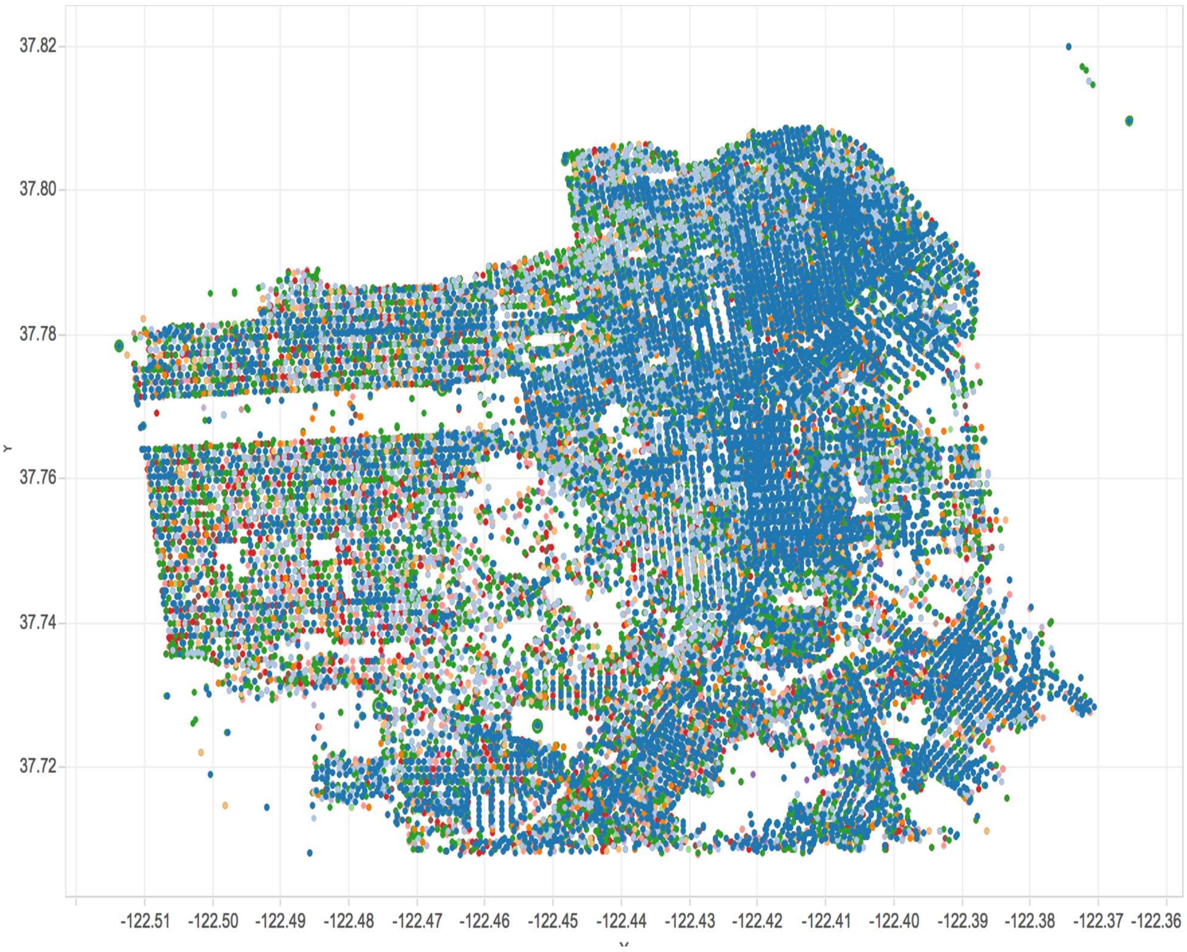
Crime Classification using Archived Data



Team Members

Balaji Jagadeesan

Chetan R Kanthala

Rahul Balu Sakore

Table of Contents

1. Introduction:

1.1 Problem description:

The problem is to classify the category of crime based on the past location and time which can be used by the San Francisco police department to reduce the crime rate. The dataset used for the prediction is based on neighborhood of San Francisco crime reports provided by city and county of San Francisco.

1.2 Motivation:

Crime can happen anywhere and at any time. Some are life threatening and some are petty theft. From the documented data and given factors, there is a possibility to predict the category of crime.

1.3 Report Organization:

The report first introduces the dataset and give brief summary on its data source and its characteristics. It also provides a brief description of attributes to understand how the category of crime is dependent on other features. Then the reports moves on to discuss the various techniques applied to the dataset to check for anomalies. It is in this step, where the data is analyzed for any missing values or outliers in order to improve the quality of the data .It also describes how methods like subset evaluation, InfoGain Analysis help us to understand our dataset. The report talks about sophisticated algorithms like J48, KNN and baseline algorithms like ZeroR and OneR to its parameters in detail to understand its working on the dataset. To gain more insight, performance metrics like kappa statistics and root mean square error are discussed for each algorithm. After the methodology, the logic of problem and its usefulness as an analytical tool is shared. Finally the algorithms are compared and the problems faced during the project are discussed in detail.

2. Data Exploration:

The crime prediction using the archived data consists of data from the San Francisco crime dataset acquired from the San Francisco Police Department. The original dataset is a compilation of 12 years of data from 2003-2015. The dataset contains old data which is of no use as the San Francisco city might have developed over the 12 years and things might have changed and areas might have improved. Also such large dataset gives more information which might have a negative effect on the model. So the dataset is reduced to contain only the recent years (2010-2015).

Data source: https://www.kaggle.com/c/sf-crime/data

Dataset Summary: Consists of data obtained during even weeks (2,4,6,.. week of a year)

Records: 335250 Date: 1/1/2010 to 5/13/2015

No of attributes: 8 and 1 class (category)

Name of the attribute	Nature of data	Description	No of Unique Values
Date	Nominal	Timestamp of the crime incident	392173
Descript	Nominal	Description about the crime	588
DayofWeek	Nominal	The day of the week	7
PdDistrict	Nominal	Police Department responded to the call	10

Resolution	Nominal	How the crime is resolved?	17
Address	Nominal	Address of the crime incident	19313
X and Y	Numeric	Longitude and Latitude	22703 and 21942
Category(Class Value)	Nominal	The kind of crime that has occurred	12

During this phase, the address field is reduced to contain only the street address in order to reduce the uniqueness of the field to some extent.

3. Methodology:

3.1 Data Preprocessing:

3.1.1 Missing Values and Outliers:

The dataset does not contain any missing values and outliers and it is checked using the weka’s preprocess tab and filter called inter-quartile range which is a measure of variability, based on dividing a data set into quartiles that is used by Weka to find the

3.1.2 Sampling:

Although simple algorithms like ZeroR, OneR run smoothly on the dataset, running sophisticated algorithms often ends up crashing Weka. So the dataset is need to be sampled to make it run in Weka. The sample percent of 5 of taken from the dataset to run all our algorithms which comes to 18000 instances. This is done by Resample filter in Weka which randomly chooses instances for each class values.

3.1.3 Attribute Selection:

The attribute selection is important preprocessing step to better understand the input features and its resulting class values. There are various attribute selection algorithms available. The algorithms like subset evaluation, correlation, principal component analysis (PCA) crashes the Weka because the features have unique values which overwhelms the Weka computational power. But InfoGain Analysis and gain ratio analysis yielded the description, date (year, month, time), address, resolution as the important feature that is responsible for classifying the crime.

