1 Introduction

This study seeks to determine the variables and conditions that cause traffic accidents and their degree of severity. This will allow us to know the conditions that cause these accidents and thus to be able to diminish its exhibition, managing to diminish fatal and nonfatal accidents. Some of the variables that are analyzed are time, road conditions, light conditions, etc. The information collected through the data will be used to develop an automatic learning algorithm, which allows the prediction of traffic accidents, and their severity given certain conditions. The main stakeholders in this project are the community, vehicles and road users, among others.

1.1 Background

The level of accidents increases with the number of cars on the road. There are multiple variables that impact on the occurrence of accidents, these factors can be the weather, lighting, etc.

1.2 Problem

Many accidents occur daily with varying degrees of severity as but are conditioned by external and internal factors. It is important then to generate a model that predicts which variables make each type of accident more likely to occur and thus be able to reduce these risks.

1.3 Interest

The main stakeholders in this project are the community, vehicles users, and road users, among others.

2 Data acquisition and cleaning

2.1 Data sources

The data used corresponds to collision information provided by the SPD and recorded by Traffic Records from 2004 onwards on a weekly basis. It consists of 194,673 records and 37 attributes each. Among the main ones in analysis will be considered the severity of the accident, for example fatal or non-fatal, the type of collision, the number of deaths, etc. The variables considered will be characteristics such as the weather, road conditions.

2.2 Data cleaning

First the data should be cleaned and records that do not provide sufficient information should be deleted.

We will consider the severity code, type of direction, weather, light conditions, road conditions and number of vehicles involved. Null values are discarded giving a total of 187.525 rows to work with.

The categories are then simplified through the following actions: For the case of road conditions we will only focus on Dry and Wet, the rest will be grouped in the category others. In the case of weather, we will consider 3 main ones, clear raining overcast and the rest in the category Other. For the Light conditions we will group the Dark types in only one category.

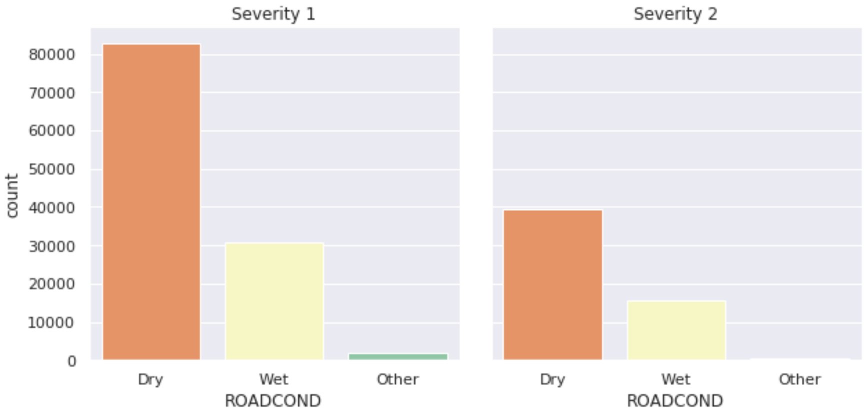
3 Exploratory Data Analysis

3.1 Calculation of target variable

The variable of interest selected corresponds to the type of severity of the accident which is defined by the code associated with each type of accident.

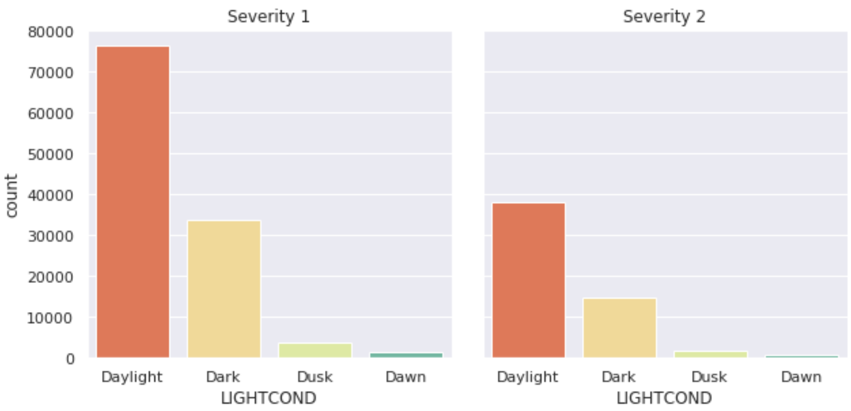
3.2 Road Conditions and Severity

It is possible to appreciate that accidents occur mainly in the presence of Dry road conditions, and secondly Wet. The totals for each category are, Dry with 123,736, Wet with 47,223 and finally Other with 16,566.



