THE BATTLE OF NEIGHBORHOODS

COURSERA CAPSTONE PROJECT

MARCH 26, 2020

SAID MANCOURI AZZOUZI

Introduction

Well in each city in the world has a lot of Neighborhood, and each one has its structure based on their avenues. Having a neighborhood doesn't mean that you can know it's city, and this is the major question we're going to ask. Can we predict the city of a neighborhood based on its avenues? Our study will be based on 3 cities: New York, Toronto and Paris.

Both cities are very diverse and are the financial capitals of their respective countries. So we will study them and after we will try to predict the city of some neighborhood.

To understand globalization and the impact of the cultural exchange in this world we need to understand the structure of a city and its neighborhoods and how avenues sometimes new ones may impact the way of living.

For people who are studying cities and their avenues and want to understand the relationship between them, for people who are interested to know to which city their neighborhood would fit. and how can we make an avenue looks like another to one based on their avenues?

Data

Data used in this project is exported from various sources:

- New York : https://geo.nyu.edu/catalog/nyu_2451_34572
- Toronto : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Paris : https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=json&timezone=Africa/Lagos&lang=fr

We will use the Foursquare API to explore neighborhoods in New York, Toronto and Paris. We will use the explore function to get the most common venue categories in each neighborhood

First we get extract from the 3 datasets these columns

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Introducing Venues on data

In this part, we will add to our datasets Venues data collected from FourSquare. for each eighborhood we will explore it and get all the venues there. We will explore data and prepare it for analyse. the data used to solve this problem is geolocation data collected from FourSquare. Data is composed of 3 dataframes. Explanation of the location data is latitude and longitude.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park , Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park , Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park , Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park , Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park , Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

Transforming Venue Category to dummy variables

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	R
0	Amérique	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	
1	Archives	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.04	0.01	0.000000	0.0	
2	Arsenal	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	
3	Arts-et-Métiers	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.01	0.02	0.00	0.000000	0.0	
4	Auteuil	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	
5	Batignolles	0.0	0.0	0.00000	0.0	0.01	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	
6	Bel-Air	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	
7	Belleville	0.0	0.0	0.02439	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	
8	Bercy	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.013333	0.0	
9	Bonne-Nouvelle	0.0	0.0	0.00000	0.0	0.00	0.0	0.0	0.00	0.00	0.00	0.000000	0.0	

Getting the most common Venue on each neighborhood

That will allow us to understand more our data and what's common thing in different Neightborhood in the 3 cities

	Neighborhood	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	Amérique	French Restaurant	Supermarket	Café	Bakery	Plaza	Park	Bed & Breakfast	Smoke Shop	Bistro	Theater
1	Archives	French Restaurant	Hotel	Coffee Shop	Clothing Store	Italian Restaurant	Bar	Art Gallery	Japanese Restaurant	Cocktail Bar	Tapas Restaurant
2	Arsenal	French Restaurant	Hotel	Plaza	Park	Tapas Restaurant	Boat or Ferry	Italian Restaurant	Cocktail Bar	Gastropub	Pedestrian Plaza
3	Arts-et-Métiers	French Restaurant	Hotel	Cocktail Bar	Wine Bar	Italian Restaurant	Chinese Restaurant	Restaurant	Moroccan Restaurant	Bar	Japanese Restaurant
4	Auteuil	Tennis Court	Museum	Garden	Botanical Garden	Bike Rental / Bike Share	Racecourse	French Restaurant	Stadium	Sporting Goods Shop	Outdoors & Recreation

After understanding the data in the 3 datasets we will create our final data set which contains data with dummy variables with the label 'City'

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	Alsatian Restaurant	American Restaurant	Animal Shelter	Antique Shop	Aquar
0	Amérique	0.0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	0	0.0	
1	Archives	0.0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	0	0.0	
2	Arsenal	0.0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	0	0.0	
3	Arts-et-Métiers	0.0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	0	0.0	
4	Auteuil	0.0	0	0.0	0.0	0	0	0	0	0	0.0	0.0	0	0.0	

Methodology

To get for each neighborhood we will use 2 algorithms:

- Logistic regression.
- Decision Tree

We will use them both to compare our prediction and also to understand with the decision tree why a neighborhood is most likely to belong to a city.

