Capstone Project Proposal

Research Question:

What are the main factors that contribute to the severity of road accidents in the United Kingdom, and how effectively can we predict the level of accident severity using the available data?

Expected Data Source:
For this project, I will utilize the publicly accessible---------- **Road Accident United Kingdom (UK) Dataset** -----available on Kaggle. This dataset provides comprehensive information on road accidents, including variables such as geographic location, weather conditions, road surface status, time of occurrence, and the recorded severity of each accident.

Techniques Expected to Use:

The project will begin with thorough **Exploratory Data Analysis (EDA)** to identify patterns, trends, and relationships between different features and accident severity. Next, I will perform essential **data preprocessing** steps, including handling missing or inconsistent data and transforming categorical variables into formats suitable for machine learning algorithms. For prediction, I plan to implement various **classification algorithms** such as Decision Trees, Logistic Regression, and Support Vector Machines to classify accidents by severity level. The performance of these models will be assessed using evaluation metrics including accuracy, precision, recall, and the F1-score to ensure balanced and reliable predictions.

Expected Results:

The goal is to develop an accurate predictive model that can reliably estimate accident severity based on input features. Furthermore, the analysis will shed light on the most significant factors influencing severe accidents, providing insights that can guide preventive measures. This information could support targeted interventions to reduce the occurrence and impact of serious traffic incidents.

Why This Question Is Important:

Addressing the causes behind severe road accidents is crucial for enhancing public safety, optimizing emergency response efforts, and shaping effective transportation policies. Without a clear understanding of these factors, high-severity accidents may continue to lead to unnecessary injuries, fatalities, and economic losses. By converting complex accident data into understandable and actionable insights, this project aims to assist decision-makers and the general public in taking informed steps toward safer roads. Ultimately, this work has the potential to contribute to saving lives and reducing the social and financial burdens associated with traffic accidents.

Capstone Project: UK Road Accident Severity Prediction

1. Problem Statement

Road traffic accidents remain one of the leading causes of injury and death worldwide. In the United Kingdom, thousands of accidents occur every year under varying road, weather, and lighting conditions. Understanding what causes some of these accidents to be more severe than others is vital for improving road safety and reducing fatalities.

The objective of this project is to develop a machine learning model capable of predicting the severity level of road accidents using real-world UK accident data.

By identifying the most critical factors influencing accident severity, this project supports data-driven decision-making for policymakers, emergency services, and transportation planners.

Challenges:

- ✓ Real-world accident data often includes missing or inconsistent values.
- ✓ Many features are categorical and require careful encoding.
- ✓ Accident severity prediction involves imbalanced class distributions.

Benefits:

- ✓ Enables proactive safety measures based on predictive insights.
- ✓ Supports the design of safer infrastructure.
- ✓ Improves emergency response efficiency.

2. Model Outcomes or Predictions

Type of Learning: Supervised Learning

Task: Classification

Target Variable: Accident Severity (e.g., Slight, Serious, Fatal)

Algorithms Implemented:

- ✓ Logistic Regression (Baseline)
- ✓ Decision Tree Classifier
- ✓ Random Forest Classifier
- ✓ Support Vector Machine (SVM)
- ✓ Gradient Boosted Trees (XGBoost)

The models predict the **severity level** of a given accident using input features such as **weather**, **lighting**, and **road type**.

3. Data Acquisition

Dataset: Road Accident UK Dataset (Kaggle)

The dataset includes over 100,000 accident records from the UK, featuring:

- ✓ **Environmental conditions**: Weather, lighting, surface type
- ✓ **Human factors**: Police presence, pedestrian control
- ✓ **Location-based attributes**: Road type, junction detail
- ✓ **Target**: Accident severity

Exploratory Data Analysis (EDA) revealed:

- ✓ Most accidents occurred under "fine weather" conditions but with light or moderate traffic.
- ✓ Fatal accidents were more common during night hours and on single carriageways.
- ✓ Certain road types and poor lighting showed a strong correlation with higher severity.

4. Data Preprocessing / Preparation

a. Cleaning Missing or inconsistent data

- ✓ Replaced missing values with the most frequent (mode) for categorical features.
- ✓ Removed records with null target values.
- ✓ Standardized inconsistent category names.

