Chicago Crime Data Analysis

Group Number: 29

First Name	Last Name	Online Students? (Y or N)	Shared with ITMD 527? (Y or N)
Jeevith	Hebbaka Mallikarjunaiah	No	No
Phanindra	Chandraprakash	No	No

Table of Contents

1. Introduction and Motivations

Crimes are anti-social activities which causes society profoundly in innumerable ways. There had been humongous expansion in the crime in the recent past. It has become the most critical concern about the national security, especially after the 9/11 attacks at World Trade Center at New York city. However, the stupendous development in information and technology overwhelmingly bottlenecks the sufficient analysis of criminal and terrorist activities. Obstructing or controlling criminal activities has become a turbulence task. In the ever-ending race between law breakers and law enforcers, the job of the police officers or law enforcers in their roles to grab the criminals, these people must have to be ahead in the race. Data mining applied in this perspective gives fruitful results to assuage criminal activities. In this paper, we visualize the crime patterns and trends by using classification based models. Results from this analysis can be used to improve safety measures for the forth coming years.

2. Data Description

The data sets are taken from "City of Chicago – Data Portal". This data set reflects the criminal activities reported in the city of. This work will be on Crime Analysis domain and this data set has around 73000 records. This data set has 22 features

Data Source URL: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Data set has 22 features and they are listed below,

ID - Unique id for the record.
Case Number - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
Date –It will indicate when was the incident happened.
Block – It is kind of some abstract address.
IUCR - The Illinois Uniform Crime Reporting code.
Primary Type - The primary description of the IUCR code.
Description - The secondary description of the IUCR code, a subcategory of the primary description.
Location Description – It will say where the incident had happened.
Arrest – It will say whether arrest had made or not.
Domestic – It will Indicates whether the incident was domestic which is given by the Illinois Domestic Violence Act.
Beat - Indicates the beat where the incident occurred.
District - Indicates the police district where the incident occurred.
Ward - The ward where the incident occurred
Community Area - Indicates the community area where the incident occurred.
FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
X Coordinate - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Y Coordinate - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Year - Year the incident occurred.

Updated On - Date and time the record was last updated.
Latitude - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Longitude - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Location - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

3. Research Problems and Solutions

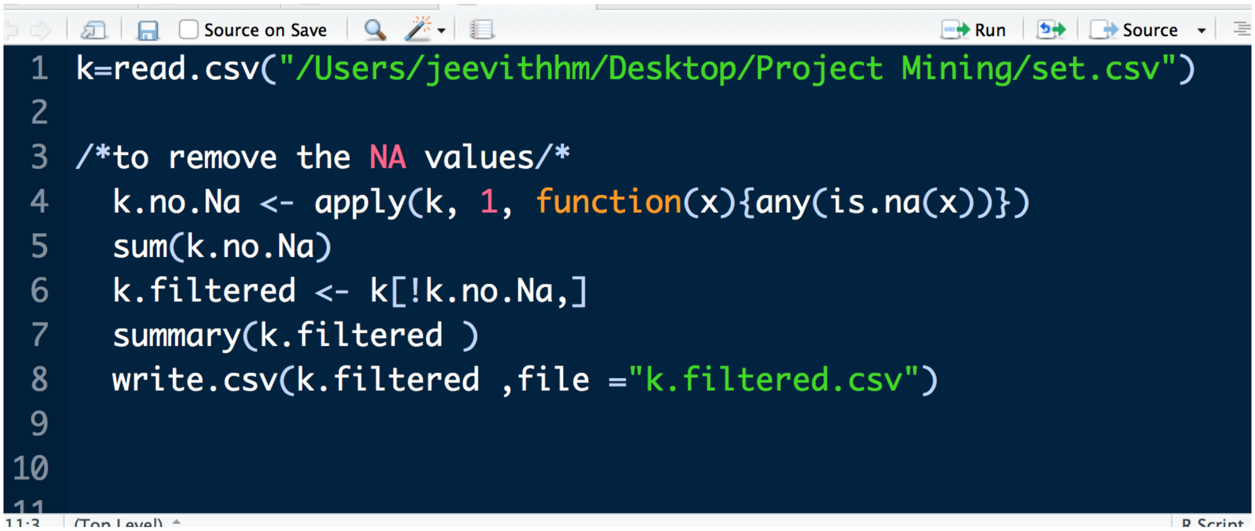
To perform classification task on Chicago Crime Data with different classification algorithms. To compare the classification algorithms based on its accuracy to predict the class label correctly and finalize a classification algorithm for classifying the Chicago City Crime Data.

4. KDD

4.1. Data Processing

a) Data cleaning (removing NA values)

We collected the data from City of Chicago – Data Portal, after getting the values we had performed the data cleaning process by eliminating the rows which contain the NA values (in R studio).



```
k=read.csv("/Users/jeevithhm/Desktop/Project Mining/set.csv")

/*to remove the NA values/*
  k.no.Na <- apply(k, 1, function(x){any(is.na(x))})
  sum(k.no.Na)
  k.filtered <- k[!k.no.Na,]
  summary(k.filtered )
  write.csv(k.filtered ,file ="k.filtered.csv")
```

b) Sampling of data sets:

After performing the data cleaning, we got 72859 entries, later we performed the sampling process in WEKA (@ 7% Resampling settings) so the data got reduced to 5100 entries, and then we split the dataset in to training and test data sets.

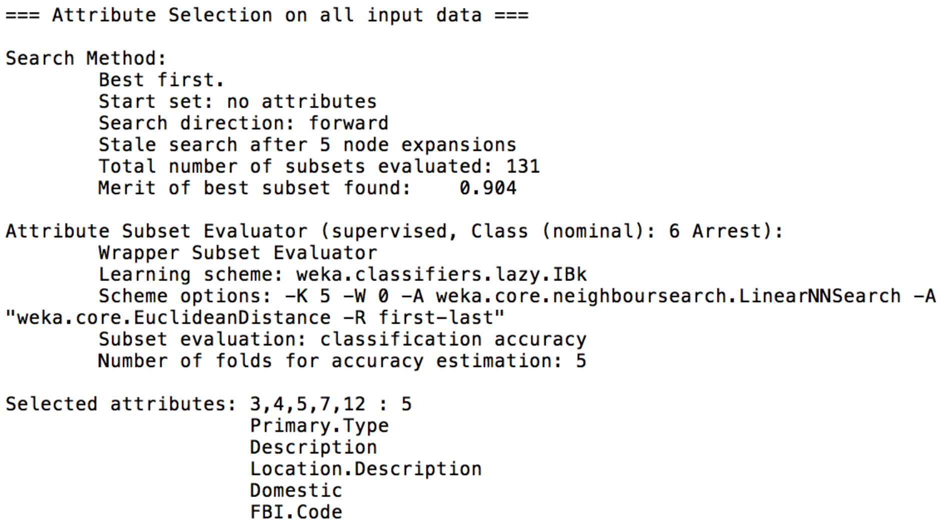
Training Set -66%-->3366 entries

Test Set-34%-->1145 entries

c) Feature Selection: After getting the training and test data we have performed the feature selection by WrapperSubsetEval method. In this method, we can specify the required classification algorithm.

