FINAL PROJECT REPORT

ROAD ACCIDENT PREDICTING SYSTEM

Abstract:

The project aims to leverage the 2012 dataset on personal injury road accidents in Great Britain to develop a data science system addressing global concerns about road safety. By analysing crucial information on accident causes and outcomes, the project seeks to inform policies and interventions through data-driven insights. The key findings are expected to uncover contributing factors to accidents, identify high-risk areas, and construct predictive models for accident prevention. The project's impact extends to saving lives, reducing injuries, and lowering societal and economic costs associated with road accidents. A specific use case involves creating a predictive model that forecasts accident likelihood based on vehicle speed in specific areas. This tool can assist traffic management authorities and law enforcement agencies in efficiently allocating resources by providing a risk score for given locations and speed limits, enabling prioritized deployment of patrol routes and safety measures. The prediction models consider Latitude, Longitude, speed limit, and accident severity as key columns.

Background/motivation:

Road safety is a critical global concern with far-reaching societal and economic implications. In Great Britain, the 2012 dataset on personal injury road accidents offers a comprehensive repository of information regarding accident causes and outcomes. The motivation behind this project stems from the need to harness the power of data science to inform and improve road safety policies and interventions. By analysing the dataset, we can gain valuable insights into the factors contributing to accidents, identify high-risk areas, and construct predictive models for accident prevention.

What makes this project particularly interesting is its potential to address a pressing global issue with a combination of data-driven insights and practical applications. The uniqueness lies in the ability to not only understand the dynamics of road accidents but also to translate these insights into tangible tools for authorities. The focus on creating a predictive model based on vehicle speed adds a distinctive dimension to the project, offering a targeted solution for allocating resources efficiently. This real-world application, allowing authorities to forecast accident likelihood in specific areas, sets the project apart by providing a proactive approach to road safety management. The project's holistic approach, combining analysis, prediction, and practical tools, makes it uniquely positioned to contribute significantly to saving lives, reducing injuries, and minimizing the societal and economic impact of road accidents.

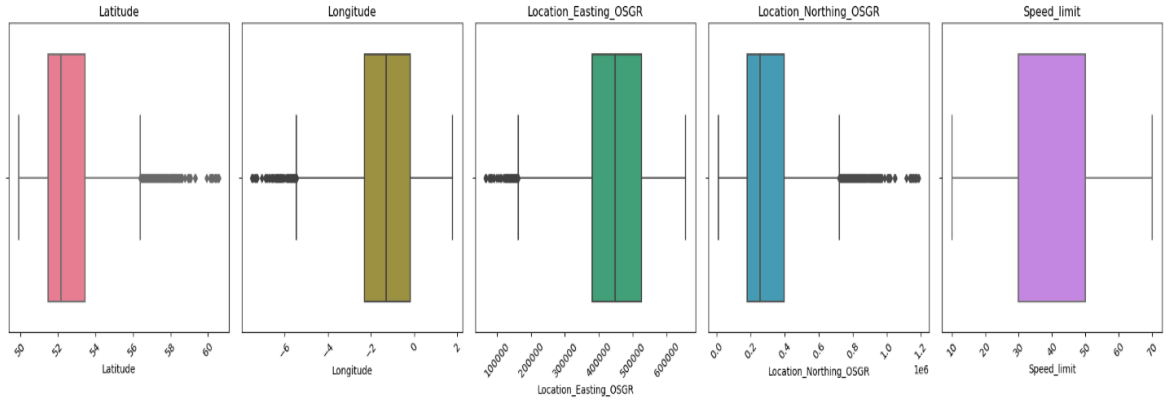
Exploratory data analysis:

Road Accident Predicting System dataset comprises 145,571 records with 32 columns of various attributes related to reported accidents. Notably, the dataset shows that LSOA (Lower Layer Super Output Area) of the Accident_Location attribute has some null values, indicating some missing data. So, we have dropped the LSOA record. The majority of attributes consist of non-null values. This dataset provides a comprehensive overview of factors related to reported accidents, offering insights into various elements surrounding these incidents, albeit with some missing locality details. The Road Accident Predicting System dataset does not have many null values. Certainly! Converting a date column to a datetime format is a

