Introduction

Road traffic accidents are estimated to cost the US economy ~$٨١٠bn per year in property damage, medical costs, legal bills and loss of earnings. Identifying the factors which influence accident severity is therefore of paramount importance.

It may seem intuitive and obvious that the occurrence and severity of road traffic accidents will be related to a number of factors including environmental conditions (weather, road surface conditions and lighting conditions), time of day/day of week, irresponsible driving (speeding, driving under the influence of alcohol/drugs, or simply not paying attention) and the angles and speeds at which vehicles collide. However, determining the manner in which these various factors interact to determine the severity of car accidents in a given state/city is not a trivial undertaking. While a number of these features may individually serve as good predictors of accident severity on their own, some of them seldom occur together: for instance, head-on collisions often involve serious injuries and fatalities, as do collisions on freeways (primarily due to the high speeds involved). However head-on collisions on freeways are relatively rare, due to the presence of barriers separating traffic travelling in different directions. How can we incorporate these features in to a cohesive understanding of road accident outcomes?

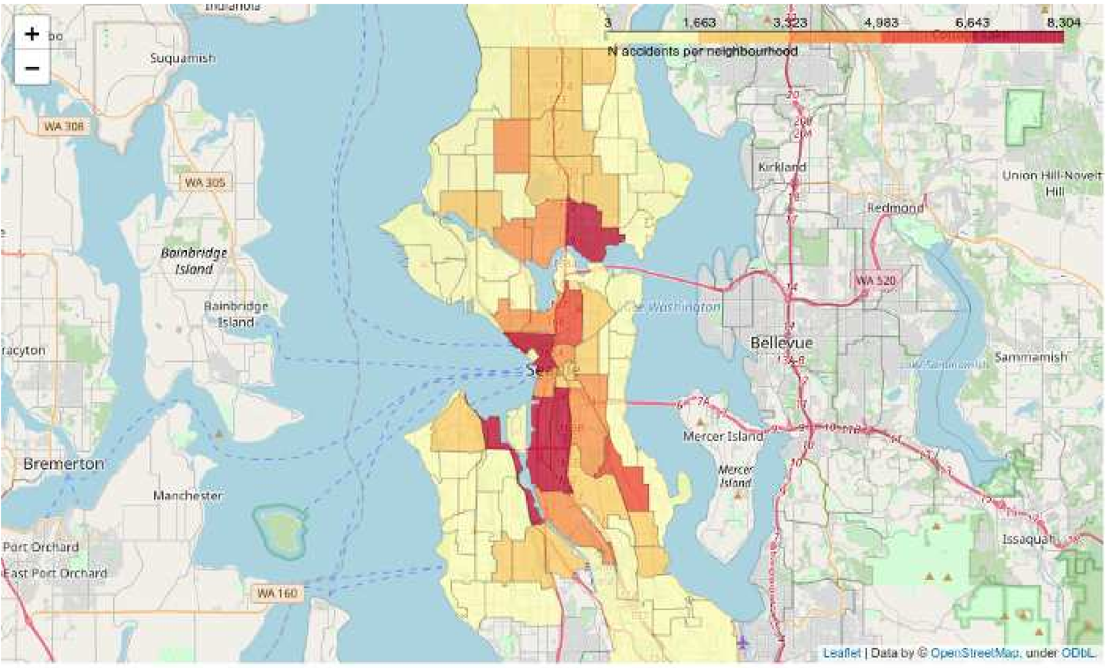
We can begin to unpick this complicated problem and gain insight into the factors which influence accident severity by studying data from past accidents, and then use Machine Learning techniques to create models to predict the severity of future/unfolding accidents based on the similarity of their initial conditions to those of other accidents in the historical record.

The two main beneficiaries of building this kind of model are (١) town/city planners, who may be able to use the model to inform their road planning and traffic calming strategies and (٢) emergency service responders, who may be able to use the model to predict the severity of an accident based on information that’s provided at the time the accident is reported in order to optimally allocate resources across the city.