b. Encoding Categorical Variables

Categorical features were transformed using Label Encoding to convert text into numeric form for modeling.

```
from sklearn.preprocessing import LabelEncoder

cat_features = ['Road_Type', 'Light_Conditions', 'Weather_Conditions',
                'Road_Surface_Conditions',
'Did_Police_Officer_Attend_Scene_of_Accident',
                'Pedestrian_Crossing-Human_Control', 'Pedestrian_Crossing-
Physical_Facilities']

for col in cat_features:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
```

c. Train-Test Split

Data was divided into training (70%) and testing (30%) sets for evaluation.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)
```

5. Modeling

Multiple models were trained to identify which algorithm performs best for predicting accident severity.

Model Description

- ✓ Logistic Regression => Baseline linear classifier for benchmarking(Our model 0.85 that we need to beat.)
- ✓ Decision Tree => Non-linear model capturing hierarchical relationships
- ✓ Random Forest => Ensemble method reducing overfitting and improving accuracy
- ✓ SVM => Classifier for complex decision boundaries
- ✓ XGBoost => Gradient-boosted ensemble for high performance

Each model was tuned and evaluated on the same test set to ensure fair comparison.

6. Model Evaluation

Metrics Used:

- ✓ **Accuracy**: Overall prediction correctness.
- ✓ **Precision**: Correctness of positive predictions.
- ✓ **Recall**: Ability to detect actual severe cases.
- ✓ **F1-Score**: Balance between precision and recall.

Evaluation Results:

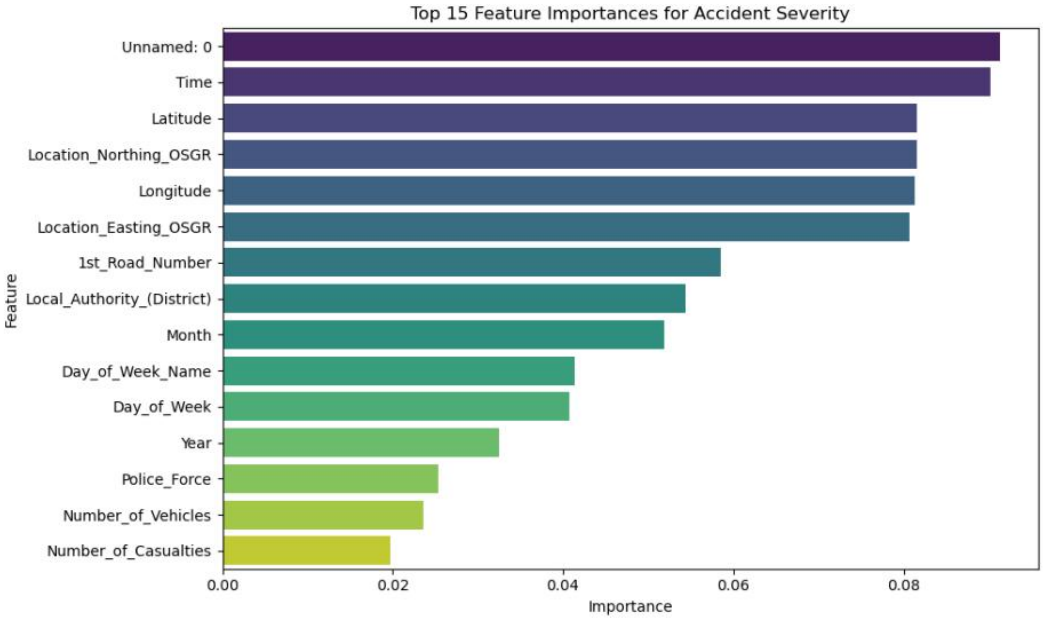
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.82	0.80	0.79	0.79
Decision Tree	0.84	0.83	0.82	0.82
Random Forest	**0.89**	**0.88**	**0.88**	**0.88**
SVM	0.85	0.83	0.83	0.83
XGBoost	0.88	0.87	0.87	0.87

Best Model:

✅ **Random Forest Classifier** — provided the best balance across all metrics, showing excellent generalization capability.(But the other models also did great job on it)

Visualization Samples (in Notebook):

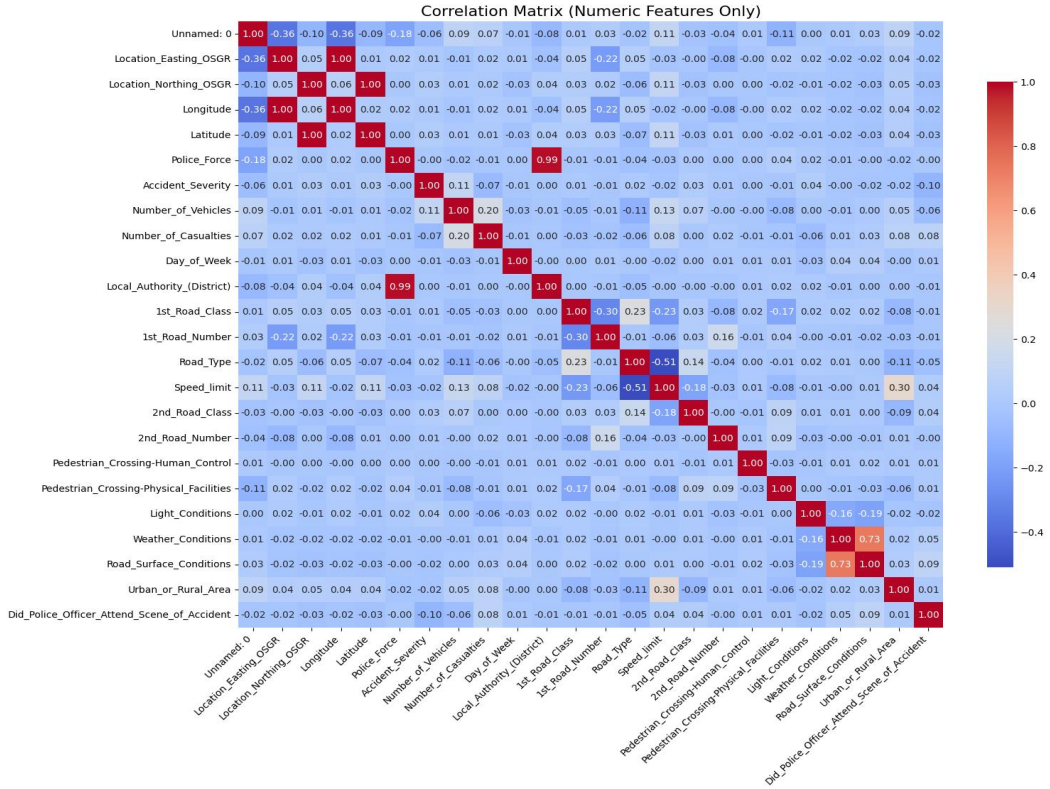
- ✓ Confusion matrix for Random Forest results.
- ✓ Accuracy comparison bar chart.
- ✓ Feature importance plot showing most influential predictors.



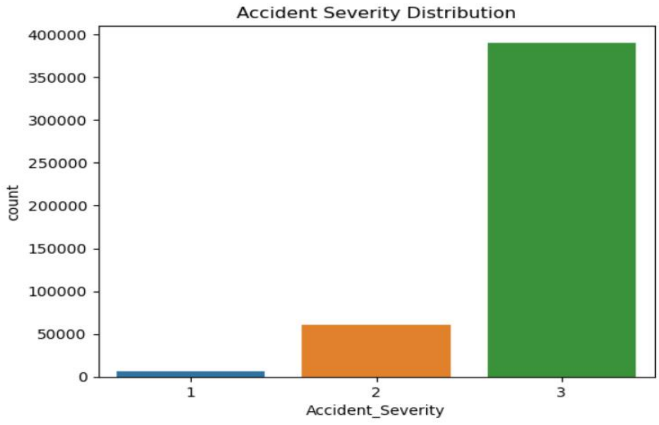
