Business Success Prediction Using Data Mining

Avinay and Dr. Muthamil Selvan T
School of Information Technology & Engineering
Vellore Institute of Technology, Vellore
avinay1165@gmail.com and tmuthamilselvan@vit.ac.in

ABSTRACT

About 70 percent of the startups fail in an average. The reasons for the failure includes poor management, not enough funds etc. The goal of work is to create a model that forecasts startup based on the several keys involved at various stages of life. It is very appealing to increase the rate of success and apart not much work had been done to address the rate. Various methods have been proposed to predict the outcome of startup based on various keys such as the location of office, product requirements, competition in the market. Each of these factors contribute to the success and failure of the startup at each milestone. Several models on data are created by us that are put together carefully through few sources like Kagal etc. To process the data with optimizations and validations, we used several data mining classifications on pre-processed data. Various techniques were used such as Random Forest, Bayesian networks by us to provide our analysis. The evaluation of accuracy in the model are based on various factors such as accuracy and the area. Through our model, startups can decide the factors that need to pay more attention for their increase in success rate.

1.INTRODUCTION

1.1. OBJECTIVE

As we know in all over the world the job opportunities are decreasing day by day and most of the talented people are seeking for job but not getting it due to lack of opportunities therefore the startups plays the major role in increasing not only the job opportunities but plays the vital role in economic growth of country. As the startups plays such important role therefore it is necessary to know the possible future outcome of the particular startups so that the investors and stakeholder's will not suffer any loses form their investment. Also, the startups companies can start their business by analyzing all the scenario's and risks can be occur in future and by using this model they can reduce the risk of failure in future. So, this model will be helpful for all investors, stakeholder's as well as the startup companies itself.

Here I am going to the supervised machine learning approach using that I will train the model on the basis of historical data of the companies which are either overtaken or closed. On the basis of these company data I will consider the factors here and try to find out which common factor is the responsible for the failure of the company.

1.2. MOTIVATION

In today's emerging technology it has made human life much easier than earlier also in the field of business we are seeing many of startups starting these day but as we see many of them fails due to lack of technology so here I am focusing on the building the system through which we can predict the scope of the startups before it starts here we can predict the success and failure rate of the startups. As we see the data plays very important role to know the behavior of the business so here we are using some data mining and machine learning models to predict the success and failure of the business.

1.3. BACKGROUND

Humans have been mining data from the earth from the centuries just to get all sort of valuable materials. It is basically the extraction of vital information and knowledge from the large sets of information. Companies use this process turn the raw data into useful information. Large amount of information is explored into meaningful patterns as it can be used in many ways such as database, credit management, detecting spam or even recognizing the opinion of users. Various complex algorithms are developed to mine the data from multiple locations and devices to generate report. It helps business to understand the marketing campaigns as well. Business data can generate insights to make a better decision than relying on intuitions and experience. Mining of data is used to understand employee behavior as well as HR policies thus improve performance. Each and every business can have advantage of data mining ranging from small to large companies. The right data helps company to increase their revenues, cut costs and customers.

Random Forest: Random forests algorithm is proposed in this paper as a new framework using data-mining technique in hybrid detection, misuse and anomaly detection. Classification and regression way of approach is used within random forest algorithm that act as an effective technique in data mining. Now a days random forests algorithm are used for different applications. Random Forest algorithms are extensively used in probability estimation and prediction.

Random forest did not apply intrusion detection automatically. In this proposition, the elements used for misused are the random algorithms for classifying intrusion while the anomaly detects mechanism of algorithm detached. Multiple decision trees are built under Random forest and it combines all of them together to generate an accurate and stable prediction.

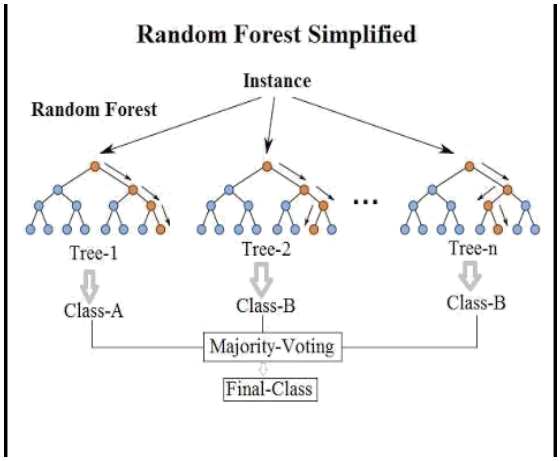


Fig.1.3.1:-Showing Random Forest in simplified way

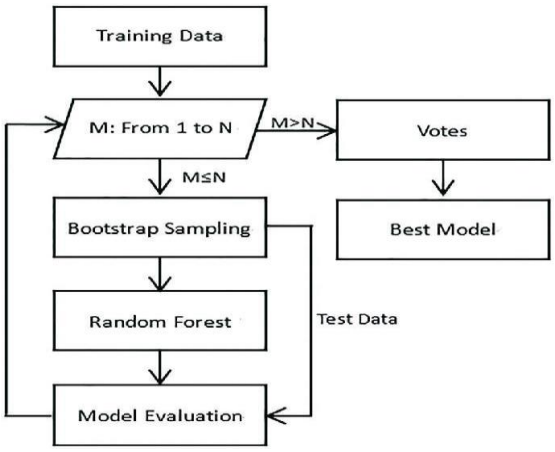


Fig.1.3.2: Workflow diagram of Random Forest

Advantage:

• To solve classification and regression problems, random forest algorithm is always the first choice.

• It is highly flexible and accurate.

• Less variant than Single decision tree as it works perfectly for large range of data.

• Does not require input preparation as there is no need to scale the data.

• Even we lose a large set of data, this algorithm maintains accuracy.

Disadvantage:

• Complexity is one of the main disadvantages of Random forests algorithm. Construction of random forest is much harder, complex and time consuming than decision tree.

• Requirement of computation is more in random forest algorithm. It always hard to get an instinctive grip of bonding present in input data when we get a large collection of decision tree.

• The intrusion predicted by this algorithm is more time consuming.

Decision Tree: Under supervised learning algorithm family Decision Tree algorithm falls. For solving regression and classification problems decision tree algorithms are used like other supervised learning algorithms. The primary motivation for implementing Decision Tree algorithm is to develop a training model that can be used for prediction of class or value for output variables by implementing decision regulations referred from earlier data (training data).

In comparison to other classification algorithms, the level of understanding of decision tree is very easy. The approach used by decision tree algorithm to solve complex problem is done using tree representation. Every interior node of the tree represents an attribute, and every leaf node represents a class label. Disjunctive Normal Form is also known as Sum of Product (SOP) form that is a working are of Decision Tree. The major challenge in decision tree is to identify the attribute used for the construction of root node in each level. And the process of doing this is called as attribute selection. Information Gain, Gini Index are two way of selecting attribute.

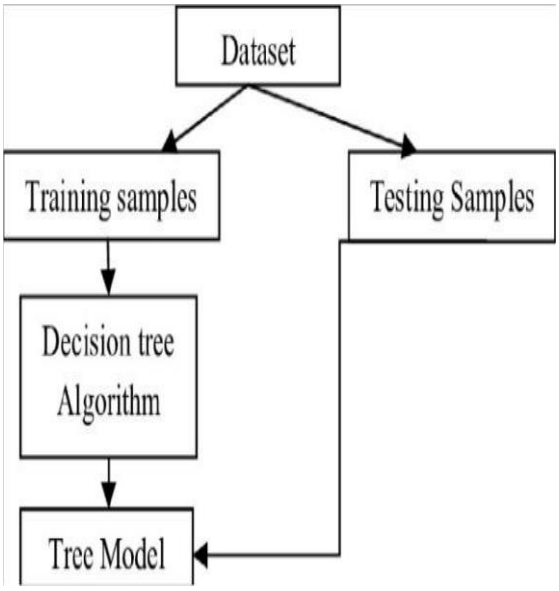


Fig.1.3.3: Workflow diagram of Decision Tree

Decision Tree Algorithm Pseudo code:

1.The best selected attribute from the dataset is placed at the root of the tree.

