Predicting traffic Accidents in South Australia

DATA6000 Capstone industry case

Assessment-2

Presented by: Bay (1816560)

Agenda

- Introduction
- About data
- Descriptive analytics
- ARIMA forecasting
- SHAP model
- Problem solved
- Random Forest – Crash severity prediction
- Conclusion
- Reference

Introduction

Traffic accidents remained significant concern, affecting public safety, economy and society.

This project explores how machine learning can be used to better understand and predict road accident severity.

Total Accident	Injured people	Death
63K	25K	494

2019-2023

“Thousands of Australians are severely injured in traffic accidents, 10% of all injury deaths”

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

About data

Data description

- South Australian accident data between 2019-2023
- Each observation represents a car accident recorded in SA

Data cleaning

- Times categorized (Morning, Day, Evening…)
- Crash type grouped into 5
- 20 unnecessary features dropped (Suburb, Post code, ACCLOC_X, ACCLOC_Y…)

Data Set

- Main dataset 63’069 observations
- 34 variables (Speed area, weather condition…)
- Target: Accident severity

Dataset reduced to 20’381 observations, 14 variables for ML model (no injury recorded data removed)

Data dictionary

Feature	Description	Data type	No of Categories	Range
Target	Indicates if the crash was fatal, minor injury or severe injury	object	3	Minor, Severe, Fatal
Area	Location of the crash	object	3	Metropolitan, Country,
Number of Cars Involved	Number of Cars Involved		11	1-15
Day	The day of the week when the crash occurred	object	7	Monday - Sunday
Time of Day	The time of the week when the crash occurred	object	5	Morning, Day, Afternoon, Evening, Night
Area Speed	Speed limit of the area where crash occurred	object	6	40km/h or under – 90km/h or above
Vertical Align	Road level	object	5	Bottom Of Hill, Crest of Hill, etc.,
Road Surface	Surface of the road	object	3	Sealed, Unsealed, Unknown
Moisture Condition	Condition of a road when crash occurred	object	3	Wet, Dry, Unknown
Day Night	Indicates if crash occurred in day or night	object	2	Day, Night
Crash Type	Type of crash (e.g., Collision, Rollover, etc.)	object	5	Collision with Person/Animal, Collision with Stationary object etc.,
Crash Involvement	Indicates who involved in crash	Int64	5	Driver, Rider, Passenger, etc.
Traffic Control	Indicates if there is a traffic control	object	6	Give Way Sign, Roundabout, etc.,
Drug and Alcohol involved	Drug and Alcohol involvement	object	2	Yes, No

