Coursera_Capstone

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

K. Giraldo, 2020

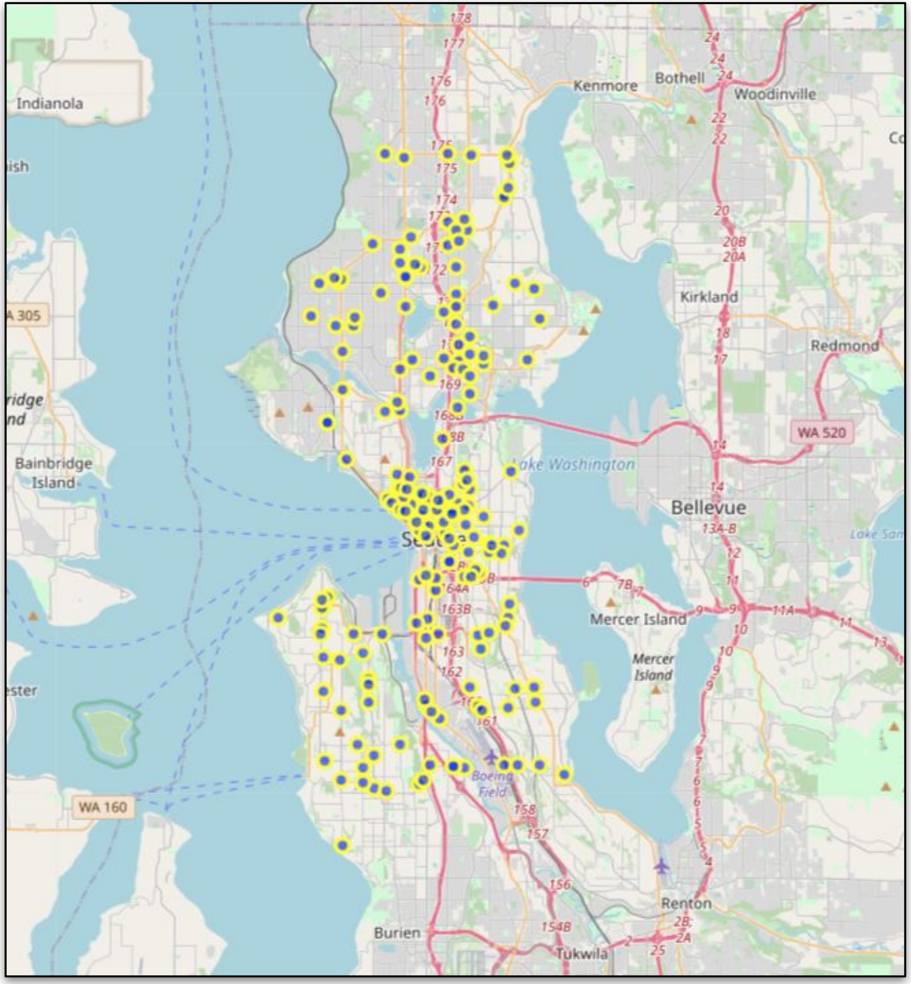
Introduction

Intro

- Traffic accidents are, in modern times, a great cause of health issues for many developed and under-developed countries
- However, weather or road conditions are not taken fully into account
- 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.