2. Subsets are spited from the training dataset. Subsets would be selected in a way such that every subset may contain data that have similar value for every attribute.

3. Until we found a leaf node in each branches of the tree step 1 and step 2 will be repeated for each subset.

2. Related Work

In [1] author follows the approach of supervised learning classifiers to get their prediction of success and failure of the startups. Accuracies of their prediction is 73.3%, 86.3%, 88.1%, 87.5%, 86.7%, 88.4%, 86.4%, 87.9%, 87%, 96.3% for their models. Also, author has studied the value of ROC area and recall. In this paper author give the model to predict the success rate of startups in their early stages also they can make their startups more effective by analyzing their previous mistakes. Author also describes about developing the web tool on this project so that is can be used by entrepreneurs and innovators.

As there are number of factors that can be responsible for success and failure of the startups. Therefore, author is analyzing these important factors which are different phases of project life cycle. As we know there are many phases of project life cycle and any mistake in these phases can result in failure of project. In this study the author is aiming to build an new Gaussian Process Interface (EGPIM). In EGPIM it is using Gaussian process, alongside with Bayesian interface and tested swarm optimization it will be helpful to optimize the hyper parameters which is required to developing the Gaussian prediction. The author trying to gain the expert knowledge from earlier data to check the relations which is important and affecting the results of the project whether it is failure or success.

In [3] author tried to predict the success rate of the startups by analyzing the kickstarter which is the one of the popular platforms for crowd funding which is used by the startups to obtain the amount from Kickstarter for their business idea. As funding plays the major role to start any business therefore if the company do not get the amount, they need for their startups this can be the major factor which plays the important role in success and failure of the startups. Also, they used here various classification and boosting algorithms like weighted random forest with the ada boost used for the sub sample datasets for best possible accuracy. They tested two models for quick execution one is XGboost and another one is CatBoost where the XGboost is quick but lack due to binary substitution but CatBoost is fast without lacking. Author uses random forest with ada boost but on categorial data to increase the accuracy of the model.

In [4] the author is developing the model which can predict the success of vendors like newspaper vendors the approach of the model is to predict the number of copies can be sold on particular day. Through this methodology the vendors can print the limited amount of copies which will decrease the price of their rate of production. Here they use the artificial neural network to predict the number of sales of particular vendor. They build the model which can predict the future trends and the sales of newspapers. It consists of 3 layers which are hidden layer, input layer and output layer. Here the input layer consists of 6 input nodes this input nodes is important factor responsible for the output or final result. Here the author predicts the accuracy based on the historical data of previous years to develop the model with high accuracy.

In [5] the author is predicting the success or failure of the business using the end user computing (EUC). The end user computing is the attracting the interest from information system in business organization and researchers. Here author discuss about how the organization can be the best approach for optimizing the task and their effectiveness for the end user application development and also IT specialist can maximize their contribution. Here they use the modeling approach which they designed to enable the effect of prediction changing the success factor on the effectiveness on the applications developed by the end users. Here the preform their prediction using the 69-business user by getting the information from them. Here the author categorized the business users into different users according to their specialization so that they can predict which business user is best suited for the customer according to their requirements.

In [6] the author is trying to predict the success of consulting business by using the application of machine learning and statistical predictive model. As in today's world the data of the customer can be the good approach to predict the interest of the customer so in 6. the author collected the customer data from the company. Here author visualize and interpret the data to support the decision made by the company as there is always the scope of advancement. There are two main aims of the prediction first is to identify the best approach and method for cost prediction in consulting business using machine learning and Second is to develop the user interface which can visualize the result of these techniques to make this more interactive for the user decision support. They collected the previous twelve-year data form the customer relationship management (CRM). They also used linear regression model, machine learning decision tree, random forest etc to get more accurate result.

In [7] the author is predicting the success rate of small and medium enterprises (SME) as these enterprises plays the crucial role in the economy of every country. Here in [7] the author is focusing in the key factors like role of innovation which is the one of the most important factors in success of business and it is measured by the patent application. Here the author applied the random forest model to sample the company and to find the attribute necessary for the prediction and success of business. In the result of [7] the show that the patent which is open source or publicly available has more accurate prediction power for predicting the business performance of SME.

In [8] the author is trying to improve the success rate of the organization by using Click- through rate (CTR) approach as to make the business successful the organization must know the interest of their customers therefore by analyzing the customers number of clicks on different social platform they predict the behavior and interest of their customers. Here author proposed the new model called R-RNN stands for (Recent Recurrent Neural Network) it can learn the behavior and user interest through his/her overall historical click. The R-RNN not only give the users interests on the basis of historical clicks but it also has a LSTM (long short-term memory) unit for checking the new trends and the basis of the user's recent clicks it will give us the new interest of the user.

In [8] they found that the R-RNN is performs better as compare to existing deep earning model and also the click through rate also perform the major role to get accuracy in their model.

In [9] author is focusing on one of the main factors for business success prediction is crowdfunding the business can be only success if they have good projects and innovative ideas. But to work on project or any idea we need funds or money this can be achieved through crowdfunding. It is done by many companies but empirical analysis shows that only 1/3rd of the crowdfunding is successful and meet their requirement therefore in [9] author is focusing on the prediction of crowdfunding whether the crowdfunding held by an organization is going to be successful or not. They collected the data from Kaggle and historical records of kick-starter campaign. Here author proposed the MLP model which can provide the accurate result and when apply to different platforms of crowdfunding which has been never used in earlier crowdfunding model also, they used many classification algorithms to get the best result with the high accuracy.

In [10] the author is providing the model for business success prediction for shareholders, banks, investors, suppliers etc. all those who get affected by business success and failure. There is already the lots of model build before to predict this and the statistical procedure like multiple discriminant analysis, logit and profit etc are mainly used techniques used in such problems. But the statistical procedure methods needed data to be formatted in specific distribution. Also, the problem of autocorrelation, multi-collinearity, heteroscedasticity and distribution involved can leads to the problem with some statistical methods. Due to these problems author used the better methods here for example multi-criteria methods or machine learning methods (decision tree, neural network).

3.Dataset Description

As the dataset plays the most vital role in this project here, I have downloaded this dataset from Kaggle. This is US based dataset in this dataset I have 1154 rows and 49 columns. In this dataset I have data about various companies and how the various attributes effect the growth of the companies from the date the company begin and become successful or get closed. Here I have various attributes like Total funding, Company Milestones, Company location etc.

In this project, I have used python language for performing the analyzation.

In python for performing data analysis we have to know about the Pandas Data Frame. In pandas, Data Structures are of two different types: -

1. Data Frame of Pandas
2. Pandas Series

Data Frame: - Pandas Data Frame is a 2-D or 2-Dimensional labelled Data-Structure having columns that are potentially different. It is an in-memory representation of an Excel sheet through a programming language python. So, Pandas Data Frames are similar to the Excel Sheets.

As Excel provides much functionality, similarly Pandas Data Frames also provides much different functionality for analyzation, changing, and extraction of the data from the provided datasets.

Pandas Data Frame are created by loading the datasets from some storage devices, not only from Excel sheets but also from MySQL database as well as from CSV file also.

For creating pandas Data Frame it should be imported from libraries which is Pandas For example: -

df= pd.DataFrame(sample)

Here sample should be a part of python dictionary which should be in the form of keys and values.

For viewing the Data Frame we can simply write df or df.head() or df.tail() command. df will return the complete rows and columns in the given dataframe. df.head() and df.tail() will return only the top and bottom data respectively.

For example:-

df.head(5):-

It will return only top 5 rows of the given dataset.

