KAPLAN

PROJECT REPORT

DATA6000 CAPSTONE
INDUSTRY CASE STUDY

Prepared for :
Dr Indu Bala

Prepared by :
Bay Bayarsaikhan (1816560)

TABLE OF CONTENTS

1. Executive summary	2
2. Industry problem	2
3. Data processing and management	6
4. Data analytics methodology	7
5. Visualization and evaluation of results	7
6. Recommendations	11
7. Data ethics and security	12
References	13

2148 words (excluding contents and references)

1. Executive summary

According to report of Government agency, there are thousands of road crashes per year in Australia[1], leading to loss of lives, serious injuries and financial hardships. This report investigates traffic accident data in South Australia to identify key factors contributing to severe injuries and fatalities. Leveraging historical trends and seasonal patterns, I apply data analytics techniques to uncover actionable insights.

The findings aim to support improved road safety measures, targeted public awareness and informed policymaking. The main goal is to reduce traffic-related injuries and fatalities by providing actionable insights that guide strategic interventions of government agencies, insurers and transport planners.

2. Industry problem

Road transport industry is a crucial part of country's economy and people's daily life, but we face persistent challenges in ensuring road safety. There are thousands of accidents annually, resulting in serious injuries, fatalities and financial hardships.

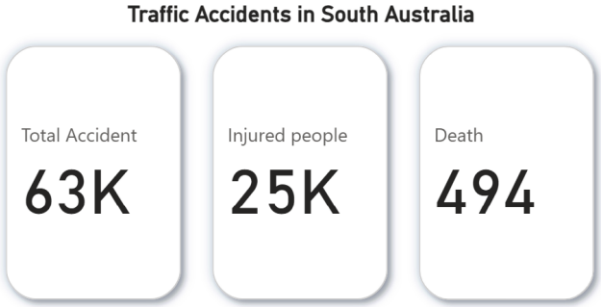

Figure 1: Total traffic accidents, injuries and fatalities in South Australia (2019–2023)

Serious road accidents remain high in numbers, despite government initiatives like awareness campaigns and improved infrastructure. A major gamechanger in the road safety industry could be the ability to effectively predict and prevent the severity of traffic accidents by leveraging historical crash data. This can lead to a reduction in severe injuries and loss of life, as well as lower insurance claims, healthcare costs, and pressure on emergency services. Therefore, I aim to find answers to this "*Which factors most significantly contribute to fatal or severe traffic injuries, and how accurately can accident severity be predicted using machine learning?*" question.

Descriptive analysis provides a comprehensive overview of traffic accident trends in South Australia from 2019 to 2023. Causes of traffic accidents are varied, including speeding, driver fatigue and experience, impaired driving and distractions such as using mobile phone. Additionally, road and weather conditions and high-risk locations contribute to accident frequency and severity.

[1] [1] AIHW, 2024, Injury in Australia: Transport Accidents, https://www.aihw.gov.au/reports/injury/transport-accidents

First graph (Figure 2) below shows total traffic accidents and injury counts monthly in 5 years (2019-2023). While accident totals remain relatively stable across years, injury counts fluctuate gradually. A notable drop during early 2020 (due to Covid-19 restrictions) is followed by a gradual rise.

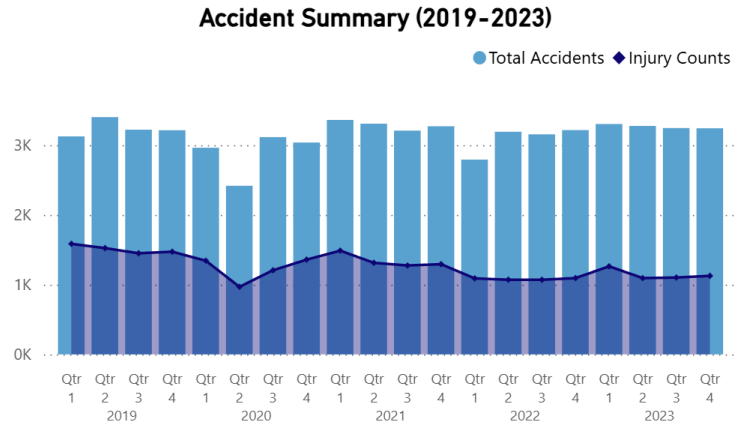

Figure 2: Monthly traffic accidents and total injuries in South Australia

The next graph (Figure 3) breaks down total injury counts into minor and severe injuries. Although minor injuries are more frequent, severe injuries are consistently present across the months, indicating a continued risk of serious trauma on roads.

