Prediction of car accident severity

Carlo Peano

September 30, 2020

1. Introduction

1.1 Background

According to the WHO[1], even though vehicles have become much safer in the last decades, every year around 1.35 million people still die because of a road traffic crash and between 20 and 50 million more people suffer non-fatal injuries with many incurring a disability.

If we take a different perspective and consider the economic impact at national level, road traffic accidents also cost around 3% of gross domestic product to most countries[2].

Therefore, there is a great interest in different parts of society (such as governments, decision-makers, carmakers, drivers, insurance companies) in changing and decreasing this trend.

1.2 Problem

A solution that would reduce the number of incidents could be the chance to warn a driver about the possibility of getting into a car accident and how severe that incident would be, given the weather and road conditions. In this way people would drive more carefully or even stay at home.

Transforming this solution into a machine learning problem, I used a dataset provided by a city and its police department (in our case Seattle City and the SPD - Seattle Police Department) to predict the severity (and its probability) of an accident based on the conditions of weather, light and the road.

[1] "Road traffic injuries", World Health Organisation (WHO), 07/02/2020, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
[2] Ibid.

2. Data

2.1 Data understanding

In this project, I used the data provided by the SPD (Seattle Police Department) and recorded by Traffic Records. This dataset - called Data-Collisions.csv - includes all types of collisions involving cars, bikes, pedestrians and others (around 200,000) from 2004 to present.

After I extracted the dataset, I looked at the columns, their meaning and their relation to the objective of the research study, i.e. to predict the probability and severity of an accident based on the conditions of weather, light and the road.

Thanks also to the description of the attributes (available together with the dataset[3]), I have been able to define the attributes and the target variable. In my opinion, it was obvious to choose SEVERITYCODE (i.e. the severity of the accident) as the dependent variable. SEVERITYCODE is a categorical variable and follows a code that corresponds to the severity of the collision: 2 (injury) and 1 (property damage).

Out of the 37 attributes available in Seattle accident dataset, I chose 7 of them as independent variables, thanks – as previously said – to their logical connection to the objective of our research study.

Variable	Description
JUNCTIONTYPE	Category of junction at which collision took place
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
SPEEDING	Whether or not speeding was a factor in the collision
PERSONCOUNT	The total number of people involved in the collision
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state

Table 2.1.1 Variables and their description

JUNCTION, WEATHER, ROADCOND, and LIGHTCOND are the main attributes since they are directly connected to the project's objective.

PERSONCOUNT and VEHCOUNT allow understanding how big the accident can be: an accident can involve a lot of vehicles and people and still have nobody injured or no property damage.

Lastly, SPEEDING has always been considered to have a direct impact on the probability of collision and is the only attribute that is a choice of the driver.

[3] "ArcGIS Metadata Form", https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

2.2 Data Preparation

2.2.1 Data cleaning

Once I chose the attributes and the target variable, I dropped the unnecessary columns and analysed more deeply the necessary ones.

At this point, there were several problems with the dataset.

Firstly, most the eight attributes had missing data because the SPD did not write all the data:

1. "SPEEDING" has 185340 missing data
2. "JUNCTIONTYPE" has 6329 missing data
3. "WEATHER" has 5081 missing data
4. "ROADCOND" has 5012 missing data
5. "LIGHTCOND" has 5170 missing data
6. "PERSONCOUNT" has 0 missing data
7. "VEHCOUNT" has 0 missing data

If most of the missing data were easily solvable by deleting them, those ones of SPEEDING seemed immediately problematic, as more than 90% of the data was missing. However, a direct observation of the attribute values showed that the police took in consideration this attribute – by writing "Y" – only when speed was one of the causes of the accident, otherwise they left the variable empty. As a consequence of this, I considered all the missing data as an "N", i.e. speed was not one of the reasons of the incident.

Secondly, four attributes had data internally classified as "Other" and/or "Unknown" which were a sort of hidden missing data as they did not actually provide a real information about the attribute and could actually provide confusion:

1. "JUNCTIONTYPE" has 9 "Unknown"
2. "WEATHER" has 832 "Other" and 15091 "Unknown"
3. "ROADCOND" has 132 "Other" and 11012 "Unknown"
4. "LIGHTCOND" has 235 "Other" and 13473 "Unknown"

I used different methods for managing the two parameters. "Other" means that the data cannot be replaced by any other available observation, therefore it was replaced with NaN and its rows were dropped together with the missing data. "Unknown" was replaced with the mode of the related attribute so as to avoid biasing the dataset.

Thirdly, there was a problem with PERSONCOUNT (the total number of people involved in the collision) and VEHCOUNT (the number of vehicles involved in the collision): PERSONCOUNT had 5544 incidents involving nobody (zero people) and VEHCOUNT had 5085 accidents involving zero vehicles. These observations were not interesting for the project as its objective implies the involvement of people and vehicles, the lack of one or both of them could bias the dataset. Therefore, I drop those rows were PERSONCOUNT and VEHCOUNT were zero.

2.2.2 Correct data format

Considering the data format of the chosen attribute (see Table 2. Data format), I did not need to change any of them.

Attribute	Type
SEVERITYCODE	int64
PERSONCOUNT	int64
VEHCOUNT	int64
JUNCTIONTYPE	object
WEATHER	object
ROADCOND	object
LIGHTCOND	object
SPEEDING	int64

Table 2.2.2.1 Data format

2.2.3 Feature selection

In order to balance the dataset, I used one hot encoding technique to convert categorical variables to binary variables and append them to the feature Data Frame. Then I defined the feature set X and the labels y and normalise the data.

