Data Mining on Startup Industry in India

Group No: 69

Instructor

Prof. P M Jat Sir

Team Members

201512006 - Chintan Sanghavi

201512062 - Hardik Bohra

201512083 - Shreya Chauhan

Prepared On:

14 November, 2016

Table of Contents

1. Abstract

Concept:

Everyone wants to be an entrepreneur now a days and government is supporting startups very well in India. Government has started Startup India campaign and not only that but also state government has also started supporting their respective states via different state policies for Startups. This is the domain when risk is high and people are confused whether they should go for startups or not, if yes then which domain they are interested in, what is the funding policy, how to apply for incubation centers and list goes on. Survey by economic times states that India has 19000 startups in 2015. Also we have 94 incubation centers in India.

Why this Topic:

Firstly this is the topic in which we as team are interested in. Secondly government is funding huge about of money for Startups but failure ratio is high. So we are trying to analyze behavior of startups and would like to predict few characteristics of startups. Also this is something challenging for us as it requires lot of research work and analysis.

Analysis:

We have dataset of all startups from 4 to 5 cities. So based on that data we will make a prediction about start-up success and failure. For that we will consider various variables like, start-up location, Investment type, Sector of Start-up and Amount of funding.

2. Objective

Using data mining tool ‘WEKA’ to do a multi-step data mining exercise. Interpreting the data well, understanding the structure of the data using one or more data mining algorithms, and presenting the findings.

Mining data to extract knowledge from available data.

To explore alternative data mining tools such as ‘Rapidminer’.

3. Data Set

In Data Mining project we are mining Start-up data of 4 to 5 cities to extract knowledge. Here we are using WEKA tool to mine the data. Original dataset consists of 980 records, but for better understanding we just take a sample data of 70 rows. Our primary goal is to use different mining tools to apply classification and clustering algorithms.

Data set is collected from government portal which contains start-up location, Investment type, Sector of Start-up and Amount of funding. From these attributes that possibly dependent with the success or failure of the startup.

Table1. Start-up Related Variables

Variable	Description	Possible Values
Location	Startup Location	{Bangalore, Chennai, Gurgaon, Mumbai, New Delhi}
Investment Type	Type of the Investment	{Seed Funding, Private Equity}
Sector	Sector of the Startup	{eCommerce, Education, HealthCare, Technology}
Amount	Amount of Funding	{1-100000000}
Status	Status of the Startup	{Success, Failure}

The domain values for some of the variables were defined for the present investigations are as follows:

- Location – In original dataset, there are many entries of startups from various cities, but we have collected the data from five cities that are Bangalore, Chennai, Gurgaon, Mumbai and New Delhi.
- Investment Type – There are many sources available for funding. In which we have considered two investment types that are Private Equity and Seed Funding. Private Equity refers to investment by private resources in private companies and seed funding refers to the first significant money received.
- Amount – This domain value consist the amount of funding.
- Sector – The value of this domain refers the sector to which start-up is belongs to. We are considering sectors like eCommerce, Education, HealthCare and Technology.
- Status – This domain contains the value which gives the idea about start-up success or failure.

4. Design Principles

The design principle of this project included data cleaning and preprocessing. The first phase of this project includes cleaning the data and makes it compatible to data mining tool, the next phase is to apply data mining algorithms to get classification and clustering results and study these algorithms.

The data is cleaned and preprocessed manually by checking all the attribute entries and made changes using Microsoft Office Excel.

