Alves, Carlos

October 14, 2020

1. Introduction

1.1 Background

According to Canadas's Road Safety Strategy 2025[1], Each year, around 2000 people are killed and 165,000 are injured in road vehicle accidents in Canada. The current annual cost of the road transportation system in the country is around $37 billion Canadian dollars, which is represents roughly, 2,2 percent of its GDP. Even with all efforts from government, lower number of deaths in comparison with other countries and new programs showing promising expectations, there is still much to be done regarding road safety.

1.2 Problem and Interest

In this study, a data set from Transport Canada[2] will be used to verify if there are that conditions may cause higher number of accidents with injuries and victims, taking the severity weather and several other attributes into consideration. With this, it’s expected to show those conditions, and to create a prediction method to enable the government to prevent accidents in dangerous conditions. This way, there could be a bigger focus on prevention, helping to minimize the number of total collisions, and their severity.

2. Data acquisition

2.1 Data sources

The data set used for this project was found in Kaggle, **here**[3]. The main source of this data is Transport Canada website. The data set presents values from all around Canada, from 1999 to 2014.

2.2 Data cleaning

The raw data had several missing and unknown values, which had to be changed. Depending on the amount of data missing, two techniques were chosen to approach it. Changing the value to the series average value, and dropping the rows, when it wasn’t possible to fill them.

On figure 1, it was possible to verify that there is a big data imbalance, that must be fixed to allow the predictions to show real values, taking into consideration both severity types of accidents (injury and fatal).

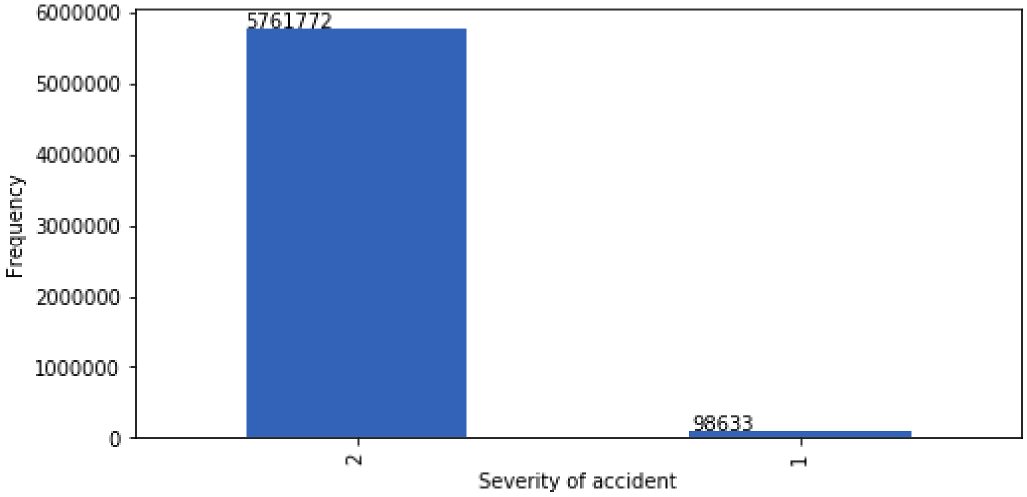


Figure 1: Imbalanced data

Therefore, it was used a down-sampling technique, balancing the data so the prediction values are balanced. Figure 2 shows the result of the down-sampling, maintaining the data structure proportional to the initial one.