Data

A comprehensive dataset of ٢٢٦،٠٠٠ accidents occurring between ٢٠٠٤–٢٠١٩ in the Seattle city area was obtained from the Seattle Open Data Portal[1]. The ٢٢٦،٠٠٠ row dataset has ٤٠ columns describing the details of each accident including the weather conditions, collision type, date/time of accident and location (latitude and longitude).

A Choropleth map (created using the Python Folium package) reveals, perhaps unsurprisingly, that accidents occur more frequently towards the centre of the city, and in neighbourhoods at either end of road bridges which straddle Seattle’s major waterways. For instance, University District and Eastlake (separated by the Ship Canal Bridge) report a higher incidence of accidents than average, as do the Industrial District and Industrial District West (separated by the West Seattle Bridge).

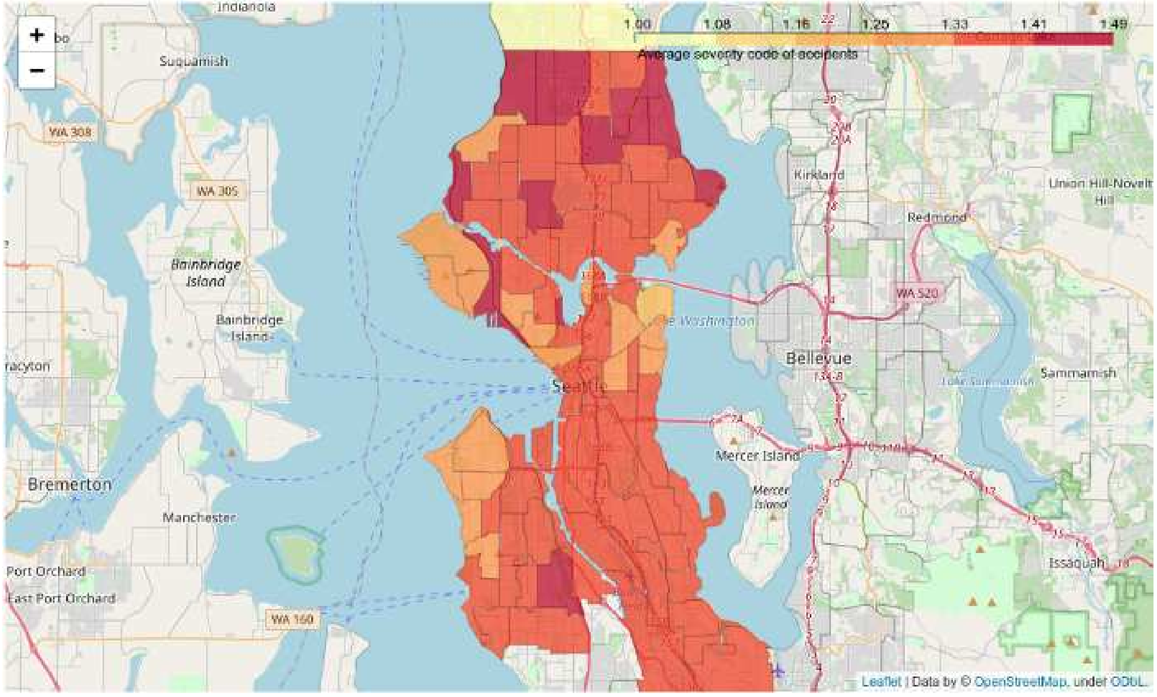


Choropleth map showing the number of accidents per neighbourhood between ٢٠٠٤–٢٠١٩ (Data source: Seattle Open data Portal)

One of the ٤٠ columns in the dataset encodes the Seattle Department of Transport (SDOT) accident severity metric, according to the following schema:

- ٠: Unknown/no data
- ١: Property damage only
- ٢: Minor injury collision
- ٢b: Major injury collision
- ٣: Fatality collision

