Predicting the Severity of Accidents

Data Science Capstone Project IBM/Coursera

September 2020

Stefanie Welcker

Table of Contents

1. Introduction

1.1 Background

In and around big cities with many inhabitants and many commuters, such as Seattle, accidents happen year-round. Sometimes accidents are severe and roads are blocked completely to rescue the injured, but even accidents with only propriety damage can cause traffic jams. It would be beneficial to have some warning beforehand so some areas could be bypassed.

1.2 Problem

This study aims to investigate whether the severity of an accident can be predicted on the basis of historical data including location, weather condition, light condition, road condition and others using supervised machine leaning.

1.3 Interest

Possible interested parties for the data can be local rescue stations, which can plan the rescue forces for example on the basis of local and weather forecasts. Better planning can both reduce costs and save lives by having the right resources readily available.
Another interested party could be radio stations, which can offer their listeners a forecast, either on traffic radio or on an online platform. This may increase the radio stations' income generated by advertisements due more listeners or clicks on a platform. Insurances can also have an interest in the data because they may be able to forecast costs based on the expected severity of an accident.

2. Data

2.1 Data Source

The city of Seattle provides data on collisions here:
https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

And a detailed description of the data here:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

The data is updated every week.

2.2 Data Cleaning and Preparation

The data covers several years of information starting at 2004-01-01 until 2020-07-29 at the time data was retrieved. Overall the original data has forty columns and 221266 rows, the SEVERITYCODE represents the label to be predicted, thirty-nine columns are possible features.

To clean the data, missing values in estimators weather, light condition and road condition were first filled with "Unknown" and subsequently deleted together with the indefinite specification "Others" because the data set is biased in favor of minor accidents and most of the "unknown" features occurred with minor accidents (Figure 1).

```
df3.groupby(['WEATHER'])['SEVERITYCODE'].value_counts(normalize=True)
```

```
WEATHER                  SEVERITYCODE
Blowing Sand/Dirt        1              0.740000
                         2              0.260000
Blowing Snow             2              1.000000
Clear                    1              0.662553
                         2              0.337447
Fog/Smog/Smoke           1              0.659537
                         2              0.340463
Other                    1              0.843829
                         2              0.156171
Overcast                 1              0.670941
                         2              0.329059
Partly Cloudy            1              0.555556
                         2              0.444444
Raining                  1              0.649170
                         2              0.350830
Severe Crosswind         1              0.680000
                         2              0.320000
Sleet/Hail/Freezing Rain 1              0.739130
                         2              0.260870
Snowing                  1              0.803532
                         2              0.196468
Unknown                  1              0.899007
                         2              0.100993
Name: SEVERITYCODE, dtype: float64
```

Figure 1: Example of checking impact of “Unknown” on feature WEATHER

Therefore I decided to drop the rows. Also missing values for location were dropped, because the location is used as an estimator.

Reclassification:
In the original data the severity is classified as:

- 3—fatality
- 2b—serious injury
- 2—injury
- 1—property damage
- 0—unknown

The data was reclassified by grouping “2”, “2b” and “3” into one group “2” for severe accidents because rescue teams will probably be needed.
The “0- unknown” category has been dropped because an “unknown” outcome has no information about severity that could be predicted.

The original data set provides the date of incident; the month and whether the day was on a weekend can be derived.

2.3 Feature selection

After cleaning and some data exploration (see chapter 3.1) the features for modeling are shown in table 1.

Feature	Explanation
X	Location as longitude and latitude. Already numerical used instead of the description
Y	Location as longitude and latitude. Already numerical used instead of the description
WEATHER	Intuitively, more severe accidents should happen under bad conditions
LIGHTCOND	
ROADCOND	
PERSONCOUNT	Number of involved persons
VEHCOUNT	Number of involved vehicles
JUNCTIONTYPE	Describes the kind of intersection like highway ramps or normal crossings
SPEEDING	Likely to influence the outcome due to higher physical forces

Table 1: Features for modeling

3. Methodology

3.1 Data Exploration

Since the date of the incident is recorded, we can see the development of accident in Seattle over the years. The proportion between minor and severe accidents is relative constant except for 2005-2009 where we can see a peak of minor accidents compared to the rest of the years (Figure 2).