Using 'WEKA' – Data Mining Tool, based on the structure and type of Database, we applied following algorithms:

1. Classification Algorithms
    a. Logistic Algorithm
    b. J48 (Decision Tree)
2. Clustering Algorithms
    a. Expected Maximization Algorithm
    b. K-Mean Algorithm

5. Implementation

To mine data we have followed following process:

1. Data Preprocessing:
    a. As it is real time data, it is noisy data and need preprocessing.
    b. To make it easy to handle, we have trimmed original data.
    c. We are using 4 attributes out of 7.
    d. Date data was not consistent throughout the data. In some records it was mention as date or number or simply a text. And we don't need that variable column so we simply ignore it from our sample data set.
    e. Some special characters were used in data which is not accepted by WEKA so we remove these characters or replace with appropriate one.
2. Import preprocessed data in WEKA
3. Applied classification and clustering algorithms as mention below:

Based on the structure of Data Set and type of DB, specific algorithms can only yields the results that interpret data well.

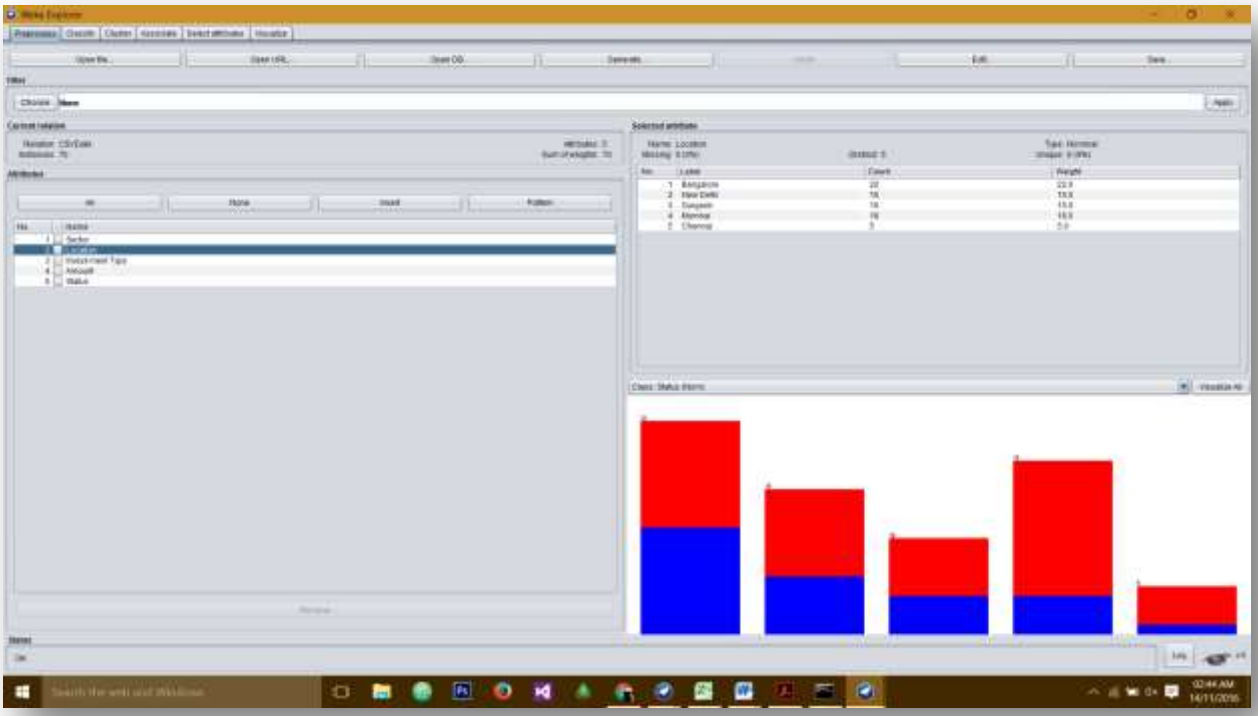
5.1 Classification Algorithms

We used same database for data mining project and data warehousing project. As the database is database is a sample data with some independent and some dependent variables. After analyzing database, we come to conclusion that to apply different data mining algorithms on different sets of attributes from the database, to interpret data well.

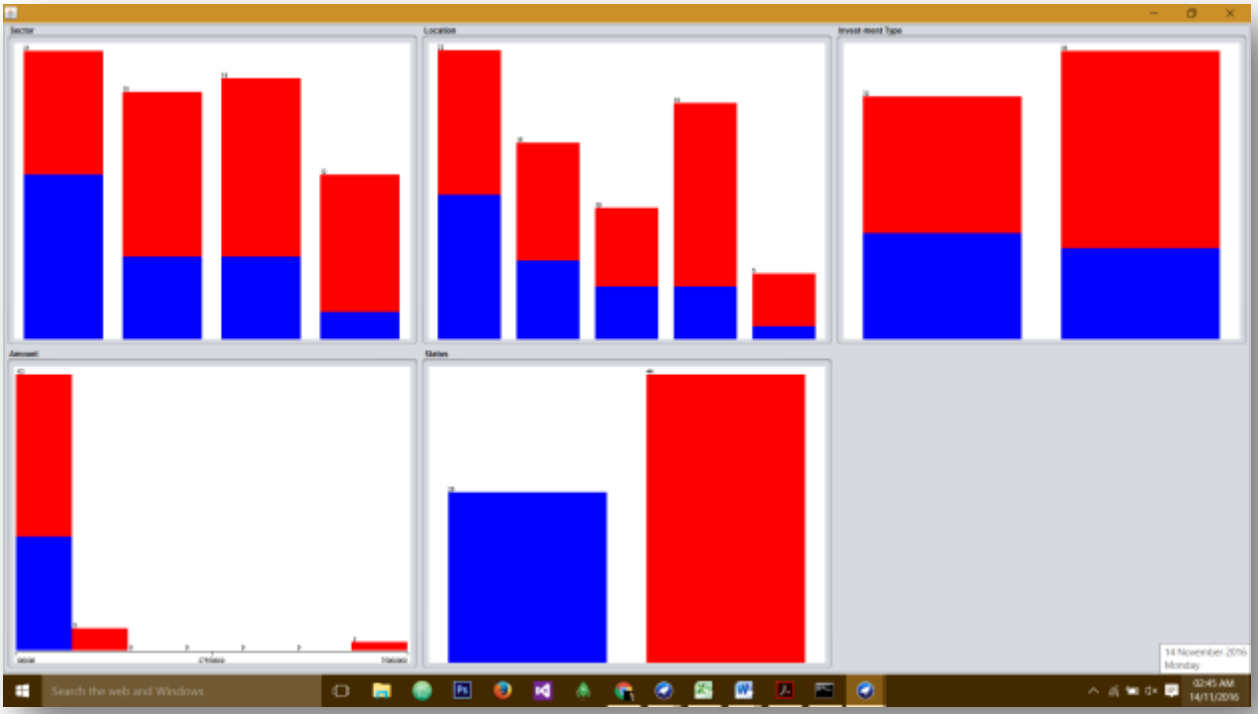
Two broad sets formed for the data mining project are:

1. Start-ups prediction and clustering:

   Location + Sector + Investment Type + Funding Amount



Basic Classification Histogram



In the above diagram we can select start-up sectors, location, investment types, amount and status.

Red – particular startup get success

Blue – particular startup fails

a. Logistic Algorithm

Highly regarded classical statistical technique for making predictions.

'Logistic Algorithm' assigns weightage to the attributes in the data set. And uses the logistic regression formula to predict how accurately a particular attribute value can be determined for the future instances. Thus using relative attribute increases prediction capability as oppose to using all the data available. Since using independent attributes would affect assignment of weightage which is used to formulate the prediction accuracy. To apply the logistic algorithm classification on 'Start-up data' set of relevant attributes are used. Logistic algorithm then assigns weightage to all attributes in dataset.

```
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                                    Class
Variable                          Success
=========================================
Sector=eCommerce                   0.9485
Sector=Healthcare                 -0.0558
Sector=Technology                 -0.5995
Sector=Education                  -0.4927
Location=Bangalore                 0.8904
Location=New Delhi                -0.0403
Location=Gurgaon                   0.1316
Location=Mumbai                   -0.6695
Location=Chennai                  -1.1056
Invest-ment Type=Private Equity    -0.456
Amount                                -0
Intercept                         -0.2192
```

