Group 12 Assignment Report

# Team Members:

|  |  |
| --- | --- |
| Shraddha Asthana | 2020CT04541 |
| Vamsi Meenavilli | 2020CT04508 |
| Harshit Gurav | 2020CT04539 |
| Ketan Mamtora | 2020CT04524 |

# Upload the dataset

Uploaded the csv file as data frame using the read\_csv method in pandas library.

# Perform Exploratory Data Analysis & Visualize to study

## Dataset statistics

|  |  |
| --- | --- |
| Number of variables | 22 |
| Number of observations | 100284 |
| Missing cells | 1154023 |
| Missing cells (%) | 52.3% |
| Duplicate rows | 37 |
| Duplicate rows (%) | < 0.1% |
| Total size in memory | 87.6 MiB |
| Average record size in memory | 915.6 B |

## Variables statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Distinct Values** | **Distinct %** | **Missing Values** | **Missing %** | **Min** | **Max** |
| Patient Number | Numerical | 99791 | 99.9% | 491 | 0.5% | 1 | 102610 |
| State Patient Number | Categorical | 18518 | 79% | 77031 | 76.8% | - | - |
| Date Announced | Categorical | 113 | 0.1% | 0 | 0% | 1/4/2020 | 9/6/2020 |
| Estimated Onset Date | - | - | - | 100,284 | 100% | - | - |
| Age Bracket | Numerical | - | - | 62443 | 62% | 0 | 99 |
| Gender | Categorical | 4 | - | 59736 | 59.6% | - | - |
| Detected City | Categorical | 1801 | 22.9% | 92,434 | 92.2% | - | - |
| Detected District | Categorical | 755 | 0.8% | 7,885 | 7.9% | - | - |
| Detected State | Categorical | 36 | <0.1% | 9 | <0.1% | - | - |
| State code | Categorical | 36 | <0.1% | 9 | <0.1% | - | - |
| Current Status | Categorical | 5 | <0.1% | 2 | <0.1% | - | - |
| Notes | Text | 2581 | 6.7% | 61894 | 61.7% | - | - |
| Contracted from which Patient (Suspected) | Text | 303 | 15.8% | 98,361 | 98.1% | - | - |
| Nationality | Categorical | 12 | 0.8% | 98730 | 98.5% | - | - |
| Type of transmission | Categorical | 5 | 0.2% | 97,294 | 97% | - | - |
| Status Change Date | Date | 55 | 0.2% | 72,504 | 72.3% | - | - |
| Source\_1 | Categorical | 3769 | 3.8% | 1824 | 1.8% | - | - |
| Source\_2 | Categorical | 638 | 12.5% | 95192 | 94.9% | - | - |
| Source\_3 | Categorical | 123 | 24.8% | 99,789 | 99.5% | - | - |
| Backup Notes | Categorical | 222 | 61.5% | 99923 | 99.6% | - | - |
| Num Cases | Real Number | 671 | 0.7% | 3 | <0.1% | -5071 | 7358 |
| Entry\_ID | Real Number | 72048 | 99.9% | 28,185 | 28.1% | 1 | 73240 |

