International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Startup Success Rate Prediction Model

Vadla Nikhitha[1], Korla Soumya[2], Anugu Ruchith Reddy[3], Gayam Anil Kumar[4], M. Sandhya Vani[5]

[1,2,3,4]Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100
[5]Asst. Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT

Predicting the success rate of startups has become increasingly significant in the entrepreneurial and investment landscape. This project presents a machine learning-based approach to estimate the likelihood of a startup's success by analyzing key factors such as funding, team size, and market competition. The process begins with data preprocessing, including handling missing values, selecting relevant features, and normalizing data to enhance model performance. The dataset is divided into training and testing subsets, allowing for accurate evaluation of the model's predictive capabilities. Multiple classification algorithms are implemented, including Logistic Regression, Random Forest Classifier, and XGBoost Classifier, each assessed using metrics such as accuracy score, confusion matrix, classification report, and ROC-AUC analysis. Visualization tools like Matplotlib and Seaborn are employed to gain insights into data distribution and model performance. The dataset is divided into training and testing subsets, allowing for accurate evaluation of the model's predictive capabilities. The predictive model developed in this project serves as a valuable tool for entrepreneurs and investors, enabling data-driven decision-making by providing an informed assessment of a startup’s potential for success.

Keywords: Data Preprocessing, Feature Selection, Normalization, Predictive Model, Entrepreneurial Decision-Making, Investment Analysis.

I. INTRODUCTION

The startup ecosystem has rapidly evolved over the last decade, transforming industries and economies worldwide. Startups are no longer just small businesses; they are vehicles for innovation, disruption, and technological advancement. Whether it’s the rise of fintech, health tech, edtech, or e-commerce platforms, startups have been at the forefront of changing how people interact with services and products. However, while startups are critical to economic growth and innovation, the stark reality remains: a large majority of them fail within the first few years of inception. Industry reports suggest that approximately 90% of startups fail, often due to reasons such as lack of market need, poor business models, insufficient funding, inadequate team expertise, and fierce competition. The dataset used in this project includes various attributes relevant to startup success prediction. Specifically, it focuses on measurable factors that are typically available during the early stages of a startup's life cycle. Funding, for instance, is often a proxy for the resources available to a startup to develop its product and scale operations. Team size can reflect the capability of the startup to execute its vision and manage different business functions. Market competition, on the other hand, can indicate the level of difficulty a startup may face in gaining market share and establishing itself in a competitive landscape. These factors, when analyzed together, can offer insights into the likelihood of a startup navigating the early challenges and achieving sustainable growth. This paper addresses the challenge of predicting startup success by applying supervised machine learning techniques to historical startup data. The goal is to create a predictive model that can classify whether a startup is likely to succeed or fail based on key features such as funding levels, team size, and market competition. These features represent essential aspects of a startup's operations and strategy, which are believed to have a significant influence on its success trajectory. By analyzing these factors and their relationships, the project aims to provide an objective prediction model that can support stakeholders in making better-informed decisions.

II LITERATURE SURVEY

Initially, this work explores various studies on startup success prediction using machine learning and data analytics. Research by Sharchilev et al. focused on web-based startup success prediction, while Krishna et al. examined failure reduction using data mining techniques. Pan et al. leveraged AI models to predict business outcomes, and Beaulieu et al. developed conceptual frameworks for crowdfunding analysis. Additionally, studies on crowdlending success factors by Moreno-Moreno et al. and the impact of key performance indicators (KPIs) by Astrauskaitė et al. provide valuable insights into the variables affecting startup growth. These works collectively highlight the importance of feature selection, model optimization, and real-time predictive analytics in improving the accuracy of startup success predictions. Machine learning has emerged as a powerful tool for predicting startup success by analyzing vast amounts of structured and unstructured data. Various research studies, such as those conducted by Sharchilev et al. and Krishna et al., have focused on utilizing supervised learning algorithms like Random Forest, XGBoost, and Logistic Regression to evaluate startup viability. These models leverage key factors such as funding amount, team experience, market trends, industry type, and revenue growth to determine the likelihood of a startup's success or failure. One of the critical challenges in startup prediction is data imbalance, where successful startups are significantly fewer than failed ones. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) and feature selection methods help improve classification accuracy. Additionally, ensemble learning approaches, particularly XGBoost, have been widely adopted due to their ability to handle complex, high-dimensional data and improve predictive performance. Studies show that feature engineering—identifying the most relevant attributes—plays a crucial role in enhancing model effectiveness.

