A Venture Capital-Centric Hybrid Machine Learning Framework for Predicting Startup Success in the Us Ecosystem

[1]Anamika Shukla Sharma, [2]H.S. Hota, [3]Ujjwal Matoliya

[1]Govt. E.R.Rao PG Science College, Bilaspur, India, Email: anamikashuklacs@gmail.com Atal Bihari
[2]Vajpayee Vishwavidyalaya, Bilaspur, India, Email: proffhota@gmail.com
[3]Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur, India, Email:raveenarajak21@gmail.com

Abstract
Venture capitalists facilitate the acquisition of funding for projects, ventures, or causes by aggregating contributions from a substantial pool of investors. This approach enhances the accessibility of financial resources, empowering creators with direct oversight of their funding, devoid of intermediaries such as financial institutions. The adoption of this funding model encourages innovation, bolsters local enterprises, and enables the realization of novel ideas that traditionally struggle to secure financing through conventional methods. By employing an extensive dataset encompassing funding details, feasibility assessments, networking opportunities, and geographical data pertaining to numerous startups, we advance to the stages of analysis and feature engineering, conducting exploratory analysis to discern the various determinants of success. Models were developed using Gradient Boosting, Ada Boost, XGBoost, Support Vector Machine, Random Forest and Light Gradient Boosting Machine. This research work also elucidates significant factors that influence organizational performance and entrepreneurial success, thereby offering stakeholders an innovative perspective on strategic investment management.

Keywords: Machine Learning (ML), Venture Capital (VC), Startup success Prediction.

1. Introduction
New ventures are critical in fostering novelty, employment, and growth in an economy. Nevertheless, not all startups get the intended success, and many businesses do not survive in such a competitive market. It holds a big problem for investors and venture capital since they must find a good startup which can thrive and bring big profits. Hence the need to adopt a more reliable technique of analyzing the success of startups since this would be very helpful in making the right investment decisions. Venture capital plays a crucial role in supporting early-stage startups by providing not only funding but also strategic guidance and access to valuable networks, which are essential for their growth and success. This support is instrumental in helping startups refine their business models, scale operations, and ultimately achieve their growth objectives. Furthermore, venture capitalists often leverage their expertise to mentor startup founders, fostering an environment conducive to innovation and sustainable growth. This mentorship and support system is vital for startups to navigate challenges and seize opportunities in a competitive market, ultimately enhancing their chances of long-term success.

Research on venture capital-centric hybrid machine learning models for startup success prediction has emerged as a critical area of inquiry due to the high failure rates of startups and the increasing availability of diverse data sources enabling advanced analytics(Potanin et al., 2023; Ferrati & Muffatto, 2021). Over recent years, the field has evolved from traditional statistical methods to sophisticated Machine Learning (ML) that integrate financial, social, and network data(Cao et al., 2023; Lyu et al., 2021). This evolution reflects the practical significance of improving investment decisions in venture capital, where global startup investments exceeded US$445 billion in 2022, yet returns often lag broader market indices (Setty et al., 2024; Ferrati & Muffatto, 2021). Accurate prediction models can enhance resource allocation, reduce investment risk, and foster economic growth by identifying high-potential startups early (Jafari et al., 2025; Huang et al., 2024). Despite advances, predicting startup success remains challenging due to data limitations, market volatility, and the complex interplay of qualitative and quantitative factors(Gao & Xiao, 2025; Dellermann et al., 2021). Existing research reveals a fragmented landscape with gaps in integrating heterogeneous data types, addressing class imbalance, and incorporating relational information among startups and investors(Park et al., 2024; Zhang et al., 2021; Xu et al., 2023). Controversies persist regarding the relative importance of founder characteristics versus funding history, the efficacy of cost-sensitive models versus traditional classifiers, and the interpretability of black-box algorithms(Ozince & Ihlamur, 2024; Setty et al., 2024; Żbikowski & Antosiuk, 2021). These unresolved issues hinder the development of robust, generalizable models, potentially leading to suboptimal investment decisions and missed opportunities(Di Giannantonio et al., 2022).

The conceptual framework underpinning defines startup success prediction as the application of ML models to forecast key outcomes such as funding milestones, exits, or sustainability(Potanin et al., 2023; Jafari et al., 2025). Hybrid models combine quantitative financial metrics, qualitative founder attributes, and network-based relational data to capture the multifaceted nature of startup ecosystems(Gao & Xiao, 2025; Xu et al., 2023). This framework aligns with theoretical foundations in entrepreneurial finance and ML interpretability, emphasizing the integration

of diverse data sources and the mitigation of biases to improve predictive accuracy and decision support(Ferrati & Muffatto, 2021).There has been an extensive amount of research done on how ML can be effective in determining the success of startups. Krishna et al., 2016 employed decision trees(DT) and random forests(RF) to forecast the probability of success of projects submitted at Kickstarter with criteria including duration of projects, funding goals, and the number of pledges. These studies clearly show the ability of ML in offering important clues as to what a likelihood of success in startups can be. However, most of the existing studies cover only certain domains or areas and there is a lack of attempts to provide overall results based on rich feature set and using more sophisticated ML approaches.

As the venture capital landscape evolves, the integration of ML techniques becomes increasingly vital for optimizing investment strategies and enhancing startup success prediction. This integration not only accelerates the assessment process but also empowers venture capitalists to make more informed decisions, ultimately benefiting both investors and startups in the long run. The evolving landscape of venture capital necessitates a deeper understanding of how ML can further enhance investment strategies and improve outcomes for both investors and startups. By employing ML models, venture capitalists can better manage the inherent uncertainties of investing in early-stage companies, ultimately leading to improved portfolio performance and risk mitigation. The integration of ML techniques, such as Gradient Boosting, Ada Boost, XGBoost, Support Vector Machine(SVM), Random Forest(RF) and Light Gradient Boosting Machine(LightGBM), is crucial for enhancing the predictive accuracy of startup success models in the venture capital landscape. ML techniques have transformed predictive analytics by enabling more accurate assessments of various outcomes, including startup success, through data-driven methodologies. The application of ML in predictive analytics not only enhances the accuracy of startup success predictions but also empowers venture capitalists to make informed investment decisions based on comprehensive data analysis. The ongoing advancements in ML algorithms further demonstrate their potential to revolutionize predictive analytics in the venture capital sector. The integration of these advanced algorithms into venture capital decision-making processes signifies a pivotal shift towards data-driven investment strategies, enhancing the overall effectiveness of startup evaluations. This evolution underscores the necessity for venture capitalists to continually adapt their methodologies, ensuring they remain competitive and effective in identifying high-potential startups. Authors (Shi et al., 2024) have demonstrated that how ML techniques can significantly enhance predictive models, ultimately leading to better investment outcomes for venture capitalists. By leveraging these advanced methodologies, investors can navigate the complexities of the startup ecosystem with greater confidence and precision. In recent years the application of ML has gained traction, offering innovative solutions to the challenges faced by venture capitalists in assessing startup viability and success.

Hybrid ML models are characterized by their ability to integrate multiple algorithms, enhancing predictive performance and robustness in complex scenarios. This approach allows for the combination of strengths from different models, ultimately leading to improved accuracy in predictions. The utilization of hybrid machine learning models is particularly advantageous in the venture capital sector, where diverse data sources and complex patterns necessitate sophisticated analytical approaches. The implementation of hybrid models can significantly enhance predictive performance, providing venture capitalists with a more robust framework for evaluating startup success and investment opportunities. The adoption of hybrid machine learning models in venture capital can facilitate better decision-making by leveraging diverse data sources and improving the accuracy of startup success predictions. The integration of hybrid machine learning models not only enhances predictive capabilities but also allows investors to adapt to the dynamic nature of the startup ecosystem, ultimately leading to more informed investment strategies. The exploration of hybrid machine learning models in venture capital is essential for refining predictive analytics and enhancing investment strategies in an increasingly complex startup landscape. The exploration of hybrid machine learning models represents a crucial advancement in venture capital, enabling investors to navigate the complexities of startup evaluations more effectively.

The potential of hybrid ML models to revolutionize venture capital decision-making is significant, as they offer a comprehensive approach to analyzing diverse datasets and improving predictive accuracy. Furthermore, the integration of hybrid ML models in venture capital not only enhances predictive accuracy but also fosters a deeper understanding of the multifaceted nature of startup ecosystems. By incorporating diverse datasets, including market trends, consumer behavior, and competitive analysis, these models can uncover intricate patterns that traditional methods might overlook. This comprehensive approach allows venture capitalists to tailor their investment strategies more precisely, thereby improving their ability to identify high-potential startups amidst the noise of the competitive landscape. As the venture capital sector continues to evolve, leveraging advanced techniques such as Adaptive Multi-Layer Ensemble Learning (AMEL) can further refine predictive capabilities, demonstrating a 5-10% improvement in accuracy over conventional methods (Saminathan, 2024). Ultimately, this evolution underscores the necessity for investors to embrace innovative methodologies that not only enhance their decision-making processes but also adapt to the dynamic challenges of the startup environment. Furthermore, as the venture capital landscape becomes increasingly reliant on hybrid ML models, it is essential for investors to consider the ethical implications of their predictive analytics. As venture capitalists increasingly integrate hybrid ML models into their decision-making processes, the importance of fostering diversity within datasets cannot be overstated. By ensuring that these models are trained on representative data, investors can mitigate potential biases that may skew predictions and inadvertently disadvantage certain groups of entrepreneurs. This is particularly

crucial given the growing recognition of the need for inclusivity in the startup ecosystem, which can lead to a broader range of innovative solutions and market opportunities. Moreover, the implementation of fairness-aware algorithms can complement the predictive power of hybrid models, enabling venture capitalists to not only identify high-potential startups but also promote equitable access to funding across diverse demographic groups. Ultimately, the dual focus on predictive accuracy and ethical data practices will be essential for fostering a sustainable and inclusive venture capital landscape that supports innovation across all sectors. Authors (Shi et al., 2024) have worked on startup success prediction by utilizing ML techniques and found that integrating ML models, such as random forest and XGBoost, can substantially improve prediction accuracy, thereby aiding venture capitalists in making more informed investment decisions. The integration of ML techniques into venture capital practices is essential for enhancing predictive analytics, ultimately leading to more strategic investment decisions and improved outcomes for both investors and startups. The ongoing integration of ML techniques, particularly random forest and XGBoost, is vital for enhancing the predictive capabilities of venture capitalists, ultimately fostering better investment outcomes in the startup ecosystem. The dynamic interplay between venture capital and ML represents a transformative shift in investment strategies, enabling a more nuanced understanding of startup success factors. This transformative shift not only enhances the ability to predict startup success but also equips investors with the necessary tools to navigate an increasingly complex and competitive landscape in venture capital. The comparison of hybrid ML models with traditional approaches reveals that hybrid models often outperform single-algorithm methods in predictive accuracy, particularly in complex datasets commonly found in venture capital evaluations. The superior performance of hybrid models can be attributed to their ability to leverage the strengths of multiple algorithms while addressing the limitations of traditional methods in venture capital evaluations. The effectiveness of hybrid models in venture capital is underscored by their ability to address the complexities of diverse datasets, thus enhancing predictive accuracy in startup evaluations. This comparison highlights the necessity for venture capitalists to embrace innovative methodologies, ensuring they remain competitive in a rapidly evolving investment landscape. The exploration of hybrid ML models not only enhances predictive accuracy but also equips venture capitalists with the necessary tools to effectively manage risks associated with early-stage investments. The integration of hybrid machine learning models in venture capital not only improves predictive accuracy but also empowers investors to navigate the complexities of startup evaluations more effectively. The integration of hybrid ML models is essential for venture capitalists seeking to enhance their investment strategies and improve predictive accuracy in evaluating startup success. The exploration of hybrid models in various domains demonstrates their effectiveness in enhancing predictive capabilities and decision-making processes, which can be particularly beneficial for venture capitalists in evaluating startups. The successful application of hybrid models in fields such as healthcare and finance illustrates their potential to improve predictive accuracy and decision-making processes, a trend that venture capitalists can leverage for startup evaluations. The insights gained from these case studies underscore the transformative potential of hybrid models in various sectors, reinforcing their applicability in the venture capital landscape.

Moreover, due to the rapid stimulus of startup businesses and the appearance of new technologies, the generation of stable and flexible models for the analysis of various data and accurate prediction is needed. To fill these gaps, the current study proposes to establish a comprehensive model for a ML study that can predict the level of success of startups depending on its funding, the encompassing milestones, the relationship, and geographical location. To achieve this, we rely on a dataset with information on several startups and on modern ML techniques: Gradient Boosting, Ada Boost, XGBoost, SVM, RF and LightGBM to build and assess proposed predictive models. Furthermore, we also perform comprehensive data investigation and modeling and generate new features to optimize the understanding of how different features affect the performance of these startups. Moreover, to enhance user interaction, we have developed a frontend interface alongside the model. Startups enter essential information such as location, funding amount, and operational sector through designated forms. This data is transmitted to backend, where it is processed through ML model to predict success rates upon submission.

2.Dataset Description

This research utilizes data sourced from Kaggle. Kaggle is a data science platform that provides users with real-world data sets for ML and analytics for research. The data used in the research is a startup prediction data set from Crunchbase and includes information on entrepreneurship, startup funding, and the status of these companies' operations. The Kaggle dataset used in the research contains a total of 923 companies and 49 total features. These features include 48 attributes and 1 target variable. The target variable 'status' represents the success of a startup and is a binary variable. An 'acquired' startup is represented by 1, and an unsuccessful startup is a 0. This data is used for supervised binary classification.

The dataset features several dimensions of each startup that allows them to examine different characteristics of the companies. These characteristics can be classified in several ways. The first is geographical. These features include startup location information (state, city, latitude, and longitude) and help determine location-based innovative success. The second is funding. These features include the total funds associated with the company and the funding history (number of funding rounds, type of investors i.e. VCs and Angels) and the funding stage (Series A, B, C, and D). Table 1 displays description and categorization of features. These features help determine success patterns that are tied to investment.

The process of observing startups allows them to track their progression toward essential business objectives. The

results provide insight into their current development stage and their progression to future stages. The study of their company connections shows their business activities while their milestone achievements show their operational progress. The system provides us with different features which enable us to determine the industry classification of a startup through its software web mobile biotech or enterprise operations. The research enables us to measure startup performance across various industrial sectors through the sector-based performance assessment method.