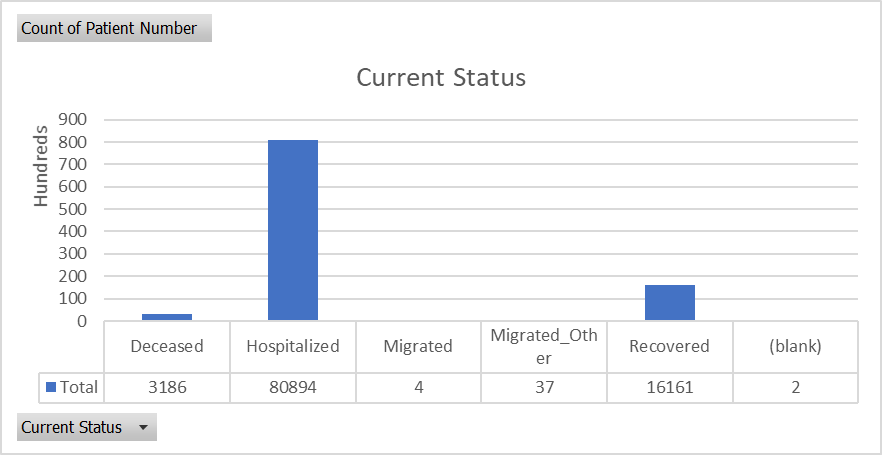
## Visualization

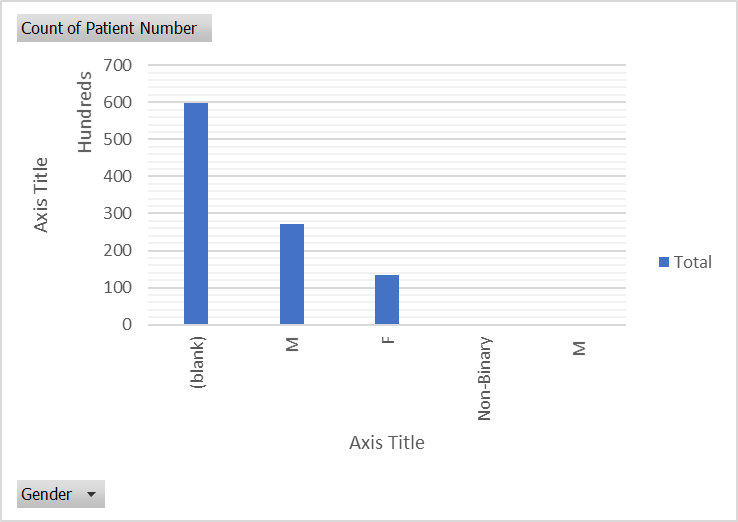
Used pivot tables in excel and pandas profiling method in Profile Report library for visualization of data

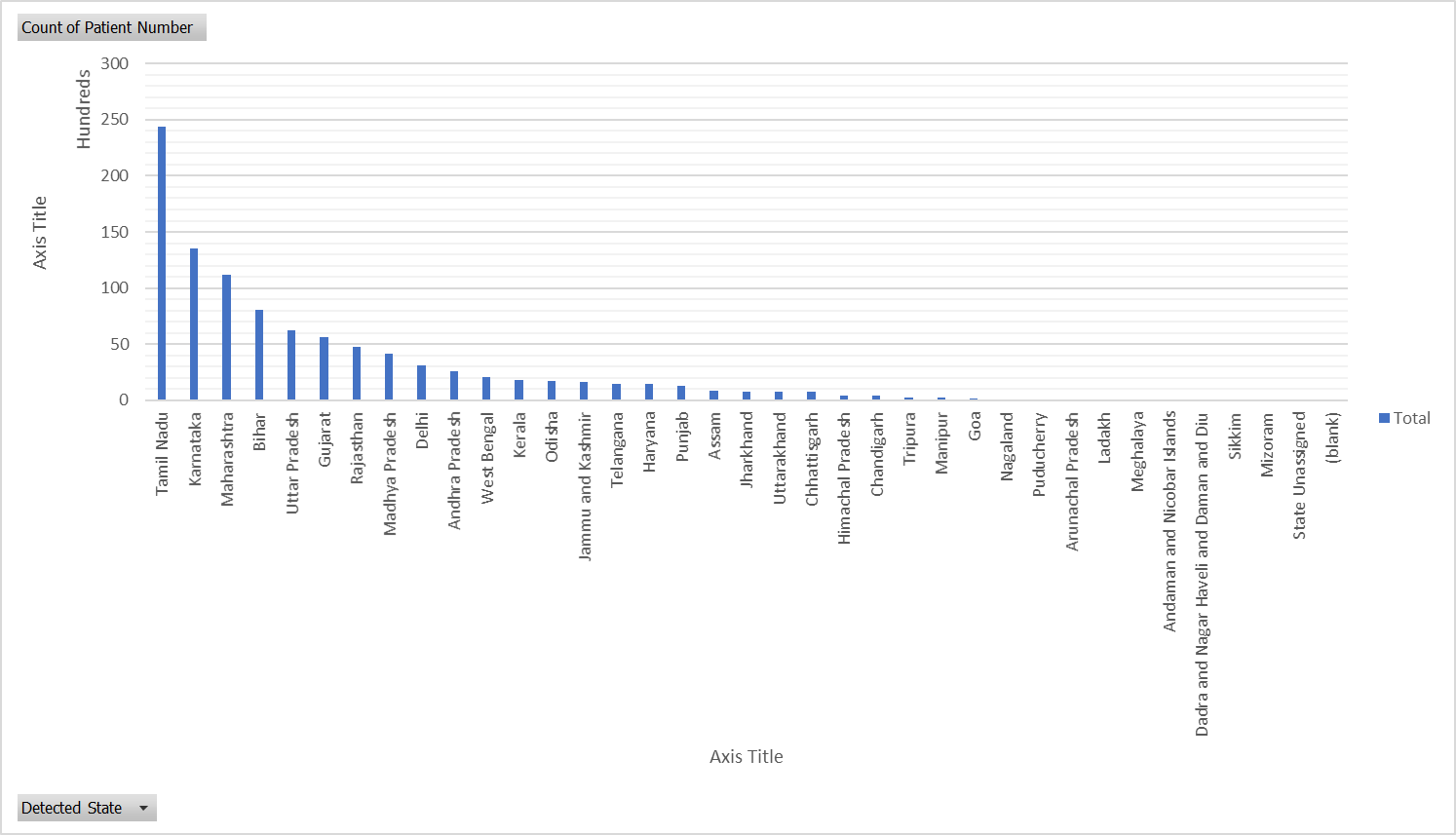
# Perform necessary pre-processing on this dataset

