# Applied Data Science Capstone: Predicting car accident severity

## I. Introduction

*Background*: The demand for vehicles rises consistently. Consequently, so does the number of vehicles on the road and the probability of involvement in traffic jams or car accidents. Traffic accidents result not only in loss of human lives but have a huge impact on economy as well. According to the Michigan Traffic Crash Decade-At-A-Glance study[1], there were over 314.000 traffic accidents in US in 2017, which translate into costs around 230 billion dollars per year. Approximately 1.3 million people die yearly because of road traffic crashes.[2]

*Problem / Objective*: By collecting and analysing relevant data on car accidents, we aim to establish to what extent can:

1) weather conditions
2) the total number of people involved in the collision or the number pedestrians
3) road conditions or light conditions
4) driving under the influence of drugs or alcohol (DUI)

help us predict car accident severity.

*Interest*: Car drivers and vulnerable road users (pedestrians, cyclists, motorcyclists) would or should be interested in an equal manner by the result of this inquiry. Having measured the impact of the abovementioned outer and inner conditions and knowing the results, one might act differently. Having established the (probable) paramount influence of drugs or alcohol in causing car accidents of great severity, authorities might decide to impose higher taxes on such vices.

## II. Data

### a. Data sources:

In order to answer the proposed questions, we shall use data from the SDOT Traffic Management Division, Traffic Records Group.

The database is updated weekly since 2004 and contains all types of collisions.

Here is the provided link to database: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The metadata is to be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

The database contains 37 attributes, including the severity code, the number of persons involved in the car accident, the number of pedestrians, road and light conditions, information on driving under the influence of drugs or alcohol.

### b. Data cleaning and feature selection

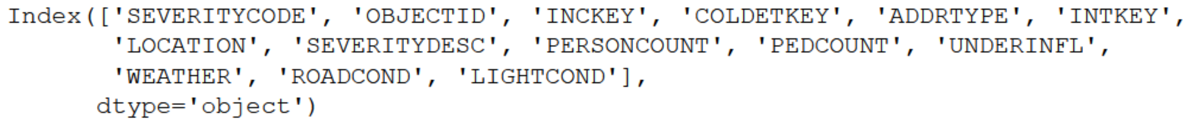
After reading the abovementioned file in a data frame, we concluded it had 194673 rows and 38 attributes. Considering our objective, we dropped the following attributes: X, Y, REPORTNO, EXCEPTRSNCODE, EXCEPTRSNDESC, INCDATE, INCDTTM, SPEEDING, ST_COLDESC,

[1] Michigan State Police, Michigan Traffic Crash Decade-At-A-Glance, 2018.

[2] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

HITPARKEDCAR, STATUS, SEVERITYCODE.1, COLLISIONTYPE, PEDCYLCOUNT, VEHCOUNT, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, INATTENTIONIND, PEDROWNOTGRNT, SDOTCOLNUM, ST_COLCODE, SEGLANEKEY, CROSSWALKKEY.

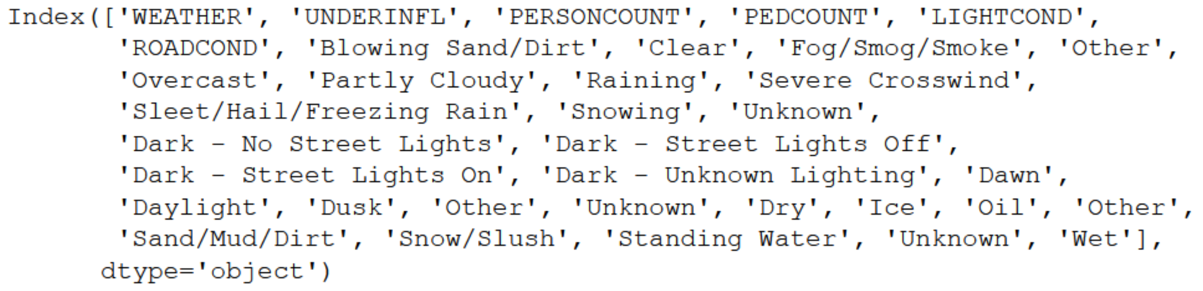
SEVERITYCODE.1 was dropped especially on the grounds that was identical with our target variable SEVERITYCODE. Additionally, we dropped all duplicate records (drop_duplicates) and all missing values (dropna). Afterwards, there were 63462 records and 14 attributes remaining, including the following columns:



```
Index(['SEVERITYCODE', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'ADDRTYPE', 'INTKEY',
       'LOCATION', 'SEVERITYDESC', 'PERSONCOUNT', 'PEDCOUNT', 'UNDERINFL',
       'WEATHER', 'ROADCOND', 'LIGHTCOND'],
      dtype='object')
```

