From Multidimensional Poverty Index by block to average by urban neighborhood using free data from foursquare.

A case Study in Medellín city, Colombia

Rodriguez, Joan

January, 2021.

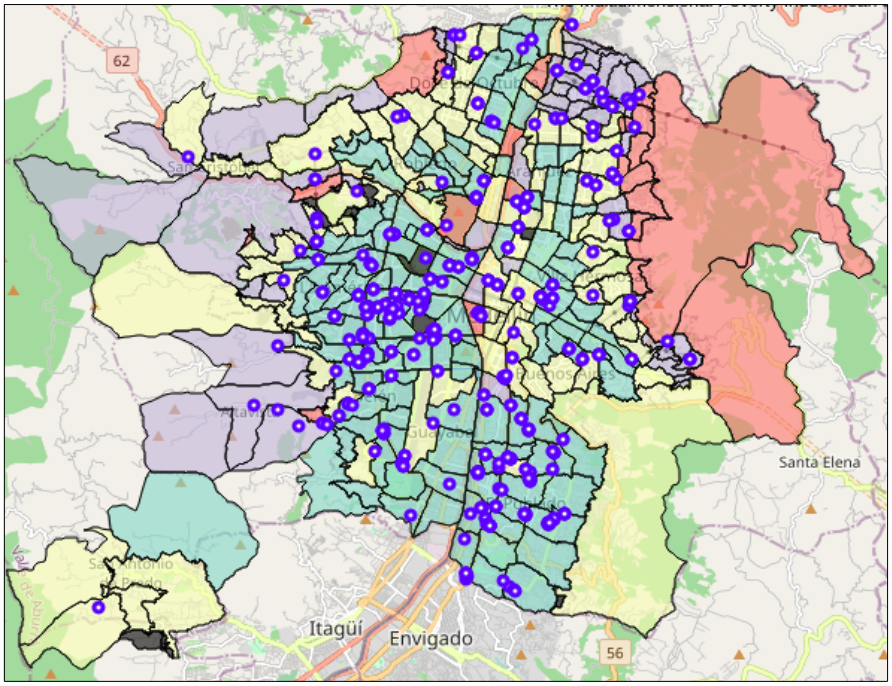

TABLE OF CONTENTS

1. INTRODUCTION

Multidimensional Poverty Index (MPI) looks beyond income to understand how people experience poverty in multiple and simultaneous ways. It identifies how people are being left behind across three key dimensions: health, education and standard of living, comprising 10 indicators. People who experience deprivation in at least one third of these weighted indicators fall into the category of multidimensional poor.

1.1 PROBLEM

The Medellin city, uses this index by every block in the urban city, according to the last census data, but it is without information about every neighborhood. In addition, it does not have data about any relationship with principal venues around every neighborhood.

Taking in count the previous information, now the major question is: could the closer venues from each neighborhood has a relationship with poverty and with this data is possible to do prediction like an approximation to Multidimensional Poverty Index?

Using Machine learning and foursquare free data, we will try to resolve this!!!

1.2 INTEREST

General population and government are interested in deep knowledge about Multidimensional Poverty Index and how new changes and effort could reduce it. Any new prediction or relationship could be a useful tool in this effort.

2. DATA

2.1 DATA SOURCE

Free data from the last census in the year 2018, Multidimensional Poverty Index (MPI) is available at Census DANE MPI like a shape file with it is index for each block in the city.(Figure 1)

Data about every neighborhood is available at Neighborhoods Medellin like a shape file.

With QGIS software (3) we change the data from shapefile to geojson to see neighborhoods and extract excel spreadsheets, to use Multidimensional Poverty Index.

Data about venues is available at [foursquare]using a free account.