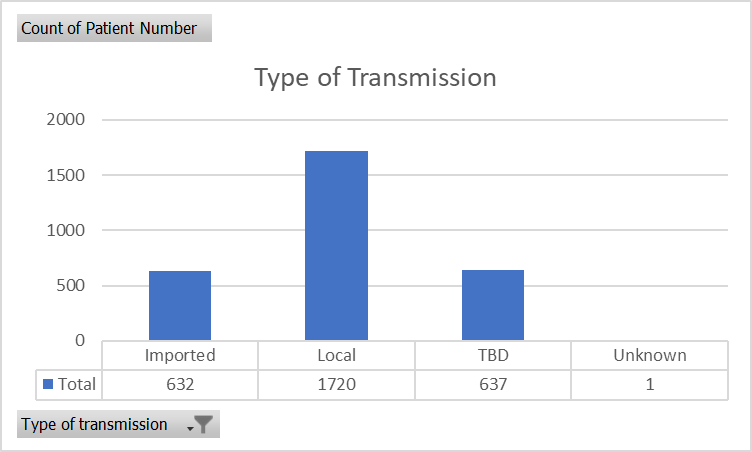
Following steps were performed to cleanup the data as part of the pre-processing

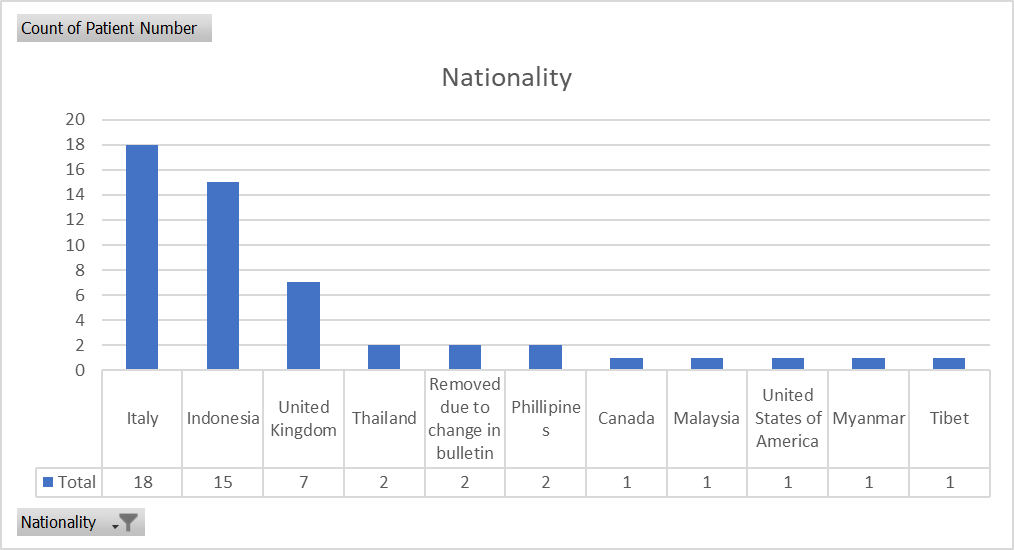
|  |  |  |
| --- | --- | --- |
| **Column** | **Observation** | **Pre-processing steps** |
| Patient Number | Very few missing values – Unique patient number | Populate the unique number automatically  *DONE* |
| State Patient Number | Derived column | Drop the column, not adding any value to data mining  *DONE* |
| Notes | Rows containing value “Correction” are mostly blank | Removed the rows containing value “Correction”  323 records removed  *DONE*  99983 records |
| Estimated Onset Date | 100% missing | Drop the column  *DONE* |
| Age Bracket | Number + Character | Replace 6 month with 0.6  ‘28-35’ – replaced with 31.5  NaN with 0  Change type to float |
| Gender | M – 27,164 (67%)  F – 13,376 (33%)  Blank – 59,300 | M<space> => M  TODO:  Blank – populate 2/3 male  And 1/3 as female  Non-binary – Others  *DONE* |
| Detected City | >90% missing | Drop the column  *DONE* |
| Detected State  State Code | High Cardinality | Only 9 records missing value –  Drop the rows with blank rows (verified – their State Code is also missing)  *DONE – 99974 records* |
| Current Status | |  |  | | --- | --- | | Deceased | 3186 | | Hospitalized | 80894 | | Migrated | 4 | | Migrated\_Other | 37 | | Recovered | 16161 | | (blank) | 2 | | 2 recrds blank are discarded row – so removing them  Migrated other is non-covid records – so removing them  *DONE - 99935 records* |
| Notes | Subset of “Contact” ~2726 records has covid spread information  Subset of data – “Travelled” ~5000 records – can be further analyzed for data mining | Cleanup the notes can keep only relevant as DataSet 2 and further analysis = “Contact”  Cleanup the notes – DataSet 3 – “Travelled” |
| Contracted from which Patient (Suspected) | >90% missing value | Drop the column  *DONE* |