Subsequently, we grouped the available data by weather, road conditions, light conditions, number of persons, number of pedestrians, whether drivers were under the influence of drugs and alcohol, and severity code of the car accident respectively in order to get a general impression of what the data consists of. There are 36088 cases of severity code 1 and 27374 case of severity code 2. Bearing in mind that the provided metadata listed a total 5 distinct codes for severity, one might conclude that the provided data set is unbalanced. Nevertheless, we intend to build a model (see section III.iii. Machine learning models) which predicts based on the 6 attributes if a particular accident is either in the property damage-only collision category (1) or in the injury collision category (2).

The 6 desired attributes were also saved in a distinct dataframe called ‘featureset’. Weather, light conditions and road conditions were one-hot coded for the subsequent modelling. This operation results in the following attributes in ‘featureset’:



```
Index(['WEATHER', 'UNDERINFL', 'PERSONCOUNT', 'PEDCOUNT', 'LIGHTCOND',
       'ROADCOND', 'Blowing Sand/Dirt', 'Clear', 'Fog/Smog/Smoke', 'Other',
       'Overcast', 'Partly Cloudy', 'Raining', 'Severe Crosswind',
       'Sleet/Hail/Freezing Rain', 'Snowing', 'Unknown',
       'Dark - No Street Lights', 'Dark - Street Lights Off',
       'Dark - Street Lights On', 'Dark - Unknown Lighting', 'Dawn',
       'Daylight', 'Dusk', 'Other', 'Unknown', 'Dry', 'Ice', 'Oil', 'Other',
       'Sand/Mud/Dirt', 'Snow/Slush', 'Standing Water', 'Unknown', 'Wet'],
      dtype='object')
```

## III. Methodology

### i. Exploratory data analysis

Our target variable is the severity code of the car accidents.

For the columns *weather conditions* (Fig. 1), *light conditions* (Fig. 2), *and road conditions* (Fig. 3) three separate bar charts were created. For *number of persons* (Fig. 4) *and number of pedestrians* (Fig. 5) respectively we used scatter plots. In order to visualize the variable ‘*UNDERINFL’* we used a form of histogram (Fig. 6), taking into account the severity of the car accidents also. In the following pages one can explore the figures.

