Applied Data Science Capstone

REPORT

Machine Learning as a tool to predict the severity of an accident: Case Study from Seattle (WA) 2004-2020 record

Dr. Daniel Rodelli

22th November, 2020

SUMMARY

1 INTRODUCTION 3
2 BUSINESS PLAN 3
3 DATAFRAME EXPLORATION 6
3.1 Severity Code 6
3.2 Time Series 6
3.3 Injuries and Vehicles 7
4 DATAFRME PREPARATION 8
4.1 Dealing with NaN values 8
4.2Data Cleaning 8
4. 4 Feature Selection 10
5 MODELING 11
5.1 K-Nearest Neighbors 11
5.2 Logistic Regression 12
5.3 Decision Tree 13
5.4 Random Forest 14
5.5 Models Comparison 15
6 DISCUSSION 16
6 CONCLUSIONS 16

1 INTRODUCTION

According to the statistics by WHO (World Health Organization) (Feb, 2020):

Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years. Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured.

This, therefore, needs serious attention, as it concerns human lives which is irreplaceable. It is possible, thanks to machine learning, to predict the severity of car accidents as a result of the complex interplay of multitudes of factors like weather, road condition, light condition, speeding etc. and also to identify which factors are more important. The information thus gathered can be used to take preventive measurements.

According to the National Safety Council, traffic collisions cause more than 40,000 deaths and injure thousands of people every year across the United States. These are not traffic accidents, but entirely preventable tragedies.

In order to reduce accidents, we need to predict it based on the external parameters.

Since the accident occurs due to very many factors (unsafe road infrastructure, light condition, vulnerable road users, speeding, driving under the influence of alcohol and other psychoactive substances, distracted driving, weather) prediction of accidental severity is a challenge. Machine Learning is ideally suited here as this is a scientific approach for modelling and predicting the parameter of interest demanding only a low budget.

The current project attempts to apply a machine learning technique to predict the severity of the accident given the parameters as stated before using car collision data for the city of Seattle, USA.

The DataBase used in this project is from the Seattle Department Of Transportation (SDOT) and can be downloaded here: https://www.kaggle.com/jonleon/seattle-sdot-collisions-data

2 BUSINESS PLAN

The ability of emergency service to manage medical assets during an accident (i.e. ambulances, personnel, material, ER and hospital beds) is a key feature for a positive outcome. Since resurces and assets are finite, there is the need to be able to dispach the required vehicle/personnel in the measure that it is needed, providing the personnel both on the field and in a hospital with the best conditions to deal

with the emergency The tempestivity of the response and the rational use of the assets can be of paramount importance for the people involved and injured in an accident.

The model i propose here is based on the information that the witness of an accident can provide in a short amount of time to the emergency service, how the emergency service decides the severity of the accident, and decides to use its assets.

The number of information and the kind of information required for the model are kept at a minimum; to minimize the time spent communicating the information. Again, a prompt response can make the difference between life and death for an injured person.

The feature selection must take into account the information that an untrained person could give about the accident during a phone call to the emergency service:

Accident characteristics: **> Number vehicles involved**

> Geometry of the accident

> Number of people involved

> Eventual injuries

Locality information: **> Location**

Local conditions such as Road, light, and weather conditions are a plus.

Weather conditions can also be obtained quickly from local weather reports. Road and light conditions can also be obtained from third parties.

First, we need to observe what information are present in the Dataframe, and rationalize about which of these information is important for our prediction model and can be easily assessed and communicated, by an untrained witness of the accident, to the emergency services.

- **SEVERITY CODE**: it is our target
- **X and Y** : geographical information, the witness could simply inform the address, and the system would automatically transform that in X and Y coordinates
- **ADDRTYPE**: whether it is an intersection or a block, easy to assess and communicate
- **COLLISIONTYPE**: geometry of the collision, relatively easy to assess and communicate
- **PERSONCOUNT**: number of people involved. Can be hard to assess in case of a high number of people
- **PEDCOUNT**: number of people involves. Not trivial to assess.
- **VEHCOUNT**: number of involved vehicles. Can be hard to assess in case of a high number of vehicles
- **INJURIES**, **SERIOUSINJURIES**, **FATALITIES**: number of injuries, gravity of the injuries, and eventual fatalities. Can be hard to assess for an untrained witness.
- **INCDATE**, **INCDTTM**: date and time of the accident. Can be calculated automatically by the system at the time of the call
- **ROADCOND**, **LIGHTCOND**, **WEATHER**: conditions of the road, illumination, and weather. Can be no trivial to assess and communicate. Weather information can be obtained almost in real time automatically from local weather services.
- **UNDERINFL**: whether one or more involved people were driving under the influence. Hard to assess by an untrained witness.
- **SPEEDING**: whether one or more vehicle was exceeding the speed limit at the time of the incident. Hard to assess by an untrained witness.
- **Information that can be hard to assess by an untrained witness will not be considered for the feature list of the model.**

Four classification algorithms will be compared to choose the best predicting model. The choice will be made based on the accuracy of the algorithm.