Descriptive analytics

Figure 1: Monthly traffic accidents and total injuries in South Australia

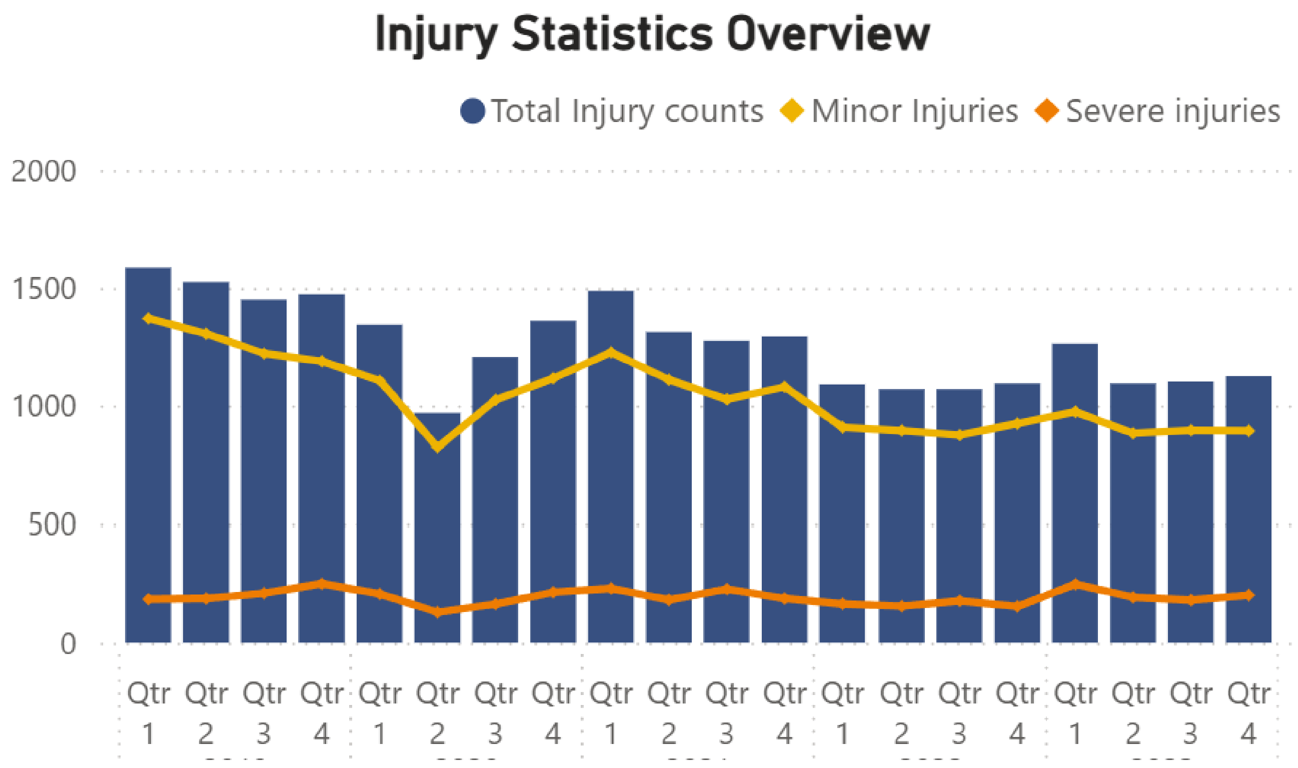


Figure 2: Monthly minor and severe injuries in South Australia

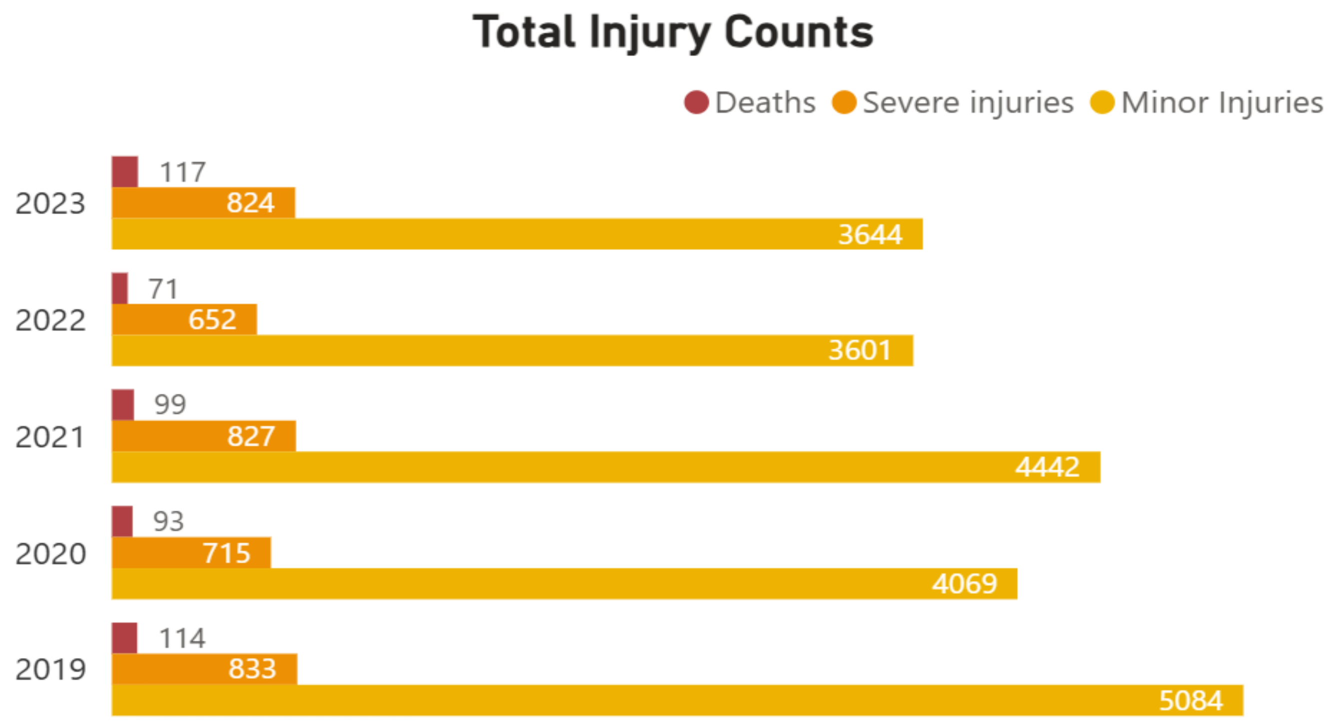


Figure 3: Annual fatalities, minor injuries, and severe injuries in South Australia

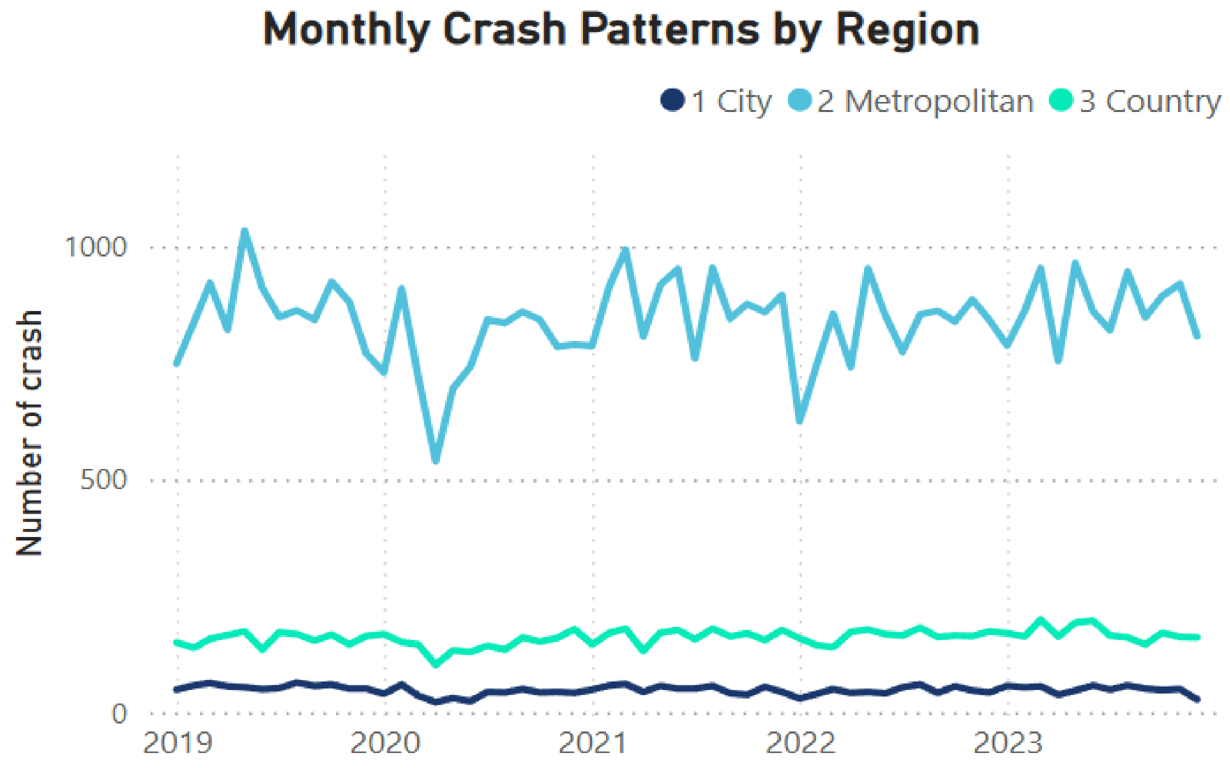


Figure 4: Monthly traffic accident pattern by region in South Australia (2019-2023)

