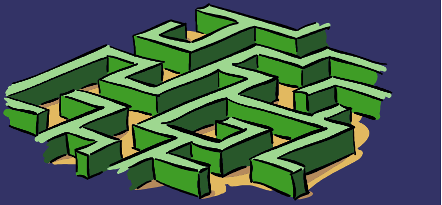
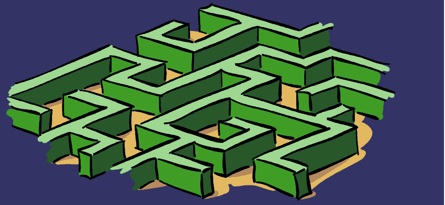
Accident Severity Prediction

Applied Data Science Capstone



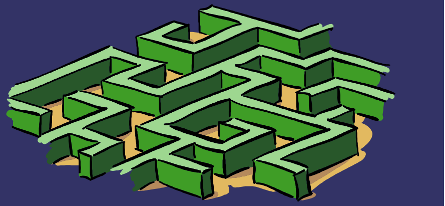
Overview

- The goal of this project is to use Supervised learning techniques to predict the severity of an accident
- The long term goal is to use the data to warn motorists of hazardous conditions that could cause an accident involving injury or death



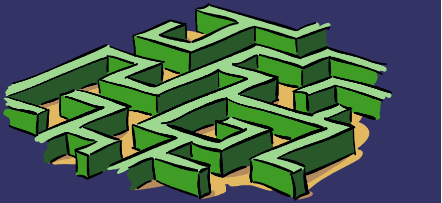
Long-term goal

- The goal is to collect and use real time data attributes that could be fed to a model such as weather, road conditions, light conditions speeding etc. to predict the likelihood of an accident.
- Motorists could then be warned of such conditions and re-routed if necessary



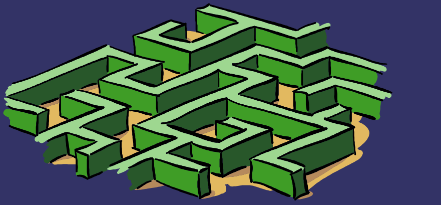
The Present Situation

- The current situation deals with a data set of accident severity in and around Seattle city.
- The data label Severity-code will be used as the target label
- The rest of the labels in the data set will be used as predictors where applicable



Development of the Data

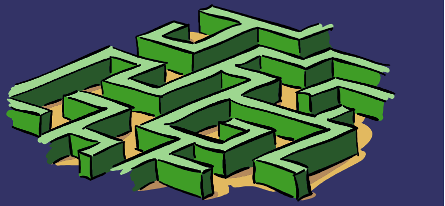
- The data set has 37 labels that could potentially be used to predict accident severity
- The data required Pre-processing balancing, and cleaning
- Most of the data types were of type “Object' and needed to be converted to integers.
- There was missing data. It was decided to remove the missing data as opposed to filling it in with the mean or Frequency



Methodologies Used

- Once the data was properly prepared three methods were used for prediction

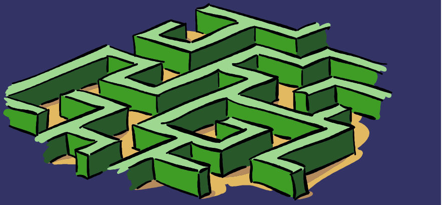
- KNN Nearest Neighbor
- Decision Tree
- Logistic Regression



Predictor Labels Used

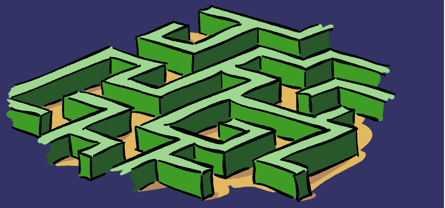
- The predictor labels were decided as

- Weather
- Road Conditions
- Lighting Conditions
- Speeding
- Address Type
- Person Count
- Vehicle Count
- Junction Type



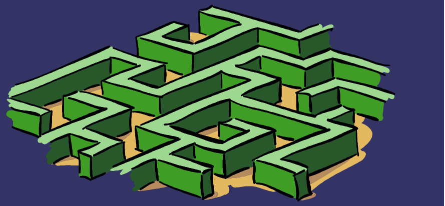
Modeling prep and Methods

- For each method the data was converted to a Numpy array and labeled. The data was-normalized
- The data was split into a training and test set
- The model was created
- The prediction was made
- The accuracy was calculated



Results

- In the end the accuracy for each method was close
- The accuracy for the decision tree was the highest at 63% I would have expected KNN to be a bit higher
- The data set, being a sample seemed in-complete and may have effect accuracy
- The inclusion of “less pure” labels may have accuracy



Conclusion

- For the purpose of accurately predicting accident severity a more complete data set should have been selected
- More attention to some of the attribute labels and how they effect the model may have improved accuracy