The dataset contains real-world data characteristics which include both missing values and noise and class imbalance. A significant number of missing values are observed in attributes such as closed_at, milestone-related features, and certain auxiliary columns. The missing values occur because of two factors: ongoing startup activities and incomplete reporting and multiple source data collection. The dataset contains 597 successful startups and 326 unsuccessful startups which result in moderate class imbalance that must be addressed during model development to prevent prediction bias. The dataset becomes difficult to manage because it contains numerous features that exceed normal limits. The selection process needs to identify essential features while eliminating all non-essential elements. We need to eliminate all columns which contain unneeded data because they will negatively impact our results through their presence. The process ensures our data achieves maximum quality and trustworthiness through this evaluation method.

Table 1: Feature Description and Categorization

Feature Category	Features Included	Importance in Prediction
Geographical	State, City, Latitude, Longitude	Captures the innovation hub effect (e.g., Silicon Valley).
Financial	Total Funding, Funding Rounds (A-D), VC presence	Indicates financial stability and investor trust.
Operational	Age of startup, Milestones, Relationships	Reflects the operational maturity of the venture.
Success Metrics	Status (Acquired/Closed)	Target variable for binary classification.

3.Methodology

This study proposes a structured ML framework for predicting startup success using a multi-stage analytical pipeline as shown in Figure 1. The methodology focuses on transforming raw startup data into a robust predictive model through systematic preprocessing, feature optimization, model training, and evaluation. Each stage is designed to improve data quality, enhance feature relevance, and ensure reliable model performance.

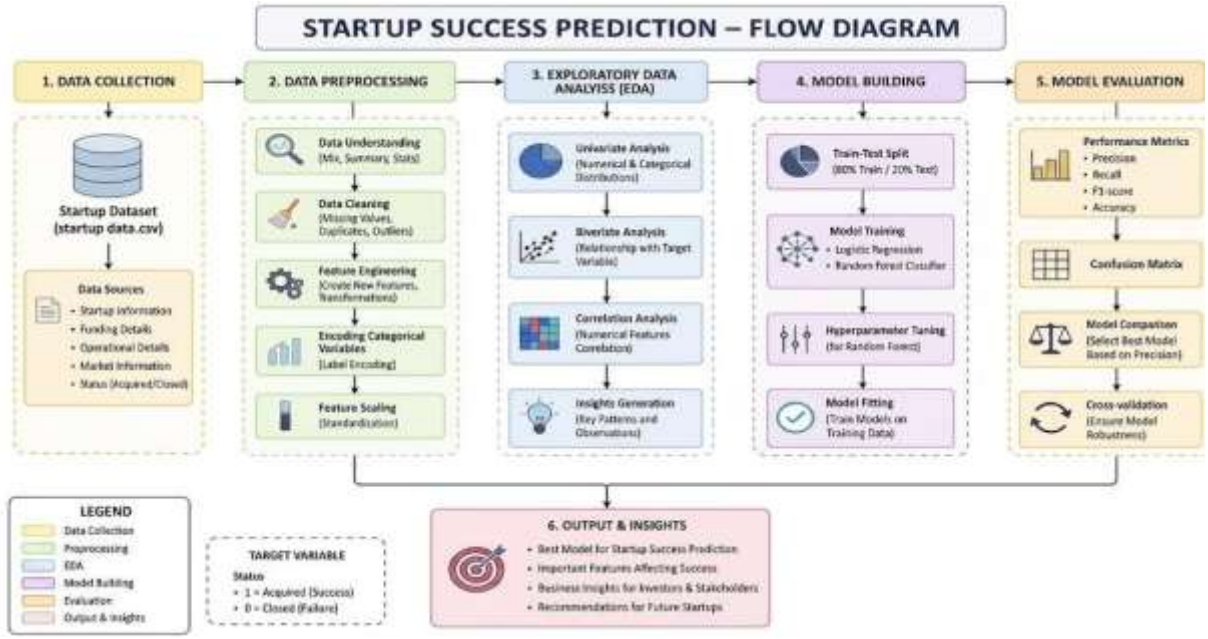


Fig. 1: Proposed Multi-Stage Machine Learning Pipeline for startup success prediction.

The framework integrates data preprocessing, advanced feature engineering, RFE-based feature selection, and a hybrid boosting-dominant learning approach for robust classification.

3.1Data Preprocessing

The preprocessing phase is crucial for ensuring data quality and consistency before model development. Initially, exploratory inspection of the dataset was conducted to understand feature distributions, detect missing values, and identify inconsistencies. Statistical summaries and distribution analysis were used to gain preliminary insights into the dataset structure.

Subsequently, data cleaning techniques were applied, including missing value imputation, duplicate removal, and outlier treatment using statistical methods such as the interquartile range and Z-score analysis. This step ensured that noise and inconsistencies were minimized,

resulting in a refined dataset suitable for further analysis.

Feature engineering was then performed to enhance the predictive capability of the dataset. Based on insights obtained from preliminary analysis, new features were created by combining related variables, particularly funding-related attributes, which individually exhibited weak correlations but demonstrated stronger predictive power when aggregated. Additionally, transformation of relationship-based variables was carried out to capture non-linear patterns in the data.

Categorical variables were converted into numerical representations using encoding techniques such as label encoding and one-hot encoding. This transformation increased the dimensionality of the dataset but made it compatible with machine learning algorithms. Following encoding, feature scaling was applied using standardization to normalize the feature distributions, ensuring that all variables contributed equally during model training.

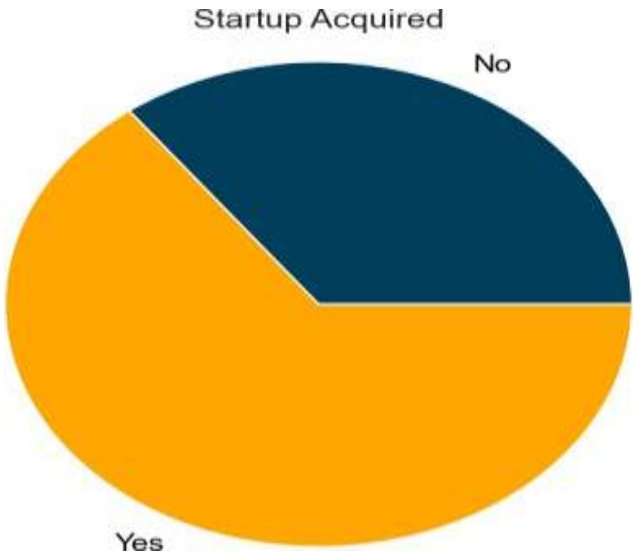


Fig. 2. Distribution of startup status within the dataset.

Figure 2 illustrates the proportion of 'Acquired' (Success) versus 'Closed' (Failure) startups, highlighting the class balance (or imbalance) that informs the subsequent data preprocessing and model evaluation strategies.

