Prosperity Prognosticator: Machine Learning for Startup Success Prediction

TEAM ID- SWTID1749880888
(YASHIKA PANDA, ANAMAY RAI, VANSH ARORA)

1. Introduction

1.1. Project overviews

The *Prosperity Prognosticator* project is a machine learning-based solution designed to instantly predict the success potential of startups by analyzing key factors such as funding history, business category, milestones, and team structure. By delivering real-time, data-driven insights, the model supports investors in identifying promising ventures, helps entrepreneurs refine their strategies, and enables policymakers to develop targeted programs that foster innovation. This project aims to reduce uncertainty, improve decision-making, and strengthen the overall startup ecosystem through fast, accurate, and actionable predictions.

1.2. Objectives

- Develop a machine learning model to predict startup success based on historical and operational data.
- Provide instant, real-time predictions to assist stakeholders in making timely and informed decisions.
- Identify key success factors that influence startup outcomes across different industries and regions.
- Support investors in optimizing funding strategies by highlighting high-potential startups.
- Help entrepreneurs improve planning and reduce risks through data-driven insights.
- Enable policymakers to design targeted support programs that promote innovation and economic growth.

2. Project Initialization and Planning Phase

2.1. Define Problem Statement

The current approach to evaluating and supporting startups presents key challenges for investors, entrepreneurs, and policymakers, impacting decision quality and confidence. Investors face difficulty identifying high-potential ventures, entrepreneurs lack clarity on success factors, and policymakers struggle to design targeted support. These gaps lead to missed opportunities, inefficiencies, and reduced trust in the startup ecosystem. By addressing these issues with a machine learning–based success prediction model, we aim to simplify decisions, reduce risk, and improve outcomes across all stakeholder journeys.

2.2. Project Proposal (Proposed Solution)

This project proposal presents Prosperity Prognosticator, a machine learning solution designed to predict the success potential of startups using key business and funding indicators. It aims to support data-driven decision-making for investors, entrepreneurs, and policymakers through real-time, accurate predictions. The project will use supervised machine learning models trained on historical startup data, supported by data preprocessing, feature engineering, and model tuning. The solution includes the implementation of a success prediction model that enables real-time decision-making, allowing for quicker investments and more informed strategic planning.

2.3. Initial Project Planning

The Prosperity Prognosticator project is divided into four sprints to ensure organized development and timely delivery.

Sprint 1: Data Collection and Preparation

Focuses on gathering the dataset, importing necessary libraries, and loading the data for processing.

Sprint 2: Exploratory Data Analysis

Includes visualizing the data, conducting univariate and multivariate analysis, preprocessing, and splitting the data for training and testing.

Sprint 3: Model Building and Testing

Covers preprocessing, training supervised machine learning models, and evaluating performance using accuracy and other metrics.

Sprint 4: Model Deployment

Involves saving the best model, integrating it with a web framework, building frontend interfaces, and deploying the application for real-time predictions.

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

The data for this project is sourced from the Kaggle Startup Success Prediction Dataset. It contains detailed information on over 9,000 startups, including funding rounds, total funding amount, milestones, team structure and final status (operating, acquired, or closed). This dataset will be used to train and evaluate the machine learning model. Data preprocessing, cleaning, and feature selection will be carried out to ensure quality and relevance for accurate predictions.

3.2. Data Quality Report

The dataset underwent a data quality check to ensure it was clean, relevant, and suitable for machine learning. Irrelevant columns were dropped, as they do not contribute to prediction and add unnecessary complexity. Columns with excessive missing values or no meaningful variation were also removed. Duplicate entries were identified and eliminated, and data types were standardized across all features. These steps helped streamline the dataset, reduce noise, and improve its overall quality for accurate and efficient model training.

