**IBM Data Science Capstone Project**

**Car Accident Severity**

**Oscar Vicente Remón**

**October 11th, 2020**

**Abstract**

Despite of the number of technological improvements, in a daily basis many car accidents still occur in the big urban areas. Seattle with 783,187 inhabitants has a relevant number of accidents. Through this project, the dataset provided by the Seattle Traffic Management Division and containing a detailed report of each of the accidents between 2004 and 2020 is going to be analysed. The first objective of the project will be to determine the influence of certain variables in the occurrence of an accident. As second objective a set of Machine Learning models will be simulated in order to foreseen under which variables conditions an accident could happen in the future. As last, a set of conclusions and recommendations are going to be draft to the stakeholders.

Table of Contents

[**Description of the problem and a discussion of the background. 3**](#_Toc53385131)

[**A description of the data and how it will be used to solve the problem. 5**](#_Toc53385132)

[**Data understanding. 5**](#_Toc53385133)

[**Data Preparation. 8**](#_Toc53385134)

[**Methodology. 10**](#_Toc53385135)

[**Results 15**](#_Toc53385137)

[**Discussion section 17**](#_Toc53385138)

[**Conclusion 19**](#_Toc53385139)

[**Recommendations 19**](#_Toc53385140)

# **Description of the problem and a discussion of the background.**

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest.

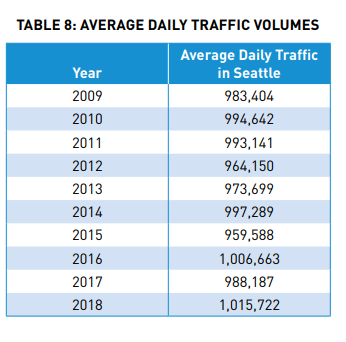
As of the 2020 census, there were 783,187 people living in Seattle.



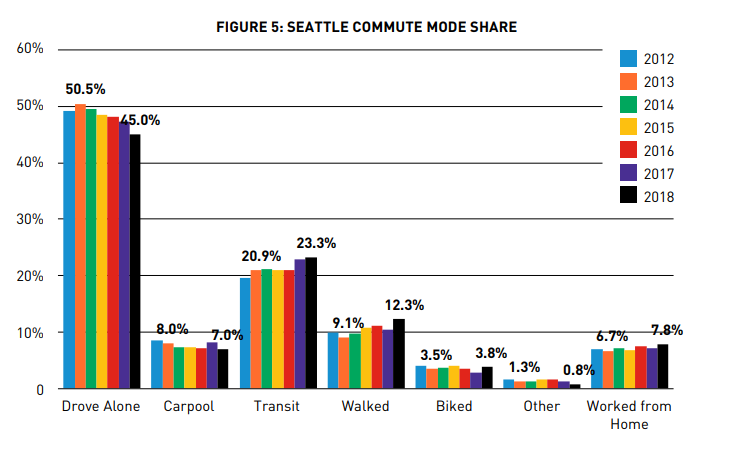
Source: https://worldpopulationreview.com/us-cities/seattle-wa-population

Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it is no surprise that Seattle sees car accidents every day.

In order to have an idea of the traffic volume order of magnitude and how is split per transportation type the tables and graphs below should give an overview to of the situation even if containing only data up to 2018.



Source:http://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019\_Traffic\_Report.pdf



Source: http://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019\_Traffic\_Report.pdf

The number of accidents according to the SDOT Traffic Management Division, Traffic Records Group dataset collisions shows a slight decrease in the last years.



Source: SDOT Traffic Management Division, Traffic Records Group

However, due to the amount of money these accidents cost to the taxpayers, the local authorities would like to understand better, which factors have an impact in the accidents and how the accidents could be predicted or better said, avoided, in the future.

With this objective, the project to be developed will:

* Evaluate the variables provided, extended in the full report version.
* Determine how they affect the accidents numbers and typology.
* Analyse the different prediction models applied to the set of selected variables.
* Proposed the best model for predicting the accidents
* Provide a lesson learnt and recommendations document to the stakeholders with the aim to improve the situation.

# **A description of the data and how it will be used to solve the problem.**

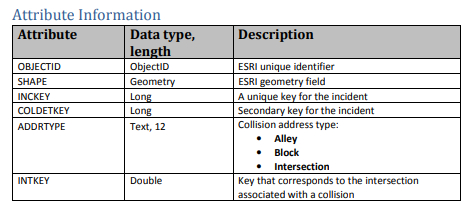
## **Data understanding.**

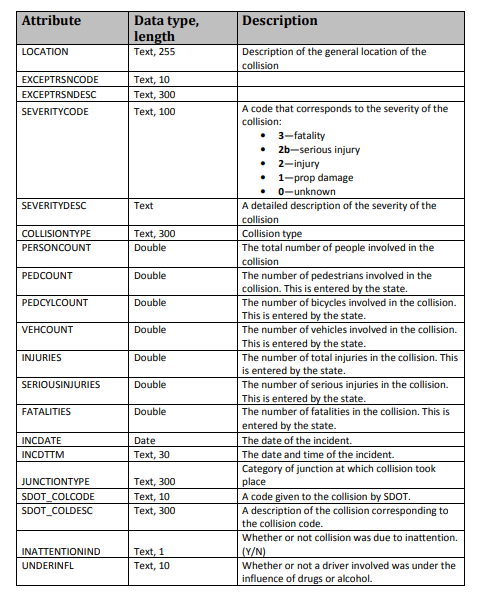
The dataset used for this project is based on car accidents occurring in the city of Seattle during the years 2004 to 2020**.**

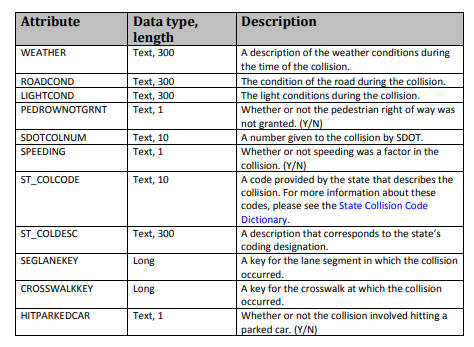
The [link](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) to download the dataset is provided in the Coursera Capstone project week 1 labs.

The Seattle Traffic Management Division, in charge to produce the dataset, categorises the accidents in two types according to their severity: 1 - Standing for “Property Damage Only” and 2 - Injury Collision.

Attached to the severity for each accident record there is a set of attributes that classifies the accident record.

