Classification Document

Vivek I(IMT2013017)
Aravind K(IMT2013019)
Nikhilesh P(IMT2013031)
Arun Prasad(IMT2013037)

Data set : FBI crime data set of California and other states.

Data preperation : The basic data preperation i.e. removing the faulty data rows, excluding extreme values while calculating averages etc, done in the descriptive analytics session is carried in this classification also. Apart from this, an attribute “is_safe” is added to the data which is described as follows,

Calculating the is_safe parameter :

The attributes 'violent crimes' and property crimes' are given weightage based on their impact of destruction i.e. 75% weightage is given to violent crimes and 25% weightage is given to property crimes. Using these weightages, we determine a value for each city which indicates wether it is safe or not and it is given by,

safety = 0.75 × (no.of.violent crimes / total population of that city) + 0.25 × (no.of property crimes / total population of that city)

After doing this operation, we get a value for each city and then we normalized these values and labeled the values greater than the average value as '1' and the values less than the average value as '0'. Which means that the the cities with 'is_safe' value as '1' are safe and the cities with 'is_safe' value as '0' are unsafe to live.

Now our aim is to classify the cities with 'is_safe' values based on the attributes present in the data set that is the predictors are the attributes Murder, Arson, Rape, Robbery, Aggrevated assault, Motor Vehicle theft, Violent crime, Burglary, Larency theft and property crime and the class variable is the 'is_safe' value of the respective cities.

To do this, we applies SVM, decision tree and naive bayesian algorithm for classification. Let 70% of the data as the training data and the rest of the 30% data as the testing data for these classification algorithms. Their results are given bellow;

SVM :

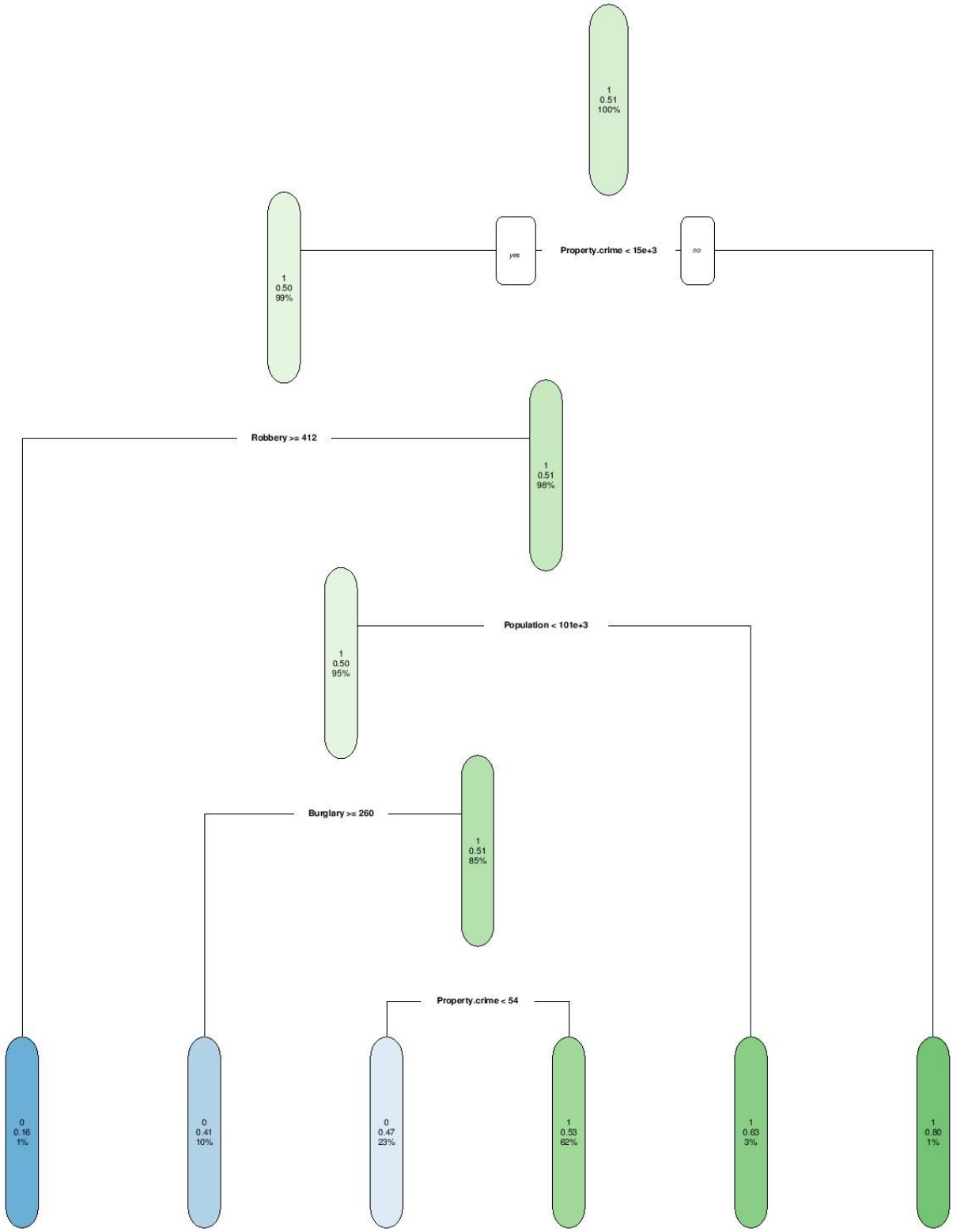
Confusion matrix :

```
            Reference
Prediction   0   1
         0 166 179
         1 105  92
```

```
Accuracy : 0.476
```

=> Accuracy is 47.6%

Decision tree :



Confusion matrix for the above decision tree :

```
          Reference
Prediction   0   1
         0  70  71
         1 220 226

      Accuracy : 0.5043
```

=>Accuracy is 50.43%

Naive Bayesian :

Confusion matrix :

```
          Reference
Prediction   0   1
         0 281 286
         1   9  11

      Accuracy : 0.4974
```

=>Accuracy = 49.74%