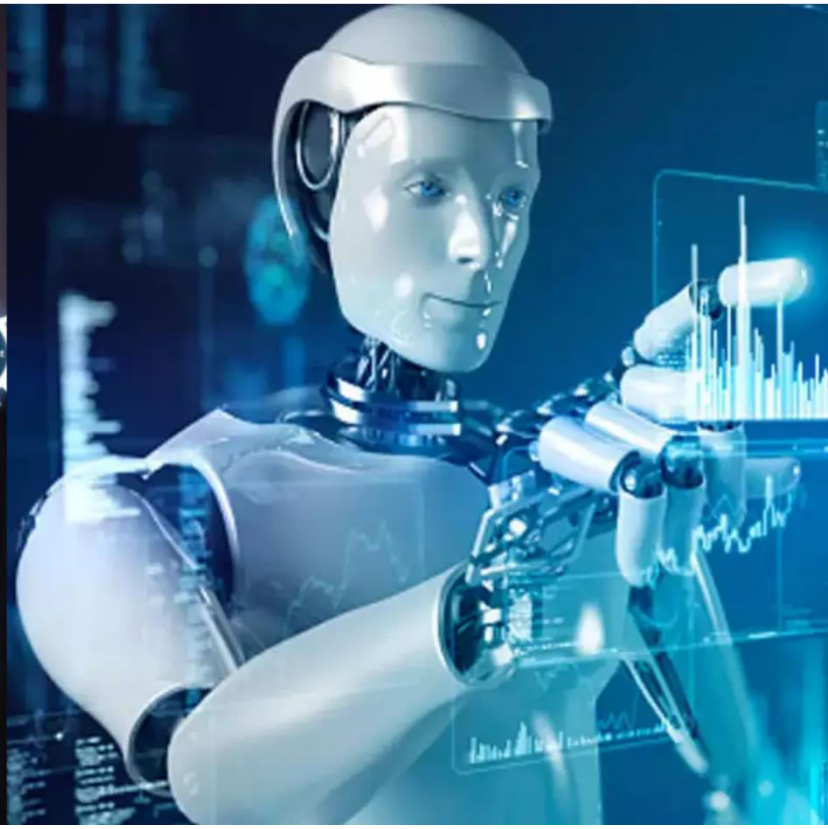
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*

**Impact Factor: 8.699**

**Volume 15, Issue 3, March 2026**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|

# Prosperity Prognosticator: Machine Learning for Start-Up Success Prediction

**Yarragunta Sri Varsha[1], Galam Lakshmi Priya[2], Lanke Naga Raju[3], Koyyalamudi Mojesh[4], Padala Venkateswara Reddy [5]**

Assistant Professor, Kallam Haranadhareddy Institute of Technology, Guntur, Andhra Pradesh, India[1]

U.G. Students, Department of CSE (Data Science), Kallam Haranadhareddy Institute of Technology, Guntur, Andhra Pradesh, India [2,3,4,5]

**ABSTRACT:** Startups play a significant role in economic growth and innovation, but predicting their success remains a challenging task due to various financial, market, and organizational factors. Investors and entrepreneurs often rely on intuition and limited analysis to make investment and strategic decisions, which may lead to financial risks and unsuccessful ventures. This paper presents a machine learning-based approach for predicting startup success using historical startup data and key business indicators. The proposed system utilizes data preprocessing, feature selection, and classification algorithms to analyze startup characteristics and predict whether a startup is likely to succeed or fail. Multiple machine learning algorithms were evaluated, and the Random Forest algorithm was selected as the final model due to its higher accuracy and better performance. The model was trained and tested using startup datasets, and performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The final model was deployed using a Flask-based web application that allows users to input startup details and receive prediction results through a user-friendly interface. The proposed system helps investors, entrepreneurs, and policymakers make data-driven decisions and reduce risks associated with startup investments. This study demonstrates how machine learning can be effectively applied in predictive analytics for business decision support systems.

## I. INTRODUCTION

Startups play a crucial role in economic development, innovation, and job creation. However, a large number of startups fail within the first few years due to various factors such as poor market analysis, lack of funding, ineffective management, and improper strategic planning. Investors and entrepreneurs often face significant challenges in predicting whether a startup will succeed or fail before making investment or business decisions. Traditional decision-making methods rely heavily on experience, intuition, and manual analysis, which may lead to inaccurate predictions and financial losses.

