Car accident severity in Canada 1999–2017

Ralph Keusch

August 26, 2020

Abstract

We analyze the Canadian database of car collsions from 1999 to 2017. Then, we study whether machine learning algorithms are able to predict from date, time, and weather conditions whether a given collision is fatal or not. It turns out that in general, the given data is not sufficiently significant, but decision trees and logistic regression are able to make reasonable 0/1-predictions on the severity of car accidents.

1 Introduction

On every day thousands of people die from traffic accidents. Often, young people lose their fruitful lives at dramatic car crashes. The costs of fatalities and injuries due to accidents on society are high, therefore governments and the industry invest millions of dollars into the safety of cars, roads, and infrastructure. However, as the world population rises and rises, roads stay congested, and there are still too many unnecessary car accidents and death cases therefrom.

Nowadays, rich data is available, and modern machine learning tools can be applied to analyze these data sources. In order to make intelligent decisions, governments need to know where money is best invested. Thus, people started to determine main factors causing severe traffic accidents and revealed patterns and clusters. This includes weather conditions, road conditions, time, driver's age, and many additional factors. Furthermore, to reduce the number of severe car crashes, it is essential to have precise models that are capable to make accurate predictions. A specific forecast model helps decision-makers to decide between different safety invests as it anticipates by how much an action may decrease the number of accidents. It could also be used at extreme weather situations to alert drivers.

In recent years, having more computaational power and more elaborated machine learning methods at hand, a lot of research has been done in this field. For instance, Abdelwahab et al. used neural networks to predict severities [1], Chong et al. applied several different machine learning methods to the problem [4], or AlMamlook et al. compared different machine learning algorithms to predict accident severities [2].

In this report, we consider the car accident severity in canada [3]. We have two main goals: first, we anaylze the data set, understand the relationships, and illustrate how the number of car collision in general and fatal collisions in particular depend on factors such as time, weather condition, or

roadway surface. Second, we apply different standard machine learning methods to our data and compare their accuracy in predicting the severity of car crashes. The main technical problem in our analysis was the size and the imbalancedness of the data source: only a 0.02-fraction of the records belong to severe collisions. We overcame this issue by sampling data sets with a large bias towards fatal cases.

The report is organized as follows. In the next section, we describe our data source, the data cleaning and preparing steps, and explore the data. In Section 3 contains the methodology and in particular the different machine learning algorithms that we apply. In Section 4 we summarize our results, which we discuss afterwards in Section 5. Finally, we conclude the report with Section 6.

2 Database

2.1 Data source and selection

As data source for this report we use the national collision database (NCDB) of Canada. It contains all police-reported motor vehicle collisions on public roads. Since Canada is a bilingual country, the data values are all numerical. The rich database contains a total of almost 7 million records. Each row corresponds to one person being involved in a car collision. Hence a collision between two cars A and B, where A has one passenger and B has two passengers, would create three records in the database.

Its 23 rows that are organized in three categories: data elements on collision level, on vehicle level, and on personal level. For the sake of this report, we drop the vehicle and personal level from the data, even though it would be possible to include them as additional features. On collision level, the database consists of the following 12 rows: "Year", "Month", "Day of week", "Collision hour", "Collision severity", "Number of vehicles involved in collision", "Collision configuration", "Roadway configuration", "Weather condition", "Road surface", "Road alignment", "Traffic control", and "Collision case". Obviously, "Collision severity" is a very significant dependent variable. We observe that the data source only distinguishes between car collision with and without death cases.

	C_YEAR	C_MNTH	C_WDAY	C_HOUR	C_SEV	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN	C_TRAF	C_CASE
0	1999	1	1	20	2	02	34	UU	1	5	3	03	752
1	1999	1	1	20	2	02	34	UU	1	5	3	03	752
2	1999	1	1	20	2	02	34	UU	1	5	3	03	752
3	1999	1	1	08	2	01	01	UU	5	3	6	18	753
4	1999	1	1	08	2	01	01	UU	5	3	6	18	753
5	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
6	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
7	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
8	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
9	1999	1	1	15	2	01	04	UU	1	5	U	UU	932

Figure 1: First rows of the data set after dropping several columns.

