Journal of Systems Thinking in Practice
JSTINP, *QUARTERLY*
Homepage: jstinp.um.ac.ir
Online ISSN: 2821-1669, Print ISSN: 2980-9460
Research Article
DOI: 10.22067/JSTINP.2025.94947.1172
JSTINP 2025; Vol. 4, No. 4

A Hybrid Feature Selection and Classification Framework for Predicting Entrepreneurial Competency Using Machine Learning and Binary Grey Wolf Optimizer

Ahmad Jafarnejad[a*], Arman Rezasoltani[a], Amir Mohammad Khani[a], Sayedeh Hoda Hosseinian[c]

[a] *Department of Industrial Management, Faculty of Management, University of Tehran, Tehran, Iran.*
[c] *Department of Industrial Management, Faculty of Kish International Campus, University of Tehran, Tehran, Iran.*

How to cite this article

Jafarnejad, A., Rezasoltani, A., Khani, A. M., and Hosseinian, H., 2025. A Hybrid Feature Selection and Classification Framework for Predicting Entrepreneurial Competency Using Machine Learning and Binary Grey Wolf Optimizer, *Journal of Systems Thinking in Practice*, *4*(4), pp.129 -155. doi: 10.22067/jstinp.2025.94947.1172.
URL: https://jstinp.um.ac.ir/article_47391.html.

ABSTRACT

Entrepreneurial competency is a fundamental element of economic development and innovation in education. Research has shown that there is a relationship between psychological, educational, and environmental factors that lead to entrepreneurial capability. However, the combination of machine learning with accurate feature selection has not been thoroughly investigated. This study offers a hybridized framework combining the Binary Grey Wolf Optimizer and machine learning classification to predict the entrepreneurial competency of university students. The Binary Grey Wolf Optimizer features flexible options while grounded in the EntreComp framework and systematically captures multidimensional aspects of entrepreneurial competence as reflected in a theoretical framework. This work applied mutual information filtering, ADASYN to resolve class imbalance, and Optuna to adjust the hyperparameters for 16 different machine learning classifiers on a student data set with 219 records. The LightBGM algorithm demonstrated the highest level of accuracy, achieving 70.7%, followed by F1 with 68.3%, and an ROC AUC measure of 70.5%. Furthermore, SHAP analysis clearly demonstrated that entrepreneurial environments, physical health, and resilience play a pivotal role in assessing an individual's entrepreneurial potential. A comparison with Sharma and Manchanda's (2020) benchmark, which utilized standard classifiers and achieved a maximum accuracy of 59.18%, showcases the clear benefits of this integrated, optimization-based method. In addition to offering greater accuracy, this study presents a scalable and interpretable framework for academic institutions to assess and encourage entrepreneurial growth and advancement. Future research may expand upon this work by conducting longitudinal studies, utilizing cross-cultural datasets, and exploring alternative metaheuristic algorithms.

Keywords

Binary grey wolf optimizer, Entrepreneurial competency, Entrepreneurship education, Feature selection, Machine learning, Predictive modeling.

Article history

Received: 2025-08-16
Revised: 2025-10-09
Accepted: 2025-10-11
Published (Online): 2025-12-31

Number of Figures: 5 **Number of Tables: 8** **Number of Pages: 28** **Number of References:74**

****Corresponding author***
Email: jafarnjd@ut.ac.ir

This is an open access article under the CC BY license
https://creativecommons.org/licenses/by/4.0/

1. Introduction

The global economy relying on innovation means it is crucial to know what skills successful entrepreneurs bring to the table. A blend of knowledge, skills, attitudes, and behaviors is part of what we call competencies, allowing someone to complete entrepreneurial tasks effectively (Martínez-Martínez et al., 2025). Shah and Lala (2022) argue that these abilities strongly indicate who will thrive as entrepreneurs, mainly in conflict zones. As a result, how entrepreneurial skills connect with education and the environment is a subject of interest among researchers; these studies confirm that they can boost motivation and accomplishments among entrepreneurs (Rotimi et al., 2022). Recently, research has brought together the main skills needed for entrepreneurship and organized them so they can be used in education and students' development (Tittel and Terzidis, 2020). Nevertheless, current models are limited because they lack robust use of machine learning methods. Even though there has been considerable progress in relating entrepreneurial competency to various studies, current research has not fully embraced exploring what machine learning algorithms can do to predict, quantify, and ideally inform entrepreneurs' decision-making process that involves competence. Traditional methods of arriving at the entrepreneurial competence measure have exclusively focused on qualitative measures and hence do not capitalize on the quantitative dominance of machine learning and feature selection in contexts containing large datasets. By employing machine learning, particularly ensemble methods and optimization algorithms, researchers have an opportunity to explore and quantify the different qualities, or components, of entrepreneurial competencies more effectively and efficiently (Sarker, 2021).

The use of traditional linear models to capture nonlinear relationships between facets and the thought of the interacting features of entrepreneurial data are strong reasons to develop interesting frameworks to incorporate structural complexities and individualized interaction in the datasets (Yi et al., 2023). Sharma and Manchanda (2020) used machine learning to predict whether university students have entrepreneurial competence. Even though this research makes an early effort to use data in entrepreneurship, it does not have a detailed set of theories, and it fails to easily arrange the different psychological, educational, and environmental elements together in one method. The present study fills these gaps by basing its study on the EntreComp approach and applying advanced machine learning techniques, like parameter optimization through Optuna, using the Binary Grey Wolf Optimizer for feature selection, and explaining model outcomes with SHAP. Thanks to these new methods, the model is more solid, easy to explain, and in line with theoretical thinking, helping to show all the different parts of entrepreneurial competence. While various researchers have examined what makes entrepreneurs successful in tough circumstances, there is still little agreement on how to predict and measure these abilities. Most existing studies depend on descriptive statistics or easy forms of regression, without making full use of modern computing tools (Ding, 2023). For that reason, attaching a hybrid feature selection method to machine learning algorithms, including the BGWO algorithm, serves as a main path to improve the accuracy and relevance of predictive models (Essien et al., 2024). Moreover, discussions about ensemble methods in literature are increasing; however, few implementations have been carried out in entrepreneurial competencies (Pintelas and Livieris, 2020).

This work targets the design and testing of a unified feature selection and classification approach for predicting the skills required for entrepreneurship through the use of machine learning and BGWO. Some of the research questions used in this study are:

1. What factors are most important for building entrepreneurial skills, and how can BGWO help choose them effectively?
2. Is a hybrid framework more accurate in dealing with entrepreneurial competencies than classic machine learning algorithms?
3. What findings from the classifications might positively shape training and education in entrepreneurship?

Its main significance is that it adds to the body of literature in entrepreneurship and applies knowledge in various business areas. It seeks to connect business theories about entrepreneurial skills with real-world ways of collecting data in order to offer structure and guidance to policymakers, teachers, and experts. It also demonstrates how new computational methods are helping to develop entrepreneurship research focused on using data more effectively (Sarker, 2021). For this reason, the results may be used to update curricula and train students in schools, helping entrepreneurial classes adjust to current and future problems in the business world. This paper is organized to systematically and clearly answer the research questions and provide a better understanding of the proposed methodology. The Introduction highlights the research context by developing the problem statement, the aims of the research, and the relevance of the study. In the Literature Review, we examine research studies of entrepreneurial competencies and the available body of research exploring the application of machine learning to predictive modeling tasks. In the Methodology section, we describe the hybrid feature selection framework, including the application of the Binary Grey Wolf Optimizer (BGWO) and several machine learning classifiers. In the Results and Discussion section, we will present the empirical findings, comparative evaluation against previously established models, and discussion of the practical application of the results obtained through this research process. The Conclusion and Future Work section summarizes the highlights of the research, our limitations, and suggestions for future research activities.

2. Theoretical foundations and research background

2.1. Predicting entrepreneurial competency

The study of entrepreneurial competence is built on an understanding of how competencies relate to each other, can be learned by experience, and can support improving other competencies. Entrepreneurship is broadly defined as the process through which individuals identify, evaluate, and exploit opportunities to introduce new goods or services, organize ventures, or create value in dynamic environments (Ahmed and Harrison, 2022; Long, 1983) . As outlined by Shane and Venkataraman (2000), entrepreneurship is the academic study of how, by whom, and with what consequences opportunities to create future goods and services are discovered, evaluated, and exploited. This definition of entrepreneurship indicates its many aspects, including cognitive, behavioral, and contextual elements, all of which are relevant in understanding behavioral competency as an entrepreneur.

