Covid-19 evolution in Santander, Colombia

Coursera Captstone project for IBM data science professional certification

By Jhon Padilla

1. Introduction

Business problem

Santander is a region of Colombia. In Colombia, covid-19 arrived in February 2020. The government of Colombia decided to decree isolation measures in March 2020. One of the regions or departments of Colombia is Santander, which is one of the most important regions in the Northeast of the country. The capital city of Santander is Bucaramanga, which is part of a metropolitan area named Metropolitan Area of Bucaramanga. This area has four cities: Bucaramanga, Girón, Floridablanca and Piedecuesta.

Now, the number of cases in Santander has increased fastly and it is important to do an analysis of how the pandemic behavior is over the people (by age, by city, by gender, etc.). A map with the cities and the number of Covid-19 cases is given to help travelers make decisions. Another aspect that should be improved is that Doctors now have to make decisions about what people should be attended at ICU (Intensive Care Units) or not; this situation is presented due to the limited number of ICU beds. Also, with this project, an ML algorithm for death probability calculation is developed based on parameters given in the data set. Such probability could be used by Doctors make decisions taking into account this parameter and other parameters on the patient's health.

Finally, as the main city of Santander, Bucaramanga has the main healthcare centers and therer are people who have relatives sick from covid, they need to know what venues are near clinics and hospitals. Then, a map with venues near to main to these healthcare centers is provided for such situations.

Target audience

Results of the project are useful for persons who need to travel to some city and need to know the number of Covid-19 cases in that place. Another audience is people that needs to know what venues are near to healthcare centers in Bucaramanga (the main city of Santander). Also, the statistical report it is important for authorities who need to make decisions over possible social, economic and health strategies. Finally, the report about the death probability is useful in medical decisions when the number of ICU beds is scarce.

2. Data background

In Colombia, health ministry has a web page with open data at www.data.gov.co. In that page there are data from covid-19 pandemic in all regions and cities within Colombia. In this project, covid-19 data published in this page was taken and studied. The number of records in such data base is bigger because it has all cases in Colombia from february of past year, then only Santander region data was taken for this study. Such data were saved on a CSV file and it is loaded and studied in the next steps. This data set have parameters such as: case ID (unique identifier given in order of appearance), diagnosis date, Department name, City name, age, gender, state (Recovery, active, dead), ethnicity, and others. The data set has 174145 rows and 23 columns.

Here are the column names (in spanish) and their types and english translation:

Spanish name	**Type**	**English translation**
fecha reporte web	object	Web report date
ID de caso	int64	Case ID number
Fecha de notificación	object	Notification date
Código DIVIPOLA departamento	int64	Department Code
Nombre departamento	object	Department Name
Código DIVIPOLA municipio	int64	City code
Nombre municipio	object	City name
Edad	int64	Age
Unidad de medida de edad	int64	Age Cuantity
Sexo	object	Gender
Tipo de contagio	object	Type of contagion
Ubicación del caso	object	Case location
Estado	object	State
Código ISO del país	float64	Country ISO code
Nombre del país	object	Country name
Recuperado	object	Recovery
Fecha de inicio de síntomas	object	Symptom onset date
Fecha de muerte	object	Death date
Fecha de diagnóstico	object	Diagnosis date
Fecha de recuperación	object	Recovery date
Tipo de recuperación	object	Recovery type
Pertenencia étnica	float64	Ethnicity
Nombre del grupo étnico	object	Ethnic group

Finally, as people who have relatives sick from covid need to know what venues are near clinics and hospitals, then, a map with venues near to these healthcare centers is provided for such situations. For this map it is necessary to obtain data from Foursquare with venues near to healthcare centers area.

3. Methodology

3.1. Exploratory Data Analysis

For this project, first, an exploratory data analysis was performed to obtain conclusions about the data behaviour respect some important variables. Thus, several graphics were obtained such as:

-Relationship between the number of Covid-19 cases and the age.
-Relatinship between the number of cases and gender
-Relationship between the number of cases and the city
-Evolution of the number of cases over the time

