1 Business problem

According to data from the Washington State Department of Transportation, there were 120,993 accidents in 2017, of which more than 500 were fatal accidents and almost 2000 had people with serious injuries. It is estimated that there are about 1 death every 20 hours in the state of Washington and 1 accident every 4 minutes. Several factors can cause accidents and other factors can influence how fatal this accident can be or not, such as track conditions, vision level, drug use, among others. [1]

With this in mind, the focus of this project will be to try to predict the severity of a future accident, based on data from past accidents. This could help transit agencies, as well as hospitals and emergency systems to help those involved.

2 Data understanding

The data we will use comes from the SDOT Traffic Management Division, Traffic Records Group, in CSV format. The data are from Seattle, Washington, and date from 2004 to the present day, they are updated weekly, so afterwards, we can make our model absorb more future data, both to improve accuracy and to look for new trends in the data. and deliver a better result. Within the CSV file we have columns that represent different types of data, such as the date of the accident, track conditions, number of accidents, number of cars involved, pedestrians, cyclists, etc. We also have many columns that serve only to identify each accident by bodies that deal with this type of work. In addition, we have many empty and unknown values that we need to deal with some techniques. We can have more information for each column in the METADATA.pdf file to know what each variable is for. The column we need to predict is the "SEVERITYCODE" which has the value of the severity of the accident, the higher, the greater the severity. Through some analysis, we can see that currently we will only deal with two values: 1 or 2 (damage only to properties and people injured in the accident, respectively). Our data set is unbalanced, so we will need to balance it to use for fairer predictions.

We will use and evaluate 3 Machine Learning models to predict the results and choose the one that has the greatest accuracy to be used.

3 Methodology

We will use the Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn libraries to analyze, organize and create our prediction models.

After organizing the data, we will use Matplotlib and Seaborn to see how the data behaves in representative graphs and to see if the variables are correlated or not with the dependent variable.

3.1 Exploratory data analysis

Our dataset has several columns that are only for identification, so we will remove it and columns such as date and location, so we will have a more general dataset with columns with greater impact. In addition, we have columns missing many values, we will remove those that have more than 50% of missing values and we will use techniques to fill in missing values in other columns, such as filling with the most recurring value.

Because the columns are categorical, we will need to clear the data and turn it into numeric values, so we will use two techniques to deal with these variables: one-hot encoding and label-encoding. [2]

3.2 Feature Selection

After processing the data, we had 21 columns, a bit high value, so we will only select the columns with values that have more influence on the dependent variable using the SelectKBest function of the Scikit-learn package.
As our dataset has more samples with a severity code of 2, we will also need to reduce it to balance. We will randomly choose a set of samples from the dataset with equal severity values, using the Under-sampling (Down sampling) technique [3]. This will prevent him from training with more data of one type than another.

3.3 Data Scaling / Standardization

After choosing the columns, we will scale the values so that they are more representative of each other, as we have columns with very high values and other columns with small values, this influences the final result of the models and also influences the time for the algorithms to run [4]. In this case, we decided to use the min-max scale technique.

3.4 Modeling

In the end, we will separate the data set into 4 smaller sets: X values for training, X values for testing, Y values for training and Y values for testing. Thus, we will be able to train our models and test the accuracy of each one to choose the best one at the end.
Running all types of Machine learning algorithm and seeing which one has the best result takes a lot of time and resources so, based on graphical representation of the data, I decided to use three that fit our data better: K-nearest neighbors, LinearSVM and Random Forest [5]. Logistic regression has been neglected due to the way the data behaves and traditional SVM is very slow and does not scale well for datasets with tens of thousands of samples. [6]
In each algorithm we use GridSearchCV, which serves to test several different parameters in each model and returns us which are the best parameters to use in the final result. [7]

4 Results

In the end, the best algorithm for this project was Random Forest, with an accuracy of around 69.86%. Due to the number of samples in the dataset, we can consider this a good result for our predictions.

5 Discussion / Conclusion

As stated earlier, this project can be used by healthcare systems, hospitals and authorities to deal with victims and even save the lives of patients by preparing everything in advance to receive injuries. This project can be retrained in the future with more data to have better accuracy and help with more accident cases. It can even be extended to other locations, based on the data received.

6 References

[1] https://www.injurytriallawyer.com/library/car-accident-statistics-seattle-washington-state.cfm

[2] https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

[3] https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5

[4] https://machinelearningmastery.com/feature-selection-with-categorical-data/

[5] https://www.youtube.com/watch?v=38SUUaMX5Rg

[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

[7] https://www.youtube.com/watch?v=CgmvAMiVKFE

[8] https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd