****

*A CAL Project Report*

on

**Startup Investment Decision Support**

*to be submitted in partial fulfilling of the requirements for the course on*

**Data Mining and Business Intelligence – ITA5007**

**(B1+TB1 / B2+TB2)**

by

**Akash Patel (21MCA0138)**

**Titas Ganguly (21MCA0152)**

**Subham Karmakar (21MCA0172)**

Under the Guidance of

**DR PRABADEVI B**

**TABLE OF CONTENTS**

**ABSTRACT**

1. Introduction …………………………………………………………….
2. Review 1 (Survey, Analysis)...………………………………….………
   1. Problem definition
   2. Dataset Description
   3. Review on Existing System
3. Review 2 (Design)…………………….…………………..
   1. Methodology
      1. Module Description
         1. Data exploration
         2. Pre-processing
         3. …….. (include other modules like that)
      2. Algorithms used
         1. Justification for choosing the models
      3. Flow diagram of your model
      4. Dataset after preprocessing
      5. Dataset split(train and test)
4. Review 3 (Code)………………………………….
   1. Implementation
      1. Software and hardware description
      2. Output screenshots
   2. Confusion Matrix
   3. Comparison of the models used
   4. Comparison graph
5. Conclusion ………………………………………………………....……
6. References ……………………………………………………………….

**ABSTRACT:-**

This research aims to explore which kinds of metrics are more valuable in making investment decisions for a venture capital firm using machine learning methods. We measure the fit of developed companies to a venture capital firm’s investment thesis with a balanced scorecard based on quantitative and qualitative characteristics of the companies. We explore the most influential factors of their balanced scorecard using their retrospective investment decisions of successful and failed startup companies.

Our study employs models and their counterparts with an additional feature selection technique. Our findings suggest that “planning strategy” and “team management” are the two most determinant factors in the firm’s investment decisions, implying that qualitative factors could be more important to startup evaluation. Furthermore, we analyzed which machine learning models were most accurate in predicting the firm’s investment decisions. Our experimental results demonstrate that the best achieve an overall accuracy of 87% in making the correct investment decisions, with an average of 87% and 77% in predicting the decision of companies the firm would and would not have invested in, respectively.

Our study provides convincing evidence that qualitative criteria could be more influential in investment decisions and machine learning models can be adapted to help provide which values may be more important to consider for a venture capital firm.

1. **INTRODUCTION:-**

Startup refers to a company that is in their first stages of operations. Startups can be founded by one or more entrepreneurs who are working on developing a product or service. Startups normally have very high cost and limited revenue. As they require a lot of capital to take it off the ground, they look for capital from a lot of sources like venture capitalists.

Startups go through multiple rounds of funding to raise capital. The different funding rounds that let outside investors the opportunity to invest cash in exchange of equity or partial ownership of the company. Other types of investments are debt, convertible note, stock or dividends. Startups can start off with “seed” funding or angel investor funding at the beginning. The next funding rounds can be followed by Series A, B, C and so on. Goal of most startups is to get acquired by a different company or become a publicly traded company.

90% of startups fail due to bad product market fit, marketing problems, team problems or other issues. They also fail within the first few years. This makes startup investment very risky. Historically only venture capitalists could invest in startups but due to the recent trend in crowdfunding sites, an average investor can easily grab a piece of an exciting startup.

1. **REVIEW-1** **(Survey, Analysis)**
2. Problem definition

Startup investment can be very risky due to the high failure rate of startups. People like angel investors and venture capitalists have a very high risk while they are investing in startups. To assist startup investors with their decisions, in this project we aim to find the important features that lead to startup success and forecast a company’s success with supervised machine learning methods.

1. Dataset Description

To train the machine learning model, we used investment data about startup companies available on [Kaggle.](https://www.kaggle.com/arindam235/startup-investments-crunchbase) The data has been collected from Crunchbase which is a leading website for company insights from early stage startups to Fortune 1000.

The data had around 54k rows and 39 columns. The dataset had company information such as name of the company, url, market, country, state, region, city, founded date, first funding date, last funding date. It also had data on different investment types such as seed, venture equity crowdfunding, undisclosed funding, convertible note, debt financing, angel, grant, private equity, post ipo equity, post ipo debt, secondary market, product crowdfunding, round A-H series funding. Detailed descriptions of the different funding types is available [here](https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types). Status of the companies were also available and segmented by acquired, operating and closed.

1. Review on Existing System

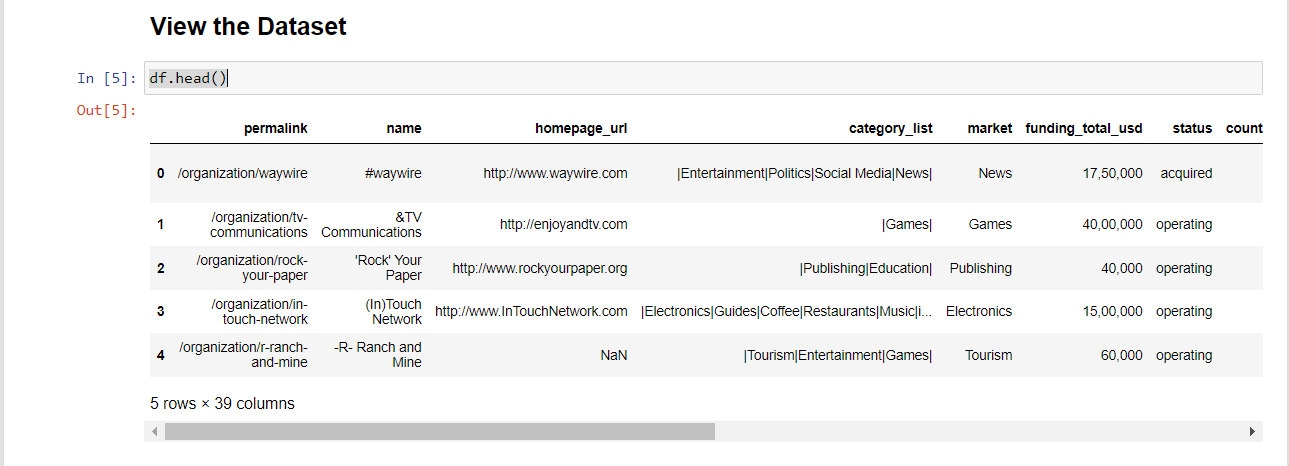
* **Every investor wants to find the hidden unicorn in a sea of potential investments.**Identifying founders, especially early in their career, with unicorn potential is extremely difficult. Though, there are many examples of attempts to predict the success of a company. I was able to find more than 20 in a quick google search. Some of these are tricks-of-the-trade from investors giving their perspective on what matters the most. Others are machine learning engineers searching for predictive insights from big data.
* **Unfortunately, these methods ultimately fall short.** There are various reasons these methods are less than ideal. The insights investors provide are valuable perspectives into what they consider important, but they are impossible to replicate. Their internal scale for measuring the “focus” of a founder isn’t something that can be calibrated (very easily) in another investors process. While I’m a firm believer that humans are excellent pattern recognizers, we can also only hold onto so much information. These investors’ “rules” may only capture a part of a much larger picture and are likely biased in some non-obvious way. Meaning the rules may not apply to all founders. I’m also a firm believer that data and machine learning/data science/ artificial intelligence/pattern recognition techniques contain the ability to predict startup success (with some degree of accuracy). However, the methods I’ve seen typically use publicly available data, making its own set of assumptions. I don’t make any references to these works because I don’t want to diminish the work they’re doing. I think it’s valuable, but it’s a problem with a lot of uncertainty surrounding it. In the rest of this post, I will focus on the data-driven approach to predicting startup success. I will give a high-level overview of the data, standard AI practices, and what we did to deliver a solution developed specifically for this purpose.

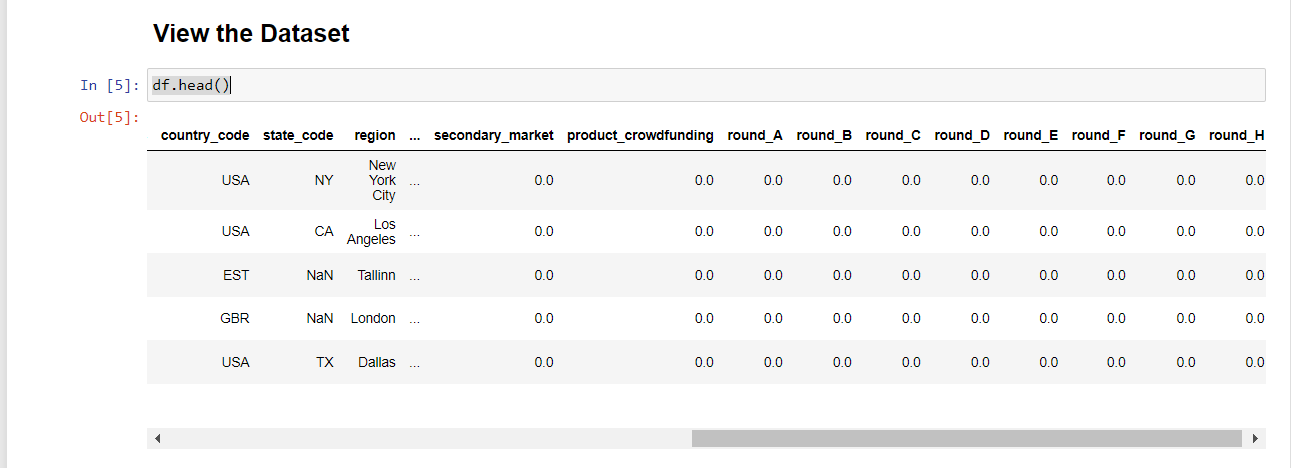
1. **REVIEW-2** **(Design)**
   1. Methodology

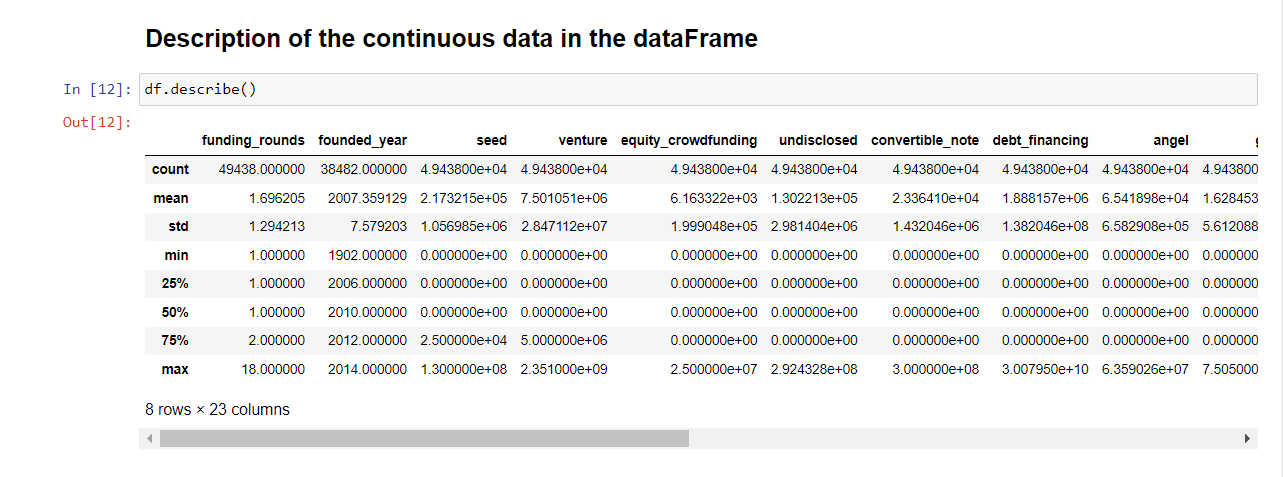
Before we could use the data to train the different models, we had to clean the data and select the most important columns to be included into the model. One of the biggest problems we had with the dataset was that it had a lot of zeros and a lot of columns to choose from.

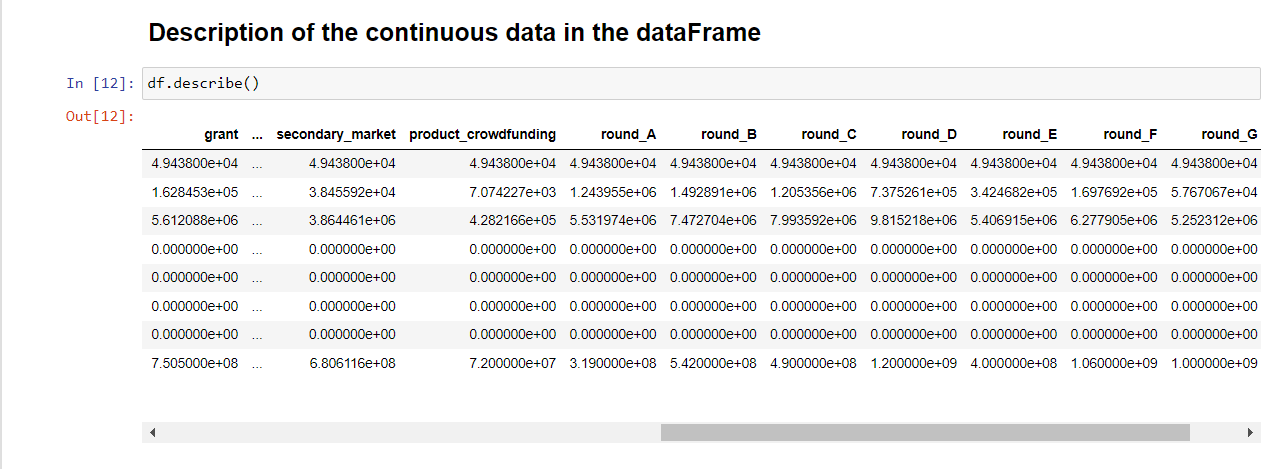
We also realized later that the status column had around 80% of the companies as operating status and the rest as closed and acquired companies.

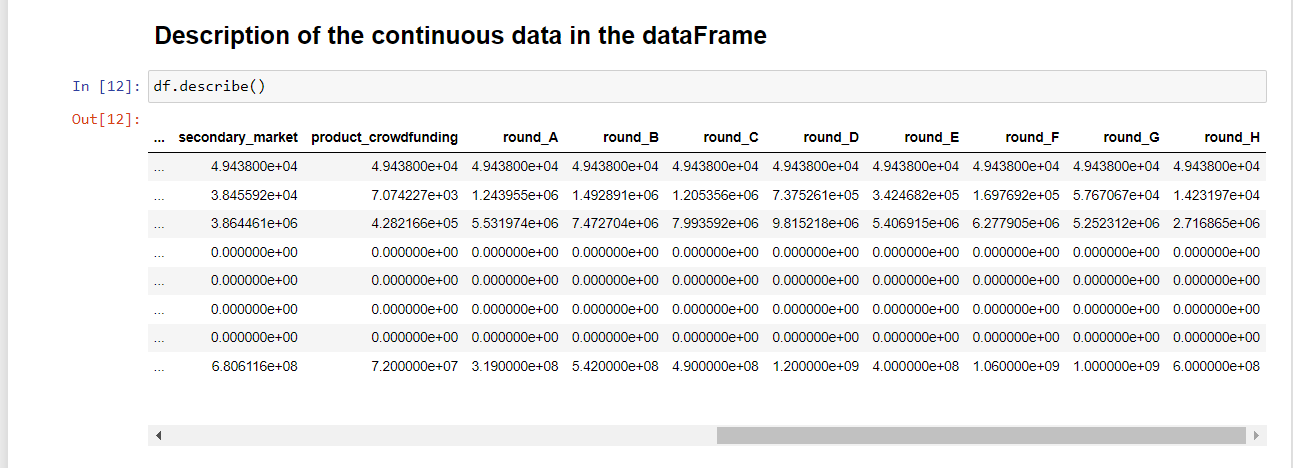
* + 1. Module Description
* The analysis of the dataset was important to understand what is important to use in the final model. It also helped us understand the data more before using them in the models. The analysis below contains some of the important EDA we did on the data.
* Most of the companies were in the Software and Biotechnology industry. Biotechnology had the highest number in total funding. Mobile companies had the second lowest number in total funding.
* A lot of the companies raised venture and seed funding. The number of companies decreased as the companies proceeded to more series funding. Round G and H have a very low number of companies compared to round A and round B.
* Most of the acquired and operating companies are from the U.S. Acquired companies had higher mean and median funding compared to closed and operating companies. Acquired companies also had more number of funding rounds compared to companies with closed and operating statuses.
* In terms of year, 2014 was the newest year and the oldest year was 1902. Most of the companies were founded quite recently around the 2000. The total investment data was very skewed, similar to the other type of funding.
  + - 1. Data exploration