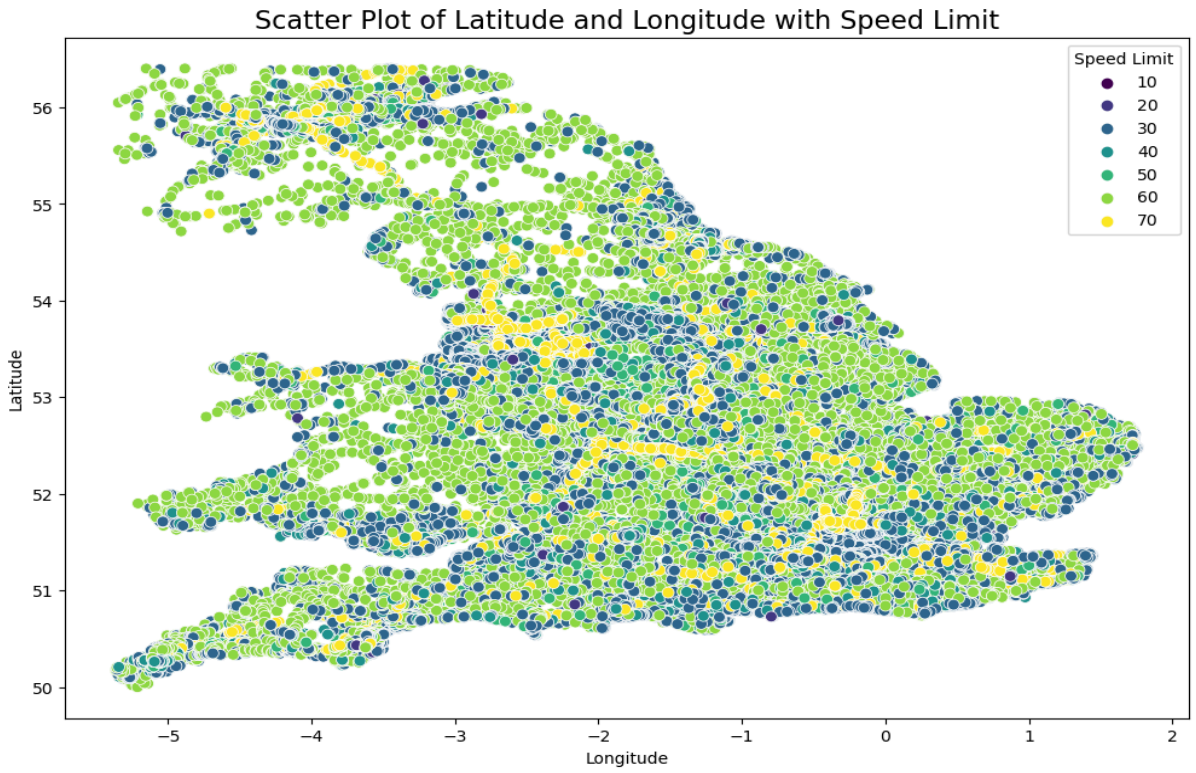
common preprocessing step in data analysis or machine learning tasks. With this the size of the dataset has become 145566 rows x 34 columns.

Model development:

Outliers are points that fall beyond the calculated bounds. Outliers can have a significant impact on the performance, causing them to overfit the training data and perform poorly on the test data. After removing the outliers, the final size of the dataset is 143178 rows x 34 columns.