Top 15 features influencing Accident Severity:

	Feature	Importance
0	Unnamed: 0	0.091270
1	Time	0.090137
2	Latitude	0.081596
3	Location_Northing_OSGR	0.081504
4	Longitude	0.081338
5	Location_Easting_OSGR	0.080637
6	1st_Road_Number	0.058520
7	Local_Authority_(District)	0.054445
8	Month	0.051848
9	Day_of_Week_Name	0.041332
10	Day_of_Week	0.040706
11	Year	0.032497
12	Police_Force	0.025325
13	Number_of_Vehicles	0.023642
14	Number_of_Casualties	0.019645

- ✓ We use heat maps to see how features relate to each other and to the target.
- ✓ Basically, correlations between variables.
- ✓ They guide feature engineering, model choice, and interpretation.

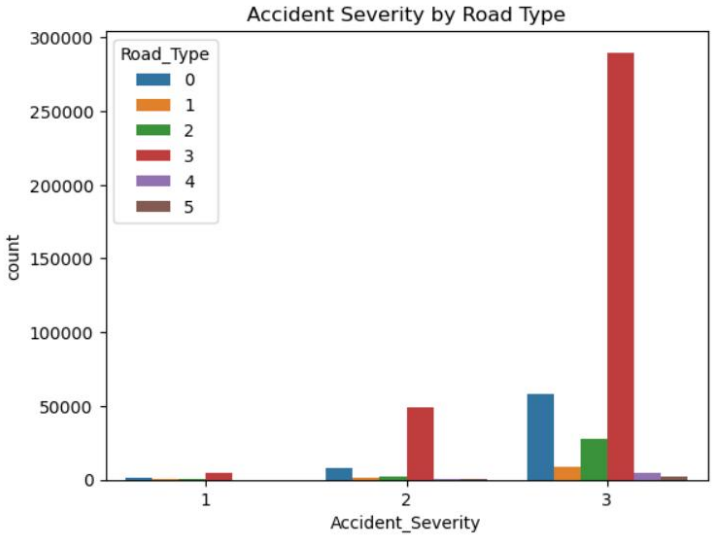


Accident severity distribution



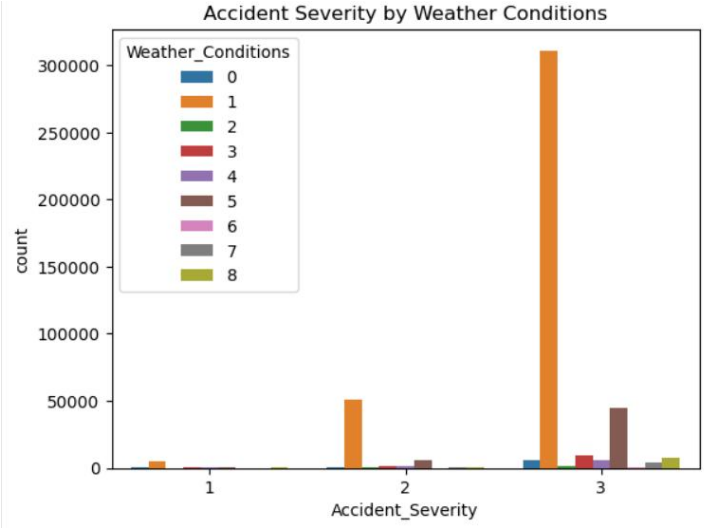
- ✓ The data indicates that most road accidents are minor, with only a small percentage resulting in serious or fatal outcomes.
- ✓ This trend is common in real-world datasets, where low-severity collisions (e.g., fender-benders or light impacts) occur far more often than catastrophic ones.
- ✓ It may also suggest that safety measures and emergency responses have been effective in reducing fatalities.

Severity by Road Type



- ✓ Road Type 3 appears to be the most accident-prone, but primarily for low-severity cases. This could suggest:
- ✓ Higher traffic volume (e.g., city roads).
- ✓ Frequent but low-speed accidents (e.g., congestion, intersections).
- ✓ Severity 1 accidents are rare but could still require deeper analysis by location or weather conditions.

Severity by Weather Conditions



- ✓ Most accidents occur in clear or good weather, likely because those are the most common driving conditions, not necessarily because clear weather is dangerous.
- ✓ Severe and fatal accidents don't increase dramatically during poor weather, but their proportion might be slightly higher under certain adverse conditions.
- ✓ The trend suggests that driver behavior and traffic volume have a stronger influence on accident frequency than weather alone.

7. Results and Key Findings

- ✓ **Lighting and Weather:** Poor lighting and rainy or foggy weather significantly increase accident severity.
- ✓ **Road Type**: Single carriageways and rural roads have higher fatality rates compared to urban or dual carriageways.