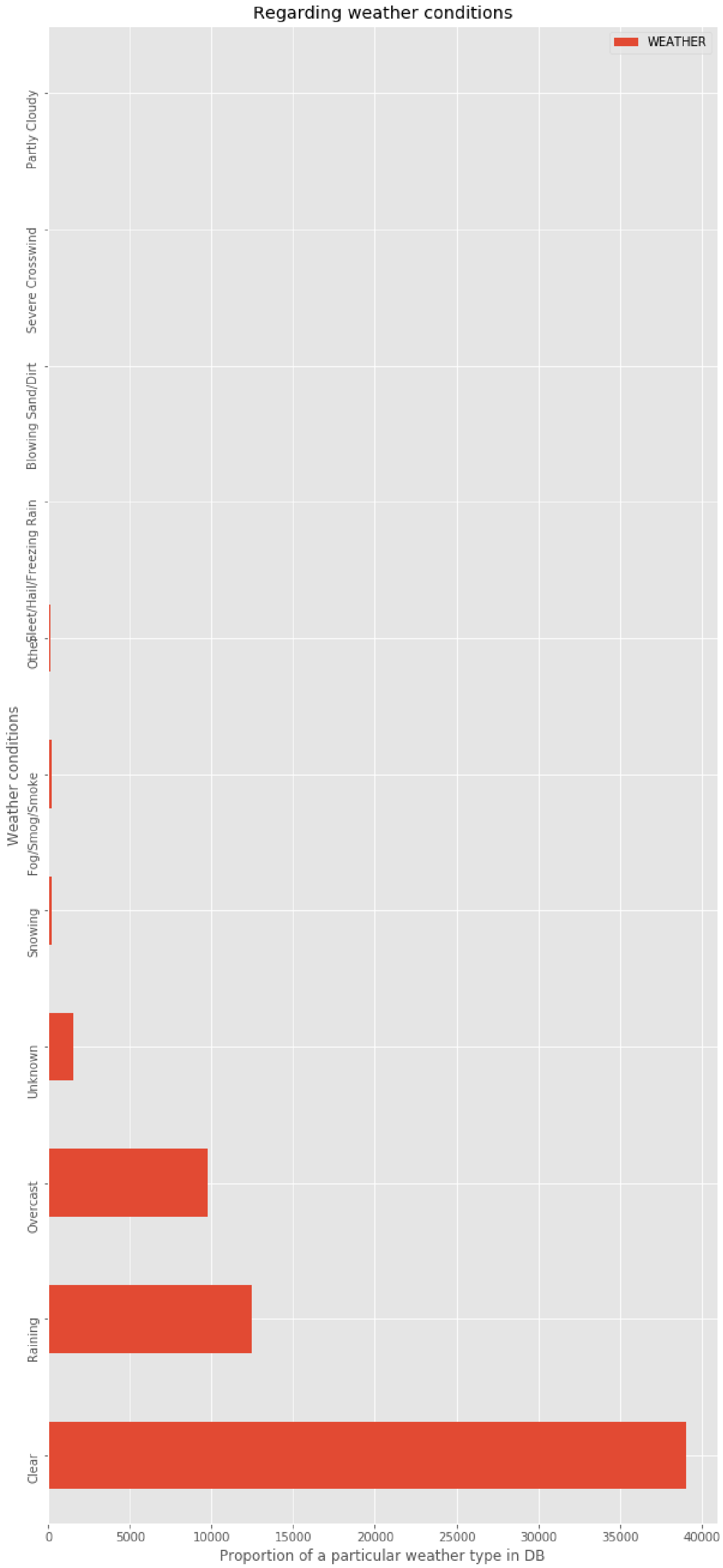


**Fig. 1. Weather conditions among car accidents data.**

Clear, raining, and overcast weather types are the most common among the available data.

(DB = database)