The four algorithms are: **K-Nearest Neighbors, Logistic Regression, Decision Tree**, and **Random Forest**.

3 DATAFRAME EXPLORATION

The DataFrame is composed of 40 columns and 221266 data entries, each representing an accident. The data encompass the time period between the years 2004 and 2020.

3.1 Severity Code

The Severity Code is divided into 4 classes, depending on the absence/presence of injuries, serious injuries, and fatalities. We can see that most of the accidents do not involve severe injuries or fatalities (Figure 1).

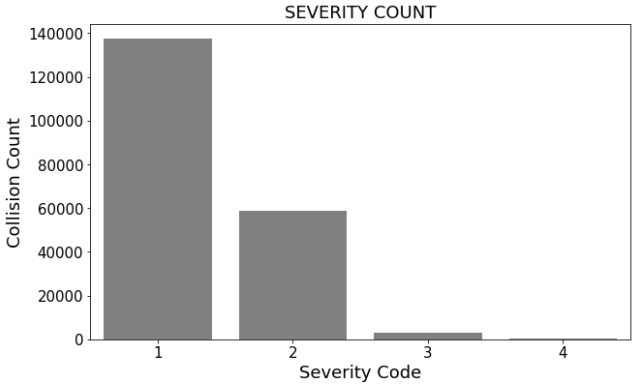

Figure 1 Cumulative number of accidents per Severity Code. The graph shown is after the dataset cleaning (see ch. 4)

3.2 Time Series

From the '**INCDATE'** and '**INCDTTM'** columns I extracted the year, month, day, day of the week and hour of the accident. There is no clear variation among the number of accidents with the year (except for 2020, having a lower total number do to being a complete year and being affected by COVID-19 limitation on people circulation), month, day, and day of the week, but there is a strong variation with the Hour of the accident. More accidents happen during the day than at night, with a peak at 17:00 (supposedly when people leave work *en masse*). As there is a variation in number of accidents during different hours of the day, we need to consider this in our prediction model, and turn this into a feature. (Figure 2). The apparent high number of accidents at 0h is an artefact of marking unknown time with '0'. This has been dealt in the data cleaning part (see. Below, cap.4.2).

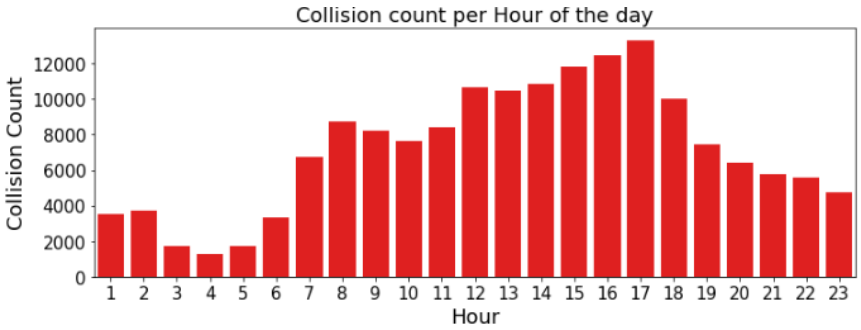

Figure 2 Cumulative numbers of accidents divided by hour of the day. The graph shown here represents the already cleaned data form.

3.3 Injuries and Vehicles

The number of injuries, severe injuries, and fatalities per accident is usually low, it amounts to no more than 2 in the vast majority of the accidents. It is very plausible that such a low number of expected injuries/fatalities can be easily assessed by an untrained witness. The same is valid for the number of vehicles involved, which is for the most accident only 2 vehicles involved (see *Figure 3*) .