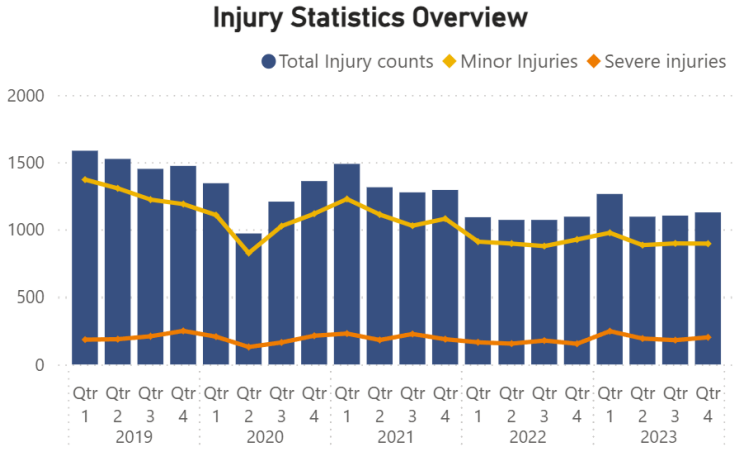

Figure 3: Monthly minor and severe injuries in South Australia

The line graph (Figure 4) shows traffic crashes in South Australia from 2019 to 2023 by region. Metropolitan areas had the most crashes, likely due to heavy traffic and congestion. Country areas had a moderate and steady number of crashes, while city areas had the fewest, possibly because of lower speed limits and better traffic control. Figure 5 illustrates crash occurrences by speed zone. Since most roads are in the 50–60 km/h range, crash numbers are relatively higher in these standard speed areas.

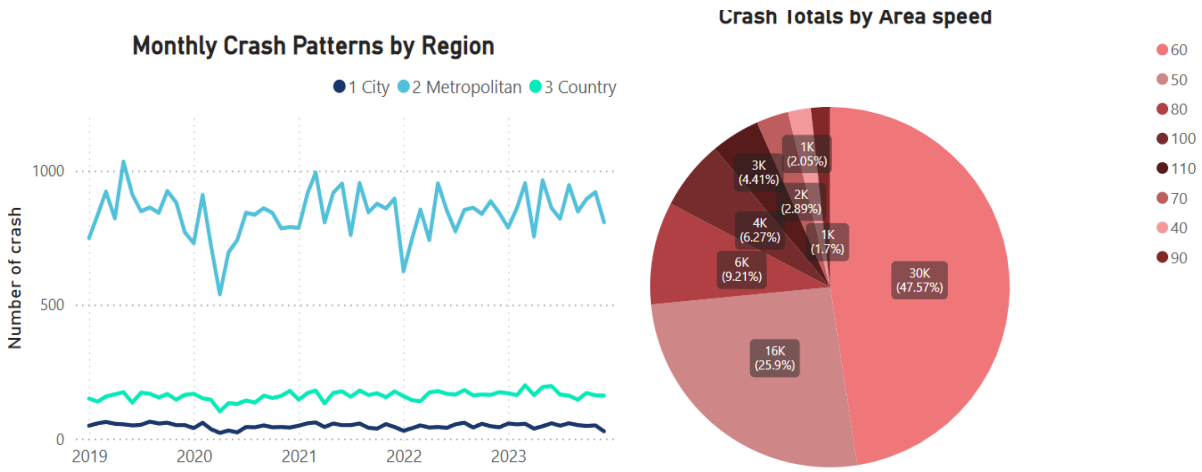

Figure 4: Monthly traffic accident pattern by region in South Australia (2019-2023)

Figure 5: Crash totals by area speed (2019-2023)

Figure 6 shows crash frequency by day of the week, with highest number of crashes occurring on Thursdays and Fridays. Figure 7 shows the total number of crashes by time of day, with highest number occurring during daytime hours (9 am–4pm), likely due to higher traffic volumes during business hours.

