	SDG goal*	Description
Social implications	SDG_soc	1-5	The entrepreneur considers social implications when making decisions about the future of their business
Environmental implications	SDG_env	6, 12-15	The entrepreneur considers environmental implications when making decisions about the future of their business
Steps to minimise environmental impact	SDG_steps1	6, 12-15	The entrepreneur has taken any steps to minimise the environmental impact of their business over the past year
Steps to maximise social impact	SDG_steps2	1-5	The entrepreneur has taken any steps to maximise the social impact of their business over the past year

Source: own study, based on GEM definitions.

Table 3. Other attributes involved in the classification model

Variable context label	Variable name	Measurement	Description / GEM question
Demographics	gender	nominal	What is your gender?
	age	Scale	What is your current age in years?
	HUreduc	ordinal	From Primary school (1) to Phd (10)
Attitudes	creativ	nominal, 3-point Likert	Other people think you are highly innovative.
	vision	Nominal, 5-point Likert	Every decision you make is part of your long-term career plan.
	consMOT2	Nominal, 5-point Likert	To build great wealth or a very high income
Business	consMKSC	Nominal	Market scope
	consCPTECH2	scale	Do you expect your business will use more digital technologies to sell your product or service in the next six months?

Source: own study, based on GEM definitions.

Based on our previous investigation of GEM data (PCA, statistical methods, correlation, data distribution) we could involve the data from Table 2 and Table 3 in the classification models. We then examined in detail the statistical parameters of the data (Table 4), their distribution (Boxplot diagrams and Density diagrams), and the correlation between them, and also ruled out multicollinearity (Variable Inflation Factor) to ensure that the models performed well.

To select suitable data for classification models in Python, researchers can employ various key strategies based on insights from research studies. Feature selection is a crucial step in data preparation for classification models, involving the removal of irrelevant or redundant features to enhance classification accuracy (Lee *et al*., 2015). Feature selection methods encompass filter, wrapper, and embedded techniques, which are fundamental in data mining and pattern recognition tasks (Chen *et al*., 2020). These methods aid in selecting the most pertinent features from the dataset, thereby boosting the performance of classification models (Peng & Liu, 2018).

Furthermore, the selection of characteristic variables is vital for developing effective classification models (Jin *et al*., 2021). By selecting the appropriate set of features, the model's accuracy can improve

significantly. Moreover, utilizing ensemble methods for feature selection can further enhance classification accuracy (Singh & Singh, 2021). Hybrid approaches that integrate different feature selection techniques can be particularly effective in improving data quality for classification tasks (Chanu *et al*., 2022).

Moreover, dimensionality reduction techniques like principal component analysis (PCA) can serve to clean noisy data and enhance the performance of artificial neural networks in classification tasks (Adolfo *et al*., 2021). Clustering techniques can also serve to improve classification accuracy by organising data into more manageable groups (Mathivanan *et al*., 2018). Furthermore, scholars have developed mutual information-based feature selection methods to identify relevant features for data classification (Bhuyan & Kamila, 2015).

For the classification model, we split the data into data and test sets in 2/3 and 1/3 proportions. In research, it is essential that the sampling is replicable, so we started the random sampling from random=12345 seed.

Several supervised classification methods (Figure 1) have been tested and are described in more detail (accuracy, confusion matrix, parameters in fitting methods, feature weights in prediction models) in the results section. Noteworthy, among the methods tested, those with the possibility to explore the built-in decision model are discussed in more explicit detail in the results section. As variables involved in the modelling require further analysis, we conduct (1) logistic regression, (2) support vector machine, (3) decision tree classifier and (4) gradient boosting classifier methods.

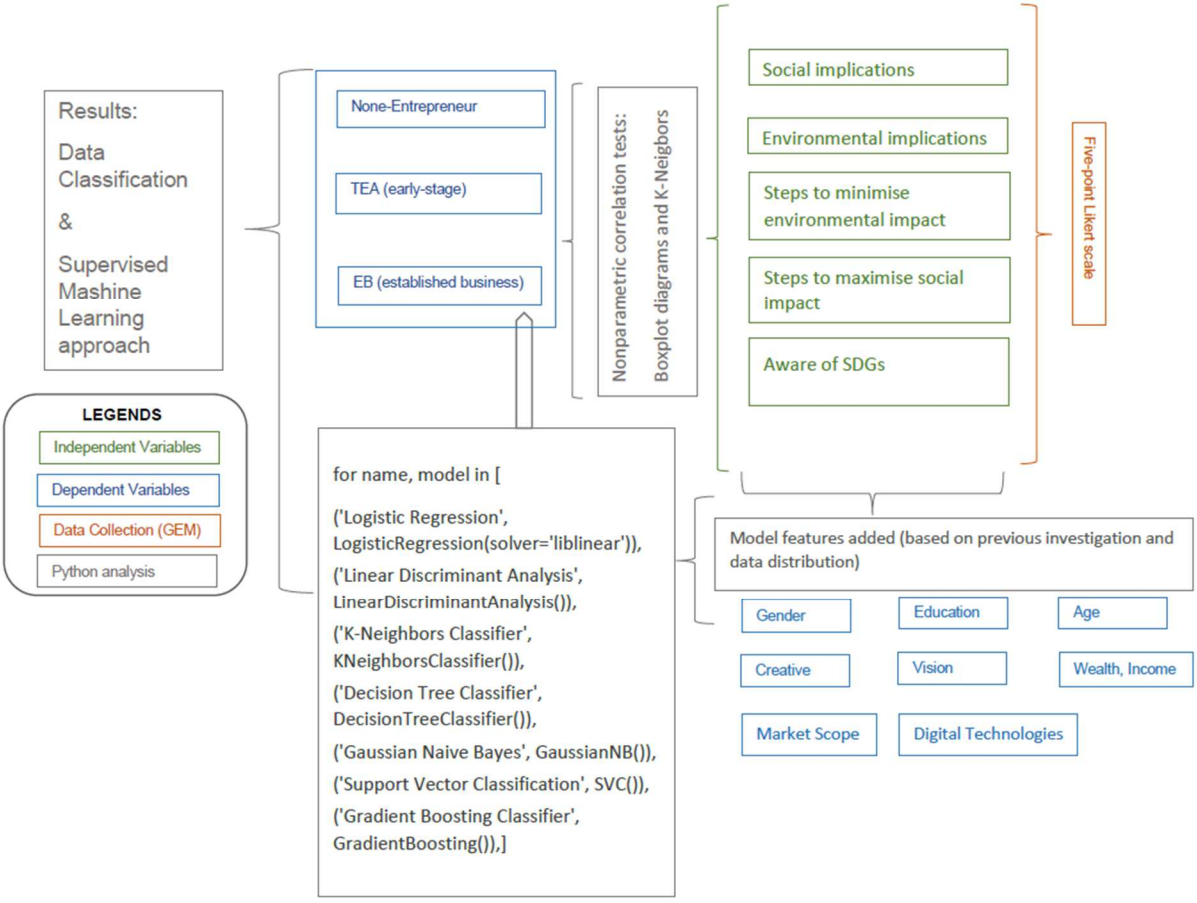

Figure 1. Research methodology with ML approach
Source: own elaboration.

RESULTS AND DISCUSSION

The gender, age, and educational attainment variables and SDG indicators had already been investigated in the GEM database (Krankovits *et al*., 2023). The vast majority (74.9%) of entrepreneurs are not aware of SDGs, but among them, it is rather likely (72.4%) that the entrepreneur identified any of the goals which are a priority for their business and defined a set of clear objectives, actions, and key performance indicators.

Table 4. Other attributes involved in the classification model

Statistics	gender	age	creativ	vision	HUreduc	consMOT2	consMKSC	consCPTECH2	SDG_soc	SDG_env	SDG_steps1	SDG_steps2	CONS_BUSO
mean	1.3724	42.5975	3.0311	3.8869	5.6919	3.1027	3.0685	1.7189	3.7272	4.1525	1.3641	1.6276	1.4523
std	0.4837	11.7023	1.8688	1.3344	2.3618	1.4742	1.1436	0.6902	1.3760	1.2324	0.4814	0.4837	0.4980
min	1.0000	18.0000	-2.0000	-2.0000	1.0000	1.0000	1.0000	-2.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.25	1.0000	33.0000	2.0000	3.0000	4.0000	2.0000	2.0000	1.0000	3.0000	4.0000	1.0000	1.0000	1.0000
0.50	1.0000	43.0000	3.0000	4.0000	5.0000	3.0000	4.0000	2.0000	4.0000	5.0000	1.0000	2.0000	1.0000
0.75	2.0000	52.0000	5.0000	5.0000	8.0000	5.0000	4.0000	2.0000	5.0000	5.0000	2.0000	2.0000	2.0000
max	2.0000	64.0000	5.0000	5.0000	10.0000	5.0000	4.0000	3.0000	5.0000	5.0000	2.0000	2.0000	2.0000
median	1.0000	43.0000	3.0000	4.0000	5.0000	3.0000	4.0000	2.0000	4.0000	5.0000	1.0000	2.0000	1.0000
iqr	1.0000	19.0000	3.0000	2.0000	4.0000	3.0000	2.0000	1.0000	2.0000	1.0000	1.0000	1.0000	1.0000
skew	0.5287	-0.0201	-0.9321	-1.2061	0.2430	-0.0470	-0.6602	-1.6157	-0.8385	-1.4330	0.5657	-0.5287	0.1920
kurtosis	-1.7241	-1.0407	-0.0604	0.8711	-1.4032	-1.4054	-1.1765	5.8343	-0.5512	0.9247	-1.6835	-1.7241	-1.9672

Source: own study.