DATASET

```
In [1]: import pandas as pd
        import plotly as py
        import plotly.graph_objs as go
        import pandasql as ps
        import numpy as np
        import seaborn as sn
        import matplotlib.pyplot as plt

        py.offline.init_notebook_mode(connected=True)
```

```
In [2]: path="E:\\New folder\\Startup-Success-Prediction-master\\d3\\data\\train-data.csv"
        df=pd.read_csv('E:\\New folder\\Startup-Success-Prediction-master\\d3\\data\\train-data.csv')
        df
```

	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	Unnamed: 6	name	status	...	is_othercategory	object_id	has_VC	has_a
0	0	CA	37.392480	-122.072612	94041	c:1001	Mountain View	NaN	FriendFeed	acquired	...	0	c:1001	0	
1	1	MA	42.368633	-71.075305	2210	c:10054	Boston	NaN	Jumptap	acquired	...	0	c:10054	1	
2	2	NY	40.745064	-73.992637	10001	c:101312	New York	NaN	SideTour	acquired	...	0	c:101312	0	
3	3	NY	40.775309	-73.983656	10001	c:10137	New York	NaN	Producteev	acquired	...	0	c:10137	0	
4	4	MO	38.703764	-90.443832	63043	c:10153	Saint Louis	NaN	ITOG, Inc.	closed	...	1	c:10153	0	
5	5	TN	36.139960	-86.796377	37212	c:10158	Nashville	NaN	StudioNow	acquired	...	1	c:10158	0	
6	6	CA	37.776246	-122.417922	94103	c:10176	San Francisco	NaN	Yammer	acquired	...	0	c:10176	0	
7	7	CA	37.798318	-122.400003	94111	c:10179	San Francisco	NaN	GoodGuide	acquired	...	0	c:10179	0	
8	8	CA	37.390501	-122.081151	94041	c:10197	Mountain	NaN	PostPath	acquired	...	0	c:10197	0	

	Unnamed: 6	name	status	...	is_othercategory	object_id	has_VC	has_angel	has_roundA	has_roundB	has_roundC	has_roundD	avg_participants	is_top500
	NaN	FriendFeed	acquired	...	0	c:1001	0	0	1	0	0	0	3.0000	1
	NaN	Jumptap	acquired	...	0	c:10054	1	0	1	1	1	1	4.5000	1
	NaN	SideTour	acquired	...	0	c:101312	0	1	1	0	0	0	2.0000	1
	NaN	Producteev	acquired	...	0	c:10137	0	1	0	0	0	0	6.0000	1
	NaN	ITOG, Inc.	closed	...	1	c:10153	0	1	0	0	0	0	1.0000	0
	NaN	StudioNow	acquired	...	1	c:10158	0	0	1	0	0	0	1.0000	0
	NaN	Yammer	acquired	...	0	c:10176	0	0	1	1	1	1	5.2000	1
	NaN	GoodGuide	acquired	...	0	c:10179	0	0	1	1	0	0	3.0000	1
	NaN	PostPath	acquired	...	0	c:10197	0	0	0	1	1	0	2.5000	1

Fig.3.1: Dataset Representation

4. SCHEDULE, TASKS AND MILESTONES

First here I will use the supervised machine learning approach using this I will train the model on the basis of the historical data obtained by the companies. I will be focusing on the historical data of the companies which is either overtaken or shutdown from this data I will find the main factors responsible for failure of company. Also I will plot the map of the different states of country where I will found out the states having maximum numbers of business success or failure. This map visualization will be helpful for prediction the most preferred state or location for setting up the industry.

4.1. TASK ANALYSIS

(I) Understand The Relationship Between Different Labels and Features:-

Here in this task the most correlated features were selected. After this the top correlated features will be separated into different ranges for example the funding and location is highly correlated when the startup is successful therefore, we can consider it as an important factor for our prediction.

(II) Discover Best Model for Predicting Success:-

Here in this step I will predict best possible machine learning model by comparing different models.

(III) Predict Startup Success or Failure on New Data and Visualize:-

The data we get after pre-processing and data cleaning using that and our machine learning model like random forest, we will predict the success and failure of business also that prediction will be grouped by state and visualize on the map of the country.

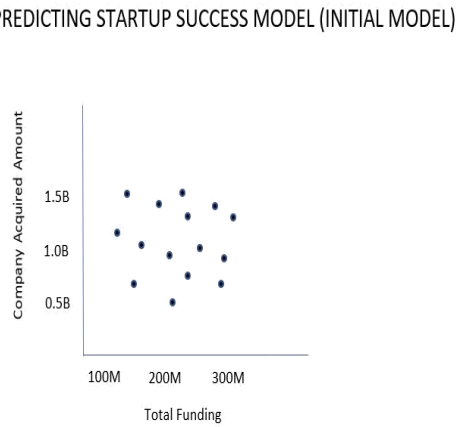


Fig.4.1: Startup success model

It is just an initial model to explain how I am going to handle my dataset and predict the best possible attribute for my final model so that I can achieve the highest accuracy possible. In Fig-4.1.1 its shows the scatter plot between the company acquired amount and total funding if the company is not acquired the acquired amount will be zero. In this scatter plot we will visualize the scatter plot on acquired amount and funding which will help us to find the correlation between the two.

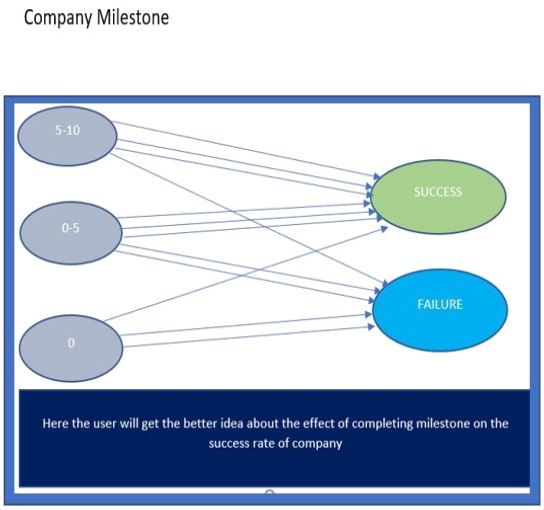


Fig.4.2: Initial Company Milestone

In figure 4.1.2 it shows the number of milestones that company has to complete for their success and failure also it will give user the better idea of the milestones that should be completed by the company for their success. There are different milestones here:-

- Companies with 0 milestones.
- Companies with 0-5 milestones.
- Companies with 5-10 milestones.

The nodes in right side represent the success and failure rate of the company with each milestone.

5. DESIGN APPROACH AND DETAILS

5.1 PERPOSED METHODOLOGY

Here in this prediction model we are using the various methodologies of data mining and machine learning models like decision tree and random forest techniques. Here I have used America based dataset which I got form Kaggle.com by using this data set and machine learning techniques I will plot the plot the prediction result in us based svg file to make it look easier and more interactive.

5.1.1. STEP WISE APPROACH

- Data Collection: - Here I have collected the data form Kaggle.com this data is based on US based startups which include the success and failures of the specific companies it includes around 4000 rows before cleaning.
- Data Pre-processing: - The data preprocessing is the process of filtering the unnecessary data from dataset this step is very important as these uncleaned data can manipulate the actual result and we will get wrong result as output the example of uncleaned data are null values etc.
- Information Filtering: -After completing of the data pre-processing now filtration of the information is to be done which is going on the basis of the requirements of the user. There are many methods are presents in the data mining through which information can be filtered or classified. Some examples like Correlation, Logistic- Regression, Random Forest and many more.
- Visualizations: - Here we also have

plotted the various visualization through which we can understand that which rows and columns are going to play the major roles in our final output this I have don't in python IDE called Jupiter notebook.

• Graphical Presentation: - After all these steps I am finally going to plot the final output on the svg file on my local developer server hosted by python flask here this will make it looks more interactive and easier to understand at the same time.