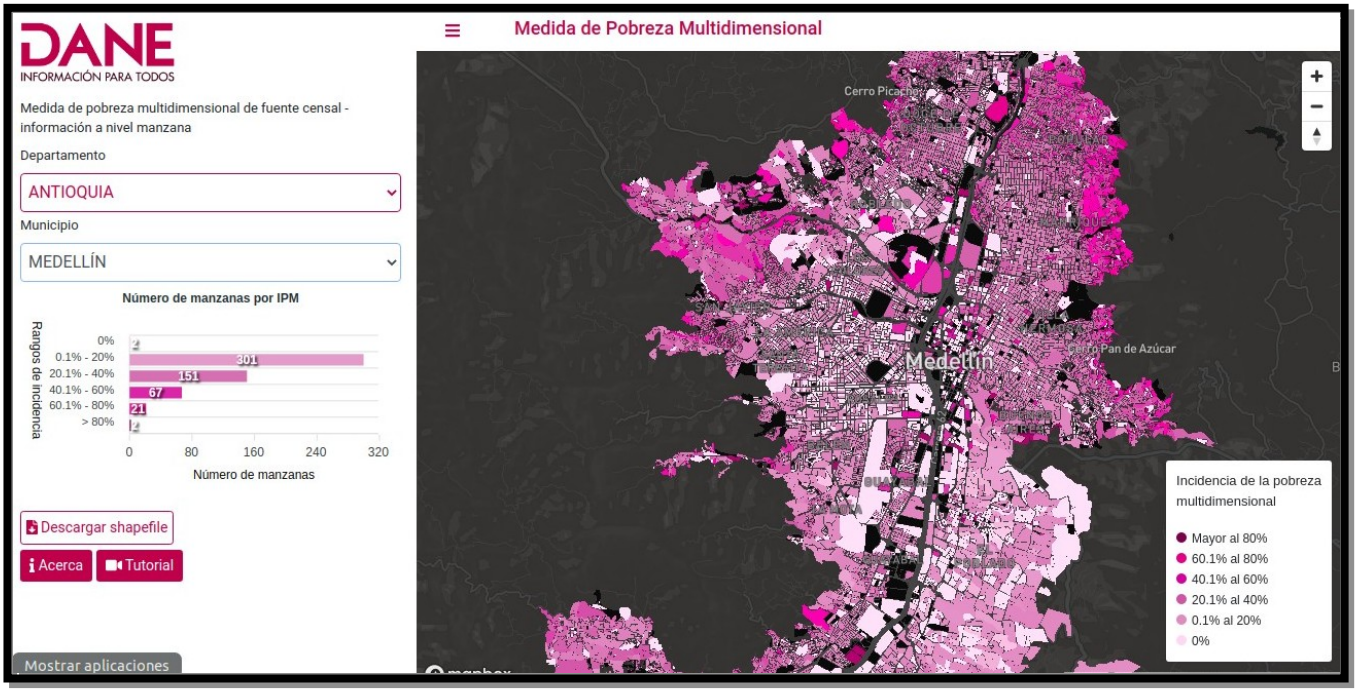

Figure 1. Multidimensional Poverty Index (MPI) by block Medellin, 2018

2.2 DATA CLEANING

Data with Neighborhood geometries was downloaded and opened with QGIS software (3) to transform data from shapefile to geojson, additionally extract coordinates inside each urban neighborhood (Table 1 and Figure 2) and excel spreadsheet with Multidimensional Poverty Index by block (Table 2)

CODE	Neighbourhoods	SUBTIPO_BA	Boroughs	SHAPEAREA	SHAPELEN	Lat	Long
1422	La Aguacatala	1	El Poblado	622090,156	3302,658	6,199	-75,577
810	El Pinal	1	Villa Hermosa	413416,805	3271,575	6,244	-75,545
719	Fuente Clara	1	Robledo	236441,173	3022,338	6,278	-75,605
102	Santo Domingo Savi	1	Popular	264750,452	2943,708	6,298	-75,539
302	Las Granjas	1	Manrique	641349,275	3964,702	6,279	-75,549

Table 1(first 5 rows from 298, Coordinates from urban Neighborhood in Medellin, Colombia)