3.3. Data Exploration and Preprocessing

Data exploration was conducted to examine the startup dataset and uncover patterns, distributions and correlations that influence success outcomes. This helped in identifying the most impactful features and understanding the structure of the data. Preprocessing involved handling missing values, encoding categorical variables, and scaling numerical features to prepare the data for model training. Additionally, feature engineering was applied to create new variables such as funding stage indicators and average investors per round, which added more depth to the model and improved its predictive performance. These steps ensured the dataset was clean, consistent, and optimized for effective machine learning.

4. Model Development Phase

4.1. Feature Selection Report

A focused feature selection process was carried out to identify the most relevant attributes that contribute to predicting startup success. Initially, the dataset contained several features, but not all were useful for modeling. Irrelevant and redundant columns were removed based on exploratory analysis, correlation checks, and domain understanding. The final selected features such as funding rounds, milestones, total funding, and investor participation were chosen for

their strong influence on startup outcomes. This selection not only reduced data complexity but also improved model accuracy and efficiency by retaining only the most impactful and meaningful variables.

4.2. Model Selection Report

Random Forest was chosen as the final model for our Startup Success Predictor due to its strong performance, reliability, and ability to handle complex datasets with mixed feature types. During model evaluation, it consistently outperformed alternatives by achieving the highest accuracy. While KNN struggled with high dimensional data and sensitivity to feature scaling, Decision Tree models, although interpretable, showed a tendency to overfit and lacked stability across different data splits. In contrast, Random Forest's ensemble approach combining multiple decision trees enabled it to capture underlying patterns more effectively while reducing the risk of overfitting. It also performed exceptionally well on structured data, offered valuable insights through feature importance metrics, and required minimal tuning. These strengths made Random Forest the most suitable and reliable choice for predicting startup success in our project.

4.3. Initial Model Training Code, Model Validation and Evaluation Report

The initial model training involved applying selected machine learning algorithms on the preprocessed startup dataset to establish baseline performance. Each model was trained using the engineered features to evaluate its predictive ability. The subsequent validation and evaluation phase used metrics such as accuracy, precision, and recall to assess model effectiveness. Among all, Random Forest demonstrated the highest accuracy and overall robustness, making it the most suitable model for predicting startup success and supporting reliable, data-driven decision-making.

5. Model Optimization and Tuning Phase

5.1. Hyperparameter Tuning Documentation

The Random Forest model was selected based on its strong performance during hyperparameter tuning, where it consistently achieved the highest accuracy score of 81% among all tested models. As an ensemble of multiple decision trees, Random Forest improves accuracy and reduces overfitting by randomly sampling data and features for each tree and averaging their predictions. During tuning, key parameters such as the number of estimators, maximum tree depth and minimum samples required for splits were adjusted to optimize model performance.

This process enhanced the model's predictive power and generalization. Its robustness, flexibility to handle various types of features, and consistent results make Random Forest the most reliable and effective choice for predicting startup success in this project.