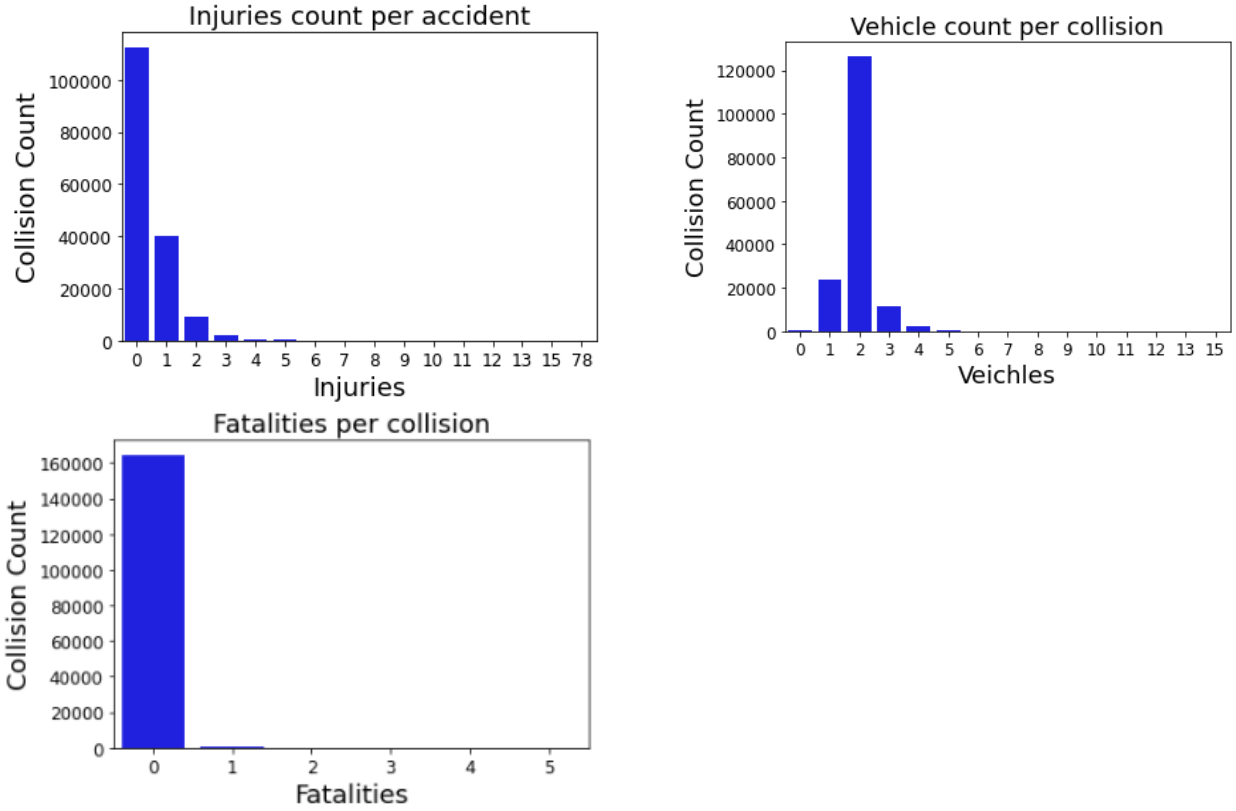

Figure 3 Cumulative number of injuries, fatalities, and vehicles involved in accidents.

4 DATAFRME PREPARATION

4.1 Dealing with NaN values

NaN values are present in most of the columns that we want to select as features.

In columns with discriptive or numerical classes, NaN values have been substituted by "0", here considered as "unknown". This allowed to keep most of the data entries.

Data entries with a NaN in either X or Y column have been discarded, as it is not possible to assign an "unknown" numerical values to a coordinate.

4.2Data Cleaning

The SEVERITYCODE has 4 classes, plus a '0' class for unknown classification. As I am not trying to predict a unknown feature, I dropped all the data entries with a SEVERITYCODE = 0. Also, I replaced the values of the original severitycode with others easier to handle: class 2b became class 3 and class 3 became class 4. Now the SEVERITYCODE has values from 1 to 4, so described:

CLASS	DESCRIPTION
1	Accident with only property damage
2	Accident with injuries
3	Accident with severe injuries
4	Accident with fatalities

Table 1 Severity Code class description after data cleaning

The descriptive classes of LIGHT, ROAD, WEATHER, and COLLISION TYPE have been converted to numerical classes (i.e. 0, 1, 2, 3 etc…). Furthermore, we see that for LIGHT, ROAD, and WEATHER conditions, only 2 or 3 classes account for most of the accidents (see *Figure 4*) . We can therefore reorganize and lower the number of classes. For example, most of the accidents occur with "Clear" weather conditions, while only a small number occur for each of the other classes: the feature has been restructured and all the classes other than "Clear" have been grouped as "Other".