FLOW CHART

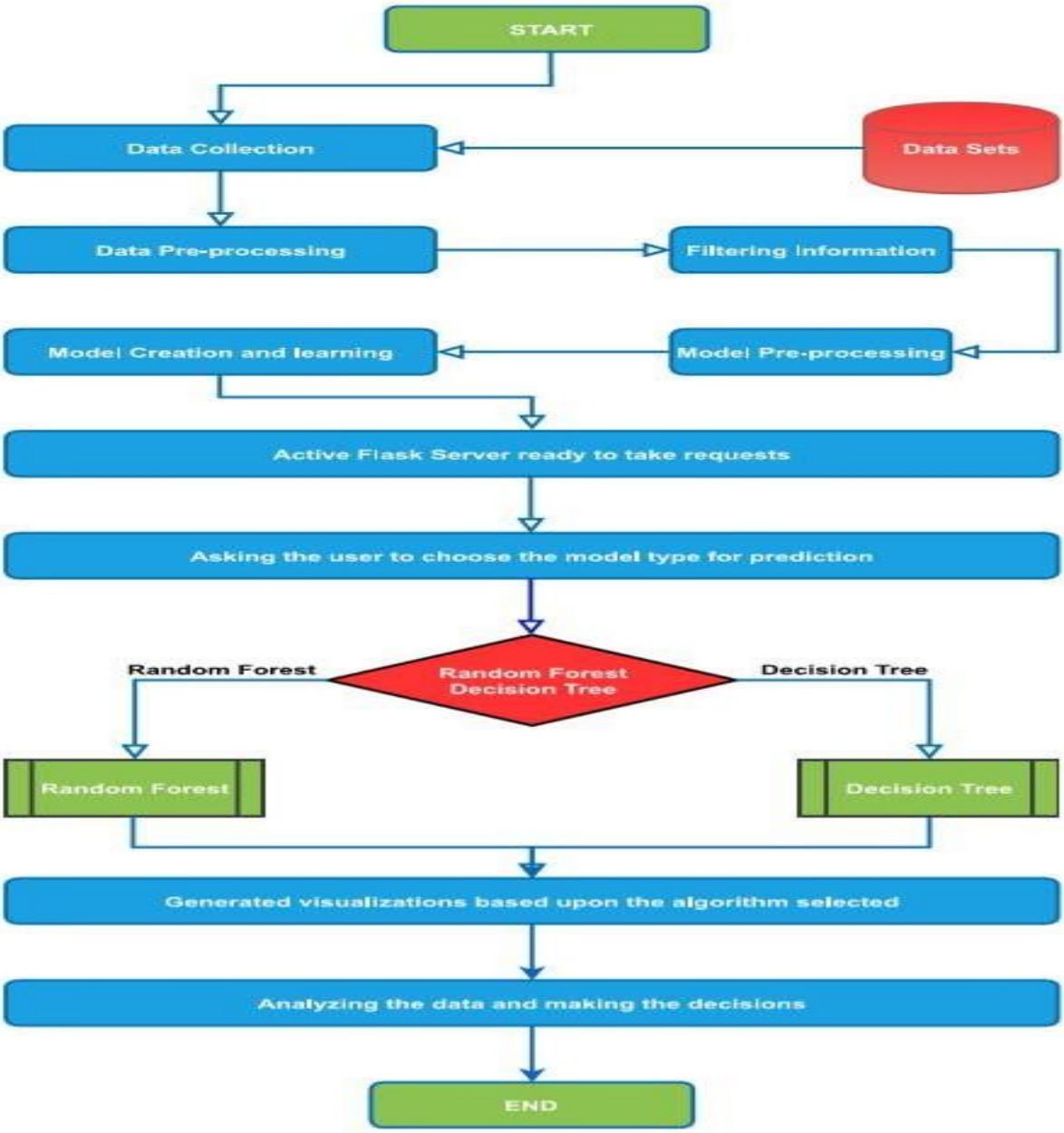


Fig.5.1. Detailed Flow Diagram

5.Visualizations

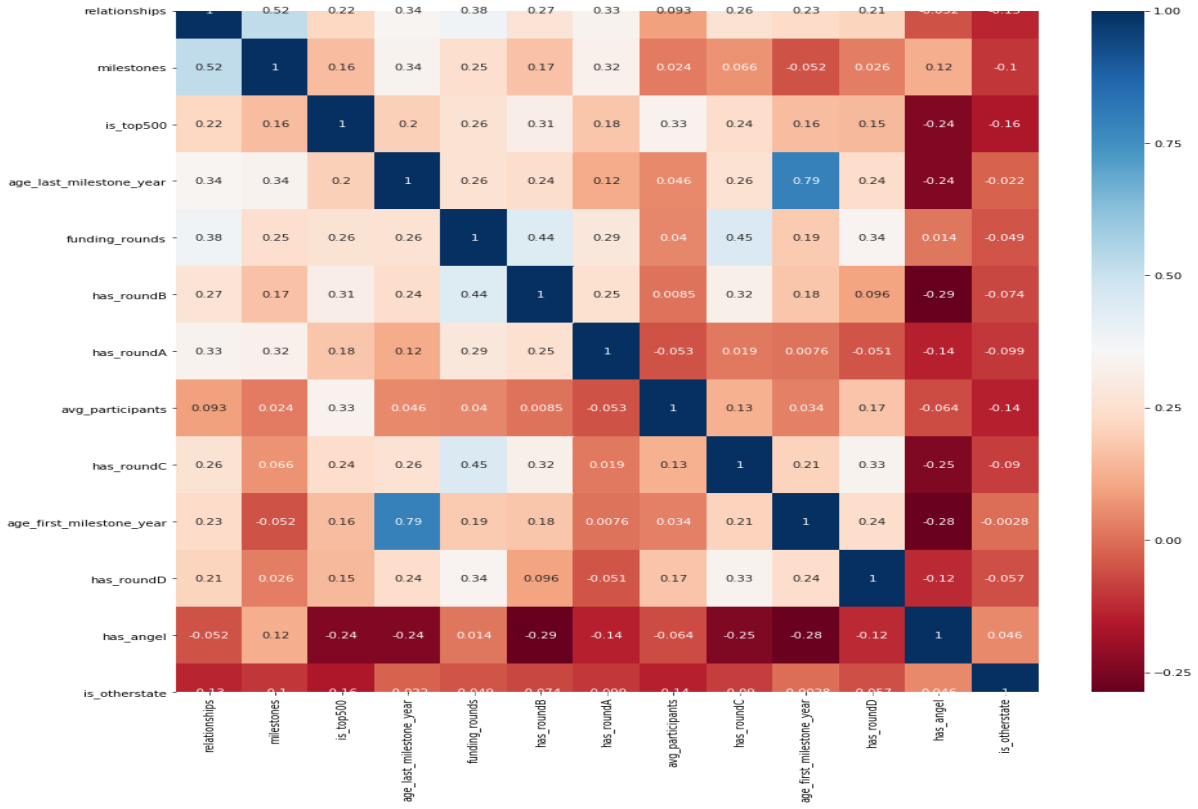


Fig.5.2. Correlation between rows and columns

Images of correlation statistics that helps in visualizing the data in correlation matrices are called Correlograms. We can gain intuitions about the high correlations between some variables through visualizing the correlations between features. There is no significant loss of information in the model if we use these insights to drop a few high correlated features in the variable section.

It would be interesting using a diverging color palette as data from both the positive and negative correlation. Positive correlation is represented by blue whereas negative correlation is represented by red. In the below Correlogram, there isn't very high correlation between the features individually. As we can see most of the features are unique and can be used for training the model.

Link Between Number of Relationships and Company success/ failure:-

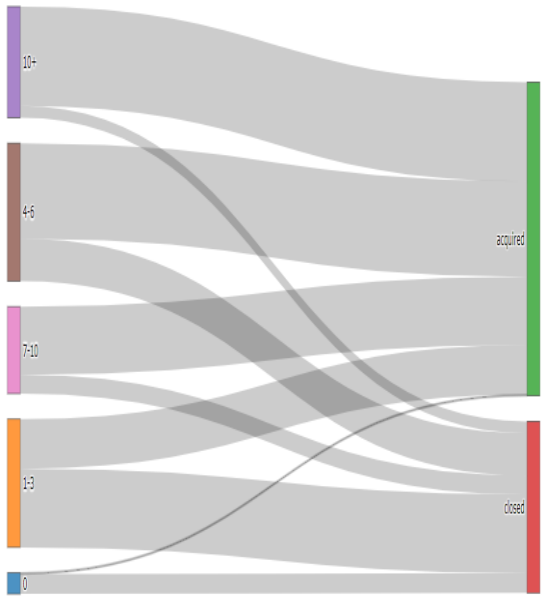


Fig.5.3. Sankey diagram showing correlation b/w success and failure

There is a Sankey diagram above showing the correlation between the companies relating to success and failure. The end node represents the number of relationships in a company during its operation which has been divided into 5 groups. Grouping is denoted by the color of node. The sum of incoming and outgoing link values connected is the value of node. They are represented by the height, with longer bars denoting more incoming or outgoing values and width is directly proportional to flow quantity. The destination nodes tell whether the company is successful represented by green or red color respectively. These visualizations have been built in Plotly. The user can run the actual code through Jupyter notebook and hover the mouse over the nodes to determine the strength and count of incoming or outgoing link values.