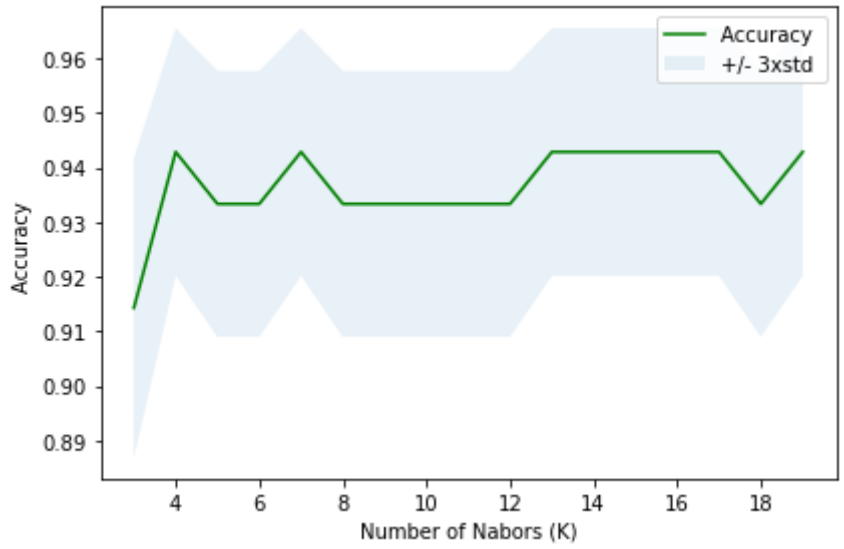
Train Test Split

To do so first we will do the Train Test Split.

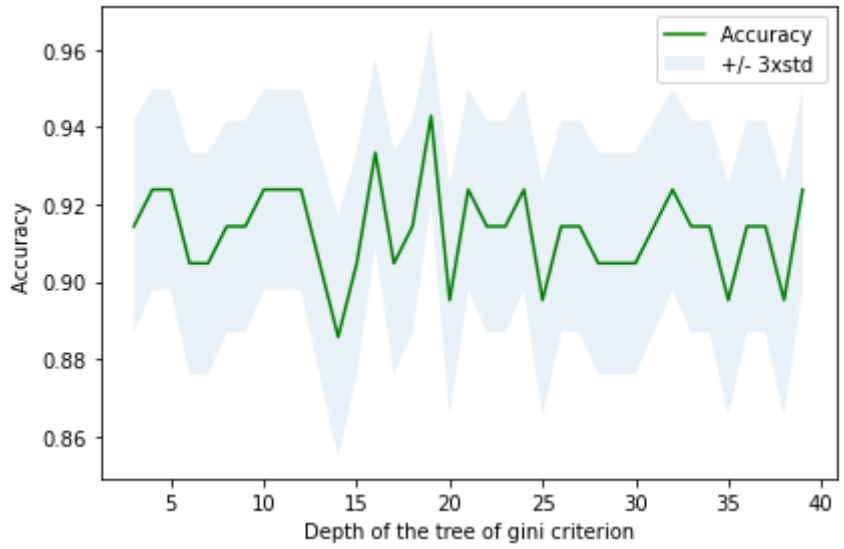
This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that have been used to train the data. It is more realistic for real world problems.