Balance of data:

The cleaned data is imbalanced, containing 65.6% minor and 34.4% severe accidents.

```
# balance check
df3['SEVERITYCODE'].value_counts()
```

```
1    111183
2     58300
Name: SEVERITYCODE, dtype: int64
```

```
df3['SEVERITYCODE'].value_counts(normalize=True)
```

```
1    0.656013
2    0.343987
Name: SEVERITYCODE, dtype: float64
```

Figure 2: Balance of data

Trend of Years:

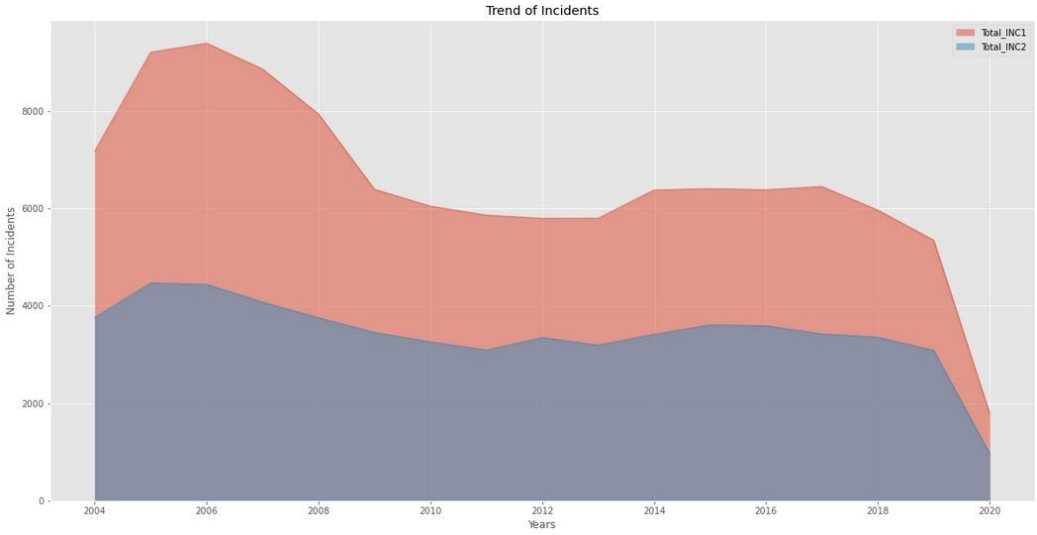

Figure 3: Trend of severe accidents (Total_INC2) and minor accidents (Total_INC1)

Day of Week:

Looking at the day of the week we can see a decrease of accidents with a slightly higher amount of minor accidents on Sundays. The day of week may influence the outcome of the accident, but it does not: for every day, the ratio is similar to that of the total data set (Figures 4 and 5).

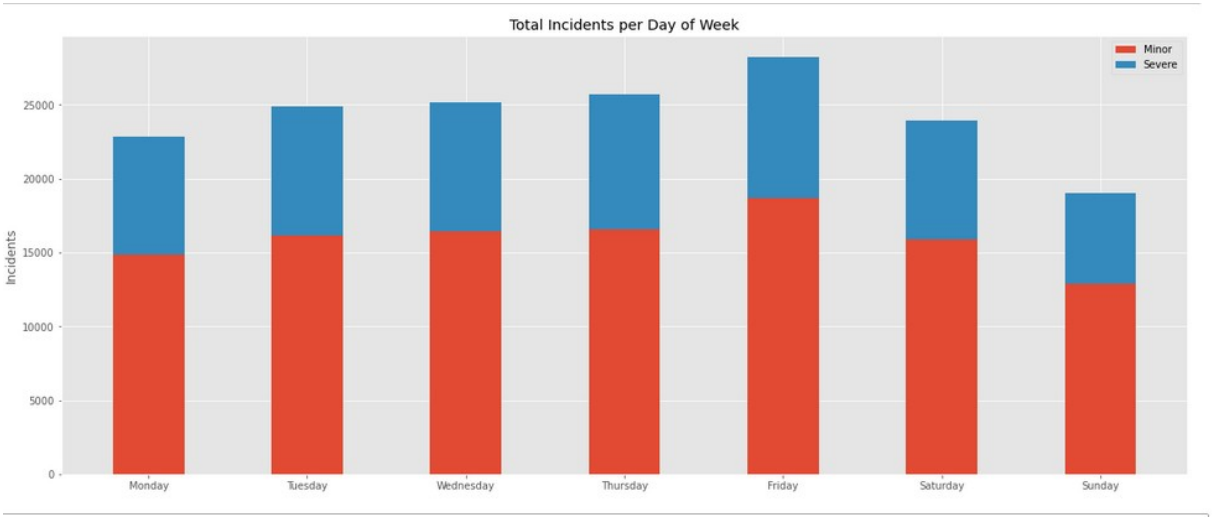

Figure 4: Total of incidents in the cleaned data set per week

```
: DAYOFWEEK  SEVERITYCODE
  0          1              64.943613
             2              35.056387
  1          1              65.034627
             2              34.965373
  2          1              65.196449
             2              34.803551
  3          1              64.482920
             2              35.517080
  4          1              66.033052
             2              33.966948
  5          1              66.287388
             2              33.712612
  6          1              67.672459
             2              32.327541
  Name: SEVERITYCODE, dtype: float64
```

