PREDICTIVE ANALYSIS OF ROAD ACCIDENT IN THAILAND: APPLICATIONS OF ARTIFICIAL INTELLIGENCE

ITTIRIT MOHAMAD

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy of Engineering in Energy
and Logistics Management Engineering
Suranaree University of Technology
Academic Year 2022

การวิเคราะห์เชิงพยากรณ์อุบัติเหตุทางถนนในประเทศไทย : การประยุกต์ใช้ปัญญาประดิษฐ์

นายอิทธิฤทธิ์ โมหะหมัด

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมการจัดการพลังงานและโลจิสติกส์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2565

PREDICTIVE ANALYSIS OF ROAD ACCIDENT IN THAILAND:
APPLICATIONS OF ARTIFICIAL INTELLIGENCE

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Doctor's Degree.

Thesis Examining Committee

.....

(Prof. Dr. Thaned Satiennam)
Chairperson

.....

(Asst. Prof. Dr. Sajjakaj Jomnonkwao)
Member (Thesis Advisor)

.....

(Prof. Dr. Vatanavongs Ratanavaraha)
Member

.....

(Asst. Prof. Dr. Somsak Siwadamrongpong)
Member

.....

(Asst. Prof Dr. Sarunya Kanjanawattana)
Member

.....

(Assoc. Prof. Dr. Chatchai Jothityangkoon)
Vice Rector for Academic Affairs and
Quality Assurance

.....

(Assoc. Prof. Dr. Pornsiri Jongkol)
Dean of Institute of Engineering

อิทธิฤทธิ์ โมหะหมัด : การวิเคราะห์เชิงพยากรณ์อุบัติเหตุทางถนนในประเทศไทย :การประยุกต์ใช้ปัญญาประดิษฐ์ (PREDICTIVE ANALYSIS OF ROAD ACCIDENT IN THAILAND : APPLICATIONS OF ARTIFICIAL INTELLIGENCE) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.สัจจากาจ จอมโนนเขวา, 120 หน้า.

คำสำคัญ: การทำเหมืองข้อมูล/การเรียนรู้ของเครื่อง/Association Rule/ปัญญาประดิษฐ์/อุบัติเหตุทางถนน/การขนส่ง/โลจิสติกส์/ข้อมูลขนาดใหญ่

อุบัติเหตุทางถนนเป็นความท้าทายระดับโลกที่สร้างปัญหาให้กับอารยธรรมมนุษย์มาช้านานแล้ว โดยเฉพาะอย่างยิ่ง ประเทศในเอเชียตะวันออกเฉียงใต้และแอฟริกา ซึ่งสองภูมิภาคที่กล่าวถึงก่อนหน้านี้ มีจำนวนอุบัติเหตุบนท้องถนนเพิ่มขึ้นอย่างต่อเนื่องอย่างน้อย 10 ปีที่ผ่านมา (พ.ศ. 2551-2561) ประเทศไทยได้รับการจัดอันดับเป็นที่หนึ่งในเอเชียและอันดับที่เก้าของโลกในปี 2561 ตามข้อมูลขององค์การอนามัยโลก คนไทยเสียชีวิตจากอุบัติเหตุทางถนนในอัตรา 32.7 ต่อประชากร 100,000 คน และตัวเลขนี้ก็เพิ่มขึ้นอย่างต่อเนื่อง ในบทความวิจัยนี้ ผู้วิจัยได้นำข้อมูลการเกิดอุบัติเหตุทางถนนจากกรมป้องกันและบรรเทาสาธารณภัยของประเทศไทย ประจำปี 2558-2563 จำนวน 168K เหตุการณ์ รวมค่าเสียหายประมาณ 1.13 พันล้านบาท เพื่อทำการวิเคราะห์ ด้วยเทคนิคปัญญาประดิษฐ์ จากผลของชุดข้อมูลอุบัติเหตุทางถนนที่ใหญ่และซับซ้อน ผู้วิจัยจึงใช้วิธีการวิเคราะห์ข้อมูลการเรียนรู้ของเครื่องเพื่อจัดการกับชุดข้อมูลนี้ การวิเคราะห์ข้อมูลขนาดใหญ่ขั้นสูงเป็นกระแสสำหรับยุค 4.0 ซึ่งสามารถขุดและระบุสัมพัทธภาพของข้อมูลด้วยวิธีที่มีประสิทธิภาพมากขึ้นในการทำความเข้าใจข้อมูลมากกว่าวิธีแบบเก่า

ประโยชน์ของงานวิจัยนี้คือ เราสามารถระบุและคาดการณ์สาเหตุหรือตัวแปรที่ทำให้เกิดอุบัติเหตุบนท้องถนนได้ ซึ่งจะเป็นข้อมูลที่เป็นประโยชน์สำหรับผู้กำหนดนโยบายในประเทศไทยในขณะที่ออกแบบนโยบายความปลอดภัยทางถนนและ/หรือเตรียมการป้องกันและแก้ไขอุบัติเหตุทางถนนใน ระยะสั้น/กลาง/ยาว.

กรณีศึกษาที่ 1 : เลือก Apriori Algorithm Mining เพื่อเชื่อมโยงรายการที่พบบ่อยในชุดข้อมูลทั้งหมด จากนั้นจึงขยายเพื่อค้นหารายการแต่ละรายการที่พบบ่อยที่สุดและขยายไปยังชุดรายการที่มีขนาดใหญ่กว่า ตราบใดที่ชุดเหล่านั้นปรากฏบ่อยเพียงพอในฐานข้อมูล

กรณีศึกษาที่ 2: แผนผังการตัดสินใจช่วยให้สามารถประมวลผลข้อมูลจำนวนมากเพื่อเจาะลึกความสัมพันธ์ระหว่างตัวแปรแต่ละตัวในชุดข้อมูลขนาดใหญ่ ในการศึกษานี้ แผนภูมิต้นไม้ตัดสินใจ (Decision Tree) ยังแยกอุบัติเหตุว่าผู้ขับขี่ ขับเกินความเร็วที่กำหนดทั้งบนทางหลวงและทางชนบทหรือไม่

กรณีศึกษาที่ 3: ความพยายามในการจัดให้มีวิธีการในการเลือกและรวบรวมเกณฑ์ที่มีอิทธิพล และพัฒนาแบบจำลองสำหรับการจัดหมวดหมู่ความรุนแรงของการบาดเจ็บ มีการใช้วิธีการเรียนรู้ของเครื่องต่าง ๆ เพื่อสร้างแบบจำลองเหล่านี้ สำหรับข้อมูลอุบัติเหตุจราจร วิธีการเรียนรู้ของเครื่องในงานวิจัยนี้ประกอบด้วย Decision Tree (DT), Support Vector Machine (SVM), Random Forests (RF), Neural Network (NW), Naive Bayes (NB) Logistic Regression (RG), K-Nearest-Neighbors (kNN) และ Gradient Boosting (GB) ใช้ในการจำแนกประเภทชุดข้อมูลแบบ หนึ่ง หรือ ศูนย์ (เสียชีวิต/ไม่เสียชีวิต) ของอุบัติเหตุทางถนน

สาขาวิชาวิศวกรรมเครื่องกล ลายมือชื่อนักศึกษา.....

ปีการศึกษา 2565 ลายมือชื่ออาจารย์ที่ปรึกษา.....

ITTIRIT MOHAMAD : PREDICTIVE ANALYSIS OF ROAD ACCIDENT IN THAILAND: APPLICATIONS OF ARTIFICIAL INTELLIGENCE. THESIS ADVISOR : ASST. PROF. SAJJAKAJ JOMNONKWAO, Ph.D., 120 PP.

Keywords: DATA MINING/MACHINE LEARNING/ASSOCIATED RULE/AI/ROAD ACCIDENT/ TRANSPORT/LOGISTICS/BIG DATA

Road accidents are a global challenge that has been troubling civilization for a long time. Specifically, the countries in Southeast Asia and Africa, the two previously mentioned regions, have the number of road accidents continuously increasing for at least the last 10 years (2008-2018). Thailand was rated first in Asia and ninth in the world in 2018, according to WHO data. Thais die in road accidents at a rate of 32.7 per 100,000 population, and this figure has been steadily rising. In this research article, the researcher has brought the information about the occurrence of road accidents from the Thailand Department of Public Disaster Prevention and Mitigation in the year 2015-2020 amounting to 168K events with total damages cost around 1.13 billion Thai baht to do an analysis with Machine Learning Technique. As a consequence of the massive and complicated road accident data set, the researcher using a machine learning data analysis approach to deal with this data set. Advanced big data analytics has been the trend for the 4.0 era, which can mine and identify the relativity of data in a more effective method to understand the data than the old fashion way.

The benefit of this research is that we can identify and predict the reasons or variables that cause road accidents, which will be useful information for policymakers in Thailand as they design road safety policies and/or prepare for preventive and corrective actions for Road accidents in the short/Med/long term.

Case Study no.1: Apriori algorithm mining was chosen to relate frequent items over entire data set. It was then expanded to discover the most common individual items and extend them to larger itemset, so long as those sets appeared frequently enough in the database.

Case Study no. 2: Decision trees allow for processing large amounts of data to drill down into associations between the individual variables in a large data set; in this

study, the tree also separated accidents into whether the driver was exceeding the speed limit on both highway and rural ways.

Case Study no. 3: This inquiry attempt provides methods for selecting a collection of influential criteria and developing a model for categorizing the severity of injuries. Various machine learning approaches are used to create these models. On traffic accident data, supervised machine learning methods such as, Decision Tree (DT), Support Vector Machine (SVM), Random Forests (RF), Neural Network (NW), Naive Bayes (NB) Logistic Regression (RG), K-Nearest-Neighbors (kNN), and Gradient Boosting (GB) are used in binary (Fatal/Nonfatal) classification of occupational accidents.

School of Mechanical Engineering

Academic Year 2022

Student's Signature.....

Advisor's Signature.....

ACKNOWLEDGEMENT

First and foremost, I'd like to thank my brain for constantly trying to get me to do things unrelated to the thesis in order to avoid the stress of writing paper, even though that it is far too much, as well as my hands and arms for always being by my side, even so, they are not always in sync with my brain. My legs for always stand for me even sometime want to lie down on the floor after several rejections from journals, and my heart for never tired of hearing feedback from Journal reviewers. I'd like to thank Assist Professor Dr. Sajjakaj JomnonKwao for his advice and guidance throughout the research process. I'm also thankful to Professor Dr. Vatanavong Ratanavaraha for advising me on how to publish a paper in a prestigious journal. Furthermore, I'd like to express my appreciation to all professors who teaches me with my coursework including random Indian programmers who created the tutorial VDO on YouTube for coding and Machine learning issue related until I was able to successfully complete my research. Finally, I'd like to thank all my colleagues in the Energy and Logistics Management Engineering program, Mechanical Engineering School, as well as my family members, for their constant encouragement and support.

Ittirit Mohamad

TABLE OF CONTENTS

Page

ABSTRACT (ENGLISH) I
ABSTRACT (THAI) III
ACKNOWLEDGEMENT V
TABLE OF CONTENTS VI
LIST OF TABLES VIII
LIST OF FIGURES IX
LIST OF ABBREVIATIONS X
CHAPTER
1 INTRODUCTION 1
1.1 The significance of the research question 1
1.2 The research objective: 1
1.3 Contribute of the research 2
1.4 Justifications for conducting research in this population 2
2 PREDICTIVE ANALYSIS OF A HIGHWAY ROAD ACCIDENT IN THAILAND: USING MACHINE LEARNING APPROACH 4
2.1 Abstract 4
2.2 Introduction 5
2.3 Data Description and Methodology 15
2.4 Descriptive Statistics and Result 21
2.5 Conclusion and Discussion 32
2.6 Study limitation and future study 34
2.7 Reference 34
3 USING A DECISION TREE TO COMPARE RURAL VERSUS HIGHWAY MOTORCYCLE FATALITIES 44
3.1 Abstract 44

TABLE OF CONTENTS (Continued)

Page

3.2 Introduction 45
3.3 Literature Review 47
3.4 Methodology 50
3.5 Results 57
3.6 Conclusion and Discussion 65
3.7 Limitations and Future Studies 67
3.8 References 68
4 COMPARISON OF MACHINE LEARNING PREDICTABILITY PERFORMANCE: THE CASE OF MOTORCYCLE ACCIDENT IN THAILAND 74
4.1 Abstract 74
4.2 Introduction 75
4.3 Literature Review 78
4.4 Methodology and Data 81
4.5 Results 89
4.6 Conclusion & Discussion 95
4.7 Limitations and Future Studies 97
4.8 Reference 98
5 CONCLUSION AND RECOMMENDATION 105
APPENDIX A LIST OF PUBLICATIONS 106
BIOGRAPHY 120