Descriptive analytics

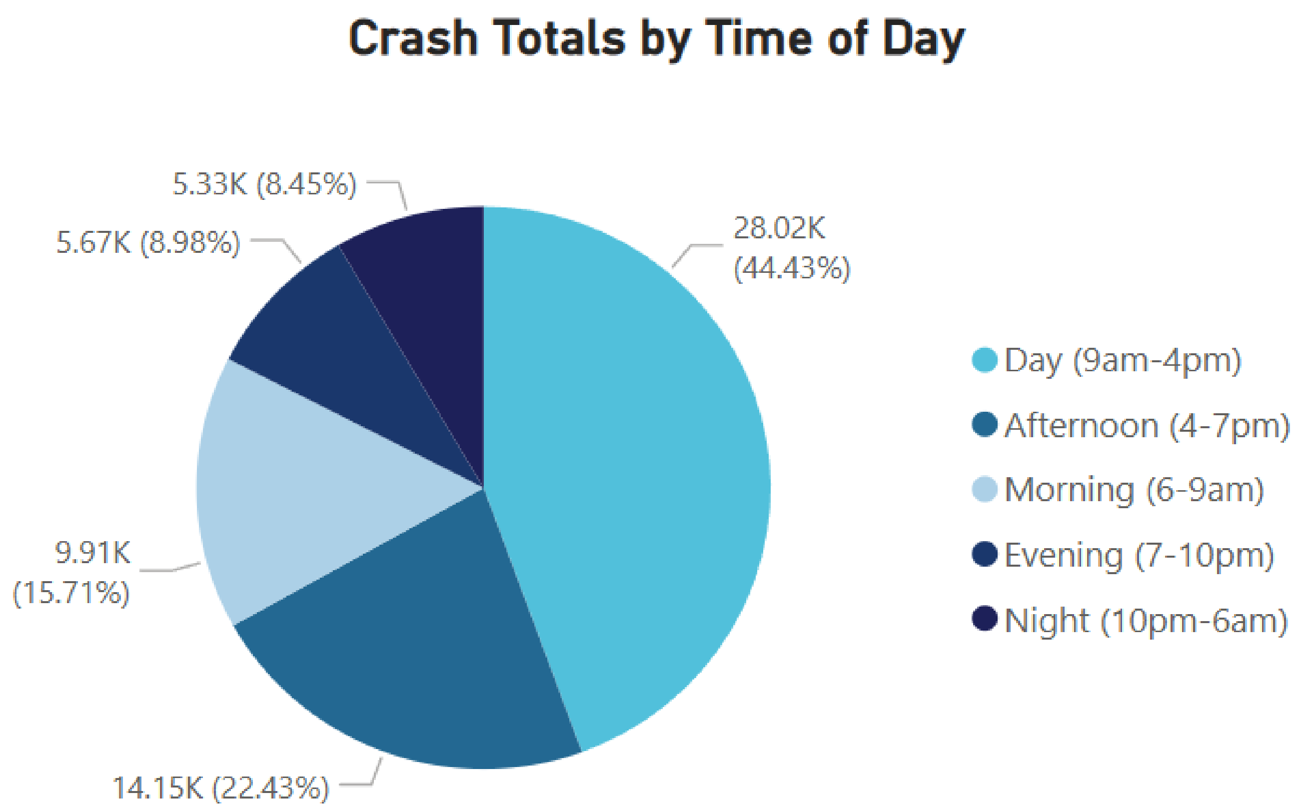


Figure 5: Crash totals by time of day

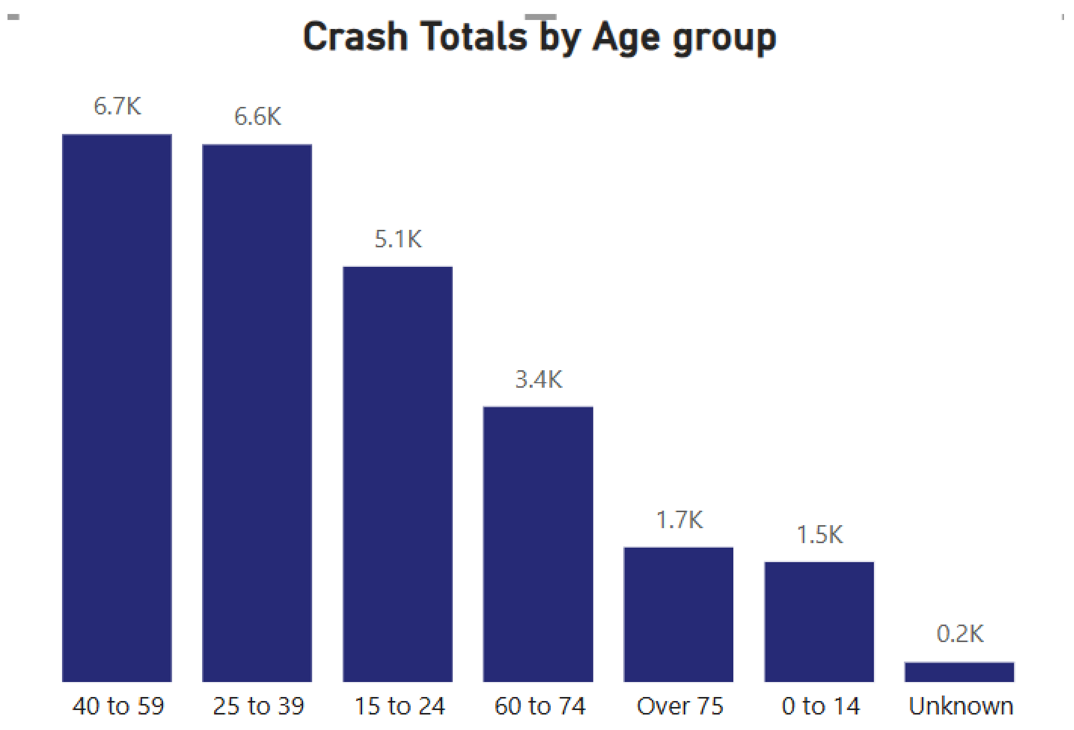


Figure 6: Crash totals by age group

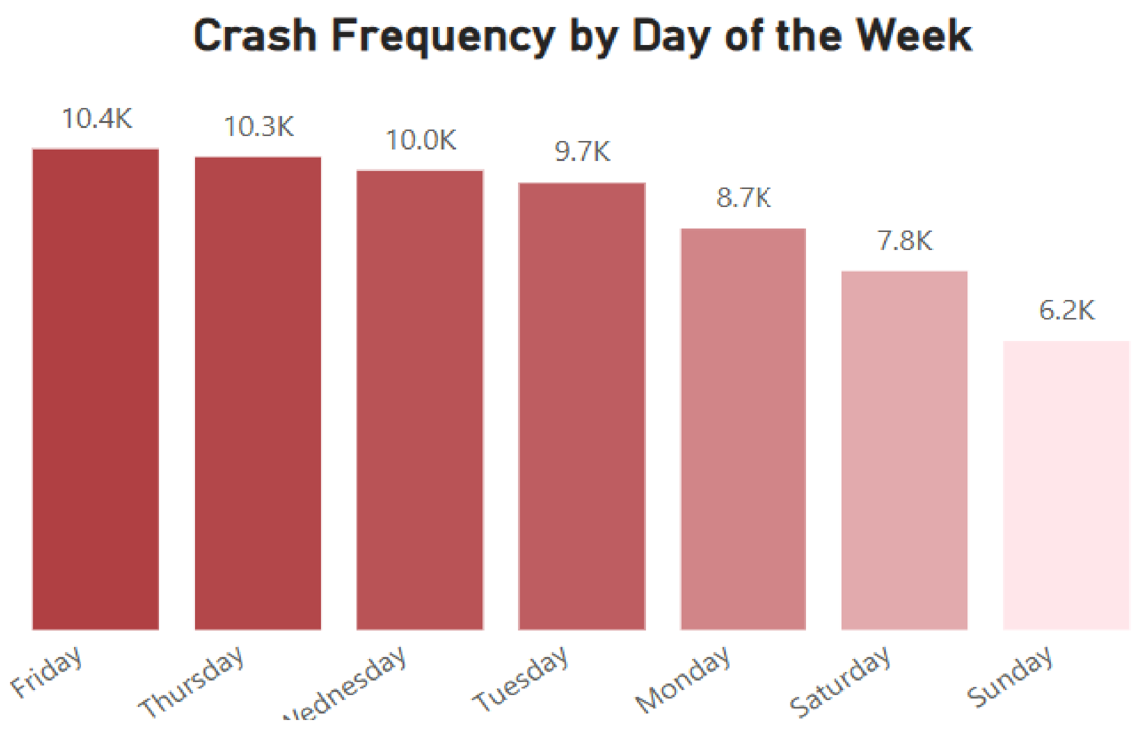


Figure 7: Total crash frequency by day of the week (2019-2023)