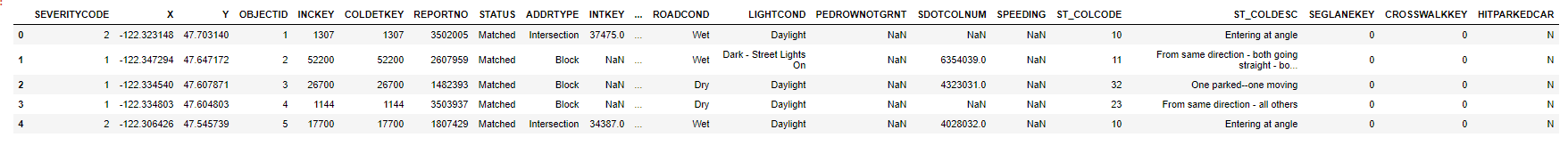






The dataset can be downloaded first in format csv and imported into pandas or directly read from the URL.

The contents can be displayed by using Head function.



A Jupiter notebook containing the final code has been submitted via Github.

Out of the total number of variables available, within this project only the following variables are initially taken into consideration:

Predicted and dependent variable: SEVERITYCODE

Independent variables:

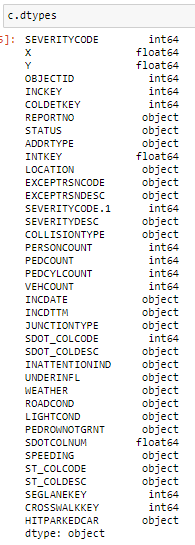
Information about the location and characteristic of the address where the location took place: ADDRTYPE, JUNCTIONTYPE, X, Y, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, ROADCOND.

Environmental factors: WEATHER, LIGHTCOND

## **Data Preparation.**

## 

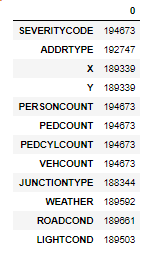
By applying the DTYPE instruction to the data frame read from the url containing the dataset, the technical structure of the dataset can be quickly displayed.



In the original format, the dataset is not suitable for quantitative analysis, because a number of categorical variables are identified, those identified as object type.

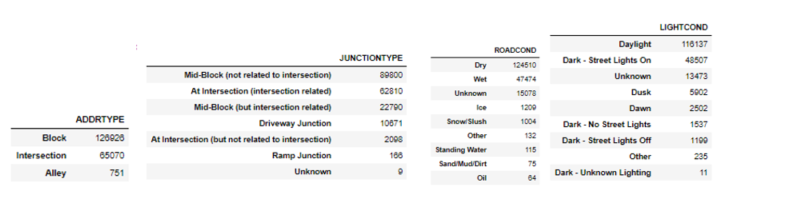
By looking further at the dataset description, some variables / columns will not provide too much information to the model, those are the variables associated to technical fields like OBJCTID, REPORTNO, INTKEY.

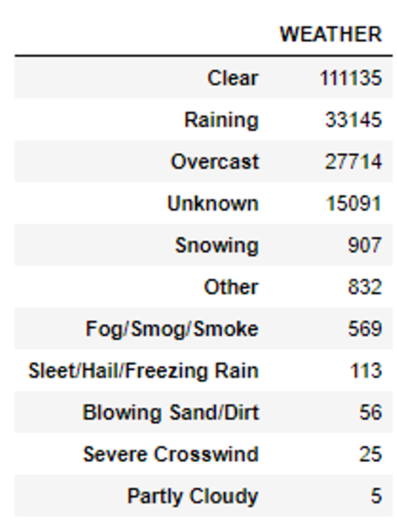
The total number of records contained in the dataset is 194,673. The drilldown in terms of populated records is displayed in the table below. Only the sort list of study variables considered in our project are shown.



This lead to the conclusion that not all variables are always populated. Some variables contain blank or empty records.

Via a quick observation to the sort list of categorical variables of study:





It can be observed that there are records associated to some of the variables showing: Unknown or Other values. This kind of values are not useful in the dataset since they do not provide any information in the model.

In order to reach an acceptable data quality to feed our models, some actions has been performed, they are further detailed in the code, but to summarise:

* + Eliminate columns not included in the sort list of study.
  + Eliminate all records with values adding potential noise into the model.
    - All accidents cases must have a value in the study variables that add value to the model, empty values or useless categories will only contribute to a worse prediction.

As a consequence of the data cleaning process, the total number or records will remain in 164,731.

# **Methodology.**

# Once the dataset has fulfilled the expected requirements in terms of data quality, the next step is to further explore the data as such.

This is a key process which will help to take some further decisions before being able to train and test the models. The following questions need to be answered:

Is the dataset balanced?

Are all variables initially in the project scope still meaningful?

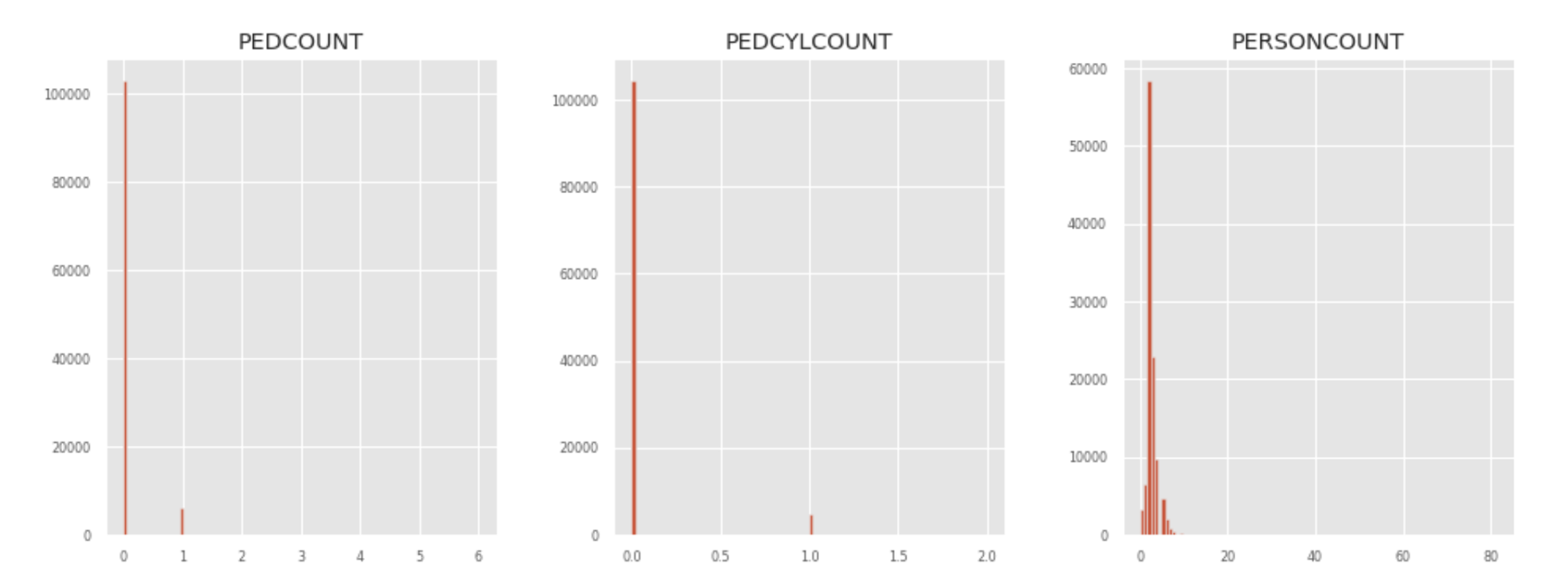
The dataset contains a distribution which in terms of records per each of the values of the target variable is unbalanced. The number of accidents with property damage severity is almost double that the injuries severity type. In order to not condition the algorithm to train the model towards one of the two values a balance method using the random down sampling technique has been performed.