LIST OF TABLES

Table **Page**

2.1 Road accident using data mining and Machine learning 7
2.2 Previous research has identified the factors that determine the severity of driving injuries..... 11
2.3 the driver who was the caused in those accident divided by highway vs Non highway..... 16
2.4 Total 34 Attribute with setting description..... 16
2.5 Focusing Rule with high lift and widely gap between support and confidence..... 28
3.1 The Machine Learning Models Used in Extant Traffic Accident Studies..... 50
3.2 The Categorical Variables and Their Descriptive Statistics 52
3.3 The Measurement Categories for the 27 Identified Motorcycle Accident Variables..... 54
3.4 The Final HW and RR Sets by Rider Speed..... 63
3.5 The Model Evaluation Results 64
4.1 Comparison of the advantages and disadvantages of ML models..... 76
4.2 Machine learning models in traffic accident study..... 78
4.3 Categorical Attribute and descriptive statistics..... 82
4.4 Total 28 Attributes with setting description 86
4.5 Info. Gain Ranking by model..... 89
4.6 evaluation result from models 91
4.7 Confusion Metrix for each model..... 94

LIST OF FIGURES

Figures **Page**

2.1 Highway accident stacked column chart by year 6
2.2 Data analysis process step 15
2.3 Associate Rules Mining Diagram 19
2.4 Highway accident distribution plot by 24-hour time series w/ Kernel density as line chart 22
2.5 Frequency itemset extraction 23
2.6 Associate Rules Mining total 1558 rules 24
2.7 and 2.8 Support and Confidence distribution from 1,558 rules discovered 25
2.8 1,558 discovered rules with scatter plot Support VS Confidence 26
2.9 Dendrogram for 1,558 rules discovered on Antecedent 27
2.10 Confidence and support chart gap trend chart by interesting rules 31
3.1 Total number of vehicles and motorcycles registered in Thailand from 2015 to 2020 45
3.2 The steps in the process for the study. 51
3.3 Diagram of the confusion matrix 57
3.4 HW and RR fatality probabilities at different times of the day. 58
3.5 The HW tree model 59
3.6 The RR tree model. 60
3.7 The confusion matrix actual and predicted results for HWs. 65
3.8 The confusion matrix actual and predicted results for RRs. 65
3.9 Key accident factors: HWs versus RRs 66
4.1 Machine learning Process flow 81
4.2 Confusion matrix diagram 88
4.3 Performance Measurement models 92
4.4 Model-specific ROC plot for predicting non-fatality 93

LIST OF ABBREVIATIONS

LHS	=	Left hand Side
RHS	=	Right Hand Side
HW	=	Highway
RR	=	Rural Road
TP	=	True Positive
FP	=	False Positive
TN	=	True Negative
FN	=	False Negative
DT	=	Decision Tree
SVM	=	Support Vector Machine
RF	=	Random Forest
kNN	=	K-Nearest-Neighbors
NN	=	Neural Network
LR	=	Logistic Regression
GB	=	Gradient Boosting
AUC	=	Area Under Curve
CA	=	Classification Accuracy

CHAPTER 1

INTRODUCTION

1.1 The significance of the research question

The research question is What factors contribute to road accidents? How many variables are present in an accident (Swiss Cheese Theory)? When these factors come together, it is possible that accidents will be significantly higher or more severe since the number of road accidents in Thailand is increasing. In 2018, the WHO reported global road deaths by country. Thailand has surpassed all other countries to take the top spot in the world. Thailand has a death rate of 32.7 people per 100,000 people. Only Thailand is from Southeast Asia, and Iran is another Asian country. Various agencies have attempted to reduce the number of fatalities in the past following the Decade of Action for Road Safety 2011-2020 (WHO, 2011), which aims to reduce fatalities. The trend of accidents in Thailand continues to rise, with less than 10 road accidents per 100,000 people in 2020. Accidents in Thailand have been on the rise for some time that we can comprehend the issues and common factors It could be one way to reduce the number of accidents in the future.

1.2 The research objective:

The goal of this study was to determine the cause or co-incidence of the most common road accidents that result in fatalities. The relationships between three factors related to road accidents were investigated using machine learning techniques: people, vehicles, roads, and the environment. This information can be used to develop policies to reduce the number of road accidents. Economic and human resources will be saved, and the overall efficiency of the country's healthcare system will be improved.

1.3 Contribute of the research

The contribute of this research is to find ways to reduce the number of accidents and fatalities caused by accidents. The model's results describe the factors and their degree of association with the risk of accidents and deaths on a large database, using artificial intelligence and machine learning principles for maximum efficiency, this is the application of modern knowledge to the existing database. Knowing the factors allows you to propose policies to reduce the number of accidents and deaths which will increase quality of life and overall health care system in the country.

1.4 Justifications for conducting research in this population

This population consists of Thai road accident victims. When considering Thailand's roads and the number of accidents, it was discovered that Thailand's roads have the highest number of accident deaths. With physical road characteristics that are designed to handle high traffic volumes. and is capable of high speed When there is an accident, the severity of the injury is also high. In Thailand, many highways are shared by vehicles of various sizes. As a result, the likelihood of death increases when an accident occurs.

According to the Thai accident research center's (TARC) study, the person/driver factor There is an 83 percent chance of causing an accident, with 36 percent being caused by a specific person. As a result, when considering the personal characteristics of this group of volunteers, to investigate the relationship between personal characteristics such as gender, age, and vehicle type. The surrounding environment and vehicles will form a link between the factors of people, vehicles, and the environment that influence the number of accidents. And the severity of the injury, etc. The information does not include specifics such as the person's or driver's name or ID card number. The data set cannot be traced back to the individual. Secondary data to be examined It is information that is freely available to the public. The data source (Department of Disaster Prevention and Mitigation) only has publicly available information only gender, age, province, type of accident, environment, and road condition are not disclosed.

Suranaree University of Technology's human research ethics committee has exempted this research, which will be carried out in compliance with international guidelines for human research protection such as the Helsinki Declaration, The Belmont Report, CIOMS guideline, International Conference on Harmonization in Good Clinical Practice (ICH-GCP), and 45 CFR 46.101(b) as project code EC-65-0013 (criteria of exemption: secondary data).

CHAPTER 2

PREDICTIVE ANALYSIS OF A HIGHWAY ROAD ACCIDENT IN THAILAND: USING MACHINE LEARNING APPROACH

2.1 Abstract

Accidents are a major obstacle to economic development and quality of life in developing countries. The same challenges are perceived today as major issues in Thailand. This research aims to assess the frequency and most common causes of road accidents that are most likely to result in fatalities. Machine learning technique is employed to examine the relation of factors in accidents, which are then applied to policymaking to lower the rate of road accidents, economic and human resource losses, as well as improve the overall efficiency of a country's healthcare system. The researcher has included information of road accidents in Thailand during the years 2015–2020; a total of 167,820 events, with total damages costing some 1.13 billion Thai baht (34 million USD). Although the overall data comprises the elements influencing the accidents, this article only considers the drivers who were the causes of fatal highway accidents. As a result, the factors that enhance the likelihood of fatality in highway road accidents are as follows: driver info, male; driver behavior, over speed limit; vehicle type, motorbike; roadway, straight, dry surface; and weather, clear. All these variables are related, as the association rule shows an increased risk of injury or death in traffic accidents.

2.1.1 Highlights:

1) Driver risk perception was discovered to have the strongest influence on road accidents.

2) The factors that enhance the likelihood of fatality in highway road accidents are as follows: driver info, male; driver behavior, over speed limit; vehicle type, motorbike; roadway, straight, dry surface; and weather, clear.

3) Most accidents occur during daytime (08.00–18.00), while peaks occur at 19.00–20.00 and 22.00–23.00 and high fatality rate at night (19.00–07.00).

4) The higher the number of elements involved, the greater the possibility of an accident.

2.2 Introduction

Road traffic accidents are a worldwide issue that have been troubling civilization for a long time. Specifically, road accidents in Southeast Asia and Africa, the two previously mentioned regions, have been continuously increasing for at least the last 10 years (2008–2018) WHO (2018). According to WHO data in 2018, Thailand was ranked No. 1 for road accidents in Asia and No. 9 in the world. An average of 32.7 Thais per 100,000 population die in road accidents every year (WHO, 2018). Not only has it caused an economic upheaval, but it has impacted the country's public health system. Road accidents have also caused the country's limited resources to be used in ways harmful to its progress. It also negatively impacts the country's human resources, resulting in the death or disability of its residents.

In Thailand, examples of road safety policies include law enforcement (e.g., for exceeding speed limits or the consumption of alcohol), road safety programs in educational institutions, the development of advertising media, an increase in the number of training hours required to obtain new drivers' licenses and their renewals, engineering solution techniques for road safety audits, and research funding. To establish these regulations, predicted data on the number of accidents was used to determine operational budgets (Jomnonkwao et al., 2020). However, the average number of roadway fatalities in Thailand from 2015 to 2020 remained consistent at 32%–35% for the fifth year in a row, as shown in Fig 2.1. The existing policy appears to be ineffective. Learning from every element recorded in the big data set and starting to predict and minimize things before they occur might be the way out.

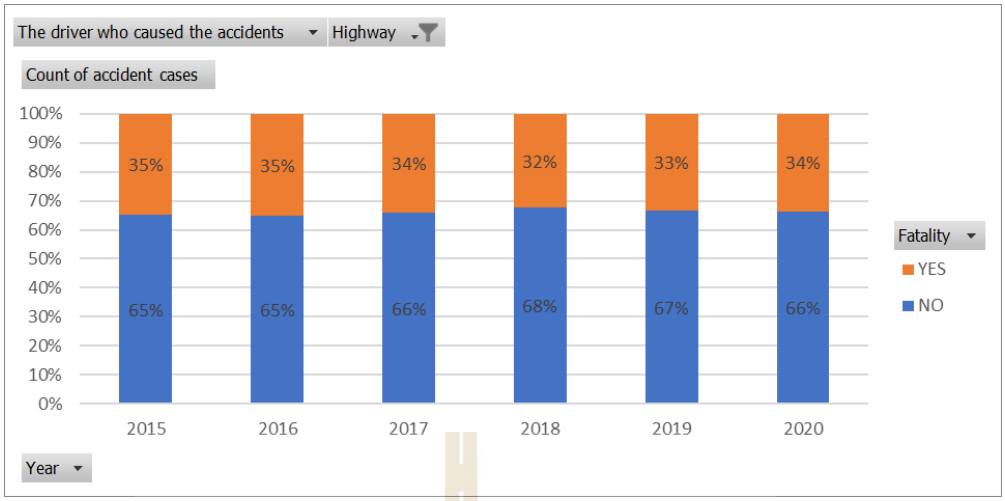


Figure 2.1 Highway accident stacked column chart by year.

Previous studies have utilized machine learning algorithms to predict injury severity. Some focus on independent factors like the environment, drivers, current weather, or road conditions, even comparing performance models, as shown in Table 2.1. However, these studies did not consider the events' coincidence for the drivers who were killed. The coincidence being discussed included type of roadway, vehicle type, external factors such as environment and weather conditions, and internal factors, e.g., driver behaviors and information, like gender and age, to understand which factors interfered with each other or any linkage between them that increased the chances of fatality. According to the Swiss cheese theory, if all the holes (factors) are aligned by chance, the accident will happen and result in death. In contrast, the risk may be decreased by controlling the primary element that has the strongest influence on fatality. For example, the researcher noted that accidents are typically caused by a combination of circumstances rather than by one or two factor(s). And, if the elements were combined, how likely is it that someone would die? However, what happens if the risk factor is reduced? That is why forecasts appear to simulate the situation. However, predicting the accident event is also essential for establishing road safety, budgeting, staffing, and policy planning.

Table 2.1 Road accident using data mining and Machine learning.

Author	Methodology														
	Apriori Algorithm	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Deep Learning	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic Regression	Neural Network	Naïve Bayes	Random Forest	Regression on python	Support Vector Machine
Sonal and Suman (2018)	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
Gutierrez-Osorio and Pedraza (2020)	-	-	-	-	-	✓	-	-	-	-	✓	-	-	-	-
Abellán et al. (2013)	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-
Al Mamlook et al. (2019)	-	-	✓	✓	✓	-	-	-	✓	-	-	✓	✓	-	✓
Mafi et al. (2018)	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Recal and Demirel (2021)	-	-	-	-	✓	-	✓	-	-	✓	✓	-	-	-	✓
Bahiru et al. (2018)	-	-	-	-	✓	-	-	-	-	-	-	✓	-	-	-

Table 2.1 Road accident using data mining and Machine learning (Continued)

Author	Methodology														
	Apriori Algorithm	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Deep Learning	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic Regression	Neural Network	Naïve Bayes	Random Forest	Regression on python	Support Vector Machine
Cuenca et al. (2018)	-	-	-	-	-	✓	✓	-	-	-	-	✓	-	-	-
Kuşkapan et al. (2021)	-	-	-	-	-	-	-	-	✓	-	-	✓	-	-	✓
Ospina-Mateus et al. (2021)	-	-	-	-	✓	-	-	-	✓	-	✓	✓	✓	-	✓
Kumar and Toshniwal (2016)	-	✓	-	-	-	-	-	✓	-	-	-	-	-	-	-
Helen et al. (2019)	-	✓	-	-	-	-	-	✓	-	-	-	-	-	-	-
El Abdallaoui et al. (2018)	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
John and Shaiba (2019)	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2.1 Road accident using data mining and Machine learning (Continued)

Author	Methodology														
	Apriori Algorithm	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Deep Learning	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic Regression	Neural Network	Naïve Bayes	Random Forest	Regression on python	Support Vector Machine
Feng et al. (2020)	-	✓	-	-	-	-	-	-	-	-	✓	-	-	-	-
Bhavsar et al. (2021)	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
Samerei et al. (2021)	-	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-
John and Shaiba (2022)	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Earlier research on road traffic accidents have also been categorized by variables in the form that are presumed to be associated in every accident, according to international research.

