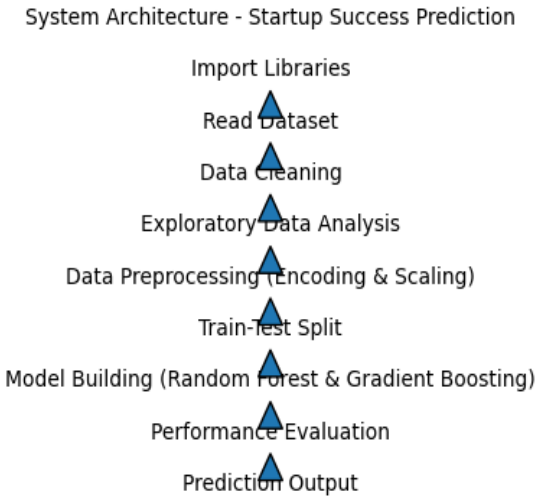
Startup Success Prediction Using Machine Learning

Abstract

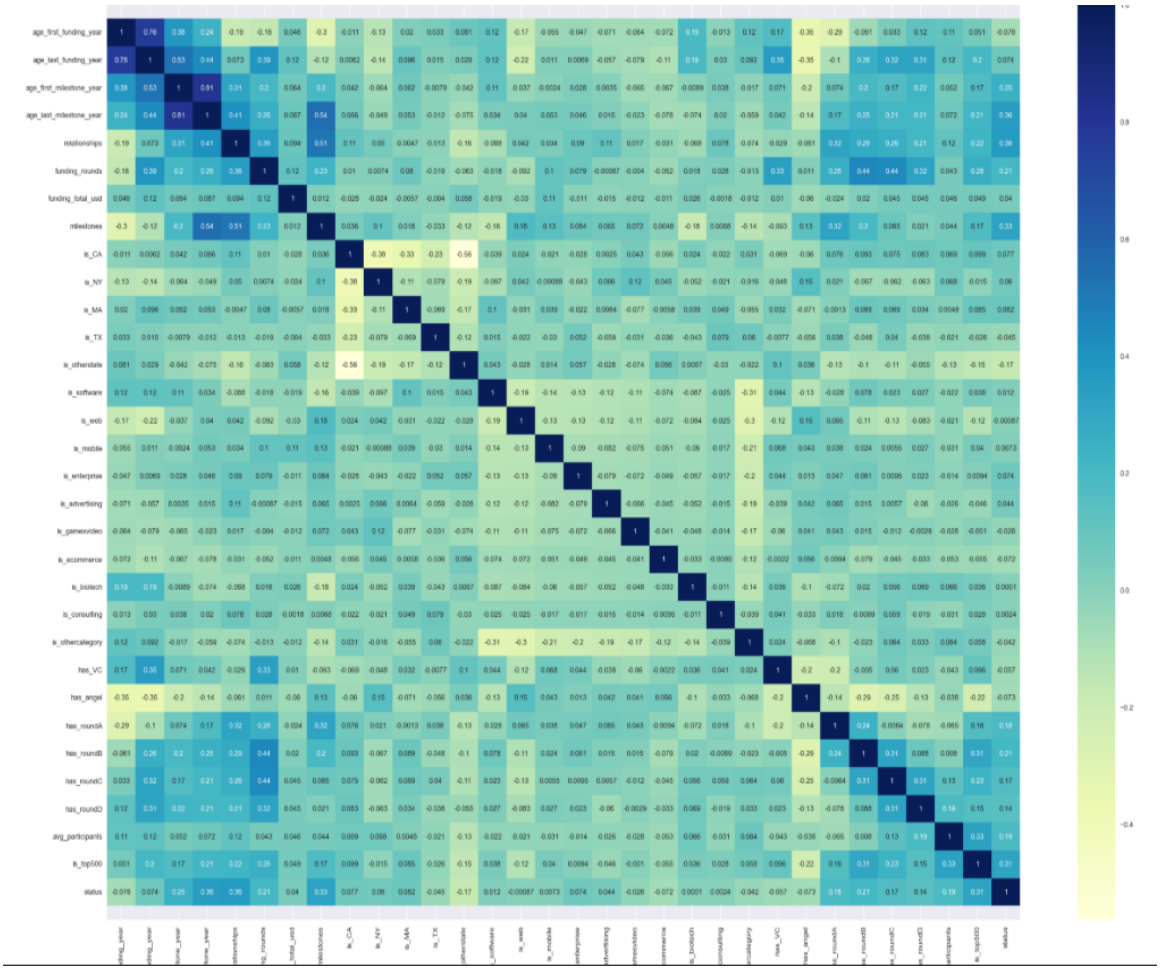
This project aims to predict startup success using machine learning algorithms. The workflow follows structured steps including data collection, preprocessing, model building, and performance evaluation. Random Forest and Gradient Boosting algorithms were implemented and compared based on evaluation metrics.

