OPTIMAL WAY FOR PREDICTING START-UP COMPANY SUCCESS RATES USING MACHINE LEARNING

Ananda R Kumar Mukkala[1], Pinisetty Hemasri[2], Panneru Yamuna[3], K Prashanth[4], Inala Akash[5]

[1]Associate Professor, Dept of CSE, Sreyas Institute of Engineering and Technology, Email: anandranjit@sreyas.ac.in

[2]Ug scholar, Sreyas Institute of Engineering and Technology, Email: pinisettyhemasri@gmail.com

[3]Ug scholar, Sreyas Institute of Engineering and Technology, Email: panneruyamuna2003@gmail.com

[4]Ug scholar, Sreyas Institute of Engineering and Technology, Email: prashanth05012003@gmail.com

[5]Ug scholar, Sreyas Institute of Engineering and Technology, Email: Panduinala18@gmail.com

Corresponding Author: - Ananda R Kumar Mukkala

Associate professor, Dept of CSE,Sreyas Institute of Engineering and Technology, Email: anandranjit@sreyas.ac.in

Abstract: *Predicting the success of start-up companies has become a critical challenge, as investors often risk significant capital without understanding potential outcomes. To address this issue, a machine learning-based approach is proposed to predict start-up success using key parameters like funding, participants, and milestones. The dataset used for this analysis is sourced from Kaggle's "Startup Success Prediction" repository, which provides valuable information for training the models. Data preprocessing steps, including handling missing values, data imbalance correction via the ADASYN algorithm, and feature normalization, were implemented. Multiple machine learning algorithms were employed and evaluated, such as Random Forest, KNN, SVM, Naïve Bayes, and Logistic Regression. Among these, Random Forest achieved the highest accuracy of over 94%, demonstrating superior performance compared to other models. The models were evaluated using metrics like accuracy, precision, recall, F1 Score, confusion matrix, and ROC curve analysis. Data visualizations were used to explore trends in funding, company status, and category distributions. The final model was deployed using Flask, allowing users to input data for real-time predictions and receive actionable recommendations for improving start-up success.*

Index Terms: *Machine Learning, start-up; sustainability; forecasting; artificial intelligence; natural language processing.*

1. INTRODUCTION

In recent years, startup entrepreneurship has emerged as a crucial driver of technological, scientific, and business innovation. Companies like AirBnb and Uber, which began as small startups, have now disrupted the global economy through their innovative business models, products, and services. As Steve Blank [11] defines it, a startup is a temporary organization in search of a viable, sustainable, and scalable business model. The growth and success of startups heavily depend on a supportive ecosystem consisting of institutional investors, such as venture capital (VC) and angel investors, along with organizations like accelerators and incubators. Additionally,

the macroeconomic environment, including interest rates and consumer sentiment, significantly influences the viability of these ventures.

However, despite the rising number of startups, only a small fraction—approximately 1 in 10—achieve long-term profitability and sustainability beyond five years [12]. This low success rate highlights the inherent risks involved in startup ventures. Researchers and practitioners have identified several key success factors, such as the quality of the founding team, the timing of the venture, the desirability of the product or service, and the market being targeted [13][14]. These factors often determine whether a startup can break through the initial stages and secure its position in the market.

As the number of startups continues to rise, they are generating vast amounts of data from landing pages, presentations, business plans, and social media activity. Moreover, startups are becoming more specialized and complex, adding another layer of challenge for investors and experts attempting to assess their potential. At the same time, personality traits play a significant role in entrepreneurship, as evidenced by models like the OCEAN (Big Five) Model. This model suggests that traits such as openness, conscientiousness, and extroversion, along with low neuroticism, are associated with successful entrepreneurship, as they enhance risk tolerance, innovation, proactivity, and resilience [15][16].

The rapid growth of startups and the massive amounts of data they produce present challenges for institutional investors. Given the increasing number of specialized companies, experts often struggle to evaluate startups effectively. The limited domain knowledge across diverse industries, coupled with the complexity of modern business models, makes startup evaluation both error-prone and costly. As a result, the process of evaluating startups has become subjective, time-consuming, and prone to human error. These challenges are not unique to startups but extend across multiple fields, including finance, healthcare, and technology. Yet, the advent of artificial intelligence (AI) offers a promising solution. AI can process large datasets rapidly, without the limitations and biases of human experts, offering data-driven insights and improving decision-making in startup evaluation [7][8]. The integration of AI in this domain presents an opportunity to overcome the challenges of traditional evaluation methods, revolutionizing the startup ecosystem.