All the results of the data preprocessing are given under **appendix**

3.1.4 Extraction of features and dimensionality reduction:

The date attribute is nothing but the timestamp that consists of different features like year, month, day and time features, so it is best to split the date to attributes into year, month and time to make the prediction more accurate. Even after splitting the date attributes, the unique values in time stamp is very large. This might not provide accurate results, so the time feature is grouped into 1 hour intervals. Example: 0:00-0:59 is made as 0:00 and 1:00-1:59 as 1:00 and so on. By this method, the dimensionality in time stamp can be reduced so that it can be efficiently used for classification.

The description and address is similar to the date attribute. So the dimensionality is reduced by aggregating the description and address. The description is based on the crime category, so similar description are joined to formed new description for each class category for example, “aggravated assault with gun” and “aggravated assault with knife” are aggregated as “aggravated assault”.

Similar looking categories are aggregated together to form a single category to reduce the misclassification during the model generation for example “vehicle theft” and “robbery” and combined together to form “theft”.

The final dataset that is used for model generation have the following features along with the values in each year (6), month (12),time(48),description (20), daysofWeek (7), resolution (17),Address (12), PdDistrict (10), Category(9). The latitude and longitude are used in tableau for visual analysis of data and are not included.

3.2 Mining the data:

The classification can be done using algorithms like J48, KNN, Naïve Bayes, clustering, association rule mining. Although there are many classification algorithms, our model is developed by using J48, KNN, ZeroR and OneR.

The working of these algorithms and parameters used are discussed below

- **KNN:** It is a simple but yet very powerful algorithm it works on the concept of K-nearest neighbors regardless of the class labels. Based on the K value, the algorithm finds the K Euclidian distance nearest neighbors and assigns the test data class value to be the same as majority classes found. The parameters used are K-value-1,2,3 and search algorithm is linear search
- **J48:** It is an algorithm used to generate a decision tree which is again generated by C4.5. The parameters are set to confidence factor =0.25 and binary splits is set to false
- **OneR:** One Rule simple, yet accurate classification algorithm. It generates one rule for each predictor in the data and then selects the rule with least total error.
- **ZeroR:** This is an algorithm which is based on frequency table. This algorithm is used as the baseline for the algorithm.

All the algorithm have used percentage (split 66% for train set and 33% for test set)

Algorithm	Accuracy	Kappa Stats	RMS	ROC
Zero R	31.87 %	0	0.3143	0.500
One R	98.2843 %	0.9807	0.0617	0.990
J48	98.6438 %	0.9847	0.0478	0.996
KNN k = 1	84.4118 %	0.8246	0.1824	0.922
KNN k = 2	84.3791 %	0.8243	0.1497	0.967
KNN k = 3	89.1503 %	0.8779	0.138	0.979

Table

It can be found that the data generated is over fitting the model, this is due to the fact the description although dimensionality reduced give more information to classify the model.

3.3 Performance Measures:

The following performance measures are used

- **RMS:** The RMSE is a quadratic scoring rule which measures the average magnitude of the error.
- **Accuracy:** It is defined as how well the different unknown attributes are classified to the correct class based on the model developed from the training dataset.
- **Kappa statistics:** The calculation is based on the difference between how much agreement is actually present (“observed” agreement) compared to how much agreement would be expected to be present by chance alone.
- **ROC:** Area under the curve

3.3.1 Motivation:

The motivation for these performance measures is that it will give how well the model is predicted and how well the class values differ from each other. Based on these values a model can be rated whether good, too good or bad.

4. Logic of the problem:

We believe logic of the problem plays a vital role in suspecting the results before actually getting into the project. During our first checkpoint it has contributed good knowledge in how to deal with different types of attributes in our data by "Purpose and Assumptions" fields in the logic of the problem. It provided an overall idea to where to begin and gather information before actually starting our project. As we moved further Point of view, Summary and concepts fields helped us to take our project to a next level. During checkpoint-2 & 3 we were able to explore and investigate to include bit more information. During this whole process it has helped in determining which steps has to considered first so that we can achieve the required end result without stepping ahead with unexpected results.

Overall the Logic of the Problem has all the clusters which makes us to think and gather information before we begin anything in this way one will achieve a precise knowledge in how to choose a better set of data and perform any changes before one gets completely involved.

4.1 Pros and Cons:

Logic of the problem has the ability to teach a variety of skills that can be applied to any situation in life that calls for reflection, analysis, recognize important relationships, make correct inferences from data, evaluate evidence and planning of any kind of problem well in advance before actually getting involved in it so that one can have a brief knowledge why and what is happening.

The major drawback is that the Logic of the Problem provides a dualistic right-versus-wrong stage where the learners basically see the world in dichotomies and they could only gather the information but could not able to pick the right one. A contextual relativism in which the learner can make intellectual commitments within a context of relative knowledge by weighing all the variables and then there should be option available to debate where the Logic of the Problem fails to include in online.

4.2 Recommendations:

The logic of the problem might contain a facts section, so that the facts from different sources can be added to the logic and can be evaluated during each checkpoint. This helps us to understand the facts from different sources which provide information about the problem that might help us to understand the problem much better.

There can also be a section called examples to relate a hypothetical situation to the problem that we are concerned with, which will help us to better understand the problem. We can also add graphs/statistics to visualize the problem in a graphical manner. If possible Critical Thinking website must have a native blog so that all the related topic users can debate and come up with the best solutions which match according to their problem of solving.

5. Conclusion:

The model generated using algorithms show exceptional results which is too good to be true. The model generated is purely based on description and address. We cannot arrive at the conclusion of choosing the best model for our dataset as the description and resolution is based on text and text mining techniques is needed to be analyzed to get best working model of the data. Excluding these two facts, the best model is the one generated through decision tree as the RMS value is low as well as the ROC is high and Kappa statistics is between the range of .4 and .9 which makes it a good model.

The basic idea of our project is to classify the crime based on the input features. Since the input features contain description which is like ID that directly identifies the crime, we tried to reduce the dimensionality by aggregation to get accurate result. Even though, aggregation is used for attributes like address and description, we are not getting accurate results. So different mining techniques should be used to get better results.

Moving on with the project, textual mining can be applied to description, resolution and address to extract the features and spatial mining can be done to the latitude and longitude attributes to better under the input features so that a better model can be created.

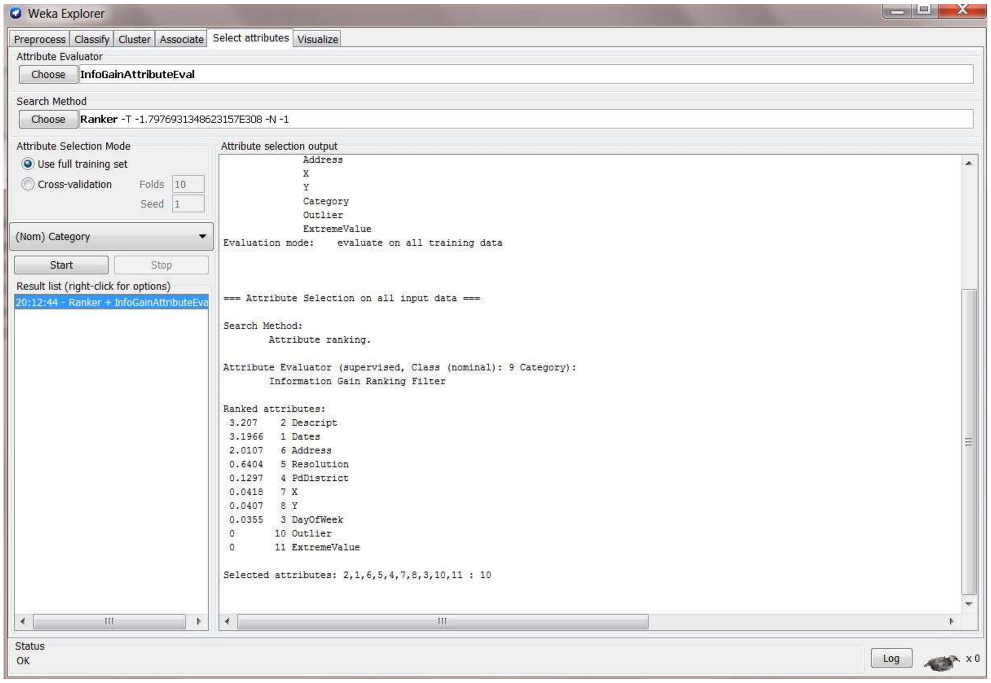
6. Appendix:

Code for Sampling the data(Alternative Code)

```
import csv,random
import numpy as np
csvR = csv.reader(open('./original.csv','rb'))
docs=[]
for k in  csvR:
   docs.append(k)
header = docs[0]
docs=docs[1:]
docs = np.asarray(docs)
clas = docs[:,-3]

clasSet = set(clas)
idx=[]
for k in clasSet:
   idx+=random.sample([i for i,x in enumerate(clas) if x==k],2000)
finalDocs = docs[idx]

csvW = csv.writer(open('./data.csv','wb'))

csvW.writerow(header)
for x in finalDocs:

   csvW.writerow(x)

csvW=[]
```

Model Output:

Feature Selection:



Weka Explorer

Attribute Evaluator: InfoGainAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) Category

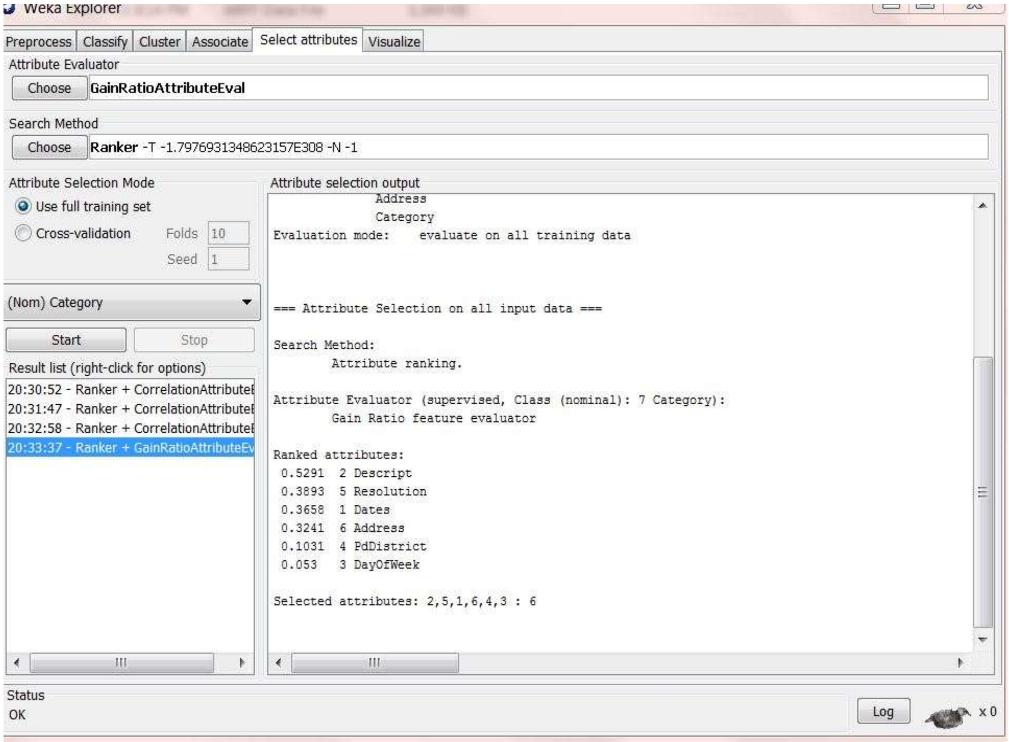
Result list (right-click for options)
20:12:44 - Ranker + InfoGainAttributeEva

Attribute selection output

```
               Address
               X
               Y
               Category
               Outlier
               ExtremeValue
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 Category):
        Information Gain Ranking Filter

Ranked attributes:
 3.207    2 Descript
 3.1966   1 Dates
 2.0107   6 Address
 0.6404   5 Resolution
 0.1297   4 PdDistrict
 0.0418   7 X
 0.0407   8 Y
 0.0355   3 DayOfWeek
 0       10 Outlier
 0       11 ExtremeValue

Selected attributes: 2,1,6,5,4,7,8,3,10,11 : 10
```

Status
OK

Gain Ratio Attribute:



Weka Explorer

Attribute Evaluator: GainRatioAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) Category

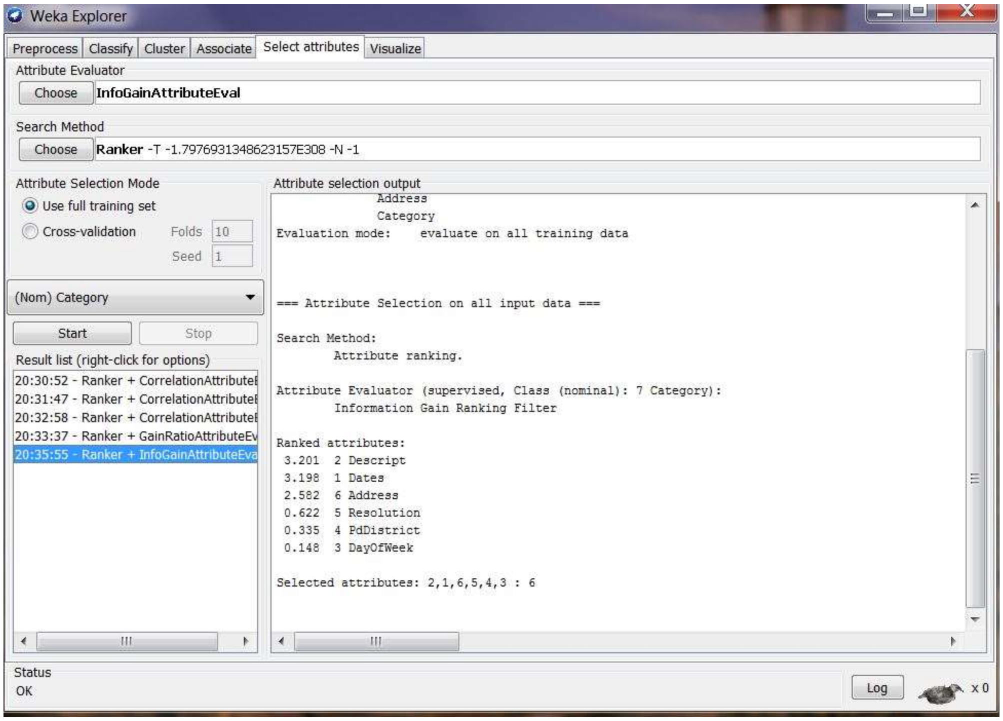
Result list (right-click for options)
20:30:52 - Ranker + CorrelationAttributeE
20:31:47 - Ranker + CorrelationAttributeE
20:32:58 - Ranker + CorrelationAttributeE
20:33:37 - Ranker + GainRatioAttributeEv

