Coursera_Capstone

Models to Predict Severity of a Traffic Accident (STA) based on Weather, Road and Light Conditions

Applied Data Science Final Project
K. Giraldo, 2020

Introduction - Business Problem

Traffic accidents are, in modern times, a great cause of health issues for many developed and under-developed countries with the latest one having the highest rates in the world and usually producing high numbers of deaths[1,2] and/or Property Damage.

Many papers have been published during the last years predicting traffic accidents severity, however, most of them come from datasets from developed countries. Available information from countries in development is limited, incomplete or segmented [3] where weather or road conditions are not taken fully into account[4] with other factors being considered as more important[4].

This work pretends to create Supervised Machine Learning Classification Models from Vehicle Collision Dataset from the City of Seattle as an example and initial starting point to show the importance of Weather and Road Conditions to estimate the STA in other regions.

The dataset to be used contains a Labeled attribute, Severity, describing the severity of an accident such as Injuries and Property Damage for the City of Seattle between the years of 2005 to 2020, however, itt doesn't contain any other information about severity such as deaths or level of Injury or damage. It includes attributes such as number of people involved, type of vehicles, type of road, date and time of incident, Weather, Road and Light conditions. The total number of records are approximately 200,000 that need to be cleaned and prepared for the models.

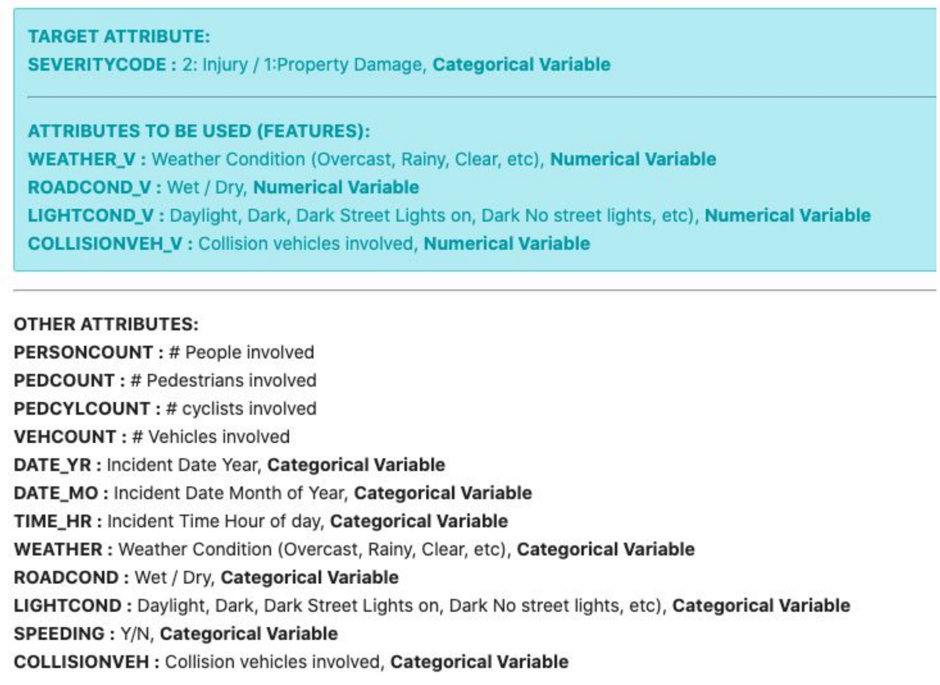
K-Nearest Neighbors, Logistic Regression, Decision Trees and Support Vector Machines will be used as classification algorithms to create the models. An effort to balance the dataset will be done.

Evaluation methods such as Jaccard Index and F1-Score will be used to assess the models.

All the information shown in this report is also available in the github page for the project that you can find here

Data Preparation

Data to be used is that proposed by the course. It is a csv file containing information about Traffic Accidents in the City of Seattle. It contains 36 attributes plus one Labeled Target named SEVERITYDESC, describing the severity of the accident. Some of the labels are not used because they are not of the interest of this work. After dropping not-used labels and removing NaN records, I finished with the next list of Attributes:



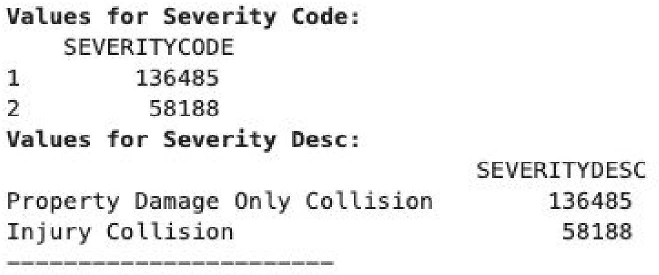
TARGET ATTRIBUTE:
SEVERITYCODE : 2: Injury / 1:Property Damage, **Categorical Variable**

ATTRIBUTES TO BE USED (FEATURES):
WEATHER_V : Weather Condition (Overcast, Rainy, Clear, etc), **Numerical Variable**
ROADCOND_V : Wet / Dry, **Numerical Variable**
LIGHTCOND_V : Daylight, Dark, Dark Street Lights on, Dark No street lights, etc), **Numerical Variable**
COLLISIONVEH_V : Collision vehicles involved, **Numerical Variable**

