IBM Capstone Project:
A Machine Learning Model for Accident Severity in the UK

Remington Oliver Sexton

Table of Contents

1 Introduction

The ability to predict vehicle accident severity is a logistical challenge for any country with modern transportation infrastructure, with the ultimate goal of reducing the human and economic costs of accidents of all types. For this project, I used publicly available traffic accident data (Kaggle) of 1.7 million accidents spanning from 2005 to 2017.

The goal of such a project is to determine predictors for accident severity to reduce the number of serious and fatal accidents, but also reduce the overall number of accidents. Reasons for performing such a study can be of interest to both government transportation and safety regulators to reduce the human cost of poorly designed infrastructure, laws, or vehicle standards. The study can also be of interest to non-government entities, either to improve safety standards in vehicles or be able to anticipate medical, legal, or business costs associated with accidents.

2 Data Acquisition

The Kaggle Dataset use for this project provides data for 1.7 million documented traffic accidents in the UK spanning from 2005 to 2017. Columns in this dataset include information on location (longitude and latitude), weather and road conditions, area types, posted speed limits, vehicle types, as well as other useful information. There are three target labels for classification, "fatal", "serious", and "slight" accident severity, which makes this a **multi-class supervised classification** problem.

The dataset is split into two separate tables, one for accident data and the other for vehicle data, which are combined into a single dataset by their unique accident identifier using the Python Data Science Library Pandas. The dataset also includes a `.geojson` file for mapping location data using folium.

3 Methodology

Below is described the exploratory data analysis, feature selection, and preparation necessary before any machine learning modeling can begin.

3.1 Exploratory Data Analysis

In this section the data is explored in its entirety to get both an understanding of the data itself and its meaning in the context of accident severity in the UK. This is necessary to identify features that would serve as good predictors of accident severity.

3.1.1 Accidents by Local Authority District

First, location data is used to determine if certain areas of the UK suffer from high occurrences of accidents, and possibly as a function of severity. To do this, the Python mapping module Folium is used to map `geojson` data by Local Authority Districts to the occurrence of accidents as a function of severity.
Figure 1 shows a choropleth map of the UK of all accidents regardless of severity. There is a surprisingly low frequency of accidents in London, compared to other populated areas such as Birmingham and Leeds. This could be due to better high-density infrastructure and more public transportation, however we can only speculate on the true reason for this.

Scaled choropleth maps are also constructed for the three types of accident severity, which are shown side-by-side in Figure 2 for comparison. Slight accidents comprise the majority of accidents in the UK, and are highly prevalent in densely-populated urban areas. Fatal accidents on the other hand are prevalent in both densely-populated urban areas as well as lower-population rural areas. The frequency of serious accidents are between both slight and fatal accidents. Here we can clearly distinguish the extremes of the target labels, with slight and fatal accidents as the extremes.

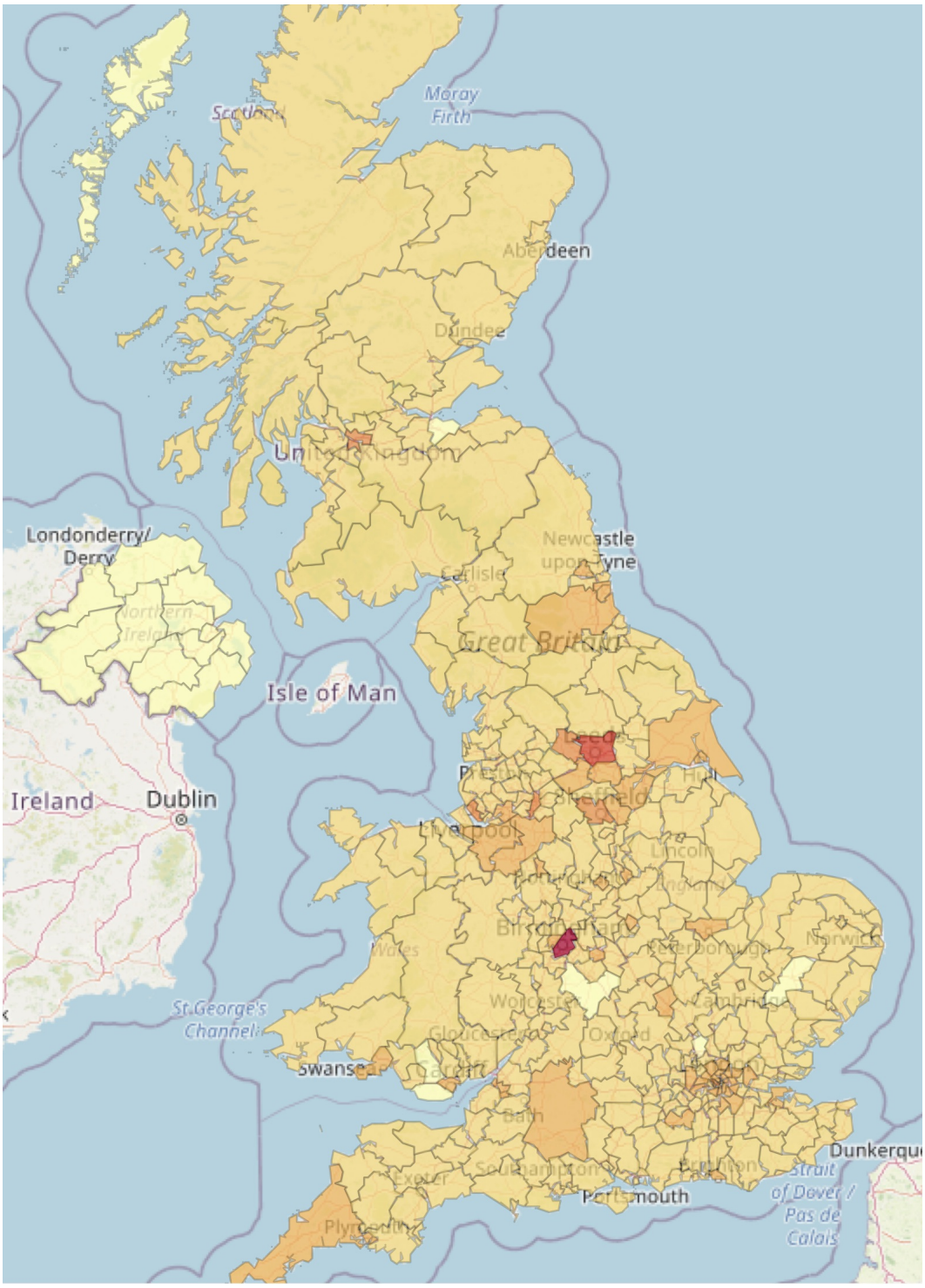

Figure 1: Choropleth map of frequency of all types of severity of accidents in the UK from 2005-2014. Surprisingly, London has a relatively few accidents compared to other populated areas such as Birmingham. This could be due to better high-density infrastructure and more public transportation, however we can only speculate.

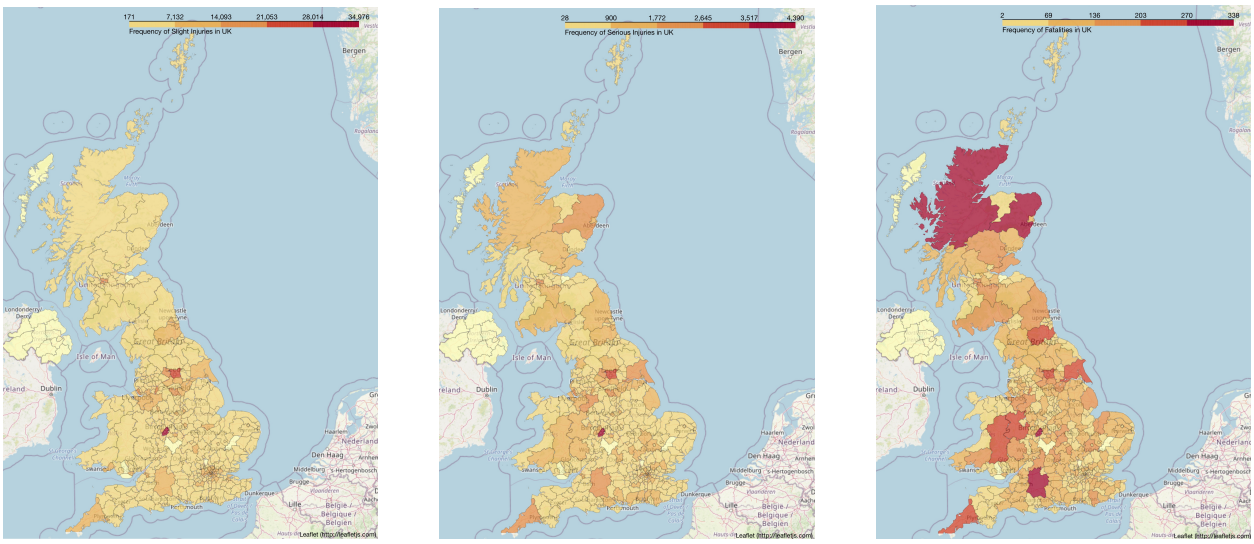

Figure 2: Scaled choropleth maps based on accident severity; (*Left*) Slight accidents, (*Middle*) Serious accidents, and (*Right*) Fatal accidents.

3.1.2 Timeline of Accidents from 2005-2017

Figure 3 shows a timeline of accidents in the UK over the course of the years from 2005 to 2017. This does not reveal any useful data regarding accident severity, but does show how the number of accidents increases over the course of 12 years. The increase may be due to an increase in the number of drivers or better accident reporting, however data to support these speculations is outside the scope of the dataset. Figure

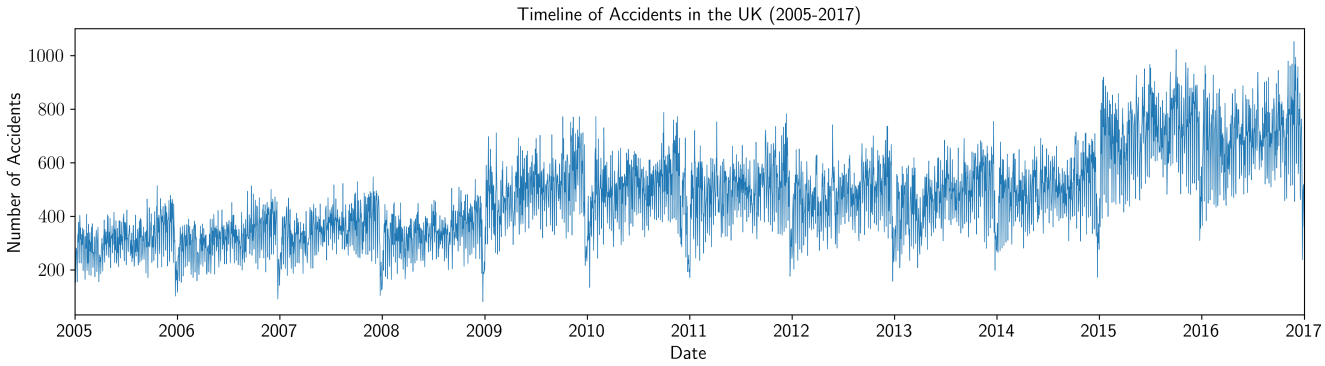

Figure 3: Total number of accidents by data from 2005 to 2017. There is a systematic increase in the number of accidents, however the reasons for this is outside the scope of the data.

4 shows the percentage of accident severity as a function of date. By reporting the number of each severity as a percentage of number of total accidents per date, we eliminate the systematic increase over the period of 12 years. This reveals that over this time period, the relative number of types of accident severity do not experience any large fluctuations, indicating that accident severity has not systematically increased or decreased in any category.

3.1.3 Accident Severity by Day-of-Week

Figure 5 shows accident severity as a function of day of the week. We calculate the percentage of each severity per day-of-week, such that the percentages for a given severity add up to 100% over the course of 7 days. We see no outstanding trends for day-of-week that would make this a good predictor of accident

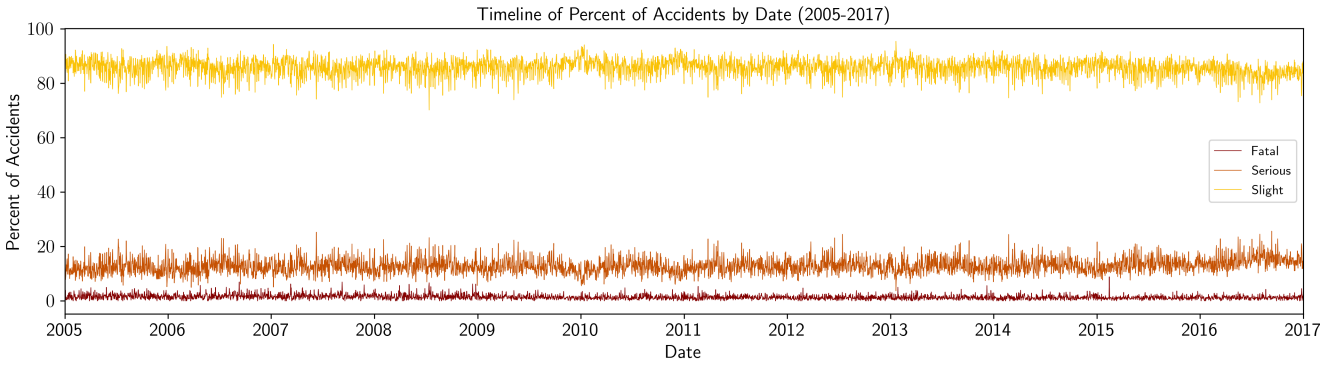

Figure 4: Percentage of accident severity as a function of date. Accident severity does not systematically increase or decrease despite the increase in the number of reported accidents over 12 years.

severity. Some noticeable trends however are that accidents are lowest on the weekend, and climb steadily as the week progresses, and peak on Friday.

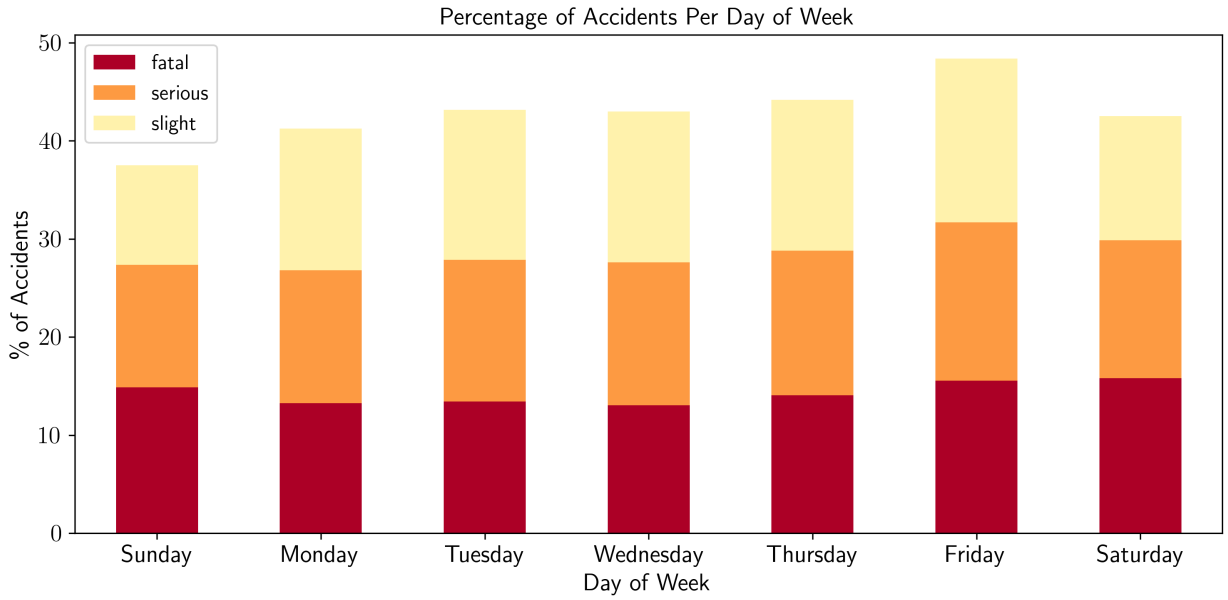

Figure 5: Accident severity as a function of day of the week.

