Predicting Car Accidents in Buenos Aires City, Argentina

Ignacio José Stivala

September 03, 2020

1. Introduction

Argentina is a country located in South America. It has an area of 2,780,400 km^2 and a population of approx. 45,000,000 inhabitants. In comparison with other countries (especially European) it has a high relation of area vs inhabitants, this means, moving inside this country takes to do many kilometers. In addition, the public transport is much worse than European systems. Also, every day there is a large quantity of people moving to and out of Buenos Aires City (CABA, Ciudad Autónoma de Buenos Aires), because of the necessity to work.

This project aims to help Argentine car drivers who usually travel to or out of CABA, preventing them to have an accident. The plan is to analyze data sets of car accidents, and search for patterns using different machine learning algorithms, to finally predict the severity of an accident that could happen.

2. Data

Data of this project was taken from an official Argentine site, that storages data sets of different topics [1]. Inside this web, there is a data set called "Intervenciones de Seguridad Vial.csv" [2] (Road Safety Interventions) recorded by "Autopistas Urbanas S.A. (AUSA)". This is the private company owner of Buenos Aires Capital City (CABA) highways. In particular, whenever an accident requires attention from Road Safety, it is recorded in the data set. It has records from January 2014 and its updated on a monthly base, but this project used data until August 2020.

As overall information, the data set has the followings columns (one row per accident): year, month, day, hour, highway name, direction circulation in the highway, kilometer number, weather, road status, quantity of people injured, quantity of people died, type of incident, quantity of motorcycles, cars, buses and trucks involved.

[1] https://datos.gob.ar/dataset

[2] https://data.buenosaires.gob.ar/dataset/seguridad-vial-autopistas-ausa

3. Methodology

3.1. Data Cleaning

At first data set was analyzed: 6,663 rows x 15 columns. There were no missing values in any row, so was not necessary to drop or replace a value. Columns that are not interested to predict were deleted: highway name, direction circulation in the highway, kilometer number, type of incident, quantity of motorcycles, cars, buses and trucks involved.

Regarding weather and road status, outliers' rows were dropped.

About quantity of people of injured, values go from 0 to 16, only rows with values from 0 to 6 were kept. And finally, about quantity of people died, rows with value of two were dropped (only kept 0 and 1).

A new column was created in base of date: day of week.

Next, there are graphics showing distribution between variables:

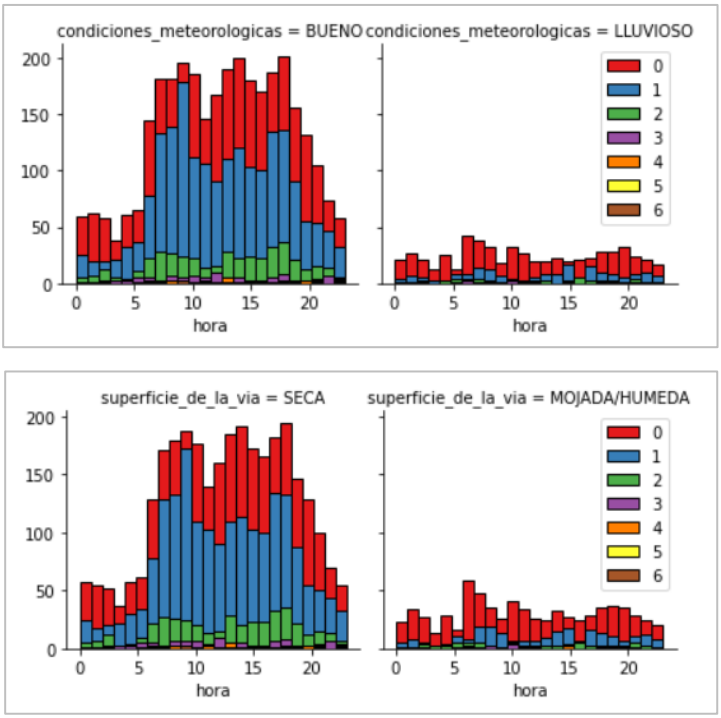

Figure 1. Relation between quantity of people injured, hour and weather vs road condition.