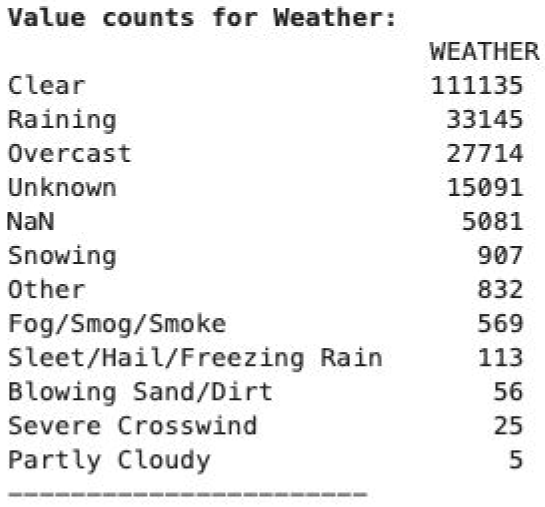
OTHER ATTRIBUTES:
PERSONCOUNT : # People involved
PEDCOUNT : # Pedestrians involved
PEDCYLCOUNT : # cyclists involved
VEHCOUNT : # Vehicles involved
DATE_YR : Incident Date Year, **Categorical Variable**
DATE_MO : Incident Date Month of Year, **Categorical Variable**
TIME_HR : Incident Time Hour of day, **Categorical Variable**
WEATHER : Weather Condition (Overcast, Rainy, Clear, etc), **Categorical Variable**
ROADCOND : Wet / Dry, **Categorical Variable**
LIGHTCOND : Daylight, Dark, Dark Street Lights on, Dark No street lights, etc), **Categorical Variable**
SPEEDING : Y/N, **Categorical Variable**
COLLISIONVEH : Collision vehicles involved, **Categorical Variable**

Other attributes such as X,Y and those listed above are used only for information. This work is focused only in the Attributes shown in Blue, corresponding to the values of the categorical variables WEATHER, ROADCONDitions and LIGHTCONDitions.

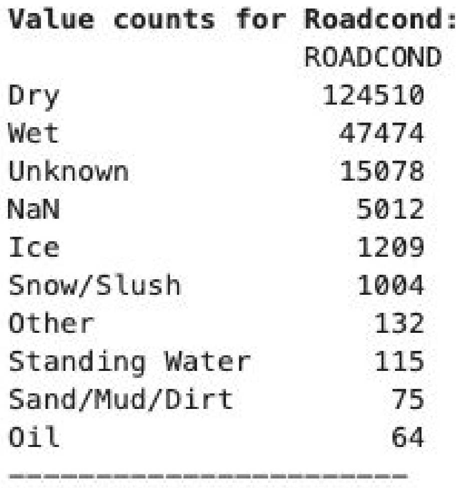
During the data preparation Frequency tables were created to understand better the attributes. Below are some of the frequency tables for the main features to be used.



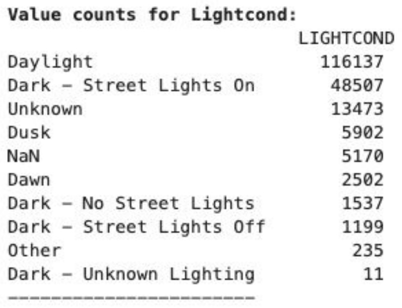
```
Values for Severity Code:
    SEVERITYCODE
1        136485
2         58188
Values for Severity Desc:
                                 SEVERITYDESC
Property Damage Only Collision        136485
Injury Collision                       58188
-----------------------
```



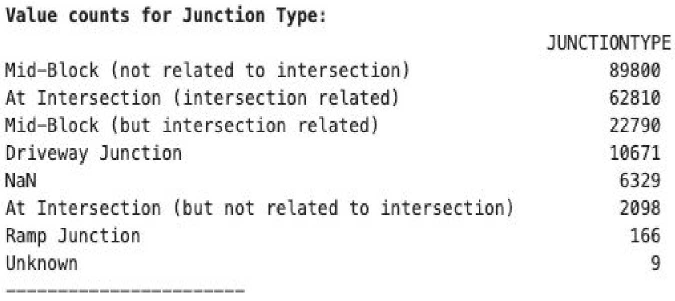
```
Value counts for Weather:
                            WEATHER
Clear                        111135
Raining                       33145
Overcast                      27714
Unknown                       15091
NaN                            5081
Snowing                         907
Other                           832
Fog/Smog/Smoke                  569
Sleet/Hail/Freezing Rain        113
Blowing Sand/Dirt                56
Severe Crosswind                 25
Partly Cloudy                     5
-----------------------
```



```
Value counts for Roadcond:
                   ROADCOND
Dry                 124510
Wet                  47474
Unknown              15078
NaN                   5012
Ice                   1209
Snow/Slush            1004
Other                  132
Standing Water         115
Sand/Mud/Dirt           75
Oil                     64
-----------------------
```



```
Value counts for Lightcond:
                          LIGHTCOND
Daylight                    116137
Dark - Street Lights On      48507
Unknown                      13473
Dusk                          5902
NaN                           5170
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Other                          235
Dark - Unknown Lighting         11
-----------------------
```



```
Value counts for Junction Type:
                                                    JUNCTIONTYPE
Mid-Block (not related to intersection)                    89800
At Intersection (intersection related)                     62810
Mid-Block (but intersection related)                       22790
Driveway Junction                                          10671
NaN                                                         6329
At Intersection (but not related to intersection)          2098
Ramp Junction                                               166
Unknown                                                       9
-----------------------
```

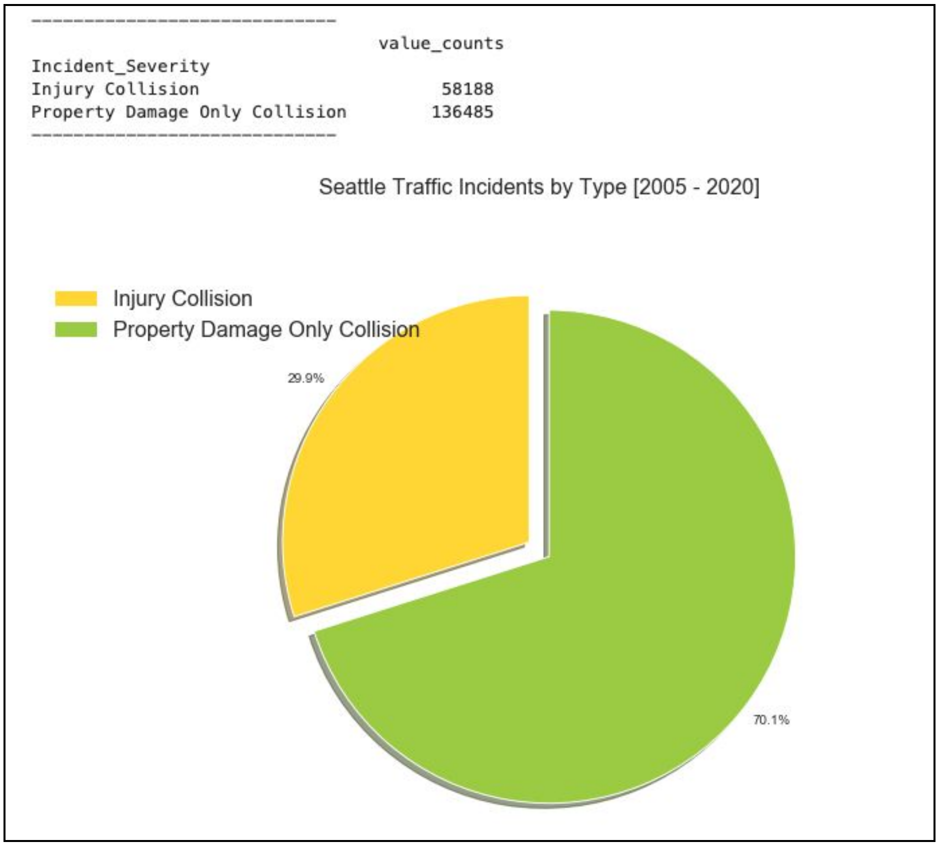
Attributes like Junction Type, Nr. People Involved or Speeding, of vital importance in a Traffic Accident Study, were not used here because the main objective is to find a relation of the Incident with the Weather Conditions.

Values in the Attributes such as NaN, Unknown, Other, were removed from the dataset. Later, final attributes will be shown in Data Exploration. Attributes such as time and date were planned to be used as an indication of the season or the natural light, however, at the end they are planned to be used for a future and deeper analysis.

Exploratory Data Analysis

PLOTS ANALYSIS

The dataset is focused in the city of Seattle. Pie Chart in figure 1 shows the Severity of the Accident. SEVERITYCODE/DESC represents a Categorical variable with only two possible values: **Injury and Property Damage** caused by the collisions. We can see that many of the reported cases are only Property Damage while Injury is only a third part of the reported cases.



```
------------------------------
                                   value_counts
Incident_Severity
Injury Collision                          58188
Property Damage Only Collision           136485
------------------------------
```

Figure 1. Severity of the Accident

Figure 2. Shows the Spatial Distribution of a 200 sample taken from the dataset. This shows that data is mainly in the city, making this dataset only appropriate for cities. We can note from the map that there are different environments such as Intersections, Highways, Residential areas, etc. Here we don't consider this information.