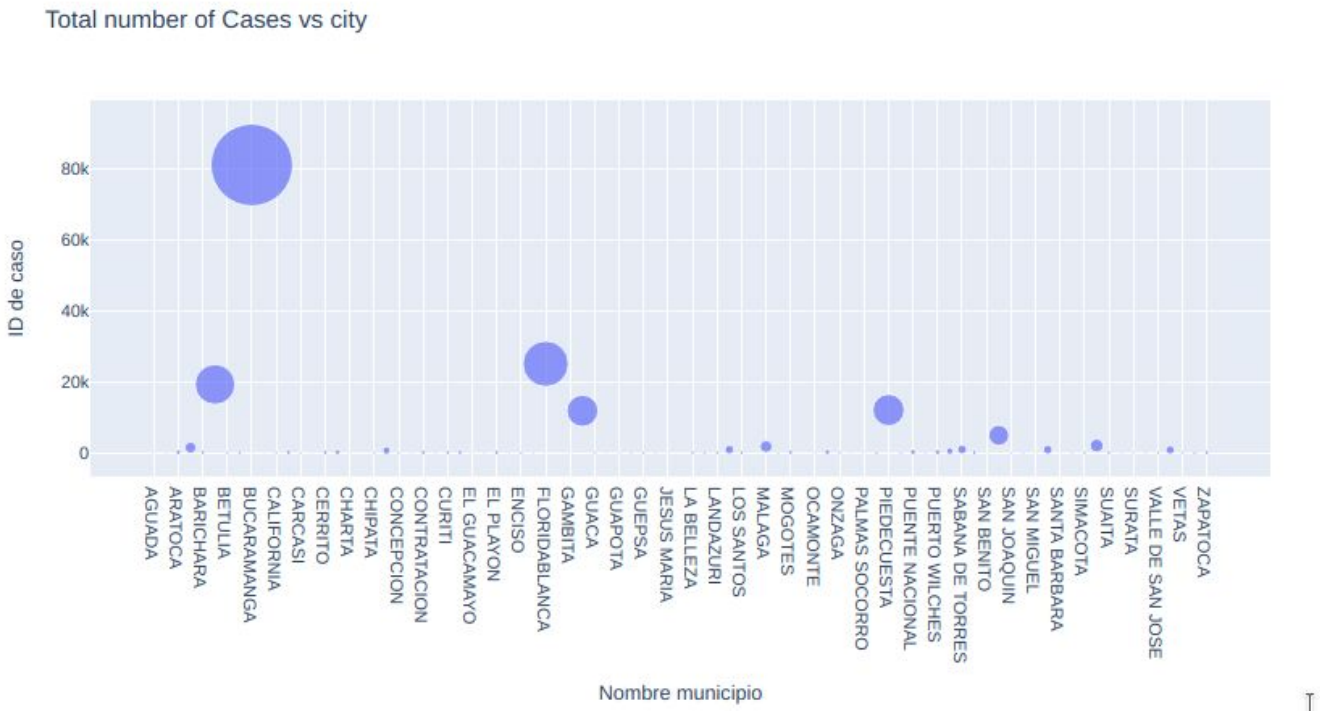
Data analysis by city

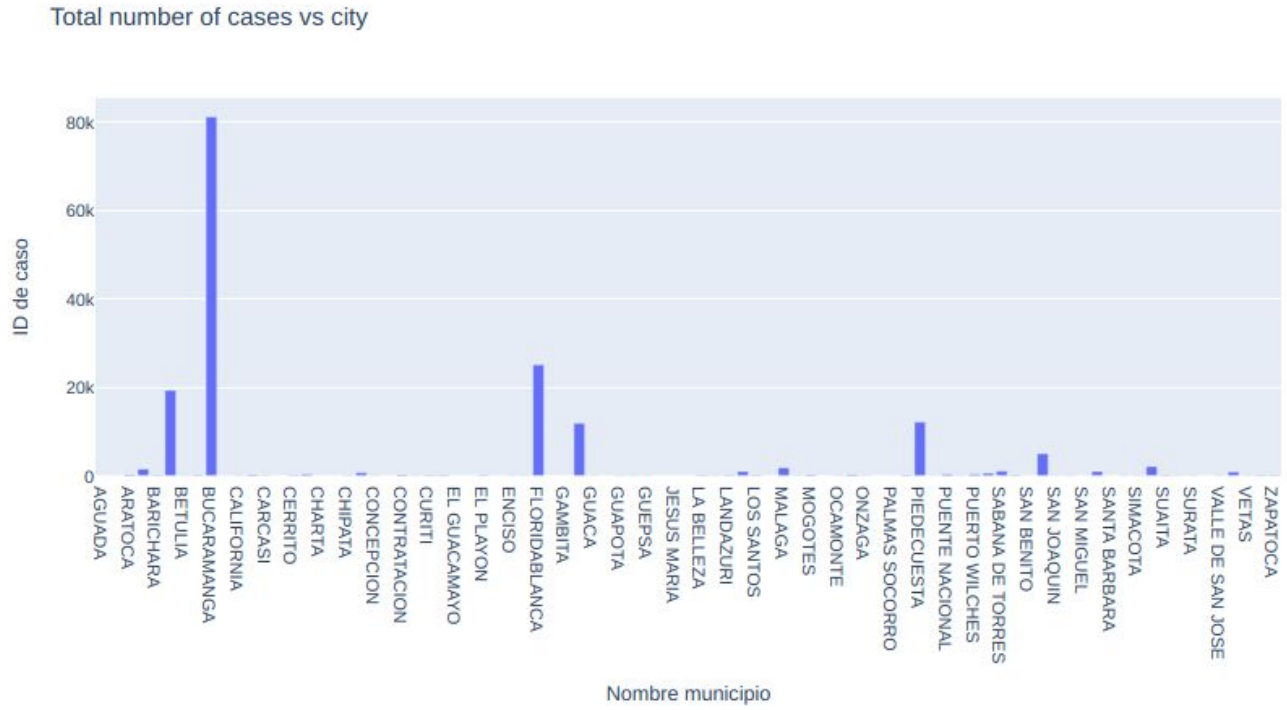
To perform this analysis, it was necessary to group the rows by cities. Results are showed in the next table:

Nombre municipio	fecha reporte web	ID de caso	Fecha de notificación	Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Edad	Unidad de medida de edad	Sexo	Tipo de contagio	...	Código ISO del país	Nombre del país	Recuperado	Fec ini sínt
AGUADA	37	37	37	37	37	37	37	37	37	37	...	0	0	37	
ALBANIA	50	50	50	50	50	50	50	50	50	50	...	0	0	50	
ARATOCA	294	294	294	294	294	294	294	294	294	294	...	0	0	290	
BARBOSA	1593	1593	1593	1593	1593	1593	1593	1593	1593	1593	...	1	1	1591	
BARICHARA	203	203	203	203	203	203	203	203	203	203	...	0	0	202	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
VELEZ	935	935	935	935	935	935	935	935	935	935	...	0	0	933	
VETAS	94	94	94	94	94	94	94	94	94	94	...	0	0	94	
VILLANUEVA	153	153	153	153	153	153	153	153	153	153	...	0	0	152	
ZAPATOCA	208	208	208	208	208	208	208	208	208	208	...	0	0	206	
barrancabermeja	1	1	1	1	1	1	1	1	1	1	...	0	0	1	

88 rows × 22 columns

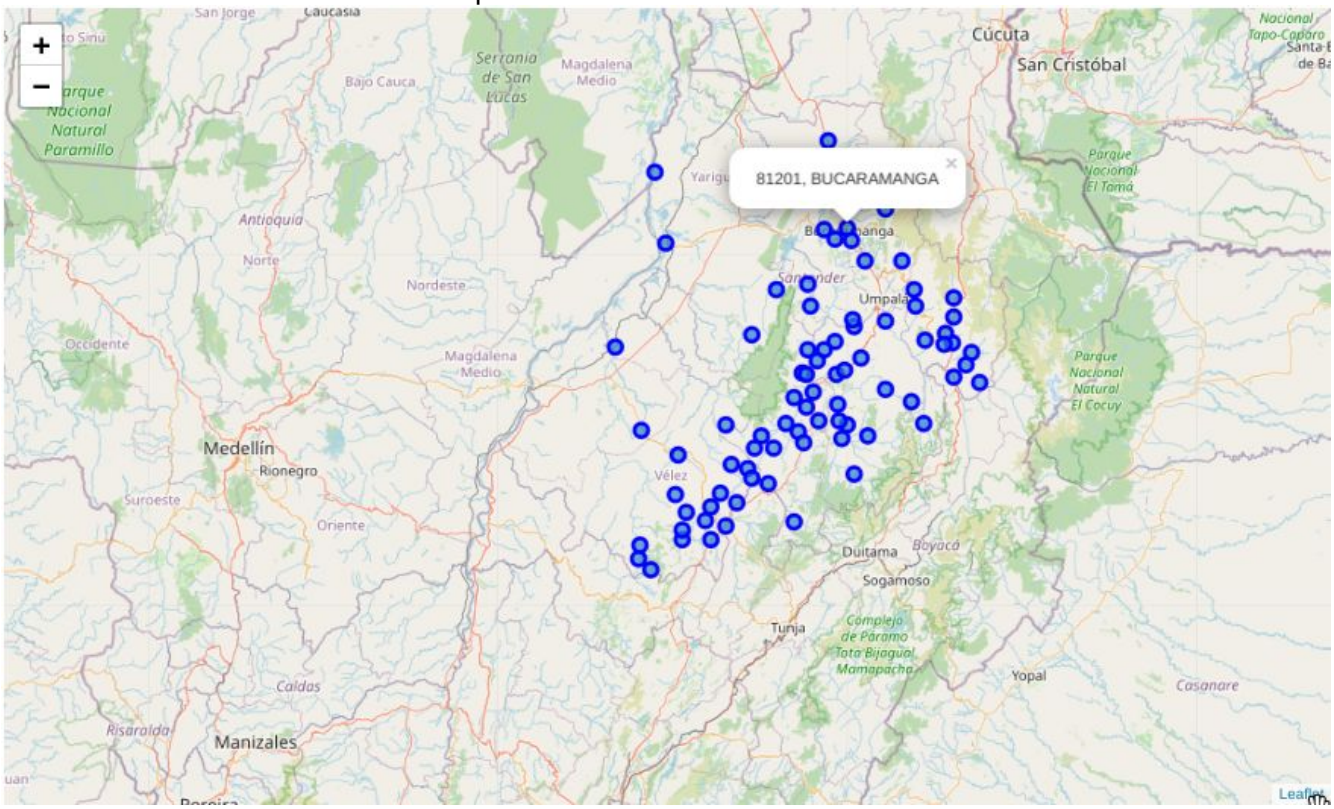
Taking into account such table, it was possible to build two graphics that show the number of cases by city, first graph is a bubble chart and second graph is a bar chart. These figures are:





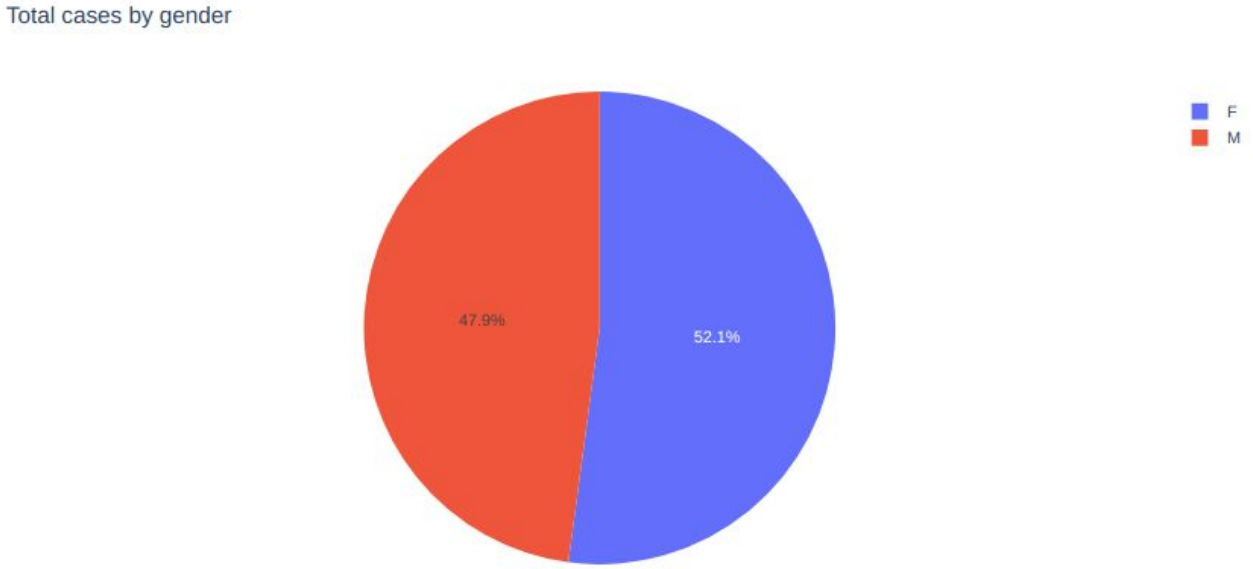
In the two figures above, we can see that high number of cases are located at main cities with a great number of people: Bucaramanga (81201 cases), Floridablanca (25193 cases), Barrancabermeja (19395 cases), Piedecuesta (12169 cases), Girón (11951 cases). All cities, except Barrancabermeja, are part of the Metropolitan Area of Bucaramanga.

Now, for travelers that needs to go to any city of Santander, a map with the cities location and the number of covid cases is printed.



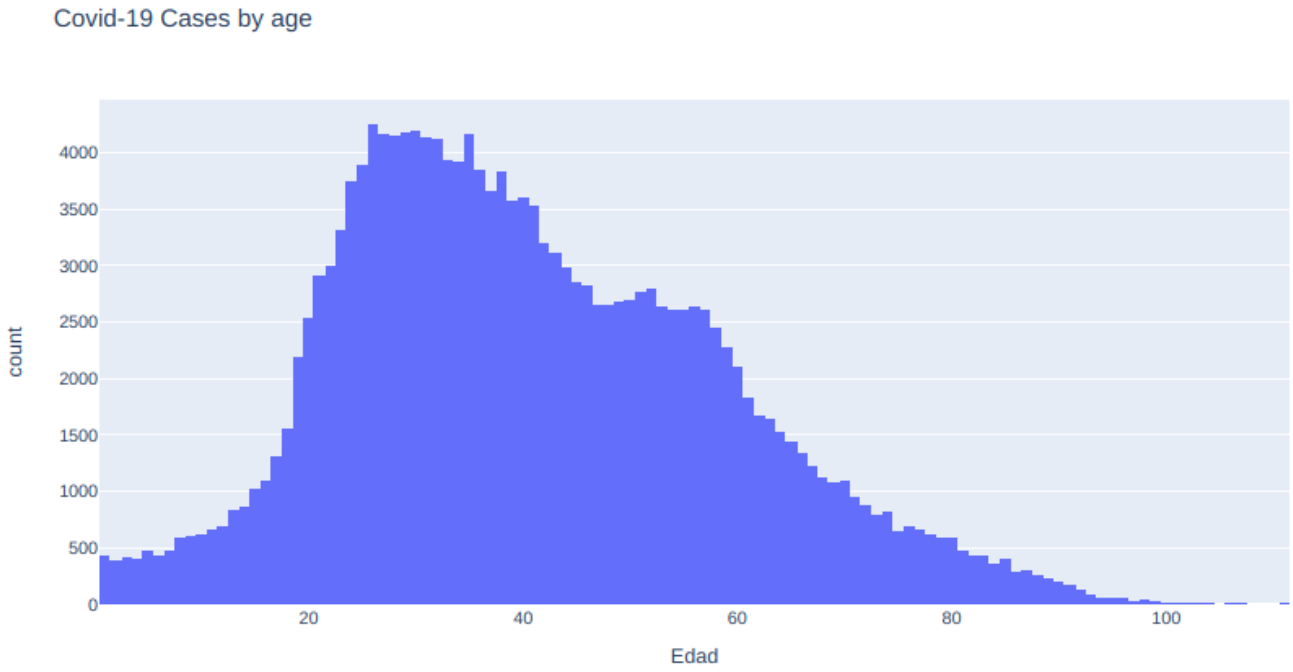
Cases by gender

Now, we will obtain the distribution of cases by gender (Female, Male). We can see that the numbers are similar, and in the pie chart below, we can see that in percentage, there is a little difference and the majority of cases are for Females with a 52.1%.



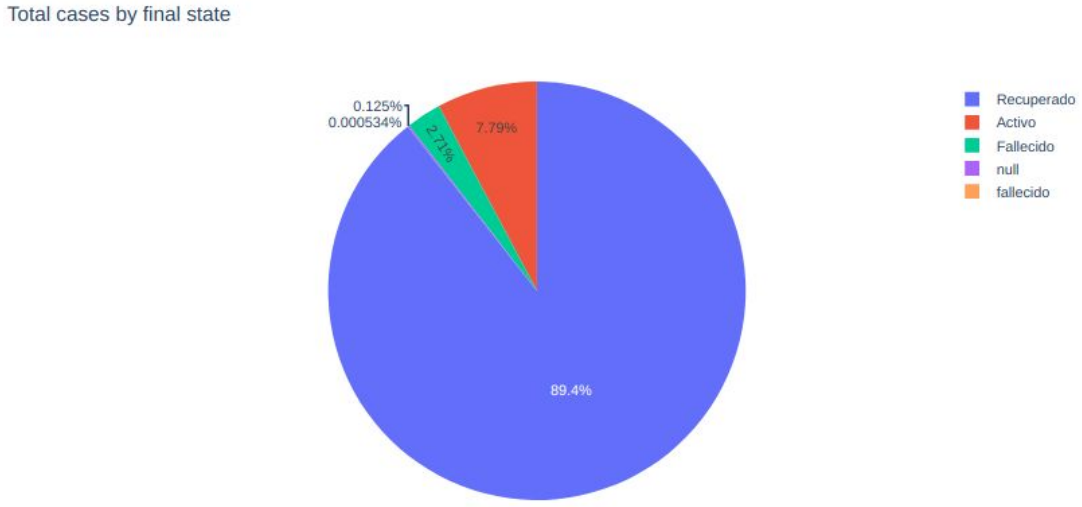
Cases by age

In figure below, we can see the number of Covid-19 cases per age, and we can see that the majority of infected people are young in ages between 20 and 40 years, which could be explained because they are people of working age. Also, these persons study in college and universities.