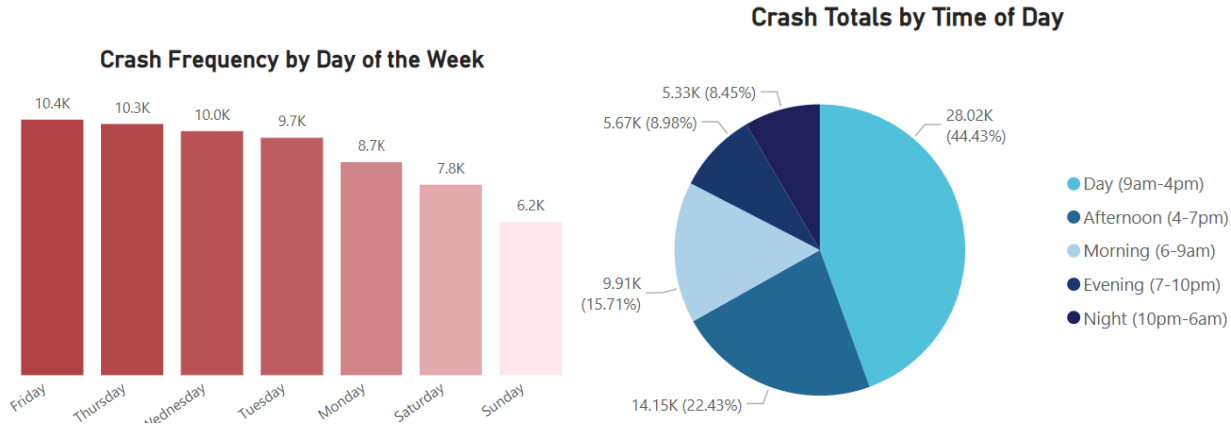

Figure 6: Total crash frequency by day of the week (2019-2023)

Figure 7: Crash totals by time of day (2019-2023)

This line graph (Figure 8) below compares road deaths across Australian states. NSW and Queensland accounted consistently high numbers. South Australia, shows moderate but steady road fatalities.

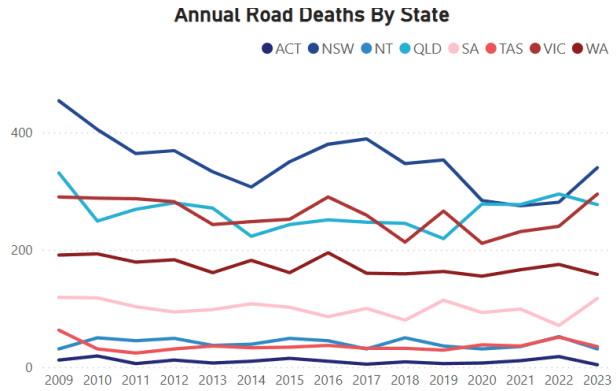

Figure 8: Annual Road death trend by State over last 15 years (2009-2023)

The bar chart (Figure 9) shows that male fatalities significantly outnumber female deaths each year. This suggests that males are more frequently involved in fatal accidents.

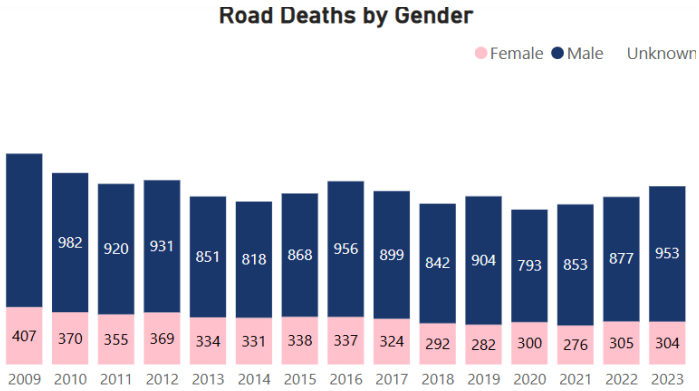

Figure 9: Road deaths by gender (2009-2023)

The pie chart (Figure 10) reveals that drivers accounted the highest proportion of road deaths with 47%, followed by passengers with 18.85% and motorcycle riders with 17.16%. Pedestrians and cyclists also make noticeable portions.