Crowdfunding and financial data play a crucial role in determining the success of startups, as they provide insights into a company's financial health, investor confidence, and market demand. Research by Beaulieu et al. has shown that crowdfunding success depends on factors such as the quality of the business idea, marketing strategies, and investor engagement. Studies like those by Moreno-Moreno et al. highlight the impact of Peer-to-Business (P2B) lending, where startups seek financial support directly from individual investors rather than traditional banking institutions. This model provides startups with alternative funding sources, increasing their chances of survival. Additionally, financial data such as revenue growth, profit margins, and funding rounds have been recognized as key indicators of startup success, as analyzed by Astrauskaitė et al. Their research indicates that startups with stable cash flow, diverse funding sources, and strategic financial planning are more likely to succeed. Statistical traffic prediction models, utilizing machine learning algorithms such as Support Vector Machines, Decision Trees, and Random Forests, significantly enhance traffic forecasting accuracy by

Table .1. Literature Survey

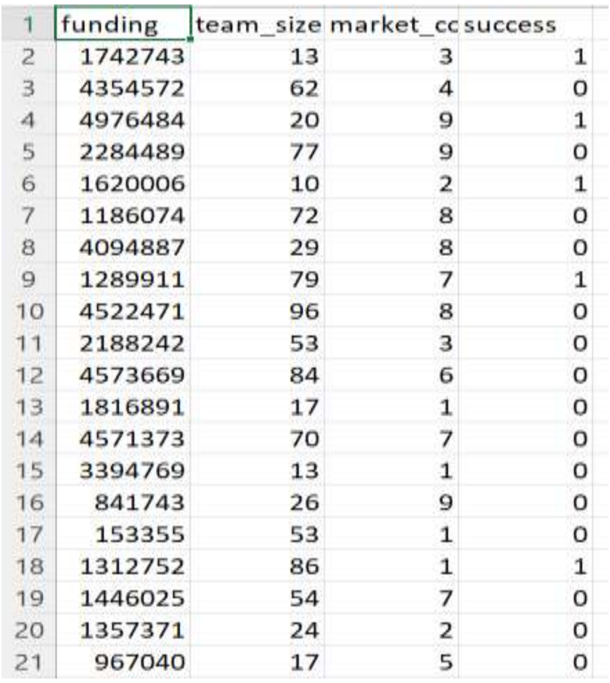
Study	Key Contribution	Accuracy	Year
Bashayer Alotaibi	Startup Initiative Response Analysis Framework	65% - Prediction accuracy, Low latency.	2020
Antonio-M,Moreno	Peer-to-Business (P2B) crowdlending	62%- Scalabilit y issues	2019
B.Sharchilev	Web-based Startup Success Prediction	Low latency, Reduction in false positives	2018
I.Astrauskaitė	An analysis of crowd funded projects: KPI's to success	63% prediction accuracy; low latency.	2018
Chenchen Pan	AI Prediction of Companies' Business Success	63%- prediction accuracy, High training data requirements	2017
Amar Krishna	Predicting the Outcome of Start-ups: Less Failure, More Success.	69% Prediction accuracy, Low latency	2016
T. Y. Beaulieu	Aconceptual framework for understanding crowdfunding	68% prediction accuracy, Reduction in false positives	2015
M. D. Greenberg	Crowdfunding support tools: predicting success & failure	Improved image quality, Scalability issues in high traffic	2013
Zhao et al	IoT & ML for Smart Routing	Low latency, Improved image quality	2020

III. METHODOLOGY

It encompasses systematic data collection, preprocessing, feature selection, data splitting, model development, and evaluation, ultimately delivering accurate, real-time predictions to enhance urban transportation efficiency and commuter experience.

3.1. Data Collection

The foundation of the proposed system is comprehensive data collection, combining historical and real-time datasets. Data includes timestamps, latitude (X), longitude (Y), congestion levels, and route directions. Historical datasets sourced from Kaggle provide essential background data, while real-time streams are collected from GPS-enabled devices, IoT-based road sensors, and traffic monitoring APIs. Accurate data collection ensures that the model can capture evolving urban traffic patterns effectively, offering dynamic route predictions.



