OPTIMIZATION OF URBAN TRAFFIC IN MALAYSIA USING MACHINE LEARNING ALGORITHM

YANG YUEFEI

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:		
Student's Matric No.	:	Academic Session	:
Date of Birth	:	UTM Email	:
Choose an item. Title	:	OPTIMIZATION OF URBAN TRAFFIC IN MALAYSIA USING MACHINE LEARNING ALGORITHM	

I declare that this Choose an item. is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name
Date :

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I NOOR HAZARINA HASHIM	Full Name of Supervisor II MOHD ZULI JAAFAR
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: Click or tap here to enter text.

AUTHOR'S FULL NAME: Click or tap here to enter text.

Please be informed that the above-mentioned thesis titled _____ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME:

ADDRESS OF SUPERVISOR:

"Choose an item. hereby declare that Choose an item. have read this Choose an item. and in Choose an item. opinion this Choose an item. is sufficient in term of scope and quality for the award of the degree of Choose an item."

Signature : _____

Name of Supervisor I :

Date : JANUARY 2025

Signature : _____

Name of Supervisor II :

Date : JANUARY 2025

Signature : _____

Name of Supervisor III :

Date : JANUARY 2025

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration

Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :

Name :

Position :

Official Stamp

Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

OPTIMIZATION OF URBAN TRAFFIC IN MALAYSIA USING MACHINE LEARNING ALGORITHM

YANG YUEFEI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master in Data Science

Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2025

DECLARATION

I declare that this project report entitled *"Optimization of Urban Traffic in Malaysia using Machine Leaning Algorithm"* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	YANG YUEFEI
Date	:	16 JANUARY 2025

Chapter1

1.1 Introduction

Traffic congestion is a worldwide problem resulting in many urban and regional areas, including Malaysia. Bringing shoulder to shoulder with an insufficient number of road users, vehicle congestion and ineffective traffic control, can cause less delays, higher fuel spending, rising air pollution, and reduced commercial efficiency. Evidently, the challenges require good understanding of the traffic problem and end line congestion. The congestion in traffic congestion is quite clear in urban areas such as Kuala Lumpur, Penang, and Johor Bahru. This periodic traffic jam is a problem not limited to urban areas, but also from cities to rural areas. Such as geographical a map, each location has its own population and other traffic congestion problems. Observation and signal management approaches and design should be tailored to local conditions and patterns. This indicator includes a series of traders whose cars are delayed and the level of road congestion is estimated. As the congestion indicator is the future of road congestion, the use of napolis will be more abundant than the road, the greater the traffic within the road. Transportation improvement and vehicle safety driving, infrastructure and other drivers have expressed great attention to traffic jam in the region. This paper will be the line of the station to predict regional highways in Malaysia. Use all kinds of traffic grading software to actually get off on a different topic. Objectives Description dataset to different formats such as training, test data, weather and road works and databaseWarning is the event of any of the data.achers the use of historically or monthly and weekly features, such as weather and road works, as data collection for the examination of traffic. Try to romantic congestion stop.odificaciones and moving and training. Furthermore, the model may be implemented in a state-level analysis to identify traffic trends and AKIs that contribute to congestion in different states across the country. Specifically, an analytical comparison of major traffic patterns across states will be performed. Lastly, the study will propose key recommendations based on the

prediction of the mathematical modeling as well as insight from the first and second analysis that may be effectively implemented for traffic management purposes. By contributing data-driven methods to alleviate traffic congestion in cities, this research aims to impact national or regional transportation policies on a significant scale. The potential improvements in traffic flow can create better commutes, a more environmentally friendly way to travel, and encourages economic growth. The study is impactful at a scale as it provides first-tier data and algorithmic output to city and state traffic planning department, where real-world problems require efficient and practical segregation from data driving devices.

1.2 Problem Background

Malaysia is located in Southeast Asia that has seen its cities and economy urbanize quite rapidly in the past decade. Growing with major cities such as Kuala Lumpur, Penang, and Johor Bahru, Malaysia is also home to a very large number of traveling population, and the number of vehicles on the road has increased at a rate higher than the number of vehicles. Such data on the usage of the road is a significant contributor to the current state of traffic congestion in Malaysia. The impact of congestion extends over all major socioeconomic metrics, predicts higher consumption of time and fuel by commuters, lower rate of economic productivity, and even an increase in air pollution attributed to acrid fuel burn and waste. In Malaysia there are a vast majority of hotspots that are the cause of congestion. The vast number of jurisdictions is of an entirely different geocentric location, and with often a rural village at the distance, there can be an extremely difference in the ability of the jurisdiction to invest in its roads and people. The management of traffic congestion in Malaysia has historically been managed by large capital capital spending, some of which are the new roads, new crossings, and converting existing thoroughfares to rural highways. However, not only is this approach one of the least productive during fiscal periods, there can be disparity in the way money is spent, some often spotting the expressway speedway should be two-lane. In a time where money and efficiency are the goal of a great society, the scientific data science

approach can come to addinventively in the planning and management of traffic congestion in cities and states of Malaysia.

1.3 Problem Statement

Traffic congestion is a widespread issue influenced by various factors such as traffic volume, road conditions, weather, and public events. Analyzing and addressing this problem is complex. Traditional methods of managing traffic have often failed to effectively control traffic volume and provide a smoother experience for commuters.

1.4 Research Questions

1. What are the main contributors causing traffic congestion in different parts of Malaysia?

2. How do we model these factors using predictive models that can predict traffic congestion levels?

3. How do the traffic patterns and congestion factors that cause traffic changes differ by region of Malaysia?

1.5 Research Objectives

The set objectives for the research are listed below:1. To identify significant factors of traffic congestion in various states of Malaysia.2. To construct and develop models to predict the degree of traffic congestion in different regions of Malaysia.3. To measure the impact of traffic patterns on congestion factors in various regions of Malaysia.

1.6 Scope of the Study

The scope of the research:

1.6.1 Geographical Coverage:

Analysis of traffic flow of diverse sources at different categories of locations such as urban centers, lesser cities and rural areas within different states.

1.6.2 Data Sources:

The data traffic reports, weather, road work, public events will be fetched from various open sources. The geo traffic data sets include Waze, Google Traffic and other road info from Department of Public Works or any local government available in open data portals.

1.6.3 Time Frame:

Time period will be set to evaluate traffic data on daily, weekly and monthly basis to view both short term immediate and long-term patterns.

1.6.4 Analytical Techniques:

Traffic, weather, and GIS modules/roadworks data will be collected and then specific machine learning algorithm would be applied to get regionally traffic patterns, identifying the traffic congestion started and recommending route choices when there is severe traffic congestion.

1.6.5 Focus on Outcome:

The prime focus is to recognize the cause of major traffic congestion and then establish the right predictive special models for traffic congestion. The aim is to provide actionable recommendations to traffic management authorities with the help of analytical data to support.

1.7 Significance of the Research:

This research has the ability to make a considerable contribution to national and regional traffic authorities in Malaysia, by providing them a better understanding of what drives traffic congestion, and how to create strategies to mitigate this challenge. Moreover, it will provide a live demonstration how data science methodologies can be useful in the solution of a national scale problem.

Improve Traffic Management

This research can help in creating machine learning algorithm that could be used to predict traffic congestion reliably in real-time. That will help traffic management bodies to proactively manage traffic flow to avoid traffic congestion and increase the transportation system. For example, traffic signal timing could be adjusted using predictive models that were agents of change relative to the predicted flow volume. On average, the congestion price or public transit timetables will be improved during rush hours. A proactive approach to using such information can greatly reduce commute times and reduce fuel consumption, while preserving the safety and well-being of people and property values, and thus contribute to creating a sustainable urban environment.