Attribute selection output

```
               Address
               Category
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 7 Category):
        Gain Ratio feature evaluator

Ranked attributes:
 0.5291  2 Descript
 0.3893  5 Resolution
 0.3658  1 Dates
 0.3241  6 Address
 0.1031  4 PdDistrict
 0.053   3 DayOfWeek

Selected attributes: 2,5,1,6,4,3 : 6
```

Status
OK

Info-Gain Attribute evaluation:



Attribute Evaluator: InfoGainAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) Category

Result list (right-click for options)

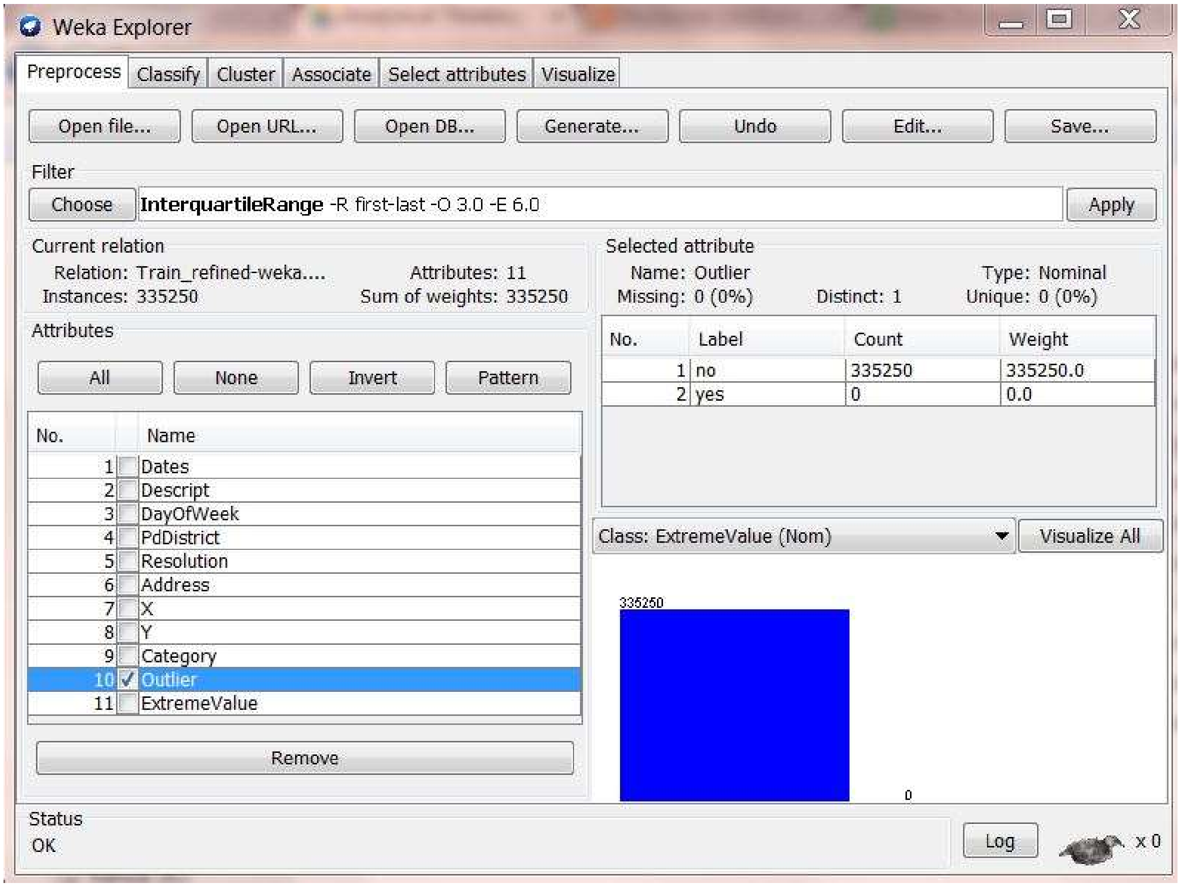
- 20:30:52 - Ranker + CorrelationAttributeE
- 20:31:47 - Ranker + CorrelationAttributeE
- 20:32:58 - Ranker + CorrelationAttributeE
- 20:33:37 - Ranker + GainRatioAttributeEv
- 20:35:55 - Ranker + InfoGainAttributeEva

Attribute selection output

```
              Address
              Category
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 7 Category):
        Information Gain Ranking Filter

Ranked attributes:
 3.201  2 Descript
 3.198  1 Dates
 2.582  6 Address
 0.622  5 Resolution
 0.335  4 PdDistrict
 0.148  3 DayOfWeek

Selected attributes: 2,1,6,5,4,3 : 6
```

Status: OK

Outlier:



Filter: InterquartileRange -R first-last -O 3.0 -E 6.0

Current relation

Relation: Train_refined-weka.... Attributes: 11

Instances: 335250 Sum of weights: 335250

Selected attribute

Name: Outlier Type: Nominal

Missing: 0 (0%) Distinct: 1 Unique: 0 (0%)

No.	Label	Count	Weight
1	no	335250	335250.0
2	yes	0	0.0

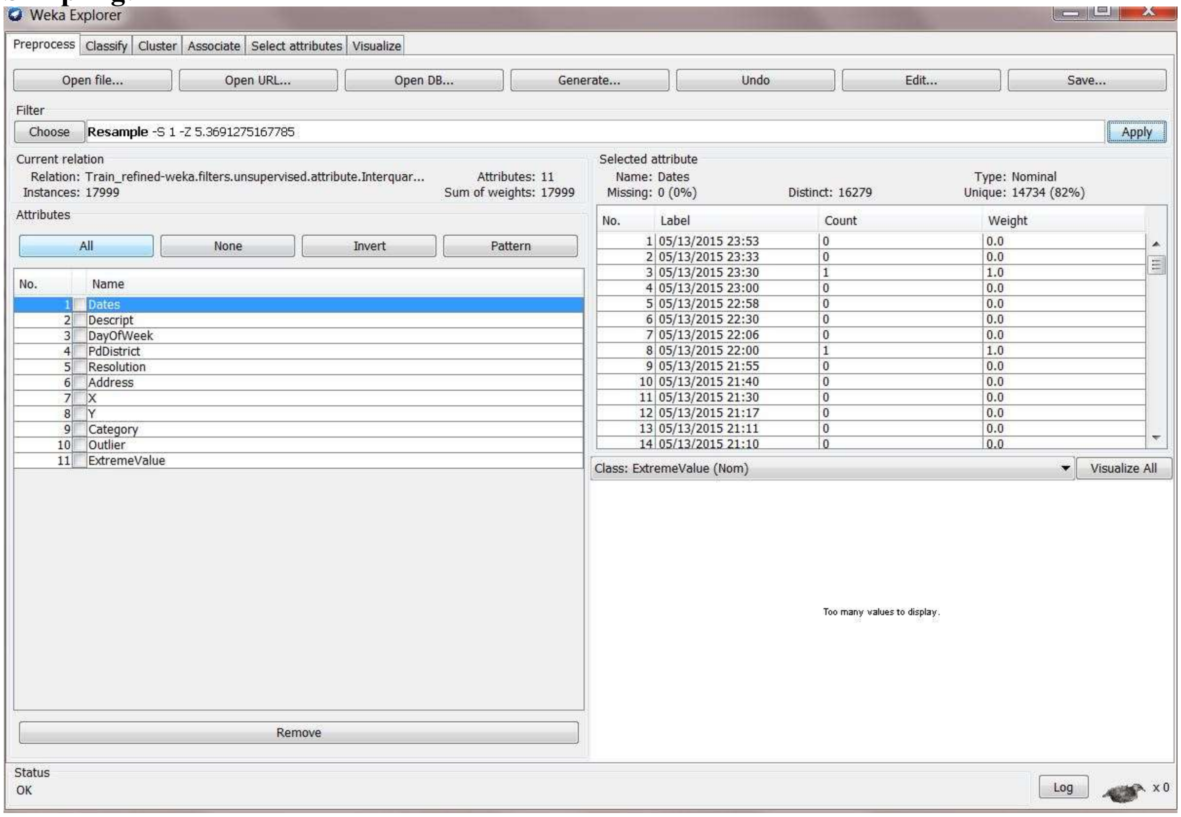
Attributes

No.		Name
1	□	Dates
2	□	Descript
3	□	DayOfWeek
4	□	PdDistrict
5	□	Resolution
6	□	Address
7	□	X
8	□	Y
9	□	Category
10	☑	Outlier
11	□	ExtremeValue