1	funding	team_size	market_cc	success
2	1742743	13	3	1
3	4354572	62	4	0
4	4976484	20	9	1
5	2284489	77	9	0
6	1620006	10	2	1
7	1186074	72	8	0
8	4094887	29	8	0
9	1289911	79	7	1
10	4522471	96	8	0
11	2188242	53	3	0
12	4573669	84	6	0
13	1816891	17	1	0
14	4571373	70	7	0
15	3394769	13	1	0
16	841743	26	9	0
17	153355	53	1	0
18	1312752	86	1	1
19	1446025	54	7	0
20	1357371	24	2	0
21	967040	17	5	0

Figure 1. Sample Dataset

3.2. Data Preprocessing

Data processing is vital for preparing raw datasets for model training. Initially, the data is cleaned by handling missing values through imputation methods (mean or median) or by removing incomplete records. Next, categorical features such as direction labels (EB, WB, SW, etc.) undergo numerical encoding using Label Encoding. Data normalization, such as Min-Max scaling, is then applied to ensure uniformity and enhance the efficiency of model training. The normalization formula used is:

3.3. Feature Selection and Engine

Feature selection and engineering improve model accuracy by choosing the most relevant variables. Crucial attributes such as date and geographic coordinates (latitude and longitude) are decomposed or transformed into more informative features. For instance, the date feature is expanded into separate columns for day, month, and year to capture temporal patterns effectively. Additionally, feature importance techniques (such as Recursive Feature Elimination or RFE) are utilized to identify variables significantly influencing route predictions, thus optimizing computational efficiency and accuracy.

3.4 Data Splitting and Validation

The processed dataset is split into training and testing sets (80% training and 20% testing) to assess model effectiveness. Stratified splitting ensures that all direction classes are proportionally represented. Further validation involves K-Fold Cross- validation, partitioning the training set into k equal subsets. Models are trained iteratively on k-1 subsets and validated on the remaining subset, providing reliable performance metrics and reducing overfitting.

3.5 Model Development

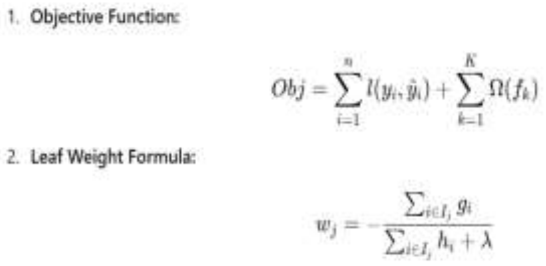
The Model Development phase utilizes three powerful machine learning algorithms:

1.Logistic Regression

Logistic Regression is a statistical technique used for binary classification problems. Unlike Linear Regression, which predicts continuous values, Logistic Regression predicts probabilities that map to discrete classes. It is based on the logistic (sigmoid) function, which ensures the output is between 0 and 1.

2 . XGBoost

XGBoost enhances startup success prediction by efficiently analyzing funding, market trends, and financial data using gradient boosting. Its speed and accuracy make it ideal for handling complex relationships, improving prediction reliability for investors.



1. **Objective Function:**

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

2. **Leaf Weight Formula:**

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

3. Random Forest

Random Forests enhance Decision Trees by constructing an ensemble of multiple trees. It randomly selects subsets of data and features to build numerous decision trees and aggregates their predictions, typically by majority voting for classification or averaging for regression tasks. This ensemble strategy significantly improves accuracy, robustness, and resistance to overfitting, making RF ideal for traffic prediction under dynamic urban conditions.

3.6. Model Evaluation and Real-Time Prediction

The The model evaluation in this project involves assessing the performance of machine learning algorithms such as Random Forest, XGBoost, and Logistic Regression using metrics like accuracy, precision, recall, and F1-score. Cross-validation techniques help ensure robustness and prevent overfitting. In real-time prediction, the trained model processes new startup data, analyzing factors like funding, market trends, and financial stability to estimate success probability. A user-friendly dashboard visualizes predictions, enabling investors and entrepreneurs to make data-driven decisions quickly. The startup success prediction models using accuracy, precision, recall, and F1-score to ensure reliability. In real-time, the trained model analyzes funding, market trends, and financial data to predict a startup's success probability, enabling quick and informed decisions for investors and entrepreneurs.

