Traffic Collision Driver Injury Analysis and Prediction with Machine Learning

Louise Heaney
School of Computing
National College of Ireland
Dublin, Ireland
x23173921@student.ncirl.ie

***Abstract*—Road traffic collisions are a significant public safety issue, leading to a vast number of injuries, fatalities, and economic losses worldwide. Despite advances in road safety, the ability to accurately predict the severity of driver injuries remains challenging. This research seeks to address this challenge by leveraging machine learning techniques to develop predictive models for driver injury severity. By analysing a comprehensive dataset of road collisions from Montgomery County, Maryland, this study evaluated five machine learning models - Random Forest, Gradient Boosting Machines (XGBoost), K-Nearest Neighbours (KNN), Naïve Bayes, and Support Vector Classifier (SVC) - to assess their performance in predicting one of five levels of driver injury severity. Initial results indicated that all models struggled with class imbalance and low predictive accuracy, with Random Forest and KNN performing slightly better in the initial evaluation. Feature enhancement through recursive feature elimination revealed the features vehicle age and distance from the hospital as significant predictors, which improved the overall accuracy of these two models on the second evaluation. Despite these improvements, the models continued to perform poorly in predicting minority classes. Future work might focus on addressing this class imbalance more effectively and exploring additional data sources and advanced algorithms to enhance prediction accuracy.**

Keywords—road traffic collisions, injury severity prediction, machine learning, predictive modelling, random forest, gradient boosting machines, k-nearest neighbours, naïve Bayes, support vector classifier, class imbalance, feature selection

I. INTRODUCTION

A. Background

Road traffic collisions are a significant public safety issue resulting in injuries, fatalities and property damage. According to the World Health Organisation (WHO) [1], almost 1.2 million people die each year as a result of these accidents, making traffic injuries the leading cause of death for children and young adults aged 5-29 years old. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. In addition, they represent a substantial economic cost to the individuals involved as well as local and, sometimes, national government authorities and agencies. Traffic collisions account for as much as 3% of a country's gross domestic product (GDP) [1], which can significantly impact a country's economic growth and development.

B. Motivation and Potential Benefits

Understanding and predicting the severity of injuries resulting from traffic collisions is critical for improving public safety and optimising resource allocation. By developing accurate predictive models, decision makers can enhance emergency response times, tailor medical interventions and inform future policy-making. For example, based on predicted injury severity, the ambulance service could prioritize resources resulting in better outcomes for injured drivers, and urban planners would be able to identify high risk locations and scenarios and plan accordingly.

C. Problem Statement

Despite recent advances in road safety, the accurate prediction of injury severity, in particular that of the driver, remains a significant challenge. Existing methodologies often lack the precision needed to provide actionable insights, leading to the afore-mentioned inefficiencies. This research aims to address these gaps by leveraging the power of machine learning to develop more accurate predictive models focusing specifically on driver injury severity.

D. Aim of the Research

The primary objective of this research is to analyse historical data about road collisions in order to develop and evaluate a selection of machine learning models that will accurately predict driver injury severity.

II. RELATED WORK

A. Introduction

Traffic collision analysis has evolved significantly over the past number of years, with machine learning emerging as a powerful tool in the analysis and prediction of driver injury severity. By leveraging large datasets and sophisticated algorithms, machine learning models can uncover complex, non-linear relationships between variables that traditional statistical methods have previously overlooked [2].

B. Literature Review

Notable case studies have highlighted the fact that driver characteristics play a significant role in influencing injury severity. A study by Bédard et al. (2002) found that older drivers were more vulnerable to severe injuries, while drivers under thirty years old were less likely to be involved in a fatal collision than drivers aged in their forties . Gender also played a role, with research by Islam and Mannering (2006) showing that male drivers generally experience more severe injuries compared to female drivers [4]. This latter study also explored interaction effects between the variables which revealed an increased risk of fatality for young, male drivers who were carrying passengers, and an increased risk of injury for middle-aged female drivers whose cars were six years old or more [4].

Vehicle features also significantly impact injury severity with a study by Savolainen et al. (2011) reporting that drivers of smaller vehicles were at a higher risk of severe injuries compared to those in larger vehicles [5]. A study by

Fu and Lee (2022) explored the validity of the safety ratings awarded to vehicles by the U.S. National Highway Traffic Safety Administration. In addition to proving that the 5-star safety rating did indeed have the intended effect of lowering the chances of serious injury, their findings revealed that female drivers were more likely to be seriously injured than males when driving vehicles with the exact same safety rating [6].

Time was another factor found to elicit an influence on the severity of driver injuries. A study by Huang et al. (2008) found that night time collisions (8pm – 7am) resulted in more severe injuries than collisions during daytime (10am – 5pm) [7]. An earlier study by Aguero-Valverde and Jovanis (2006) explored how time of year affected injury severity and concluded that collisions in winter months had a higher likelihood of resulting in severe injuries, due to seasonal weather conditions such as slippery road surfaces and poor visibility [8].

Location also plays a major role in the severity of injuries. Studies [3], [7], [9] have shown that rural areas exhibit higher injury severity while urban areas, conversely, exhibit a higher collision frequency but lower injury severity. Khorashadi et al. (2005) found that collisions at an intersection in a rural area resulted in an increased likelihood (725%) of severe or fatal injury whereas similar collisions in urban areas were 10.3% less likely to be serious or fatal [10].