Age – Zhang and Fan (2013) found that accidents are more likely to occur among junior drivers (≤25 yrs.) who have a lack of discipline, are inexperienced with traffic regulations, as well as having less driving experience. Most traffic accidents in Dubai are caused by a lack of space between vehicles, with youth (≤35 yrs.) being the most usually involved; the peak hour(s) are late at night, and the overwhelming majority of drivers were discovered to be inebriated. (John & Shaiba, 2019). Young (18–24 years old) drivers lack experience at controlling speeding or adjusting well while driving (Bucsuházy et al., 2020). John and Shaiba (2022) found that most alcohol-involved accidents are caused by youths (≤35 yrs.) late at night.

Gender – Ospina-Mateus et al. (2019) and Mohamad et al. (2022) observed that men are more likely to be involved in serious accidents than women.

Driver behaviors – When compared to other drivers, intoxicated drivers have a higher accident rate (Helen et al., 2019). The most important aspect in predicting the severity of an injury is its driving over speed limit (Al Mamlook et al., 2019).

Driver – Drivers are more likely to be injured or killed in accidents than other passengers (El Abdallaoui et al., 2018).

Time – Traveling at night increases the chances of car accidents (Mphela, 2020).

Road and light conditions – Chen et al. (2016) observed that road slope and visibility were predictors of driver injuries. Highway intersections are riskier for all accident types. Poor road conditions increase the likelihood of accidents, especially on motorways (Malin et al., 2019). Road type, lighting, speed limits, and road surface all play key roles in accident incidence (Feng et al., 2020). Most fatal injuries occur as a result of aggressive driving, inattentiveness, and speeding. However, compared to other situations, dark or dim roads also played significant roles (Shweta et al., 2021).

Weather conditions – (Kumar & Toshniwal, 2016) Sonal and Suman (2018) observed that external factors, like weather conditions such as fog, rain, and snow, have greater impacts on road accidents than internal factors, such as the driver.

Type of vehicles – Chen et al. (2015) mentioned this factor as significant for driver injuries and fatalities in rear-end accidents involving trucks, lighting, wind, and multiple vehicles involved. The analysis revealed that the most essential and impactful traffic accident elements are speed limit, weather conditions, number of lanes, lighting conditions, and accident timing, while gender, age, accident location, and vehicle type have less of an impact on severity (Bahiru et al., 2018)

The researchers are continuing to evaluate the literature on road accidents and the factors involved. It will cover a wide range of research from across the world, but Table 2.2 will concentrate on research from the same region as this study.

Table 2.2 Previous research has identified the factors that determine the severity of driving injuries.

Variables	Finding
Driver Characteristics	
Gender	Decrease injury-severity: male. (Xie & Huynh, 2012), (Behnood & Mannering, 2017), (Li, Wu, et al., 2019a), Increase injury-severity: female (Wu et al., 2016), (Osman et al., 2018), (Behnood & Mannering, 2017) (Hou et al., 2019), Male (Kim et al., 2013), (Li et al., 2018), (Champahom et al., 2020)

Table 2.2 Previous research has identified the factors that determine the severity of driving injuries (Continued)

Variables	Finding
Age	Decrease injury-severity: less than 25. (Behnood & Mannering, 2017), (Li, Ci, et al., 2019) Increase injury-severity: Less than 25 (Li et al., 2018) more than 65 (Kim et al., 2013), (Wu et al., 2016), (Li, Wu, et al., 2019b), (Zhou & Chin, 2019), (Hou et al., 2019), (Wei et al., 2021) (Champahom et al., 2020),
Speeding	Increase injury-severity: speeding vehicle. (Kim et al., 2013), (Osman et al., 2018), (Krull et al., 2000), (Xie & Huynh, 2012), (M. Yu et al., 2020)
Drunk	Increase injury-severity: drunk driving. (Krull et al., 2000), (Xie & Huynh, 2012), (Kim et al., 2013), (Wu et al., 2016), (Zhou & Chin, 2019), (John & Shaiba, 2019), (Helen et al., 2019) ,(Champahom et al., 2020)
Fatigue	Increase injury-severity: Doze off. (Champahom et al., 2020)
Overtaking	Increase injury-severity: improper overtaking. (Jafari Anarkooli et al., 2017), (Li, Wu, et al., 2019a)

Table 2.2 Previous research has identified the factors that determine the severity of driving injuries (Continued)

Variables	Finding
Vehicle characteristics	
Vehicle type	Decrease injury-severity: SUV/van (Chamroeun Se et al., 2021) Pick-up truck (Wu et al., 2016), (Chamroeun Se et al., 2021) passenger car (Huo et al., 2020) Increase injury-severity: rollover SUV/van (Jafari Anarkooli et al., 2017) large truck (Jafari Anarkooli et al., 2017), (Li et al., 2018), (Huo et al., 2020) Pickup (Li et al., 2018),
External Factor (Environment and road condition)	
Light status	Decrease injury-severity: darkness without light. (Xie & Huynh, 2012), Increase injury-severity: daylight. (Krull et al., 2000) darkness without light (Kim et al., 2013), (Jafari Anarkooli et al., 2017) (Zhou & Chin, 2019)

Table 2.2 Previous research has identified the factors that determine the severity of driving injuries (Continued)

Variables	Finding
	after midnight (Zhou & Chin, 2019) Nighttime (Mphela, 2020), (Osman et al., 2018)
Dry/wet road surface	Decrease injury-severity: wet road. (Zhou & Chin, 2019), (H. Yu et al., 2020) Increase injury-severity: Wet road (Li, Wu, et al., 2019a), (Li et al., 2018) dry road (Krull et al., 2000)
Weather	Decrease injury-severity: raining. (Jung et al., 2010) Increase injury-severity: raining. (Shweta et al., 2021), (Jafari Anarkooli et al., 2017), (Li, Wu, et al., 2019a) Fog, Rainfall, Snowfall (Shweta et al., 2021),
Time	Increase injury-severity: Daytime. (Shaheed et al., 2013) Nighttime (Champahom et al., 2020), (Chamroeun Se et al., 2021)

2.3 Data Description and Methodology

2.3.1 Data Description

The occurrence of road accidents from the Thailand government organization during the years 2015–2020 amounted to 167,820 events PDPM (2020). This study focuses on drivers who caused their accidents. Those came to 129,015 total, of which 95,249 were nonfatal and 33,766 fatal (24,559 for highway and 9,207 for nonhighway). Using the data analysis technique to execute the following steps in Fig. 2.2.

Data cleaning – missing and incompletely captured data detection and correction.

Data validation – validation the quality of the data after the data set has been cleansed.

Data converting – data partitioning to binary mode.

Data analysis and interpretation – discovering the data for informing conclusion.

Data visualization – creating a visual to represent information and data.

Figure 2.2 Data analysis process step

The data in Table 2.3 is classified into four different categories, consisting of fatalities (HW & NHW) and nonfatalities (HW&NHW) to find the link between those and the type of roads. However, this study focuses on highway fatalities. The authors converted the total data to binary to represent Yes or No in each accident event and fed it through Python base software. Table 3 presented data divided by road type and fatality. The large number, 24,599, drew our attention and encouraged us to investigate.

Table 2.3 the driver who was the caused in those accident divided by highway vs Non highway.

Count of Road Accident Case	Fatality		
Road Type	No	Yes	Grand Total
Non-Highway	47,136	9,207	56,343
Highway	48,113	24,559*	72,672
Grand Total	95,249	33,766	129,015

In every event, aspects of 34 attributes from accident data collection appeared, including roadway, vehicle type, environment, weather conditions, driver behavior, driver info, and driver status in Table 2.4.

Table 2.4 Total 34 Attribute with setting description

Attribute Name	Attribute Description
Roadway	
Highway	1 - Yes
Dry Surface Road	1 - Yes, 0-Otherwise
Straight Way	1 - Yes, 0-Otherwise
Obstruction	1 - Yes, 0-Otherwise
Road condition	1 - Yes, 0-Otherwise
Vehicle condition	1 - Yes, 0-Otherwise

Table 2.4 Total 34 Attribute with setting description (Continued)

Attribute Name	Attribute Description
Vehicle Type	
Motorcycle	1 - Yes, 0-Otherwise
Mini truck/ Pick up (4 wheels)	1 - Yes, 0-Otherwise
Sedan	1 - Yes, 0-Otherwise
Light Truck (6 wheels)	1 - Yes, 0-Otherwise
Heavy Truck (10+ wheels)	1 - Yes, 0-Otherwise
Other Type of car	1 - Yes, 0-Otherwise
External Factor (Environment and Weather Condition)	
Day Time (06.00-18.00)	1 - Yes, 0-Otherwise
Night with Light	1 - Yes, 0-Otherwise
Night without Light	1 - Yes, 0-Otherwise
Low visibility	1 - Yes, 0-Otherwise
Clear Weather	1 - Yes, 0-Otherwise
Internal Factor (Driver Behavior)	
Drunk	1 - Yes, 0-Otherwise
Over Speed limit	1 - Yes, 0-Otherwise
Break Through Traffic lights	1 - Yes, 0-Otherwise
Break Through Traffic Signs	1 - Yes, 0-Otherwise

Table 2.4 Total 34 Attribute with setting description (Continued)

Attribute Name	Attribute Description
Overtake	1 - Yes, 0-Otherwise
Use Mobile Phone	1 - Yes, 0-Otherwise
Short Cut off	1 - Yes, 0-Otherwise
Drug	1 - Yes, 0-Otherwise
Drive in opposite direction	1 - Yes, 0-Otherwise
Doze off	1 - Yes, 0-Otherwise
Overweight Carry	1 - Yes, 0-Otherwise
Cannot Conclude	1 - Yes, 0-Otherwise
Driver info	
Gender	1- Male, 0-Otherwise
Youth 15-35	1 - Yes, 0-Otherwise
Adult 36-60	1 - Yes, 0-Otherwise
Senior 61-90+	1 - Yes, 0-Otherwise
Driver Status	
Fatality (Death)	1 - Yes

2.3.2 Methodology

Apriori algorithm (Srikant, 1994) was picked to mine for frequent items set over the entire massive relational data set to discover the most common individual items and extend them to larger itemset as long as the sets appeared frequently

enough in the database. Apriori's frequent itemset can be used to generate association rules that highlight overall trends.

Association rule learning is a *rule-based, machine learning* method for discovering key relations between variables in large databases. It is intended to identify strong rules using various measures of attraction (William J. Frawley, 1992). To detect correlations and co-occurrences between data sets, association rules are utilized. They are best suited for explaining data patterns from among seemingly unrelated information sources, such as relational and transactional databases. The act of employing association rules is known as *association rule mining*, or *mining associations*. See Fig. 2.3:

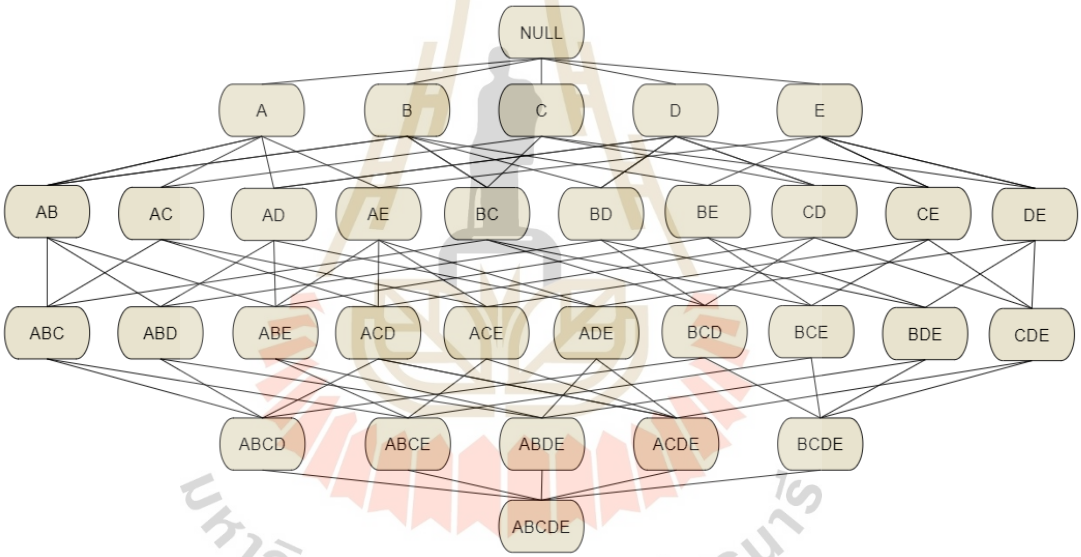


Figure 2.3 Associate Rules Mining Diagram

Rule definition and measurement

An association rule is determined by two factors: support and confidence. The frequency with which a specific rule appears in the database being mined is referred to as *support*. The number of times a particular rule turns out to be true in practice is referred to as a *confidence*.

Let *I* = {...} represent a collection of "n" binary characteristics known as items.

Let *J* = {...} be a set of transactions referred to as a database.

Each transaction in J has a distinct transaction ID and includes a subset of the items in I. A rule is defined as an implication of the type XY in which *X*, *Y* ⊆ I if and only *X* ≠ ∅, *Y* ≠ ∅, *X* ⋂ *Y* = ∅. The sets of objects *X* and *Y* are referred to as the rule's antecedent and consequent, respectively.

Support is an indicator of how frequently the itemset appears in the data set.

$$\text{Support } (x) = \frac{Frequent\ item(x)}{N(Total\ Number\ of\ transaction)}$$

Confidence is an indication of how often the rule has been found to be true.

$$Confidence\ [LHS(x) \Rightarrow\ RHS(y)] = \frac{Support\ (LHS, RHS)}{Support\ (LHS)}$$

The ratio of the observed support to the support expected if X and Y were independent.

$$Lift\ [LHS(x) \Rightarrow\ RHS(y)] = \frac{Support\ (LHS, RHS)}{Support\ (LHS) \times Support\ (RHS)}$$

A rule may have a significant association in a data collection because it frequently appears, but it may occur considerably less frequently when implemented. This would be an example of strong support but low confidence.

Step to perform associated rule mining.

1. Sequence the transaction accident by event (binary) – If minimum support, measure the effectiveness of the accident. If >50% (threshold), then others below 50% will be removed.

1.1 Use frequency itemset from 1 to build item new itemset (length 2). Using join command, if all are set, the sequencing does not matter.

1.2 Recalculate the support score, using transaction in 1.1 to intersection such as

Transaction {Road wet} = {1,1,1,0,1, 0...}

Transaction {Darkness} = {1,1,1,1,0,0...}

Transaction {Road wet, Darkness} = {1,1,1,0,0,0...}

If minimum support < threshold will get removed

1.3 Use frequency itemset from 1.2 to create item new itemset (length 3). However, remember that the initial item must be the same (using the join command), and only one linkage can join:

Transaction {Road wet, Darkness} = {1,1,1,0,0,0...}

Transaction {Road wet, Drunk} = {1,1,1,0,1,0...}

Transaction {Road wet, Darkness, Drunk} = {1,1,1,0,0,0...}

1.4 Frequency all Itemset

2. Consider the following two items or more and then calculate for confidence and lift

2.4 Descriptive Statistics and Result

To comprehend the data pattern and how data distribution works, a distribution chart was created using 72,672 highway accident incidents over 24-h fitted with kernel density as a time series as descriptive statistics shown in Fig. 2.4. To determine a difference between day and night:

1 – Representing fatalities from highway accidents; μ =13.19, σ = 7.03

0 – Representing nonfatalities from highway accidents; μ =13.57, σ =6.37

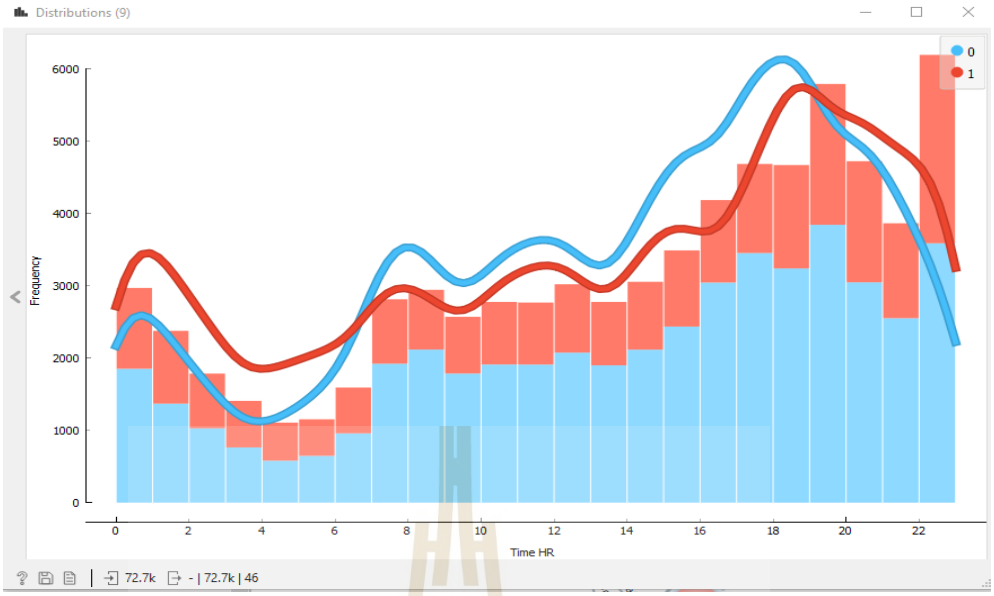


Figure 2.4 Highway accident distribution plot by 24-hour time series w/ Kernel density as line chart

Most accidents occur during daytime (08.00–18.00), while peaks occur at 19.00–20.00 and 22.00–23.00 and high fatality rate at night (19.00–07.00).

Later, they started to frequent items set on fatality as a precondition for the extraction of rules emphasizing causal linkages (Fig. 2.5). Knowing which elements occur together aids in identifying the linkages between them (minimum support at 50%). According to Fig. 5, the most often discovered itemset in the 2018 set is connected to the item found: dry road (95.98%), clear weather (87.33%), male (86.42%), motorcycle (80.77%), straightaway (71.99%), and over speed limit, (69.03%), respectively.

Frequent Itemsets (2)

Itemsets	Support	%
Dry Surface Road=1	23572	95.98
Cannot Conclude =0	22709	92.47
Break Through Traffic lights=0	23364	95.13
Break Through Traffic Signs=0	23238	94.62
Drive in opposite direction=0	23188	94.42
Overtake=0	22917	93.31
Use Mobile Phone=0	23548	95.88
Drug=0	23569	95.97
Doze off=0	22724	92.53
Overweight Carry=0	23542	95.86
Obstruction=0	23190	94.43
Vehicle condition=0	23329	94.99
Road condition=0	23340	95.04
Light Truck =0	23335	95.02
Heavy Truck =0	23409	95.32
Other Type of car=0	23232	94.6
Sedan=0	22037	89.73
Low visibility=0	21652	88.16
Mini truck/ Pick up =0	21267	86.6
Drunk=0	20805	84.71
Clear Weather=1	21308	86.76
61-90=0	20480	83.39
Gender=1	20364	82.92
Short Cut off=0	19025	77.47
Motorcycle=1	18992	77.33
Night without Light=0	18299	74.51
Straight Way=1	17078	69.54
Night with Light=0	16920	68.9
Over Speed limit=1	16369	66.65
36-60=0	14475	58.94
Clear Weather=1	21448	87.33
Gender=1	21224	86.42
Drug=0	21220	86.4
Use Mobile Phone=0	21201	86.33
Overweight Carry=0	21195	86.3
Cannot Conclude =0	20458	83.3
Break Through Traffic lights=0	21046	85.7
Drive in opposite direction=0	20863	84.95
Break Through Traffic Signs=0	20949	85.3
Vehicle condition=0	21005	85.53
Obstruction=0	20875	85
Overtake=0	20639	84.04
Doze off=0	20454	83.29
Road condition=0	20955	85.33
Drunk=0	18569	75.61
Low visibility=0	19253	78.39
Short Cut off=0	17409	70.89
Motorcycle=1	19665	80.07
61-90=0	17005	69.24
Night without Light=0	16183	65.89
Straight Way=1	17679	71.99
Night with Light=0	14944	60.85
Over Speed limit=1	16952	69.03
36-60=0	12554	51.12

24.6k -

Figure 2.5 Frequency itemset extraction

After frequent itemset, the first result came from a highway with 24,559 fatalities. The association rule discovered 1,558 rules (lift ≥ 1 containing 1,377 rules), all of which had been configured to obey the threshold (support 50%, confidence 95%) using Orange 3.30 software (Demšar et al., 2013) (Fig. 2.6). The support distribution (Fig. 2.7) has μ = 0.680263, $\boldsymbol{\sigma}$ = 0.0954974, while confidence distribution (Fig. 2.8) has μ = 0.972597, $\boldsymbol{\sigma}$ = 0.0126851.

Association Rule for Fatality on HW

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.551	0.966	0.570	1.404	1.206	0.094	Over Speed limit=1, Mini truck/ Pick up =0, Sedan=0	→	Motorcycle=1
0.577	0.964	0.599	1.337	1.203	0.098	Straight Way=1, Mini truck/ Pick up =0, Sedan=0	→	Motorcycle=1
0.773	0.962	0.803	0.997	1.202	0.130	Dry Surface Road=1, Mini truck/ Pick up =0, Sedan=0	→	Motorcycle=1
0.704	0.962	0.732	1.093	1.201	0.118	Clear Weather=1, Mini truck/ Pick up =0, Sedan=0	→	Motorcycle=1
0.689	0.958	0.719	1.113	1.196	0.113	Gender=1, Mini truck/ Pick up =0, Sedan=0	→	Motorcycle=1
0.548	0.995	0.551	1.742	1.037	0.019	Clear Weather=1, Over Speed limit=1, Mini truck/ Pick up =0	→	Dry Surface Road=1
0.527	0.995	0.529	1.813	1.037	0.019	Clear Weather=1, Road condition=0, 36-60=0	→	Dry Surface Road=1
0.514	0.995	0.517	1.856	1.037	0.018	Clear Weather=1, Doze off=0, 36-60=0	→	Dry Surface Road=1
0.566	0.995	0.569	1.686	1.037	0.020	Clear Weather=1, Over Speed limit=1, Sedan=0	→	Dry Surface Road=1
0.516	0.995	0.519	1.849	1.036	0.018	Clear Weather=1, Overtake=0, 36-60=0	→	Dry Surface Road=1
0.523	0.995	0.526	1.824	1.036	0.018	Clear Weather=1, Obstruction=0, 36-60=0	→	Dry Surface Road=1
0.529	0.995	0.532	1.805	1.036	0.019	Clear Weather=1, 36-60=0, Heavy Truck =0	→	Dry Surface Road=1
0.527	0.995	0.530	1.812	1.036	0.019	Clear Weather=1, Break Through Traffic lights=0, 36-60=0	→	Dry Surface Road=1
0.526	0.995	0.529	1.816	1.036	0.018	Clear Weather=1, Vehicle condition=0, 36-60=0	→	Dry Surface Road=1
0.531	0.995	0.534	1.797	1.036	0.019	Clear Weather=1, 36-60=0	→	Dry Surface Road=1
0.531	0.995	0.534	1.797	1.036	0.019	Clear Weather=1, Drug=0, 36-60=0	→	Dry Surface Road=1
0.531	0.995	0.533	1.799	1.036	0.019	Clear Weather=1, Use Mobile Phone=0, 36-60=0	→	Dry Surface Road=1
0.531	0.995	0.533	1.799	1.036	0.019	Clear Weather=1, Overweight Carry=0, 36-60=0	→	Dry Surface Road=1
0.513	0.995	0.516	1.859	1.036	0.018	Clear Weather=1, Cannot Conclude =0, 36-60=0	→	Dry Surface Road=1
0.527	0.995	0.530	1.810	1.036	0.019	Clear Weather=1, 36-60=0, Light Truck =0	→	Dry Surface Road=1
0.525	0.995	0.528	1.818	1.036	0.018	Clear Weather=1, 36-60=0, Other Type of car=0	→	Dry Surface Road=1
0.524	0.995	0.527	1.821	1.036	0.018	Clear Weather=1, Break Through Traffic Signs=0, 36-60=0	→	Dry Surface Road=1
0.523	0.995	0.525	1.827	1.036	0.018	Clear Weather=1, Drive in opposite direction=0, 36-60=0	→	Dry Surface Road=1
0.604	0.995	0.608	1.580	1.036	0.021	Clear Weather=1, Over Speed limit=1, Heavy Truck =0	→	Dry Surface Road=1
0.694	0.995	0.698	1.376	1.036	0.024	Clear Weather=1, Road condition=0, Motorcycle=1	→	Dry Surface Road=1
0.602	0.995	0.605	1.586	1.036	0.021	Clear Weather=1, Over Speed limit=1, Obstruction=0	→	Dry Surface Road=1
0.511	0.995	0.514	1.868	1.036	0.018	Straight Way=1, Clear Weather=1, Motorcycle=1	→	Dry Surface Road=1
0.593	0.995	0.597	1.609	1.036	0.021	Clear Weather=1, Over Speed limit=1, Overtake=0	→	Dry Surface Road=1
0.604	0.995	0.608	1.580	1.036	0.021	Clear Weather=1, Over Speed limit=1, Road condition=0	→	Dry Surface Road=1
0.601	0.995	0.604	1.589	1.036	0.021	Clear Weather=1, Over Speed limit=1, Break Through Traffic Signs=0	→	Dry Surface Road=1
0.608	0.995	0.611	1.570	1.036	0.021	Clear Weather=1, Over Speed limit=1	→	Dry Surface Road=1
0.608	0.995	0.611	1.570	1.036	0.021	Clear Weather=1, Cannot Conclude =0, Over Speed limit=1	→	Dry Surface Road=1
0.608	0.995	0.611	1.570	1.036	0.021	Clear Weather=1, Over Speed limit=1, Drug=0	→	Dry Surface Road=1
0.608	0.995	0.611	1.571	1.036	0.021	Clear Weather=1, Over Speed limit=1, Overweight Carry=0	→	Dry Surface Road=1
0.608	0.995	0.611	1.571	1.036	0.021	Clear Weather=1, Over Speed limit=1, Use Mobile Phone=0	→	Dry Surface Road=1