With the advancement of data science and machine learning, predictive analytics has become an effective approach for analysing large datasets and identifying patterns that can support decision-making. Machine learning algorithms can analyse historical startup data, funding details, market trends, and business characteristics to predict the probability of startup success. This enables investors, entrepreneurs, and policymakers to make informed decisions based on data rather than assumptions.

This project proposes a machine learning-based system for predicting startup success using classification algorithms. The system uses historical startup data and applies data preprocessing, feature selection, and model training techniques to build a predictive model. Multiple machine learning algorithms were evaluated, and the Random Forest algorithm was selected due to its higher accuracy and ability to handle structured data effectively.

The developed model is integrated into a web-based application using the Flask framework, where users can input startup-related parameters and receive prediction results indicating whether a startup is likely to succeed or fail. The system aims to provide a decision support tool that helps reduce investment risk and improve strategic planning for new ventures.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|

## II. LITERATURE SURVEY

Startup success prediction has gained significant attention in recent years due to the high failure rate of startups and the increasing availability of business, financial, and market data. Researchers and analysts have attempted to use data mining and machine learning techniques to predict whether a startup will succeed or fail based on various factors such as funding amount, team size, business model, market competition, and growth indicators. Traditional methods for evaluating startup success relied heavily on expert opinions, manual financial analysis, and market research. However, these methods are often time-consuming, subjective, and prone to human bias. With the advancement of machine learning and predictive analytics, automated prediction systems have been developed to analyse historical data and identify patterns that influence startup success. Machine learning models can process large amounts of structured and unstructured data and provide predictions that support investors, entrepreneurs, and policymakers in making data-driven decisions.

Several machine learning algorithms have been used in previous studies for classification and prediction tasks related to business success and financial forecasting. Logistic Regression is commonly used for binary classification problems and provides interpretable results, but it may not perform well when the relationship between variables is nonlinear. Support Vector Machine (SVM) is another powerful classification algorithm that can handle high-dimensional datasets and create optimal decision boundaries, but it requires careful parameter tuning and may not be efficient for very large datasets. Decision Tree algorithms are widely used because they are easy to understand and interpret, but they often suffer from overfitting when trained on complex datasets. To overcome these limitations, ensemble learning methods such as Random Forest have been introduced, which combine multiple decision trees to improve prediction accuracy and reduce overfitting. Random Forest is particularly effective for structured datasets and can handle missing values, feature importance evaluation, and nonlinear relationships between variables.

Recent research indicates that ensemble methods such as Random Forest and Gradient Boosting outperform traditional machine learning algorithms in many prediction problems involving business analytics and financial risk analysis. These models provide higher accuracy, better generalization, and improved performance when dealing with complex datasets. Many studies have also focused on using machine learning models for investment decision support systems, where predictive models analyze startup features and estimate the probability of success or failure. These systems help reduce investment risk and improve decision-making by providing objective, data-driven insights.

## III. METHODOLOGY

The startup success prediction system using machine learning is developed through a series of well-defined steps. The complete workflow of the project is explained below:

**1. Data Collection:**
The first step involves collecting startup-related data from a reliable source such as the Startup Success Prediction dataset available on Kaggle. The dataset includes important startup parameters such as funding amount, number of funding rounds, startup category, company age, team size, and market-related features. These parameters are selected because they play a significant role in determining whether a startup will succeed or fail.

**2. Data Preprocessing:**
Before training the model, the collected data is pre-processed to ensure data quality and consistency. Missing values are handled using appropriate techniques such as mean or median imputation. Duplicate records and irrelevant features are removed to improve data quality. Categorical features such as startup category and location are converted into numerical values using encoding techniques. Feature scaling is also applied to normalize numerical values so that all features are within a similar range, which improves the performance of machine learning algorithms.