2.2 Data cleansing, first step

We started with 6.772.563 rows, each value stored as string. Our data preparing should ensure numeric analysis, hence we converted the columns "Year", "Month", "Day of week", "Collision hour", "Number of vehicles involved in collision" to numeric values. For each column, there are values for *unknown* and for *not provided by jurisdiction*. In a first cleaning step, we removed all rows that have such values in a numeric row. This left us still with 6.705.062 records, thus we didn't lose many records. Next, we observed that some values are not stored consistently. For example, the month January is stored either as 1 or 01. We converted the numeric rows to integers and thereby got rid of this inconsistency.

There were still too many categorial variables, for example 21 different collision configurations. We dropped the columns "Collision configuration", "Road alignment", and "Traffic control" to reduce the size of the data, and gave the remaining columns intuitive names. There were 15 different road configuration categories which we simplified by a binary variable, indicating whether the accident occured at an intersection or not. Next, to improve the readability we applied a dictionary to encode the values for weather conditions and road surfaces. Thereby, we simplified the road surface column: there were 12 different categories before, which we compressed to *normal*, *wet*, *snow*, *other*.

	Year	Month	Weekday	Hour	Fatal	Nbr Vehicles	Intersection	Weather	Road Surface
0	1999	1	1	20	0	2	0	Clear	snow
1	1999	1	1	20	0	2	0	Clear	snow
2	1999	1	1	20	0	2	0	Clear	snow
3	1999	1	1	8	0	1	0	Hail	snow
4	1999	1	1	8	0	1	0	Hail	snow
5	1999	1	1	17	0	3	0	Clear	wet
6	1999	1	1	17	0	3	0	Clear	wet
7	1999	1	1	17	0	3	0	Clear	wet
8	1999	1	1	17	0	3	0	Clear	wet
9	1999	1	1	15	0	1	0	Clear	snow

Figure 2: First rows of the data set after the cleaning steps.

2.3 Data exploration

Our key dependent variable is the crash severity. We observed that there are 111.302 fatal cases and 6.593.760 cases without any fatality. Hence, with respect to crash severity the data set is very *unbalanced*.

First, we looked how the total number of cases evolved during time. As the portion of collisions with fatality is rather small, we created a separate plot for the fatal cases. In general, the numbers are decreasing over the years, as indicated by the linear regression line. We also created plots of the total number of collisions against month, against weekday, and against daily hour.