2. LITERATURE SURVEY

The success of startups has been an area of extensive research due to the inherent challenges of new ventures and the growing importance of entrepreneurship in economic and technological innovation. In their 2017 study, Spender et al. [1] explore how startups are increasingly leveraging open innovation models, wherein they collaborate with external entities such as universities, research centers, and other organizations. This shift toward open innovation allows startups to access resources, knowledge, and technologies beyond their internal capabilities, enhancing their chances of success. The authors also note the importance of aligning a startup's innovation strategy with market needs, which is crucial for long-term growth.

Krishna et al. [2] focus on predicting the success or failure of startups by analyzing key indicators that contribute to startup outcomes. Their research highlights the predictive power of factors such as early-stage funding, team dynamics, market demand, and the entrepreneurial ecosystem. The authors propose using machine learning algorithms to predict the future performance of startups, a notion that ties into the growing role of data analytics

and AI in entrepreneurship. According to their findings, startups that receive adequate early-stage funding and that have a well-balanced, experienced founding team are more likely to succeed. This insight underscores the need for a comprehensive evaluation framework that incorporates various parameters that determine startup success.

Another important aspect of startup success is the quality of the founding team, which is discussed by Diakanastasi et al. [3]. They emphasize that entrepreneurial team dynamics, including the skills, experience, and communication abilities of the founding team members, play a significant role in the creation and sustainability of new ventures. Their study, conducted within a startup incubator, reveals that teams with complementary skills and experience are better positioned to adapt to challenges and capitalize on opportunities. This idea resonates with the findings of Eliakis et al. [4], who further emphasize the importance of leadership, team composition, and adaptability in ensuring a startup's survival and growth in the competitive tech entrepreneurship space.

Personality traits also significantly impact startup success, as noted by Jang et al. [5]. In their study, they discuss the heritability of the Big Five personality traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—and their influence on entrepreneurial success. The study suggests that individuals with high openness, conscientiousness, and emotional stability tend to be more successful entrepreneurs, as these traits enhance their ability to deal with uncertainty, take risks, and execute business ideas effectively. Rauch [6] extends this discussion by conducting a meta-analysis that highlights the correlation between business owners' personality traits and business success. His findings support the notion that entrepreneurs who exhibit strong leadership skills, resilience, and adaptability are more likely to build successful ventures, which has profound implications for understanding the role of personal characteristics in entrepreneurship.

The importance of the macroeconomic environment in shaping startup success is highlighted in Cader and Leatherman's [7] research on small business survival. Their study focuses on how economic factors such as interest rates, inflation, and consumer sentiment influence the ability of startups to survive and grow. They argue that these macroeconomic indicators can provide important insights into the likelihood of a startup's success, particularly in industries that are highly sensitive to economic fluctuations. For instance, startups operating in sectors such as technology and healthcare may be more affected by economic downturns compared to those in more stable industries. This insight is particularly valuable for investors and policymakers who need to understand the external factors that affect the viability of new ventures.

Delmar [8] investigates the relationship between the experience of founding teams and the survival and success of startups. His study concludes that prior experience in founding ventures significantly enhances the chances of success. Experienced founders are better equipped to handle the complexities of starting and running a business, as they have learned from past failures and successes. Delmar also suggests that the experience of the team members in relevant industries or fields of technology is crucial for the innovation and competitive advantage of the startup. This supports the notion that not just the entrepreneurial drive, but also the technical expertise and industry knowledge of the founders, plays a critical role in the sustainability of a startup.

In addition to these studies, various other researchers have investigated the role of institutional support in enhancing the likelihood of startup success. Incubators, accelerators, and venture capitalists have been shown to provide crucial mentorship, funding, and access to networks that can significantly increase the chances of a

startup's success. For example, recent studies have examined how startups that are part of an incubator or accelerator program benefit from shared resources and the guidance of experienced entrepreneurs. These institutions often help mitigate the high risk associated with startup ventures by providing structured support and fostering an environment conducive to innovation.

Another area of interest in the literature is the increasing use of data analytics and artificial intelligence (AI) to predict startup success. Researchers have explored how machine learning models can be trained to analyze startup data—such as funding amounts, team composition, and market trends—to identify patterns that correlate with future success. Such AI-driven approaches are gaining traction, as they can provide more objective, data-driven insights into the likelihood of a startup's success, reducing the subjectivity and biases often associated with traditional methods of evaluation.

