Coursera Capstone Report:

Collision severity of car accidents in Seattle

Julian Withöft

September 10, 2020

1 Introduction

An important part of today's transport policy is focused on improving the safety for road users. Therefore, cars are experiencing a continuous improvement concerning the safety aspect, which can be seen by added features starting from seat belts over anti blocking systems and electric power steering up to forward-collision and blind spot warning. These features have led to less accidents and more safety on roads, but there are still accidents especially with passengers involved leading to road fatalities. This report focuses on the collision severity of car accidents and analyzes connections between the collision severity code indicating the severity of an accident and the other related accident information. Based on the gathered details and influencing factors of the accident different models based on machine learning algorithms are trained to classify accidents into the collision severity code based on the other relevant information. Finally, a statement on the danger of travelling under specified circumstances can be constructed and therefore a warning under very dangerous circumstances can be formulated to prevent further accidents and improve the safety for road users even more.

2 Data

The data source for this project is the "Data-Collisions" dataset, which was made available in the IBM Data Science Capstone course. This dataset consists out of 38 columns, which contain the features and accident information of 194673 accidents that occurred in the city of Seattle and are displayed in the single rows. The features include geographical, accident course, weather, road & light condition, address and most importantly collision severity code information. The collision severity code information is in this case the dependent variable. Only some of the other information is useful and relevant to analyze and predict the dependent variable, namely the following features:

X, Y, ADDRTYPE, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, WEATHER, ROADCOND, LIGHTCOND, SPEEDING, ST_COLDESC, HITPARKEDCAR

Data cleaning included the deletion of rows with NaN values and especially relabeling and aggregating similar entries for some of the features. For example for the ST_COLDESC feature has 62 unique entries, which makes this feature very unwieldly and confusing, especially because many entries are basically describing a similar event like 'From opposite direction – one left turn – one straight' and 'From opposite direction – both going straight – sideswipe'.

The dependent variable, the collision severity code, describes the outcome of the accident and has two characteristics: Severity code 1 stands for an accident, where the collision only affects the vehicle but not the people therein, while severity code 2 implicates human collision.

3 Methodology

For analyzing, visualizing and predictive modelling on basis of the dataset various methods within the Python 3 Jupyter Notebook environment are used. At the beginning and for some interim recaps of the data the df.head()-function from the pandas library is used to explore the dataset.

Looking at the data analysis, the relevant features that were selected in Chapter 2 are viewed individually. Statistical testing is performed upon each feature based on the relevant unique entries and the collision severity code. Thereupon the number of cases for each collision severity code are calculated with the df.loc()- and df.count()-function from the pandas library, which lead to the percentages for each unique entry and each collision severity code. Moreover, the same procedure is also applied to only the collision severity code, to assess the general amount of accidents in each collision severity code.

For visualization purposes the violin plot from the seaborn library is used to illustrate the categorical entries of the selected features. The violin plot is constructed upon a matplotlib subplots figure to control the size parameter of the final plot and enable a tight layout saving option for the plots. In addition to that the figure 'axes' object is also used to assign x and y axis labels, set major & minor locators, create a grid and a limitation for the y axis to make the plot more appealing and easier to understand. After the calculation of the above-mentioned numbers and percentages for each entry of the features, the 'axes' object is used to write this information into the plot to provide every important insight of a feature at a glance.