Figure 5: Percentage of minor (1) and severe (2) accidents per weekday

Influence of Month:
The ratio of minor to severe accidents per month represents roughly the distribution in the whole data set (Figures 6 and 7). Therefore, the month was ultimately not included in the modeling features.

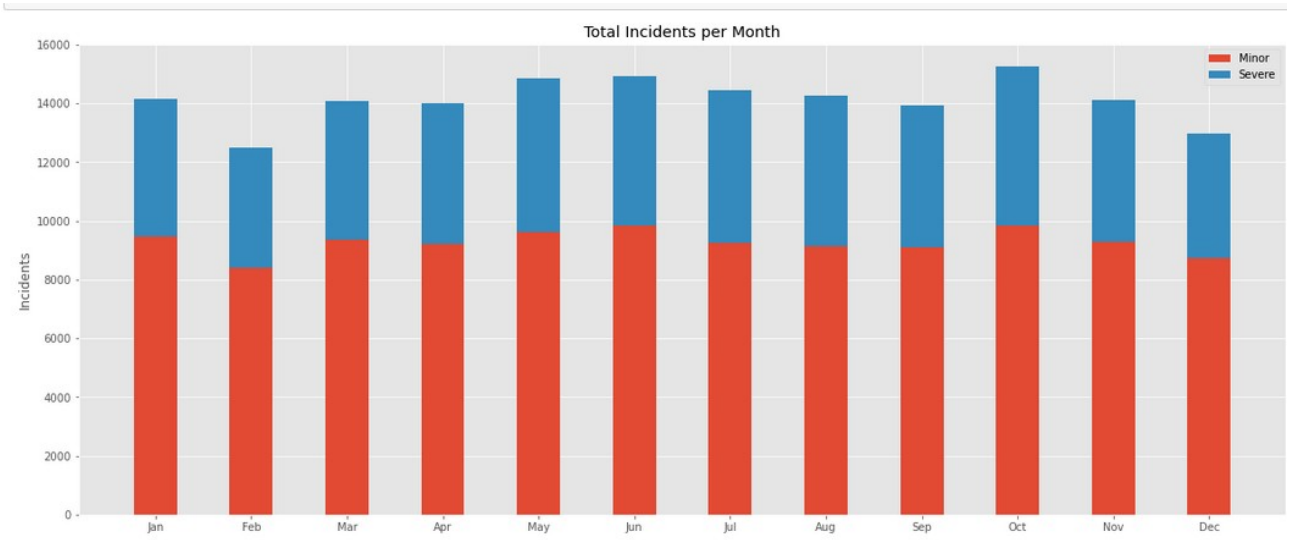

Figure 6: total accidents per month

```
: MONTH  SEVERITYCODE
  1      1               66.932580
         2               33.067420
  2      1               67.366565
         2               32.633435
  3      1               66.333049
         2               33.666951
  4      1               65.925661
         2               34.074339
  5      1               64.726489
         2               35.273511
  6      1               66.016358
         2               33.983642
  7      1               63.894846
         2               36.105154
  8      1               63.937758
         2               36.062242
  9      1               65.242983
         2               34.757017
  10     1               64.377373
         2               35.622627
  11     1               65.730496
         2               34.269504
  12     1               67.244701
         2               32.755299
Name: SEVERITYCODE, dtype: float64
```

Figure 7: Percentage of minor (1) and severe (2) accidents per month

Location:
When looking on the map where the accidents (starting with 2020-06-01) are clustered, the location is an indicator as to whether an accident is likely. Two accident black spots are detected (Figure 8).

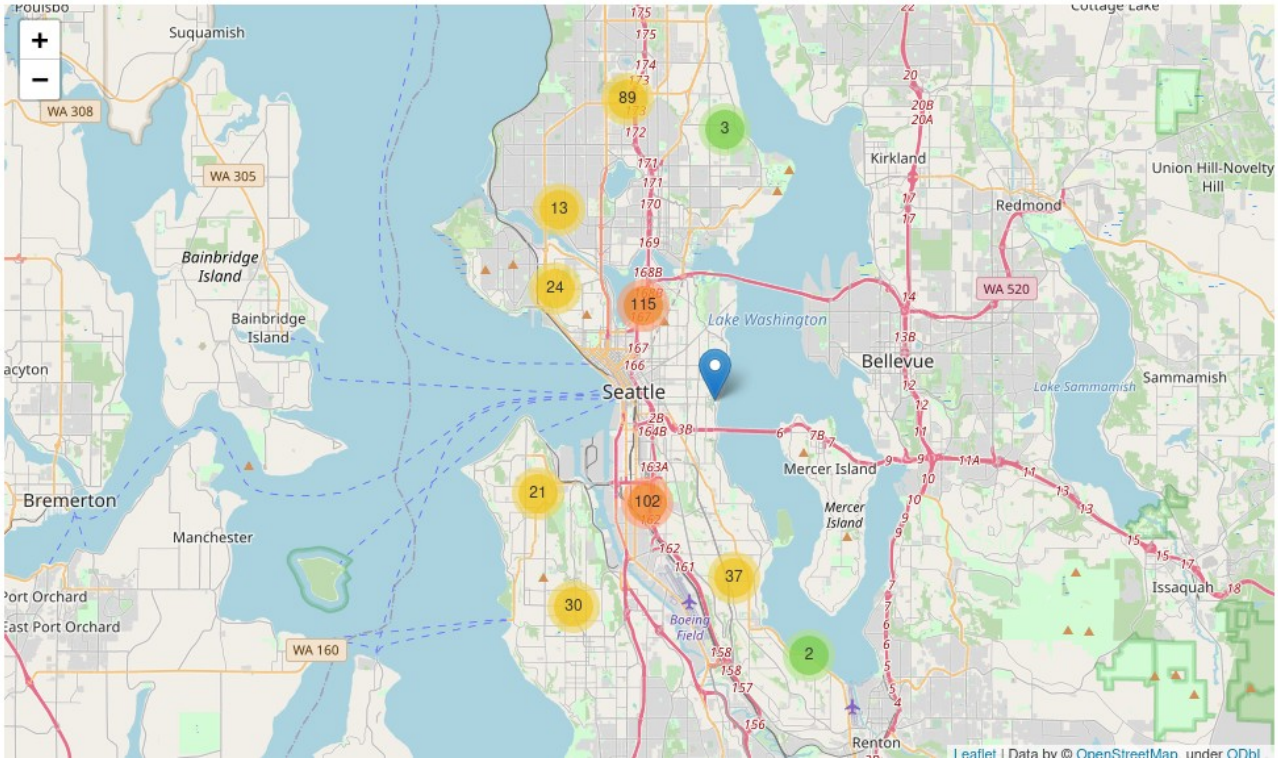

Figure 8: Accident black spots

Weather Conditions:
Regarding the weather conditions, the percentage of severe accidents in partly cloudy weather is significant higher than usual, whereas people appear to drive more carefully in snowy weather (Figure 9).