While traditional approaches to evaluating startups have relied heavily on the experience and intuition of investors, the growing complexity of startups and the vast amounts of data they generate have made it increasingly difficult for human experts to keep up. As a result, AI and machine learning are being proposed as critical tools to assist investors in making informed decisions. These technologies can process vast amounts of data in real time, allowing for faster and more accurate assessments of startups. Furthermore, machine learning models can continuously adapt to changing market conditions, providing ongoing insights into the factors that influence startup success.

3. MATERIALS AND METHODS

The proposed system aims to predict the success of start-up companies using machine learning algorithms, providing investors with insights to make informed decisions. The system uses a dataset from Kaggle's "Startup Success Prediction" repository, containing features such as funding, participants, and milestones. After preprocessing the data to handle missing values, address class imbalance using the ADASYN algorithm, and normalize features, the system trains several machine learning models, including Random Forest, KNN, SVM, Naïve Bayes, and Logistic Regression. The performance of these models is evaluated based on accuracy, precision, recall, F1 Score, and confusion matrix. The Random Forest model, with the highest accuracy, is then deployed through a Flask web application, enabling real-time predictions and providing recommendations for improving start-up success.

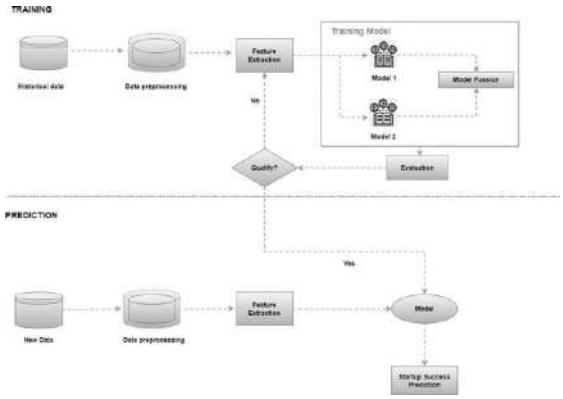

Fig.1 Proposed Architecture

The architecture outlines a machine learning model for startup success prediction. During training, historical data undergoes preprocessing and feature extraction, feeding into two models. Model fusion combines their outputs, followed by evaluation. For new data, similar preprocessing and feature extraction are performed, and the qualified model predicts startup success. The system iterates between training and prediction, refining the model over time.

a) Dataset Collection:

The dataset used for this study was sourced from the Kaggle repository, specifically the "Startup Success Prediction" dataset. It includes information about various startup companies, such as industry, funding, location, and performance metrics. The data helps in training and testing machine learning algorithms to predict the success or failure of startups. This dataset is crucial for identifying key factors that influence a startup's likelihood of success in a competitive and dynamic business environment.

	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	Unnamed: 6	name	labels	...	object_id	has_VC	has_angel	has_round
0	1005	CA	42.358880	-71.056820	92101	c:6669	San Diego	NaN	Bandsintown	1	...	c:6669	0	1	
1	204	CA	37.238916	-121.973718	95032	c:1628	Los Gatos	NaN	TriCipher	1	...	c:1628	1	0	
2	1001	CA	32.901049	-117.192656	92121	c:6620	San Diego	San Diego CA 92121	Plixi	1	...	c:6620	0	0	
3	738	CA	37.320309	-122.050040	95014	c:4268	Cupertino	Cupertino CA 95014	Solidcore Systems	1	...	c:4268	0	0	
4	1002	CA	37.779281	-122.419236	94105	c:6806	San Francisco	San Francisco CA 94105	Inhale Digital	0	...	c:6806	1	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
918	352	CA	37.749594	-122.376471	94107	c:2134	San Francisco	NaN	DoTweet	1	...	c:2134	0	0	

Fig.2 Dataset

b) Pre-Processing:

i) Handling Missing Values; In the initial stage of the pre-processing pipeline, we examined the dataset for missing values. Since the dataset did not contain any missing data, no additional steps were required for imputation or removal of missing values. This ensured that the algorithms could operate without the need for any further adjustments, streamlining the process for more efficient training and evaluation of machine learning models.

ii) Data Imbalance Handling: The dataset exhibited an imbalance between the success and failure classes, which could negatively impact the model's performance. To address this, we employed the ADASYN (Adaptive Synthetic Sampling) algorithm. This technique generates synthetic data for the minority class, thereby balancing the class distribution. After applying ADASYN, the dataset size increased from 923 to 1172 instances, enhancing the model's ability to generalize and reducing bias towards the majority class.

iii) Feature Normalization: For optimal performance of the machine learning models, it was crucial to normalize the feature values. The dataset's features, such as funding amount and number of participants, had varying scales, which could distort the performance of certain algorithms. Using standard normalization techniques, we scaled the training features to a uniform range. This step improved the convergence rate and accuracy of the algorithms, ensuring a more reliable prediction outcome.