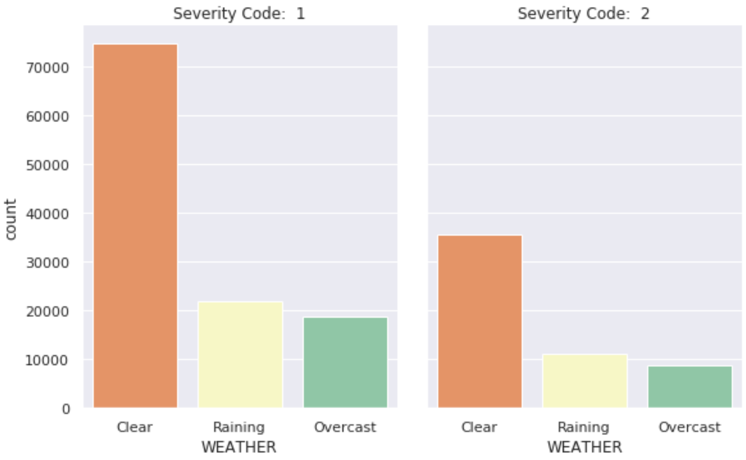
3.3 Light Conditions and Severity

Most accidents occur in daylight for both cases of severity, then with lower values when it is dark. The total for Daylight is 114,577, for Dark it is 48,491 for Dusk 5,612 and the lowest values for Dawn with 2,346.



3.4 Weather and Severity

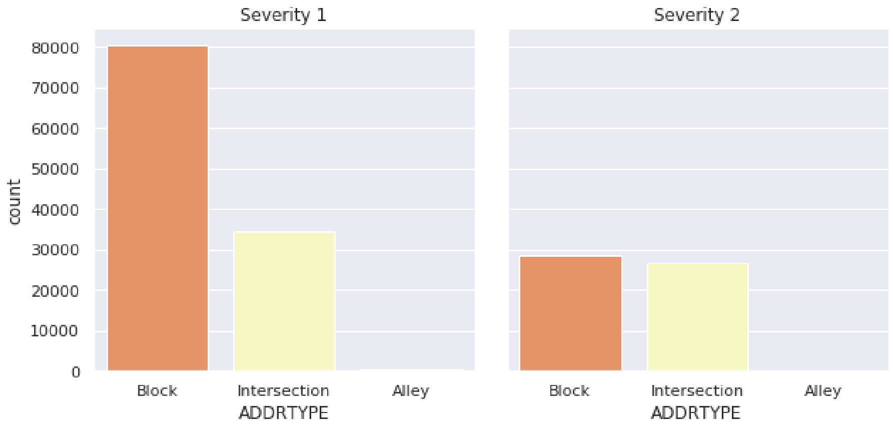
When the weather is Clear it is when more accidents occur although with a higher number of accidents of severity type 1, the rest of the conditions have similar values. According to the total data we have the following composition, Clear with 110.499, Raining with 32.976, Overcast with 27.551 and finally Other with 16.499.



3.5 AddrType and Severity

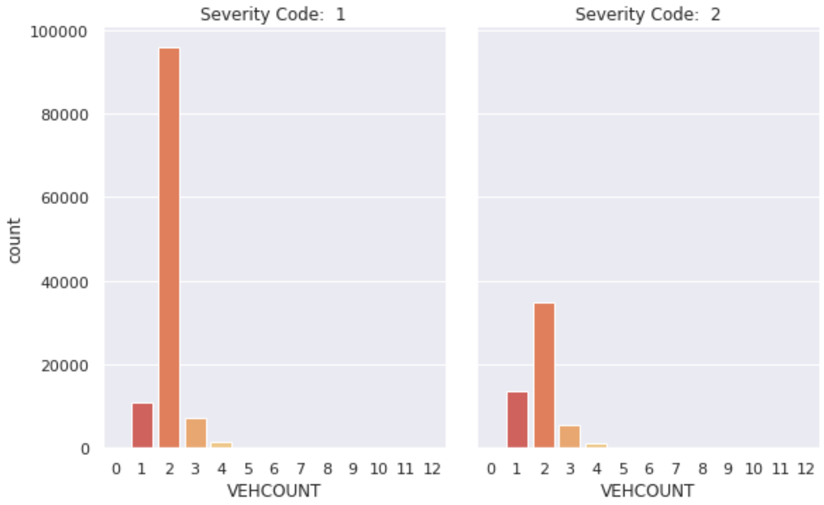
There is an appreciable difference according to the type of direction when the severity is 1, since in cases where the direction is Block it exceeds by more than double the case of Intersection. However, the values in the case of severity 2, are very similar.

According to the total data we have the following composition, Block with 123,321, Intersection with 63,462 and finally Alley with 742.