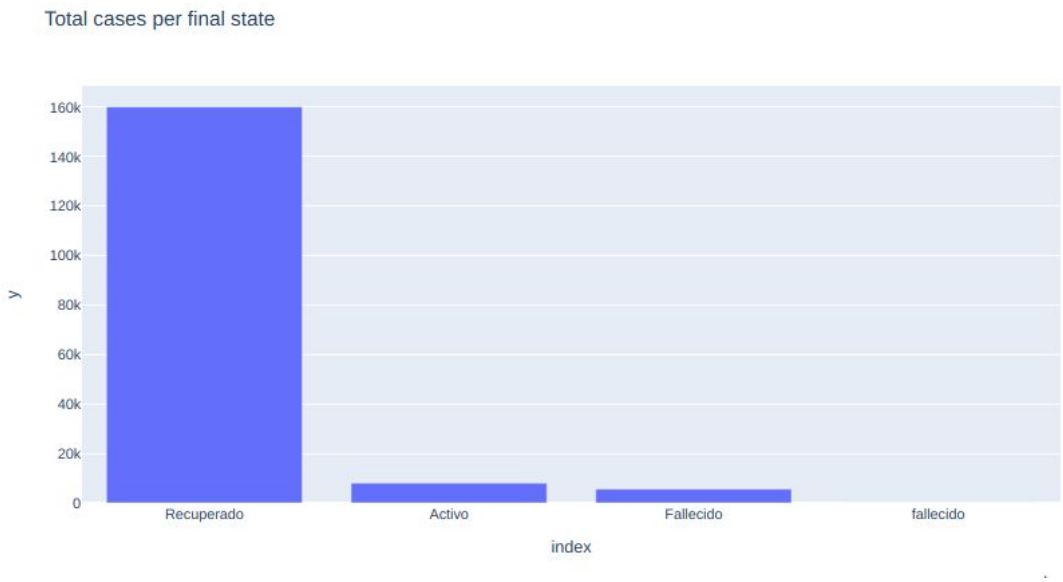


Analysis of deaths

In the two figures below, we can see an histogram and a pie chart that shows the number and the percentage of persons at the final state of the Covid-19 disease: Recovered (Recuperado), Active (Activo), Dead (Fallecido). We can see that the percentage of deaths is 2.71%, meanwhile recovered persons are 89.4% and active are 7.79%. Thus, the number of deaths

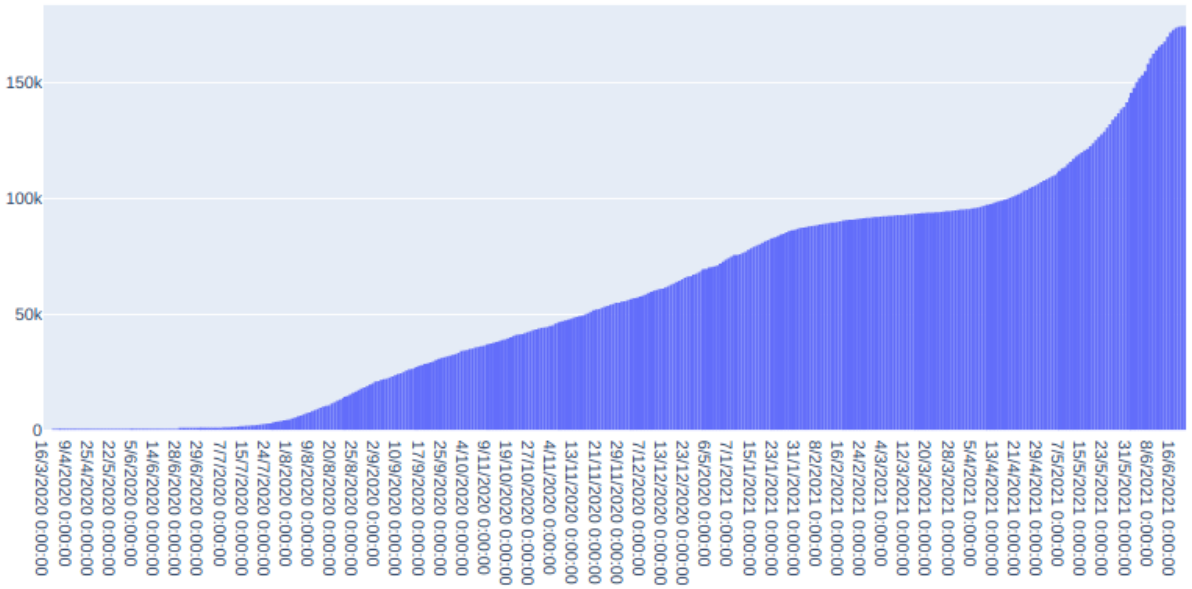


are 103849 in Colombia and 4719 in Santander, which is an important number compared with other countries in the world.