Geographic Location

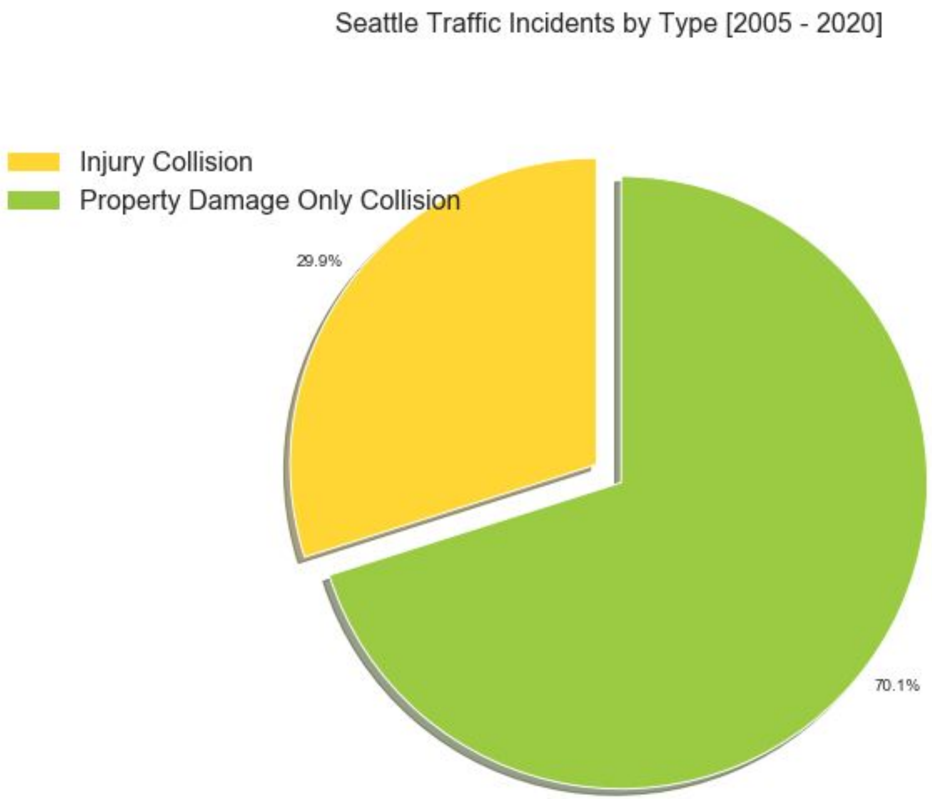
A sample of accidents in the City of Seattle, WA, between the years 2005-2020



Traffic Incidents Distribution

Pie Chart showing Severity of the Accident. Only two values included in the dataset

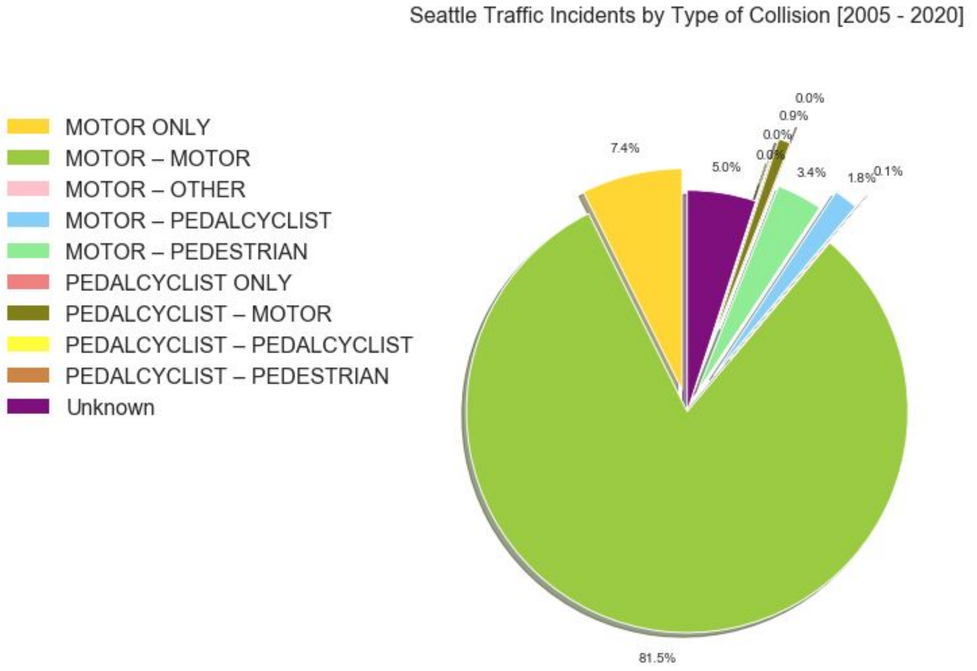
```
-----------------------------
                                value_counts
Incident_Severity
Injury Collision                       58188
Property Damage Only Collision        136485
-----------------------------
```



Usual Types

In the Accidents usually are involved bicycles, pedestrians and vehicles. Here it is the distribution, but not used in this work.