Tuning our prediction

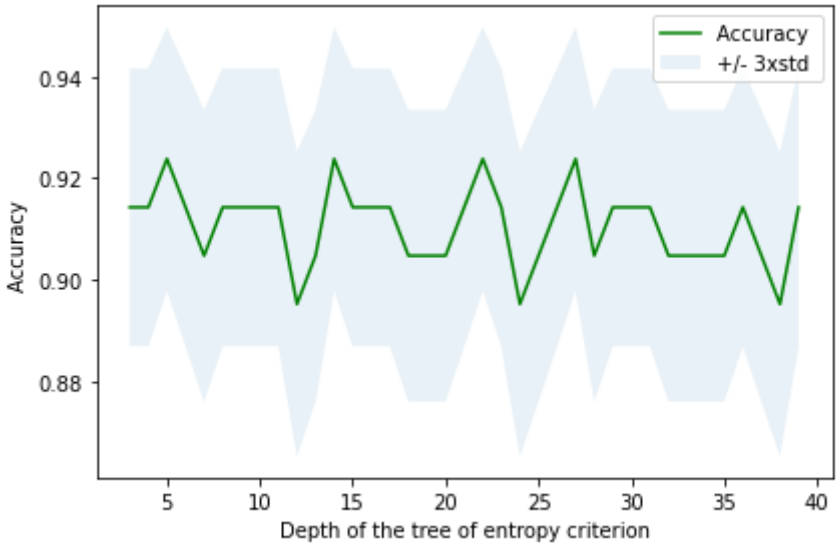
KNN Algorithm



Decision Tree gini Criterion



Decision Tree entropy criterion



Results

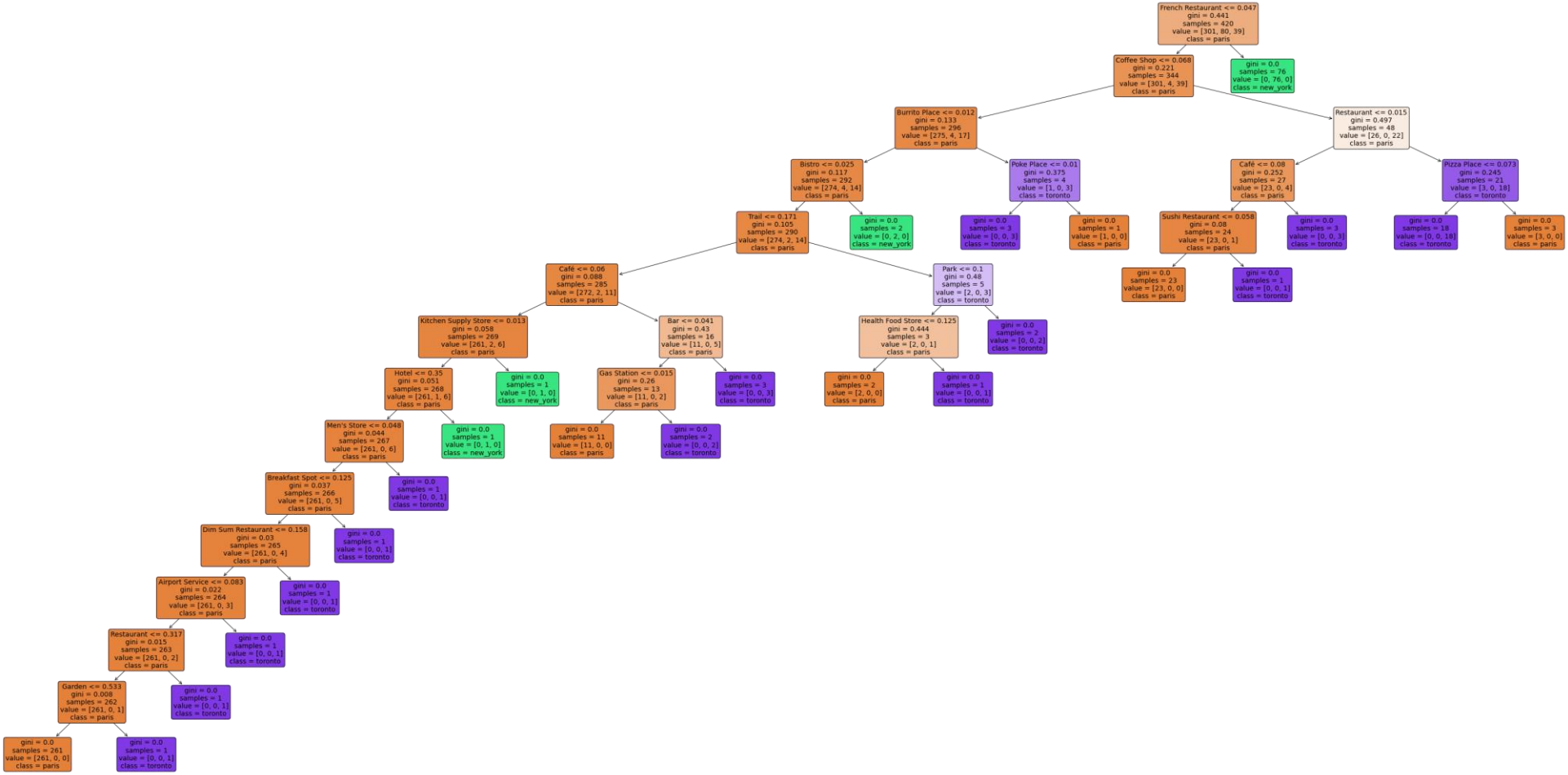
For logistic regression we will use K=16

Those are the false predictions.