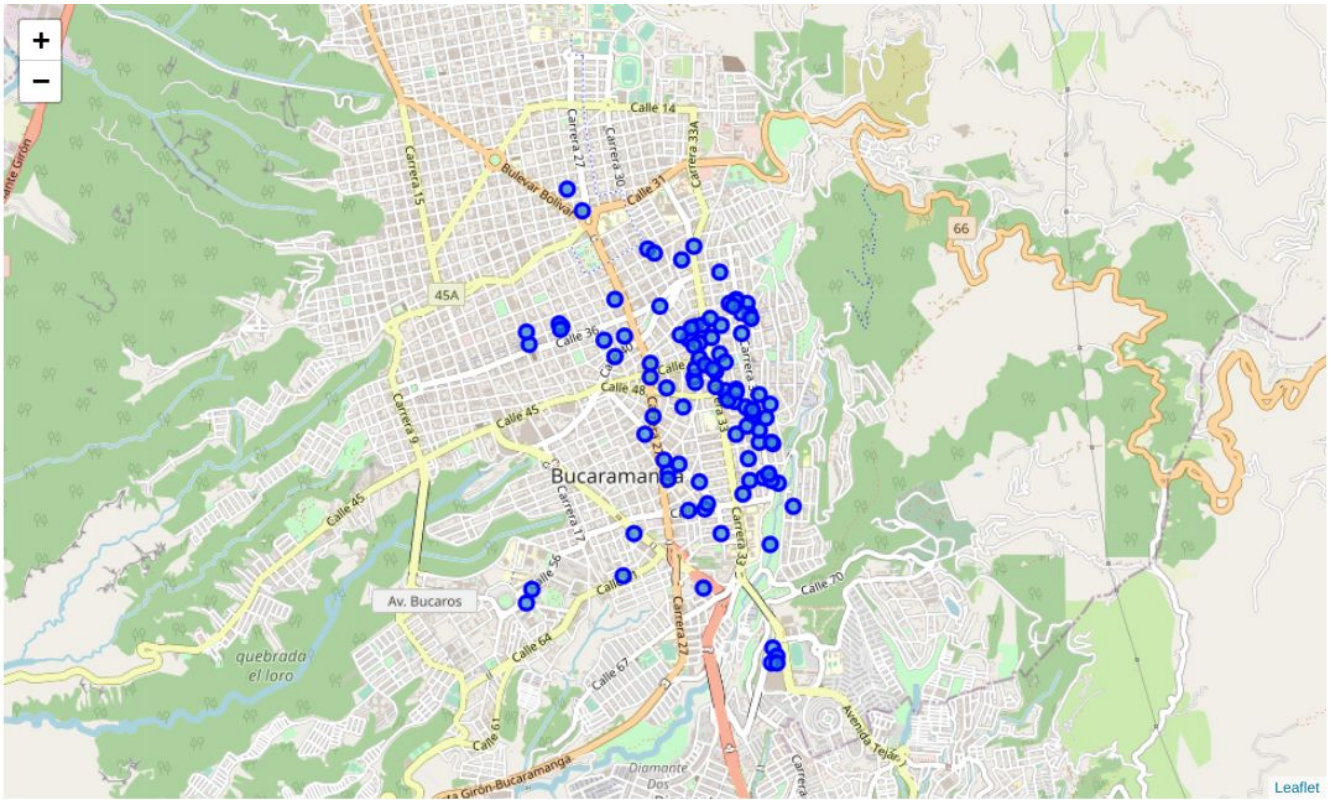
Number of cases in time

In figure below, we can see that an important increasing in number of cases initiated from second week of april and then, the slope increase in may an june. This is due to two reasons mainly, first, because of the holly week vacations, and the second reason is because of protests ocurred due to economic recession caused by social isolation and the closure of businesses for long periods of time decreed by the government.



Venues near to healthcare centers

Finally, people who have relatives sick from covid need to know what venues are near clinics and hospitals. Then, a map from Bucaramanga with venues near to main to these healthcare centers is provided for such situations. Data for this map were obtained from Foursquare. Bucaramanga was selected because is the main city and it has the main health centers for this Region. The map is showed as follows:



3.2. ML Model Development

In these days, the death rate is higher and many people are dying due to Covid-19. Due to the shortage of ICU beds, doctors are having to make ethical decisions about who lives and who dies. Then, It would be important to have an algorithm that tells the doctor how likely a covid-19 patient is to die based on certain parameters given in the data table that we have. To obtain such information, it was decided to use a Logistic Regression model because it can calculate the probability of occurrence of a given event. In our case, the event is the death of a patient.

Data Cleaning

For purposes of Machine Learning modeling, firsts, a subset Data set columns was taken. Thus, a table with city code, age, gender, ethnicity and Final State of the pacient, was obtained.

	Código DIVIPOLA municipio	Edad	Sexo	Pertenencia étnica	Recuperado
0	68276	24	M	6.0	Recuperado
1	68001	39	F	6.0	Recuperado
2	68001	49	F	6.0	Recuperado
3	68001	33	F	6.0	Recuperado
4	68001	80	F	6.0	Recuperado

Second, we replace the gender values by M=0 and F=1:

	Código DIVIPOLA municipio	Edad	Sexo	Pertenencia étnica	Recuperado
0	68276	24	0	6.0	Recuperado
1	68001	39	1	6.0	Recuperado
2	68001	49	1	6.0	Recuperado
3	68001	33	1	6.0	Recuperado
4	68001	80	1	6.0	Recuperado

Now, the column "Recuperado" is changed to numeric categories as: 'Recuperado'=1 (survivors), 'Activo'=1 (active with covid-19) and 'Fallecido'=0 (dead)and 'fallecido'=0 (also is dead). Then, if patient is alive, the label category is 1 and if patient is dead, label category is 0.

