Dallas Police Public Data - RMS Incidents

Predicting the whether Drugs are involved in Crime

—

Sebastian, Jules 040 983 119
Simpson, Skyler 040 951 590

Data Collection

Source: Dallas Police Public Data - RMS Incidents

https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7

How the data was collected

The data was sourced from preliminary reports given to the police after June 1st 2014.

Preprocessing

The Initial Data was Enormous, approx 875 000 Rows of Data, denoted by 86 Attributes.

- Lots of data does not mean the data is good.

Duplicate data was Removed

Redundant Attributes were Merged

Outliers were Removed

Correlating Attributes were filled based on the dataset

Missing Data led to removed rows: Where 6 or more attributes had missing values, the rows are ignored

Analysis

Classification

Vector distances are calculated between neighboring data points and classifying the dataset based on those distances.

kNearestNeighbors

We attempted kNN but a vector based approach was very costly, when nominal fields have such a large breadth of options.

- Long Calculations [>10 hours] and unreliable results

Analysis

Clustering

Grouping objects in such a way, that objects falling in a group would be more similar to those in other groups

- **Simple kMeans**
- **Farthest First**

Results

Clustering predicted approx 8% of incidents were drug related

With 10.85% of the data incorrectly clustered.

By projecting from the original dataset, we would expect the results to land closer to 3.85% of incidents are drug related.

This means our model overestimates the number of incidents.

kNN

1.91%

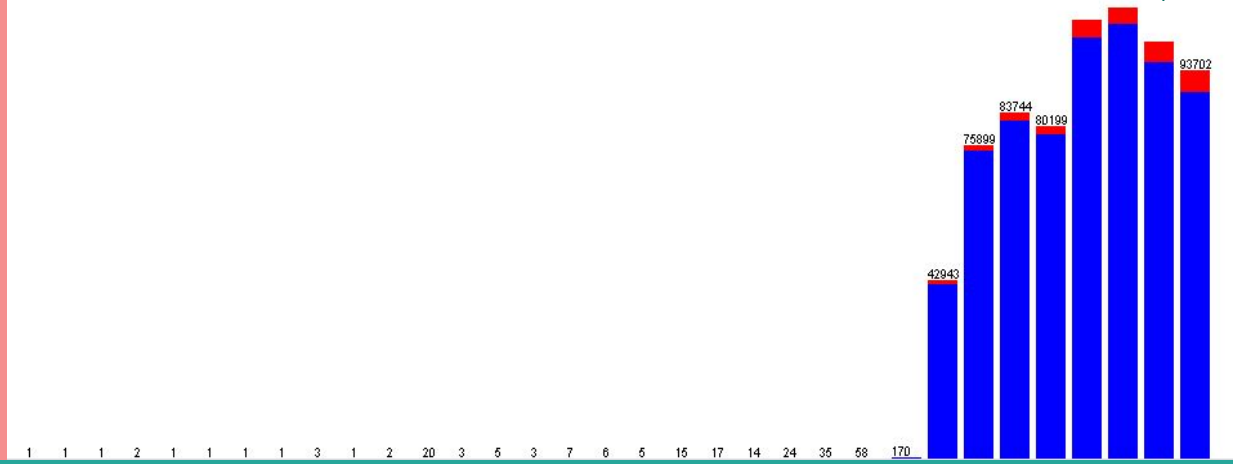
```
Time taken to test model on training data: 51823.52 seconds

=== Summary ===

Correctly Classified Instances      679686               98.1469 %
Incorrectly Classified Instances     12833                1.8531 %
Kappa statistic                          0.6609
Mean absolute error                      0.0192
Root mean squared error                  0.098
Relative absolute error                 27.632  %
Root relative squared error             52.6366 %
Total Number of Instances           692519

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.999     0.477     0.982      0.999     0.990      0.690    0.995     1.000     No
               0.523     0.001     0.932      0.523     0.670      0.690    0.995     0.838     Yes
Weighted Avg.  0.981     0.460     0.981      0.981     0.979      0.690    0.995     0.994

=== Confusion Matrix ===

     a      b   <-- classified as
 666678    946 |      a = No
  11887  13008 |      b = Yes
```