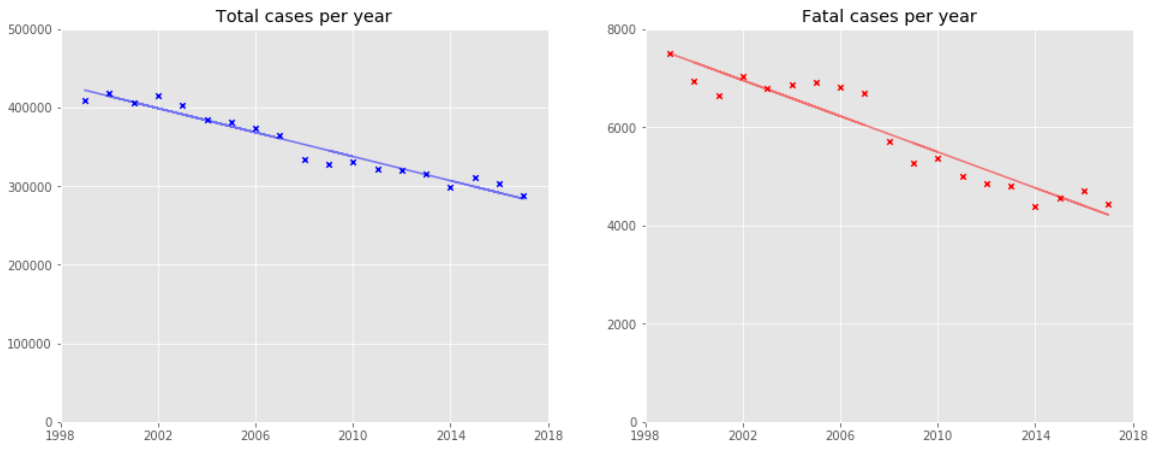

Figure 3: Total cases and total fatal cases per year. The drop around 2008 may be caused by the most recent amendments by the Canadian parliament to the law on drinking and driving.

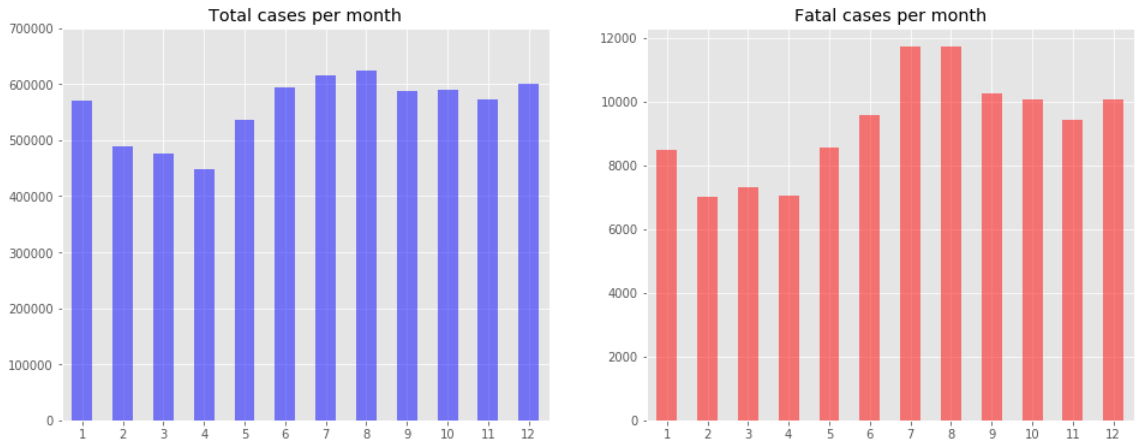

Figure 4: Total cases and total fatal cases per month. We observe that in July and August, the number of fatal collisions is over-average.

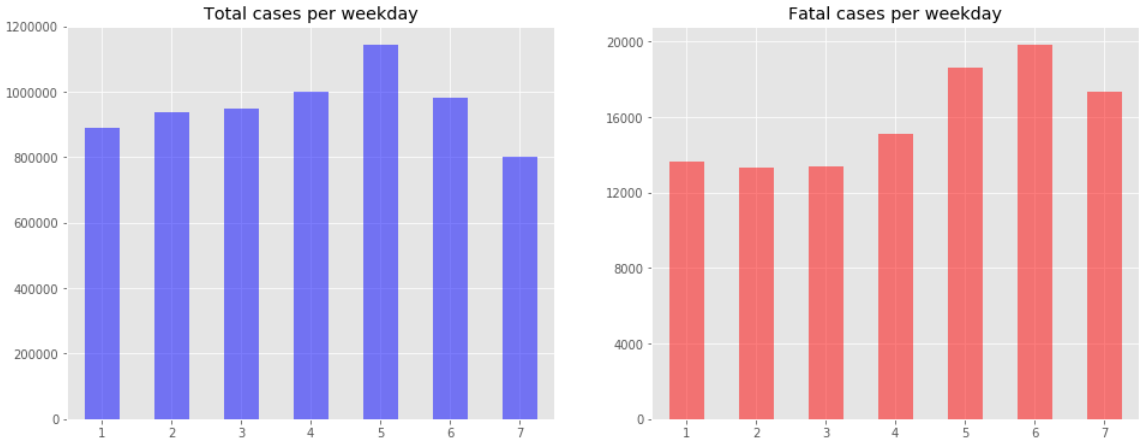

Figure 5: Total cases and total fatal cases per weekday. We see that most collisions occur on Fridays and the least on Sundays. However, the most fatal collisions happen on Fridays and Saturdays.

