Crime and the city:
Spatial planning for crime reduction

Ruben Sharpe

March 22, 2021

Abstract

It is hypothesised that crime in a neighbourhood is correlated, albeit in a complex manner with the availability of public venues. Rather than hypothesising about causal relations, a machine learning approach is proposed in which high-crime, medium-crime and low-crime neighbourhoods are characterised by their public venues. These characterisations may serve as templates for good and bad practices in spatial planning. As a test case, the city of Tornonto (Ontario, Canada) was characterised based on information about public venues, obtained from Foursquare, and the City government of Toronto Open Dataset on crime.

Neighbourhoods were clustered by similarity of their composition in terms of venues using K-means clustering. Spatial information on crime was visualised in choropleth maps of Toronto. Visual comparison of neighbourhood clusters with crime maps shows the feasibility of the approach.

Decision trees and support vector machines were trained to classify neighbourhoods as high-, medium- or low-crime. With such models, city planners may try to predict how modifications would impact the crime classification or,in other words, which modifications correlate with what crime level.

Contents

1 Introduction

It is intuitively clear that public venues, such as bars and restaurants, but also public service locations such as transport hubs or banks impact the wellbeing of local communities. However, it is not straightforward to predict whether this impact will be positive or negative. Bars and restaurants can enliven a neighbourhood and may contribute to greater safety because of increased presence of people in the streets because potential witnesses deter crime. On the other hand, the presence of bars and restaurants is also correlated with alcohol abuse related crime. Similarly, there is a duality in the social impact for the presence of shops and banks. These may have a positive impact because of associated prosperity, but may also attract criminal activity at night because of reduced traffic compared to residential areas.

The complexity and the interrelatedness of these issues carries over to the complexity of city planning. The difficulty for city planners, moreover, is that they rarely get to design a neighbourhood from scratch. Typically, they have to deal with neighbourhoods with for the most part a fixed set of public facilities, and which may be suboptimal from a design point of view. The question for a city planner, therefore, is often : "What would be, *for a certain neighbourhood and given all its other attributes*, the impact of building (or removing) a shopping mall? Or a park? Or adding, or removing a bus stop?"

There is, of course, a lot of scientific literature on this topic but that is out of the scope of this project. This report is aimed at city planners and is intended to show the feasibility of machine learning as a tool for spatial planning.
Machine learning allowes an automated approach to compare neighbourhoods both in terms of their public venues and in terms of social markers, such as crime rates. This provides city planners, apart from their professional expertise, with examples of good and bad practices that may help them in their decision making. For reasons of convenience, the analysis in this report will look at the city of Toronto (Ontario, Canada). Crime rates will be used as the proxy for the wellbeing of communities, although other markers such as employment rate or health statistics could also be used.

2 Data

Location data, information about public venues and crime statistics will be used to classify neighbourhoods and map them with respect to the prevalance of crime. From this, generalised conclusions may be drawn with respect to the makeup of 'high-crime' and 'low-crime' neighbourhoods.

2.1 Location data

Location data for the city of Toronto can be acquired by scraping postal codes, boroughs and neighbourhoods from Wikipedia.[1] This data needs to be enriched with the GPS coordinates of the respective neighbourhoods. These coordinates can be obtained from the internet in the form of a '.csv' file from the Coursera website for the Data Science Capstone.[2] The enriched dataset consists of the postal codes, borough, and neighbourhood names and the neighbourhoods' latitude and longitude (Figure 1).

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 1: Example of data enriched with coordinates.

2.2 Public venues

Foursquare is used to obtain information about the public venues, specifically their category, in each neighbourhood (Figure 2). For this purpose, the neighbourhood is arbitrarily defined as the area within a 1.500 m radius around a postal code's geographical centre. The top 10 most prevalent venues will be used to characterise the neighbourhood (Figure 3).