Link Between Number of Milestone and Company Success Failure:-

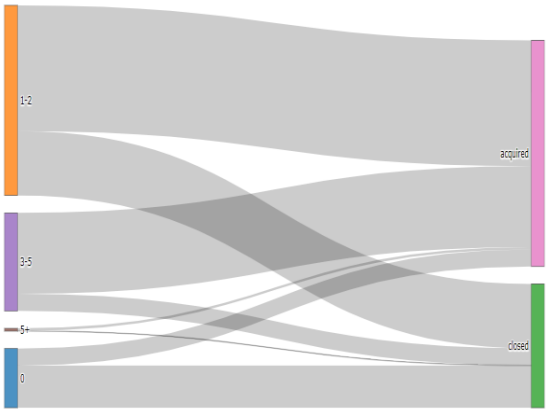


Fig.5.4. Sankey Diagram showing correlation between company milestones

Above Sankey diagram is showing the correlation between companies milestone to its success and failure. The designs were made similar to the previous one. We can see companies with more milestones are more likely to succeed as compared with fewer milestones.

Scatter Plot Between Id and Funding Total USD:-

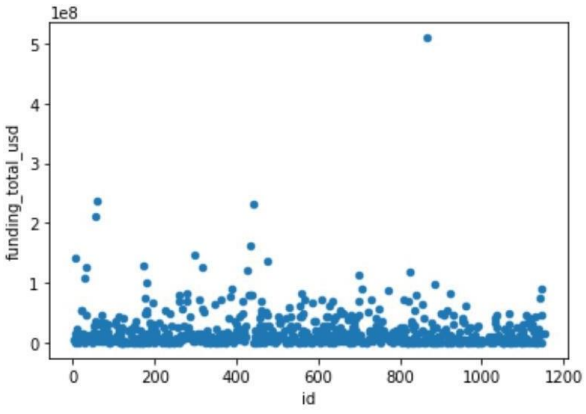


Fig.5.5. Scatter plot between ID and Funding's

6.FINAL RESULTS

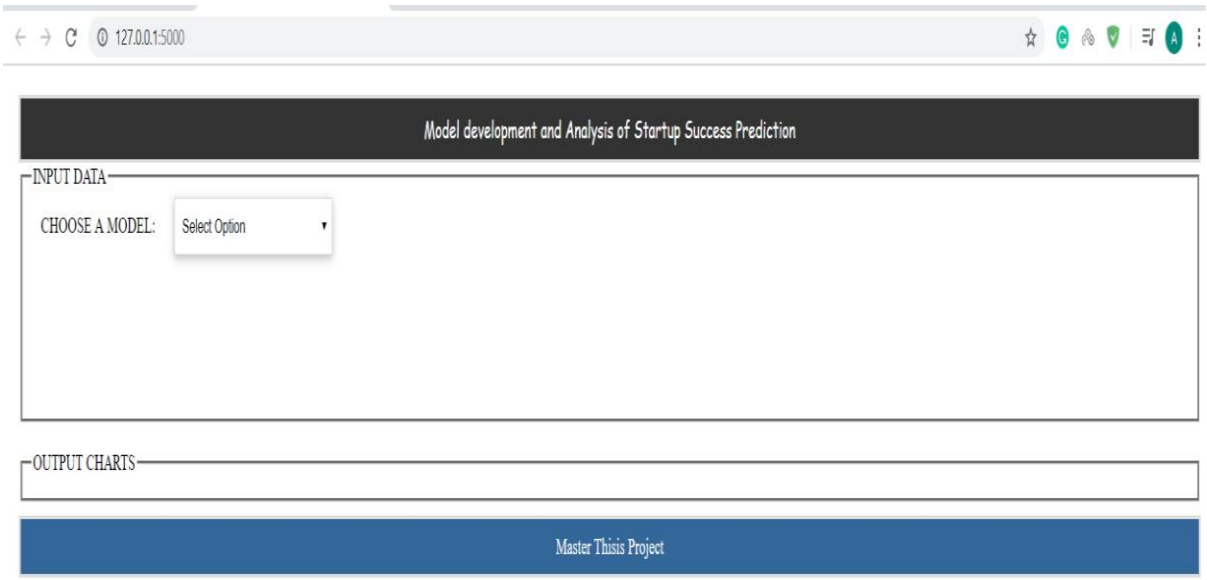
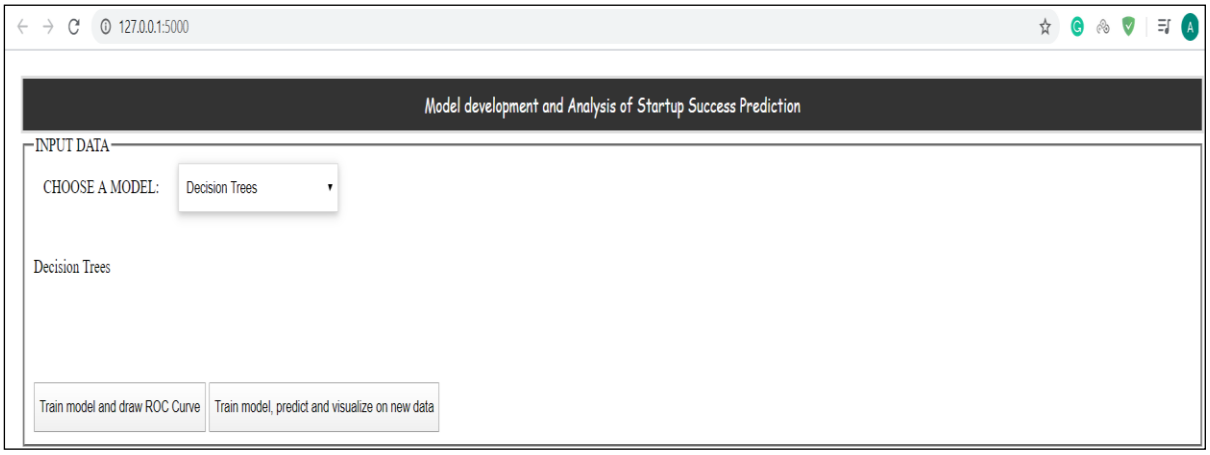


Fig.6.1. Front end diagram

Here you can see the model development and analysis of this project. This is running on the developer server provided by the flask IDE in python as you can see above link 127.0.0.1.5000. Here you can see the two labels input data and output chart. In input data you can choose the models and among the decision tree and random forest after selecting one of the model you will see the two button to draw ROC curve and train model predict and visualize on new data. After click on the button you will see the Roc curve or visualization on map in the output section.