24.6k | 12.9k | 1558

Figure 2.6 Associate Rules Mining total 1558 rules.

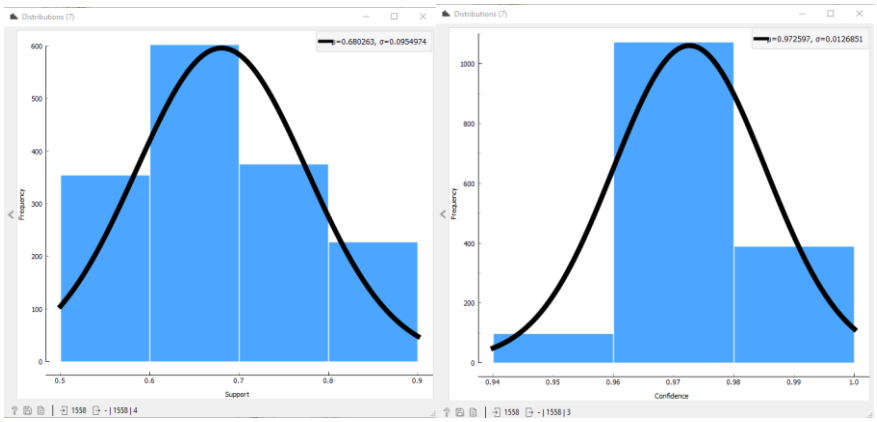


Figure 2.7 and 2.8 Support and Confidence distribution from 1,558 rules discovered.

Overall, 1,558 rule mining was discovered and divided by confidence clustering with color shades representing a Confidence Zone. The *y*-axis represents confidence, while the x-axis represents support. It becomes apparent that:

Group 1 Confidence 0.95–0.965 – Blue shade majority rule containing antecedent as male and dry surface as consequence.

Group 2 Confidence 0.965–0.98 – Green shade majority rule containing motorcycle and over speed limit as antecedent and dry surface road as consequence.

Group 3 Confidence 0.98–0.995 – Yellow shade is always high confidence, although with low support, since Cluster 3 contains clear weather as an antecedent and dry surface. Consequently, it implies that these two elements have a significant role in road accident mortality (Fig. 2.8) and that extreme caution should be taken during clear weather on dry surfaces.

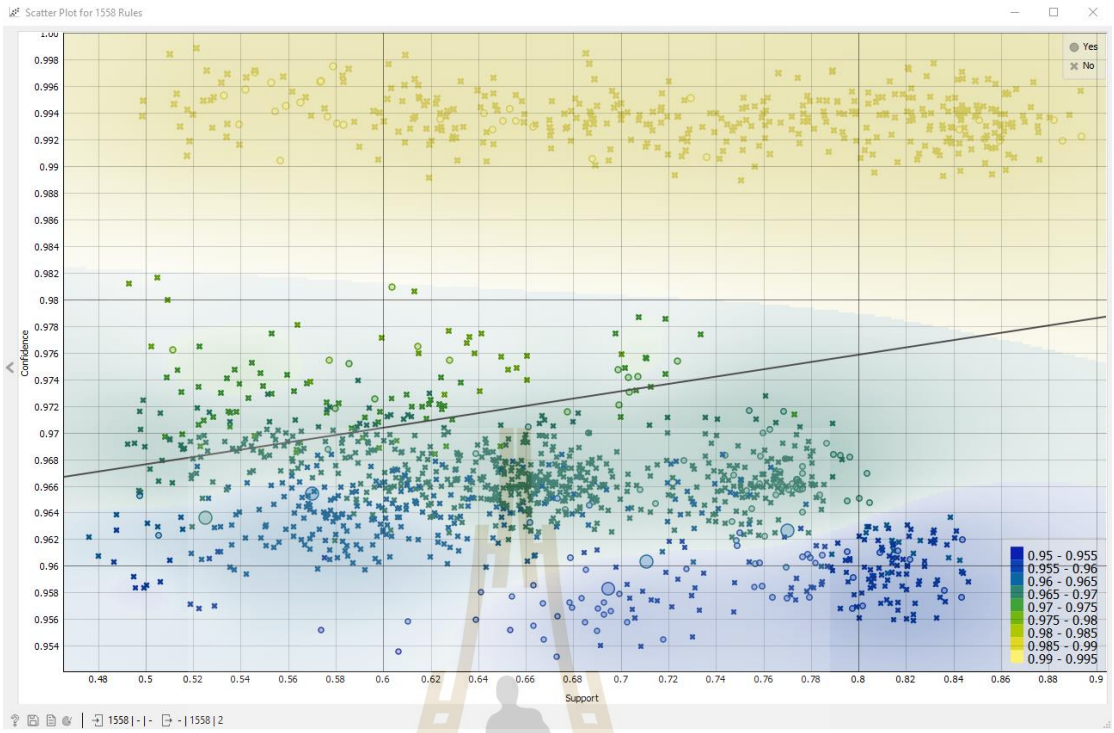


Figure 2.8 1,558 discovered rules with scatter plot Support VS Confidence

Later, start building a hierarchy cluster (HCA) by applying the agglomerative on 1,558 rules to arrange related antecedents into similar groups as a cluster with distancing. The distance between clusters was calculated using Euclidean distance as a complete linkage criterion. The dendrogram (Fig. 2.9) shows a C1–C3 cluster for the antecedent:

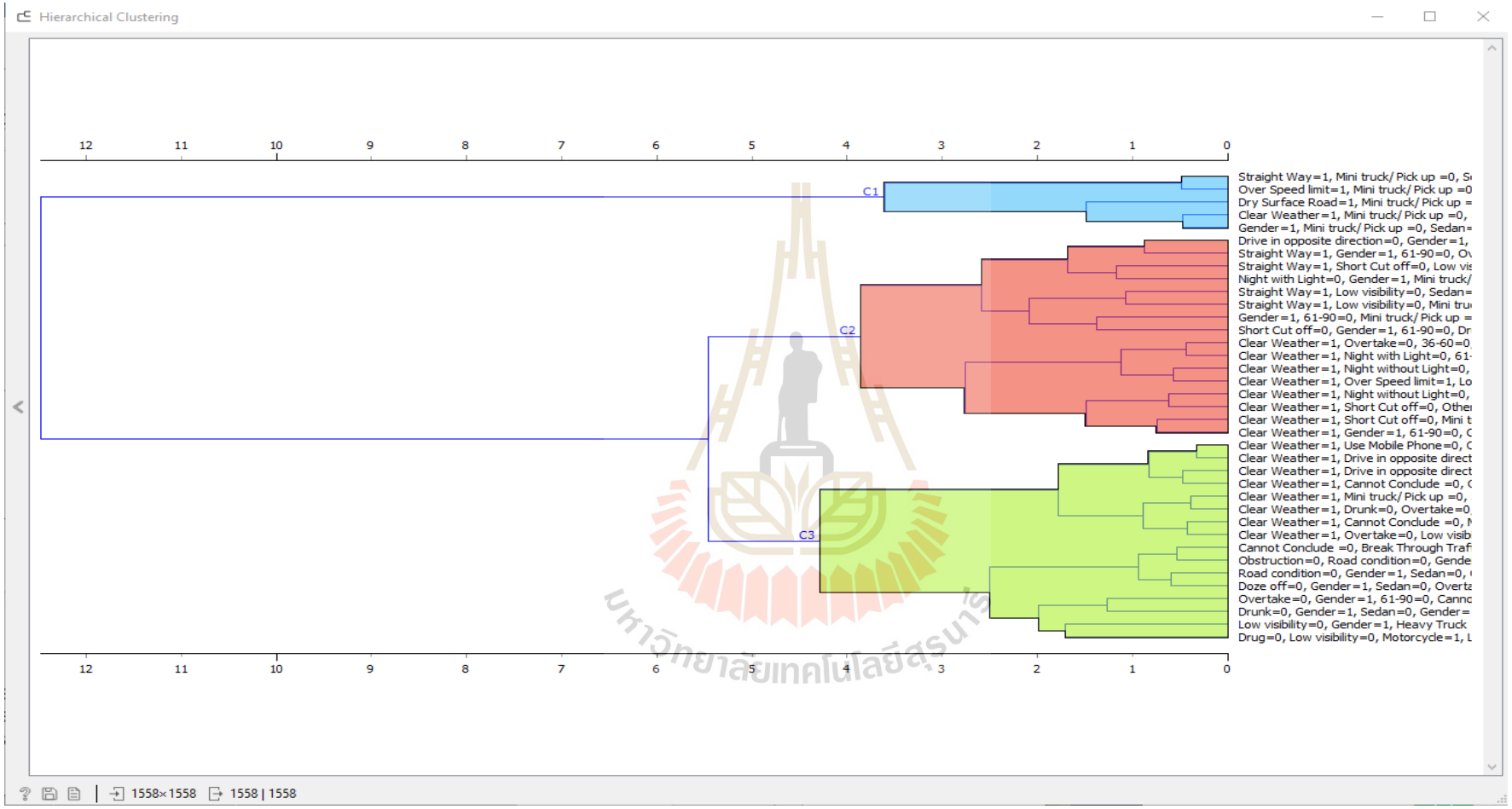


Figure 2.9 Dendrogram for 1,558 rules discovered on Antecedent.

C1 contains straightaway, over speed limit, dry surface road, clear weather, and male.

C2 contains straightaway, over speed limit, clear weather, and male.

C3 contains motorcycle, over speed limit, clear weather, and male.

Regarding the C1 cluster, all elements indicate the same consequence point to motorcycles, implying that the C1 cluster has most motorcycle fatalities, while C2 and C3 have consequence points to the dry surface road. That makes more sense when motorcyclists ride at higher speeds in clear weather on dry surface roads with less care than on wet road surfaces in poor weather conditions.

Table 2.5 Focusing Rule with high lift and widely gap between support and confidence.