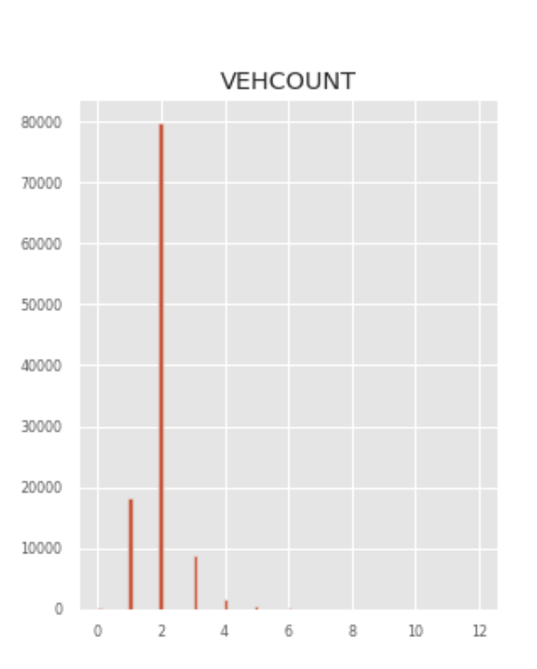
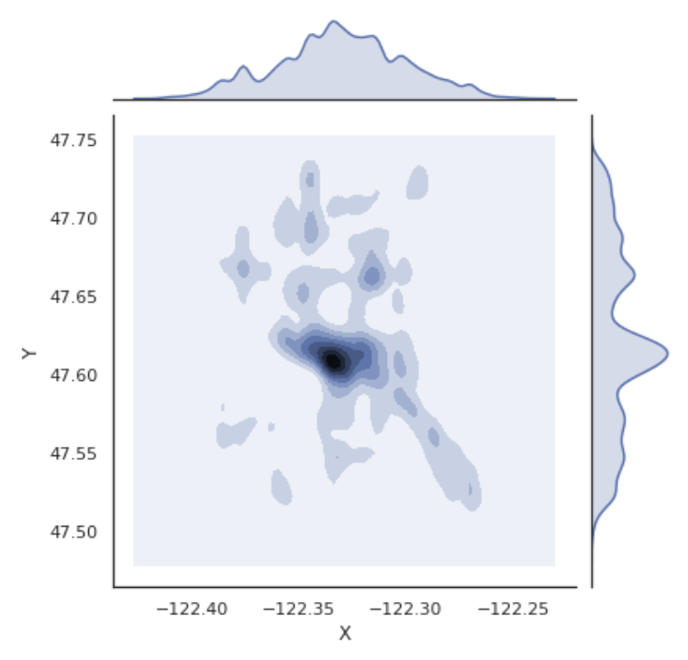
As a consequence, the number of records for each of the variables states in 54,544.

The next step will be to split the dataset between numerical and categorical variables. And explore them carefully.

Numerical variables: X, Y, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT.

First step will be to display the distribution in terms of number of records and possible values for each of the variables. With this purpose the following graphs have been generated.



In addition to the number of records a correlation study between the variables has been carried out. Internally, among the independent variables and external between the independent variables towards the target variable. The methods of Pearson and Kendall has been used to study the correlation. The outcome can be displayed in the tables below.