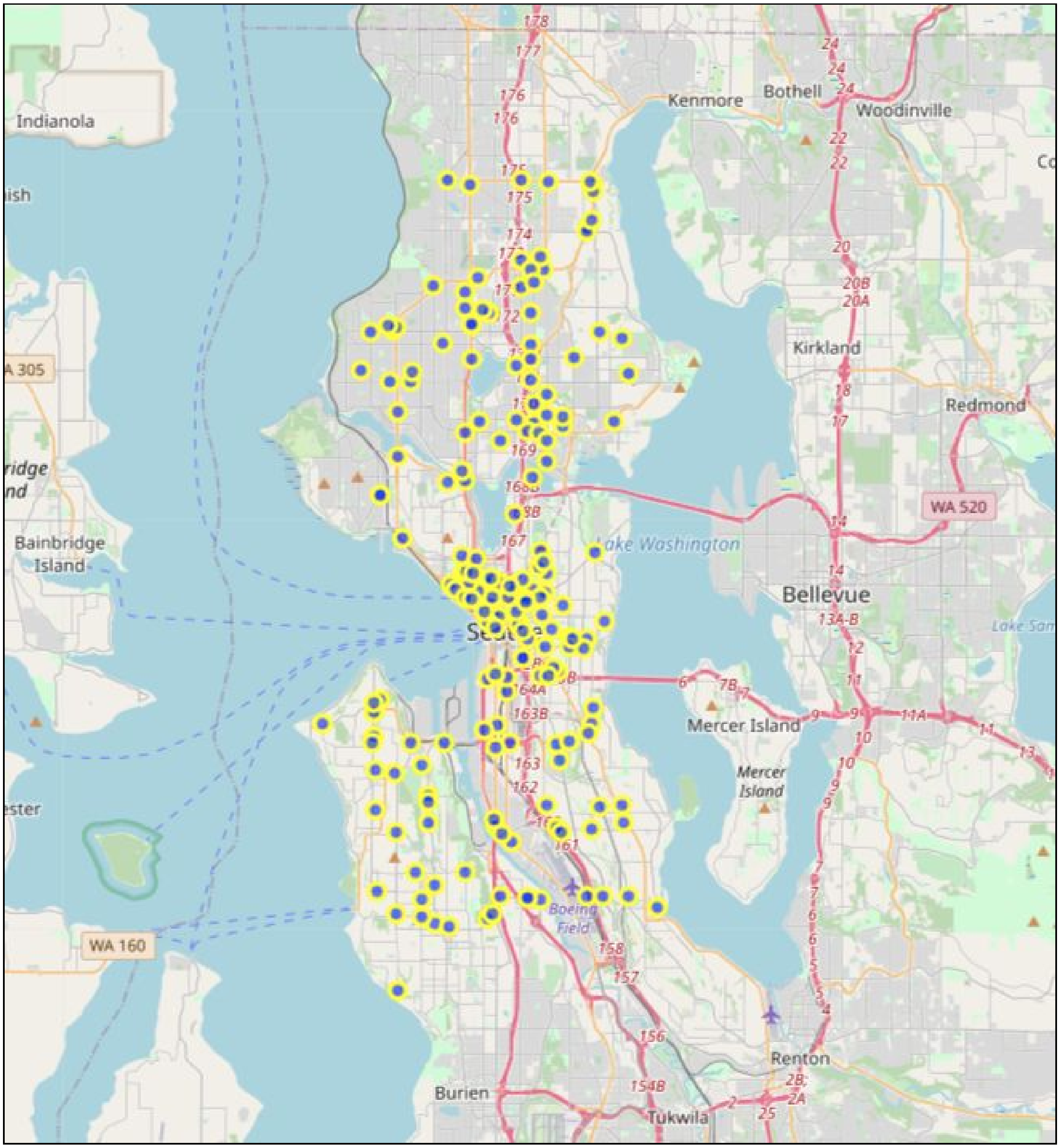
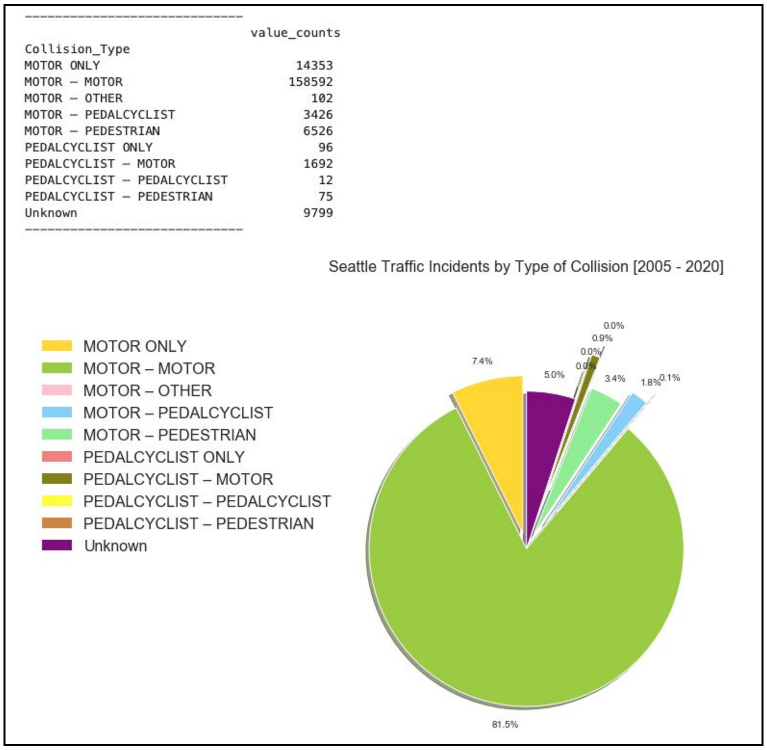


Figure 2. Incident Locations

There are also different types of incidents that are shown below. **Motor - Motor and Motor-Only** are dominating. Some others are also very important but not dominating the samples. We don't consider this type of feature in the analysis.



```
-----------------------------
                              value_counts
Collision_Type
MOTOR ONLY                           14353
MOTOR – MOTOR                       158592
MOTOR – OTHER                          102
MOTOR – PEDALCYCLIST                  3426
MOTOR – PEDESTRIAN                    6526
PEDALCYCLIST ONLY                       96
PEDALCYCLIST – MOTOR                  1692
PEDALCYCLIST – PEDALCYCLIST             12
PEDALCYCLIST – PEDESTRIAN               75
Unknown                               9799
-----------------------------
```