Table 5. Correlation matrix between attributes

Variables	gender	age	creativ	vision	HUreduc	consMOT2	consMKSC	consCPTECH2	SDG_soc	SDG_env	SDG_steps1	SDG_steps2	CONS_BUSO
gender	1.0000	0.0159	-0.0220	-0.0457	0.0824	-0.0348	0.0158	0.0278	-0.0188	-0.0327	-0.0255	0.0075	-0.0533
age	0.0159	1.0000	-0.0476	-0.1801	0.0313	-0.1690	-0.3933	0.0855	-0.0372	-0.0216	-0.0600	0.0584	0.4149
creativ	-0.0220	-0.0476	1.0000	0.1651	0.0321	0.0934	0.0131	-0.0480	0.0102	0.0214	-0.1176	-0.1101	-0.0230
vision	-0.0457	-0.1801	0.1651	1.0000	-0.1000	0.1046	0.0779	0.0060	0.1189	0.0762	-0.0635	-0.1200	-0.1074
HUreduc	0.0824	0.0313	0.0321	-0.1000	1.0000	-0.0300	0.0097	-0.0857	-0.0131	-0.0630	-0.0483	-0.0315	0.0374
consMOT2	-0.0348	-0.1690	0.0934	0.1046	-0.0300	1.0000	0.1239	-0.0788	0.1746	0.1194	-0.0308	-0.0949	-0.1185
consMKSC	0.0158	-0.3933	0.0131	0.0779	0.0097	0.1239	1.0000	-0.0821	0.1392	0.0884	0.0924	-0.0440	-0.8968
consCPTECH2	0.0278	0.0855	-0.0480	0.0060	-0.0857	-0.0788	-0.0821	1.0000	-0.1563	-0.0301	-0.0135	0.0967	0.0803
SDG_soc	-0.0188	-0.0372	0.0102	0.1189	-0.0131	0.1746	0.1392	-0.1563	1.0000	0.4600	-0.2261	-0.2574	-0.1395
SDG_env	-0.0327	-0.0216	0.0214	0.0762	-0.0630	0.1194	0.0884	-0.0301	0.4600	1.0000	-0.3282	-0.1990	-0.0702
SDG_steps1	-0.0255	-0.0600	-0.1176	-0.0635	-0.0483	-0.0308	0.0924	-0.0135	-0.2261	-0.3282	1.0000	0.3198	-0.1115
SDG_steps2	0.0075	0.0584	-0.1101	-0.1200	-0.0315	-0.0949	-0.0440	0.0967	-0.2574	-0.1990	0.3198	1.0000	0.0318
CONS_BUSO	-0.0533	0.4149	-0.0230	-0.1074	0.0374	-0.1185	-0.8968	0.0803	-0.1395	-0.0702	-0.1115	0.0318	1.0000

Source: own study.

Distributions of TEA and EB were homogenous (Pearson correlation with p>=0.05) in the case of age and education. We may explain these results by the fact that entrepreneurs are generally older and have higher education than the total population, and in addition, they are mostly male (Csákné Filep *et al*., 2023). In our analysis, we fitted the variables to the machine learning model using statistical distributions (Table 4).

After the descriptive statistics (Table 4) and correlation matrix (Table 5), we plot the distribution for each variable on density diagrams and boxplot diagrams (Figure 2) for the target variable (TEA or EB).

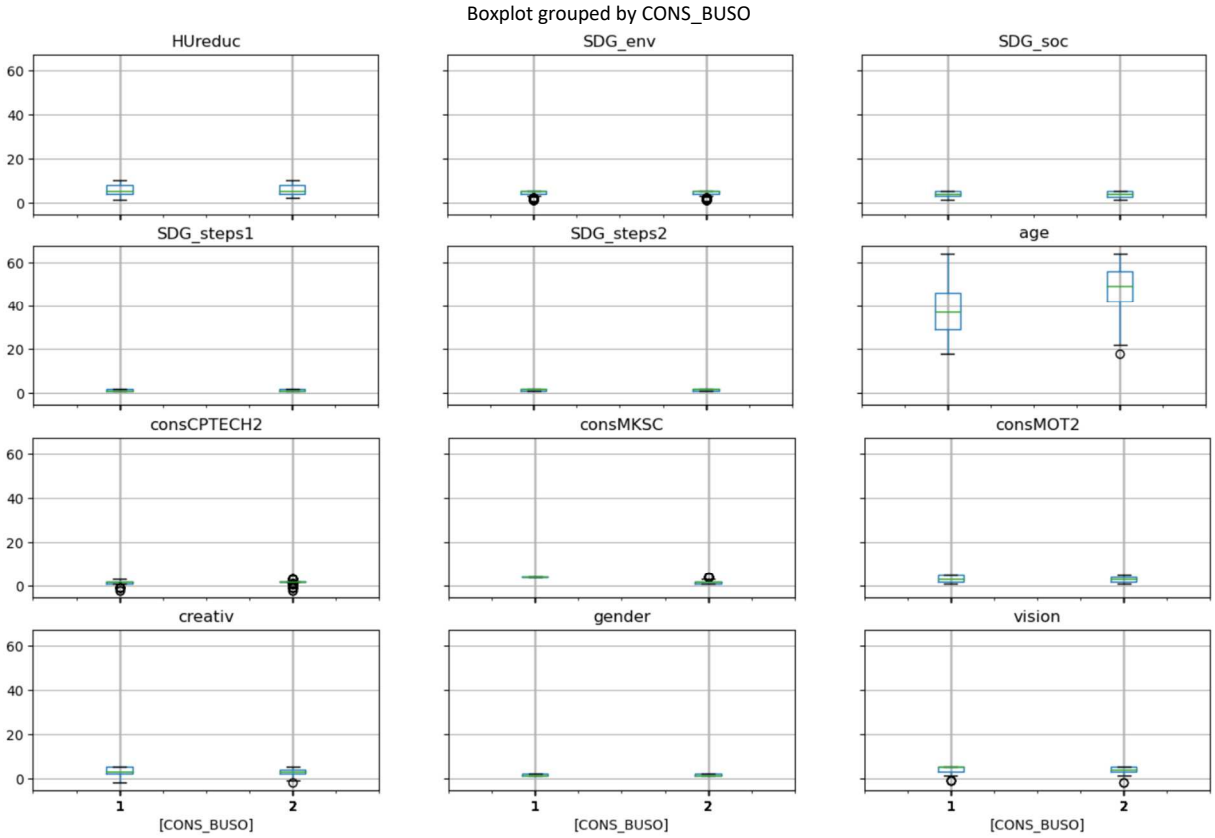

Figure 2. Boxplot diagrams for predictor variables
Source: own elaboration.

A density diagram provides a convenient way to explore the relationships between multiple variables in our dataset, making it easier to identify patterns, correlations, and potential outliers (Figure 3 and Figure 4).

The variance inflation factor (VIF) is a well-established metric for quantifying multicollinearity, a potential issue in regression models and other statistical analyses. Multicollinearity can lead to the distortion of estimated parameters and a reduction in the predictive accuracy of models (Table 6).

The elevated VIF value of consMKSC signifies its capacity to exhibit a robust linear relationship with other predictors, including the target variable itself (CONS_BUSO). Nevertheless, this observation does not negate its potential as a significant predictor. The presence of a substantial relationship between a target variable and a predictor is an anticipated feature of a robust model.

These results are only a suggestion for determining which variables are likely to play an important predictive role in the following models. The main diagonal clearly shows the distributions that can be used for classification (*i.e.*, age, gender, education), but the role of a variable may be important even if its distribution alone does not show encouraging signs. In the total sample, the average age of TEA entrepreneurs was 38.19 years and that of EB entrepreneurs was 47.98 years. Among TEA entrepreneurs, there were just under 4% more males (33.16%) than EB entrepreneurs (29.75%).