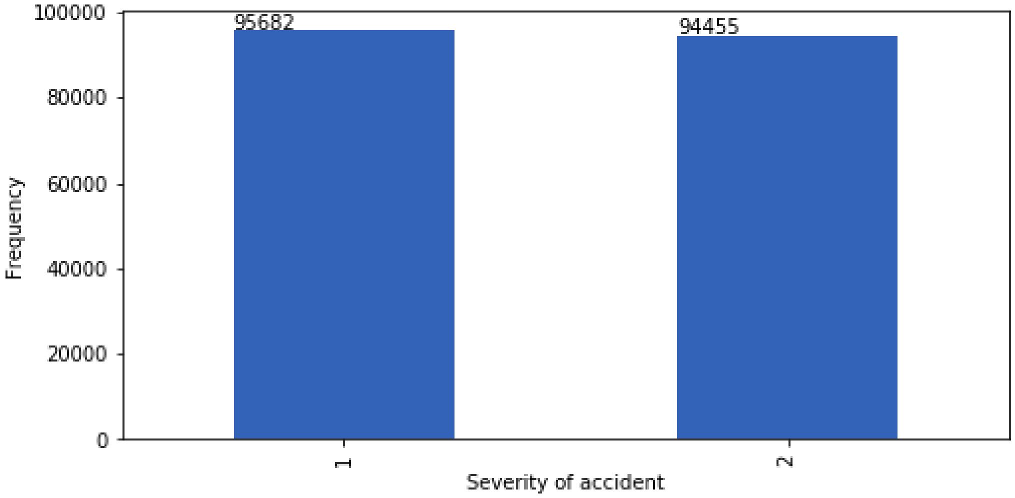
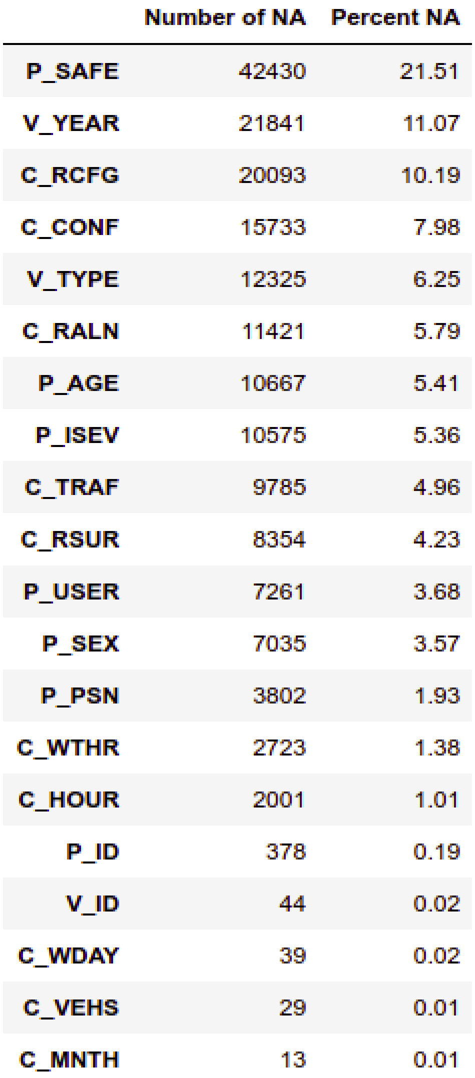


Figure 2: Balanced data

After Balanced, the data was divided to verify the number of unknown values and Figure 3 was the result.



	Number of NA	Percent NA
P_SAFE	42430	21.51
V_YEAR	21841	11.07
C_RCFG	20093	10.19
C_CONF	15733	7.98
V_TYPE	12325	6.25
C_RALN	11421	5.79
P_AGE	10667	5.41
P_ISEV	10575	5.36
C_TRAF	9785	4.96
C_RSUR	8354	4.23
P_USER	7261	3.68
P_SEX	7035	3.57
P_PSN	3802	1.93
C_WTHR	2723	1.38
C_HOUR	2001	1.01
P_ID	378	0.19
V_ID	44	0.02
C_WDAY	39	0.02
C_VEHS	29	0.01
C_MNTH	13	0.01

Figure 3: NA values table

For values that represented a considerable amount of the data, the means technique was used to complete the NA values. For the lower portion, the rows were dropped.

2.3 Feature Selection

After data cleaning, there were 190137 samples and 22 features in the data.
The attributes legend can be found here[4], showing the meaning of all attributes used in this project, and the possible values.

3. Exploratory Data Analysis

3.1 Calculation of target variable

The target of this project, is to be able to calculate the severity of the accident (C_SEV), taking into consideration the attributes shown in the section above, to predict possible accident conditions, so that those conditions can be avoided or mitigated.

3.2 Relationships found in data attributes

Some important relationships were found using visualization methods, such as relations between weather and accidents, hours of the day, months and others that will be shown below. Using matplotlib, those relations will be shown by charts.