The EntreComp[1] framework captures three interconnected competence areas: “Ideas and Opportunities,” “Resources,” and “Into Action,” divided into 15 competence areas and descriptors. These competencies represent the amalgamation of knowledge, skills, and attitudes in the process of converting ideas into action within professional and personal environments (Stenholm et al., 2021; Margherita et al., 2016). This framework is a commonly used way within entrepreneurship education to facilitate curriculum development and competency-based assessment, as they relate to creativity, ethical thinking, sustainability, resource mobilization, taking initiative, and learning or being empowered by experience (Botha, 2023). Thus, it offers a theoretical extension for predictive models aiming at measuring entrepreneurial potential in student populations(Stenholm et al., 2021). Research efforts have helped create a list of necessary entrepreneurial abilities to use in forming predictive models. Draksler and Širec (2018) built a model that merges 17 constructs of entrepreneurial competencies from the earlier work of Botha (2023) and Yin et al. (2022). The model represents a detailed set of abilities required for being an entrepreneur in the modern-day economy (Draksler and Širec, 2018). All of these skills, as described in the frameworks, point out that entrepreneurs use both knowledge

[1] Entrepreneurship Competence

and behavioral methods, mainly in areas such as communication, self-confidence, and flexibility (Morris et al., 2013; Morris and König, 2020). Bolzani and Luppi point out that assessing these competencies is tough because they are quite varied (Bolzani and Luppi, 2021). The authors are summarizing the common view among scholars that entrepreneurship education needs to show both the methods and the place where such skills are developed. Since development is important with assessment, there should be a way in the model to measure how much someone's abilities progress and how they self-assess their skills.

Besides this, the awareness of sustainability's place in entrepreneurship has opened up studies of what it means to be a sustainable entrepreneur (SEC) and how that impacts entrepreneurial intentions (Joensuu-Salo et al., 2022; Ploum et al., 2017). There is a new focus in this area on creating entrepreneurial skills that also align with social and environmental goals, which in turn influence what is taught and how it is measured in schools (Reis et al., 2020). According to Reis et al., (2020), developing entrepreneurial skills should consider the whole economic and social environment, leading to a broader understanding of what it means to be entrepreneurial. In particular, adding self-directed learning to important competencies means that entrepreneurs should always be willing to adapt. In today's quickly changing marketplace, learning and putting what you know into practice are very important for being successful. Integrating these skills into several educational models proposes that we promote more than knowledge sharing by also nurturing the ability to face changes as entrepreneurs. All in all, assessing entrepreneurial competence calls for a wide theoretical basis that shows how various skills are related, addresses the evaluation concerns, and responds to leading tendencies towards sustainability. In schools, developing different competencies helps explain what individuals need to do well as entrepreneurs.

2.2. Machine learning and binary grey wolf optimizer

Machine learning represents a fundamental barrier of artificial intelligence, employing computational methods to recognize patterns, automate decisions, and improve efficiency across multiple applications. ML is founded on ideas from various mathematical and statistical domains, which influence how algorithms are constructed and used in practice (Jafarnejad Chaghoshi et al., 2024a). ML models have been developed to read data and perform prediction and support methods of supervised learning, unsupervised learning, and reinforcement learning (Mehregan et al., 2025a). Algorithms function this way because they rely on strong mathematical fields, like probabilitya theory and statistics, that allow us to derive models that we can use in general cases from data in sample sets (Qu et al., 2019). Much of the current research is focused on addressing the combination of ML, data science, and computational methods, while emphasizing more complex big data, now focused on improving the learning algorithms and addressing their main weaknesses of overfitting and underfitting.

The Binary Grey Wolf Optimizer (BGWO) is not a core element of machine learning but serves as a metaheuristic optimization algorithm inspired by the social hierarchy and hunting behavior of grey wolves. It is often employed in specific ML tasks, particularly for feature selection and hyperparameter tuning. BGWO enhances optimization tasks by simulating leadership dynamics in wolf packs to explore the solution space (Wang, 2024) efficiently. There are heuristic features of BGWO, and it draws on elements of competitive swarm intelligence that facilitate advancing solutions to challenging problems in finance (Chen, 2024), neuroinformatics (Glaser et al., 2020), and operational research (Mills et al., 2022). Incorporating BGWO, in conjunction with learning algorithms, creates multiple opportunities for strengthening and improving the efficiency of algorithms. The notion and application of information-theoretic principles in optimization processes demonstrate variations in algorithmic behavior and effectiveness (Bashir et al., 2020). Metaheuristic algorithms like BGWO have gained attention as complementary tools to enhance specific machine learning tasks such as feature selection and hyperparameter tuning, especially in complex problem domains (Schuld and Killoran, 2019). Accordingly, engaging in BGWO requires consideration of the main principles underlying ML and deciding on the algorithm depending on the organizational methods and styles of data (Adjodah et al., 2019). All in all, with machine learning, change continues as we develop, in large part, because of the advancements in algorithm design with natural heuristic approaches in systems such as BGWO. This rapid change provides opportunities for prospective theoretical principles to be considered with the now-expanding range of machine learning applications.

2.3. Research background

In order to situate this research within existing research, a review of existing literature that has used machine learning methods to predict entrepreneurship is summarized in Table 1. The most notable studies are categorized by their aim, methodology, target population, and findings. This comparison shows the different ways machine learning has been used to model key entrepreneurial outcomes: entrepreneurial intention, spirit, competence, or startup success. In addition, this comparison starts to surface methodological gaps while setting the stage for the study at hand.

Table 1. Research background

Authors	Article title	Goals	Methods	Population	Conclusion
Chen et al. (2025)	What makes you entrepreneurial? Using machine learning to predict technology entrepreneurship	Prediction accuracy of technology entrepreneurship	Model development and comparison	Global Entrepreneurship Monitor in 2020	The XGBoost+SMOTE algorithm best predicted technology entrepreneurship, with education and opportunity perception as key factors.
Suryadi and Kurniawan (2024)	Predicting Entrepreneurial Spirit: A Machine Learning Approach	Entrepreneurial spirit levels	Model development and comparison	400 students	Random Forest predicted entrepreneurial spirit with a macro-averaged F1-score of 0.656.
Aristizábal et al. (2025)	Predicting entrepreneurial intention in Colombian academics: a machine learning approach	Entrepreneurial intention	Survey and machine learning-based data analysis	Colombian academic researchers	Entrepreneurial intention was higher among researchers connected to the private sector and lacking university support.
Bidgoli et al. (2024)	Predicting the success of startups using a machine learning approach	Prediction accuracy of startup success	Prediction accuracy of startup success	startups	Random Forest and Gradient Boosting predicted startup success.
Gosztonyi and Judit (2022)	Profiling (Non-)Nascent Entrepreneurs in Hungary Based on Machine Learning Approaches	Prediction accuracy of nascent entrepreneur identification.	Model development and comparison	Nascent entrepreneurs in Hungary	Machine learning algorithms can accurately predict nascent entrepreneurs in Hungary.
Kaur and Bhatt (2023)	Forecasting Student Entrepreneurial Competency: Machine Learning Approach	Predictive accuracy of entrepreneurial competency models	Model development and comparison	University students	XGBoost predicted entrepreneurial competency with 79% accuracy, outperforming prior models.
Rivera-Kempis et al. (2021)	Entrepreneurial Competence: Using Machine Learning to Classify Entrepreneurs	Entrepreneurial competence attributes	Model development and comparison	6649 individuals from the Latin American	Machine learning models accurately classified individuals as entrepreneurs based on 20 competence attributes.
Jabeur et al. (2021)	Forecasting the macro-level determinants of entrepreneurial opportunities using artificial intelligence models	Determinants of entrepreneurial opportunity, predictive accuracy of machine learning models	Model development and comparison	Panel data set from 149 countries covering the years 2007 to 2018	CatBoost regression best predicted entrepreneurial opportunity, outperforming other models.
Wei et al. (2020)	Predicting Entrepreneurial Intention of Students: An Extreme Learning Machine With Gaussian Barebone Harris Hawks Optimizer	Prediction accuracy of entrepreneurial intention	Model development and comparison	Students	GBHHO-KELM achieved superior classification performance and stability compared to other methods.

Machine learning is an effective means of predicting entrepreneurial abilities, intentions, and performance based on the assessment of intricate behavioral, sociocultural, demographic, and contextual data. Recent studies utilize a variety of ML algorithms and analysis techniques in order to estimate and predict which individuals will become entrepreneurs; the outcomes of their startups are predicted; and to inform the development of educational, training, and policy interventions aimed at increasing entrepreneurship. Simple to complex ML models have been used to study entrepreneurial mode, including random forest, support vector machines, XGBoost, kernel extreme learning machines, and ensemble methods. Studies consistently show that advanced ensemble models, such as XGBoost and gradient boosting, demonstrate more accuracy and stability than traditional statistical methods and simpler ML algorithms to predict entrepreneurial intention and activity (Gosztonyi and Judit, 2022; Jabeur et al., 2021; Suryadi and Kurniawan, 2024).

Key predictors of entrepreneurial ability comprise self-efficacy, opportunity identification, past entrepreneurial experience, education, social capital, and personality. Student factors may consist of major, campus innovation experience rated by students and/or staff, and competencies around leading teams. For emerging entrepreneurs and startups, social media use, access to funding and additional team members were important. Predictions that rely on ML inform talent discovery, curriculum development, and investment decisions to help policymakers and educators adjust interventions to nurture entrepreneurial ecosystems and identify gaps (e.g., gender gaps or lack of institutional support).