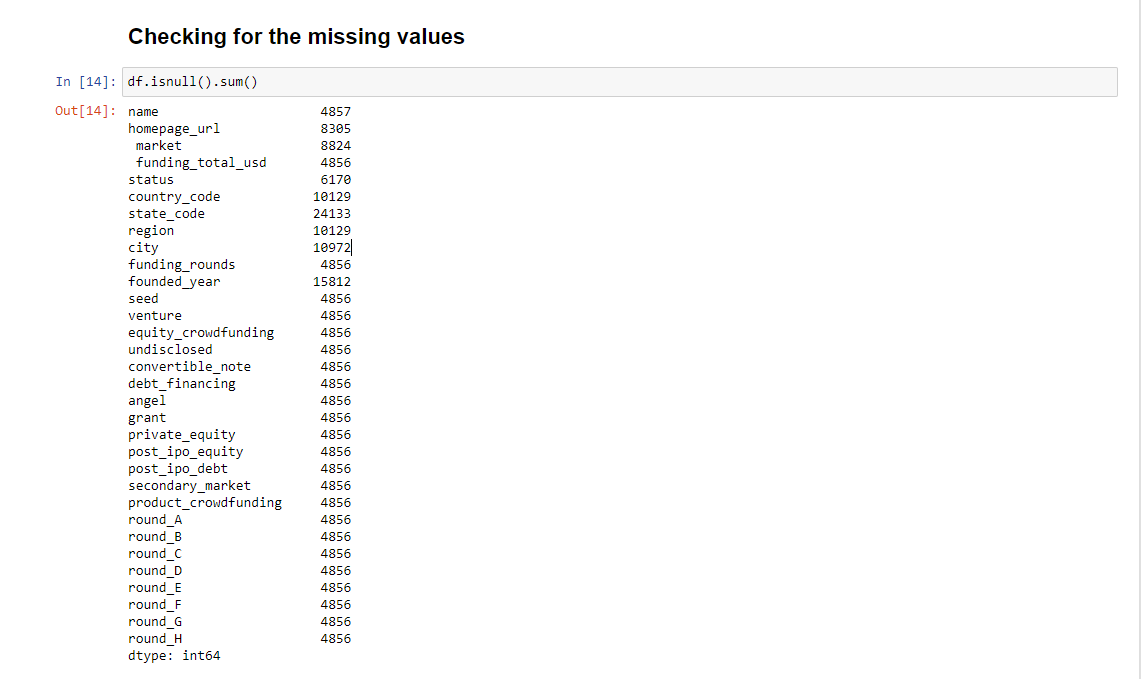


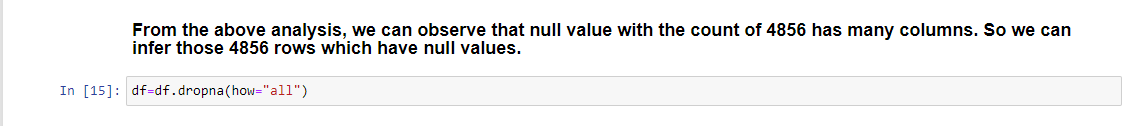


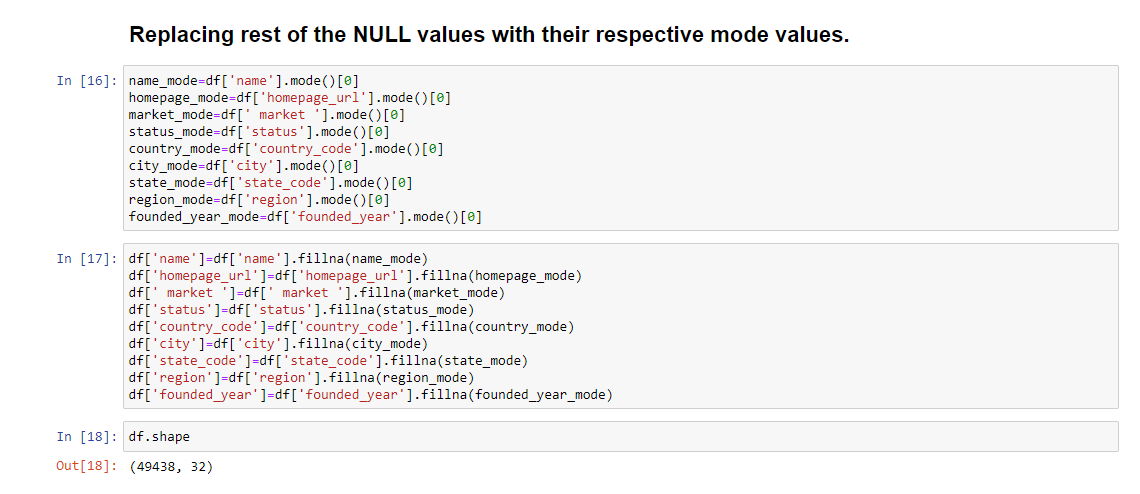


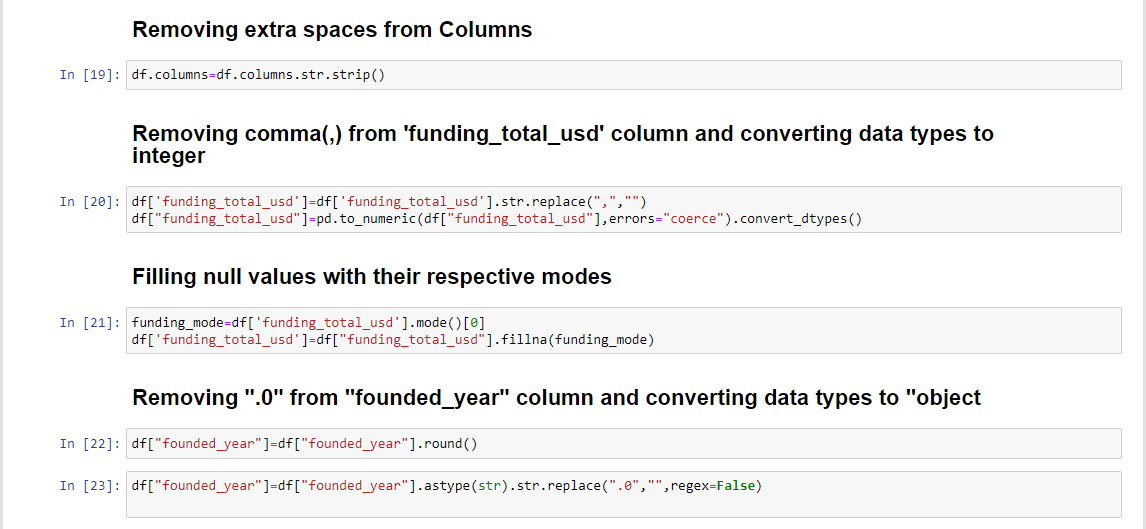
* + - 1. Pre-processing











* + 1. Algorithms Used
* KNN classifier
* Decision tree classifier

## Random forest classifier

* + - 1. Justification for choosing the models

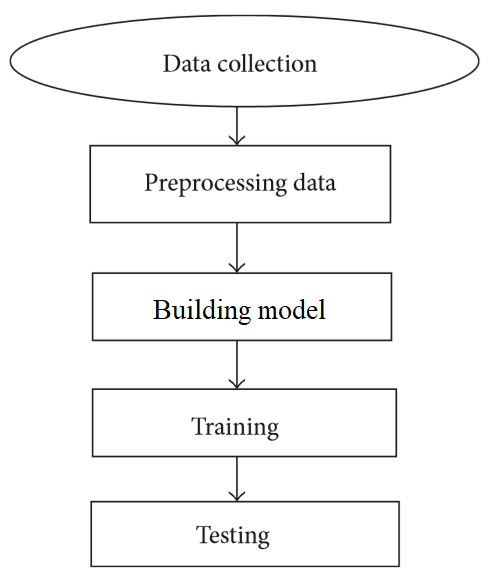
**KNN classifier:-**

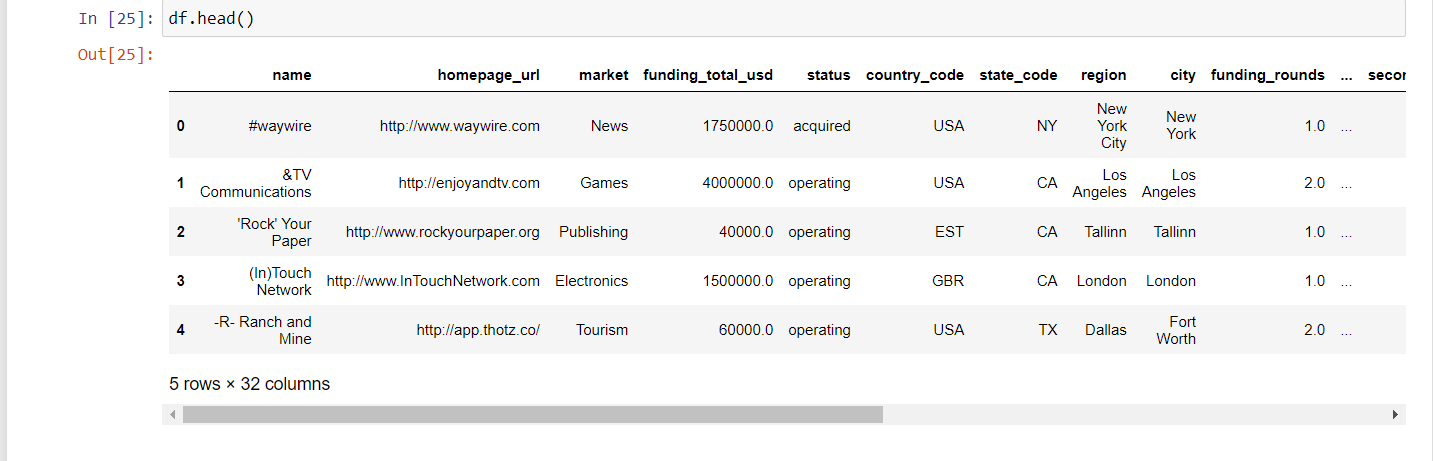
* **No Training Period**- KNN modeling does not include training period as the data itself is a model which will be the reference for future prediction and because of this it is very time efficient in term of improvising for a random modeling on the available data.
* **Easy Implementation**- KNN is very easy to implement as the only thing to be calculated is the distance between different points on the basis of data of different features and this distance can easily be calculated using distance formula such as- Euclidian or Manhattan
* As there is no training period thus new data can be added at any time since it wont affect the model.

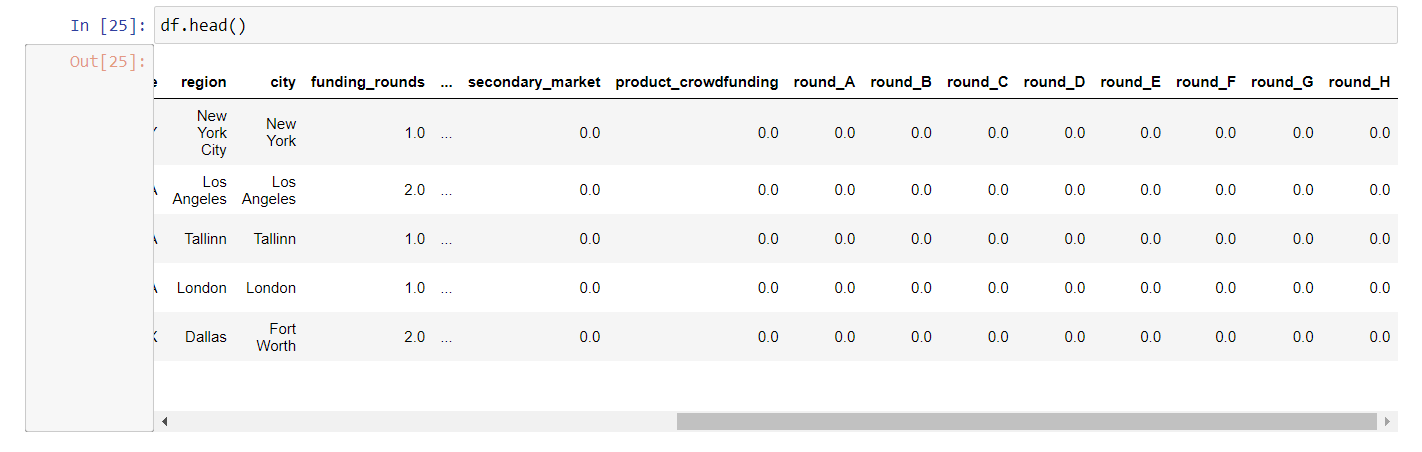
**Decision tree classifier:-**

* Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
* A decision tree does not require normalization of data.
* A decision tree does not require scaling of data as well.
* Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.

**Random Forest classifier:-**

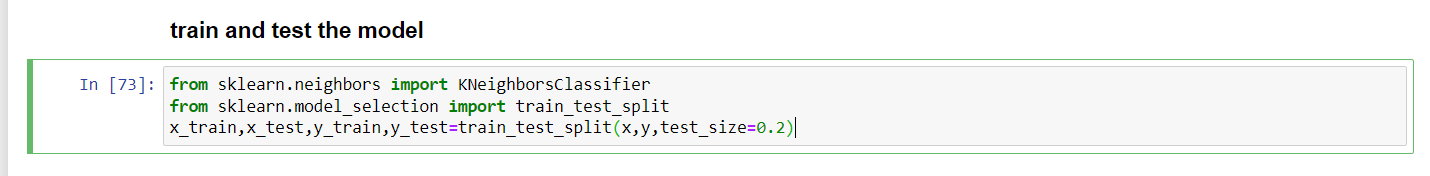
* Accuracy of Random forest is generally very high
* Its efficiency is particularly Notable in Large Data sets
* Provides an estimate of important variables in classification
* Forests Generated can be saved and reused
* Unlike other models It does nt overfit with more features
  + 1. Flow diagram of your model
    2. Dataset after preprocessing







* + 1. Dataset split(train and test)