3.3 Month and severe accidents

Creating a bar chart that relates the months with the number of severe and fatal accidents, we can verify that August and July are the months with the most severe accidents, and this can be linked to the Summer, and Summer vacations. December comes in third place, and the holidays might be a cause of the increase in relation to other months.

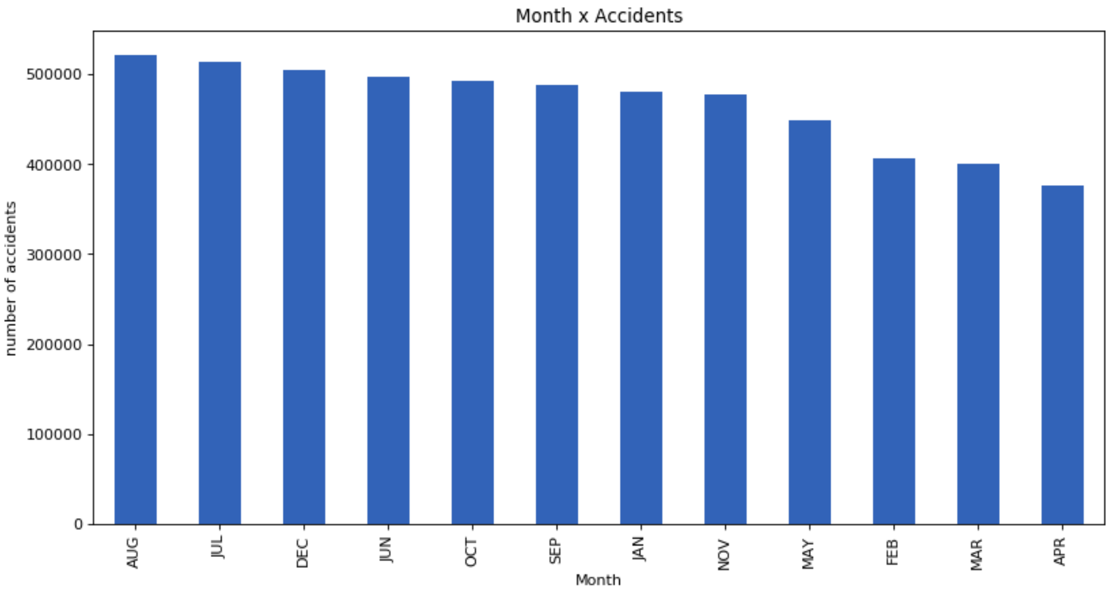


Figure 4: Months and accidents

3.4 Week day and severe accidents

Using the same method as above, we can relate the week days to numbers of accidents, and verify that the weekends present higher values of severe accidents, specially Fridays, as shown in figure 5.

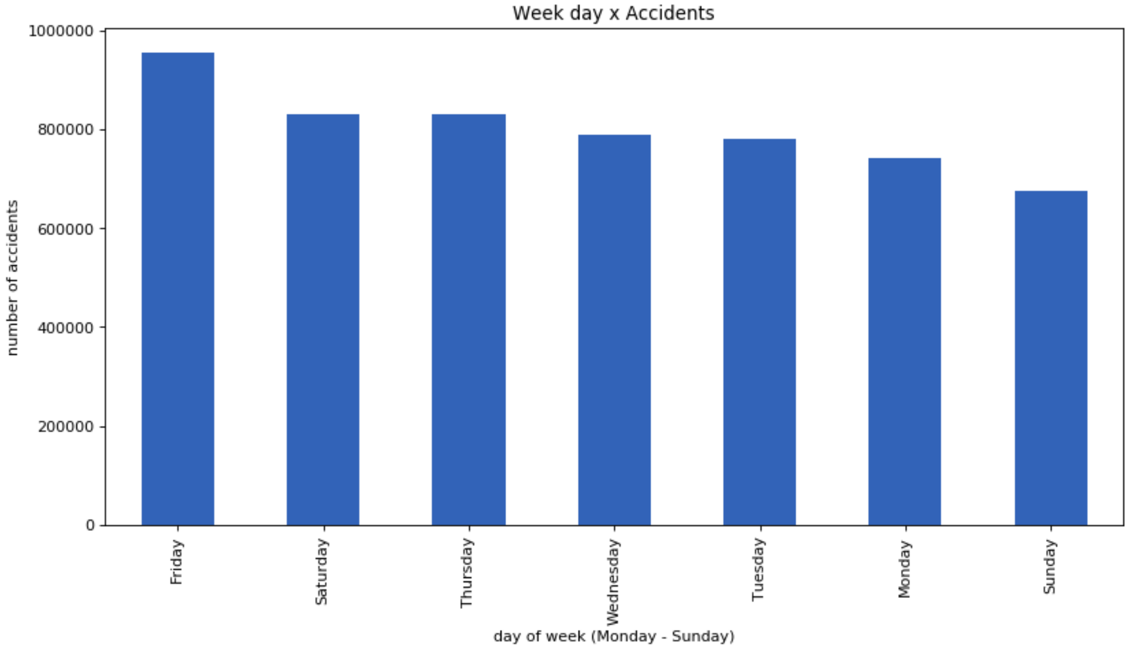


Figure 5: Days of week and accidents

3.5 Hours of the day and severe accidents

The chart that relates the hours of the day with accidents might change the paradigm of sever accidents happening at night, when it's more common to see speeding cars occurrence and other issues such as driving under the influence of alcohol and other drugs. It's possible to verify that most accidents occurs between 15:00 and 17:00.

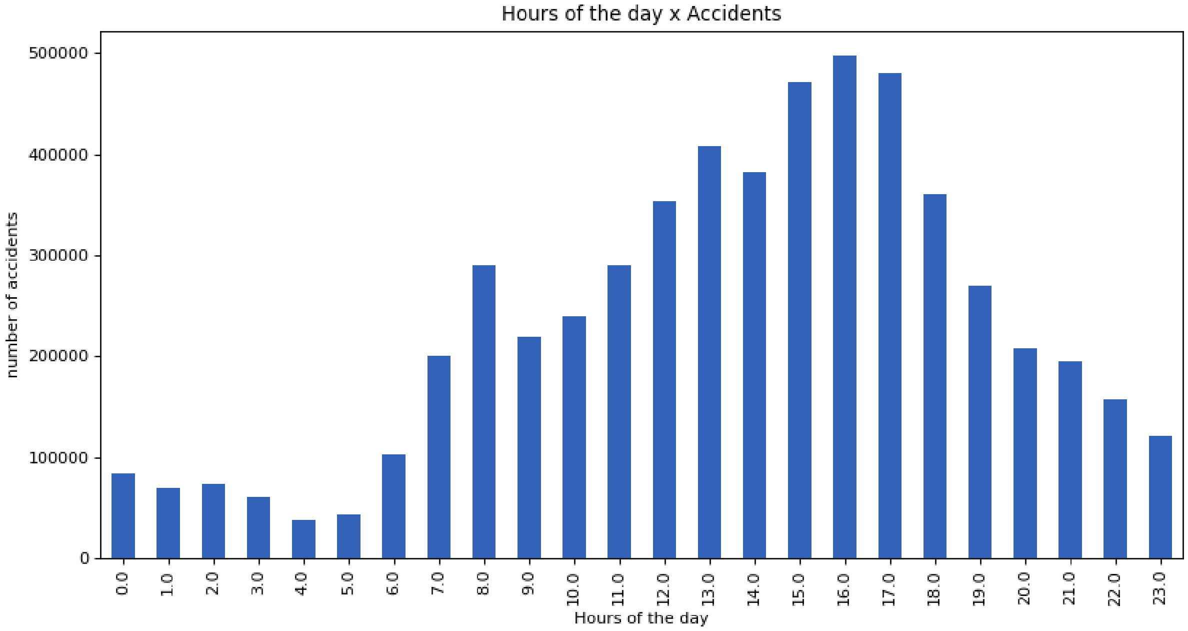


Figure 6: Hours of day and Accidents

3.6 Weather conditions and severe accidents

Corroborating with the hypothesis of the summer, and going a little against the main sense of more accidents in the winter, and snow days, we see that severe and fatal accidents actually occurs in a much larger scale in sunny days. That most be given since in snowy and cold days, people drive slower, and even if there are more accidents, they are with no injuries, and small collisions.

