Data Sheet for Accident Data

# Dataset Overview

This dataset contains records of road accidents in the United Kingdom, capturing various attributes related to accidents, weather conditions, light conditions, road surface conditions, and accident severity. It aims to provide insights into accident patterns, contributing factors, and potential areas for road safety improvements.

# Source information.

The UK government amassed traffic data from 2000 and 2018, recording over 1.8 million accidents in the process and making this one of the most comprehensive traffic data sets out there. It's a huge picture of a country undergoing change.

<https://www.kaggle.com/datasets/devansodariya/road-accident-united-kingdom-uk-dataset>

# Data Composition

The data is made up of 32 columns indicating various attributes and parameters collected as part of the collision investigation. Details below:

| **Column** | **Non-Null Count & Dtype** |
| --- | --- |
| Unnamed: 0 | 1504150 non-null, int64 |
| Accident\_Index | 1504150 non-null, object |
| Location\_Easting\_OSGR | 1504049 non-null, float64 |
| Location\_Northing\_OSGR | 1504150 non-null, float64 |
| Longitude | 1504049 non-null, float64 |
| Latitude | 1504150 non-null, float64 |
| Police\_Force | 1504150 non-null, int64 |
| Accident\_Severity | 1504150 non-null, int64 |
| Number\_of\_Vehicles | 1504150 non-null, int64 |
| Number\_of\_Casualties | 1504150 non-null, int64 |
| Date | 1504150 non-null, object |
| Day\_of\_Week | 1504150 non-null, int64 |
| Time | 1504033 non-null, object |
| Local\_Authority\_(District) | 1504150 non-null, int64 |
| Local\_Authority\_(Highway) | 1504150 non-null, object |
| 1st\_Road\_Class | 1504150 non-null, int64 |
| 1st\_Road\_Number | 1504150 non-null, int64 |
| Road\_Type | 1504150 non-null, object |
| Speed\_limit | 1504150 non-null, int64 |
| Junction\_Control | 901315 non-null, object |
| 2nd\_Road\_Class | 1504150 non-null, int64 |
| 2nd\_Road\_Number | 1504150 non-null, int64 |
| Pedestrian\_Crossing-Human\_Control | 1504133 non-null, object |
| Pedestrian\_Crossing-Physical\_Facilities | 1504116 non-null, object |
| Light\_Conditions | 1504150 non-null, object |
| Weather\_Conditions | 1504150 non-null, object |
| Road\_Surface\_Conditions | 1504150 non-null, object |
| Special\_Conditions\_at\_Site | 36582 non-null, object |
| Carriageway\_Hazards | 27250 non-null, object |
| Urban\_or\_Rural\_Area | 1504150 non-null, int64 |
| Did\_Police\_Officer\_Attend\_Scene\_of\_Accident | 1504150 non-null, object |
| LSOA\_of\_Accident\_Location | 1395912 non-null, object |
| Year | 1504150 non-null, int64 |

# Column Description detail

For the sake of this project, we will consider only the below columns.

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Accident\_Severity | Integer | Categorical value representing the severity of the accident. Possible values:  - 1: Fatal accident  - 2: Serious accident  - 3: Slight accident |
| Light\_Conditions | Categorical | Describes the light conditions at the time of the accident. Possible values include:  - Daylight: Street light present  - Darkness: Street lights present and lit |
| Weather\_Conditions | Categorical | Describes the weather conditions during the accident. Includes conditions like:  - Fine without high winds  - Raining without high winds  - Fog or mist |
| Road\_Surface\_Conditions | Categorical | Describes the condition of the road surface at the time of the accident. Possible values:  - Wet/Damp  - Dry  - Frost/Ice |

## Data Summary:

### Accident Severity:

This column has 3 unique values corresponding to accident severity:

* 3 (Slight): 1,280,205 instances
* 2 (Serious): 204,504 instances
* 1 (Fatal): 19,441 instances

The majority of accidents are of slight severity.

### Light Conditions:

This column has 5 unique values describing the lighting at the time of the accident. Most accidents occurred under the following conditions:

* Daylight: Street light present: 1,102,222 instances
* Darkness: Street lights present and lit: 296,340 instances

### Weather Conditions:

This column has 9 unique weather conditions:

* The most common is "Fine without high winds" (1,203,943 instances), followed by "Raining without high winds" (177,663 instances).

### Road Surface Conditions:

This column contains 6 unique values describing the condition of the road surface:

The majority of accidents occurred on dry roads (1,034,670 instances), followed by wet/damp roads (423,477 instances).