Table 2. Timeline of key studies on ML for entrepreneurial prediction

Focus area	Best-Performing algorithms	Key Predictors/Features	Citations
Student entrepreneurial intention	KELM, XGBoost, Random Forest	Major, innovation experience, personality	(Wei et al., 2020) (Suryadi and Kurniawan, 2024) (Kaur and Bhatt, 2023)
Startup success	Random Forest, Gradient Boosting	Social media, funding, team size	Bidgoli et al. (2024)
Entrepreneurial activity (general)	XGBoost, SVM, Ensemble Methods	Self-efficacy, opportunity, and social capital	(Jabeur et al., 2021) (Gosztonyi and Judit, 2022) (Chen et al., 2025)

As illustrated in Figure 1, machine learning (ML) techniques have recently been implemented successfully in a variety of entrepreneurial prediction problems, such as entrepreneurial intention (Wei et al., 2020; Aristizábal et al., 2025), startup success (Bidgoli et al., 2024), and entrepreneurial spirit (Suryadi and Kurniawan, 2024). However, most of these contributions have looked at either binary Classification Problems (e.g., Entrepreneur vs. Non-Entrepreneur) or wider macro-level factors (e.g., national indicators) (Jabeur et al., 2021), often ignoring competency-based contributions to enhance prediction at the individual level significantly. Moreover, some studies simply do not include hybrid modeling approaches that take advantage of advanced feature selection and ensemble learning, which has implications for limiting interpretability and generalizability.

This study addresses these limitations by focusing exclusively on predicting entrepreneurial competencies using individual-level data using the EntreComp framework conceptually. The approach taken combines the use of a Binary Grey Wolf Optimizer for feature selection as well as benchmarked predictive performance of 16 machine learning classifiers, thereby combining optimization and performance evaluation. This hybrid approach improves the predictive accuracy as well as identifies which competencies matter. Accordingly, this study adds to the literature by generating a novel, data-driven approach to identify entrepreneurial competencies in students so that they can experience value with their education or career development programs.

3. Research methodology

A complete and precise approach for selecting features and predicting entrepreneurial skills with machine learning and meta-heuristic methods is described in this study. Sharma and Manchanda (2020) provide the data used initially, including demographic, educational, psychological, and behavioral information about students. After all the data was preprocessed, encoded, and standardized, we fixed the class imbalance issue by applying the ADASYN algorithm. First, mutual information was used for a rapid screening of features, and next, the LightGBM model was optimized using both Optuna and the Binary Grey Wolf Algorithm (BGWO) to identify the best features. In addition, the model was examined using SHAP analysis to understand the effect of the chosen features. Subsequently, 16 machine learning algorithms were refined using Optuna and assessed using five evaluation criteria and five-fold cross-validation.

3.1. Data characteristics and preprocessing

In this study, we analyzed 219 samples of university students to assess their entrepreneurial competence. The data in this case study is from Sharma and Manchanda's (2020) work titled

"Predicting and Improving Entrepreneurial Competency in University Students using Machine Learning Algorithms." Every person represented in the study has specific personal, psychological, educational, and environmental attributes connected to their choice to become an entrepreneur. Along with the other features, the target variable (y) is also defined as a binary label specifying whether a person is an entrepreneur. As detailed in Table 3, the dataset features are described along with their types and corresponding categories or value ranges.

Table 3. Description of features

Feature	Description	Data Type	Range / Categories
EducationSector	Student's academic discipline	Nominal (categorical)	e.g., Economics, Humanities, Arts, Medicine, etc.
IndividualProject	Whether the student has done an individual project	Binary (text)	Yes / No
Age	Age of the student	Numeric	Between 17 and 26
Gender	Gender of the student	Binary (text)	Male / Female
City	Type of residence	Binary (text)	Yes / No
Influenced	Whether the student has been influenced by an entrepreneurial environment	Binary (text)	Yes / No
Perseverance	Level of perseverance	Ordinal (Likert scale)	Scale from 1 to 5
DesireToTakeInitiative	Desire to take initiative	Ordinal (Likert scale)	Scale from 1 to 5
Competitiveness	Degree of competitiveness	Ordinal (Likert scale)	Scale from 1 to 5
SelfReliance	Degree of self-reliance	Ordinal (Likert scale)	Scale from 1 to 5
StrongNeedToAchieve	Strong need for achievement	Ordinal (Likert scale)	Scale from 1 to 5
SelfConfidence	Self-confidence level	Ordinal (Likert scale)	Scale from 1 to 5
GoodPhysicalHealth	Physical health status	Ordinal (Likert scale)	Scale from 1 to 5
MentalDisorder	Whether the student has any mental disorder	Binary (text)	Yes / No
KeyTraits	Key personal traits (e.g., vision, positivity, resilience, work ethic)	Nominal (categorical)	Selected trait
ReasonsForLack	Reasons for not pursuing entrepreneurship	Textual (qualitative)	Descriptive text
y	Target label: Whether the person became an entrepreneur	Binary (numeric)	0: No, 1: Yes

Because it lacked structure, contained outliers, and had unbalanced content, the Reasons For Lack text variable—which provided a descriptive explanation of why some people did not wish to pursue entrepreneurship—was excluded from the analysis at this point. Label Encoding was used to convert IndividualProject, City, Influenced, Mental Disorder, and Gender into numerical values that could be used with machine learning algorithms. Lastly, One-Hot Encoding was used when working with KeyTraits and EducationSector, which contain more than two different values. In order to maintain clear and comprehensive results, no features were removed in this case. The ADASYN (Adaptive Synthetic Sampling) method was used to add balance to datasets. A useful technique for resolving imbalanced classes in classification problems is ADASYN. Although SMOTE and related techniques produce data in the same manner, this method chooses examples close to the boundary that are more difficult for the model to process (Mujahid et al., 2024; Tekkali and Natarajan, 2024).

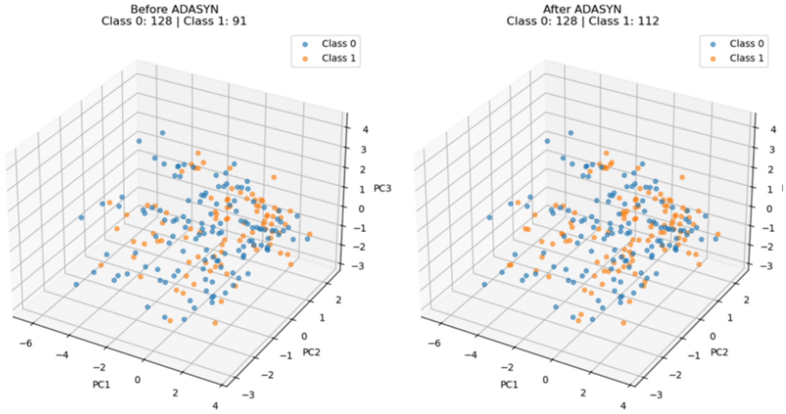

Figure 1. The impact of the ADASYN method on the data

Figure 1 illustrates the impact of the ADASYN method on class distribution. The initial data is presented in the left graph with 128 class 0 samples and only 91 class 1 samples(Toufighi et al., 2025). This can cause the model to prioritize the larger class and struggle to find attributes of the minority class. You can see in the right-hand graph that ADASYN has already applied to the data before presenting it in this manner. The new data class 1 samples increased to 112, and the distribution of samples across classes is now closer to equal. In order to provide an easy-to-see approach, PCA was applied to plot on both graphs in three-dimensional space. Clearly, ADASYN improved the way that new minority samples were added so that the distribution of class 1 samples more closely resembled that of the main class without changing the overall shape of the data. It is an important step to improving machine learning model performance. In this research, the StandardScaler was applied to the data to normalize the features so that the model would not depend mainly on larger numerical features.

3.2. Correlation analysis

In this work, the linear links between the dataset's variables were investigated using correlation analysis (Khani et al., 2022). This study sought variables with overlapping information with other features. A correlation matrix combined with a Heatmap diagram helped one to grasp the correlation structure between variables better intuitively (Malik et al., 2025).

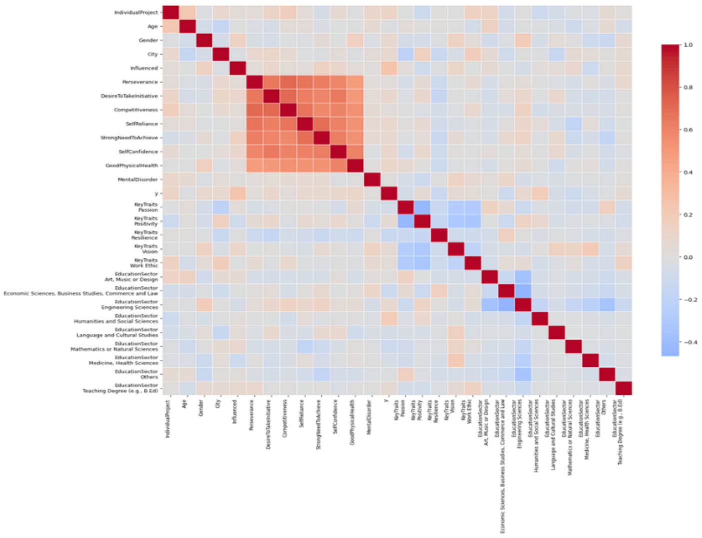

Figure 2. Correlation heat map

You can see the correlation matrix for all features of the dataset in Figure 2. It is easy to see the strength of the linear link between each pair of factors because the graph shades one color for positive (red) and another for negative (blue) dependencies. Red colors in the heat map are for high positive correlations, blue colors mark areas of negative correlations, and gray indicates there is no or a barely noticeable relation (Dudáš, 2024). On careful examination of the matrix, no correlation coefficient between the features goes above the value 0.8 in either direction. Consequently, since there were no strong and repeated correlations, we do not need to remove any features, so all remain in place for the following analyses(Hemdanou et al., 2024).

3.3. Selecting features and analyzing the importance and role of selected features

This work used an initial filter, metaheuristic optimization, and a machine learning model in concert for feature selection. Initially, the mutual information approach was used to determine if every feature included useful information connected to the target variable. Mutual information describes the link between two random variables that deviate from a straight line and is explained as:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

p(x,y) is the probability of both feature X and label Y together, and p(x) and p(y) are the marginal probabilities of just feature X and label Y. Only those 25 features were chosen that had the largest level of mutual information. Following this, the Binary Grey Wolf Optimizer was used to find the best combination of attributes for optimum classification. Like wolves, which find food by cooperating in large groups, this algorithm runs by updating the population's position after considering the alpha, beta, and delta leaders at every iteration (Ataş, 2025). This version treats each position in the search space as a binary vector for selecting or deselecting an associated feature. Having updated the position of wolves in the algorithm, the model uses this process (Kitonyi and Segera, 2021):

$$X(t+1) = \frac{1}{3}\left(X_\alpha + X_\beta + X_\delta\right) \tag{2}$$

Where:

X_α, X_β and X_δ are the positions of the best, second, and third wolf,

The update value is adjusted using random and control parameters A_i and C_i, which are defined as follows:

$$A = 2a \cdot r - a, C = 2 \cdot r \tag{3}$$

where $r \in [0,1]$ is a random value and a is a linearly decreasing parameter:

$$a = 2 - \frac{2t}{T} \tag{4}$$

Where t is the iteration number and T is the total number of iterations.

The LightGBM model was applied to assess every wolf's performance—selected subsetwise. Before beginning the GWO process, this model was refined on the balanced data using the Optuna optimization tool. Optuna's objective function was defined in line with minimizing the mean F1 complement value in the five-fold cross-validation (Shao et al., 2024):

$$Objective = 1 - \frac{1}{k}\sum_{i=1}^{k} F1_i \tag{5}$$

Table 4. Optimized parameters for LightGBM

Hyperparameter	Best Value	Search Range
learning_rate	0.004822835283446929	[0.001, 0.2] (log-uniform)
num_leaves	57	[15, 64] (int)
max_depth	3	[3, 10] (int)
min_data_in_leaf	20	[20, 120] (int)
feature_fraction	0.8332671412310059	[0.6, 1.0] (float)
bagging_fraction	0.9172201780267553	[0.6, 1.0] (float)
bagging_freq	6	[1, 10] (int)
lambda_l1	0.17739736393705147	[0.0, 5.0] (float)
lambda_l2	0.23964992272362257	[0.0, 5.0] (float)
n_estimators	581	[100, 600] (int)

Using a population of 80 wolves (POP_SIZE = 80), the Binary Grey Wolf Optimizer algorithm was implemented with 500 (EPOCHS = 500) number of algorithm iterations set. The wolves' starting position vector was generated at random from a uniform distribution spanning [0.1]. The top three wolves were found in each iteration with positions α, β, and δ, respectively, and the position update process of the other wolves was carried out by applying the conventional relations of the GWO technique. Every iteration, the value of the linear parameter a dropped from 2 to 0 to progressively transition from exploration to exploitation behavior. The values of the random coefficients A and C were computed for every position; the formula of the average of the three modified positions relative to wolves α, β, and δ found the new position of every feature (in binary space). Every position's fitness value was computed with a LightGBM model with Optuna optimal parameters. Emphasizing maximum F1 score in the five-fold cross-validation, the objective function was defined to minimize the value of 1 - F1 Score. The Binary Grey Wolf Optimizer (BGWO) algorithm applied in this work has settings shown in Table 5:

Table 5. Binary grey wolf optimizer (GWO) algorithm settings

Parameter	Value	Description
Population Size	80	Number of candidate solutions (wolves) in the population
Number of Iterations	500	Number of optimization iterations (epochs)
a (initial value)	Linearly decreases from 2 to 0	Controls balance between exploration and exploitation
Position Initialization	Uniform in [0, 1]	Random initialization of the feature selection vector
Binarization Threshold	0.5	Real-valued positions are converted to binary using a threshold 0.5
Fitness Function	1 - F1 Score (5-fold CV)	Optimization objective using 5-fold cross-validated F1-score
Base Classifier	LightGBM (tuned by Optuna)	LGBM with hyperparameters optimized by Optuna

The Binary Grey Wolf Optimizer algorithm was used, and 15 of those features were picked to be retained. The data set consisted of the initial variables, plus the features added as new independent variables from the One-Hot Encoding process. In the process, every column produced by One-Hot encoding was regarded as a unique feature and added to the analysis of its impact on the model’s performance. These features were tested through the modeling process: EducationSector_Art, Music or Design, Influenced, Age, KeyTraits_Resilience, EducationSector_Humanities and Social Sciences, GoodPhysicalHealth, Competitiveness, KeyTraits_Positivity, Perseverance, DesireToTakeInitiative, EducationSector_Others, KeyTraits_Vision, KeyTraits_Work Ethic, EducationSector_Economic Sciences, Business Studies, Commerce and Law, EducationSector_Mathematics or Natural Sciences.

This work applies the SHAP (SHapley Additive exPlanations) approach to investigate the significance and function of particular features in the output of a machine learning model. Based on the idea of cooperative games, this approach aims to clarify the contribution of every feature in the prediction of the model for every particular example (Wang et al., 2024a). SHAP shows its negative or positive contribution to the model's decision-making by valuing every feature, so that the total of these contributions equals the difference between the model's prediction and the average value of the whole output of the model (Santos et al., 2024). Mathematically, the SHAP value for a given feature i in a sample x is defined as follows:

$$\phi_i(f,x) = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x) - f_S(x)\right] \quad (6)$$

1. F is the total set of features.
2. S is a subset of features without feature i.
3. $f_S(x)$ is the predicted value of the model using the features S.

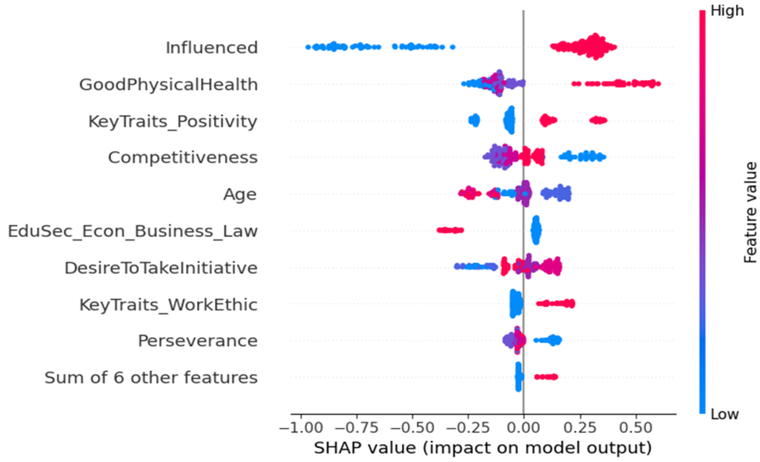

Figure 3. SHAP beeswarm plot

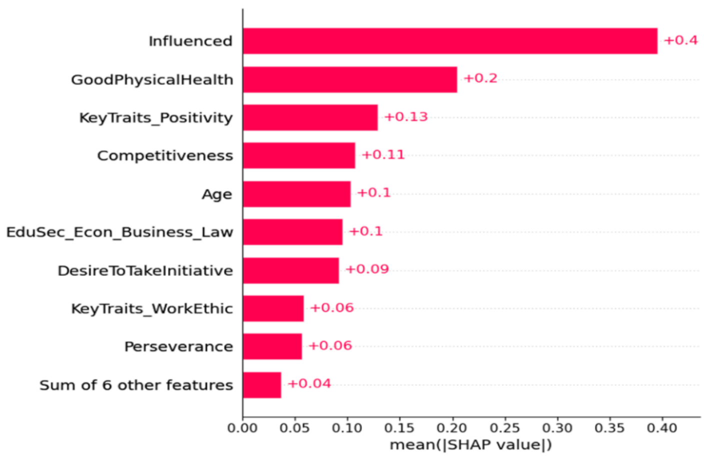

Figure 4. SHAP bar plot (mean absolute values)

Using color to indicate the degree of each feature—from light blue for low values to dark red for high values—the "Bee Swarm" plot in the figures shows the influence of each feature for every sample. To show the general significance of each feature in the model, the bar plot also shows its absolute value of the average SHAP for every feature. The features Influenced, GoodPhysicalHealth, and KeyTraits_Positivity dominated the model output based on the SHAP analysis. These results not only show the statistical relevance of the variables but also support an interpretable knowledge of the model's decisions. Consequently, the SHAP approach has been applied in this work as a useful instrument in analyzing the behavior of the model and clarifying its reasoning in feature application.