Then these weightage are run through 'logistic regression formula' to predict the attribute under consideration in the example.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          47               67.1429 %
Incorrectly Classified Instances        23               32.8571 %
Kappa statistic                          0.2675
Mean absolute error                      0.4081
Root mean squared error                  0.4771
Relative absolute error                 87.1068 %
Root relative squared error             98.6506 %
Total Number of Instances               70

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.462    0.205    0.571      0.462    0.511      0.271    0.649     0.578     Success
                 0.795    0.538    0.714      0.795    0.753      0.271    0.649     0.731     Failure
Weighted Avg.    0.671    0.414    0.661      0.671    0.663      0.271    0.649     0.674

=== Confusion Matrix ===

  a  b   <-- classified as
 12 14 |  a = Success
  9 35 |  b = Failure
```

Logistic Algorithm

Thus from the above diagram we interpret that using ‘Logistic Classification Algorithm’ can predict next/future accuracy will be 67.10 %, given the dependent relation among all the attributes, that we used for this example.

b. J48 Algorithm (Decision Tree)

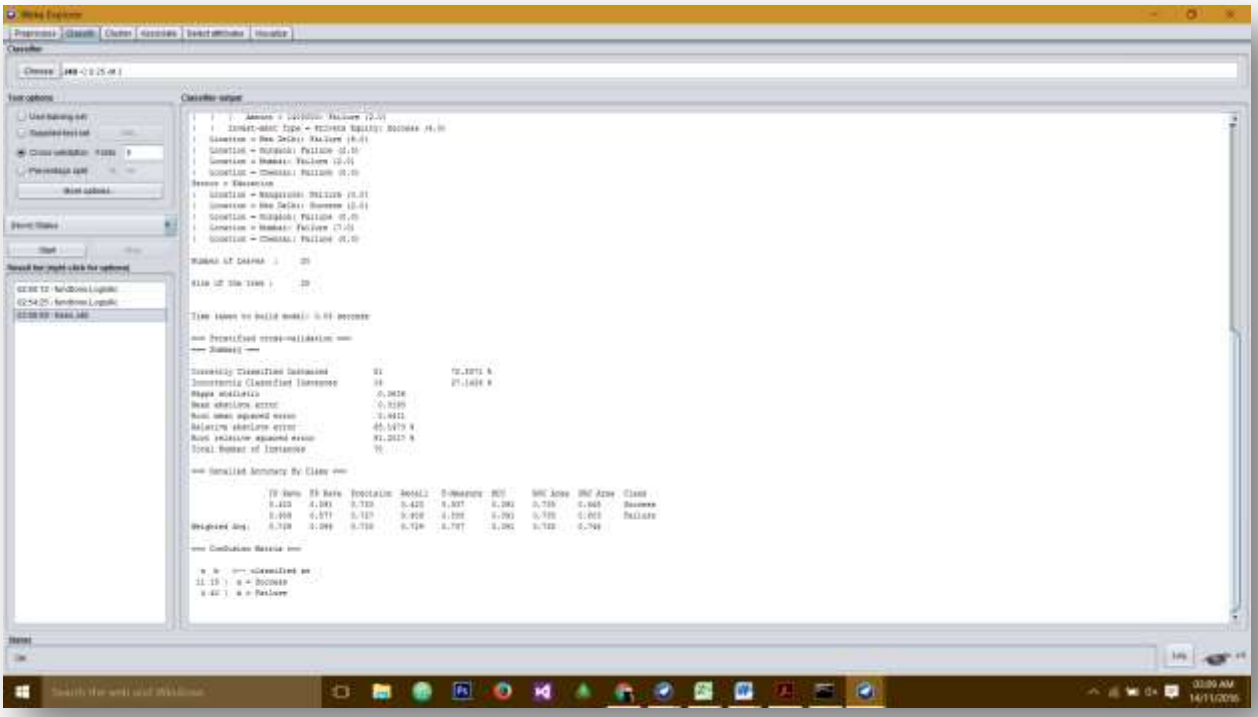
Logistic Algorithm cannot predict numeric values. Whereas J48 algorithm can predict both ‘nominal’ and ‘numeric’ attribute values.

J48 algorithm uses ‘most relevant attribute’ from the dataset to determine the prediction values, thus it’s better to have all the attributes rather that only relevant attributes, as we did in logistic algorithm.

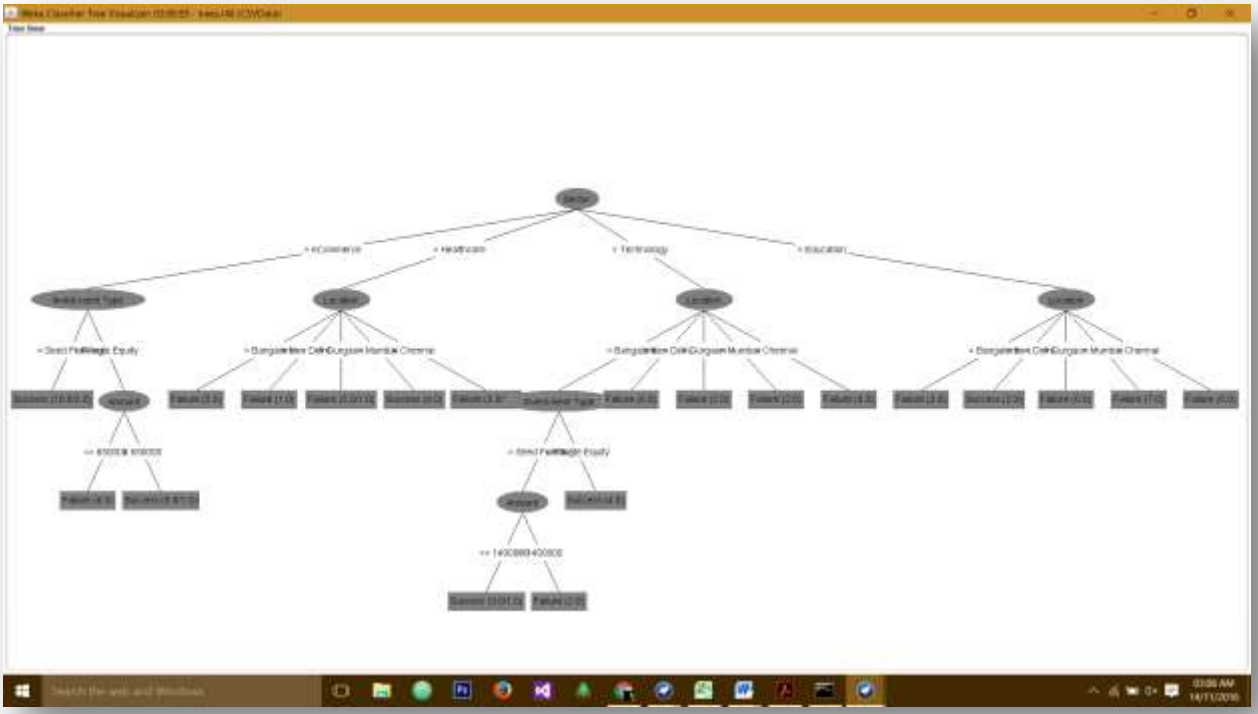
Using all the sample dataset for J48 algorithm, the prediction efficiency increases.