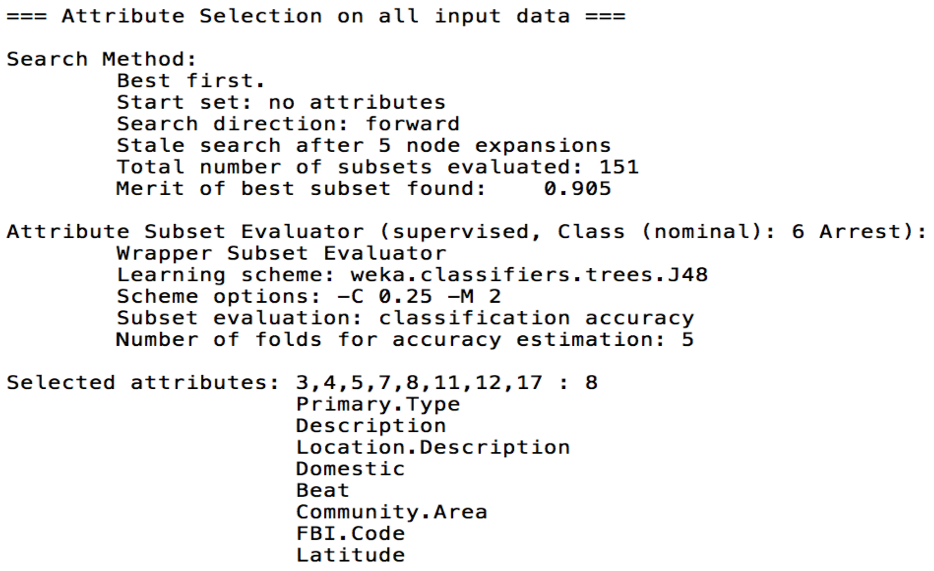
Feature Selection by considering class attribute as Arrest

1)Feature Selection for KNN



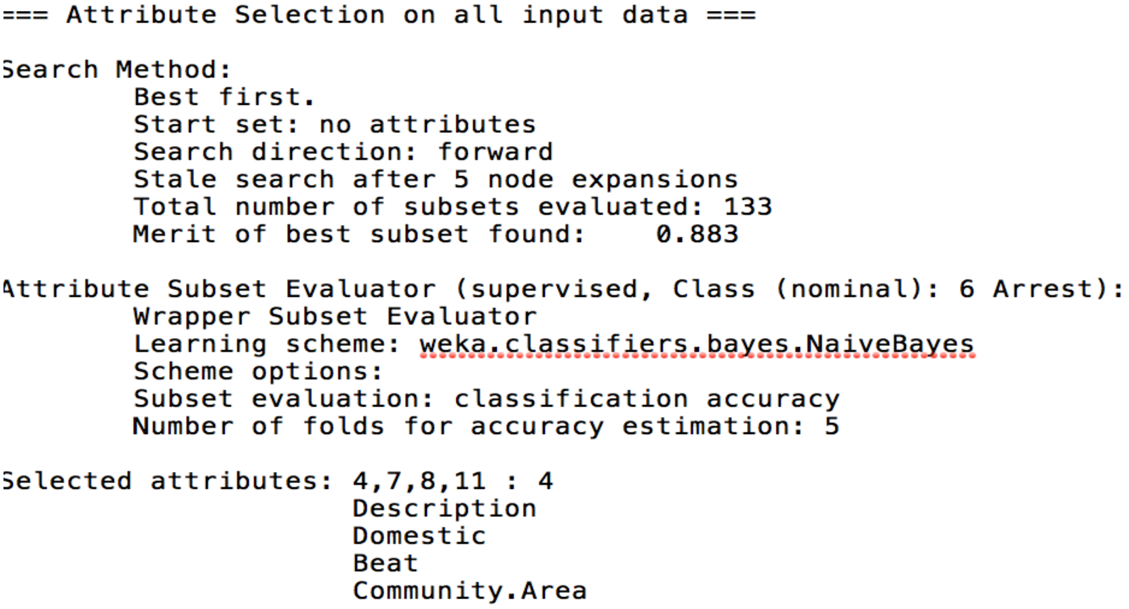
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 131
        Merit of best subset found:    0.904

Attribute Subset Evaluator (supervised, Class (nominal): 6 Arrest):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.lazy.IBk
        Scheme options: -K 5 -W 0 -A weka.core.neighboursearch.LinearNNSearch -A
"weka.core.EuclideanDistance -R first-last"
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 3,4,5,7,12 : 5
                     Primary.Type
                     Description
                     Location.Description
                     Domestic
                     FBI.Code
```

2) Feature Selection for J48



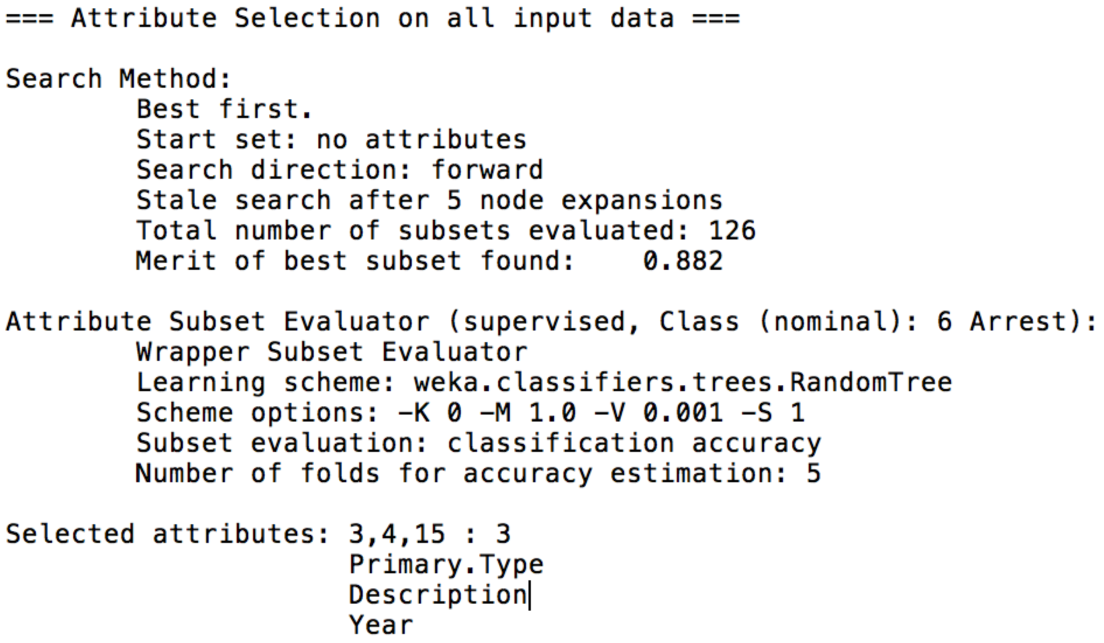
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 151
        Merit of best subset found:    0.905

Attribute Subset Evaluator (supervised, Class (nominal): 6 Arrest):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 3,4,5,7,8,11,12,17 : 8
                     Primary.Type
                     Description
                     Location.Description
                     Domestic
                     Beat
                     Community.Area
                     FBI.Code
                     Latitude
```

3) Feature Selection for Naïve Bayes



```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 133
        Merit of best subset found:    0.883

Attribute Subset Evaluator (supervised, Class (nominal): 6 Arrest):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.bayes.NaiveBayes
        Scheme options: 
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,7,8,11 : 4
                     Description
                     Domestic
                     Beat
                     Community.Area
```

4) Feature Selection for Random Tree



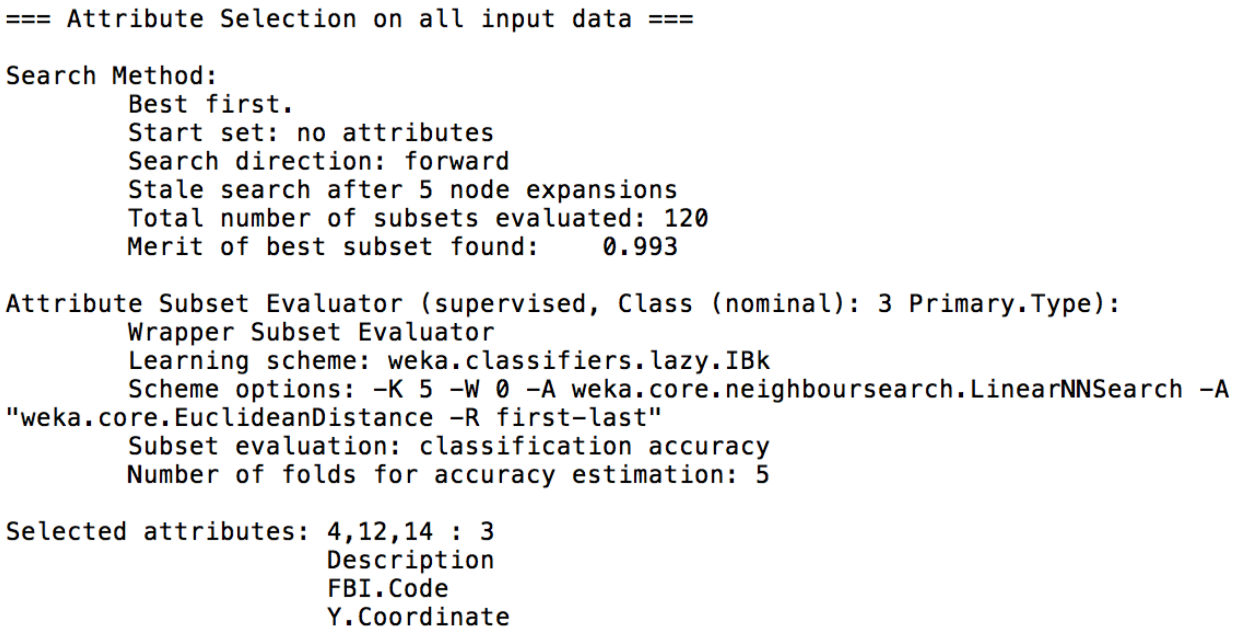
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 126
        Merit of best subset found:    0.882

Attribute Subset Evaluator (supervised, Class (nominal): 6 Arrest):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.RandomTree
        Scheme options: -K 0 -M 1.0 -V 0.001 -S 1
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 3,4,15 : 3
                     Primary.Type
                     Description
                     Year
```

5) Feature Selection for Decision Table

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 99
        Merit of best subset found:    0.876

Attribute Subset Evaluator (supervised, Class (nominal): 6 Arrest):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.rules.DecisionTable
        Scheme options: -X 1 -S weka.attributeSelection.BestFirst -D 1 -N 5
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4 : 1
                     Description
```