3.4. Machine learning models

This work employs a set of 16 different machine learning methods to find the most effective algorithm to predict entrepreneurial competency. Various methods are included in these algorithms, including linear, tree, reinforcement, statistical, and neural network. A hyperparameter optimization approach was used with the Optuna library to increase the exactness and performance of each algorithm. For each model, up to 2000 iterations of Bayesian search were completed to find the best mix of hyperparameters using the F1 criterion during 5-fold cross-validation. Table 6 indicates the list of best understood hyperparameters for all presented models.

Table 6. Machine learning models and optimal hyperparameters

Model	Description	Best Parameters	Reference
Logistic Regression	A simple linear model for binary classification, assuming feature independence	'C': 0.029, 'penalty': 'l2', 'solver': 'liblinear'	(Friedrich et al., 2022)
SVM	A margin-based classifier using the RBF kernel for non-linear separation	'C': 22.865, 'kernel': 'rbf', 'gamma': 'scale'	(Amaya-Tejera et al., 2024)
Decision Tree	A rule-based model that splits data into subsets based on feature thresholds	'criterion': 'gini', 'max_depth': 19, 'min_samples_split': 17, 'min_samples_leaf': 6	(Aaboub et al., 2023)
Random Forest	An ensemble of decision trees trained on bootstrapped samples	'n_estimators': 231, 'max_depth': 14, 'min_samples_split': 3, 'min_samples_leaf': 1, 'max_features': 'log2'	(Zhu et al., 2022)
Gradient Boosting	An iterative model that minimizes error using successive trees	'n_estimators': 325, 'learning_rate': 0.126, 'max_depth': 5, 'subsample': 0.508	(Fan et al., 2022)
AdaBoost	A boosting algorithm emphasizing difficult samples through adaptive weights	'n_estimators': 525, 'learning_rate': 0.024	(Khan et al., 2023)
Extra Trees	A randomized ensemble of decision trees for variance reduction	'n_estimators': 590, 'max_depth': 30, 'min_samples_split': 2, 'min_samples_leaf': 1	(Kaur et al., 2024)

Model	Description	Best Parameters	Reference
KNN	A non-parametric method based on nearest neighbors	'n_neighbors': 6, 'weights': 'distance', 'p': 2	(Halder et al., 2024)
MLP	A multi-layer feedforward neural network with nonlinear activation	'hidden_layer_sizes': (150, 100, 50), 'alpha': 0.000124, 'learning_rate_init': 0.0277	(Jafarnejad Chaghoshi et al., 2024b)
Naive Bayes	A probabilistic model with strong independence assumptions	default	(Gohari et al., 2023)
XGBoost	A gradient boosting model with efficient regularization and pruning	'n_estimators': 561, 'max_depth': 7, 'learning_rate': 0.016, 'subsample': 0.516, 'colsample_bytree': 0.875	(Mehdary et al., 2024)
LightGBM	A fast, histogram-based gradient boosting model optimized for performance	'n_estimators': 273, 'learning_rate': 0.298, 'max_depth': 9, 'min_data_in_leaf': 19, 'feature_fraction': 0.500, 'bagging_fraction': 0.607, 'bagging_freq': 3	(Lai et al., 2023)
CatBoost	A boosting model that handles categorical features internally	'iterations': 314, 'learning_rate': 0.0315, 'depth': 5	(Wang et al., 2024b)
LDA	A linear classifier assuming normal distribution and equal class covariances	default	(Gyamfi et al., 2019)
QDA	A quadratic version of LDA allowing for unequal class covariances	default	(Araveeporn, 2022)
GPC	A kernel-based probabilistic model using Gaussian processes	default	(Yang and Zhang, 2022)

In this research, four basic indicators (TP, TN, FP, and FN) were applied to assess machine learning models' performance in predicting entrepreneurial competence of students (Richardson et al., 2024). TP is the indicator of the successful prediction of entrepreneurial competence, while TN is the accurate identification of the incompetent (students lacking competence). FP is the prediction of competence where the student does not actually possess competence. FN is also the case where the model fails to identify the competent student. Based on the four basic indicators, accuracy, precision, recall, and the F1 score (weighted average of precision and recall) were calculated (Mehregan et al., 2025b). Furthermore, to assess the ability of models to discriminate groups of students (competent/incompetent), the Area Under the Curve—Receiver Operating Characteristic (AUC-ROC) index was used, which is most helpful in unbalanced datasets (Rainio et al., 2024). To avoid overfitting and maintain result validity, the 5-fold cross-validation method was used. In this system, the total data set is divided into five equal parts, and each time one of the five sections is used as the evaluation pool and the other four are used as the training pool. All five scenarios are run in a cyclic fashion where all of the data takes its turn as a test, and the mean of the report is the final answer to the performance of the model. This system also improves the reliability of the results and increases the performance of the parameters for the generalizability of models relative to ignored data. Table 7 presents the evaluation indicators for the machine learning models, including Accuracy, Precision, Recall, F1 Score, and ROC AUC.

Table 7. Evaluation indicators for machine learning models

Index	Definition	Formula
Accuracy	The ratio of correct predictions to total samples.	$\frac{TP + TN}{TP + FP + FN + TN}$
Precision	The ratio of correct predictions for a class to the total samples predicted as that class.	$\frac{TP}{TP + FP}$
Recall	The ratio of correct predictions for a class to the total actual samples of that class.	$\frac{TP}{TP + FN}$
F1 Score	The harmonic mean between Precision and Recall to balance them.	$\frac{2 \times Precision \times Recall}{Precision + Recall}$
ROC AUC	The area under the receiver operating characteristic curve; indicates the ability of the model to discriminate between classes. In this measure, for different thresholds, TPR (True Positive Rate = TP / (TP + FN)) and FPR (False Positive Rate = FP / (FP + TN)) are calculated and the area under their graph (ROC Curve) is calculated.	$\int_0^1 \mathrm{TPR(FPR)}\ d\mathrm{(FPR)}$

4. Results and discussion

This section describes the performance of machine learning models. The study utilized the Python programming language, and the models were all run on a computer with an Intel Core i7-13700H processor, 16 GB of RAM, and Python 12.3. Table 8 indicates the performance of machine learning models based on the evaluation criteria.

Table 8. Comparison of models based on evaluation criteria

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
LogisticRegression	0.602571	0.572693	0.576285	0.573517	0.600450
SVM	0.669858	0.640993	0.676680	0.654396	0.670648
DecisionTree	0.644858	0.615631	0.613043	0.610155	0.642522
RandomForest	0.628280	0.620874	0.550593	0.605828	0.642375
GradientBoosting	0.623582	0.602078	0.586957	0.599811	0.629087
AdaBoost	0.619415	0.607137	0.531225	0.562860	0.613613
ExtraTrees	0.648848	0.632245	0.632016	0.623344	0.656707
KNN	0.632358	0.588668	0.668379	0.622317	0.635574
MLP	0.649025	0.656935	0.640316	0.642478	0.672744
NaiveBayes	0.535993	0.502074	0.675889	0.574405	0.545637
XGBoost	0.677926	0.657198	0.658498	0.651329	0.677403
LightGBM	0.707447	0.685383	0.684585	0.682897	0.705831
CatBoost	0.678280	0.658603	0.660079	0.648454	0.678347
LDA	0.577571	0.547254	0.532016	0.539325	0.574316
QDA	0.586082	0.533121	0.893676	0.665493	0.607607
GPC	0.623848	0.586029	0.640711	0.610972	0.625433

