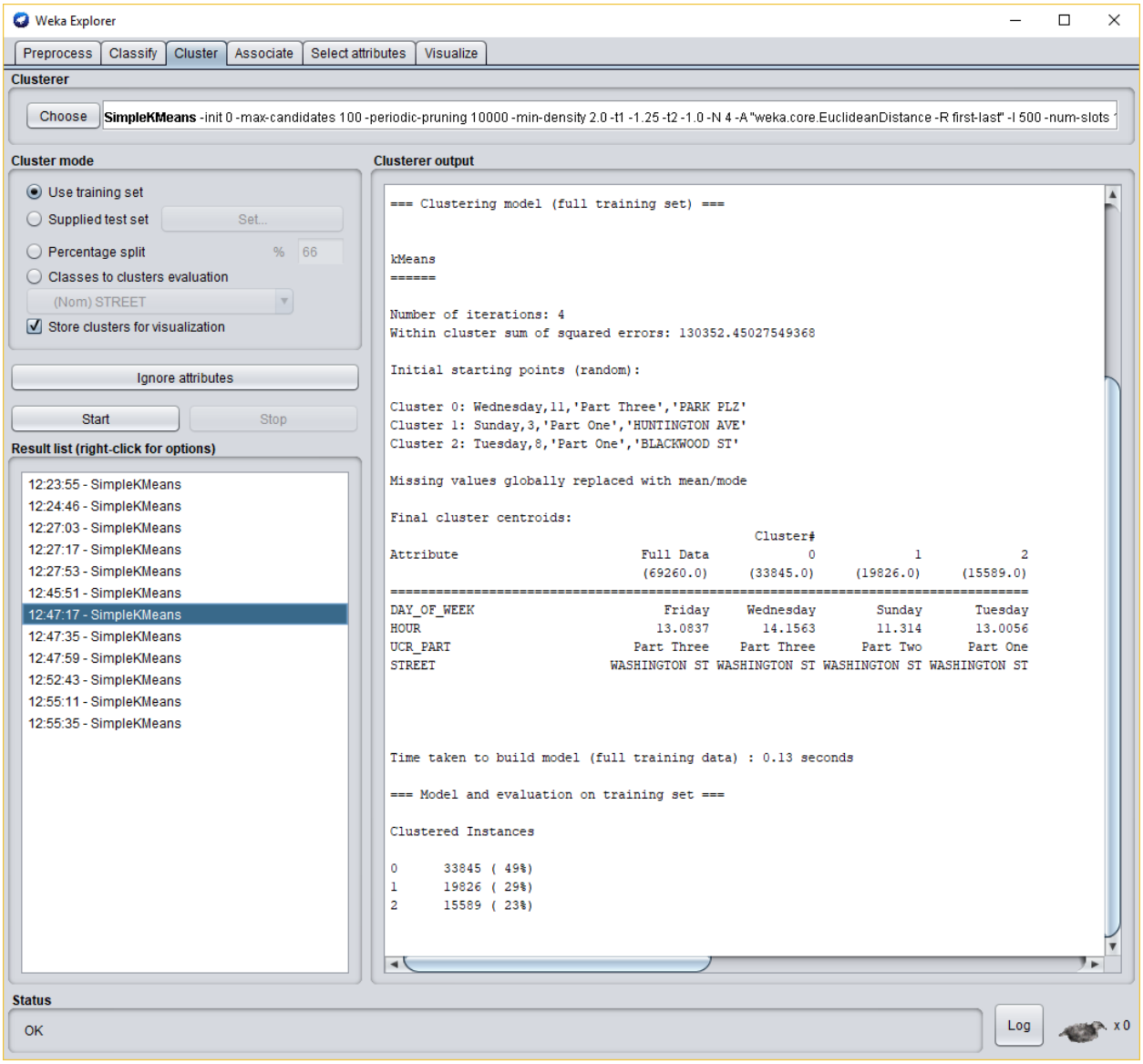
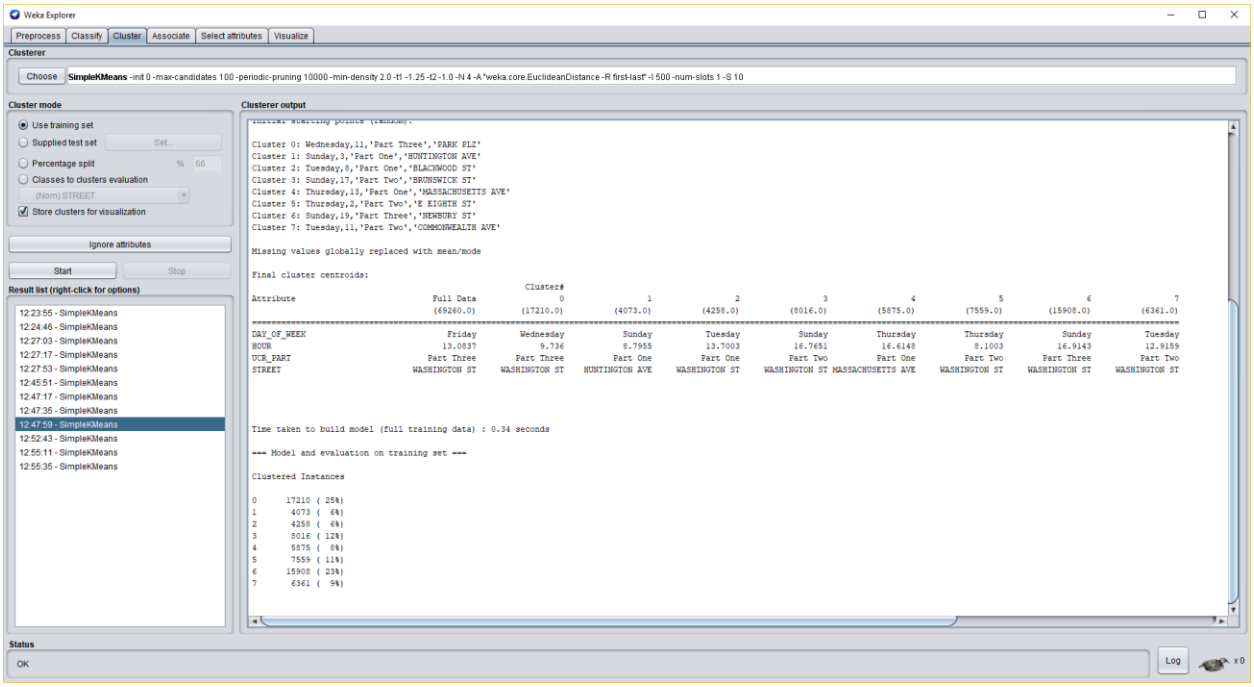
What are the streets with the most crime? What type of crime and when do they happen?

Removed all records with missing location data because the impact of these missing locations was so significant that all the results ended in missing data.

To find the location with the most crime, by type of crime, hour, and day of the week a Simple K Means algorithm was used. Multiple runs were made using different numbers of clusters. The result using 3 clusters was that Washington St had the most crime for all UCR parts. Surprisingly most incidents happen during daylight between 11 am and 2 pm and between Sunday, Tuesday, and Wednesday. Part 1 crimes are the most serious type of crime happening Tuesdays around 1 pm.