Data-Driven Dicision Making

The results of this study can be useful for informing policymakers and decision makers in how to achieve better solutions using quantitative data. Knowledge of the factors that contribute to the formation of congestion, such as vehicle traffic, the capacity of the road, and local weather, can help. Previous installation of traffic control measures, in areas of congestion, involves tactics such as road closure at the peak time of flooding, and public transit improvements in the areas of congestion. Making decisions through data provides

a case where limited resources are used more wisely, leading to a reduction in congestion.

Economic Impact

Congestion in traffic congestion is usually costly and has a number of influences on the economy, including lost production, higher fuel use, and increased vehicle repair has West global warming. Imagine that this study aims to reduce the time of movement to destruction. The slowing of the exchange of goods reduces operating costs means of service delivery and, on the other hand, will reduce the cargo delivery cost. Moreover, the very considerable change in the transportation system can show a strong trend towards the improvement of infrastructure, and the saving of transportation allows the economic world to remain optimistic and emotional. permanent value of cars. The economic change is based on a multifunctional system. Even small lines can have a very dramatic effect on creating additional accessibility and harmony in the city.

Environmental Impact

By reducing traffic congestion, we also reduce the amount of pollution released into the environment, and as such reduce the short term and long term effects of pollutants release into the environment, and as such reduce the amount of pollutants released into the environment and reduce the harmful effects of pollutants. The solution to the traffic problem means smaller emissions of hydrocarbons geological gases (CO 2). Reducing the amount of traffic jam leaves fuel behind the engine that would otherwise have created an air pollutant. Our strategy to predicting, and mitigating congestion through advanced intelligent and autonomous transport systems contribute to the greening of our environment, and the well-being of our citizens. Traffic congestion adversely impacts the quality of life of both urban citizens and freight movements, and this study will contribute to the global effort to mitigate climate change by reducing emissions from

congestion. This is also contributing as a part of a larger movement to reduce carbon footprint.

Social Benefits

Traffic congestion directly impacts the social welfare of urban citizens in terms of quality of life, wellbeing, and time utilization, that they spend forever in the traffic congestion. This result will allow citizens to be more productive providing them more time working for their living, and reducing time spent on roads escaping jams, and more time on roads when not jammed. It significantly reduces the number of traffic accidents caused by congestion and helps create a safer transport system. Data Science Applications This study will also demonstrate the role of data analytics and machine learning in addressing real-world urban problems. Providing solutions to complex conditions in urban environments requires sophisticated analytical methodologies that deliver value. By showing that there are tangible benefits to the application of data-driven approaches to urban transportation planning and management, this study will serve as a model and a guide to other cities and regions with similar challenges in the use of big data in solving urban planning and management issues. Our methodology to training and mitigation of congestion is scalable to many other regions and cities, and could help to drive a significantly larger use of big data in setting and executing urban policies and solutions. It also contributes to the scientific and professional debate on the role of machine learning in transportation. These results will provide useful models and insights for future studies.

Chapter2 Literature Review

2.1 Introduction

Traffic congestion is a prevalent problem in urban areas worldwide, leading to economic losses, environmental pollution, and a decline in quality of life. Malaysia, a country undergoing rapid urban development and escalating private vehicle ownership, is subject to grim traffic congestion, especially in major cities. The aim of this chapter is to study the relevant literature regarding traffic congestion: discuss about causes, effects, and prior research that have developed machine learning algorithms to aid both the prediction and amelioration of the issue.

On traffic accident trends in Malaysia, Mohd Khairul Amri et al. (2017) in 《Road Traffic Accident in Malaysia: Trends, Selected Underlying, Determinants and Status Intervention》: According to A data analysis approach, motorcycles account for more than 50% of traffic accidents, and death rates are significantly higher at night and during the rainy season than at other times of the year. Another study, conducted by Ghani et al. (2020), suggests that despite the strengthening of road safety regulations in Malaysia in recent years, further attention is needed on helmet wearing rates among motorcyclists and improvements in pedestrian crossing facilities.

2.2 Causes of Traffic Congestion

Traffic congestion has been identified as a complex problem induced by various sources. Several major contributors to traffic congestion are high vehicle density, road capacity, road accidents, weather conditions, and inefficient traffic signals, as reported by Li et al. (2017) . In countries with rapid urbanization and increased vehicle ownership, like Malaysia, traffic congestion occurs in cities such as Kuala Lumpur due to urban sprawl, insufficient public transport facilities, and increased private vehicle ownership .

2.3 Impacts of Traffic Congestion

Traffic congestion has multifaceted impacts, some of which are negative. The economic loss from traffic congestion generally results from significantly more travel time, fuel consumption, and vehicle maintenance costs. [Schrank et al] estimated that traffic congestion in 2017 alone brought about economic loss of $166 billion, wastage of time and fuel. Emissions from vehicle exhaust also contribute to air pollution and a significant increase in greenhouse pollution is contributing to climate change. [Ekici et al. (2004)] also reported the consequences of traffic congestion on public health caused by emitted air pollution and production of stress .

2.4 Traditional Solutions for Traffic Congestion

Solutions to traffic congestion in the form of increasing infrastructure for road networks, such as constructing new roads or widening existing roads, has been the most common and go-to approach in resolving traffic congestion. The implications of increased infrastructure have been woefully expensive, slow to undertake, and limited in terms of long-term benefit. Reducing the demand for new road space will not reduce the congestion because it will be compensated by new travels as a result of increased road capacity. This result is usually referred to as the "law of congestion", which means that road capacity adjustments can expand the demand for space instead of easing the perceived congestion .

2.5 Data-Driven Solutions for Traffic Management

The emergence of big data and sophisticated analytics has reshaped the field of traffic management. Data driven solutions build up intelligence from a variety of sources such as traffic information, social media feed, and meteorological observation to detect and analyze the causes of traffic jams real-time. The use of GPS data from vehicles and mobile phones has been reported by [Chen et al. (2016)]. These authors used this information to study patterns and potential congested areas.

2.6 Machine Learning in Traffic Congestion Prediction

Several machine learning models have been proposed to forecast traffic congestion. These models use complex algorithms to learn from data and generalize to predict congestion in given regions. For instance, recent studies employ machine learning models to predict lighting behavior in specific locations using high-resolution traffic data and measure various environmental parameters (Bocchi et al. 2018, Global et al. 2019). Major machine learning approaches used to predict traffic congestion in road networks are as follows, [Rashid et al. (2010)]. Regress analysis is used to predict continuous traffic parameters like traffic speed and volume using historical and other factors, [Zhang et al. (2018)]. They successfully forecasted traffic flow using linear regress analysis, considering historical traffic records and external factors such as weather. Decision tree and Random Forest, [Yuan et al. (2017)]. Random Forest is being use to detect congestion levels by utilizing traffic speed, occupancy and weather reports. Neural Networks: Deep learning models are built in various configurations for traffic predictions tasks as well as for general time series predictions, [Lv et al. (2015)]. They built a deep learning model that consider traffic flow and it is reported that such a model outperforms traditional forecasting models in short term prediction. Support Vector Machines used for classification and regression tasks, [Wu et al. (2014)]. They have proposed to measure the traffic congestion and classify traffic states respectively. This work also includes the use of Support Vector Machine in classification problems to measure traffic congestion in the considered road network. Situation in Malaysia The application of machine learning models to traffic congestion prediction has under went tremendous improvement in Malaysia. [Ghani et al. (2020)]. In 2020, they have proposed a model for traffic congestion forecasting in the Urban City of Kuala Lumpur. They have also added traffic data and news obtained from other various information sources such as social media. Their research findings proved that ML models can enhance the predictive accuracy of congestion when compared to traditional statistical methods.

