1 Business Problem:

Traffic accidents and congestion are one of not that old problems big cities must confront in modern days, where more and more people acquire vehicles and their use is contantly increasing. The reference to this topic as a problem is due to not only the high density of citizens living in big cities, adding the subsequent high number of cars, but the amount of people that visit the city for various purposes as working, tourism, etc. In the end, these situations provoke from traffic congestion, as the least severe consequence, to accidents with material losses, injured people or even deaths, in the worst case scenario. In general basis, governments are concerned about these tragic events because they consumed resources, as money for reparations or healthcare material, human lives in some cases and, in the case of companies, it can also mean a decrease of productivity, due to the workers arriving late due to congestion. For that reason, dealing with the issues that come with traffic is a hot topic in areas with high density of people.

In that regard, being able to understand the features that trigger those mentioned traffic problems is key in the effort of finding solutions to those issues in the future, saving money for the governments, problems to the rest of the citizens and even lives.

2 Data:

The main source of information that has fed the attempt to solve the previously mentioned problem of traffic issues is a database of given by the state of Seattle that is actually a record of traffic incidents that happened in the city. The complete database consists of almost 200000 traffic incidents that were recorded alongside 37 features of information about each of them, inside the "information" can be found the "degree" of severity of the incidents, the location, the traffic offenses that may have occurred (as speeding or being influenced by drugs or alcohol), the weather, the condition of the road, etc. The usefulness of the attributes, and which ones will be relevant for the project, is discussed in the next section.

3 Methodology:

After a glimpse to the database and a logical thinking approach, considering the nature of the incidents, a small subset of attributes was chosen where some of them, on one hand, the driver or drivers have no control over as, for example, the weather conditions, the collision type or the condition of the road. On the other hand, several attributes reflect the variables that can fall under the control of the driver in an accident and must be taken into consideration, those could be the driver being influenced by alcohol or drugs, speeding on the road or being distracted, for example. All of the mentioned circumstances can play a significant role in an accident and the model needs to take them into account.

Reviewing the project and after taking a quick look to the database, it makes clear that this is a classification problem for a machine learning model. In fact, due to the non-binary classification label, two models were built for comparison purposes: a *K-Nearest Neighbors* classification algorithm (*KNN* henceforth) and a Support Vector Machine (referred as SVM) one are chosen.

4 Results:

After evaluating the best k value for the model (between a sample of 50) the value of k that gave the highest accuracy test for the KNN algorithm was $k = 44$. After training and testing the model with the highest performance, the displayed result was accuracy score of ≈ 0.753 and, just for consistency purposes, a Jaccard index of ≈ 0.753 and a F1-score of ≈ 0.843.

Regarding the SVM model, the model accuracy is ≈ 0.755 alongside a Jaccard index of 0.755 and a F1-score of 0.847.

5 Discussion:

Despite being showed exclusively the best models achieved for both algorithms, several additional tests were made including and excluding attributes, as the importance that the number of vehicles or the number of people involved may have. However, in all the extra tests that were performed, the accuracy score always lied between 0.725 and 0.755 for both models, which can lead to the conclusion that these are variables with little impact in an

accident.

6 Conclusion:

In the attempt of getting a model that could accurately predict the severity of a traffic accident, based on the database of traffic incidents provided by the state of the city of Seattle, two machine learning models of classification were built: a *KNN* and a *SVM* algorithm. Both models achieved an overall accuracy score of $\approx 0.753 - 0.755$ which, far from being ideal, gives an acceptable success predicting the severity of an accident.