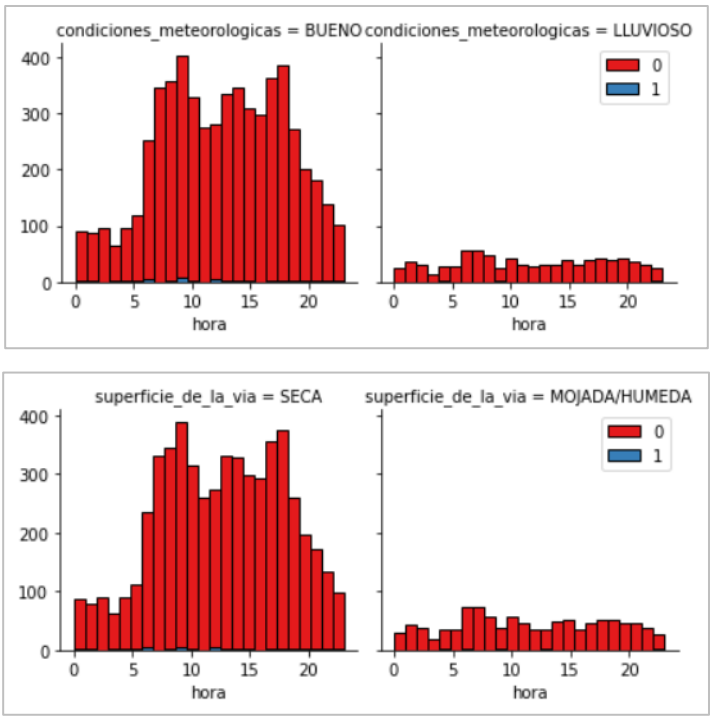

Figure 2. Relation between quantity of people died, hour and weather vs road condition.

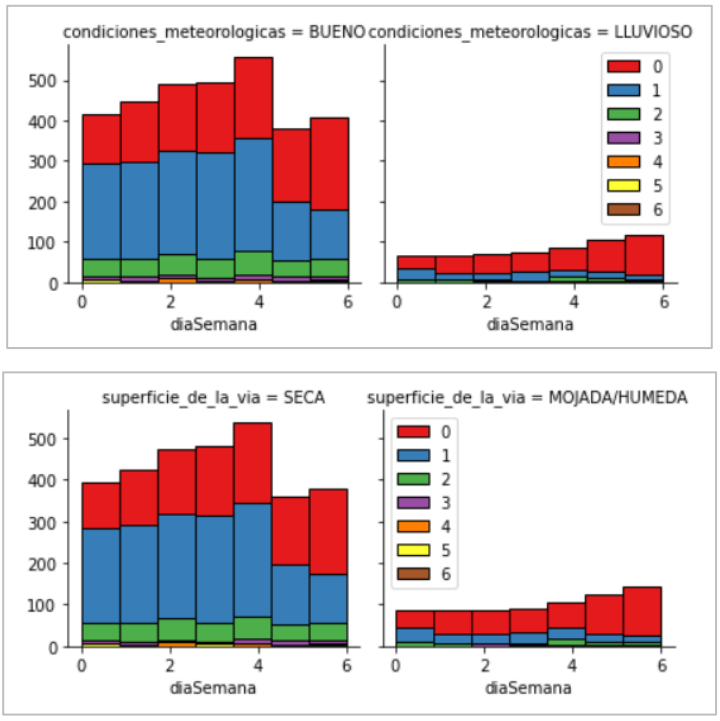

Figure 3. Relation between quantity of people injured, day of week and weather vs road condition.

