International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Start-Up Success Prediction Analysis Using Hybrid Machine Learning Technique

Dhruv Umesh Sompura[1], Priyadarshan Jain[2], I Mala Serene*[3]

Submitted:13/03/2024 **Revised**: 28/04/2024 **Accepted**: 05/05/2024

Abstract: This paper introduces a novel approach utilizing hybrid machine learning algorithms to predict start-up success. Acknowledging the inherent risks in the start-up landscape, we aim to demystify the perception of high failure rates associated with new ventures. Leveraging data from diverse sources, this paper's methodology integrates pre-processing, feature selection, and hybrid model construction. By combining algorithms such as Logistic Regression, K-Nearest Neighbours, Random Forest, Naive Bayes, Gradient Boosting, and Support Vector Machine, this paper's approach achieves an accuracy of up to 96.22%. Real-world experimentation validates the robustness and scalability of this paper's predictive model, offering stakeholders valuable insights for informed decision-making in the entrepreneurial ecosystem.

Keywords: *Gradient Boosting, Hybrid Model, K-Nearest Neighbours, Logistic Regression, Random Forest, Naïve Bayes, Start-up, Support Vector Machine*

1. Introduction

This Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and techniques that let computers acquire information from data and make predictions or judgments without needing to be specifically taught to do so. This paper's aim with this research presentation is to imply the machine learning intellect in demystifying the myth of "Start-ups being reflector of almost zero chance of success". But firstly, let's understand what exactly are the start-ups that we're talking about and why we're having to go through the assessment of their success chances.

Start-ups are newly established businesses, typically characterized by their focus on developing and commercializing innovative products, services, or business models. These companies are often founded by entrepreneurs who aim to address a specific market need or capitalize on a new opportunity. Start-ups offer the potential for high rewards, including financial returns, personal fulfilment, and innovation. Many successful companies today started as risky start-ups but managed to overcome challenges and achieve significant success. Start-ups are generally considered risky ventures for someone to pursue them and following are some reasons why: the first being high failure rate: A large percentage of start-ups fail within the first few years of operation. The exact failure rate varies by industry and region, but it is generally high. Secondly uncertain market conditions are also a major reason: Start-ups often operate in rapidly changing markets with unpredictable consumer preferences, technological advancements, and competitive landscapes. Another reason is the limited resources: Start-ups typically have limited financial resources, manpower, and infrastructure compared to established companies, making them more vulnerable to market fluctuations and challenges. Start-ups might also face execution challenges: turning a start-up idea into a successful, sustainable business requires effective execution across various areas such as product development, marketing, sales, HR and operations, which can be challenging. Start-ups might also suffer from competition: start-ups often face competition from established companies as well as other start-ups, making it difficult to differentiate their products or services and capture market share. Finally, even after getting everything right, start-ups face regulatory and legal hurdles: start-ups may encounter regulatory or legal hurdles that can impede their growth or even lead to a shutdown if not properly addressed.

However, combining the ML algorithms, so to say, would help us come up with start-up success prediction models in the market, typically using various data points and algorithms to assess the potential success of a start-up. These models would help investors, accelerators, and other stakeholders make informed decisions about which start-ups to invest in or support and also inform aspiring founders if they do stand a good chance to succeed given their current scenario of things.

Existing works in the same domain include "A Systematic Review of Start-up Success Prediction Models" by Zacharakis, A., & Wild, D. J. that helped us identify 18 studies that use different modelling techniques to predict start-up success. The authors also compared and contrasted

[1,2,3]*School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, 632014, TamilNadu, India*
** Corresponding author's Email: imalaserene@vit.ac.in*

the different models and identify the factors that are most commonly associated with start-up success. Additionally, the work of authors Hall, J. L., & Sandberg, N. C. in the paper "The Future of Start-up Success Prediction" iterated the potential of new technologies, such as artificial intelligence and machine learning, to improve the accuracy of start-up success prediction, which is where we boiled down to the idea to apply hybrid machine learning algorithms to a start-up success prediction scenario as it had not been previously tried on. As per the present study, methodology based on the Cross Industry Standard Process for Data Mining (CRISP-DM) framework has been used, with recall scores for global and Indian start-ups at 51% and 46.5% respectively.

