1. Introduction

For an entrepreneur, choosing the location of a new establishment within a city can be a very important and also very difficult task. For this, it is advisable to have as much information as possible from each neighborhood.

Similarly, the city government also needs as much information as possible from each neighborhood to manage them properly.

Amongst all information about neighborhoods, one that stands out is its prevailing social class. The needs and opportunities of a neighborhood are often associated with this information.

Here, we will seek to develop a model capable of predicting the prevailing social class of each neighborhood, based on the categories of venues there. This model will be trained with data from the set of reports of venues in each neighborhood, retrievable from the Foursquare API and with the data a report from UFMG (University of Minas Gerais) that informs the majority social class of each neighborhood.

If the model works well, we may use it to find out a valuable information of neighborhood on cities similar to Belo Horizonte that hasn't a report about their neighborhood prevailing social class.

2. Data acquisition and cleaning

The report about the Belo Horizonte neighborhood prevailing social classes is published in PDF format. Fortunately, it is very easy to copy the data contents and past into a csv file. The resulting columns are "Neighborhood" and "Class". Let's see the head of this data set.

	Neighborhood	**Class**
0	AARAO REIS	low
1	ALTO DOS PINHEIROS	low
2	ALTO PARAISO	low
3	ALVARO CAMARGOS	low
4	ALVORADA	low

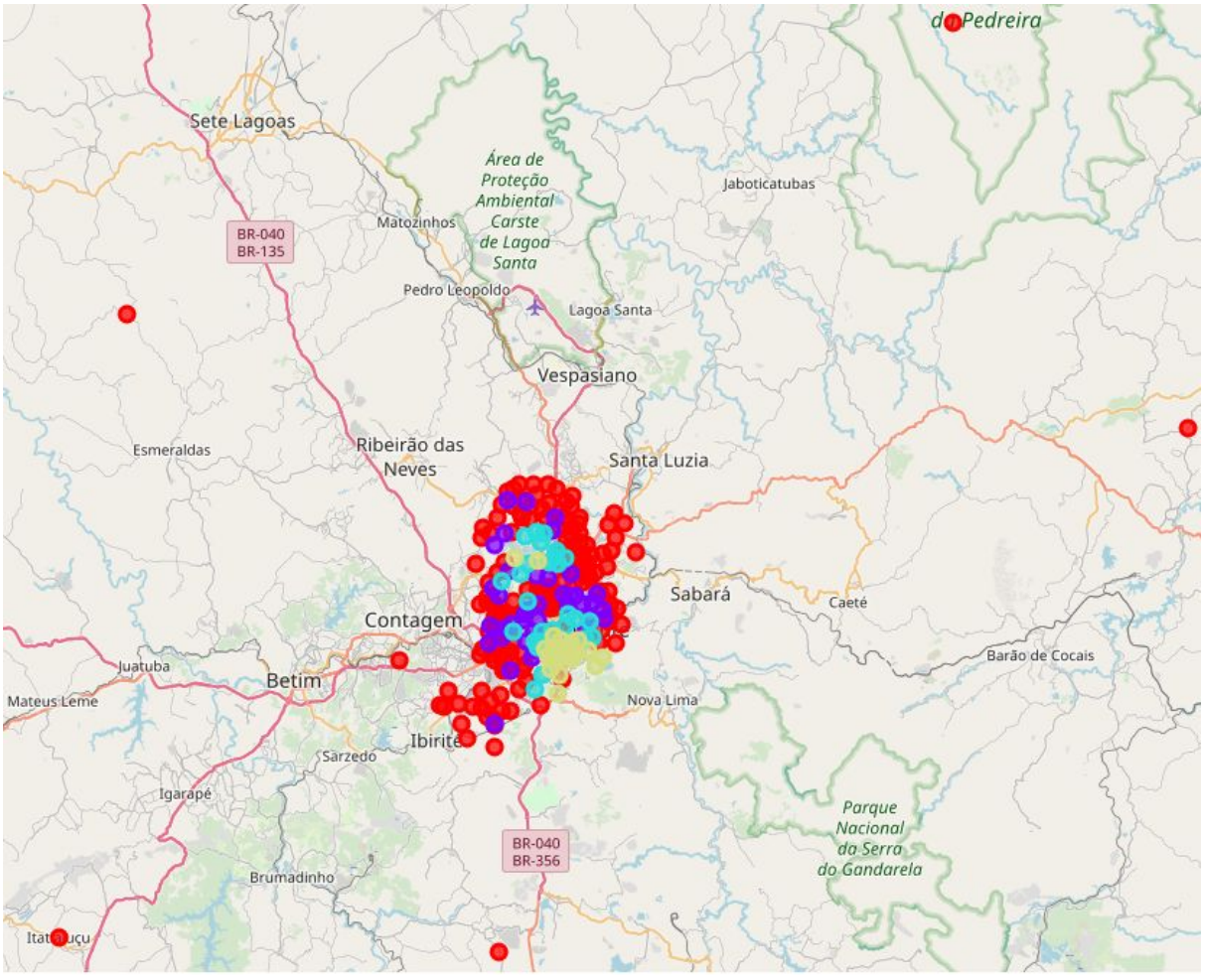
...