The donut chart (Figure 11) shows the total number of road deaths by age group. The 40–64 age group accounted for the highest proportion at 31%, followed by the 17–25 and 26–39 age groups, each with 20%.

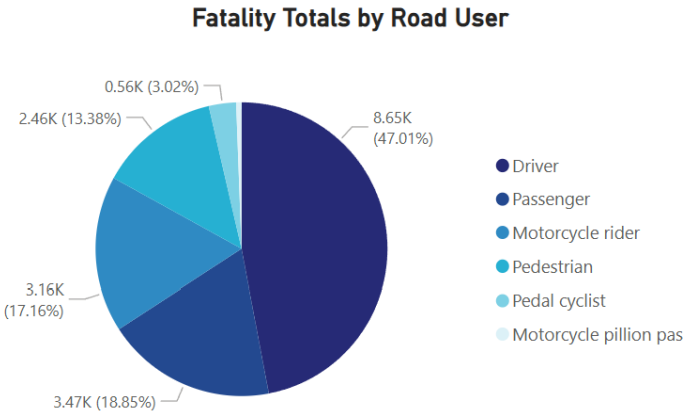

Figure 10: Road deaths by Road user (2009-2023)

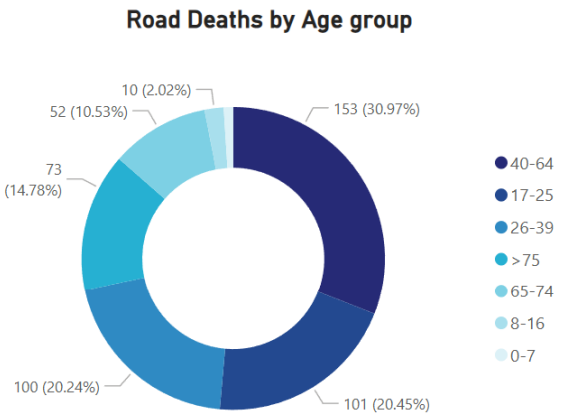

Figure 11: Road deaths by Age group (2009-2023)

Australian government set an ambitious goal for 2050[2], based on the belief that no one should die or be seriously injured on the roads. To achieve this goal, data analysis plays a role by offering insights into trends, high-risk areas and possible preventive actions. By leveraging data analytics, government agencies can develop more effective strategies to enhance road safety and reduce fatalities and severe injuries from traffic accidents.

3. Data processing and management

The road safety issue was explored using detailed crash data from trusted government sources like SA Police and Bureau of Infrastructure and Transport Research Economics. The data includes key information such as crash type, time, road conditions, weather and speed limits. This rich data helped in understanding what causes serious accidents more clearly. The dataset I used covers daily traffic accidents in South Australia from 2019 to 2023. Since it comes from reliable sources, it allowed machine learning models to be trained with real-world information, making the insights and safety recommendations more useful and accurate.

Originally, the dataset contained 34 variables and 63,069 records with no missing values, each records represent a road crash. To improve model accuracy, I reduced dataset to 20,381 observations with 14 relevant features, excluding cases with no recorded injuries. I also removed features such as suburb, postcode, location, LGA name, and report ID, which were not meaningful for prediction purposes. Following data cleaning, I further categorized some variables such as grouping crash times into five periods (morning, day, afternoon, evening and night) and grouping speed zones and collision types into manageable categories.
Please visit to this link for the dataset: https://github.com/BayAus/Data6000-Capstone-project

[2] Think Road Safety, n/a, Action plan 2023-2025, https://www.thinkroadsafety.sa.gov.au/road_safety_strategy/road-safety-action-plan

Data sources: *https://data.sa.gov.au/data/dataset/road-crash-data*

https://www.aihw.gov.au/reports/injury/transport-accidents

4. Data analytics methodology

I started with applying Random Forest algorithm, to predict crash severity (minor, severe and fatality). But it resulted in poor accuracy because my dataset was heavily imbalanced, meaning the majority of traffic accidents related to the "minor injury" class. This class dominance made it difficult for the models to learn from minority classes such as severe or fatal accidents.

To address this issue, I applied the Smote technique with up-sampling method that synthetically increased number of severe injury and fatality records, balancing them with the minority class. This allowed the model to better learn from the underrepresented classes.