3. Exploratory Data Analysis (Methodology – part 1)

3.1 Relation between speeding and accidents

Speeding	Number of accidents	% of accidents
No	167476	94.83%
Yes	9122	5.17%

Table 3.1 Relation between speeding and accidents

When I examined the value counts of speeding and considered what impact speeding had on accidents, I saw that this attribute would not have been a good predictor variable for the severity of an accident (and its probability). This is because only 5.17% of the accidents are caused by driving too fast: this result is skewed. Thus, we are not able to draw any conclusions about this attribute.

3.2 Relation between the type of junction and accidents

Accidents happen in junctions since passing through one of them puts the drivers in a more stressful situation. Comparing the different types of junctions, I discovered that most accidents (47,16%) take place in mid-block crossing (not related to intersections).

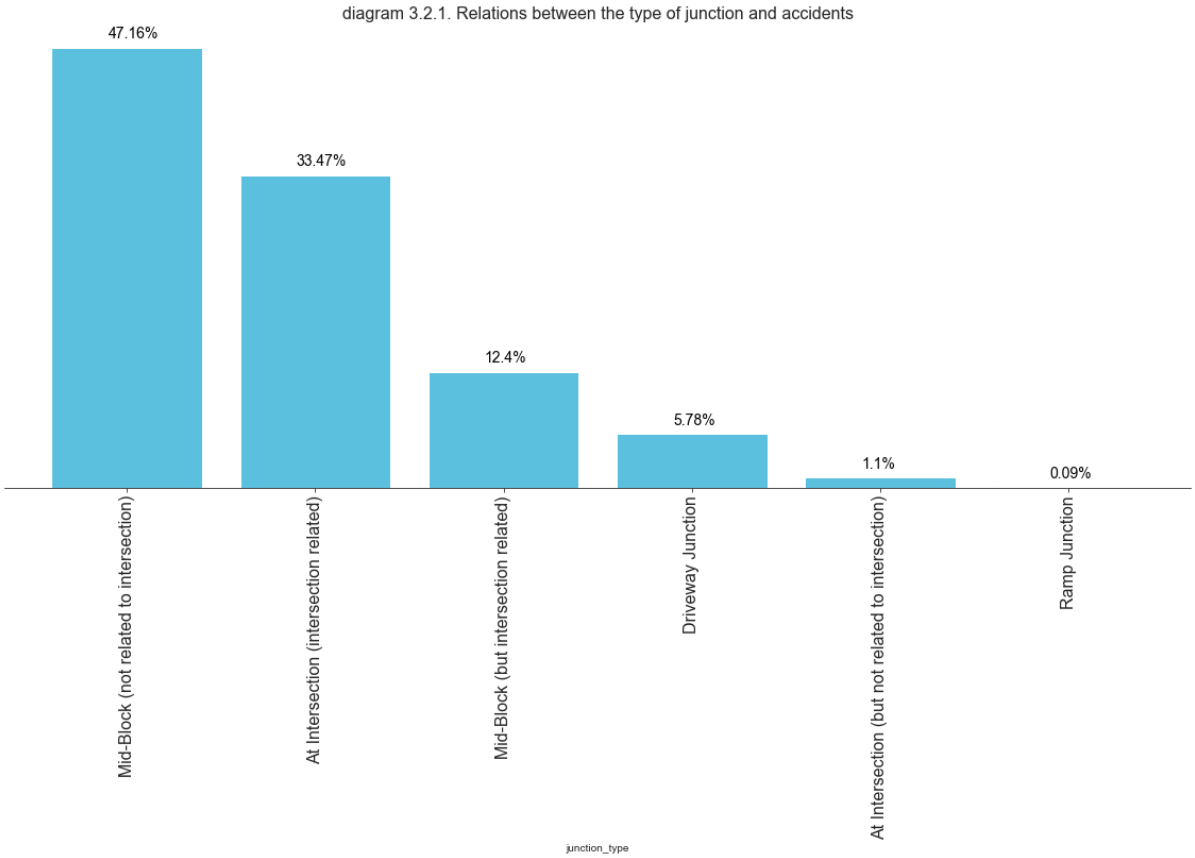
Type of junction	Number of accidents	% of accidents
Mid-Block (not related to intersection)	83286	47.16 %
At Intersection (intersection related)	59106	33.47 %
Mid-Block (but intersection related)	21890	12.40 %

Type of junction	Number of accidents	% of accidents
Driveway Junction	10215	5.78 %
At Intersection (but not related to intersection)	1944	1.10 %
Ramp Junction	157	0.09 %

Table 3.2.1 Relation between the type of junction and accidents

Accidents at intersections (related to the intersection) are the second most usual incidents (33.47%) and happen more than the double of the time than the incidents in mid-block crossing related to intersections (12.40%).

These first three types of junction together are where most of the accidents happen (more than 90%), and this is evident if we look at the chart below (diagram 3.2.1).



If we consider the kind of severity that the incident can have, accidents with property damages are always more often than those ones with injuries.

Type of junction	Severity (1 = property damage, 2 = injury)	Number of accidents
At Intersection (but not related to intersection)	1	1365
	2	579
At Intersection (intersection related)	1	33337
	2	25769
Driveway Junction	1	7105
	2	3110
Mid-Block (but intersection related)	1	14838
	2	7052
Mid-Block (not related to intersection)	1	65293
	2	17993
Ramp Junction	1	107
	2	50

Table 3.2.2 Relation between the type of junction and the severity of the accidents (1 = property damage, 2 = injury)

The majority of the accidents that takes place in mid-block crossings (not related to intersections) are incidents with property damages (i.e. 65,293): a number higher than the sum of incidents with property damages in the other junctions.