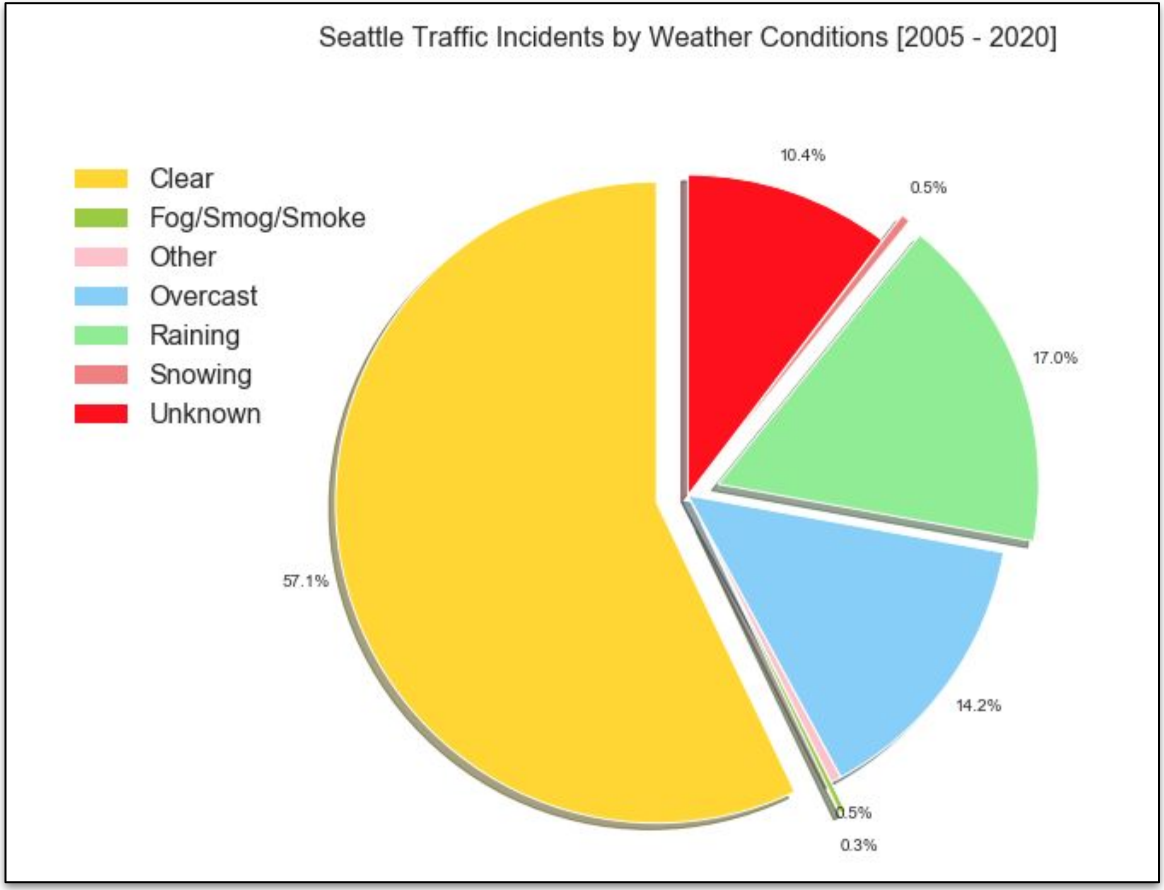
Collision_Type	value_counts
MOTOR ONLY	14353
MOTOR – MOTOR	158592
MOTOR – OTHER	102
MOTOR – PEDALCYCLIST	3426
MOTOR – PEDESTRIAN	6526
PEDALCYCLIST ONLY	96
PEDALCYCLIST – MOTOR	1692
PEDALCYCLIST – PEDALCYCLIST	12
PEDALCYCLIST – PEDESTRIAN	75
Unknown	9799