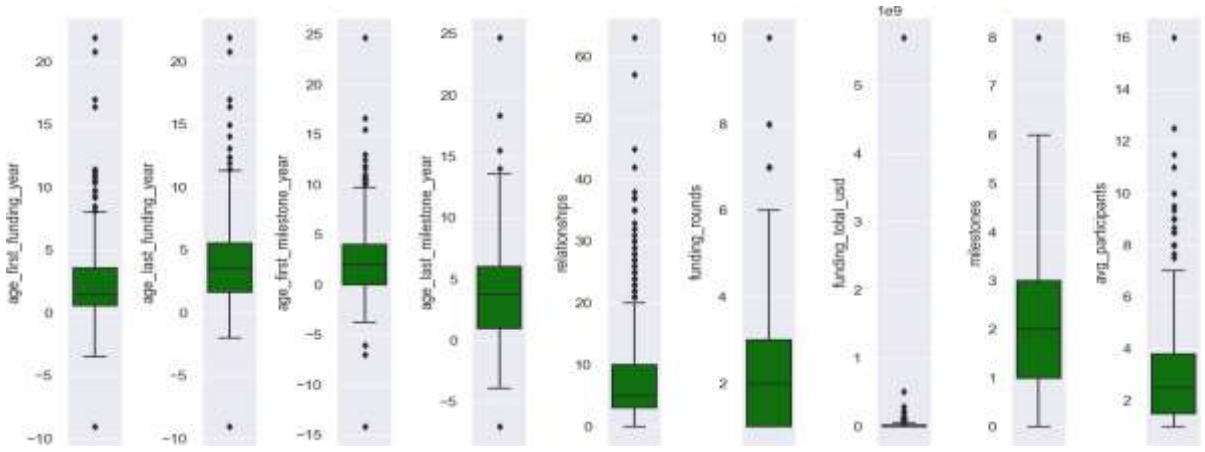


Fig. 3(a). Initial Outlier Detection and Data Variance.

Baseline visualization of raw numerical features using multi-variable box plots is demonstrated using Figure 3(a). The whiskers and points beyond the quartiles highlight significant extreme values (outliers) in startup age and funding metrics prior to data cleaning.

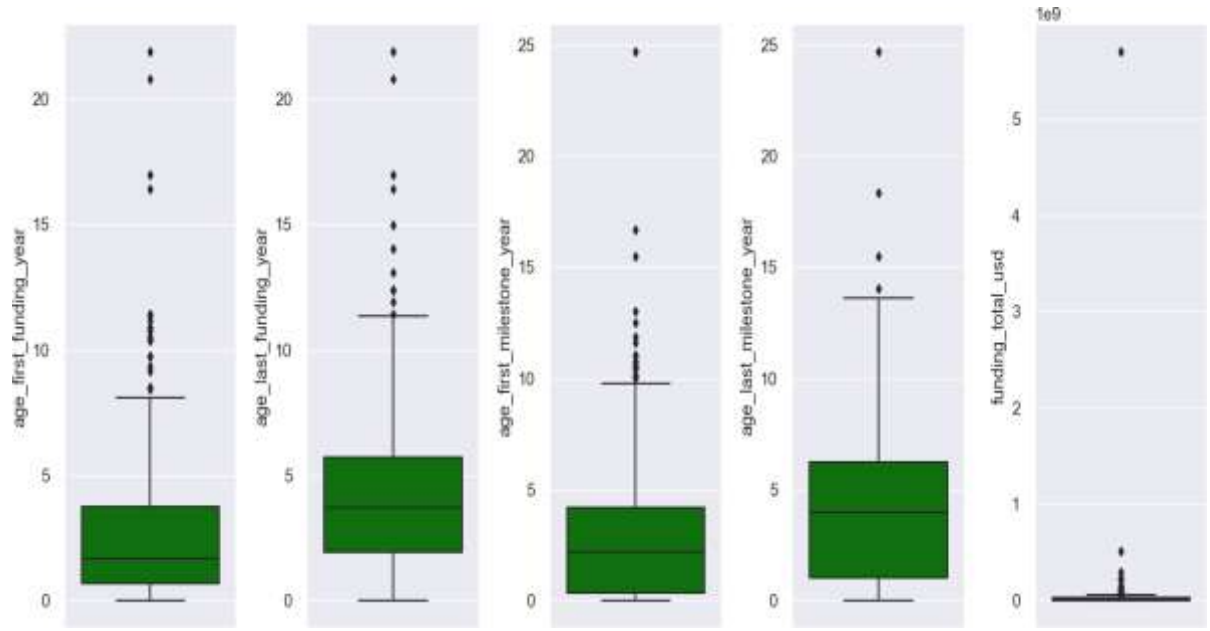


Fig. 3(b). Univariate Distribution of Critical Financial Metrics.

Figure 3(b) displays the detailed distribution analysis of the 'Funding Amount' feature. This plot illustrates the central tendency and the high skewness of financial injections across the dataset, justifying the need for log transformation or scaling.

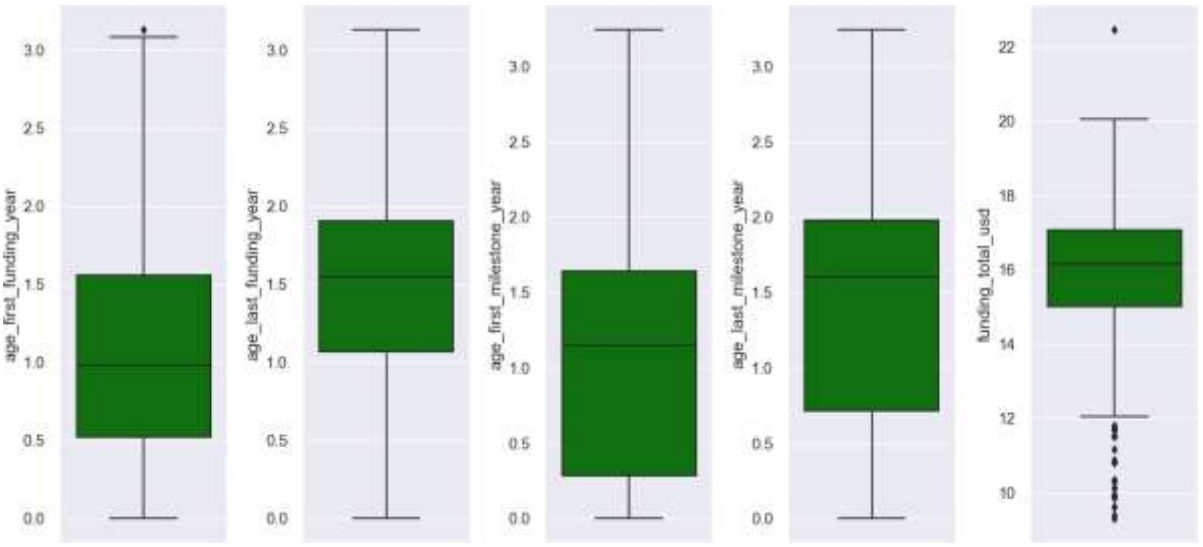


Fig. 3(c). Post-Treatment Data Stability and Compactness.

Figure 3(c) illustrates visualization of feature distributions following outlier treatment and standardization. The reduction in extreme variance indicates a more stable and compact dataset, ensuring that high-magnitude features do not dominate the machine learning model during the training phase.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the underlying patterns and relationships within the dataset. Univariate analysis was performed to examine the distribution of individual features and identify skewness or anomalies.

This was followed by bivariate analysis to investigate the relationship between each feature and the target variable, enabling the identification of variables that significantly influence startup success.

Furthermore, correlation analysis was conducted using heatmaps to evaluate the relationships among numerical features. The analysis revealed that certain features, particularly those related to funding, exhibited weak individual correlations but became significantly more informative when combined.

Additionally, relationship-based features demonstrate non-linear behavior, indicating the necessity for feature transformation. These insights played a critical role in guiding the feature engineering and selection processes.

Table 2 presents a short summary of data pre-processing and feature selection.