The 'Hour' feature shows that most of the accidents happen durign peak hours in the afternoon, but as unknown time was marked as "hour 0", we have an apparent surplus of accidents happening at 0h, due to the sum of real accidents and unknown time. I chose to drop all the data entries with a '0h' timestamp.

The number of INJURIES, SERIOUSINJURIES and FATALITIES are used to assign the severity code of the accident, therefore there is a very high correlation between these features and the SEVERITYCODE. As stated in the business plan (cap.2) I suppose that all the information are to be collected and communicated by untrained witnesses to an emergency service, and, being untrained in the field, the witness might not be able to distinguish at a glance between an injury, a severe injury, and even

a fatality (e.g. fainted person vs. dead person). I chose to sum the values of the three classes for each accident. The new column was called ‘TOTALINJURIES”

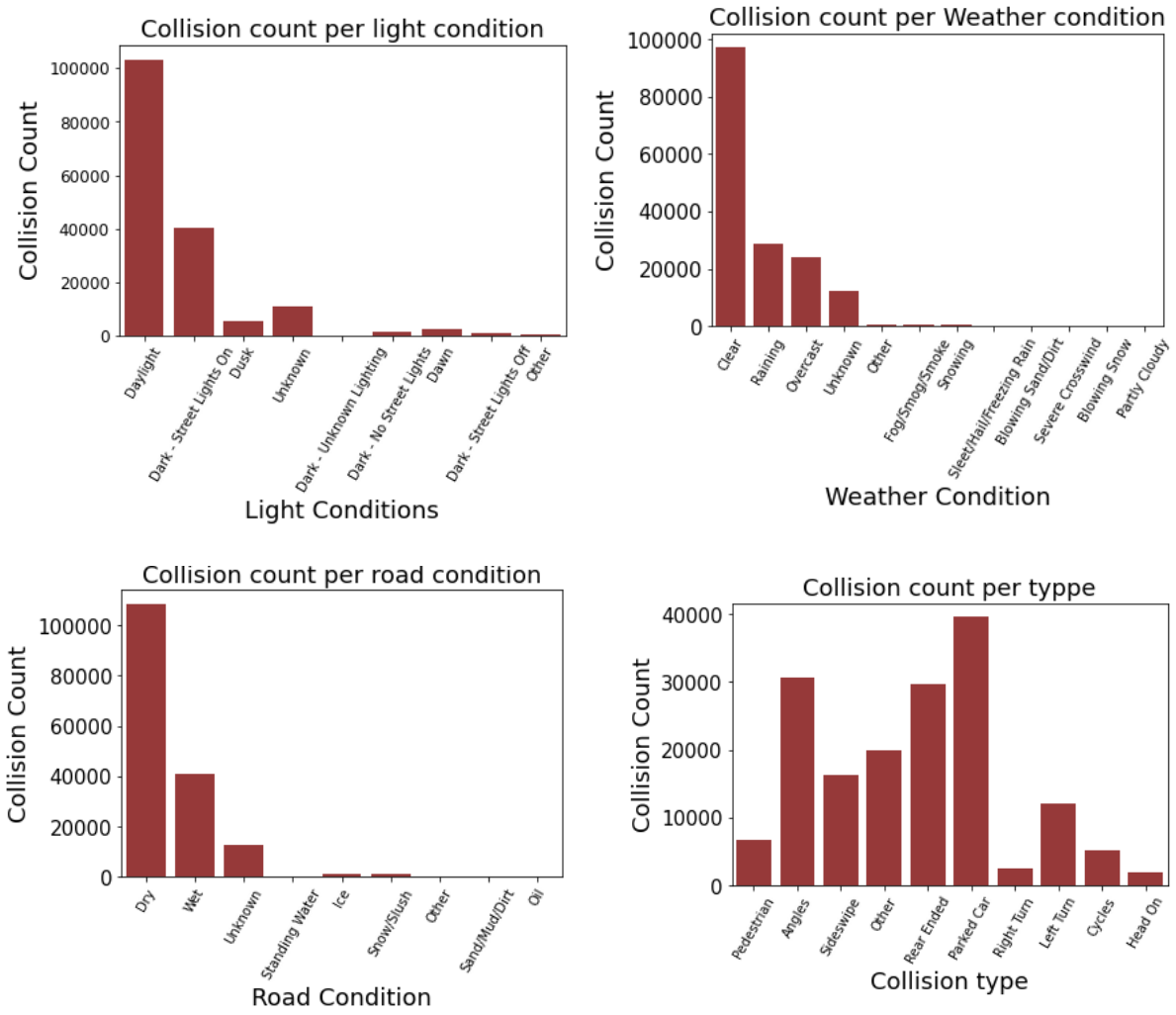

Figure 4 Cumulative number of accidents per Light, Weather, Road conditions, and Collision Type

The correlation matrix graph (Figure 5) shows how strongly correlated are the injuries (TOTALINJURIES) and the SEVERITY CODE, as expected, given that the severity code is assigned based on the presence and gravity of injuries.