The existing body of literature contains a diverse array of machine learning prediction models for driver injury severity. The following examples highlight a notable sample of this literature, and summarises their key methodologies and respective findings.

The first study by Abdelwahab and Abdel-Aty (2002) developed two artificial neural networks (ANN) to predict injury severity at signalized intersections. The models analysed various factors such as vehicle type, driver characteristics, and environmental conditions with the results demonstrating that the ANNs were successful at capturing complex, non-linear relationships and could effectively predict injury severity with high accuracy [11].

The second study by de Oña et al. (2011) applied Bayesian Networks to analyse injury severity in accidents on rural highways in Spain. The model considered variables like road conditions, vehicle speed, and driver behaviour. The findings indicated that Bayesian Networks provided a solid framework for identifying predictors of injury severity based on a probabilistic understanding of the factors involved [9].

The third study by Chang, L. Y., & Wang, H. W. (2006) applied non-parametric classification and regression trees (CART). The study included a variety of predictor variables such as driver age, road type, and weather conditions. While not as accurate as SVMs or ANNs, the CART models efficiently identified factors influencing injury severity and displayed substantial predictive performance that proved very effective in this analysis [12].

Another noteworthy study by Boulieri et al. (2017) employed a space-time multivariate Bayesian model to analyse road traffic accidents by severity. This model incorporated both spatial and temporal data to provide a more comprehensive understanding of accident severity patterns over time and across different locations. The study found that this approach significantly improved the accuracy of severity predictions and highlighted the importance of considering both spatial and temporal factors in traffic accident analysis [8].

Finally, a review by Santos et al. (2022) provided a detailed roundup of fifty-six studies dating from 2001 to 2021 and an assessment of the 20 statistical and machine learning prediction techniques that those studies comprised [13]. The key findings were that Random Forest performed the best overall for predicting driver injury severity, followed by Support Vector Machine (SVM), Decision Tree and K-Nearest Neighbours [13].

Based on these findings, a comparative analysis will be conducted on the performance of Random Forest, Gradient Boosting Machines (XGBoost), K-Nearest Neighbours, Naïve Bayes and Support Vector Classifier (SVC). Each model will be trained and evaluated using metrics such as accuracy, precision, recall, F1-score and ROC curve. The comparison will involve a thorough analysis of each model's strengths and weaknesses in predicting the target feature that should guide the development of a reliable prediction system for driver injury severity.

III. RESEARCH METHODOLOGY

A. Research Design

The Cross Industry Process for Data Mining method (CRISP-DM) was chosen for this analysis and modelling project because of its well-defined, structured approach and adaptability, making it highly effective for data mining projects [14]. The six stages of CRISP-DM are shown in the figure below:

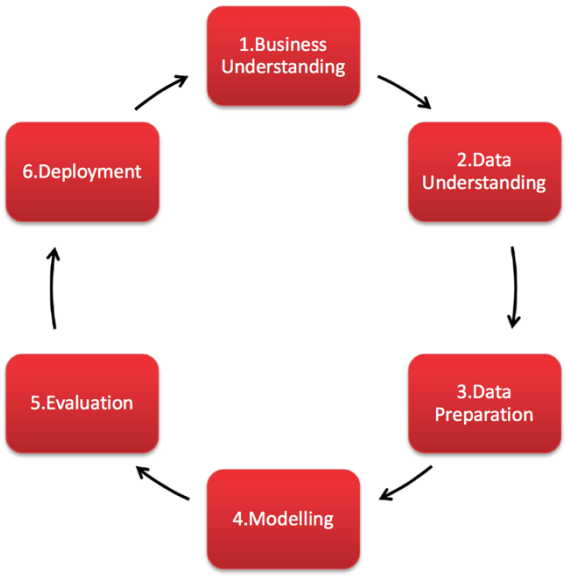

Figure 1: Flow chart of the six stages of CRISP-DM [15]

Stage 1, Business Understanding, has already been discussed in the preceding chapters. For stage 2, Data Understanding, a thorough exploratory data analysis (EDA) was conducted to identify patterns and understand data distribution. The quality of data was also assessed by

checking for missing and inconsistent data to be addressed in Stage 3, Data Preparation. At this third stage, the median and mode was used to fill missing values and the median was also used to replace a large proportion of erroneous outliers. Additionally, feature engineering was implemented to create new variables such as distance from the largest urban centre and a selection of time variables including Day of Week and Quarter. Data transformation in the form of numeric standardization ensured that the data was ready for Stage 4, Modelling, which will be discussed in Chapter IV.

B. Revised Timeline and Milestones

The project was initially scheduled for completion over a 9-week period, with specific phases designed to ensure thorough development and review. The timeline was structured as follows:

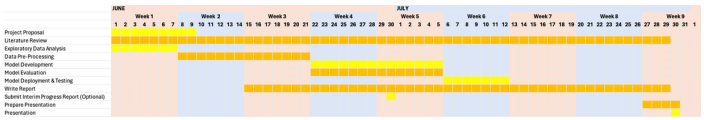

Figure 2: Gant chart of delivery timeline

- Exploratory Data Analysis (EDA) and Pre-processing: Week 1 - Week 3. This stage involved essential data exploration and preparation, which would set the foundations for the development of the predictive models.

- Model Development: Week 4 - Week 5. During this phase, the five models were created, tuned and validated. The intentional overlap with the model integration and testing phase allowed for a continuous feedback loop.

- Integration and Testing: Week 7 - Week 8. The models were integrated, and rigorous testing was conducted to ensure optimal performance and accuracy.