iv) Feature Extraction: Feature extraction involved identifying the relevant attributes that would contribute to the model's predictive power. We extracted key features from the dataset such as the amount of investment, industry category, and company status. These features were selected based on their significance to the success or failure of startups. After extracting these features, they were normalized to prepare the data for training and testing machine learning models.

c) Training and Testing:

To assess the performance of the machine learning models, the dataset was divided into training and testing subsets. We used an 80-20 split, where 80% of the data was used for training the models, and the remaining 20% was reserved for testing. This approach allowed us to evaluate how well the models generalized to unseen data, providing a reliable estimate of their performance and ensuring that the models were not overfitting.

d) Algorithms:

Random Forest: This ensemble method builds multiple decision trees to improve prediction accuracy. It handles high-dimensional data well and minimizes overfitting, delivering robust predictions for start-up success classification.

KNN: The K-Nearest Neighbors algorithm classifies start-ups based on the proximity to the closest training instances. It's used to predict success by evaluating similarities between data points in the feature space.

SVM: Support Vector Machine constructs hyperplanes to separate success and failure categories. It's effective for high-dimensional data, focusing on maximizing margins between classes for more accurate predictions.

Naïve Bayes: This probabilistic classifier predicts success by calculating the likelihood of each class based on feature independence assumptions. It's efficient for handling large datasets with categorical features.

Logistic Regression: Logistic Regression models the probability of success using a linear combination of input features. It's used for binary classification and provides insights into the influence of individual features on the outcome.

4. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} (1)$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (2)$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} (3)$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1\ Score = 2 * \frac{Recall\ \mathrm{X}\ Precision}{Recall + Precision} * 100 (1)$$

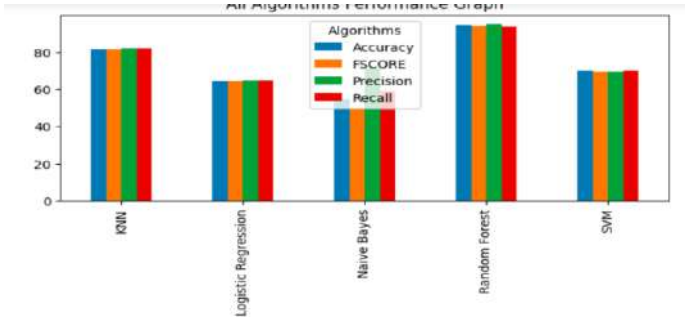

Fig.3 Comparison Graphs

	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	Random Forest	94.468085	95.212766	93.901099	94.335144
1	Logistic Regression	64.255319	64.654545	64.761905	64.239130
2	SVM	69.787234	69.722544	69.945055	69.679623
3	Naive Bayes	54.893617	71.307798	59.047619	49.468517
4	KNN	81.702128	81.952706	82.271062	81.680898