2.3 Crime data

The city of Toronto provides a lot of open data related to the city.[3] This data ranges from polls conducted by the government, to inventory lists of street furniture, to 'bicycle count and locations' and is ever increasing. From this

---

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
[2] https://cocl.us/Geospatial_data
[3] https://www.toronto.ca/city-government/data-research-maps/open-data/

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	Malvern, Rouge	43.806686	-79.194353	Harvey's	43.800020	-79.198307	Restaurant
2	Malvern, Rouge	43.806686	-79.194353	Canadiana exhibit	43.817962	-79.193374	Zoo Exhibit
3	Malvern, Rouge	43.806686	-79.194353	RBC Royal Bank	43.798782	-79.197090	Bank
4	Malvern, Rouge	43.806686	-79.194353	Staples Morningside	43.800285	-79.196607	Paper / Office Supplies Store

Figure 2: Example of which venue information can be obtained.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Chinese Restaurant	Cantonese Restaurant	Caribbean Restaurant	Park	Gym / Fitness Center	Shopping Mall	Breakfast Spot	Coffee Shop	Bakery	Hong Kong Restaurant
1	Alderwood, Long Branch	Park	Coffee Shop	Grocery Store	Light Rail Station	Bank	Pizza Place	Burger Joint	Discount Store	Café	Moroccan Restaurant
2	Bathurst Manor, Wilson Heights, Downsview North	Park	Coffee Shop	Bank	Gas Station	Pizza Place	Ski Chalet	Sandwich Place	French Restaurant	Gym	Shopping Mall
3	Bayview Village	Bank	Trail	Park	Japanese Restaurant	Gas Station	Grocery Store	Café	Chinese Restaurant	Intersection	Restaurant
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Coffee Shop	Sushi Restaurant	Bakery	Bagel Shop	Pizza Place	Café	Pub	Restaurant	Sandwich Place

Figure 3: Example of neighbourhoods with top 10 most prevalent venues.

repository, a list of crime rates can be obtained.[4] This list contains, for each neighbourhood the number of incidences of certain types of crime (Table 1). Population information from the 2016 census, included in the file, can be used to normalise crime rates relative to population size. Conveniently, the file also contains geographical information so that this information can be easily mapped. The crime categories that are accessible from the data files are given in Table 1.

Category	available data
Assault	from $2014 - 2019$
Auto Theft	from $2014 - 2019$
Breaking and Entering	from $2014 - 2019$
Homicide	from $2014 - 2019$
Robbery	from $2014 - 2019$
Theft	from $2014 - 2019$

Table 1: Crime data categories.

---

[4]https://open.toronto.ca/dataset/neighbourhood-crime-rates/

3 Methodology

3.1 Exploration of crime information

Crime information is available for six categories. This information may be visualised in choropleth maps (Figure 4). Since there are six crime categories, a neighbourhood classification in 'high-', 'medium-' or 'low-crime' must either relate to one of the six categories, or a new crime index must be constructed.