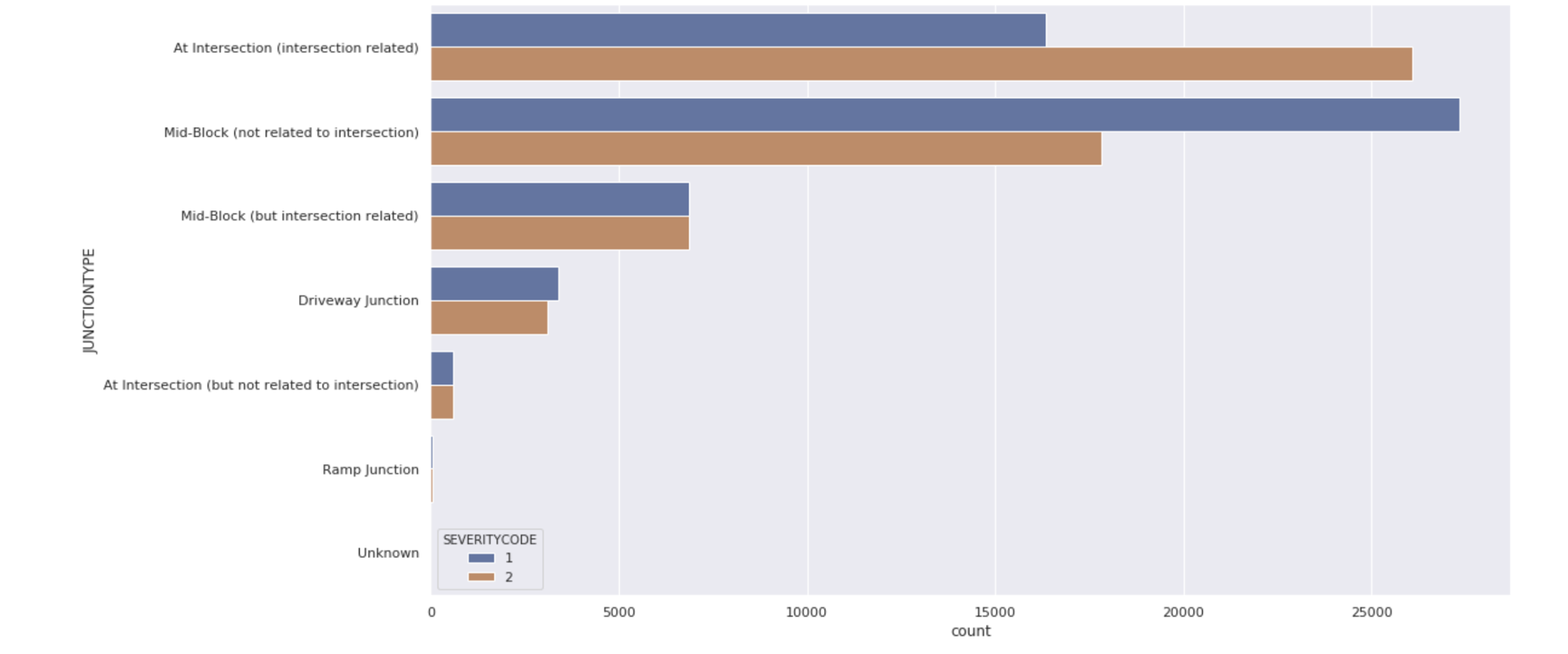
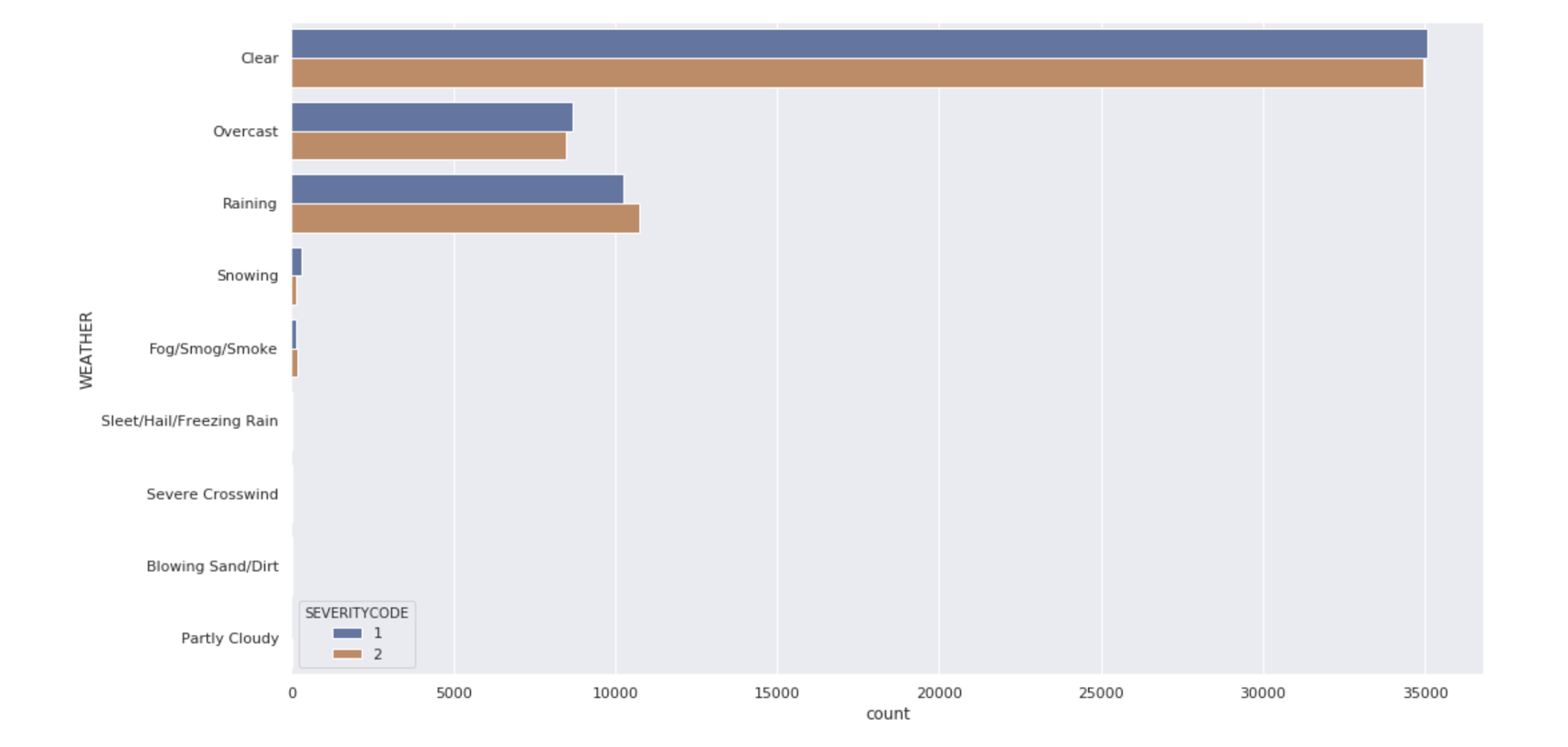