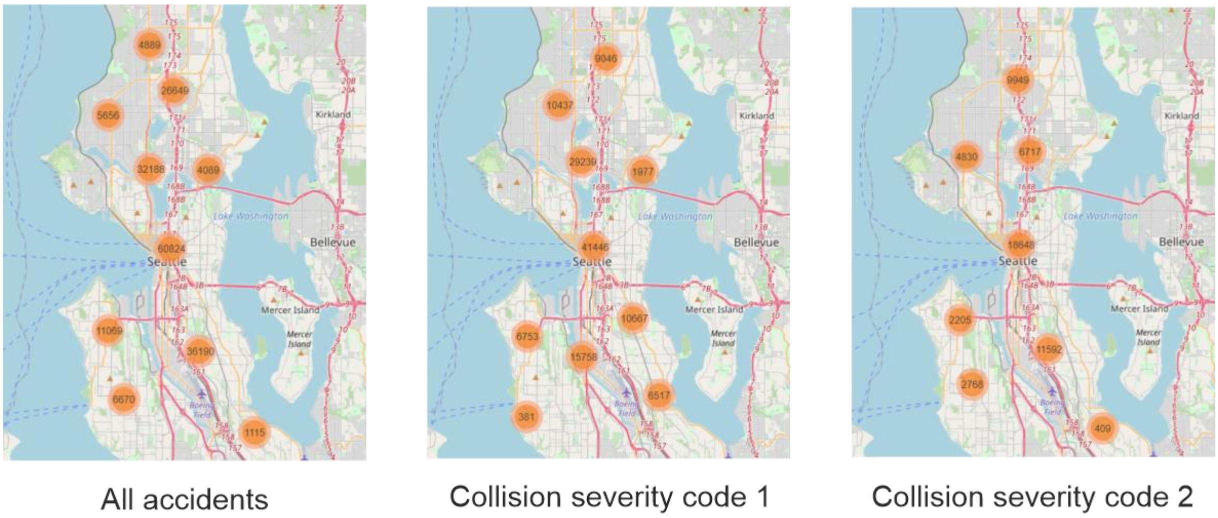
Apart from that for visualization of the geographical information with the x and y coordinates a map is generated with the folium library. This map is equipped with clustered markers, which are based on the x and y coordinates to show where and how many accidents exactly happened in Seattle.

Finally, for predictive modelling four different machine learning classification algorithms are used to predict the collision severity code of an accident based on the above-mentioned features. The four classification algorithms implemented with the scikit-learn library consist out of the decision tree, the k-nearest neighbor, the logistic regression and the support vector machine algorithm. The predictive modelling data was split into training data with a percentage of 70 % and an inverse testing data with a percentage of 30 % by the train_test_split()-function from the scikit-learn library. Before the feature data could be used to train and test the algorithms a transformation of the categorical and non-numeric features had to be applied. In that regard the LabelEncoder()-function from the scikit-learn preprocessing library was used to transform the categorical or string values into numeric values. Moreover, for the k-nearest neighbor algorithm the k-value with the highest accuracy was determined graphically by plotting the accuracy for 19 different k-values. After the application of the machine learning algorithms, the accuracy, F1 score and MSE were calculated with the scikit-learn metrics library and printed out afterwards.

4 Results

In the following the different features and indicators and their influence on the collision severity code are analyzed and visualized showing the relationships within the dataset. Based on the recognized knowledge about the data machine learning classification algorithms are then applied to predict the collision severity code from the feature data.

Firstly, the geographical information, namely the x and y coordinates, are considered. In that regard a clustered map with all accidents and two more maps showing accidents with a specific severity code, namely 1 or 2, are shown in Figure 1. The maps show that overall, over 30 % of the accidents occurred in the city center of Seattle, as roughly 60000 accidents happened in that area and a total of around 195000 accidents are contemplated. Also, there is a clear trend that the further away the location is from the main traffic roads, the fewer accidents occur. The amount of cases considered to be classified as collision severity code 1 clearly exceeds the amount of cases classified as collision severity code 2. This aspect is also backed by the statistical analysis, which shows that 70 % of the cases belong to collision severity code 1 and only 30 % are classified as collision severity code 2.



All accidents

Collision severity code 1

Collision severity code 2

Figure 1: Clustered accident map for all accidents and for specified collision severity codes 1 or 2

When looking at another feature concerning the geographical and surrounding circumstances of the accident the address type seems like a reasonable detail. This feature contains the three main characteristic entries intersection, block and alley as shown in figure 2. From the figure it can be observed that accidents related to an intersection tend to belong to the collision severity code 2 more often with a percentage of 42,25 % than accidents that happened in an alley with only 10,92 %. This means that intersection related accidents have a higher percentage of leading to a human collision, while accidents in an alley have a much lower risk of a human collision and lead more often to vehicle collision with a percentage of 89,08 %. In between these two extremes lies the address type with 23,71 % of the cases belonging to collision severity code 2 and the inverse 76,29 % being classified into the collision severity code 1.