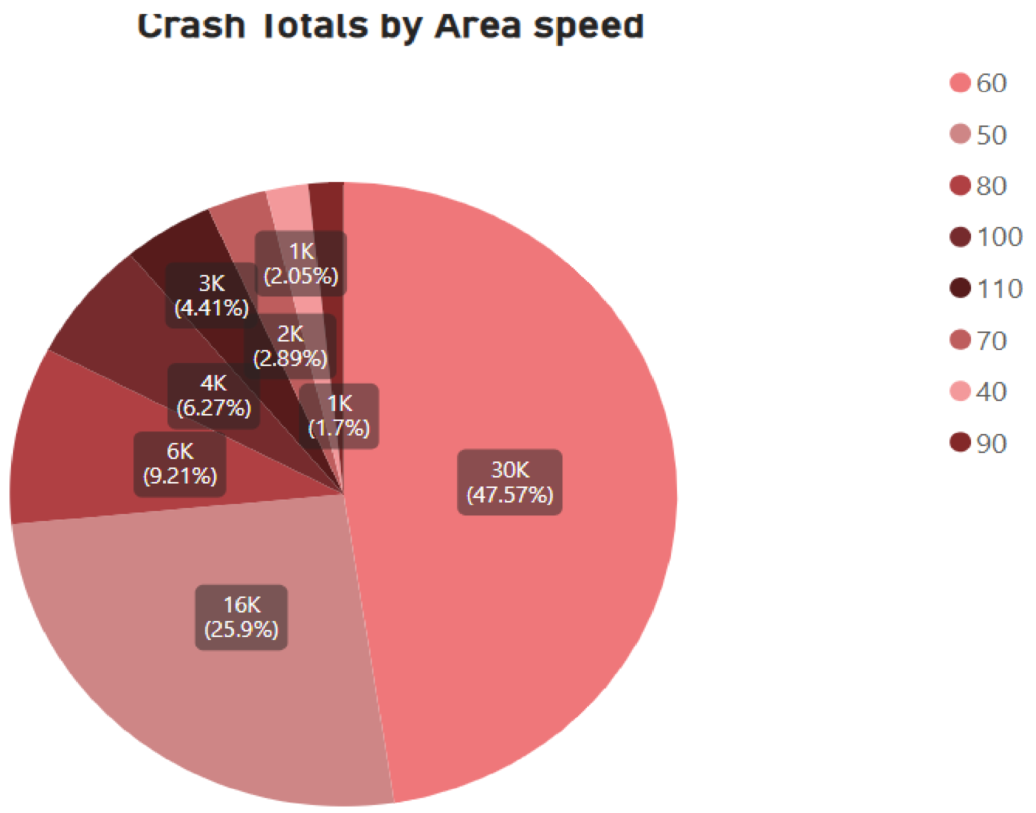


Figure 8: Crash totals by area speed (2019-2023)

Descriptive analytics (Fatality)