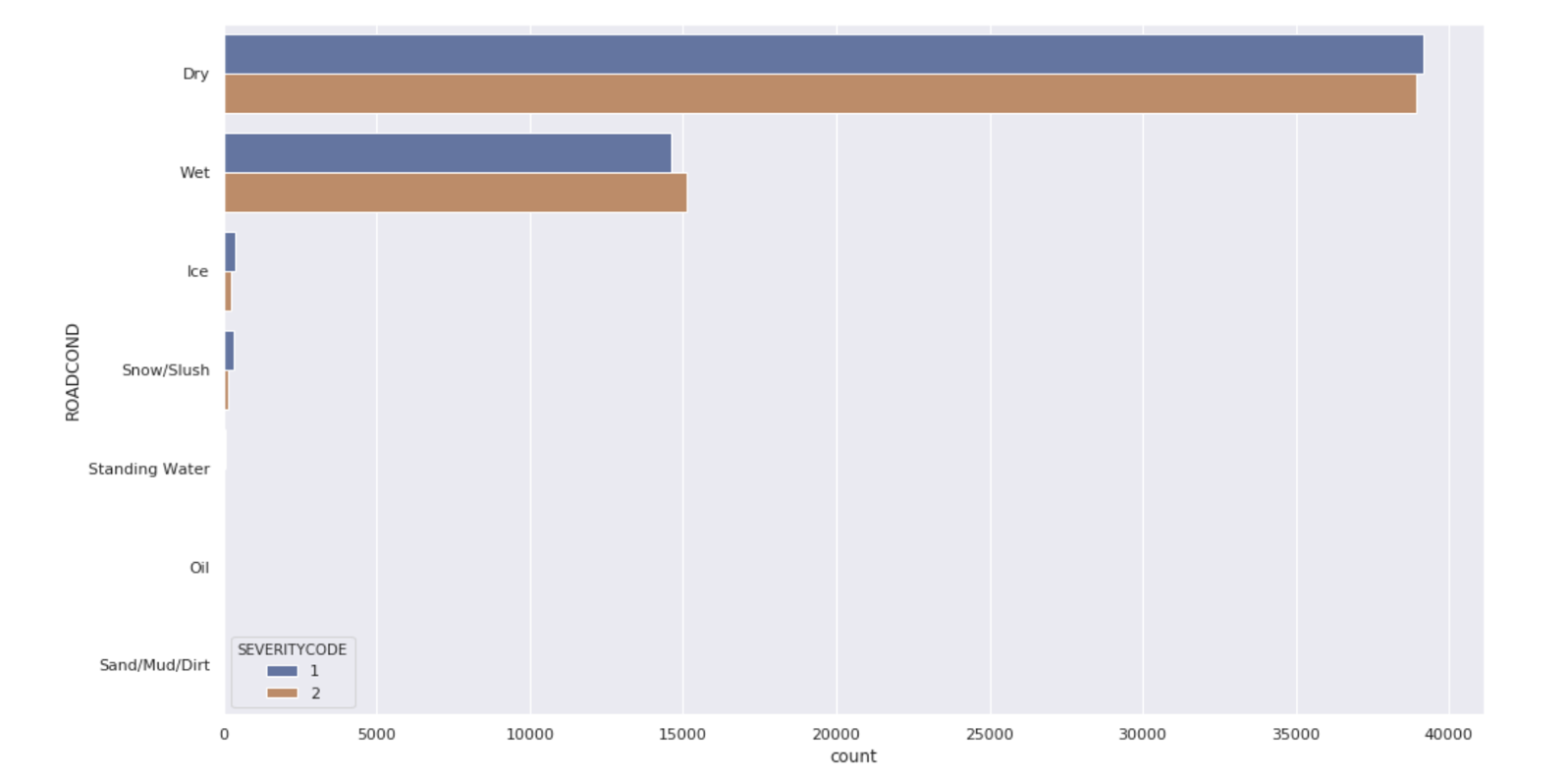
As a result of the analysis one can conclude that the numerical variables as such, do not influence massively in the variable categorizing the type of accidents.

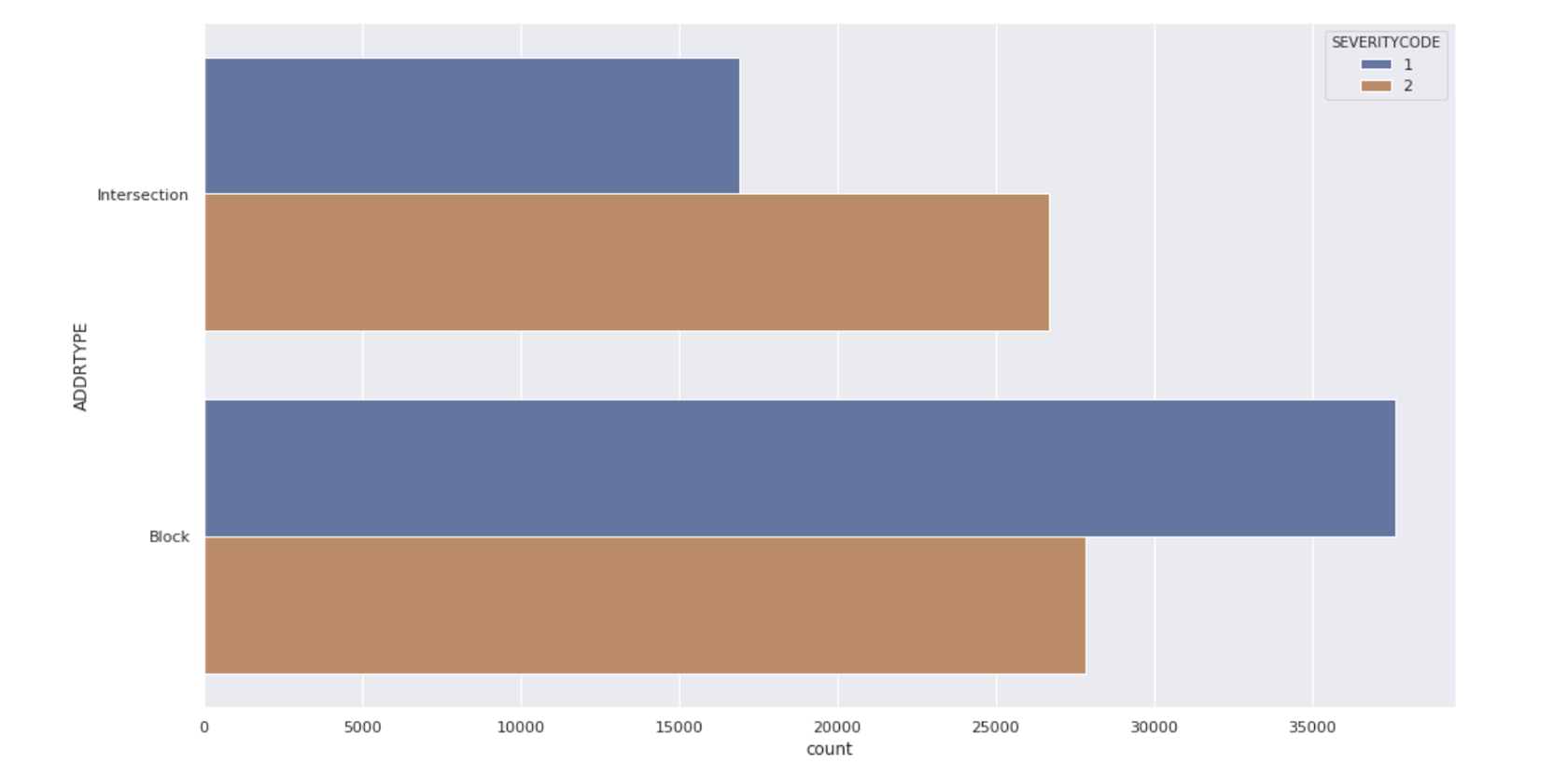
The big majority of the accidents always occur with no pedestrians and no bicycles involved. A low number of people involved and with 1 or 2 vehicles. On top of that, the location of the accidents is basically around the centre where the majority of the cars are usually condensed. The correlation values, always below 0.5, between these set of independent variables and the target variable to be predicted, using Pearson and Kendall methods depict that is irrelevant for the study.