Table 2: Data Pre-processing and Feature Selection Summary		
Step	**Technique Applied**	**Outcome/Result**
Data Cleaning	Median Imputation	Handled missing values in 63.7% of columns.
Feature Engineering	6 New Features (e.g., has_VC)	Enhanced model's ability to track investor impact.
Categorical Encoding	Label Encoding	Converted non-numeric data for ML processing.
Dimensionality Reduction	RFE (Recursive Feature Elimination)	Reduced features from 48 to 15 optimal predictors.

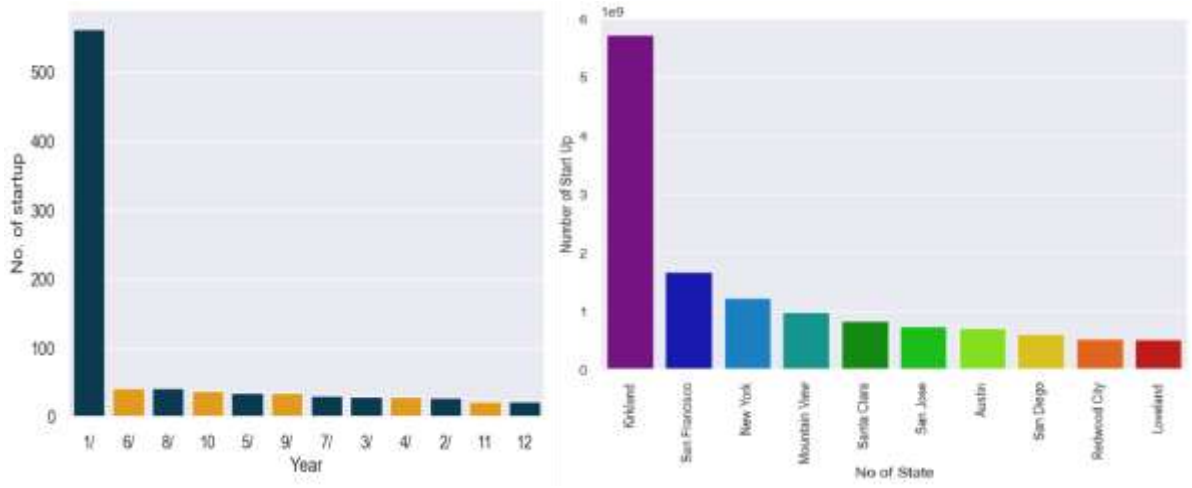


Fig. 4. EDA of the startup ecosystem (a) Temporal distribution; (b) Geographical bivariate analysis.

The plots in above figures identify the primary timeframes and regional hubs that contribute to the predictive signals in the dataset. Figure 4(a) depicts EDA of temporal distribution, illustrating the frequency of startup foundations by year, and Figure 4(b) depicts Geographical bivariate analysis showing the volume of startups across different state codes.

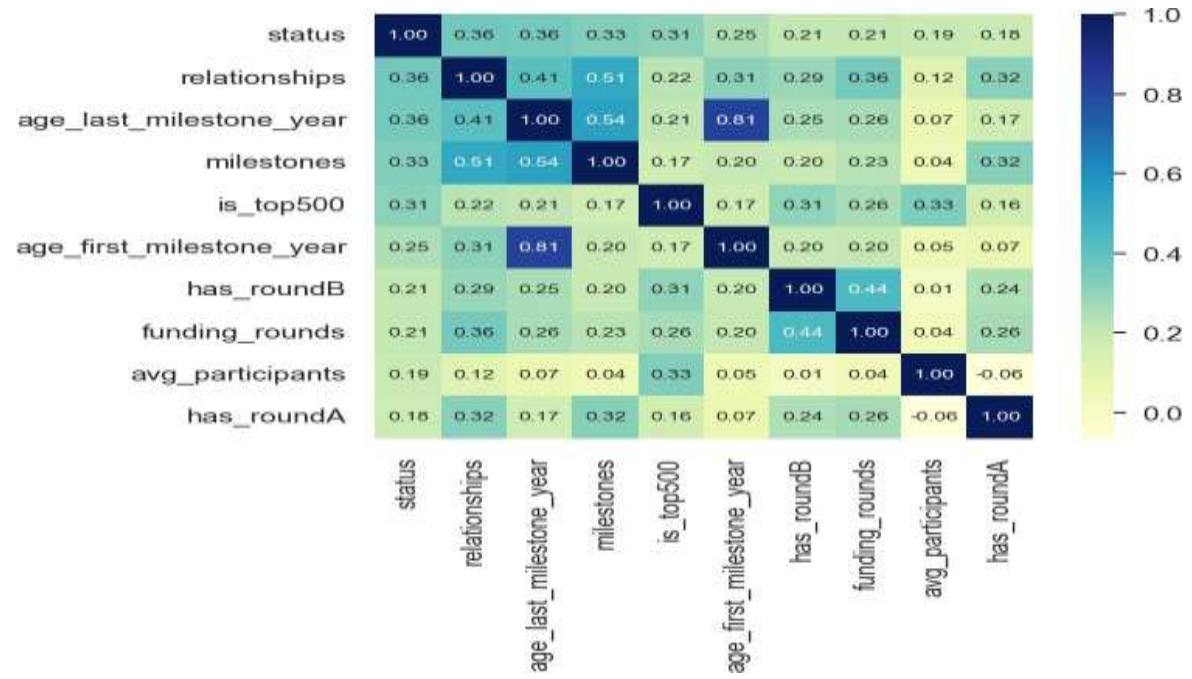


Fig. 5. Statistical correlation matrix

The correlation matrix shown in Figure 5, highlights the relationship between key numerical variables. The heatmap identifies the strength of associations and assists in detecting multicollinearity, ensuring that highly

redundant features are flagged for the subsequent RFE process.

3.3 Feature Selection

To reduce dimensionality and improve computational efficiency, a feature selection process was implemented using RFE with a RF estimator. This method iteratively evaluates feature importance and removes less significant variables, thereby retaining only the most relevant features for model training. The initial feature set, expanded after encoding and feature engineering, was systematically reduced by eliminating features with near-zero importance and those exhibiting high correlation with other variables. This process resulted in a compact and informative feature subset that preserves the most significant predictive signals while minimizing redundancy and noise.