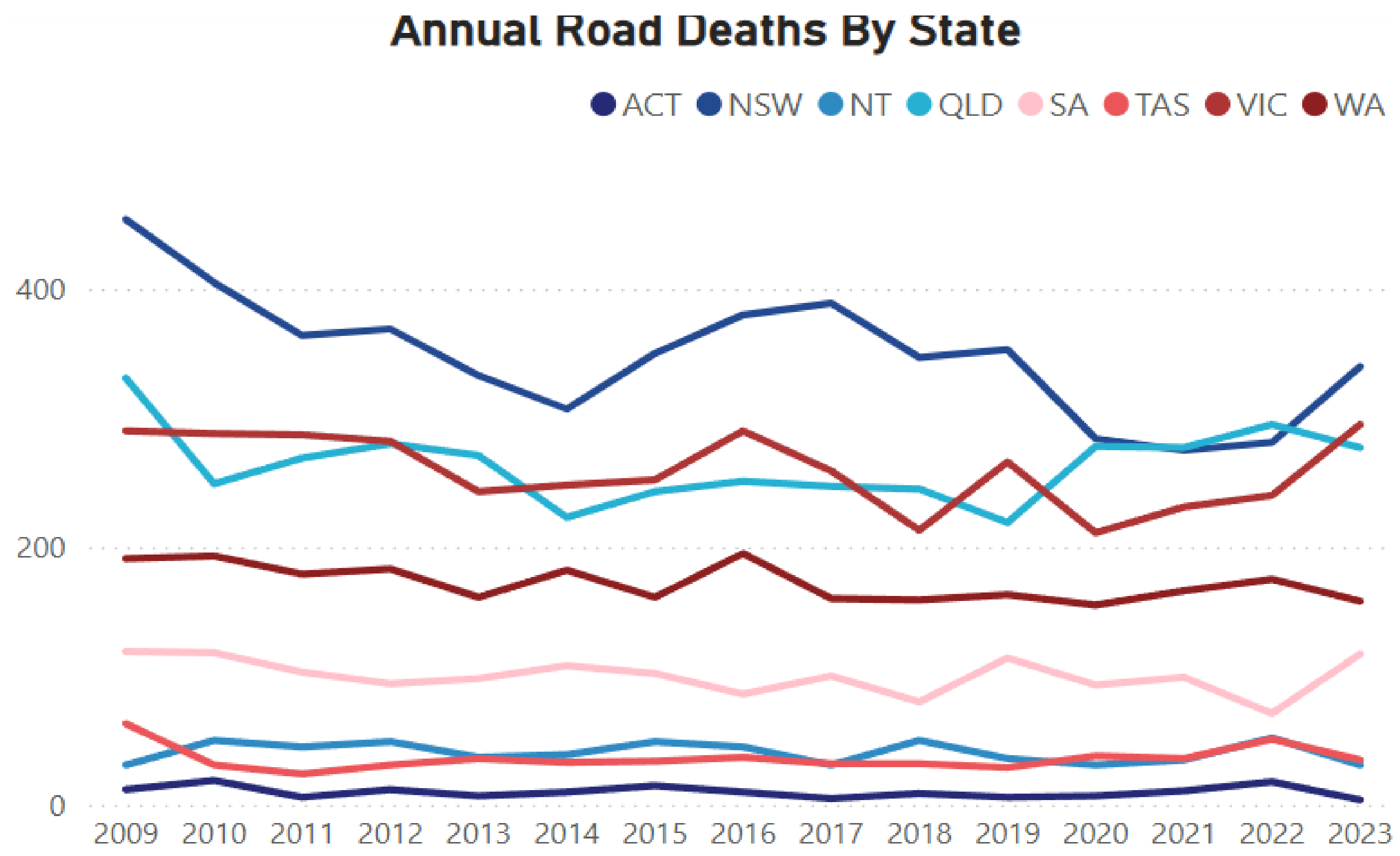


Figure 9: Annual Road death trend by State over last 15 years

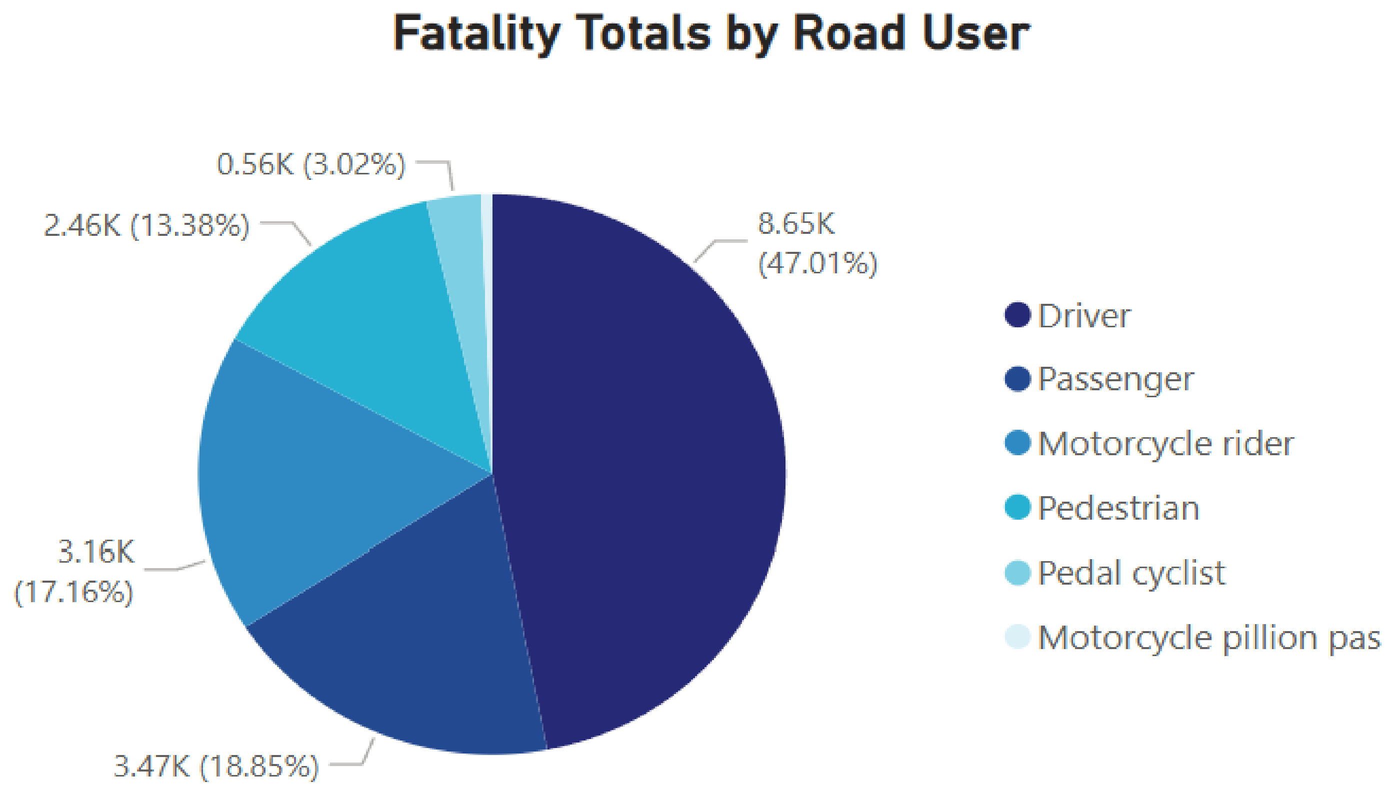


Figure 10: Road deaths by Road user