We retrieve the Neighborhood venues with Foursquare API, by calling the "query" endpoint for each Neighborhood, which requires localization data. The *geopy* library is used to retrieve localization data of each Neighborhood.

	Neighborhood	Class	Latitude	Longitude
0	AARAO REIS	low	-19.847221	-43.919508
1	ALTO DOS PINHEIROS	low	-19.932567	-44.004875
2	ALVARO CAMARGOS	low	-19.916339	-44.007857
3	ALVORADA	low	-30.031715	-51.049711
4	ANA LUCIA	low	-19.887783	-43.906368

…

Unfortunately, the geopy library sometimes can't be accurate, so we had to remove the Neighborhoods that couldn't have its localization data accurately retrieved. The identification of such cases was map manually with a map visualization support.



By viewing the map it's clear that geopy defined many points out of bounds of Belo Horizonte. Besides that, as I live in Belo Horizonte, I could detect some neighborhoods far from downtown that were inacurately defined by geopy.

So, we chose to restrict the analysis to neighborhoods not far from downtown.

Also, the location of "Pindorama" neighborhood, near from downtown, is remarkably wrong. So it will be removed too.
Now we have the Latitude and Longitude for each neighborhood, so we are ready to make Foursquare API calls on "query" endpoint.

Actually, for each neighborhood we made 4 API calls, one for each category: "food", "stores and services" and "professional", and one for all categories combined.

So there will be 4 resulting datasets, one for each category, plus one for all categories combined, and it will look like this:

	Neighborhood	Venue	Latitude	Longitude	Venue Category
0	AARAO REIS	Chapa Mágica	-19.845448	-43.921754	BBQ Joint
1	AARAO REIS	Burger King	-19.846823	-43.919360	Fast Food Restaurant
2	AARAO REIS	Celo Burguer	-19.847524	-43.919394	Burger Joint
3	AARAO REIS	bobs	-19.846710	-43.917326	Burger Joint
4	AARAO REIS	Padaria Vila Verde	-19.847362	-43.921778	Bakery

3. Methodology

So now we have 4 clean datasets:

- bh_food_venues
- bh_stores_venues
- bh_pro_venues
- bh_all_venues

They will be our asset for training our model using classification algorithms - the venues categories will be its features (after applying *onehot encoding*) and the social class will be the target variable.

The resulting models will be evaluated and we will show the best dataset and best classification algorithm for our goal.

As we are going to use the Venues categories as features of our classification algorithms, it's appropriate to avoid neighborhoods with small number of venues, because it's high potential to become outliers.

Unfortunately, the dataset bh_stores_venues become too small after that restriction, so it will be discarded.

Now we apply *onehot encoding* and drop columns that are not features or target on each remaining datasets.

Then we split training set with test set and build KNN, SVM and Logistic regression models.

The target (y) will be tested in the following formats:

- the actual class
- if the class == 'luxury'
- if the class == 'high'
- if the class == 'regular'
- if the class == 'low'

4. Results

As mentioned before, we built KNN, SVM and Logistic regression models, but we will show only the results obtained by Logistic regression, because it gets the better scores (jaccard index score) in most cases.

4.1 All venue categories dataset scores

- the actual class: 0.6666
- if the class == 'luxury': **0.9**
- if the class == 'high': 0.8666
- if the class == 'regular': 0.3333
- if the class == 'low': **0.9**

4.2 Food venue categories dataset scores

- the actual class: 0.6923
- if the class == 'luxury': 0.8
- if the class == 'high': **0.96**
- if the class == 'regular': 0.3076
- if the class == 'low': **0.923**

4.3 Professional venue categories dataset scores

- the actual class: 0.5714
- if the class == 'luxury': 0.8571
- if the class == 'high': 0.8928
- if the class == 'regular': 0.4286
- if the class == 'low': 0.8214

5. Discussion

It's interesting to see that the food venue categories dataset got the best overall results but closely followed by the all categories which indicates that may be possible to combine two or more categories to get optimistic scores, as it's clear that there are categories that disturbs the score (see the professional venue categories dataset).

Besides that, it's also interesting to see that the model built with foods venues categories dataset can predict the 'high' and 'low' class remarkably well, and definately could be used in a different city, similar to 'Belo Horizonte'.

6. Conclusion

Even though we couldn't get a great model to predict the actual class of a Neighborhood, we could get interesting results on predicting the 'high' and 'low' classes using the Food categories dataset, and predicting 'luxury' and 'low' casses using the All categories dataset.