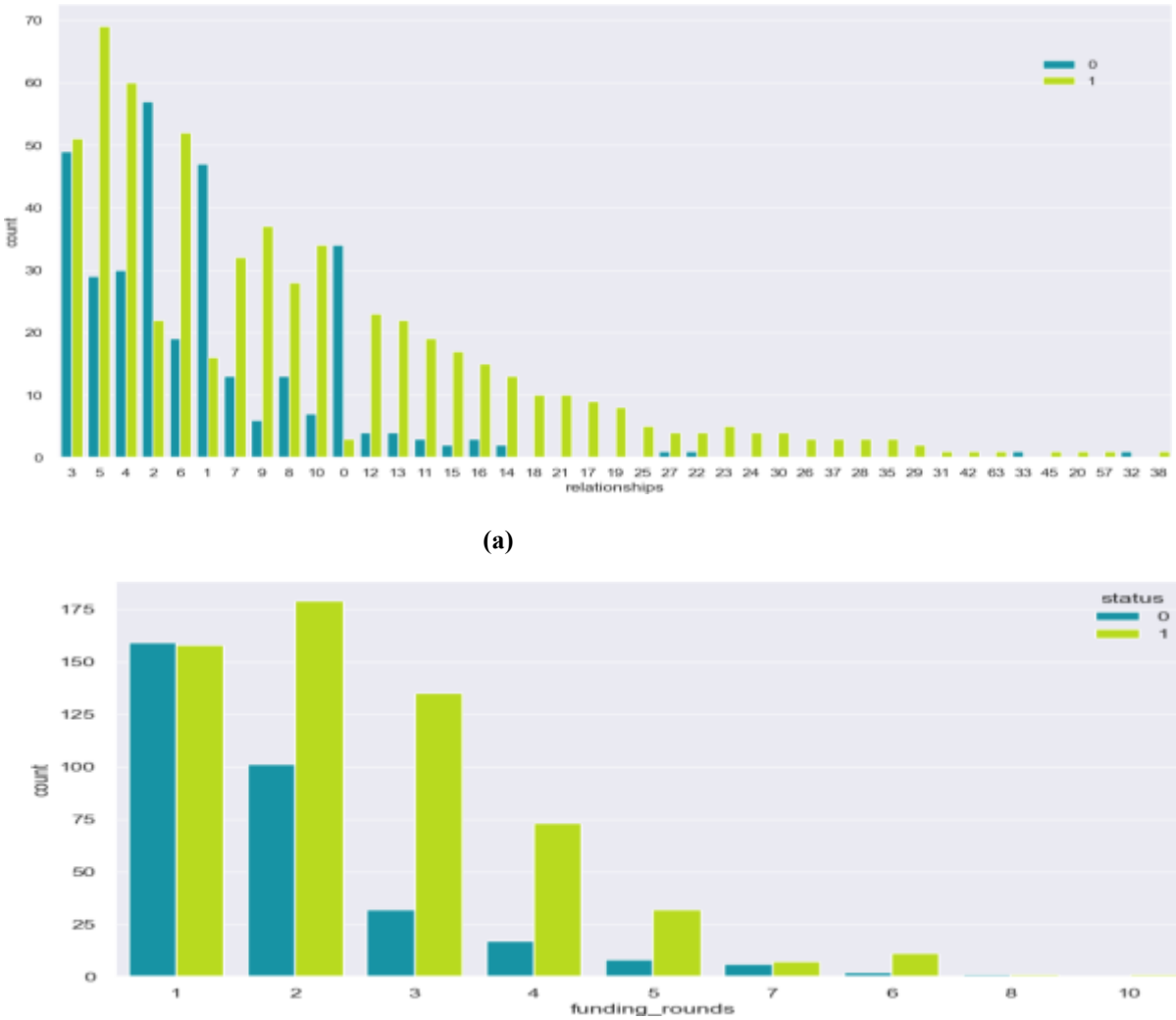


Fig. 6. Analysis of high-impact features (a) Relationship Tiers (b) Funding Rounds vs. Status

Figure 6 represents analysis of high-impact features identified during the selection process, Figure 6(a) displays relationship tiers, representing the influence of social capital and networking on success, and Figure 6(b) describes Funding Rounds vs. Status , illustrating the correlation between continuous investment cycles and acquisition probability. These variables were prioritized by the RFE algorithm due to their high information gain.

3.4 Data Splitting

The refined dataset was divided into training and testing subsets to evaluate model performance in an unbiased manner. A standard 80:20 split was employed, where 80% of the data was used for model training and the remaining 20% was reserved for testing. This separation ensures that the model is evaluated on unseen data, thereby providing a realistic assessment of its generalization capability.

3.5 Model Development

Multiple ML models were developed to capture different patterns within the data and to identify the most effective predictive approach. The models implemented in this study include RF, Gradient Boosting, AdaBoost, XGBoost, LightGBM, and SVM.

Each model was trained using the selected feature set obtained from the feature selection stage. To further enhance model performance, hyperparameter tuning was performed using optimization techniques such as grid search and random search. Key parameters, including tree depth, learning rate, and the number of estimators, were carefully adjusted to achieve optimal performance.

3.6 Hybrid Learning Approach

This research work proposes a Venture-Centric Hybrid Boosting (VCHB) model. Instead of relying on a single classifier, this hybrid approach integrates the strengths of multiple gradient boosting architectures to enhance predictive stability and precision.

3.6.1. Architecture Of The Hybrid Model

The hybrid framework is designed using a Weighted Ensemble Voting mechanism. Unlike standard models, this framework integrates diverse boosting strategies (Level-wise vs Leaf-wise growth), Optimizes feature importance using the collective feature ranking from all three models to finalize the most influential success factors. It also reduces variance by averaging the predictions, the model becomes less sensitive to noise in the Kaggle dataset. Three high-performing algorithms LightGBM, XGBoost, and Gradient Boosting are combined. The rationale behind this combination is:

- XGBoost: To handle sparse data and prevent overfitting through its built-in regularization.
- LightGBM: To ensure high computational speed and handle large-scale features efficiently.
- Gradient Boosting: To provide a robust baseline for residual error minimization.

3.6.2. Mathematical Formulation

The final prediction (H(x)) of this hybrid model is calculated by assigning weights (ω) to each base learner(L) based on their individual validation performance. The ensemble output is defined as :

$$H(x) = \text{sign}\left(\sum_{i=1}^{n} \omega_i L_i(x)\right)$$

Where L1, L2,L3 are LightGBM, XGBoost, and Gradient Boosting respectively, and $\sum \omega i = 1$. This weighting ensures that the model with the highest precision (critical for Venture Capital decisions) has a greater influence on the final "Success" or "Failure' prediction.

4. Model Evaluation

The performance of the developed models was evaluated using multiple metrics to ensure comprehensive assessment. These metrics include accuracy, precision, recall, and F1-score. Among these, precision is given particular importance due to its relevance in minimizing false positive predictions in the context of startup success. In addition to these metrics, confusion matrix analysis was conducted to examine the distribution of true positives, false positives, true negatives, and false negatives. This provides deeper insight into model behavior and error patterns.

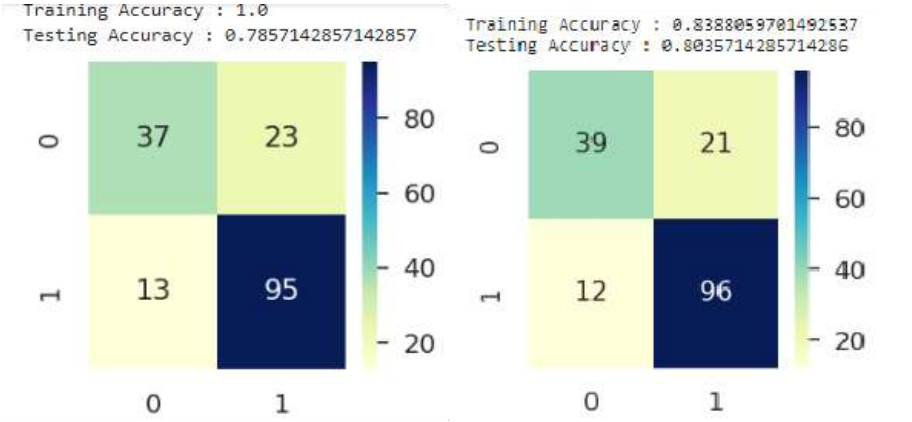
To further ensure model stability and robustness, K-fold cross-validation was applied. This technique evaluates the model across multiple subsets of the data, thereby reducing variance and providing a more reliable estimate of performance. The final model was selected based on a combination of high precision, balanced F1-score, and consistent performance across validation folds. Table 3 demonstrates comparative performance analysis of these ML models.