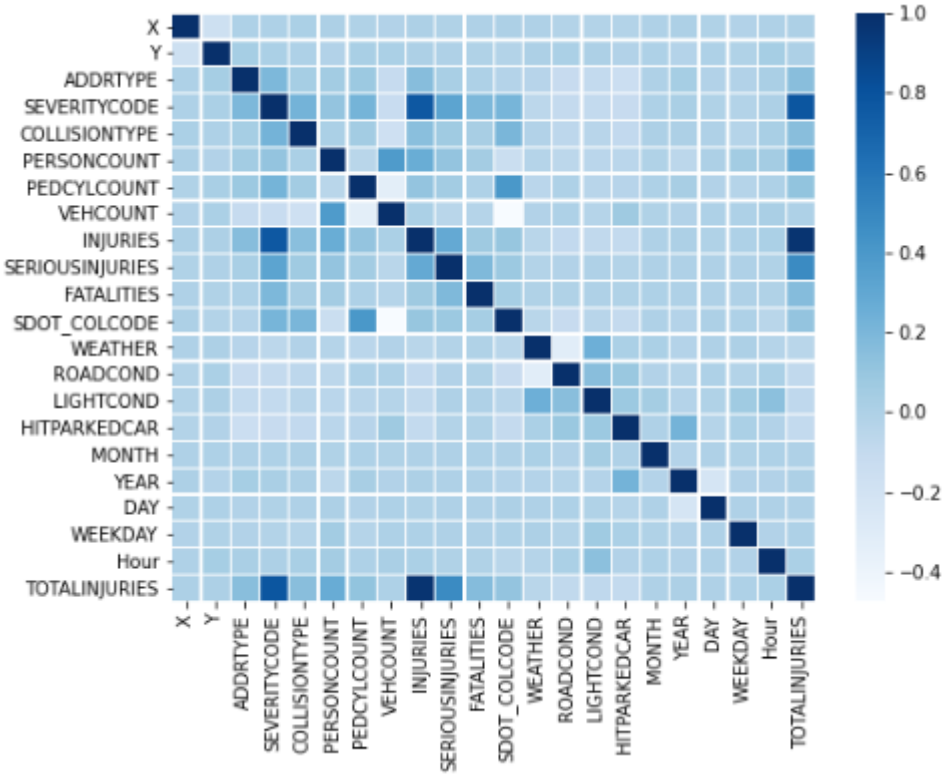

Figure 5 Graphical representation of the Correlation matrix of the variables in the cleaned dataset. Darker color represents a higher correlation.

4. 4 Feature Selection

After the exploration and the cleaning of the DataFrame, and after taking in consideration the information that a witness could be able to quikly communicate during an emergency service phone call, the following columns have been selected as features for the model:

X, Y, ROADCOND, WEATHER, LIGHTCOND, TOTALINJURIES, HOUR, COLLISIONTYPE, ADDTYPE, VEHCOUNT.

The target of our model is the column **'SEVERITYCODE'**

5 MODELING

5.1 K-Nearest Neighbors

The best accuracy (**0.9774**) was reached with a **K=5** (Figure 6).

The confusion matrix shows that the model predict with 100% accuracy the Severity Code 1 (only property damage, without injuries).

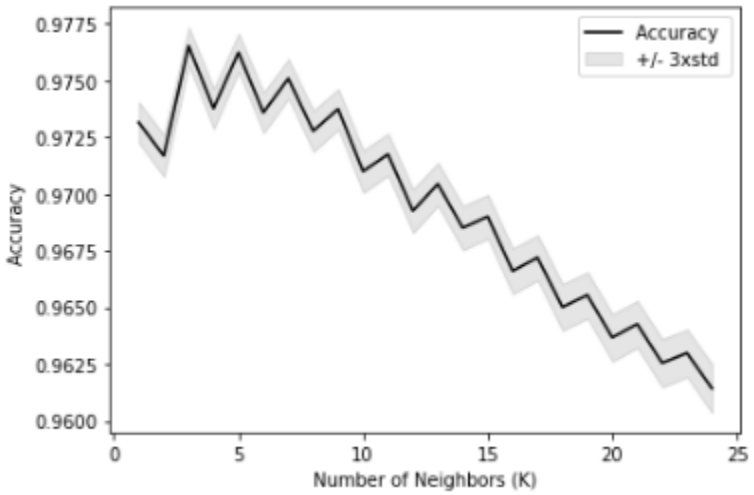

Figure 6 Variation of accuracy relative to the number of Ks

The test accuracy of the K-NN model is **97.62%.** The confusion matrix shows that the model predict with 100% accuracy the Severity Code 1 (only property damage, without injuries).