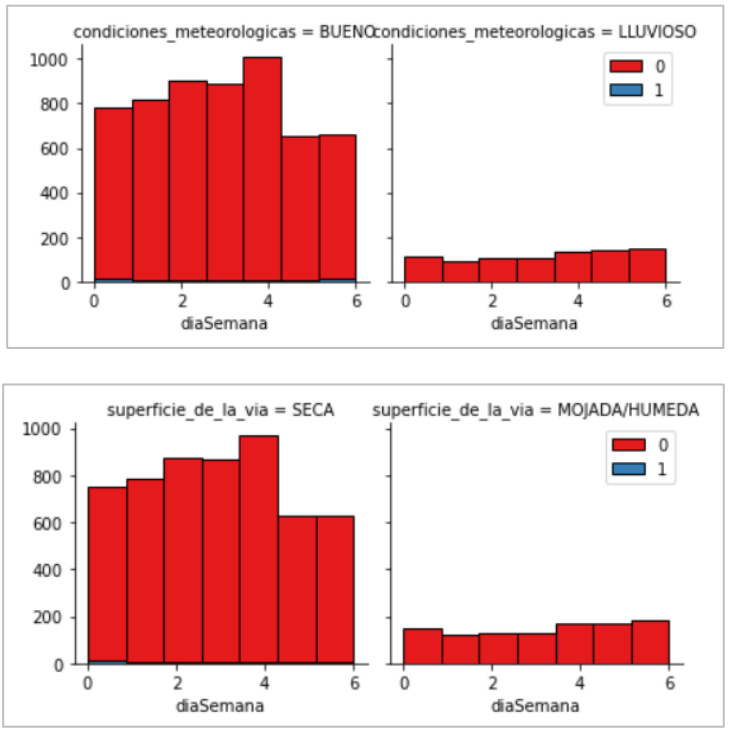

Figure 4. Relation between quantity of people died, day of week and weather vs road condition.

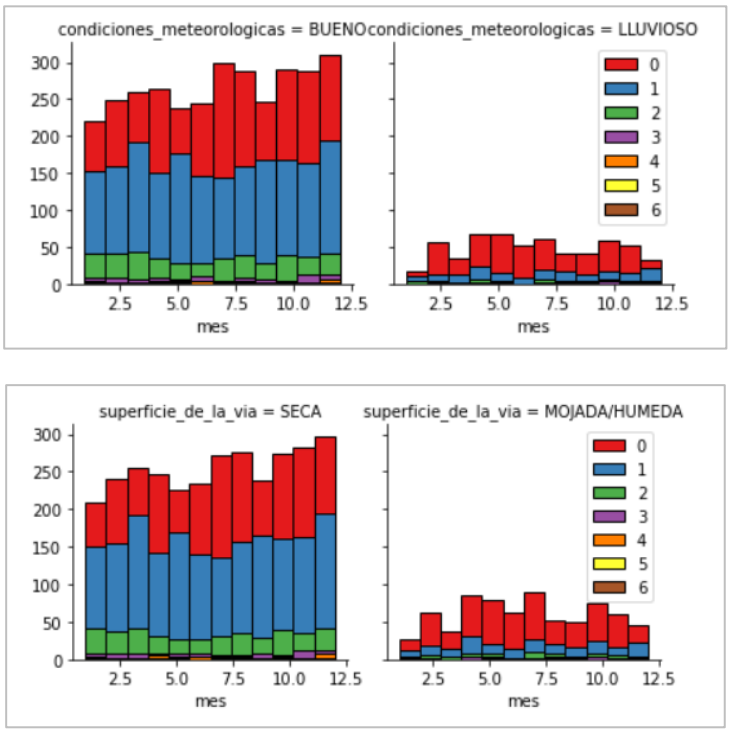

Figure 5. Relation between quantity of people injured, month and weather vs road condition.

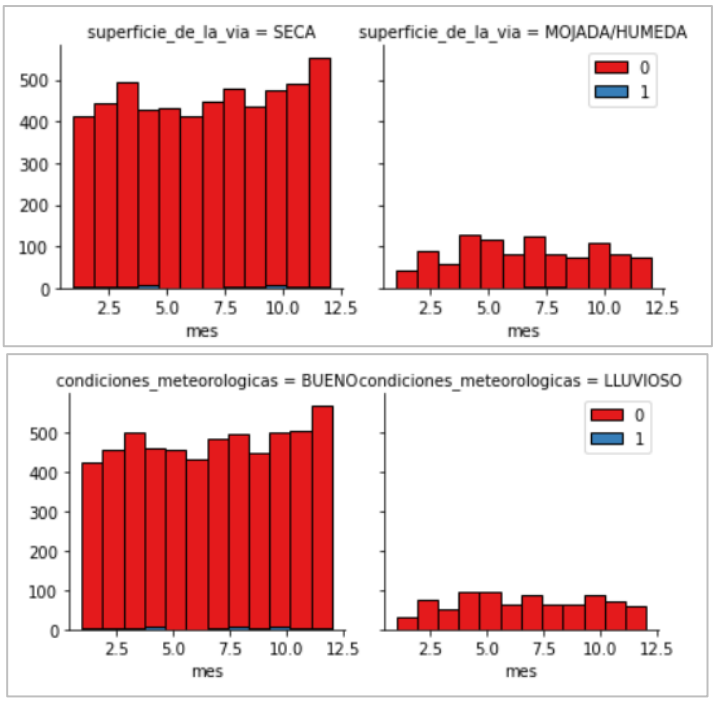

Figure 6. Relation between quantity of people died, month and weather vs road condition.