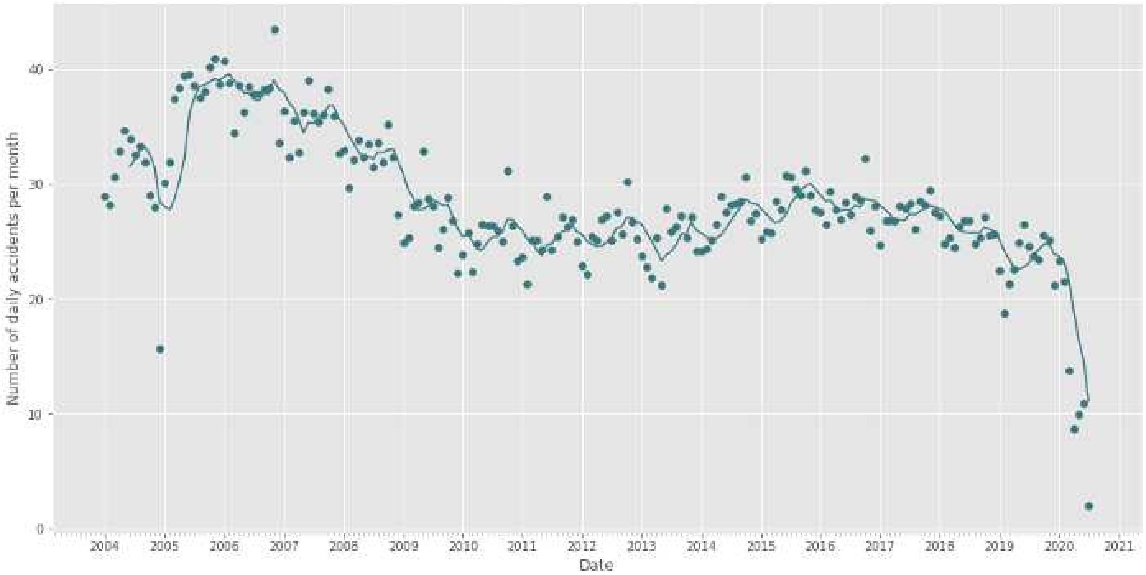
Using the severity codes assigned to each accident, we can identify the neighbourhoods with the highest mean accident severity codes:



Choropleth map showing the mean accident severity per neighbourhood (Data source: Seattle Open Data Portal)

We see that while the suburban neighbourhoods to the North of the city (e.g. Bitter Lake, Pinehurst, Olympic Hills) have lower rates of accidents than downtown neighbourhoods, the mean accident severity in these less densely populated neigbourhoods is higher. From this we may infer that traffic congestion in the city centre limits the potential for severe accidents to occur, whereas on less congested roads (with higher average speeds), accidents are more likely to result in injury or death.

Using the timestamp data, we can search for temporal variations in the rate of road accidents in Seattle:



Number of accidents per day for each month in the Seattle accident dataset (six month rolling average shown with a solid green line). Note the apparent drop-off in the early months of ٢٠٢٠ is likely due to a delay in reporting (Data source: Seattle Open Data Portal).

We see that the daily accident rate reached a peak of ~٣٨ per day in ٢٠٠٦–٢٠٠٧ before declining to ~٢٥ per day between ٢٠١٠–٢٠١٩. From the six-month rolling average we see a tendency for the accident rate to peak in Q٤/Q١ of each year, when the daylight hours are shortest, and reach a minimum in Q٢/Q٣ each year, when the daylight hours are longest.