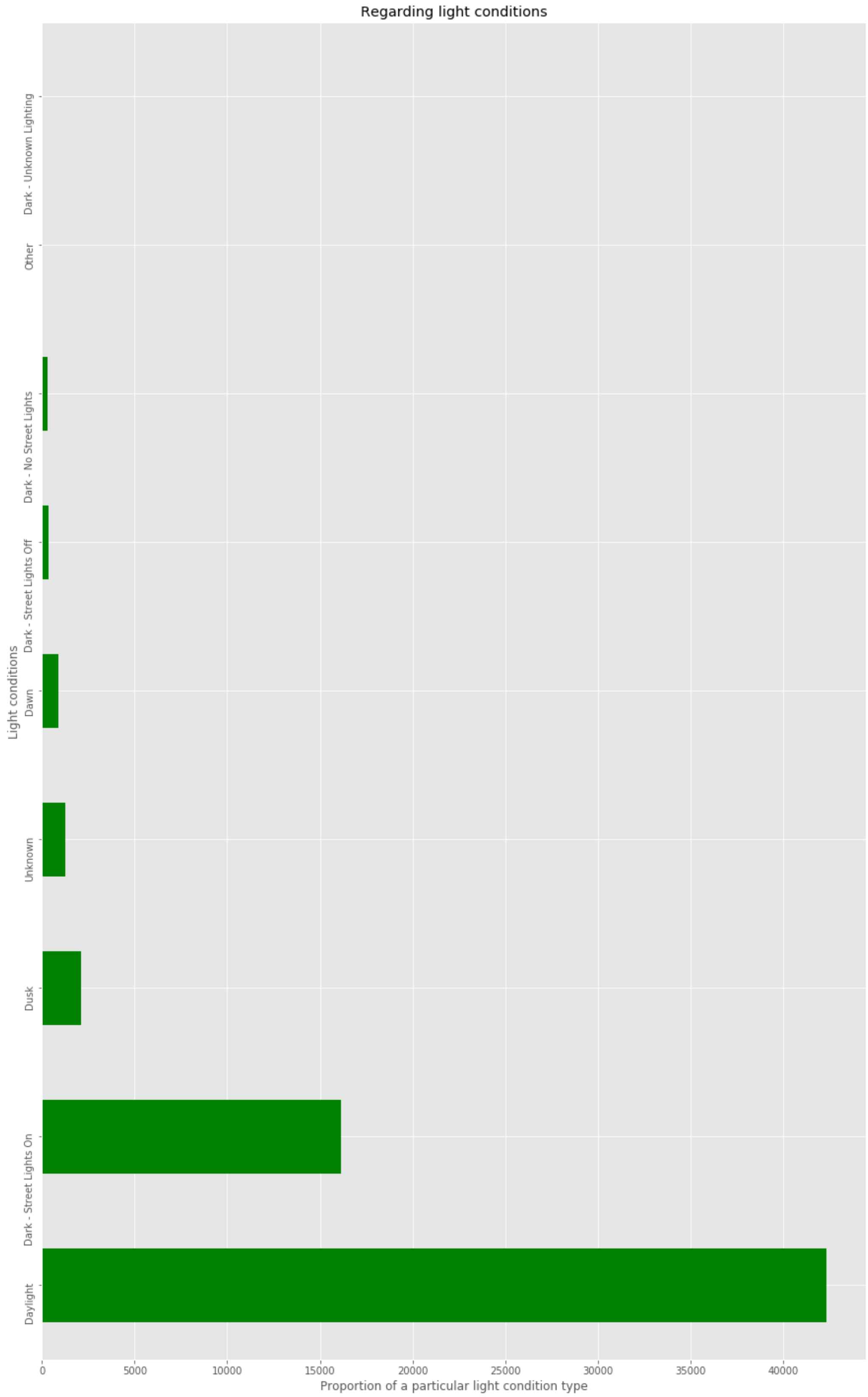


**Fig. 2. Light conditions among car accident data.**
The most common light conditions are: daylight, dark with street lights on, and dusk.

