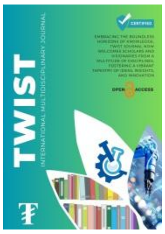
DOI https://doi.org/10.5281/zenodo.10049652#84

TWIST, 2024, Vol. 19, Issue 1, pp. 203-208

TWIST

Journal homepage: www.twistjournal.net



Supervised and Unsupervised Machine Learning Approaches in Predicting Startup Success

Carmen Sherellyn L. Biol*
College of Development Management, University of Southeastern Philippines, Mintal Campus,
Davao City 8000, Philippines
[*Corresponding author]

John Vianne B. Murcia
College of Business Administration Education, University of Mindanao, Bolton Street,
Davao City 8000, Philippines

Abstract

The alarming rates of failure that startups are facing highlight the need to promptly discover the crucial factors that determine success. This study investigates the forecasting of start-up success using both supervised and unsupervised machine learning techniques. Unsupervised instance filters were employed to address missing values and excessive standard deviations. In addition, the issue of data imbalance was addressed by implementing the Synthetic Minority Oversampling Technique (SMOTE). The investigation, conducted using the correlation and Info Gain attribute evaluator, revealed that relationships and milestones have the highest correlation with startup success. In order to forecast this achievement, we employed classifiers such as NaïveBayes, Functions.Logistics, Lazy.lBk, and Trees.J48. Out of all the options, the Trees.J48 model had the highest accuracy rate of 94.3%, with a confidence factor of 0.75. The accuracy of the Lazy.lBK (k-NN) variations decreases from 87.1% to 80.9% when the k-NN values increase from 3 to 7. Trees.J48 consistently exhibited strong prediction ability across different confidence levels in comparison to the other classifiers.

Keywords

Machine learning, Weka, Startup success, SMOTE, Cross-validation, Classifiers

INTRODUCTION

The significant rates of failure among companies during their initial years emphasize the crucial necessity to uncover the fundamental factors that contribute to their success or failure. Janáková (2015) found that around 50% of nascent enterprises terminate their operations during the first five years, a period commonly referred to as the 'valley of death' in the realm of company growth. The occurrence has stimulated a concentrated emphasis on entrepreneurial investigation, prompting an examination of the complex dynamics involved (Antretter et al., 2019). These dynamics extend beyond traditional measurements, exploring further into the complex network of elements that influence the future of startups.

Predicting the success of a startup is a complex and difficult task. Startups function within dynamic and frequently uncertain contexts (Von Gelderen, Frese & Thurik, 2000), which makes precise forecasting particularly intricate. This endeavor is further complicated by the difference in success rates among various sectors and geographical regions. Conventional predictive models sometimes depend on a narrow set of financial and operational measures (Svabova et al., 2020), which may not comprehensively encompass the complex range of elements that impact the future path of a firm. Furthermore, the dynamic and fast-changing characteristics of technology, industry trends, and consumer habits can diminish the predictive value of historical data for future results (Bharadiya, 2023a). Furthermore, the problem of data availability and quality arises, since some businesses, particularly in their first phases, may lack comprehensive records or data points commonly employed in predictive modeling (Martinez, Viles & Olaizola, 2021). To address these issues, a sophisticated strategy is required that takes into account various factors affecting the business, both from within and outside, and utilizes advanced analytical methods capable of managing the intricacy and unpredictability inherent in startup ecosystems.

The incorporation of machine learning into this field signifies a significant and transformative change, providing a fresh perspective to decipher and forecast the paths of startups (Jordan & Mitchell, 2015). Machine learning technologies, including as regression, classification, neural networks, and ensemble methods, show potential in identifying patterns in data connected to startups. Through the utilization of extensive datasets, machine learning algorithms have the ability to detect patterns, interrelationships, and crucial elements that impact the trajectory of a firm. The coupling of machine learning techniques with entrepreneurship studies enables avenues to create predictive models capable of projecting a venture's probable success or failure with exceptional accuracy.