5.2. Performance Metrics Comparison Report

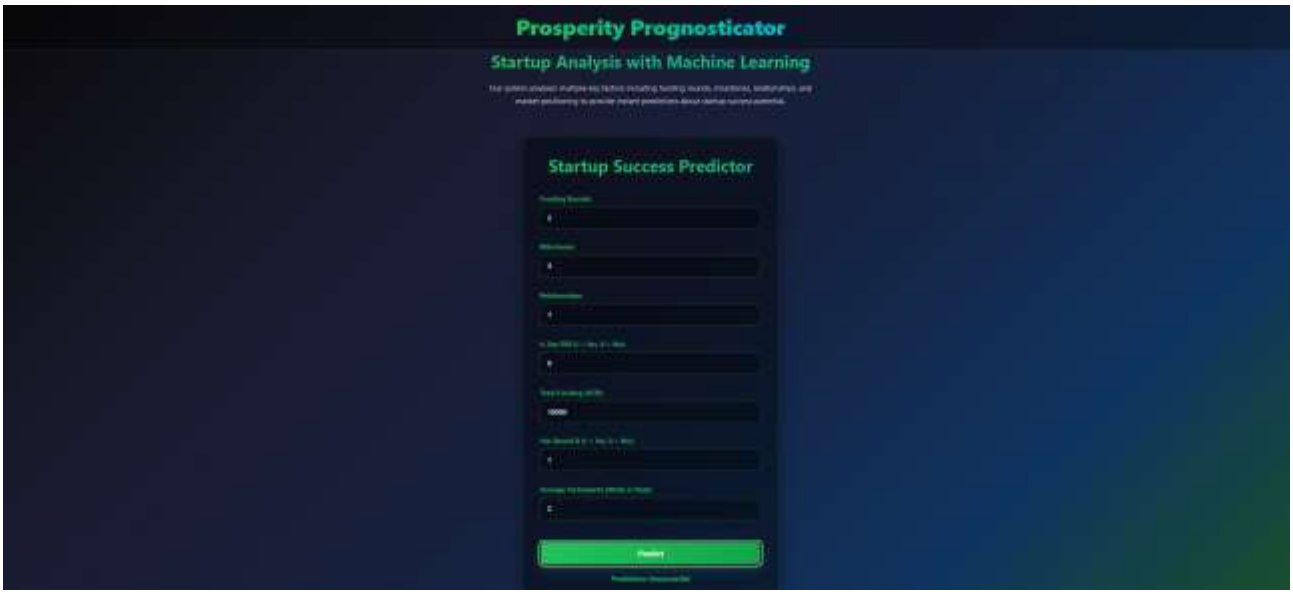
The Performance Metrics Comparison Report presents a detailed analysis of baseline versus optimized metrics across different models, including K-Nearest Neighbors, Decision Tree and Random Forest. While all models showed improvement after tuning, the report specifically highlights the superior performance of the Random Forest model, which achieved the highest accuracy and balanced precision-recall scores. This comparison underscores the effectiveness of hyperparameter tuning in enhancing predictive accuracy and demonstrates Random Forest's reliability as the final model for startup success prediction.

5.3. Final Model Selection Justification

The Final Model Selection Justification explains the rationale for choosing Random Forest as the final model for predicting startup success. Its superior accuracy, robustness against overfitting, and effective handling of both categorical and numerical features made it the most reliable choice. Through successful hyperparameter tuning, Random Forest demonstrated consistent performance and strong generalization, aligning well with the project's goal of delivering accurate, real-time predictions to support data-driven decisions across stakeholders.

6. Results

6.1. Output Screenshots



7. Advantages & Disadvantages

Advantages

- High accuracy in predicting startup outcomes.
- Handles complex feature interactions like funding and milestones effectively.
- Provides feature importance, helping identify key success factors.
- Robust against overfitting, ensuring reliable predictions.

Disadvantages

- Less interpretable, making it harder to explain individual predictions.
- Higher resource usage during training and deployment.
- Slightly slower than simpler models for real-time predictions.

8. Conclusion

The Prosperity Prognosticator project successfully developed a machine learning model to predict the success of startups using key business, funding, and operational features. Through systematic data collection, exploration, feature engineering, and model evaluation, Random Forest emerged as the most accurate and reliable algorithm. The final model enables real-time predictions, offering valuable insights for investors, entrepreneurs, and policymakers. By leveraging data-driven decision-making, the project enhances startup evaluation, reduces risk, and contributes to building a stronger, more informed startup ecosystem.

9. Future Scope

The Prosperity Prognosticator project has strong potential for future enhancement and expansion. Additional datasets can be integrated to improve prediction accuracy and generalizability across different regions and industries. Advanced models like XGBoost or deep learning can be explored for improved performance. The web application can be extended to include interactive dashboards, personalized recommendations, and real-time data updates. Further, integrating external data sources such as market trends, social media sentiment, and economic indicators could provide deeper insights. Ultimately, the model can evolve into a comprehensive startup evaluation tool for global use.