Antecedent_1	Antecedent_2	Antecedent_3	Consequence	Support	Confidence	Lift
Over Speed Limit=1	Mini truck/ Pick up=0	Sedan=0	Motorcycle=1	0.551	0.966	1.206
Straight Way=1	Mini truck/ Pick up=0	Sedan=0	Motorcycle =1	0.577	0.964	1.203
Dry Surface Road=1	Mini truck/ Pick up=0	Sedan=0	Motorcycle =1	0.773	0.962	1.202
Clear Whether=1	Mini truck/ Pick up=0	Sedan=0	Motorcycle =1	0.704	0.962	1.201
Gender=1	Mini truck/ Pick up=0	Sedan=0	Motorcycle =1	0.689	0.958	1.196
Clear Weather=1	Over Speed Limit=1	Sedan=0	Dry Surface Road=1	0.566	0.995	1.037
Clear Weather=1	Over Speed Limit=1	Mini truck/ Pick up=0	Dry Surface Road=1	0.548	0.995	1.037
Clear Whether=1	Drunk=0	Motorcycle =1	Dry Surface Road=1	0.620	0.994	1.036
Clear Whether=1	Gender=1	Motorcycle =1	Dry Surface Road=1	0.599	0.994	1.036

Table 2.5 Focusing Rule with high lift and widely gap between support and confidence (Continued)

Antecedent_1	Antecedent_2	Antecedent_3	Consequence	Support	Confidence	Lift
Clear Whether=1	Over Speed Limit=1	Gender=1	Dry Surface Road=1	0.527	0.994	1.036
Straight Way=1	Clear Weather=1	Motorcycle =1	Dry Surface Road=1	0.511	0.995	1.036
Clear Weather=1	Gender=1	Dry Surface Road=1	Dry Surface Road=1	0.746	0.993	1.035
Straight Way=1	Clear Weather=1	Gender=1	Dry Surface Road=1	0.546	0.993	1.035
Over Speed Limit=1	Motor Bike=1		Dry Surface Road=1	0.535	0.972	1.013
Straight Way=1	Motor Bike=1		Dry Surface Road=1	0.56	0.97	1.011
Road Condition=0	Gender=1	Motorcycle =1	Dry Surface Road=1	0.659	0.968	1.008
Over Speed Limit=1	Road Condition=0	Gender=1	Dry Surface Road=1	0.576	0.966	1.007
Drunk=0	Gender=1	Motorcycle =1	Dry Surface Road=1	0.577	0.966	1.006
Gender=1	Motorcycle =1	Sedan=0	Dry Surface Road=1	0.665	0.966	1.006
Gender=1	Motorcycle =1	Mini truck/ Pick up=0	Dry Surface Road=1	0.665	0.966	1.006
Gender=1	Motorcycle =1	Other Type of car=0	Dry Surface Road=1	0.665	0.966	1.006
Gender=1	Motorcycle =1	Light Truck (6 wheels) =0	Dry Surface Road=1	0.665	0.966	1.006

Table 2.5 Focusing Rule with high lift and widely gap between support and confidence (Continued)

Antecedent_1	Antecedent_2	Antecedent_3	Consequence	Support	Confidence	Lift
Gender=1	Motorcycle =1	Heavy Truck (10+ wheels) =0	Dry Surface Road=1	0.665	0.966	1.006
Gender=1	Motorcycle=1		Dry Surface Road=1	0.665	0.966	1.006
Vehicle condition=0	Gender=1	Motorcycle =1	Dry Surface Road=1	0.659	0.966	1.006
Straight Way=1	Vehicle condition=0	Gender=1	Dry Surface Road=1	0.596	0.965	1.006

The following Table 2.5 and Fig. 2.10 display the association rules with a high lift and a wide gap between support and confidence with the antecedents 1–3 and the consequences, followed by the support score, confidence, and lift. The study established a minimum support score of more than 50%, a confidence threshold of more than 95%, and a lift threshold of more than one (1). For example, the rule with the widest gap between support and confidence is antecedent (straightaway, clear weather, motorcycle) => consequence (dry surface road), which increases 0.484 from support 0.511 to confidence 0.995. The rule with the highest lift is contained by motorcycles with different antecedents. All the interesting rules have been plotted, as shown in Fig. 2.10.

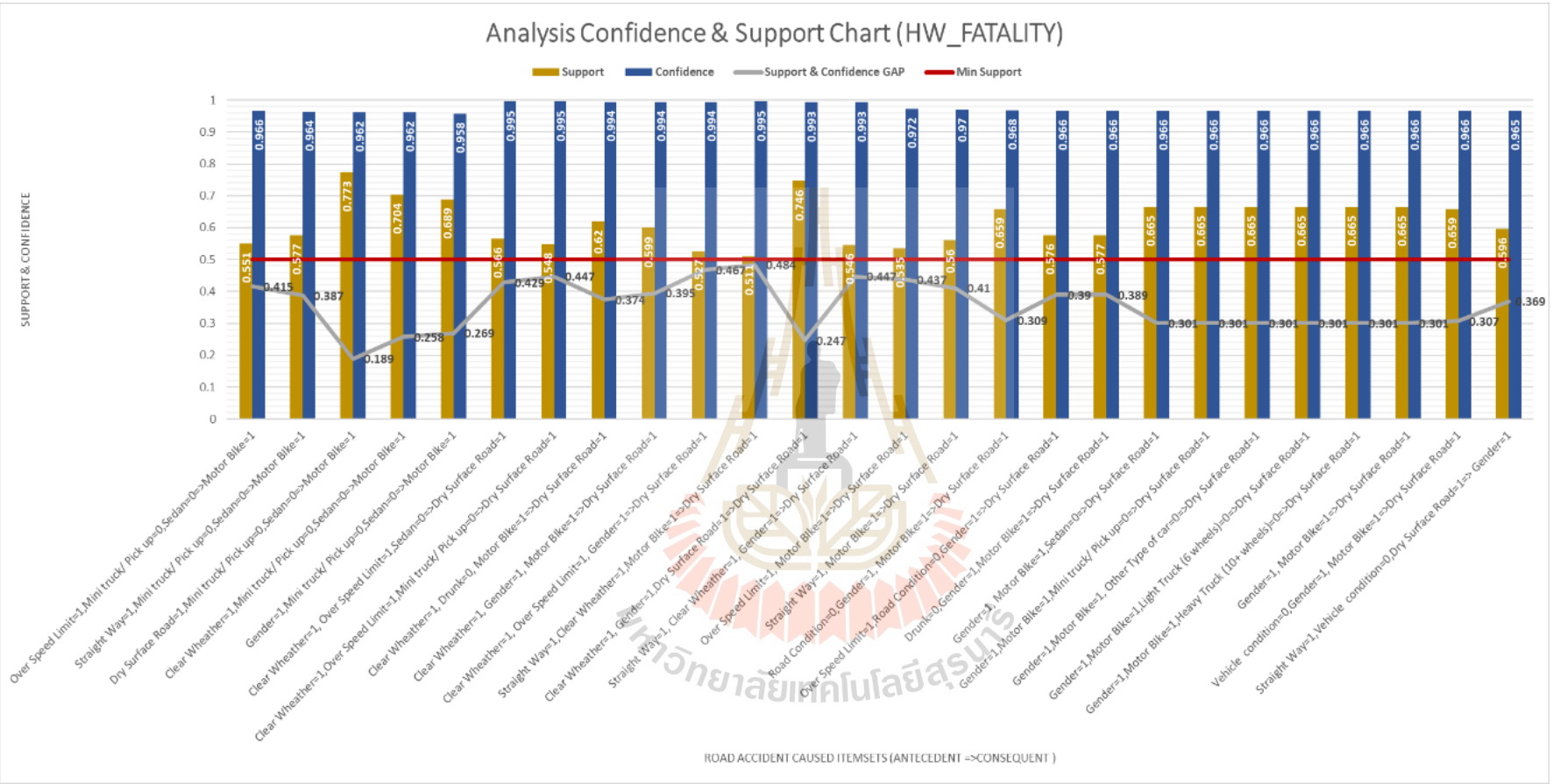


Figure 2.10 Confidence and support chart gap trend chart by interesting rules

2.5 Conclusion and Discussion

As a result of the association rule, the factors that enhance the likelihood of fatalities in highway road accidents are as follows:

1) Driver info - male
2) Driver behavior - over speed limit
3) Vehicle type - motorbike
4) Roadway - dry surface and straightaway
5) Weather - clear weather

When an accident occurs, all of these variables have a relationship and are linked to one another as the associated rule shows a potential cause of road accident fatalities, such that males riding motorcycles at speeds over the limit on straight roads in clear weather show increased risk for injury or death in traffic accidents, more than other conditions, with confidence levels increased from 0.5x, 06x, and 0.7x to 0.99x regarding if the consequences are motorcycle and dry surface road with high lift. As described at the opening pages, the higher the number of elements involved, the greater the possibility of an accident. Furthermore, the newly discovered straightway is a significant contributor, while transportation authority's exercise caution at intersections and on curve roads.

This might simply be due to the fact that when there is clear weather and a straight road with no curves, junctions, or turns, drivers frequently violate the speed limit as a result, which is more likely to cause accidents than when the weather is inclement, and it appears that males are driving faster than females. However, as of 2021, the current number of vehicles registered in Thailand is over 42 million, with motorbikes accounting for 50% of the total (DLT, 2021), potentially contributing to the largest number of fatalities from significant accidents. As Jomnonkwao et al. (2020) observed, motorcyclists are responsible for the vast majority of road fatalities, while prior studies showed different types of cars and motorcycles, such as rollover SUV/vans (Jafari Anarkooli et al., 2017), large truck (Huo et al., 2020; Jafari Anarkooli et al., 2017; Li et al., 2018), and pick-ups (Li et al., 2018). Additional research on motorcycle riders specifically, as well as other types of road users, may be conducted in the future. Aside from motorcycles, Sonal and Suman (2018) observed that external factors, such as

weather conditions like fog, rain, and snow, show greater impacts in road accidents than internal factors, such as the drivers themselves. Meanwhile, Thailand's climate has no snow or ice, with rain contributing only roughly 5 months a year (June to October) and the chilly season taking 4 months (November to February). The remainder of the year is summer, with clear weather conditions and dry road surfaces contributing approximately 7 months a year. The rule discovered that fatalities have a high chance in clear weather on dry surfaces, which correlate to the chilly and summer seasons.

Highway junctions were determined to be the riskiest for all accidents (Kumar & Toshniwal, 2016). However, this study discovered that a major risk exists even on straightaways, since drivers usually violate the law about exceeding the speed limit on straightaways with no junctions. Bahiru et al. (2018) observed internal factors, such as gender, age, accident location, and vehicle type. Those were discovered to have less of an influence on the severity of road accidents, although being male is still one of the primary factors leading to highway fatalities.

With all the rules discovered from this study; policymakers may eliminate some of the factors implicated in highway traffic accidents. At least it should raise awareness of risky driver behaviors. Authorities are considering proposed laws to control speed limits on long straightaways by using light signs, warning signs, and cameras that closely monitor driving speeds, especially motorcycles.

The study used data from 2015-2020, although the last 2 years (2019–2020) of the COVID-19 pandemic, the government issued an order ordering people across the country to lock down and not allow cross-provincial travel, particularly between 10PM – 4AM. People are also apprehensive about travelling to separate zones on their own, which means they are not travelling much. As such, the numbers for 2019–2020 may not accurately reflect the real number of accidents and fatalities for country.

As a related rule for future research, further analysis may be extended to all types of roads, particular automobile types, criminal data, medical data, or nonhighway data to aid policymakers in formulating the best option feasible with solid data backup.

2.6 Study limitation and future study

The study used accident data from the COVID-19 pandemic, which caused the government to lock down and prohibit travel between provinces. People are also cautious to travel to the separated zones on their own, implying that they have not traveled extensively. As such, the numbers for 2019–2020 may not accurately reflect the real number of accidents and fatalities for country.

As a related rule capability, for future research, the further analysis may be extended to all types of roads, particular automobile types, criminal data, medical data, or nonhighway data to aid policymakers in choosing the most feasible options with solid data backup.

2.7 Reference

Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. *Expert Systems with Applications*, *40*(15), 6047-6054. https://doi.org/10.1016/j.eswa.2013.05.027

Al Mamlook, R. E., Ali, A., Hasan, R. A., & Mohamed Kazim, H. A. (2019). Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation. Proceedings of the IEEE National Aerospace Electronics Conference, NAECON,

Anvari, M. B., Tavakoli Kashani, A., & Rabieyan, R. (2017). Identifying the Most Important Factors in the At-Fault Probability of Motorcyclists by Data Mining, Based on Classification Tree Models. *International Journal of Civil Engineering*, *15*(4), 653-662. https://doi.org/10.1007/s40999-017-0180-0

Bahiru, T. K., Kumar Singh, D., & Tessfaw, E. A. (2018). Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018,

Behnood, A., & Mannering, F. (2017). The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. *Analytic Methods in Accident Research*, *14*, 41-53. https://doi.org/https://doi.org/10.1016/j.amar.2017.04.001

Ben-David, S. S.-S. a. S. (2014). <understanding-machine-learning-theory-algorithms.pdf>. *Cambridge University Press.* http://www.cs.huji.ac.il/~shais/UnderstandingMachine Learning

Bhavsar, R., Amin, A., & Zala, L. (2021). Development of Model for Road Crashes and Identification of Accident Spots [Article]. *International Journal of Intelligent Transportation Systems Research*, *19*(1), 99-111. https://doi.org/10.1007/|s13177-020-00228-z

Breiman, L. (2001). Mach Learn.