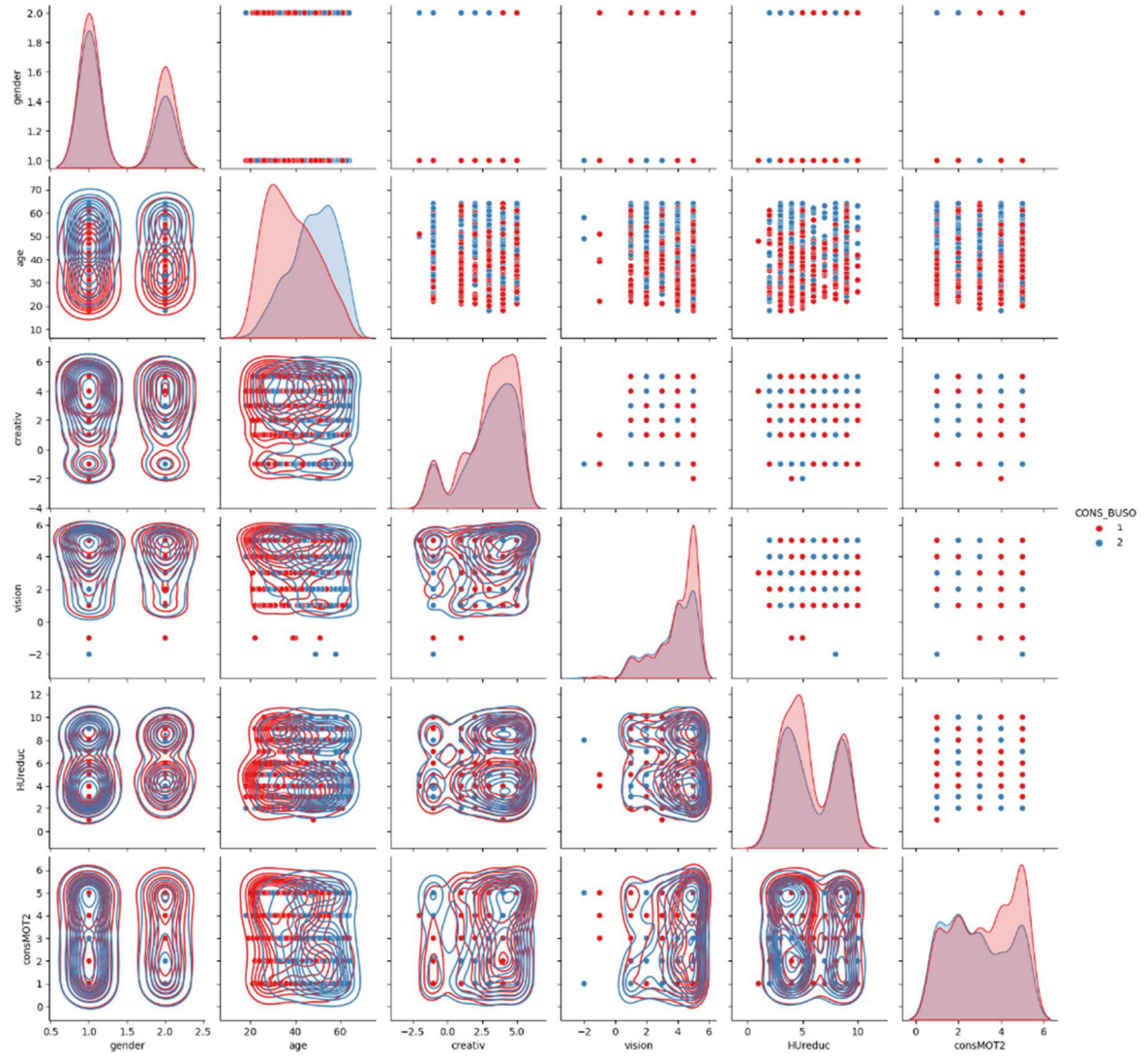

Figure 3. Density diagrams for predictor variables (1)
Source: own elaboration.

Table 6. Multicollinearity (VIF) of the parameters

Variable	VIF
gender	1.0249
age	1.2703
creativ	1.0588
vision	1.1002
HUreduc	1.0485
consMOT2	1.0787
consMKSC	4.2506
consCPTECH2	1.0562
SDG_soc	1.3971
SDG_env	1.3817
SDG_steps1	1.2708
SDG_steps2	1.1881
CONS_BUSO	5.4039

Source: own study.

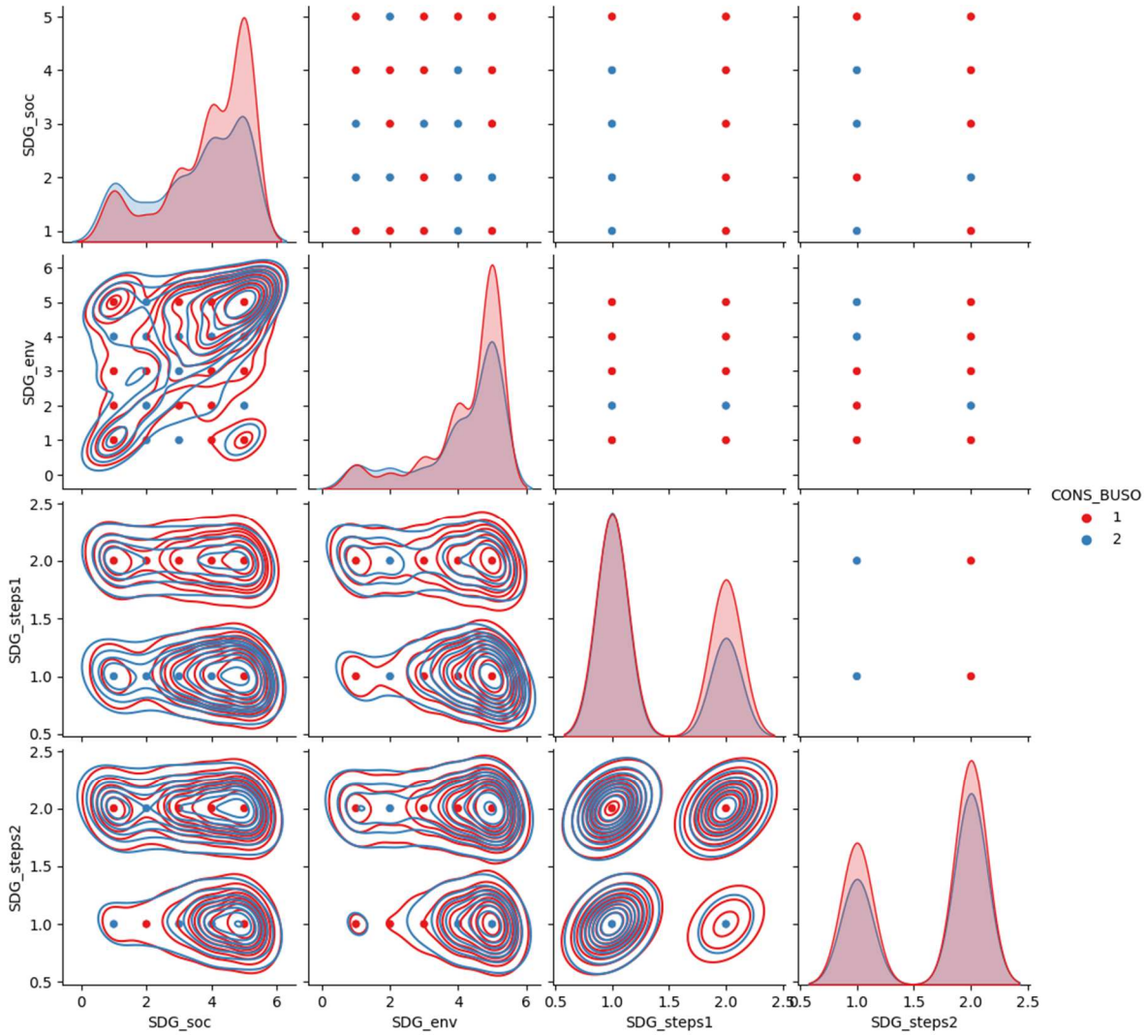

Figure 4. Density diagrams for predictor variables (2)
Source: own elaboration.

Results of Classification Algorithms

Below, we will briefly present the applicability of the applied algorithms in the enterprise data environment, as well as the implementation of the algorithms on GEM data and the results obtained.

As discussed in the method section, we tested a total of seven basic classification methods on the GEM data:

- Nearest neighbour (KNN);
- Linear discriminant analysis (LDA);
- Gaussian naive bayes (GNB);
- Logistic regression (LG);
- Support vector machine (SVM);
- Decision tree classifier (DTC);
- Gradient boosting classifier (GBC).

