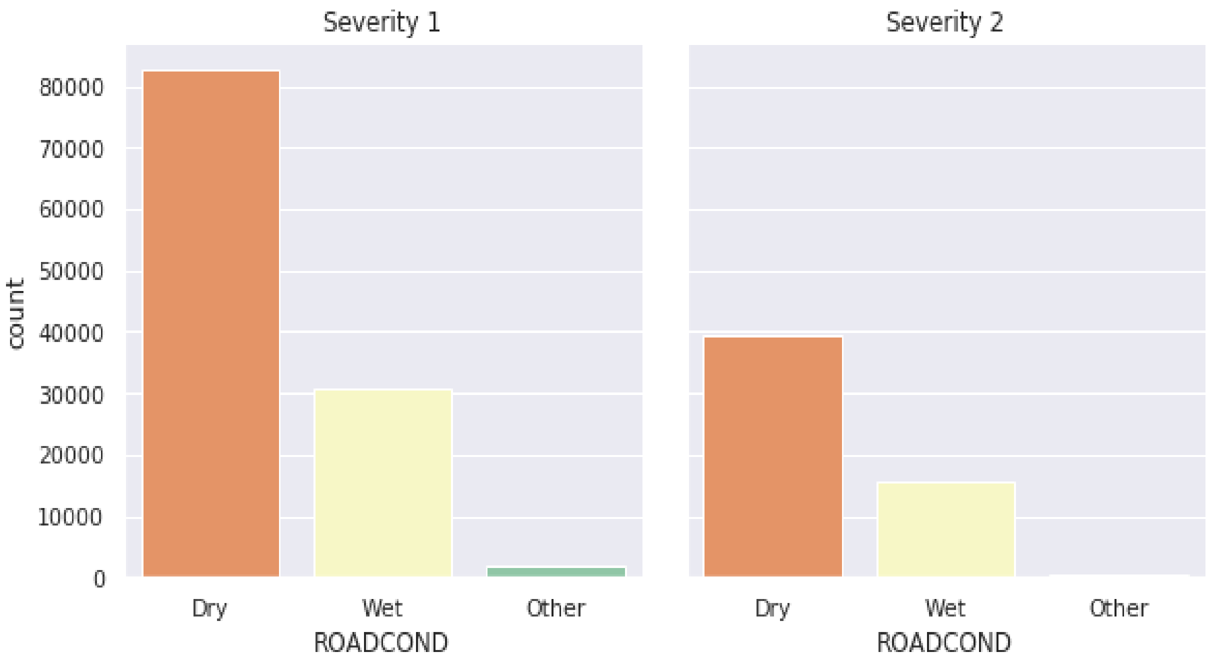
Background and Problem

- 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.