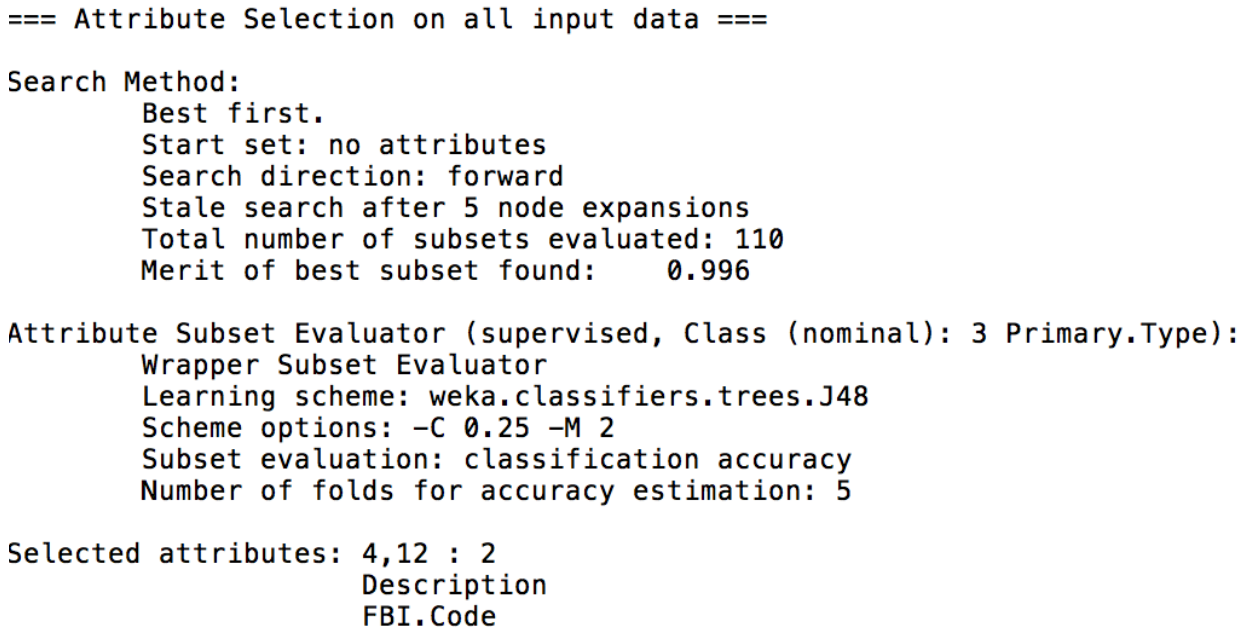
Feature Selection by considering class attribute as Primary. Type

1) Feature Selection for KNN



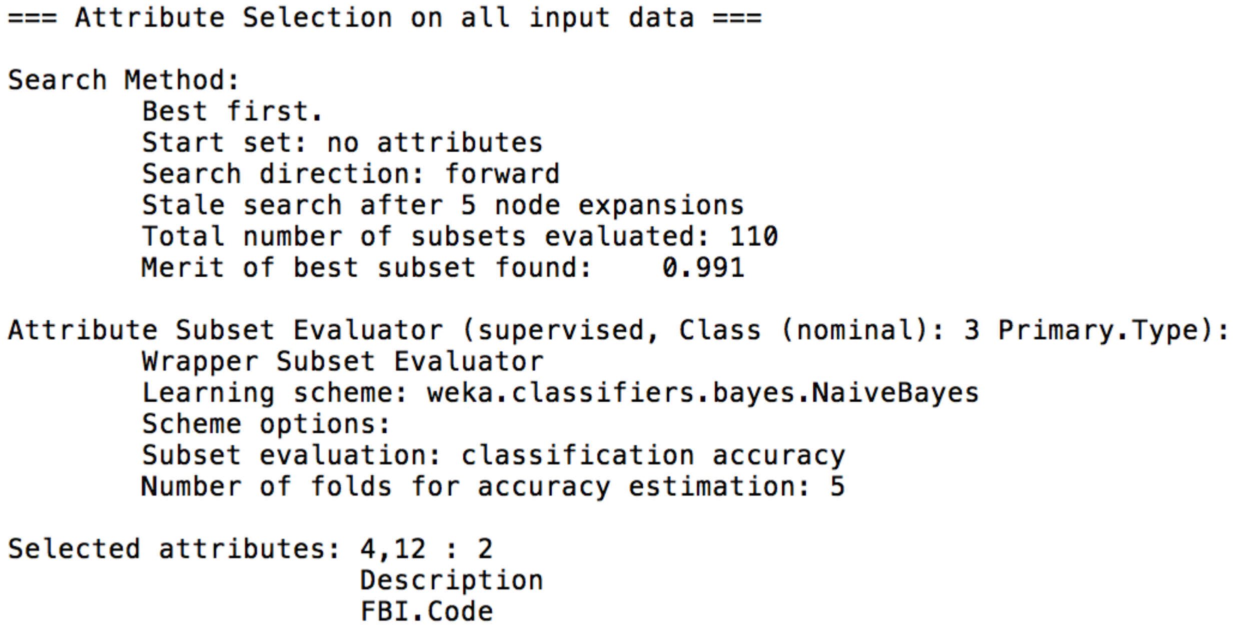
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 120
        Merit of best subset found:    0.993

Attribute Subset Evaluator (supervised, Class (nominal): 3 Primary.Type):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.lazy.IBk
        Scheme options: -K 5 -W 0 -A weka.core.neighboursearch.LinearNNSearch -A
"weka.core.EuclideanDistance -R first-last"
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,12,14 : 3
                     Description
                     FBI.Code
                     Y.Coordinate
```

2) Feature Selection for J48



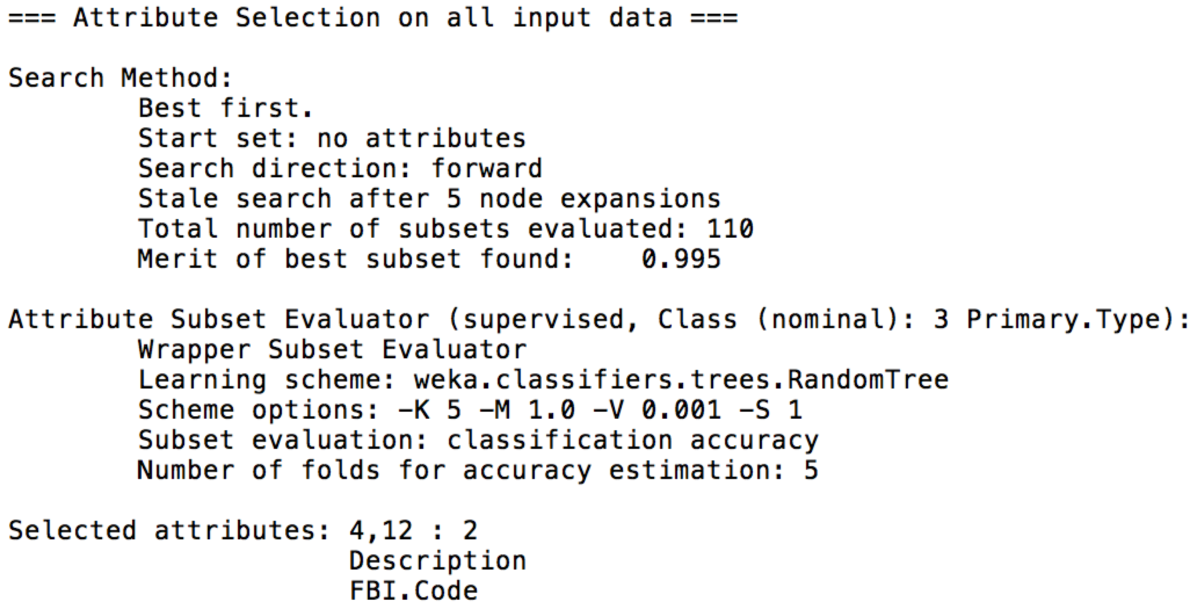
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 110
        Merit of best subset found:    0.996

Attribute Subset Evaluator (supervised, Class (nominal): 3 Primary.Type):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,12 : 2
                     Description
                     FBI.Code
```