| Nationality | >90% missing value  Only 51 records identified as Foreign National from 13 different country | Drop the column  *DONE* |
| Type of transmissions | >90% missing value  To be combined with Notes – Data Subset for effective use | Cleanup Imported<space>  *DONE* |
| Status Change Date | Mostly identical to “Date Announced” | Drop the column – not adding value  *DONE* |
| Source\_1 | Mostly of the values are URLs | Cleanup url and use Text values. E.g. <https://twitter>.com/handle => Twitter <Handle>  Update missing value as “Unknown”  *DONE* |
| Source\_2 | Mostly of the values are URLs | Cleanup url and use Text values. E.g. <https://twitter>.com/handle => Twitter <Handle>  Update missing value as Unknown |
| Source\_3 | >90% missing value | Drop the column  *DONE* |
| Backup Notes | Combine this field with Notes and extract derived column as per requirement of “Travel” and “Contact” | TODO: Derived column |
| Num Cases | Count of case column – Consider each row as one case | Drop the column  *DONE* |
| Entry\_ID | Unique value, however, Patient Number has less missing value | Use Patient Number as unique identifier, Drop this column  *DONE* |

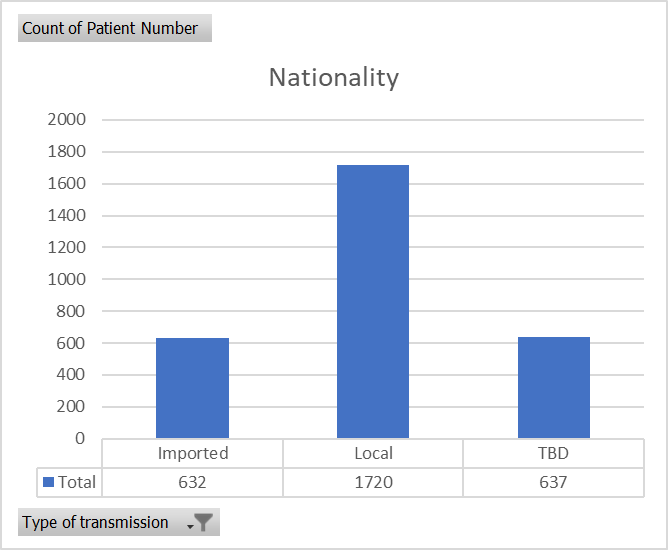












# Perform the data mining activity

## CLASSIFICATION

Justify your design choices by filling below:

### Requirement Design:

* Data set was found to be highly imbalanced as follows considering the target attribute as “Current Status”
* Explore Decision Tree and Random Forest classifier to generate the classification tree
* SMOTE and SMOTEENN synthetic oversampling was applied for over sampling to balance the minority dataset
* Considered only 3 features - gender, age and state for classification for simplicity

### Justification for Classification:

* With synthetic oversampling data, Decision Tree classification provided improved precision, recall and f1 score
* Random Forest Classification algorithm gave very similar results for the selected performance metric (precision, recall, and f1 score)