The necessity of implementing a machine learning approach for predicting the success of startups is emphasized by the constraints of conventional analytical techniques in managing the extensive and intricate datasets that typify contemporary business landscapes (Allioui & Mourdi, 2023). The capacity of machine learning to analyze vast amounts of data and reveal concealed patterns (Basole, Park & Seuss, 2023) provides a fresh outlook on comprehending the complex characteristics of startup ecosystems (Graham & Bonner, 2022). Machine learning algorithms differ from traditional statistical methods in their ability to handle non-linear and interaction effects of several predictors (El Hajj & Hammoud, 2023). This enables a more sophisticated and dynamic comprehension of the factors that contribute to the success of startups. This strategy is especially vital in a company environment that is becoming more data-driven, as the capacity to make well-informed, predictive decisions typically determines competitive advantage. Using machine learning, businesses can acquire valuable insights into aspects such as market trends, consumer behavior, and operational efficiencies, which are not easily discernible using conventional analysis methods (Bharadiya, 2023b; Ma & Sun, 2020). This innovative methodology makes a substantial contribution to the area of entrepreneurship by providing a more accurate and anticipatory structure for assessing the potential of startups. As a result, it assists investors, policymakers, and entrepreneurs in making well-informed strategic choices based on data-driven observations.

In their study, Kim et al. (2023) employed machine learning techniques to investigate the features of various industries. They specifically focused on media exposure, funding dynamics, industry convergence, and association as crucial elements that influence the success of startups. In a similar manner, Bangdiwala et al. (2022) employed Decision Trees, Random Forest, Gradient Boost, Logistic Regression, and MLP Neural Networks. They found that all models achieved an accuracy of around 92%. Meanwhile, Vasquez et al. (2023) conducted a study that specifically examined 265 startups operating in the information technology industry in Australia. By utilizing seven machine learning algorithms and three hybrid models that incorporate the Voting strategy and the GreedyStepwise method, the researchers optimized variables and achieved notable improvements, resulting in a precision rate of 82% and an accuracy rate of 88%.

Piskunova et al. (2022) employed machine learning techniques to predict the likelihood of success for 123 start-ups in Ukraine. They discovered that the Decision Tree model, albeit less accurate overall with efficacy rates of approximately 61%, 55%, and 52%, exhibited an AUC level of 58%, which was lower than that of the Logistic Regression and Random Forest models. Their suggestion was to prioritize future study on the adaptation of these methodologies to other markets in similar evolutionary stages, as well as the exploration of alternative algorithms to gain a deeper understanding of the correlation between different types of start-up activity and their level of success. Hasan and Jain (2023) outlined the ultimate characteristics examined in their research, which include the list of categories, groups within those categories, gender, attainment of a degree, geographical region, city, number of years between graduation and establishment, and the years of study and graduation of the founder. While these internal determinants maintain value, the omission of external variables potentially overlooks crucial factors determining start-up success.

Ünal and Ceasu (2019) conducted a study on machine learning models, namely ensemble approaches such as random forests and extreme gradient boosting. They found that the timing of funding and the age of the company are important factors for determining the success of startups. Nevertheless, their analysis was restricted to some characteristics such as fundraising rounds, firm age, funding date, total money, social media utilization, continent, and sectors. They failed to include important factors such as interpersonal connections, the influence of age on financial support, significant accomplishments, forms of funding, and other necessary characteristics. Yin et al. (2021) discovered that LightGBM and XGBoost were the most effective models, achieving prediction accuracies of 53.03% and 52.96% respectively. However, their analysis was deficient in incorporating crucial factors such as relationships, milestones, and funding forms. Their suggestions for future research centered on the integration of other data sources and the exploration of supplementary features to augment prediction skills.