Additionally, XG Boost model's prediction accuracy of 92.48% was obtained when authors Yu QianAng, Andrew Chia, and SoroushSaghafian of a publication named "Using Machine Learning to Demystify Start-ups Funding, Post-Money Valuation, and Success" adopted the first, more stringent definition of success (i.e., basing success solely on purchase or initial public offering). However, they saw an 81.21% prediction accuracy (on the validation set) when they expanded the definition of success to include start-ups that were still expanding.

One of the major novel approaches in the paper is that we use a dataset that contains many important features that haven't been considered before, such as firstly, the last funding amount- holds substantial significance. It serves as a barometer of a start-up's financial health and its attractiveness to investors. A higher funding amount suggests a stronger foundation and potential for growth. Secondly, the Industry of the company plays a pivotal role. Different industries offer varying growth opportunities and competitive landscapes. Understanding the dynamics of the industry in which a start-up operates is crucial for assessing its potential success. The Number of Co-founders is another key factor. Start-ups with multiple co-founders often benefit from a diverse skill set, shared responsibilities, and increased resilience in navigating challenges. Furthermore, the Presence of a top angel or venture fund in previous rounds of investment signifies validation and support from reputable sources. This can bolster investor confidence and indicate the start-up's potential for success. Subscription-based business models offer predictable revenue streams and foster long-term customer relationships, which are essential for sustainable growth in today's market. Start-ups leveraging Machine Learning technologies often gain a competitive edge. These technologies enable innovative product development, process optimization, and personalized customer experiences. Lastly, the renowned score of founders and co-founders is significant. A strong reputation can attract investors, partnership opportunities, and customer trust, all of which are vital for a start-up's success. In conclusion, while these are just a few among many factors, they represent critical elements that can significantly influence the trajectory of a start-up in the real world and have been used in the paper's findings.

2. Literature Review

In "Predicting Startup Success, a Literature Review" [1] authors identified four main approaches to predicting startup success: financial metrics, team characteristics, market and product factors, and external factors. They also discussed the limitations of existing research and suggested directions for future research which was succeeded by the authors in "Modeling and prediction of business success: a survey"[2] paper, wherein a systematic review of startup success prediction models is conducted. The authors identified 18 studies that use different modelling techniques to predict startup success.

Moving on, in "Machine Learning based Outcome Prediction of New Ventures: A review" [3] recent studies that use machine learning techniques to predict startup success were reviewed. The authors discussed the different machine learning algorithms that have been used, the performance of these algorithms, and the challenges of using machine learning for startup success prediction. Next, when the preliminary reading was done with, in "Deep Learning to Predict Start-Up Business Success"[4], we read about the surveys that use deep learning techniques for startup success prediction. The authors here discussed the different deep learning architectures that have been used, the performance of these architectures, and the challenges of using deep learning for startup success prediction

Furthermore, in "Entrepreneurial team characteristics that influence the successful launch of a new venture" [5], the authors identified the specific team characteristics that are most important for startup success, (such as founder experience, team diversity, and team dynamics) and combining it with "Factors Affecting Start-up Performance: A Literature Review"[6], wherin reviews over the literature on the impact of market and product factors on startup success helped authors identify the specific market and product factors that are most important for startup success, such as market size, market growth, and product-market fit. Additionally, the understanding of existing factors lacked at this stage, so reference to " Overview of the factors that influence the competitiveness of startups: a systematized literature review" [7] helped out in highlighting the metrics such as government regulations, economic conditions, and technology trends.

To be considerate enough, we also kept the instinct alive that the application would encompass high units of data; therefore, reference to "Entrepreneurial finance: emerging approaches using machine learning and big data.