However, even though the accidents with injuries in mid-block crossings (not related to intersections) is still high (i.e. 17,993), it is smaller than the incidents with injuries at intersections (i.e. 25.769).

This means that intersections are the most dangerous junctions.

These observations are evident if we look at the diagram 3.2.2, that shows the relations between the severity of the accidents and the type of junctions.

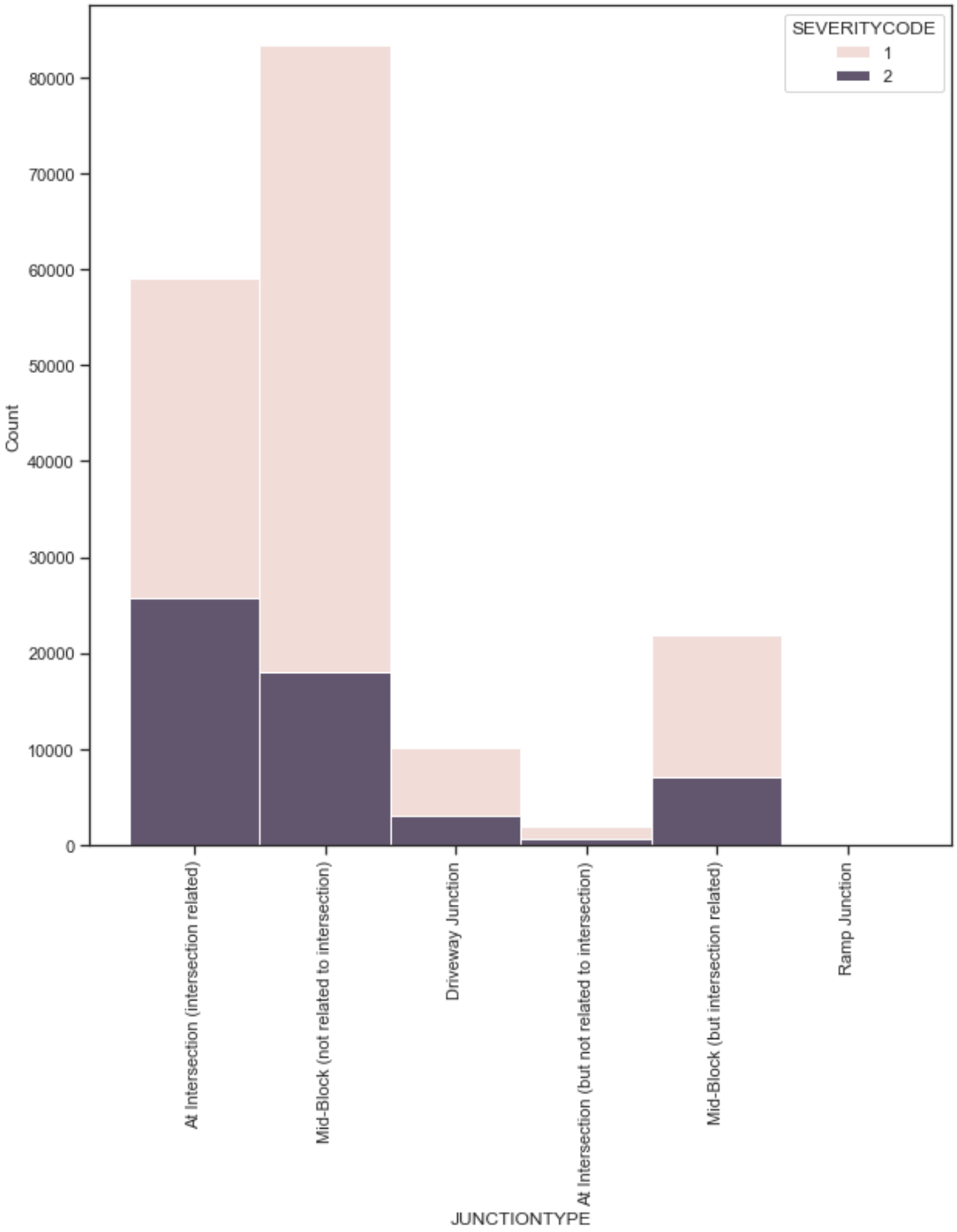


Diagram 3.2.2. Severity of the accidents and type of junction

3.3 Relation between the weather and accidents

Weather is another variable that can influence the driver and create conditions that increase the difficulty of driving.

Even though we may think that severe crosswind, hail, snow or fog are the main causes of the accidents, this is actually not true.

Weather	Number of accidents	% of accidents
Clear	116714	66.09
Raining	31793	18.00
Overcast	26509	15.01
Snowing	865	0.49
Fog/Smog/Smoke	533	0.30
Sleet/Hail/Freezing Rain	106	0.06
Blowing Sand/Dirt	48	0.03
Severe Crosswind	25	0.01
Partly Cloudy	5	0.00

Table 3.3.1 Relation between the weather and accidents

The larger number of accidents takes place with clear weather (66.09%), rain (18%) and overcast (15.01%) and together they are 99.10% of all the incidents.

Weather	Severity (1 = property damage, 2 = injury)	Number of accidents
Blowing Sand/Dirt	1	36
	2	12
Clear	1	81758
	2	34956
Fog/Smog/Smoke	1	357
	2	176
Overcast	1	18074
	2	8435
Partly Cloudy	1	2
	2	3
Raining	1	21017

Weather	Severity (1 = property damage, 2 = injury)	Number of accidents
	2	10776
Severe Crosswind	1	18
	2	7
Sleet/Hail/Freezing Rain	1	79
	2	27
Snowing	1	704
	2	161

Table 3.3.2 Relation between the weather and severity of the accidents

Like in the case of the junctions, the weather always has more incidents with property damages than with injuries. The chart below (Diagram 3.3.1) shows even better how the major part of the accidents – both with property damages and injuries – occurs with clear sky.

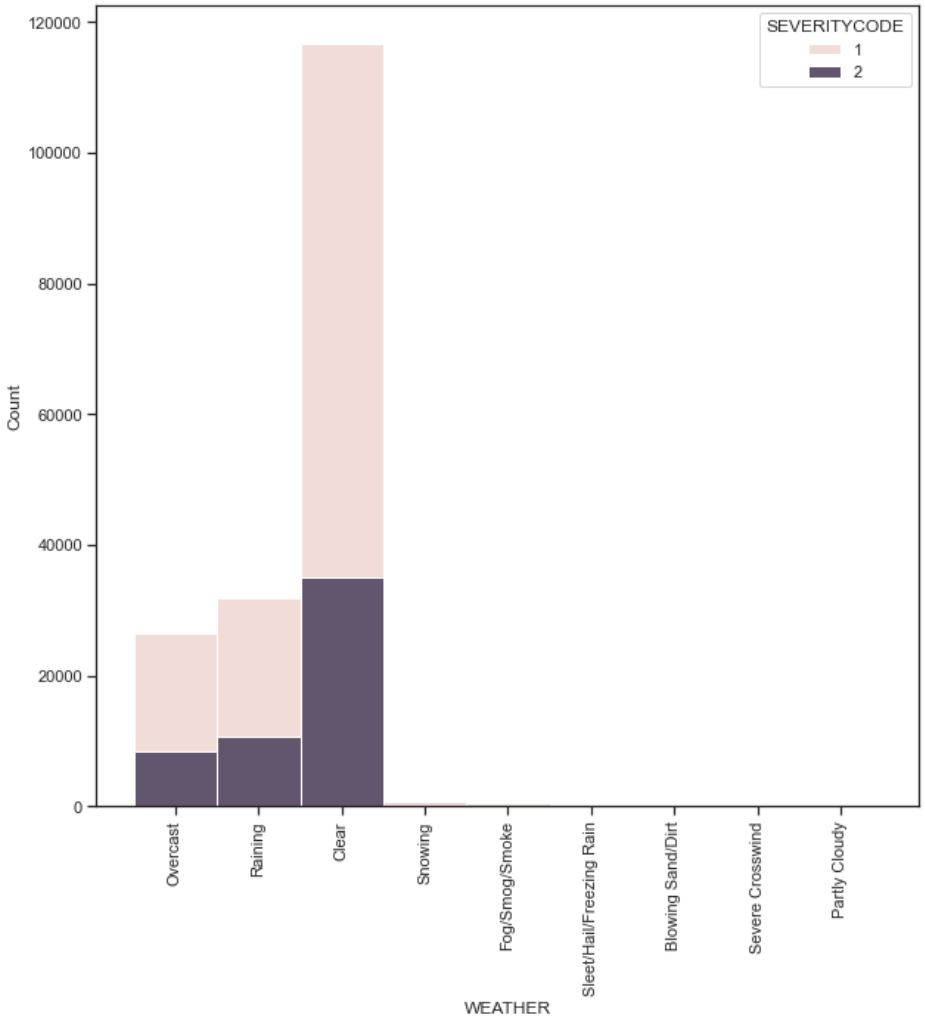


Diagram 3.3.1 Relation between weather and the type of severity (1 = property damage, 2 = injury)

3.4 Relation between the road conditions and accidents

Road conditions also have an impact on the accidents and their severity as they may put the driver in complicated and stressful situations. Circumstances, where the road has ice, snow, oil or water, are dangerous and may not be properly managed.

Road conditions	Number of accidents	% of accidents
Dry	128735	72.90
Wet	45538	25.79
Ice	1143	0.65
Snow/Slush	956	0.54
Standing Water	103	0.06
Sand/Mud/Dirt	65	0.04
Oil	58	0.03

Table 3.4.1 Relation between the road condition and accidents

Surprisingly, the vast majority of the accidents takes place when roads are dry, in 72.90% of the cases. Wet conditions of the road are the variable with the second higher percentage of incidents (25.79%).

These two variables together collect 98.69% of the accidents while the other apparently more complicated conditions are correlated to only 1.31%.

Road conditions	Severity (1 = property damage, 2 = injury)	Number of accidents
Dry	1	89879
	2	38856
Ice	1	888
	2	255
Oil	1	34
	2	24
Sand/Mud/Dirt	1	43

Road conditions	Severity (1 = property damage, 2 = injury)	Number of accidents
Snow/Slush	2	22
	1	795
	2	161
Standing Water	1	77
	2	26
Wet	1	30329
	2	15209

Table 3.4.2 Relation between the road conditions and the severity of the accidents (1 = property damage, 2 = injury)

Like with other previous attributes, road condition always has more incidents with property damages than with injuries. The chart below (Diagram 3.4.1) makes even more visible how the larger number of the accidents – both with property damages and injuries – occurs with dry conditions of the road.