3.1.4 Accident Severity by Time-of-Day

Finally, Figure 6 shows the kernel density estimation (KDE) of accident severity as a function of time-of-day. Again, there are no trends that would lend time-of-day as a good predictor of severity, since severity tend to track one another with time. We do see some interesting trends however, such as fatal accidents become more frequent than both slight and serious accidents between the hours of 12:00 AM and 7:00 AM. There are two peaks to all severity occurring during the morning rush hour (between 7:00 AM and 10:00 AM) and the evening rush hour (between 4:00 PM and 7:00 PM).

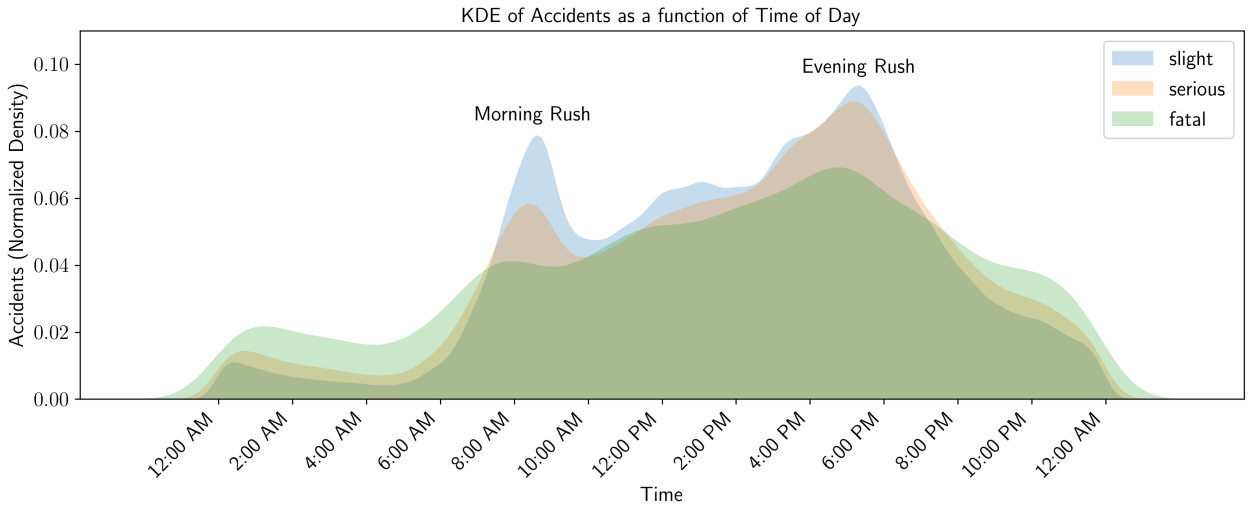

Figure 6: Accident severity as a function of time-of-day.

3.2 Continuous Features

The most useful continuous variable features in our data are time-of-day, engine size (CC), and age of vehicle. In Figure 7 we show box plots of each of these continuous variables as a function of accident severity. The box plots for vehicle age and engine size both show significant outliers, rendering these continuous features useless for predicting accident severity. The box plots for time-of-day, while show no significant outliers, do not show any significant statistical differences as a function of accident severity (as mentioned in the previous section). We therefore conclude that these continuous variables would not be suitable for our machine learning analysis.

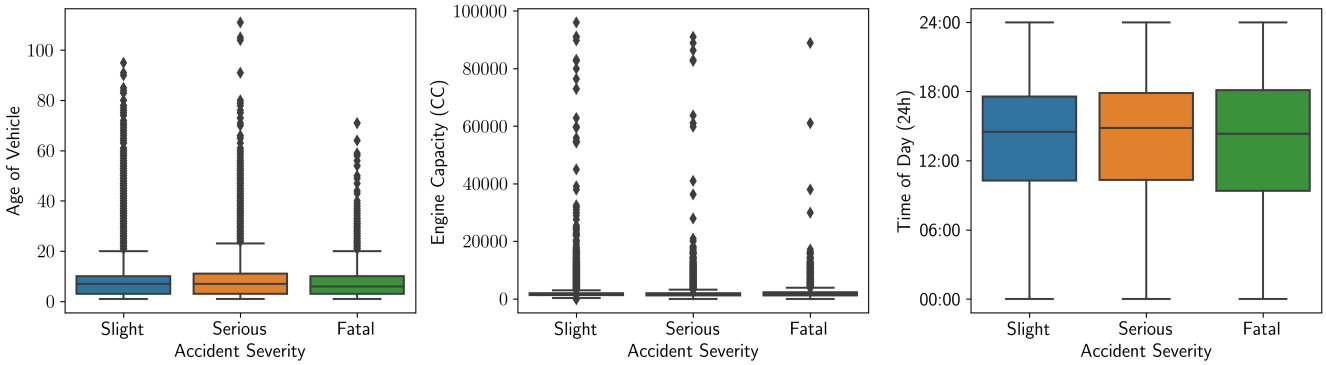

Figure 7: Continuous variables in our dataset. There are significant outliers or no significant differences as a function of accident severity.

3.3 Categorical Features

Below we investigate categorical features as predictors of accident severity. Because the dataset is very large (1.7 million accidents), it is feasible to remove missing data, "nan", or "unknown" conditions. Missing and unknown data constitutes a negligible fraction of the dataset.

3.3.1 Light Conditions

Figure 8 shows the percentage of accidents by light conditions as a function of accident severity. There is no difference in severity when it is dark and lights are lit. The largest difference can be seen when its dark and there is no lighting, showing a 15% difference between fatal and slight accidents. The opposite trend occurs in daylight, with more slight accidents than fatal ones.