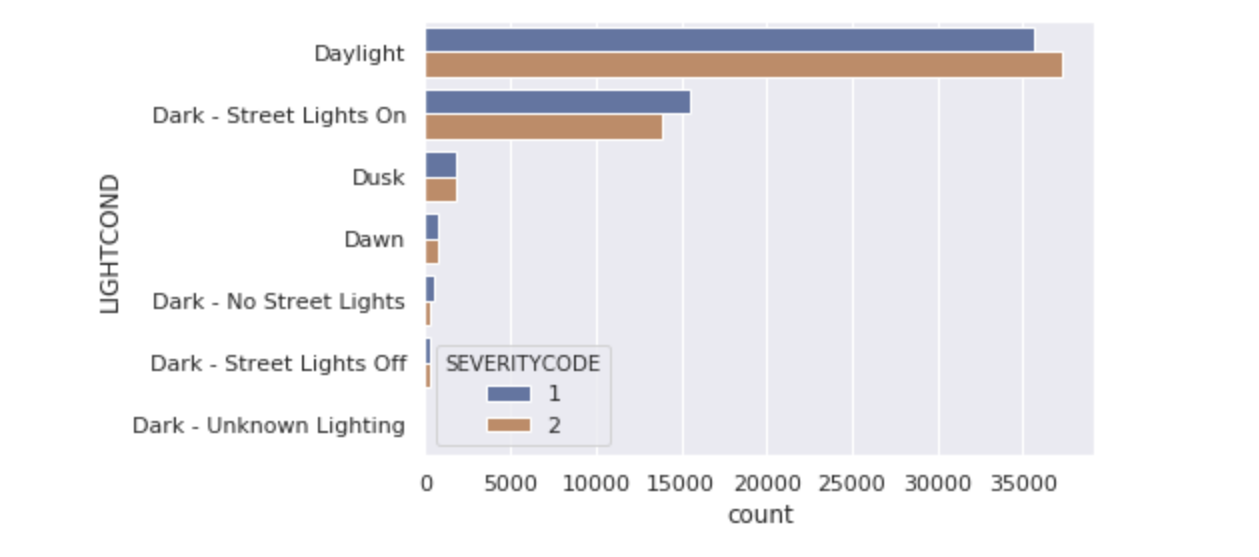
Categorical variables:

The distribution in terms of number of records for each of the values can be observed in the graphs below. In this case it has been drilldown for each of the values of the forecasted value.









Contrary of what one could think, the majority of the accidents occur when the conditions of are better in terms of weather, road and light conditions. And under this scenario both type of accidents are affected in a similar manner.

One could explain this situation by saying that when the conditions of weather, road or light are worse the drivers are more concentrated or in tension, so they get less distracted and focused on the road and the environment. The variable junction type, does not provide much more information further than the information already provided in the variable address type, blocks or intersection. For this reason the junction type variable will not be consider to train the model but address type only instead.

The information depicted under the address type graph, indicates that the accidents with personal damage occur more often in the intersections while the ones near to blocks affects more to accidents relevant to infrastructure only damages.

As a result of this exploratory analysis, the variables which are going to be considered for to train and test our models are: Address type, Light Conditions, Weather, Road conditions. So the models a priori should be able to predict what type of accident could occur under a given set of location, infrastructure and weather conditions.

Prior to describe the potential models to be used, the categorical variables need to be encoded as a numerical variables.

**Models.**

The project objective will be to find the model which predict better by using these set of selected variables the severity of a potential future accident.

The Machine Learning models used are K – nearest neighbors, Decision Tree, Support Vector Machine and Logistic regression.

K – nearest neighbours is one of the simplest algorithms that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.

Decision Tree is a non-parametric supervised learning method. It classifies the dataset by using a tree structure, composed of nodes and leaves. At any node one of the features of the data is evaluated in order to split the training observations or to make a data point to follow a given path when making a prediction.

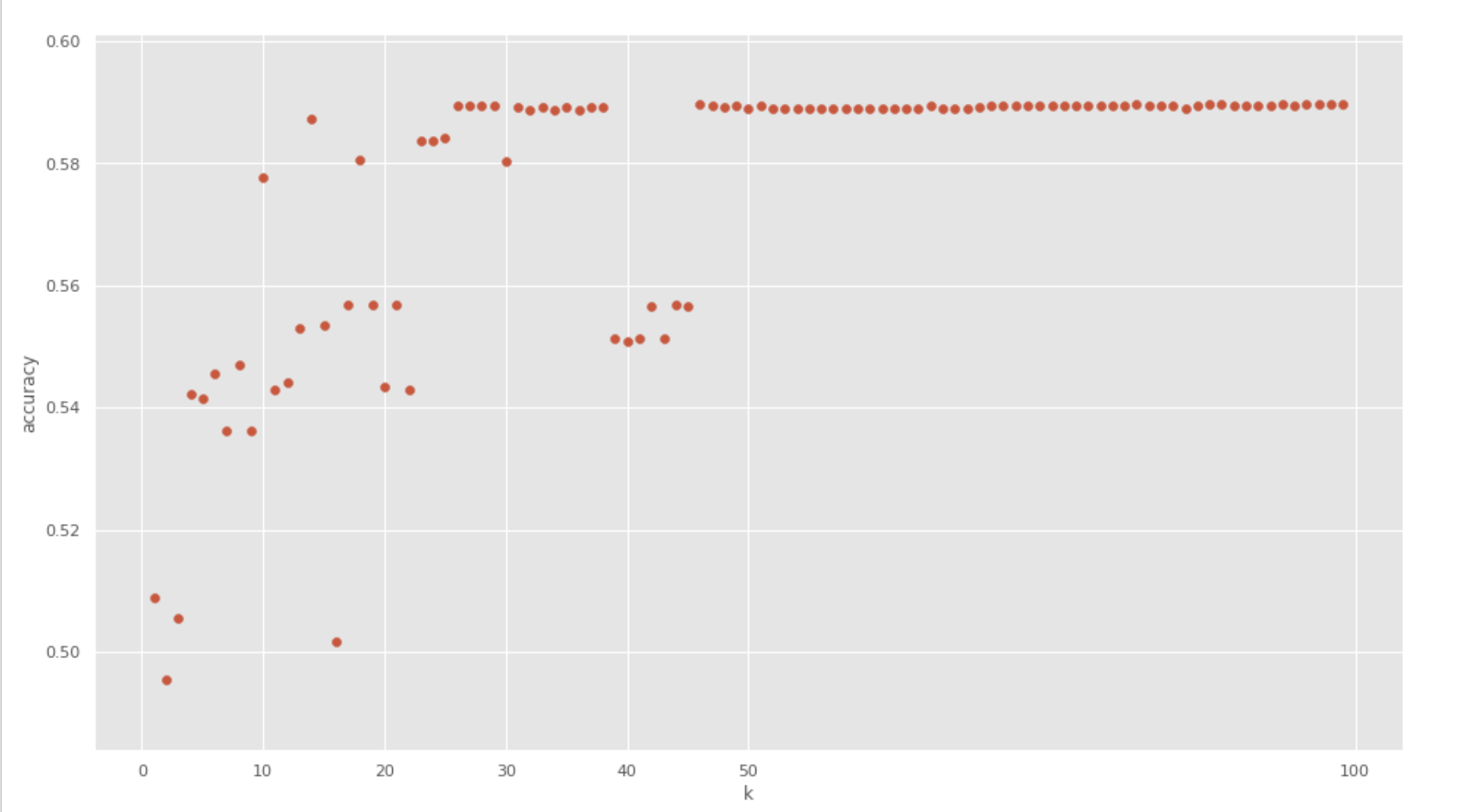
Support vector machine algorithm tries to find an hyperplane in an N dimensional space, being N the number of features that is able to classify in a unique way the datapoints.