(a) Assault

(b) Auto theft

(c) Breaking and entering

(d) Homicide
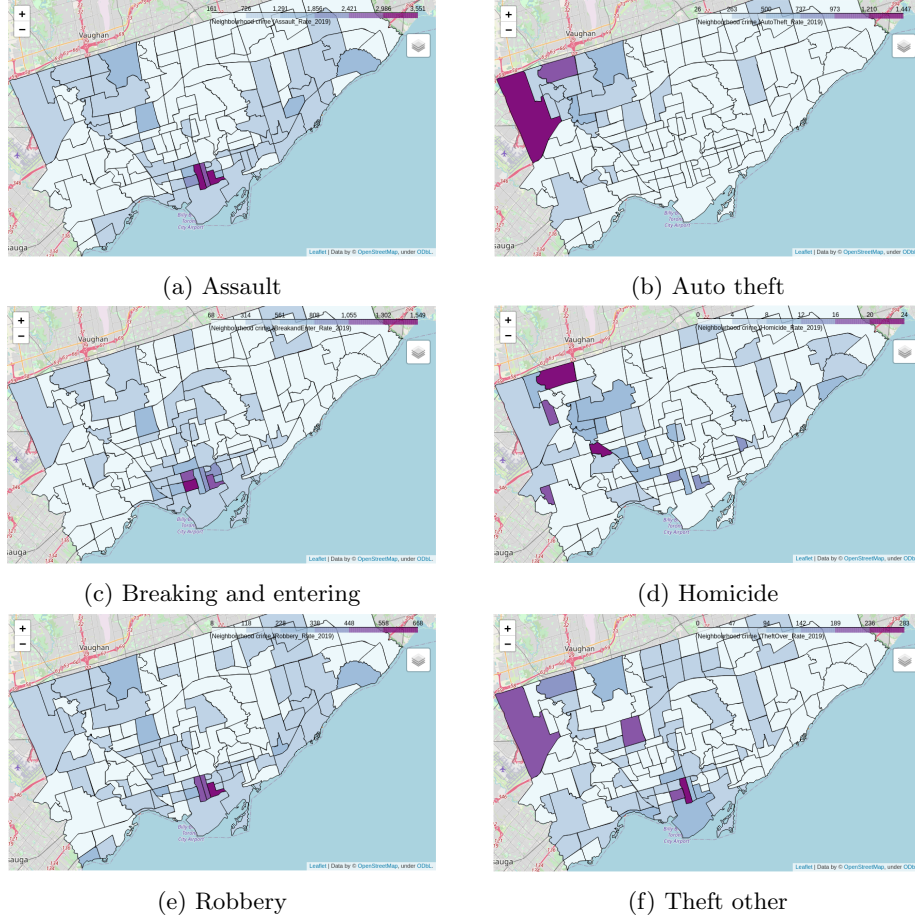
(e) Robbery

(f) Theft other

Figure 4: The six crime categories and their spatial distribution (2019 crime rates).

In order to create a crime index, each crime category is first normalised with respect to the maximum number of occurences across all neighbourhoods. Then these indices are summed over all the six crime categories and normalised again with respect to the maximum of this new 'overall' category. The index can be interpreted as follows: an index of '1' means that the neigbourhood has scored,

on average, the worst in all crime categories, whereas an index of '0' means that there have been no incidences of crime in any of the categories in that neighbourhood. Therefore, areas with indices up to 0.3 are labelled 'low-crime', areas with indices between 0.3 and 0.7 are labelled 'medium-crime' and areas with indices higher than 0.7 are labelled 'high-crime'.

For convenience, the 2019 crime rate for each crime category is used. These are already represented as incidents per 100.000 population and therefore the population size of each neigbourhood does not need to be taken into account.

By visually comparing the crime index choropleth with the spatial distribution of the neigbourhoods in the venue dataset, the latter can be manually labeled with respect to the crime index (Figure 5).
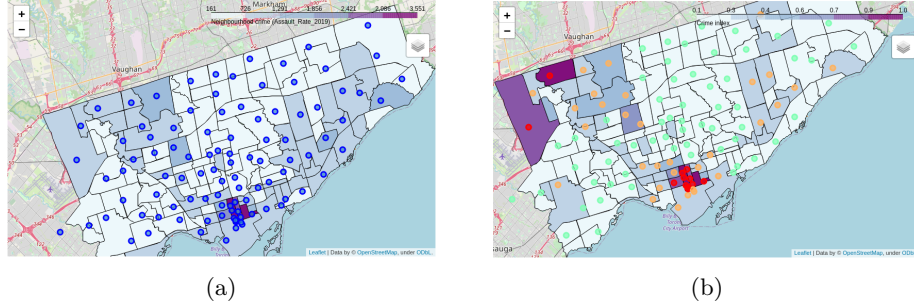

(a)   (b)

Figure 5: Spatial correlation of crime with neighbourhood information (a). Neighbourhoods that lie on purple areas may be manually labeled 'high-crime' (red dots), neighbourhoods on blue areas 'medium-crime' (yellow dots) and neighbourhoods on light blue areas 'low-crime' (light blue dots)(b).

3.2 Machine learning

3.2.1 Unsupervised clustering of neighbourhoods and visual correlation with crime

Similarity of neighbourhoods with respect to their public venues, as well as the spatial distribution of similar neighbourhoods is explored using K-means clustering. A qualitative correlation between the neighbourhood clusters and crime can be obtained by visual inspection of crime and cluster maps. Different clusters that spatially correlate with different crime indices indicate a correlation between neighbourhood composition and crime. The more dissimilar two clusters are, the more intuitive it is to make future predictions for a given neigbourhood composition.