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., & Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia,

Champahom, T., Jomnonkwao, S., Chatpattananan, V., Karoonsoontawong, A., & Ratanavaraha, V. (2019). Analysis of Rear-End Crash on Thai Highway: Decision Tree Approach. *Journal of Advanced Transportation*, *2019*, 1-13. https://doi.org/10.1155/2019/2568978

Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., & Ratanavaraha, V. (2020). Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. *Accident Analysis & Prevention*, *141*, 105537. https://doi.org/https://doi.org/10.1016/j.aap.2020.105537

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., & Guan, H. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis & Prevention*, *80*, 76-88. https://doi.org/https://doi.org/10.1016/j.aap.2015.03.036

Chen, C., Zhang, G., Yang, J., Milton, J. C., & Alcántara, A. D. (2016). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident Analysis & Prevention*, *90*, 95-107. https://doi.org/https://doi.org/10.1016/j.aap.2016.02.002

Chen, M.-Y. (2012). Comparing Traditional Statistics, Decision Tree Classification And Support Vector Machine Techniques For Financial Bankruptcy Prediction. *Intelligent Automation & Soft Computing*, *18*(1), 65-73. https://doi.org/10.1080/10798587.2012.10643227

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cuenca, L. G., Puertas, E., Aliane, N., & Andres, J. F. (2018). Traffic Accidents Classification and Injury Severity Prediction. 2018 3rd IEEE International Conference on Intelligent Transportation Engineering, ICITE 2018,

Cunto, F. J. C., & Ferreira, S. (2017). An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. *Journal of Transportation Safety & Security*, *9*(sup1), 33-46. https://doi.org/10.1080/19439962.2016.1162891

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Zupan, B. (2013). Orange: Data mining toolbox in python [Article]. *Journal of Machine Learning Research*, *14*, 2349-2353. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885599052&partnerID=40&md5=75d2df52a0c46b5ab58ab08e1576114e

DLT. (2021). Department of Land Transportation. *https://www.dlt.go.th/th/public-news/view.php?_did=2806*.

Dongare, A., Kharde, R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

El Abdallaoui, H. E. A., El Fazziki, A., Ennaji, F. Z., & Sadgal, M. (2018). Decision Support System for the Analysis of Traffic Accident Big Data. Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018,

Feng, M., Zheng, J., Ren, J., & Xi, Y. (2020). Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. In *Advances in Brain Inspired Cognitive Systems* (pp. 520-529). https://doi.org/10.1007/978-3-030-39431-8_50

Geedipally, S. R., Turner, P. A., & Patil, S. (2011). Analysis of Motorcycle Crashes in Texas with Multinomial Logit Model. *Transportation Research Record*, *2265*(1), 62-69. https://doi.org/10.3141/2265-07

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. http://oreilly.com/catalog/errata.csp?isbn=9781492032649

Guido, A. C. M. S. (2017). Introduction to machinelearning with python. http://oreilly.com/catalog/errata.csp?isbn=9781449369415 (Third Release) (O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.)

Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review [Review]. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(4), 432-446. https://doi.org/10.1016/j.jtte.2020.05.002

Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests [Article]. *Accident Analysis and Prevention*, *41*(1), 98-107. https://doi.org/10.1016/j.aap.2008.09.009

Helen, W. R., Almelu, N., & Nivethitha, S. (2019). Mining Road Accident Data Based on Diverted Attention of Drivers. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018,

Hou, Q., Huo, X., Leng, J., & Cheng, Y. (2019). Examination of driver injury severity in freeway single-vehicle crashes using a mixed logit model with heterogeneity-in-means. *Physica A: Statistical Mechanics and its Applications*, *531*, 121760. https://doi.org/10.1016/j.physa.2019.121760

Huo, X., Leng, J., Hou, Q., & Yang, H. (2020). A Correlated Random Parameters Model with Heterogeneity in Means to Account for Unobserved Heterogeneity in Crash Frequency Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*, 036119812092221. https://doi.org/10.1177/0361198120922212

Jafari Anarkooli, A., Hosseinpour, M., & Kardar, A. (2017). Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. *Accident Analysis & Prevention*, *106*, 399-410. https://doi.org/10.1016/j.aap.2017.07.008

John, M., & Shaiba, H. (2019). Apriori-Based Algorithm for Dubai Road Accident Analysis. Procedia Computer Science,

John, M., & Shaiba, H. (2022). Analysis of Road Accidents Using Data Mining Paradigm. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 68, pp. 215-223).

Jomnonkwao, S., Uttra, S., & Ratanavaraha, V. (2020). Forecasting Road Traffic Deaths in Thailand: Applications of Time-Series, Curve Estimation, Multiple Linear Regression, and Path Analysis Models. *Sustainability*, *12*(1). https://doi.org/10.3390/su12010395

Jou, R. C., Yeh, T. H., & Chen, R. S. (2012). Risk factors in motorcyclist fatalities in Taiwan. *Traffic Inj Prev*, *13*(2), 155-162. https://doi.org/10.1080/15389588.2011.641166

Jung, S., Qin, X., & Noyce, D. A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis & Prevention*, *42*(1), 213-224. https://doi.org/https://doi.org/10.1016/j.aap.2009.07.020

Khorashadi, A., Niemeier, D., Shankar, V., & Mannering, F. (2005). Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. *Accid Anal Prev*, *37*(5), 910-921. https://doi.org/10.1016/j.aap.2005.04.009

Kim, J.-K., Ulfarsson, G. F., Kim, S., & Shankar, V. N. (2013). Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention*, *50*, 1073-1081. https://doi.org/https://doi.org/10.1016/j.aap.2012.08.011

Kim, J. H., Kim, J., Lee, G., & Park, J. (2021). Machine Learning-Based Models for Accident Prediction at a Korean Container Port. *Sustainability*, *13*(16), 9137. https://www.mdpi.com/2071-1050/13/16/9137

Krull, K. A., Khattak, A. J., & Council, F. M. (2000). Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes. *Transportation Research Record*, *1717*(1), 46-54. https://doi.org/10.3141/1717-07

Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations [Article]. *Journal of Modern Transportation*, *24*(1), 62-72. https://doi.org/10.1007/s40534-016-0095-5

Kuşkapan, E., Çodur, M. Y., & Atalay, A. (2021). Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms [Article]. *Accident Analysis and Prevention*, *155*, Article 106098. https://doi.org/10.1016/j.aap.2021.106098

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P. D., & Ma, D. T. (2018). Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. *Analytic Methods in Accident Research*, *20*, 1-14. https://doi.org/https://doi.org/10.1016/ j.amar.2018.08.001

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z., Ma, D. T. (2019). Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. *Accident Analysis & Prevention*, *124*, 219-229. https://doi.org/https://doi.org/10.1016/j.aap.2018.12.020

Li, Z., Wu, Q., Ci, Y., Chen, C., Chen, X., & Zhang, G. (2019a). Using latent class analysis and mixed logit model to explore risk factors on driver injury severity in single-vehicle crashes. *Accident; analysis and prevention*, *129*, 230-240.

Li, Z., Wu, Q., Ci, Y., Chen, C., Chen, X., & Zhang, G. (2019b). Using latent class analysis and mixed logit model to explore risk factors on driver injury severity in single-vehicle crashes. *Accident Analysis & Prevention*, *129*. https://doi.org/10.1016/j.aap.2019.04.001

Mafi, S., AbdelRazig, Y., & Doczy, R. (2018). Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. In *Transportation Research Record* (Vol. 2672, pp. 171-183).

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accid Anal Prev*, *122*, 181-188. https://doi.org/10.1016/j.aap.2018.10.014

Mohamad, I., Jomnonkwao, S., & Ratanavaraha, V. (2022). Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand. *Case Studies on Transport Policy*, *10*(4), 2165-2174. https://doi.org/https://doi.org/10.1016/j.cstp.2022.09.016

Mphela, T. (2020). Causes of road accidents in botswana: An econometric model [Article]. *Journal of Transport and Supply Chain Management*, *14*, 1-8, Article a509. https://doi.org/10.4102/jtscm.v14i0.509

Osman, M., Mishra, S., & Paleti, R. (2018). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*. https://doi.org/10.1016/j.aap.2018.05.004

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists [Article]. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051-10072. https://doi.org/10.1007/s12652-020-02759-5

Ospina-Mateus, H., Quintana Jiménez, L. A., López-Valdés, F. J., Morales-Londoño, N., & Salas-Navarro, K. (2019). Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia. In *Communications in Computer and Information Science* (Vol. 1052, pp. 309-320).

Pakgohar, A., Tabrizi, R. S., Khalili, M., & Esmaeili, A. (2011). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. *Procedia Computer Science*, *3*, 764-769. https://doi.org/10.1016/j.procs.2010.12.126

PDPM. (2020). Thailand Department of Public Disaster Prevention and Mitigation. *https://www.disaster.go.th/en/*

Recal, F., & Demirel, T. (2021). Comparison of machine learning methods in predicting binary and multi-class occupational accident severity [Article]. *Journal of Intelligent and Fuzzy Systems*, *40*(6), 10981-10998. https://doi.org/10.3233/JIFS-202099

Rezapour, M., Mehrara Molan, A., & Ksaibati, K. (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. *International Journal of Transportation Science and Technology*, *9*(2), 89-99. https://doi.org/10.1016/j.ijtst.2019.10.002

RSC, T. (2019). Thailand Accident Research Center *Thailand Accident Research Center* https://www.thairsc.com/

Samerei, S. A., Aghabayk, K., Mohammadi, A., & Shiwakoti, N. (2021). Data mining approach to model bus crash severity in Australia [Article]. *Journal of Safety Research*, *76*, 73-82. https://doi.org/10.1016/j.jsr.2020.12.004

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. *Computers*, *10*(12), 157. https://www.mdpi.com/2073-431X/10/12/157

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Se, C., Champahom, T., Jomnonkwao, S., Chaimuang, P., & Ratanavaraha, V. (2021). Empirical comparison of the effects of urban and rural crashes on motorcyclist injury severities: A correlated random parameters ordered probit approach with heterogeneity in means. *Accid Anal Prev*, *161*, 106352. https://doi.org/10.1016/j.aap.2021.106352

Se, C., Champahom, T., Jomnonkwao, S., Karoonsoontawong, A., & Ratanavaraha, V. (2021). Temporal stability of factors influencing driver-injury severities in single-vehicle crashes: A correlated random parameters with heterogeneity in means and variances approach. *Analytic Methods in Accident Research*, *32*, 100179. https://doi.org/https://doi.org/10.1016/j.amar.2021.100179

Shaheed, M. S., Gkritza, K., Zhang, W., & Hans, Z. (2013). A mixed logit analysis of two-vehicle crash seventies involving a motorcycle. *Accident; analysis and prevention*, *61*. https://doi.org/10.1016/j.aap.2013.05.028

Shweta, Yadav, J., Batra, K., & Goel, A. K. (2021). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. Journal of Physics: Conference Series,

Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., & Hanks, H. (2011). Risk factors for fatal crashes in rural Australia. *Accident Analysis & Prevention*, *43*(3), 1082-1088. https://doi.org/https://doi.org/10.1016/j.aap.2010.12.016

Sonal, S., & Suman, S. (2018). A framework for analysis of road accidents. 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research, ICETIETR 2018,

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Srikant, R. A. a. R. (1994). *Fast algorithms for mining association rules* Proceedings of the 20th International Conference on Very Large Data Bases, VLDB,, Santiago, Chile.

Tolles, J., & Meurer, W. J. (2016). Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, *316*(5), 533-534. https://doi.org/10.1001/jama.2016.7653

Webb, G. I. (2010). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 713-714). Springer US. https://doi.org/10.1007/978-0-387-30164-8_576

Wei, F., Cai, Z., Liu, P., Guo, Y., Li, X., & Li, Q. (2021). Exploring Driver Injury Severity in Single-Vehicle Crashes under Foggy Weather and Clear Weather. *Journal of Advanced Transportation*, *2021*, 9939800. https://doi.org/10.1155/2021/9939800

WHO. (2018). World Health Organization: Global status report on road safety 2018. . *https://extranet.who.int/roadsafety/death-on-the-roads/*.