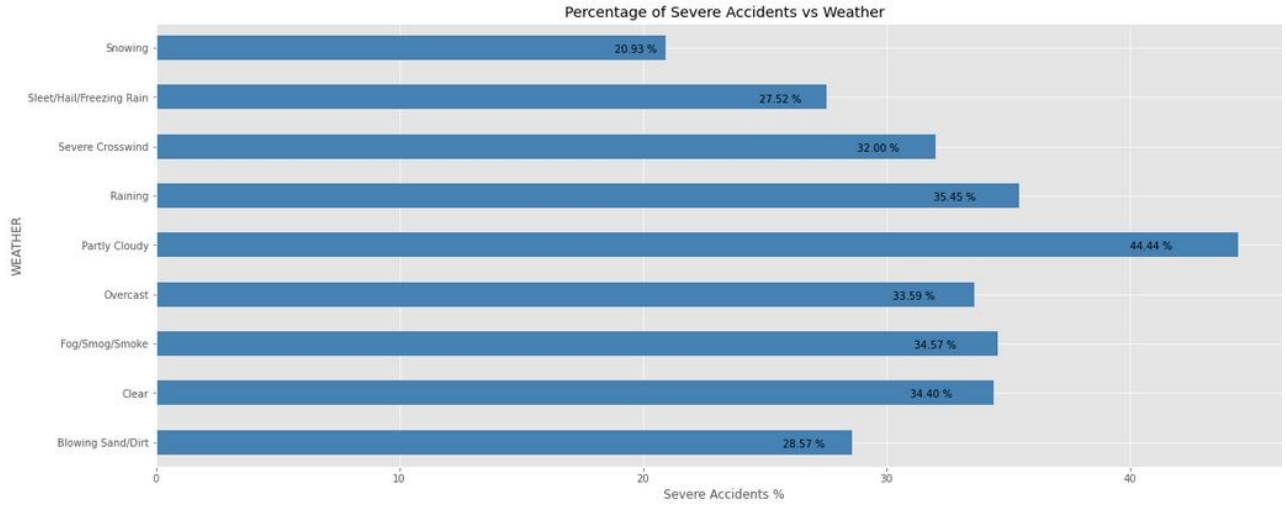

Figure 9: Influence of weather condition on severity of accidents

Junction Type:
Nearly half of the accidents related to intersections are severe (Figure 10).

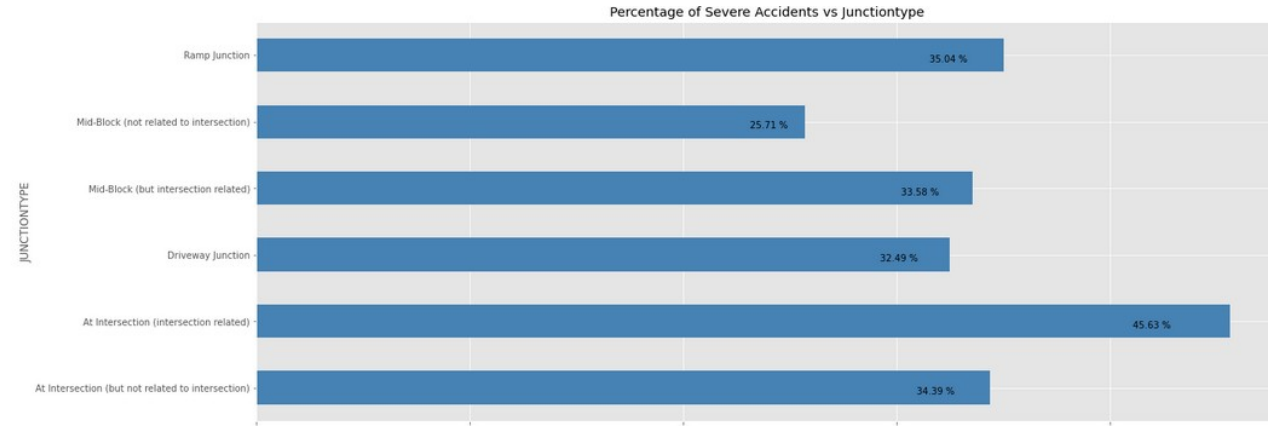

Figure 10: Percentage of severe accidents per junction type

Light Condition:
When regarding light condition it seems that people are more cautious when it is dark and there are no street lights, therefore fewer severe accidents are recorded. The most severe accidents appear to happen in the dark when lighting conditions are unknown (Figure 11)