An idea of the intra- and inter-cluster similarities can be obtained by inspecting the output of the clustering algorithm for increasing numbers of clusters. From this, it can be observed which clusters break up first and which stay intact. Clusters that break off from larger clusters are siblings and share some inter-cluster similarity. Clusters that resist breaking up, indicate high intra-cluster

cohesion (i.e. greater internal similarity). Because the most distinct clusters are the most intuitively useful, it is best to focus on non-sibling clusters with large intra-cluster similarity.

3.2.2 Modelling

Whereas K-means clustering may be used to qualitatively classify clusters of varying venue composition by crime index, it is not necessarily obvious how changes in the neighbourhood composition (of a specific neighbourhood) cause its classification to change. In other words, it is from K-means clustering not obvious which changes would cause a certain neighbourhood to become more or less similar to neighbourhoods with a more desirable crime rating. For this reason, a decision tree model will be built. Such a model will allow parsing of changes to a neighbourhood's composition and a prediction of the new crime rating. Decision trees are also very intuitive and its explicit structure facilitates the up front estimation of which changes the outcome will be most sensitive to. For reasons of comparison, however, a support vector machine classifier will also be built.

4 Results

4.1 Visual inspection

The K-means algorithm has been run to create 3, 4 , 5 and 7 clusters of neighbourhoods (Figure 6). From this, it can be seen that there are basically two large clusters with high inter-cluster similarity: one centred around the harbour and one broad band from east to west. Increasing the number of clusters causes smaller break-away clusters, mostly from the second ('broad band') cluster. This indicates that for higher numbers of clusters, the main division will still be between the 'harbour cluster' and the other clusters that share some commonality with the original 'broad band' cluster. The nature of the dissimilarity that causes this division, however, is not intuitively clear from inspection of the most prevalent venues in each group ('harbour cluster' versus 'broad band offspring clusters') (Figure 7).

Spatially, the clusters correlate somewhat with the crime map, insofar that a relatively high crime index is found in the neighbourhoods around the harbour and in the North-West corner of Toronto. This agrees reasonably well with the spatial distribution of the harbour cluster, which is marked with red dots in Figure 8. A significant amount of neighbourhoods belonging to the harbour cluster, moreover are found in low-crime areas. The cluster distribution, furthermore, does not clearly discriminate between medium and low crime indices which both are equally well covered by the broad band cluster, which is marked with blue dots in Figure 8.