3.6 VehCount and Severity

Most accidents with severity 1 occur between 2 vehicles, followed by 1 and then 3. For severity 2, the same order is maintained but the difference between the categories decreases. According to the number of vehicles, the highest frequency occurs for 2 involved with 130,720, followed by 1 with 24,496, then 3 with 12,533.



4 Predictive Modeling

4.1 Classifications models

Classification models are supervised learning algorithms that seek to classify data into a set of discrete value classes.

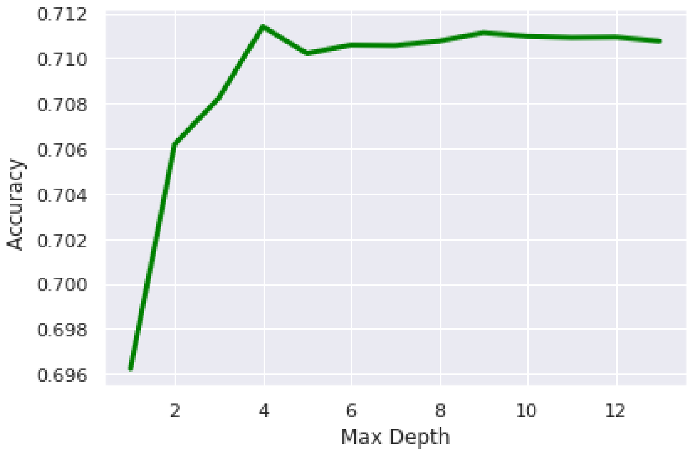
This type of algorithm learns from the relationship between a set of categorical variables and a variable of interest, in which the latter is a categorical variable with discrete values.

This study will consider 3 types of classification models, Decision tree, K neighbors and logarithmic regression.

4.1.1 Decision Tree

As a starting point you must choose an attribute within the dataset, then calculate the significance of the attribute when separating the data. This should be done to find the best attribute After obtaining the model with the best significance we can predict based on input information.

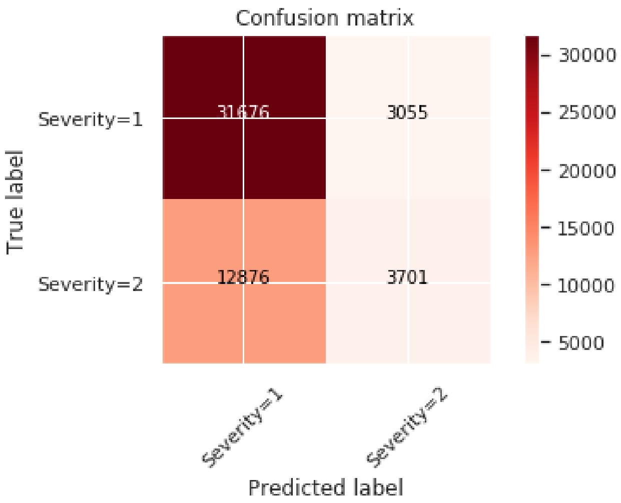
The level of depth that generates the highest precision as analyzed in this case corresponds to 4.



4.1.2 Logistic Regression

Logistic regression aims to classify data from a set considering the input variables in order to predict a target variable which must be binary. It is important to note that the independent variables must be continuous.

When making the confusion matrix we can see that in general the model correctly predicts the values.

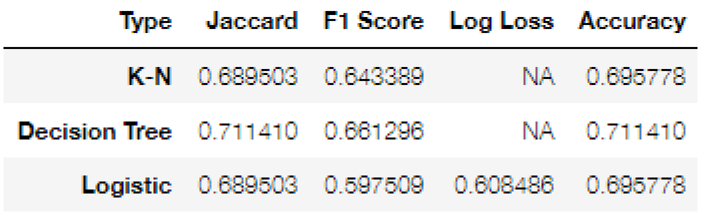


4.1.3 K-Nearest Neighbors

The K-Nearest Neighbors algorithm is a classification method which groups diverse points with common characteristics to learn from the relationships between them to label unknown information. Those data that are close to each other are called neighbors.

4.2 Evaluation of models

When evaluating each model according to indicators such as accuracy, f1 score, log loss or jac card we have the following table.



Type	Jaccard	F1 Score	Log Loss	Accuracy
K-N	0.689503	0.643389	NA	0.695778
Decision Tree	0.711410	0.661296	NA	0.711410
Logistic	0.689503	0.597509	0.608486	0.695778

Based on this we can notice that the model that offers greater precision with respect to the projections is the decision tree model, followed by K-Nearest Neighbors and finally the logistic regression. It is possible to appreciate that the models offer similar levels to each other, but the best option lies in the decision tree.

5 Conclusions

Finally, by way of conclusion, each model offers a medium high level of precision. The model that is recommended, however, corresponds to the decision tree since it offers the best indicators, surpassing the other models.