### Target Attribute

* Used “Current Status” as target attribute outcome of the classification
* Considered only two outcomes as Hospitalization and Deceased
* Used Label encoder to map value to 0 and 1

|  |  |  |
| --- | --- | --- |
| **Current Status** | **Label** | **Count** |
| Hospitalized | 1 | 34,737 |
| Deceased | 0 | 338 |

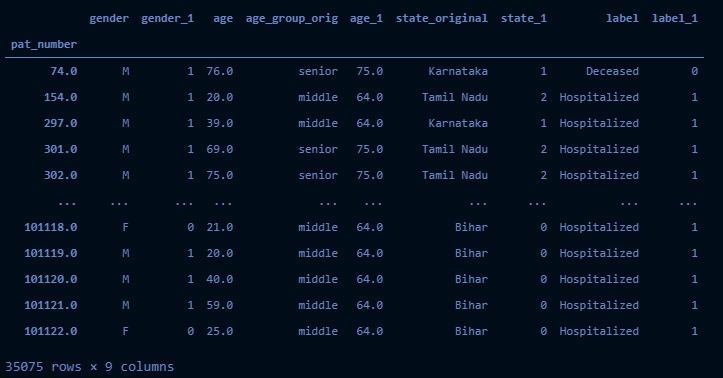
### Classification Type/ Model Selected with reason

* Decision Tree classification with gini index was chosen as the model. Based on the imbalanced data and use of SMOTE oversampling data provided better performance metrics and model accuracy

### Input Parameter settings

Used following attributes as features for the classification:

|  |  |  |  |
| --- | --- | --- | --- |
| **Target Feature** | **Data Set** | **Encoding (Label)** |  |
| Gender | M – Male  F – Female | 1 – Male  0 – Female |  |
| Age | <15 years  Between 15 and 65 years  >=65 years | 14  64  75 |  |
| State (selected only top 3 states with highest patient count) | Karnataka  Tamil Nadu  Bihar | 1  2  0 |  |



While doing encoding, we tried using Label Encoder and One Hot Encoder, the results with Label encoder provided meaningful decision tree.

### Approach or Results expected:

* Use the cleansed data set
* Further refine the target attribute and input parameters selected
* Use SMOTE oversampling technique
* Use 80% data for training and 20% data for Testing
* Create Decision Tree using gini classifier and maximum depth as 3
* Trian the model using training data set
* Predict the outcome/label for the test data set
* Calculate and verify the metrics selected – Confusion matrix, precision, recall and f1 score
* Create an image for the decision tree represented by the model

### Metric chosen for analysis:

Confusion matrix, precision, recall and f1 score are chosen for data analysis

## ASSOCIATION ANALYSIS

Justify your design choices by filling below:

### 1. Requirement Design:

To identify some common patterns in the covid dataset.

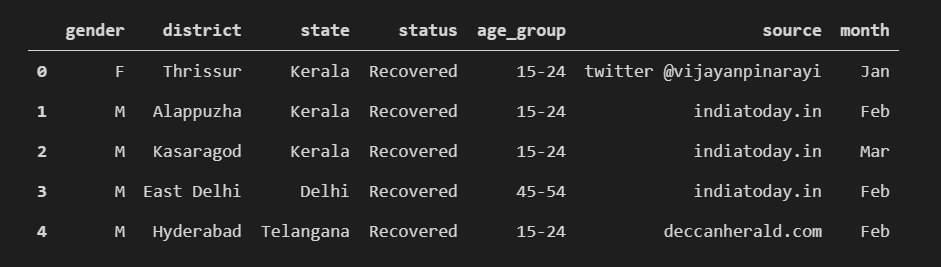
### 2. Justification for Association Analysis:

To identify the cases spike pattern in a month, place, which age group or gender

where most likely to survive, etc.

### 3. Attributes of interest:

gender, district, state, status, age\_group, source, month



### 4. Association Type/ Model Selected with reason:

Since all the data is categorical using apriori algorithm to find the association rules.

### 5. Input Parameter settings:

Clean the data set replace all the null values and convert all the columns to string data type.

### 6. Approach or Results expected:

Dataset is having a total of 113 days records. To identify 20 common patterns happening a day.