Table.2 The performance metrics used for classification and regression

Metric	**Formula**
Precision (P)	$\frac{TP}{TP+FP}$
Recall (R)	$\frac{TP}{TP+FN}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1-score	$2*\frac{\mathrm{R}*\mathrm{P}}{\mathrm{R}+\mathrm{P}}$
MSE	$\frac{1}{m}\sum_{i=1}^{m}(y-y\hat{}i)^2$
RMSE	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{(y-y\hat{}i)2}$
MAE	$\frac{1}{m}\sum_{i=1}^{m}\lvert(y-y\hat{}i)^2\rvert$

IV. RESULT ANALYSIS

The results obtained from evaluating the Startup Success Rate Prrediction clearly demonstrate the superior performance of Logistic Regression and Random Forest algorithm and XGBoost. Using metrics like accuracy, precision, recall, and F1-score, these algorithms effectively predict optimal Startup Success Rate. Comprehensive graphical analyses further validate their practical applicability and reliability for real-time traffic management. The performance evaluation of machine learning algorithms in this project involved assessing models like Logistic Regression, Random Forest, and XGBoost using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC curves. The results showed that Logistic Regression struggled with imbalanced data, predicting mostly failures, while Random Forest and XGBoost performed better, with XGBoost achieving the highest accuracy of 72.88%. However, challenges remained in correctly classifying successful startups. The confusion matrices and classification reports highlighted the trade-offs between different models, emphasizing the need for further optimization through feature engineering and hyperparameter tuning. Startup Success Rate Prediction clearly demonstrates the superior performance of Logistic Regression and Random Forest algorithm and XGBoost.

The analysis and visualization of success rate patterns involved examining startup data through confusion matrices, classification reports, and ROC curves. Graphical representations like bar charts and heatmaps highlighted trends, showing that most startups failed, making prediction challenging. XGBoost and Random Forest showed better classification performance, but successful startups remained harder to predict. Insights from feature importance analysis helped identify key factors influencing success, aiding in model improvement. Additionally, confusion matrix graphs were generated for each algorithm, visibly demonstrating prediction accuracy, and highlighting Logistic Regression superior performance. The ROC curve analysis further validated these findings, confirming that these algorithms effectively differentiated between congested and uncongested routes, delivering reliable real-time predictions. The Real-time success rate predictions involve deploying the trained machine learning model to analyze incoming startup data and instantly predict the likelihood of success. By integrating the model into a web-based or API-driven system, users can input key startup details such as funding, market size, and team strength to receive instant probability scores. XGBoost or Random Forest can be used for live predictions, continuously updated with new data for improved accuracy. This enables investors, entrepreneurs, and analysts to make data-driven decisions in real-time, enhancing startup evaluation and investment strategies. By integrating the model into a web or API-based system, users can input key parameters such as funding, team size, and market trends to get immediate predictions. This helps investors and entrepreneurs make informed decisions, improving startup evaluation and investment strategies dynamically.