- Finalization and Report Preparation: Week 8 - Week 9. The final report was written during this period, which went in tandem with the finalising the presentation.

Due to a scheduling adjustment, the presentation was moved forward by one week, which necessitated an earlier completion of the project. Despite the compressed timeline, the report was successfully finalized and delivered one week ahead of the original deadline which allowed ample time for the presentation, and ensured that all project deliverables were completed on schedule.

C. Data Understanding

The dataset was obtained from the Montgomery County of Maryland's data repository [16] and the analysis was conducted on a Macbook with an Apple M1 chip and 8GB of unified memory, ensuring efficient data processing and compatibility with the latest software tools necessary for analysis and model prediction. In total, the dataset contained 6,795,163 cells of original data, consisting of 172,105 records with 43 attributes. The data was recorded at incident level, meaning that each row contains the record of a single vehicle involved in a road traffic collision in the county from January 1st 2015 to December 31st 2023. The dataset was loaded into the Python environment, Jupyter Noteboook, for initial EDA which revealed several attributes with missing values.

D. Data Preparation

Data cleaning and preparation involved dropping any attributes that did not provide information that could be used to predict driver injury. These redundant features included Local Case Number, Agency Name, Road and Cross Street Name, Municipality, Person ID, Vehicle ID, Drivers License State and Related Non-Motorist.

#	Column	Non-Null Count	Dtype
0	Report Number	172105 non-null	object
1	Local Case Number	172105 non-null	object
2	Agency Name	172105 non-null	object
3	ACRS Report Type	172105 non-null	object
4	Crash Date/Time	172105 non-null	object
5	Route Type	155132 non-null	object
6	Road Name	156168 non-null	object
7	Cross-Street Type	155099 non-null	object
8	Cross-Street Name	156154 non-null	object
9	Off-Road Description	15935 non-null	object
10	Municipality	19126 non-null	object
11	Related Non-Motorist	5463 non-null	object
12	Collision Type	171520 non-null	object
13	Weather	158751 non-null	object
14	Surface Condition	151987 non-null	object
15	Light	170660 non-null	object
16	Traffic Control	146636 non-null	object
17	Driver Substance Abuse	140781 non-null	object
18	Non-Motorist Substance Abuse	4317 non-null	object
19	Person ID	172105 non-null	object
20	Driver At Fault	172105 non-null	object
21	Injury Severity	172105 non-null	object
22	Circumstance	31359 non-null	object
23	Driver Distracted By	172105 non-null	object
24	Drivers License State	162155 non-null	object
25	Vehicle ID	172105 non-null	object
26	Vehicle Damage Extent	171789 non-null	object
27	Vehicle First Impact Location	171949 non-null	object
28	Vehicle Second Impact Location	171849 non-null	object
29	Vehicle Body Type	169456 non-null	object
30	Vehicle Movement	171719 non-null	object
31	Vehicle Continuing Dir	169416 non-null	object
32	Vehicle Going Dir	169416 non-null	object
33	Speed Limit	172105 non-null	int64
34	Driverless Vehicle	172105 non-null	object
35	Parked Vehicle	172105 non-null	object
36	Vehicle Year	172105 non-null	int64
37	Vehicle Make	172081 non-null	object
38	Vehicle Model	172039 non-null	object
39	Equipment Problems	137964 non-null	object
40	Latitude	172105 non-null	float64
41	Longitude	172105 non-null	float64
42	Location	172105 non-null	object

Figure 3: Null value count for every column

In the majority of cases, missing values for categorical data were filled with the mode, while numerical variables were filled with the median. In a minority of cases, for example the Equipment Problem variable, missing values were understood to imply that there was no problem so this was filled with 'No Issues' instead of the mode. Attributes with negative or implausible values were corrected by replacing them with the median value, while obvious data input errors were handled on a case by case basis with, for example, Vehicle Year entries such as 99 and 9999 both being replaced by 1999. Three features, Route Type, Cross Type and Off-Road Description, were found to have corresponding and/or supplementary data. As such, data from Route Type and Off Road Description were used to fill missing data in Cross Street Type, before the mode was applied to the final 1071 null-value rows.

E. Feature Engineering

The dataset contained a large number of categorical features with very few numeric features, so feature engineering was applied to the Crash Date/Time variable to generate numeric data for analysis. A number of new features were generated

including a count of the number of vehicles involved in crashes in the previous hour, as well as day of the week, month and quarter for each collision. The location co-ordinates were also used to generate the new feature Distance which measured the distance (in miles) between the collision location and Holy Cross Germantown Hospital Emergency Room.

IV. DATA ANALYSIS

F. Feature Analysis

Statistical analysis and visualization were performed to understand the overall data distribution and relationships. Key insights from the trend analysis include:

The number of collisions dropped dramatically in 2020, but despite rising steadily every year since, have not quite reached the levels of 2015-2019:

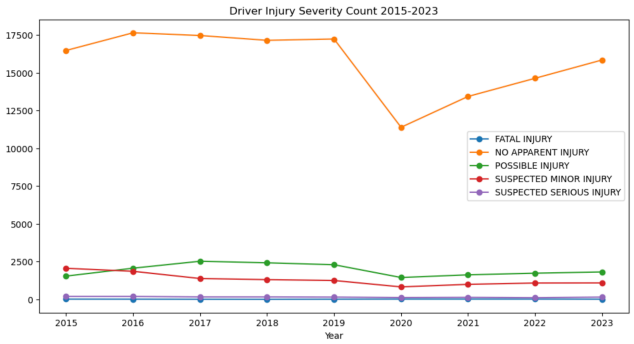

Figure 4: Number of collisions by injury severity 2015-2023

Peak crash times are 7-9am in the morning and 3-7pm in the afternoon/evening for all injury severities:

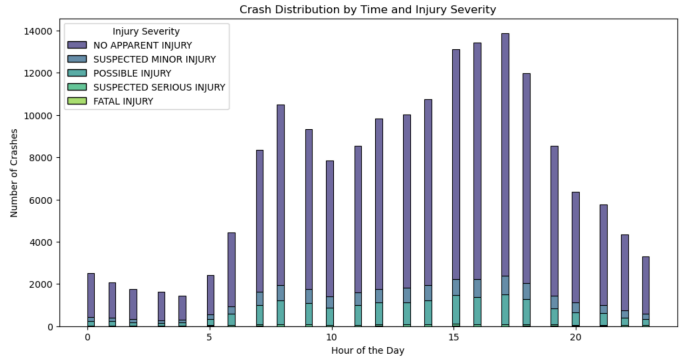

Figure 5: Crash distribution by time of day and injury severity

Collisions with a higher injury severity tended to cluster around urban centres compared to more rural locations:

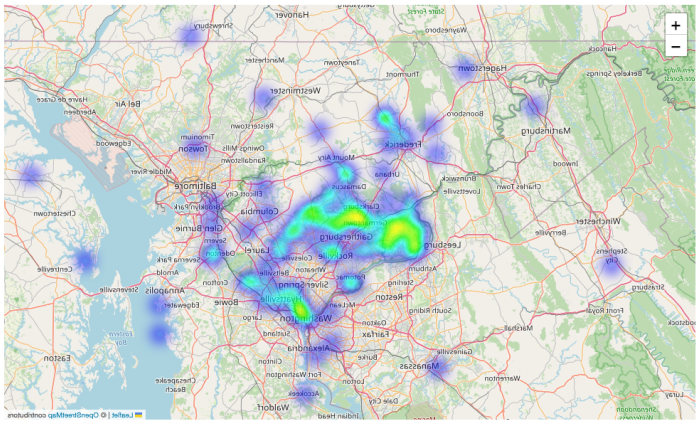

Figure 6: Geographic heatmap of crash locations based on their score for injury severity

The summary data, in the barchart and heatmap below, indicates that the majority of traffic collisions, regardless of injury severity, occur within 15-19 miles of the hospital. Possible injuries is the most common injury type across all distances, while fatal injuries is the least. Overall, as the distance increases the number of incidents decreases, particularly for both possible and minor injuries, which exhibit a notable drop off at a distance of 20-49 miles from the hospital.

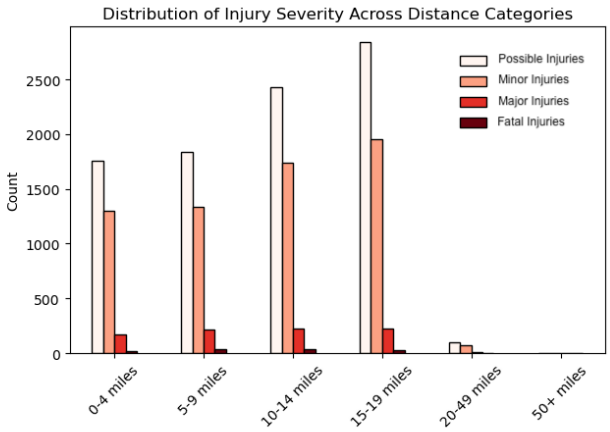

Figure 7: Distribution of injuries by severity and distance

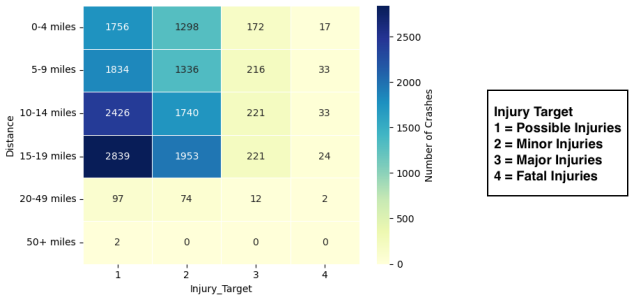

Figure 8: Heatmap of injuries by severity and distance

On average, collisions occurred approximately 11.89 miles away from Holy Cross Germantown Hospital Emergency Room with a standard deviation from the mean of 5.85 miles. The closest recorded collision to the hospital was 0.02 miles and the furthest recorded collision was 159.09 miles away. 25% of collisions were within 6.74 miles of the hospital and 75% were within 17.24 miles, with a median distance of 12.87 miles:

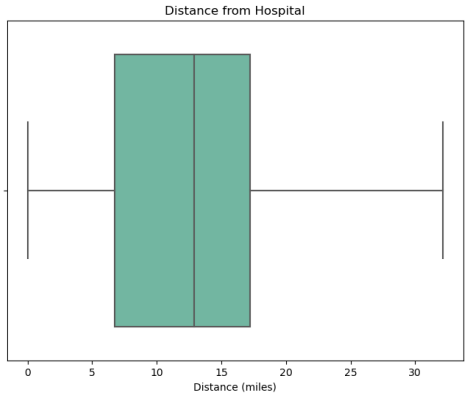

Figure 9: Distribution of collisions by distance from hospital

As a result of this analysis, the dataset was refined to include time series features such as Crash Quarter, Crash Month, Crash Day of Week, Crash Hour, Is Holiday, Is Peak Hour, and Is Weekend that were extracted from the Crash Date/Time column. These features were chosen based on the documented correlation with driver injury severity [7], [8], [9]. Other engineered features such as Distance (from hospital), Total Vehicles and Vehicle Age were also retained due to their effect on driver injury severity [3], [10]. Finally, Speed Limit was retained due to its well-established relationship with crash severity, where higher speed limits often correlate with more severe injuries [3], [5]. The intention was that this reduced dataset would enhance the accuracy and efficiency of the predictive models, while also ensuring that the features themselves were grounded in evidence-backed research.