Data

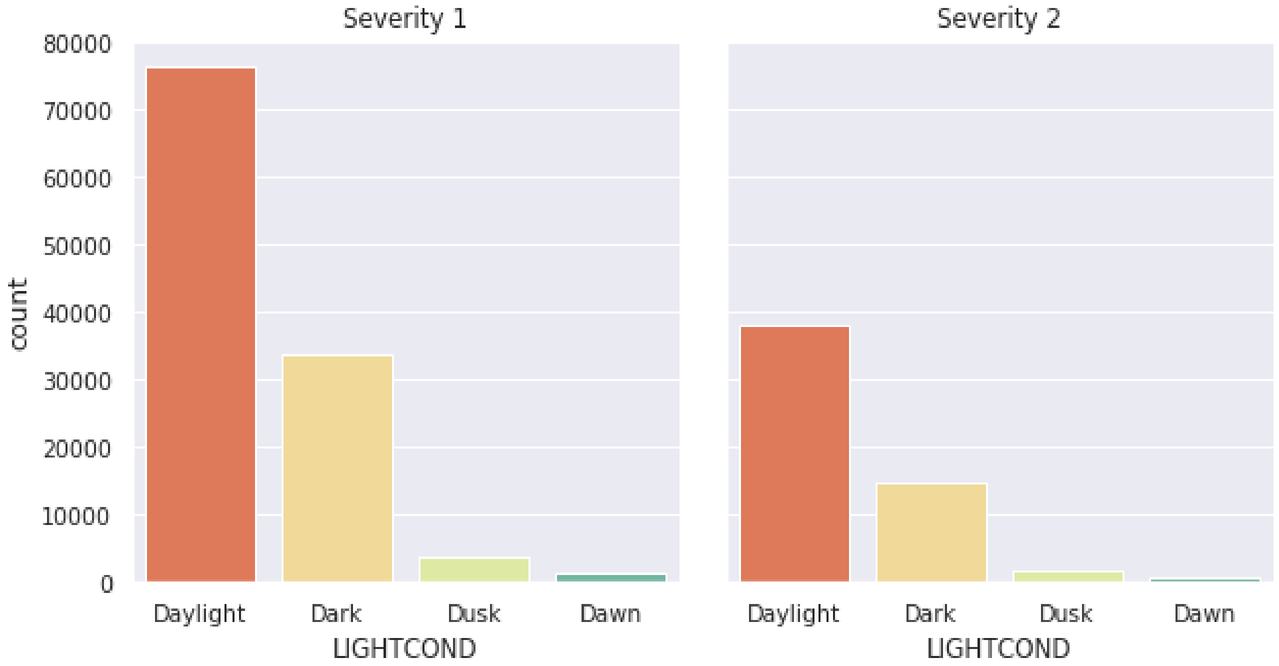
- The data used corresponds to collision information provided by the SPD and recorded by Traffic Records from 2004 onwards on a weekly basis. It consists of 194,673 records and 37 attributes each.
- Null values are discarded giving a total of **187.525** rows to work with.
- The categories are then simplified.
- 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.

Road Conditions and Severity



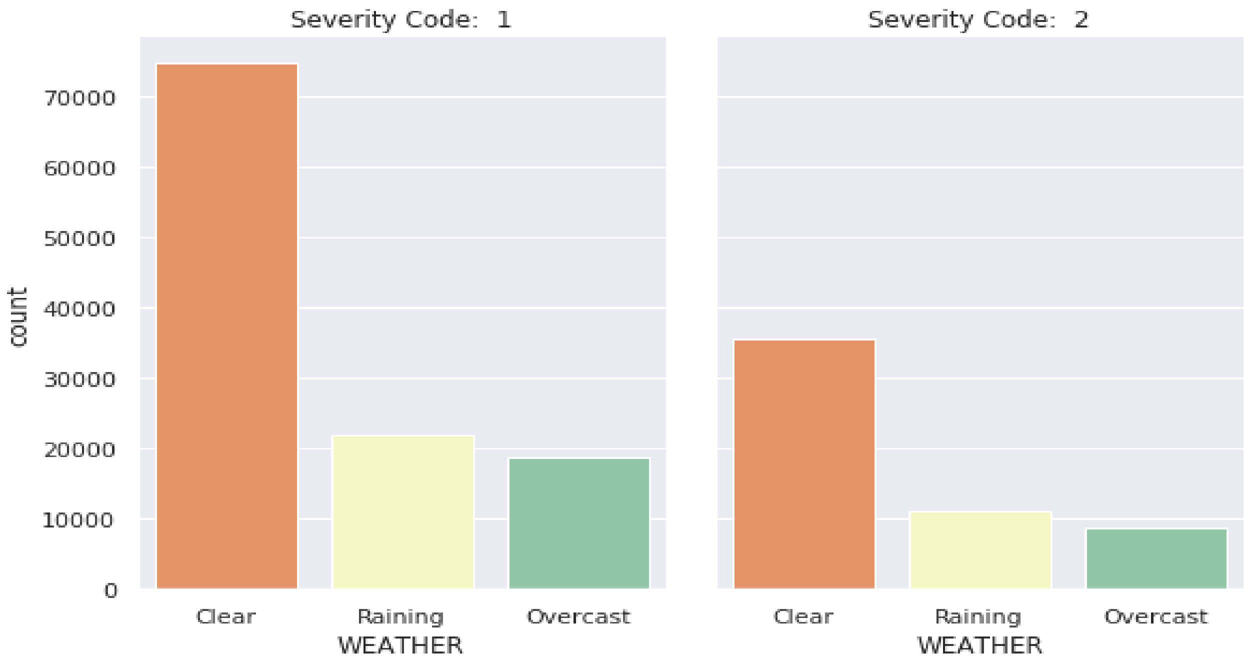
It is possible to appreciate that accidents occur mainly in the presence of Dry road conditions, and secondly Wet.