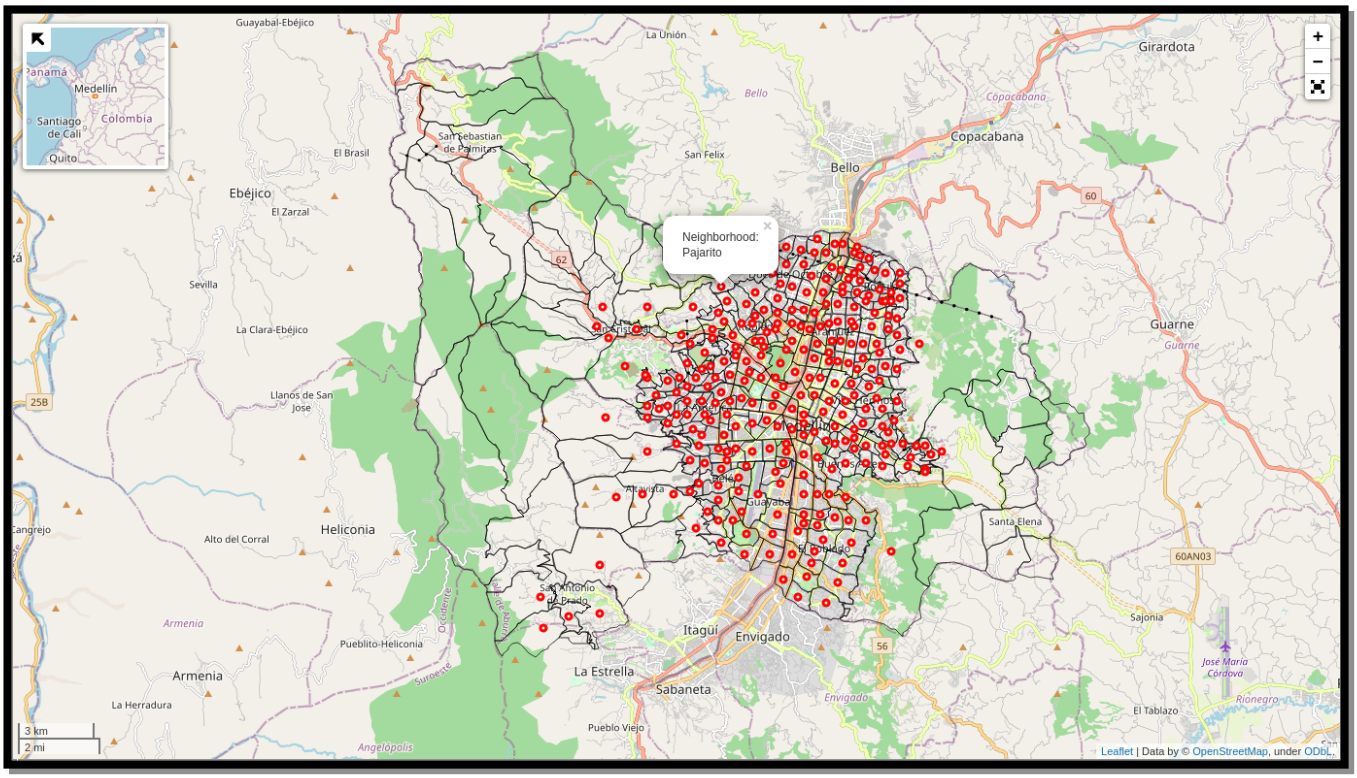

Figure 2. Urban Neighborhoods points (Red points) with coordinates in a Folium Map.

CODE	Neighbourhoods	SUBTIPO_BA	Boroughs	IPM
1509	Cristo Rey	1	Guayabal	9,6
1422	La Aguacatala	1	El Poblado	1
1414	El Castillo	1	El Poblado	0
1509	Cristo Rey	1	Guayabal	10,5
1421	Patio Bonito	1	El Poblado	0

Table 2. (First 5 rows from 13998, Multidimensional Poverty Index (IPM) by block)

With Multidimensional Poverty Index (IPM) by block, now group by average IPM for each neighborhood and reduce it from 13998 blocks to 294 neighborhoods (Table 3).