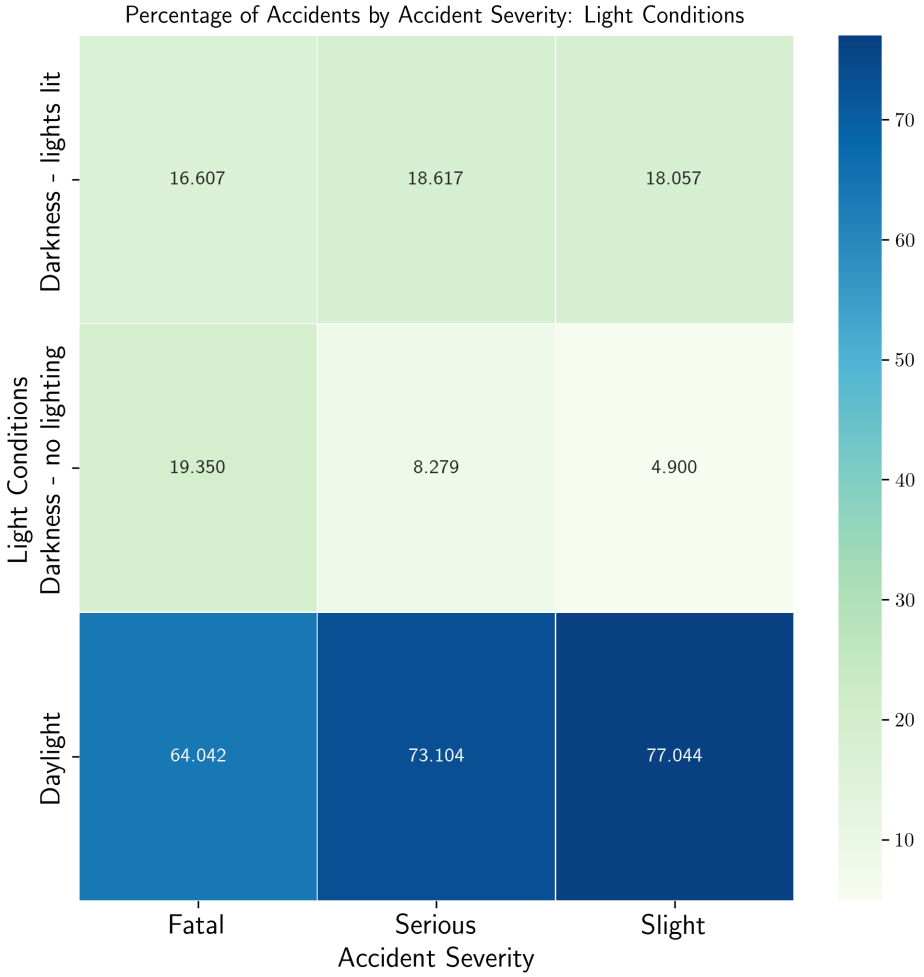

Figure 8: Percentage of accidents based on light conditions as a function of accident severity.

3.3.2 Weather Conditions

Commons sense would make it seem like weather conditions would play a large role in accident severity, however Figure 9 shows otherwise. In fact, during inclement weather (fog, rain, or snow), accident frequency actually decreases, and there is no trend with accident severity for any weather condition. This may be because drivers tend to drive more cautiously during inclement weather conditions, but drive less carefully when the weather is fine. Therefore, weather is *not* a good predictor of accident severity.

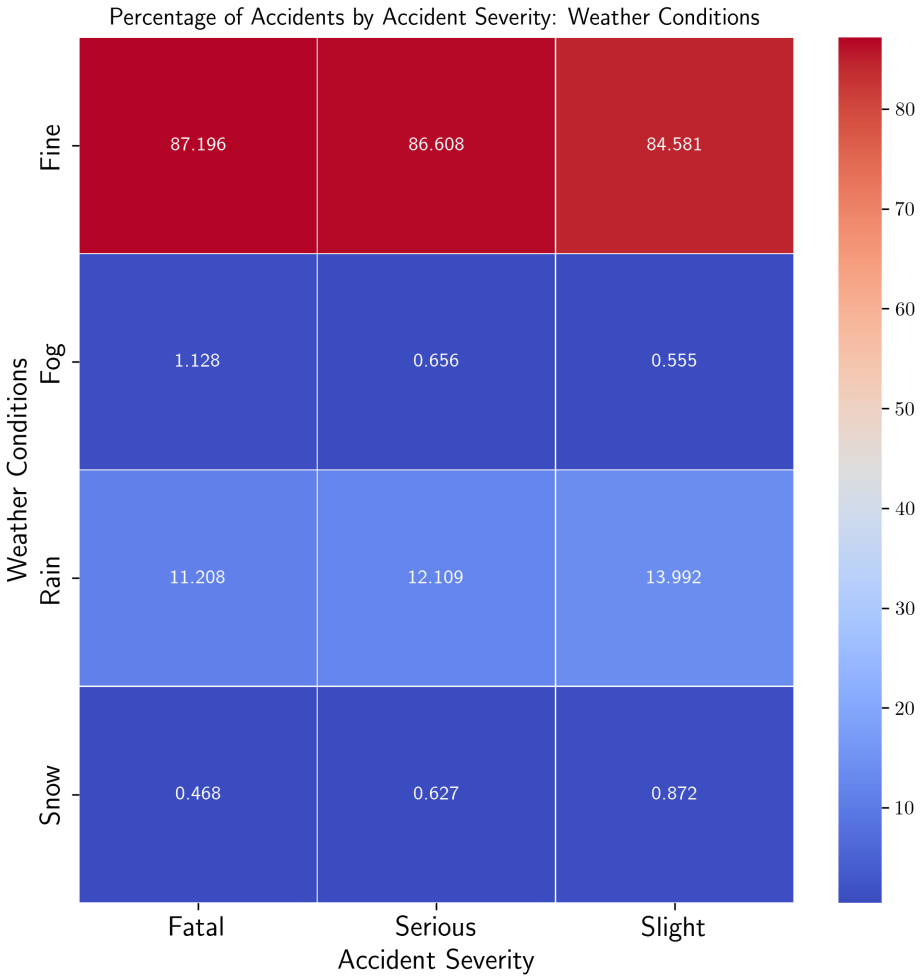

Figure 9: Percentage of accidents based on weather conditions as a function of accident severity.

3.3.3 Road Surface Conditions

Road surface conditions, which can correspond with weather conditions, also does not show any trend with accident severity, as shown in Figure 10. That is, accident rate by severity is relatively consistent across all road surface conditions. Therefore, road surface conditions, like weather, is *not* a good predictor of accident severity.

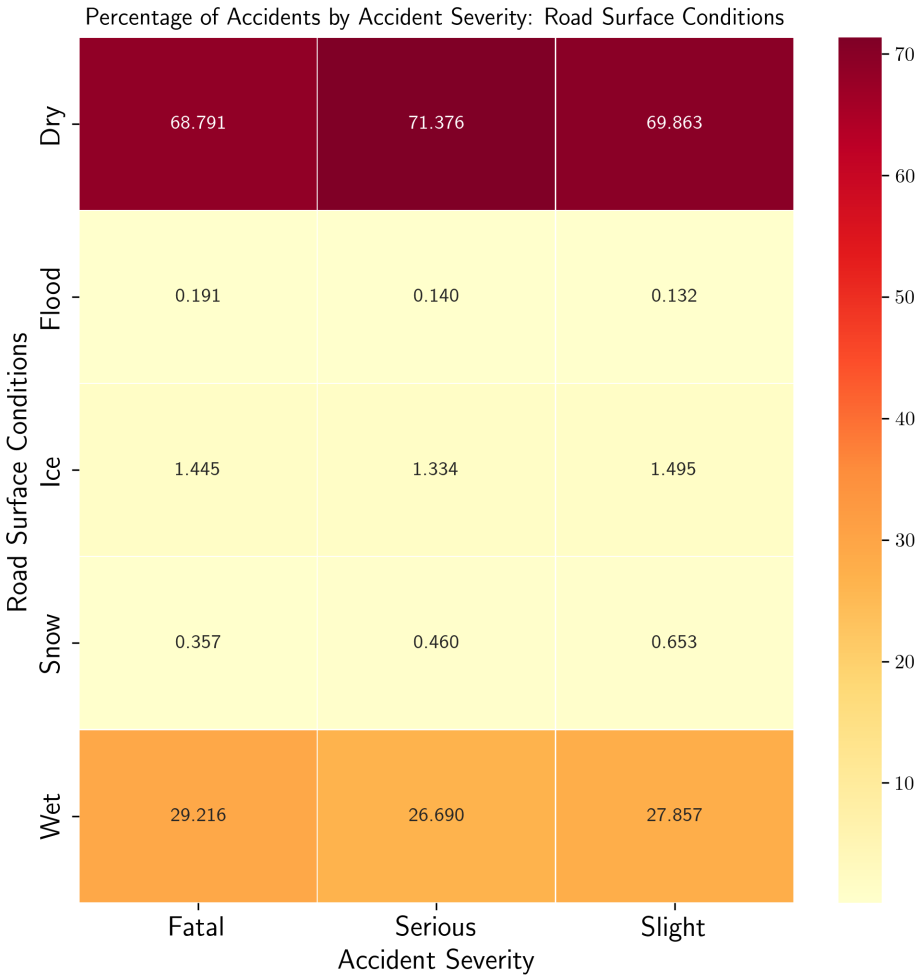

Figure 10: Percentage of accidents based on road surface conditions as a function of accident severity.

