**Capstone project report**

1. **Introduction**

***Task: Clearly define a problem or an idea of your choice. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.***

The aim of this study is to predict the severity of a car accident. This would be a tool that, given the weather, the road conditions and other variables, could warn surrounding hospitals about the possibility of getting a severe and be prepared for it. It would also warn people to drive more carefully or even change their travel if they are able to.

In order to study this problem, I used the dataset given in the Coursera Capstone project: Data-Collisions.csv that contains Data of collisions in Seattle.

1. **Data**

##### *Task: Describe the data that you will be using to solve the problem or execute your idea. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using.*

We have just two classes. This will simplify our job with the classifier later. However, in the meta data there is an explanation for the severity code:

1) property damage

2) injury

Now let's analyse the features and select the ones that are appropriate for the problem. In order to be suitable for classification problems, features should be categorical variables. In case we have some useful feature that is not categorical, we could still get dunny variables from them, so we do not exclude non-categorical variables a priori.

I have selected and corrected for the NaN values these following columns:

PERSONCOUNT 194673 non-null int64

PEDCOUNT 194673 non-null int64

PEDCYLCOUNT 194673 non-null int64

VEHCOUNT 194673 non-null int64

INCDTTM 194673 non-null object

SDOT\_COLCODE 194673 non-null int64

INATTENTIONIND 194673 non-null int64

UNDERINFL 194673 non-null int64

WEATHER 194673 non-null object

ROADCOND 194673 non-null object

LIGHTCOND 194673 non-null object

PEDROWNOTGRNT 194673 non-null int64

SPEEDING 194673 non-null float64

1. **Methodology**

I decided to use the logistic regression model, since our target value is binary and it is much faster than the other classification models.

1. **Results**

The model gives a good prediction given the features. I varied the c parameter in order to get the best result. For small datasets, ‘liblinear’ is a good choice for the solver, so I kept it that way. I tried to solve this problem also with KNN, but the computational time was terribly slow.

1. **Discussion and Conclusion**

In this study I analysed the relationship between the severity of an accident and some features of the accident such as for example, light condition, weather, number of car involved, some causes of the accident like speeding.

I reached a good result with an accuracy of 0.75. this model could be very useful to alert hospitals that that day, in those particular conditions, it is expected a larger number of severe accidents and thay can prepare in advance toold and personals required.