**3. Feature Selection:**
In this step, the most relevant features that significantly influence startup success are selected. Feature selection helps in reducing noise, improving model performance, and reducing computational complexity. Correlation analysis and feature importance techniques are used to identify the most important features affecting startup success prediction.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|

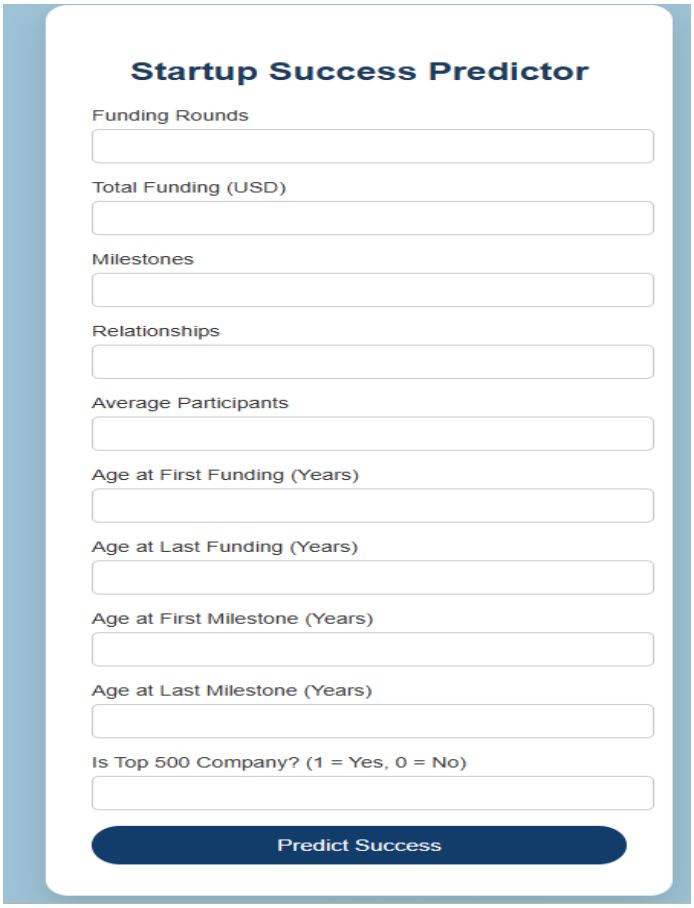
**4. Splitting the Dataset:**
The processed dataset is divided into two parts: training data and testing data. Generally, 80% of the dataset is used for training the machine learning model, and the remaining 20% is used for testing the model. This helps in evaluating the model performance on unseen data and prevents overfitting.

**5. Model Training:**
Multiple supervised machine learning algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest are trained using the training dataset. Each model learns patterns from the input features and their relationship with the target variable, which indicates whether the startup is successful or unsuccessful. Among these algorithms, Random Forest is selected because it provides higher accuracy and handles structured data efficiently.

**6. Model Evaluation:**
After training the models, they are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. These evaluation metrics help in comparing the performance of different models. The model that provides the best accuracy and balanced performance across all evaluation metrics is selected as the final model.

**7. Prediction and Deployment:**
The final step involves using the trained Random Forest model to predict startup success. The model is deployed using a Flask web application where users can enter startup details through a web interface. The system processes the input data and provides the prediction result indicating whether the startup is likely to succeed or fail. This makes the system user-friendly and useful for investors and entrepreneurs for decision-making.

## IV. RESULTS AND DISCUSSION



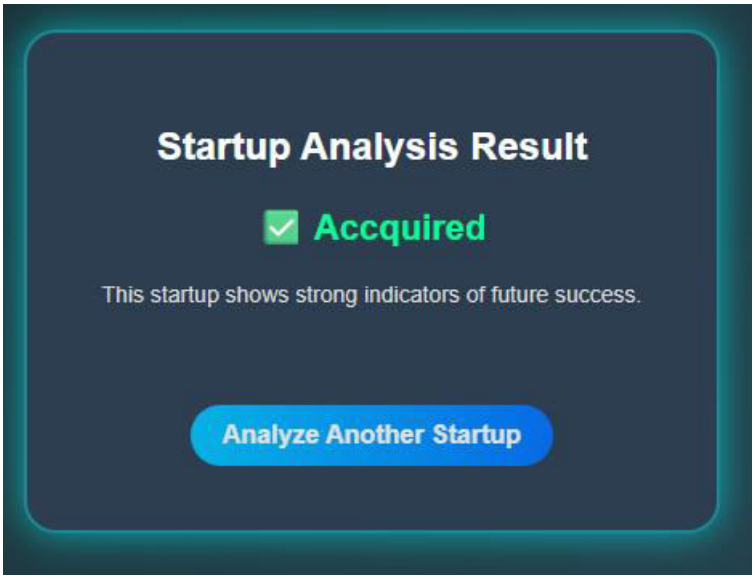
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|




**Accuracy and Precision:**
The trained machine learning model achieved an overall accuracy of approximately 81% on the test dataset. The class-wise performance metrics showed strong predictive performance for successful startups (Class 1) with higher precision and recall compared to failed startups (Class 0). The classification report indicated that the model achieved balanced precision, recall, and F1-score values, demonstrating that the model can effectively classify startup success and failure based on the selected features. The weighted average precision, recall, and F1-score values were also satisfactory, indicating that the model performs well even when the dataset contains class imbalance. These results show that the Random Forest model can successfully identify important patterns related to startup success prediction.