Data Preprocessing

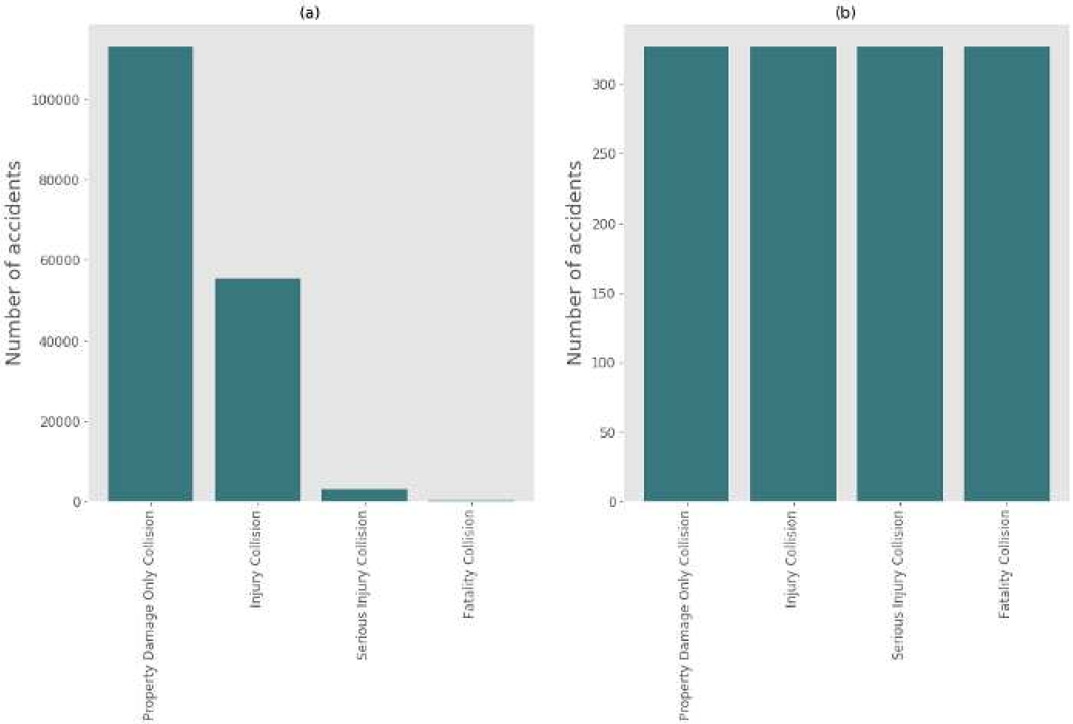
In their original form, the Seattle accident data are not suitable for quantitative data analysis. The main reasons for this are as follows:

١. **Data incompleteness:** around ١٥٪ of the accidents in the dataset are missing one or more key features, including in some cases the target variable (accident severity code) and in others, are missing information about weather or road conditions. As the purpose of building the model is to see how these various features interact and influence the overall accident severity, data entries which are missing one or more of these key features are not useful, and were removed from the dataset.

٢. **Dataset includes unnecessary/redundant columns:** the accident dataset includes many columns of metadata (such as incident report numbers) and columns which duplicate information which is already included in other columns (such as a text field “SEVERITYDESC” which provides a written definition of the accompanying accident severity code, the target variable). Columns which include unnecessary/redundant information were removed from the dataset.

٣. **Presence of categorical/non-numeric data:** Machine Learning models require numerical data, and cannot handle alphanumeric strings. For example, each entry in the “WEATHER” column contains a text string which takes one of eleven values (e.g. “Clear”, “Rain”, “Snow”, etc) which describes the prevailing weather conditions at the time of the accident. Columns such as this were re-engineered to allow data analysis using the “One-Hot Encoding” technique[2], which replaces the input textual column with a series of binary columns (containing ١ or ٠) relating to each possible value in the input column.

٤. **Data are imbalanced and non-standardised:** the target variable for this model is the accident severity code. Fortunately, the majority of accidents in Seattle (>٩٧٪) involve either no or minor injuries, and so the number of accidents with SEVERITYCODE=٣ is very much less than the number of accidents with SEVERITYCODE=١. However this imbalance between the real-life occurrences of different accident outcomes may bias the model if not accounted for. To create the least-biased model, we re-sample the dataset by randomly-selecting a number of accidents (٣٢٧) with severity codes ١, ٢ and ٢b which matches the total number of accidents in the dataset with severity code ٣. Furthermore, the data are non-standardised: most of the data columns after data engineering have values of the order ~١, however a few columns (e.g. X, Y, giving the longitude and latitude of the accident, and YEAR) have numerical values which are orders of magnitude higher. To allow the model to identify the key features without being biased by the imbalance in data values, the data were standardised (i.e. every column re-scaled to have ~zero mean and ~unit variance)



Accident severity codes before and after resampling (Data source: Seattle Open Data Portal)

In randomly re-sampling accidents with damage-only, minor injuries or major injury outcomes to match the number of fatal accidents, the frequency distributions of the road/weather/light conditions in which accidents happened were not altered, nor were the ratio of accidents occurring on week days versus weekends, or on different months of the year altered significantly[3]. We can therefore be confident that the re-sampled data is not biased.