### 7. Metric chosen for analysis:\_

Minimum\_suppport = (patterns per day \* no. of days)/no.of rows

= (20 \* 113)/99935 = 0.022

Chosen confidence as 20% , so minimum\_confidence = 0.2

Minimum\_lift = 2

## CLUSTERING

Justify your design choices by filling below:

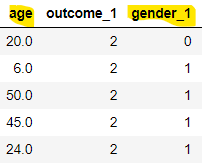
### 1. Requirement Design:

* Dataset consisted of categorical data, converted into numerical data and used the updated dataset for clustering
* Used K-Means algorithm for clustering
* Standardization, normalization and PCA helped before providing to K-means clustering algorithm

### 2. Justification for clustering:

* Used Standardization which standardizes a feature by subtracting the mean and then scaling to unit variance and used Normalization which scales each input variable separately to the range 0-1, which is the range for floating-point values where we have the most precision.
* Principal Component Analysis, or PCA, is a dimensionality-reduction method was used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
* Forming the clusters based on columns “age” and “gender”

### 3. Attributes of interest:

* Used “Age” and “gender\_1” as attributes of interest
* 

### 4. Clustering Type/ Model Selected with reason:

* K- Means clustering is chosen by division of the dataset into groups in which the members in the same group possess similarities in features.

### 5. Input Parameter settings:

* Before applying the K-means algorithm, clean the data by removing NAN values.
* Used Standardization, Normalization and PCA for scaling, normalizing and reducing the dimensions of data respectively.

### 6. Approach or Results expected:

* Numerical dataset was used after removing NAN values.
* Used Standardization, Normalization and PCA techniques
* Applied K-means algorithm on age and gender\_1 column
* Used elbow method for finding optimal value of clusters
* After applying the K-means clustering, we found that 2 clusters representing male and female

### 7. Metric chosen for analysis:

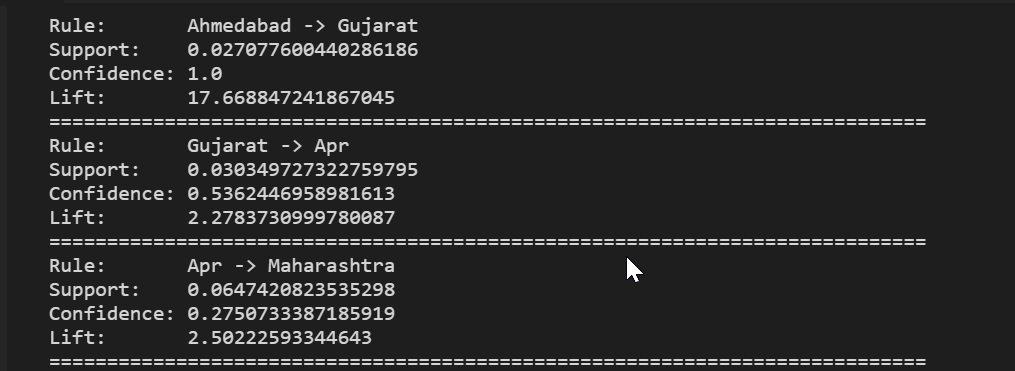
* Elbow method was used for getting the optimal K value or an optimal number of clusters
* silhouette\_score was used for clustering validation

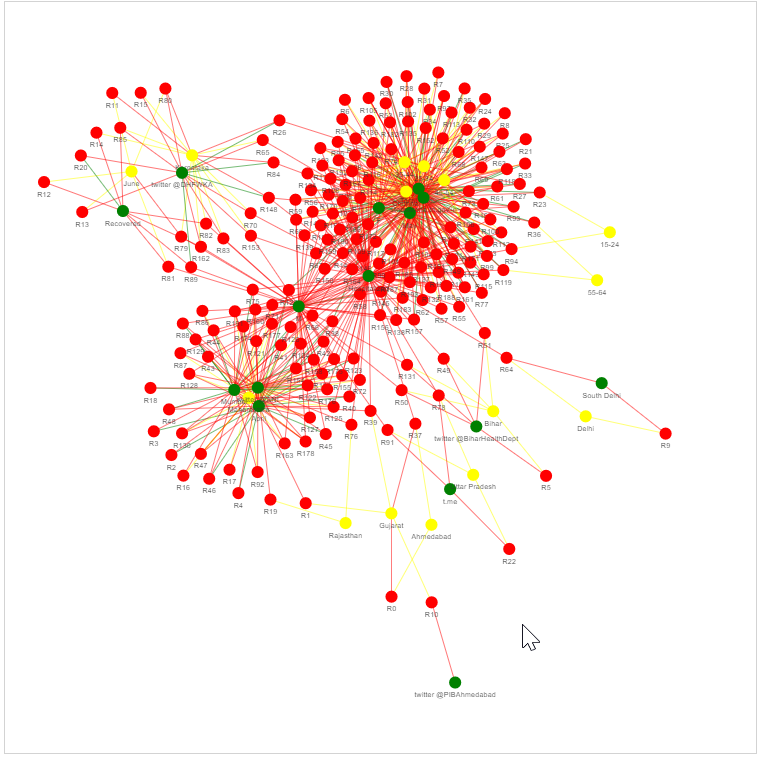
# Visualize to interpret the observations