These supervised learning methods fit well with our selection database, as we had a training set with predefined and controlled outputs for TEA and EB enterprises.

K-nearest neighbour (KNN) classification is a widely used method in entrepreneurship databases due to its simplicity and effectiveness in predicting outcomes based on similar cases. Studies have ap-

plied KNN in various contexts within entrepreneurship, such as road risk assessment for accident prediction (Saranyadevi *et al*., 2019), predicting the penetration rate of tunnel boring machines (Xu *et al*., 2019), and analysing startup trends (Savin *et al*., 2023). The parameter settings configure the KNN classifier to make predictions based on the 5 nearest neighbours, with closer neighbours having more influence, using a KD-tree data structure for efficient search, and using the Euclidean distance metric. With default settings, the classification accuracy of the KNN algorithm was 84.13%, which has been improved to 85.71% with the changed parameters.

Linear discriminant analysis (LDA) is commonly used in business databases for classification tasks. It can be combined with feature selection methods like principal component analysis (PCA) and variable selection algorithms such as genetic algorithm (GA) to separate groups or samples within the database (Alves *et al*., 2023).

The LDA statistical model employs Bayes' theorem to achieve linear separation between classes. The efficacy of the model is optimised when utilising data that is free from contamination and exhibits a normal distribution, with equal covariance. While the model demonstrated a commendable performance in the normal setting, with an accuracy of 94.61%, further investigation was deemed unnecessary due to previous statistical analyses that demonstrated the absence of novel insights derived from linear correlations between data in the context of GEM data.

The Gauss-Naive Bayes (GNB) model is a probabilistic model also based on Bayes' theorem, assuming a normal distribution of characteristics. The characteristics are assumed to be independent of each other, a simplistic assumption that renders the model well-suited to simpler problems where the characteristics are nearly independent. The GNB model also performs well (98.86%), with only a fine-tuning parameter that can be used for variance smoothing. The Gaussian Naive Bayes (GNB) model does not utilise explicit weights, in contrast to linear models (*e.g.*, logistic regression), as GNB calculates probabilities based on the independence of the characteristics (naive assumption). Consequently, the GNB model does not provide direct variable weights. We did not investigate the first three methods in greater depth due to the article's length.

Table 7 shows the hyperparameter tuning for the models.

Table 7. Summarized accuracy and hyperparameters by classification models

Classification model	Default settings accuracy	Hyperparameter	Parameterized accuracy
Logistic Regression (LG)	97.62%	penalty='l2', C=10, solver='liblinear', max_iter=100, random_state=42	98.13%
Linear Discriminant Analysis (LDA)	94.61%		
K-Neighbors Classifier (KNN)	84.13%	n_neighbors=5, weights='distance', algorithm='kd_tree', leaf_size=30, p=2, metric='minkowski'	85.71%
Gaussian Naive Bayes (GNB)	98.86%		
Support Vector Classification (SVC)	91.39%	kernel='linear', C=10, gamma='scale', degree=3	98.44%
Decision Tree Classifier (DTC)	97.61%		
Gradient Boosting Classifier (GBC)	98.76%	n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42	98.44%

Source: own study.

In the following part, we focus on the LG, SVC, DTC and GBC models, with the results presented in detail.

Logistic regression is a commonly used statistical method in entrepreneurship databases for classification purposes. It has been utilized in various studies to analyse factors influencing entrepreneurial activities (Urbano *et al*., 2013), predict business takeover intentions (Joensuu-Salo *et al*., 2021), and assess the likelihood of youth entrepreneurship (Damoah, 2020). Logistic regression has also been applied in healthcare settings to predict physical function upon discharge of older adults (Chu *et al*., 2023). Furthermore, it has been employed in research focusing on social entrepreneurship

Kachlami *et al*. (2017) and entrepreneurial attitudes Puga and García (2012). The method's efficiency in handling binary classification tasks makes it a valuable tool for understanding and predicting entrepreneurial behaviours and outcomes. The flexibility and interpretability of logistic regression make it a popular choice for analysing complex relationships within entrepreneurship databases.

With default settings, the classification accuracy of the logistic regression algorithm was 97.62%, which improved to 98.13% with the changed parameters. We used the logistic regression with parameter C=10 and a liblinear solver. Setting `C=10` instead of `C=1.0` (default settings) means reducing the regularization strength, which may lead to a more flexible model that fits the training data more closely. However, it also increases the risk of overfitting, especially if the model becomes too complex for the given dataset.

The LG algorithm is one of the few methods thanks to which we can look behind the model and determine the weight of the variables involved in the model.

Some key variables received more weight in the model, such as gender (0.6081), vision (0.3433), motivation (0.289), market scope (6.965), digital technology skills (0.3768), and SDG variables(Figure 5). We may explain this latter finding with the fact that it is much easier to find low or no-cost measures (*e.g.*, recycling, using more efficient devices, *etc*.) and even grants to implement investments reducing energy usage when a business fosters environmental friendliness, but it is not the case when attempting to maximise social impact (*e.g.*, employing disadvantaged people or even women with small children). The vast majority (74.9%) of entrepreneurs are not aware of SDGs, but among them, it is rather likely (72.4%) that the entrepreneur identified any of the goals which are a priority for their business and defined a set of clear objectives, actions, and key performance indicators. The model constructed the weights from this correlated SDG construct to maximize explanatory power.

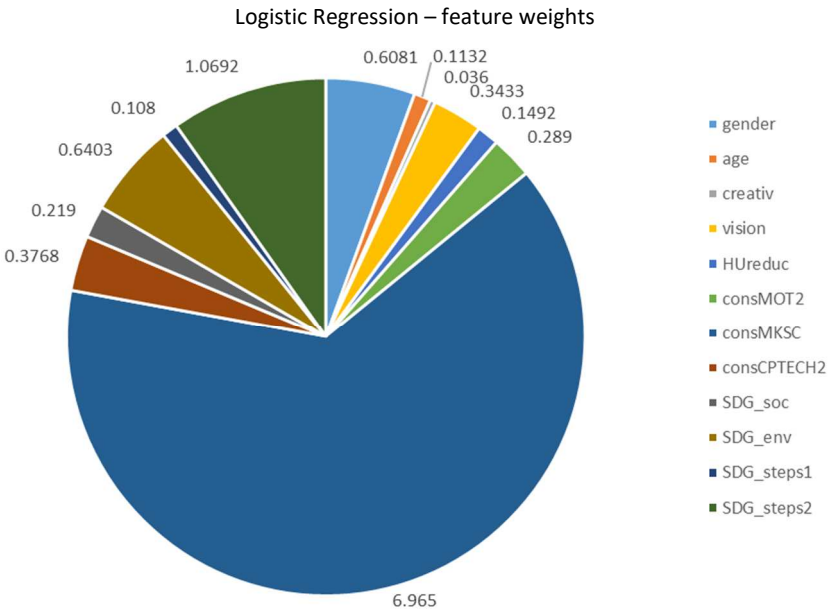

Figure 5. Feature weights in the logistic regression model
Source: own elaboration.

Support vector machine (SVM) classifiers have found significant utility in various business applications, particularly in database management. SVM, known for its ability to effectively categorize and analyse data, has been employed in diverse fields such as healthcare, finance, and e-commerce (Jameel, 2023). For instance, scholars have utilised SVM in the classification of heart rate variability for medical diagnostics (Ashtiyani *et al*., 2018), air quality monitoring systems (Sattar *et al*., 2019), and even in predicting hyperlipidemia (Lakshmi *et al*., 2018). The robustness of SVM lies in its optimization-based approach and its capability to incorporate various kernel methods, making it a versatile tool for businesses to extract valuable insights from their databases (Tuncer *et al*., 2020). One can achieve greater accuracy with the following parameters instead of the default settings.

The Gaussian kernel maps data into an infinite-dimensional space. The 'gamma' parameter affects the shape of the decision boundary. For 'rbf,' gamma determines how much influence a single training example has. Higher gamma values make the decision boundary more flexible and can lead to overfitting.

With default settings, the classification accuracy of the SVC algorithm was 89.13%, which has been improved to 98.44% with the changed parameters.

The SVM did not take gender and education variables into account in the model construction, but market scope and SDG variables were included in the decision, even if to a much lesser extent.

Decision tree classifiers, such as J48, are commonly utilized in entrepreneurship databases for their simplicity, interpretability, and performance in supervised learning (Cañete-Sifuentes *et al*., 2021; Obeidat *et al*., 2019).