CODE	IPM	Neighbourhoods	Boroughs	Lat	Long
101	23.347692	Santo Domingo Savio No.2	Popular	6298	-75544
102	23.133333	Santo Domingo Savio No.1	Popular	6298	-75539
103	29.224786	Popular	Popular	6299	-75548
104	22.317722	Granizal	Popular	6292	-75546
105	22.534454	Moscú	Popular	6290	-75550

Table 3. (First 5 rows from 294, Multidimensional Poverty Index (IPM) average by Neighborhood)

Venues data was downloaded from foursquare using a free account and taking in count restriction about amount of data using coordinates inside each neighborhood and a radius of 600 meters. Despite this, data available for the Medellin city in this study have 4058 venues and 236 categories (Table 4).

Making dummies venues categories and ordering the venues data, now such as columns. The data available in foursquare, reduce data from 294 to 273 neighborhoods. (Table 5).

In this project, Multidimensional Poverty Index (IPM) has a new categorical level: Low (0-10), Medium (10-20), HIGH (20-30) and very high >30. (Table 6 and Figure 3)

CODE	IPM	Neighbourhoods	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
101	23,34769	Santo Domingo Savio	6,298	-75,544	La Mesa del Barri	6,294748560948	-75,5440434137111	South American Restaurant
101	23,34769	Santo Domingo Savio	6,298	-75,544	Mirador Somos T	6,294683789031	-75,5448678675302	Bar
101	23,34769	Santo Domingo Savio	6,298	-75,544	Mi Cebi & Chela	6,293481	-75,541798	Caribbean Restaurant
101	23,34769	Santo Domingo Savio	6,298	-75,544	Metrocable Linea	6,29509961273	-75,5481883799316	Cable Car
102	23,13333	Santo Domingo Savio	6,298	-75,539	Placa La Avanzad	6,294780362067	-75,5396083421013	Playground

Table 4. (First 5 rows from 4058, venues from Medellin city downloaded from foursquare)

Venue	Venue Latitude	Venue Longitude	Advertising Agency	Airport	Airport Lounge	Airport Service	American Restaurant
La Mesa del	6,294748560948	-75,5440434137111	0	0	0	0	0
Mi Cebi &	6,293481	-75,541798	0	0	0	0	0
Metrocable	6,29509961273	-75,5481883799316	0	0	0	0	0
Donde sab	6,2939332	-75,5380763	0	0	0	0	0
Mi Cebi &	6,293481	-75,541798	0	0	0	0	0

Table 5. (First 5 rows from 273, and first 5 categories as columns from 236 total categories)

IPM	POVERTY_LEVEL	LEVEL
23,347	High	3
23,133	High	3
29,224	High	3
22,317	High	3
22,534	High	3
17,978	Medium	2

Table 6. (First 5 rows from 277, categorical data from Multidimensional Poverty Index (IPM), by neighborhood)

3. METHODOLOGY

3.1 EXPLORATORY ANALYSIS

The amount of venues was about 4058 venues and 236 categories by 273 neighborhoods in Medellin city. In the Table 7, we have the top ten categories in the city. The most important or frequent venues are Cafe and food places, hotel, and shopping. However, this data was available for a free foursquare account and not with all data.

Total	Count
Café	201
Hotel	142
Burger Joint	127
Shopping Mall	127
Restaurant	121
Sandwich Place	121
Bar	118
Pizza Place	115
Italian Restaurant	112
BBQ Joint	95

Table 7. Top ten categories in Medellin city.