The objective of this study is to fill these gaps by undertaking a thorough investigation of multiple aspects that impact the success of startups. It leverages both supervised and unsupervised machine learning techniques to develop prediction models giving actionable insights for entrepreneurs, investors, and governments. The study employs various machine learning methods, including Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Decision Trees, using the dataset from Kaggle, a well-known data platform. Each algorithm boasts various techniques and strengths, all ready to uncover the subtle interplay of factors driving company success. Through the utilization of these algorithms, the research aims to not only reveal the specific factors that influence the success or failure of companies, but also to develop more precise predictive models that can accurately forecast success.

MATERIALS AND METHODS

Dataset

The dataset, acquired from Kaggle, comprises 923 instances with 40 attributes. Among these attributes, one is designated as the class attribute, while the remaining 39 serve as explanatory variables. The classification pertains to startup success, delineating two classes: "**acquired**" indicates successful acquisition by investors and "**closed**" represents project failure in the startup venture. The Acquired classification signifies successful project outcomes, whereas “closed” signifies unsuccessful or failed projects.

Table 1 List of Attributes of the Start-Up Success Prediction Dataset

Attribute	Type	Description
Age_first_funding_year	Numeric	The duration between the founding year and the year when it received its first funding round
Age_last_funding_year	Numeric	The duration between the founding year and the year of its most recent or last funding round.
Age_first_milestone_year	Numeric	The duration between the founding year and the year when it achieved its first significant milestone.
Age_last_milestone_year	Numeric	The duration between the founding year and the year when it achieved its most recent or latest milestone.
relationships	Numeric	The number of connections or associations established with other entities, such as other companies, investors, partners, or individuals
funding_rounds	Numeric	The number of funding rounds successfully secured from investors or funding sources.
funding_total	Numeric	The total amount of funding raised across all its funding rounds.
milestones	Numeric	The number of significant achievements, events, or key objectives reached during its development and growth
category_code	Nominal	The classification or categorization of the start-up firm based on the industry or sector it operates in.
has_VC	Nominal	If a startup has received venture capital funding or investment, coded 0-No, 1-Yes
has_angel	Nominal	If a startup has received funding from angel investors, coded 0-No, 1-Yes
has_roundA	Nominal	If a startup has successfully secured funding in a Series A round, coded 0-No, 1-Yes
has_roundB	Nominal	If a startup has successfully secured funding in a Series B round, coded 0-No, 1-Yes
has_roundC	Nominal	If a startup has successfully secured funding in a Series C round, coded 0-No, 1-Yes
has_roundD	Nominal	If a startup has successfully secured funding in a Series D round, coded 0-No, 1-Yes
avg_participants	Numeric	The average number of participants or investors involved in a funding round
is_top500	Nominal	If a startup falls within a list or category of top companies, often within a specific industry or ranking criteria.
status	Nominal	Classification whether the start-up has been 1-acquired or 2- closed

Data Preparation

The initial phase of data preparation involved a strategic process of identifying and removing irrelevant attributes to refine the dataset for subsequent analysis. Specifically, attributes such as state code, latitude, longitude, zip code, city identifiers (is_NY, is_MA, is_TX, is_otherstate), and 11 columns for the category list were excluded from the analysis. This curation resulted in a more concise dataset with 18 attributes, which are detailed in Table 1. Upon inspection, further adjustments were deemed necessary to enhance its suitability for analysis. Attributes category_code, has_VC, has_angel, has_roundA, has_roundB, has_roundC, has_roundD, and is_top500 were identified as requiring a transformation from their existing numeric format to a nominal representation. To achieve this, an unsupervised attribute filter, 'numeric to nominal', was applied. Following this adjustment, the dataset composition was refined to include a total of 9 nominal attributes, inclusive of the class attribute, along with 9 remaining attributes in their numeric format.