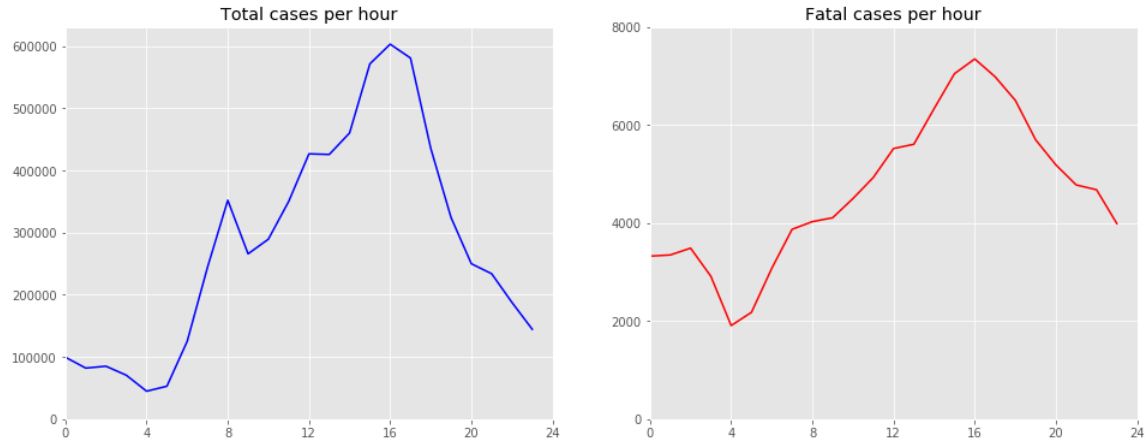

Figure 6: Total cases and total fatal cases per hour. Clearly, there is a peak around 4 PM. At night, the fraction of fatal cases is higher than at day.

Furthermore, we investigated the proportions of weather conditions and road surfaces. Here, it turned out that for the fatal cases, the proportions were rather similar, so we only illustrate the numbers for the overall collisions. The majority of collisions happens at clear weather, on normal road conditions. Minor but still significant weather conditions are "cloudy", "raining", and "snowing". For the sake of overview, we grouped all other weather conditions ("hail", "fog", "storm", "other", "unknown") to a single category.

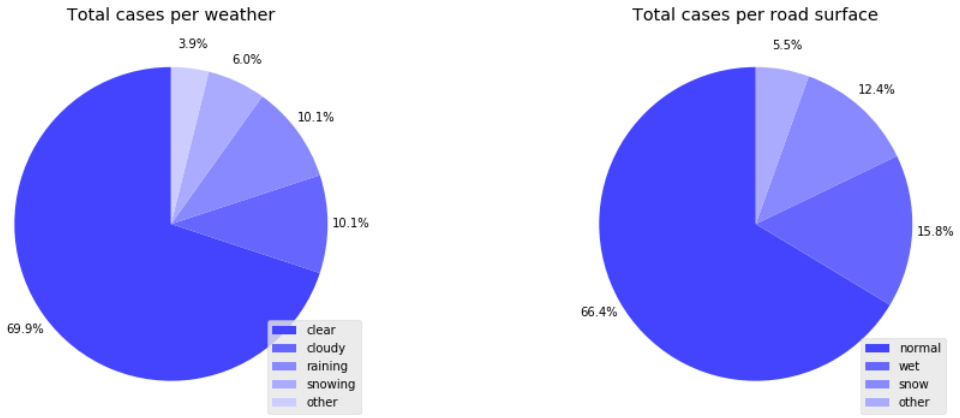

Figure 7: The total number of collision cases per weather category and per road surface category.