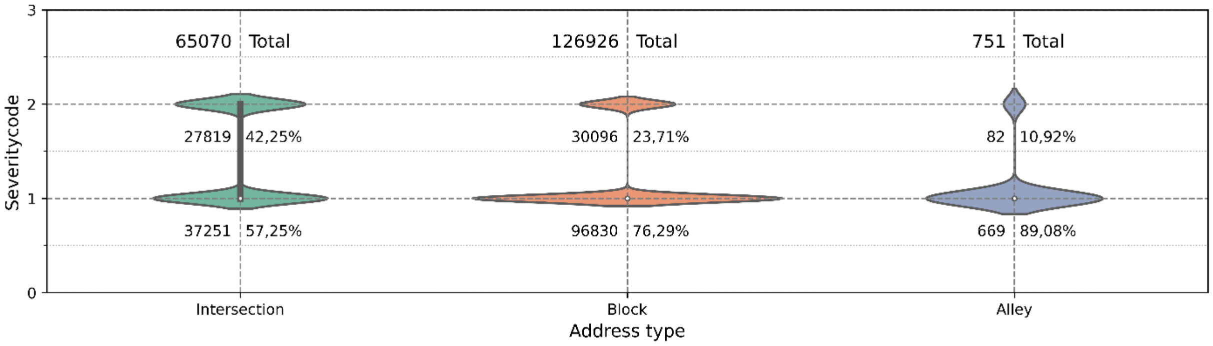


Figure 2: Violin plot of the collision severity code versus the address type of the accidents

Another relevant influence factor on the collision severity code is the predominant weather at which the accident emerged. Figure 3 shows, when the weather is defined "Clear" the collision severity code is very obviously more likely to belong to class 2 with 32,25 % than for weather declared as "Snowing" with only 18,85 %, "Hail" with 24,78 %, "Sand/Dirt" with 26,79 % or "Wind" with 28 %. For weather defined as "Raining" or "Fog/Smoke" the amount of cases belonging to collision severity code 2 is only slightly higher by a few percent, while for weather declared as "Overcast" it is slightly lower. When the weather is "Other" or "Unknown" it is with 94,15 % very likely, that the collision severity code has a value of 1.

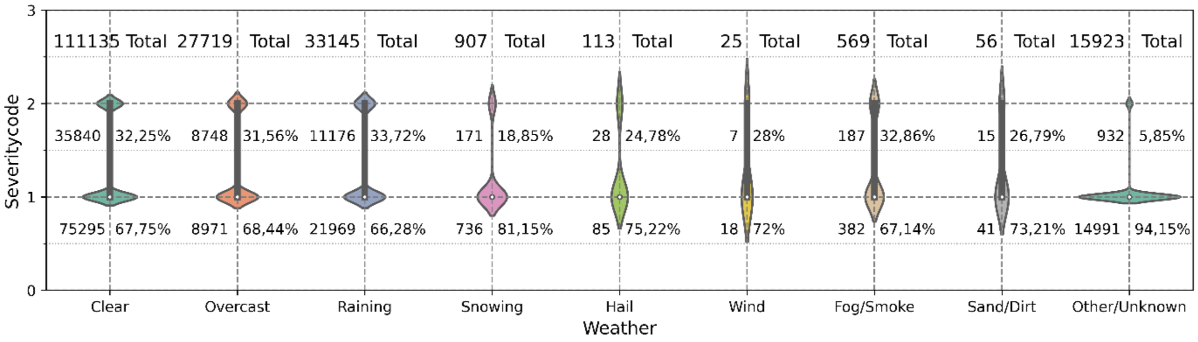


Figure 3: Violin plot of the collision severity code versus the weather of the accidents

The next relevant factor is the light condition at which the accident occurred. This feature is divided into four characteristics as shown in figure 4: Day, Night, Dusk and Dawn. It can be observed that more accidents with collision severity code 2 happen during the daytime with 33,19 % than at nighttime with only 29,52 %. Moreover, dusk and dawn share a similar amount of collision severity code 2 accidents with 32,94 and 32,93 % and therefore lie in between the values for daytime and nighttime. Still the results are more comparable to the daytime light condition as the values are much closer to the daytime than the nighttime values. Lastly, the accidents where other or unknown light conditions are predominant, tend to be once again more likely to belong to collision severity code 1 than 2 with a percentage of 95,21 %.