3.3.4 Road Type

Figure 11 shows some useful trends by accident severity. Most notably, that slight accidents are much more frequent on on/off ramps, one way streets, and roundabouts. Fatalities are more frequent on dual and single carriageways. Note that we have still not encountered a good predictor for serious accident severity as of yet.

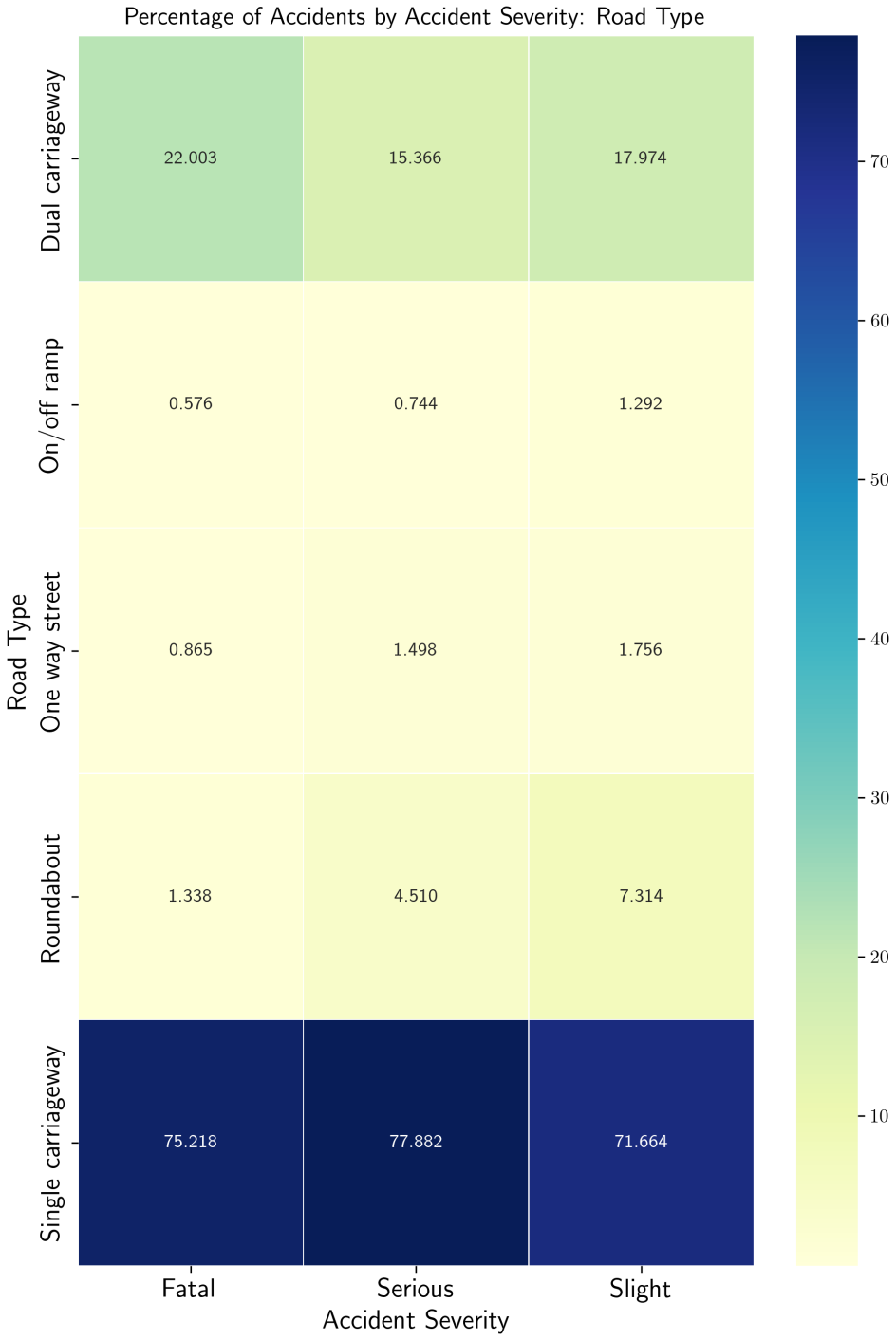

Figure 11: Percentage of accidents based on road type as a function of accident severity.

3.3.5 Posted Speed Limit

For posted speed limit, the data originally was split into eight speed categories (10,15,20,30,40,50,60,70), however there was a clear dichotomy in accident severity we could preserve if we consolidated speeds to "Under 40 km/s" and "Over 40 km/s", corresponding to non-highway and highway speeds. This clear divergence in accident severity is seen in Figure 12, which shows that fatal accidents are much more frequent at speeds over 40 km/s, while slight accidents are much more frequent at speeds under 40 km/s, and serious accidents somewhere in between.

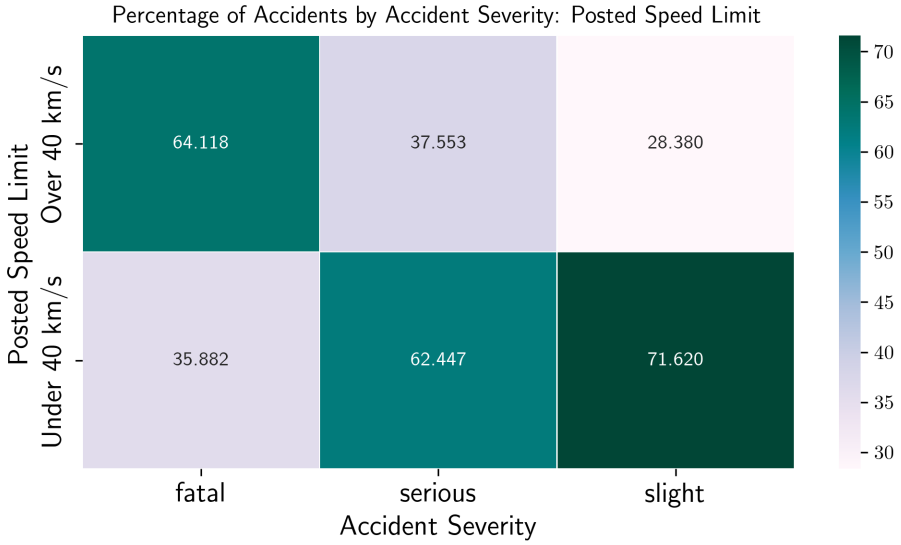

Figure 12: Percentage of accidents based on speed limit as a function of accident severity.

3.3.6 Urban or Rural Area

As we did with speed limit, we also see a strong separation between fatal and slight accident severity with urban and rural area, seen in Figure 13. We also saw this to a degree in Figure 2 for fatal severity accidents, which show much higher frequency in not just urban, but many rural areas of the UK. This could be due to less access to or delays emergency services.

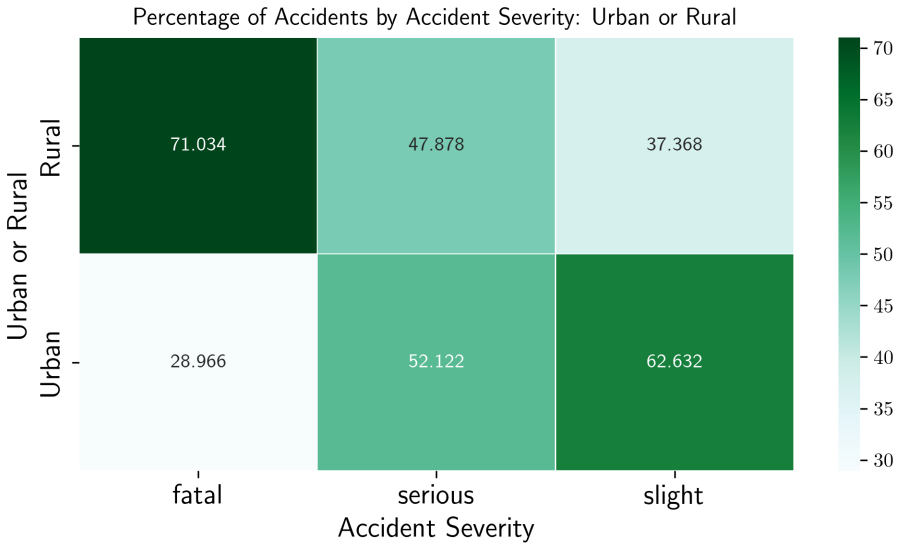

Figure 13: Percentage of accidents based on urban or rural area as a function of accident severity.