2.4 Data cleansing, second step

As discussed above, our data set is highly imbalanced. In fact, only 1.69% of the records are marked as fatal severity. Directly using this data would be highly problematic: a trivial algorithm could always predict "not fatal", achieving a brillant Jaccard-score of 0.9831. There are different ways to handle unbalanced datasets: oversampling fatal cases, undersampling cases without fatality, using modified evaluation scores, or using weighted cost functions in the algorithms. Since the data set is rather large and we suffered from performance and memory issues, we decided to undersample cases without fatality.

More precisely, we sampled an 0.1-fraction of the fatal cases and from the non-fatal records a random subset of fraction 0.002. This reduced the dataset to a total of 24.318 records, 11.130 being fatal cases, leaving us with a rather balanced dataset. Finally, we replaced the two columns for the road surface and the weather by dummy 0-1-columns. This variant is more suitable when training machine learning algorithms.

	Year	Month	Weekday	Hour	Fatal	Nbr Vehicles	Intersection	Clear	Cloudy	Fog	Hail	Raining	Snowing	Stormy	Unknown	normal	other	snow	wet
2610316	2005	6	2	17	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
4346196	2010	4	4	9	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0
624422	2000	7	4	7	0	3	1	1	0	0	0	0	0	0	0	1	0	0	0
3668039	2008	3	5	10	0	3	1	0	1	0	0	0	0	0	0	1	0	0	0
2062666	2003	12	5	12	0	3	1	1	0	0	0	0	0	0	0	1	0	0	0
6073014	2015	9	2	16	0	2	1	1	0	0	0	0	0	0	0	1	0	0	0
5521204	2013	11	4	10	0	2	1	1	0	0	0	0	0	0	0	1	0	0	0
3057329	2006	7	7	2	0	2	0	1	0	0	0	0	0	0	0	1	0	0	0
4630273	2011	2	2	7	0	11	0	0	0	0	0	0	1	0	0	0	0	1	0
826725	2000	12	6	16	0	2	1	1	0	0	0	0	0	0	0	1	0	0	0

Figure 8: First rows of the reduced data set with dummy variables for the weather and road surface categories.

3 Methodology

We have seen in Section 2 that the total number of cases is not depending linearly on features such as "hour". Using polynomial features in general would make the data set far too large, therefore we just introduced a small number of additional, higher order features: h^2, h^3, d^2, d^3, m^2, m^3, where $h =$"hour", $d =$"weekday", $m =$"month".

We split the data randomly into a train set (60%), a validation set (20%) for tuning the parameters of the methods, and a test set(20%) for testing the performance of our methods and comparing their accuracy. The severity is our only dependent variable. For the independent variables we applied standard feature normalization. We then applied *logistic regression*, the *k-nearest-neighbor algorithm*, *decision trees*, and *support vector machines* to predict the severity of a collision. In addition, we used logistic regression to predict the *probability* of a collision being fatal. We used the validation set to tune the corresponding model parameters, and finally compared the performance of the algorithms on the test set.

4 Results

4.1 Summary

The performance of the four different methods was not overwhelming. It turned out that the given data (date and time, weather, road surface, number of involved vehicles, and whether the collision happend at an intersection) is not sufficiently significant to make reliable predictions. While the algorithms mostly recognized non-fatal collisions as such, their performance on fatal cases was rather dissatisfying. Concretely, support vector machines and the k-nearest-neighbor-method made a wrong guess on more than half of the fatal cases. In contrast, logistic regression and decision trees reconized at least the majority of fatal cases as such. Decisions trees were not only the fastest algorithms but also achieved the best accuracy scores. Therefore, we believe that for this type of collision data, decision trees are the most suitable machine learning method to

forecast the severity of car accidents, based on weak data. The accuracy of logistic regression was also acceptable, it achieved a log-loss score of 0.64.