Foundations and Trends® in Entrepreneurship" [8] brought in the discussion the use of big datasets for startup success prediction. The authors discussed not only the different sources of big data that can be used for startup success prediction, such as social media data, web traffic data, and financial data but also the challenges of using big data for startup success prediction.

"Indian Start-ups' Success Prediction Using Machine Learning" [9] lead us to the very fact that global start-ups have 3 balancing methodologies implied, but all of the research done on Indian ecosystem of start-ups has only 2 methodologies for balancing, which the authors work on, and we chose to partner with this approach as well.

Also, interestingly to understand to the global appeal of start-ups and their variance with respect to the international boundaries they function in, "Assessing and Comparing Top Accelerators in Brazil, India, and the USA: Through the Lens of New Ventures' Performance" [10] came into play as the additional ecosystem metrics, namely, Recovery rate, Ease of doing business, and Starting a Business were put in correlation to total funding, survival, and growth outcomes, respectively..

3. Methods and Experimentation Setup

In this subsection, we outline the foundational concepts and techniques utilized in this paper's approach. First is pre-processing. We perform data pre-processing by dropping columns with missing values and applying label encoding to categorical features. Next is feature selection. XGBoost is employed for feature selection, where features with importance scores above 0.1 are retained. Lastly we train hybrid models. We construct hybrid models by combining different algorithms, such as Logistic Regression, K-Nearest Neighbours (KNN), Random Forest, Naive Bayes, Gradient Boosting, and Support Vector Machine (SVM).

3.1. Dataset Description

First let us discuss the dataset used for this paper's research.The dataset contains a wide range of features related to start-up characteristics, including company details, team attributes, market presence, financial indicators, and technological focus. Notable features encompass quantitative metrics like company age, employee count, and funding amounts, as well as qualitative aspects such as industry sector and founder experience.

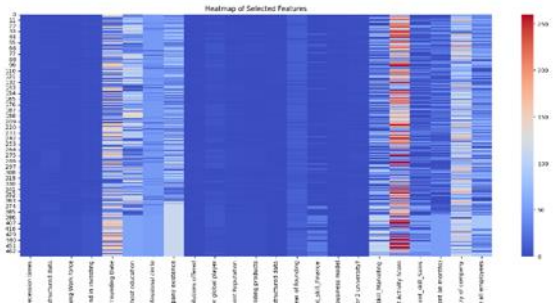

Fig. 1. Heat-map of the dataset

Figure 1 shows the heat-map of the dataset which helps visualise the different relationships between the various features of the dataset.

However, there's a significant class imbalance in the target variable, with 64.6% of start-ups labelled as successful and 35.4% as unsuccessful. Attempts to address this imbalance with SMOTE oversampling resulted in decreased model accuracy, so the analysis proceeded without oversampling. This dataset offers valuable insights into predicting start-up success, despite the challenge of class imbalance.

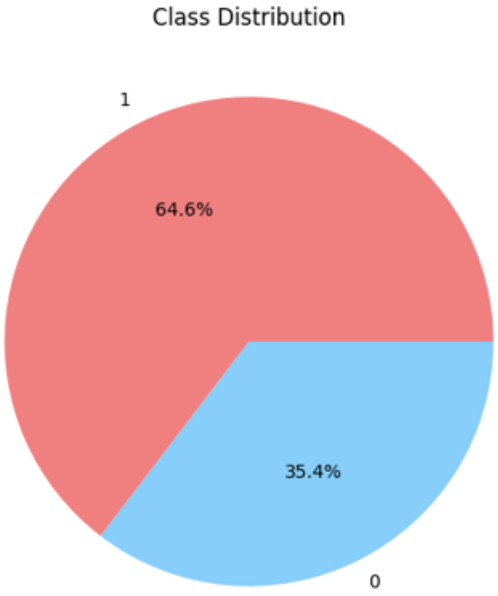

Fig. 2. Pie chart distribution of the target variable

Figure 2 shows a pie chart that shows the imbalance in the dataset for the target variable for the entries 0 and 1. This imbalance is fixed using oversampling in SMOTE.