Fig.4 Performance Evaluation Table

Fig.5 Home Page

Fig.6 Start-Up Success Prediction Screen

Fig.7 Final Outcome

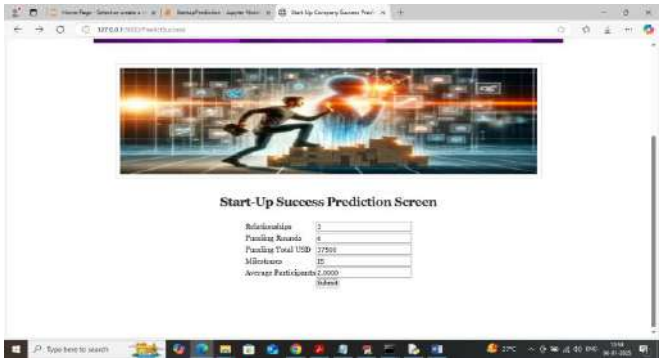

Fig.8 Upload Input Data

Fig.9 Predicted Results

5. CONCLUSION

In conclusion, the proposed machine learning-based approach effectively predicts the success of start-up companies, offering valuable insights for investors to make data-driven decisions. By utilizing key parameters like funding, participants, and milestones, the system analyzes a comprehensive dataset from Kaggle's "Startup Success Prediction" repository. Through meticulous preprocessing, including handling missing values, balancing the dataset with the ADASYN algorithm, and normalizing features, the system ensures reliable data for model training. The evaluation of multiple machine learning algorithms—Random Forest, KNN, SVM, Naïve Bayes, and Logistic Regression—demonstrated that the Random Forest algorithm achieved the highest accuracy, surpassing 94%, making it the most reliable model for predicting start-up success. Other performance metrics such as precision, recall, and F1 Score further supported the superiority of Random Forest. The model was successfully deployed using Flask, enabling real-time predictions and offering actionable recommendations for improving start-up success. This system provides a robust tool for start-up founders and investors, helping them understand key factors influencing success and mitigating risks associated with blind investments in new ventures. The **future scope** of this system includes expanding the dataset to incorporate additional features such as market trends, competitor analysis, and geographical factors to enhance prediction accuracy. Integration of deep learning techniques, like neural networks, could further improve model performance. Additionally, incorporating real-time data updates and interactive dashboards could allow investors to monitor evolving start-up metrics. The system could also be extended to predict long-term success beyond initial funding stages.

REFERENCES

[1] Spender, J.C.; Corvello, V.; Grimaldi, M.; Rippa, P. Startups and open innovation: A review of the literature. Eur. J. Innov. Manag. 2017, 20, 4–30.
[2] Krishna, A.; Agrawal, A.; Choudhary, A. Predicting the Outcome of Startups: Less Failure, More Success. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 798–805.
[3] Diakanastasi, E.; Karagiannaki, A.; Pramatari, K. Entrepreneurial Team Dynamics and New Venture Creation Process: An Exploratory Study within a Start-Up Incubator. Sage Open 2018, 8, 2158244018781446.
[4] Eliakis, S.; Kotsopoulos, D.; Karagiannaki, A.; Pramatari, K. Survival and Growth in Innovative Technology Entrepreneurship: A Mixed-Methods Investigation. Adm. Sci. 2020, 10, 39.
[5] Jang, K.L.; Livesley, W.J.; Vernon, P.A. Heritability of the big five personality dimensions and their facets: A twin study. J. Personal. 1996, 64, 577–591.
[6] Rauch, A. Let's Put the Person Back into Entrepreneurship Research: A Meta-Analysis on the Relationship between Business Owners' Personality Traits, Business Creation, and Success. Eur. J. Work. Organ. Psychol. 2007, 16, 353–385.
[7] Cader, H.; Leatherman, J. Small business survival and sample selection bias. Small Bus. Econ. 2011, 37, 155–165.
[8] Delmar, F. Does Experience Matter? The Effect of Founding Team Experience on the Survival and Sales of Newly Founded Ventures. Strateg. Organ. 2006, 4, 215–247.
[9] Liapis, G.; Zacharia, K.; Rrasa, K.; Liapi, A.; Vlahavas, I. Modelling Core Personality Traits Behaviours in a Gamified Escape Room Environment. In Proceedings of the 16th European Conference on Games Based Learning, Lisbon, Portugal, 6–7 October 2022; Volume 16, p. 731.
[10] Liapis, G.; Zacharia, K.; Rrasa, K.; Vlahavas, I. Serious Escape Room Game for Personality Assessment. In Proceedings of the 12th International Conference on Games and Learning Alliance, Dublin, Ireland, 29 November–1 December 2023; pp. 420–425.
[11] Żbikowski, K.; Antosiuk, P. A machine learning, bias-free approach for predicting business success using Crunchbase data. Inf. Process. Manag. 2021, 58, 102555.
[12] Sadatrasoul, S.M.; Ebadati, O.; Saedi, R. A Hybrid Business Success Versus Failure Classification Prediction Model: A Case of Iranian Accelerated Start-ups. J. Data Min. 2020, 8, 279–287.
[13] Dellermann, D.; Lipusch, N.; Ebel, P.; Popp, K.M.; Leimeister, J.M. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. arXiv 2021, arXiv:2105.03360.
[14] Pan, C.; Gao, Y.; Luo, Y. Machine learning prediction of companies' business success. In CS229: Machine Learning, Fall 2018; Stanford University: Stanford, CA, USA, 2018.
[15] Ünal, C.; Ceasu, I. A Machine Learning Approach towards Startup Success Prediction. IRTG 1792 Discussion Papers 2019-022, Humboldt University of Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series". 2019. Available online: https://www.econstor.eu/bitstream/10419/230798/1/irtg1792dp2019-022.pdf (accessed on 2 September 2024).
[15] Byworth, C. Measuring Personality Constructs: The Advantages and Disadvantages of Self-Reports, Informant Reports and Behavioural Assessments. Enquire 2008, 1, 75–94.

[16] Jospin, L.V.; Laga, H.; Boussaid, F.; Buntine, W.; Bennamoun, M. Hands-on Bayesian neural networks—A tutorial for deep learning users. IEEE Comput. Intell. Mag. 2022, 17, 29–48.