# Data Splitting

The dataset was split into training, validation, and test sets as follows:

* Training Set: 56% of the original data
* Validation Set: 14% of the original data
* Test Set: 30% of the original data

# Data Sampling

A 10% random sample of the original data (150,415 rows) was used for modelling due to the large size of the dataset.

# Potential Risks and ethical considerations:

* Bias in Reporting: The dataset might not capture all accidents, especially minor ones that go unreported.
* Ethical Use: Any predictive modelling based on this dataset should prioritise reducing road accidents and improving safety rather than penalising drivers based on accident predictions.

Model Card for K-Nearest Neighbors (KNN) Classifier for Accident Severity Prediction

# Model Details

* **Model Type:** K-Nearest Neighbors (KNN) Classifier
* **Developed by:** Yunus Jabalpurwala
* **Date:** 6 Oct 2024
* **Framework:** scikit-learn
* **Language:** Python
* **Model Objective:** Predict the severity of road accidents based on environmental factors (light conditions, weather, and road surface).

# Intended Use

### **Primary Use Case**

The KNN model is designed to predict the **severity of road accidents** based on the following features:

* **Light\_Conditions**: Daylight, darkness with/without street lights, etc.
* **Weather\_Conditions**: Rain, snow, fog, high winds, etc.
* **Road\_Surface\_Conditions**: Dry, wet, icy, snow-covered roads, etc.

The model classifies accident severity into three categories:

1. **Slight (3)**
2. **Serious (2)**
3. **Fatal (1)**

### **Intended Users**

* **Traffic safety researchers**: To analyse and understand the contributing factors to accident severity.
* **City planners and policymakers**: To make informed decisions regarding road safety improvements.
* **Machine learning practitioners**: To experiment with classification models in safety-related datasets.

### **Primary Context of Use**

The model can be used to analyse historical accident data to provide insights and aid in developing preventative measures for traffic accidents. It is not intended for real-time decision-making or use in autonomous vehicle systems.

# Training Data

Refer to Data sheet (Above).

**Data Preprocessing**

* **One-hot encoding** was applied to the categorical features(Light\_Conditions, Weather\_Conditions, and Road\_Surface\_Conditions).
* A **10% random sample** of the dataset was used for faster model training.
* The dataset was split into **training (56%)**, **validation (14%)**, and **test (30%)** sets.

### **Feature Overview**

* **Light\_Conditions**: Categorical, encoded into multiple binary features.
* **Weather\_Conditions**: Categorical, encoded into multiple binary features.
* **Road\_Surface\_Conditions**: Categorical, encoded into multiple binary features.
* **Target**: Accident\_Severity (1 = Fatal, 2 = Serious, 3 = Slight).

# Model Performance

### **Evaluation Metrics**

* **Validation Accuracy**: 84%
* **Test Accuracy rate:**

### **Performance Summary**

* The model showed good performance in predicting accident severity for different environmental conditions, with accuracy in the range of [accuracy range] on the test set.

# Ethical Considerations

## Fairness

* **Bias in reporting**: The model is trained on data that may under report certain types of accidents, especially minor ones or accidents in rural areas. This could introduce bias in predictions, especially for "Slight" accident severity.
* **Feature imbalances**: There are significantly more data points for certain conditions, like daylight or dry roads, which may lead to biased predictions in underrepresented categories (e.g., accidents during snow or floods).

## Mitigations

* Consider gathering more balanced datasets or applying techniques like **oversampling** or **undersampling** to address the class imbalance in accident severity or environmental conditions.

# Limitations

1. **Data limitations**: The dataset covers historical road accident data from a specific region (UK) and time period. Predictions may not generalise well to other geographic locations or changing road conditions.
2. **Non-Real-Time**: The KNN classifier is not suitable for real-time prediction or autonomous driving systems due to the computational cost and potential delay in making predictions.
3. **No Causal Inference**: The model is designed for prediction, not for inferring causality between environmental conditions and accident severity.

# Caveats and Recommendations

* **Not for real-time use**: This model should not be used in safety-critical systems such as autonomous driving.
* **Interpretability**: Use caution when interpreting predictions, especially in regions where certain weather or lighting conditions are underrepresented.
* **Further validation**: This model should be tested on additional datasets to ensure its robustness across diverse geographical locations and conditions.

# Future Improvements

* **Hyperparameter tuning**: Further exploration of the number of neighbors and distance metrics (Euclidean, Manhattan) in KNN could improve model performance.
* **Feature engineering**: Additional features such as time of day, traffic density, or road types could enhance the model’s predictive capabilities. This would require higher computational power.
* **Model alternatives**: Consider experimenting with other classifiers such as Random Forests for improved accuracy and robustness.

Contact Information

Yunus Jabalpurwala

yunus.jabalpurwala@gmail.com