Figure 3. Kind of Incident

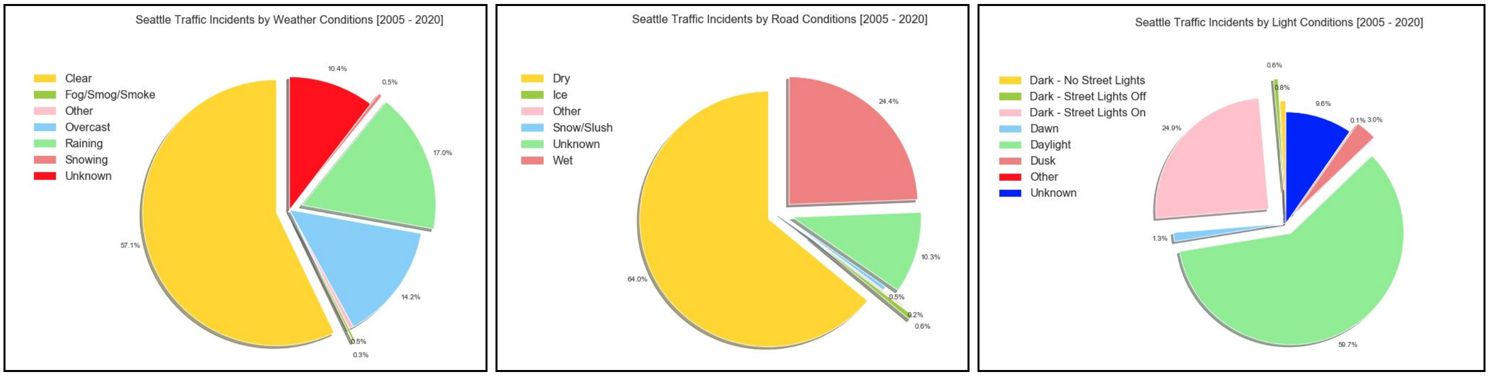
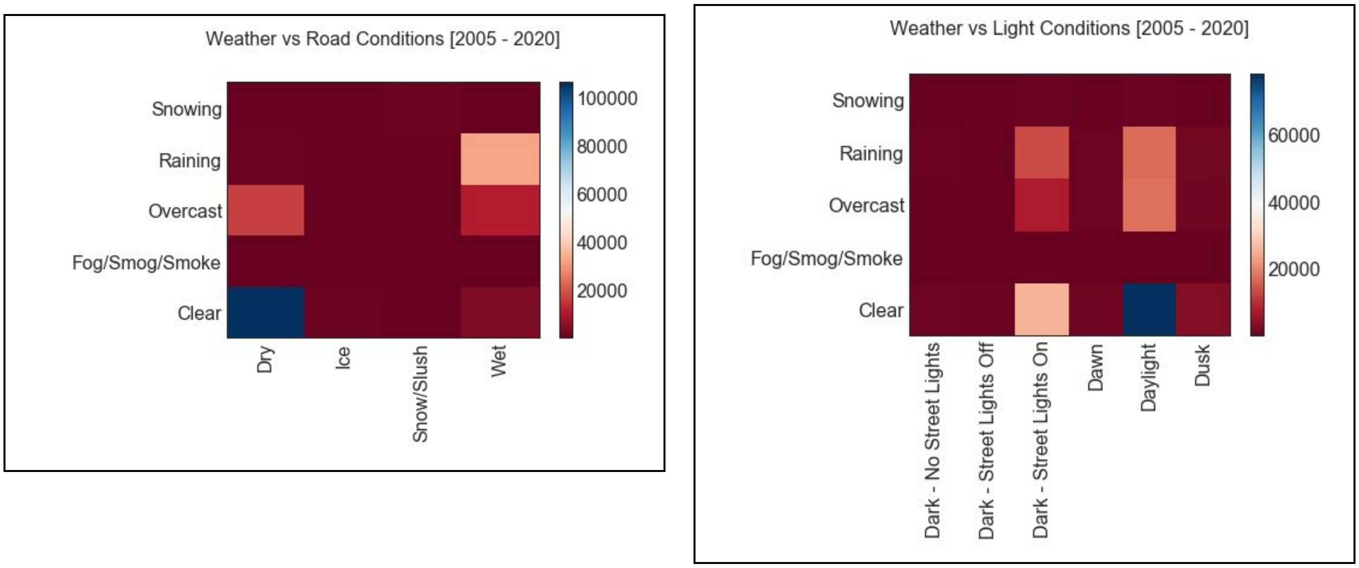


Fig. 4. (a) Weather Distr. (b) Road Conditions Distr. (c) Light Conditions Distr.

As we can see the main characteristics for the incidents are related to Clear, Dry and Daylight conditions, however, it could be considered as a bias since these are the most predominant conditions during the year. One of the first analyses is that the results could be biased by these conditions since they represent, in all the cases, more than 50% of the cases with other cases representing a very low portion of the samples.

Pivot tables are also biased by these conditions as we follows:

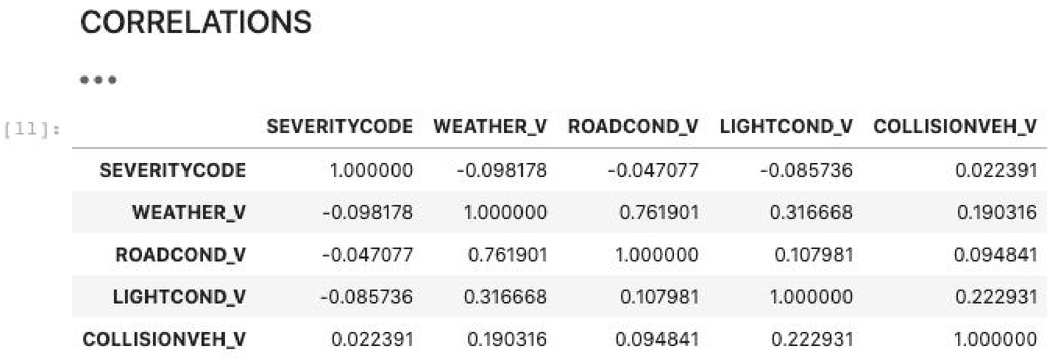


(a) (b)

Figure 5. Cross-Correlated Plots for Weather vs Road Conditions (a); and Weather vs Light Conditions. We can note that obvious results are the high correlation between Clear and Dry, Raining and Wet, Clear and Daylight, Clear and Dark but with Lights on.

Other non obvious correlations such as Overcast and Dry or Wet, However many other features are overshadowed by the high number of obvious results. This is also a very interesting situation for this dataset where unbalancing needs to be considered, however, outside of the scope of this work.

Cross correlations between the attributes are shown below. Note that mostly all of them show a very low correlation.