Model Development and Evaluation

In order to develop a model for predicting accident severity, the re-sampled, cleaned dataset was split in to testing and training sub-samples (containing ٣٠٪ and ٧٠٪ of the samples, respectively) using the scikit learn “train_test_split” method. In total, four models were trained and evaluated.

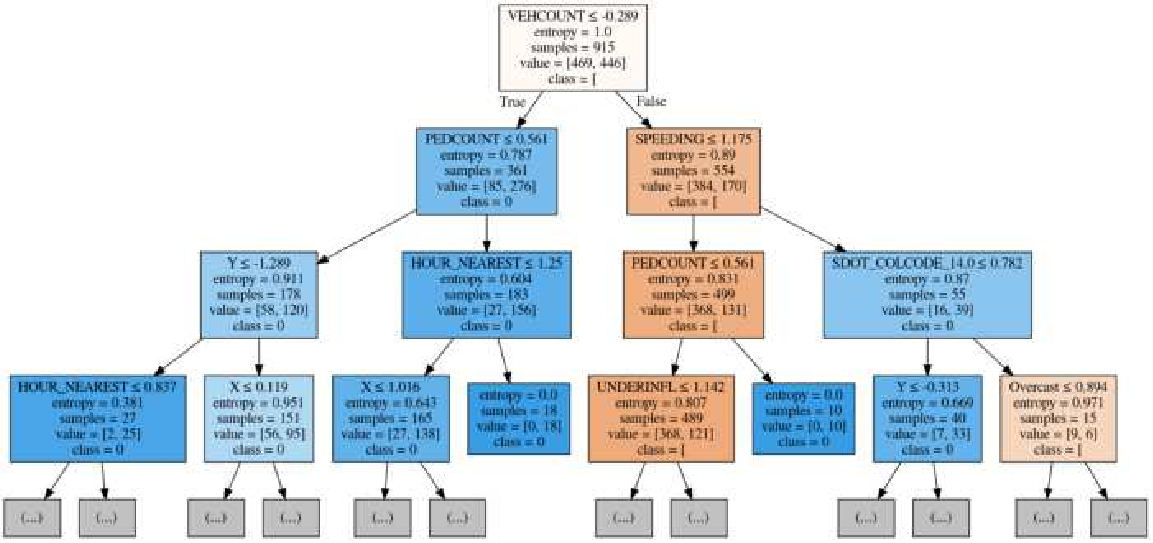
١. **Logistic Regression Model:** by converting the accident severity to a binary variable (٠ for no/minor injuries, ١ for major injuries/fatalities) we can employ Logistic Regression techniques to attempt to classify accident outcomes based on the properties in the feature set. A Logistic Regression model was trained using an inverse-regularisation strength C=٠,٠١, and tested on the testing subset. The Logistic Regression model correctly predicts SEVERITYCODE=١ accidents ٧٠٪ of the time, correctly predicts SEVERITYCODE=٠ accidents ٨٤٪ of the time, and has F١ scores of ٠,٧٧ and ٠,٧٦ for the two outcomes.