Multidimensional Poverty Index (IPM) has a classification from Low to very high, according to the all 273 neighborhoods, Low is the most frequent 129, Medium 80, High 46, and Very high 18. Fortunately about 76 % of neighborhood are low to medium values from urban zones. (Figure 3)

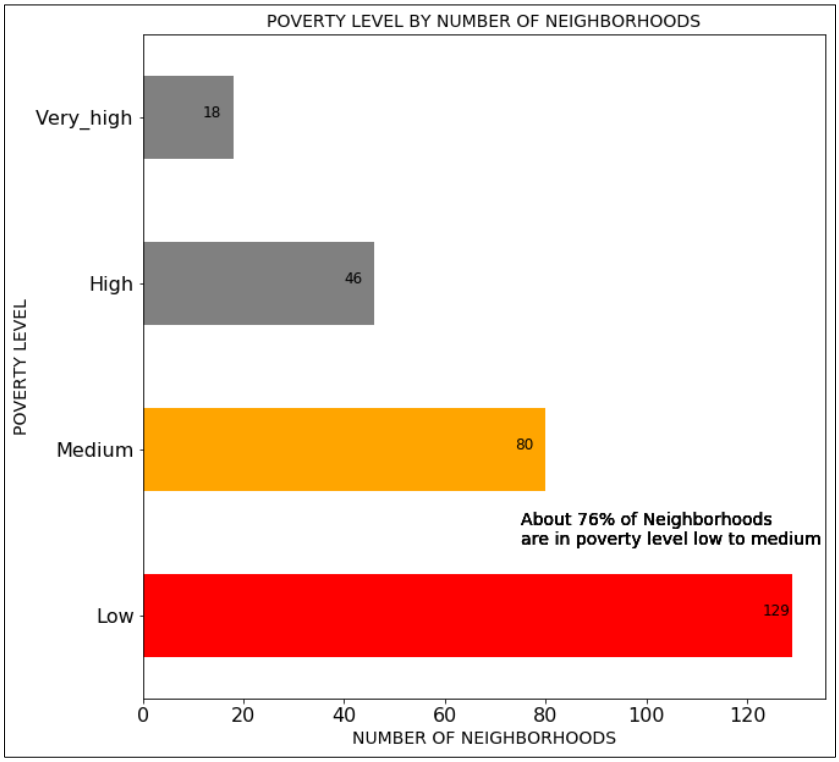

Figure 3. Poverty level by number of neighborhoods. Low and Medium have about 80% of total neighborhoods.

Making Pearson correlation between venues categories and Multidimensional Poverty Index (IPM) , with about 7 to 30% in Pearson values for positive and negatives top 15 correlations like show (Table 8). Cable Car and Pizza Place are the high values with positive and negative relationship respectively, but all 30 categories were used to seek a possible prediction.

Positive	**IPM**	**Negative**	**IPM**
Categories	**Pearson**	**Categories**	**Pearson**
Cable Car	0,325716	Pizza Place	-0,317615
Print Shop	0,242016	Supermarket	-0,306148
Rental Service	0,163193	Burger Joint	-0,303793
Construction & Landscaping	0,129225	Mexican Restaurant	-0,284867
Locksmith	0,127111	Restaurant	-0,260067
Other Great Outdoors	0,12428	Gym / Fitness Center	-0,247494
Playground	0,119828	Bar	-0,24739

Mountain	0,1109	Italian Restaurant	-0,239921
Science Museum	0,108608	Gym	-0,239022
Summer Camp	0,095169	Shopping Mall	-0,238117
Soccer Field	0,095079	Steakhouse	-0,232709
Campground	0,082331	Bakery	-0,229419
Aquarium	0,07931	Fast Food Restaurant	-0,225579
Planetarium	0,078812	Hotel	-0,215947
Mobile Phone Shop	0,074887	Seafood Restaurant	-0,213963

Table 8. Positive and negative Pearson correlation 30 categories with Multidimensional Poverty Index (IPM).

3.2 PREDICTIVE MODELING

Using first 25 from 30 categorical data get the best results

Several kind of algorithms were used in this study starting with K-Nearest Neighbors, Decision Tree, Support Vector Machine Evaluation and Logistic Regression.

K-Nearest Neighbors is an algorithm for supervised learning. Where the data is 'trained' with data points corresponding to their classification. Once a point is to be predicted, it takes into account the 'K' nearest points to it to determine it´s classification (1). With early 30 categories the best result was using the first 25 categories and the accuracy from 0,74 to K=6. (Figure 1)

Training with 80% and test with 20%, and random state = 5.