These classifiers are valuable for understanding patterns within entrepreneurial datasets and providing clear insights essential for decision-making (Cañete-Sifuentes *et al*., 2021). Decision tree classifiers have demonstrated success in various domains including healthcare for disease classification, financial risk assessment, and consumer behaviour prediction in e-commerce applications (Idris & Ismail, 2021; Sharma & Sharma, 2019). Their adaptability and effectiveness make them a valuable asset in entrepreneurship databases for tasks like customer analysis and predicting loan defaults (Akanmu & Gilal, 2019; Subramanian *et al*., 2021).

The DTC performed well with the default settings (97.61%), but its real value comes from the fact that the decision tree model can be generalized, making the variables visible (Figure 6).

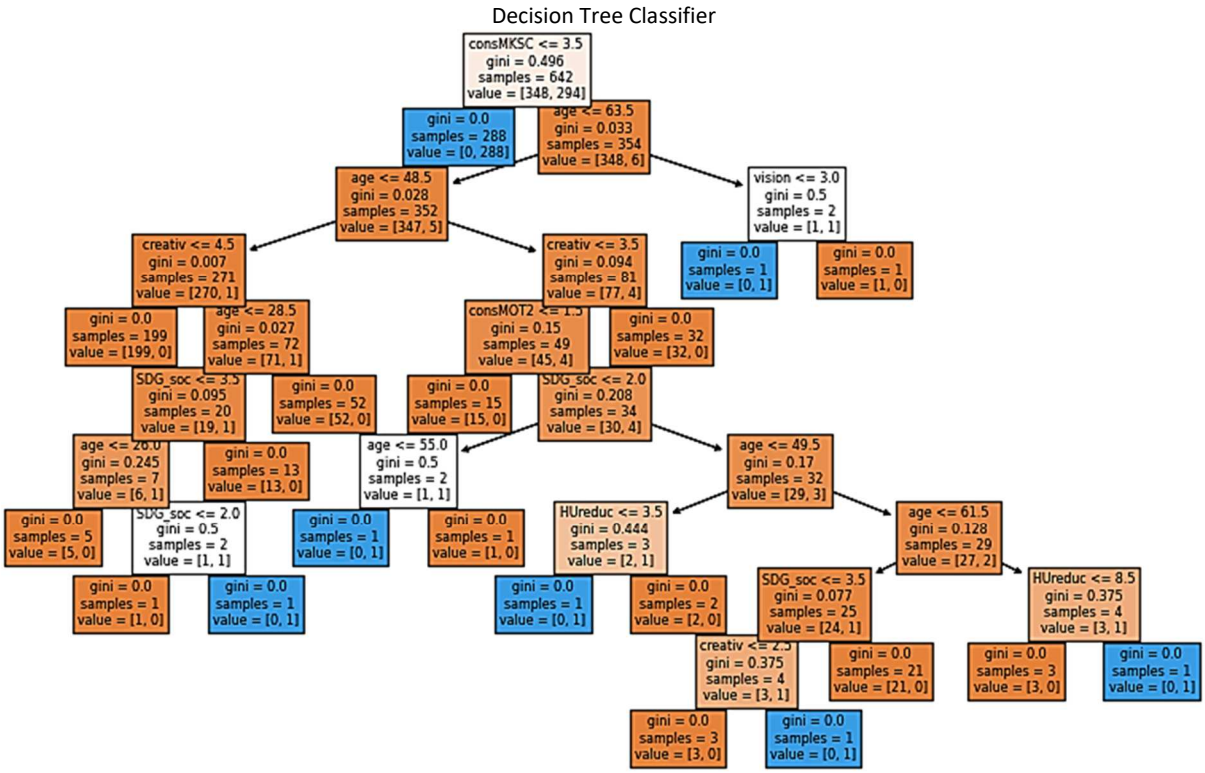

Figure 6. Feature roles in the DTC model
Source: own elaboration.

Gini impurity measures how impure the information in a node is. It helps determine which questions to ask at each node to classify categories effectively. The goal is to minimize Gini Impurity during tree construction.

The complexity of the decision tree is illustrated by the fact that even at the lower levels, the relationship with SGD variables, education and creativity are repeatedly mentioned. The left end of the decision tree is for more creative and younger entrepreneurs, the next branch on the left is for

less creative but more financially stable entrepreneurs. The top right branch is for older entrepreneurs with less SGD awareness but with career plans.

Gradient Boosting Classifier, a member of the Classification and Regression Trees (CART) family, is recognized for its ability to handle complex datasets and enhance prediction accuracy by iteratively reducing errors (Georganos *et al*., 2018). Recent advancements in Gradient Boosting have led to the development of Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, further improving the algorithm's efficiency and scalability (Mienye & Sun, 2022). Businesses are increasingly utilizing Gradient Boosting for diverse tasks such as urban mapping, soil erosion prediction, and healthcare risk prediction models, highlighting its versatility and effectiveness across various domains (Jozdani *et al*., 2019; Patel *et al*., 2023; Wang *et al*., 2023). The adaptability of gradient boosting to ensemble learning approaches and its robustness in handling imbalanced data make it a valuable asset for businesses seeking precise and reliable insights from their databases (Malek *et al*., 2023; Muhathir *et al*., 2023).

Like DTC, gradient boosting also uses decision trees for classification, but here there are several trees nested under each other. Fortunately, the weights of the variables in the model can be determined and displayed here.

Compared to LG, the weight of the other variables are orders of magnitude smaller (Figure 7), but the age variable is also prominent, as well as the SDG_soc-SDG step 2 pair.

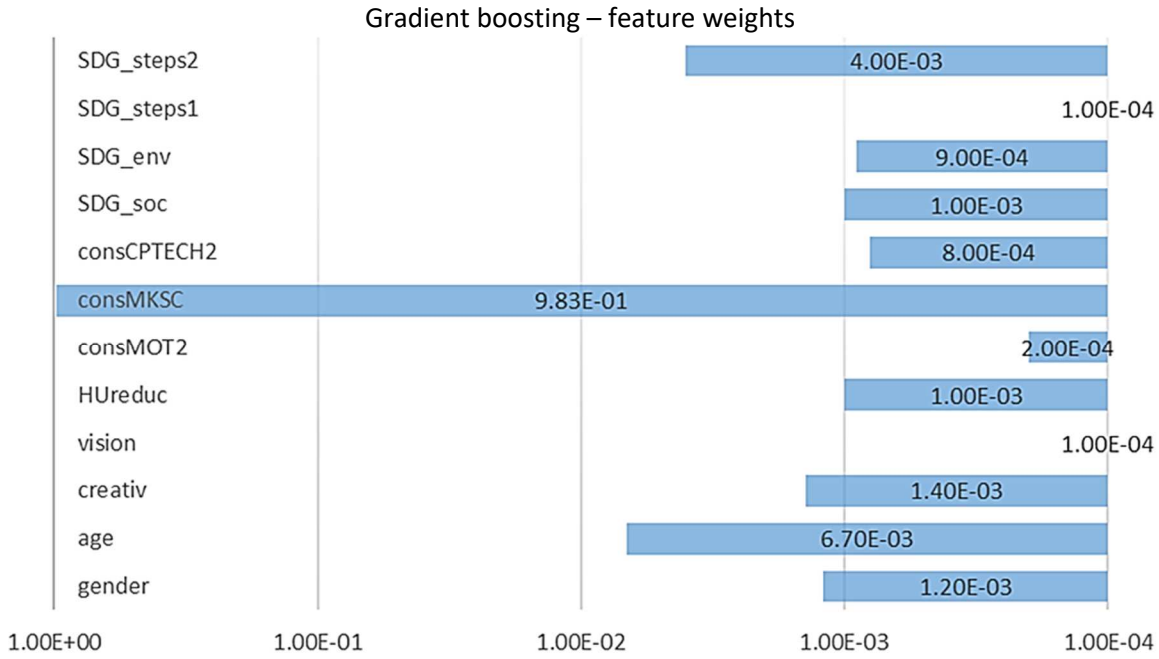

Figure 7. Feature weights in the gradient-boosting model
Source: own elaboration.

It is insufficient to employ merely the accuracy value for model comparison; specific values such as specificity (proportion of actual negatives correctly identified), sensitivity (proportion of actual positives correctly identified), precision (proportion of correct positive predictions), the F1-score (harmonic mean of precision and recall, balancing their trade-off), error rate (proportion of incorrect predictions)and Cohen's kappa (measures inter-annotator agreement, adjusting for chance agreement) must also be taken into consideration (Table 7). Table 9 presents the confusion matrix of the models.

For both hypotheses, it was important to apply new machine learning methods to the GEM database that would allow exploring deeper relationships in the data than with traditional statistical methods. The analysis used supervised learning on a small sample (964 items), which is why the inclusion of additional annual data, and the prior identification of more potential variables would be necessary for a better understanding. With a more thorough methodological analysis, the DTC or Gradient methods

can be further explored and if not the more accurate result (because it is above 98%), the goal may be to understand why these variables are the determinants.