A thorough examination of the performance table of machine learning models with respect to five evaluation metrics has confirmed that models performed differently, demonstrating a range of performances in each facet of predicting applicants' performance. The LightGBM with an accuracy of 0.707 provided the best accuracy overall amongst all models, demonstrating it was able to make ~71% of the predictions correctly. These models also lead in terms of three of the main criteria: The precision value of 0.685 indicates that a high percentage of its positive predictors were correct; its recall value is also 0.684 which demonstrates that the model was able to correct a high percent of true positives; the F1 score for this model is 0.683, which has a good trade-off in correct detection when accounting for the trade-off in precision and recall; finally, the ROC AUC value of 0.705 for LightGBM demonstrates its relatively good resolving power being able to discriminate between two classes correctly. The XGBoost and CatBoost also performed well. With an accuracy of 0.678, a precision of 0.657, a recall of 0.658, an F1 of 0.651, and an ROC AUC of 0.677, XGBoost specifically ranks lower than LightGBM across all metrics. This demonstrates that XGBoost is a very competitive choice that has successfully balanced all the metrics. With an accuracy of 0.678, a precision of 0.658, and an ROC AUC of 0.678, CatBoost is also comparable to XGBoost and occasionally even outperforms it, despite having a slightly lower F1 value (0.648). SVM has performed reasonably well among the classical models. With an accuracy of 0.669, precision of 0.640, recall of 0.677, and F1 of 0.654, it comes in at number four. Its acceptable discrimination power between two classes is indicated by its ROC AUC value of 0.670. Additionally, tree-based models with intermediate performance included RandomForest, ExtraTrees, and GradientBoosting. Specifically, ExtraTrees, with an F1 of 0.623 and an accuracy of 0.648, performs more steadily than other tree models. In contrast, RandomForest achieves a relatively good accuracy (0.628), but recall was even worse (0.551), which would indicate a rise in the False Negative rate. MLP, which is a multilayer perceptron neural network, performed relatively well. The F1 of this model was 0.642. Although this value improves most classical models. And although this model was better in final accuracy than KNN or NaiveBayes models, it does not match the level of LightGBM. For all other models run at default parameters, such as NaiveBayes, LDA, QDA, and GPC, average or poor results were achieved. NaiveBayes in particular performed the worst with the lowest accuracy of 0.535 and ROC AUC of 0.546, indicating that their assumption of independence between features is not suited for this problem. Although QDA reached a very high recall (0.894), this model suffers from a low precision (0.533) and moderate F1 (0.665), therefore also suffers from a high Final Positive Rate (FPR), which can lead to wrong decisions in practice.

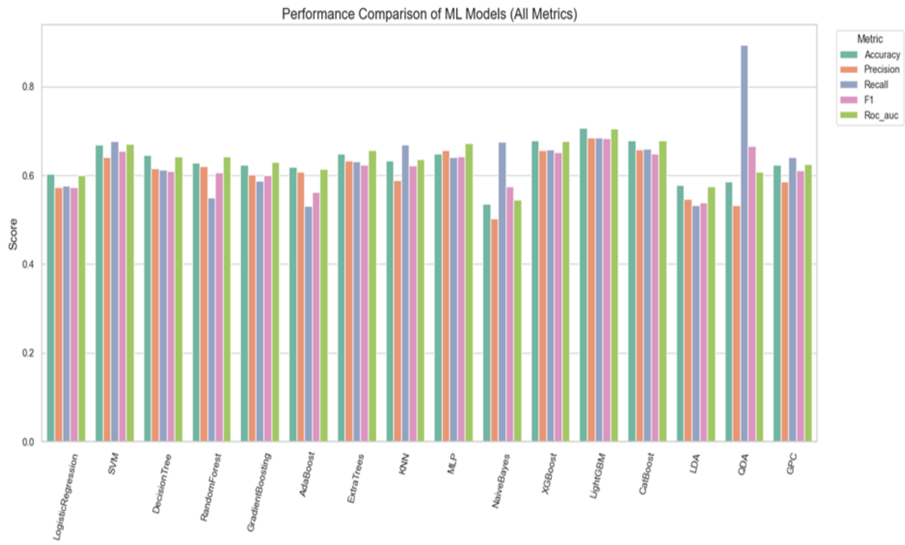

Figure 5. Comparison of model performance

According to Figure 5, LightGBM was measured to be the appropriate model for this classification problem. Out of 16 models observed, it had the best overall performance across all evaluation criteria, and the LightGBM model outperformed the next closest models. The next appropriate models were XGBoost, CatBoost, and SVM, which can be looked at as additional options.

5. Conclusion

The framework being proposed here unites machine learning and metaheuristic optimization to predict how well entrepreneurial skills are developed in university students. The study improves performance by applying the Binary Grey Wolf Optimizer for feature selection and applying Optuna to tune the hyperparameters, all while using SHAP analysis to maintain clarity. EntreComp is used as the basis for the predictive framework, helping to describe entrepreneurial skills in different areas such as thought, behavior, and circumstances. The research results point to the hybrid solution being superior to existing models. Of the 16 machine learning classifiers analyzed, LightGBM—which was hyperparameter-tuned via Bayesian search— subsequently performed the best overall on all assessment metrics with an accuracy of 70.7%, an average F1 score of 0.683, and an ROC AUC of 0.705. Overall, our results demonstrate the robustness, generalizability, and real-world applicability of our proposed model to consider entrepreneurial potential in an academic context.

An evaluation of the algorithms shows that the present study is stronger in predictive power and research methods when compared with the earlier work by Sharma and Manchanda (2020). The authors found that the highest accuracy was 59.18% with SVM, compared to the lowest at

40.81% from Random Forest when using 500 trees. However, for our study, we chose to include sixteen diverse machine learning models, backing them up with modern approaches such as hyperparameter tuning using Optuna, correcting unbalanced classes with ADASYN, and feature selection using Binary Grey Wolf Optimizer. Because of this, the LightGBM model displayed the best accuracy of 70.7%, far outperforming the prior study's models. As well, our assessment made use of F1 Score and ROC AUC in addition to accuracy to evaluate how the model truly performed. Our framework managed to improve the results of Naïve Bayes and Decision Tree, which previously had accuracies of 48.97% and 57.14% for Sharma and Manchanda (2020). Overall, our results prove that a hybrid optimization-based framework makes our approach stronger and more practical. By way of contrast, the current study employs a number of methodological innovations in this field. First, we take advantage of a three-step feature selection process that includes mutual information filtering, a BGWO, and SHAP to retain only the maximum relevance features with interpretability in mind. Secondly, ADASYN is used to address the class imbalance to balance the entrepreneurial and non-entrepreneurial cases. Thirdly, the ensemble models (LightGBM and CatBoost) that we optimized using extensive cross-validation and tuning provide better single-night prediction accuracy and stability. Finally, the SHAP values allow transparent interpretation that brings clarity of important features, undertaking (in-depth) responsibility, and giving understanding into the relevance of resilient, physical health, and motivational aspects.

In addition to the technical contributions offered, this work is a meaningful contribution to entrepreneurship education, both theoretically and practically. The linkage between predictive analytics and various existing models of competency development may provide educators, policy makers, and trainers with a reliable mechanism to assess entrepreneurial potential and suggest the steps necessary to develop it. Systematically identifying high efficacy competencies serves as a basis for data-informed curricular design and other interventions. After finishing this study, multiple directions for future research appear. To start, using longitudinal data allows us to capture the way entrepreneurial abilities change with time. Furthermore, by adapting the existing design, it could be known if entrepreneurship is treated similarly or differently in schools around the world. In addition to using DE, genetic algorithms, ant colony optimization, and particle swarm optimization can be tested within the same structure to examine how each one works and how fast it converges. Further, if statisticians further analyzed the model by differentiating between levels or styles of entrepreneurial orientation, they could discover new and valuable findings. Incorporating this predictive framework into applications in educational settings or online learning environments may encourage some form of real-time feedback, self-assessment tools, and personalized learning paths customized to students' own entrepreneurial strengths and developmental needs. Therefore, this study is the first step in understanding what it means to have an intelligent, theory-based, and computationally inspired assessment of entrepreneurial competency in education for the 21st century.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Aaboub, F., Chamlal, H. and Ouaderhman, T., 2023. Statistical analysis of various splitting criteria for decision trees. *Journal of Algorithms & Computational Technology*, *17*, p.17483026231198181. https://doi.org/10.1177/17483026231198181.

Adjodah, D., Calacci, D., Dubey, A., Goyal, A., Krafft, P., Moro, E. and Pentland, A., 2019. Leveraging communication topologies between learning agents in deep reinforcement learning. *arXiv preprint arXiv:1902.06740*. https://doi.org/10.48550/arxiv.1902.06740.

Ahmed, F and Harrison, C., 2025. Entrepreneurial leadership development in teams: A conceptual model. *The International Journal of Entrepreneurship and Innovation*, *26*(4), pp.267-279.https://doi.org/10.1177/14657503221143977.

Amaya-Tejera, N., Gamarra, M., Vélez, J.I. and Zurek, E., 2024. A distance-based kernel for classification via Support Vector Machines. *Frontiers in Artificial Intelligence*, *7*, p.1287875. https://doi.org/10.3389/frai.2024.1287875.

Araveeporn, A., 2022. Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity. *International Journal of Mathematics and Mathematical Sciences*, *2022*(1), p.7829795. https://doi.org/10.1155/2022/7829795.