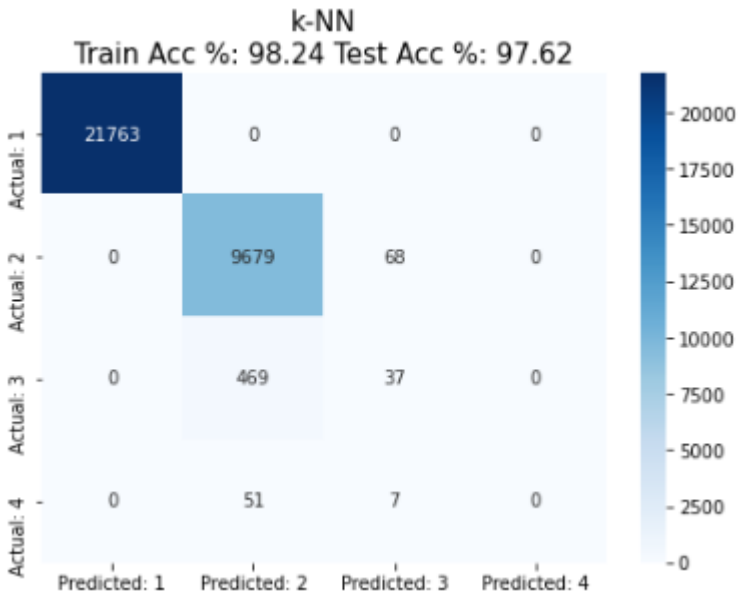

Figure 7 Confusion matrix for the KNN model

5.2 Logistic Regression

Logistic Regression modelling was performed with a C=5 and maximum number of iterations of 1000.

The test accuracy of the logistic resgression model was 98.14% (Figure 8).

The confusion matrix shows that the model predict with 100% accuracy the Severity Code 1 (only property damage, without injuries).

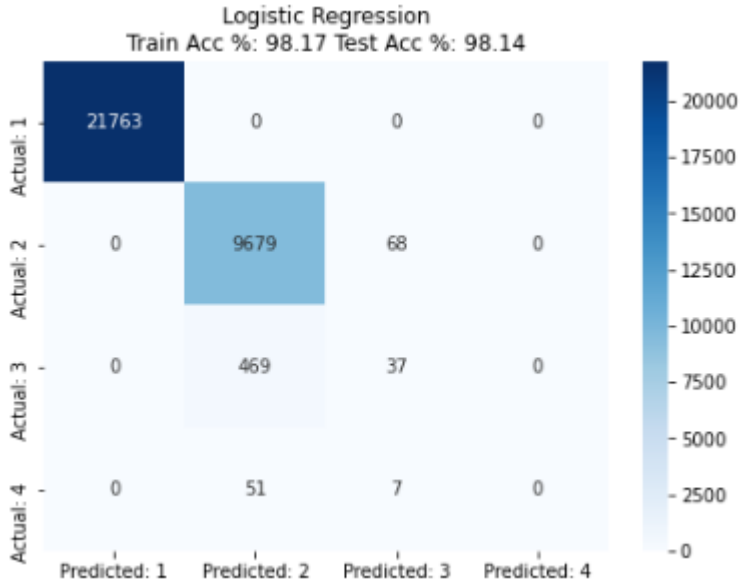

Figure 8 Confusion matrix for the Logistic Regression model

5.3 Decision Tree

The Decision Tree algorithm was calculated using as criterion ‘Entropy’ and a depth of 5. The best value of depth was choosen based on the highest accuracy score with a c value between 1 and 50 (Figure 9).

The algorithm has a test precision of 98.77% (Figure 10). The confusion matrix shows that the model predict with 100% accuracy the Severity Code 1 (only property damage, without injuries).