After analyzing these graphs some conclusions can be taken:

- The most likely hours for a car accident to occur are 9 am, 2 pm and 6 pm.
- The most likely day of the week for a car accident to occur is Friday.
- The most likely months for a car accident to occur are December and July.

Also, it is easy to see that the distribution between the variables "condiciones_meteorologicas" (weather) and "superficie_de_la_via" (road condition) are similar. To have both variables in the model will not add any rich information, so only one is kept and the other is dropped.

3.2. Define Target

At this point, there are two columns related with the severity of an accident: quantity of people injured and quantity of people died. The plan is to have only one column to predict, the reason is that it will simplify the algorithm. So the following rule was defined:

- Variable “severidad” (severity) was created
- If quantity of people injured and quantity of people died is equal to cero (0), “severidad” is cero (0), this means that if there is an accident there would be no people injured or died.

- If quantity of people injured is mayor or equal to one (1) and quantity of people died is equal to cero (0), “severidad” is one (1), this means that if there is an accident there would be at least one injured person.
- If quantity of people died is mayor or equal to one (1) no matter the value of quantity of people injured, “severidad” is two (2), this means that if there is an accident there would be at least one died person.
- Columns quantity of people injured and quantity of people died were dropped.

Distribution graph with severity are:

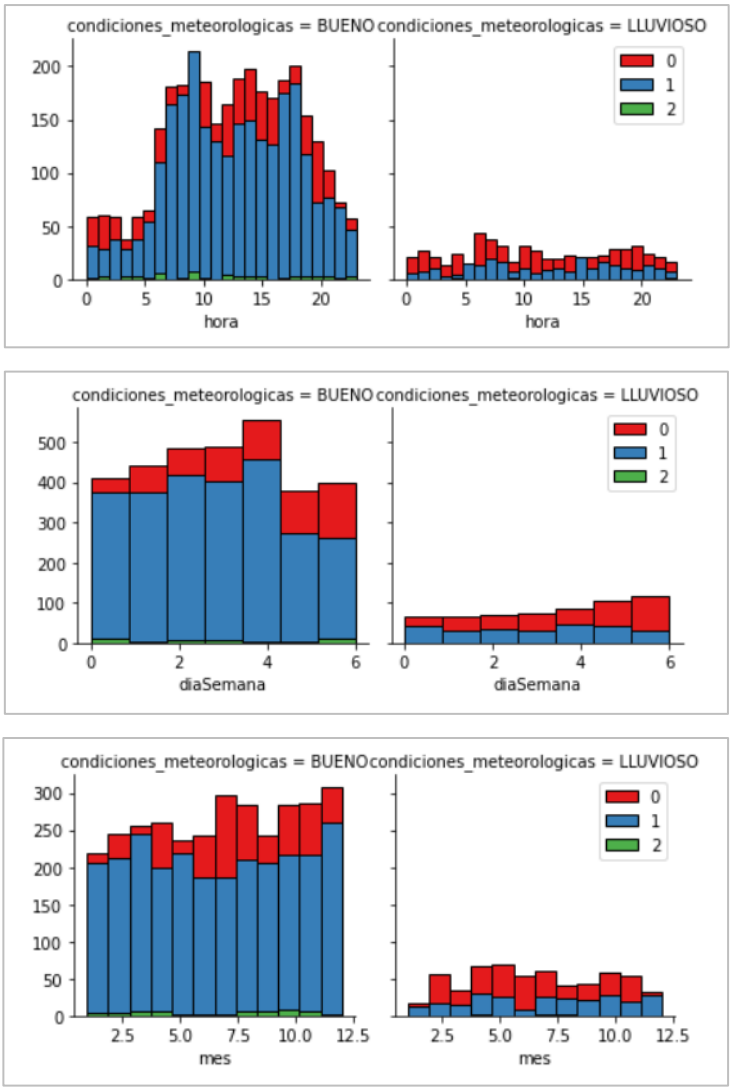

Figure 7. Relation between severity, weather and hour, day of week and month.

At last, dummy variables are set for weather, so the model learning algorithms can use as input variables only numerical values. After cleaning 6615 rows remain in total.