	Speed Limit	Crash Quarter	Crash Month	Crash DayOfWeek	Crash Hour	Is Holiday	Is Peak Hour	Is Weekend	Distance	Total Vehicles	Injury_Target	Vehicle Age
0	35	1	1	3	0	1	0	0	14.88	2	0	7
1	35	1	1	3	0	1	0	0	14.88	2	0	22
2	40	1	1	3	1	1	0	0	1.64	2	0	8
3	40	1	1	3	1	1	0	0	1.64	2	0	5
4	40	1	1	3	1	1	0	0	10.48	2	2	13

Figure 10: First 5 rows of the reduced feature dataset used for modelling

G. *Target Analysis*

The data revealed that the vast majority of crashes (82.03%) resulted in no apparent injuries, while 10.16% of crashes resulted in possible injuries. More severe injuries occurred less frequently, with only 0.91% of crashes resulting in serious or fatal injuries:

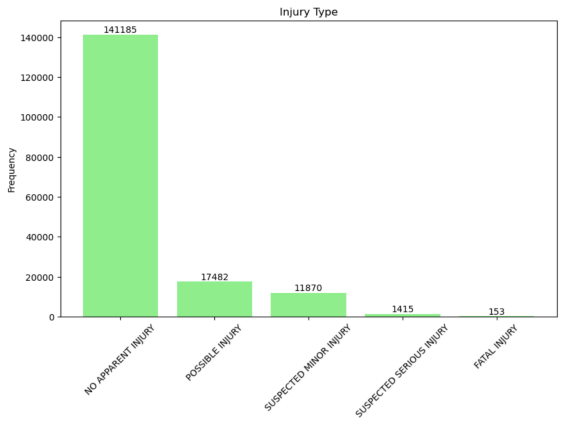

Figure 11: Distribution of driver injury severity

This represented a significant class imbalance of the target feature which was essential to address before training the model to ensure that all levels of injury severity were predicted accurately. Two techniques were employed to create a more balanced dataset: oversampling the minority classes and undersampling the majority class. SMOTE (Synthetic Minority Over-sampling Technique) was used to generate synthetic examples for the underrepresented classes, while random undersampling was used to reduce the number of majority class instances.

V. RESULTS

The following section is a summary of the performance of each of the five machine learning models' predictive capability on the test dataset consisting of 34,421 samples.

A. *Random Forest*

The results from the Random Forest model indicated a low overall accuracy and a varied performance across the 5 different classes. The model performed well for class 0 but struggled significantly with the other classes, particularly classes 3 and 4, which had very low precision, recall and F1-scores.

```
Accuracy: 0.5181139420702479
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.57      0.68     28214
           1       0.13      0.30      0.18      3553
           2       0.10      0.24      0.14      2364
           3       0.02      0.11      0.04       254
           4       0.00      0.00      0.00        36

    accuracy                           0.52     34421
   macro avg       0.22      0.25      0.21     34421
weighted avg       0.72      0.52      0.59     34421
```

Figure 12: Classification report for random forest model

The ROC AUC scores of 0.58 are only slightly better than the score expected by random chance (0.5), indicating that the model has some ability to distinguish between classes but it is not particularly effective.

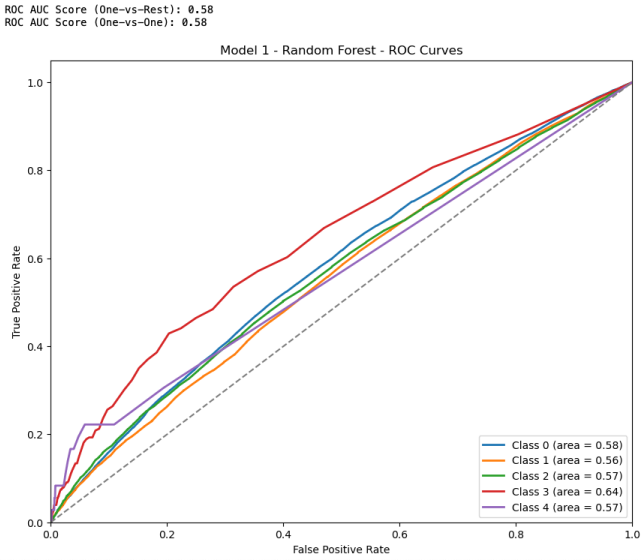

Figure 13: ROC curve for random forest model

B. *Gradient Boosting Machine (XGBoost)*

The XGBoost model demonstrated a n under-par performance, achieving an overall accuracy of 46.06%.