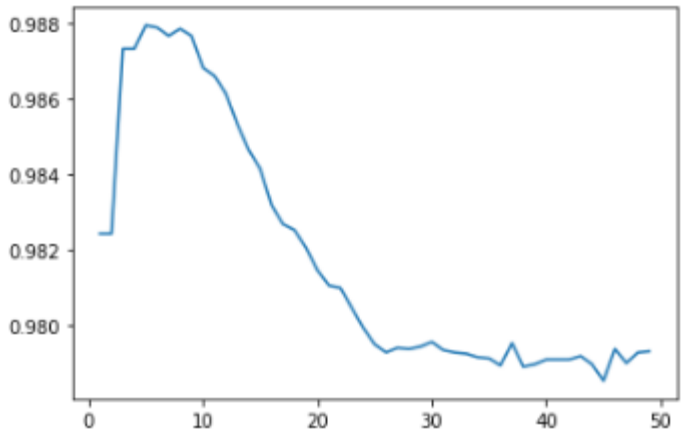

Figure 9 Variation of Accuracy score with respect to depth for the Decision Tree algorithm

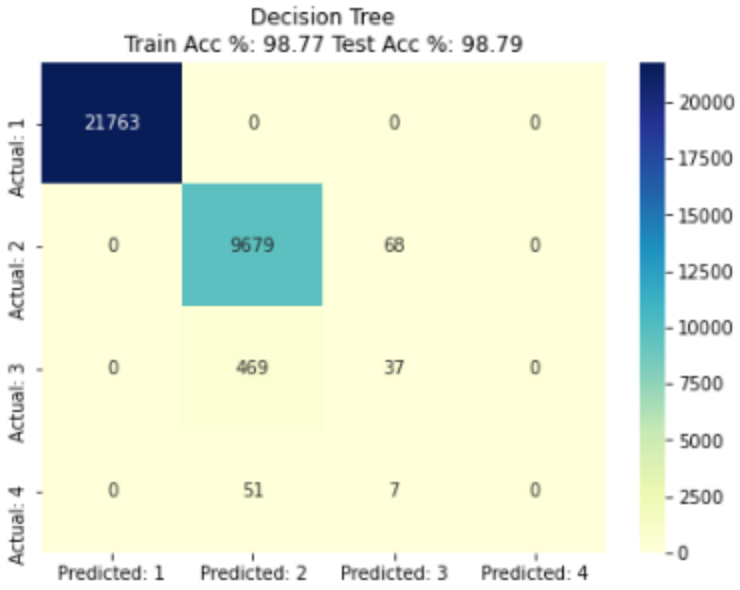

Figure 10 Confusion matrix of the Decision Tree model

5.4 Random Forest

The following parameters were used for the Random Forest model: bootstrap=**False**, max_depth=40, max_features='sqrt', n_estimators=50, random_state=42.

This algorithm performed very poorly compared with the other choosen models, with a test accuracy of 67.94 % (Figure 11). The confusion matrix shows that the model completely fails at predict any accident with a severity code 2, 3, 4. All the accidents are predicted to fall in the severity code 1.

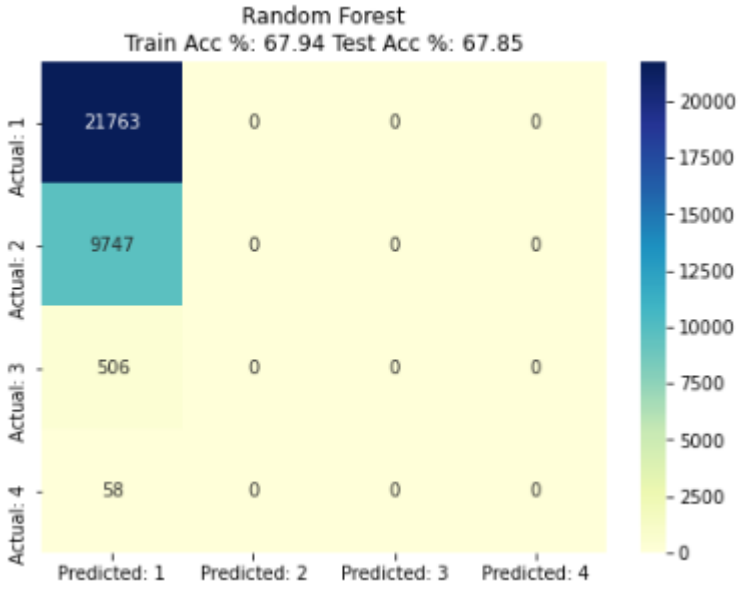

Figure 11 Confusion Matric of the Random Forest model

5.5 Models Comparison

After running the 4 algorithms I compared the accuracies of each one to select the best one to apply to my case study. The accuracy scores of the algorithm are reported in Table 2 and Figure 12.