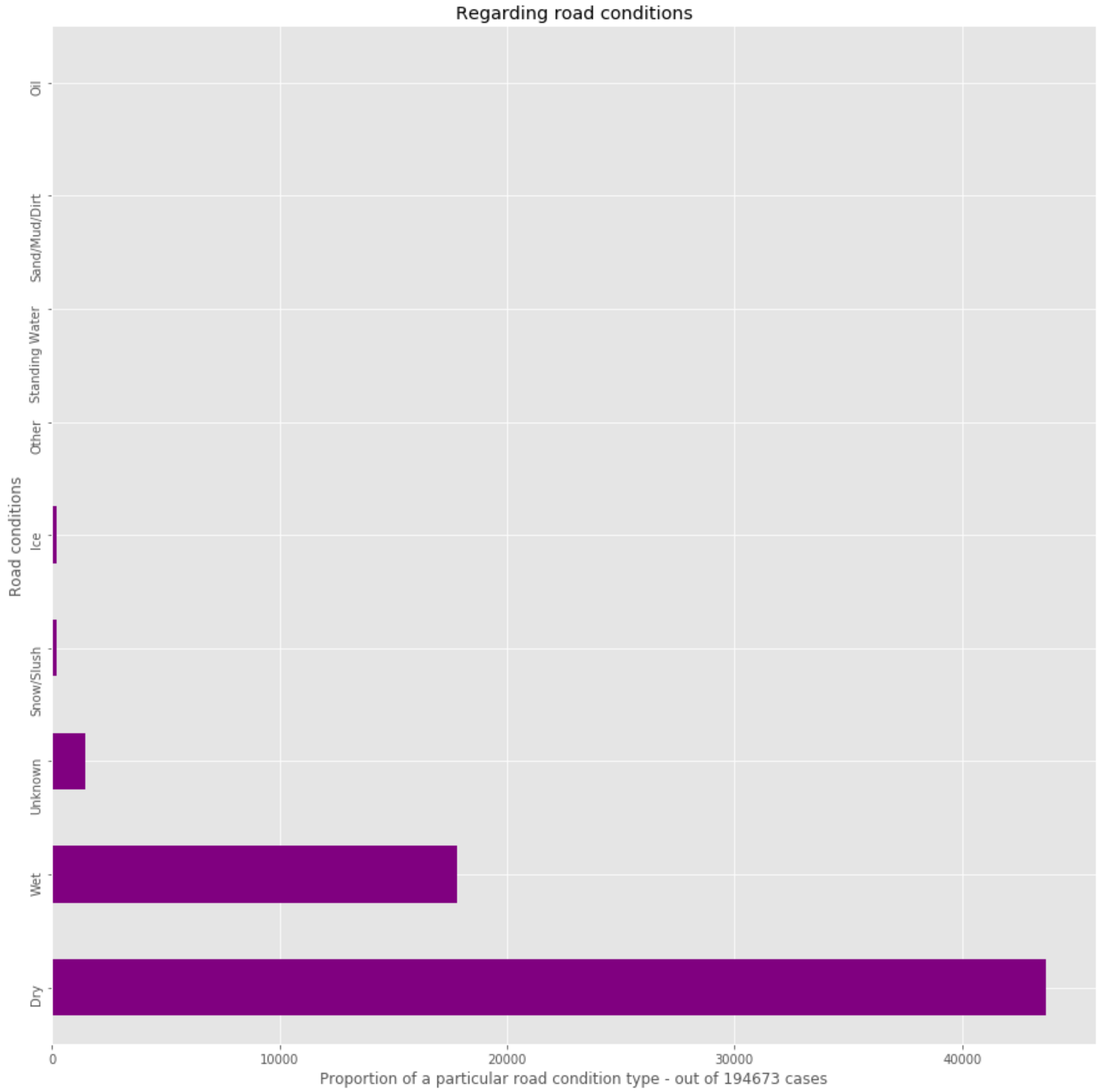


**Fig. 3. Road conditions among car accident data.**

Dry, wet, and unknown road conditions are the most common.