DECISION TREE MODEL VISUALITION



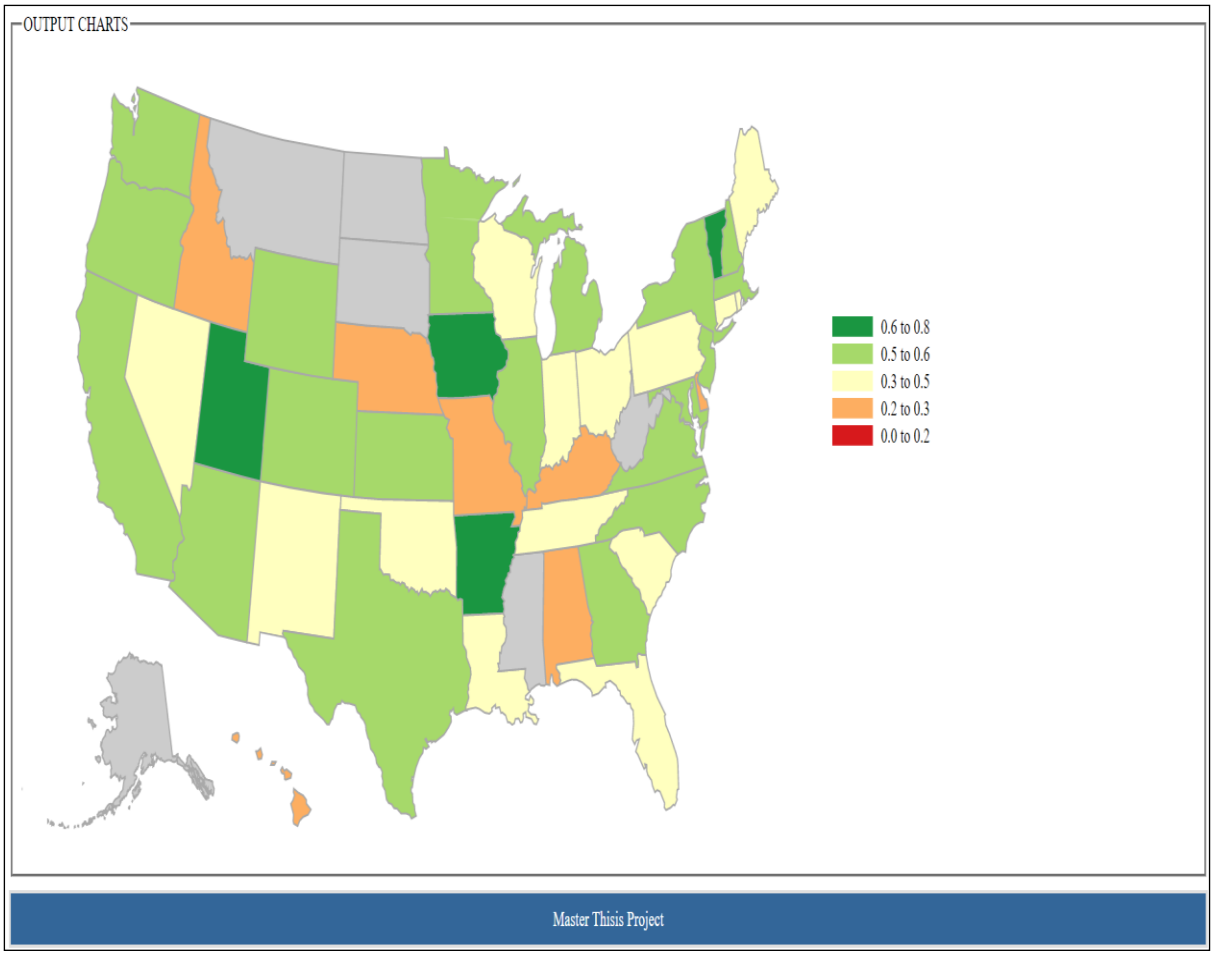


Fig.6.2. Decision Tree Model Visualization

Now here in above visualization 5.3.2 I have given the max depth value equal to 5 and max leaf node equal to 10 too now as per using the decision tree classifier in my flask It is returning me the output in the form of the success and failure probability. So here this visualization you can se above is my svg image which is my json file this svg is also an image but is build by connecting various points and by using those points I am displaying the map format. Here I am visualizing my result in the 5 different colors where the green color represent the higher rate of success and the yellow color display the almost 50 to 60 percent chances of success and the red color shows the less probability of business success.

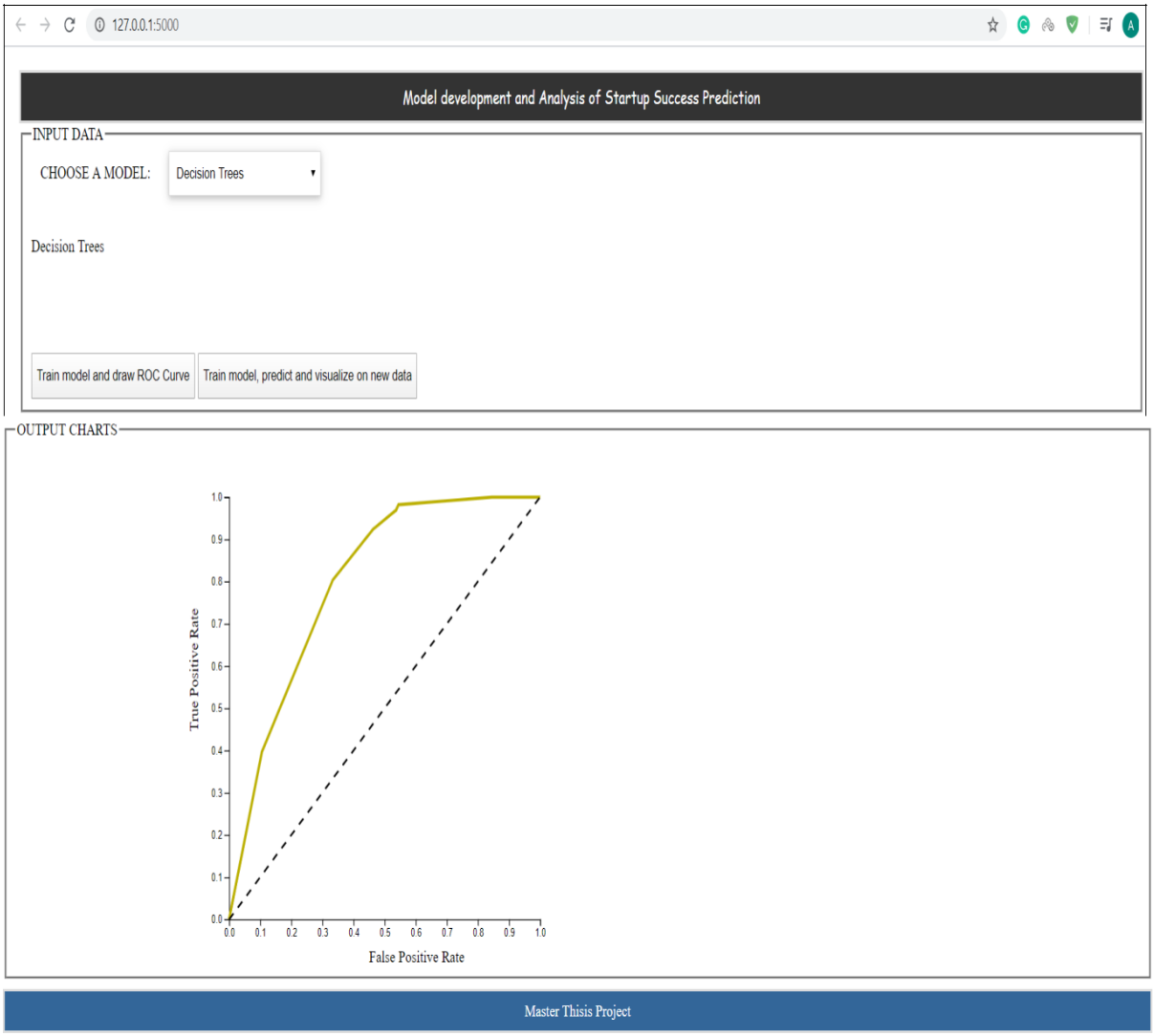


Fig.6.3.ROC Curve for Decision Tree

In visualization 5.3.3 we can see that the AUC and TRP (True Positive Rate) is high as our curve almost tends to 1 and the area under the curve us high this means that our decision tree model is predicting the result with higher accuracy.

About ROC Curve

In ML, the performance of the measurement is a necessary task. That's why when a classification problem appears AUC-ROC Curve can be used. When it is required to check or visualize the multi-classification performance, AUC which is Area under the Curve ROC which is Receiver Operating Characteristics Curve. Which is the most essential metrics evaluations which is required to check the performance of any classification model. It can also be written as AUROC or Area under the Receiver Operating Characteristics.