(a) Three clusters   (b) Four clusters

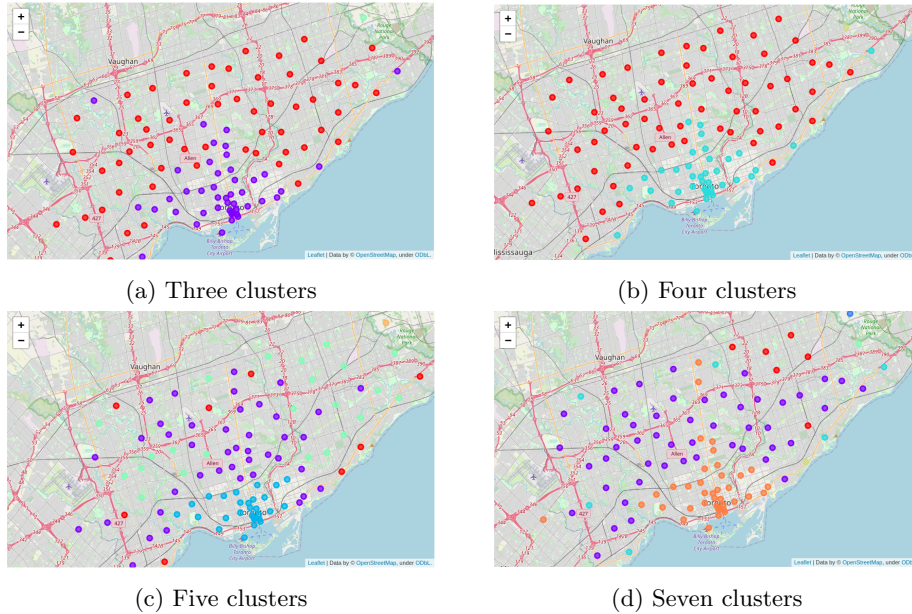(c) Five clusters   (d) Seven clusters

Figure 6: Different distributions of three, four, five and seven clusters. Note that there are basically two stable clusters, concentrically oriented around the harbour. Increasing the number of clusters only causes small split offs.

Accuracy type	Decision tree	SVM
Overall	0.76	0.57
Jaccard 'high'	0.25	0.0
Jaccard 'medium'	0.56	0.29
Jaccard 'low'	0.77	0.59

Table 2: Accuracy of the decision tree predictions (on a test set).

4.2 Decision tree and SVM

A decision tree is built with an overall test accuracy of 0.76. This indicates that 76% of a test set has been correctly predicted. However, the number of high-crime neighbourhoods is far smaller than the number of medium- and low-crime neighbourhoods so that a high overall accuracy can still be obtained with a high number of mislabelling of the high-crime neighbourhoods. In fact, the Jaccard accuracy for high-crime is only 0.25 while the Jaccard accuracy for medium-crime is 0.56 and that for low-crime is 0.77 (Table 2). This is also clearly visible when the decision tree is trained on the full set of neigbourhoods (with an overall training accuracy of 0.92) and the newly predicted labels are plotted on the crime index choropleth (Figure 10).

For reasons of comparision, also a support vector machine (SVM) has been

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Rouge Hill, Port Union, Highland Creek	0	Park	Hotel	Italian Restaurant	Burger Joint	Neighborhood	Grocery Store	Gym	Gym / Fitness Center	Breakfast Spot	Optical Shop
9	Birch Cliff, Cliffside West	0	Park	Ice Cream Shop	Filipino Restaurant	Gym Pool	Thai Restaurant	Golf Course	Café	General Entertainment	Restaurant	Diner
40	East Toronto, Broadview North (Old East York)	0	Greek Restaurant	Café	Coffee Shop	Pizza Place	Park	Ice Cream Shop	Ethiopian Restaurant	Bakery	Thai Restaurant	Brewery
41	The Danforth West, Riverdale	0	Greek Restaurant	Café	Park	Coffee Shop	Bakery	Pub	Pizza Place	Ice Cream Shop	Flower Shop	Vietnamese Restaurant
42	India Bazaar, The Beaches West	0	Park	Coffee Shop	Brewery	Indian Restaurant	Bakery	Café	Grocery Store	Beach	BBQ Joint	Restaurant

(a) Venue composition of first five neighbourhoods in the 'harbour' cluster.

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Malvern, Rouge	2	Zoo Exhibit	Pizza Place	Fast Food Restaurant	Restaurant	Spa	Gas Station	Liquor Store	Caribbean Restaurant	Chinese Restaurant	Bank
2	Guildwood, Morningside, West Hill	2	Pizza Place	Breakfast Spot	Fast Food Restaurant	Coffee Shop	Bank	Sports Bar	Food & Drink Shop	Greek Restaurant	Grocery Store	Gym / Fitness Center
3	Woburn	2	Coffee Shop	Pizza Place	Fast Food Restaurant	Pharmacy	Sandwich Place	Bakery	Bus Line	Supermarket	Beer Store	Filipino Restaurant
4	Cedarbrae	2	Restaurant	Coffee Shop	Clothing Store	Indian Restaurant	Sandwich Place	Gas Station	Bakery	Pharmacy	Bank	Grocery Store
5	Scarborough Village	2	Sandwich Place	Pharmacy	Bank	Ice Cream Shop	Breakfast Spot	Wings Joint	Grocery Store	Pizza Place	Coffee Shop	Bookstore

(b) Venue compostition of the first five neighbourhoods in the 'broad band' cluster.