	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.618421	0.552555	NA
1	Decision Tree	0.644326	0.577843	NA
2	SVM	0.634868	0.545315	NA
3	Logistic Regression	0.625411	0.557982	0.642561

Figure 9: Jaccard-score and F1-score of the four classification methods, indicating their accuracy.

4.2 Performance of the methods in more detail

In the following, we illustrate the accuracy of the methods by considering their confusion matrices.

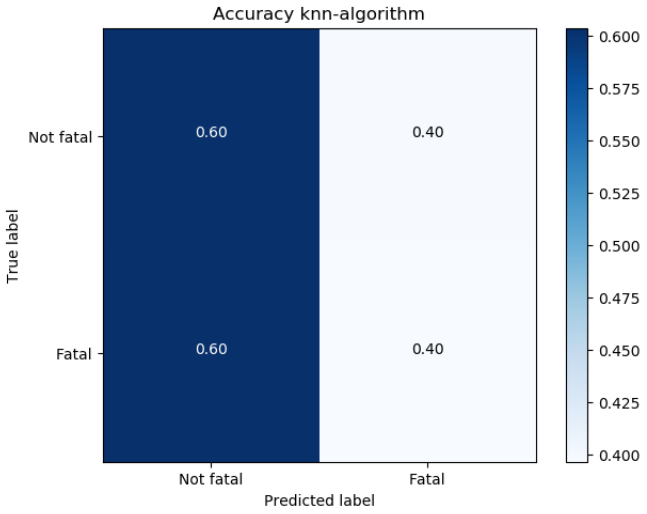

Figure 10: Accuracy of the knn-algorithm with $k = 9$.

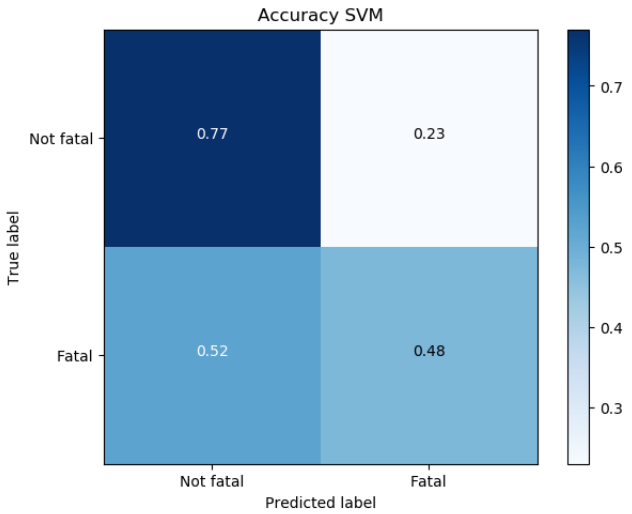

Figure 11: Accuracy of the SVM method with $C = 1$ and rbf-kernels.

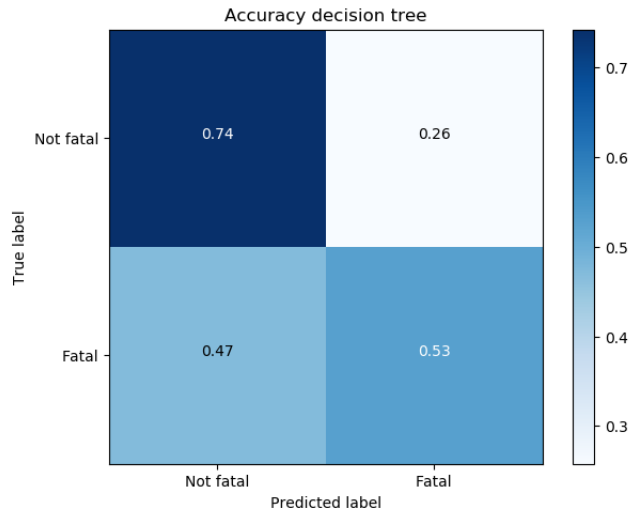

Figure 12: Accuracy of decision trees of depth 9.