Class: ExtremeValue (Nom)

Status: OK

Sampling:



Filter

Resample -S 1 -Z 5.3691275167785

Current relation

Relation: Train_refined-weka.filters.unsupervised.attribute.Interquar... Attributes: 11

Instances: 17999 Sum of weights: 17999

Attributes

No.	Name
1	Dates
2	Descript
3	DayOfWeek
4	PdDistrict
5	Resolution
6	Address
7	X
8	Y
9	Category
10	Outlier
11	ExtremeValue

Selected attribute

Name: Dates Type: Nominal

Missing: 0 (0%) Distinct: 16279 Unique: 14734 (82%)

No.	Label	Count	Weight
1	05/13/2015 23:53	0	0.0
2	05/13/2015 23:33	0	0.0
3	05/13/2015 23:30	1	1.0
4	05/13/2015 23:00	0	0.0
5	05/13/2015 22:58	0	0.0
6	05/13/2015 22:30	0	0.0
7	05/13/2015 22:06	0	0.0
8	05/13/2015 22:00	1	1.0
9	05/13/2015 21:55	0	0.0
10	05/13/2015 21:40	0	0.0
11	05/13/2015 21:30	0	0.0
12	05/13/2015 21:17	0	0.0
13	05/13/2015 21:11	0	0.0
14	05/13/2015 21:10	0	0.0

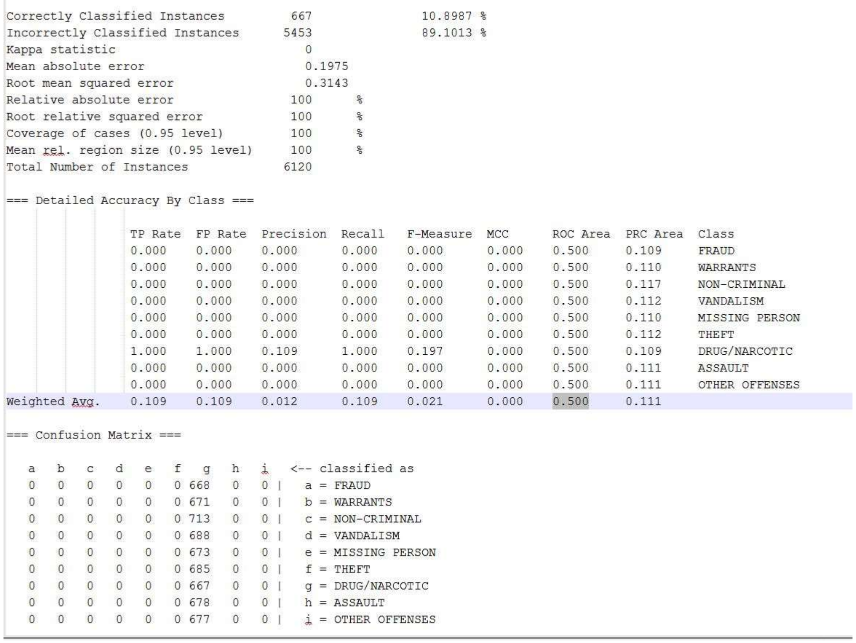
Class: ExtremeValue (Nom)

Too many values to display.

Status

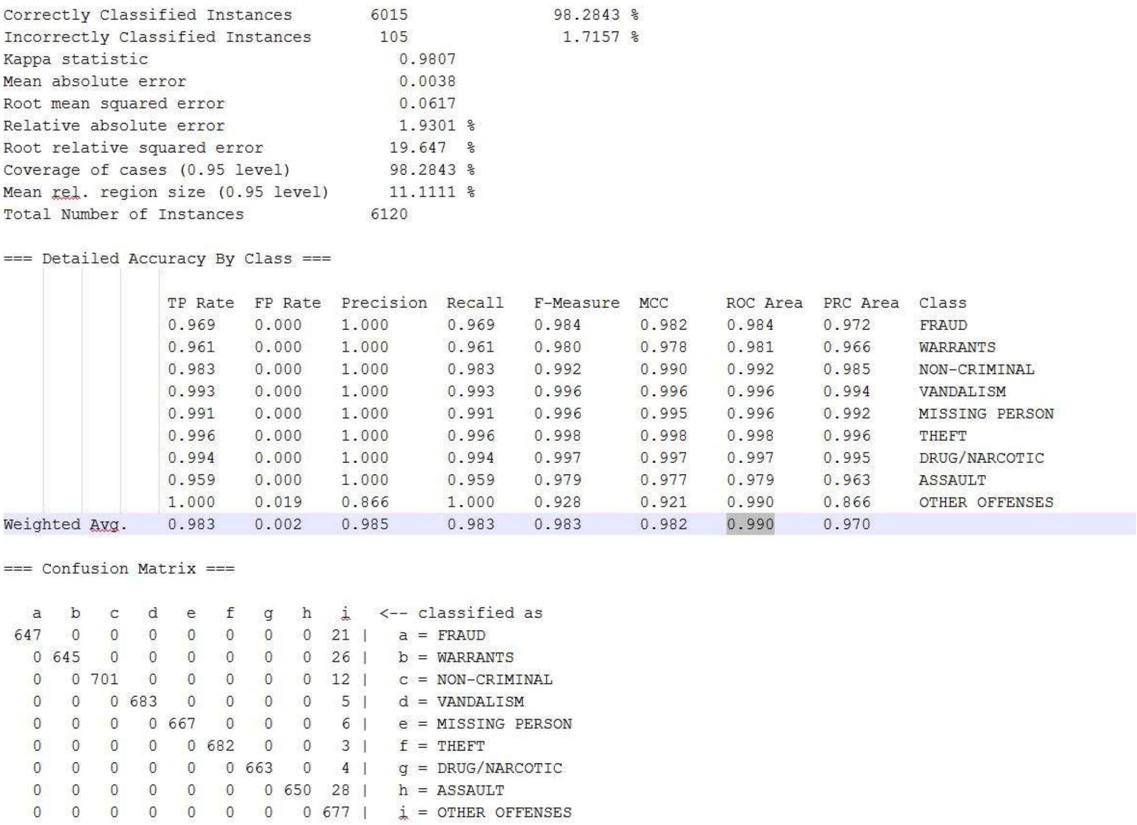
OK

Model Output For: ZeroR:



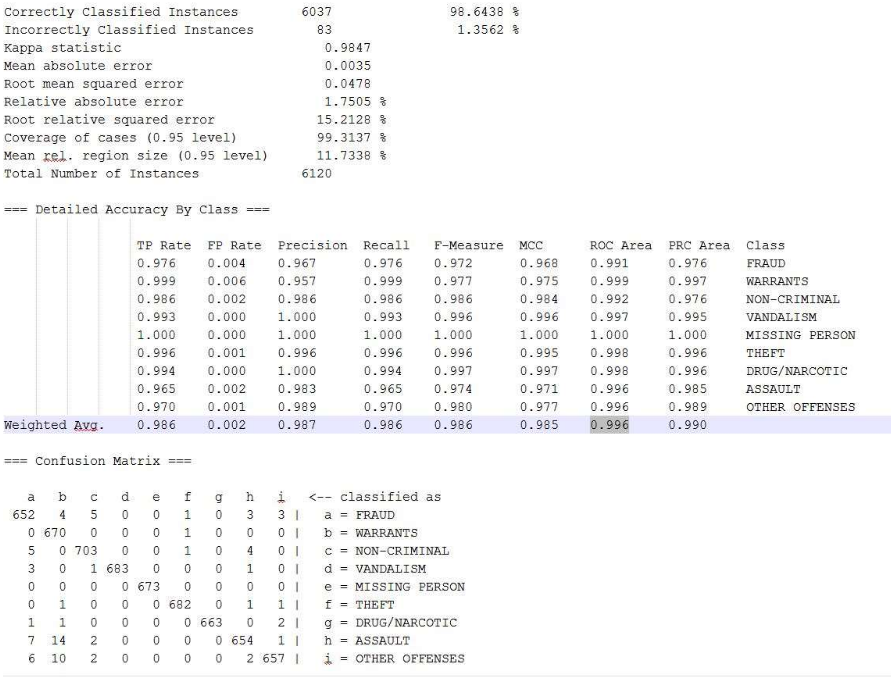
```
Correctly Classified Instances         667               10.8987 %
Incorrectly Classified Instances      5453               89.1013 %
Kappa statistic                          0
Mean absolute error                      0.1975
Root mean squared error                  0.3143
Relative absolute error                100      %
Root relative squared error            100      %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances             6120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.109     FRAUD
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.110     WARRANTS
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.117     NON-CRIMINAL
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.112     VANDALISM
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.110     MISSING PERSON
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.112     THEFT
                 1.000    1.000    0.109      1.000    0.197      0.000    0.500     0.109     DRUG/NARCOTIC
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.111     ASSAULT
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.111     OTHER OFFENSES
Weighted Avg.    0.109    0.109    0.012      0.109    0.021      0.000    0.500     0.111

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
   0   0   0   0   0   0 668   0   0 |   a = FRAUD
   0   0   0   0   0   0 671   0   0 |   b = WARRANTS
   0   0   0   0   0   0 713   0   0 |   c = NON-CRIMINAL
   0   0   0   0   0   0 688   0   0 |   d = VANDALISM
   0   0   0   0   0   0 673   0   0 |   e = MISSING PERSON
   0   0   0   0   0   0 685   0   0 |   f = THEFT
   0   0   0   0   0   0 667   0   0 |   g = DRUG/NARCOTIC
   0   0   0   0   0   0 678   0   0 |   h = ASSAULT
   0   0   0   0   0   0 677   0   0 |   i = OTHER OFFENSES
```

Model Output for One R:



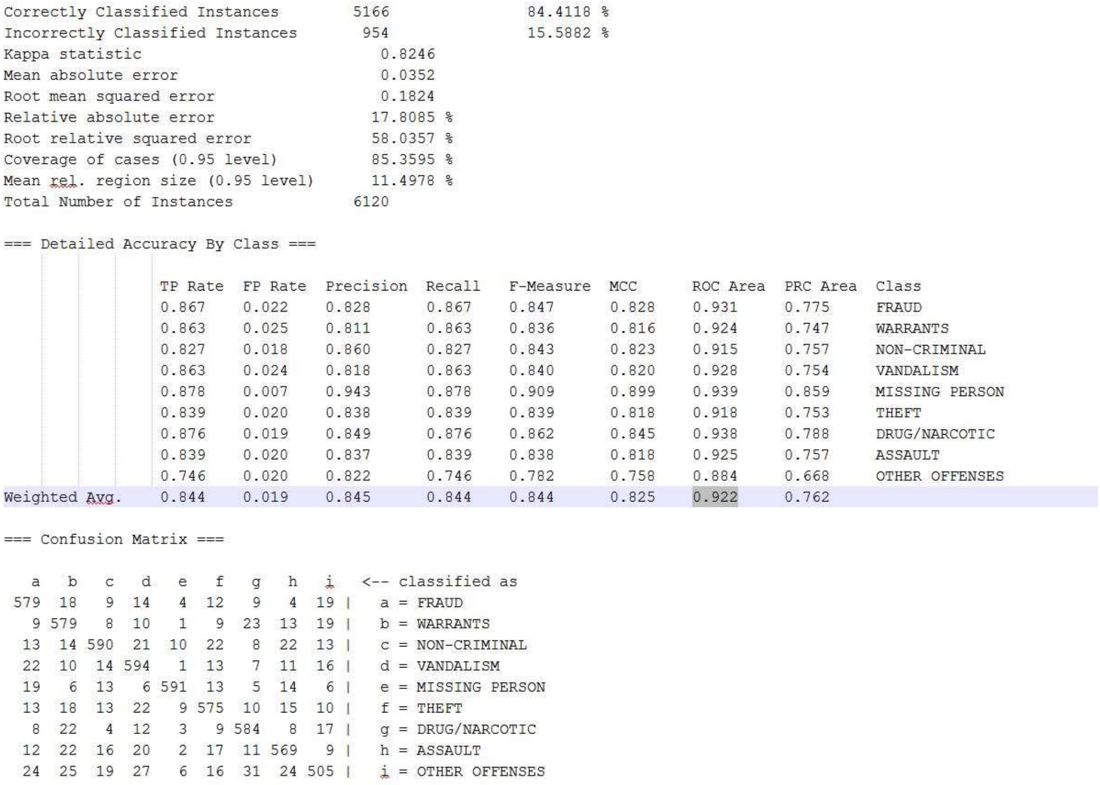
```
Correctly Classified Instances        6015               98.2843 %
Incorrectly Classified Instances       105                1.7157 %
Kappa statistic                          0.9807
Mean absolute error                      0.0038
Root mean squared error                  0.0617
Relative absolute error                  1.9301 %
Root relative squared error             19.647  %
Coverage of cases (0.95 level)          98.2843 %
Mean rel. region size (0.95 level)      11.1111 %
Total Number of Instances             6120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.000    1.000      0.969    0.984      0.982    0.984     0.972     FRAUD
                 0.961    0.000    1.000      0.961    0.980      0.978    0.981     0.966     WARRANTS
                 0.983    0.000    1.000      0.983    0.992      0.990    0.992     0.985     NON-CRIMINAL
                 0.993    0.000    1.000      0.993    0.996      0.996    0.996     0.994     VANDALISM
                 0.991    0.000    1.000      0.991    0.996      0.995    0.996     0.992     MISSING PERSON
                 0.996    0.000    1.000      0.996    0.998      0.998    0.998     0.996     THEFT
                 0.994    0.000    1.000      0.994    0.997      0.997    0.997     0.995     DRUG/NARCOTIC
                 0.959    0.000    1.000      0.959    0.979      0.977    0.979     0.963     ASSAULT
                 1.000    0.019    0.866      1.000    0.928      0.921    0.990     0.866     OTHER OFFENSES
Weighted Avg.    0.983    0.002    0.985      0.983    0.983      0.982    0.990     0.970

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 647   0   0   0   0   0   0   0  21 |   a = FRAUD
   0 645   0   0   0   0   0   0  26 |   b = WARRANTS
   0   0 701   0   0   0   0   0  12 |   c = NON-CRIMINAL
   0   0   0 683   0   0   0   0   5 |   d = VANDALISM
   0   0   0   0 667   0   0   0   6 |   e = MISSING PERSON
   0   0   0   0   0 682   0   0   3 |   f = THEFT
   0   0   0   0   0   0 663   0   4 |   g = DRUG/NARCOTIC
   0   0   0   0   0   0   0 650  28 |   h = ASSAULT
   0   0   0   0   0   0   0   0 677 |   i = OTHER OFFENSES
```