Light Conditions and Severity



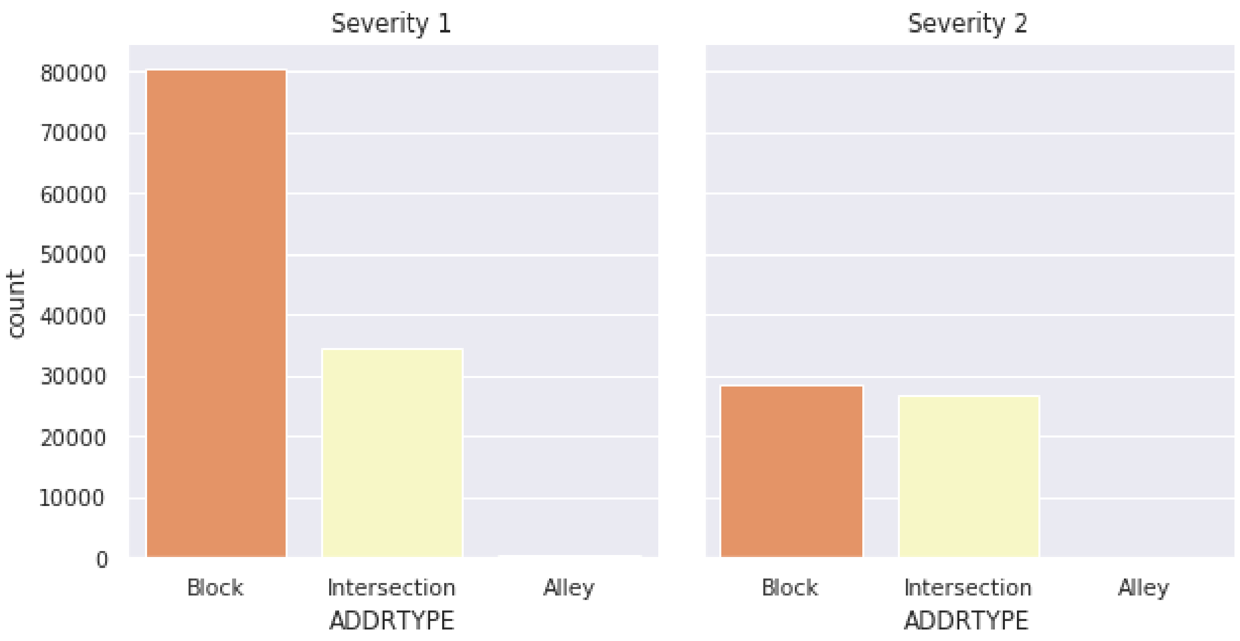
Most accidents occur in daylight for both cases of severity, then with lower values when it is dark.

