**Introduction**

In Seattle exists a lot of accidents by year. That include a various collision types, in different localization of the city, people involved, transport modes, datetime, etc. We need to understand which variables influence in accidents so we can take actions. For example, a certain intersection can present a lot of accidents with a lot of mortality, so we can advertising to modify it. For these reason we want to predict the severity of a collision. The objective of that, is to know how affect a new street (or an old modified), in the accidents so we can decrease the mortal accidents.

It will be used a Supervised Machine Learning Algorithm to do that. In specific i'll use classification models like K Nearest Neighbor, Decision Tree, SVM and Logistic Regression. With that, i'll choose that present the best index parameters.

**Data**

I had around 200.000 rows in which one represents the accidents in Seattle and a categorization of him severity.
In specific I'll use the following variables:

- Severity Code: Dependent variable, which one I will find to predict (Property damage or Injury damage)
- Weather: Weather conditions in the event
- Road conditions: Conditions of the road in the event
- Light conditions: Conditions of the light in the event
- Under influence: Represents if a driver involved in the event was under the influence of drugs or alcohol
- Collision type: Type of collision.

With the filter dataset, I worked with the NaN values and balanced the problem to improve the models. So, I had around 100.000 rows to construct the prediction.

**Methodology Section**

I analyzed the data. I found that the "Property damage severity" appears a lot (140.000 rows approximately). In the other hand a "Injury damage severity" had less information (60.000 rows approximately). So, I decided balanced the data.
Then analyzed the time. The behavior of the accidents in a day of the week is not too much different of the weekend behavior, so I decided not to separate the data with this parameter. Also, I made an analysis of hour by accidents. The number of accidents increase a lot near at 00:00 of midnight (like four-six times higher than other hours). That is strange, because in the night are less cars, so I decided keep calm with this.
I had some NaN values, that represent the 2% of the data, so I decided to eliminate, because is a small number.
Finally, I convert the categorical variables into numerical with the objective of build the models.

In the model section, I normalize the data so a certain number can't influence more than another variable. I divided the data in a train set (represents the 70%) and a test set (represent the 30% of the total).
I worked with a Supervised Machine Learning. Specifically, algorithms of classification like Logistic Regression, Decision Tree Classifier, Support Vector Machine and K-Nearest Neigbour (KNN).
I trained the models with the training dataset (fit) and validate with the test set (predict). The index that I used to compare the results were Jaccard index, F1-score and LogLoss.

**Results Section**

The results of the models are in the next table:

| Algorithm | Jaccard | F1-Score | Log Loss |
|---|---|---|---|
| KNN | 0.65 | 0.64 | - |
| Decision Tree | 0.70 | 0.70 | - |
| SVM | 0.66 | 0.65 | - |
| Logistic Regression | 0.53 | 0.53 | 0.67 |

Table: Parameters of the models

As we can see, the best models is that has a best fit of the data, in other words, a high Jaccard index (near to 1) and F1-Score (near to 1). The Logistic Regression is the worst model.
In the other hand, the Decision Tree Model is the best with a performance of 0.7 for both parameters.

**Discussion Section**

We conclude that the best way of construct a predict model is using a Decision Tree Algorithm. With that we use this model to save lives. For example, if someone decide to modify certain road, we can predict if the possible accidents are" property damage" or "injury damage" based in the lights that the road should be. Also, this model helps to predict some police action, like control and check if drivers are under the influence of drugs or alcohol. For example, ¿How much invest in road controls? Well, we can valorize the benefits of convert into severity 2 to 1 and compare with the cost of control a road.

**Conclusion Section**

This project helps me to improve my writing English skills and my software skills. Now, because the capstone is a real practice case, I see more opportunities to make things and gives solution to the society.

The Machine Learning Algorithms efficient, increase a lot with the number of columns (I wanted to use more variables but the KNN and SVM doesn't support less than 15 minutes). But is a very powerful tool to use.

The best model founded was the Decision Tree with parameters index of 0.7, that represents the performance of the algorithm. This is a high number and I conclude that we predict with a certain grade of probability real cases in the real world.

This can help to take decisions: Invest in some actions? What if the benefits if a change some variable? Need I be carefully in certain weather conditions?