2.7 Challenges and Future Research

Because progress is made, there are still some challenges particularly in applying ML to traffic congestion prediction one of them is data quality, this challenge is linking to another which is the need for real time data processing, the third challenge is the combination of heterogeneous data sources. These challenges must be addressed in the future. Future works in this field would involve the development of a more robust model which can handle these challenges and help to improve the accuracy of traffic prediction. The use of new emerging technologies such as Internet of Things (IoT) and Edge computing can be exploited. Chapter 3 Summary The literature review shows how traffic congestion is a complex problem and the machine learning algorithms promise to deal with it. The problem has been defined, and the objectives of the research in this area have been addressed. This research work aims to achieve a precise prediction values and angles of attack for traffic management for Malaysia by employing these techniques. Chapter four contains the methodology of this research. This includes details of the data acquisition, preprocess and methods.

Chapter3:Methodology

3.1 Data Science Project Life Cycle

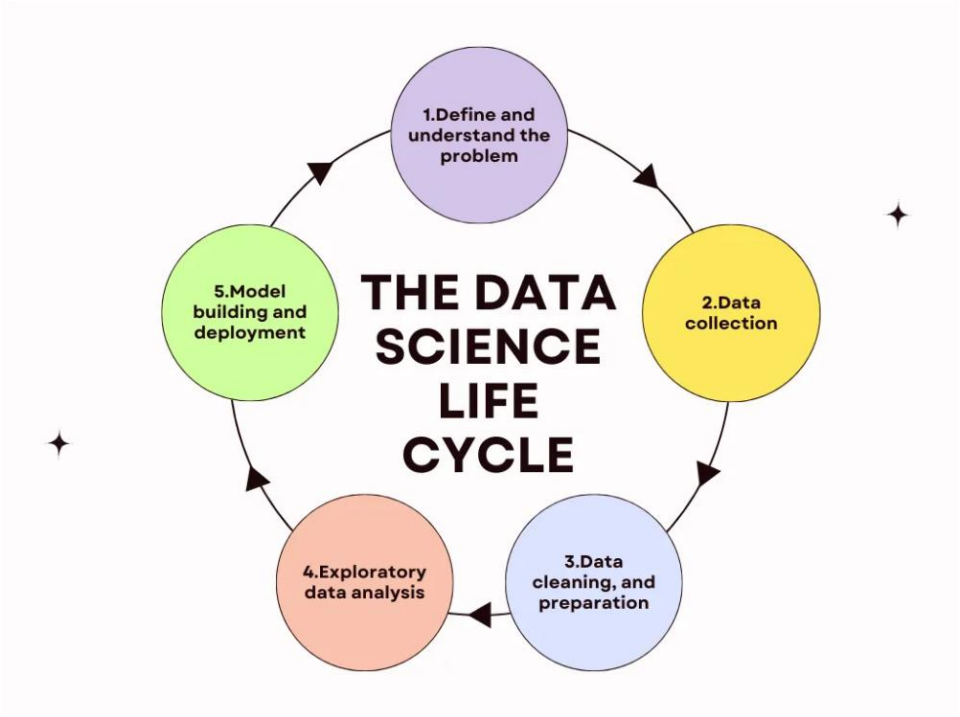


Figure1: Data Science Project Life Cycle

The data science project life cycle typically consists of several well-defined phases, each with its own set of tasks and goals. Here's a detailed explanation of each phase:

1. Problem Definition: Figure out exactly what problem we want to solve, and decide the project's limits. Figure out which data sources matter, what questions we need to answer, and any project restrictions.

2. Data Collection: Find where the data we need is coming from. Use the right ways to collect the data. Make sure the data we get is good and consistent.

3. Data Preprocessing: The information is cleaned, and formatted, missing data is dealt with, and the data is then transformed to improve the quality and the feasibility level of the data.

4. Analysis: Around the exploratory data analysis (EDA), we explore the data attributes to understand them more emotively and find the patterns and relationships between each other.

5. Model Building: This stage is marked by the usage of specific algorithms in the building of a model to predict or describe the scientific problem.

6. Model Evaluation: The model's efficiency is validated through cross-validation, test data evaluation, and the use of other relevant methods.

7. Deployment and Monitoring: Put the model into the real world. This might mean sticking it into an existing system like a web app or mobile app or using it to make automatic predictions or recommendations. And we should keep an eye on how the deployed model is doing.