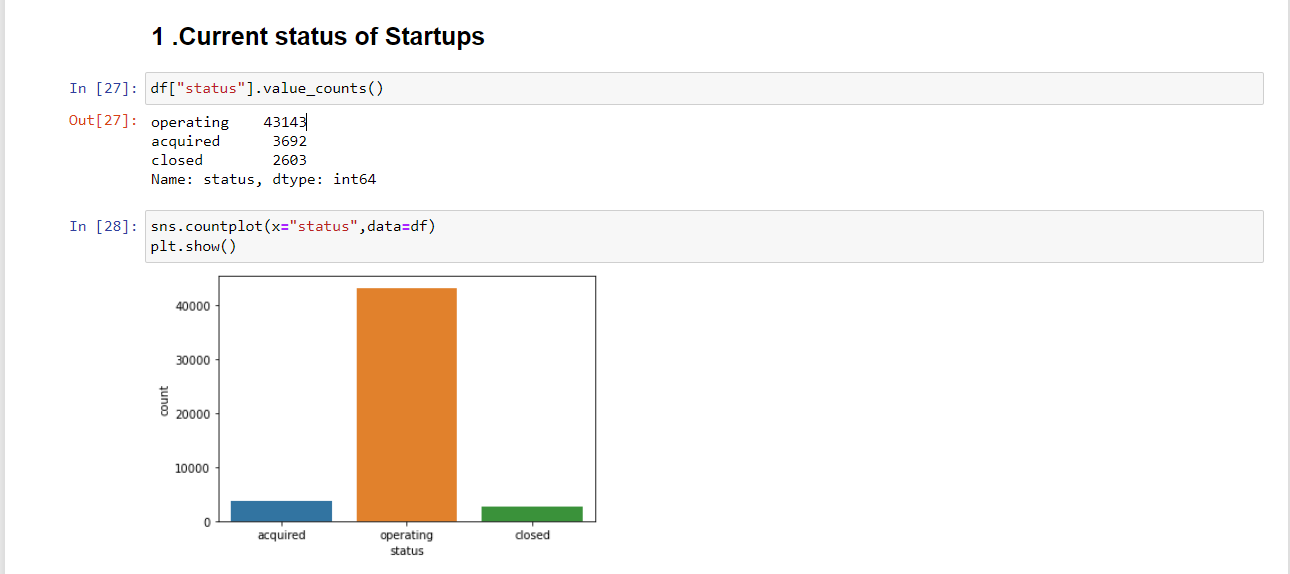


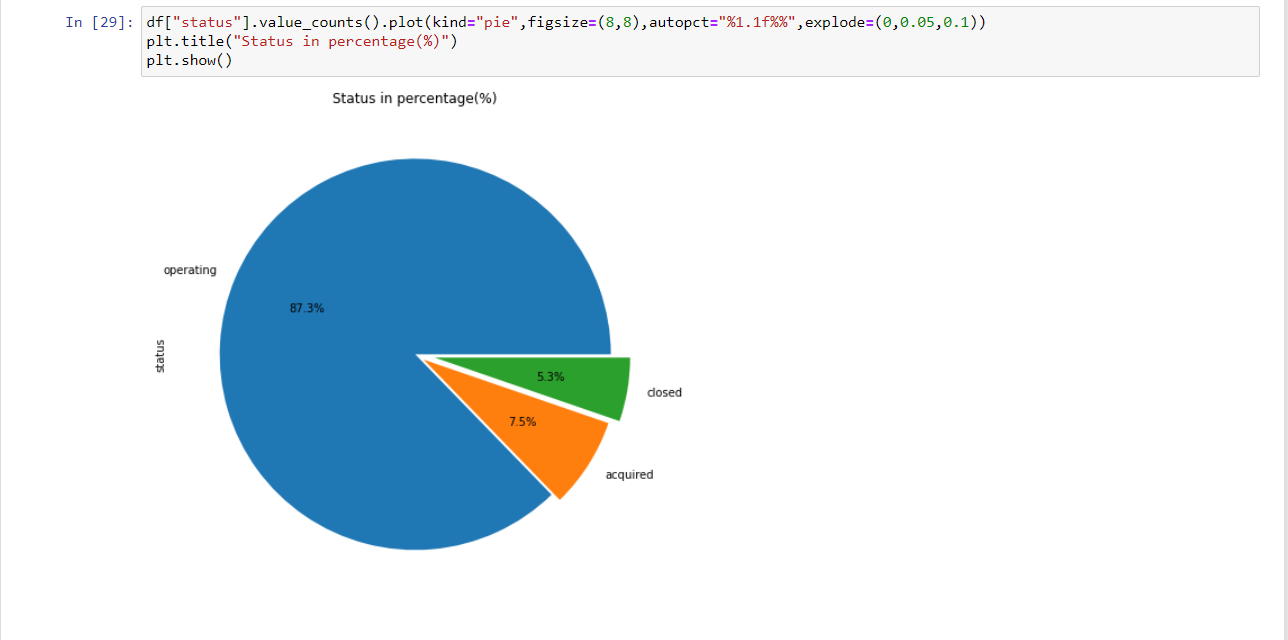
4. **Review 3 (Code)**

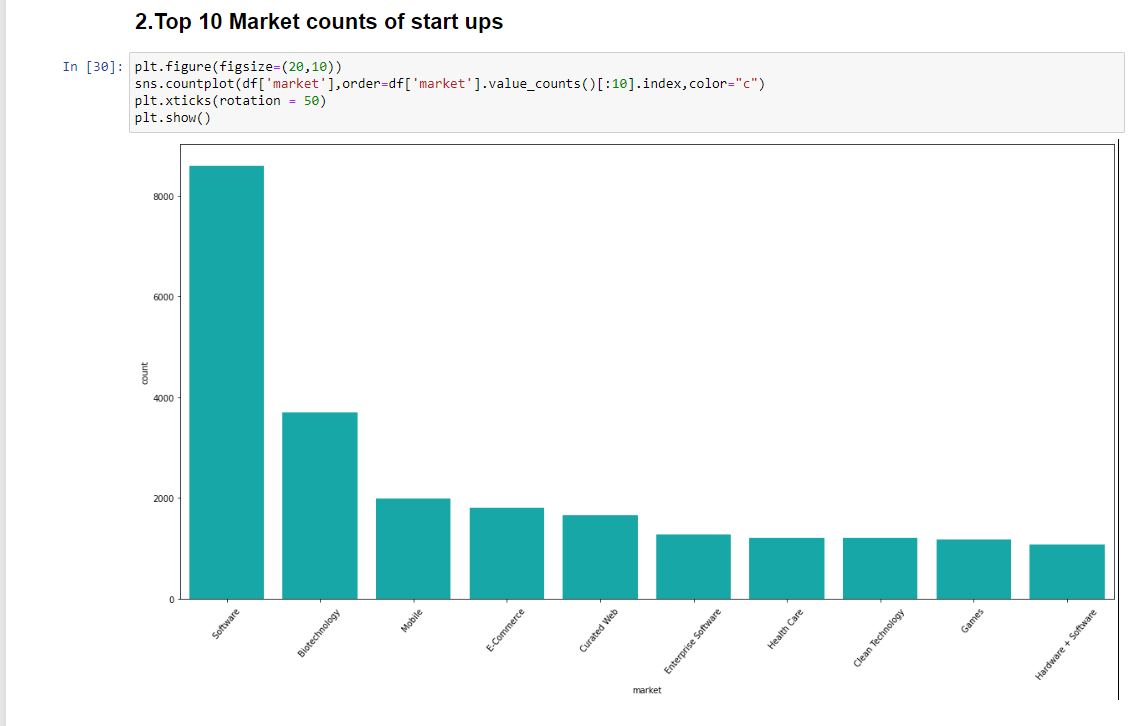
* 1. Implementation:
     + 1. Software and Hardware Description:-

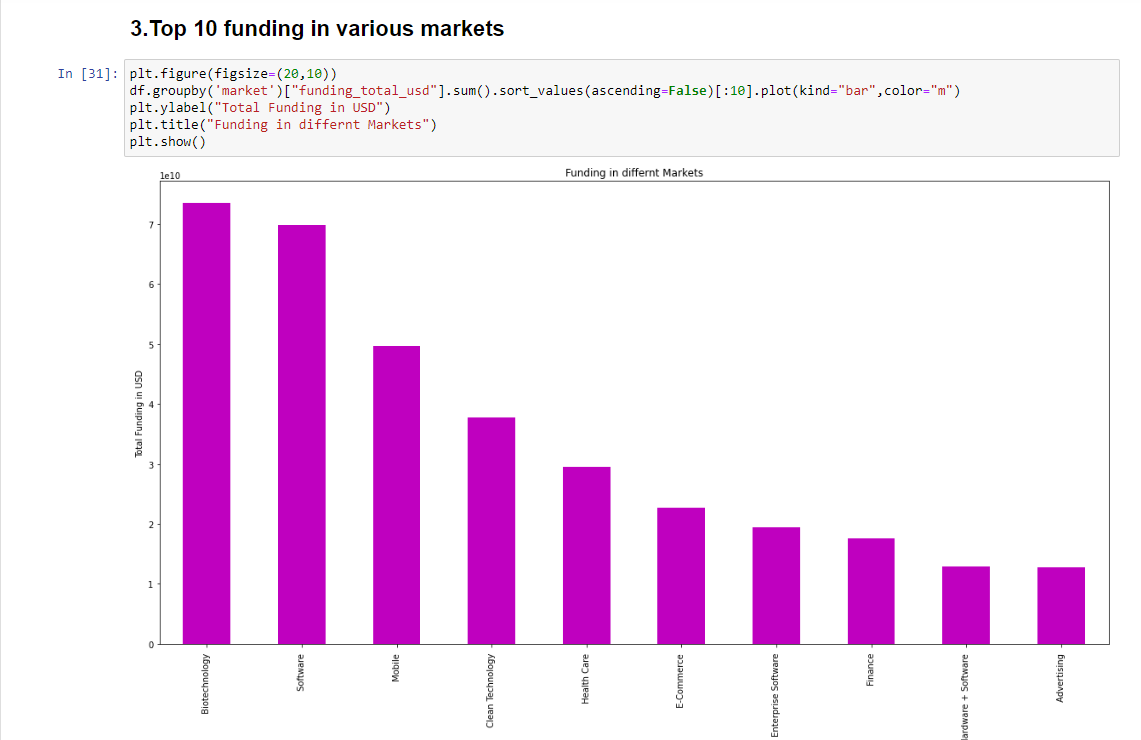
The entire analysis was done using Python and its ML frameworks: numpy , pandas, matplotlib, seaborn. Jupyter notebook was used.