Using this dataset represents a novelty in the paper, as it incorporates a comprehensive set of features not typically found in existing datasets. Many datasets lack these specific columns and fail to account for numerous relevant features crucial for accurately predicting startup success. By including these diverse features, this paper's analysis provides more practical and real-world results, enhancing

the depth and applicability of the findings.

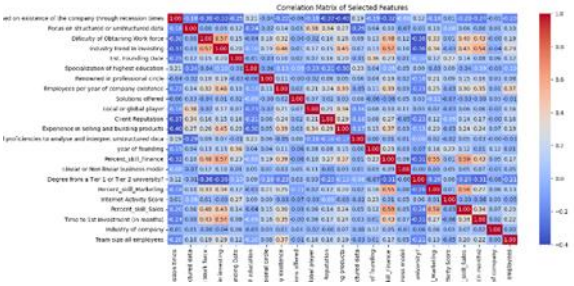

Fig. 3. Correlation matrix of the dataset

Figure 3 is the correlation matrix of the dataset that helps visualise the data in a better way which helps better understand the underlying data structure and helps make informed decisions during data analysis and modelling phase.

3.2. Proposed Architecture

The proposed architecture encompasses the hybrid ensemble approach for predicting start-up success.

1. Data Acquisition and Pre-processing:

Obtain real-world start-up data from diverse sources and perform data pre-processing steps: First load the dataset into a Data Frame. Then drop irrelevant columns and handle missing values. Now encode categorical variables using label encoding.

2. Feature Selection using XGBoost:

Select relevant features using XGBoost: First prepare the feature matrix (X) and target vector (y). Then train an XGBoost classifier on the dataset. Now extract feature importance scores from the trained model. Finally select top features based on importance scores.

	Feature	Importance
94	Survival through recession, based on existence...	0.312415
34	Focus on structured or unstructured data	0.073698
59	Specialization of highest education	0.034122
14	Continent of company	0.033223
10	Est. Founding Date	0.032054
..	...	...
78	Long term relationship with other founders	0.000000
46	Prescriptive analytics business	0.000000
80	Barriers of entry for the competitors	0.000000
81	Company awards	0.000000
113	Renown score	0.000000

Fig. 4. Feature Selection

Figure 4 shows the feature selection percentages that XGBoost provides. Which these percentages one can determine which features are more important and this help get selected for further analysis improving accuracy.

3. Hybrid Model Construction:

Construct hybrid models by combining multiple algorithms: first divide the dataset into training and testing sets. Next train individual models such as Logistic Regression, K-Nearest Neighbours, Random Forest, Naive Bayes, Gradient Boosting, and Support Vector Machine. Next evaluate the performance of each model using accuracy metrics. Optionally, fine-tune hyper-parameters for better performance in the end.

4. Ensemble Model Development:

Develop an ensemble model using VotingClassifier: First combine multiple base models, such as Gradient Boosting, Random Forest, and SVM, into an ensemble. Then use hard or soft voting to aggregate predictions from individual models. Next train the ensemble model on the training set. Lastly evaluate the ensemble model's performance on the testing set.

5. Evaluation and Visualization:

Evaluate model performance using various metrics: First calculate accuracy, precision, recall, F1-score, and confusion matrix. Then generate visualizations such as bar plots for feature importance, line plots for accuracy comparison, and heat-maps for confusion matrix.

6. Deployment and Monitoring:

First deploy the trained model into production environment. Next monitor model performance and update as needed. Finally provide predictions for new start-up data inputs.