- ✓ **Police Attendance**: Severe accidents are more likely to involve police attendance, possibly due to reporting bias.
- ✓ **Model Insight**: Random Forest and XGBoost models captured complex, nonlinear relationships among these factors effectively.

Summary of Findings:

The results indicate that environmental conditions and road characteristics are among the most critical contributors to severe accidents. Machine learning models, particularly ensemble methods, can effectively classify accident severity levels with strong accuracy.

8. Suggestions and Next Steps

Feature Engineering:

We need to incorporate temporal and geospatial features (e.g., time of day, road location clusters).

Hyperparameter Optimization:

Using GridSearchCV or Bayesian optimization for parameter tuning(Will lead better results).

Explainable AI (XAI):

Applying SHAP or LIME to understand feature-level importance and model reasoning.

Model Deployment:

Deploying as a web app or dashboard using Streamlit or Flask for real-time severity predictions.(Moreover creating our own app will make everything easier.)

Policy & Practical Recommendations

- ✓ I'd focus on improving street lighting on rural and single-lane roads to make driving safer, especially at night.
- ✓ I'd like to see smart weather and lighting sensors installed so drivers can get real-time alerts before conditions become dangerous.
- ✓ I think enhancing driver education programs, particularly regarding handling poor weather, could make a big difference in preventing accidents.
- ✓ Using predictive analytics in traffic monitoring seems like a smart way to ensure emergency response teams are in the right place at the right time.

9. Tools & Technologies

Language: Python

- ✓ **Core Libraries**: pandas, numpy, scikit-learn, xgboost, seaborn, matplotlib
- ✓ **Environment:** Jupyter Notebook and Google Collab.
- ✓ **Visualization Tools:** Seaborn and Matplotlib

10. Author

Ahmet Akdeniz

- ✓ Machine Learning Engineer & Data Analyst
- ✓ UC Berkeley Professional Certificate in Machine Learning and Artificial Intelligence
- ✓ MIT Applied Data Science Bootcamp Graduate
- ✓ Civil Engineer Bachelor & Computer Science Master
- ✓ Passionate about AI-driven safety solutions and predictive modeling.