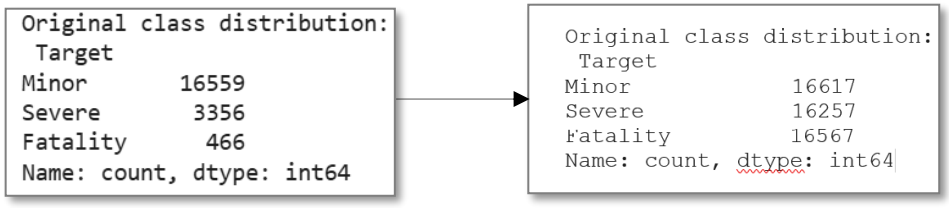

Figure 12: How Smote with up-sampling synthetically increased minority class

After applying class balancing, both models showed improved recall for the minority classes. Random Forest algorithm is a powerful and interpretable machine learning model. My dataset includes both categorical (road type, time of day, weather) and numerical (speed limit, vehicle count) variables, which Random Forest handles well. It is ideal for capturing complex, non-linear crash patterns, like how road type, time and area speed interact. It also provides feature importance scores, making it easier to identify key risk factors, a useful tool for public safety planning. Its ensemble structure helps prevent overfitting, leading to strong, reliable predictions.

5. Visualization and evaluation of results

I used ARIMA forecasting on five years of crash data (2019–2023) to predict future trends in South Australia. While the model showed moderate accuracy, its lower R-squared and MAE indicate limited precision. Still, it revealed seasonal patterns that can help policymakers plan and allocate resources more effectively.

RMSE▲	MAE	MAPE (Ratio)	R Squared	AIC
95.9183997236	71.3335199622	0.071151139078	0.14316832042	728.3154835101

