STARTUP SUCCESS RATE PREDICTION

Amogh Varsh Raju[1], Anugam Sai Kiran[2], Asmath Fathima[3]

[1]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
[2]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
[3]Department of Electronics & Communication Engineering SR University, Warangal, Telangana, India. [4]

*Email: varshrajuambati@gmail.com

Abstract: Startups have bought a huge change in the world of economy. Over the past few years there has been an vast growth. Its all about the plan of action and organization of the ideas that make it successful so as to attract the new investors and entrepreneurs. The success prediction is the vital criteria for new ventures in the way for investments , since the first step in the path of success is the potential of growing . With the developments in the model with new features using different data-set, we can make it more effectively . The results are used by enriching data and deploying these machine learning models through research and making its performance more accurate and dynamic to work.

1. INTRODUCTION

This paper focuses primarily on Startup and its success prediction with help of different machine learning models. Startup and new ventures have become an essential element of innovation. They are rapidly growing, and need monitoring of the performance and enhancing it. Prior to the startup it's important to analyze what makes it successful. In the approach of goal we have to go beyond the data previously used with different success factors.
Predicting success is defined based on the money invested by founders and investors and the in turn benefit they achieve for the shares.
With a focus on how to explore and make a better decision of making investment strategy and gain by applying improved machine learning we can make the process more effective.

To attain success accuracy we can work on the following factors :
a. Explore new data variables, by brain storming with stakeholders and investors.
b. Implement more accurate models to enhance prediction on startup.
The insights provided will be valuable for stakeholders and venture capitals. The predictive model to explain the phenomena is the indicator of data mining that allow to achieve full potential of data.

2. PROBLEM DEFINATION

using various models of machine learning. This predicts whether or not the patient has a heart attack. We use numerous Machine learning frameworks including pandas, matplotlib, sci-kit-learn, Keras etc. to analyze such a model.

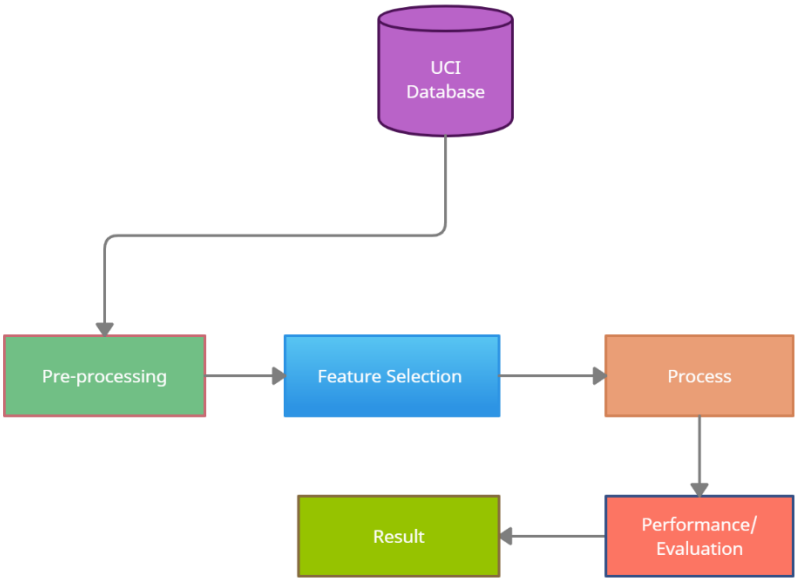


Figure 2. Processing Steps for Machine Learning for Heart Disease Detection.

3. DATASET AND ATTRIBUTES

We collected the data from the Kagil Startup success prediction question. We use 7 attributes to predict whether or not a startup is successful or failure.