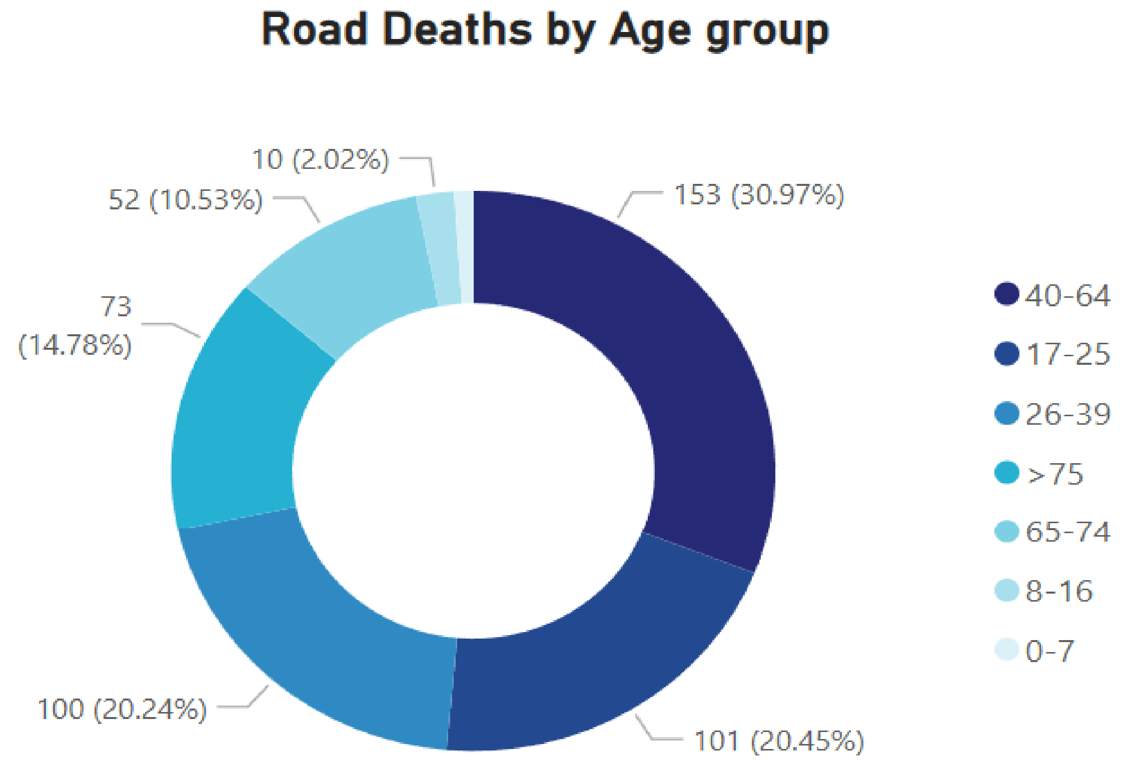


Figure 11: Road deaths by Age group

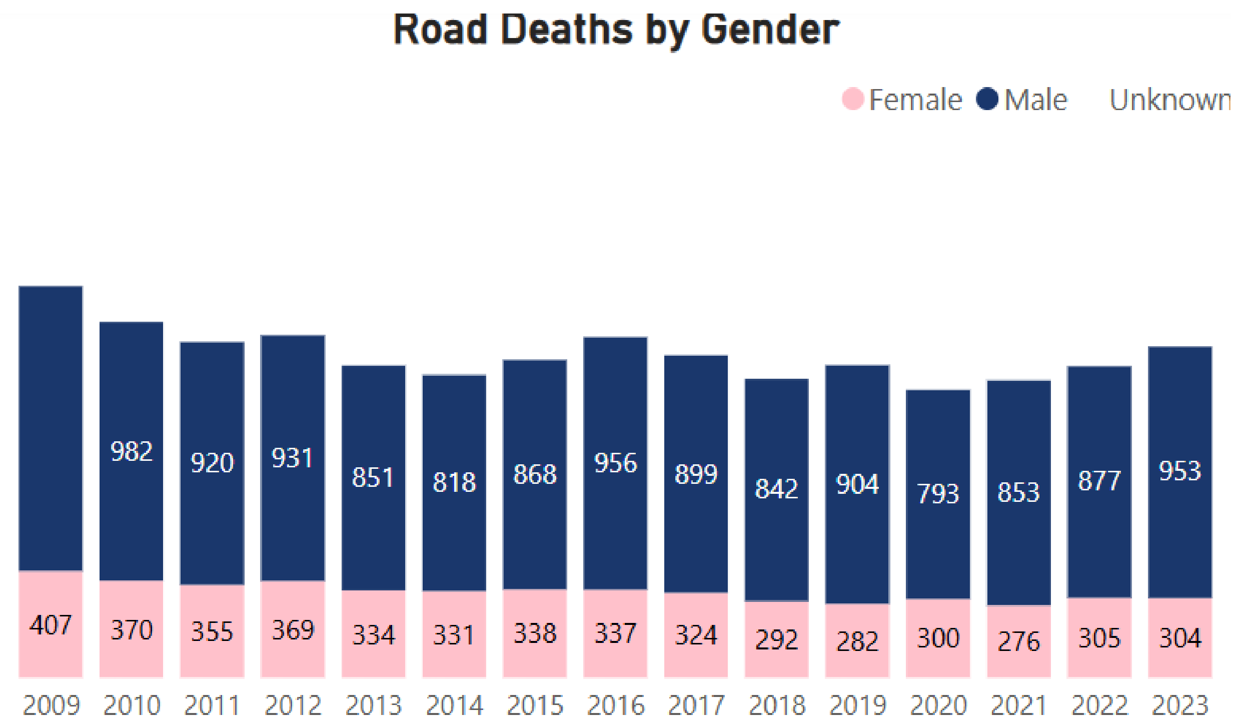
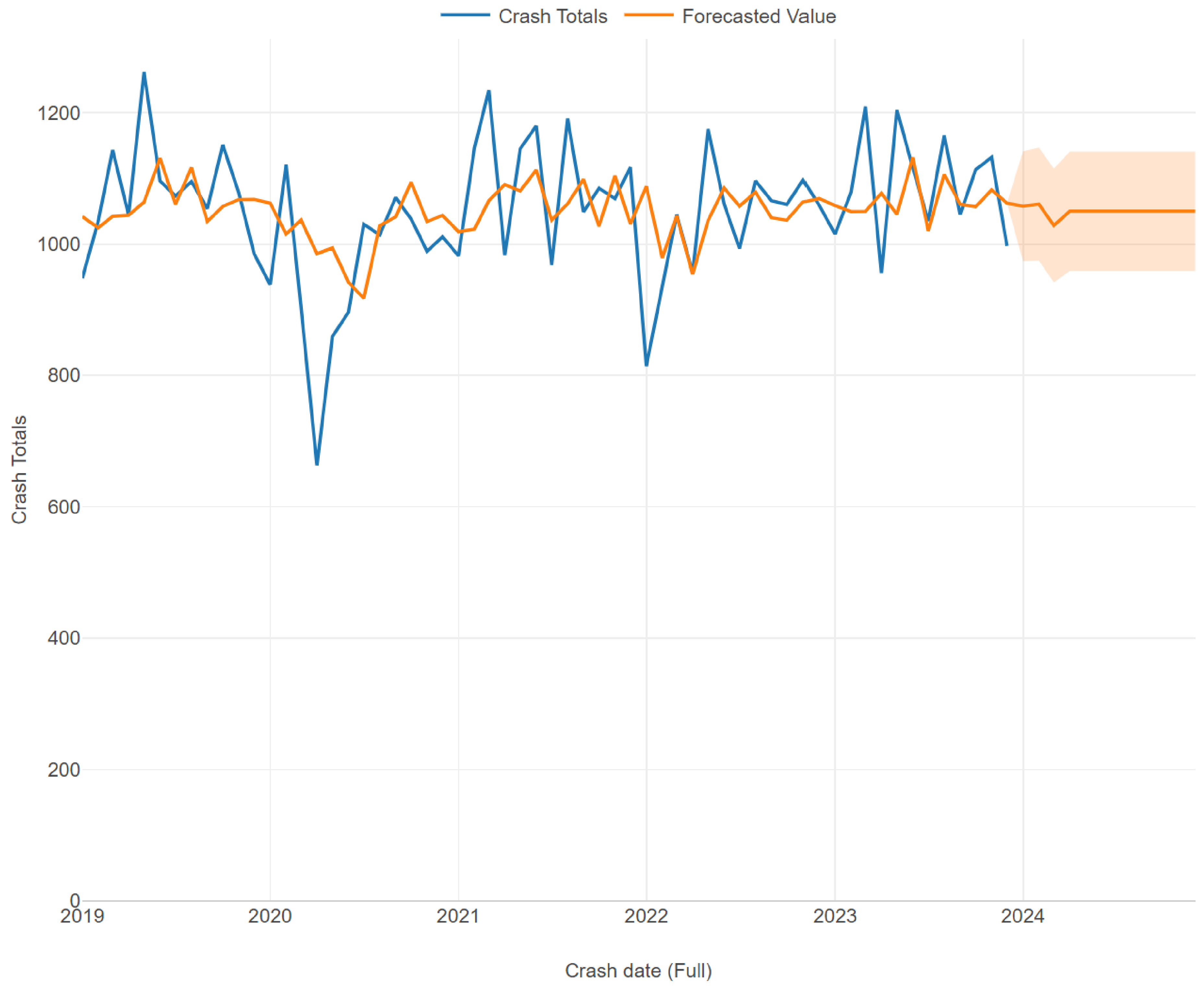