Out[60]:

	real_data	predicted_data	Neighborhood
383	toronto	new_york	Business reply mail Processing CentrE
47	paris	new_york	Parc-de-Montsouris
408	toronto	new_york	Roselawn
403	toronto	new_york	Parkdale , Roncesvalles
206	new_york	toronto	Heartland Village
417	toronto	new_york	The Danforth West , Riverdale

For Decision Tree we will use Depth = 19


French Restaurant <= 0.047
gini = 0.441
samples = 420
value = [301, 80, 39]
class = paris
Coffee Shop <= 0.068
gini = 0.221
samples = 344
value = [301, 4, 39]
class = paris
gini = 0.0
samples = 76
value = [0, 76, 0]
class = new_york
Burrito Place <= 0.012
gini = 0.133
samples = 296
value = [275, 4, 17]
class = paris
Restaurant <= 0.015
gini = 0.497
samples = 48
value = [26, 0, 22]
class = paris
Bistro <= 0.025
gini = 0.117
samples = 292
value = [274, 4, 14]
class = paris
Poke Place <= 0.01
gini = 0.375
samples = 4
value = [1, 0, 3]
class = toronto
Café <= 0.08
gini = 0.252
samples = 27
value = [23, 0, 4]
class = paris
Pizza Place <= 0.073
gini = 0.245
samples = 21
value = [3, 0, 18]
class = toronto
Trail <= 0.171
gini = 0.105
samples = 290
value = [274, 2, 14]
class = paris
gini = 0.0
samples = 2
value = [0, 2, 0]
class = new_york
gini = 0.0
samples = 3
value = [0, 0, 3]
class = toronto
gini = 0.0
samples = 1
value = [1, 0, 0]
class = paris
Sushi Restaurant <= 0.058
gini = 0.08
samples = 24
value = [23, 0, 1]
class = paris
gini = 0.0
samples = 3
value = [0, 0, 3]
class = toronto
gini = 0.0
samples = 18
value = [0, 0, 18]
class = toronto
gini = 0.0
samples = 3
value = [3, 0, 0]
class = paris
Café <= 0.06
gini = 0.088
samples = 285
value = [272, 2, 11]
class = paris
Park <= 0.1
gini = 0.48
samples = 5
value = [2, 0, 3]
class = toronto
gini = 0.0
samples = 23
value = [23, 0, 0]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
Kitchen Supply Store <= 0.013
gini = 0.058
samples = 269
value = [261, 2, 6]
class = paris
Bar <= 0.041
gini = 0.43
samples = 16
value = [11, 0, 5]
class = paris
Health Food Store <= 0.125
gini = 0.444
samples = 3
value = [2, 0, 1]
class = paris
gini = 0.0
samples = 2
value = [0, 0, 2]
class = toronto
Hotel <= 0.35
gini = 0.051
samples = 268
value = [261, 1, 6]
class = paris
gini = 0.0
samples = 1
value = [0, 1, 0]
class = new_york
Gas Station <= 0.015
gini = 0.26
samples = 13
value = [11, 0, 2]
class = paris
gini = 0.0
samples = 3
value = [0, 0, 3]
class = toronto
gini = 0.0
samples = 2
value = [2, 0, 0]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
Men's Store <= 0.048
gini = 0.044
samples = 267
value = [261, 0, 6]
class = paris
gini = 0.0
samples = 1
value = [0, 1, 0]
class = new_york
gini = 0.0
samples = 11
value = [11, 0, 0]
class = paris
gini = 0.0
samples = 2
value = [0, 0, 2]
class = toronto
Breakfast Spot <= 0.125
gini = 0.037
samples = 266
value = [261, 0, 5]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
Dim Sum Restaurant <= 0.158
gini = 0.03
samples = 265
value = [261, 0, 4]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
Airport Service <= 0.083
gini = 0.022
samples = 264
value = [261, 0, 3]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
Restaurant <= 0.317
gini = 0.015
samples = 263
value = [261, 0, 2]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
Garden <= 0.533
gini = 0.008
samples = 262
value = [261, 0, 1]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto
gini = 0.0
samples = 261
value = [261, 0, 0]
class = paris
gini = 0.0
samples = 1
value = [0, 0, 1]
class = toronto

Discussion

This method can be used for different goals:

- we can predict the city based on the avenues
- the error of prediction can give us idea about how similar an avenue to others from in another city
- From this Tree we can understand more how we can make a neighborhood similar to one in a country. If this method used with many cities we can go further and why not make a city that have neighborhood that would make you feel like if you're living in another one.

If we add more cities we will have more understanding of how cities are constructed.