Model Output for J48:



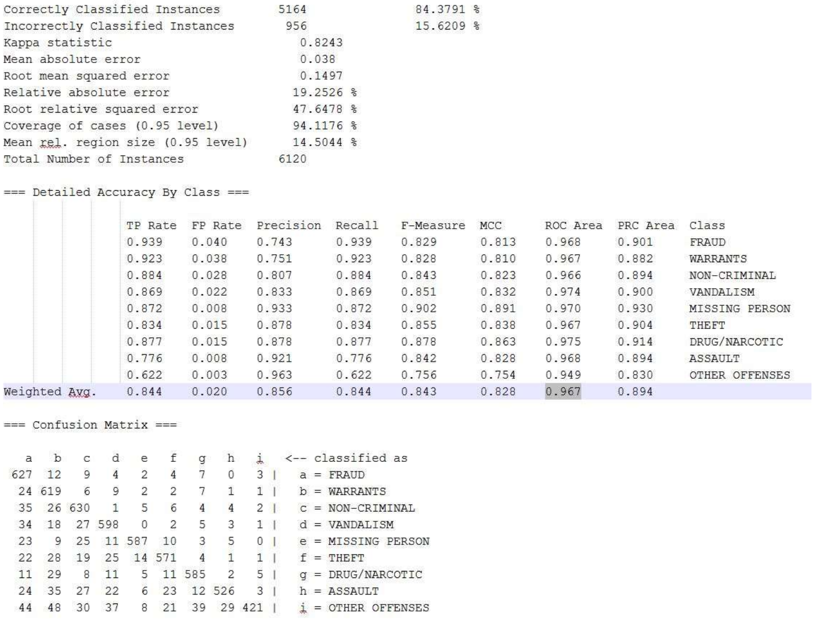
```
Correctly Classified Instances        6037               98.6438 %
Incorrectly Classified Instances        83                1.3562 %
Kappa statistic                          0.9847
Mean absolute error                      0.0035
Root mean squared error                  0.0478
Relative absolute error                  1.7505 %
Root relative squared error             15.2128 %
Coverage of cases (0.95 level)          99.3137 %
Mean rel. region size (0.95 level)      11.7338 %
Total Number of Instances             6120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.976    0.004    0.967      0.976    0.972      0.968    0.991     0.976     FRAUD
                 0.999    0.006    0.957      0.999    0.977      0.975    0.999     0.997     WARRANTS
                 0.986    0.002    0.986      0.986    0.986      0.984    0.992     0.976     NON-CRIMINAL
                 0.993    0.000    1.000      0.993    0.996      0.996    0.997     0.995     VANDALISM
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     MISSING PERSON
                 0.996    0.001    0.996      0.996    0.996      0.995    0.998     0.996     THEFT
                 0.994    0.000    1.000      0.994    0.997      0.997    0.998     0.996     DRUG/NARCOTIC
                 0.965    0.002    0.983      0.965    0.974      0.971    0.996     0.985     ASSAULT
                 0.970    0.001    0.989      0.970    0.980      0.977    0.996     0.989     OTHER OFFENSES
Weighted Avg.    0.986    0.002    0.987      0.986    0.986      0.985    0.996     0.990

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 652   4   5   0   0   1   0   3   3 |   a = FRAUD
   0 670   0   0   0   1   0   0   0 |   b = WARRANTS
   5   0 703   0   0   1   0   4   0 |   c = NON-CRIMINAL
   3   0   1 683   0   0   0   1   0 |   d = VANDALISM
   0   0   0   0 673   0   0   0   0 |   e = MISSING PERSON
   0   1   0   0   0 682   0   1   1 |   f = THEFT
   1   1   0   0   0   0 663   0   2 |   g = DRUG/NARCOTIC
   7  14   2   0   0   0   0 654   1 |   h = ASSAULT
   6  10   2   0   0   0   0   2 657 |   i = OTHER OFFENSES
```

Model Output for KNN K = 1:



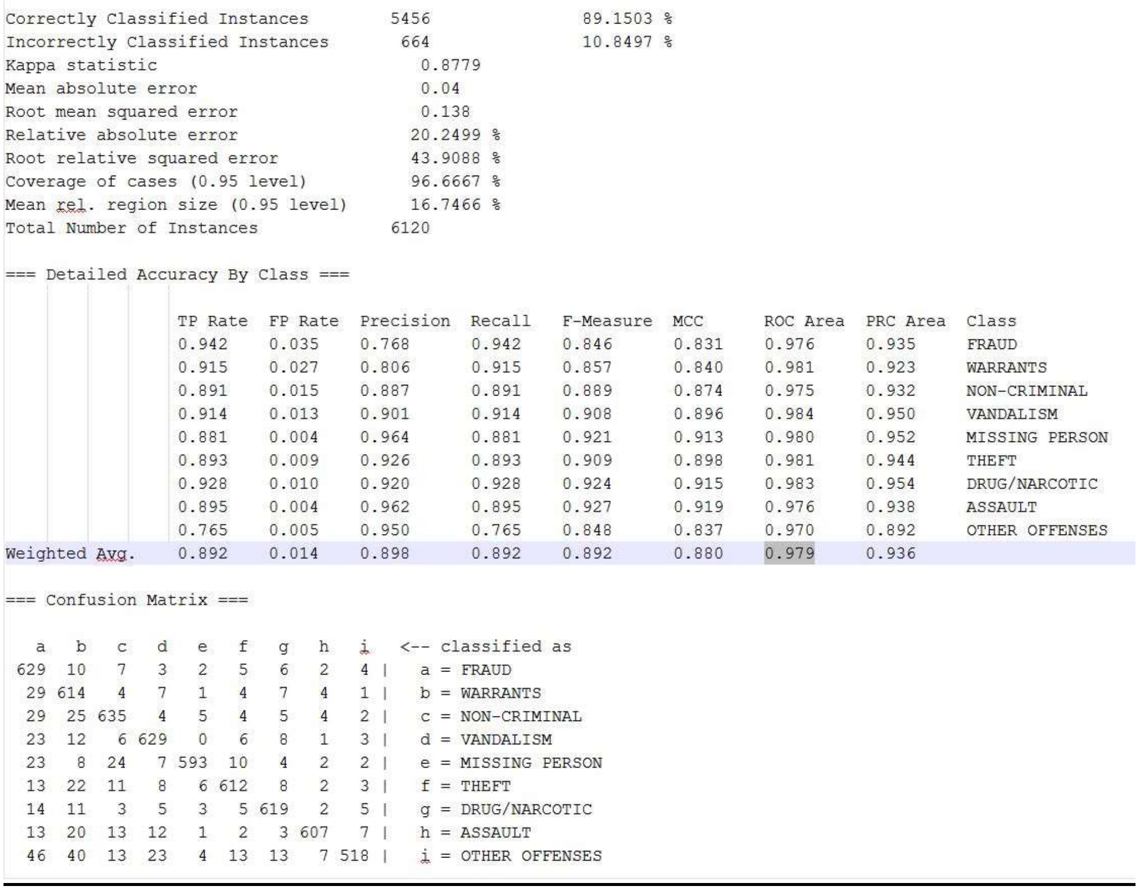
```
Correctly Classified Instances        5166               84.4118 %
Incorrectly Classified Instances       954               15.5882 %
Kappa statistic                          0.8246
Mean absolute error                      0.0352
Root mean squared error                  0.1824
Relative absolute error                 17.8085 %
Root relative squared error             58.0357 %
Coverage of cases (0.95 level)          85.3595 %
Mean rel. region size (0.95 level)      11.4978 %
Total Number of Instances             6120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.867    0.022    0.828      0.867    0.847      0.828    0.931     0.775     FRAUD
                 0.863    0.025    0.811      0.863    0.836      0.816    0.924     0.747     WARRANTS
                 0.827    0.018    0.860      0.827    0.843      0.823    0.915     0.757     NON-CRIMINAL
                 0.863    0.024    0.818      0.863    0.840      0.820    0.928     0.754     VANDALISM
                 0.878    0.007    0.943      0.878    0.909      0.899    0.939     0.859     MISSING PERSON
                 0.839    0.020    0.838      0.839    0.839      0.818    0.918     0.753     THEFT
                 0.876    0.019    0.849      0.876    0.862      0.845    0.938     0.788     DRUG/NARCOTIC
                 0.839    0.020    0.837      0.839    0.838      0.818    0.925     0.757     ASSAULT
                 0.746    0.020    0.822      0.746    0.782      0.758    0.884     0.668     OTHER OFFENSES
Weighted Avg.    0.844    0.019    0.845      0.844    0.844      0.825    0.922     0.762

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 579  18   9  14   4  12   9   4  19 |   a = FRAUD
   9 579   8  10   1   9  23  13  19 |   b = WARRANTS
  13  14 590  21  10  22   8  22  13 |   c = NON-CRIMINAL
  22  10  14 594   1  13   7  11  16 |   d = VANDALISM
  19   6  13   6 591  13   5  14   6 |   e = MISSING PERSON
  13  18  13  22   9 575  10  15  10 |   f = THEFT
   8  22   4  12   3   9 584   8  17 |   g = DRUG/NARCOTIC
  12  22  16  20   2  17  11 569   9 |   h = ASSAULT
  24  25  19  27   6  16  31  24 505 |   i = OTHER OFFENSES
```

Model Output for KNN k = 2



```
Correctly Classified Instances        5164               84.3791 %
Incorrectly Classified Instances       956               15.6209 %
Kappa statistic                          0.8243
Mean absolute error                      0.038
Root mean squared error                  0.1497
Relative absolute error                 19.2526 %
Root relative squared error             47.6478 %
Coverage of cases (0.95 level)          94.1176 %
Mean rel. region size (0.95 level)      14.5044 %
Total Number of Instances             6120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.939    0.040    0.743      0.939    0.829      0.813    0.968     0.901     FRAUD
                 0.923    0.038    0.751      0.923    0.828      0.810    0.967     0.882     WARRANTS
                 0.884    0.028    0.807      0.884    0.843      0.823    0.966     0.894     NON-CRIMINAL
                 0.869    0.022    0.833      0.869    0.851      0.832    0.974     0.900     VANDALISM
                 0.872    0.008    0.933      0.872    0.902      0.891    0.970     0.930     MISSING PERSON
                 0.834    0.015    0.878      0.834    0.855      0.838    0.967     0.904     THEFT
                 0.877    0.015    0.878      0.877    0.878      0.863    0.975     0.914     DRUG/NARCOTIC
                 0.776    0.008    0.921      0.776    0.842      0.828    0.968     0.894     ASSAULT
                 0.622    0.003    0.963      0.622    0.756      0.754    0.949     0.830     OTHER OFFENSES
Weighted Avg.    0.844    0.020    0.856      0.844    0.843      0.828    0.967     0.894

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 627  12   9   4   2   4   7   0   3 |   a = FRAUD
  24 619   6   9   2   2   7   1   1 |   b = WARRANTS
  35  26 630   1   5   6   4   4   2 |   c = NON-CRIMINAL
  34  18  27 598   0   2   5   3   1 |   d = VANDALISM
  23   9  25  11 587  10   3   5   0 |   e = MISSING PERSON
  22  28  19  25  14 571   4   1   1 |   f = THEFT
  11  29   8  11   5  11 585   2   5 |   g = DRUG/NARCOTIC
  24  35  27  22   6  23  12 526   3 |   h = ASSAULT
  44  48  30  37   8  21  39  29 421 |   i = OTHER OFFENSES
```

Model Output for KNN k = 3



```
Correctly Classified Instances        5456               89.1503 %
Incorrectly Classified Instances       664               10.8497 %
Kappa statistic                          0.8779
Mean absolute error                      0.04
Root mean squared error                  0.138
Relative absolute error                 20.2499 %
Root relative squared error             43.9088 %
Coverage of cases (0.95 level)          96.6667 %
Mean rel. region size (0.95 level)      16.7466 %
Total Number of Instances             6120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.942    0.035    0.768      0.942    0.846      0.831    0.976     0.935     FRAUD
                 0.915    0.027    0.806      0.915    0.857      0.840    0.981     0.923     WARRANTS
                 0.891    0.015    0.887      0.891    0.889      0.874    0.975     0.932     NON-CRIMINAL
                 0.914    0.013    0.901      0.914    0.908      0.896    0.984     0.950     VANDALISM
                 0.881    0.004    0.964      0.881    0.921      0.913    0.980     0.952     MISSING PERSON
                 0.893    0.009    0.926      0.893    0.909      0.898    0.981     0.944     THEFT
                 0.928    0.010    0.920      0.928    0.924      0.915    0.983     0.954     DRUG/NARCOTIC
                 0.895    0.004    0.962      0.895    0.927      0.919    0.976     0.938     ASSAULT
                 0.765    0.005    0.950      0.765    0.848      0.837    0.970     0.892     OTHER OFFENSES
Weighted Avg.    0.892    0.014    0.898      0.892    0.892      0.880    0.979     0.936

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   <-- classified as
 629  10   7   3   2   5   6   2   4 |   a = FRAUD
  29 614   4   7   1   4   7   4   1 |   b = WARRANTS
  29  25 635   4   5   4   5   4   2 |   c = NON-CRIMINAL
  23  12   6 629   0   6   8   1   3 |   d = VANDALISM
  23   8  24   7 593  10   4   2   2 |   e = MISSING PERSON
  13  22  11   8   6 612   8   2   3 |   f = THEFT
  14  11   3   5   3   5 619   2   5 |   g = DRUG/NARCOTIC
  13  20  13  12   1   2   3 607   7 |   h = ASSAULT
  46  40  13  23   4  13  13   7 518 |   i = OTHER OFFENSES
```

7. Reference

- Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKA---Experiences with a Java Open-Source Project. The Journal of Machine Learning Research, 11, 2533-2541.
- San fransisco crime classification. (n.d.). Retrieved from https://www.kaggle.com/c/sf-crime
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining (Vol. 1). Boston: Pearson Addison Wesley.