AUC-ROC Curve is used for measurement of performance which is used to solve the classification problems like various thresholds settings. ROC is actually a probability curve and AUC show degree or the measure of separability. It tells the capability of the model for differentiating between the classes. AUC is higher, model will be better for prediction at 0s as 0s and 1s as 1s.

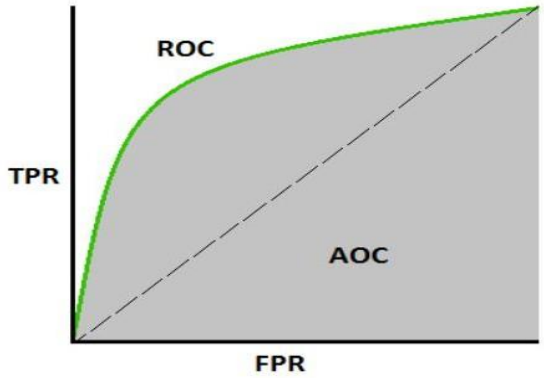


Fig.6.4. Basic Structure of ROC Curve

This curve is plotted with TRP and FRP where TRP against FPR is on X-axis and TRP is on y-axis.

Terms in RUC Curve

• TPR (True Positive Rate)/ Recall/ Sensitivity: -

TPR/ Recall/ Sensitivity = $\frac{TP}{TP+FN}$

• Specificity: -

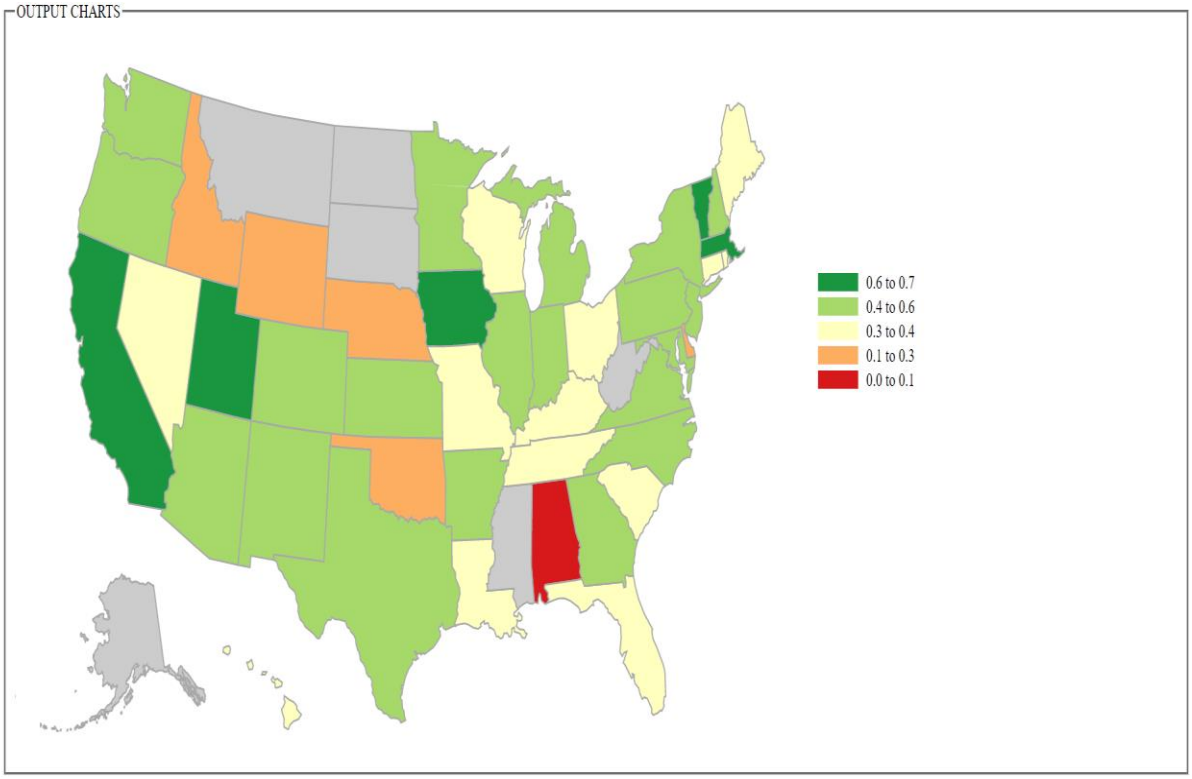
Specificity = $\frac{TN}{TN+FP}$

• FPR: -

FPR=1-Specificity

= $\frac{FP}{TN+FP}$

RANDOM FOREST MODEL VISUALITION



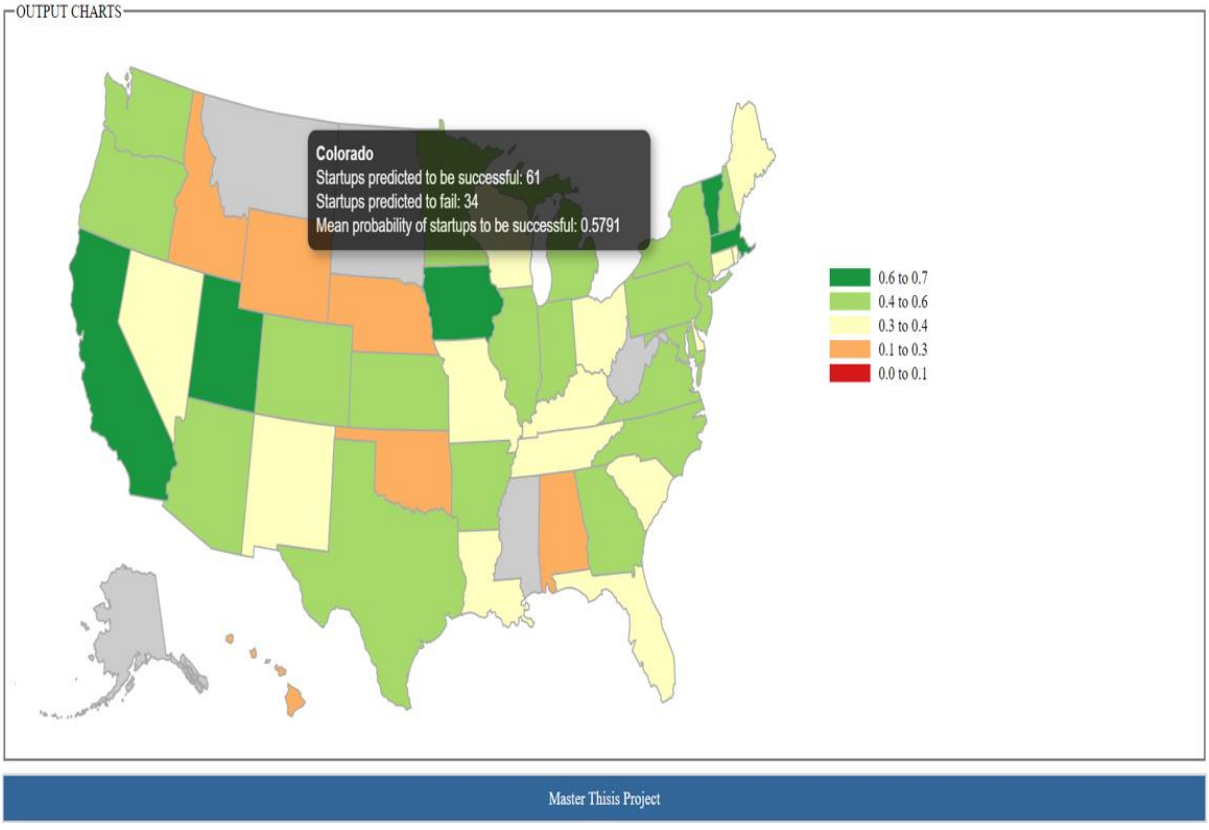


Fig.6.5. Random Forest Model Visualization

Now here in above visualization 5.3.5 I have given the max depth value equal to 5 and max leaf node equal to 10 also here I have number of estimator here as I am using random forest here in my flask It is returning me the output in the form of the success and failure probability. So here this visualization you can see above is my svg image which is my json file this svg is also an image but is built by connecting various points and by using those points I am displaying the map format. Here I am visualizing my result in the 5 different colors where the green color represents the higher rate of success and the yellow color display the almost 50 to 60 percent chances of success and the red color shows the less probability of business success. Also as you can see in above figure we can also hover over any state to get data and prediction displayed about it as above we can see Colorado state which predicts the startups successful there is 64 and failed startups is 34 which leads to our final prediction of startups to be successful is 0.5759 like this we can hover over any state and know about the what is this prediction of success in that particular state.