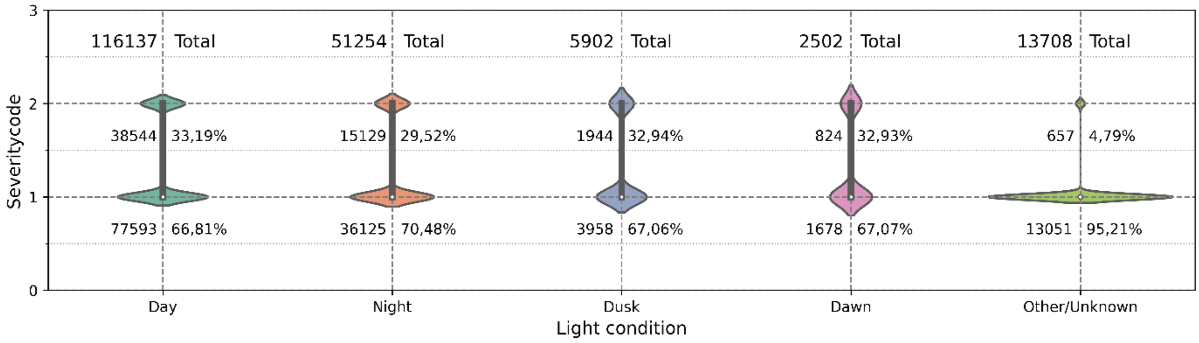


Figure 4: Violin plot of the collision severity code versus the light condition of the accidents

Besides the light condition, the road condition also plays an important role for the collision severity of the accident. As shown in figure 5, six characteristics are specified: Dry, Wet, Snow, Ice, Dirt and Oil. For a dry road condition in 67,82 % of the cases the collision severity code is 1. This value is slightly lower, when a wet road condition is present with 66,83 %, and significantly lower when an oily road condition is present with 62,50 %. In contrast this value is slightly higher for dirty road conditions with 69,33 % and significantly higher for icy and snowy road conditions with 77,42 % and 83,37 %. Like before in the case of other or unknown road conditions it is likely to be a collision severity code of 1 with a probability of 94,79 %.

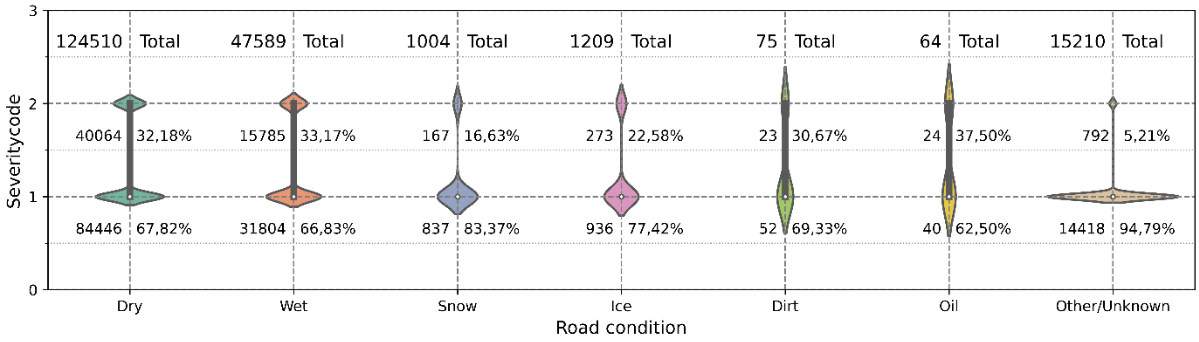


Figure 5: Violin plot of the collision severity code versus the road condition of the accidents

After viewing the external factors of the accident, a more to the accident itself relating parameter, the collision type, is analyzed regarding the collision severity code. As shown in figure 6, in case of a head on collision, with a probability of 43,08 % a collision severity code 2 is extracted. A similar probability can be extracted for collision types described as “rear ended” with 43,04 %, while for accidents described as “left turn” or “angles” the probability is a bit lower with around 39 %. A “right turn” collision type leads a lower probability with 20,60 % and a “sideswipe” collision type leads to even lower probability with 13,47 %. In contrast a collision type classified as “Parked” leads to an extremely high probability for collision severity code 1 with 94,45 %. A reverse high probability for collision severity code 2 is noticed with collision types defined as “Cycles” and “Pedestrians” with almost 90 %. Finally, a collision type classified as “Other” has a probability of 25,79 % to belong to collision severity code 2.