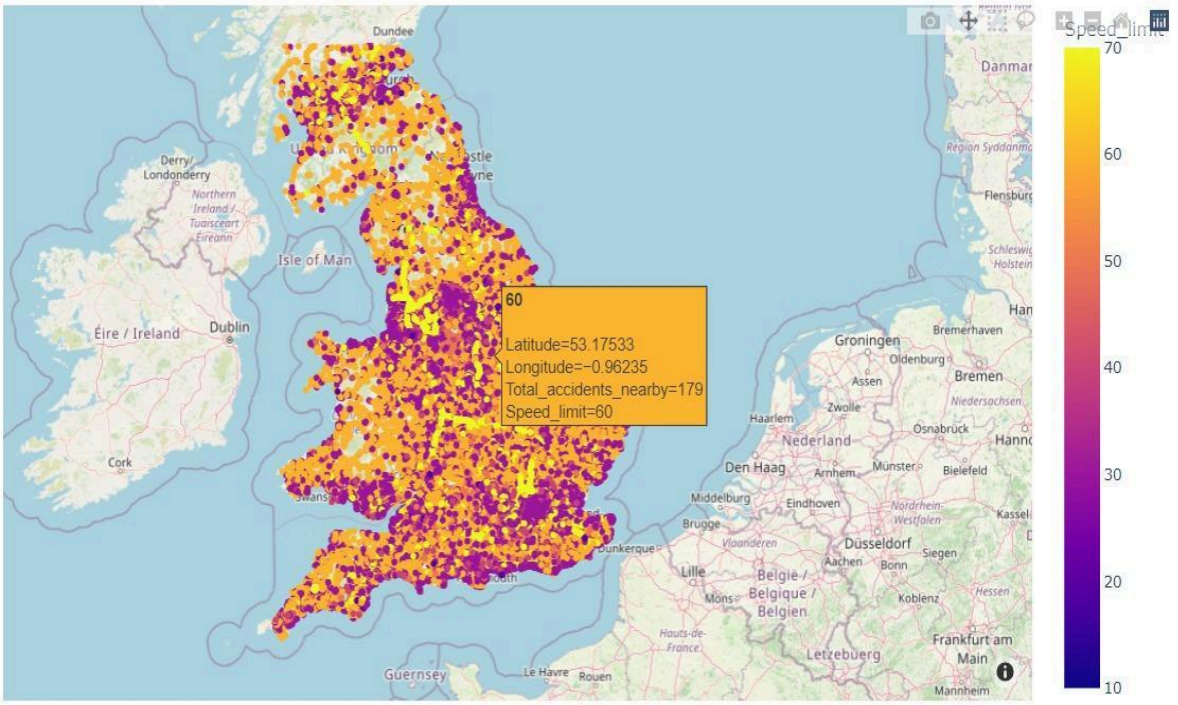


The scatter plot for Latitude and Longitude is given below with the speed limit aside. And the outliers are removed using z-score and interquartile range method



Result and Insights:

The spatial analysis of the Great Britain 2012 Road Accident Severity dataset according to speed limit is as follows.