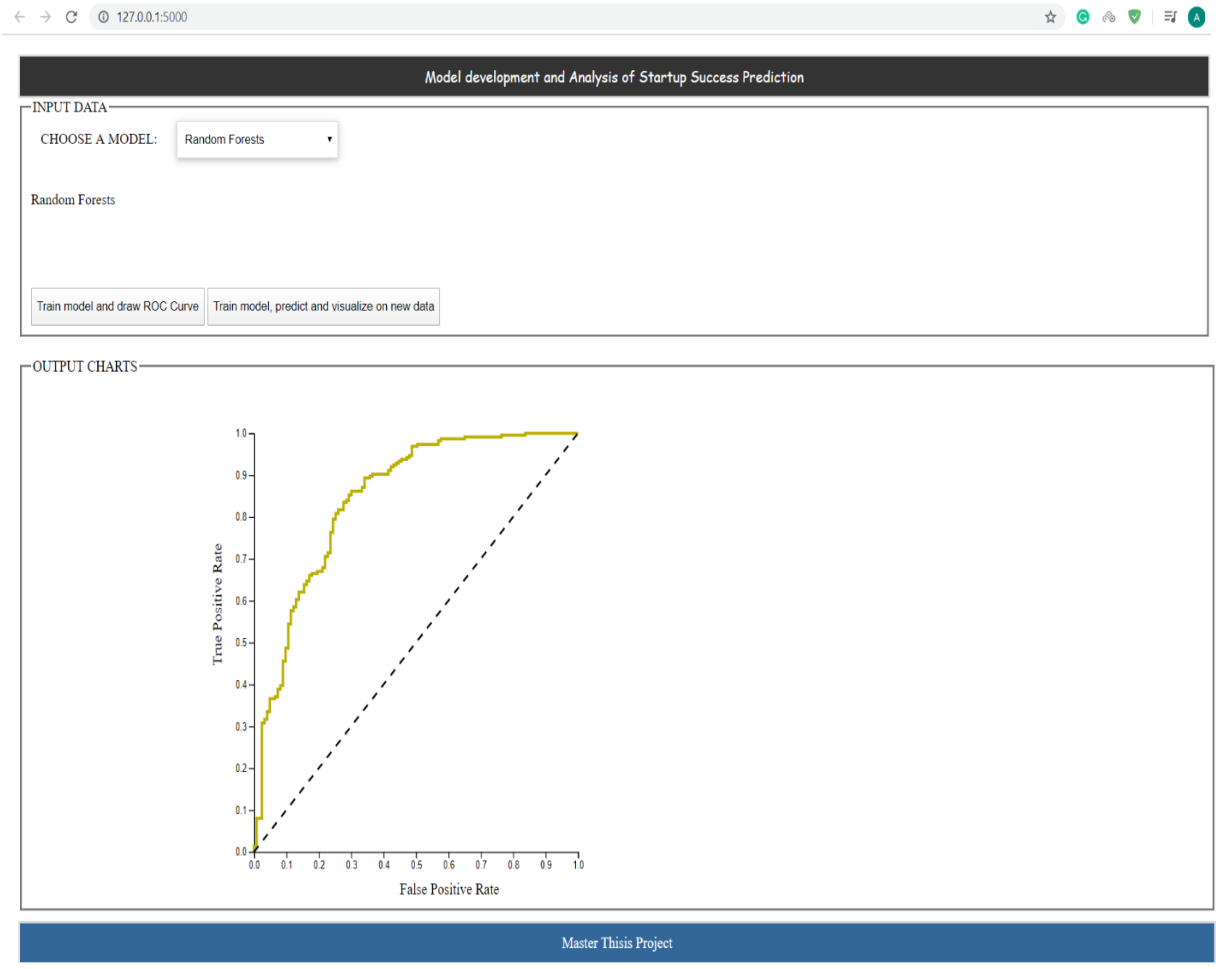


Fig.6.6. Random Forest ROC Curve

In visualization 5.3.6 we can see the ROC curve drawn for random forest visualization as we can see here the area under the curve is also high this means the accuracy is also high in random forest model so we can go with the visualization we get from random forest model.

6. RESULT AND CONCLUSION

Finally we have a model which is able to predict the success and failure of the company on the basis of our machine learning model decision tree and random forest as we seen in our final visualizations it show the diverging color scale to show the success and failure of the startups on the basis of states. Where green color highlights the higher success probability and red color predicts the more chance of failure. Also there are states like South Dakota and Alaska where no startups has been found till now so we can predict the probability in such states therefore the state has been displayed in grey color. Also our user can get additional information like number of startups predicted, number of startups failed in particular state and the mean probability of startup success this provide user a more interactive way to get his/her output by just hover on particular state and we can also draw the ROC curve for each model through which we can see the accuracy of our model. Therefore, we have a model which provide the result to the user in very interactive manner and with high accuracy.

FUTURE SCOPE

As per now our prediction model is complete and we are getting the desired output from our model we are able to predict the success and failure of the company and display our result through graphical representation and ROC curve. But as we know there is always a scope of improvement therefore in this model we also have some scope of improvement too we can improve our model by adding various features in it like:-

• As per know our model is predicting the success and failure of the project on the basis of states in future we can predict the company success in particular city in the state through the our user will get the best possible result on specific area.

• This model is currently running on the python developer server we can implement this project to the production server where everyone can use it.

References

[1] Ankit Agrawal , Alok Choudhary "Predicting the Outcome of Startups: Less Failure, More Success", In proceeding of IEEE International Conference on Science and Information, pp 0,2016.

[2] Min-Yuan CHENG, Chin-Chi HUANG, Andreas Franskie Van ROY "Predicting Project Success in Construction using an Evolutionary Gaussian Process Inference Model" The National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C. 2013.

[3] Siddharth Jhaveri , Ishan Khedkar , Yash Kantharia , Shree Jaswal "Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns", In proceeding of IEEE International Conference on Science and Information, pp 0-10,2019.

[4] Abdul Sahli Fakharudin , Mohd Azwan Mohamad, Mohd Usaid Johan "Newspaper Vendor Sales Prediction Using Artificial Neural Networks" , In proceeding of IEEE International Conference on Science and Information, pp 0-10,2009.

[5]D.R.Lawrence, H.U.Shah, P.Golder, "End user computing:how organisation can maximise potential" , In proceeding of IEEE International Conference on Science and Information, 1997.

[6]AmyCook, PaulWu, KerrieMengersen "Machine Learning and Visual Analytics for Consulting Business Decision Support", In proceeding of IEEE International Conference on Science and Information, ,2015.

[7] Daniel Müller, Yiea-Funk Te, Pratiksha Jain "Predicting Business Performance through Patent Applications", In proceeding of IEEE International Conference on Science and Information, 2017.

[8] Mingxin Gan , Kejun Xiao "R-RNN: "Extracting User Recent Behavior Sequence for Click-Through Rate Prediction", In proceeding of IEEE International Conference on Science and Information,volume:7, 2019.

[9] Hong Zhou ; Luzhuang Wang ; Qinqin Xia ; Shuting Zhan "Business Failure Prediction: A Review and Analysis of the Literature", In proceeding of IEEE International Conference on Science and Information, 2010.

[10] M. Daubie, N. Meskens "Business Success Prediction Model for Startups and Stackholders", In proceeding of IEEE International Conference on Science and Information,2014.