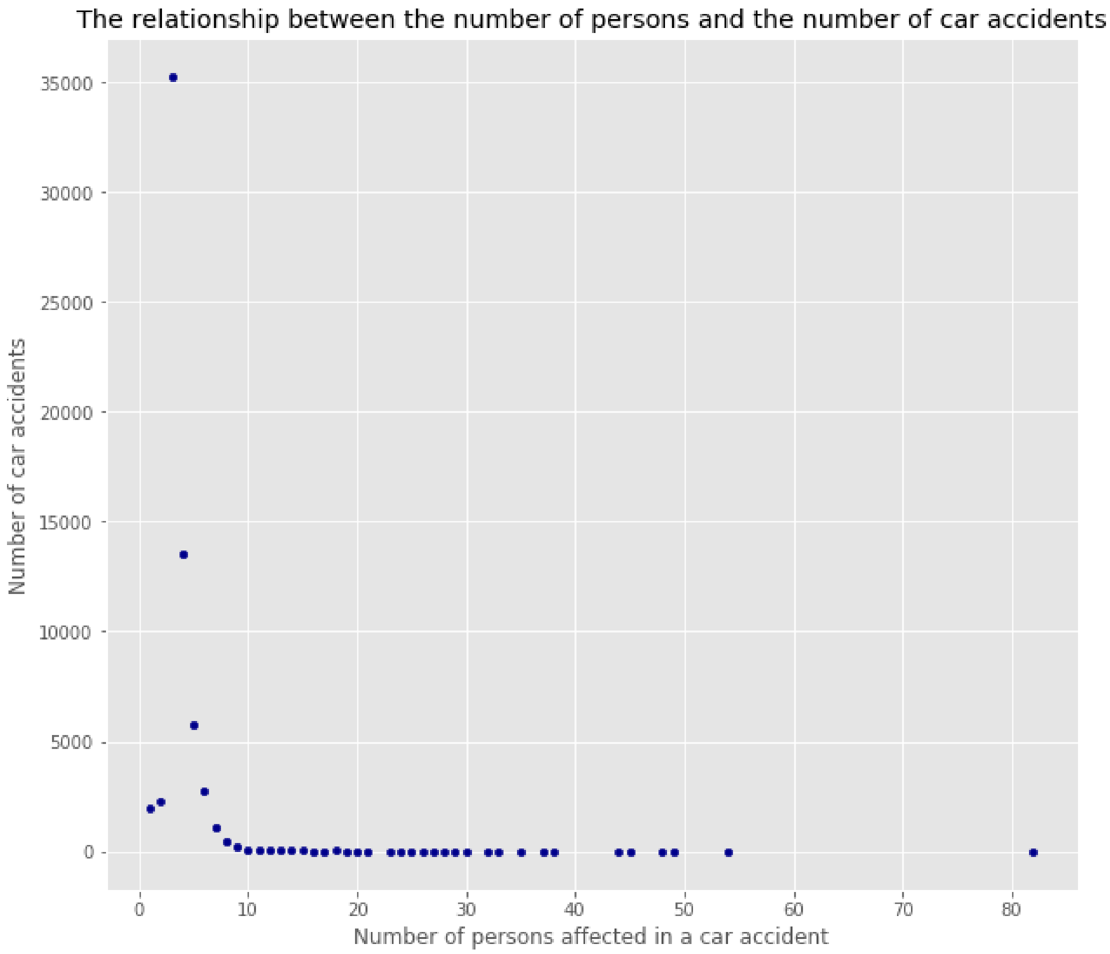


**Fig. 4. The relationship between the number of persons involved and the number of car accidents.**

The majority of cases involve under 10 persons.

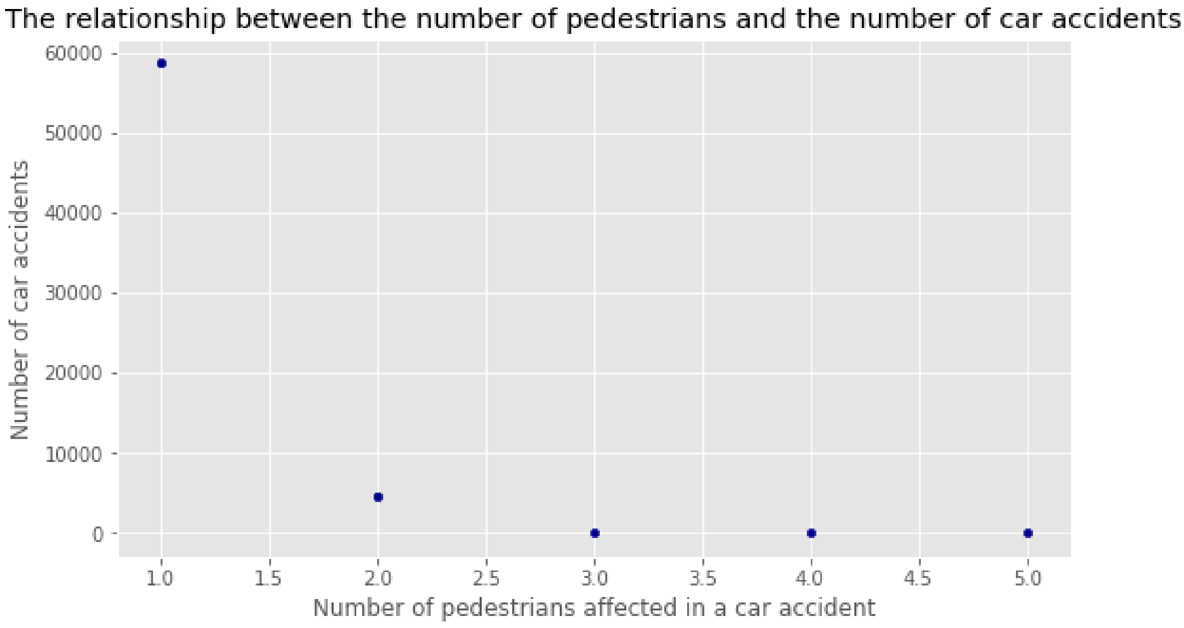


**Fig. 5. The relationship between the number of pedestrians and the number of car accidents.**

One can tell the most the accidents involve one pedestrian.