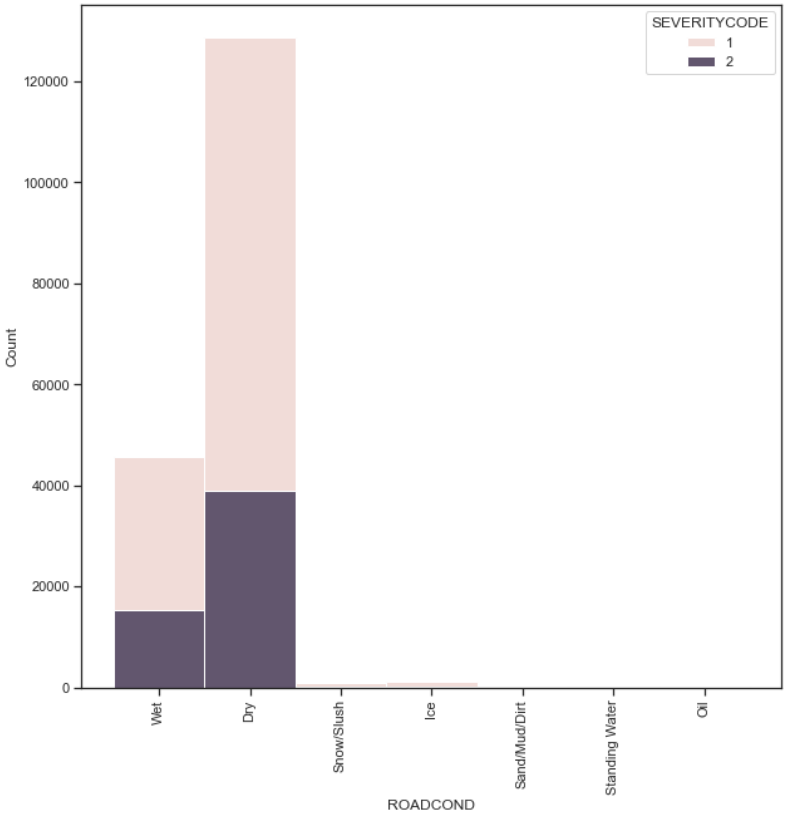


Diagram 3.4.1 Relation between road conditions and the type of severity (1 = property damage, 2 = injury)

3.5 Relation between the light conditions and accidents

Light conditions have an impact on the accidents and their severity as well. Situations like driving in the dark with no street lights may be dangerous and stressful for a driver.

Light conditions	Number of accidents	% of accidents
Daylight	119798	67.84
Dark - Street Lights On	46296	26.22
Dusk	5599	3.17
Dawn	2360	1.34
Dark - No Street Lights	1407	0.80
Dark - Street Lights Off	1129	0.64
Dark - Unknown Lighting	9	0.01

Table 3.5.1 Relation between the light condition and accidents

However, the vast majority of the accidents does not take place in the dark, or during the dusk or dawn, but in daylight (in 67.84% of the cases). Dark – with lights street on – is the variable with the second higher percentage of incidents (26.22%). Together they cover 94.06% of the accidents.

Road conditions	Severity (1 = property damage, 2 = injury)	Number of accidents
Dark - No Street Lights	1	1096
	2	311
Dark - Street Lights Off	1	821
	2	308
Dark - Street Lights On	1	32346
	2	13950
Dark - Unknown Lighting	1	6
	2	3
Dawn	1	1576
	2	784

Road conditions	Severity (1 = property damage, 2 = injury)	Number of accidents
Daylight	1	82461
	2	37337
Dusk	1	3739
	2	1860

Table 3.5.2 Relation between the light conditions and the severity of the accidents (1 = property damage, 2 = injury)

Like in other previous attributes, light condition always has more accidents with property damages than with injuries. The chart below (Diagram 3.5.1) shows how the major part of the accidents – both with property damages and injuries – occurs in daylight and dark (with street lights on), the proportions between them and the other conditions.