J48 algorithm visualizes result in the form of ‘Decision Tree’, where most relevant attributes are used for prediction of particular attribute’s future-instance value. Using this tree, rules can be formed.

J48 Algorithm on Startup Dataset



From the above image, Successful start-ups can be predicted with 72.85% accuracy using the attribute ‘Sector’ which is determined as ‘most relative’ by J48.



Where attribute ‘Sector’ is not alone used to predict the success or failure, but other relevant attributes such as ‘location’ and ‘Amount’ etc.

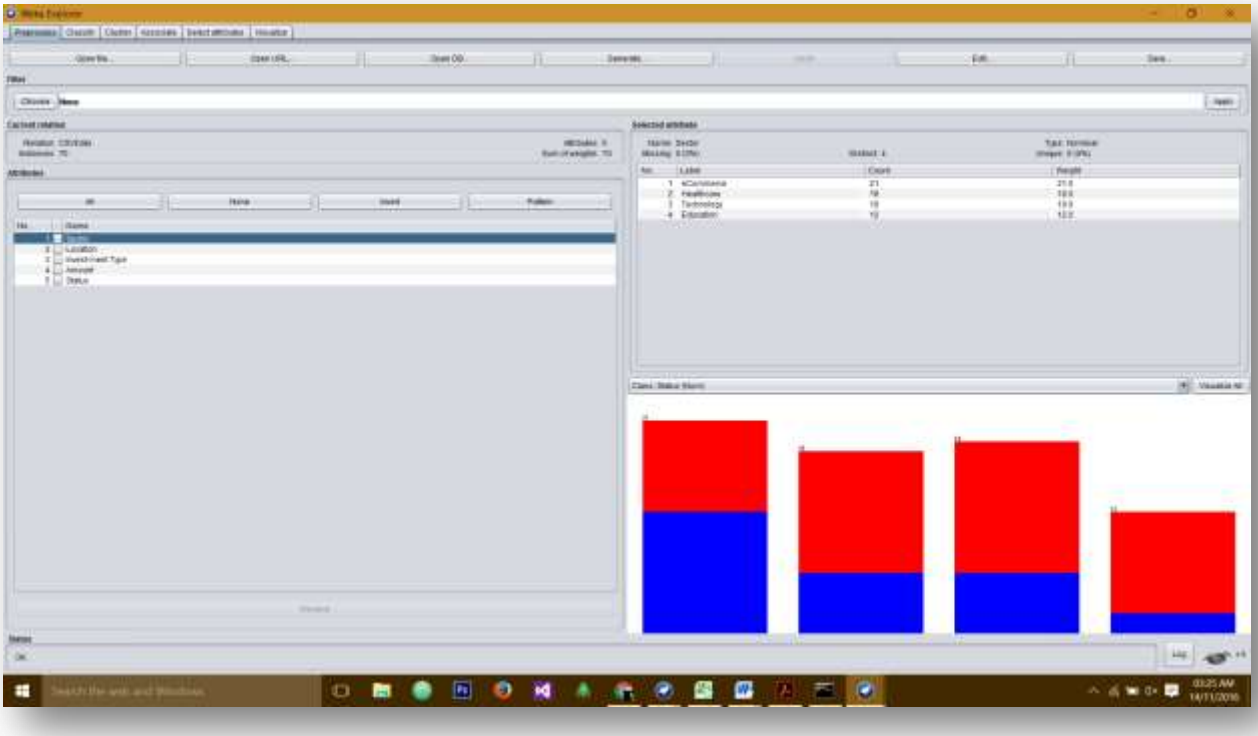
Rules that can be formed from the above decision tree are:

1. IF Sector=eCommerce And Investment-Type=Seed Funding Then Status=Success
2. IF Sector=eCommerce And Investment-Type=Private Equity And Amount>65000 Then Status=Success
3. IF Sector=eCommerce And Investment-Type=Private Equity And Amount<=65000 Then Status=Failure
4. IF Sector=HealthCare And Location=Mumbai Then Status=Success