CORRELATIONS

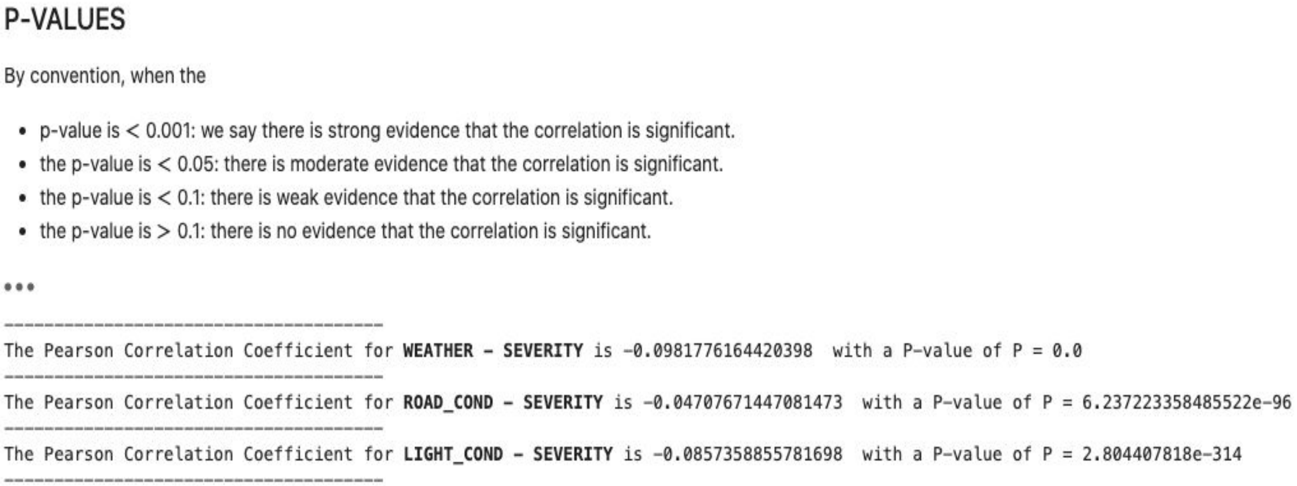
•••

[11]:

	SEVERITYCODE	WEATHER_V	ROADCOND_V	LIGHTCOND_V	COLLISIONVEH_V
SEVERITYCODE	1.000000	-0.098178	-0.047077	-0.085736	0.022391
WEATHER_V	-0.098178	1.000000	0.761901	0.316668	0.190316
ROADCOND_V	-0.047077	0.761901	1.000000	0.107981	0.094841
LIGHTCOND_V	-0.085736	0.316668	0.107981	1.000000	0.222931
COLLISIONVEH_V	0.022391	0.190316	0.094841	0.222931	1.000000

Table 1. Attributes correlations

The correlation between the attributes doesn't show high values, in particular between the independent attributes and the target. However, a fair correlation between Weather and Road Conditions and Light Conditions is an indicator of the importance of these features for the occurrence of an accident.



P-VALUES

By convention, when the

- p-value is < 0.001: we say there is strong evidence that the correlation is significant.
- the p-value is < 0.05: there is moderate evidence that the correlation is significant.
- the p-value is < 0.1: there is weak evidence that the correlation is significant.
- the p-value is > 0.1: there is no evidence that the correlation is significant.

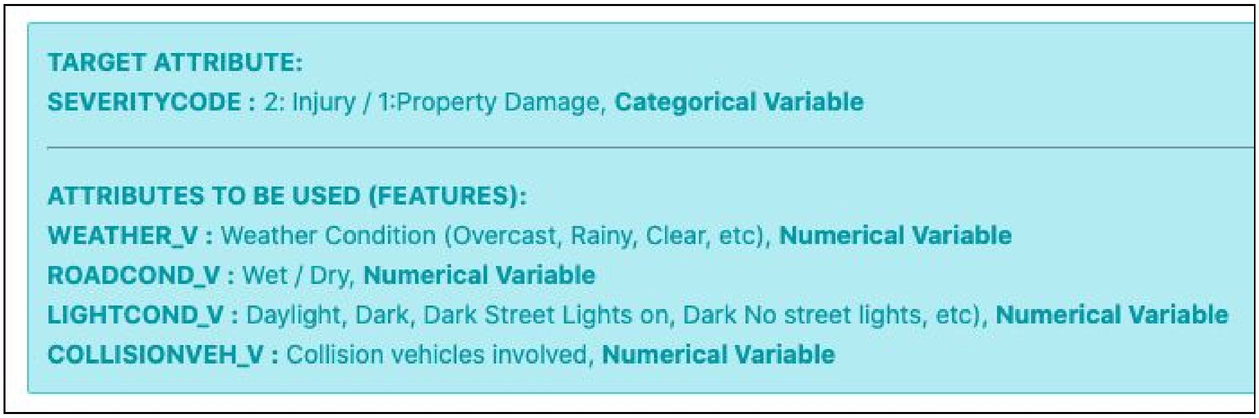
•••

```
---------------------------------------
The Pearson Correlation Coefficient for WEATHER - SEVERITY is -0.0981776164420398  with a P-value of P = 0.0
---------------------------------------
The Pearson Correlation Coefficient for ROAD_COND - SEVERITY is -0.04707671447081473  with a P-value of P = 6.237223358485522e-96
---------------------------------------
The Pearson Correlation Coefficient for LIGHT_COND - SEVERITY is -0.0857358855781698  with a P-value of P = 2.804407818e-314
---------------------------------------
```

Table 2. P-Values

We can note, however P-values. I include the table of values with the understanding of a strong evidence of the correlation.

Finally, the attribute selection for the Models are:



TARGET ATTRIBUTE:
SEVERITYCODE : 2: Injury / 1:Property Damage, **Categorical Variable**

ATTRIBUTES TO BE USED (FEATURES):
WEATHER_V : Weather Condition (Overcast, Rainy, Clear, etc), **Numerical Variable**
ROADCOND_V : Wet / Dry, **Numerical Variable**
LIGHTCOND_V : Daylight, Dark, Dark Street Lights on, Dark No street lights, etc), **Numerical Variable**
COLLISIONVEH_V : Collision vehicles involved, **Numerical Variable**