## Association Rules:

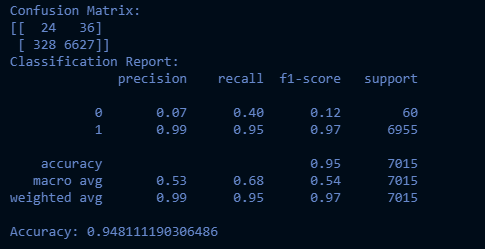
Created a network graph to show the relations between the attributes and also printed the

Rules as plain text with support, confidence and lift values.

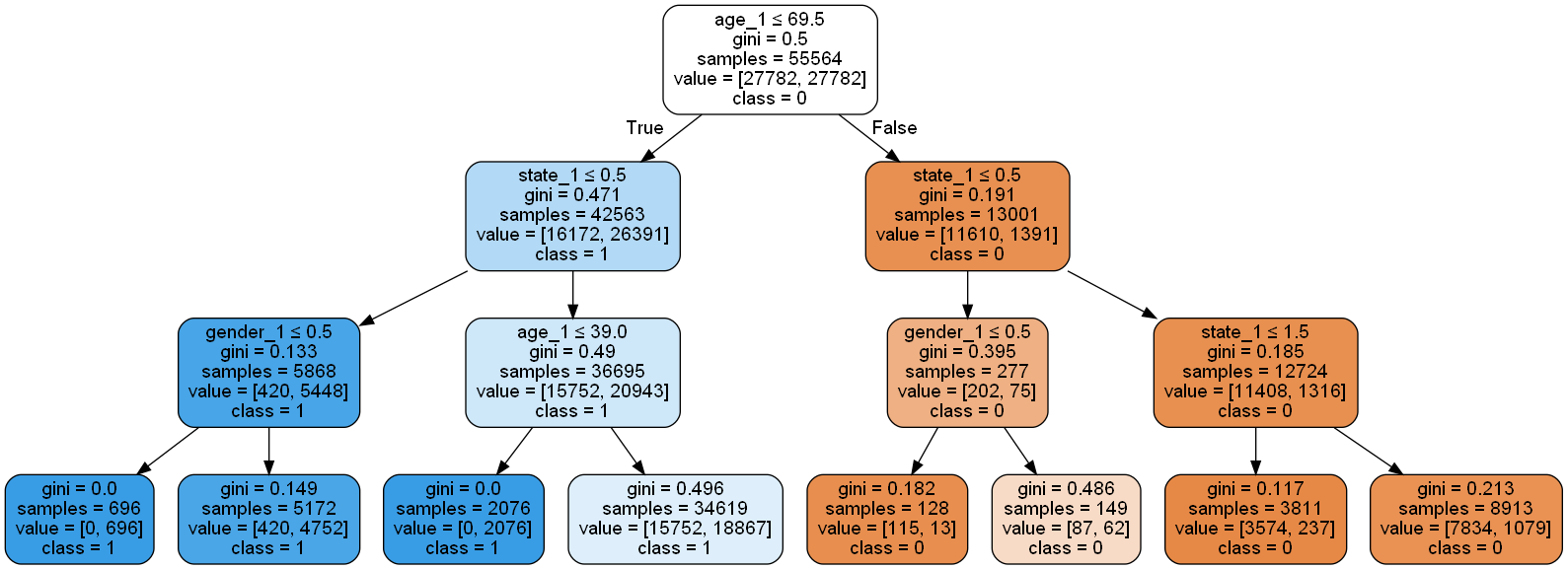




## Classification:



F1 score is low for minority class, however, this was best among all other classification combinations.