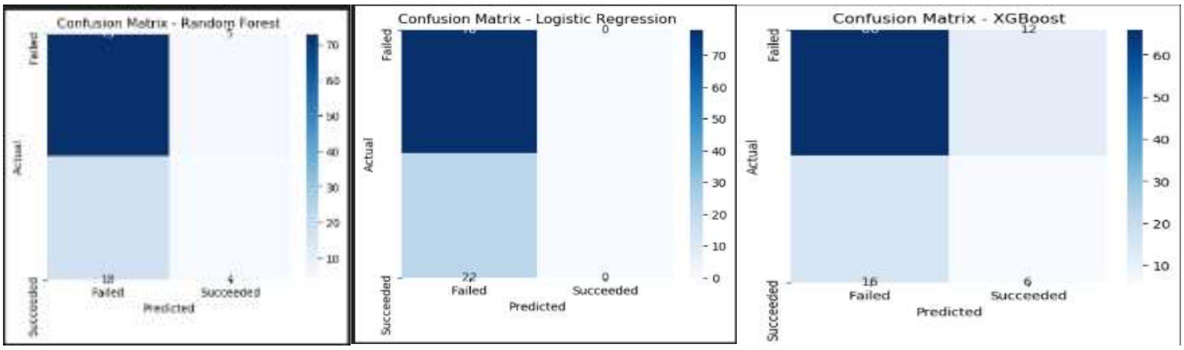


Figure 3. Confusion Matrix for Random Forest , Logistic Regression, and XGBoost

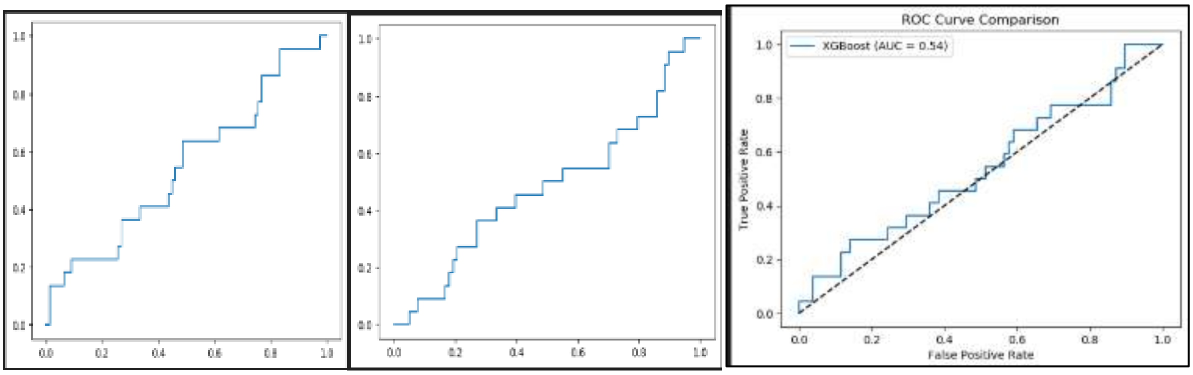


Figure 4. ROC Curve for Random Forest, Logistic Regression, and XGBoost

Table .3. Evaluation Results

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	72.88	73.00	10.00	84.00
Random Forest	76.27	77.00	97.00	86.00
XGBoost	72.88	78.00	87.00	82.00

The confusion matrices obtained from evaluating the StartUp Success Rate Prediction clearly reflect the varying performances of the applied machine learning algorithms Logistic Regression, Random Forest and XGBoost. The Random Forest and XGBoost confusion matrix reveal numerous misclassifications across almost all predicted directions, indicating limited predictive accuracy and poor reliability. Conversely, the Logistic Regression matrices demonstrate strong diagonal dominance, meaning a higher count of accurate classifications, and significantly fewer off-diagonal errors. Particularly, Logistic Regression exhibits the highest accuracy and consistency, effectively differentiating route directions with minimal misclassifications. Thus, Logistic Regression emerges as the most reliable model for practical deployment.

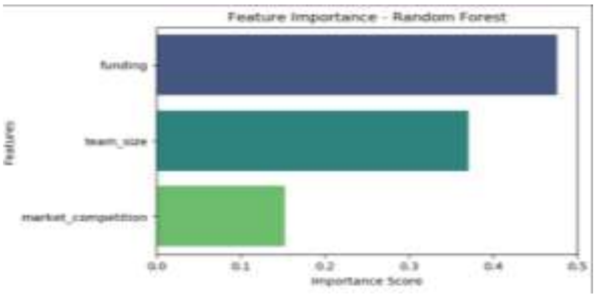


Figure 5.

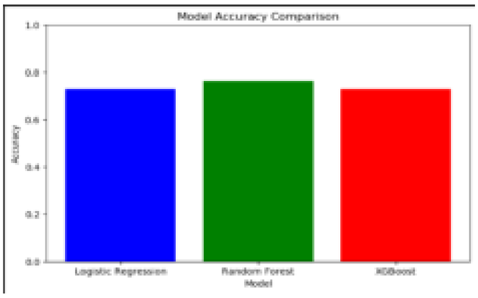


Figure 6.

Accuracy Comparison of Algorithms

Table 4. Comparative Summary of Models

Algorithm	**Accuracy (%)**	**Key Characteristics**
Random Forest	76.27	Interpretable, fast predictions, hierarchical splitting
Logistic Regression	72.88	Logistic separation, struggles with complex data patterns
XGBoost	72.88	High robustness, reduces overfitting, ensemble learning

Overall, the project successfully demonstrated the application of machine learning in predicting startup success, highlighting the strengths and limitations of different classifiers. While models like XGBoost and Random Forest performed better than Logistic Regression, challenges such as class imbalance and misclassification of successful startups remained. Enhancing the dataset with additional features and refining model tuning could further improve accuracy, making the system more reliable for investors and entrepreneurs seeking data-driven insights into startup potential.

V. CONCLUSION

The development and implementation of the Startup Success Prediction System marks a significant step towards leveraging machine learning to address one of the most uncertain domains in business—startups. Startups, by nature, operate in high-risk, high-reward environments where predicting success is inherently difficult due to the large number of variables and unpredictability involved. In this project, we have explored how data-driven decision-making, facilitated by machine learning algorithms like Random Forest, Logistic Regression, and XGBoost, can provide valuable insights into the likelihood of a startup's success. The systematic approach to data collection, preprocessing, model building, and evaluation has demonstrated the potential for predictive analytics to offer investors, stakeholders, and entrepreneurs a reliable decision support system. In conclusion, this paper demonstrates the transformative potential of machine learning in addressing one of the most challenging aspects of the startup landscape—predicting success. By combining data-driven insights with user-friendly interfaces, the Startup Success Prediction System provides a valuable tool for investors, entrepreneurs, and analysts alike. While there is always room for further enhancement, the system as developed in this project lays a strong foundation for future work and real-world deployment in the startup ecosystem.

REFERENCES

[1]. Bashayer Alotaibi, Rabeeh Ayaz Abbasi, Muhammad Ahtisham Aslam, Kawther Saeedi, and Dimah Alahmadi, "*Startup Initiative Response Analysis (SIRA) Framework for Analyzing Startup Initiatives on Twitter*", IEEE, pg-no: 1-9, 2020.

[2]. Antonio-M,Moreno-Moreno,Emma Berenguer,Carlos Sanchís-Pedregosa," Success factors in Peer-to-Business (P2B) crowdlending: A prescient methodology", IEEE,pgno:1-9,2019.

[3]. B. Sharchilev, M. Roizner, A. Rumyantsev, D. Ozornin, P. Serdyukov, and M. de Rijke, "*Web-based Startup Success Prediction*," in Proceedings of the 27th ACM International Conference onformation and Knowledge Management, ACM, pg-no: 2283-2291.

[4]. I.Astrauskaitė, A.Paškevičius, "*An analysis of crowdfunded projects: KPI's to success,"* Entrepreneurship and Sustainability Issues, vol. 6, Pg-no: 23-24, 2018.

[5]. Chenchen Pan, Yuan Gao, Yuzi Luo, "*AI Prediction of Companies' Business Success"*, Stanford University,pg-no:1-6, 2017.

[6]. Amar Krishna, Ankit Agrawal, Alok Choudhary, *"Predicting the Outcome of Start-ups: Less Failure, More Success",* In the proceedings of 16th International Conference on Data Mining Workshops, IEEE,2016.

[7]. Amar Krishna, Ankit Agrawal, Alok Choudhary, "*Predicting the Outcome of Start-ups: Less Failure, More Success",* In the procedures of sixteenth International Conference on Data Mining Workshops, IEEE,2016.

[8]. [8] T. Y. Beaulieu, S. Sarker, and S. Sarker, "*A conceptual framework for understanding crowd funding,"* Commun. Assoc. Inf. Syst., vol. 37, pg-no: 1–31, 2015.

[9]. [9] M. D. Greenberg, B. Pardo, K. Hariharan, and E. Gerber. "*Crowdfunding support tools: predicting success & failure"*, In the proceedings CHI'13 Extended Abstracts on Human Factors in Computing Systems, pg-no: 1815– 1820. ACM, 2013.

[10]. [10] Itziar Landa-Torres, Emilio G. Ortiz-García, Sancho Salcedo-Sanz, María J. Segovia-Vargas, Sergio GilLópez, Marta Miranda, Jose M. Leiva-Murillo, and Javier Del Ser, "*Evaluating the Internationalization Success of Companies Through a Hybrid Grouping Harmony Search—Extreme Learning Machine Approach",* IEEE, pg-no:388-399, 2012.

[11]. [11] Sönke Albers, Michel Clement, "*Analyzing the Success Drivers of e-Business Companies",* M. Van Gelderen, R. Thurik, and N. Bosma. Success and risk factors in the pre-start-up phase. Small Business Economics, pg-no:365–380, 2005.

[12]. [12] Hermann, B. L., Gauthier, J., Holtschke, D., Bermann, R. D., & Marmer, M. (2015). *"The Global Startup Ecosystem Ranking 2015. The Startup Ecosystem Report Series, (August)".*

[13]. [13] Marco Felgueiras, Fernando Batista, Joao Paulo Carvalho, "*Creating Classification Models from Textual Descriptions of Companies Using Crunch base",* IPMU 2020, CCIS 1237.

[14]. Ma et al., in "Dynamic Route Optimization Based on Real-time Traffic and Environmental Data", published in Transportation Research Part C: Emerging Technologies, Vol. 130, 2021.

[15]. [15] D. C. McClelland. "Characteristics of successful Entrepreneurs". The journal of creative behavior, 21(3):219–233.