Table 3. Features and Target for Supervised Machine Learning Models

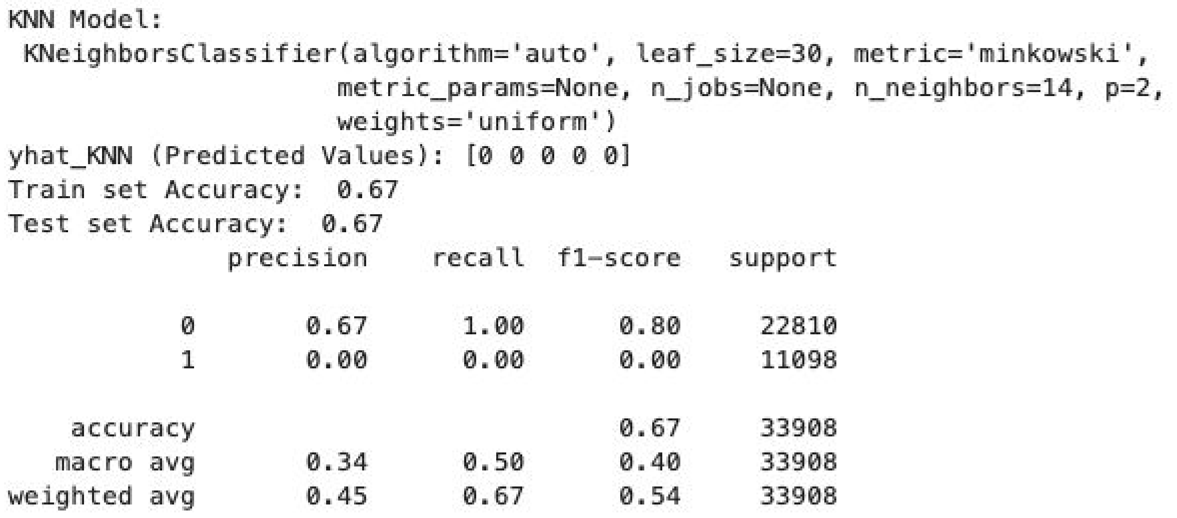
Predictive Models - Results and Evaluation

For the Supervised Machine Learning Models I select those used in the course. Although the author studied some other models, deadlines for these tests were proposed for future work.

The models used in this work for the Predictive Models are:

1. KNN Classifier
2. Decision Tree Classifier
3. Logistic Regression
4. Supported Vector Machine SVM

The results, when all the feature are used, leads to strange results with the following images showing them for KNN:

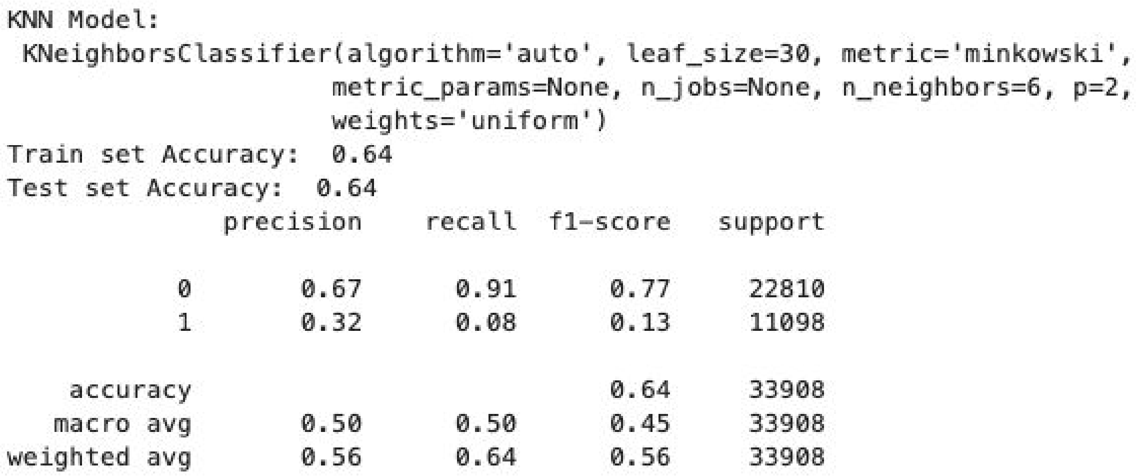


```
KNN Model:
 KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=14, p=2,
                     weights='uniform')
yhat_KNN (Predicted Values): [0 0 0 0 0]
Train set Accuracy:  0.67
Test set Accuracy:  0.67
              precision    recall  f1-score   support

           0       0.67      1.00      0.80     22810
           1       0.00      0.00      0.00     11098

    accuracy                           0.67     33908
   macro avg       0.34      0.50      0.40     33908
weighted avg       0.45      0.67      0.54     33908
```

Using X_feature with **WEATHER, ROADCOND, LIGHTCOND** and y=SEVERITYCODE, the results for the model. Note a 0.0 precision, recall and f1-score for Property Damage. This is not a valid result. After several analyses, I found k=15, the recommended value, was affecting the results. For k=6, results improved. Please see results next page.



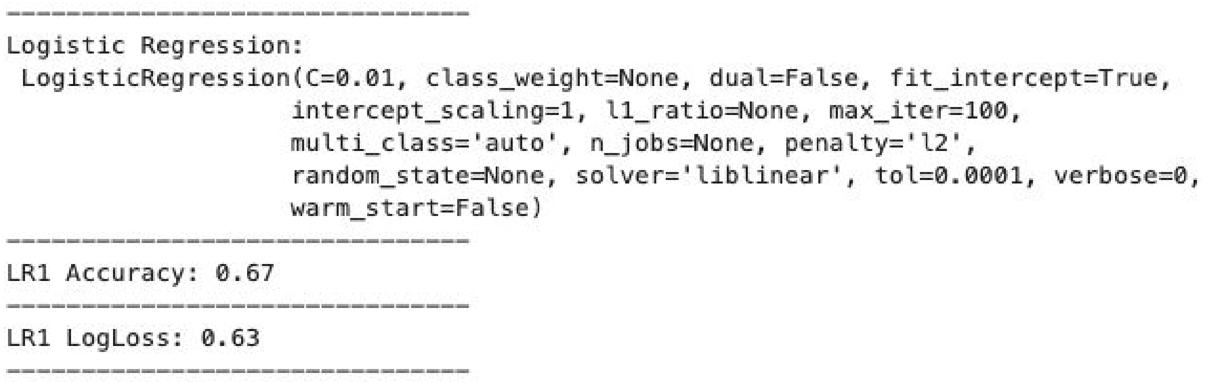
```
KNN Model:
 KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                      metric_params=None, n_jobs=None, n_neighbors=6, p=2,
                      weights='uniform')
Train set Accuracy:  0.64
Test set Accuracy:  0.64
              precision    recall  f1-score   support

           0       0.67      0.91      0.77     22810
           1       0.32      0.08      0.13     11098

    accuracy                           0.64     33908
   macro avg       0.50      0.50      0.45     33908
weighted avg       0.56      0.64      0.56     33908
```

