1. Introduction

Nowadays, we are witnessing a rapid growth in the number of motor vehicles in circulation, and as a consequence, a continuous increase in the number of road accidents, despite the efforts made to implement modern infrastructures meeting international standards. These accidents are now one of the leading causes of death in the world according to World Health Organization (WHO) .The problem is not specific to a given country, but the whole world undergoes this phenomenon which cannot be inevitable. Terrifying statistics were published by WHO that indicate the following; About 1.25 million deaths per year, accidents are the Leading cause of death among young people aged 15 to 29, Almost half of those killed on the roads are “vulnerable users” (pedestrians, cyclists and motorcyclists). Without sustained action, road accidents are projected to become the seventh leading cause of death by 2030. The consequences of road accidents go beyond threatening human lives, it also has a considerable economic impact for the relatives of the victims and the countries concerned. The treatment of victims very often requires large sums of money and the cost of repairing damaged public funds which can be costly to the government. Accurate prediction of accident severity can be helpful to provide proactive solutions and test the readiness of road practitioners and local governments. In this project, we develop machine learning models to predict severity of road accidents in UK based on data collected in 2017.

2. Data description

Our goal in this project is to build a machine learning solution to predict the severity of traffic accidents. Data that will be used for analysis was obtained from the UK government open data portal website. Dataset was downloaded as CSV file and includes records of traffic accident that took place in Leeds, UK in 2017. The raw sample has a total of 2203 accidents with 15 columns. The main target variable is “Severity” which is a categorical variable that supports 3 classes: “Fatal”, “Serious”, and “slight”. The attributes include the time and date of accident, road type where the accident occurred, the state of road surface, weather conditions, as well as age and gender of drivers.

3. Methodology

3.1 Data preprocessing

Once the raw data is obtained, a pre-processing step is required to clean data and finalize the feature that will be used for modelling. First, we check for missing data by creating a simple heatmap using “seaborn” library. Figure 1 shows the heatmap for our dataset. Yellow marks would appear to indicate the missing values. However, in this case, we can clearly see that there are no missing values.

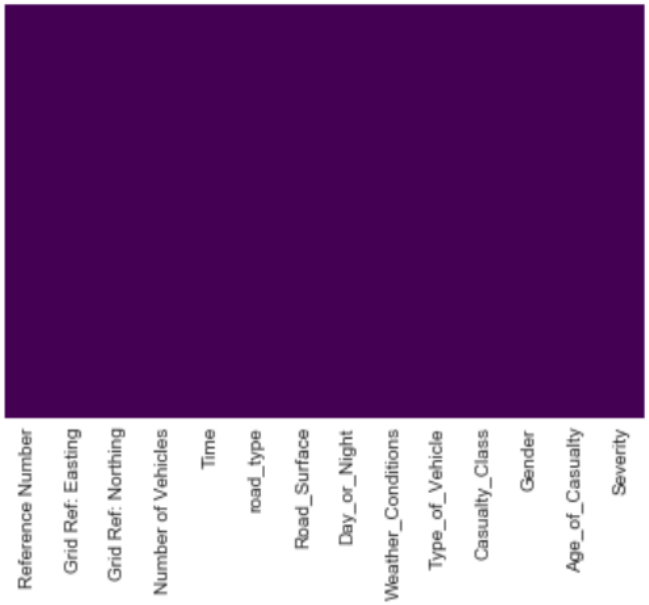

Figure 1 Missing value Heatmap

Then, we do some data exploration with a focus on the the target variable which is “Severity”. We can see that “Fatal” accidents are less than 1% of the total accidents. Figure 2 shows a barplot of the count in each severity class. This critical class imbalance would be approached in further work with some treatment techniques like oversampling or undersampling. But for now, we use the data it is.

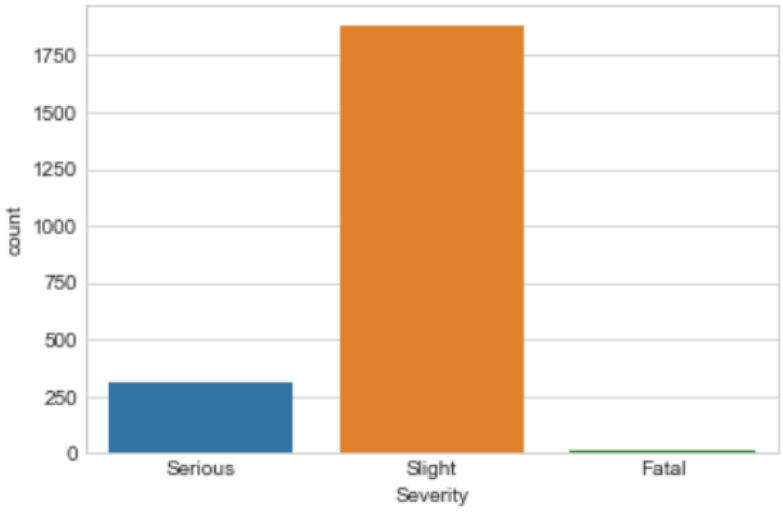

Figure 2 Severities count by for each class

By looking at accidents severity according to gender in figure 3, male accidents are exceeding female accidents in all 3 classes. This requires a further investigation to identify the reasons and suggest practical recommendations.

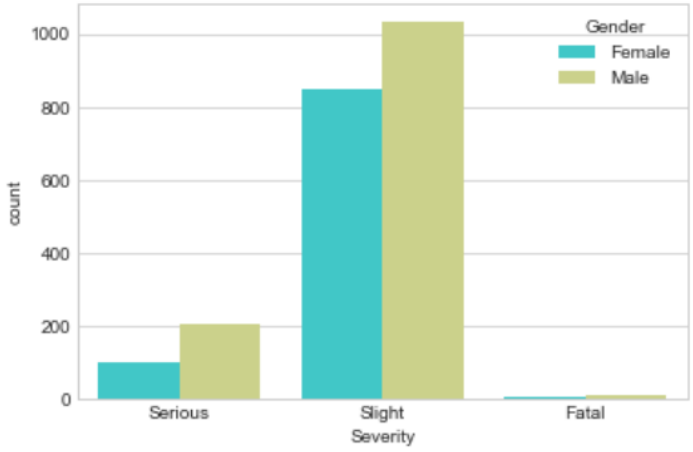

Figure 3 Accident severity by gender

After examining the attributes of the dataset, we can see that some columns are irrelevant to our problem. So we simply drop them from the dataset. A total of 5 attributes were excluded from further analysis such as “Reference number”, ”Grid reference”, and so on. The first 5 rows of our dataset is shown in Table 1.

Table 1 First 5 rows of accident dataset after cleaning

	Time	road_type	Road_Surface	Day_or_Night	Weather_Conditions	Casualty_Class	Gender	Age_of_Casualty	Severity
0	815	A	Dry	Day	Other	Pedestrian	Female	61	Serious
1	1330	A	Dry	Day	Fine without high winds	Driver or rider	Male	36	Slight
2	805	A	Wet/Damp	Day	Fine without high winds	Driver or rider	Male	32	Slight
3	805	A	Wet/Damp	Day	Fine without high winds	Driver or rider	Male	30	Slight
4	1705	U	Wet/Damp	Night	Raining without high winds	Vehicle or pillion passenger	Female	26	Slight

Before moving to model building, some feature engineering work is thought to be useful. For the time, I think it is better to convert it categorical variable. Since the data is obtained from UK, a quick internet search indicates that there are 2 time intervals that are considered as peak hour: 6:30 -> 9:30 and 16:00 -> 19:00. So, “Time” column will be converted to 2 values : Peak and Offpeak.

In the final stage of the data preprocessing, we'll need to convert categorical features to dummy variables using pandas. Otherwise, our machine learning algorithm won't be able to directly take in those features as inputs. The final table of input variables can be seen in the Jupyter notebook.

3.2 Model Building

In order to build the optimal model, 6 Machine learning algorithms will be used, evaluated, and compared to choose the classifier that shows the best performance.

The classifiers are:

- Logistic Regression (LR)
- LinearDiscriminantAnalysis (LDA)
- K-Nearest Neighbor (Knn)

- Decision Tree (CART)
- Gaussian Naive Bayes (NB)
- Support Vector Machine (SVM)

Evaluation was performed using K-fold cross validation with 10 splits. For the best model, further evaluation will be performed using precision, recall, F1-score and support to evaluate the models.

4. Results

All the code implementation was performed in python using a Jupyter notebook. Various packages were used for data preprocessing as well as build the machine learning models. These packages include: Numpy, Pandas, Matplotlib, Seabron, and sklearn.

Data were splitted into 80% training data and 20% validation data. We can clearly see that Logistic Regression model and Support Vector Machine model are the best. The results are shown in table 2.

Table 2 Accuracy of candidate models

Model	**Accuracy**
LR	0.8524
LDA	0.8411
KNN	0.8382
CART	0.7434
NB	0.0590
SVM	0.8536

SVM slightly outperforms, so we choose it for further validation. So we choose it for further analysis. This time we will use other evaluation metrics such as precision, recall, F1-score, and support. Please see Table 3.

Table 3 Evaluation of classification performance of SVM model

	Precision	Recall	F1-Score	Support
Fatal	0.00	0.00	0.00	2
Serious	0.00	0.00	0.00	64
Slight	0.85	1.00	0.92	375
Avg/total	0.72	0.85	0.78	441

5. Discussion

In previous section, we found that SVM model performs the best, however, with further analysis, SVM does not seem to show optimal performance. One reason could be the critical imbalance between the 3 severity classes. Other data treatment techniques can be used to improve the input data by using oversampling or undersampling techniques to create balance between the different classes.

Another problem can be the hyper parameter tuning which can improve the classification performance of the models and provide better results through optimization of parameters of each model.

6. Conclusion

In this project, we propose a machine learning model to predict the severity of traffic accidents in the city of Leeds, UK. A comparative approach was followed to identify the best performing model among 6 models. The best model was support vector machine. Further improvements are required to optimize the model performance.