3.3.7 Driver Age

Finally, Figure 14 shows age band of driver as a function of accident severity. Originally, ages we categorized in 8 bands from 16 to over 70, however we did not see any strong trends with accident severity. We re-categorized age bands into three categories: adolescent (16-20), adult (21-65), and senior (65 - over 70). Doing this allowed us to find some weak trends with accident severity, but nothing that would lead age bands to be a strong predictor. For instance, fatal accidents are $\sim 3\%$ more frequent than slight for seniors, and the opposite is true for adults.

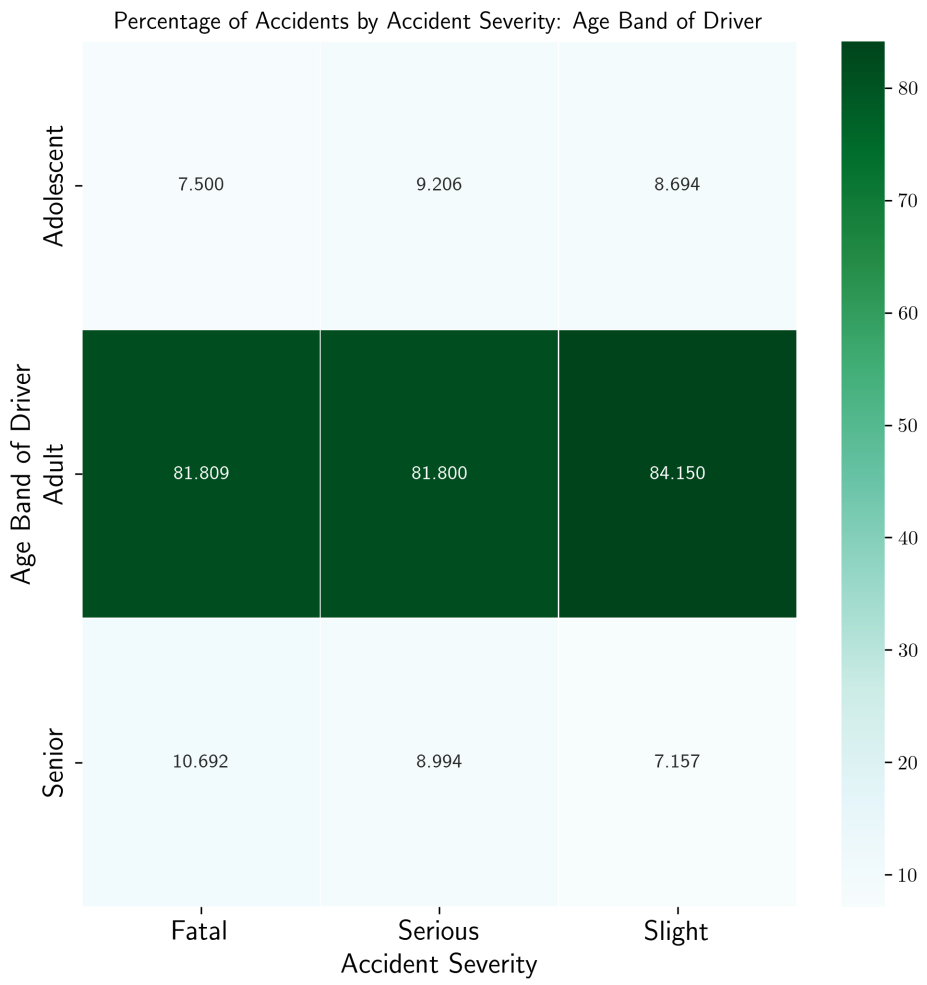

Figure 14: Percentage of accidents based on driver age as a function of accident severity.

3.4 Feature Selection

Table 2 lists the features and their values use for the machine learning modeling. We have omitted weather conditions and road surface conditions because they do not appear to be good predictors of accident severity.

Feature	Values
Light conditions	Darkness - lights lit Darkness - no lighting Daylight
Road type	Dual carriageway On/Off ramp One way street Roundabout Single carriageway
Speed limit	Over 40 km/s Under 40 km/s
Area	Urban Rural
Age	Adolescent Adult Senior
Vehicle	Bike Car Goods Motorcycle

Table 1: Features selected for machine learning analysis and their corresponding values.

3.5 Data Preparation

Below we summarize steps in preparing our data for machine learning modeling.

3.5.1 One-Hot Encoding Categorical Variables

Because all of the features of our dataset are categorical, we employ one-hot encoding of all features using the Pandas `get_dummies()` function. This increased the number of feature columns in the table from 6 to 19. The effect of this increases computation time, but is a necessary step for categorical variables.

3.5.2 Imbalanced Dataset

After feature selection, the number of slight severity labels vastly outnumbers the number serious and fatal labels. In order to address this imbalance in categorical labels, and to prevent any bias in categorization by the machine learning models, we perform sub-sampling of the majority classes (slight and serious) to match the number of the minority class (fatal). We justify this approach by the shear number of data for both slight and serious labels, therefore a random sampling from each class should adequately represent each class. Also, the minority class (fatal) is still very large, therefore, the sub-sampled classes should be adequately represented by random sampling.

We perform this sub-sampling using `sklearn.utils.resample()` function, and set `replace=False`. This is done to match the minority class such that the resulting number of slight and serious labels is equal to the number of fatal labels.

Label	Count
Slight	1,430,139
Serious	224,816
Fatal	23,723

Table 2: Target label counts before sub-sampling the dataset to correct for imbalance.

3.6 Results

3.6.1 k-Nearest Neighbors Classifier

To determine the value of k for the k-Nearest Neighbor (kNN) classifier, we implement `sklearn.model_selection.GridSearchCV()` to determine the value of k that results in the best accuracy score. The best accuracy achieved was 0.49 at $k = 64$. The performance metrics for this algorithm are given in Table **??**.

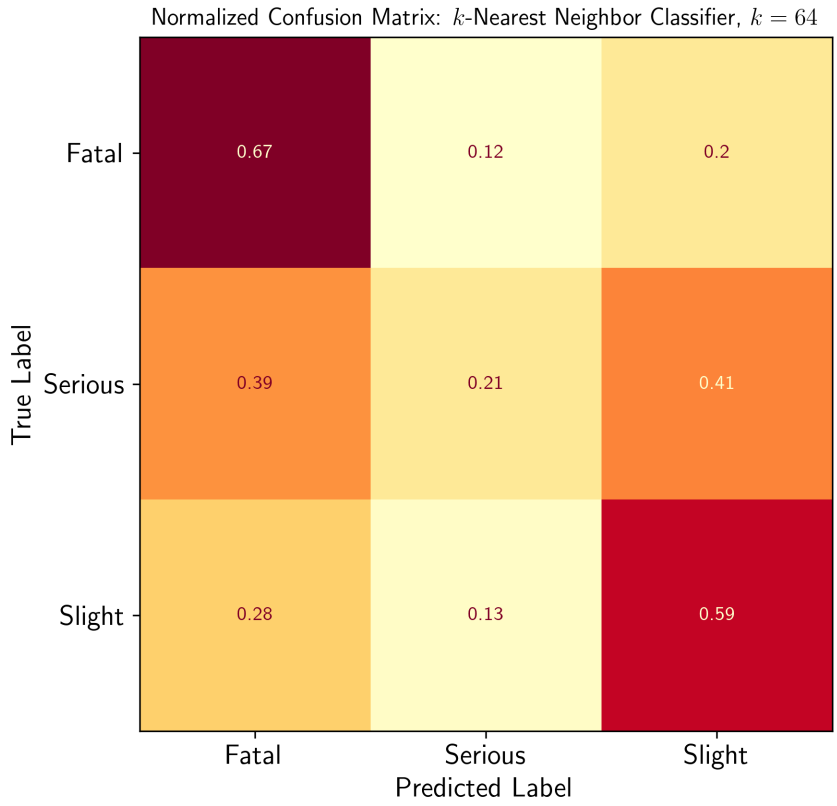

Figure 15: Confusion matrix resulting from the k-Nearest Neighbor algorithm, where $k = 64$.