Data Exploration

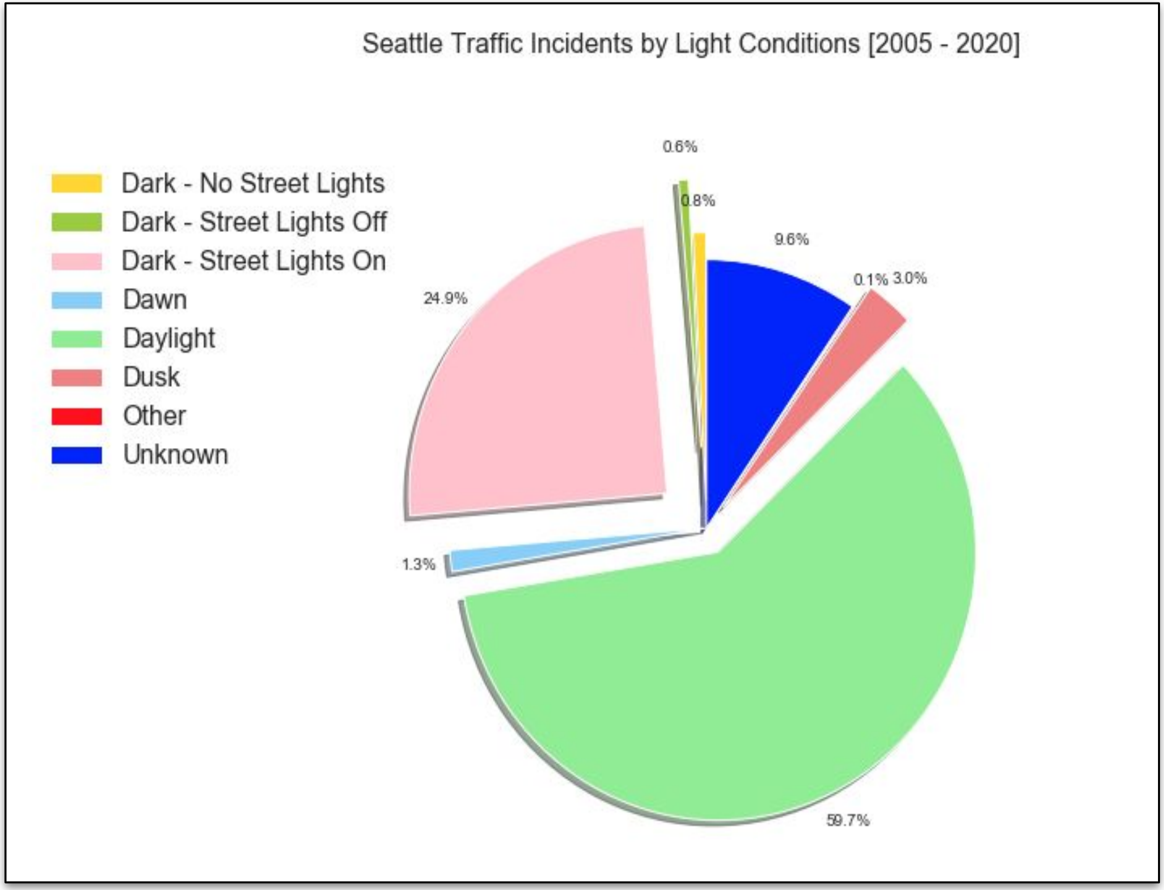
Traffic Accidents by Weather Condition

Pie Chart showing the distribution of the values of the categorical variable Weather within the dataset. Note a highly unbalanced distribution for Clear Weather