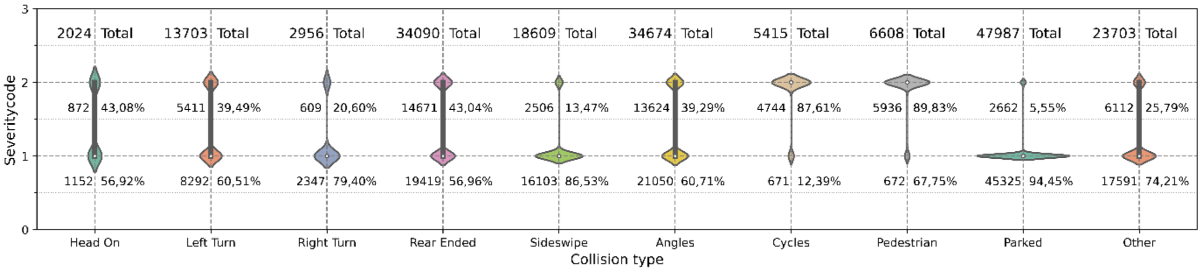


Figure 6: Violin plot of the collision severity code versus the collision type of the accidents

Now an even closer look at the course of the accident is taken. Figure 7 indicates that percentagewise there are more collision severity code 2 accidents for “Opp. Direction” with 40,16 % than for “Same direction” with 30,67 %. A similar value is obtained for a “Driveway” accident course with 29,89 % and “Multi-Vehicle” with 37,50 %. As before, for “Parking” a very high probability for collision severity code 1 is present, while for “Vehicle-Cyclist” and “Vehicle-Pedestrian” as well as “Cyclist-Pedestrian” the reverse probability for collision severity code 2 is very high. “Vehicle-Road” and “Vehicle-Animal” tend to be more likely collision severity code 1 related with 87,71 % and 86,11 %, while a slightly lower value for “Railway” and “Objects” with 79,19 % and 75,90 % is stated. Lastly, for “Non-Collision” the majority of cases belongs to collision severity code 2 with 65,76 %.

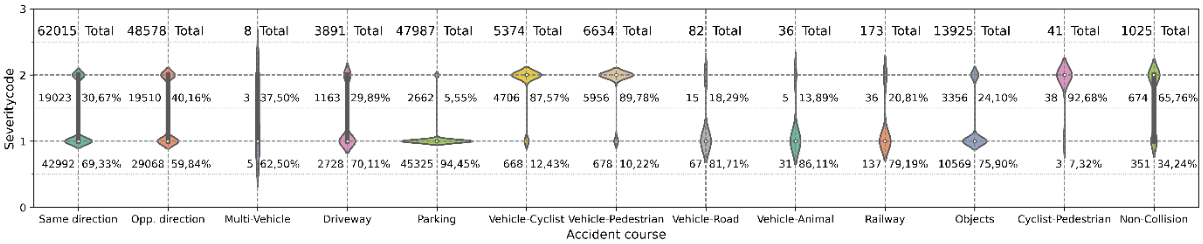


Figure 7: Violin plot of the collision severity code versus the accident course

The aspect of a parked car collision is analyzed with the next feature, which describes if a parked car was hit during the accident. Figure 8 shows, that if a parked car was hit it is very likely that the collision severity code is 1 and therefore no human collision occurred as it is for 93,79 % of the cases. If a parked car was not hit during the accident, the probability for collision severity code 1 is much lower with 69,20 %.

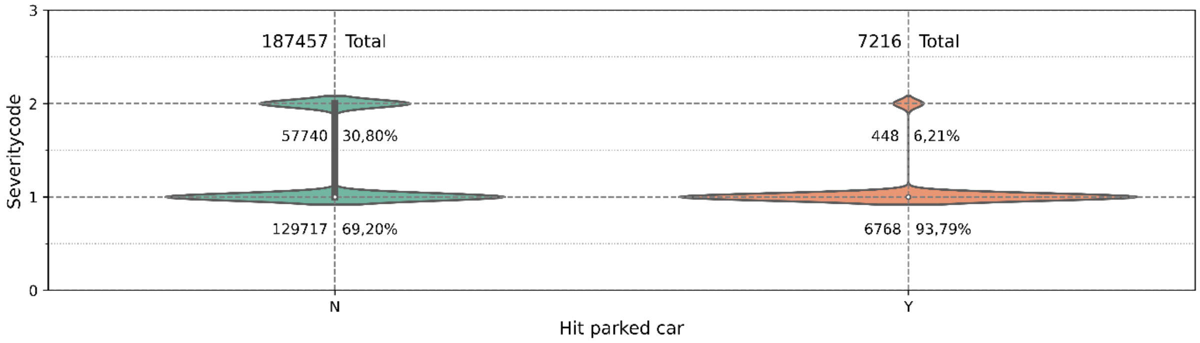


Figure 8: Violin plot of the collision severity code versus an indicator signalizing if a parked car was hit