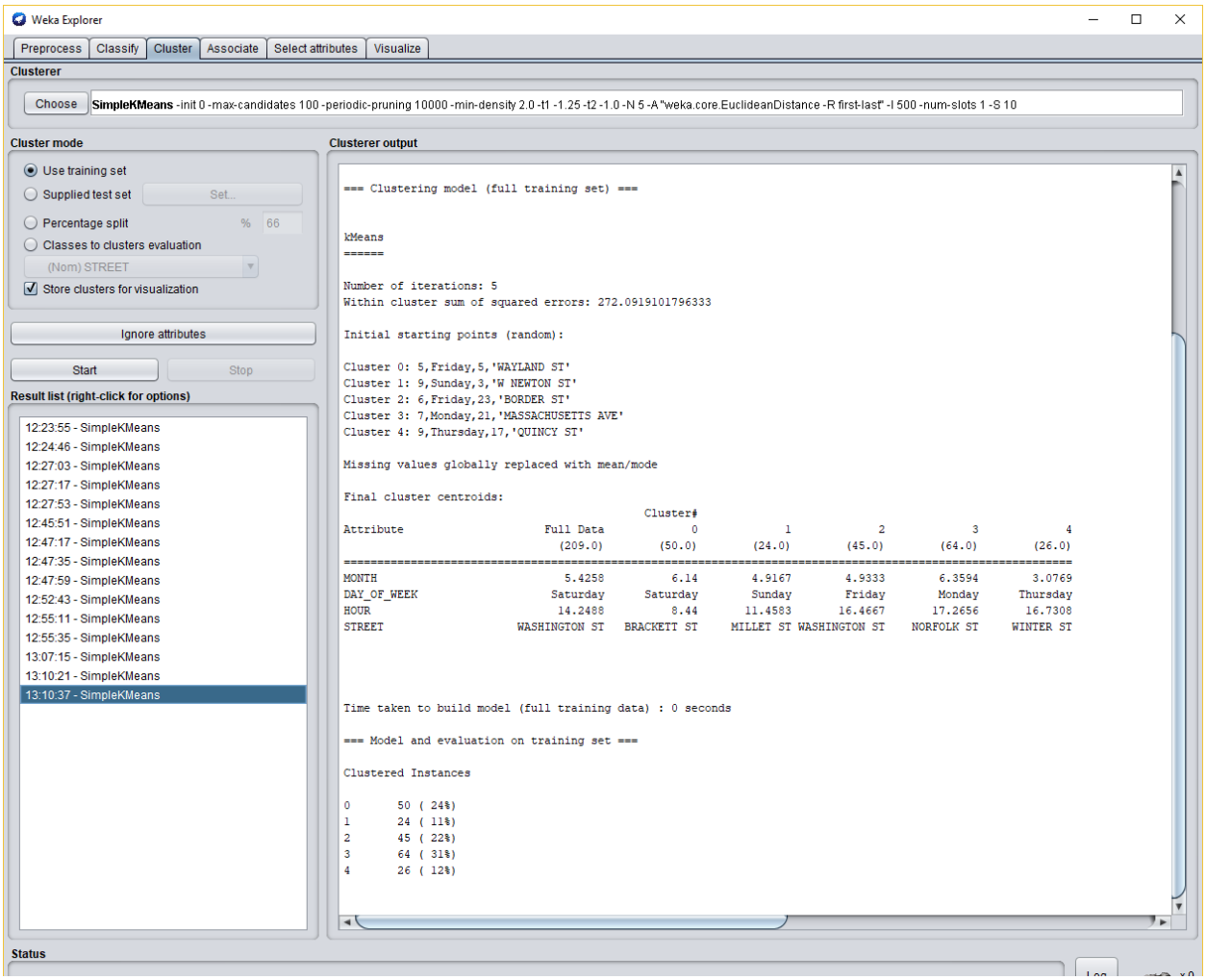
After running the algorithm with 8 clusters, the results were somewhat similar since Washington St appeared in 6 out of the 8 clusters with the highest number of records. Again, most crimes seem to happen during daylight and on weekdays.



When and where do crimes involving shootings happen?

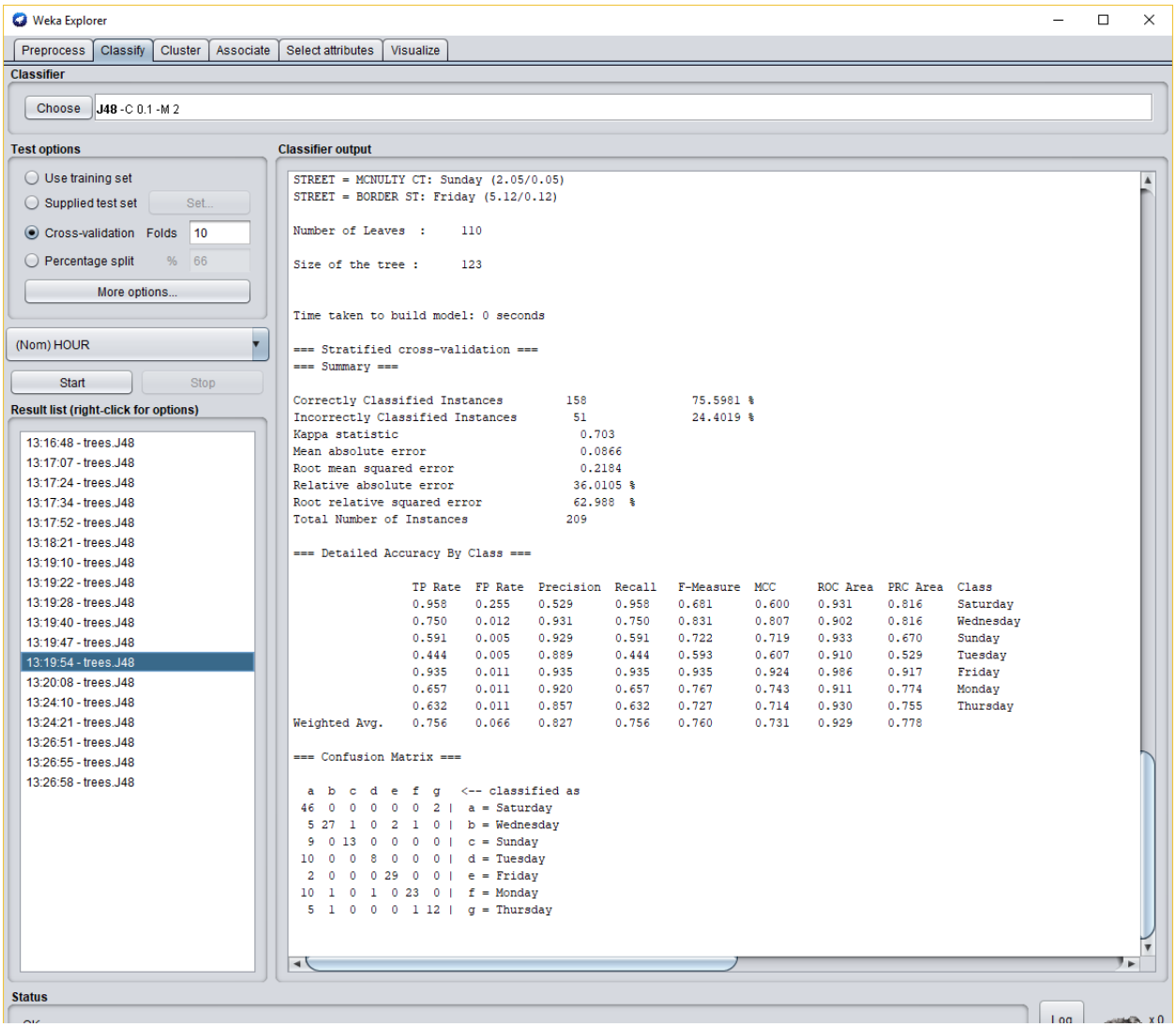
***After removing all records with No shooting, the dataset was reduced to 210. This was done to only focus our attention on records where there was a shooting and find out the day, hour, month, and street where the shootings occurred. By keeping the data where there was no shooting, all clusters would result in no shooting clusters.