5. IF Sector=HealthCare And Location=Bangalore Then Status=Failure
6. IF Sector=HealthCare And Location=New Delhi Then Status= Failure
7. IF Sector=HealthCare And Location=Gurgaon Then Status= Failure
8. IF Sector=HealthCare And Location=Chennai Then Status= Failure
9. If Sectoe=Technology And Location=Bangalore And Investment-Type=Seed Funding And Amount<=1400000 Then Status=Success
10. If Sectoe=Technology And Location=Bangalore And Investment-Type=Seed Funding And Amount>1400000 Then Status=Failure
    Etc.

5.2 Clustering Algorithm

Clustering algorithm is applied to set of similar data, to interpret data well. We created one set of attributes that is start-up sample data set.

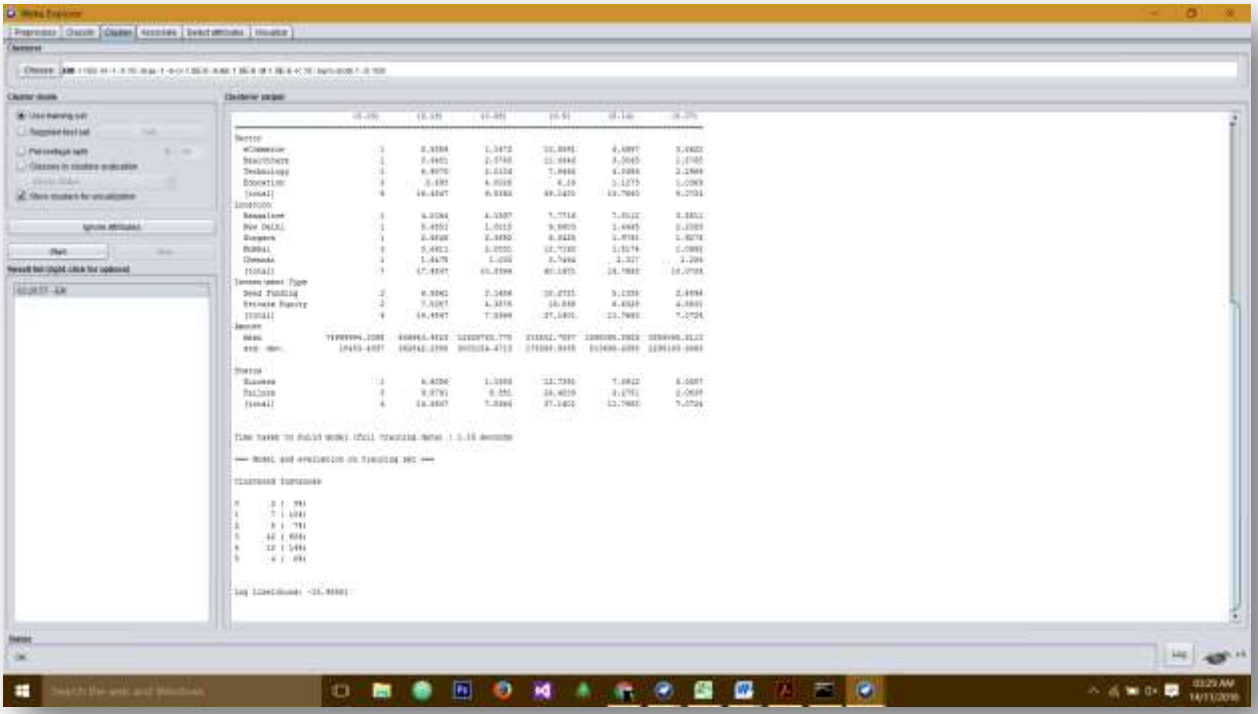
Numbers of distinct values for attributes are two, Success/Failure. Thus number of clusters used for both EM and K-means algorithm are two.



Basic Clustering Histogram for Start-up Data set Sectors

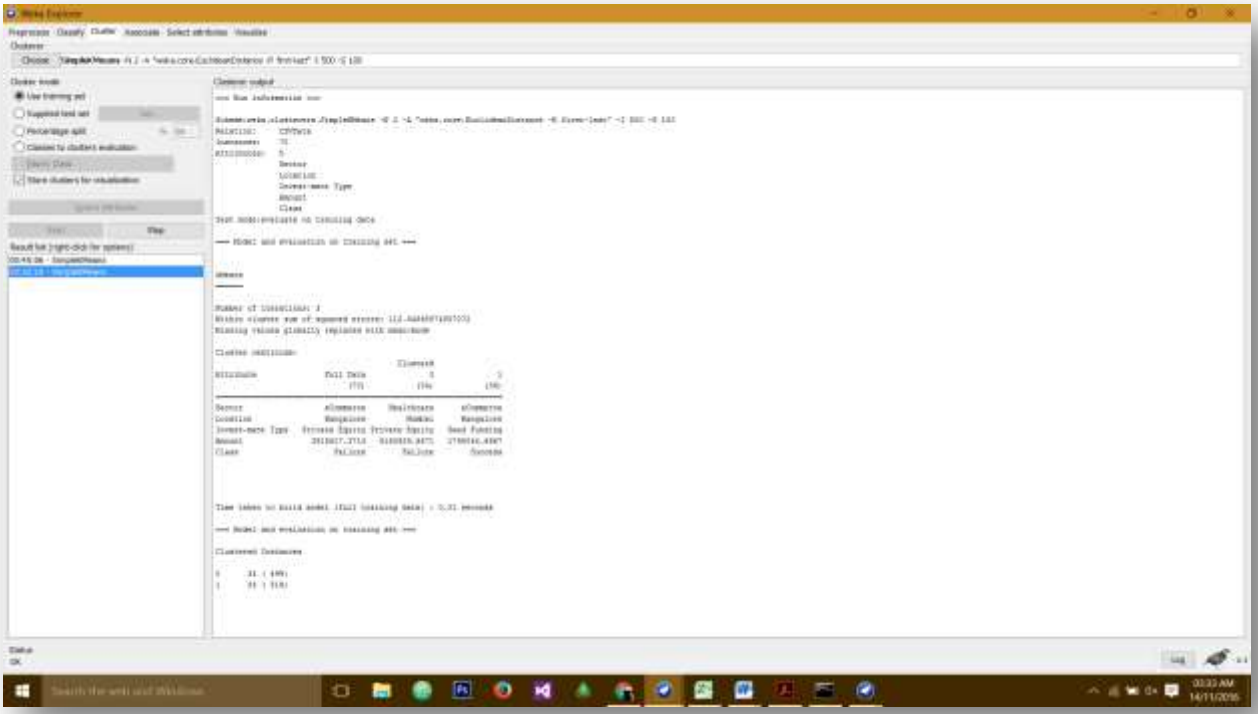
a. EM Algorithm

Thus from the below diagram, EM forms six clusters, the reason for six clusters might be based on various distinct values in dataset.



b. K-Mean Algorithm

Second clustering algorithm we used is simple K-Mean algorithm.



Simple K-Means applied for Start-up Dataset

By comparing both the clustering results for the data, we get nearly similar result with '~50%' instances in one cluser and '~50%' instances in another clusters.

Summary of Learning experience such as experiments and readings

- Learned Data Mining Tool such as WEKA
- Got better understanding of classification algorithm such as J48 and Logistic Algorithm
- Learned different clustering algorithms such as EM and K-Mean Algorithms.
- Read many articles to get clear idea of how to do data mining.

References

- WEKA Tutorial: https://www.youtube.com/watch?v=vlhUNeoNDR4
- Data Source: https://data.gov.in/startups-in-india/data-2016