Accuracy: 0.46061997036692715
Classification Report:

	precision	recall	f1-score	support
0	0.85	0.50	0.63	28214
1	0.13	0.31	0.18	3553
2	0.09	0.21	0.12	2364
3	0.02	0.19	0.03	254
4	0.00	0.08	0.01	36
accuracy			0.46	34421
macro avg	0.22	0.26	0.20	34421
weighted avg	0.72	0.46	0.55	34421

Figure 14: Classification report for gradient boosting machine (XGBoost) model

This time, the model performed well in predicting class 0 with high precision but it struggled significantly with other classes. The model's precision and recall were particularly poor for classes 1, 2, 3, and 4, reflecting issues with class imbalance and poor classification for minority classes.

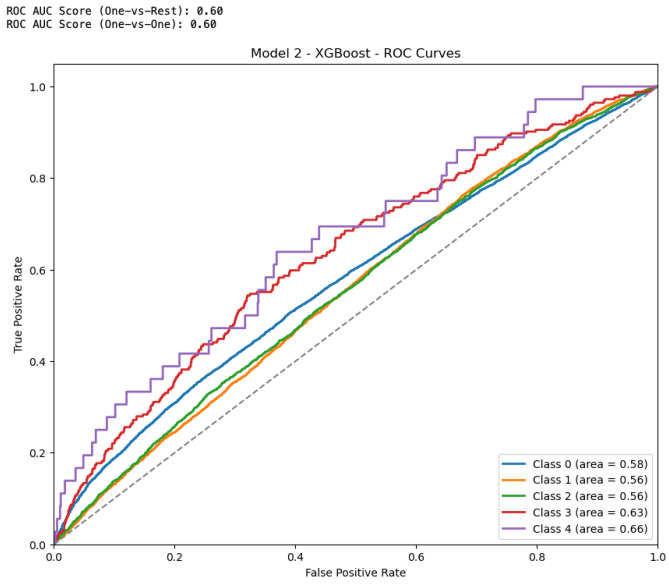

Figure 15: ROC curve for gradient boosting machine (XGBoost) model

The ROC AUC scores of 0.60 suggest some ability to differentiate between the classes but there is definite room for improvement.

C. K-Nearest Neighbours

The K-Nearest Neighbours (KNN) classifier with 2-neighbours only achieved an accuracy of 52.8% on the dataset.

Accuracy: 0.5284564655297638
Classification Report:

	precision	recall	f1-score	support
0	0.83	0.60	0.70	28214
1	0.11	0.24	0.15	3553
2	0.08	0.15	0.10	2364
3	0.01	0.07	0.02	254
4	0.00	0.03	0.01	36
accuracy			0.53	34421
macro avg	0.21	0.22	0.20	34421
weighted avg	0.70	0.53	0.59	34421

Figure 16: Classification report for k-nearest neighbours model

Once again, the model is capable of classifying the majority class 0 with a precision of 0.83 and a recall of 0.60 but the classification report shows that it has difficulty predicting the less frequent classes, especially class 4.

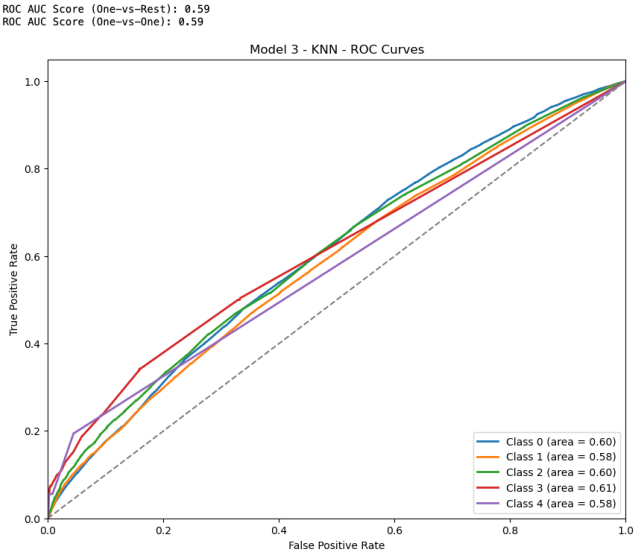

Figure 17: ROC curve for k-nearest neighbours model

The ROC AUC scores for both one-vs-rest and one-vs-one are 0.59, which are no better than the previous models.

D. Naive Bayes

The Naïve Bayes model performed the poorest out of all the models, achieving an accuracy of only 17%, indicating huge challenges in prediction across all classes.

Accuracy: 0.16887946311844512
Classification Report:

	precision	recall	f1-score	support
0	0.87	0.18	0.30	28214
1	0.12	0.14	0.13	3553
2	0.07	0.02	0.03	2364
3	0.01	0.24	0.02	254
4	0.00	0.58	0.00	36
accuracy			0.17	34421
macro avg	0.21	0.23	0.10	34421
weighted avg	0.73	0.17	0.26	34421

Figure 18: Classification report for naïve bayes model

These results suggest that the model is heavily biased towards the majority class and may not be suitable for this prediction model without considerable tuning and pre-processing. The ROC AUC scores of 0.53 indicate no real improvement over random guessing.