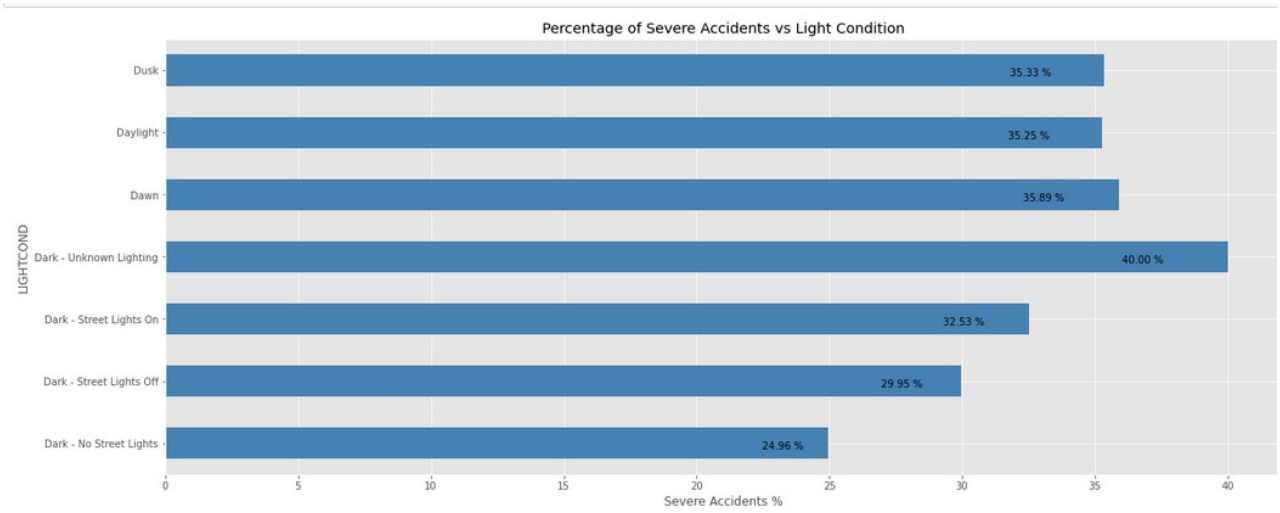

Figure 11: Influence of light condition

Speeding:
Recorded Accidents where speeding (Speeding = 1) is involved show a higher percentage of severe accidents, which makes SPEEDING a candidate for modeling (Figure 12).

```
SPEEDING  SEVERITYCODE
0.0       1              65.996860
          2              34.003140
1.0       1              58.538213
          2              41.461787
Name: SEVERITYCODE, dtype: float64
```

Figure 12

Number of involved Vehicles:
Most accidents involve two vehicles, roughly 72% of those are minor (Per-Inc1), but the ratio changes when looking at accidents only involving one vehicle (40% minor to 70% severe) and to 56% minor to 44% severe for three involved vehicles. This shows the dependency on the number of involved vehicles and makes VEHCOUNT a feature for modeling.

VEHCOUNT	Total_INC1	Total_INC2	Total_VC	Per_Inc2	Per_Inc1
0	3.0	222.0	225.0	98.666667	1.333333
1	10081.0	15081.0	25162.0	59.935617	40.064383
2	92371.0	35976.0	128347.0	28.030262	71.969738
3	7063.0	5520.0	12583.0	43.868712	56.131288
4	1278.0	1120.0	2398.0	46.705588	53.294412
5	264.0	271.0	535.0	50.654206	49.345794
6	83.0	64.0	147.0	43.537415	56.462585
7	22.0	25.0	47.0	53.191489	46.808511
8	11.0	7.0	18.0	38.888889	61.111111
9	3.0	7.0	10.0	70.000000	30.000000
10	0.0	2.0	2.0	100.000000	0.000000
11	3.0	2.0	5.0	40.000000	60.000000
12	1.0	0.0	1.0	0.000000	100.000000
13	0.0	1.0	1.0	100.000000	0.000000
14	0.0	1.0	1.0	100.000000	0.000000
15	0.0	1.0	1.0	100.000000	0.000000

Figure 13: influence of involved vehicles

Number of involved Persons:
The highest number of involved persons is 93, but only one such incident is recorded, so we only inspect the top 15 most accident-prone numbers. Depending on the number of persons the ratio in % between severe (Per_Inc2) to minor (Per_inc1) changes which makes PERSONCOUNT a feature for modeling.

PERSONCOUNT	Total_INC1	Total_INC2	Total_PC	Per_Inc2	Per_Inc1
2	65934.0	27884.0	93818.0	29.721375	70.278625
3	20697.0	13612.0	34309.0	39.674721	60.325279
4	7936.0	6323.0	14259.0	44.343923	55.656077
1	7438.0	3059.0	10497.0	29.141660	70.858340
5	3498.0	3030.0	6528.0	46.415441	53.584559
0	3406.0	1724.0	5130.0	33.606238	66.393762
6	1310.0	1389.0	2699.0	51.463505	48.536495
7	469.0	662.0	1131.0	58.532272	41.467728
8	240.0	287.0	527.0	54.459203	45.540797
9	81.0	132.0	213.0	61.971831	38.028169
10	55.0	76.0	131.0	58.015267	41.984733
11	21.0	35.0	56.0	62.500000	37.500000
12	13.0	21.0	34.0	61.764706	38.235294
13	8.0	13.0	21.0	61.904762	38.095238
14	11.0	9.0	20.0	45.000000	55.000000

Figure 14: Influence of involved persons

3.2 Predictive Modeling

For modeling five different types of supervised machine leaning algorithms will be tested:

- Decision Tree
- Logistic Regression
- Random Forest Classifier [1]
- SMOTE + Random Forest Classifier
- SMOTE + Logistic Regression
- Balanced Random Forest Classifier (BRFC) [3]

The feature data set is split into a training set (80%) and a test (20%) set. All models are trained with the training set and the prediction is made with the test set. For each algorithm the best parameters are determined by calculating the accuracy.