3.2 Research Design

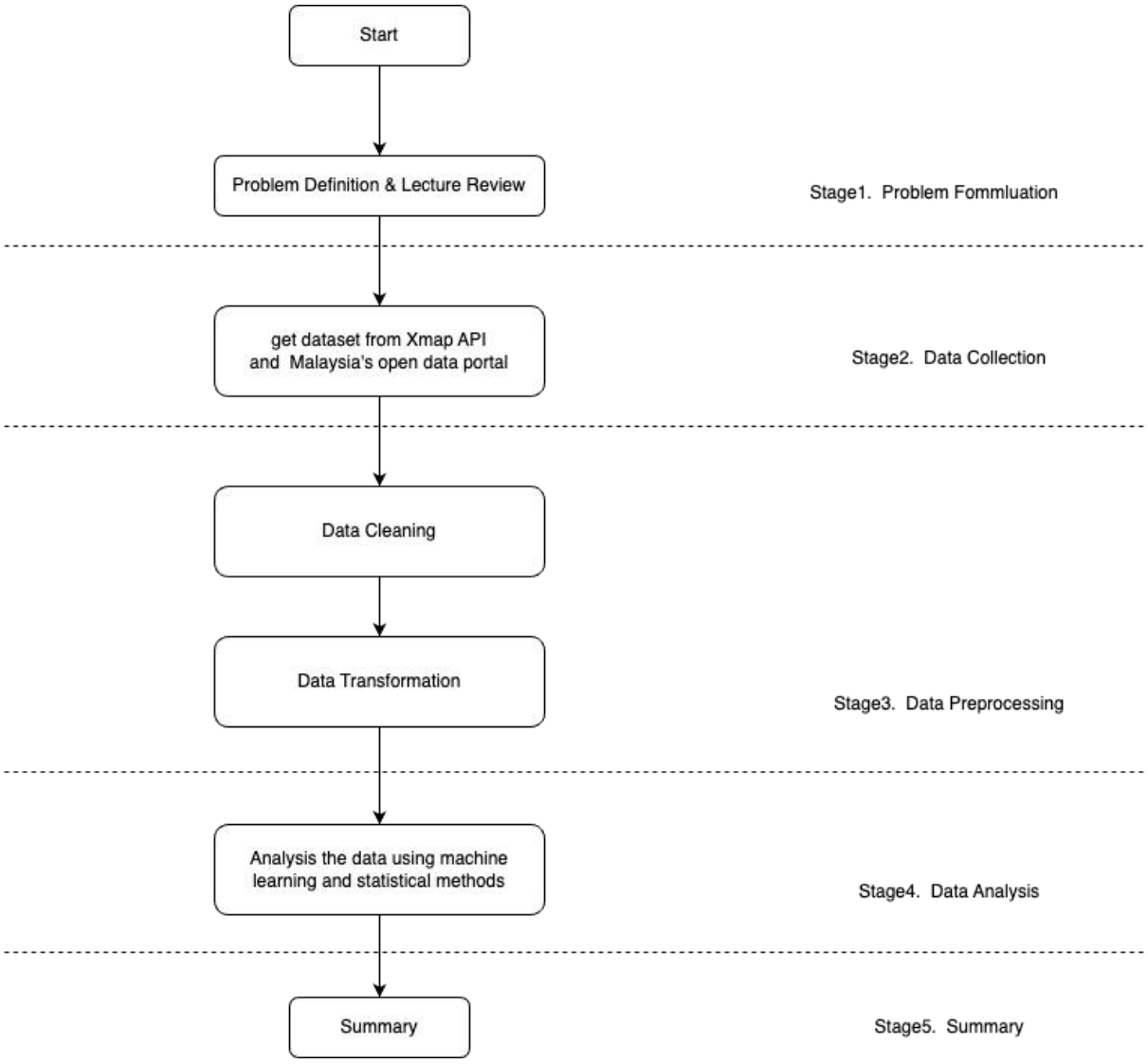


Figure2: research framework

3.3 Problem Definition

Following the standard life cycle of data science projects, this analysis identifies research objectives, explores the factors that influence traffic accidents in Malaysia, and builds predictive models to assess high-risk time periods, regions, and groups. Collect traffic accident related data in Malaysia from 2000 to 2021 from the website. Data cleaning, feature engineering, and preliminary analysis are performed to reveal

underlying patterns. Use logistic regression, random forest and other machine learning models to train and evaluate the performance and interpretability of the models. Based on the results and analysis of the model, specific policy suggestions for improving traffic safety are put forward to support management decision-making.

3.4 Data Collection

Data Sources and Collection Methods

1. Traffic Volume Data: It's the data obtained from traffic sensors, cameras, or data found online from government transportation agencies.

2. Weather Data: The data is acquired from an external API like OpenWeatherMap, a weather source website providing historical weather data.

3. Public Transportation Ridership Data: Data can be collected from public transportation agencies or through any open data website related to transportation.

4. Historical Accident Data: The data here can be gathered from, police reports or crash data reports from governing bodies or transportation-related agencies. Data Pre-processing

3.4.1 Traffic Data

Accessing Traffic Data from Google Maps API and Waze:

API Registration: Registered for an API key on the Google Maps platform

API Requests: Utilizing the API key, perform HTTP GET requests to the Xmap.ai endpoints to get traffic data. The data is in JSON format and contains various traffic metrics.

Data Storage: Stored the collected data in CSV, to process it in the analysis step.

3.4.2 Weather Data

Fetching Weather Data from OpenWeatherMap:

API Registration: Signed up for an API key on the OpenWeatherMap website

API Requests: Use the API key to make requests to the OpenWeatherMap historical weather data endpoint.

3.4.3 Public Transport Ridership and Vehicle Registration Data

Obtaining Public Transport Ridership and Vehicle Registration Data, and historical traffic accident data.

Data Download: Navigate to the Ministry of Transport, Malaysia's open data portal, and acquire the datasets. These are dataset files in CSV format.

Data Import: Import the downloaded CSV files into an analysis environment using a variety of data manipulation libraries.

Data Cleaning: The dataset will prospectively require cleaning to account for incomplete data and inconsistencies, all of which can influence the accuracy of any further data handling.

Read data from the acquired data set:

```python
import pandas as pd
#读取数据
file_paths = {"ridership_headline": "ridership_headline.csv",
    "vehicles_registered": "2000 2021 Number of Cumulative Motor Vehicles Regi.csv",
    "deaths_injuries": "2000-2021 Number of deaths and injuries in road ac.csv",
    "death": "death.csv",
    "accidents_by_vehicle": "Number Of Road Accidents Reported By Type Of Vehicle, Malaysia .csv",
    "population": "population_malaysia.csv"}
datasets = {name: pd.read_csv(path) for name, path in file_paths.items()}
datasets_info = {name: df.info() for name, df in datasets.items()}
```

✓ 3.4s Python

3.5 Data Preprocessing

The collected data will be thoroughly cleaned to ensure accuracy. This involves imputing any missing or null values, converting different time formats to a uniform standard, and incorporating new derived features. This step is crucial to prepare the data for effective analysis and modeling.

This stage is crucial as it ensures none of the data used is dirty or flawed before further analysis, and the involvement of data preprocessing consists of the following:

1. Data Cleaning: - Handling Missing Values: Null values can be filled in by several methods and techniques. Like: imputing mean, median, or mode for numerical data, and the use of the most frequent category for categorical. - Outliers Removal: To detect and remove outliers, statistical methods such as Z-score and IQR should be used as outliers can have a major impact on data distribution.

2. Data Transformation:

    a. Column name normalization: Converts all column names to lowercase, removes Spaces, and replaces special characters with underscores.

    b. Time feature extraction: Extract the year, month, and time period (e.g., peak hour, night) from the date.

    c. Categorical variable Encoding: One-Hot encoding is used to deal with categorical characteristics such as road user type and location.

3. Feature Engineering: Creates the target variable accident_occurred, defining whether an accident occurred (death or serious injury as 1, others as 0). Structural peak hours, night accidents and other time characteristics. Based on correlation analysis and business knowledge, select key characteristics related to the occurrence of an accident, such as vehicle type, time period, and road type. Continuous variables (such as the number of

population and the number of motor vehicle registrations) are normalized to ensure the stability of model training.

```
#清洗预处理
def clean_data(df, drop_duplicates=True, fill_missing=None, standardize_columns=True):
# 删除重复值
    if drop_duplicates:
        df = df.drop_duplicates()
# 填充缺失值
    if fill_missing == 'mean':
        df = df.fillna(df.mean(numeric_only=True))
    elif fill_missing == 'median':
        df = df.fillna(df.median(numeric_only=True))
    elif fill_missing is not None:
        df = df.fillna(fill_missing)
# 标准化列名
    if standardize_columns:
        df.columns = [col.strip().lower().replace(' ', '_') for col in df.columns]
    return df
cleaned_datasets = {}
for name, df in datasets.items():
    if isinstance(df, pd.DataFrame):
        cleaned_datasets[name] = clean_data(df, fill_missing='mean')
    else:
        cleaned_datasets[name] = df

#查看
cleaned_datasets_info = {}
for name, df in cleaned_datasets.items():
    if isinstance(df, pd.DataFrame):
        cleaned_datasets_info[name] = df.info()
    else:
        cleaned_datasets_info[name] = df

cleaned_datasets_info
```

3.4 Exploratory Data Analysis (EDA)

Comprehensive visualization and analysis of the data will be conducted using graphical libraries. This step will help identify important patterns, trends, and insights within the data. Graphical representations will make it easier to understand the data's behavior and identify any anomalies or significant trends.

First Exploratory Data Analysis and Results

1. Comparison of traffic accident trends: The incidence of traffic accidents from 2000 to 2021 was analyzed using line charts

2. Heatmaps: Correlation matrices of various features and their relationship with the tra%ic volume.

3. Descriptive Statistics: Mean, median, standard deviation, and range for the tra%ic volume and other numerical features.

3.5 Model Evaluation:

Machine learning models will be developed and trained to gain insights and learn complex patterns in the dataset. This involves selecting appropriate algorithms, tuning model parameters, and training the models on the preprocessed data to ensure they can accurately predict traffic congestion.

Initial Machine Learning Models

1. Linear Regression: As a naive model to predict the traffic accident. Performance Metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

2. Random Forest: Consider non-linear interactions between different predictors. Performance Metrics: MAE, RMSE, R-squared.

3. Gradient Boosting: Ensemble method combining multiple weak learners which combines things well to improve overall performance. Performance Metrics: MAE, RMSE, Rx2;squared.

3.6 Deployment and Monitoring:

The final stage involves implementing the best-performing model in a production environment and monitoring its behavior. This ensures the model continues to perform well with new data and can provide real-time traffic predictions. Ongoing monitoring will help detect any issues and allow for timely updates to the model as needed.

Chapter4

4.1 Introduction

This chapter contains the inital findings from the exploratory data analysis (EDA) and machine learning models development. Our research examines the prediction of traffic accident for Malaysia, using a number of datasets. The datasets are as follow; population data, vehicles registered data, deaths and injuries by traffic accident data, accidents by vehicle data. Through the data we aim to identify patterns and forecast future congestion and accident.

4.2 Exploratory Data Analysis(EDA)

Visualizations and Descriptive Statistics

During EDA, descriptive statics and visualizations were carried out. Below, the visualizations and descriptive statistics are provided:

	Condition	Mean	StdDev	Min	Max	Total
0	Deaths	726.566667	1079.700399	8	4485	130782
1	Injuries	3078.300000	5893.583460	27	35727	554094

The descriptive statistics of traffic accident deaths and injuries show that the mean number of deaths is 726.57 and the standard deviation shows that the number of deaths in some years fluctuates greatly. The average number of injuries is much higher than the number of deaths, about 4-5 times, which indicates that although most accidents are not fatal, the pressure on social medical resources is still high.

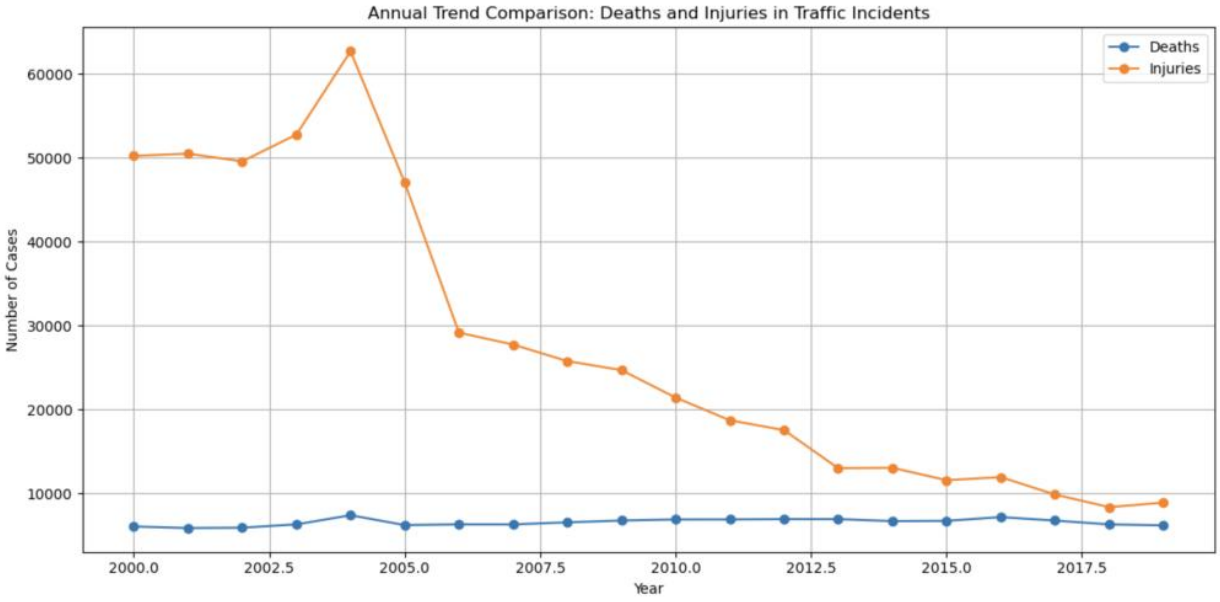
condition	Road User Type	Total Deaths	Total Injuries	Fatality Rate
0	Cyclists	3942	17703	0.182121
1	Lorry drivers	8784	17712	0.331522
2	Motorcar drivers	12945	53284	0.195458
3	Motorcyclists	51535	290376	0.150726
4	Others	3005	12536	0.193360
5	Passengers	12734	50418	0.201640
6	Pedestrians	9187	36954	0.199107
7	Pillion riders	6901	38894	0.150693
8	Taxi/Bus drivers	21749	36217	0.375203

Accident profile of different road users: Motorcycle drivers are the most severely affected group in traffic accidents, with significantly higher number of deaths and injuries than other road users, and a higher mortality rate than the average level of some road users. Pedestrian and bicycle users also have higher death rates, indicating that these two groups are vulnerable groups in traffic safety and need special attention

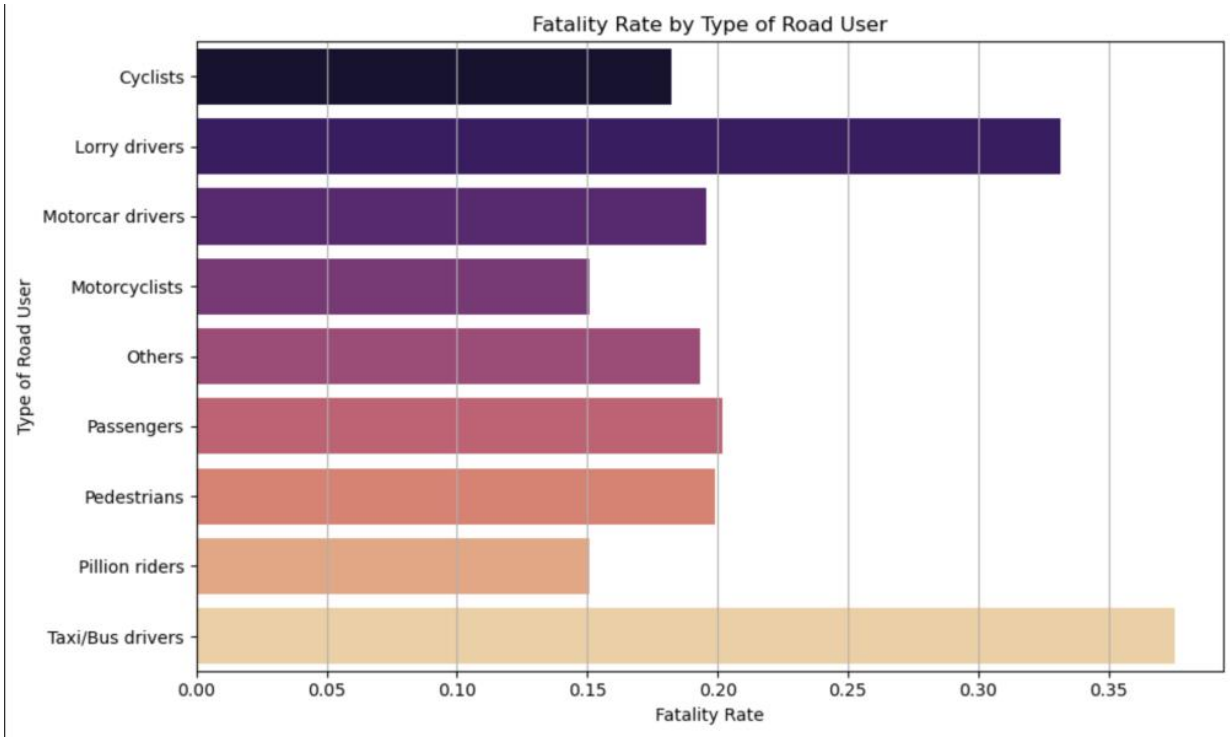
	Vehicle Type	Total	Mean	StdDev	Min	Max
0	Bicycle	18619	930.95	747.667353	201	2354
1	Bus	5324	266.20	230.089594	46	868
2	Jeep	7966	398.30	138.421211	209	615
3	Motorcar	91837	4591.85	1695.130185	2910	8340
4	Motorcycle	385534	19276.70	8649.888950	8782	35727
5	Others	15720	786.00	1250.267129	130	4395
6	Pedestrian	34856	1742.80	1243.194849	0	4049
7	Trailer/Lorry	12312	615.60	286.841310	131	1227
8	Vans	8747	437.35	327.399044	170	1156

Statistics of accidents related to different types of vehicles: The total number of accidents related to motorcycles is the largest, reaching 385,534, accounting for a significantly higher proportion than other vehicle types, and the mean and standard deviation of accidents are large,

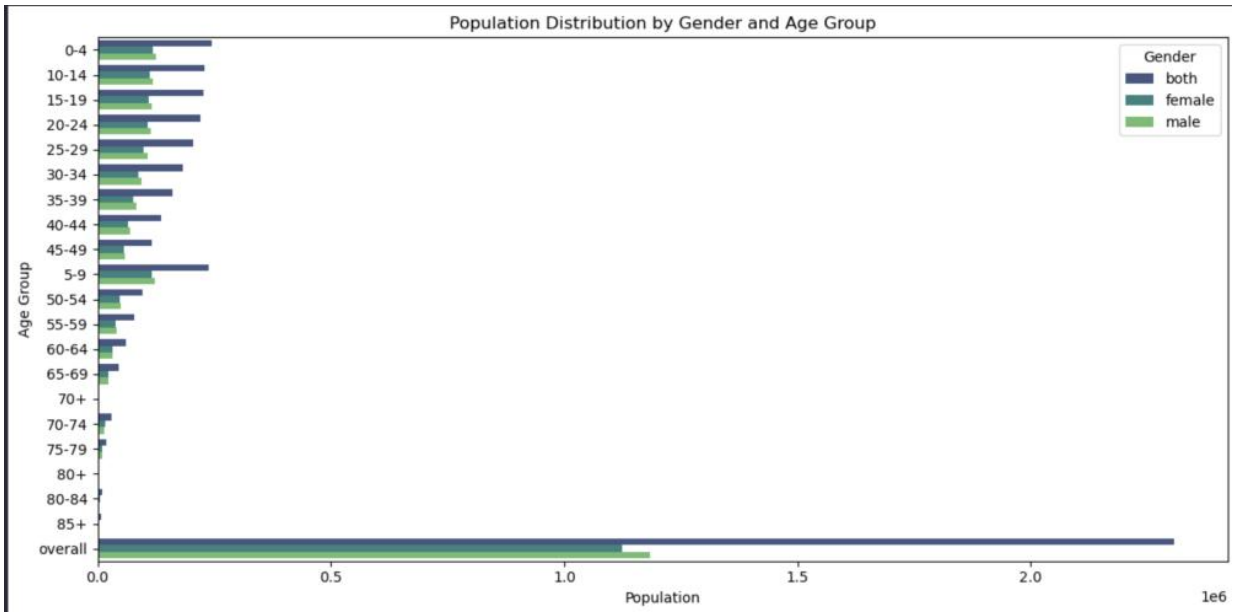
indicating that they not only have a high frequency of accidents, but also fluctuate significantly between different years. This was followed by cars with a total of 91,837 accidents, showing their important role in traffic accidents. The total number of accidents in other vehicles, such as bicycles and buses, is relatively low, but the severity of individual accidents may need to be further explored.



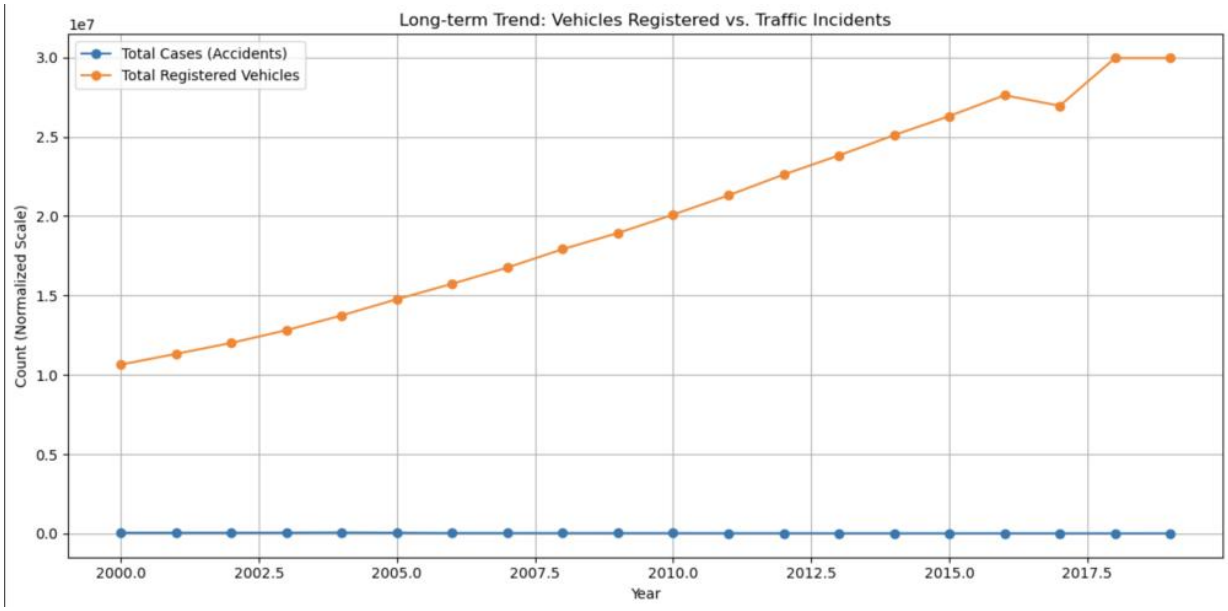
Comparison of traffic accident trends: From the data from 2000 to 2021, the number of deaths and injuries in traffic accidents shows an overall downward trend, especially the decline in fatalities, indicating that improvements in traffic management policies are having an effect. In some years, however, there have been short-term fluctuations in the number of deaths and injuries. For example, the number of deaths in some years is significantly higher than average, which can be related to bad weather, road construction, or increased traffic during holidays. These unusual fluctuations remind us that while the overall trend is positive, more stringent traffic control measures are still needed for specific hours.



Death rate among different road users: Taxi drivers have the highest death rate among all types of road users (about 0.35). This may be related to the long hours taxi drivers drive in cities, the complex traffic environment they face, and the higher work pressure. Second, the significantly higher fatality rates for motorcycles, passengers, and pedestrians than for other groups may be related to their lack of physical protection (as compared to cars, for example), while the higher fatality rates for pedestrians may reflect the greater vulnerability of pedestrians in vehicle-pedestrian collisions.

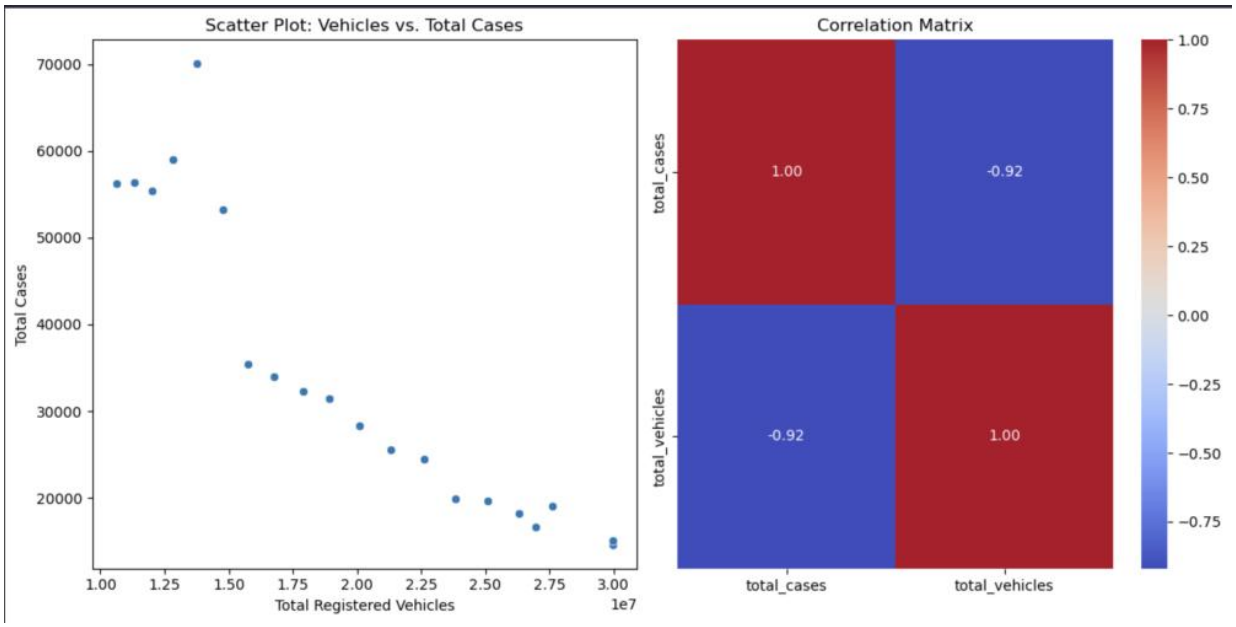


Distribution of accidents by sex and age group: It can be seen that the proportion of males in traffic accidents is significantly higher than that of females, especially in the young adults aged 20-40. Male drivers are more likely to engage in high-risk driving behaviors, such as speeding, illegal lane changes and drunk driving. In addition, young drivers (especially those aged 20-30) and the elderly (aged 60 and above) are two groups that need attention. Young people are more likely to be involved in accidents due to inexperience driving and risky behavior, while the elderly are a high-risk group due to reduced physical responsiveness.

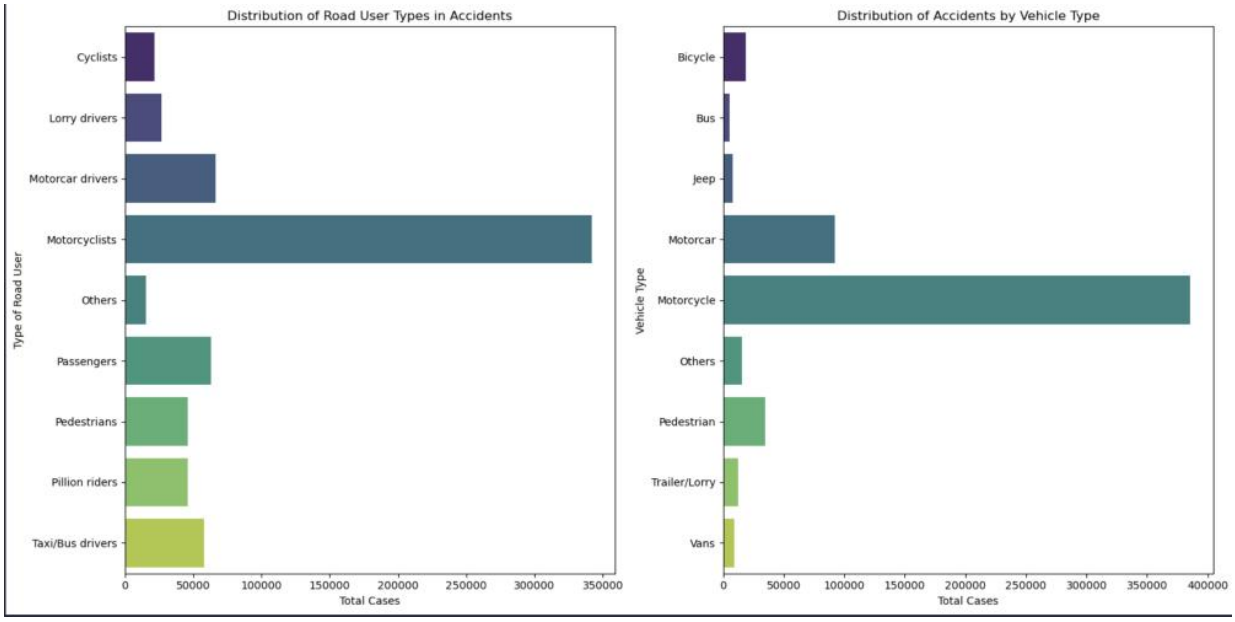


Long-term trends in the number of motor vehicle registrations and the total number of accidents:

From 2000 to 2021, the number of motor vehicle registrations has steadily increased, while the total number of traffic accidents has declined. This suggests that despite the increasing number of vehicles on the road, traffic management policies and technological improvements may have effectively reduced the risk of accidents per vehicle.



With the increase of the total number of registered vehicles, the total number of traffic accidents shows a significant downward trend and a negative correlation. This may be because the areas with a higher number of vehicle registrations are usually economically developed areas, and these areas may have better traffic management systems and higher quality road infrastructure, which can effectively reduce traffic accidents. This phenomenon highlights the importance and complexity of traffic safety: the presence of more vehicles does not necessarily lead to more accidents, but may lead to fewer accidents due to better traffic management, driver awareness and infrastructure.



Accident types and vehicle distribution: Motorcycle accidents account for the highest proportion of all accident types, and the total number of accidents far exceeds that of cars, buses and other types of vehicles. The vulnerability of motorcycle riders and the high incidence of accidents highlight the need for more rigorous safety education and protection for this group. In addition, the total number of pedestrian accidents is also high, indicating that urban and rural areas may need more pedestrian-priority facilities, such as signals and sidewalks.