I. Project Workflow

II. System Architecture



III. Graphical Analysis (Correlation Heatmap)



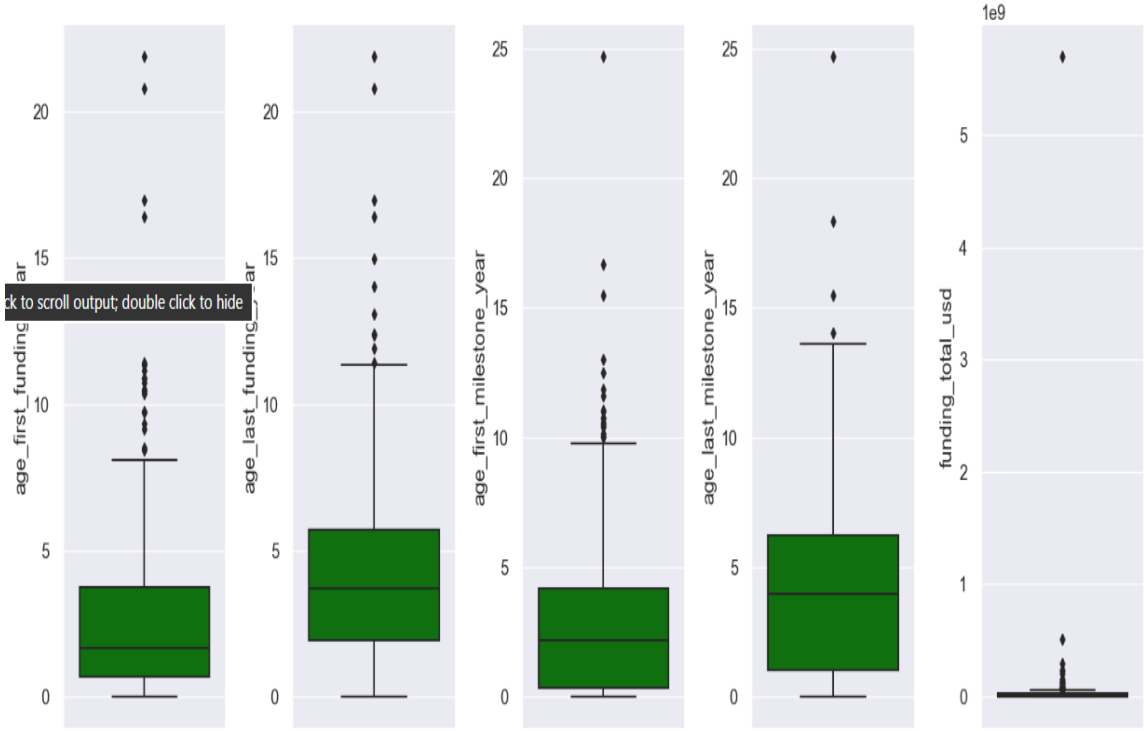
IV. Missing Values Analysis

```
Missing Values
```

```
null=pd.DataFrame(df.isnull().sum(),columns=["Null Values"])
null["% Missing Values"]=(df.isna().sum()/len(df)*100)
null = null[null["% Missing Values"] > 0]
null.style.background_gradient(cmap='viridis',low =0.2,high=0.1)
```

	Null Values	% Missing Values
Unnamed: 6	493	53.412784
closed_at	588	63.705309
age_first_milestone_year	152	16.468039
age_last_milestone_year	152	16.468039
state_code.1	1	0.108342

V. Outlier Detection



VI. Conclusion

The comparative analysis between Random Forest and Gradient Boosting revealed that Random Forest achieved a higher Recall score, making it more effective in identifying successful startups. On the other hand, Gradient Boosting achieved a higher Precision score, indicating better performance in minimizing false positives. Based on the evaluation, Random Forest is more suitable when recall is prioritized, whereas Gradient Boosting is preferable when precision is the main concern.