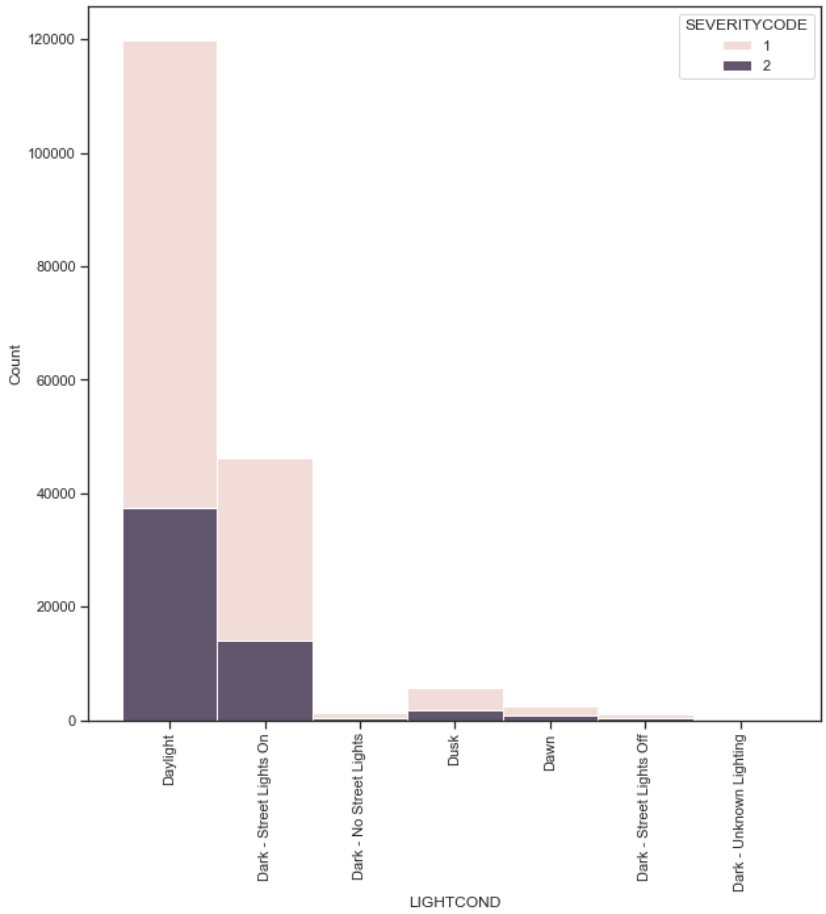


Diagram 3.5.1 Relation between light conditions and the type of severity (1 = property damage, 2 = injury)

3.6 Relation between accidents and people involved

The number of people involved in an incident is another attribute to take in consideration. Its mean per accident is 2.55 people, its mode is 2 and the 75% of the incidents include between 1 and 3 people. This means that even though the highest number of people in an accident is 81, the vast majority of the incidents involves very few people.

People involved	% of accidents
2	59.48
3	19.67
4	8.10
1	6.27
5	3.70
6	1.52
7	0.63
8	0.30
9	0.12
10	0.07

Table 3.6.1 Relation between accidents and people involved

The major part of the accidents involves 2 people (59.48%). Incidents involving 3 people have the second-highest percentage (19.67%). Together they cover 79.15% of the accidents and the percentage increase to 98,74% if we consider all the incidents that involve between 1 and 6 people. Over 10 people the percentage of the accidents becomes so small that is rounded to zero.

This is even more visible in the diagram below (3.6.1) that shows how accidents (both with property damages and injuries) involve a very small number of people for most of the times and the rest of the cases is mainly composed by outliers.

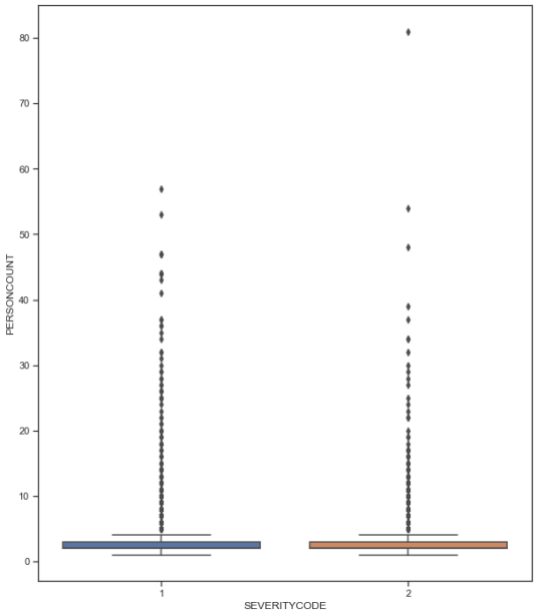


Diagram 3.6.1 Relation between people involved and the type of severity (1 = property damage, 2 = injury)

3.7 Relation between accidents and vehicles involved

The number of vehicles involved in incidents is the last attribute that I took in consideration. Its mean per accident is 1.97 vehicles, its mode is 2 and the 75% of the incidents include 1 or 2 vehicles. This means that even though the highest number of vehicles in an accident is 12, the vast majority of the incidents involves very few vehicles – as it is visible in the diagram below (diagram 3.7.1).

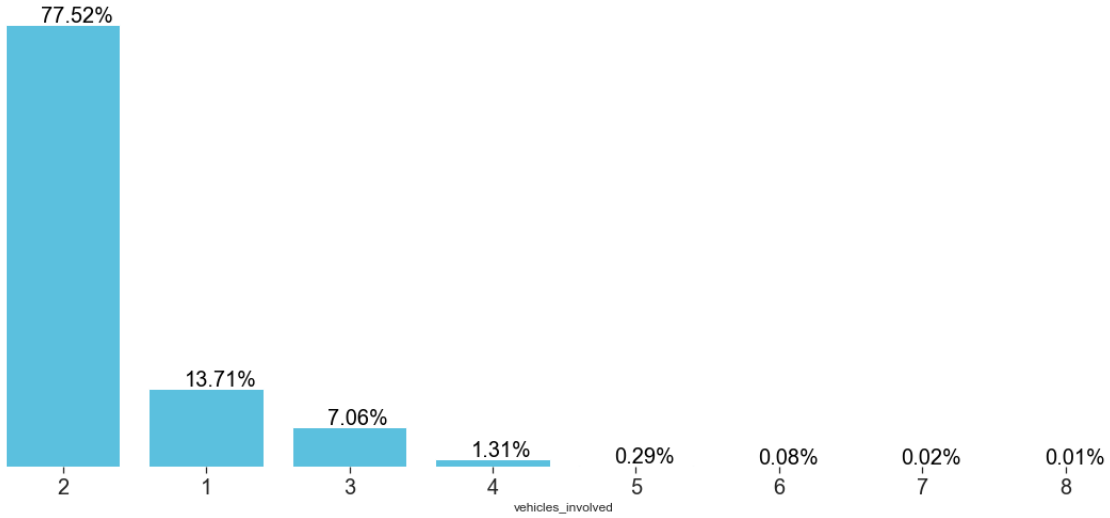


Diagram 3.7.1 Percentage of vehicles involved in accidents

The 91.23% of the accidents involve 1 or 2 vehicles and 98.29% involves between 1 and 3 of them. Over 8 vehicles the percentage of the accidents is so low that is rounded to zero.