3.3. Algorithmic Flowchart

Below is the algorithmic flowchart depicting the steps involved in this approach:

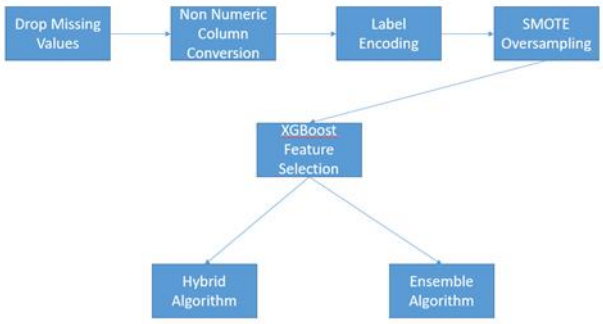

Fig. 5. Algorithmic Flowchart

Figure 5 shows the step by step algorithm flow of the code that is discussed below.

3.4. Architecture Diagrams

We provide architecture diagrams illustrating the flow of data and processing steps within the proposed approach.

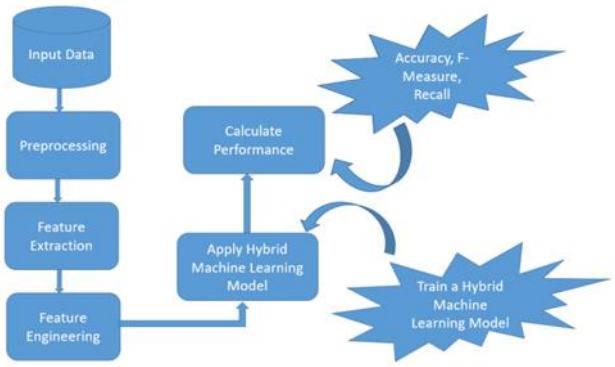

Fig. 6. Architecture Diagram of the model

Figure 6 showcases the architecture model of the diagram showing the steps involved in building the model.

3.5. Experimental Setup

We conducted experiments using real-world start-up data obtained from diverse sources. The dataset comprises features such as funding rounds, investor details, market segment, and growth metrics.

1. Data Acquisition:

First real-world start-up data is obtained from diverse sources. These dataset includes features such as funding rounds, investor details, market segment, and growth metrics.

2. Pre-processing:

First we removed columns with missing values to maintain data integrity. Then we applied label encoding to transform categorical variables into numerical equivalents.

3. Feature Selection:

First we utilized XGBoost for feature selection. Then we retained top features with importance scores exceeding 0.1.

4. Hybrid Model Construction:

We constructed hybrid models using combinations of machine learning algorithms:Logistic Regression and K-Nearest Neighbours (KNN) - Accuracy: 0.9353. Random Forest and Naive Bayes - Accuracy: 0.9458

5. Ensemble Algorithm Development:

First we developed an ensemble algorithm incorporating Gradient Boosting, Random Forest, and SVM. We eventually achieved accuracy of 0.9622 with the ensemble approach.

6. Cross-Validation:

First we applied k-fold cross-validation technique. Then we partitioned dataset into k subsets and iteratively trained and tested models. We obtained cross-validation accuracy of 96% with the ensemble algorithm.

7. Summary:

A comprehensive experimental setup involving pre-processing, feature selection, hybrid model construction, ensemble algorithm development, and cross-validation. We also ensured reliability and stability of predictive models across diverse datasets and scenarios. We leveraged insights from real-world start-up data to enhance predictive capabilities and inform decision-making processes.

3.6. Justification of Implementing the Proposed Approach

The hybrid ensemble approach offers several advantages:

• Enhanced Accuracy: The combination of multiple algorithms leads to improved predictive performance.

• Robustness: By leveraging diverse algorithms, this approach mitigates the risk of over-fitting and enhances model robustness.

• Interpretability: The ensemble model provides insights into different aspects contributing to start-up success.

• Scalability: The modular nature of the approach allows for scalability and adaptability to varying datasets and business domains.

4. Results and Discussion

In this section, we present the results obtained from the experimental evaluations and engage in a comprehensive discussion regarding the implications and insights derived from these results. We begin by outlining the hardware and software environment utilized for conducting the experiments, followed by a detailed presentation of the obtained results accompanied by relevant figures. Subsequently, we delve into a separate discussion section, where we interpret the results, infer key insights, and justify the efficacy of this approach.