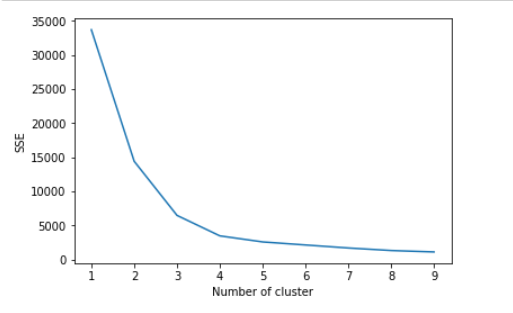


Analysis:

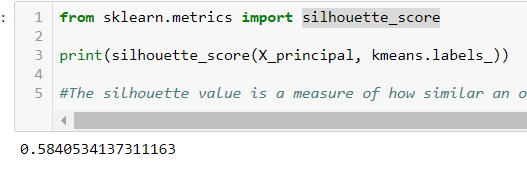
* Root node has age attribute with highest gini score of 0.5
* State attribute has second highest gini score and thus child of the root node
* Gender attributes has lowest gini score and is appearing at depth three
* Label class – 0 and 1 are leaf nodes at the depth four
* The classifier is not very accurate due to imbalanced data set, use of synthetic data has improved the accuracy of the model, however the decision tree is incorrectly classifying the minority data class

## Clustering:

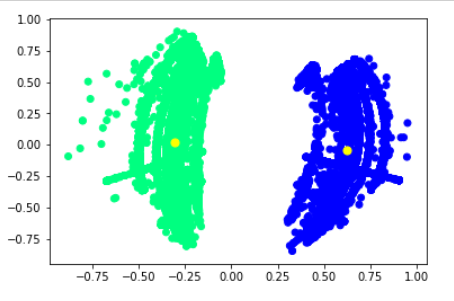
* Elbow method for finding clusters: Optimal clusters were 2



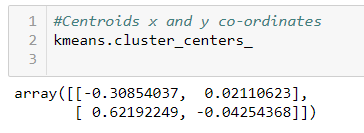
* silhouette score value



* Visualizing the clusters



* Centroid's co-ordinates



# OBSERVATION SECTION

Two distinct learnings by individual student specifying the name of the student

## Shraddha Asthana

### Learning 1: Application of various pre-processing and cleaning methods to fetch significant columns in the dataset which contain sound values in terms of quality and quantity, deducing a criterion to drop the rest of the insignificant columns, and determining suitable values to replace the Null or NaN values with in a necessary column

### Learning 2: Data may need to be converted into suitable groups or formats for better results as and when required, and newer and more useful columns can be composed from the given data upon reviewing it

## Vamsi Meenavilli

### Learning 1:

Data preprocessing is the key time consuming factor.

The cleaner the data the easier it becomes to model the data.

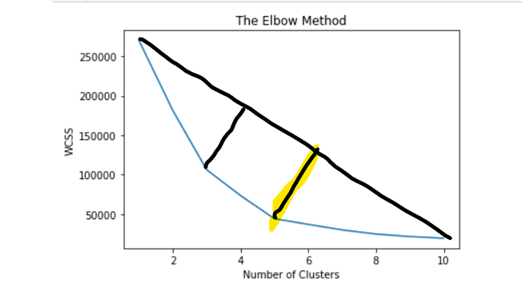
### Learning 2:

Learned about network graphs to visualize the relationship between the attributes.

## Harshit Gurav

### Learning 1:

For finding the value of K, elbow method was used. When we observe the elbow method graph, we can clearly see that when x - axis is on 5, graph clearly has an elbow. My finding was that if we draw a straight line from point 1 to 10, and if we calculate the distance from each point to this line, point with the largest distance should be the point that contains the elbow.



### Learning 2:

During clustering, for our dataset it was required to apply standardization, normalization and PCA algorithm for scaling, normalizing and dimension reduction of dataset respectively before applying clustering algorithm. Prior to applying above mentioned methods, clusters were not accurately formed

## Ketan Mamtora

### Learning 1:

During classification using Decision Tree method, it was found that sklearn library does not support categorical data as it is converts values to float during data mining activity. Upon further digging it was found that Categorical data needs to be converted to either of the following different formats (Numerical Encoding, One Hot Encoding, Binary Encoding). Each of these encoding format has its impact on accuracy and performance of the classification model training and validation.

<https://medium.com/data-design/visiting-categorical-features-and-encoding-in-decision-trees-53400fa65931>

### Learning 2:

Highly imbalanced data set along with Categorical attributes are very challenging in terms of classification. With One Hot Encoder, the decision trees (images) had “meaningless” branches and leaf nodes, e.g. age group – senior node appears towards to the root and age group – middle appeared as child nodes/lead nodes. Converting he Categorical data into Label (numerical) generated meaningful decision trees, however, as per theory, Decision Tree algorithm treats numerical data with order, which leads to incorrect calculations at times.

[How to handle Imbalanced Classification Problems | by Mukesh Chaudhary | Medium](https://medium.com/@cmukesh8688/how-to-handle-imbalanced-classification-problems-4a96f42ae4c4)