٢. **Decision Tree model:** decision tree models identify the key features on which the data can be partitioned (and the thresholds at which to partition the data) in the hope of arriving, after some iterations, at “leaves” which contain only accidents belonging to one target variable value (in this case, accident severity code). A decision tree model was trained on the data according to the “entropy” criterion[4], and allowed to run until covergence. The final decision tree is ٣٠ branch-layers deep (from the data feature set of ٦٠ features), the top four layers of which are shown below. The decision tree correctly predicts SEVERITYCODE=١ ٦٢٪ of the time, correctly predicts

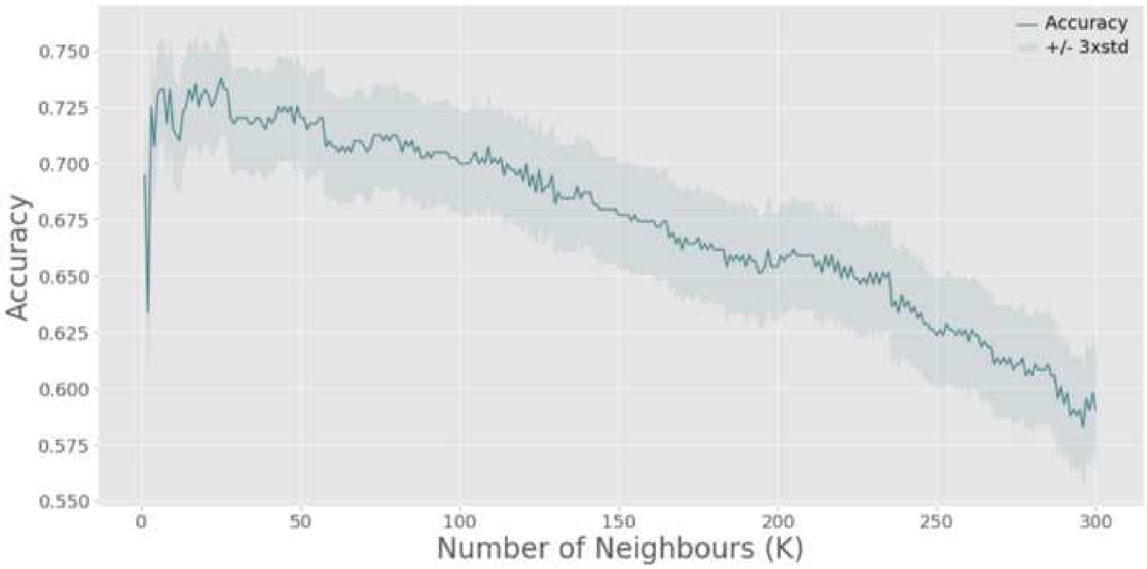
SEVERITYCODE=٠ ٦٩٪ of the time and has F١ scores of ٠,٦٦ and ٠,٦٧ for the two accident classifications. According to the decision tree model, the top three features for determining the likelihood of an accident having severe consequences are (i) the number of vehicles involved, (ii) the number of pedestrians involved, and (iii) whether or not one or more drivers were speeding.

٣. **Support Vector Machine (SVM) model:** SVM models seek to separate data based on different values of the target variable by mapping the dataset to a higher-dimension space and identifying the support vectors which best-describe the hyper planes that most effectively partition the data. An SVM model was built using the scikit learn C-Support Vector Classification method (svm.svc), with a linear mapping kernel employed in order that the model could return a list of the features with the most diagnostic power for determining accident severity. The SVM model correctly predicts SEVERITYCODE=١ ٧٠٪ of the time, correctly predicts SEVERITYCODE=٠ ٨٢٪ of the time, and has F١ scores of ٠,٧٦ and ٠,٧٥ for both accident outcomes. Like the decision tree model, SVM finds that two of the top three most crucial features are (i) the number of pedestrians involved, and (ii) whether or not a driver was speeding. SVM also identifies the accident type “head on collision; both vehicles moving” to be a major factor in accident severity.