Drug Related Incidents are Increasing Over Time

Name: Year of Occurrence
Missing: 0 (0%)
Distinct: 33
Type: Nominal
Unique: 8 (0%)

No.	Label	Count	Weight
16	2004	7	7.0
17	2005	6	6.0
18	2006	5	5.0
19	2007	15	15.0
20	2008	17	17.0
21	2009	14	14.0
22	2010	24	24.0
23	2011	35	35.0
24	2012	58	58.0
25	2013	170	170.0
26	2014	42943	42943.0
27	2015	75899	75899.0
28	2016	83744	83744.0
29	2017	80199	80199.0
30	2018	105971	105971.0
31	2019	108899	108899.0
32	2020	100765	100765.0
33	2021	93702	93702.0

Class: Drug Related Incident (Nom)



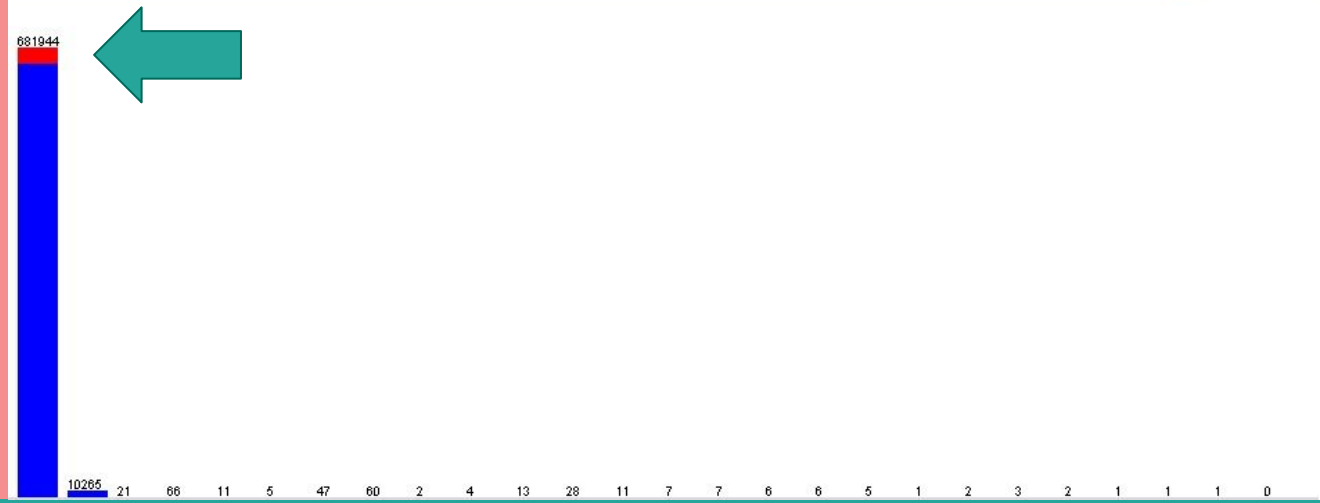
Drugs and Hate Crimes Don't Mix

Name: Hate Crime Description
Missing: 0 (0%)
Distinct: 25
Type: Nominal
Unique: 4 (0%)

No.	Label	Count	Weight
1	None	681944	681944.0
2	Unknown	10265	10265.0
3	Anti Male Homosexual (Gay)	21	21.0
4	Anti White	66	66.0
5	Anti Other Ethnicity/Natl Origin	11	11.0
6	Anti Transgender	5	5.0
7	Anti Homosexual (Gays and Lesbians)	47	47.0
8	Anti Black Or African American	60	60.0
9	Anti Catholic	2	2.0
10	Anti-Hawaiian Or Other Pacific Islander	4	4.0
11	Anti Jewish	13	13.0
12	Anti Hispanic	28	28.0
13	Anti Female Homosexual (Lesbian)	11	11.0
14	Anti Am. Indian/Alaskan Native	7	7.0
15	Anti Multi-Racial Group	7	7.0
16	Anti Islamic (Muslim)	6	6.0
17	Anti Protestant	6	6.0
18	Anti Arab	5	5.0