The last relevant feature is an indicator if speeding was involved in the accident. If this indicator is true there is a greater likelihood that the accident belongs to collision severity code 2 as the probability is about 8 % larger than for a false speeding indicator, where 29,49 % of the accidents belong to collision severity code 2.

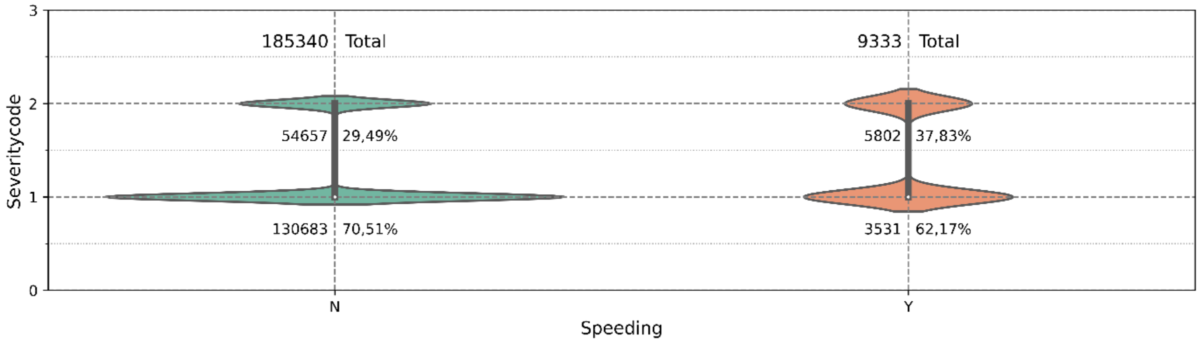


Figure 9: Violin plot of the collision severity code versus an indicator signalizing if speeding was identified

Besides the statistical analysis of the most relevant influence factors on the collision severity of car accidents, predictive modelling in form of machine learning classification algorithms is performed and compared. The goal of this is to predict the dependent variable, the collision severity code, based on the relevant accident information and features. The four relevant machine learning classification algorithms are the decision tree, the k-nearest neighbor, the logistic regression and the support vector machine. In case of the k-nearest neighbor algorithm the k with the highest accuracy was identified graphically with an optimum of 16.

Table 1 shows that the highest accuracy is achieved with the support vector machine with 75,81% of the testing data labelled correctly. However, the other algorithms are only slightly less accurate as the maximum difference in terms of accuracy is determined from the logistic regression, which is about one percent less accurate than the support vector machine algorithm. The accuracy of the k-nearest neighbor and decision tree algorithms lie in between these values with an accuracy of about 75,30 %.

The F1 score, which is also found in Table 1, is maximized with the decision tree algorithm with a value of 0,7239. This value is closely followed by the k-nearest neighbor algorithm, which realizes a F1 score of 0,7234. The support vector machines F1 score is also in this area with a value of 0,7219. Far behind with a F1 score of 0,7030 is the logistic regression, which is therefore again delivering the worst results.

Finally, the MSE shows that the lowest value of 0,2419 for the support vector machine algorithm, which is then followed by the k-nearest neighbor algorithm with a value of 0,2469 and the decision tree algorithm with a value of 0,2475. The largest MSE is stated for the logistic regression with a value of 0,2503.

Table 1: Accuracy, F1 and Jaccard score of the machine learning classification algorithms

Algorithm | Metric	Accuracy	F1 score	MSE
Decision tree	0,7525	0,7239	0,2475
K-nearest neighbor	0,7531	0,7234	0,2469
Logistic regression	0,7497	0,7030	0,2503
Support vector machine	0,7581	0,7219	0,2419

5 Discussion

The results of the clustered accident map in figure 1 are expectable as the very dense and stimulated traffic in the city center is usually accompanied with more accidents, while the less populated areas outside of the city center in the suburban regions tend to have a lower volume of traffic and therefore less frequent accidents.