Therefore, both the severities (i.e. property damages and injuries) are focus on the accidents that involve 2 vehicles – as shown in the diagram below (diagram 3.7.2) – and the rest of the cases is composed by outliers.

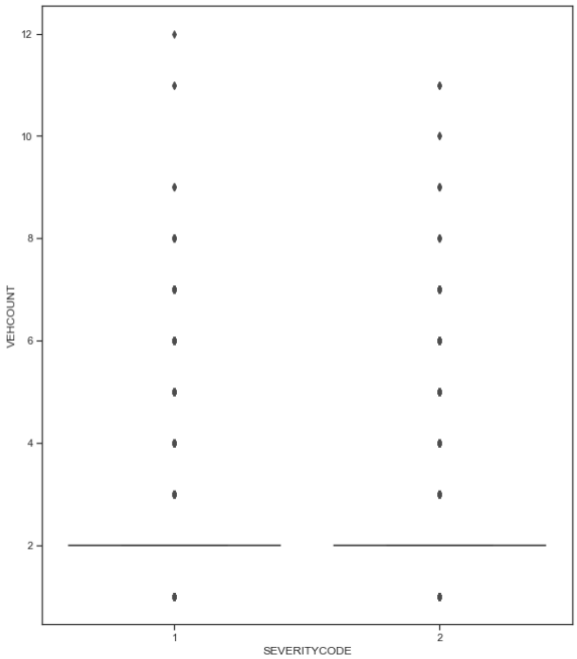


Diagram 3.7.2 Relation between vehicles involved and the type of severity (1 = property damage, 2 = injury)

4. Predictive Modelling (Methodology – part 2)

Since the target (i.e. dependent variable) is a categorical data, the logical model is classification so as to predict the severity of an accident.

The application of classification models follows a specific path. I divided the samples into two classes (80% training data, 20% test data, giving a random state of 4).

I then used three approaches to create three different predictive models:

1. Decision Tree
2. Support Vector Machine (SVM)
3. Logistic Regression.

Among the three models, logistic regression was the one that had the worst accuracy with all the measures that I used, as it is visible in Table 4.1.

Algorithm	Jaccard	F1-score	Accuracy	LogLoss
Decision Tree	0.7171	0.6953	0.7396	NA
SVM	0.7216	0.6828	0.7395	NA
Logistic Regression	0.6984	0.6640	0.7171	0.5775

Table 4.1 Performance of classification models - Accuracy (red best results)

Even though the decision tree has a lower Jaccard score than the one of SVM, its F1-score and accuracy (using metrics.accuracy_score) are slightly higher. This situation is well shown by their confusion matrixes below (Diagram 4.1, 4.2, 4.3).

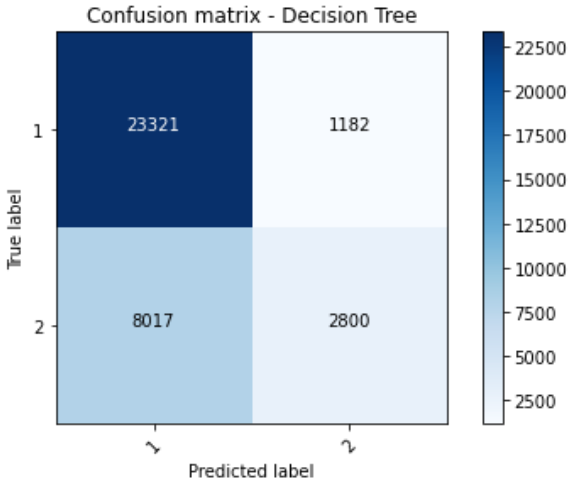


Diagram 4.1 Confusion matrix – Decision Tree (1 = property damages; 2 = injuries)

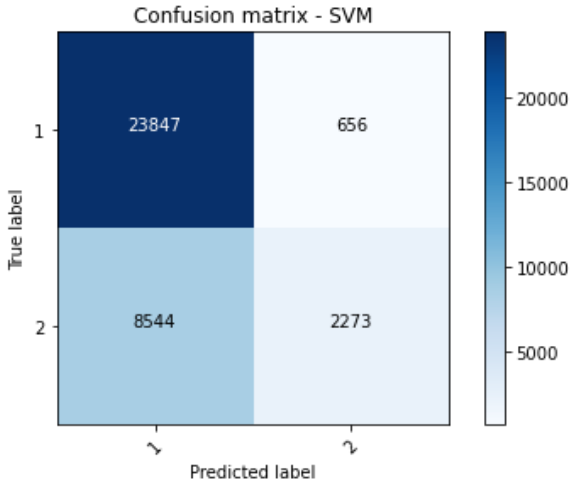


Diagram 4.2 Confusion matrix – SVM (1 = property damages; 2 = injuries)

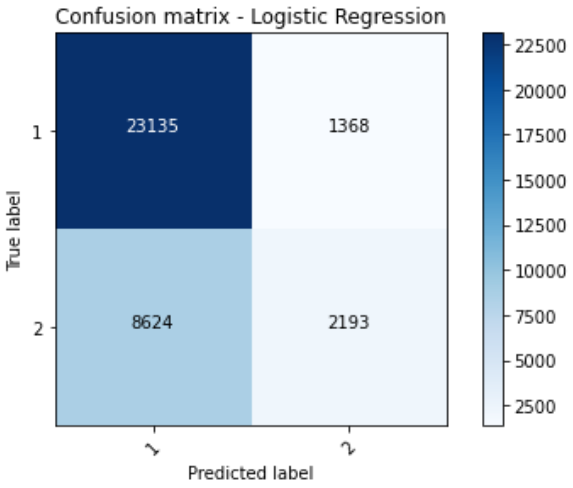


Diagram 4.3 Confusion matrix – Logistic Regression (1 = property damages; 2 = injuries)

The three confusion matrixes show how the logistic regression model has a higher mistake with both severity 1 (i.e. property damages) and severity 2 (i.e. injuries). However, the decision tree has the highest true negatives while the SVM has the highest true positives (see table below, Table 4.2).