Subsequently, the identification of missing values in the dataset prompted the use of an unsupervised instance filter named 'replace missing values'. This method replaced missing instances with appropriate values, ensuring the integrity of the dataset for subsequent analysis. Specifically, for numeric attributes, the algorithm calculated the mean value based on available data and filled in missing entries with this computed mean. On the other hand, for nominal attributes, which refer to categorical data, the mode or the most frequently occurring value within the attribute was used to replace the missing instances. This step is crucial in maintaining the dataset's reliability and integrity, minimizing the impact of missing data on the accuracy and reliability of the analytical results derived from it (Maharana et al., 2022).

Further examination revealed that certain numeric data exhibited notably high standard deviations. To rectify this, the unsupervised instance filter Normalize was implemented. This method orchestrates a rescaling of the numeric attributes contained within the dataset, effectively adjusting their values to fit within a standardized range from 0 to 1. The principal objective underlying this normalization technique revolves around mitigating the impact of substantial standard deviations or variances observed in the numeric data (Vafaei et al., 2018). By normalizing the numeric features to a consistent scale, this procedure curtails the influence of outlier values and standardizes the numeric attributes.

Consequently, this strategy aids in creating a more uniform and comparable set of numeric attributes, addressing discrepancies originating from differing scales or magnitudes in the original dataset. Ultimately, this process significantly contributes to bolstering the stability and efficacy of subsequent analyses or modeling endeavors applied to the dataset, ensuring more reliable and consistent outcomes.

Data Imbalance

Addressing the imbalance in class attributes within the training set was crucial, with 597 instances belonging to class label 1 and 326 instances for class label 2. To counteract potential biases stemming from this imbalance, the Synthetic Minority Oversampling Technique (SMOTE), a supervised instance filter, was employed. SMOTE works by creating synthetic samples that are strategically placed in proximity to existing minority class instances in the feature space (Chawla et al., 2002). These newly generated instances are not merely duplicates but are synthetically created by considering the characteristics of existing minority instances, thereby expanding the representation of the minority class.

By oversampling the minority class (Class 2) using SMOTE, the dataset was augmented with new synthetic instances. The goal was to bring about a more balanced distribution between the classes, helping the model learn from a more representative dataset and prevent biases toward the majority class during training. In this case, the addition of 271 synthetic instances for Class 2 increased its representation by 83.3%, aligning both classes more evenly. The resulting dataset, comprising 1,194 instances after the SMOTE procedure, provided a more equitable representation of both classes, enabling the machine learning model to learn from a more diverse set of examples and make more accurate predictions for both the majority and minority classes.

Selection of Attributes

The attribute selection process was conducted using three prominent feature selection algorithms in Weka. This approach encompassed techniques based on correlation (CorrelationAttributeEval), information gain (InfoGainAttributeEval), and learning (WrapperSubsetEval). The attributes identified as most significant using CorrelationAttributeEval include relationships (*r*=0.4118), milestones (*r*=0.3519), is_top500 (*r*=0.3083), age_last_milestone_year (*r*=0.2554), and funding_rounds (*r*=0.2526). Notably, funding_total_usd (*r*=0.0486) and has_VC (*r*=0.0287) exhibit the lowest correlation. In the case of InfoGainAttributeEval, top-ranked attributes consist of relationships (*r*=0.234921), milestones (*r*=0.219051), funding_rounds (*r*=0.140209), and age_last_milestone_year (*r*=0.130464). Conversely, has_angel (*r*=0.002877) and has_VC (*r*=0.000595) are among the attributes ranked lowest. The WrapperSubsetEval approach determined the optimal number of folds for estimation accuracy, suggesting a five-fold cross-validation based on the results, which aided in refining the attribute selection process.

Data Classification and Cross-Validation

The chosen classification algorithms, NaiveBayes, Functions.Logistics, Lazy.lBK (k-NN), and Trees.J48 represent diverse approaches to handling data and making predictions. Each algorithm offers unique methodologies, from probabilistic reasoning in NaiveBayes to decision trees in Trees.J48, allowing for a comprehensive exploration of different modeling techniques. Within the Lazy.lBK (k-NN) classifier, experiments were conducted using different values of k (k-NN3, k-NN5, and k-NN7), altering the number of nearest neighbors considered in the classification process. Similarly, the Trees.J48 classifier underwent evaluations with varying confidence factors (0.25, 0.5, and 0.75), influencing the decision-making process within the tree-based model.