To deal with the imbalanced data Imblearn's SMOTE (Synthetic Minority Over-sampling Technique) algorithm [2] was used on the train data. The algorithm was used to enrich the minor class by using a k-nearest-neighbor technique to generate synthetic data.
Balanced Random Forest Classifier is specifically designed to deal with imbalanced data.

4. Results

Classification-Metrics:

Algorithm	**Jaccard**	**F1-score**	**Log Loss**
Decision Tree	0.5474478619002631	0.6845902075222968	NA
Logistic Regression (LR)	0.508996786771024	0.6424616253016807	0.6085578323148951
Random Forest Classifier (RFC)	0.5174818500608895	0.66963022803068310	NA
SMOTE + RFC	0.48970528530860175	0.6501701188984736	NA
SMOTE + LR	0.49867590040148363	0.6331256145583903	0.6119197084616842
BRFC	0.45507781723203866	0.6215009477114664	NA

5. Discussion

Regarding the performance metrics, the decision tree is the model which performed best with the highest values for Jaccard and F1-score, followed by the random forest classifier.

Trying to deal with the imbalance of the data using SMOTE for training, results in lower values for Jaccard and F1-score. But comparing the confusion matrix for RFC without SMOTE with the one with using SMOTE, we see a shift in the correct prediction of severe accidents from 44.55% to 52.89% (Figure 15). Using SMOTE with Logistic Regression does not have the same effect, the correct prediction of severe accidents even drops slightly from 22.8% to 22.31% (Figure 16).