The same reasoning can be consulted for the address type feature shown in figure 2. The high rate of human collisions at intersections is due to the simultaneous crossing of pedestrians and turning of vehicles and the fact that drivers must watch out for oncoming traffic and pedestrians at once leading to more human errors and more collisions with humans taking damage, while also leading to heavier vehicle collisions when turning at the intersection. At a block, fewer human collisions are observed due to less pedestrians and less heavy collisions and in an alley even fewer human collisions due to similar reasons are extracted.

The results from the weather feature shown in figure 3 realized fewer human collisions for “Snow” and “Hail” entries as these circumstances rather lead to vehicle collision, because less pedestrians are on the streets and the accidents are typically less devastating and include smaller incidents. Similar reasoning can be given for the fewer human collisions from “Sand/Dirt” and “Wind” features. For “Clear”, “Overcast” and “Rain” similar distributions were found as this is the normal weather condition that appears in most cases. The highest rate of collisions for “Fog/Smoke” weather is due to the impaired visibility conditions and the more severe collisions and the overlooking of pedestrians.

Looking at the light condition feature results from figure 4, a clear trend is visible. Daytime light conditions lead to more and nighttime light conditions to fewer human collisions, while dusk and dawn are in between these extremes, although the values are much closer to the daytime than to the nighttime results. This can be explained by a simple relation as more pedestrians are on the streets at daytime than at nighttime, which leads to more human collisions. At the same time dusk and dawn are the phase between these two main parts. Light is still available at this time, but there is less of it and still there are much more people on the streets than at nighttime, which is why the values rather correspond to the daytime values.

When comparing the road condition from figure 5 with the weather results from figure 3, they are clearly in line with each other and therefore the above-mentioned reasoning for the weather results are also relevant here.

For the collision type and accident course results from figure 6 and 7 a pattern can be recognized. Generally parked entries implicate a distribution with a lot of vehicle collisions and only very few human collisions, which makes sense as parked cars are usually not occupied and therefore only the velocity of the collision causing vehicle is present and normally relatively low. Moreover, any accident that involved pedestrians or cycles has a very high probability for human collision, which is self-explanatory. Also, in vehicle-vehicle accidents it is important to note from which direction the vehicles collided as coming from the opposite direction leads to more human collision than when coming from the same direction. Opposite direction especially includes head-on, left turn, sideswipe and angles accidents, while same direction consists out of right turn and rear ended accidents. The reasoning behind this implicates that opposite direction accidents involve contrary forces, while in same direction accidents forces are usually pointing into the same direction. This makes opposite direction accidents much more severe than same direction accidents leading to more human collisions.

As seen before the hit parked car indicator shown in figure 8 demonstrates that a parked car accident leads to a much lower probability for a human collision due to mentioned reasons.

Similarly, the speeding indicator from figure 9 shows that when speeding is involved, human collision is more likely going to be the case. This is due to the fact that a higher velocity than allowed makes the driver loose control on the vehicle and leads to more severe collisions in general.

Originating from the observations from the machine learning classification algorithms and the raised metrics from table 1, it can be noted that the support vector machine algorithm delivers the highest accuracy and the lowest error, while the decision tree and k-nearest neighbor algorithms share similar characteristics and reach the highest F1 scores and the second-best accuracy and errors. The worst performing algorithm is the logistic regression, which has the poorest accuracy, the highest error and the lowest F1 score.

6 Conclusion

In this report the collision severity of car accidents in Seattle was analyzed based on the relevant accident information. This consists out of geographical and address data, weather data, road and light condition data and address course information. The influence of these factors on the collision severity code was analyzed statistically and graphically and a reasoning for the detected relationships within the data was stated. The collision severity code defines if the accident caused human collision or only resulted in vehicle collision. Moreover, predictive modelling in form of machine learning classification algorithms were developed and compared to find the best classification algorithm for this use case. The use case was namely to classify an accident into human or vehicle collision based on the above-mentioned relevant accident information. With this predictive modelling accidents can be pre-classified based on the accident information and are therefore a valuable tool to predict and prevent severe accident or to intelligently distribute first aid responders onto the occurring accidents.