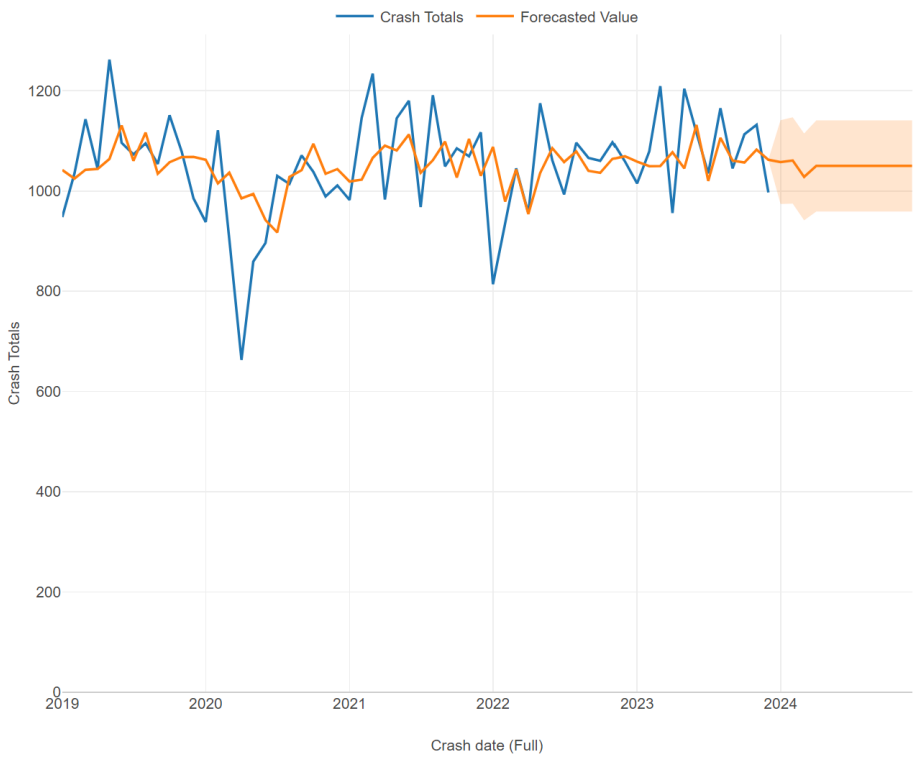

Figure 13: Trend of traffic accident in the next 12 months by ARIMA model

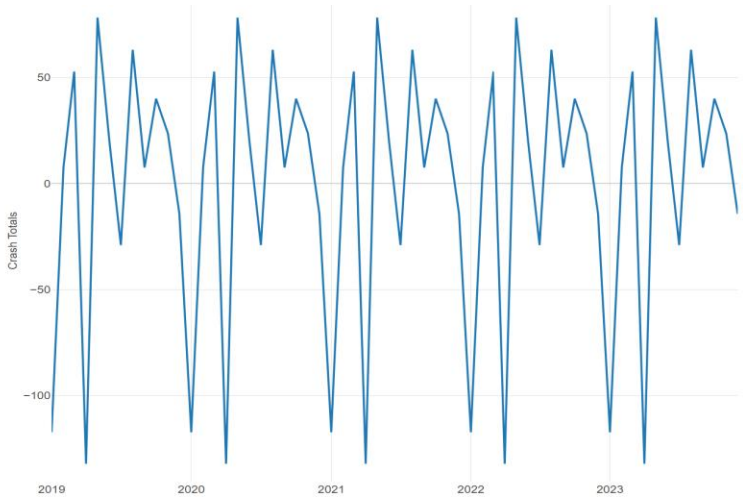

Figure 14: Seasonality of traffic accident by ARIMA model

Then I used SHAP values (Figure 15) with the Random Forest model to clearly show which factors have the biggest impact on crash severity. Key features like crash type, speed zone, time of day, day of the week, and the number of vehicles involved were identified as the most influential. These insights can help transport authorities focus on safety improvements—such as adding better lighting in high-risk areas or running awareness campaigns for certain driver groups.

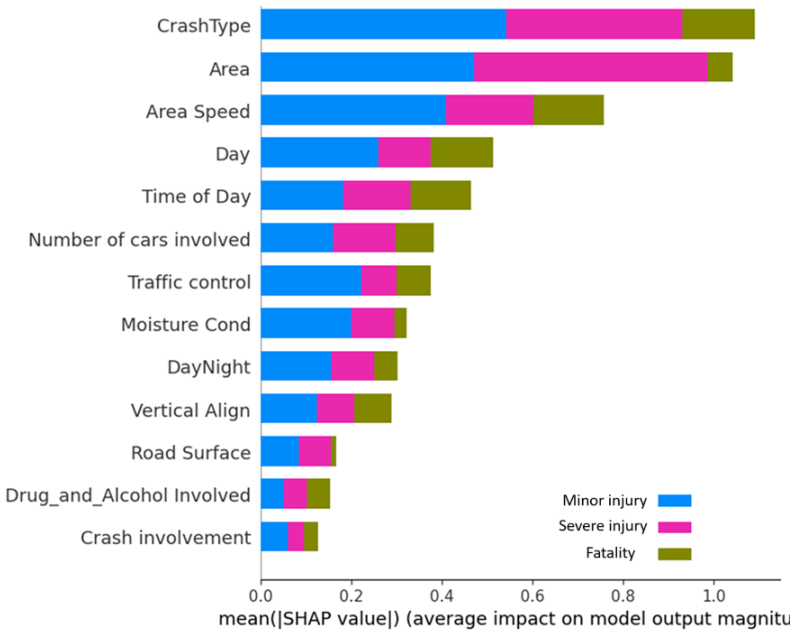

Figure 15: SHAP model shows most influenceable factors

To predict crash severity in road accidents, I implemented a Random Forest classifier. This model performed multi-class classification, categorizing crashes into three classes such as Minor Injury, Severe Injury, and Fatality. This model helps identify high-risk situations and estimate the chance of serious or fatal outcomes.

Below is a flow chart of Random Forest model:

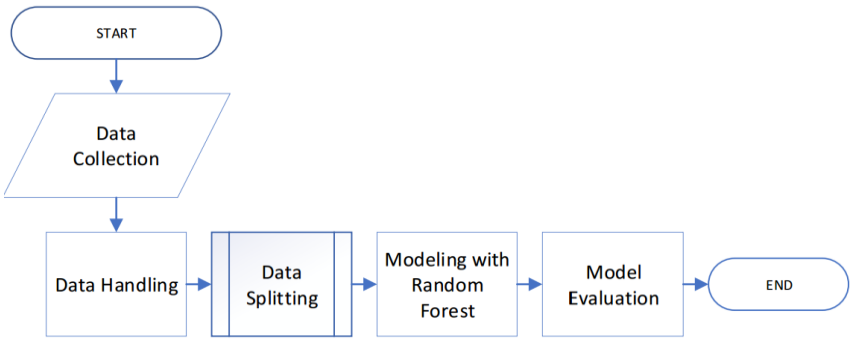

Figure 16: Flow chart of Random Forest model

The model achieved an overall accuracy of 78% and performed best on Minor injury class with a high recall (93%) and F1-score (88%), while performance slightly dropped on other classes, severe injury with recall (72%), F1-score of 74% and fatality recall 70%, F1-score of 71%. Below is classification report:

Classification Report:

	precision	recall	f1-score	support
Minor injury	0.84	0.93	0.88	4970
Severe injury	0.76	0.72	0.74	4985
Fatality	0.73	0.70	0.71	4949
accuracy			0.78	14904
macro avg	0.78	0.78	0.78	14904
weighted avg	0.78	0.78	0.78	14904

Figure 17: Classification report of Random Forest

The model predicted minor injury class (0) most accurately, with 4,623 true positives. Only 130 cases were misclassified as severe injuries and 217 as fatalities. However, the model had some difficulty distinguishing between severe injuries (1) and fatalities (2). For example, 972 fatal cases were predicted as severe injuries, while 1,088 severe injuries were predicted as fatalities.

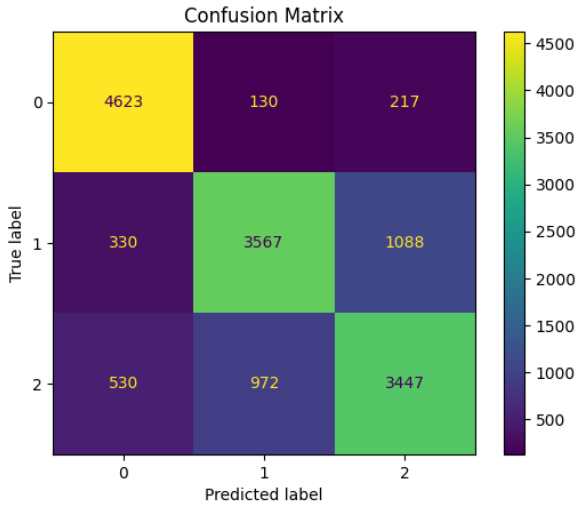

Figure 18: Confusion matrix of Random Forest

To improve model training and accuracy, the following strategies are recommended:

- Use a larger dataset and analyze feature importance to focus on key predictors.
- Eliminate irrelevant or redundant features to reduce noise.
- Clean or cap extreme values, such as speeds well beyond legal limits.
- Replace Label encoder with One-hot encoding for categorical variables with low cardinality, as it may improve performance.

In conclusion, combining ARIMA, SHAP and Random Forest models shows data can be used to extract actionable insights for road safety in South Australia by predicting crash trends, identifying key risk factors, and supporting proactive decision-making to reduce accidents and save lives.

6. Recommendations

The analysis of South Australian traffic accident data provided valuable insights for improving road safety. Descriptive analytics identified patterns such as weekday crashes during business hours, higher accident rates in metropolitan areas and risks related to certain age groups, road types and speed zones.

The ARIMA model was moderately accurate, revealed seasonal crash trends that is useful for planning emergency services and safety campaigns. SHAP analysis highlighted key factors influencing severe injuries and fatalities like crash type, speed zones and time of day and made the findings easy to understand through visuals. The Random Forest model predicted accident severity, helping to identify high-risk cases and support quicker, better-informed responses by emergency services.

Overall, data analytics plays a vital role in moving from reactive to proactive road safety strategies. Based on the outcomes of analysis, I suggest following recommendations:

- Leveraging predictive analytics:
  - Develop incident prediction dashboard for ambulance or emergency service,
  - Use time-series forecasting like ARIMA to identify seasonal crash peaks and provide targeted campaigns in particular months where data shows crash increases,
  - Use predictive models to test “what if” situations, like reducing speed limits or adding pedestrian crossings, to see how those changes might reduce accidents,
  - Random forest can predict how serious crash might be based on time, location and road conditions. Share these insights with emergency services and traffic control centers to help them prepare during high-risk times.