Table 8. Accuracy and metrics for the models

Classification model	Accuracy	Sensitivity	Specificity	Precision	F1-score	Error Rate	Cohen's Kappa
Logistic Regression (LG)	0.9813	1.0000	1.0000	0.9677	0.9836	0.0186	0.9620
Support Vector Classification (SVC)	0.9844	1.0000	1.0000	0.9729	0.9863	0.0155	0.9683
Decision Tree Classifier (DTC)	0.9813	0.9944	0.9944	0.9728	0.9835	0.0186	0.9620
Gradient Boosting Classifier (GBC)	0.9844	1.0000	1.0000	0.9729	0.9863	0.0155	0.9683

Source: own study.

Table 9. Confusion matrix

Models	Actual	Predicted		Recall	Error rate
		C1	C2		
LG	C1	180	0	1.00	0.00
	C2	6	136	0.96	0.04
	Precision	0.96	1	NaN	NaN
SVC	C1	180	0	1.00	0.00
	C2	5	137	0.96	0.03
	Precision	0.97	1	NaN	NaN
DTC	C1	179	1	0.99	0.01
	C2	5	137	0.96	0.03
	Precision	0.97	0.99	NaN	NaN
GBC	C1	180	0	1.00	0.00
	C2	5	137	0.96	0.03
	Precision	0.97	1	NaN	NaN

Source: own study.

We subjected hypothesis H1 to empirical testing using supervised machine learning (ML) classification algorithms. During the model construction process, the algorithms demonstrated an exceptional capacity for identifying variables that can effectively determine the status of enterprises (TEA or EB), exhibiting a remarkable level of accuracy of 98%. As hypothesis H1 was accepted, we can state that the characteristics of early-stage and established enterprises are different. Although variables in the final models may differ in each method, the models provide impressive explanatory power, which means that we may explain the distinction between the two entrepreneurial phases with different variable sets. This finding suggests that entrepreneurs are evolving over time and thus, incentives and policies should also reflect these differences. However, as the weights identified are vehicles of purely mathematic modelling, explaining their exact meaning needs further research.

We could partially confirm hypothesis H2, as we can determine the entrepreneurial phase with above 90% accuracy with only six out of the seven methods tested. However, the KNN method also provides fairly good accuracy as its parameterized accuracy also lies at 85.71%. Conventional methods failed to determine the characteristics of entrepreneurs, so this finding has implications primarily for researchers. First, the results of machine learning techniques are encouraging even in the case of rather small datasets (n=964) in classifying entrepreneurs based on their attributes. Second, one can replicate modelling using the data of another country or even countries.

CONCLUSIONS

In Central and Eastern European countries, such as Hungary, the establishment of businesses faced substantial constraints during the decades of socialism. Consequently, this region experiences a nota-

ble lack of entrepreneurial experience as well as academic and policy-related knowledge about businesses. This underscores the importance of research aimed at enhancing our understanding of the region's businesses and providing a foundation for comprehending their life cycles and behaviours.

We tested seven methods and identified those that performed well on the Global Entrepreneurship Monitor (GEM) data. Specifically, four methods – logistic regression (LG), support vector machine (SVM), decision tree classifier (DTC), and gradient boosting classifier (GBC) – were highlighted, with the variables used in the models explicitly defined. These methods present opportunities for further refinement and testing on larger samples. The practical significance of our work lies in the confirmation of a deeper relationship in the data beyond statistical correlations, offering concrete insights into these patterns.

The machine learning aspect of the research demonstrates the capability to classify businesses as early-stage entrepreneurs (TEA) or established businesses (EB) based on the examined data. Importantly, we employed supervised learning methods, achieving an accuracy exceeding 98% in distinguishing between TEA and EB entrepreneurs using the training data in the GEM database. The research is replicable, as the process for separating the test and training datasets has been clearly outlined.

Our findings show that the characteristics of early-stage and established businesses differ, and through the application of machine learning methods, it is possible to determine the category to which a business belongs.

This study also identifies several promising directions for future research. Firstly, the application of machine learning techniques to uncover deeper patterns across countries holds considerable potential for gaining a nuanced understanding of entrepreneurship in Central and Eastern Europe. Such techniques could reveal latent trends and interconnections that shape the region's entrepreneurial landscape. Furthermore, analysing additional datasets could strengthen the robustness of the current findings, offering a more comprehensive perspective on entrepreneurial behaviour. Extending the analysis to other countries would provide comparative insights, broadening our understanding of how contextual factors influence entrepreneurship across diverse regions. These future research directions could significantly advance knowledge in the field and support evidence-based policymaking and practice.

A key limitation of this research is that it relies on data collected through a pre-designed questionnaire, the content of which could not be modified by the researchers. This limitation restricts the ability to incorporate additional criteria for distinguishing businesses, as highlighted in the existing literature.

The findings of this research have practical implications for enterprise development professionals, as they demonstrate that one can effectively achieve business classification and categorization using machine learning methodologies. In Hungary's entrepreneurial ecosystem, which predominantly consists of micro, small, and medium-sized enterprises, segmentation is critical for providing targeted support. Segmentation enables policymakers to identify and prioritize specific groups for support and, based on the unique characteristics of these groups, to implement tailored legislative changes and support programs designed to meet their specific needs.

REFERENCES

Acs, Z. (2006). How Is Entrepreneurship Good for Economic Growth?. *Innovations: Technology, Governance, Globalization*, *1*(1), 97-107. https://doi.org/10.1162/itgg.2006.1.1.97

Adolfo, C.M.S., Chizari, H., Win, T.Y., & Al-Majeed, S. (2021). Sample Reduction for Physiological Data Analysis Using Principal Component Analysis in Artificial Neural Network. *Applied Sciences*, *11*(17), Article 17. https://doi.org/10.3390/app11178240

Ahn, K., & Winters, J.V. (2023). Does education enhance entrepreneurship?. *Small Business Economics*, *61*(2), 717-743. https://doi.org/10.1007/s11187-022-00701-x

Akanmu, S.A., & Gilal, A.R. (2019). A Boosted Decision Tree Model for Predicting Loan Default in P2P Lending Communities. *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*(1), 1257-1261. https://doi.org/10.35940/ijeat.a9626.109119

Alves, M.V.S., Maciel, L.I.L., Passos, J.O.S., Morais, C.L.M., dos Santos, M.C.D., Lima, L.A.S., Vaz, B.G., Pegado, R., & Lima, K.M.G. (2023). Spectrochemical approach combined with symptoms data to diagnose fibromyalgia

through paper spray ionization mass spectrometry (PSI-MS) and multivariate classification. *Scientific Reports*, *13*(1), 4658. https://doi.org/10.1038/s41598-023-31565-0

Amit, R., MacCrimmon, K.R., Zietsma, C., & Oesch, J.M. (2001). Does money matter?: Wealth attainment as the motive for initiating growth-oriented technology ventures. *Journal of Business Venturing*, *16*(2), 119-143. https://doi.org/10.1016/S0883-9026(99)00044-0

Ashtiyani, M., Navaei Lavasani, S., Asgharzadeh Alvar, A., & Deevband, M.R. (2018). Heart Rate Variability Classification using Support Vector Machine and Genetic Algorithm. *Journal of Biomedical Physics and Engineering*, *8*(4), 423-434. https://doi.org/10.31661/jbpe.v0i0.614

Bhukya, D.P., & Ramachandram, S. (2010). Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. *International Journal of Computer and Electrical Engineering*, 660-665. https://doi.org/10.7763/IJCEE.2010.V2.208

Bhuyan, H.K., & Kamila, N.K. (2015). Privacy preserving sub-feature selection in distributed data mining. *Applied Soft Computing*, *36*, 552-569. https://doi.org/10.1016/j.asoc.2015.06.060

Cañete-Sifuentes, S., Monroy, R., & Medina-Pérez, M.A. (2021). A Review and Experimental Comparison of Multivariate Decision Trees. *IEEE Access9*, 110451-110479. https://doi.org/10.1109/ACCESS.2021.3102239

Celbiş, M.G. (2021). A machine learning approach to rural entrepreneurship. *Papers in Regional Science*, *100*(4), 1079-1105. https://doi.org/10.1111/pirs.12595

Chanu, U.S., Singh, K.J., & Chanu, Y.J. (2022). An ensemble method for feature selection and an integrated approach for mitigation of distributed denial of service attacks. *Concurrency and Computation: Practice and Experience*, *34*(13), e6919. https://doi.org/10.1002/cpe.6919

Chen, C.-W., Tsai, Y.-H., Chang, F.-R., & Lin, W.-C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, *37*(5), e12553. https://doi.org/10.1111/exsy.12553