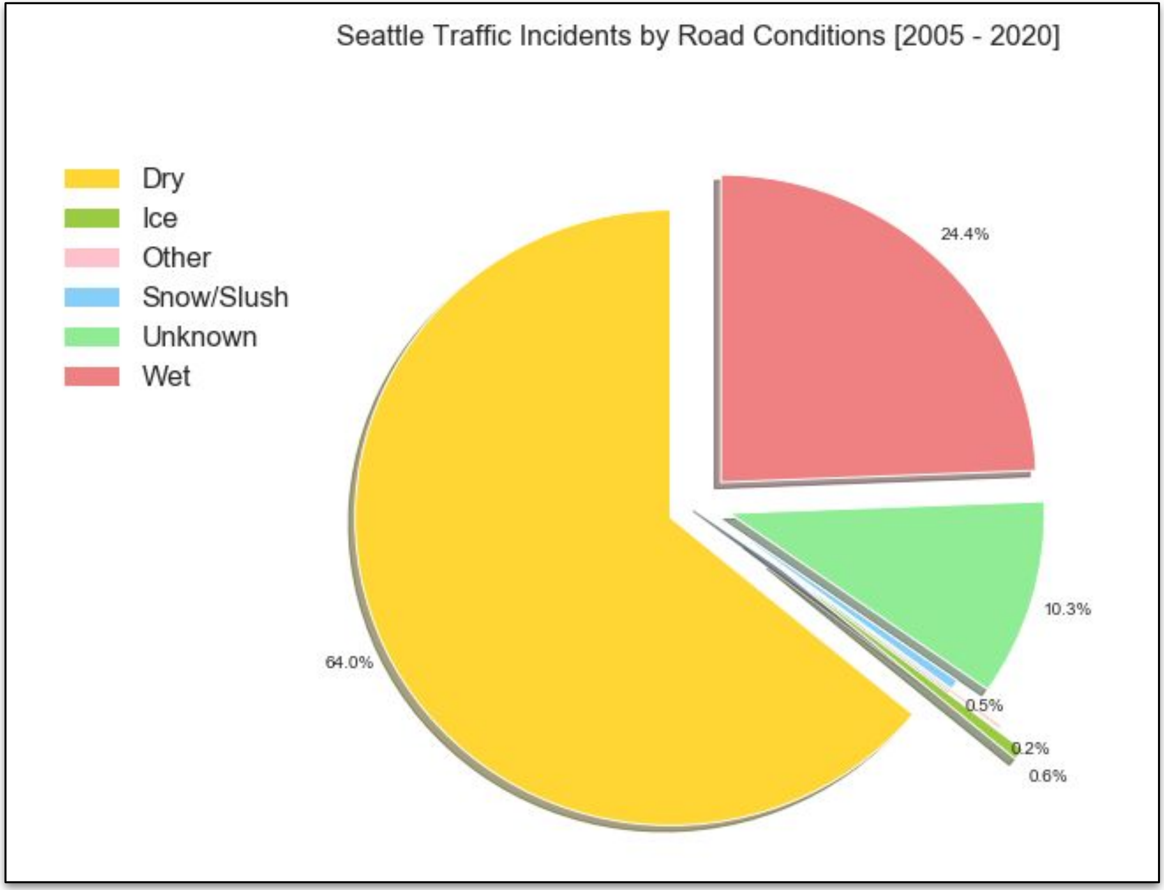
Traffic Accidents by Road Condition

Here Daylight dominates the distribution.