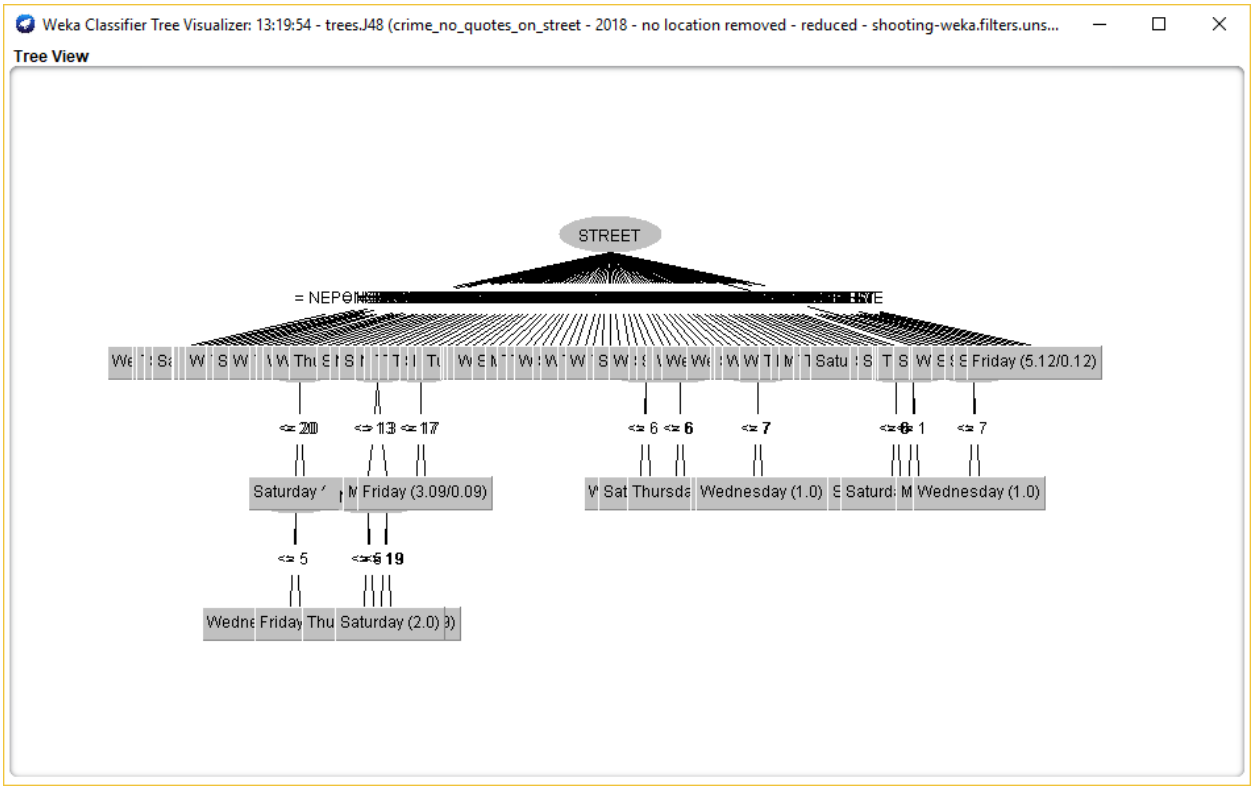
Clusters for the shooting data with 5 clusters were distributed thru 5 different streets. Most shootings occurred during the months of March to June, the shootings happened mostly during daylight and were distributed between weekdays and weekends. It is of no surprise that the shootings mostly happened during months of warmer weather, rather than during the winter.



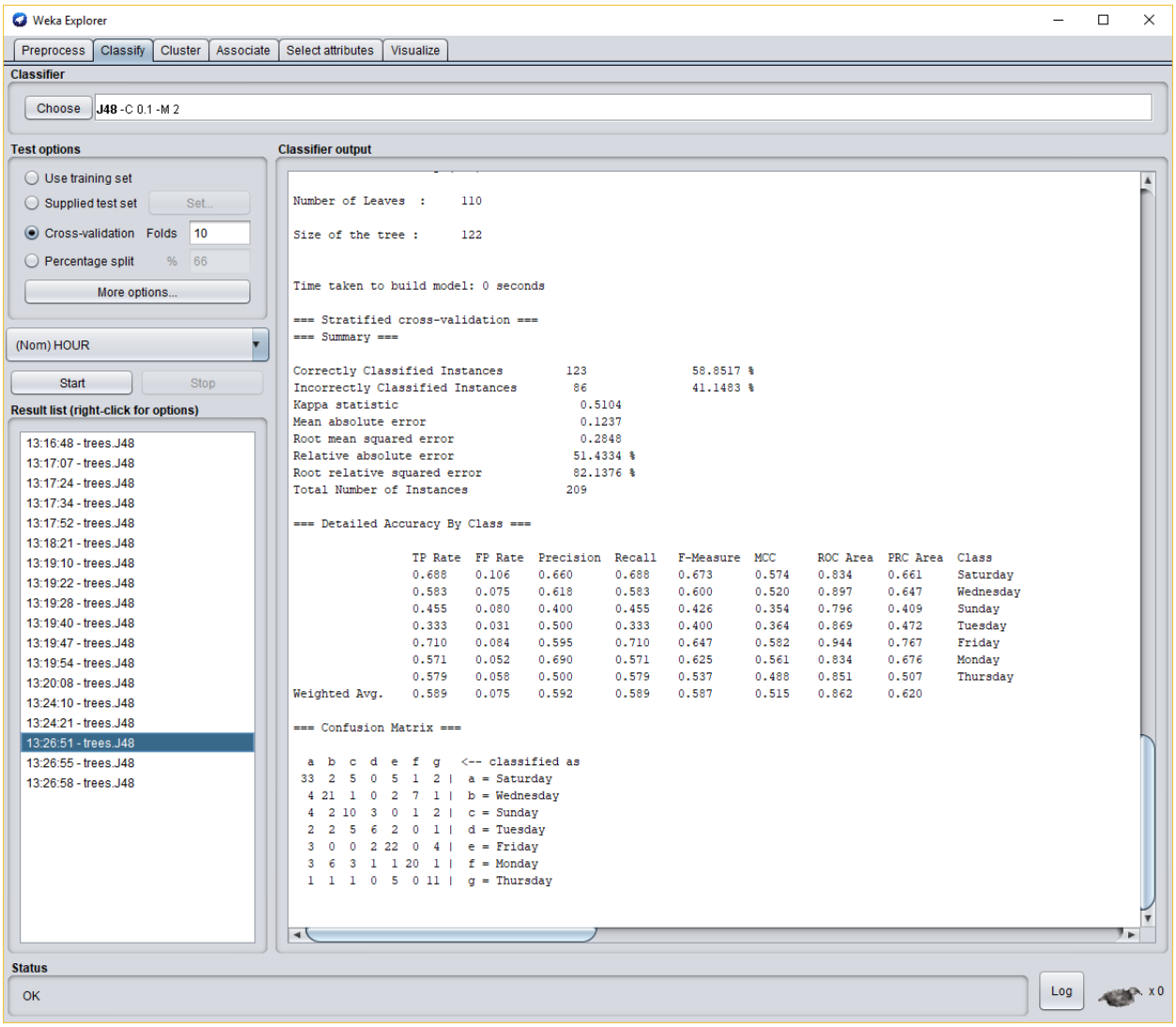
Classifying shooting data

Using hour, month, day of the week, and street to classify the possibility of a shooting showed a 75% correct classification but the rules were very hard to obtain because of the large number of Street names. The tree is difficult to read.