Figure 7: Compositions of the most stable clusters. Note that although these are stable and therefore very dissimilar, the dissimilarity is not intuitively obvious.

trained. The overall test accuracy of the SVM was 0.57, which is significantly lower than that of the decision tree. The Jaccard scores for high-, medium- and low-crime were 0(!), 0.29 and 0.59 respectively. Again, those are clearly worse than the decision tree predictions.

5 Discussion

5.1 Correlating neighbourhood composition with high-crime areas

By visual inspection of similar neighbourhoods and crime data, a correlation between neigbourhood composition and crime rate may be inferred (Figure 6). The cluster with the most internal cohesion (i.e. which is most resistant to breaking up) is situated concurrend with high-crime neigbourhoods and therefore its composition may be qualitatively interpreted as indicative of higher occurences of crime.

Although the cluster that correlates most strongly with high-crime areas is clearly distinct from the other clusters that correlate with medium- and low-crime areas, the distinction is not intuitive (Figure 7). This means that a lot of professional expertise is required in order to use these characterisations to predict which manipulations of the neigbourhood composition might be favorable.

Medium- and low-crime areas correlate less strongly with neigbourhood clusters than high-crime areas. Moreover, any such clusters that do correlate with
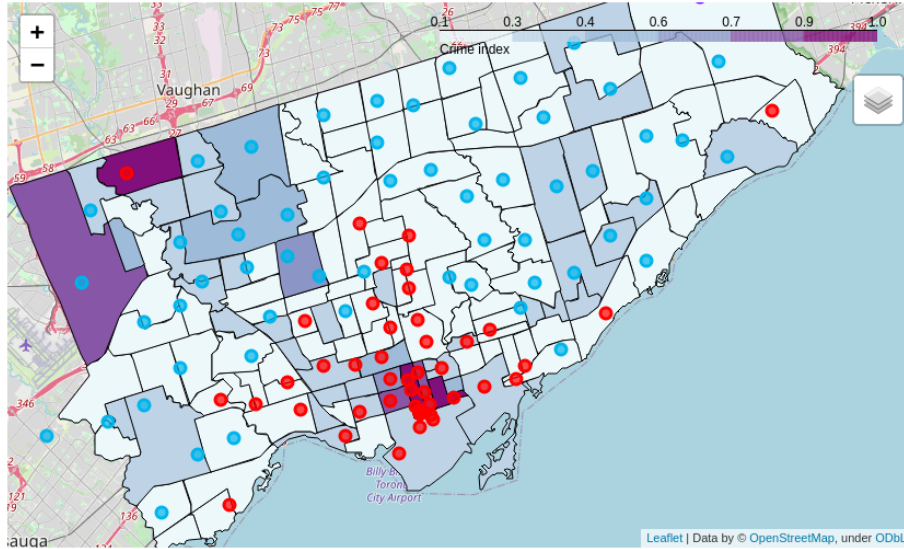
9

Figure 8: Spatial distribution of crime and of clusters with a similar public venue composition.

medium- and low-crime areas are siblings, i.e. they are more similar to each other than they are to the high-crime cluster. It may be expected that these are therefore even harder to distinguish between each other and less intuitive to help guide city planning decisions.

5.2 Correlating neighbourhood composition with medium- and low-crime areas

Whereas K-means clustering and visual inspection favour a qualitative correlation between neighbourhood composition and high-crime areas, the oposite is true for decision trees and SVMs. Because the vast majority of the neigbourhoods in Toronto can be classified as either low-, or medium-crime, the algorithm is skewed toward those. Specifically low crime neigbourhoods can be predicted with relatively high accuracy but this accuracy quickly deteriorates toward high-crime. This is, however, not necessarily a problem for city planners. Since they most likely aim to improve a neigbourhood, they would be most interested in how to increase its similarity with low-crime areas (starting from an already known higher crime composition).