Logistic Regression predictions are mapped to be between 0 and 1 through the logistic function, which means that predictions can be interpreted as class probabilities.

# **Results**

Cross validation process has been performed, so forth all the models are trained with the same apportion of samples for test (70%) and training (30 %) cases.

K nearest neighbour method from the sklearn library has been used. Prior to train the model the set has been scaled so that all of them can be uniformly evaluated. The best value for K starts to be in 26 but starts to be fully stable around 50, and with this value the simulation has been done. In the graph below you can see how the accuracy of the model will vary according to the value of K.



Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1 - score** |
| **Property damage** | **0.69** | **0.57** | **0.62** |
| **Injury** | **0.49** | **0.62** | **0.55** |
| **Accuracy** | **0.59** |  |  |
| **Micro Average** | **0.59** | **0.59** | **0.59** |
| **Macro Average** | **0.59** | **0.59** | **0.59** |
| **Weighted Average** | **0.61** | **0.59** | **0.59** |

Decision Tree algorithm from the sklearn library has been used. In this simulation the entropy criteria has been chosen together with the entropy as 6.

Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1 - score** |
| **Property damage** | **0.69** | **0.57** | **0.62** |
| **Injury** | **0.49** | **0.62** | **0.55** |
| **Accuracy** | **0.59** |  |  |
| **Micro Average** | **0.59** | **0.59** | **0.59** |
| **Macro Average** | **0.59** | **0.59** | **0.59** |
| **Weighted Average** | **0.61** | **0.59** | **0.59** |

Support Vector Machine algorithm from the sklearn library has been used.

Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 - score |
| **Property damage** | **0.57** | **0.69** | **0.63** |
| **Injury** | **0.62** | **0.49** | **0.55** |
| **Accuracy** | **0.59** |  |  |
| **Micro Average** | **0.59** | **0.59** | **0.59** |
| **Macro Average** | **0.59** | **0.59** | **0.59** |
| **Weighted Average** | **0.59** | **0.59** | **0.59** |

Logistic Regression from the sklearn library has been used. The parameters for this simulation has been the set up of C , regularization strength as 0.01 and Libilineal as Algorithm to used in the optimization problem.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 - score |
| **Property damage** | **0.57** | **0.69** | **0.63** |
| **Injury** | **0.62** | **0.49** | **0.55** |
| **Accuracy** | **0.59** |  |  |
| **Micro Average** | **0.59** | **0.59** | **0.59** |
| **Macro Average** | **0.59** | **0.59** | **0.59** |
| **Weighted Average** | **0.59** | **0.59** | **0.59** |
| **LogR** | 0.68 |  |  |

# **Discussion section**

Usually after performing the model of a problem with several algorithm options one would like to know how each of the models are performing and being able to evaluate them.

Prior to this, some concepts which will be used during the evaluation will need to be explained and adapted to the case study.

**Accuracy:** Correctly classified points by a total number of points in the test set.

**Confusion matrix**: a step further to measure the performance of our project. Via this grid you classify out of the total of test observations how many cases has been predicted well towards the actual values.

Actual Values ( Test Sample)

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted values  (model outcome) |  | Material damage | Personal injury |
| Material damage | TN | FN |
|  |  |  |
| Personal injury | FP | TP |
|  |  |  |

Apply to our case we will have :

True positives(TP): Number of accidents contained in the test sample with human losses impact and predicted as such.

False Positive(FP): Number of accidents that being only related to material damage are wrongly predicted as human damage.

False Negative(FN): Number of accidents that being accidents with human losses impact are predicted as material damage.

True Negative (TN): Number of accidents contained in the test sample related to accidents with material damages and predicted as such.

**Precision.**

Cases that are declared to be creating human injuries but what percentage of them are actually creating human injuries.

**Recall**

All the points that are actually creating human injuries but what percentage declare creating human injuries.

**F1 – score.**

The average F1 is a measure to evaluate test accuracy in a binary classification system. It is a weighted average of the precision and recall. When F1 score is 1 it’s best and on 0 it is worst.

Summary report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Avg F1 – score** | **Severity value** | **Precision** | **Recall** |
| **K nearest neighbou** | **0.59** | **Property damage** | **0.69** | **0.57** |
| **Injury** | **0.49** | **0.62** |
| **Decistion Tree** | **0.59** | **Property damage** | **0.69** | **0.57** |
| **Injury** | **0.49** | **0.62** |
| **Support Vector Machine** | **0.59** | **Property damage** | **0.57** | **0.69** |
| **Injury** | **0.62** | **0.49** |
| **Logistic Regression** | **0.59** | **Property damage** | **0.57** | **0.69** |
| **Injury** | **0.62** | **0.49** |

What can be observed out of the simulations and results of the four models, is that the four present the same F1 – Score 0.59.

On the other hand, it is possible to distinguish a parallelism in terms of behaviour about the precision and recall indicators.

The KNN and Decision Tree algorithms present 0.61 and 0.59 precision and recall weighted averages, while Logistic regression and Support Vector Machine do with 0.59.

# **Conclusion**

Based on the results section the four models under the conditions of the dataset produce a very similar outcome, there are not huge differences in terms of accuracy precision and recall.

If one needs to choose a model, will go to the one where there is a better precision averaged concerning the two potential values, which represent a more balanced outcome.