3) Feature Selection for Naïve Bayes



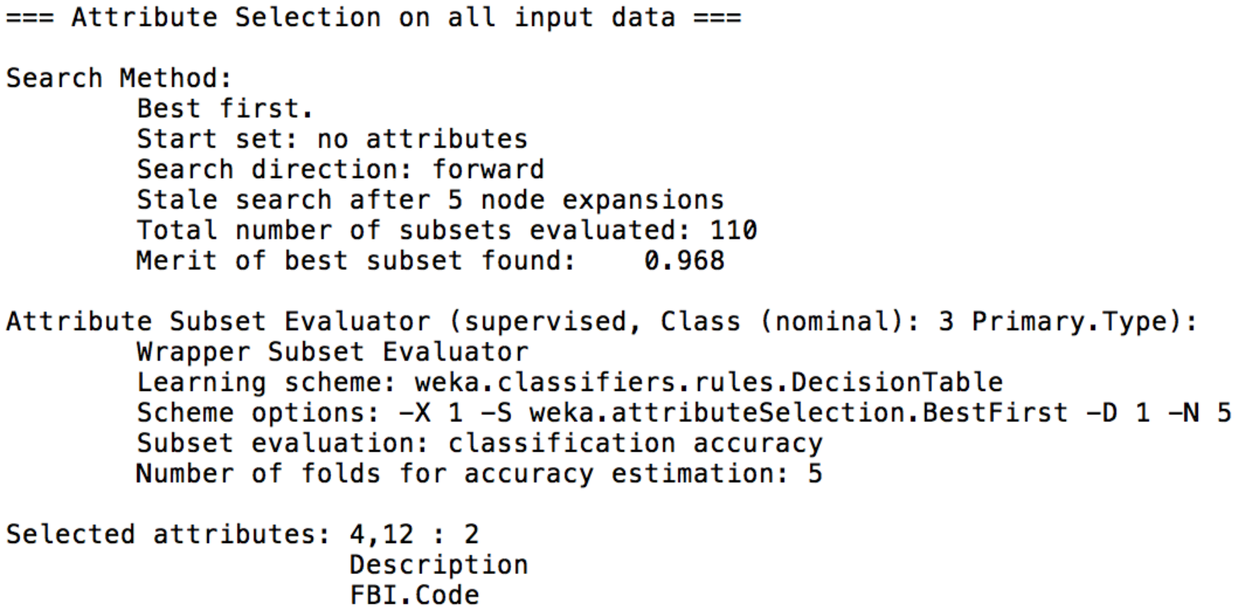
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 110
        Merit of best subset found:    0.991

Attribute Subset Evaluator (supervised, Class (nominal): 3 Primary.Type):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.bayes.NaiveBayes
        Scheme options: 
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,12 : 2
                     Description
                     FBI.Code
```

4) Feature Selection for Random Tree



```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 110
        Merit of best subset found:    0.995

Attribute Subset Evaluator (supervised, Class (nominal): 3 Primary.Type):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.RandomTree
        Scheme options: -K 5 -M 1.0 -V 0.001 -S 1
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,12 : 2
                     Description
                     FBI.Code
```

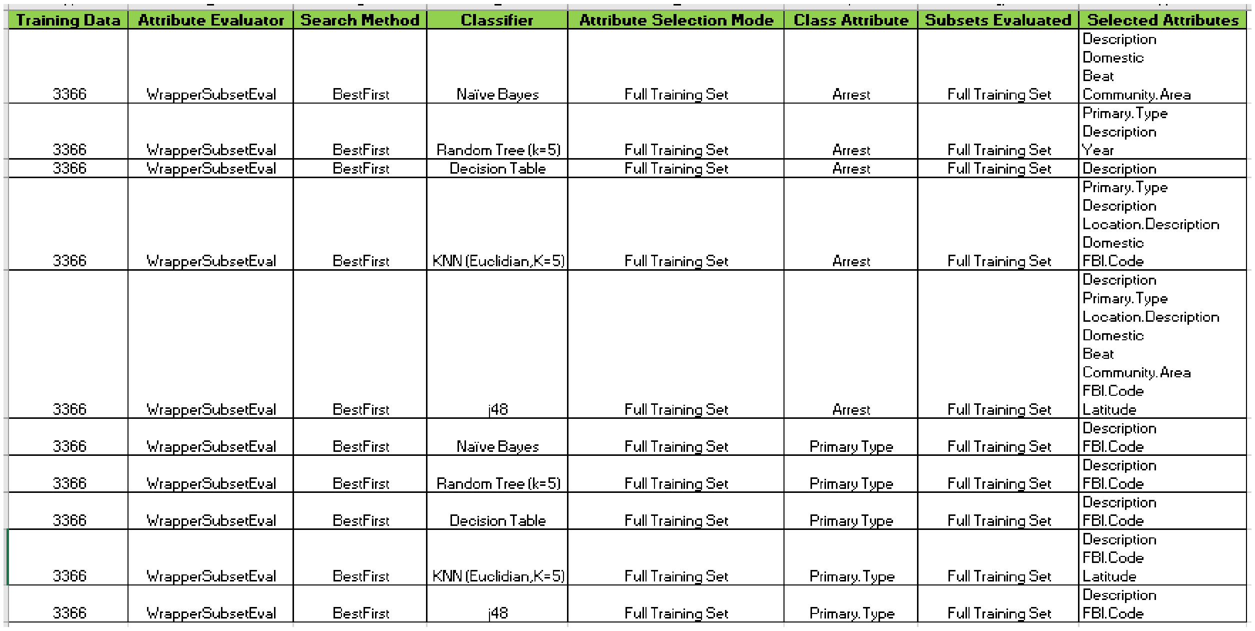
5) Feature Selection for Decision Table