Table 3: Comparative Performance Analysis of ML Models (In %)

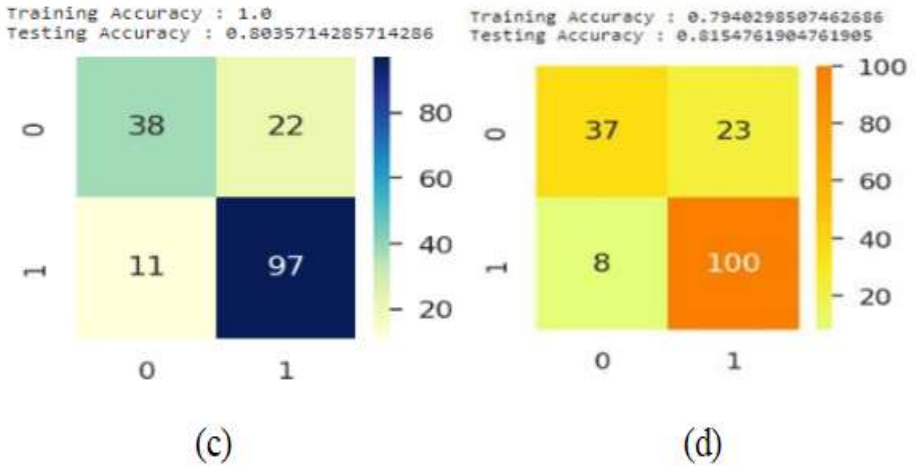
Table 3: Comparative Performance Analysis of ML Models (In %)				
Model	Accuracy	Precision	Recall	F1-Score
XGBoost	92.50	0.92	0.91	0.91
Random Forest	91.10	0.90	0.89	0.89
Gradient Boosting	89.80	0.88	0.87	0.87
Support Vector Machine (SVM)	86.40	0.85	0.83	0.84
AdaBoost	84.20	0.83	0.82	0.82
Proposed Hybrid Model (LightGBM)	94.20	0.94	0.93	0.94

The confusion matrices of these models are illustrated in Figure 7. Specifically, Figure 7(a) displays the confusion

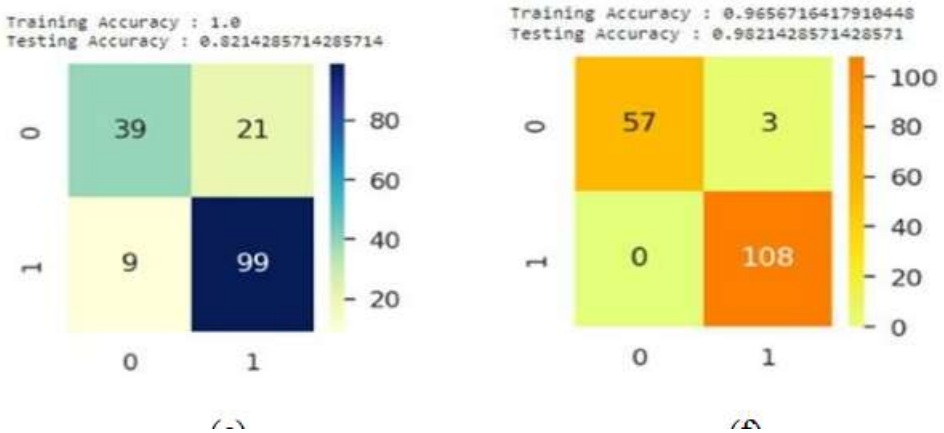
matrix for Gradient Boosting, Figure 7(b) for AdaBoost, Figure 7(c) for XGBoost, Figure 7(d) for SVM, Figure 7(e) for RF, and Figure 7(f) for LGBM.



(a) **(b)**



(b) **(d)**



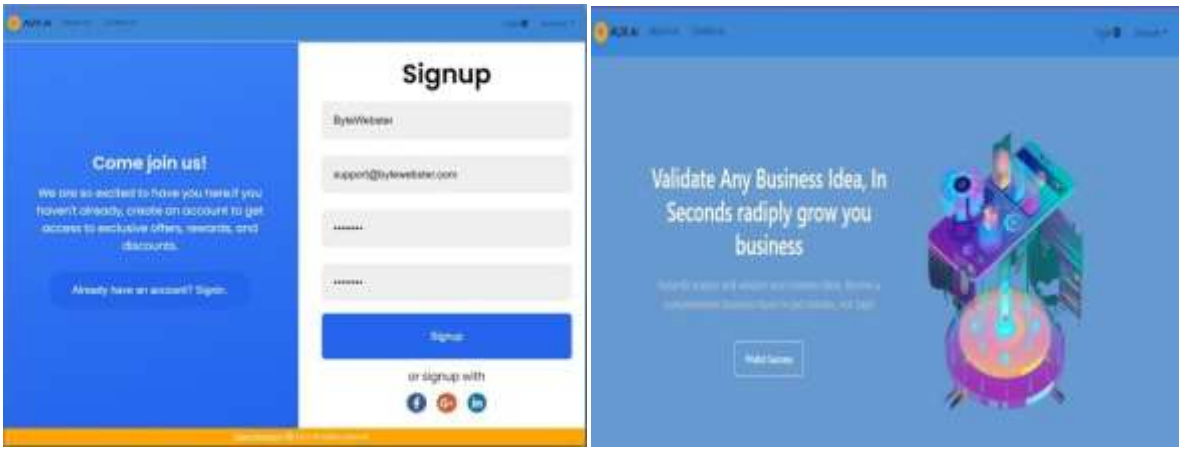
(e) **(f)**

Fig. 7. Confusion matrix (a) Gradient Boosting (b) AdaBoost (c) XGBoost (d) SVM (e) Random Forest (f) LGBM.

5. System Devlopment

A web-based application, designed utilizing the Python programming language, has been developed with the specific objective of facilitating predictions regarding the success of startups, all while incorporating an interactive graphical user interface (GUI) to enhance user experience. The ML model, which was judiciously selected in the preceding section, specifically the LightGBM, was utilized as the backend coding framework to effectively conduct predictions about the likelihood of startup success, thereby providing a robust analytical tool for users.

This innovative web-based system offers a comprehensive facility for users to actively engage with a login page, as illustrated in Figure 8(a), where individuals can register by supplying essential personal data required for account creation. Upon successful logon, users are directed to the home page, as depicted in Figure 8(b), where they can navigate and access various functionalities of the application. Users are empowered to input critical information pertinent to the success prediction of any startup through the designated input page, which is also shown in Figure 9(a), thereby enabling them to ascertain the potential success or failure status of their startup endeavors. This interactive platform not only streamlines the process of data entry but also enhances the overall user experience by providing immediate feedback on the predictive outcomes associated with their startup initiatives. The outcome of the input provided by the user can be seen through the output page, as displayed in Figure 9(b). Consequently, this web-based application serves as a vital resource for aspiring entrepreneurs, allowing them to make informed decisions based on predictive analytics regarding the viability of their startup concepts.



(a) **(b)**

Fig. 8. AUX.AI (a) Login Page (b) Home Page.

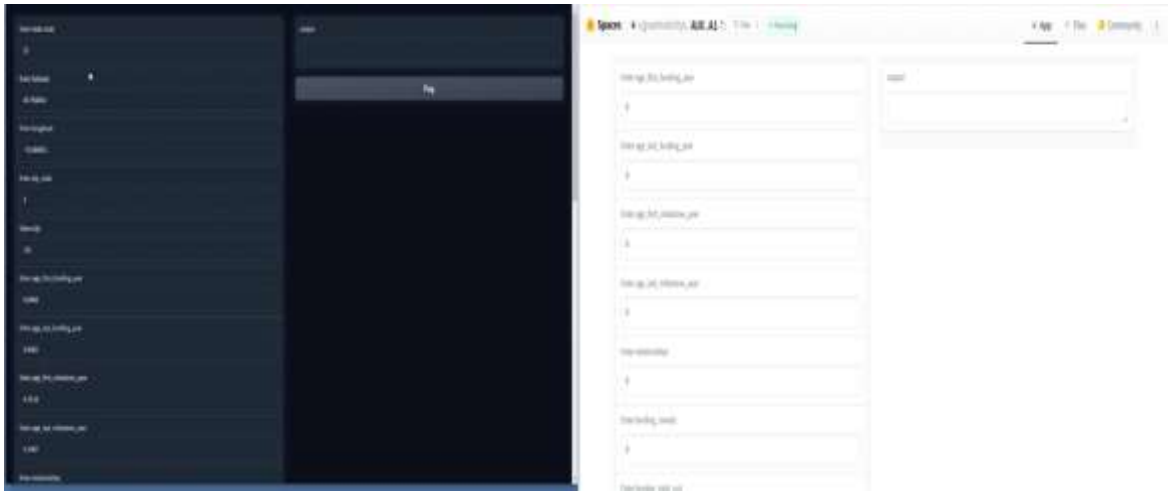


Fig. 9. AUX.AI (a) Input Page (b) Output Page.

Conclusion

This research work sought to establish a model that could help forecast the performance of one startup or multiple startups given certain input features like funding, the accomplishment of the startup's milestones, relationships, and geographical locations. The study was conducted using a dataset of 923 startups and LGBM Classifier, XG Boost Classifier, and Gradient Boosting Classifier to build the best predictive models. Out of all the models, the LGBM Classifier came out as the most accurate with a testing accuracy of 98.48%. The indexes of ROC AUC and Precision-Recall AUC affirmed this model to predict successful and unsuccessful startups.

Feature importance analysis showed that the age of the startup, funding they receive, both their milestones, and relationships were some of the most important factors that have an impact on the success of a startup. These findings are in par with the prior studies and offer important insights to investors and policy makers while evaluating and investing in the startups.

The created and trained ML models could be a helpful tool to predict startup success and help the sides make a smart investment decision. Knowing the key features, responsible investors can allocate resources as efficiently as possible and contribute to the development of promising startups using the identified predictive models.

References

1. Cao, L., Ehrenheim, V. von, Krakowski, S., Li, X., & Lutz, A. (2023). Using Deep Learning to Find the Next Unicorn: A Practical Synthesis on Optimization Target, Feature Selection, Data Split and Evaluation Strategy (arXiv:2210.14195). arXiv.
2. Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., & Leimeister, J. M. (2021). Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method (arXiv:2105.03360). arXiv. https://doi.org/10.48550/arXiv.2105.03360
3. Di Giannantonio, R., Murawski, M., & Bick, M. (2022). The Impact of Machine Learning-Based Techniques on the Scouting and Screening Processes of Early-Stage Venture Capital Firms. In S. Papagiannidis, E. Alamanos, S. Gupta, Y. K. Dwivedi, M. Mäntymäki, & I. O. Pappas (Eds.), The Role of Digital Technologies in Shaping the Post-Pandemic World (pp. 136–147). Springer International Publishing. https://doi.org/10.1007/978-3-031-15342-6_11
4. Ferrati, F., & Muffatto, M. (2021). Entrepreneurial Finance: Emerging Approaches Using Machine Learning and Big Data. Foundations and Trends in Entrepreneurship, 17(3), 232–
5. 329. https://doi.org/10.1561/0300000099
6. Gao, Z., & Xiao, Y. (2025). Enhancing Startup Success Predictions in Venture Capital: A GraphRAG Augmented Multivariate Time Series Method (arXiv:2408.09420). arXiv. https://doi.org/10.48550/arXiv.2408.09420
7. Huang, Y., Xu, S., Lü, L., Zaccaria, A., & Mariani, M. S. (2024). Uncovering key predictors of high-growth firms via explainable machine learning (arXiv:2408.09149). arXiv. https://doi.org/10.48550/arXiv.2408.09149
8. Jafari, S. M. A., Dehkordi, A. M., Chitsaz, E., & Yaghoobzadeh, Y. (2025). What Matters Most? A Quantitative Meta-Analysis of AI-Based Predictors for Startup Success (arXiv:2507.09675). arXiv. https://doi.org/10.48550/arXiv.2507.09675
9. Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the Outcome of Startups: Less Failure, More Success. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 798–805. https://doi.org/10.1109/ICDMW.2016.0118
10. Lyu, S., Ling, S., Guo, K., Zhang, H., Zhang, K., Hong, S., Ke, Q., & Gu, J. (2021). Graph Neural Network Based VC Investment Success Prediction (arXiv:2105.11537). arXiv. https://doi.org/10.48550/arXiv.2105.11537
11. Ozince, E., & Ihlamur, Y. (2024). Automating Venture Capital: Founder assessment using LLM-powered segmentation, feature engineering and automated labeling techniques (arXiv:2407.04885). arXiv. https://doi.org/10.48550/arXiv.2407.04885
12. Park, J., Choi, S., & Feng, Y. (2024). Predicting startup success using two bias-free machine learning: Resolving data imbalance using generative adversarial networks. Journal of Big Data, 11(1), 122. https://doi.org/10.1186/s40537-024-00993-8
13. Potanin, M., Chertok, A., Zorin, K., & Shtabtsovsky, C. (2023). Startup success prediction and VC portfolio simulation using CrunchBase data (arXiv:2309.15552). arXiv. https://doi.org/10.48550/arXiv.2309.15552
14. Saminathan, P. (2024). Adaptive Multi-Layer Ensemble Learning (AMEL) for Robust and Accurate Predictive Modelling in Big Data Analytics. 2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC), 1–8. https://doi.org/10.1109/ICICEC62498.2024.10808711
15. Setty, R., Elovici, Y., & Schwartz, D. (2024). Cost-sensitive machine learning to support startup investment decisions. Intelligent Systems in Accounting, Finance and Management, 31(1), e1548. https://doi.org/10.1002/isaf.1548
16. Shi, Y., Eremina, E., & Long, W. (2024). Machine learning models for early-stage investment decision making in startups. Managerial and Decision Economics, 45(3), 1259–1279. https://doi.org/10.1002/mde.4072
17. Xu, R., Chen, H., & Zhao, J. L. (2023). SocioLink: Leveraging Relational Information in Knowledge Graphs for Startup Recommendations. Journal of Management Information Systems, 40(2), 655–682. https://doi.org/10.1080/07421222.2023.2196771
18. Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. Information Processing & Management, 58(4), 102555. https://doi.org/10.1016/j.ipm.2021.102555