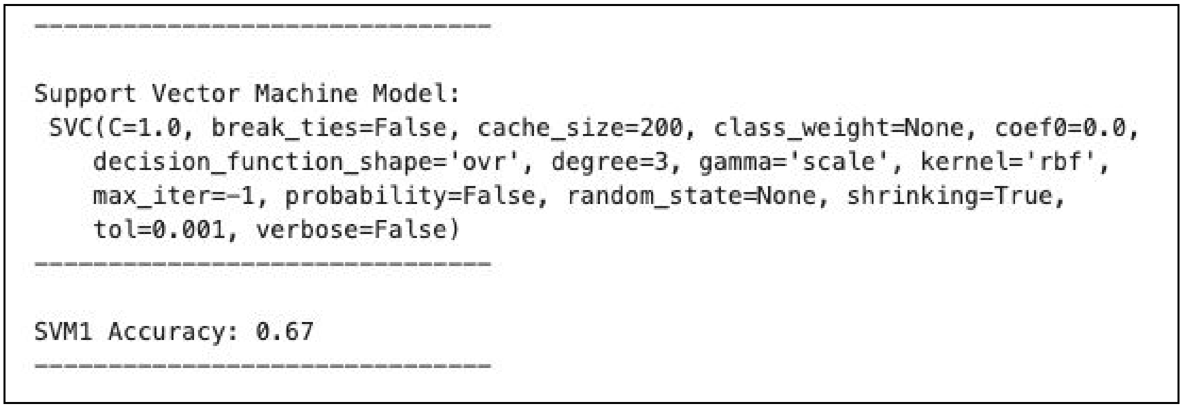
For Decision Tree, Computed Accuracy was 0.67. No meaning results for Precision and Recall.

For Logistic Regression, the results are shown below with an Accuracy of 0.67, very similar to Decision Trees. LogLoss = 0.63, that is not a bad result.



```
-------------------------------
Logistic Regression:
 LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='liblinear', tol=0.0001, verbose=0,
                   warm_start=False)
-------------------------------
LR1 Accuracy: 0.67
-------------------------------
LR1 LogLoss: 0.63
-------------------------------
```

For SVM was obtained:

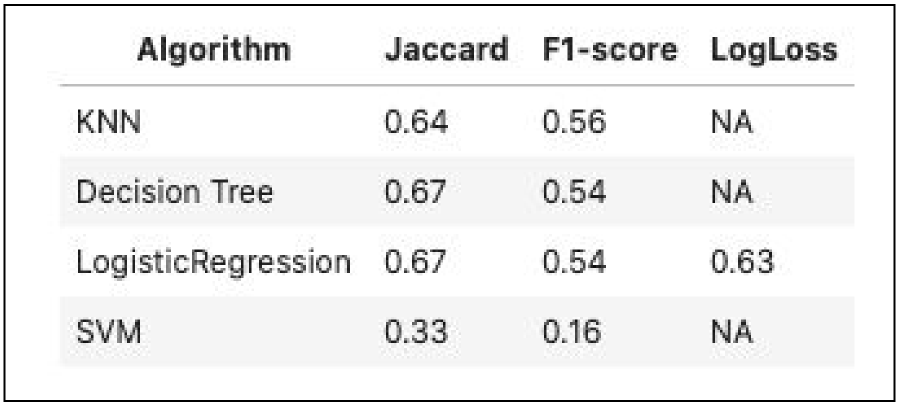


```
-------------------------------


Support Vector Machine Model:
 SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
-------------------------------


SVM1 Accuracy: 0.67
-------------------------------
```

The accuracy of the models are shown below for the Jaccard Index, f1-score and for LogLoss in Logistic Regression.



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.64	0.56	NA
Decision Tree	0.67	0.54	NA
LogisticRegression	0.67	0.54	0.63
SVM	0.33	0.16	NA

Several combinations of features (3,2 & 1) were tested to recover the f1-score for all the models, however, changes were very small. KNN was strongly affected by the parameter K. SVM showed a very poor result that needs to be investigated.

However, given that the information of the weather conditions are very subjective, in terms of categorical variables, other resources need to be included in the models for a more objective result.

Conclusions

- Models behavior is similar in terms of accuracy, except for SVM. Not all reported f1-score for all the values of the target variable. It indicates an issue with the feature attributes that needs to be reviewed in a later work.

- As shown in the exploratory data analysis, Weather is an important factor in traffic accidents, however, because of its imbalance, it is hard to manage with simple methodologies (like those presented here). Imbalance methods, relation of more datasets (weather, junctions, etc), and also the segmentation of the problem (for example only highways) will lead to improved predicted results.

- It is a very interesting topic worth-value to continue investigating it.

Future Work

- To work in integrating more datasets for better features analysis
- Include imbalance methods such as downsampling, oversampling, using methods such as SMOTE or using unbalanced machine learning algorithms.
- Investigate more on how to use satellite images to include in the data integration.

Main References

[1] Accident Severity Prediction Using data Mining Methods. S Ramya et al; International Journal of Scientific Research in Computer Science, Vol 5, Issue 2, 2019

[2] High-Resolution Road Vehicle Collision Prediction for the City of Montreal, Antoine H'ebert et al, Preprint, 2019

[3] After personal research on the web.

[4] Ibero American Report for Traffic Safety, 2015,2016, In Spanish.