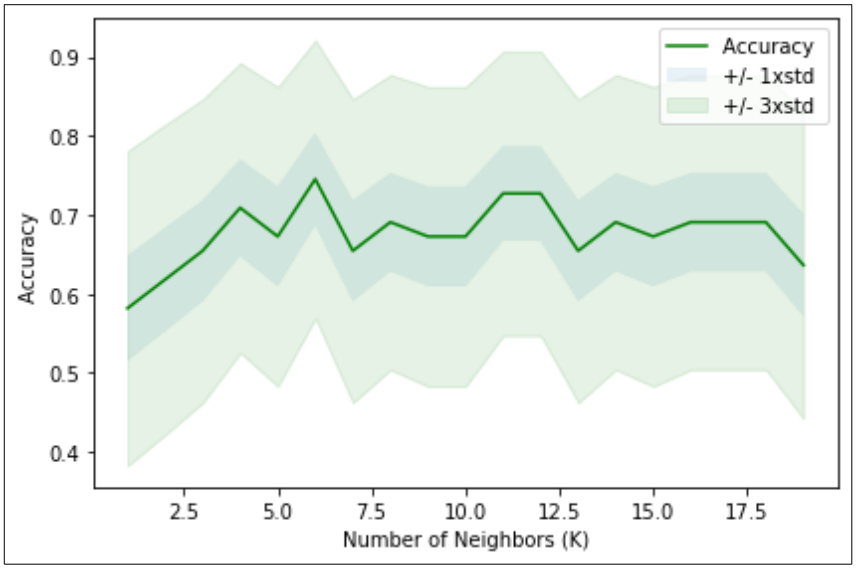

Figure 1. Number of K and the best accuracy with K=6.

Decision Tree Algorithm

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. (2)

Data= with decision tree, 80% were training data and 20 % test data, with random state = 5 and max depth = 5, and entropy criterion.

SVM (Support Vector Machines)

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data is transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.(1)

Data= Training with 80% and test with 20%, and random state = 5.

Logistic Regression

While Linear Regression is suited for estimating continuous values (e.g. estimating house price), it is not the best tool for predicting the class of an observed data point. In order to estimate the class of a data point, we need some sort of guidance on what would be the **most probable class** for that data point. For this, we use **Logistic Regression**.

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables. (1).

Data=Training with 80% and test with 20%, and random state = 5.

4. RESULTS

After use 25 from 30 categorical data with the best results and several kind of algorithms such as K-Nearest Neighbors, Decision Tree, Support Vector Machine Evaluation and Logistic Regression, we have final table with Jaccard and F1 score for each. (Table 9).

The best accuracy result was K-Nearest Neighbors with 73% F1 score using K= 6, follow by Logistic Regression with 64 %.

Algorithm	**Jaccard**	**F1_score**
KNN	0,607338	**0,736754**
Decision Tree	0,34609	0,542054
SVM	0,420294	0,502424
Logistic_Regression	0,495667	**0,64768**

Table 9. Accuracy table to several kind of algorithms.

Taking in count 25 categories (Table 10) and 1073 venues the map show these venues inside Medellin city and it is Multidimensional Poverty Index like categorical color. (Figure 2)

25 venue categories
Cable Car
Print Shop
Construction & Landscaping
Caribbean Restaurant
Other Great Outdoors
Business Service
Betting Shop
Mountain
Home Service
Playground
Science Museum
Financial or Legal Service
Jewelry Store

Fabric Shop
Farm
Supermarket
Pizza Place
Burger Joint
Gym
Mexican Restaurant
Shopping Mall
Gym / Fitness Center
Bar
Bakery
Italian Restaurant

Table 10. 25 venue categories using to predict with 73 % of accuracy categorical poverty level in Medellin by Neighborhood.