To ensure robust model assessment, a five-fold cross-validation strategy was adopted across all five classifiers. This cross-validation technique partitions the dataset into five subsets, iteratively using four subsets for training and the remaining subset for validation. Repeating this process five times with different subsets allows for comprehensive model evaluation, reducing the risk of overfitting and providing more reliable estimates of predictive performance across different algorithms.

RESULTS AND DISCUSSION

Assessing various classifiers' performance in predicting startup success yielded diverse performance outcomes. NaiveBayes demonstrated moderate accuracy, hovering at around 68.5%, with a fair agreement beyond chance (κ = 0.3702). In contrast, Functions.Logistics showcased improved accuracy, reaching approximately 75% with a moderate level of agreement (κ = 0.4992). The Lazy.lBK (k-NN) variants, adjusting the number of neighbors, displayed consistent classification but a slight decrease in accuracy from 87.1% to 80.9% as k-NN values progressed from 3 to 7, maintaining substantial to moderate agreement beyond chance. The highest correctly identified instances occurred at the k-NN 3 variant with1,040, followed by k-NN 5 with 988 instances and k-NN 5 with 967 instances. Conversely, Trees.J48 exhibited exceptional performance, achieving accuracy rates between 89.2% to 94.3% as the confidence factor rose from 0.25 to 0.75, demonstrating substantial or the highest agreement beyond chance as Kappa statistic ranges from 0.7839 to 0.8861. Notably, the highest correctly identified instances occurred at the 0.75 confidence factor, totaling 1126, followed by 1095 instances at 0.50 confidence and 1065 instances at 0.25 confidence. These findings highlight Trees.J48's robust predictive capabilities across varying confidence levels compared to other classifiers. The result of the classification accuracy of classifiers is displayed in Table 2.

Table 2 Classification accuracy of classifiers on the training dataset

Classifiers	Variant	Correctly Classified Instances (%)	κ
NaiveBayes		818 (68.5092%)	0.3702
Functions.Logistics		895 (74.9581%)	0.4992
Lazy.lBK (k-NN)	3	1040 (87.1022%)	0.742
Lazy.lBK (k-NN)	5	988 (82.7471%)	0.6549
Lazy.lBK (k-NN)	7	967 (80.9883%)	0.6198
Trees.J48	0.25	1065 (89.196 %)	0.7839
Trees.J48	0.50	1095 (91.7085%)	0.8342
Trees.J48	0.75	1126 (94.3049%)	0.8861

Using the five-fold cross-validation technique, the classifiers exhibited varying levels of performance in predicting startup success. NaiveBayes displayed an accuracy of 68.258%, accompanied by a Kappa statistic of 0.3652, suggesting a fair agreement beyond chance in its predictions. Functions.Logistics showed improved accuracy at 72.8643% with a Kappa statistic of 0.4573, indicating a moderate agreement in its predictive capabilities. The Lazy.lBK (k-NN) variants, employing different numbers of neighbors (3, 5, and 7), demonstrated accuracies ranging from 73.8693% to 74.3719% and Kappa statistics hovering around 0.4774, showcasing a moderate agreement in predictions. The highest accuracy was found in the k-NN 5, followed by k-NN 7, and k-NN 3. Meanwhile, Trees.J48 variations, with confidence factors of 0.25, 0.50, and 0.75, depicted accuracies between 73.3668% to 74.2881%, and Kappa statistics consistent around 0.4673 to 0.4858, indicating moderate agreement beyond chance in their predictive capacities. The highest accuracy was found in the confidence factor of 0.25, followed by 0.50, and 0.75. The result of Five folds cross-validation is displayed in Table 3.