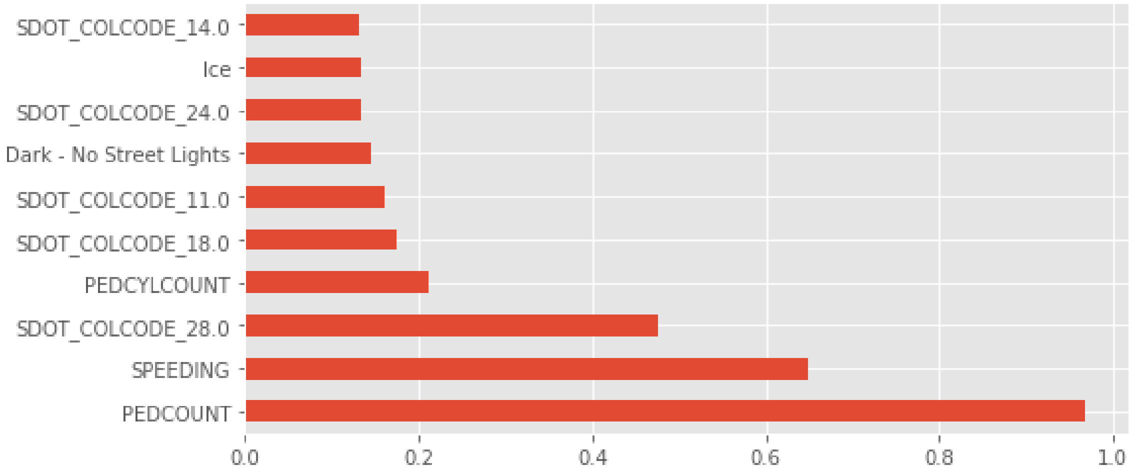
٤. **k-Nearest Neighbours (kNN) model:** kNN models seek to categorise the outcome of an unknown data sample based on its proximity in the multi-dimensional hyperspace of the feature set to its “k” nearest neighbours, which have known outcomes. Establishing the value of “k” which optimises the model’s accuracy (between ١ and the total number of samples in the dataset) is an empirical undertaking: if too-few neighbouring datapoints are used, the model is susceptible to being dominated by noise, however if too many neighbours are included in the classification, the model risks losing all diagnostic power completely. kNN models were built for k=١–٣٠٠ using the kNeighborsClassifier function from scikit learn. The model is optimised at k=٢٥ (see below), at which the model correctly predicts accident severity code ١ and ٠ ٦٢٪ and ٨٧٪ of the time, respectively. The F١ scores of the two accident outcomes are ٠,٧٦ and ٠,٧١.



Top four layers of the decision tree model, highlighting the features on which the data are most cleanly segregated in to “No/minor injury” and “Major injury/fatality” outcomes.



k-Nearest Neighbour model accuracy for ٠ < k < ٣٠٠. The model’s performance is optimsed at k=٢٥.



SVM model output showing the top ten features with the most diagnostic power for determining accident severity (ranked from least to most significant).

Conclusions and Future Work

Car accident data for the city of Seattle between ٢٠٠٤–٢٠١٩ have been used to train and evaluate machine learning models for predicting accident severity based on the circumstances of the accident. Four classes of models have been trained and evaluated: (i) Logistic Regression, (ii) Decision Tree, (iii) Support Vector Machine ant (iv) k-Nearest Neighbours. The Logistic Regression and SVM models perform best, having average F١ scores of ٠,٧٦٥ and ٠,٧٥٥, respectively. The kNN model has an average F١ score of ٠,٧٣٥. The Decision Tree model performs poorest, with an average F١ score of ٠,٦٦٥.

One of the major advantages of the SVM model is its ability (if using a linear mapping kernel) to de-project the fitted support vectors back to the original parameter space in order to determine which features in the initial dataset offer the most diagnostic power. The SVM model highlights that accidents involving pedestrians, accidents involving excess speed and head-on collisions between moving vehicles have particularly severe outcomes. Despite the tiny fraction of accidents in the Seattle area which occurred on icy roads, the SVM model was still able to determine that when conditions *are* icy, the risk of severe accident outcomes is increased.

This work highlights that machine learning techniques can be used to probe historical data in order to make reliable predictions about the outcome of road traffic accidents, given information which is available at the time when an accident is reported. This model can be extended to include new features or applied to accident databases in other cities/regions. By doing so, city planners can gain insight into the road conditions/features which are associated with high accident severity, and use this insight to improve road design. Additionally, by predicting accident severity as functions of weather, date, location and road conditions, this model may be able to help aid the decision making of emergency services call handlers, by allowing them to prioritise resources toward collisions with a greater likelihood of severe consequences.