**Prediction Performance and Evaluation Metrics:**
Additional evaluation metrics such as precision, recall, F1-score, and ROC-AUC score were calculated to evaluate the model performance more accurately. The ROC-AUC score indicated good classification capability of the model in distinguishing between successful and unsuccessful startups. The model also showed consistent performance across both training and testing datasets, although slightly higher training accuracy suggested minor overfitting. However, the model still performed well on unseen data, indicating good generalization ability.

**Model Comparison:**
A comparison of different machine learning algorithms was performed to identify the best model for startup success prediction. Algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest were trained and evaluated. The Random Forest model achieved the highest accuracy compared to other algorithms, while Logistic Regression and Decision Tree showed moderate accuracy. Support Vector Machine performed reasonably well but required more parameter tuning. The comparison results indicate that ensemble learning methods such as Random Forest perform better than individual classification models for structured business datasets.

**Confusion Matrix Analysis:**
The confusion matrix illustrates the distribution of predicted and actual startup outcomes. The majority of successful startups were correctly classified, indicating good prediction capability of the model for positive outcomes. However, some unsuccessful startups were misclassified as successful due to similarities in feature patterns such as funding and company size. These misclassifications indicate that additional feature engineering and data balancing techniques could further improve model performance. Overall, the confusion matrix shows that the model performs effectively in predicting startup success and failure with acceptable error rates.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|

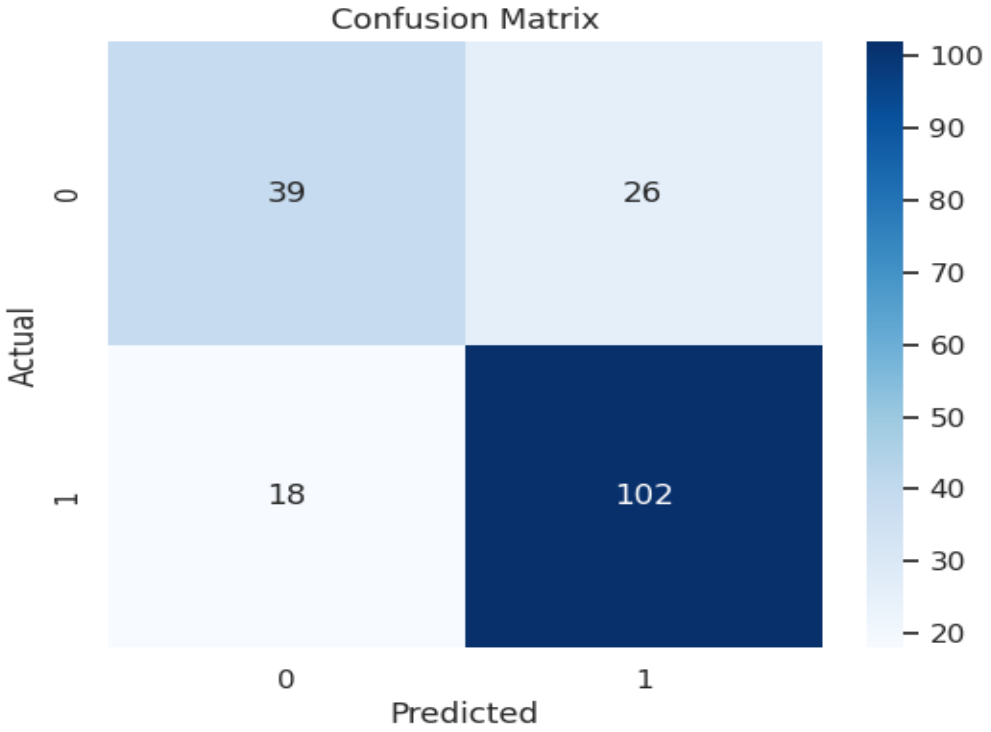



Fig 3 Confusion matrix

## V. LIMITATIONS

**1. Limited and Imbalanced Dataset:**
The dataset used for training the startup success prediction model contained an uneven distribution between successful and unsuccessful startups. This class imbalance affected the model's ability to accurately predict minority class outcomes, which may lead to lower recall values for certain categories.

**2. Dependence on Historical Data Quality:**
The performance of the machine learning model heavily depends on the quality and completeness of the historical startup data. Missing values, incorrect funding information, or inconsistent records in the dataset can negatively impact the prediction accuracy.

**3. Limited Feature Availability:**
The dataset used in this project contains only a limited number of startup-related features such as funding amount, category, and company age. Other important factors such as founder experience, market competition, business strategy, and customer growth were not included, which may affect the overall prediction performance.

**4. Model Overfitting Risk:**
The Random Forest model showed very high accuracy on the training data compared to the testing data, indicating a possibility of overfitting. This means the model may perform very well on known data but slightly less accurately on completely new data.

**5. Prediction Uncertainty:**
Startup success depends on many external factors such as economic conditions, market trends, and management decisions, which cannot be fully captured using historical data alone. Therefore, the model predictions should be considered as supportive decision-making tools rather than absolute predictions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|

**6. Deployment and Real-Time Data Limitations:**
The current system uses a static dataset and does not update automatically with real-time startup data. For real-world applications, continuous data updates and model retraining would be required to maintain prediction accuracy

## VI. FUTURE WORK

**1. Dataset Expansion and Improvement:**
Future work will focus on collecting larger and more comprehensive startup datasets that include additional features such as founder experience, employee growth, revenue trends, and market competition. Expanding the dataset will help improve the accuracy, reliability, and generalization capability of the prediction model.

**2.Implementation of Advanced Machine Learning Models:**
Future research can explore the use of advanced machine learning and deep learning algorithms such as Gradient Boosting, XGBoost, Artificial Neural Networks, and other deep learning models to further improve prediction performance and model efficiency.

**3. Real-Time Prediction System:**
The system can be enhanced to support real-time data input and continuous model updating, allowing predictions to be generated using the most recent startup data and current market trends. This will make the system more practical and applicable in real-world scenarios.

**4. Web Application Enhancement:**
The web-based system can be further improved by adding user dashboards, prediction history, graphical analysis, and automated report generation features to enhance usability and provide a better user experience.

**5. Integration with Financial and Market Data Sources:**Future systems can integrate external financial APIs, market trend data, and economic indicators to improve prediction accuracy and provide more comprehensive decision support for investors, entrepreneurs, and stakeholders.

**6. Model Optimization and Automation:**
Future work can include automatic hyperparameter tuning, automated model retraining, and deployment using cloud platforms. This will make the system more scalable, efficient, and accessible to a larger number of users.

## VII. CONCLUSION

In this paper, a machine learning-based approach for predicting startup success has been presented. The main objective of this project was to develop a predictive model that can analyse startup-related features and determine whether a startup is likely to succeed or fail. The system was developed using historical startup data and various machine learning techniques including Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest algorithms. After comparing the performance of different models, the Random Forest algorithm was selected as the final model due to its higher accuracy and better performance in handling structured datasets.

The methodology involved data collection, data preprocessing, feature selection, model training, and model evaluation using performance metrics such as accuracy, precision, recall, and F1-score. The trained model was then deployed using a Flask-based web application where users can input startup details and obtain prediction results. The results demonstrate that machine learning techniques can be effectively used for predictive analytics in business decision-making and investment analysis.

The proposed system helps investors, entrepreneurs, and policymakers make data-driven decisions by providing startup success predictions based on historical data patterns. This system can reduce investment risk, improve strategic planning, and support better decision-making for startup investments. Overall, the project demonstrates the practical application of machine learning in predictive business analytics and decision support systems.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503320|

## VIII. ACKNOWLEDGMENT

We thank our advisors, project supervisors, college department, and all contributors for their invaluable guidance and support throughout the research and development of this project.

## REFERENCES AND RESOURCES

1. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
2. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
3. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.
4. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.
5. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
6. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.
7. Kaggle, "Startup Success Prediction Dataset," [Online]. Available: https://www.kaggle.com/
8. W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, 2010.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details