	Algorithm	Accuracy_Score
3	Random Forest	0.678525
0	K-Nearest Neighbors	0.976242
1	Logistic Regression	0.981449
2	Decision Tree	0.987934

Table 2 Accuracy scores for the 4 algorithms

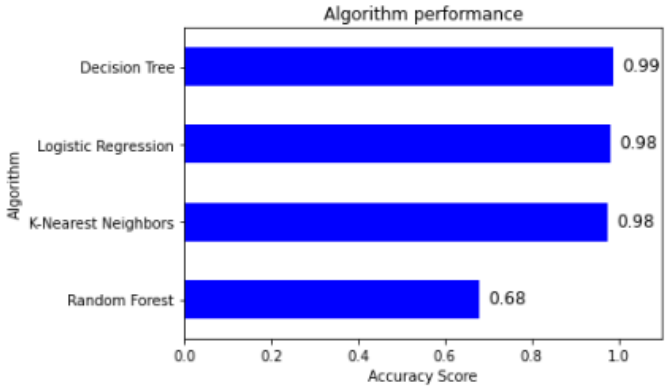

Figure 12 Histogram plot of the accuracy scores for the 4 algorithms

The results show that Random Forest, K-Nearest Neighbors and Decision Tree have a very similar accuracy. The best model is the Decision Tree, with an accuracy of 99%. Such high accuracies are expected, as the feature that most correlates with the target is the number and severity of injuries, which is also the parameter used by the SDOT to define the severity code. This can be seen in Table 3, where the importance of each feature in the Decision Tree model is shown. The Random Forest model performs worse than the others (accuracy of only 67%).

	FEATURE	IMPORTANCE
1	TOTALINJURIES	0.982
2	VEHCOUNT	0.015
3	ST_COLCODE	0.02
4	COLLISIONTYPE	0.01
5	X	0.00
6	Y	0.00
7	WEATHER	0.00
8	ROADCOND	0.00
9	LIGHTCOND	0.00
10	ADDRTYPE	0.00
11	HOUR	0.00

Table 3 Importance of each feature in the Decision Tree model

6 DISCUSSION

After comparing the 4 algorithms, the best one, based on the accuracy on the Test set, is the Decision Tree. The extremely high accuracy of this model (and the others) is because injuries is the parameter used by the SDOT to assign a severity code, and it also is a feature of the model. The selection of the features took in consideration only information that a witness to an accident could observe and quickly and concisely report to an operator during an emergency call. Selecting the injuries information as part of the features resulted in a clearly biased model, but I considered, based on common sense, a fundamental information in a real world scenario to be communicated to the emergency services.

The confusion matrices for K-Nearest Neighbors, Logistic Regression, and Decision Tree predict with 100% accuracy all accidents with a Severity Code 1 (only property damage), while usuallyt slightly underestimating the Severity Code of accidents that involved injured people. This suggests that the model can be completely trusted when it predicts only property damage, but when injuries are predicted, it must taken into consideration the possibility of a more serious situation. In the real life application, this means that the operator of the emergency service must be ready to dispatch more emergency vehicles/personnel if the necessity should arise. This part of the prediction model is the one that might require refinement, to increase the accuracy.

The importance of each feature tells us which information is most important, not just for the model *per se*, but also for the real case scenario, when a witness must inform the emergency service about the accident. Given the importance values, the information that a witness must pass are first, the presence of injured people, then the number of vehicles involved, and then the geometry of the accident. Of course, in a real case scenario knowing the location of the accident is fundamental to dispatch the proper emergency vehicles, even if in the model the location (X and Y coordinates of the accident) are of negligible importance. The other features have such a low importance in the prediction model that there is no necessity for them to be communicated during an emergency phone call.

6 CONCLUSIONS

I analyzed the Database of road accidents provided by the Seattle Department of Transportation and generated a prediction model using a Decision Tree classification algorithm to predict the severity of an accident given basic, easy to assess information. The chosen model was extremely effective in predicting the severity of accidents where no damage to people happened. The prediction of injuries is still not 100% accurate, but considering that I limited by design the number of features of the model, it should still considered a valid prediction model for the business case. The insights gained from analyzing the results indicate also that for a valid prediction only a very limited number of characteristics of an accident must be communicated to the emergency service, namely the presence of injuries, the number of vehicles, and the geometry of the accident/collision (beside, of course, the location or address of the accident).

This limited number of characteristics necessary to assess the gravity of an accident means that this information can be passed extremely quick from the witness to the operator, and this saved time could make the difference between a serious injury and death for an eventual person involved in the accident.