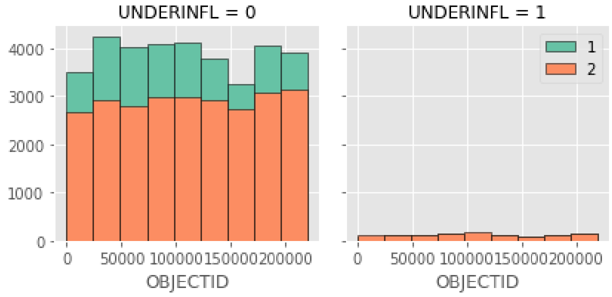


**Fig. 6. Representation of cases involving driving under the influence of drugs or alcohol.** With green are represented the property damage cases (1) and with orange the injury cases (2). Additionally, one notices that most cases do not involve DUI. In fact, 3.15% are DUI-caused car accidents (2231 cases).

### ii. Inferential statistical testing

The attributes we took into consideration are mostly categorical (apart from number of pedestrians and number of persons). Moreover, the distributions of the variables *number of pedestrians* and *number of persons* are presumably non-normal. This is also to be confirmed by the fact that most cases lie at the beginning of the spectrum (maximum 10 people involved) – see the scatter plots in Fig. 4 and Fig. 5.

Because of that, we assessed the correlation between some aspects of the attributes and the severity code of the car accident using the Spearman correlation from the module scipy.stats.

### iii. Machine learning models

In this project, the support vector machine (SVM) and logistic regression (LR) were used to predict the severity of car accidents. Afterwards, the performance of the models was assessed with Jaccard's index and F1-score for the support vector machine, and with log loss, Jaccard's index and F1-score for logistic regression. Additionally, a confusion matrix was computed for SVM.

The feature set was transferred in the variable X, which was afterwards pre-processed. Using train_test_split from sklearn.model_selection we trained and tested the model. Finally, we predicted values for severity code in yhat and then compared those values with the real ones saved in the y variable.

## IV. Results

### i. Inferential statistical testing

The following table (Table 1) illustrates the Spearman coefficient, the p value for each tested attribute, as well the interpretation:

| **Attribute** (in correlation with severity code) | **Spearman coefficient** | **p value** | **Interpretation** |
|---|---|---|---|
| **DUI** | 0.030 | 1.468 | not correlated; not statistically significant |
| **No. of persons** | 0.079 | 6.383 | not correlated; not statistically significant |
| **No. of pedestrians** | 0.270 | 0.0 | slightly correlated; probably statistically significant |
| **Raining (weather condition)** | 0.006 | 0.103 | not correlated; not statistically significant |
| **Dark - Street Lights Off (light conditions)** | -0.0002 | 0.952 | not correlated; not statistically significant |
| **Standing water (road conditions)** | -0.004 | 0.235 | not correlated; not statistically significant |
| **Daylight (weather condition)** | 0.017 | 1.332 | not correlated; not statistically significant |
| **Sand/Mud/Dirt (road conditions)** | -0.001 | 0.733 | not correlated; not statistically significant |

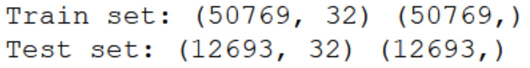
**Table 1. Inferential statistical testing: Spearman coefficients and p values for a series of selected attributes.**

The set level of statistical significance was $p<0.05$. A Spearman coefficient of 1 or -1 (or close to those values) denote a correlation.

Among the tested variables, only number of pedestrians was slightly correlated and probably of statistical significance (the obtained value for p was 0.0, with no further decimals).

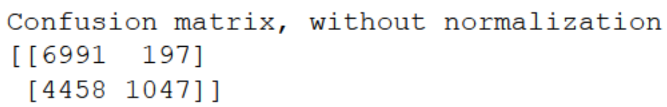
### ii. Machine learning models

The train set and the test set had the following shapes (test_size was set to 2, and random_state to 4):