```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 110
        Merit of best subset found:    0.968

Attribute Subset Evaluator (supervised, Class (nominal): 3 Primary.Type):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.rules.DecisionTable
        Scheme options: -X 1 -S weka.attributeSelection.BestFirst -D 1 -N 5
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 4,12 : 2
                     Description
                     FBI.Code
```

Feature Selection Overview

The below table contains all the details of feature selection process that was used during the analysis.



Training Data	Attribute Evaluator	Search Method	Classifier	Attribute Selection Mode	Class Attribute	Subsets Evaluated	Selected Attributes
3366	WrapperSubsetEval	BestFirst	Naïve Bayes	Full Training Set	Arrest	Full Training Set	Description Domestic Beat Community.Area
3366	WrapperSubsetEval	BestFirst	Random Tree (k=5)	Full Training Set	Arrest	Full Training Set	Primary.Type Description Year
3366	WrapperSubsetEval	BestFirst	Decision Table	Full Training Set	Arrest	Full Training Set	Description
3366	WrapperSubsetEval	BestFirst	KNN (Euclidian,K=5)	Full Training Set	Arrest	Full Training Set	Primary.Type Description Location.Description Domestic FBI.Code
3366	WrapperSubsetEval	BestFirst	j48	Full Training Set	Arrest	Full Training Set	Description Primary.Type Location.Description Domestic Beat Community.Area FBI.Code Latitude
3366	WrapperSubsetEval	BestFirst	Naïve Bayes	Full Training Set	Primary Type	Full Training Set	Description FBI.Code
3366	WrapperSubsetEval	BestFirst	Random Tree (k=5)	Full Training Set	Primary Type	Full Training Set	Description FBI.Code
3366	WrapperSubsetEval	BestFirst	Decision Table	Full Training Set	Primary Type	Full Training Set	Description FBI.Code
3366	WrapperSubsetEval	BestFirst	KNN (Euclidian,K=5)	Full Training Set	Primary.Type	Full Training Set	Description FBI.Code Latitude
3366	WrapperSubsetEval	BestFirst	j48	Full Training Set	Primary.Type	Full Training Set	Description FBI.Code

4.2. Data Mining Tasks and Processes

After performing the Data Preprocessing, the below classification algorithms were used for both training and test sets for classification task.

- KNN
- J48
- Naive Bayes
- Random Tree
- Decision Table

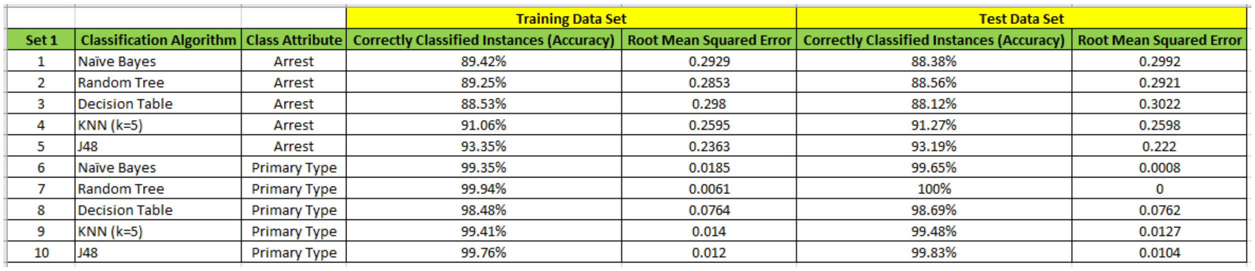
5. Evaluations and Results

5.1. Evaluation Methods

Splitting of data is done through sampling as we mentioned above in the data preprocessing stage. We have considered Correctly Classified Instances accuracy and Root mean squared error metrics to analyze and conclude the classifiers. If the accuracy is higher and Root mean squared error is lower then, we will consider that classification algorithm as the best.