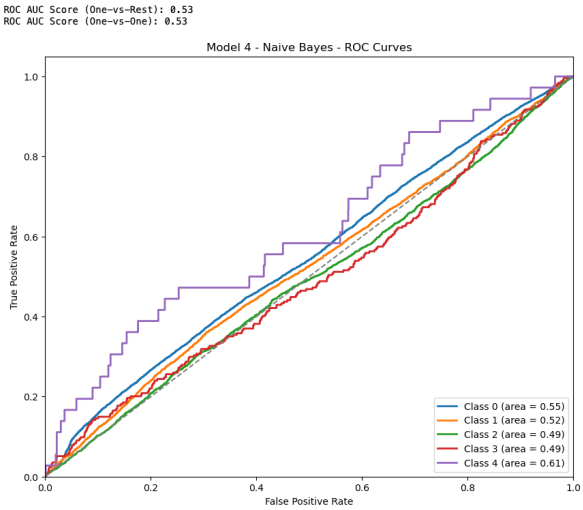

Figure 19: ROC curve for naïve bayes model

E. Support Vector Classifier

The Support Vector Classifier (SVC) model displayed an overall accuracy of 48.3%, indicating moderate performance. However, the results followed the same pattern as before with strong results for predicting class 0 (0.84) and reasonable recall (0.56), but poor results for the other classes.

Accuracy: 0.48299003515295896
Classification Report:

	precision	recall	f1-score	support
0	0.84	0.56	0.67	28214
1	0.13	0.22	0.16	3553
2	0.05	0.01	0.01	2364
3	0.01	0.13	0.02	254
4	0.00	0.33	0.00	36
accuracy			0.48	34421
macro avg	0.21	0.25	0.17	34421
weighted avg	0.71	0.48	0.57	34421

Figure 20: Classification report for support vector classifier model

The ROC AUC scores of 0.5 – 0.59 again suggest that the model has poor ability at differentiating between the classes.

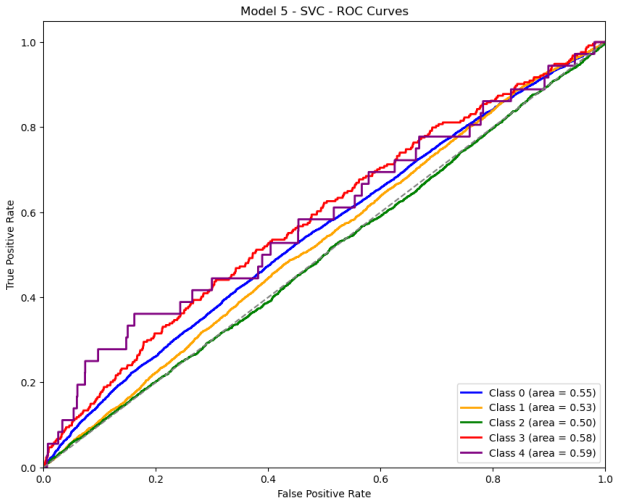

Figure 21: ROC curve for support vector classifier model

VI. ENHANCED FEATURE SELECTION

Given that all five models displayed consistently poor performances, it was clear that the approach was not capturing the complexities of the data. As such, recursive feature elimination was employed to identify the most significant features, which turned out to be vehicle age and distance.

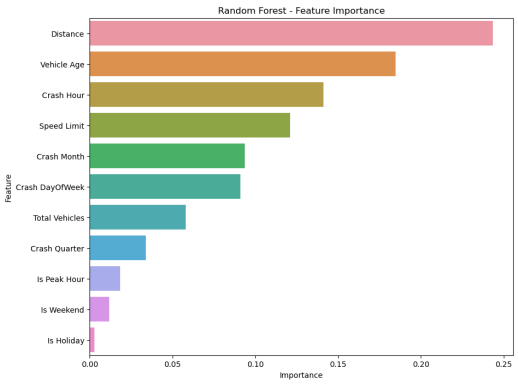

Figure 22: Bar plot of feature importance for random forest model

The two best performing models from the initial round of evaluation, Random Forest and K-Nearest Neighbours (KNN) were then re-evaluated using only these features, with the results displaying a considerable improvement in overall model accuracy for both Random Forest and KNN, scoring 0.78 and 0.81 respectively.

Accuracy: 0.78
Classification Report:

	precision	recall	f1-score	support
0	0.82	0.94	0.88	28214
1	0.11	0.04	0.06	3553
2	0.09	0.03	0.04	2364
3	0.01	0.00	0.01	254
4	0.00	0.00	0.00	36
accuracy			0.78	34421
macro avg	0.21	0.20	0.20	34421
weighted avg	0.69	0.78	0.73	34421

Figure 23: Second classification report for random forest model

Accuracy: 0.81
Classification Report:

	precision	recall	f1-score	support
0	0.82	0.98	0.89	28214
1	0.10	0.01	0.02	3553
2	0.06	0.00	0.01	2364
3	0.00	0.00	0.00	254
4	0.00	0.00	0.00	36
accuracy			0.81	34421
macro avg	0.20	0.20	0.18	34421
weighted avg	0.69	0.81	0.74	34421

Figure 24: Second classification report for k-nearest neighbours model

However, the performance for the minority classes did not improve, with low precision, recall and F1-scores indicating that all four classes are still poorly represented by the models.

VII. CONCLUSIONS

A. Conclusion

This research aimed to enhance the prediction of driver injury severity in road traffic collisions by evaluating five machine learning models: Random Forest, Gradient

Boosting Machines (XGBoost), K-Nearest Neighbours (KNN), Naïve Bayes, and Support Vector Classifier (SVC). Initially, the models demonstrated poor performance in the prediction of driver injuries with an overall accuracy of approximately 0.5 to 0.6, which was not much better than random guessing. The problem stemmed from the fact that the original dataset displayed considerable class imbalance in the target feature, injury severity. Despite employing corrective measures to the training dataset such as under sampling the major class and oversampling the minor classes, none of the models managed to significantly improve the prediction of these minority classes in the test set, which remained unbalanced.