Weather and Severity



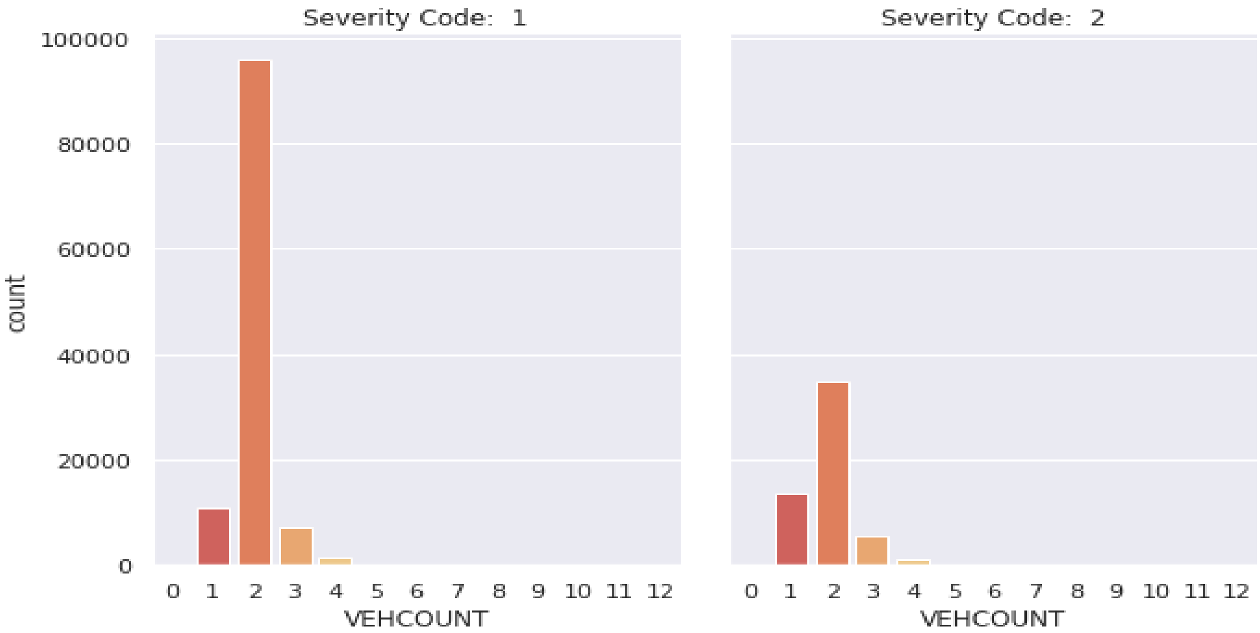
When the weather is Clear it is when more accidents occur although with a higher number of accidents of severity type 1, the rest of the conditions have similar values.

Address 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.

Vehicle Count 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.

Classifications Models

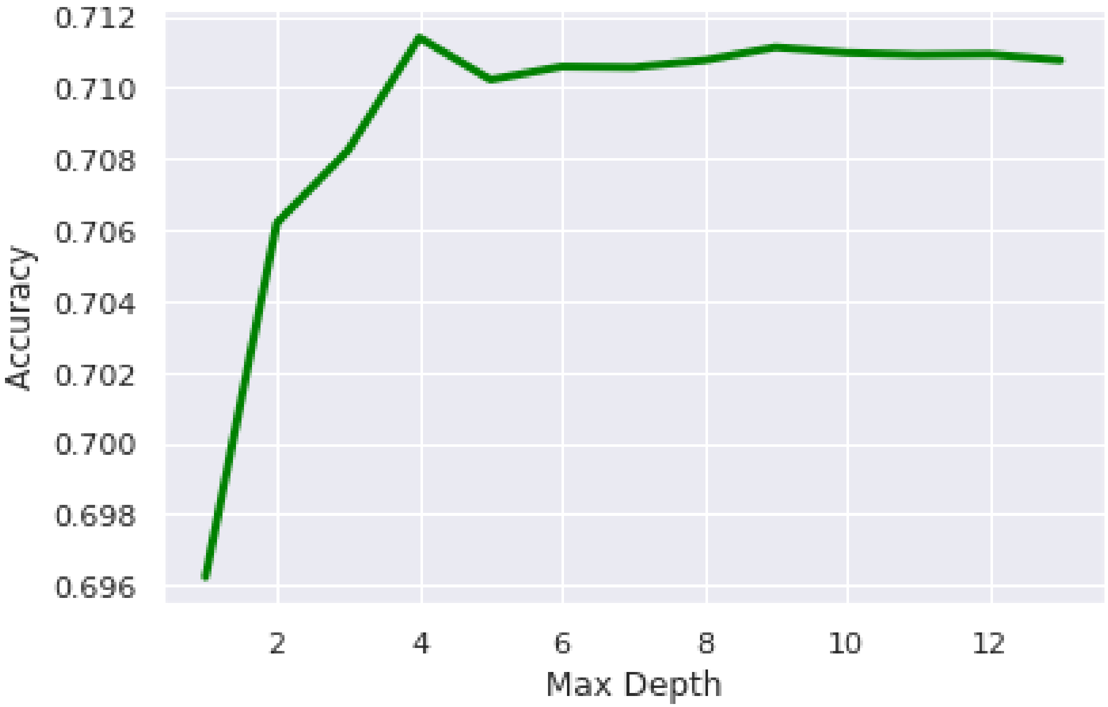
Classification models are supervised learning algorithms that seek to classify data into a set of discrete value classes.