Table 3 Result of five-fold cross-validation

Classifiers	Variant	Classification Accuracy (5 folds)	κ
NaiveBayes		68.258%	0.3652
Functions.Logistics		72.8643%	0.4573
Lazy.lBK (k-NN)	3	73.8693%	0.4774
Lazy.lBK (k-NN)	5	74.3719%	0.4874
Lazy.lBK (k-NN)	7	74.1206%	0.4824
Trees.J48	0.25	74.2881%	0.4858
Trees.J48	0.50	73.8693%	0.4774
Trees.J48	0.75	73.3668%	0.4673

CONCLUSION

The results of this study represent a notable progress in utilizing machine learning methods to forecast the success of startups. The varied performance results of different classifiers highlight the intricate intricacy involved in simulating the progress of startup enterprises. Trees.J48 demonstrated remarkable accuracy rates, varying from 89.2% to 94.3%, depending on the confidence factor. The exceptional performance of Trees.J48 is demonstrated by its best agreement beyond chance in kappa statistics. This emphasizes the strong predictive potential of Trees.J48 and establishes it as a standard for startup success prediction models.

The study employed classifiers such as NaiveBayes and Functions.Logistics, albeit displaying a reasonable level of precision, further exemplifies the complex and diverse characteristics of startup ecosystems. The fair to moderate agreement values of these classifiers indicate the difficulties encountered in capturing the dynamic interplay of elements that influence startup results. The Lazy.lBK (k-NN) variants, by varying the number of neighbors, provided valuable information about how the complexity of the model affects the quality of classification. This analysis revealed a subtle trade-off between precision and processing efficiency.

RECOMMENDATIONS

This study provides a strong argument for decision-makers and investors in the startup ecosystem to incorporate machine learning algorithms into their strategic evaluations and investment choices. Due to the impressive precision rates exhibited by models such as Trees.J48, it is advisable for stakeholders in the startup industry to embrace these sophisticated predictive tools in order to obtain more profound understanding of the potential success of startup enterprises. By harnessing the predictive capabilities of these algorithms, investors may effectively evaluate the feasibility of startups, thereby improving their investment portfolios and reducing risks associated with initiatives in their early stages. This approach not only increases the likelihood of successful investments but also adds to a more dynamic and sustainable startup ecosystem.

For startup decision-makers, such as founders and management teams, utilizing these machine learning models can play a crucial role in strategic planning and enhancing operational efficiency. Startups can enhance their performance by prioritizing the essential characteristics that contribute to success, as determined by these models. These factors may include effective management practices, strategic market positioning, innovative product development, and customer engagement methods. Moreover, the knowledge obtained from machine learning analyses might stimulate advancements, directing new businesses in creating distinctive value propositions and investigating unexplored market segments.

Startups should utilize these results to effectively convey their promise to investors and stakeholders, hence increasing their likelihood of obtaining funding and strategic alliances. Essentially, the incorporation of machine learning into the decision-making process signifies a transition towards a data-centric, analytical approach in the startup realm, promoting a culture of innovation and well-informed decision-making.