4.1 Final map

Geojson data has a column called 'CODIGO' with one zero before numbers when lenght =3, for example our column called ‘CODE’ for one Neighborhood is 101, but in Geojson is number is 0101. Then is necessary a new column with codes with equal codes in Geojson and draw IPM value for each neighborhood.

CODE	IPM	Neighborhoods	Boroughs	Lat	Long	POVERTY_LEVEL	LEVEL	CODIGO
101	23,347692	Santo Domingo Savio No.1	Popular	6,298	-75,544	High	3	0101
102	23,133333	Santo Domingo Savio No.2	Popular	6,298	-75,539	High	3	0102
103	29,224786	Popular	Popular	6,299	-75,548	High	3	0103
104	22,317722	Granizal	Popular	6,292	-75,546	High	3	0104
105	22,534454	Moscú No.2	Popular	6,29	-75,55	High	3	0105
106	17,978	Villa	Popular	6,288	-75,551	Medium	2	0106

		Guadalupe						

Table 11. New column with the same name ‘CODIGO’ in geojson.

Now, 55% of venues could be showing over IPM map.

We can see that the very high values are in North and Northeast (Red colors) and low values from South to North, following almost a line (Green color), medium and high values around it with higher values to the west (Purple color. Center to South east corresponds to the best zone according to more values low and medium.