Algorithm	True Positives	False Positives	False Negatives	True Negatives
Decision Tree	23,331	1,182	8,017	2,800
SVM	23,847	656	8,544	2,273
Logistic Regression	23,135	1,368	8,624	2,193

Table 4.2 Performance of classification models – True/False Positive/Negative (red best results)

In this situation, both the decision tree and the SVM are good models to use.

However, if we consider the objective of our project, we are more interested in saving life and avoiding injuries than avoiding property damages. Thus, predicting injuries (i.e. 2, true negative) is more important than property damages (i.e. 1, true positive).

Therefore, the Decision Tree is the best model to use for two reasons:

1. It has better F1-score and accuracy
2. It has the highest true negatives, thus it predicts accidents with injuries with greater accuracy

5. Results

In this research I discovered that accidents occur much more often in situations that are – at least in theory – safer, i.e. conditions with clear sky, dry road, and daylight.

The larger number of the accidents takes place with clear sky (66.09%), rain (18%) and overcast (15.01%) and together they are the 99.10% of all the incidents.

The vast majority of the incidents takes place when the roads are dry, in the 72.90% of the cases. Wet conditions of the road are the situation with the second higher percentage of incident (25.79%). These two variables together collect the 98.69% of the accidents while the other apparently more complicated conditions are correlated to only 1.31%.

The major part of the accidents takes place in daylight (in 67.84% of the cases). Dark (with lights street on) is the situation with the second highest percentage of incident (26.22%). Together they cover the 94.06% of the accidents.

Considering the conditions of weather, road and light, the amount of incidents with property damages is always larger than those ones with injuries.

Speeding is not a considerable attribute since it is responsible for only a small percentage of accidents. Thus, it was not possible to use it in the research study.

Whether we compare the different types of junctions, we discover that most of the accidents (47.16%) take place in mid-block crossing (not related to intersections). Accidents at intersections (related to the intersection) are the second most often incidents (33.47%) and happen more than the double of the time that the incidents in mid-block crossing related to intersections (12.40%). These three types of junction together are the places where more of the accidents happen (more than 90%).

If we take in consideration the type of severity that an incident can have, accidents with property damages are always more often than those ones with injuries – like in weather, light and road conditions.

We can also see that the larger number of the accidents that occur in mid-block crossings (not related to intersections) have the highest number of incidents with property damages (i.e. 65,293): a number higher than the sum of incidents with property damages in the other junctions. However, the accidents with injuries at intersections (related to the intersections) are higher (i.e. 25.769) than those ones in mid-block crossing (not related to intersections), i.e. 17,993 incidents. This makes the intersections the most dangerous junctions.

The major part of the accidents involves 2 people (59.48%). Incidents involving 3 people have the second-highest percentage (19.67%). Together they cover 79.15% of the accidents and the percentage increase to 98,74% if we consider all the incidents that involve between 1 and 6 people. Over 10 people the percentage becomes so small that is rounded to zero. Its mean per accident is 2.55 people, its mode is 2 and the 75% of the incidents include between 1 and 3 people. This means that even though the highest number of people involved in an accident is 81, the vast majority of the incidents involves few people. This is true for both of the severities (i.e. property damages and injuries).

The 91.23% of the accidents involve 1 or 2 vehicles and 98.29% involves between 1 and 3 vehicles. The percentage of incidents with over 8 vehicles is so small that is rounded to zero. The mean per accident is 1.97 vehicles, its mode is 2 and 75% of the accidents include 1 or 2 vehicles. This means that even though the highest number of vehicles involved in an accident

is 12, the larger number of incidents involves very few vehicles. This consideration is also valid for both of the severities (i.e. property damages and injuries).

Regarding the model to use for the predictions, both the Decision Tree (best F1-score and accuracy) and the Support Vector Machine (best Jaccard score) had good results.

However, considering the objective of the study, we are more interested in saving life and avoiding injuries than avoiding property damages. Thus, predicting injuries (i.e. 2) is more important than property damages (i.e. 1).

Therefore, the Decision Tree is the best model to use for two reasons:

1. It has better F1-score and accuracy
2. It has higher true negatives, thus it predicts accidents with injuries with greater accuracy.

6. Discussion

In this study I showed that accidents occur much more often in situations that are – at least in theory – safer, such as clear sky, dry road, and daylight.

It looks like drivers are more careful when situations are more stressful and they undervalue the risk of accidents when there are good conditions of the weather, the light and the road.

This is also consistent with the small number of people and vehicles usually involved in accidents.

7. Conclusion

Most of the people drive a vehicle for moving inside and outside cities everyday. It can be for commuting, going on holidays, visiting someone or something else.

Having a reminder showing that most of the accidents happen in the easiest conditions (such as clear sky or daylight) could be useful. It would keep the driver in the same state of alert that they have when the conditions are worst (such as snow).

The models developed could also be useful for a municipality that wants to decrease the number of accidents in their district.