REFERENCES

1. Allioui, H., & Mourdi, Y. (2023). Unleashing the potential of AI: Investigating cutting-edge technologies that are transforming businesses. *International Journal of Computer Engineering and Data Science*, 3(2), 1-12. https://ijceds.com/ijceds/article/view/59
2. Antretter, T., Blohm, I., Grichnik, D., & Wincent, J. (2019). Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy. *Journal of Business Venturing Insights, 11*, e00109. https://doi.org/10.1016/j.jbvi.2018.e00109
3. Bangdiwala, M., Mehta, Y., Agrawal, S., & Ghane, S. (2022). Predicting Success Rate of Startups using Machine Learning Algorithms. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, 1–6. https://doi.org/10.1109/ASIANCON55314.2022.9908921
4. Basole, R. C., Park, H., & Seuss, C. D. (2023). Complex business ecosystem intelligence using AI-powered visual analytics. *Decision Support Systems*, 114-133. https://doi.org/10.1016/j.dss.2023.114133
5. Bharadiya, J. P. (2023a). A comparative study of business intelligence and artificial intelligence with big data analytics. *American Journal of Artificial Intelligence*, *7*(1), 24. ttps://doi.org/10.11648/j.ajai.20230701.14
6. Bharadiya, J. P. (2023b). Leveraging machine learning for enhanced business intelligence. *International Journal of Computer Science and Technology*, *7*(1), 1-19. https://www.ijcst.com.pk/index.php/IJCST/article/view/234
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953
8. El Hajj, M., & Hammoud, J. (2023). Unveiling the influence of artificial intelligence and machine learning on financial markets: A comprehensive analysis of AI applications in trading, risk management, and financial operations. *Journal of Risk and Financial Management*, *16*(10), 434. https://doi.org/10.3390/jrfm16100434
9. Graham, B., & Bonner, K. (2022). One size fits all? Using machine learning to study heterogeneity and dominance in the determinants of early-stage entrepreneurship. *Journal of Business Research*, *152*, 42-59. https://doi.org/10.1016/j.jbusres.2022.07.043
10. Hasan, Y., & Jain, A. (2023). Anticipating company success or failure using machine learning abstract: Model. *Journal of Emerging Technologies and Innovative Research*, *10*(5), m584–m604.
11. Janáková, H. (2015). The Success Prediction of the Technological Start–up Projects in Slovak Conditions. *Procedia Economics and Finance*, *34*, 73–80. https://doi.org/10.1016/S2212-5671(15)01603-2
12. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415
13. Kim, J., Kim, H., & Geum, Y. (2023). How to succeed in the market? Predicting startup success using a machine learning approach. *Technological Forecasting and Social Change*, *193*, 122614. https://doi.org/10.1016/j.techfore.2023.122614
14. Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481-504. https://doi.org/10.1016/j.ijresmar.2020.04.005
15. Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020
16. Martinez, I., Viles, E., & Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research*, *24*, 100183. https://doi.org/10.1016/j.bdr.2020.100183
17. Piskunova, O., Ligonenko, L., Klochko, R., Frolova, T., & Bilyk, T. (2022). Applying Machine Learning Approach to Start-up Success Prediction. *Scientific Horizons*, *24*(11), 72–84. https://doi.org/10.48077/scihor.24(11).2021.72-84
18. Svabova, L., Michalkova, L., Durica, M., & Nica, E. (2020). Business failure prediction for Slovak small and medium-sized companies. *Sustainability*, *12*(11), 4572. https://doi.org/10.3390/su12114572
19. Ünal, C., & Ceasu, I. (2019). A Machine Learning Approach Towards Startup Success Prediction. *IRTG 1792 Discussion Paper*, 2019–022. https://www.econstor.eu/bitstream/10419/230798/1/irtg1792dp2019-022.pdf
20. Vafaei, N., Ribeiro, R. A., & Matos, L. M. C. (2018). Data normalisation techniques in decision making: Case study with TOPSIS method. *International Journal of Information and Decision Sciences*, *10*(1), 19. https://doi.org/10.1504/IJIDS.2018.090667
21. Vasquez, E., Santisteban, J., & Mauricio, D. (2023). Predicting the success of a startup in information technology through machine learning. *International Journal of Information Technology and Web Engineering*, *18*(1), 1–17. https://doi.org/10.4018/IJITWE.323657
22. Von Gelderen, M., Frese, M., & Thurik, R. (2000). Strategies, uncertainty and performance of small business startups. *Small Business Economics*, *15*, 165-181. https://doi.org/10.1023/A:1008113613597
23. Yin, D., Li, J., & Wu, G. (2021). Solving the Data Sparsity Problem in Predicting the Success of the Startups with Machine Learning Methods. *arXiv:2112.07985v1 [cs.LG] Preprint*, 1-24. https://doi.org/10.48550/ARXIV.2112.07985.