Labels	The Startup has marketing labels or not.	Nominal
Age first Year	Age of the company when it first received its funding	Nominal
Age first milestone	Age of the company when it reached its first milestone.	Numeric
Relationships	The company relationship rating between its users and customers.	Numeric
Milestones	The number of milestones the company has reached until now	Nominal
Avg-Participants	Amount of employees in the company including the founder and Directors.	Nominal
Is top 500	Did the startup or company reach the top places globally in top 500	Numeric

Table 1. kagil Data-set attributes detailed information

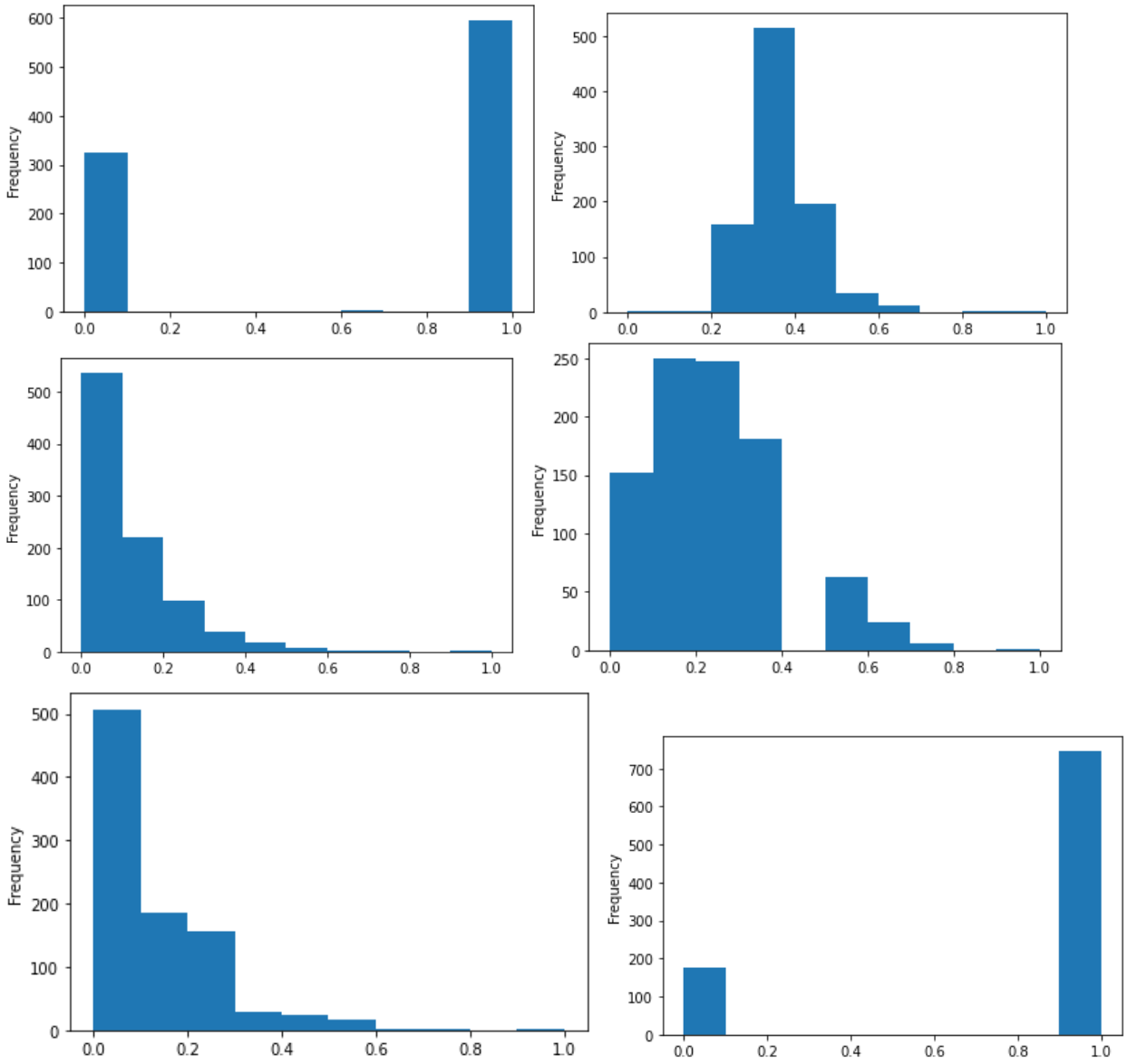


Figure 3 visualizing all attributes

1. DATA PRE-PROCESSING

After an assortment of data, The total number of rows available data is 923 rows of data for the attributes of Marketing labels,age_first_year,age_first_milestone_year, relationships, milestones, avg_participants,is_top500 and these 8 columns provide the exact necessary data which is required to be calculated , these columns have been filtered from 44 columns available to 23 columns and the best columns have been selected those are these 8 columns