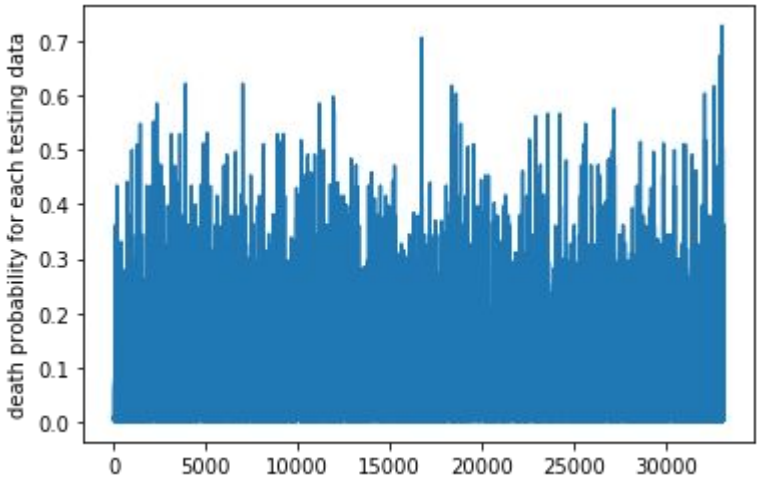
	Código DIVIPOLA municipio	Edad	Sexo	Pertenencia étnica	Recuperado
0	68276	24	0	6.0	1.0
1	68001	39	1	6.0	1.0
2	68001	49	1	6.0	1.0
3	68001	33	1	6.0	1.0
4	68001	80	1	6.0	1.0

Next, a check for NaN values in dataframe was performed. Then, rows with NaN values were dropped.

ML model development

As were described above, the Machine Learning model that was selected is Logistic Regression. Such algorithm can calculate the probability of occurrence of a given event based on several parameters given as input. In our case, the event is the death of a patient. Input parameters to Logistic Regression model are: city code, age, gender and ethnicity, meanwhile the output is the “Recuperado” column (Recovery or final state in this situation).

Next, input parameters were normalized and then, data were divided into training set and test set. Then, Logistic Regression algorithm was fitted and, as the next step, predict_proba algorithm was executed to return estimates for all classes, ordered by the label of classes. So, the first column is the probability of class 0, P(Y=0|X), and second column is probability of class 1, P(Y=1|X). Finally, the probability of death was graphed for all cases.



The model for Regression was not executed for classification because the data set does not have important health parameters. Then, this model is only a tool (based on parameters like age, gender, ethnicity and location) to help to doctors, but they should take into account another relevant aspects as health parameters to make decisions.

4. Results and Discussion

On this project, a data set from www.data.gov.co about covid-19 cases in Colombia was taken as source to perform an analysis of Covid-19 pandemic in Santander, a region of Colombia. Also, data from Foursquare was taken to display a map with venues around the Bucaramanga health centers area. Besides, a file with geographical coordinates for all cities in Santander was loaded to make another map that shows markers with the number of cases at each city of the Santander Region.

In the data analysis performed in this project, we first obtain the relationship between the cities and the number of cases and we observe that the main cities, which have higher population, have the majority of cases. These cities are: Bucaramanga (81201 cases), Floridablanca (25193 cases), Barrancabermeja (19395 cases), Piedecuesta (12169 cases), Girón (11951 cases). All cities, except Barrancabermeja, are part of the Metropolitan Area of Bucaramanga.

Also, about the relationship between the number of cases and the age, it is clear that most cases are located between 20 and 40 years old, this may be because these people are of working age and also are students and young people with a hectic social life.

If we see the comparison by gender, the number of cases for both, male and female, are similar, with 52,1% for females and 47,9% for males.

Another aspect evaluated was the number of cases that led to death. The percentage of deaths is 2,71%. Thus, the number of deaths are 103849 in Colombia and 4719 in Santander, which is an important number compared with other countries in the world.

About the number of cases over the time, we can see that in the last two months (May and June), the number of cases was increased fastly. This is consequence of two main aspects: the holly week vacations in April, and the protests ocurred due to economic recession caused by social isolation and the closure of businesses for long periods of time decreed by the government.

Finally, in this project, a Machine Learning model was developed to make predictions about the death probability based on parameters such as: age, city, ethnicity and gender. This model was developed with a Logistic Regression algorithm. The output of the algorithm is the probability of death. This results should be taken as another tool but not the final answer to this question because health parameters that doctors knows about patient should be the main reasons. However, this probability could be useful when the doctors should make ethical decisions when ICU beds are scarce.

Conclusion

In this project,an important situation that is happening in Colombia with the increasing of covid-19 cases, was studied with base in the data set that is feeded by Colombia government and is located at www.data.gov.co web page. First, an exploratory data analysis was developed, which showed important results about the distribution by cities, the cases distributed by age and gender, and the number of cases that finish in death. Besides, the number of cases over the time. Also, another two tools were developed: a map with the number of cases per city, which could be useful for travelers, and another map with he main venues around the health centers area at Bucaramanga, the capital city of Santander, which could be useful for people with relatives hosted in Clinics and hospitals in that city. Finally, in this project, a Machine Learning model was developed to make predictions about the death probability based on parameters such as: age, city, ethnicity and gender.