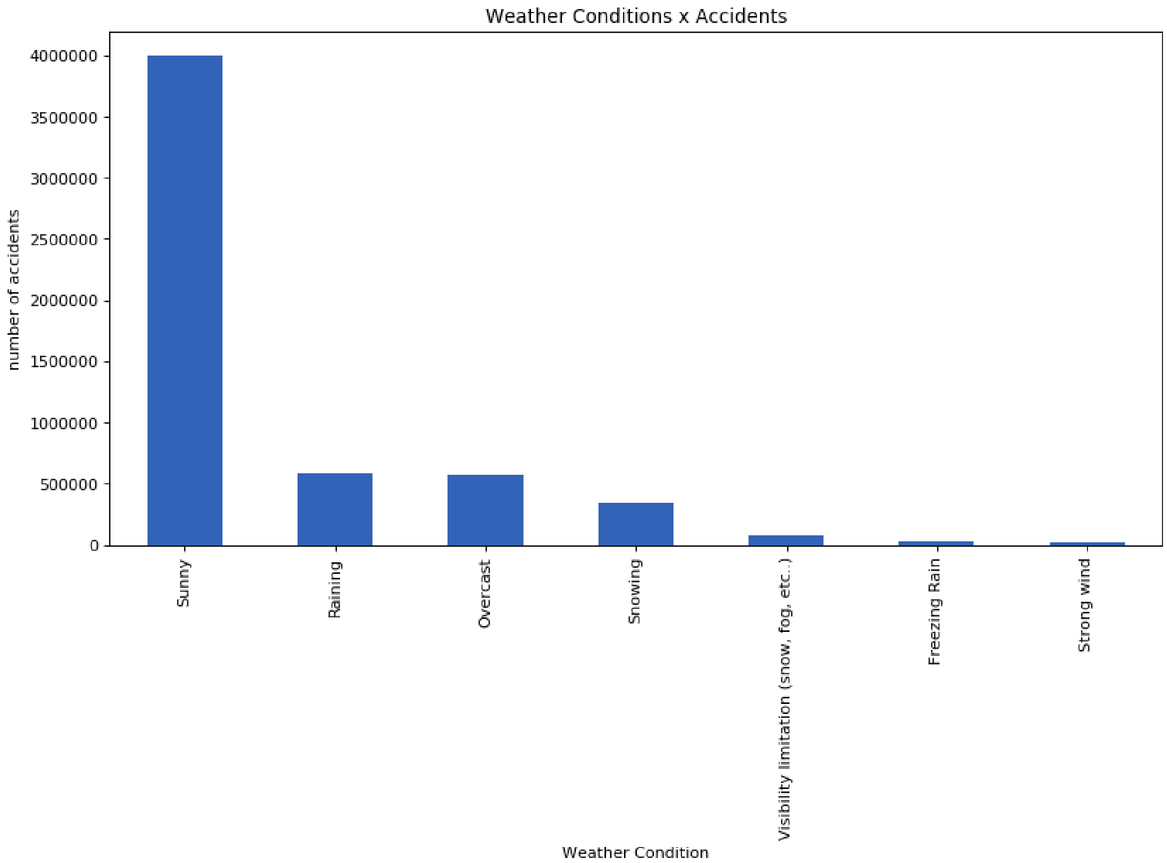


Figure 7: weather condition and Accidents

4. Predictive Modeling

4.1 Algorithms used

The project is a classification problem, therefore, the models used were the K-Nearest-Neighbor, Logistic Regression and Decision Tree. Each one will be presented with their results, and the highest accuracy model will be the chosen one.

4.2 K-Nearest-Neighbors

The first step to using KNN method was normalizing the data. That was done using skit-learn library, creating the training set with 80% of the data, and the rest was used for the testing set.

Initially, the k value used was four, but to improve the test accuracy of 77% found, a range of ks = 12 was used. And re-evaluating the accuracy, it was found that with a k of 12, the value of 78.95% of accuracy was found. The pot below shows the accuracy found in the ks= 12 range.