- Targeting high-risk periods
  - Increase police patrol and speed monitoring during high-risk periods,
  - Place road safety awareness campaigns tailored to peak hours,
  - Encourage use of public transports by reducing fee during peak hours.
- Improve safety in Metropolitan areas where the most accidents occur
  - Invest in better traffic flow management like smart traffic lights and roundabouts,
  - Reduce public transport fee within metropolitan area.

7. Data ethics and security

The dataset I used is publicly available through official Australian government agency websites, ensuring it is legally accessible and approved for public use. Importantly, the data does not include any personally identifiable information, which minimizes ethical or privacy concerns. However, there are still several considerations and potential risks to keep in mind outlined below:

Data should be used to benefit society, for improving road safety. Avoid misrepresenting insights or drawing conclusions that could unfairly blame specific groups, for example, based on age or region without evidence. Also, data usage must comply with data use agreements and relevant laws, ensuring the dataset is used only for approved purposes like research and public safety. Finally, datasets should be stored securely, especially when accessed via platforms like Google-colab, ensuring files are not shared with unauthorized users and are properly deleted when no longer needed.

Reference

- *Australian Institute of Health and Welfare, 2024, Injury in Australia: Transport Accidents, https://www.aihw.gov.au/reports/injury/transport-accidents*
- *DataSA, 2024, Road Crash Locations in SA, https://data.sa.gov.au/data/dataset/road-crash-data*
- *Think Road Safety, n/a, Action plan 2023-2025, https://www.thinkroadsafety.sa.gov.au/road_safety_strategy/road-safety-action-plan*
- *Bitre, n/a, Australian Road Deaths Database, https://www.bitre.gov.au/statistics/safety/fatal_road_crash_database*
- *Office of Road Safety, n/a, National Road Safety Data Hub, https://www.officeofroadsafety.gov.au/data-hub?_gl=1*lbe452*_ga*MTM1MjI5MjcyOC4xNzQzMjQ4NTM2*_ga_XV4JMWELH5*MTc0NDYzMjQyOS4xLjEuMTc0NDYzMjUzMy4wLjAuMA..*
- *Bitre, 2022, Road Trauma Australia 2022 Statistical Summary, https://www.bitre.gov.au/sites/default/files/documents/road_trauma_2022.pdf*
- Bitre, https://www.bitre.gov.au/forecasts
- *Abdulaziz H. Alshehri, Fayez Alanazi, PLOS.one, 2024, Comparing fatal crash risk factors by age and crash type by machine learning techniques, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0302171*
- *TingTing Huang, Shuo Wang, Anuj Sharma, 2020, Highway crash detection and risk estimation using deep learning, https://www.sciencedirect.com/science/article/abs/pii/S000145751930555X*
- *Road safety, 2024, Annual Trauma, https://datahub.roadsafety.gov.au/progress-reporting/annual-trauma*
- *Road safety, 2025, Monthly road deaths, https://datahub.roadsafety.gov.au/progress-reporting/monthly-road-deaths*
- *Adam Shafi (2024), Random forest classification with Scikit learn, https://www.datacamp.com/tutorial/random-forests-classifier-python*
- *Abdulaziz H. Alshehri, Fayez Alanazi (2024), Comparing fatal crash risks factors by age and crash type by using machine learning techniques, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0302171*
- *Son Truong Nguyen (2022), Road Traffic Severity Prediction, https://github.com/sonnguyen129/Accident-Severity-Prediction/blob/main/README.md*
- *Emmett Zhao (2020), Machine Learning: Predicting Car Crash Severity for South Australia, https://www.linkedin.com/pulse/machine-learning-predicting-car-crash-severity-south-australia-zhao/*

- *ProudJiao (2023), Car-Accident-Severity-Prediction-Using-Random-Forest-and-XGBoost, https://github.com/proudjiao/Car-Accident-Severity-Prediction-Using-Random-Forest-and-XGBoost/blob/main/README.md*
- *SHAP (2018), An introduction to explainable AI with Shapley values, https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html*
- *A Data Odessey (2023), SHAP with Python (Code and explanations), https://www.youtube.com/watch?v=L8_sVRhBDLU*
- *Data professor (2022), How to handle imbalanced datasets in Python, https://www.youtube.com/watch?v=4SivdTLIwHc*