3.6.2 Decision Tree Classifier

The decision tree classifier implemented used the "entropy" criteria with a `max_depth=5`. Testing with different criteria and depths did not change results considerably.

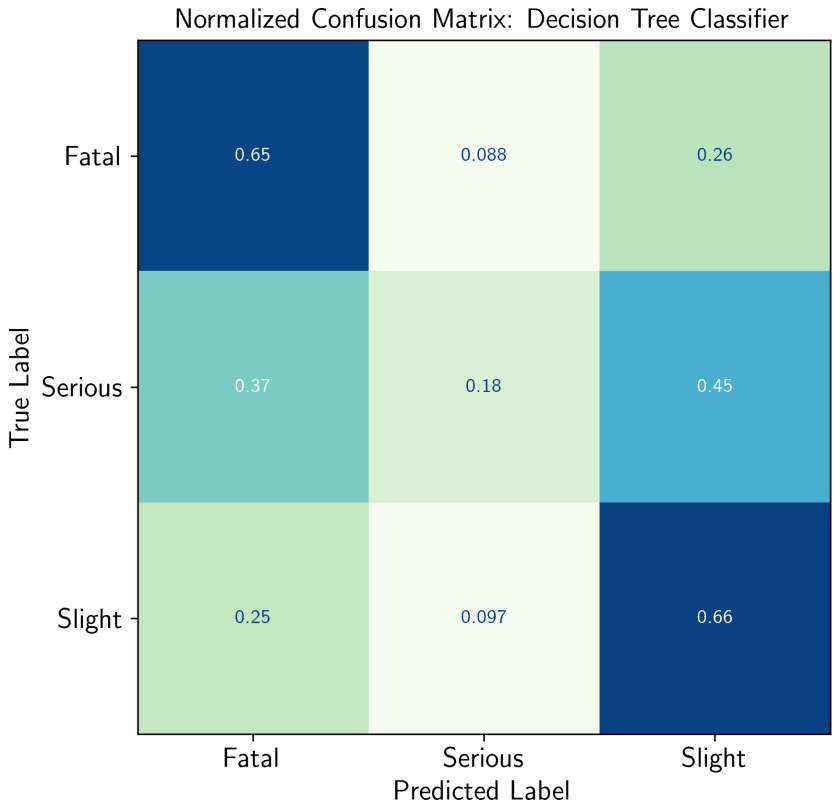

Figure 16: Confusion matrix for the decision tree classifier.

3.6.3 Random Forest Classifier

We used a random forest classifier using a `max_depth=4`.

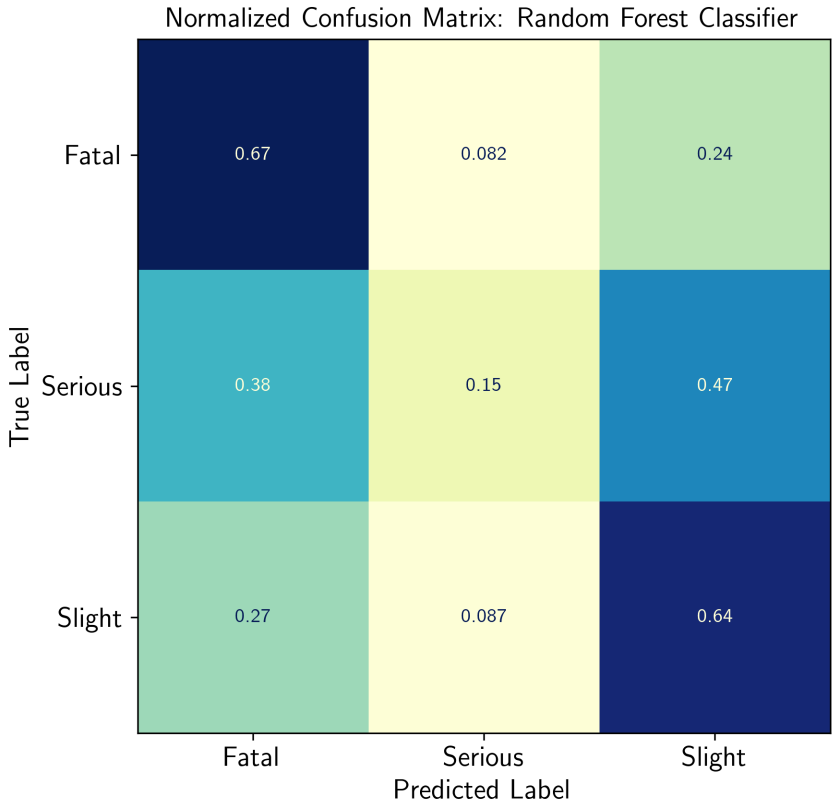

Figure 17: Confusion matrix for the random forest classifier.

3.6.4 Linear Support Vector Classifier

We used a linear support vector classifier using hyperparameter values.

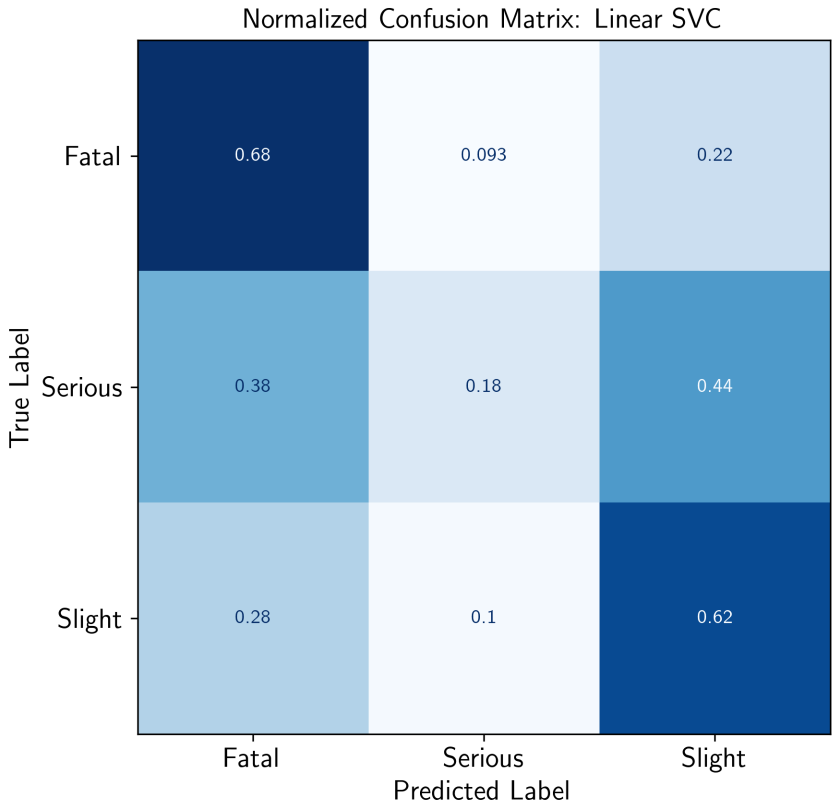

Figure 18: Confusion matrix for the linear support vector classifier.