Chen, Z. (2023). Medical Image Segmentation Based on U-Net. *Journal of Physics: Conference Series*, *2547*(1), 012010. https://doi.org/10.1088/1742-6596/2547/1/012010

Chu, W.-M., Tsan, Y.-T., Chen, P.-Y., Chen, C.-Y., Hao, M.-L., Chan, W.-C., Chen, H.-M., Hsu, P.-S., Lin, S.-Y., & Yang, C.-T. (2023). A model for predicting physical function upon discharge of hospitalized older adults in Taiwan—A machine learning approach based on both electronic health records and comprehensive geriatric assessment. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1160013

Chung, D. (2023). Machine learning for predictive model in entrepreneurship research: Predicting entrepreneurial action. *Small Enterprise Research*, *30*(1), 89-106. https://doi.org/10.1080/13215906.2022.2164606

Csákné Filep, J., Radácsi, L., Szennay, Á., & Timár, G. (2023). *Taking initiative and earning a living – Entrepreneurial motivations and opportunity perception in Hungary*. Budapesti Gazdasági Egyetem. Retrieved from https://budapestlab.hu/wpcontent/uploads/2023/08/GEM-BGE_beliv_2023_angol_webre.pdf on November 21, 2023.

Damoah, O.B.O. (2020). Strategic factors predicting the likelihood of youth entrepreneurship in Ghana: A logistic regression analysis. *World Journal of Entrepreneurship, Management and Sustainable Development*, *16*(4), 389-401. https://doi.org/10.1108/WJEMSD-06-2018-0057

Filser, M., & Eggers, F. (2014). Entrepreneurial orientation and firm performance: A comparative study of Austria, Liechtenstein and Switzerland. *South African Journal of Business Management*, *45*(1), Article 1.

GEM. (Global Entrepreneurship Monitor). (2022). *Global Entrepreneurship Monitor 2021/2022. Global Report: Opportunity Amid Disruption*. Retrieved from https://gemconsortium.org/report/gem-20212022-global-report-opportunity-amid-disruption on March 3, 2023.

GEM. (Global Entrepreneurship Monitor). (2023). *Global Entrepreneurship Monitor 2022/2023 Global Report: Adapting to a "New Normal"*. Retrieved from https://gemconsortium.org/file/open?fileId=51147 Retrieved on March 3, 2023.

GEM. (Global Entrepreneurship Monitor). (2024). *Global Entrepreneurship Monitor 2023/2024 Global Report: 25 Years and Growing*. Retrieved from https://www.gemconsortium.org/report/global-entrepreneurship-monitor-gem-20232024-global-report-25-years-and-growing Retrieved on March 3, 2023.

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., & Wolff, E. (2018). Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geoscience and Remote Sensing Letters*, *15*(4), 607-611. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2018.2803259

Idris, N.F., & Ismail, M.A. (2021). Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: Automatic fuzzy database definition. *PeerJ Computer Science*, *7*, e427. https://doi.org/10.7717/peerj-cs.427

Jameel, M.M. (2023). Enhancement of E-Banking System in Iraq by web application-based authentication system using face recognition. *Wasit Journal for Pure Sciences*, *2*(4), https://doi.org/10.31185/wjps.252

Jin, Y., Guo, J., Ye, H., Zhao, J., Huang, W., & Cui, B. (2021). Extraction of Arecanut Planting Distribution Based on the Feature Space Optimization of PlanetScope Imagery. *Agriculture*, *11*(4), https://doi.org/10.3390/agriculture11040371

Joensuu-Salo, S., Viljamaa, A., & Varamäki, E. (2021). Understanding Business Takeover Intentions—The Role of Theory of Planned Behavior and Entrepreneurship Competence. *Administrative Sciences*, *11*(3), https://doi.org/10.3390/admsci11030061

Jozdani, S.E., Johnson, B.A., & Chen, D. (2019). Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification. *Remote Sensing*, *11*(14). https://doi.org/10.3390/rs11141713

Kachlami, H., Yazdanfar, D., & Öhman, P. (2017). Regional demand and supply factors of social entrepreneurship. *International Journal of Entrepreneurial Behavior & Research*, *24*(3), 714-733. https://doi.org/10.1108/IJEBR-09-2016-0292

Kautonen, T., Down, S., & Minniti, M. (2014). Ageing and entrepreneurial preferences. *Small Business Economics*, *42*(3), 579-594. https://doi.org/10.1007/s11187-013-9489-5

Kelley, D., Singer, S., Herrington, M., & Entrepreneurship Research Association (GERA). (2016). *Global Entrepreneurship Monitor 2015/2016 Global Report*. Retrieved from https://www.gemconsortium.org/file/open?fileId=49480 on March 3, 2023.

Ključnikov, A., Civelek, M., Čech, P., & Kloudová, J. (2019). Entrepreneurial orientation of SMEs? Executives in the comparative perspective for Czechia and Turkey. *Oeconomia Copernicana*, *10*(4), https://doi.org/10.24136/oc.2019.035

Krankovits, M., Filep, J.C., & Szennay, Á. (2023). Factors of Responsible Entrepreneurial Behaviour: Empirical Findings from Hungary. *Chemical Engineering Transactions*, *107*, 25-30. https://doi.org/10.3303/CET23107005

Kurczewska, A., Doryń, W., & Wawrzyniak, D. (2020). An Everlasting Battle between Theoretical Knowledge and Practical Skills? The Joint Impact of Education and Professional Experience on Entrepreneurial Success. *Entrepreneurial Business and Economics Review*, *8*(2), https://doi.org/10.15678/EBER.2020.080212

Lafuente, E.M., & Vaillant, Y. (2013). Age driven influence of role-models on entrepreneurship in a transition economy. *Journal of Small Business and Enterprise Development*, *20*(1), 181-203. https://doi.org/10.1108/14626001311298475

Lakshmi, K.S., Vadivu, G., & Subramanian, S. (2018). Predicting hyperlipidemia using enhanced ensemble classifier. *International Journal of Engineering & Technology*, *7*(3), https://doi.org/10.14419/ijet.v7i3.10693

Lee, Y.-C., Hsiao, Y.-C., Peng, C.-F., Tsai, S.-B., Wu, C.-H., & Chen, Q. (2015). Using Mahalanobis–Taguchi system, logistic regression, and neural network method to evaluate purchasing audit quality. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *229*(1_suppl), 3-12. https://doi.org/10.1177/0954405414539934

Lévesque, M., & Minniti, M. (2011). Age matters: How demographics influence aggregate entrepreneurship. *Strategic Entrepreneurship Journal*, *5*(3), 269-284. https://doi.org/10.1002/sej.117

Malek, N.H.A., Yaacob, W.F.W., Wah, Y.B., Nasir, S.A.M., Shaadan, N., & Indratno, S.W. (2023). Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(1), https://doi.org/10.11591/ijeecs.v29.i1.pp598-608

Mathivanan, N.M.N., Md.Ghani, N.A., & Janor, R.M. (2018). Improving Classification Accuracy Using Clustering Technique. *Bulletin of Electrical Engineering and Informatics*, *7*(3). https://doi.org/10.11591/eei.v7i3.1272

Mienye, I.D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

Muhathir, M., Pangestu, R.T., Safira, I., & Melisah, M. (2023). Performance Comparison of Boosting Algorithms in Spices Classification Using Histogram of Oriented Gradient Feature Extraction. *Journal of Computer Science, Information Technologi and Telecommunication Engineering (JCoSITTE)*, *4*(1). https://doi.org/10.30596/jcositte.v4i1.13710

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1. https://doi.org/10.1186/s40537-014-0007-7

Nath, P., Saha, P., Middya, A.I., & Roy, S. (2021). Long-term time-series pollution forecast using statistical and deep learning methods. *Neural Computing and Applications*, *33*(19), 12551-12570. https://doi.org/10.1007/s00521-021-05901-2

Obeidat, I., Hamadneh, N., Alkasassbeh, M., Almseidin, M., & AlZubi, M.I. (2019). Intensive Pre-Processing of KDD Cup 99 for Network Intrusion Classification Using Machine Learning Techniques. *International Journal of Interactive Mobile Technologies (iJIM)*, *13*(01), Article 01. https://doi.org/10.3991/ijim.v13i01.9679

Park, R.C., & Hong, E.J. (2022). Urban traffic accident risk prediction for knowledge-based mobile multimedia service. *Personal and Ubiquitous Computing*, *26*(2), 417-427. https://doi.org/10.1007/s00779-020-01442-y

Patel, S., Wang, M., Guo, J., Smith, G., & Chen, C. (2023). A Study of R-R Interval Transition Matrix Features for Machine Learning Algorithms in AFib Detection. *Sensors*, *23*(7), https://doi.org/10.3390/s23073700