Class: Drug Related Incident (Nom)

Visualize All



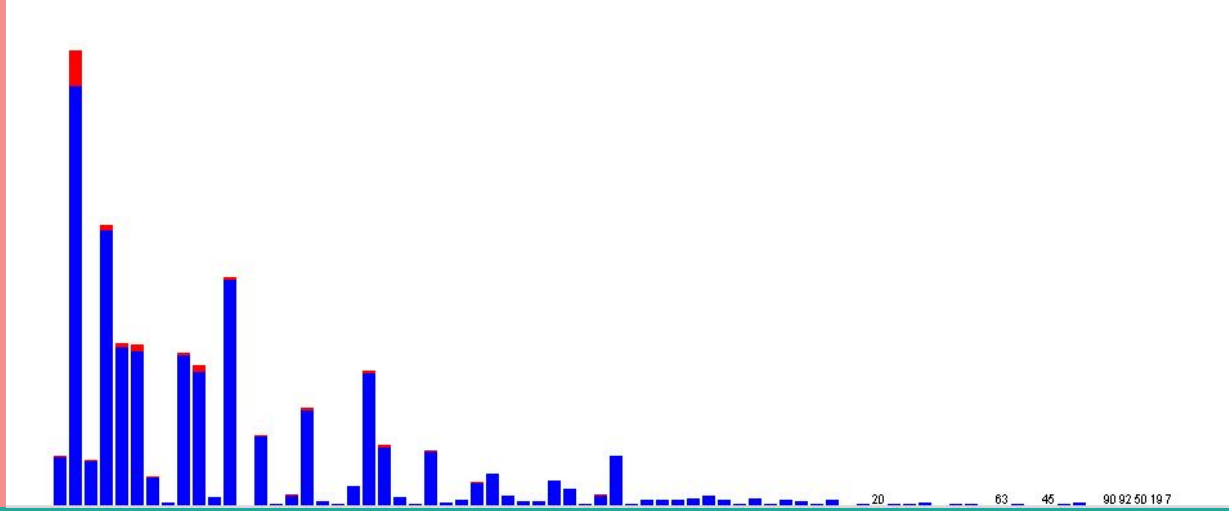
Most Drug Related Incidents Occur On Highways, Streets or in Alleyways

Name: Type of Location
Missing: 0 (0%)
Distinct: 73
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Commercial Property Occupied/Vacant	13778	13778.0
2	Highway, Street, Alley ETC	128819	128819.0
3	Gas or Service Station	12837	12837.0
4	Single Family Residence - Occupied	79603	79603.0
5	Apartment Complex/Building	45800	45800.0
6	Outdoor Area Public/Private	45346	45346.0
7	Hotel/Motel/ETC	8155	8155.0
8	Government/Public Building	775	775.0
9	Parking Lot (All Others)	43420	43420.0
10	Parking (Business)	39418	39418.0
11	Medical Facility	2632	2632.0
12	Apartment Parking Lot	64719	64719.0
13	Jail/Prison/Penitentiary/Corrections Fac	169	169.0
14	Retail Store	19677	19677.0
15	Rental Storage Facility	203	203.0
16	Government Facility	2877	2877.0
17	Other	27536	27536.0
18	Condominium/Townhome Building	1054	1054.0

Class: Drug Related Incident (Nom)

Visualize All



Conclusion

Our results had higher accuracy than we thought.

The data should be better cleaned.

- The dataset contains too many sources of error.
- Most of the information has no impact of our search
- The data being populated is too unreliable for critical fields