Aristizábal, J.M., Tarapuez, E. and Astudillo, C.A., 2025. Predicting entrepreneurial intention in Colombian academics: a machine learning approach. *Journal of Entrepreneurship in Emerging Economies*, *17*(2), pp.260-288. https://doi.org/10.1108/jeee-04-2023-0141.

Ataş, P., 2025. A novel clustered-based binary grey wolf optimizer to solve the feature selection problem for uncovering the genetic links between non-Hodgkin lymphomas and rheumatologic diseases. *Health Information Science and Systems*, *13*(1), pp.1-22. https://doi.org/10.1007/s13755-025-00350-w.

Bashir, D., Montañez, G.D., Sehra, S., Segura, P.S. and Lauw, J., 2020, November. An information-theoretic perspective on overfitting and underfitting. In *Australasian Joint Conference on Artificial Intelligence* (pp. 347-358). Cham: Springer International Publishing. https://doi.org/10.48550/arxiv.2010.06076.

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

Bolzani, D. and Luppi, E., 2021. Assessing entrepreneurial competences: Insights from a business model challenge. *Education+ training*, *63*(2), pp.214-238. https://doi.org/10.1108/ET-04-2020-0072.

Botha, M., 2023. Using entrepreneurial competencies and action to profile entrepreneurs: a CHAID analysis approach. *Journal of Research in Marketing and Entrepreneurship*, *26*(2), pp.337-367. https://doi.org/10.1108/JRME-07-2022-0091.

Chen, C., Huang, Y., Wu, S., Zhao, Y. and Xu, L., 2025. What makes you entrepreneurial? Using machine learning to predict technology entrepreneurship. *Baltic Journal of Management*. https://doi.org/10.1108/bjm-09-2024-0552

Chen, Z., 2024. Research on Portfolio Optimization Model based on Machine Learning Algorithm in Stock Market. *Transactions on Economics, Business and Management Research 2024*, *6*, pp.50-56. https://doi.org/10.62051/sdqv4p21.

Ding, R., 2023. Which network is stronger? Le Net, Alex Net and VGG on image classification. *Applied Computing and Engineering*, *4*(1), pp.294–300. https://doi.org/10.54254/2755-2721/4/20230476.

Draksler, T.Z. and Širec, K., 2018. Conceptual research model for studying students' entrepreneurial competencies. *Naše gospodarstvo/Our economy*, *64*(4), pp.23-33. https://doi.org/10.2478/ngoe-2018-0020.

Dudáš, A., 2024. Graphical representation of data prediction potential: correlation graphs and correlation chains. *The Visual Computer*, *40*(10), pp.6969-6982. https://doi.org/10.1007/s00371-023-03240-y.

Essien, U.D., Ansa, G.O. and Akpnobong, A., 2024. EnsembleForge: A Comprehensive Framework for Simplified Training and Deployment of Stacked Ensemble Models in Classification Tasks. *Journal of Information Systems and Informatics*, *6*(1), pp.68-82. https://doi.org/10.51519/journalisi.v6i1.643.

Fan, M., Xiao, K., Sun, L., Zhang, S. and Xu, Y., 2022. Automated hyperparameter optimization of gradient boosting decision tree approach for gold mineral prospectivity mapping in the Xiong'ershan area. *Minerals*, *12*(12), p.1621. https://doi.org/10.3390/min12121621.

Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J. and Friede, T., 2022. Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research*, *32*(2), pp.425-440. https://doi.org/10.1177/09622802221133557.

Glaser, J.I., Benjamin, A.S., Chowdhury, R.H., Perich, M.G., Miller, L.E. and Kording, K.P., 2020. Machine learning for neural decoding. *eneuro*, *7*(4). https://doi.org/10.1523/eneuro.0506-19.2020.

Gohari, K., Kazemnejad, A., Mohammadi, M., Eskandari, F., Saberi, S., Esmaieli, M. and Sheidaei, A., 2023. A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients. *BMC Medical Research Methodology*, *23*(1), p.190. https://doi.org/10.1186/s12874-023-02013-4.

Gosztonyi, M. and Judit, C.F., 2022. Profiling (non-) nascent entrepreneurs in Hungary based on machine learning approaches. *Sustainability*, *14*(6), p.3571. https://doi.org/10.3390/su14063571.

Gyamfi, K.S., Brusey, J., Hunt, A. and Gaura, E., 2019. A dynamic linear model for heteroscedastic LDA under class imbalance. *Neurocomputing*, *343*, pp.65-75. https://doi.org/10.1016/j.neucom.2018.07.090.

Halder, R.K., Uddin, M.N., Uddin, M.A., Aryal, S. and Khraisat, A., 2024. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), p.113. https://doi.org/10.1186/s40537-024-00973-y.

Hemdanou, A.L., Sefian, M.L., Achtoun, Y. and Tahiri, I., 2024. Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. *Computers and Education: Artificial Intelligence*, *7*, p.100301. https://doi.org/10.1016/j.caeai.2024.100301.

Jabeur, S.B., Ballouk, H., Mefteh-Wali, S. and Omri, A., 2021. Forecasting the macrolevel determinants of entrepreneurial opportunities using artificial intelligence models. *Technological Forecasting and Social Change*, *175*, p.121353. https://doi.org/10.1016/j.techfore.2021.121353.

Jafarnejad Chaghoshi, A., Khani, A.M. and Rezasoltani, A., 2024b. Risk Modeling in Banking Services for the Blind Using Fuzzy FMEA and Graph Neural Network (GNN). *Journal of Industrial Management Perspective*, *14*(4), pp.223-255. https://doi.org/10.48308/jimp.14.4.223.[in Persian].

Jafarnejad Chaghoshi, A., Rezasoltani, A., & Khani, A. M., 2024a. Unleashing the power of ensemble learning: Predicting national ranks in Iran's university entrance examination. *Industrial Management Journal*, *16*(3), 457–481. https://doi.org/10.22059/imj.2024.381521.1008178. [in Persian].

Joensuu-Salo, S., Viljamaa, A. and Varamäki, E., 2022. Sustainable entrepreneurs of the future: The interplay between educational context, sustainable entrepreneurship competence, and entrepreneurial intentions. *Administrative Sciences*, *12*(1), p.23. https://doi.org/10.3390/admsci12010023.

Kaur, B.P., Singh, H., Hans, R., Sharma, S.K., Sharma, C. and Hassan, M.M., 2024. A Genetic algorithm aided hyper parameter optimization based ensemble model for respiratory disease prediction with Explainable AI. *Plos one*, *19*(12), p.e0308015. https://doi.org/10.1371/journal.pone.0308015.

Kaur, S. and Bhatt, P.C., 2023, December. Forecasting student entrepreneurial competency: Machine learning approach. In *2023 IEEE 3rd International Conference on Social Sciences and Intelligence Management (SSIM)* (pp. 184-189). IEEE.

Khan, A.A., Chaudhari, O. and Chandra, R., 2023. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, p.122778. https://doi.org/10.1016/j.eswa.2023.122778.

Khani, A. M., Kazazi, A., and Taqhavi Fard, M. T., 2022. Evaluating the quality of services of the cultural and social deputy of Tehran municipality in the field of culture and art. *Social Development & Welfare Planning*, *13*(50), 205–250. https://doi.org/10.22054/qjsd.2021.58035.2110. [in Persian].

Kitonyi, P.M. and Segera, D.R., 2021. Hybrid gradient descent grey wolf optimizer for optimal feature selection. *BioMed Research International*, *2021*(1), p.2555622. https://doi.org/10.1155/2021/2555622.

Lai, J.P., Lin, Y.L., Lin, H.C., Shih, C.Y., Wang, Y.P. and Pai, P.F., 2023. Tree-based machine learning models with optuna in predicting impedance values for circuit analysis. *Micromachines*, *14*(2), p.265. https://doi.org/10.3390/mi14020265.

Long, W., 1983. The meaning of entrepreneurship. *American Journal of small business*, *8*(2), pp.47-59.https://doi.org/10.1177/104225878300800209.

Malik, S., Patro, S.G.K., Mahanty, C., Hegde, R., Naveed, Q.N., Lasisi, A., Buradi, A., Emma, A.F. and Kraiem, N., 2025. Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports*, *15*(1), p.8738. https://doi.org/10.1038/s41598-025-92324-x.

Margherita, M., Kampylis, P., Punie, Y. and Van den Brande, G., 2016. *EntreComp: het Entrepreneurship Competence-raamwerk*. Europese Commissie-Joint Research Centre. https://doi.org/10.2791/160811.

Martínez-Martínez, S.L., Ventura, R. and Santos-Jaén, J.M., 2025. Creativity Vs Grit: key competences to understand Entrepreneurial Intention. *The International Journal of Management Education*, *23*(2), p.101143. https://doi.org/10.1016/j.ijme.2025.101143

Mehdary, A., Chehri, A., Jakimi, A. and Saadane, R., 2024. Hyperparameter optimization with genetic algorithms and XGBoost: a step forward in smart grid fraud detection. *sensors*, *24*(4), p.1230. https://doi.org/10.3390/s24041230.