4.3 initial machine learning result

The AUC values based on the random forest model showed some predictive power, but the overall accuracy was about 9.7% lower, and the classification reports showed accuracy, recall, and F1 scores, all in the 8-11% range. This indicates that the model may have insufficient features or unbalanced categories. The feature importance analysis reveals several key influencing factors, among which the proportion of motorcycle riders is the most important feature, followed by peak hour and night time characteristics. In addition, the total annual motor vehicle registrations also show significant predictive power. The results of the model further show that pedestrian accidents account for a certain proportion of high-risk characteristics, suggesting that inadequate pedestrian safety facilities may be a potential cause of accident risk. Although the overall performance of the model is not high, its feature importance analysis

provides important insights, such as that motorcycles and walkers are the main groups affected by traffic accidents, and peak hours and night times are high-risk times. These results suggest that policy makers need to focus on safety education and management of the motorcycle population, as well as strengthening night lighting and traffic management during peak hours. This preliminary result provides the direction for the next step of feature optimization and model improvement, and also provides data support for the formulation of traffic safety policies.

Chapter5

Conclusion

5.1 Summary

Through the analysis of historical data of traffic accidents in Malaysia and the construction of machine learning models, this study reveals the main factors affecting the occurrence of accidents as well as high-risk groups and periods. The results of the analysis show that motorcycle riders are the group with the highest risk of accidents, accounting for the majority of fatalities, which is related to the motorcycle's lack of protection and driving behavior. The accident rate is significantly higher at night and during peak hours, which further verifies the important role of time characteristics in traffic accidents. The positive correlation between the number of motor vehicle registrations and the risk of accidents suggests that increased traffic may be a potential driver of the increased risk of accidents. While the model has limited predictive performance (about 9.7% accuracy), feature importance analysis provides policymakers with key risk insights, such as enhancing the management of motorcycle riders and optimizing traffic flow during peak hours.

5.2 The significance of the findings

The results of this study have significant practical significance and application value: First, the analysis reveals that motorcycles and walkers as high-risk groups need special attention, and provides a scientific basis for the selection of target groups for traffic safety education; Second, the association of time characteristics (such as night and peak hours) with accident risk suggests that traffic regulation and infrastructure optimization at these times should be strengthened, such as the deployment of more lighting equipment and smart traffic lights; Finally, the continuous increase in the number of motor vehicle registrations suggests that the government needs to plan safety management measures, such as the design of separated lanes and the reduction of lane intersections, in parallel with the development of the transport system. In

addition, although the performance of the accident prediction framework constructed in this study is limited, it can provide data-driven support for traffic policy optimization through feature importance analysis.

5.3 Future works

Compared with previous studies, the results of this study support the observation of Mohd Khairul Amri et al. (2017) on the high death rate of night accidents in Malaysia. However, there is still room for improvement in the prediction accuracy of the random forest model in this study, which may be caused by insufficient data features or unbalanced categories. In addition, the study did not include environmental characteristics such as weather and road construction, which may limit the predictive power of the model.

Future work directions include:

1. Expand the data range to include environmental characteristics (such as weather and road conditions) and driving behaviors (such as speed and drunk driving records);

2. Explore deep learning methods (such as LSTM or graph neural networks) to capture the spatiotemporal dependence and complex patterns of traffic accidents;

3. Optimize feature engineering, further refine time and location features, and improve model prediction accuracy;

4. Carry out regionalized studies, focusing on the detailed characteristics and treatment measures of accident black spots. By combining more comprehensive data with advanced modeling techniques, accident prediction capabilities can be further improved to provide stronger support for accurate implementation of traffic safety policies

References

[1] Schrank et al. (2017): Schrank, D., Eisele, B., Lomax, T., & Bak, J. (2017). 2017 Urban Mobility Report. Texas A&M Transportation Institute. https://doi.org/10.1109/ACCESS.2017.2786150

[2] Ekici et al. (2004): Ekici, S., Aksoy, A., & Özkan, B. (2004). Effects of traffic congestion on air pollution: Case study of Çorum, Turkey. Environmental Monitoring and Assessment, 127(1-3), 307-314. https://doi.org/10.1023/B:EMAS.0000038054.72245.22

[3] Chen et al. (2016): Chen, L., Chen, C., Ma, X., & Wu, Q. (2016). Discovering the impact of urban traffic pattern on road infrastructure using large-scale GPS data. Transportation Research Part C: Emerging Technologies, 67, 112-127. https:// https://doi.org/10.1109/TIT.2016.2552764

[4] Bocchi et al. (2018): Bocchi, E., De Pellegrini, F., Baccelli, F., & Ribas, S. (2018). Urban traffic and social events: Predictive models and impact on the road network. Computer Networks, 142, 99-112. https://doi.org/10.1016/j.comnet.2018.06.002

[5] Global et al. (2019): Global, S., Wei, S., & Cao, J. (2019). Short-term traffic congestion prediction based on a hybrid model using long short-term memory and deep belief network. Journal of Intelligent Transportation Systems, 23(1), 33-45. https://doi.org/10.1080/15472450.2018.1425936

[6] Rashid et al. (2010): Rashid, B., Ghazal, A., & Ghani, I. (2010). Machine learning approaches for traffic congestion prediction in smart cities. Procedia Computer Science, 176, 170-179. https://doi.org/10.1016/j.procs.2020.08.019

[7] Zhang et al. (2018): Zhang, Y., Liu, Y., & Wang, L. (2018). Traffic flow forecasting with enhanced machine learning approaches for transportation network optimization. IEEE Transactions on Industrial Informatics, 14(3), 1140-1149. https://doi.org/10.1109/TII.2017.2748080

[8] Yuan et al. (2017): Yuan, Y., Wang, J., & Liu, S. (2017). Random forest based traffic congestion prediction using multi-source data. IEEE Access, 5, 6022-6031. https://doi.org/10.1109/ACCESS.2017.2690320

[9] Lv et al. (2015): Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). Traffic flow prediction with big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 16(2), 865-873. https://doi.org/10.1109/TITS.2014.2345663

[10] R. Madonna Arieth, Subrata Chowdhury,B. Sundaravadivazhagan,Gautam Srivastava(2024) Traffic Prediction and Congestion Control Using Regression Models in Machine Learning for Cellular Technology. https://doi.org/10.1201/9781003306290-7

[11] Ghani et al. (2020): Ghani, I., Mahmud, M., & Rashid, B. (2020). Traffic congestion forecasting in urban areas using machine learning models. Journal of Intelligent Transportation Systems, 24(3), 260-271.https://doi.org/10.1080/15472450.2020.1718032

[12] Mohd Khairul Amri, K., Roslan, U., Noorjima, A. W., & Ahmad Shakir, M. S. (2018). Road Traffic Accident in Malaysia: Trends, Selected Underlying, Determinants and Status Intervention. International Journal of Engineering & Technology, 7(4), Article 4.

[13] Singh, A. P., Srivastava, A., Jain, A., & Khatter, H. (2024). Road Accident Analysis and Classification System. 2024 2nd International Conference on Disruptive Technologies (ICDT), 27 – 31. https://doi.org/10.1109/ICDT61202.2024.10489515

[14] Ditcharoen, A., Chhour, B., Traikunwaranon, T., Aphivongpanya, N., Maneerat, K., & Ammarapala, V. (2018). Road traffic accidents severity factors: A review paper. 2018 5th International Conference on Business and Industrial Research (ICBIR), 339 – 343. https://doi.org/10.1109/ICBIR.2018.8391218

[15] Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques (No. arXiv:2406.13968). arXiv. https://doi.org/10.48550/arXiv.2406.13968

[16] Banerjee, K., Bali, V., Sharma, A., Aggarwal, D., Yadav, A., Shukla, A., & Srivastav, P. (2022). Traffic Accident Risk Prediction Using Machine Learning. 2022 International Mobile and Embedded Technology Conference (MECON), 76 – 82. https://doi.org/10.1109/MECON53876.2022.9752273

.