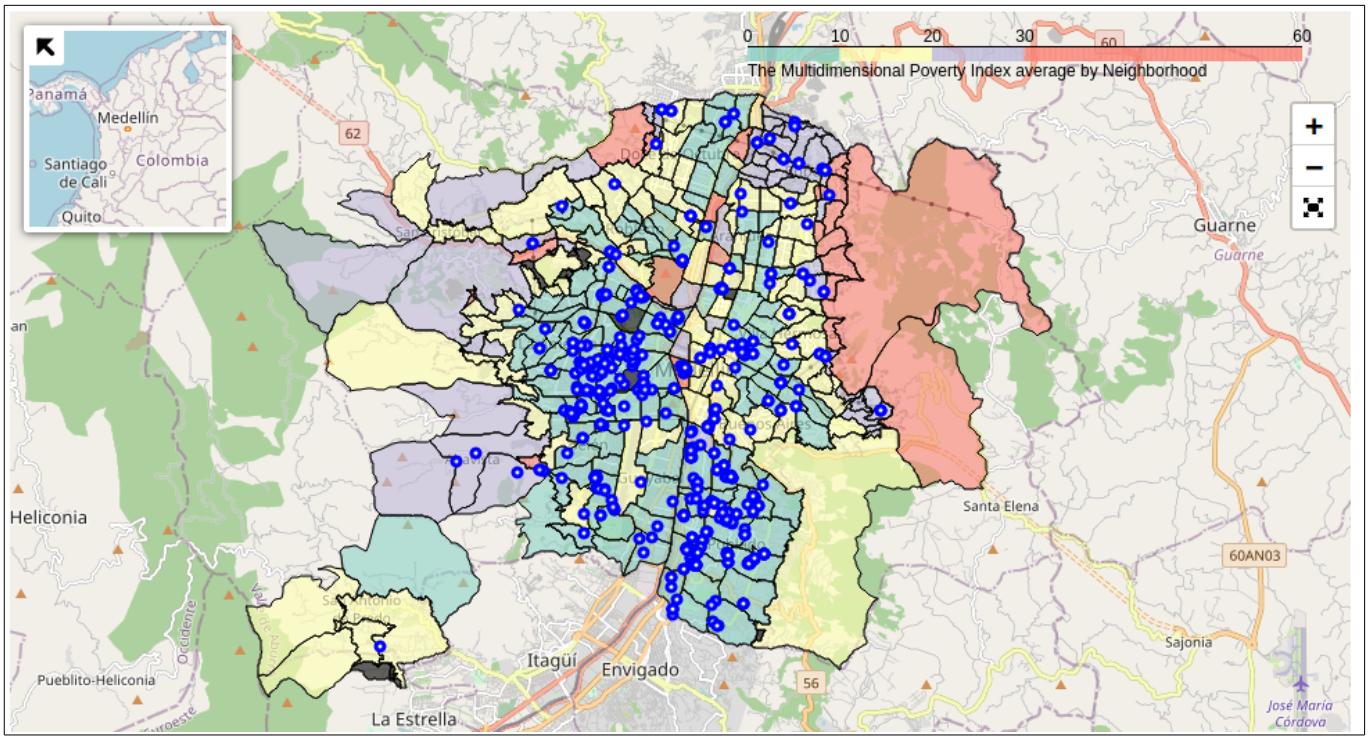

Figure 2. Map with 55% venues using to training and test model (blue) and categorical Multidimensional Poverty Index: Low (0-10) (Green), Medium (10-20) (Yellow), High (20-30) (Purple) and Very high >30(Red) in Medellin city, Colombia.

4. DISCUSSIONS

Clearly the worst zones correspond to mountain zones and boroughs around the urban city. These neighborhoods need especial attention. Prediction only with venues and categories is a good way to see the categorical poverty and local situation as an approximation, but not like a unique reference, because the lack of information, decreases better results and accuracy.

However, K-Nearest Neighbors model is a useful tool and could be the first step to have an idea about Multidimensional Poverty Index and using this information to possible new neighborhood constructions, marketing, territorial ordainment and set new strategies to reduce poverty to lowest values.

According to previous information recommendations are:

- Extract all possible data from venues and improve model and accuracy to have better results.
- Avoid lack of information like in the south west from the map with only one venue in five neighborhoods.
- Many places are not clearly identified they are the black zones from the map, like parks and green zones and other places.
- Places without houses or buildings does not have Multidimensional Poverty Index, and could create biases in a few neighborhoods
- Mountain zones and big neighborhoods have a problem with coordinates, because we only use one point inside the neighborhood and with 600 meters of radio is not the best way, maybe those big neighborhoods needs two or more points
- Choose categories is not an easy task, here is important evaluate categories and which ones are the better with entire data, according to correlation here was used 25 categories but with full data this could be changed.

5. CONCLUSIONS

Using a free foursquare account and coordinates from each neighborhood with radius of 600 meters, it could extract 4058 venues and 236 unique categories from Medellin city, Colombia. With a final data from 25 categories and 1073 venues use to training and test to 273 urban neighborhoods.

Several kind of algorithms were used in this study starting with K-Nearest Neighbors, Decision Tree, Support Vector Machine Evaluation and Logistic Regression. The best result was K-Nearest Neighbors with 73% of F1 score accuracy, using K= 6, follow by Logistic Regression with 64%.

Multidimensional Poverty Index (IPM) has a classification from Low to very high, according to the all 277 neighborhoods, Low is the most frequent 129, Medium 80, High 46, and Very high 18.

Very high values are in the North and Northeast and low values from South to North, medium and high values around it, with higher values to the west. Center to South east zone corresponds to the best zone according to more values low and medium in Medellin city.

Prediction to Multidimensional Poverty Index (IPM) only with venues and categories is a good way to see the categorical poverty and local situation as an approximation, but not like a unique reference, because the lack of information, decreases better results and accuracy. However, K-Nearest Neighbors model is a useful tool and could be the first step to have an idea about Multidimensional Poverty Index and use this information to possible new neighborhood constructions, marketing, territorial ordainment and set new strategies to reduce poverty to lowest values.

To next researches and improve this study is necessary to extract all possible data from venues in the entire city with a premium account at foursquare and building new models having accuracy above 73% and taking in count categories with higher weight in the full data.

5. REFERENCES

1-IBM Developer skill Network. Data Science labs.

2-https://www.kdnuggets.com

Free software:

3-QGIS: https://www.qgis.org/es/site/

Free data:

- Multidimensional Poverty Index Medellin, Colombia: https://geoportal.dane.gov.co/visipm/
- Shapes Medellin Neighborhoods: https://geomedellin-m-medellin.opendata.arcgis.com/datasets/barrio-vereda
- Venues data download site: www.foursquare.com