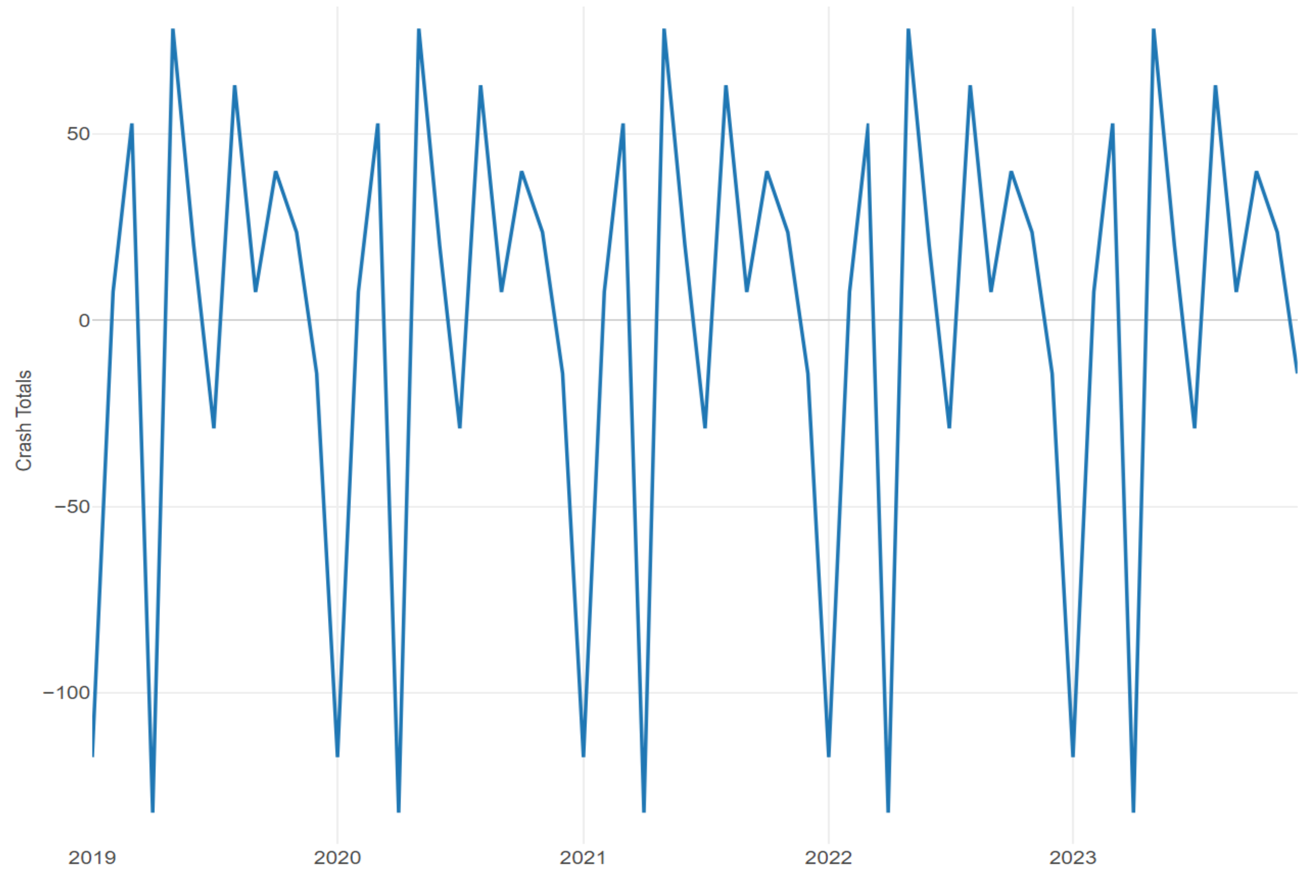
Figure 12: Road deaths by gender

ARIMA Forecasting

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

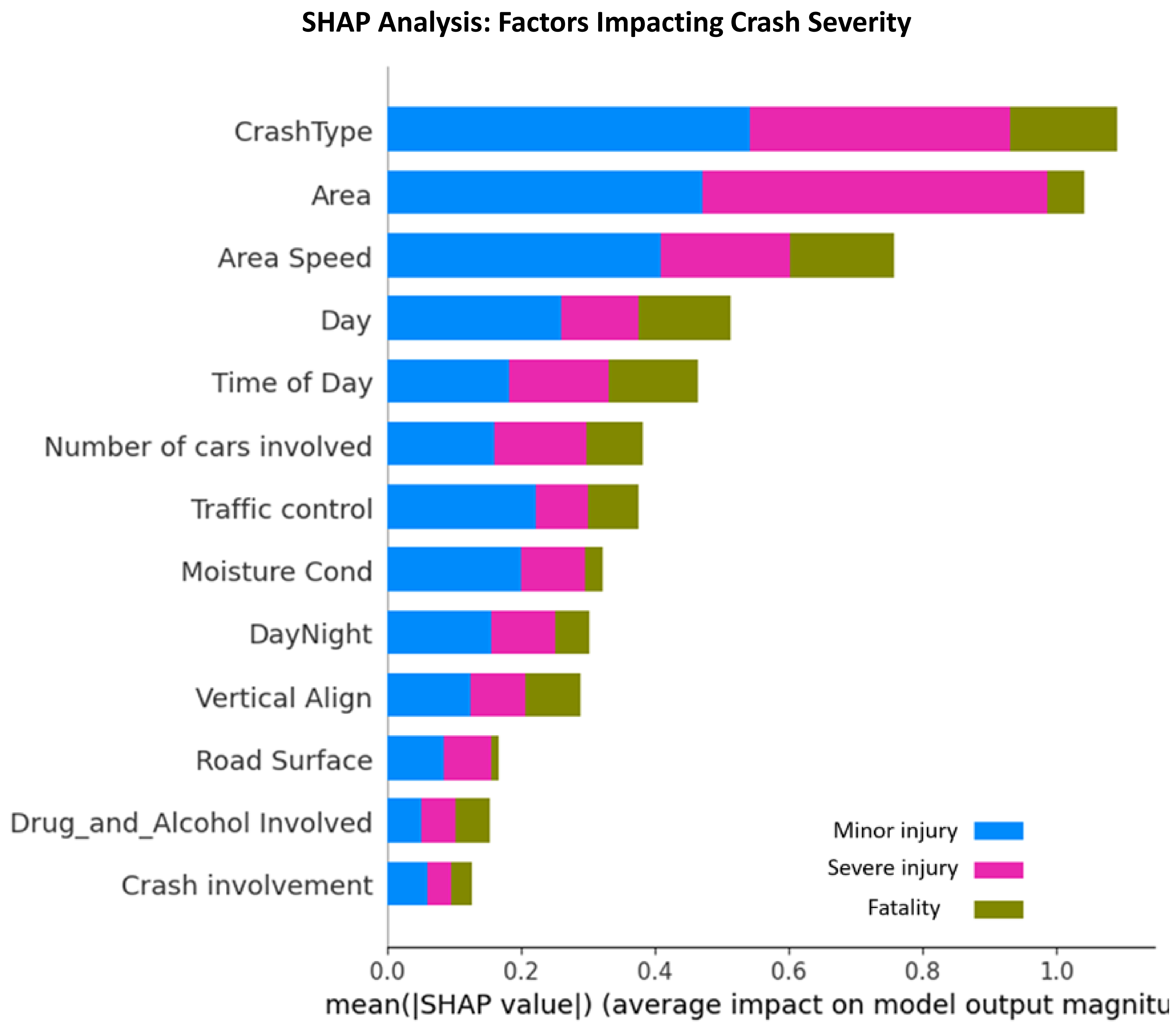


Seasonality



- The **forecasted values** stand stable around 1060, 70 per month,
- **Seasonal pattern** suggest that traffic accidents tend to increase during certain months (possibly during winter seasons).
- The **model evaluation** (such as MAE, RMSE, MAPE, and R^2) showing a acceptable **fit**.

SHAP model



- SHAP model can be used to explain feature importance,
- It tells us which factors most influence the crash severity,
- Factors like crash type, road user type, and age group stand out as most impactful.

Problem solved before ML algorithm

Solution: Class balancing

Using up-sampling techniques to handle imbalanced dataset:

- Increase the number of samples in the minority class (Severe, Fatality) by duplicating existing rows or generating synthetic ones,
- Helps the model learn patterns in small classes (Fatality),
- Might lead to overfitting due to repeated data.