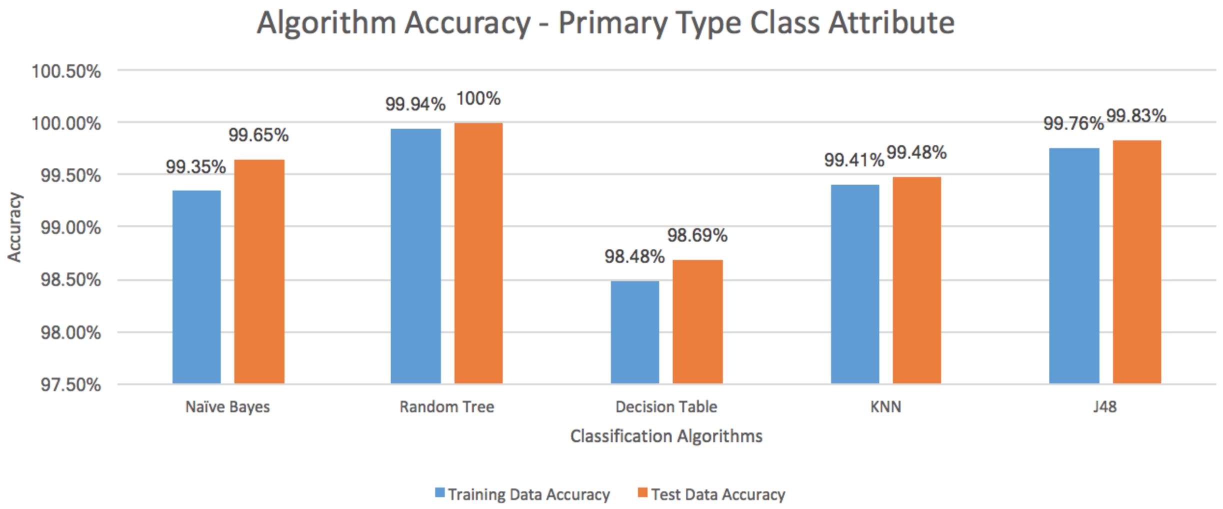
5.2. Results and Findings

The below table shows the result of different classification algorithms with accuracy and RMSE for both training and test data sets

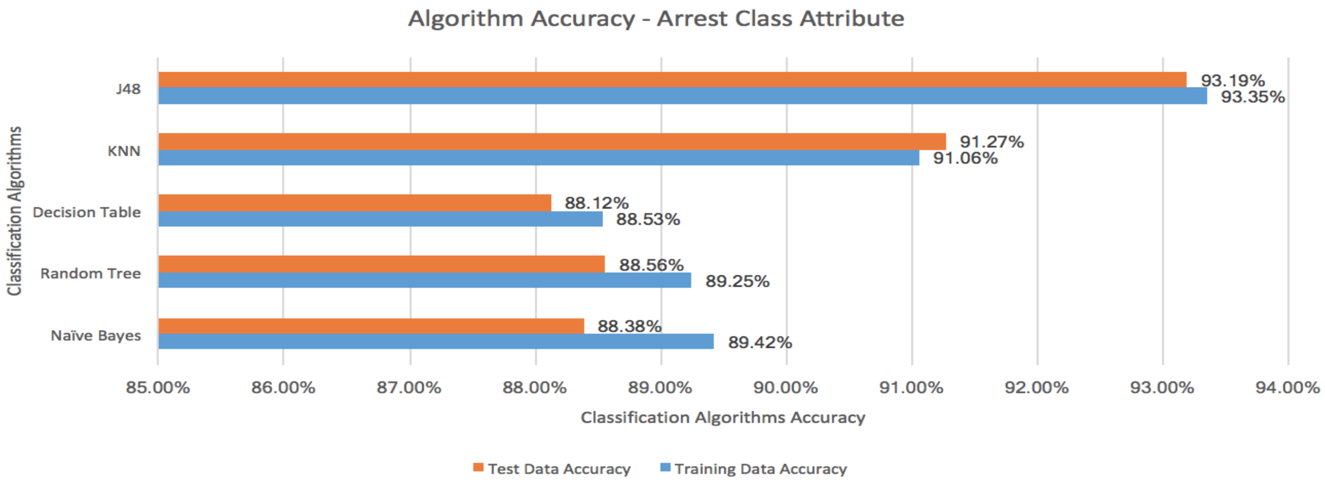


			Training Data Set		Test Data Set	
Set 1	Classification Algorithm	Class Attribute	Correctly Classified Instances (Accuracy)	Root Mean Squared Error	Correctly Classified Instances (Accuracy)	Root Mean Squared Error
1	Naïve Bayes	Arrest	89.42%	0.2929	88.38%	0.2992
2	Random Tree	Arrest	89.25%	0.2853	88.56%	0.2921
3	Decision Table	Arrest	88.53%	0.298	88.12%	0.3022
4	KNN (k=5)	Arrest	91.06%	0.2595	91.27%	0.2598
5	J48	Arrest	93.35%	0.2363	93.19%	0.222
6	Naïve Bayes	Primary Type	99.35%	0.0185	99.65%	0.0008
7	Random Tree	Primary Type	99.94%	0.0061	100%	0
8	Decision Table	Primary Type	98.48%	0.0764	98.69%	0.0762
9	KNN (k=5)	Primary Type	99.41%	0.014	99.48%	0.0127
10	J48	Primary Type	99.76%	0.012	99.83%	0.0104

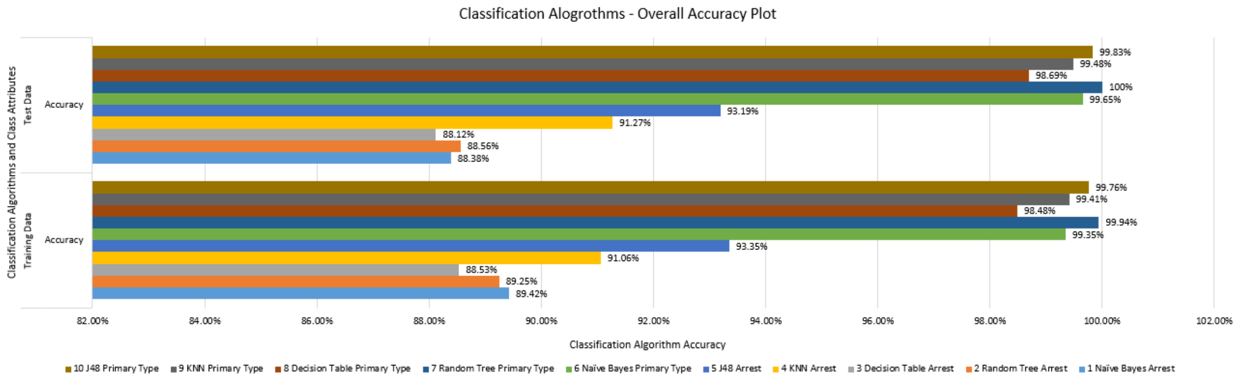
Accuracy of algorithms for Primary Type Class Attribute



Accuracy of algorithms for Arrest Class Attribute



Classification Algorithms Consolidated Output



We could not put the classification results in this report since the confusion matrix were too huge and it would distort the report structure. Hence, we request you to go through the Save Buffer files enclosed in Codes.zip file.

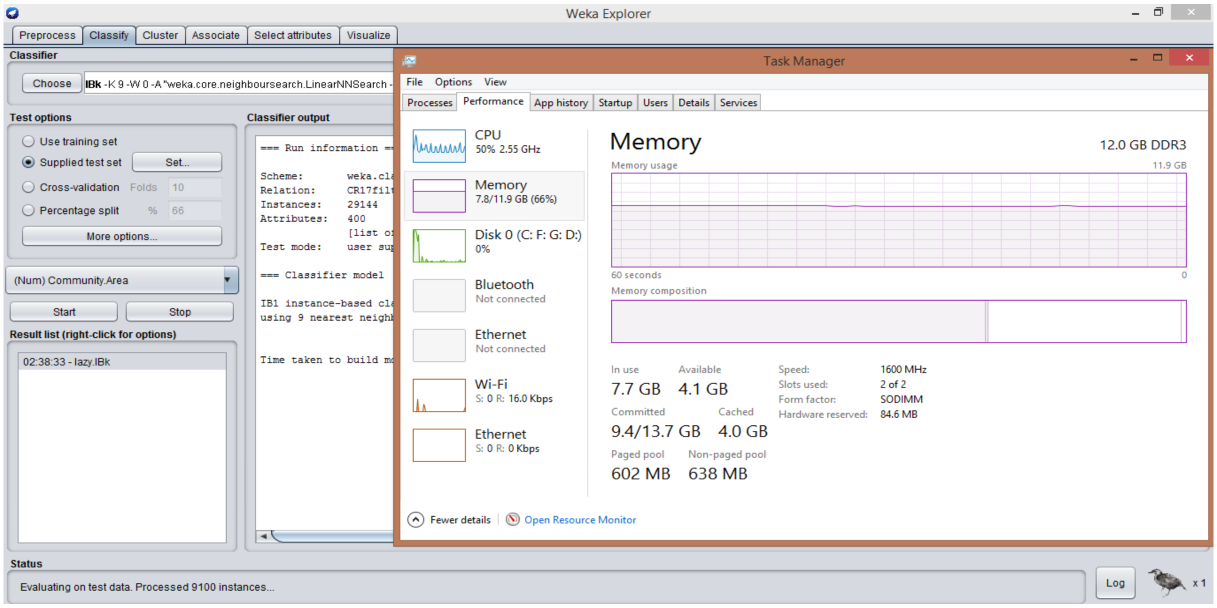
6. Conclusions and Future Work

6.1. Conclusions

Based on the analysis, we have chosen Random Tree classification algorithm to classify the Chicago City Crime Data, since accuracy was 100% and Root Mean Squared Error was zero.

6.2. Limitations

We have faced some challenges while performing this classification task, data set was too huge, initially we tried to consider the entire data set from 2001 to till date, and the records were around 68 lacs and we faced problems while preprocessing data in RStudio, we could not completely eliminate the NA or NULL values inside the data. Later we considered only 2017 data set and it came up to approximately 73000 records. We easily preprocessed this data set in RStudio to remove the NA or NULL values, but we could not perform the analysis for this data set in Weka, since we landed with heap memory and RAM issues. Please find the below images for more information.



Classifier

IBk -K 9 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -

Test options

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) Community.Area

Result list (right-click for options)

02:38:33 - lazy.IBk

Classifier output