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.1 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) HOUR
Start
Stop
Result list (right-click for options)
13:16:48 - trees.J48
13:17:07 - trees.J48
13:17:24 - trees.J48
13:17:34 - trees.J48
13:17:52 - trees.J48
13:18:21 - trees.J48
13:19:10 - trees.J48
13:19:22 - trees.J48
13:19:28 - trees.J48
13:19:40 - trees.J48
13:19:47 - trees.J48
13:19:54 - trees.J48
13:20:08 - trees.J48
13:24:10 - trees.J48
13:24:21 - trees.J48
13:26:51 - trees.J48
13:26:55 - trees.J48
13:26:58 - trees.J48
Classifier output
STREET = MCNULTY CT: Sunday (2.05/0.05)
STREET = BORDER ST: Friday (5.12/0.12)
Number of Leaves : 110
Size of the tree : 123
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 158 75.5981 %
Incorrectly Classified Instances 51 24.4019 %
Kappa statistic 0.703
Mean absolute error 0.0866
Root mean squared error 0.2184
Relative absolute error 36.0105 %
Root relative squared error 62.988 %
Total Number of Instances 209
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.958 0.255 0.529 0.958 0.681 0.600 0.931 0.816 Saturday
0.750 0.012 0.931 0.750 0.831 0.807 0.902 0.816 Wednesday
0.591 0.005 0.929 0.591 0.722 0.719 0.933 0.670 Sunday
0.444 0.005 0.889 0.444 0.593 0.607 0.910 0.529 Tuesday
0.935 0.011 0.935 0.935 0.935 0.924 0.986 0.917 Friday
0.657 0.011 0.920 0.657 0.767 0.743 0.911 0.774 Monday
0.632 0.011 0.857 0.632 0.727 0.714 0.930 0.755 Thursday
Weighted Avg. 0.756 0.066 0.827 0.756 0.760 0.731 0.929 0.778
=== Confusion Matrix ===
a b c d e f g <-- classified as
46 0 0 0 0 0 2 | a = Saturday
5 27 1 0 2 1 0 | b = Wednesday
9 0 13 0 0 0 0 | c = Sunday
10 0 0 8 0 0 0 | d = Tuesday
2 0 0 0 29 0 0 | e = Friday
10 1 0 1 0 23 0 | f = Monday
5 1 0 0 0 1 12 | g = Thursday
Status
Log
x 0



Results without the Street data and using only hour, day of the week, and month, resulted in 58% correct classification, therefore these are not reliable.



To run the Apriori algorithm the Month attribute was changed from Numeric to Nominal using the NumerictoNominal filter in Weka.