3.6.5 Radial Basis Function Support Vector Classifier

We used a Radial Basis Function (RBF) support vector classifier using default hyperparameter values.

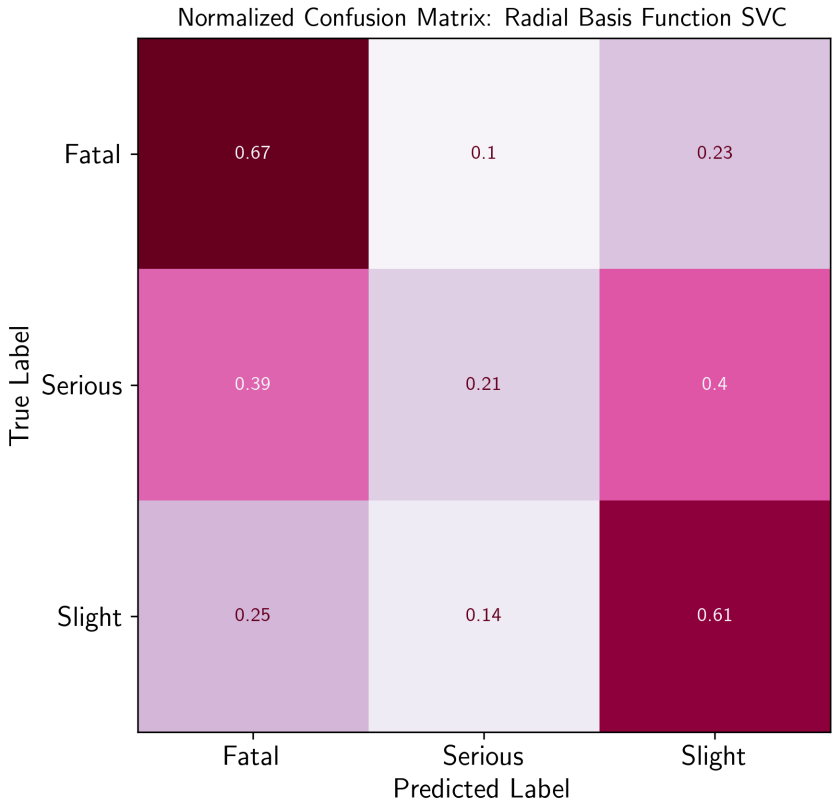

Figure 19: Confusion matrix for the radial basis function (RBF) support vector classifier.

3.6.6 Gaussian Naive Bayes Classifier

We used a Gaussian naive Bayes classifier using hyperparameter values.

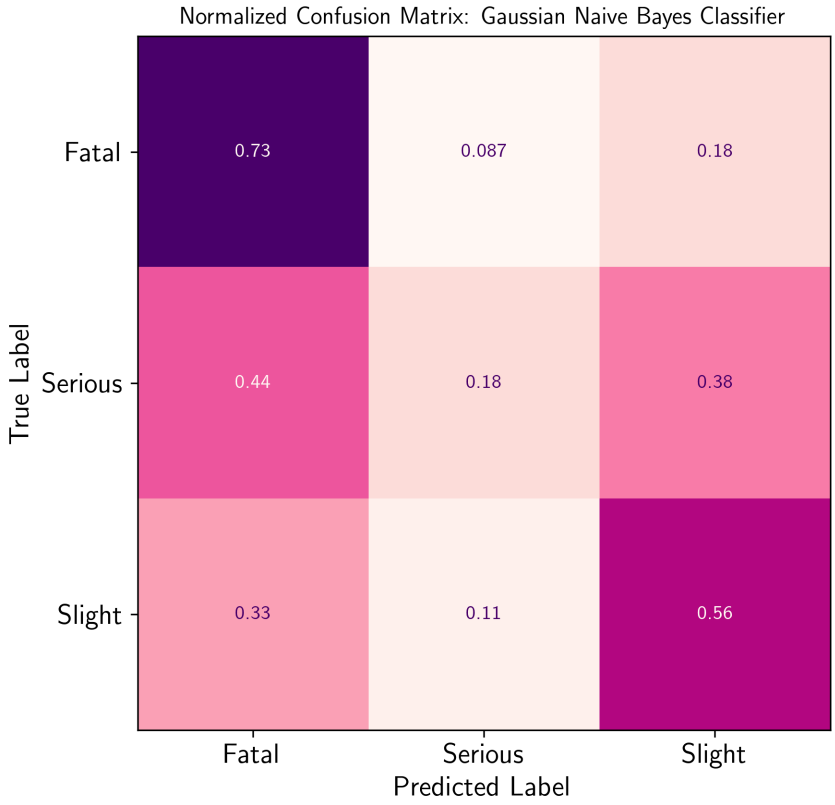

Figure 20: Confusion matrix for the Gaussian Naive Bayes classifier.

3.6.7 Summary of Performance Metrics

Classifier	Precision	Recall	Jaccard	F1	Fatal TP	Serious TP	Slight TP
k-NN	0.48	0.49	0.31	0.47	0.67	0.21	0.59
Dec. Tree	0.49	0.50	0.31	0.46	0.65	0.18	0.66
RBF SVC	0.48	0.49	0.32	0.47	0.67	0.21	0.61
Rand. Forest	0.48	0.49	0.30	0.45	0.67	0.15	0.54
GNB	0.49	0.49	0.31	0.46	0.73	0.18	0.56
Linear SVC	0.49	0.50	0.31	0.46	0.68	0.18	0.62

Table 3: Summary of performance metrics for the six different algorithms used in the machine learning analysis.

4 Discussion

We can see that performance measures are generally poor across the board. This is due to the fact that although we have three distinct labels for classification, there is no single good predictor for "serious" accidents. This can be seen in all confusion matrices as well as the heatmaps generated for data exploration. There is also no consistent way of combining "serious" labels with the other two classes, since often times serious accidents occur at frequencies between the two others.

Overall the Radial Basis Function Support Vector Classifier performs best, with high true positive rate for "serious" cases. The k-Nearest Neighbors classifier performs similarly to the RBF SVC model. The Gaussian Naive Bayes model results in the highest true positive rate for "fatal" accidents, at the expense of performance on the other two label classifications. The Decision Tree classifier performs the best in identifying true positive cases for slight accidents, but at the expense of the other two target labels.

This exercise reveals that while there are good a number of good predictors for slight and fatal accidents, there is no single good predictor for serious accidents. The reasons for this are that slight and fatal accidents are on two extremes of many predictors, with serious cases split between the two extremes. This makes it difficult for the machine learning model to distinguish serious cases from slight or fatal accident cases.

Possible solutions for future studies could be to limit the classification to binary instead of multi-class, either by studying slight and fatal independently from serious, or some variation of two classes independent from the other. However in this exercise, we explicitly set out to perform multi-class (3) classification. A good predictor for serious cases may not currently exist, or cannot be accounted for with current data. It is very likely that serious cases could be caused by multiple conditions, as well as random chance, making it very difficult to predict.

It may be an interesting project to try and classify "serious" accidents as either "slight" or "fatal," but it would require an unsupervised approach rather than supervised. Then it would be possible to revisit this data using binary classification of accidents as "non-fatal" or "fatal".

5 Conclusion

In summary, we performed machine learning analysis on accident severity data in the UK. Our goal was to develop a model to predict accident severity using supervised multi-class classification. In doing so, we found successful predictors for both "slight" and "fatal" accident severity. However, we found that accidents classified as "serious" lack any good predictors. The lack of a good predictor for serious accident severity can be seen in all heatmaps generated as part of the exploratory data analysis, which showed that serious accidents do not have any strong predictors while fatal and slight accidents do. This leads to poor classification of serious accidents. The lack of a good predictor could be due to a lack of data that can adequately describe serious accidents, or that serious accidents cannot be controlled for (are random events). These are supported by the fact that serious accidents usually fall at frequencies between slight and fatal accidents. Future studies of accident severity should attempt to further quantify accident types and their causes to better model serious accident severity.