```
=== Run information ==

Scheme:       weka.cla
Relation:     CR17filt
Instances:    29144
Attributes:   400
              [list of
Test mode:    user sup

=== Classifier model

IB1 instance-based cla
using 9 nearest neighb


Time taken to build mo
```

Status

Evaluating on test data. Processed 9100 instances...

CPU 50% 2.55 GHz

Memory 7.8/11.9 GB (66%)

Disk 0 (C: F: G: D:) 0%

Bluetooth Not connected

Ethernet Not connected

Wi-Fi S: 0 R: 16.0 Kbps

Ethernet S: 0 R: 0 Kbps

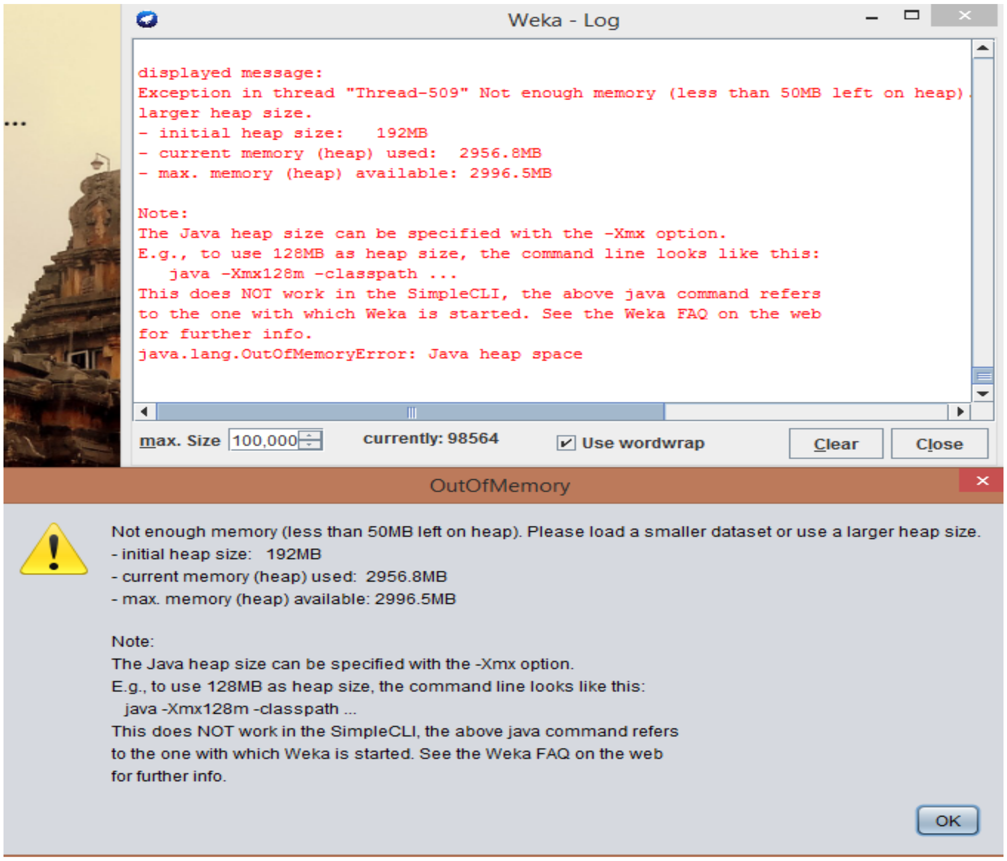
Memory 12.0 GB DDR3

Memory usage 11.9 GB

60 seconds 0

Memory composition

In use	Available	Speed:	1600 MHz
7.7 GB	4.1 GB	Slots used:	2 of 2
Committed	Cached	Form factor:	SODIMM
9.4/13.7 GB	4.0 GB	Hardware reserved:	84.6 MB
Paged pool	Non-paged pool		
602 MB	638 MB		



```
displayed message:
Exception in thread "Thread-509" Not enough memory (less than 50MB left on heap).
larger heap size.
- initial heap size:   192MB
- current memory (heap) used:  2956.8MB
- max. memory (heap) available: 2996.5MB

Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
   java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers
to the one with which Weka is started. See the Weka FAQ on the web
for further info.
java.lang.OutOfMemoryError: Java heap space
```

```
Not enough memory (less than 50MB left on heap). Please load a smaller dataset or use a larger heap size.
- initial heap size:   192MB
- current memory (heap) used:  2956.8MB
- max. memory (heap) available: 2996.5MB

Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
  java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers
to the one with which Weka is started. See the Weka FAQ on the web
for further info.
```

6.3. Potential Improvements or Future Work

We can enhance this analysis further to detect crime pattern and suggest relevant safety measures depending upon the trend and pattern of the crime which has occurred over a period. More precisely, we will can use clustering based models to help in identification of crime pattern. We can predict or foresee future crimes based on the training data set.