In a final attempt to address this, recursive feature elimination (RFE) was employed, which identified vehicle age and distance as the most significant features for predicting injury severity. When the two best-performing models, Random Forest and KNN, were re-evaluated using only these features, a notable improvement in overall accuracy was observed, with results of 0.78 for Random Forest and 0.81 for KNN. Despite this improvement, the models continued to struggle with accurately predicting the minority classes, indicating that feature selection alone was unable to resolve the issue of class imbalance in the test set.

B. Future Work

As such, future research might focus on addressing the class imbalance inherent in this dataset. This might involve integrating additional datasets to enrich the current data source and improve the representation of minority classes or include using sophisticated algorithms such as Adaptive Synthetic Sampling (ADASYN), to generate synthetic samples to balance class distributions more effectively. Another consideration might be to employ cyclical feature encoding [17] to better represent the cyclical nature of time. Cyclical feature encoding would allow models to recognize that, for example, December (month 12) is as close to January (month 1) as it is to November (month 11). Initial trials suggest that cyclical feature encoding is a valuable pre-processing step that warrants further exploration.

REFERENCES

[1] ‘Road traffic injuries’. Accessed: Jul. 07, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] ‘Traffic Crash Severity: Comparing the Predictive Performance of Popular Statistical and Machine Learning Models Using the Glasgow Coma Scale | Journal of The Institution of Engineers (India): Series A’. Accessed: Jun. 29, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s40030-023-00710-3

[3] M. Bédard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, ‘The independent contribution of driver, crash, and vehicle characteristics to driver fatalities’, *Accid. Anal. Prev.*, vol. 34, no. 6, pp. 717–727, Nov. 2002, doi: 10.1016/S0001-4575(01)00072-0.

[4] S. Islam and F. Mannering, ‘Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence’, *J. Safety Res.*, vol. 37, no. 3, pp. 267–276, Jan. 2006, doi: 10.1016/j.jsr.2006.04.003.

[5] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, ‘The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives’, *Accid. Anal. Prev.*, vol. 43, no. 5, pp. 1666–1676, Sep. 2011, doi: 10.1016/j.aap.2011.03.025.

[6] W. Fu and J. Lee, ‘Relationship between Vehicle Safety Ratings and Drivers’ Injury Severity in the Context of Gender Disparity’, *Int. J. Environ. Res. Public. Health*, vol. 19, no. 10, p. 5885, May 2022, doi: 10.3390/ijerph19105885.

[7] H. Huang, H. C. Chin, and Md. M. Haque, ‘Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis’, *Accid. Anal. Prev.*, vol. 40, no. 1, pp. 45–54, Jan. 2008, doi: 10.1016/j.aap.2007.04.002.

[8] A. Boulieri, S. Liverani, K. Hoogh, and M. Blangiardo, ‘A Space–Time Multivariate Bayesian Model to Analyse Road Traffic Accidents by Severity’, *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 180, no. 1, pp. 119–139, Jan. 2017, doi: 10.1111/rssa.12178.

[9] J. de Oña, R. O. Mujalli, and F. J. Calvo, ‘Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks’, *Accid. Anal. Prev.*, vol. 43, no. 1, pp. 402–411, Jan. 2011, doi: 10.1016/j.aap.2010.09.010.

[10] A. Khorashadi, D. Niemeier, V. Shankar, and F. Mannering, ‘Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis’, *Accid. Anal. Prev.*, vol. 37, no. 5, pp. 910–921, Sep. 2005, doi: 10.1016/j.aap.2005.04.009.

[11] H. Abdelwahab and M. Abdel-Aty, ‘Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections’, *Transp. Res. Rec.*, vol. 1746, pp. 6–13, Jan. 2001, doi: 10.3141/1746-02.

[12] L.-Y. Chang and H.-W. Wang, ‘Analysis of traffic injury severity: An application of non-parametric classification tree techniques’, *Accid. Anal. Prev.*, vol. 38, no. 5, pp. 1019–

1027, Sep. 2006, doi: 10.1016/j.aap.2006.04.009.
[13] K. Santos, J. P. Dias, and C. Amado, ‘A literature review of machine learning algorithms for crash injury severity prediction’, *J. Safety Res.*, vol. 80, pp. 254–269, Feb. 2022, doi: 10.1016/j.jsr.2021.12.007.
[14] ‘The CRISP-DM Model: A Comprehensive Guide to the Main Stages of Data Analysis - Tech Hyme’. Accessed: Jul. 06, 2024. [Online]. Available: https://techhyme.com/the-crisp-dm-model-a-comprehensive-guide-to-the-main-stages-of-data-analysis/
[15] ‘Crisp DM methodology’, Smart Vision Europe. Accessed: Jul. 06, 2024. [Online]. Available: https://www.sv-europe.com/crisp-dm-methodology/
[16] ‘Crash Reporting - Drivers Data’. data.montgomerycountymd.gov, Jun. 28, 2024. Accessed: Jul. 06, 2024. [Online]. Available: https://catalog.data.gov/dataset/crash-reporting-drivers-data
[17] A. Kud, ‘Why We Need Encoding Cyclical Features’, Medium. Accessed: Jul. 19, 2024. [Online]. Available: https://medium.com/@axelazara6/why-we-need-encoding-cyclical-features-79ecc3531232