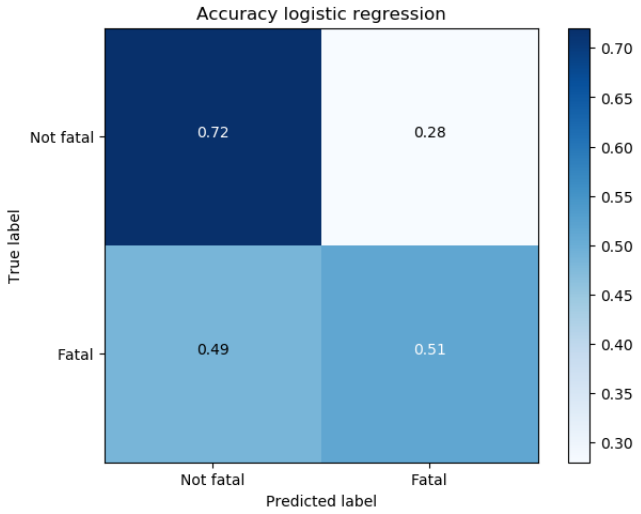

Figure 13: Accuracy of logistic regression with $C = 1$

5 Discussion

For all methods, we tried to tune the parameters, but mostly changing the parameters didn't change the accuracy by much. Only for decision trees, we observed a peak around depth 9.

As briefly discussed above, the performance of the methods was rather dissatisfying. In particular, the knn-algorithm was not at all able to distinguish between fatal and non-fatal cases. Most likely, the reason for this behaviour is that for each combination of weather and time, there are both fatal and non-fatal cases, leading to very close data points, preventing the neighborhood-based mechanism to make any reasonable predictions. Vice-versa, in the training phase decision trees and logistic regression were able to detect differences between fatal and non-fatal cases, and therefore recognized the majority of fatal cases in the test set.

Since decision trees was by far the fastest method, we tried to improve its performance by using general polynomial features of degree 2. Unfortunately, this led to overfitting and henceforth didn't improve the performance of the algorithm.

As seen in Section 2, categories such as time or weather heavily influence the frequency of car accidents. But we believe that these categories do not significantly influence the *severity* of the accidents. It seems that the entropy in the present data is too large to learn accurate rules. We thus believe that in order to learn the distribution of accident severity more relevant features are necessary. This certainly includes the speed of the cars, whether a driver was drunken or consumed other substances, the geographic location, or the violation of other traffic regulations. However, this information was not available from our data source.

Nevertheless, we could observe that the performance of our four methods was different and that decision trees were surprisingly fast compared to the other algorithms while at the same time achieving the best accuracy. This may be caused by the data structure as we didn't have much

numerical data available and most features were categorial. Therefore, we believe that for a source providing more relevant and more significant data sets, decision trees would be the suitable machine learning method to analyze it.

6 Conclusion

We used four different machine learning methods to predict the severity of car accidents in Canada, given only weak data such as weather, time, or weekday. We thereby faced the problem of a very imbalanced data sourced which we solved by undersampling. It turned out that decision trees performed the best, but still their accuracy was rather low. We arrived at the conclusion that essentially the available data was not sufficiently significant.

However, it would be worth to study the same data by additional methods such as neural networks. Another option would be to select other features, for instance the size of involved vehicles or the age of drivers. The data source includes as well a column for collision types (e.g., single car, two vehicles in same direction, two cars in opposite direction, ...), but it is questionable whether this column is a dependent or an independent variable.

References

[1] Hassan Abdelwahab and Mohamed Abdel-Aty. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record*, 1746:6–13, 01 2001.

[2] Rabia Emhamed AlMamlook, Keneth Morgan Kwayu, Maha Reda Alkasisbeh, and Abdulbaset Ali Frefer. Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 272–276, 2019.

[3] Transport Canada. National collision database. https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a, 2017.

[4] Miao Chong, Ajith Abraham, and Marcin Paprzycki. Traffic accident data mining using machine learning paradigms. *Informatica*, 01 2005.