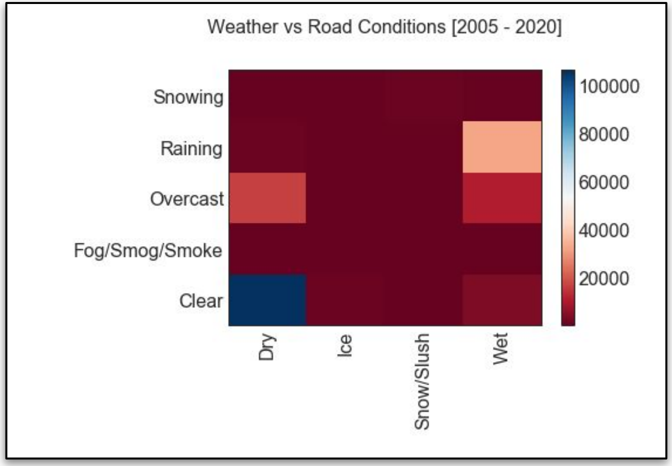
Traffic Accidents by Light Conditions

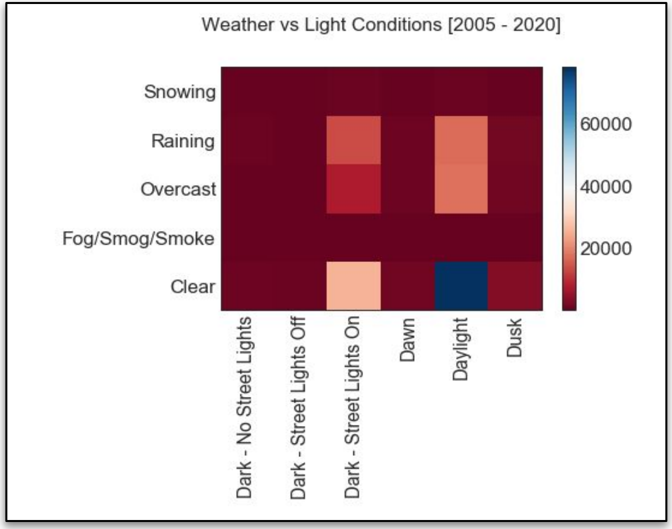
... and here is Dry that creates the imbalance



Imbalance Represented in the Pivot Tables

Weather vs Road Conditions and Weather vs Light Conditions





Poor Correlations with the Target Variable

Only few correlations are of relevance in this table

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

Models & Evaluation

Machine Learning Models (Supervised)

- K-Nearest Neighbors (KNN) Classifier
- Decision Trees Classifier
- Logistic Regression
- Supported Vector Machine (SVM)

KNN Classifier Results

1. X_Features = [Weather, RoadCond, LightCond]
2. Accuracy: 0.64
3. Precision, Recall, F1-Score: 0.77/0.13

```
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
```

Decision Trees Results

1. X_Features = [Weather, RoadCond, LightCond]
2. Accuracy: 0.67

```
Decision Tree Model:
 DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=4, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
-------------------------------

DecisionTrees's Accuracy: 0.67
-------------------------------

              precision    recall  f1-score   support

           1       0.67      1.00      0.80     22810
           2       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
```

Logistic Regression Results

1. X_Features = [Weather, RoadCond, LightCond]
2. Accuracy: 0.67
3. LogLoss: 0.63

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

	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

SVM Results

1. X_Features = [Weather, RoadCond, LightCond]
2. Accuracy: 0.67

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

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

0	0.00	0.00	0.00	22810
1	0.33	1.00	0.49	11098
accuracy			0.33	33908
macro avg	0.16	0.50	0.25	33908
weighted avg	0.11	0.33	0.16	33908

Results Comparison for All ML Models

1. Jaccard Index
2. F1-Score
3. LogLoss

```
------------------------------------
KNN Jaccard index: 0.64
KNN F1-score: 0.56
------------------------------------
DT Jaccard index: 0.67
DT F1-score: 0.54
------------------------------------
LR Jaccard index: 0.67
LR F1-score: 0.54
LR LogLoss: 0.63
------------------------------------
SVM Jaccard index: 0.33
SVM F1-score: 0.16
------------------------------------
```

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

Conclusions

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.