Peng, L., & Liu, Y. (2018). Feature Selection and Overlapping Clustering-Based Multilabel Classification Model. *Mathematical Problems in Engineering*, *2018*(1), 2814897. https://doi.org/10.1155/2018/2814897

Puga, J.L., & García, J.G. (2012). A Comparative Study on Entrepreneurial Attitudes Modeled with Logistic Regression and Bayes Nets. *The Spanish Journal of Psychology*, *15*(3), 1147-1162. https://doi.org/10.5209/rev_SJOP.2012.v15.n3.39404

Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024). Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship*, *13*(1), 80. https://doi.org/10.1186/s13731-024-00436-x

Reynolds, P., Bosma, N., Autio, E., Hunt, S., De Bono, N., Servais, I., Lopez-Garcia, P., & Chin, N. (2005). Global Entrepreneurship Monitor: Data Collection Design and Implementation 1998-2003. *Small Business Economics*, *24*(3), 205-231. https://doi.org/10.1007/s11187-005-1980-1

Rezende, P.M., Xavier, J.S., Ascher, D.B., Fernandes, G.R., & Pires, D.E.V. (2022). Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings in Bioinformatics*, *23*(4), bbac216. https://doi.org/10.1093/bib/bbac216

Saranyadevi, S., Murugeswari, R., & Bathrinath, S. (2019). Road risk assessment using fuzzy Context-free Grammar based Association Rule Miner. *Sādhanā*, *44*(6), 151. https://doi.org/10.1007/s12046-019-1136-7

Sattar, H., Bajwa, I.S., & Shafi, U.F. (2019). An Intelligent Air Quality Sensing System for Open-Skin Wound Monitoring. *Electronics*, *8*(7), https://doi.org/10.3390/electronics8070801

Savin, I., Chukavina, K., & Pushkarev, A. (2023). Topic-based classification and identification of global trends for startup companies. *Small Business Economics*, *60*(2), 659-689. https://doi.org/10.1007/s11187-022-00609-6

Sharma, S., & Sharma, P. (2019). Predictive Risk Factors of Heart Disease using an Efficient Classification based Approach. *International Journal of Computer Applications*, *178*(27), 27-30. https://doi.org/10.5120/ijca2019919028

Singh, N., & Singh, P. (2021). A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemometrics and Intelligent Laboratory Systems*, *217*, 104396. https://doi.org/10.1016/j.chemolab.2021.104396

Soria, L.M., Ortega, F.J., Álvarez-García, J.A., Velasco, F., & Fernández-Cerero, D. (2020). How efficient deep-learning object detectors are?. *Neurocomputing*, *385*, 231-257. https://doi.org/10.1016/j.neucom.2019.10.094

Staartjes, V.E., Serra, C., Muscas, G., Maldaner, N., Akeret, K., Niftrik, C.H.B. van, Fierstra, J., Holzmann, D., & Regli, L. (2018). Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: A pilot study. *Neurosurgical Focus*, *45*(5), E12. https://doi.org/10.3171/2018.8.FOCUS18243

Stel, A. van, Carree, M., & Thurik, R. (2005). The Effect of Entrepreneurial Activity on National Economic Growth. *Small Business Economics*, *24*(3), 311-321. https://doi.org/10.1007/s11187-005-1996-6

Sternberg, R., & Wennekers, S. (2005). Determinants and Effects of New Business Creation Using Global Entrepreneurship Monitor Data. *Small Business Economics*, *24*(3), 193-203. https://doi.org/10.1007/s11187-005-1974-z

Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., & Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology*, *37*(8), 953-961. https://doi.org/10.1038/s41587-019-0202-3

Subramanian, R.S., Prabha, D., Maheswari, B., & Aswini, J. (2021). Customer Analysis Using Machine Learning Algorithms: A Case Study Using Banking Consumer Dataset. In *Recent Trends in Intensive Computing* (pp. 689-694). IOS Press. https://doi.org/10.3233/APC210263

Szerb L. (2004). A vállalkozás és a vállalkozói aktivitás mérése. *Statisztikai Szemle*, *82*(6-7), 545-566.

Tuncer, T., Dogan, S., Özyurt, F., Belhaouari, S.B., & Bensmail, H. (2020). Novel Multi Center and Threshold Ternary Pattern Based Method for Disease Detection Method Using Voice. *IEEE Access*, *8*, 84532-84540. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2992641

United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*. Retrieved from https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E on March 3, 2023.

Urbano, D., Alvarez, C., & Turró, A. (2013). Organizational resources and intrapreneurial activities: An international study. *Management Decision*, *51*(4), 854-870. https://doi.org/10.1108/00251741311326617

Vaghela, B.V., Vandra, H.K., & Modi, K.N. (2012). Analysis and Comparative Study of Classifiers for Relational Data Mining. *International Journal of Computer Applications*, *55*(7), 11-21. https://doi.org/10.5120/8765-2685

Wach, K., & Głodowska, A. (2021). How do demographics and basic traits of an entrepreneur impact the internationalization of firms?. *Oeconomia Copernicana*, *12*(2), Article 2. https://doi.org/10.24136/oc.2021.014

Wang, Z., Xu, C., Liu, W., Zhang, M., Zou, J., Shao, M., Feng, X., Yang, Q., Li, W., Shi, X., Zang, G., & Yin, C. (2023). A clinical prediction model for predicting the risk of liver metastasis from renal cell carcinoma based on machine learning. *Frontiers in Endocrinology*, *13*. https://doi.org/10.3389/fendo.2022.1083569

Weber, M. (1982). *A protestáns etika és a kapitalizmus szelleme [The Protestant Ethic and the Spirit of Capitalism]*. Gondolat.

Wood, D.E., & Salzberg, S.L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xu, H., Zhou, J., G. Asteris, P., Jahed Armaghani, D., & Tahir, M.M. (2019). Supervised Machine Learning Techniques to the Prediction of Tunnel Boring Machine Penetration Rate. *Applied Sciences*, *9*(18).

Authors

The contributions of the co-authors are unequal and can be expressed as 40%, 15%, and 45%. Á. Szennay prepared the dataset, conception, introduction, literature review (conceptual framework), and discussion. J. Csákné Filep supervised the entire project, secured the funding, and reviewed the results, while M. Krankovits prepared a literature review (machine learning techniques), statistical calculations, and materials and methods.

Aron Szennay (corresponding author)
Senior Research Fellow at the Budapest LAB Office for Entrepreneurship Development, Budapest University of Economics and Business, PhD in regional sciences, and author of publications on entrepreneurship and its concerns regarding sustainability. His research interests include entrepreneurship, digitalisation, sustainability.
Correspondence to: Dr Aron Szennay, Budapest University of Economics and Business, HU-1087 Berzsenyi str. 6, Budapest, Hungary, e-mail: szennay.aron@uni-bge.hu
ORCID https://orcid.org/0000-0003-3567-9394

Judit Csákné Filep
Senior Research Fellow at the Budapest LAB Office for Entrepreneurship Development, Budapest University of Economics and Business. She holds a PhD in Management and Business Administration. She is the author of publications on family business and entrepreneurship. As the National Team Leader for Hungary in the Global Entrepreneurship Monitor, she is a major contributor to national and international entrepreneurship research. She also heads the Family Business Research Programme at the Budapest Business University. She is committed to the development of Hungarian family businesses and is a frequent contributor to conferences, podcasts, and other media engagements focused on the sector. Her research and interests include family business and entrepreneurship.
Correspondence to: Dr Judit Csákné Filep, Budapest Business University, HU-1087 Berzsenyi str. 6, Budapest, Hungary, e-mail: csaknefilep.judit@uni-bge.hu
ORCID https://orcid.org/0000-0002-5902-5195

Melinda Krankovits
Assistant professor at Széchenyi István University, Department of Mathematics and Computer Science, PhD in regional sciences. Author of publications on distance learning in higher education and a quality assurance expert. She also contributes to papers on other fields as an expert in AI and machine learning. Her research and interests include machine learning, data processing, and data mining.
Correspondence to: Dr Melinda Krankovits, Széchenyi István University, HU-9026 Egyetem sqr. 1, Győr, Hungary, e-mail: kmelinda@math.sze.hu
ORCID https://orcid.org/0000-0002-6722-782X

Acknowledgements and Financial Disclosure

Project no. TKP2021-NKTA-44 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NKTA funding scheme. The authors would like to thank the anonymous referees for their useful comments, which allowed to increase the value of this article.

Use of Artificial Intelligence

The Authors declare that AI tools, namely Grammarly and DeepL, were used to improve the manuscript's grammatical correctness and conciseness. The authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.