However, being KNN and Decision Tree the potential candidates, it is relevant to highlight that for KNN the best accuracy is reached when the K factor is around 50. This constraint requires a longer computational performance and in the long run when the dataset evolves during the years will cause a problem. All in all, the best option could be to use the Decision tree under the simulations constraints.

Despite the accuracy of the 4 models is above 50% there is still in all of them a 40% probability of wrong prediction. This is high in a so delicate matter like accidents which could cause personal injuries. Under the recommendation section some advices will be provided in order to reduce these problems and try to produce in the future algorithms with better performances in terms of accuracy.

# **Recommendations**

**Data collection:**

The Traffic Management Division will need to improve the data collection process. As observed during the data preparation process the data have incompleteness and a lot of records were not useful for the analysis. This lead in the end to problems at the time to produce predictions, since the models were not enough accurate for such a delicate matter.

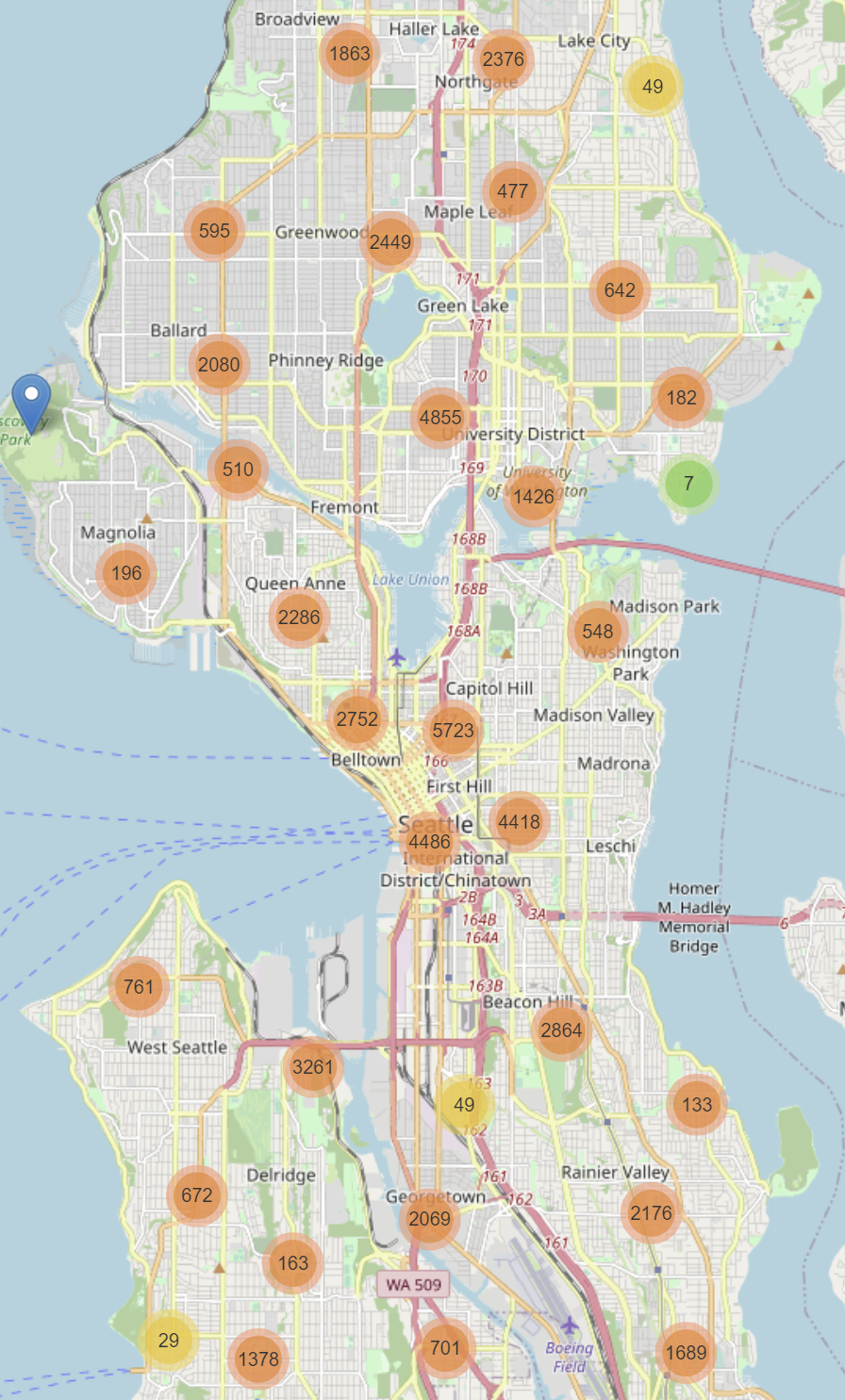
A good procedure could be that every time an accident occur, via an app an electronic form resuming the accident details is submitted by the competent party either local police on site or insurance companies responsible parties.

In there the fields which could help to clarify the accident occurrence, like location, distraction factor, overspeed etc should be mandatory. In this way we could add further variables into the models and get a clearer picture of the situation when the accident occur.

**Alerts when driving**

From the data exploratory outcome and the map below. It is clear, that a lot of accidents are occurring when the environmental factors are good, and specially in certain spots of the city. The Traffic Management Division should develop an alert system to communicate with the drivers in their cars highlighting its action specially when conditions are good.

This can be a system that sent a push communication to the driving dashboard control system or covering older car models segment too via a panels system along the road that really alert the drivers to be focus. Together with action, an increase of police surveillance around the critical spots.

Accidents distribution

# References

General:

https://worldpopulationreview.com/us-cities/seattle-wa-population

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

http://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019\_Traffic\_Report.pdf

<https://www.colburnlaw.com/seattle-traffic-accidents/>

https://www.wsdot.wa.gov/mapsdata/crash/pdf/2015\_Annual\_Collision\_Summary.pdf

Algorithms

<https://en.wikipedia.org/wiki/Decision_tree>

<https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248#:~:text=Decision%20tree%20builds%20classification%20or,decision%20nodes%20and%20leaf%20nodes>.

<https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7#:~:text=KNN%20algorithm%20is%20one%20of,the%20most%20used%20learning%20algorithms.&text=KNN%20is%20a%20non%2Dparametric,of%20a%20new%20sample%20point>.

<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

Evaluation:

<https://medium.com/analytics-vidhya/accuracy-in-machine-learning-bb3ddbc984b>

<https://medium.com/datadriveninvestor/confusion-matric-tpr-fpr-fnr-tnr-precision-recall-f1-score-73efa162a25f>