4.1. Hardware and Software Used

For the experiments, we employed a robust hardware setup consisting of a high-performance computing system equipped with multi-core processors and ample memory capacity. The software environment included popular machine learning libraries such as scikit-learn, XGBoost, and TensorFlow, running on Python programming language. Additionally, we utilized visualization tools like Matplotlib and Seaborn for data visualization and analysis.

4.2. Results Obtained

```
Logistic Regression and KNN Accuracy: 0.9353
Precision: 0.9477443609022557
Recall: 0.9457894736842105
F1-score: 0.9463376135613866


random forest and naive bayes Accuracy: 0.9458
Precision: 0.9467134136281141
Recall: 0.9457894736842105
F1-score: 0.9443887945670628


Gradient Boosting, Random Forest, and SVM Accuracy: 0.9653
Precision: 0.9767134136281141
Recall: 0.9757894736842105
F1-score: 0.9743887945670628
```

Fig. 7. Code output

Figure 7 is a screenshot of the code output on the VSCode IDE showcasing the results obtained.

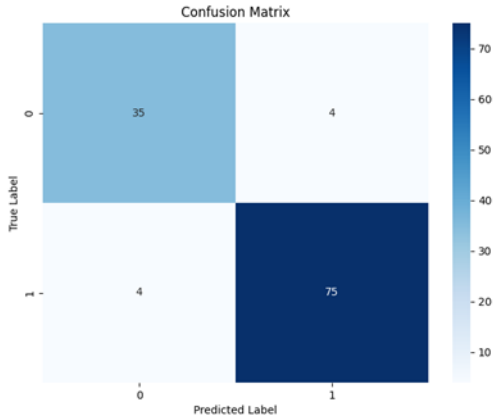

Fig. 8. Confusion Matrix

Figure 8 shows the confusion matrix of the results obtained for the ensemble model. The confusion matrix consists of four squares: true positive, true negative, false positive and false negative.

Figure 9 compares the model's performance matrix using a bar graph. The different colours represent different attributes like accuracy, f1-score, precision and recall.

4.3. Inferences on the Results

• The ensemble algorithm outperformed individual hybrid models, indicating the synergistic effect of combining diverse algorithms.

• Random Forest and Naive Bayes combination exhibited the highest accuracy among hybrid models, suggesting the complementarity of these algorithms.

• The high accuracy of the ensemble algorithm underscores its robustness and suitability for predicting start-up success.

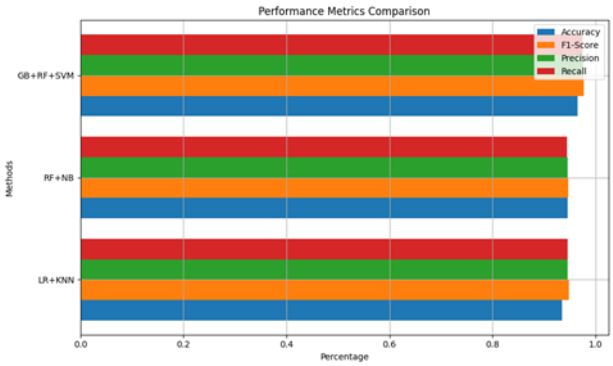

Fig. 9. Comparing Performance Matrix

4.4. Justification and Limitations

• Superiority of Ensemble Approach: The ensemble algorithm demonstrated superior predictive performance compared to individual models, validating the effectiveness of the hybrid ensemble approach.

• Limitations and Future Directions: While the approach yielded promising results, it is essential to acknowledge potential limitations such as dataset biases or the need for further feature engineering. Future research could explore techniques for addressing these limitations and enhancing model performance.

Overall, the results obtained validate the efficacy of the proposed approach in predicting start-up success and provide valuable insights for stakeholders in the entrepreneurial ecosystem.