This study will consider 3 types of classification models:

- Decision tree
- K-Near Neighbors
- Logarithmic regression.

Decision Tree

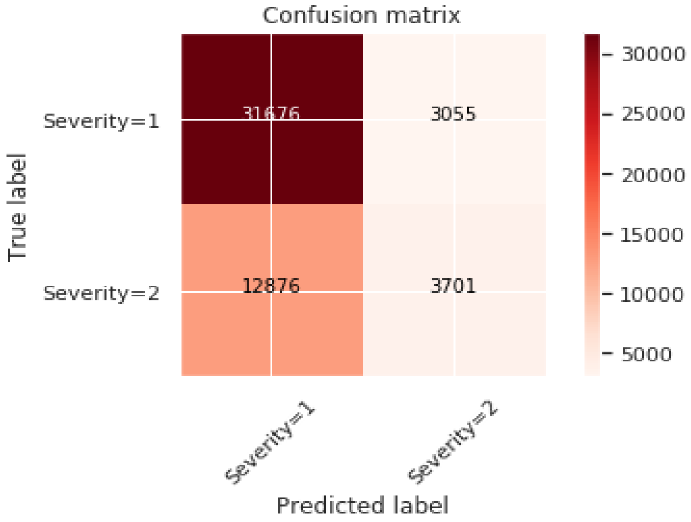
- As a starting point you must choose an attribute within the dataset, then calculate the significance of the attribute when separating the data.
- The level of depth that generates the highest precision as analyzed in this case corresponds to 4.



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.

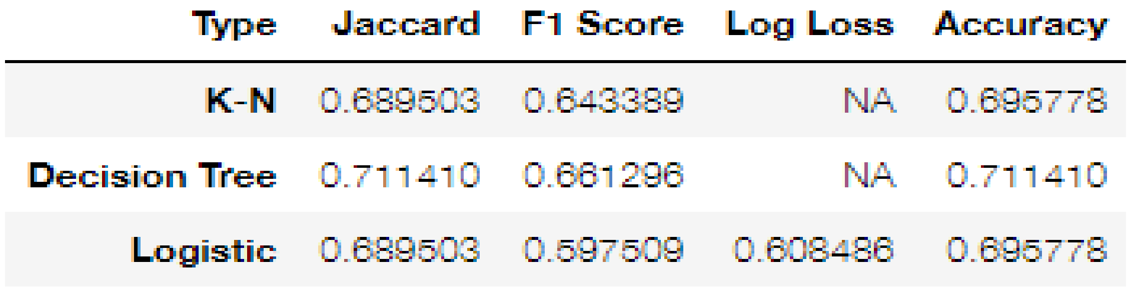


K-Near Neighbor

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.

Evaluation

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



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

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.