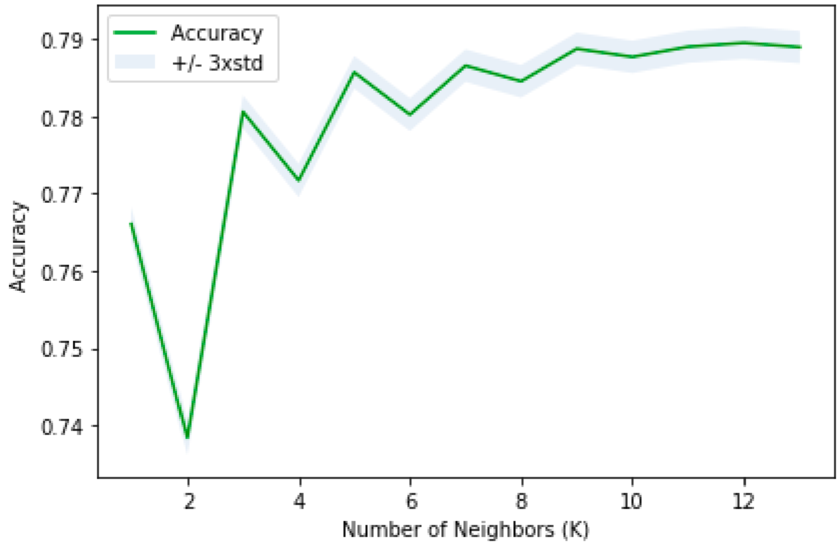


Figure 8: KNN accuracy

4.3 Logistic Regression

On Logistic Regression, C = 0.01 was used, and the solver was ‘liblinear’. With that it was possible to train and test the sets. Using the jaccard index as a tool, it was found a Jaccard similarity score of 73.51%. A Confusion Matrix was generated to better represent the results, such as shown in Figure 9.

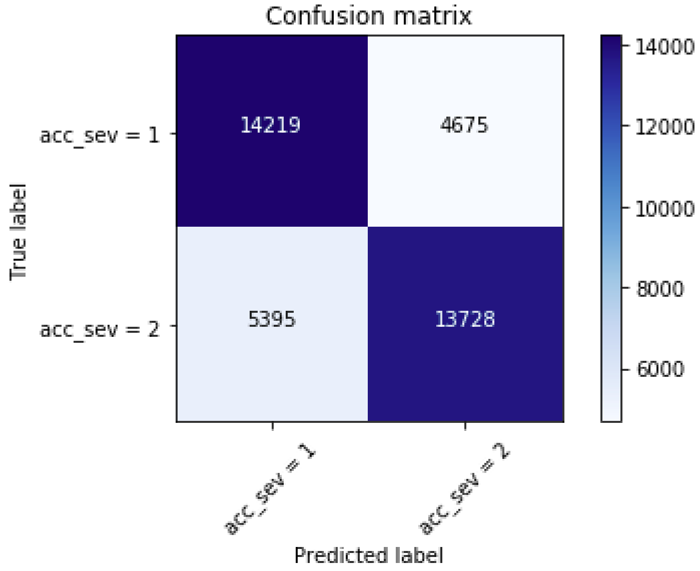


Figure 9: Confusion matrix

4.4 Log Loss

In logistic regression, the output can be the probability of the accident severity to be 1 or 2. This probability is a value between 0 and 1. Log loss(Logarithmic loss) measures the performance of a classifier where the predicted output is a probability value between 0 and 1. the value found was 0.5288 .

4.5 Decision Trees

For the Decision Trees, it was used the “DecisionTreeClassifier”, with the entropy criterion. The test and train set were created and fitted, and the accuracy found was of 76.74 %.

4.5 Solution

It was decided to use the model acquired by the K-Nearest-Neighbor method, with k = 14, since this method shown the highest accuracy among all others. This way, it’s possible to predict events with almost 80% of accuracy, and from there, develop prevention campaigns all around Canada.

5 Conclusion

In this study, the accident with severe and fatal victims in Canada and its relations to several factors were created. It was possible to create a model with a high accuracy to predict those events, and through data visualization, it was possible to verify several aspects that impact the number of sever or fatal accidents. With that information, it would be possible to create campaigns with focus on specific months, days of week, hours and weather conditions to prevent and reduce number of accidents.

6 Future Directions

The data-set have more possibilities available to analyze, such as age of conductor, type of car, and many more, making it possible to create more data visualization that would allow better comprehension. It’s also possible, to try and acquire location information from the accidents, to generate models by location, which can improve the efficacy of prevention campaigns and a better prediction taking location into consideration.