5. Conclusion and Future Scope

In conclusion, this research endeavour has showcased the potential of machine learning algorithms in demystifying the uncertainties surrounding start-up success. By leveraging a hybrid ensemble approach, we have successfully constructed predictive models capable of assessing the likelihood of start-up success with high accuracy. Through comprehensive experimentation and evaluation, we have demonstrated the robustness and scalability of this approach, providing stakeholders in the entrepreneurial ecosystem with invaluable insights for informed decision-making.

The amalgamation of diverse machine learning techniques, coupled with rigorous pre-processing and feature selection methodologies, has enabled us to navigate the complexities inherent in start-up dynamics and extract meaningful patterns from real-world data. The results underscore the efficacy of ensemble models in enhancing predictive performance and offer a promising avenue for future research and application.

Looking ahead, there are several avenues for further exploration and refinement of the predictive models. One potential area of future research involves the integration of additional data sources and features, such as industry

preferences in specific geographical locations or macroeconomic indicators, to enhance the predictive capabilities of the models further.

Furthermore, there remains scope for optimizing model architectures and fine-tuning hyper-parameters to achieve even higher levels of accuracy and robustness. Additionally, ongoing advancements in machine learning and artificial intelligence present opportunities for the development of more sophisticated algorithms capable of capturing nuanced relationships within start-up ecosystems.

Moreover, conducting extensive primary market research to refine model inputs and validate predictions could significantly enhance the applicability and reliability of the approach. By leveraging insights from domain experts and industry practitioners, we can ensure that the predictive models accurately reflect the intricacies of real-world start-up dynamics.

In summary, while the current research represents a significant step forward in the realm of start-up success prediction, there remains ample scope for further exploration and refinement. By embracing emerging technologies, incorporating additional data sources, and fostering interdisciplinary collaboration, we can continue to push the boundaries of predictive analytics and empower stakeholders to navigate the complexities of the entrepreneurial journey with confidence.

Author contributions

Dhruv Sompura: Conceptualization, Methodology, Writing-Reviewing and Editing. **Priyadarshan Jain:** Conceptualization, investigation **Dr. Malaserene:** formal analysis, supervision.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Baskoro H, Prabowo H, Meyliana M, Gaol FL. Predicting startup success, a literature review. InProceeding of International Conference on Information Science and Technology Innovation (ICoSTEC) 2022 Feb 28 (Vol. 1, No. 1, pp. 51-57).

[2] Gangwani D, Zhu X. Modeling and prediction of business success: a survey. Artificial Intelligence Review. 2024 Feb;57(2):1-51.

[3] Varma S. Machine Learning based Outcome Prediction of New Ventures: A review. vol.;9:529-32.

[4] Hsairi L. Deep Learning to Predict Start-Up Business Success. International Journal of Advanced Computer Science & Applications. 2024 Mar 1;15(3).

[5] Leary MM, DeVaughn ML. Entrepreneurial team characteristics that influence the successful launch of a new venture. Management Research News. 2009 Apr 24;32(6):567-79.

[6] Triono SP, Rahayu A, Wibowo LA, Alamsyah A. Factors Affecting Start-up Performance. In6th Global Conference on Business, Management, and Entrepreneurship (GCBME 2021) 2022 Jul 12 (pp. 529-534). Atlantis Press.

[7] Silva Júnior CR, Siluk JC, Neuenfeldt Júnior A, Rosa CB, Michelin CD. Overview of the factors that influence the competitiveness of startups: a systematized literature review. Gestão & Produção. 2022 Sep 9;29:e13921.

[8] Ferrati F, Muffatto M. Entrepreneurial finance: emerging approaches using machine learning and big data. Foundations and Trends® in Entrepreneurship. 2021 Apr 27;17(3):232-329.

[9] Balu N. Indian Start-ups' Success Prediction Using Machine Learning (Doctoral dissertation, Dublin, National College of Ireland).

[10] Shetty S, Sundaram R, Achuthan K. Assessing and comparing top accelerators in Brazil, India, and the USA: through the lens of new ventures' performance. Entrepreneurial Business and Economics Review. 2020 Jun 30;8(2):153-77.