Mehregan, M. R., Rezasoltani, A., and Khani, A. M., 2025b. A novel hybrid machine learning model for defect prediction in industrial manufacturing processes. *Contributions of Science and Technology for Engineering*, *2*(4), pp.43–58. https://doi.org/10.22080/cste.2025.29099.1037. [in Persian].

Mehregan, M.R., Taghavifard, M.T., Khani, A.M., Rezasoltani, A. and Nikkhah, M.A., 2025a. A Hybrid Machine Learning Model Based on Deep Learning for Air Quality Prediction. *Pollution*, *11*(4), pp.1199-1215. https://doi.org/10.22059/poll.2025.388743.2750.[in Persian].

Mills, A.W., Goings, J.J., Beck, D., Yang, C. and Li, X., 2022. Exploring potential energy surfaces using reinforcement machine learning. *Journal of Chemical Information and Modeling*, *62*(13), pp.3169-3179. https://doi.org/10.1021/acs.jcim.2c00373.

Morris, M.H., Webb, J.W., Fu, J. and Singhal, S., 2013. A competency-based perspective on entrepreneurship education: conceptual and empirical insights. *Journal of small business management*, *51*(3), pp.352-369. https://doi.org/10.1111/jsbm.12023.

Morris, T.H. and König, P.D., 2020. Self-directed experiential learning to meet ever-changing entrepreneurship demands. *Education+ Training*, *63*(1), pp.23-49. https://doi.org/10.1108/ET-09-2019-0209.

Mujahid, M., Kına, E.R.O.L., Rustam, F., Villar, M.G., Alvarado, E.S., De La Torre Diez, I. and Ashraf, I., 2024. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(1), p.87. https://doi.org/10.1186/s40537-024-00943-4.

Pintelas, P. and Livieris, I.E., 2020. Special issue on ensemble learning and applications. *Algorithms*, *13*(6), p.140. https://doi.org/10.3390/a13060140.

Ploum, L., Blok, V., Lans, T. and Omta, O., 2017. Toward a validated competence framework for sustainable entrepreneurship. *Organization & environment*, *31*(2), pp.113-132. https://doi.org/10.1177/1086026617697039.

Qu, K., Guo, F., Liu, X., Lin, Y. and Zou, Q., 2019. Application of machine learning in microbiology. *Frontiers in microbiology*, *10*, p.827. https://doi.org/10.3389/fmicb.2019.00827.

Rainio, O., Teuho, J. and Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), p.6086. https://doi.org/10.1038/s41598-024-56706-x.

Reis, D.A., Fleury, A.L. and Carvalho, M.M., 2020. Consolidating core entrepreneurial competences: toward a meta-competence framework. *International Journal of Entrepreneurial Behavior & Research*, *27*(1), pp.179-204. https://doi.org/10.1108/IJEBR-02-2020-0079.

Richardson, E., Trevizani, R., Greenbaum, J.A., Carter, H., Nielsen, M. and Peters, B., 2024. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, *5*(6). https://doi.org/10.1016/j.patter.2024.100994.

Rivera-Kempis, C., Valera, L. and Sastre-Castillo, M.A., 2021. Entrepreneurial competence: Using machine learning to classify entrepreneurs. *Sustainability*, *13*(15), p.8252. https://doi.org/10.3390/su13158252.

Rotimi, M.E., IseOlorunkanmi, J.O., Rotimi, G.G. and Doorasamy, M., 2022. Does Entrepreneurship Education Determine Entrepreneurial Motivation among University Graduates: An Empirical Investigation. *Jurnal Ekonomi, Bisnis & Entrepreneurship*, *16*(2), pp.107-123. https://doi.org/10.55208/jebe.v16i2.257.

Santos, M.R., Guedes, A. and Sanchez-Gendriz, I., 2024. Shapley additive explanations (shap) for efficient feature selection in rolling bearing fault diagnosis. *Machine Learning and Knowledge Extraction*, *6*(1), pp.316-341. https://doi.org/10.3390/make6010016.

Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), p.160. https://doi.org/10.1007/s42979-021-00592-x.

Schuld, M. and Killoran, N., 2019. Quantum machine learning in feature Hilbert spaces. *Physical review letters*, *122*(4), p.040504. https://doi.org/10.1103/PhysRevLett.122.040504.

Shah, M.S. and Lala, M.F., 2022. Are competencies of an entrepreneur determinant of success in a conflict zone: an evidence from conflict zone Kashmir. *Deleted Journal*, *1*(2), pp.77–83. https://doi.org/10.26480/mbmj.02.2022.77.83.

Shane, S. and Venkataraman, S., 2000. The promise of entrepreneurship as a field of research. *Academy of management review*, *25*(1), pp.217-226. https://doi.org/10.5465/amr.2000.2791611.

Shao, H., Liu, X., Zong, D. and Song, Q., 2024. Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nutrition & Diabetes*, *14*(1), p.63. https://doi.org/10.1038/s41387-024-00324-z.

Sharma, U. and Manchanda, N., 2020, Predicting and improving entrepreneurial competency in university students using machine learning algorithms. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 305-309). IEEE. https://doi.org/10.1109/Confluence47617.2020.9058292.

Stenholm, P., Ramström, J., Franzén, R. and Nieminen, L., 2021. Unintentional teaching of entrepreneurial competences. *Industry and Higher Education*, *35*(4), pp.505-517. https://doi.org/10.1177/09504222211018068.

Suryadi, D. and Kurniawan, D., 2024, August. Predicting Entrepreneurial Spirit: A Machine Learning Approach. In *2024 International Conference on Information Management and Technology (ICIMTech)* (pp. 399-403). https://doi.org/10.1109/icimtech63123.2024.10780930.

Tekkali, C.G. and Natarajan, K., 2024. An advancement in AdaSyn for imbalanced learning: An application to fraud detection in digital transactions. *Journal of Intelligent & Fuzzy Systems*, *46*(5-6), pp.11381-11396. https://doi.org/10.3233/JIFS-236392.

Tittel, A. and Terzidis, O., 2020. Entrepreneurial competences revised: developing a consolidated and categorized list of entrepreneurial competences. *Entrepreneurship Education*, *3*(1), pp.1-35. https://doi.org/10.1007/s41959-019-00021-4.

Toufighi, S. P., Khani, A. M., Rezasoltani, A., Sahebi, I. G., & Vang, J. (2025). Forecasting stock market anomalies in emerging markets: An OPTUNA-optimized isolation forest and K-means approach. Machine Learning With Applications, 22, 100770. https://doi.org/10.1016/j.mlwa.2025.100770

Wang, H., Liang, Q., Hancock, J.T. and Khoshgoftaar, T.M., 2024a. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), p.44. https://doi.org/10.1186/s40537-024-00905-w.

Wang, Q., Yan, C., Zhang, Y., Xu, Y., Wang, X. and Cui, P., 2024b. Numerical Simulation and Bayesian Optimization CatBoost Prediction Method for Characteristic Parameters of Veneer Roller Pressing and Defibering. *Forests*, *15*(12), p.2173. https://doi.org/10.3390/f15122173.

Wang, Y., 2024. Strengthen the foundation and practice orientation: exploration of teaching reform in machine learning courses. *Academic Journal of Science and Technology*, *12*(1), pp.145–148. https://doi.org/10.54097/a4xeav89

Wei, Y., Lv, H., Chen, M., Wang, M., Heidari, A.A., Chen, H. and Li, C., 2020. Predicting entrepreneurial intention of students: An extreme learning machine with Gaussian barebone Harris hawks optimizer. *Ieee Access*, *8*, pp.76841-76855. https://doi.org/10.1109/access.2020.2982796.

Yang, N. and Zhang, Y., 2022. A Gaussian process classification and target recognition algorithm for SAR images. *Scientific programming*, *2022*(1), p.9212856. https://doi.org/10.1155/2022/9212856.

Yi, F., Yang, H., Chen, D., Qin, Y., Han, H., Cui, J., Bai, W., Ma, Y., Zhang, R. and Yu, H., 2023. XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. *BMC medical informatics and decision making*, *23*(1), p.137. https://doi.org/10.1186/s12911-023-02238-9.

Yin, Z., Liu, Z. and Tong, P., 2022. Core entrepreneurial competences of Chinese college students: expert conceptualisation versus real-life cases. *The Asia-Pacific Education Researcher*, *31*(6), pp.781-801. https://doi.org/10.1007/s40299-022-00656-3.

Zhu, N., Zhu, C., Zhou, L., Zhu, Y. and Zhang, X., 2022. Optimization of the random forest hyperparameters for power industrial control systems intrusion detection using an improved grid search algorithm. *Applied Sciences*, *12*(20), p.10456. https://doi.org/10.3390/app122010456.