Labels.	Boolean(The company has,marketing label or not,marketing partner or not, marketing partner is present or not.)
age_first_year.	Float(The age of the company:When it received its first funding,first investment ,first transaction)
Age_milestone_year.	Float(The age of the company when it received:its first success,its first threshold limit,its first profit,its first pitch,its first user base.)
Relationships	Float(The number of people who are working with the company and are purchasing from the company.)
Milestones	Float(The number of milestones reached by the company)
Avg_participants	Float(The number of employees along with the founder and management team of the company)
Is_top500	The company is in top 500 companies world wide or is available in the magazines or is it well known.

Table 2. Attributes Validation.

So, from the above graph, we can predict that the startup can be successful or not.

Correlation Matrix

A correlation matrix is simply a table that displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. coefficients for different variables.

How it is calculated?

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (X_i) in the table is correlated with each of the other values in the table (X_j)... The diagonal of the table is always a set of ones because the correlation between a variable and itself is always 1. Let's perform the Correlation matrix to understand the relation between the dependent variable and the independent variable and within the independent variable.

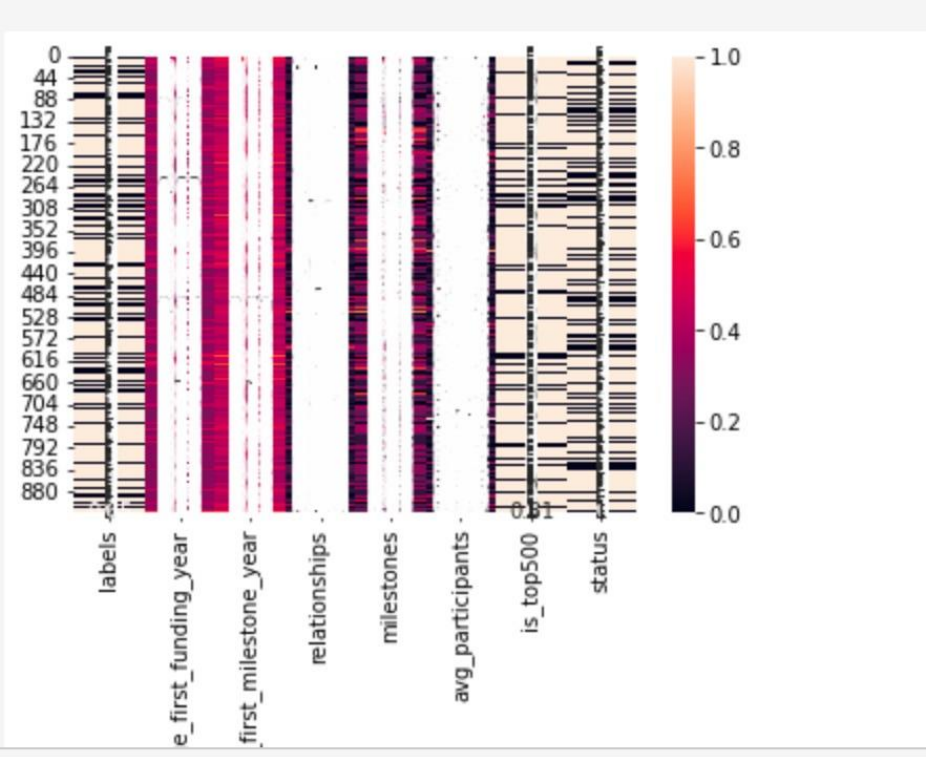


Figure 7 Correlation Matrix

Enough EDA performs on the data to evaluate the data-set and gather knowledge about the data. Let's perform some Machine Learning model and Experimentation to create a model that helps us to achieve our goal we state in the problem definition.

2. ALGORITHMS

We use different machine learning model to solve our classification problem:

1. Logistic Regression
2. K-Neighbour Classifier
3. Random Forest Classifier
4. SVC- Support vector classifier.

So, let us make our data ready for training and testing our machine learning model.