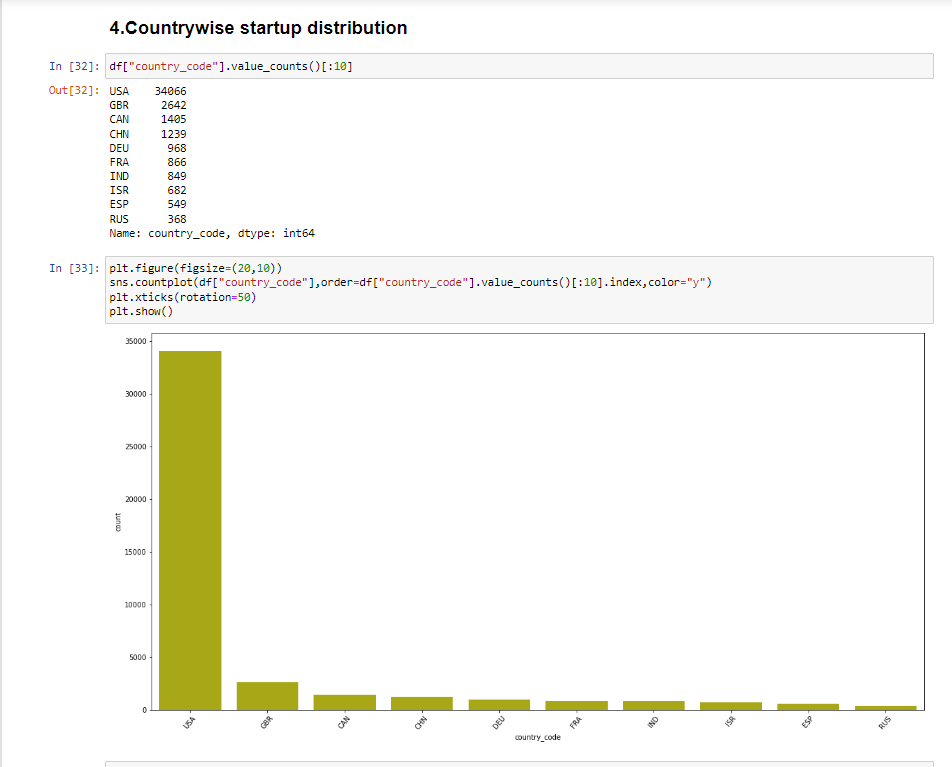
* + - 1. Output screenshots

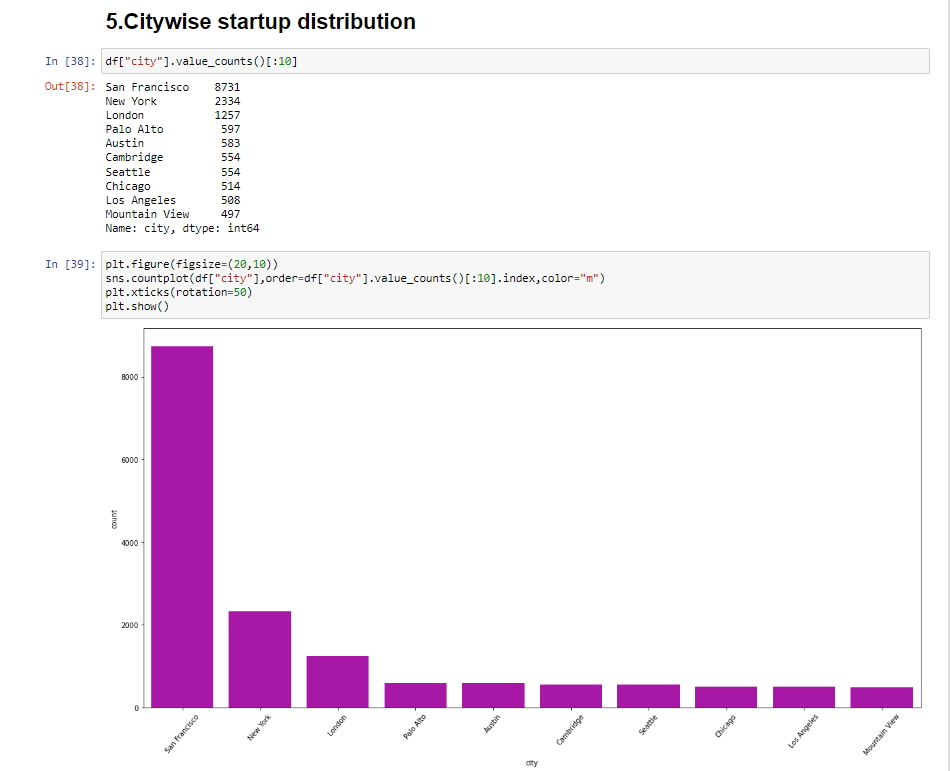


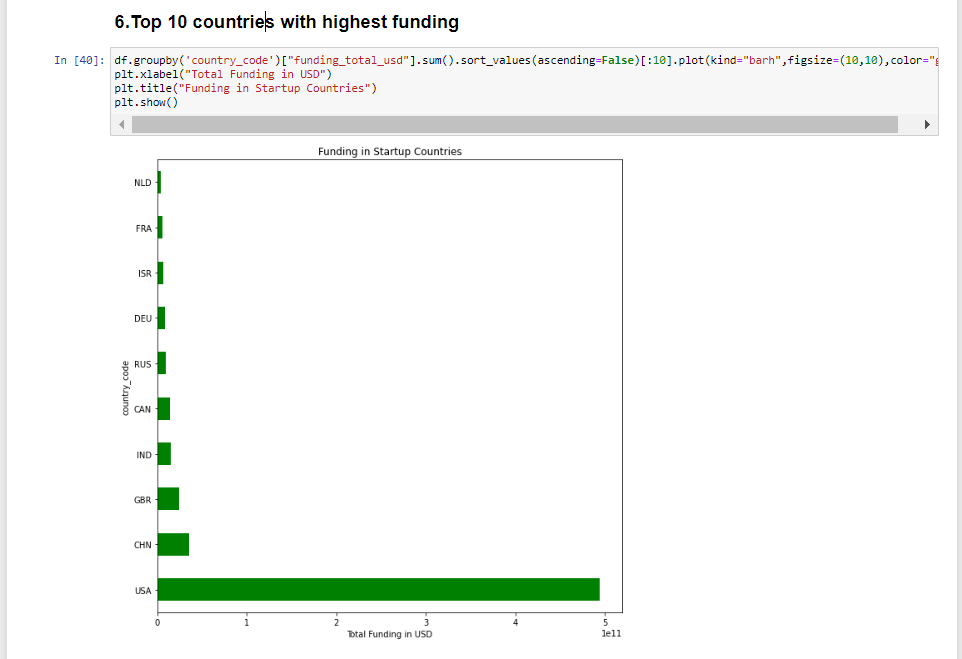


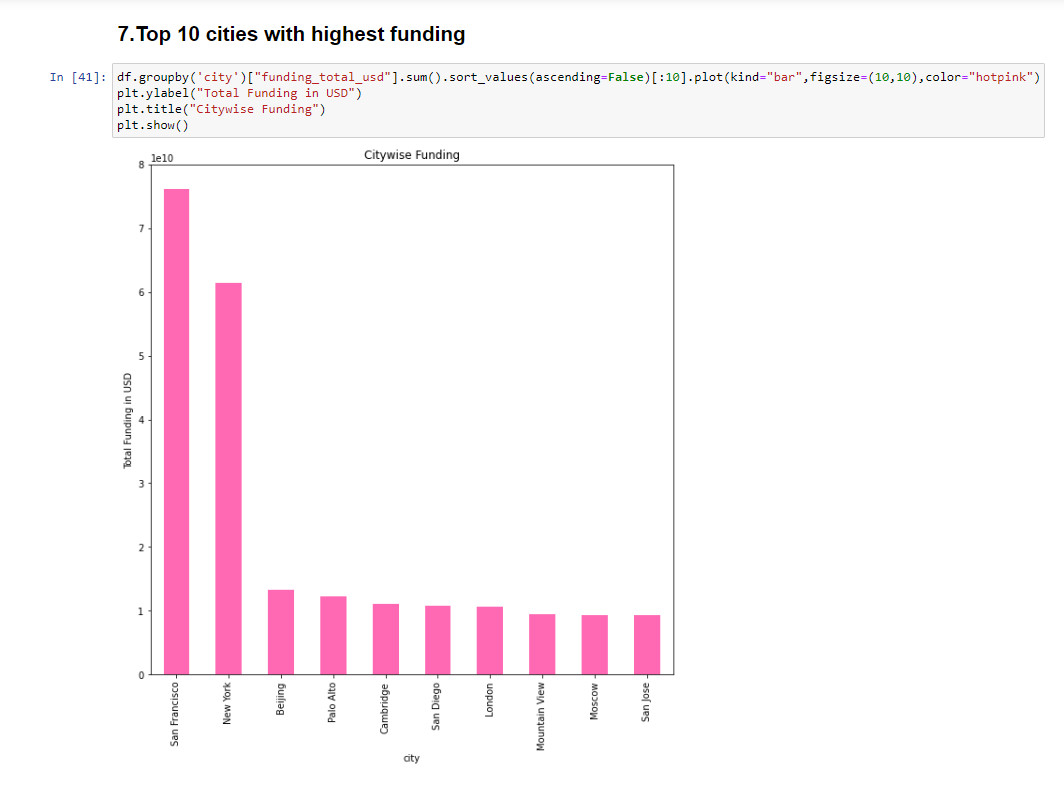


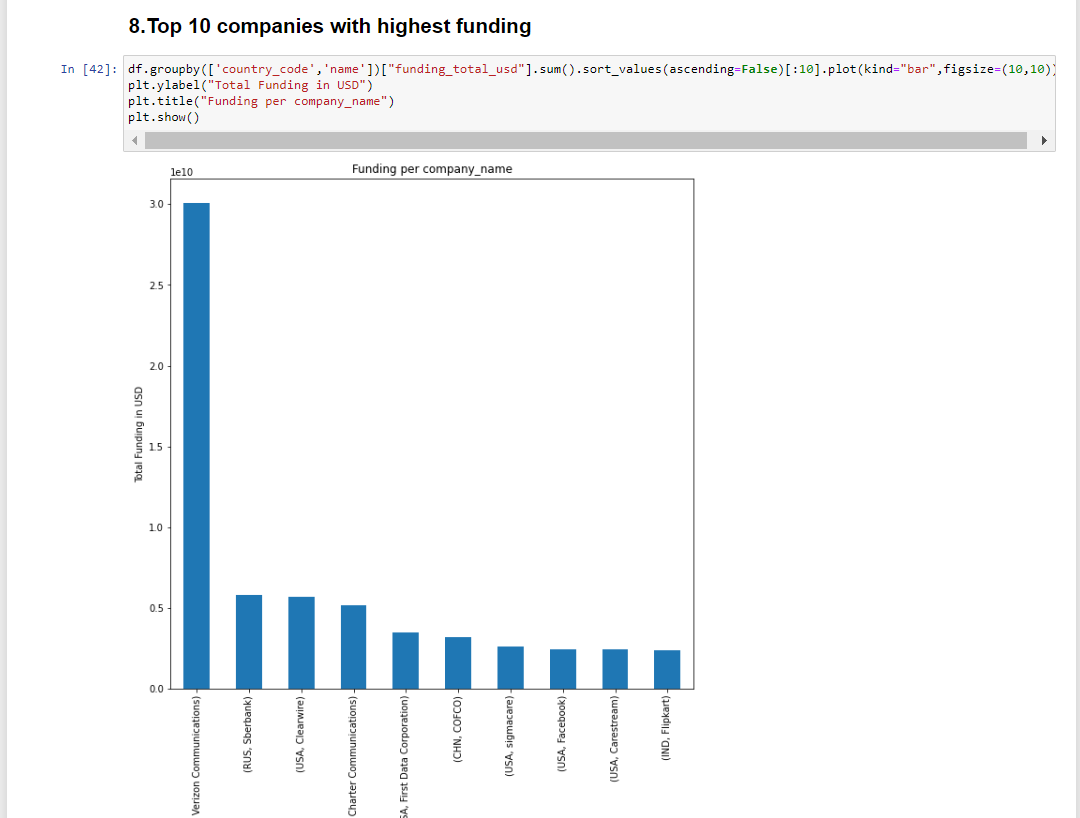




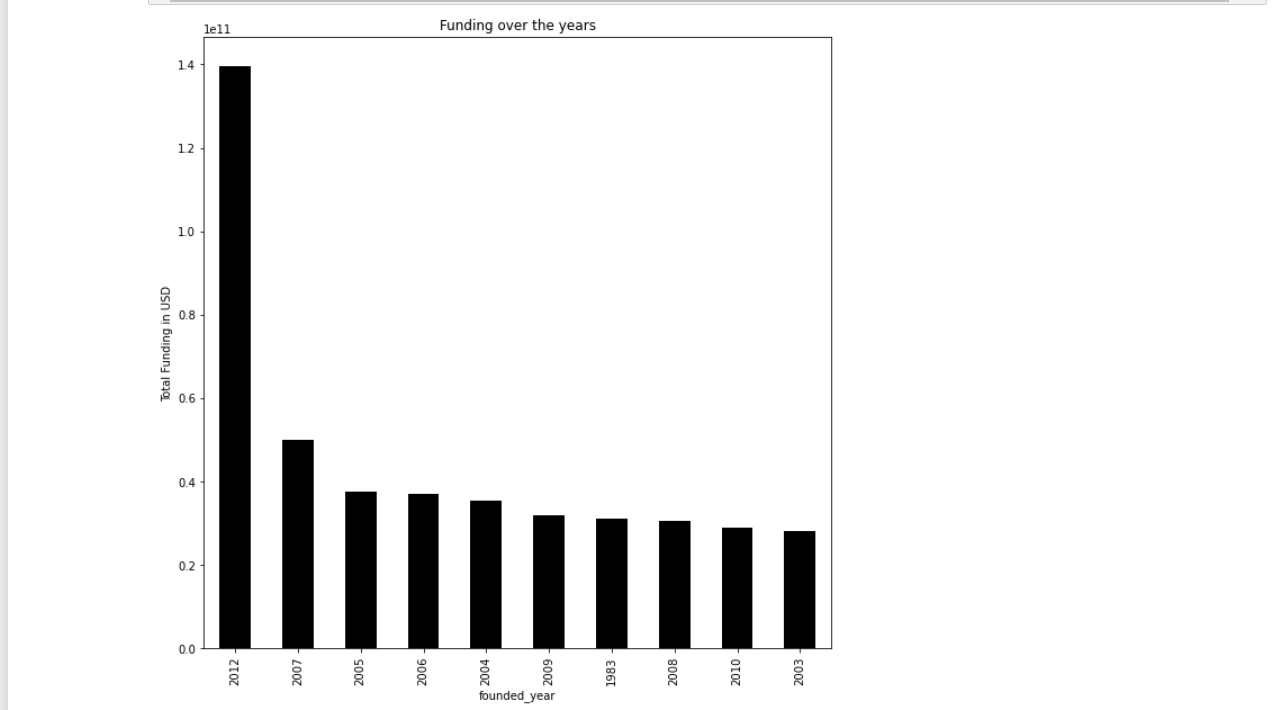


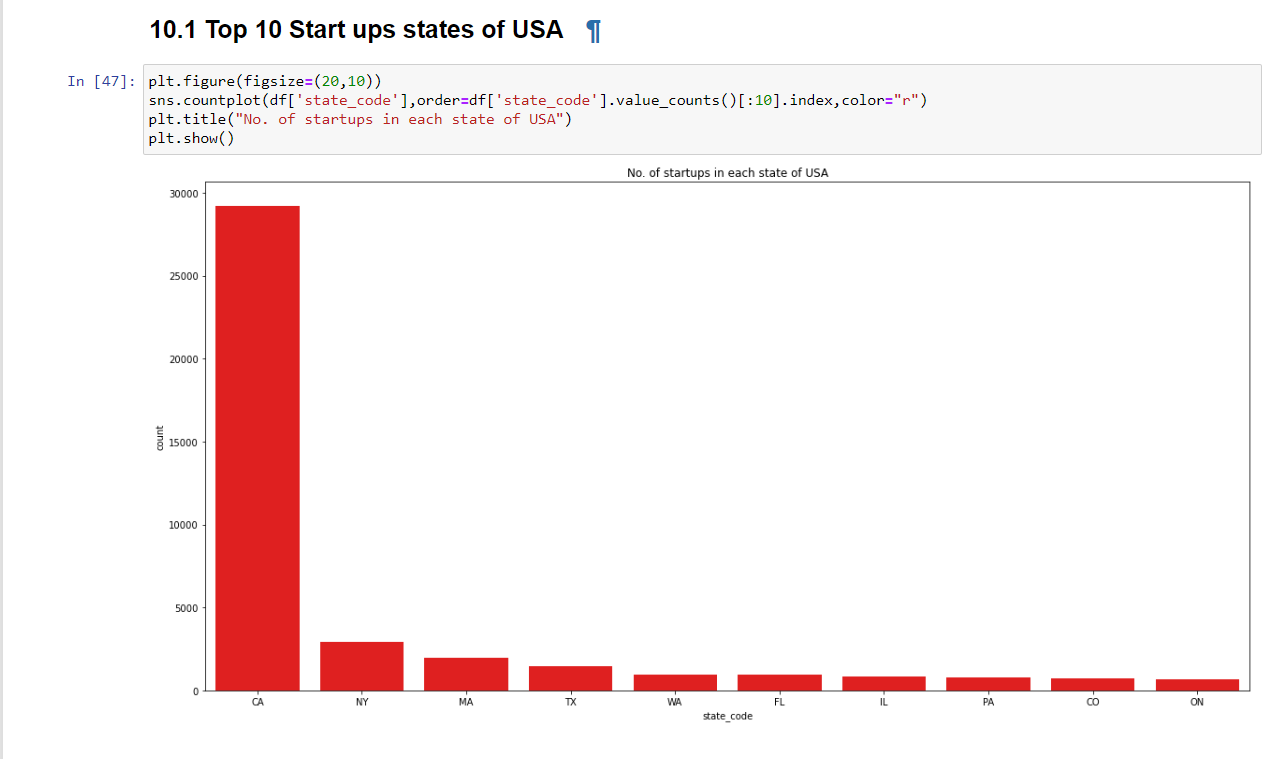


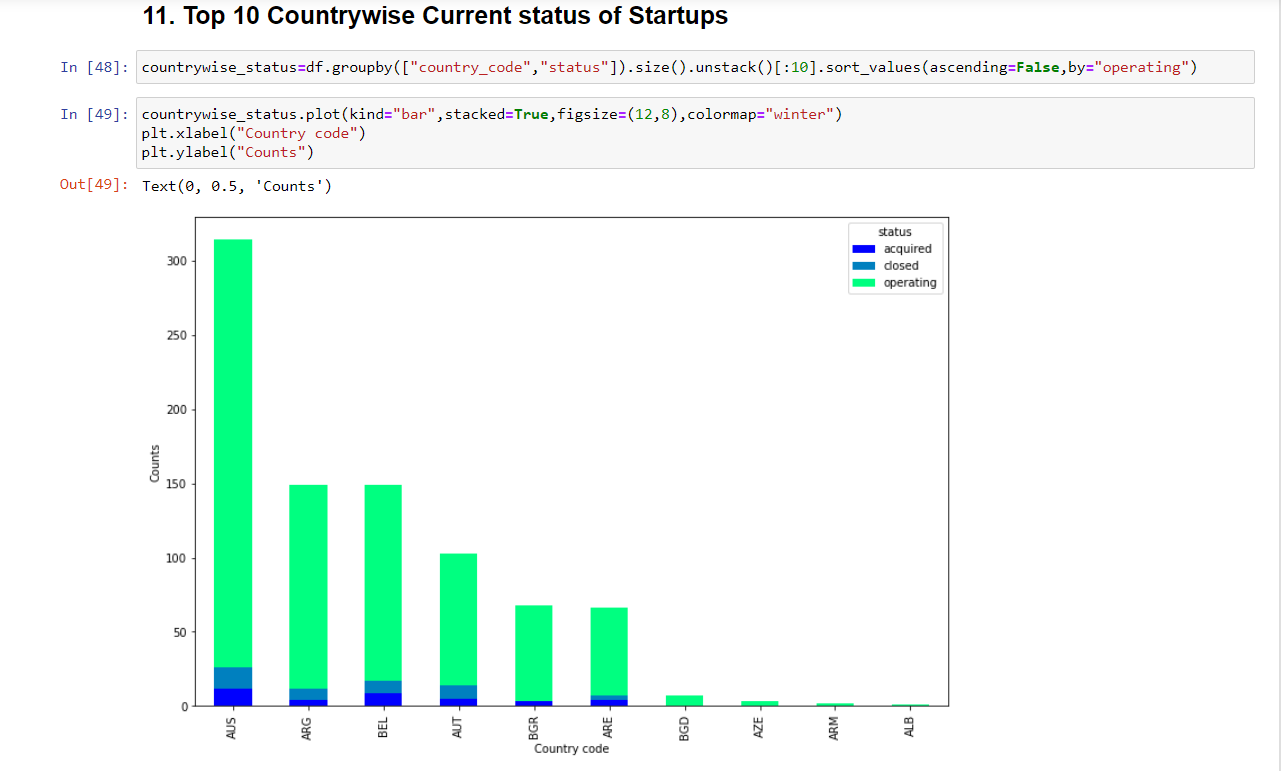


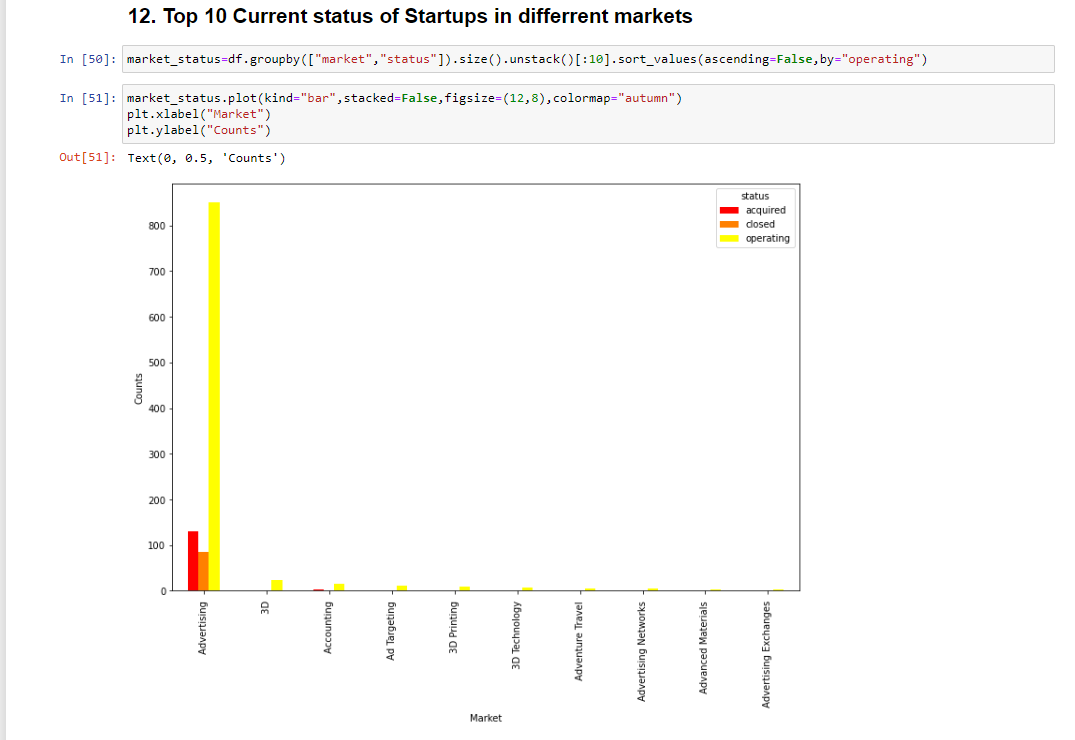


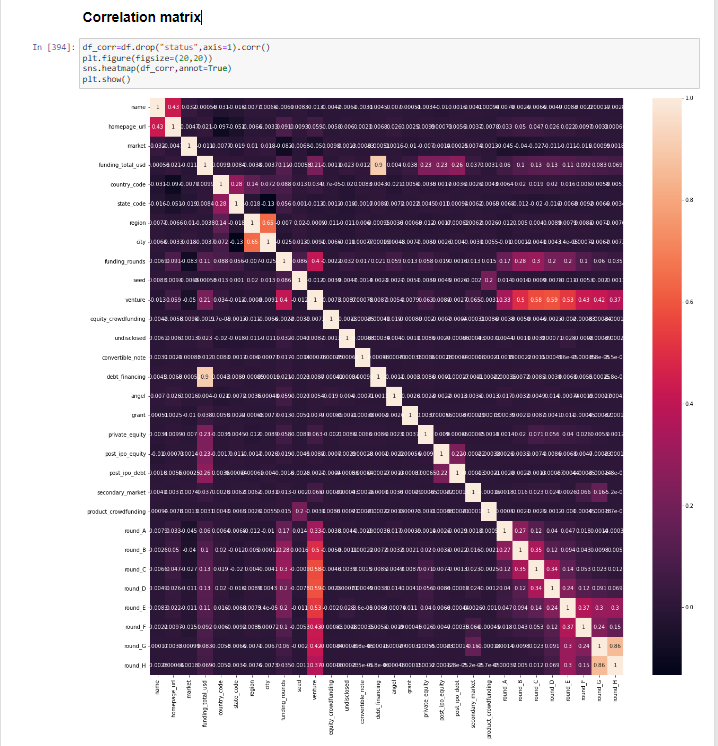






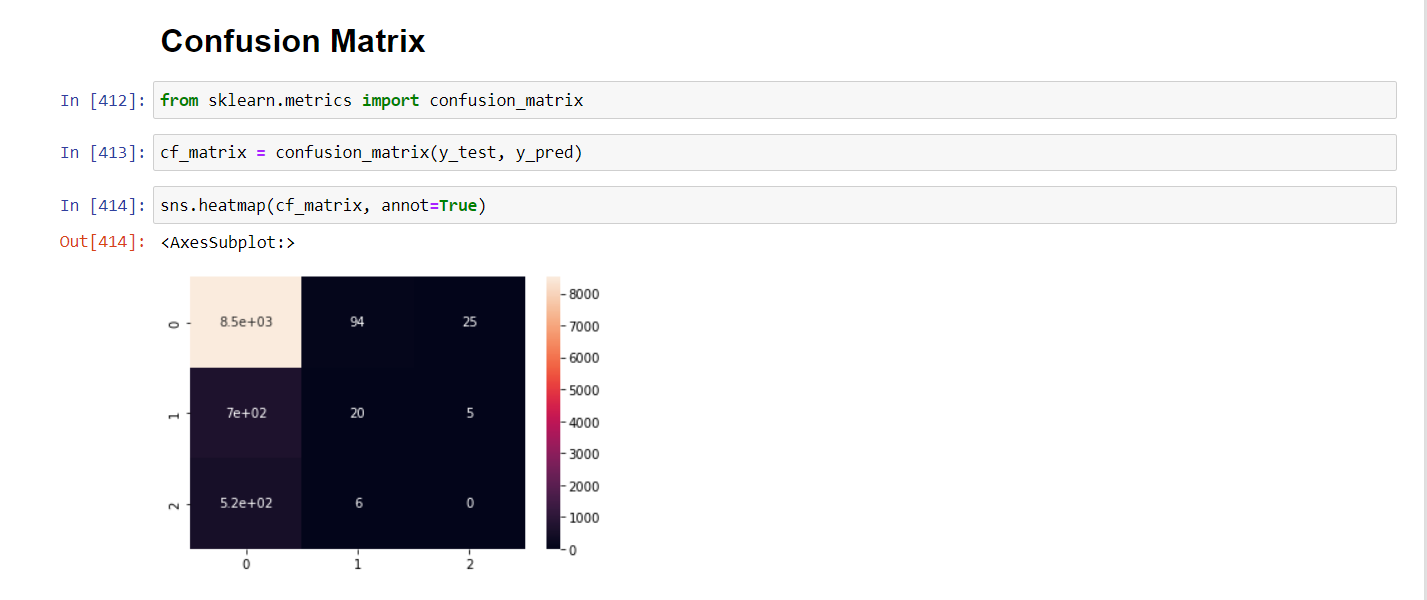