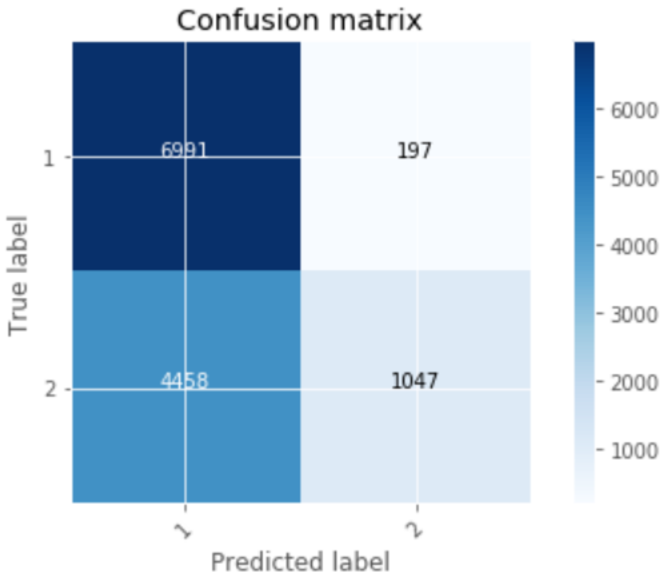


```
Train set: (50769, 32) (50769,)
Test set: (12693, 32) (12693,)
```

The support vector machine produced the following confusion matrix (without normalization):

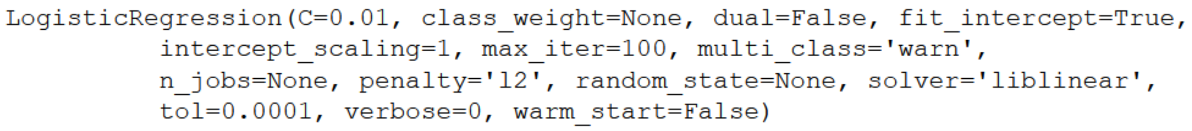


```
Confusion matrix, without normalization
[[6991  197]
 [4458 1047]]
```



The F1-score for SVM was 0.5594158617301305 and the Jaccard's score was 0.6332624281099819. Considering that the ideal value for these accuracy measures is around 1, one might interpret the built model as not perfectly accurate.

The model for logistic regression had the following characteristics:



```
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='liblinear',
          tol=0.0001, verbose=0, warm_start=False)
```

The accuracy indicators for logistic regression are:
Log Loss = 0.6326192709245222
F1-score = 0.5658557811404462
Jaccard's index = 0.6339714803434964

The values for F1-score and Jaccard's index are comparable with those obtained through support vector machine. Log Loss of 0.632 in a case of binary classification (severity code of 1 or 2) showcases a considerable level of uncertainty or entropy of the model.

## V. Discussion

Many studies tackled the issue of traffic fatalities. This study[3] by Gu et. al. also provides a literature review on the matter. One of the important points made are that SVM is more accurate than traditional negative binomial models in predicting motor vehicle collisions. However, the chosen parameters are of utmost importance since they influence the model's performance. The study used Particle Swarm Optimization (PSO-SVM) and proves that such a model has a higher prediction precision and smaller

[3] Gu, X., Li, T., Wang, Y., Zhang, L., Wang, Y., & Yao, J. (2018). Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization. Journal of Algorithms & Computational Technology, 20–29. https://doi.org/10.1177/1748301817729953

errors in comparison with backpropagation neural network, Bayesian network, and K Nearest Neighbour.

Our project deals with an unbalanced set of data (judging by the aspect of the plots and the reduced types of severity codes – just two) and attempts to build a machine learning model using support vector machine and logistic regression. We limited the number of attributes to 6 for the feature set; some of these attributes were subsequently one-hot coded. Thus, we can formulate the following recommendations for future studies or projects: choosing or providing a balanced dataset, increasing the number of attributes in the future set. In the latter case, the issue of overfitting needs to be considered, as well. The resulted model should have better accuracy parameters than those we listed in the Results section of this report.

## VI. Conclusion

All things considered, our project highlighted the fact that the number of involved pedestrians is slightly correlated with car accidents' severity code; that in the large majority of cases the drivers were not driving under the influence of drugs or alcohol; that most car accidents seem to occur having clear weather and daylight. The last aspect emphasizes the need to built performant and accurate machine learning models, taking the abovementioned recommendations into account (Section V).

The stakeholders (traffic participants, traffic police) should be aware of all these facts in order to reduce traffic fatalities.