Compared to SVMs, decision trees have been found to be the most accurate. Because of their explicit decision structure they are also the most helpful for planning interventions.
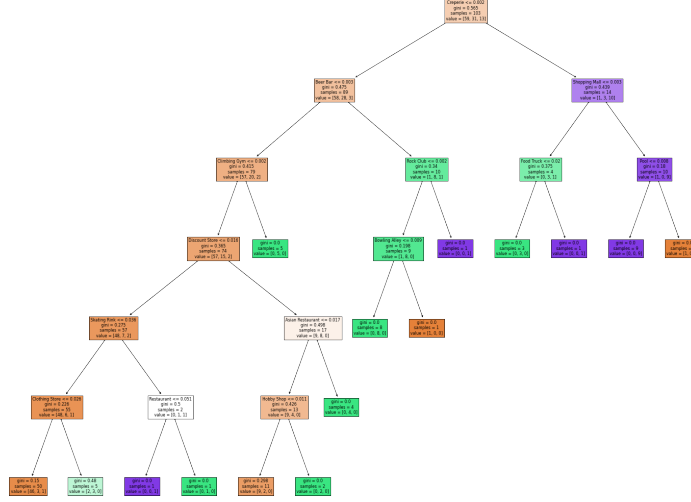
10

Figure 9: Although illegible, the graphical representation of the decision tree illustrates the explicit nature of the decisions that lead to a certain classification. These may therefore be used to predict to which tweaks this decision is particularly sensitive.

6   Conclusion

Machine learning, specifically K-means clustering, decision trees and SVMs have been explored for their feasibility to support city planners to reduce crime. It has been shown that this is indeed the case, albeit with limitations.

The useability of K-means clustering relies on the ability of a human interpreter to make useful distinctions. In this study, the meaningful differences were found to be small. This maybe due to a lack of distinctive features in the top 10 most prevalent venues for each neigbourhood. For further study it is advisable to look deeper into the dataset for features that may be less prevalent but more distinctive.

Decision trees, although relatively accurate, are heavily skewed in favour of the most abundant category in the training set. In this study, the training set was a random sample of the actual Toronto neighbourhoods and therefore reflected the city's low-crime bias. For future study it may be interesting to selectively sample for a more proportional training set.
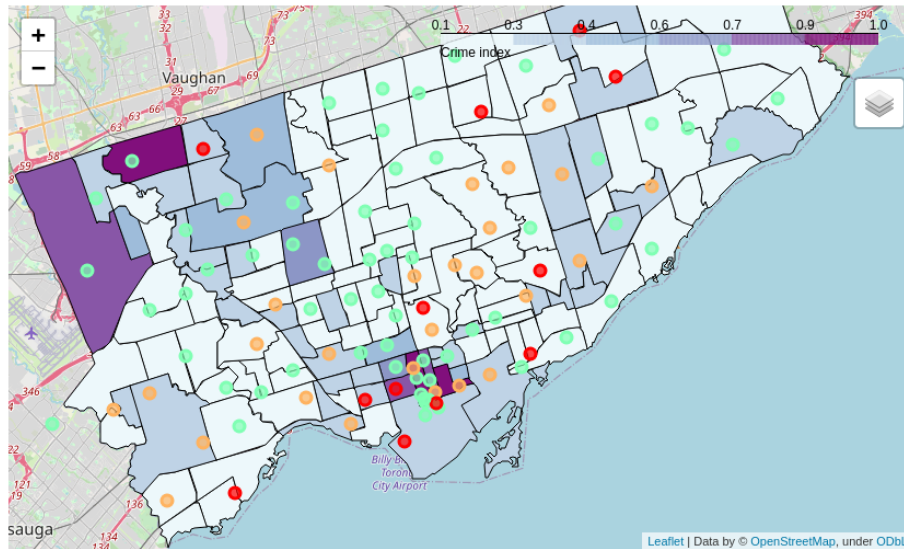
11

Figure 10: Graphical representation of the correlation between the predicted (via decision tree) and observed crime indices. Mislabelling of high-crime neighbourhoods is most frequent (red dots), whereas low crime neighbourhoods were most accurately predicted (light blue dots).

Overall it may be concluded that machine learning may be a powerfull tool in a city planner's toolkit. For now, however, a solid professional expertise is still required.