Logistic Regression

Logistic regression is a supervised algorithm for study classification. The likelihood of a destination variable was predicted. The nature of the target or dependent variable is dichotomous, meaning that only two possible classes are available

K-Nearest Neighbour

The K-Nearest-Neighbors (KNN) method of classification is one of the simplest methods in machine learning and is a great way to introduce yourself to machine learning and classification in general. At its most basic level, it is essentially classification by finding the most similar data points in the training data and making an educated guess based on their classifications. Although very simple to understand and implement, this method has seen wide application in many domains, such as in **recommendation systems**, **semantic searching**, and **anomaly detection**

Random Forest

Random forest is used for both regression and classification-based applications. This algorithm is flexible and easy to use. Most of the times this algorithm gives accurate results even without hyper tuning the parameters. It builds many decision trees which on merging forms as a forest. While building the decision trees, adds more randomness to the model. This algorithm searches for the best feature in the random subset of features, which results in the formation of a better model.

Support Vector Machine SVC

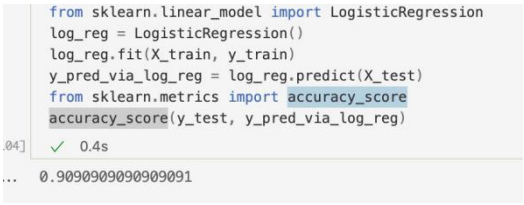
The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N - the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified

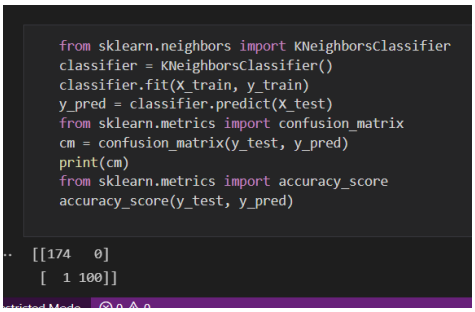
with more confidence.

3. RESULTS

ALGORITHM	ACCURACY RATE
KNN	0.72(72%)
SVC	0.70 (70%)
Decision Tree	0.89(89%)
Random Forest	0.90(90%)
Logistic regression	0.90(90%)

Figure 8. Result of various models with the proposed model





```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)
```

```
[[158  16]
 [ 13  88]]
0.8945454545454545
```

```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred))
print(classifier.get_params())
```

```
[[164  10]
 [ 15  86]]
0.9090909090909091
```

The algorithms that are used in the given model are Logistic regression, K-Nearest Neighbor algorithm, random forest, SVC,Decision tree. On comparision of the above algorithms accuracy rate for the SVC is more than the other algorithms. The above table shows the accuracy rates of the algorithms used.

7. CONCLUSION

The main purpose of the project is to make an efficient model that would classify success of start up companies based on the input features provided .With the use of binary classification we can predict whether a company would be successful or not. We have used 5 models-Logistic Regression, KNN, Random Forest Classifier, Decision Tree Classifier and SVC.SVC provides better accuracy compared to others we used 4 main metrics for prediction : age of first funding , age of last funding year, age at which company achieved its first milestone, is it in top 500 or not. The model with best accuracy is used to find the success rate.

5. REFERENCES

[1] https://startup-prediction.herokuapp.com/

[2]https://www.google.com/url?sa=t&source=web&rct=j&url=https://github.com/RamkishanPanthena/Startup-Success-Prediction&ved=2ahUKEwi91ZfQgtH0AhWV_XMBHVoXCvQQFnoECDIQAQ&usg=AOvVaw298k4PtURYlpuEIQfVG8JG

[3]https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3878-2019.pdf&ved=2ahUKEwi91ZfQgtH0AhWV_XMBHVoXCvQQFnoECA4QAQ&usg=AOvVaw3TTNRVoEVuMz7qCWqODIuk

[4] Graham,P.(2012). Startups equal growth. Web page.

[5]https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.kaggle.com/manishkc06/startup-success-prediction&ved=2ahUKEwjN96DNn9H0AhUWIbcAHTW9CpsQFnoECC8QAQ&usg=AOvVaw0eN-0y_TinT96XAKORBM5C

[6] https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148