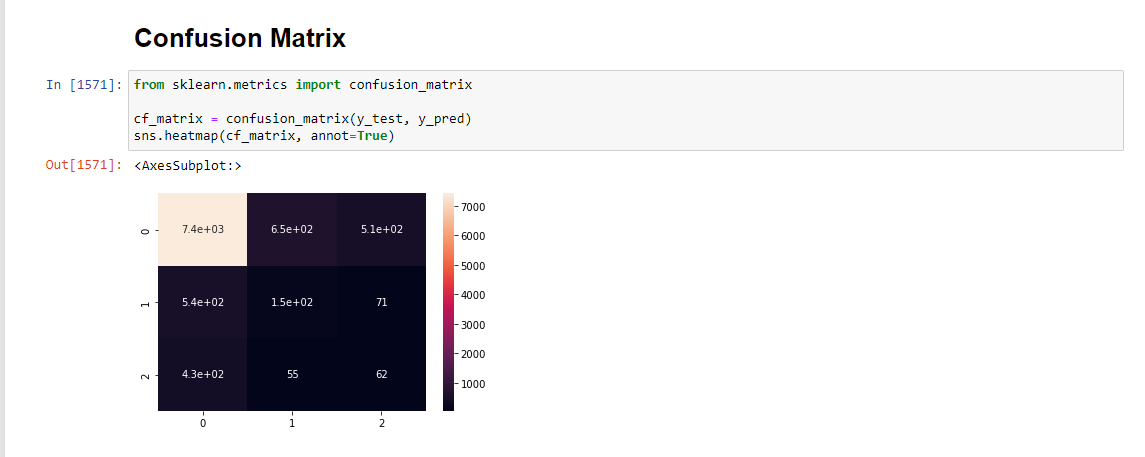
* 1. i) Correlation Matrix

Iii) Confusion Matrix

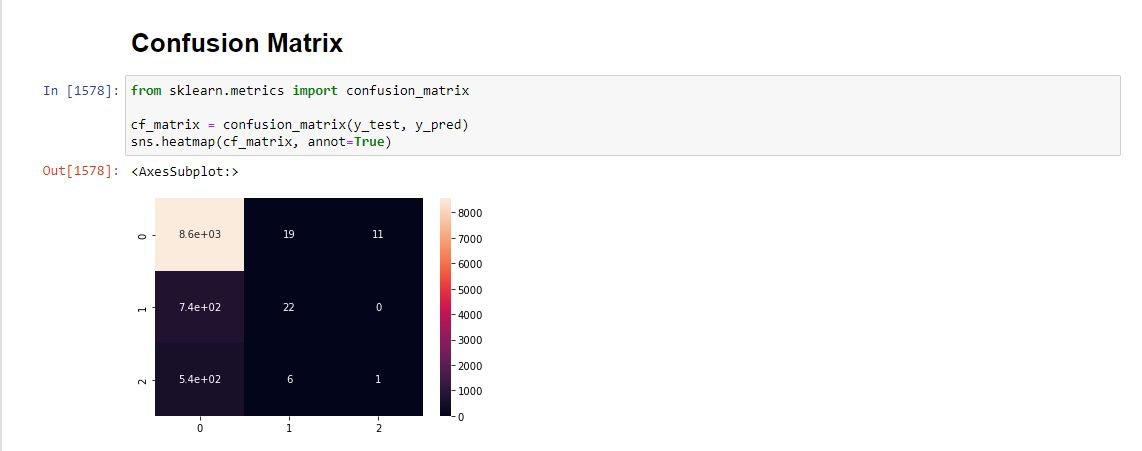
**KNN**



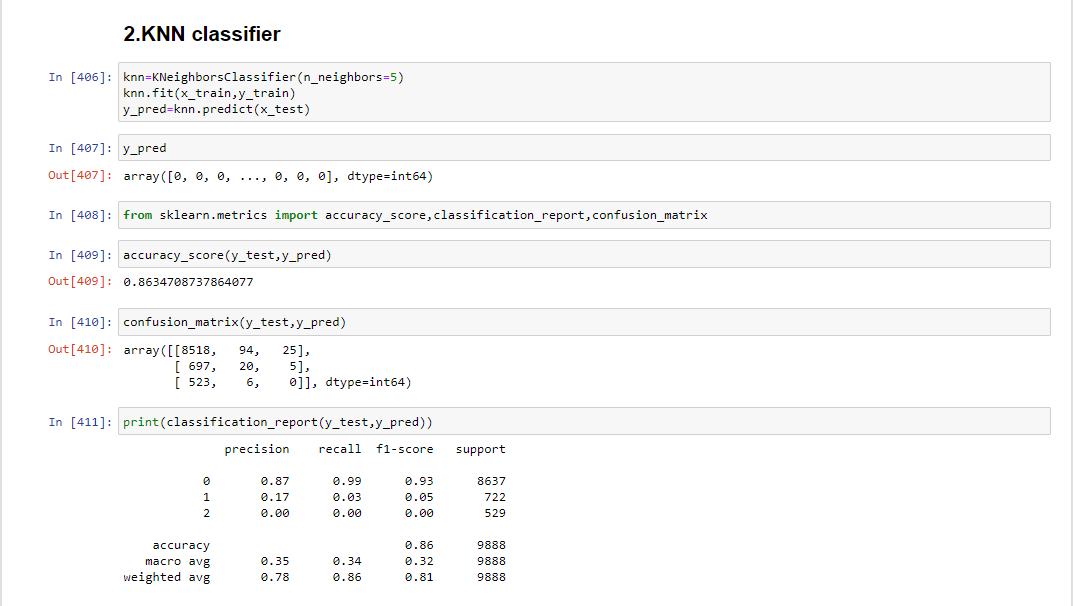
**Decision Tree**

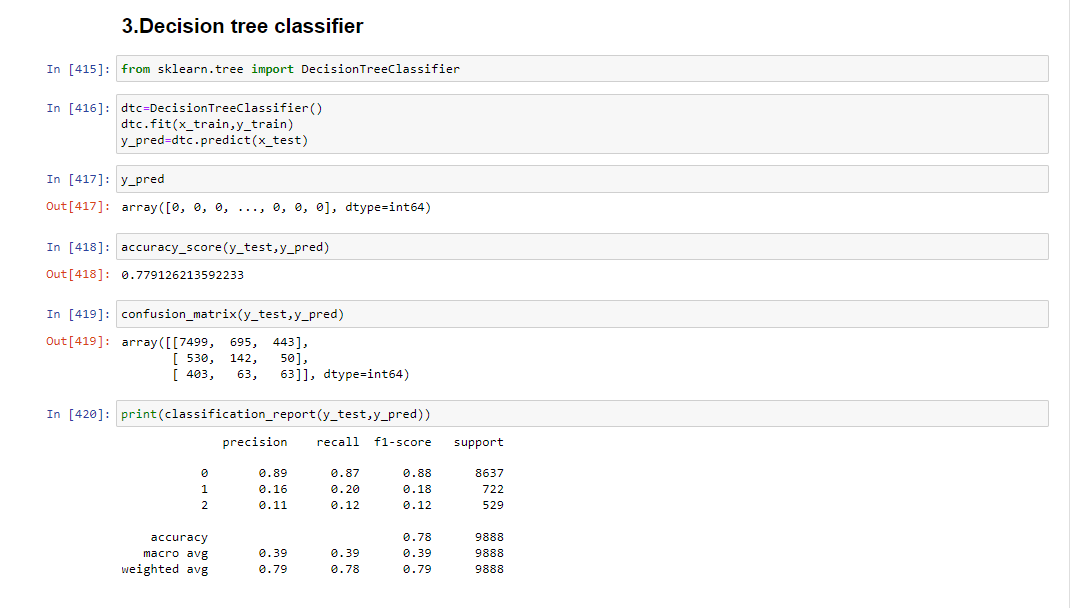


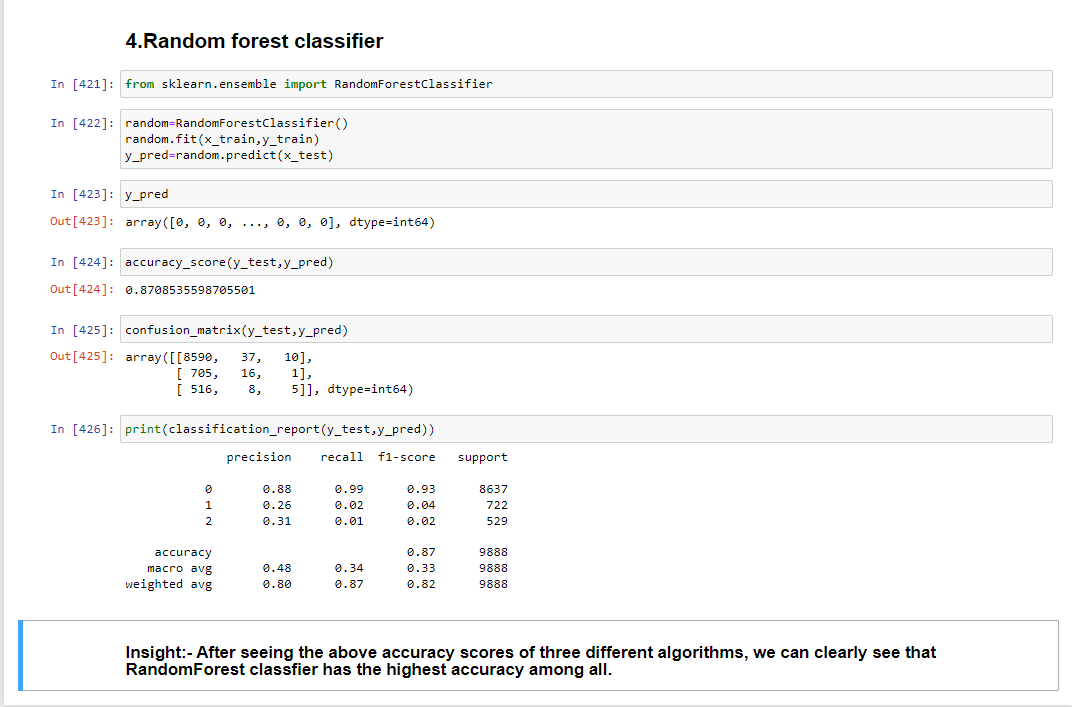
**Random Forest**



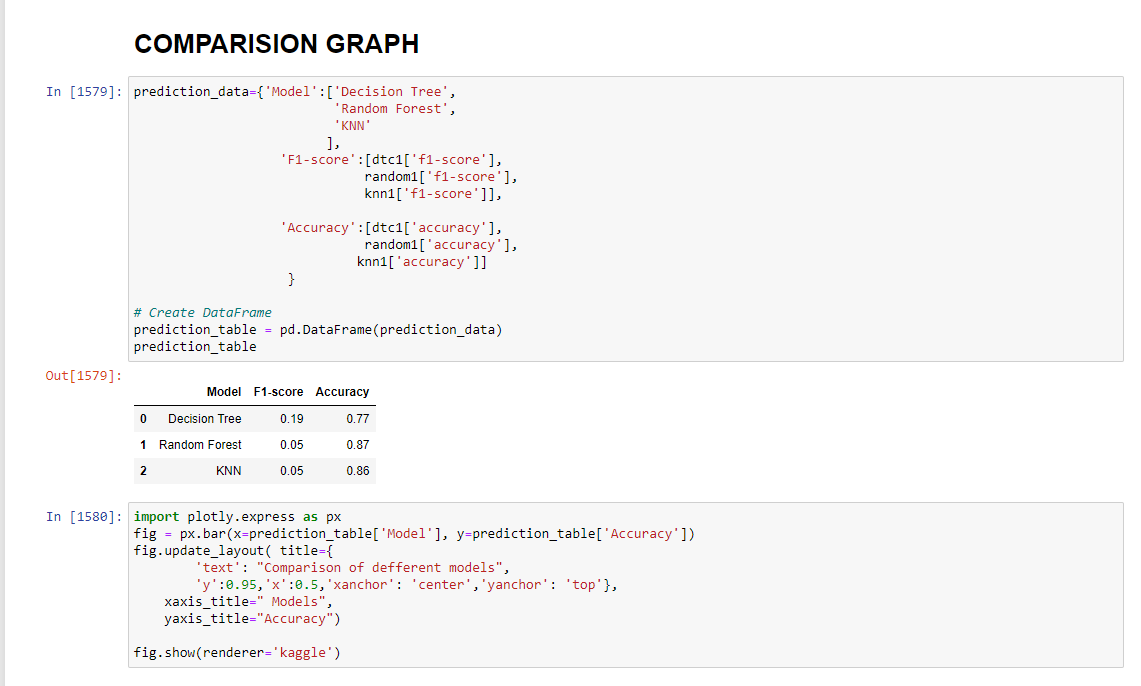
* 1. Comparison of the models used

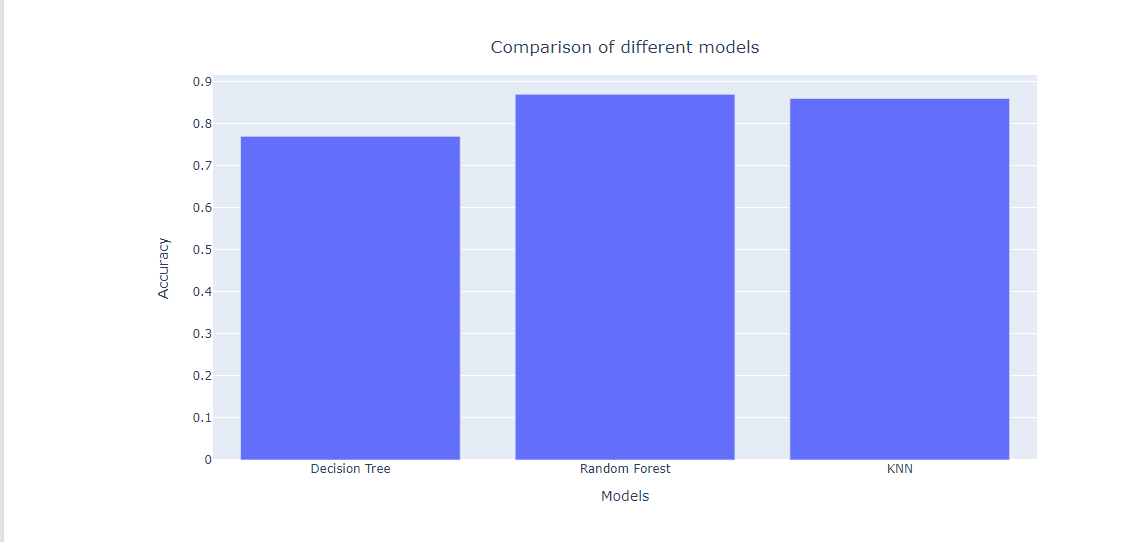






* 1. Comparison graph





**GITHUB LINK**

<https://github.com/Titas152000/ITA5007JCOMPONENT.git>

* 1. Conclusion

Industry, continent and total investment are important features. We have used algorithms like KNN classifier, Decision tree classifier, Random forest classifier. We received the best result when we used Random Forest classifier. For future scope, we would like more data for closed and acquired companies, test model with one-hot encoding, test with other models like SVM, Naive Bayes classifier. Using Crunchbase API, we can also make a real time dashboard and deploy a model so that it can assist investors and founders.

* 1. References:-
* <https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase?datasetId=517018>
* Yeh, J.Y.; Chen, C.H. A machine learning approach to predict the success of crowdfunding fintech project. J. Enterp. Inf. Manag. **2020**.
* Bai, S.; Zhao, Y. Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches. Systems **2021**, 9, 55. https://doi.org/10.3390/systems9030055