The final variables are:

Input: hour, day of week, month, Bueno (good weather), Lluvioso (bad weather).

Target / variable to predict: severity.

The data frame is:

	hora	diaSemana	mes	año	severidad	BUENO	LLUVIOSO
0	1	2	1	2014	1	1	0
1	3	2	1	2014	0	1	0
3	7	5	1	2014	0	1	0
4	21	5	1	2014	1	1	0
5	9	1	1	2014	0	1	0

Figure 7. First five rows of the data frame to analyze.

Column "año" (year), was not used.

3.3. Predictive Modeling

For this project supervised machine learning algorithms are the chosen one, because the target to predict is already know for the data set. Four classification models were used: K-Nearest Neighbor, Decision Tree, Support Vector Machine and Logistic Regression.

"X" variables (hour, day of week, month, Bueno, Lluvioso) were normalized. "Y" variable (severity) is the target (to predict). Also, data set was split into train and test data. 80% of the total for the first (5292 rows) and 20% of the total for the second (1323 rows).

K-Nearest Neighbor

The following graph shows the best K:

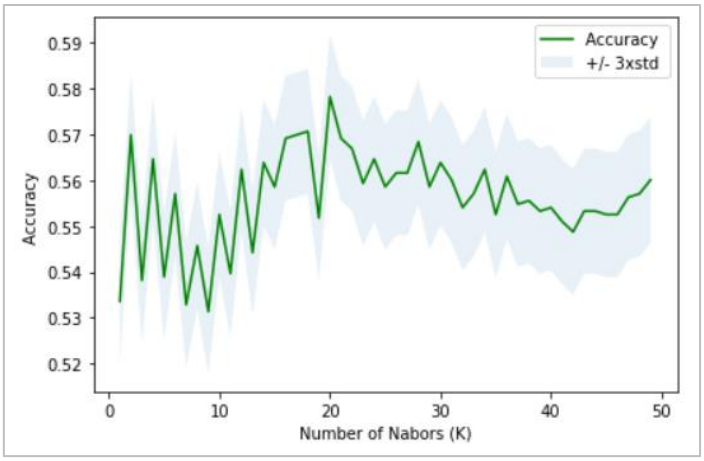

Figure 7. Accuracy of the model for different K.

Best value is K = 20.

Decision Tree

Criterion was set to “entropy”.

Support Vector Machine

Kernel was set to “rbf”.

Logistic Regression

Solver was set to “liblinear” and C = 0.01.

4. Results

After applying the models, the prediction of each one was compared with the real “Y” train value, and with the real “Y” test value with two criteria: Jaccard Index and F1-score. Also, for Logistic Regression, Log loss was used.

	K-Nearest Neighbor	Decision Tree	Support Vector Machine	Logistic Regression	K-Nearest Neighbor	Decision Tree	Support Vector Machine	Logistic Regression
Data Set	Train				Test			
Jaccard Index	61.04%	75.76%	56.50%	55.78%	57.82%	54.27%	58.50%	58.28%
F1-score	58.45%	75.05%	45.53%	44.49%	54.12%	53.33%	47.94%	48.30%
Log loss	-	-	-	72.22%	-	-	-	75.50%

Table 1. Accuracy of each model.

5. Discussion

As is it shown in *Table 1*, regarding train data, the best performance is Decision Tree model with an accuracy of around 75%, but for test data has an accuracy of around 54%. This is a poor performance. All other three models have also low performance, accuracy below 60%. So in conclusion, with this conditions, those models are not a good estimator for severity.

6. Conclusion

In this study, data about car accidents in Buenos Aires City and the relation between its variables was analyzed: hour, month, day of the week and weather. A cleaning of the data was done, which mainly consisted of deleting variables not needed, deleting outliers and using two input variables (quantity of people injured and died in an accident) to define the new target: severity. To predict it, four machine learning classification models were used (K-Nearest Neighbor, Decision Tree, Support Vector Machine and Logistic Regression). The idea of the project is to help people: depending the condition, if there is a car accident the model predict the severity and each car driver who travel to or out of Buenos Aires City eventually could pay more attention while driving.

Unfortunately, the accuracy obtain was poor (below 60%), so models are not good estimators. I personally think that quantity of data lines is not enough, with more lines a better model should be develop. But the quantity of data set available in Argentina is not too much. So, the first step to improve this study is to record more quantity of data, but this takes an huge effort and needs a monetary inversion.