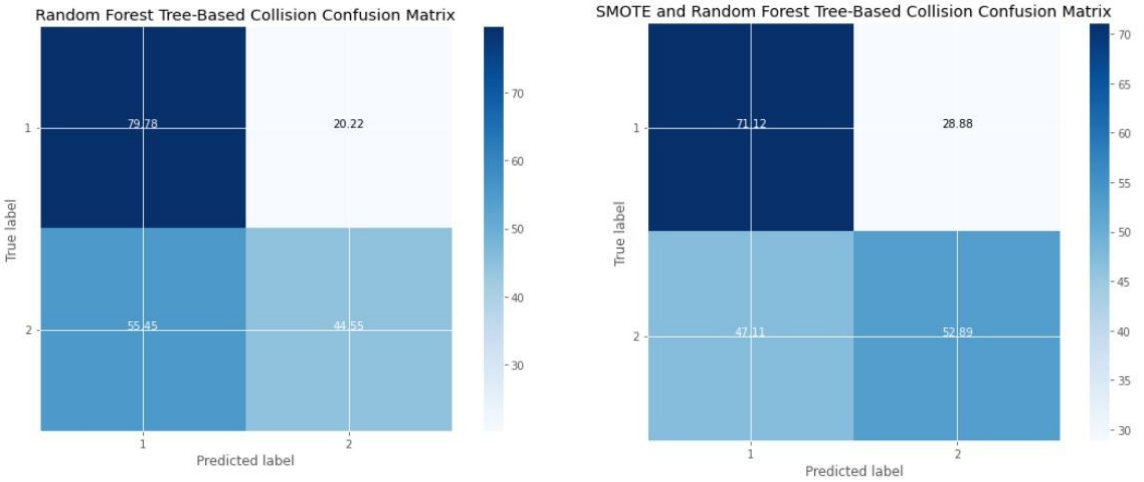

Figure 15: RFC Confusion matrix without (left) and with (right) use of SMOTE

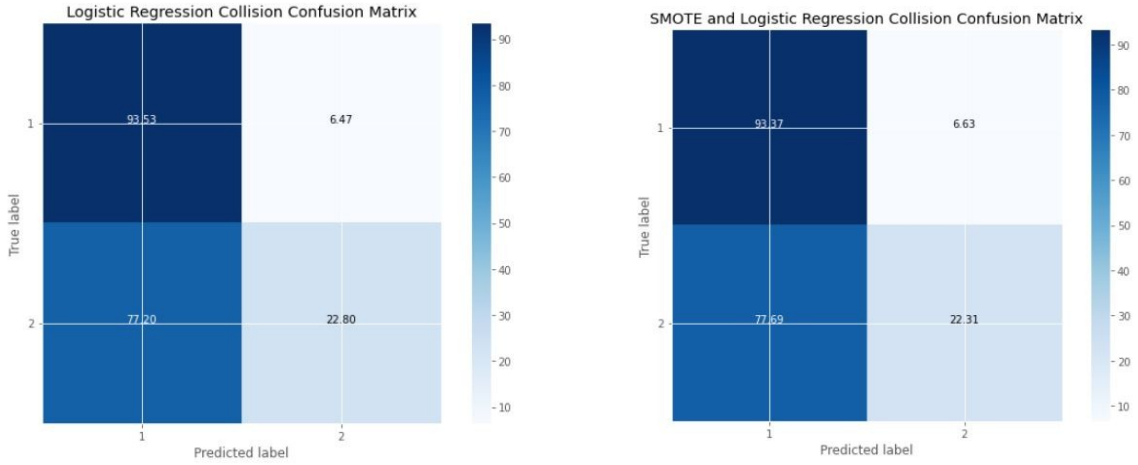

Figure 16: LR Confusion matrix without (left) and with (right) use of SMOTE

The Decision Tree, which performed best in terms of classification metrics, tends to falsely predict the majority class with 68.93% severe accidents being predicted as minor, but performance is better than both Logistic Regression and LR with use of SMOTE. In comparison the algorithm for imbalanced data BRFC predicts the minority class best with 60.69% true for severe accidents (Figure 17) which outperforms RFC and the use of oversampling with SMOTE.

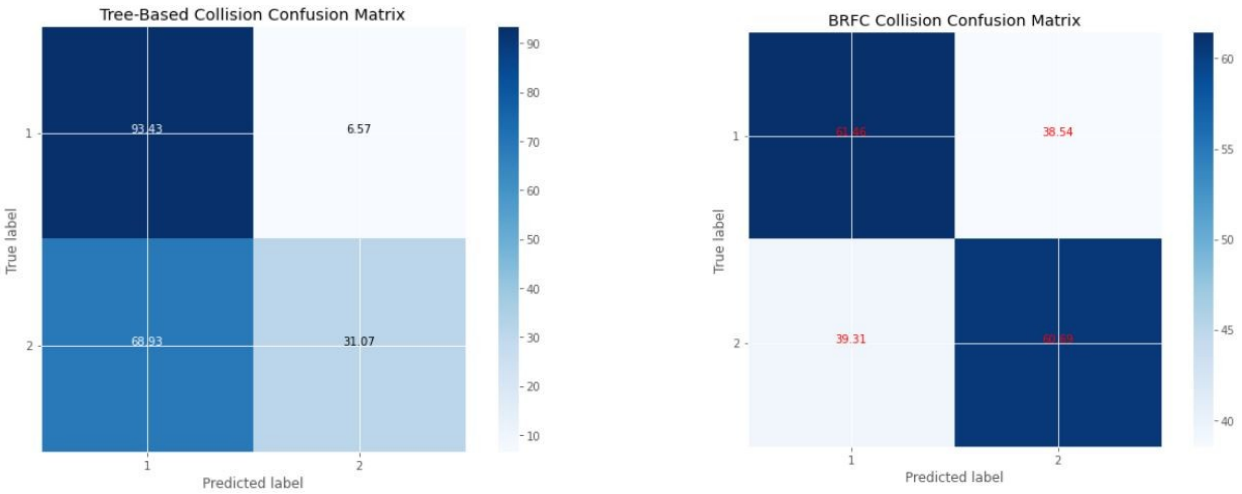

Figure 17 Confusion matrix Decision Tree (left) and BRFC (right)

6. Conclusions

Going back to the initial problem, machine leaning can help to predict the severity of accidents using historical data. Fortunately most of the recorded accidents are minor, but this is a drawback in terms of prediction with classification. With the oversampling technique SMOTE the prediction of the minor class (severe outcome) improves, but it depends on the algorithm used, best prediction of severe outcome of accidents were made with the algorithm specialized on imbalanced data Balanced Random Forest Classifier on this data and the chosen features. In context of the problem at hand it is arguably more important to correctly predict the severe accidents than to have a good prediction of the minor accidents using the other algorithms.

Looking on the practical use, since values like number of involved persons and vehicles were needed as features to make a prediction, you would need this information before the accident occurs to “feed” the model. Thinking about rescue stations planning human resources beforehand or radio stations making forecasts, this is not likely. But using prediction to evaluate future costs seems possible.

Since no influence of day of week or month was detected, it may be a good idea to look on the time of accident occurrence in future studies.

7. References

[1]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[2]https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html
[3]https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.ensemble.BalancedRandomForestClassifier.html