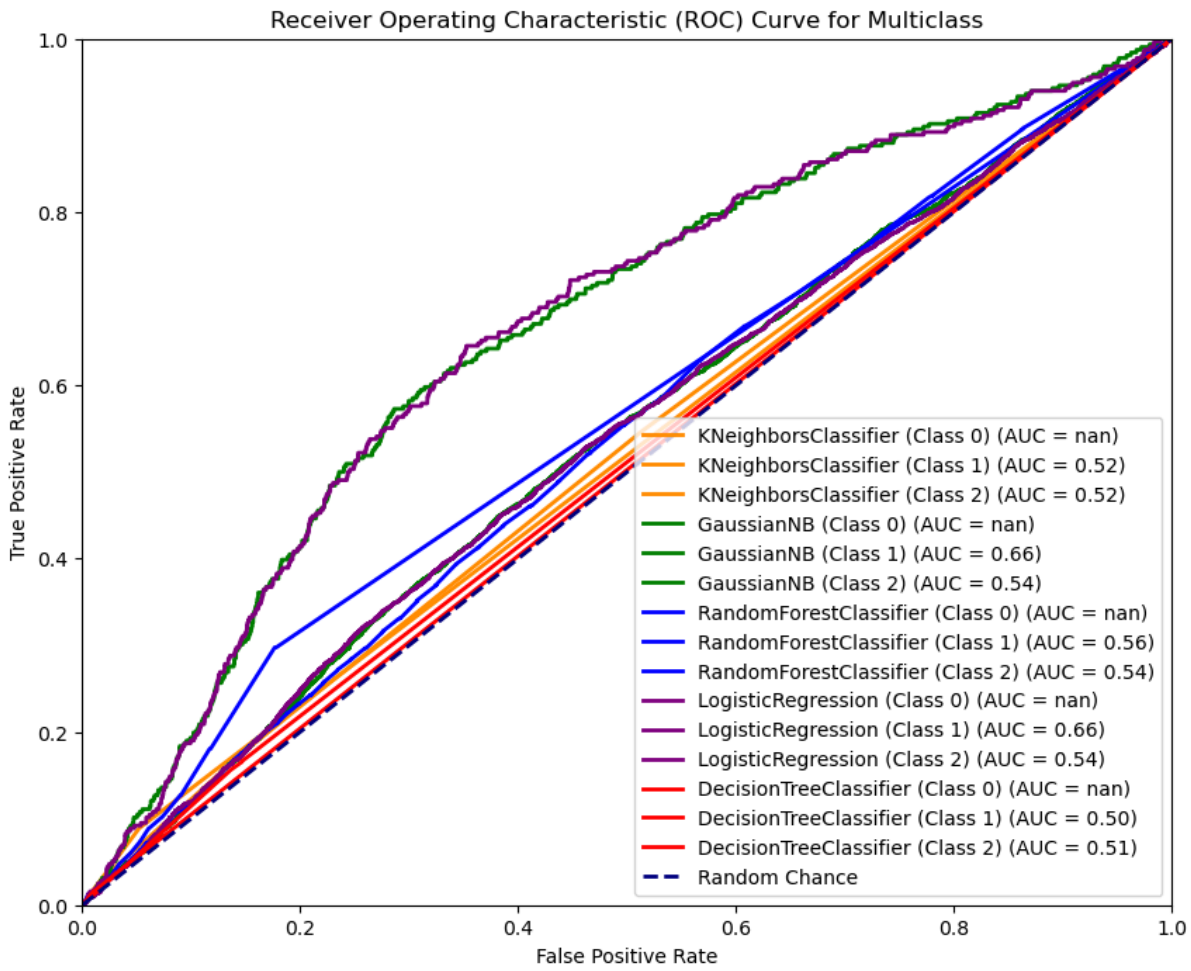


The final result is as follows with a videography of our observation.

https://drive.google.com/drive/u/1/folders/0ACL1c4NbIwBpUk9PVA

Evaluation Metrics:

	Precision	Recall	F1 – score	Support	Accuracy
Logistic Regression	0.66	1.00	0.79	18809	0.6568
Random Forest	0.85	0.91	0.88	18809	0.7751
Decision Tree	0.87	0.87	0.87	18809	0.753
Gaussian NB	0.66	0.97	0.79	18809	0.6513
KNN	0.87	0.88	0.87	18809	0.7616



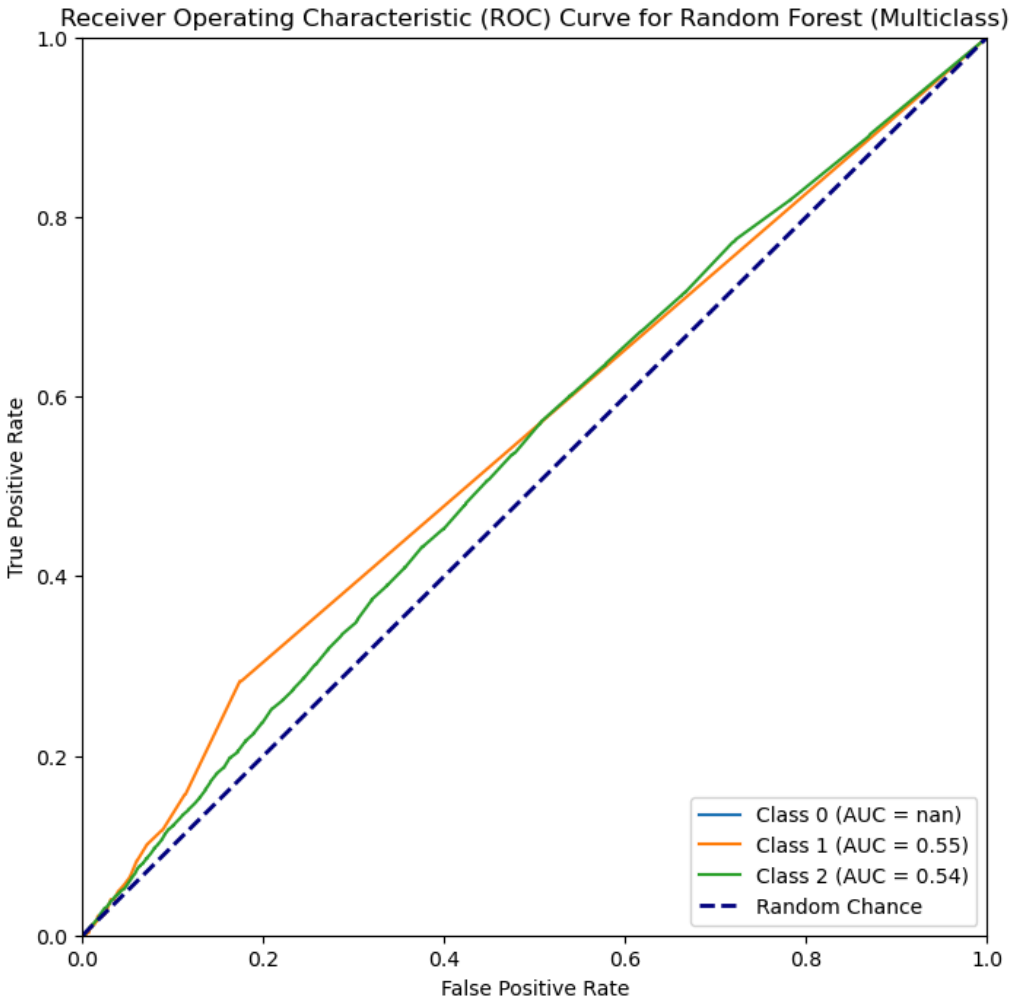
Conclusion:

We have chosen the Random Forest model among all other evaluation metrics. Because the accuracy for this model is high when compared to other models. The confusion matrix for Random Forest is as follows:

```
Accuracy: 0.7751431764212879
Confusion Matrix:
[[    0     0     1     0     0     0     0]
 [    0    91   276    14     5     9     4]
 [    0    85 17198   505   137   678   206]
 [    0     7  1123   837    60   247    99]
 [    0     2   377    70   337   175    71]
 [    0     6   927   165   110  2603   245]
 [    0     1   438    77    56   263  1131]]
Classification Report:
              precision    recall  f1-score   support

          10       0.00      0.00      0.00         1
          20       0.47      0.23      0.31       399
          30       0.85      0.91      0.88     18809
          40       0.50      0.35      0.41      2373
          50       0.48      0.33      0.39      1032
          60       0.65      0.64      0.65      4056
          70       0.64      0.58      0.61      1966

    accuracy                           0.78     28636
   macro avg       0.51      0.43      0.46     28636
weighted avg       0.76      0.78      0.76     28636
```



In conclusion, leveraging the 2012 dataset on road accidents in Great Britain, this project aims to enhance global road safety. Through data-driven insights, it identifies accident causes, high-risk areas, and constructs predictive models. The emphasis on forecasting accident likelihood based on vehicle speed provides a targeted solution, enabling efficient resource allocation and prioritized safety measures.

Future Work:

The future scope of this project could involve integrating weather conditions as a predictive factor for road accidents. By incorporating weather data, such as precipitation, visibility, and road surface conditions, the model could provide more comprehensive insights. This enhancement would contribute to a more robust accident prediction system, allowing authorities to implement preventive measures tailored to specific weather-related risks. Additionally, the expanded model could facilitate a deeper understanding of the dynamic interplay between weather and road safety, thereby advancing the overall effectiveness of accident prevention strategies.

References:

- Reported Road Casualties Great Britain: Annual Report 2012: Published by the Department for Transport, this report provides detailed statistics on personal injury accidents, including types of vehicles involved, resulting casualties, and factors contributing to accidents.

- Reported Road Collisions, Vehicles, and Casualties Tables for Great Britain: This resource offers detailed statistics about reported personal injury road collisions in Great Britain, with data on vehicles, casualties, and other relevant factors. It also includes various statistical tables in Excel format, providing a comprehensive data analysis.

- Road Safety Data on data.gov.uk: This platform provides a wide range of road safety data, including information on vehicles, casualties, and collisions. The data is presented in various formats, including CSV, and covers several years, making it a valuable resource for longitudinal studies.

Comments on Feedback on Milestone report:

1. The section on motivation of the project is added.
2. The percentage of null values are very negligible like 3% and that too only one column contains null values which we don't use.
3. How outliers are handled is mentioned in this report.
4. Provided detailed analysis on performance metrics.
5. Problem formulation was stated.
6. Confusion Matrix was shown in this report.
7. Visualizations are provided with great understanding.

Online Platforms:

Platforms like GitHub can be valuable for finding code snippets, and project ideas. Search for relevant notebooks or repositories related to road safety or predictive modeling.

Programming Languages:

Python: Widely used for data science. Libraries like Pandas, NumPy, Scikit-Learn are commonly employed.

Machine Learning Libraries:

Scikit-Learn: Comprehensive library for machine learning in Python.

Notebooks and IDEs:

Jupyter Notebooks: For interactive development and data exploration.

Github:

https://github.com/NikhilChaganti/733-FINAL-CODE/tree/main

Link to YouTube:

https://youtu.be/OS0H-HbnUGE