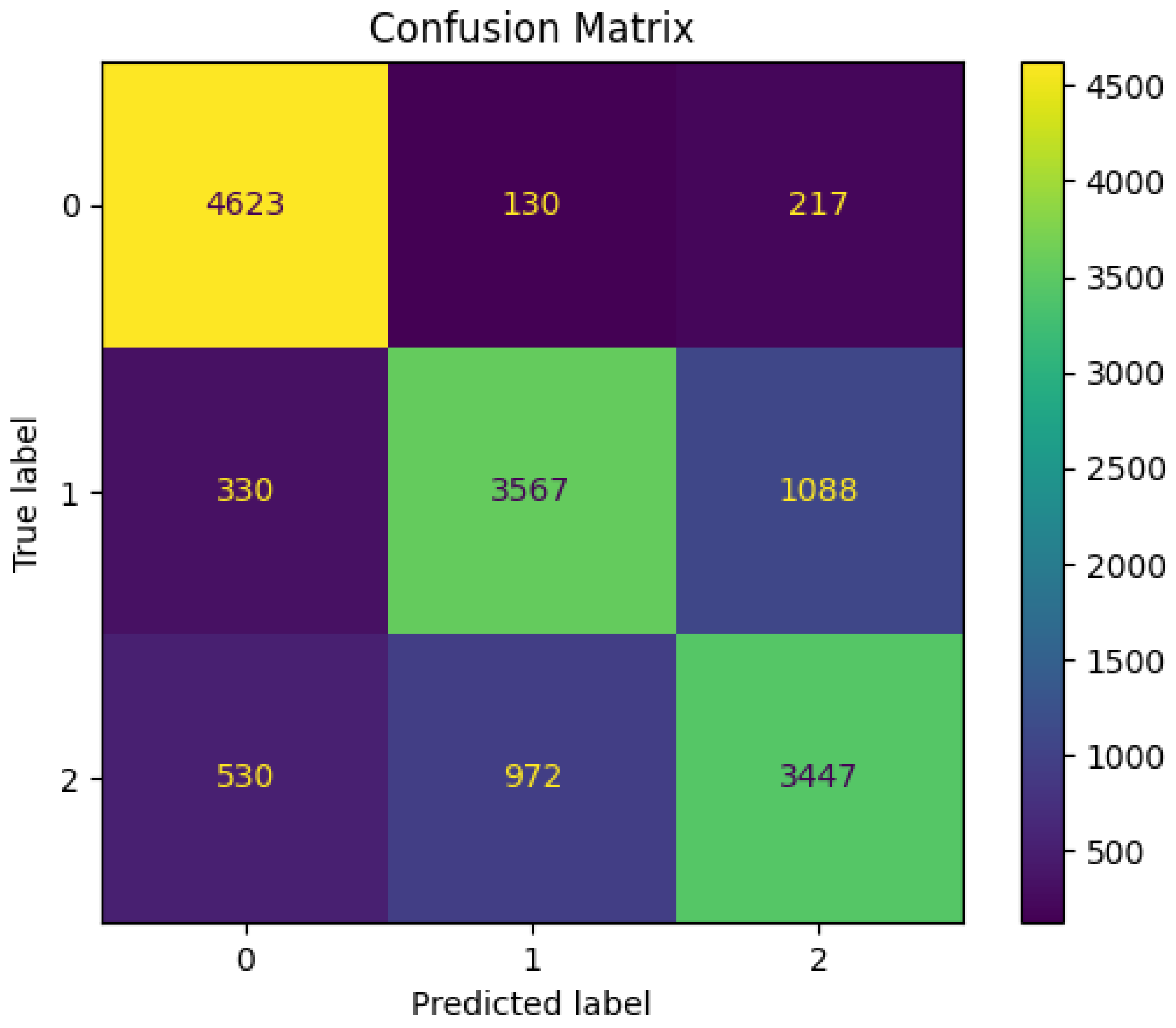
```
Original class distribution:
 Target
Minor        16559
Severe        3356
Fatality       466
Name: count, dtype: int64
```

```
Original class distribution:
 Target
Minor                16617
Severe               16257
Fatality             16567
Name: count, dtype: int64
```

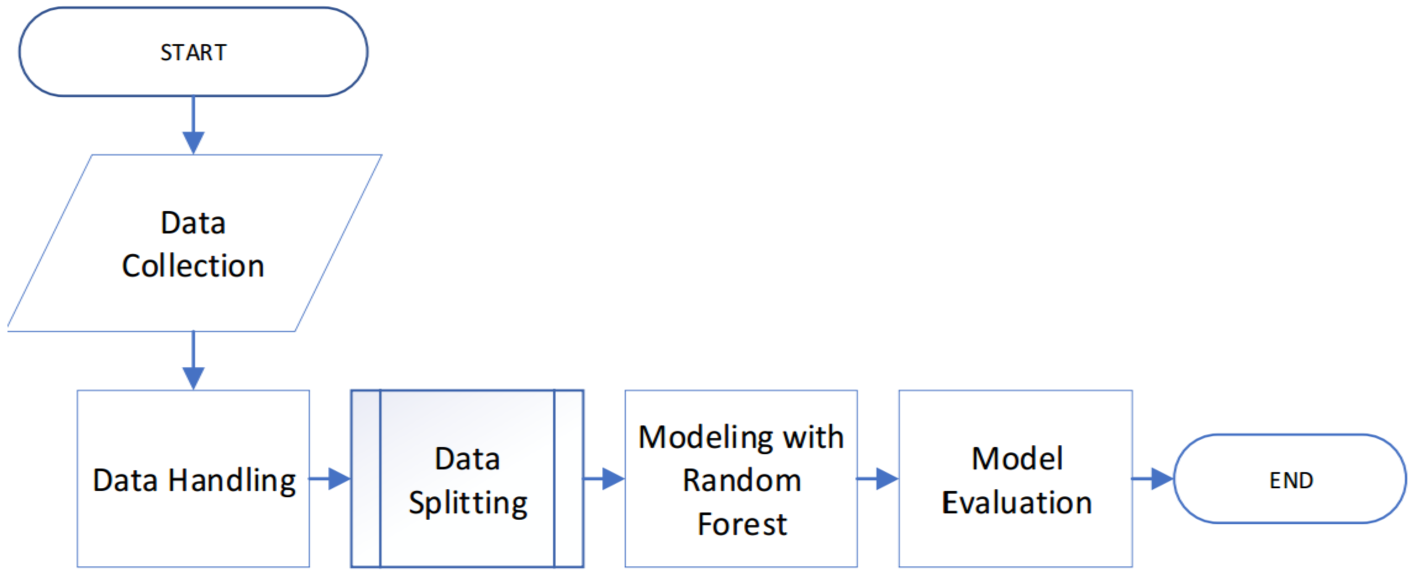
Random Forest – Crash severity prediction

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



Implementing analytic methodology FLOW chart



Original Data:

Color	Size	Price
Blue	L	100
Green	M	150
Red	S	200
Green	XL	120
Red	M	180

Label Encoding

Label Encoded Data:

Color	Size	Price
0	0	100
1	1	150
2	2	200
1	3	120
2	1	180

Conclusion

In this project, I applied machine learning and forecasting techniques to analyze 5 years of South Australian traffic accident data.

I identified key risk factors using SHAP, forecasted future crash trends with ARIMA and built predictive models using Random Forest.

While class (Minor, Severe, Fatal) imbalance posed a challenge, up-sampling techniques helped improve model performance. These insights can support road safety authorities in making data-driven decisions to reduce severe accidents and fatalities.

Peer review

- Developed dynamic visualizations,
- Investigated influence of alcohol and drug involvement in accidents.

Reference

- 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

Thank you
for your attention