William J. Frawley, G. P.-S., and Christopher J. Matheus. (1992). Knowledge Discovery in Databases. https://doi.org/DOI: https://doi.org/10.1609/aimag.v13i3.1011 (An Overview. AI Magazine, 13(3), 57) (AAAI/MIT Press, Cambridge, MA)

Wu, Q., Zhang, G., Zhu, X., Liu, X. C., & Tarefder, R. (2016). Analysis of driver injury severity in single-vehicle crashes on rural and urban roadways. *Accident Analysis & Prevention*, *94*, 35-45. https://doi.org/https://doi.org/10.1016/j.aap.2016.03.026

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Xie, Y., & Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. *Accident; analysis and prevention*, *47*, 36-44. https://doi.org/10.1016/j.aap.2011.12.012

Yu, H., Yuan, R., Li, Z., Zhang, G., & Ma, D. T. (2020). Identifying heterogeneous factors for driver injury severity variations in snow-related rural single-vehicle crashes. *Accident Analysis & Prevention*, *144*, 105587. https://doi.org/https://doi.org/10.1016/j.aap.2020.105587

Yu, M., Zheng, C., & Ma, C. (2020). Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *27*, 100126. https://doi.org/https://doi.org/10.1016/j.amar.2020.100126

Zhang, X. F., & Fan, L. (2013). A decision tree approach for traffic accident analysis of Saskatchewan highways. Canadian Conference on Electrical and Computer Engineering,

Zhou, M., & Chin, H. C. (2019). Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. *Accident Analysis & Prevention*, *124*, 104-112. https://doi.org/10.1016/j.aap.2019.01.009

CHAPTER 3
USING A DECISION TREE TO COMPARE RURAL VERSUS HIGHWAY MOTORCYLE FATALITIES

3.1 Abstract

Thailand ranks first in Asia and ninth in the world in term of road accident. As of 2020, the number of vehicles registered in Thailand was over 41 million, with motorcycles accounting for half of all vehicles. This study aimed to determine the cause of fatalities to reduce motorcycle accidents. The research entailed separating the accidents and fatalities into those occurring on highways (HWs) versus those occurring on rural roadways (RRs) and focused solely on rider at fault accidents to involve any confounding factors related to passengers or others involved. In Thailand, HWs have higher speed limits and allow more vehicle types than some RRs. Thailand's Department of Public Disaster Prevention and Mitigation recorded 115,154 motorcycle accidents from 2015 to 2020. Decision trees allow for processing large amounts of data to drill down into associations between the individual variables in a large data set; in this study, the tree also separated accidents into whether the driver was exceeding the speed limit. The model's performance for HWs, predicted misclassifications were found to be 28.3% (fatality to nonfatality) and a 44.5% (nonfatality to fatality) while predicted misclassification for RRs were 15.5% (fatality to nonfatality) and 60% (nonfatality to fatality). At all ages, the most fatalities were among male riders on dry straightaways in clear daytime weather; notably, however, on RRs, even when the rider was driving responsibly, fatalities were high at night on roads with no light. Following the presentation of the study findings, suggestions are made for ways the Thai government can improve the motorcycle accident and fatality statistics, including increasing the age limit for a motorcycle license, with engine size limits further divided according to age; proper enforcement of the existing rules will also improve the country's accident statistics. It will also be highly effective to improve road lighting, particularly on RRs.

3.1.1 Highlight

1) Gender was the most significant variable in fatalities: Most fatalities were among male riders on both Highway and Rural irrespective of the rider speed limit.

2) At all ages, the most fatalities were among male riders on dry straightaways in clear daytime weather notably.

3) Short cutoffs were among the most common causes of fatalities on both Highway and Rural, but excess speed was also a factor only on Rural.

4) On rural roadway even when the rider was driving responsibly, fatalities were high at night on roads with no light

5) There is a higher probability of a fatality on a HW than on a RR: HW, and RR, However, both RRs and HWs have higher fatality rates at night (00.00–07.00).

3.2 Introduction

Thailand is one of the countries with a high rate of fatalities from road accidents, ranking first in Asia and ninth in the world. Thais are killed in traffic accidents at a rate of 32.7 per 100,000 persons (WHO, 2018). As of 2020, the total number of vehicles registered in Thailand was over 41 million, with motorcycles accounting for half of all vehicles (DLT, 2021).

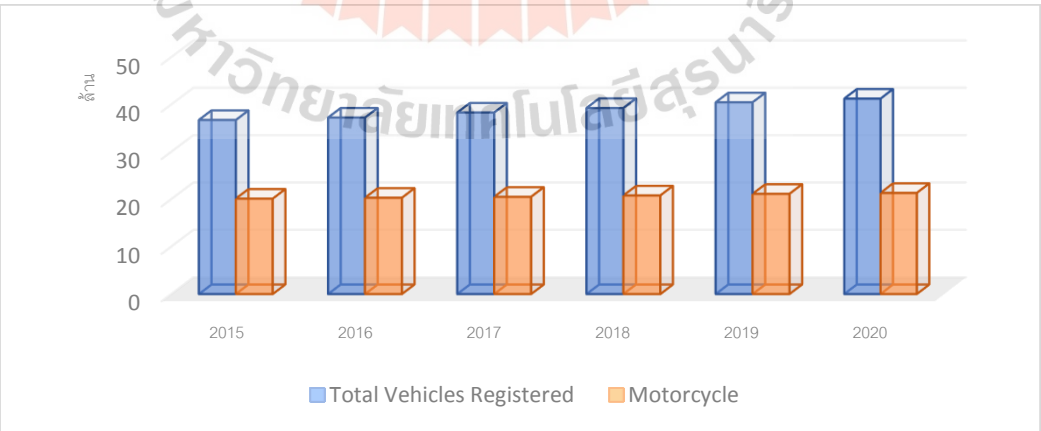


Figure 3.1 Total number of vehicles and motorcycles registered in Thailand from 2015 to 2020.

Figure 3.1 shows that the number of motorcycles registered in Thailand grew continually from 2015 to 2020, when there were 41,471,135 vehicles registered in the country, of which 21,396,980 were motorcycles; these were followed by lightweight 4-wheeled drive vehicles (10,446,505), mini-trucks (6,878,050), and others (2,749,600) (DLT, 2021). However, although motorcycles only account for half the vehicles in Thailand, motorcyclists account for most road accident fatalities (Jomnonkwao et al., 2020). According to the Thailand Accident Research Center, in 2019, there were 4,802 motorcycle fatalities, an average of 13.15 people per day, with most occurring among people aged 20 years and up, also known as the working age group. In terms of causes, 54% of accidents were the fault of the motorcyclists, drivers were at fault in 40% of the accidents, and the road and vehicle accounted for 4% and 2%, respectively (RSC, 2019). Worldwide research has identified elements that appear to be common to every accident. Thailand has six road types:

1) A motorway is a HW designed for high mobility with low access based on limited entrances and exits to designated points, and no two-wheeled vehicles are permitted. Motorways are supervised by the Department of Highways.
2) The national HWs link regions, provinces, and districts; they emphasize mobility, but access is not limited, and rules in general are less strict than those for motorways are. It is difficult to travel through cities, and the HWs were designed to bypass the cities; they are also supervised by the Department of Highways.
3) Rural roadways (RRs) are located outside of municipalities and connect with the national HWs. They are supervised by the Department of Rural Roads.
4) Municipal roadways provide the streets in municipalities and are maintained by the municipalities.
5) Subdistrict roadways serve as the streets for those areas, and they are supervised by subdistrict organizations.
6) Concession roadways are privately owned; the government grants concessions to private entities that are then responsible for supervising the roads.

For this study, all roads managed by the Department of Highways are designated HWs and rural and subdistrict roads are RRs. The motorcycle speed limits are lower on RRs—80 km/hr. for engines larger than 400cc and 60 km/hr. for smaller cycles; for HWs, the limits are 90 km/hr. for engines larger than 400cc and 70 km/hr. otherwise (DLT, 2021) and the study was distinctive in that it separated accidents and fatalities into those that occurred on highways (HWs) and those that occurred on rural roadways (RR), and it focused solely on rider-at-fault accidents in order to eliminate any confounding factors related to passengers or others involved.

In statistical side, Discriminant analysis (DA) is a technique commonly used in Statistical Algorithms to classify a set of observations into predefined classes and LDA (linear Discriminant analysis) is a diagnostic method for detecting potentially influential observations. The usual assumptions relevant to discriminant analysis are linearity, normality, and homoscedasticity of within-group variances of independent variables. However, due to violations of these assumptions, discriminant analysis has been supplanted by LR (logistic Regression), which requires fewer assumptions, produces more robust results, and is easier to use and comprehend than discriminant analysis (Chen, 2012). A regression-type model is a CART model that predicts the value of continuous variables using a set of continuous or categorical predictor variables. For this study, we selected a decision tree (CART regression Tree) to drill down into a set of big data to identify the relevant variables and analyze the relationships among them. Decision tree mining is among the most popular machine learning techniques (Wu et al., 2007) for its comprehensibility and ease of interpretation. One of the primary advantages of a decision tree is the ability to derive decision rules; these rules can aid in identifying safety issues and developing performance metrics (Abellán et al., 2013).

3.3 Literature Review

Previous studies on road accidents have been classified by group components that are suspected to be involved in every accident, according to international research.

3.3.1 Age and Gender

Motorcycle accidents are more likely to occur among young people because they are less disciplined, are unfamiliar with traffic laws, and have less driving

experience (Zhang & Fan, 2013). Men and women aged 20–39 years who ride motorcycles are more likely to be involved in major accidents, whereas when no motorcycle or cyclist is involved in the incident, the severity is likely to be minor (Ospina-Mateus et al., 2019). Jou et al. (2012) found that being older, male, and unlicensed; not wearing a helmet; riding after drinking; and driving heavy motorcycles (above 550cc) were linked to higher motorcycle fatality rates. Additionally, rider age was the most important factor when the rider was not at fault (Champahom et al., 2019). Pakgohar et al. (2011) found that most fatalities were among young persons who were in good health before the accident. Riders between the ages of 18 and 24 years have insufficient experience to adjust while driving including adjusting their speed to the road conditions (Bucsuházy et al., 2020).

3.3.2 Weather and Road Conditions

Research has established that external conditions such as fog, rain, and snow have a greater influence on road accidents than rider-related internal factors and that the drivers/riders are more likely than passengers to be injured or killed in an accident (El Abdallaoui et al., 2018). According to the findings, the most important and influential road accident variables are speed limit; weather conditions; road factors such as type, surface, and number of lanes; lighting conditions; and time of the accident. Factors that had less influence on accidents were gender, age, accident site, and vehicle type (Feng et al., 2020). Highway (HW) intersections have been identified as the most dangerous for all accidents (Kumar & Toshniwal, 2016), Malin et al. (2019). As noted above, however, there are still significant accidents on straightaways with no intersections, in part because riders disobey the speed limit and in part because of poor road conditions.

3.3.3 Other Important Factors

Vehicle speed is the most critical determinant of an accident's severity (Al Mamlook et al., 2019), followed by factors such as speed limit, age, and road type (Rezapour et al., 2020). Travel at night increases the risk of an accident (Mphela, 2020) and increases the severity of any injuries, particularly when there is no light Shaheed et al. (2013), (Kim et al., 2013), (Jafari Anarkooli et al., 2017) and after midnight (Zhou & Chin, 2019). Xie and Huynh (2012) determined that the severity of injuries from

accidents on dark roads decreases when riders are more cautious. Motorcycles are riskier in rural areas. Male riders, pillion riders, speeding, improper overtaking, and fatigue are all important factors that influence severe and fatal injuries. (C. Se et al., 2021)

Additionally, the risk of motorcycle death increases for single-vehicle accidents that occur on nonurban roads at night, and the major factors that affect rear-end crashes are passenger characteristics and the rider's age, whereas side collisions are most commonly the result of lighting conditions and landscape (Anvari et al., 2017), Siskind et al. (2011). Focusing on driver factors, researchers discovered that high-speed driving, driving while intoxicated. And traffic violations all contributed to high rates of fatalities on RRs (Khorashadi et al., 2005). Researchers have used many tools in accident analysis, including measuring the accuracy between models or methods (Table 3.1), but few have studied the same model or method to compare two road types.

Table 3.1 The Machine Learning Models Used in Extant Traffic Accident Studies

Author	Methodology											
	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-Nearest Neighbor	K-Means	Multinomial Logistic	Neural Network	Naïve Bayes	Random Forest	SVM
Ospina-Mateus et al. (2021)	-	-	-	✓	-	✓	-	-	✓	✓	✓	✓
Harb et al. (2009)	-	-	-	✓	-	-	-	-	-	-	✓	-
Kuşkapan et al. (2021)	-	-	-	-	-	✓	-	-	-	✓	-	✓
Abellán et al. (2013)	-	-	-	✓	-	-	-	-	-	-	-	-
Mafi et al. (2018)	-	-	-	-	-	-	-	-	-	-	✓	-
Al Mamlook et al. (2019)	-	✓	✓	✓	-	✓	-	-	-	✓	✓	✓
Recal and Demirel (2021)	-	-	-	✓	✓	-	-	✓	✓	-	-	✓
Kumar and Toshniwal (2016)	✓	-	-	-	-	-	✓	-	-	-	-	-
Helen et al. (2019)	✓	-	-	-	-	-	✓	-	-	-	-	-
Feng et al. (2020)	✓	-	-	-	-	-	-	-	✓	-	-	-
Bhavsar et al. (2021)	✓	-	-	-	-	-	-	-	-	-	-	-
Bahiru et al. (2018)	-	-	-	✓	-	-		-	-	✓	-	-

3.4 Methodology

The research begins with motorcycle accident data from Thailand's Department of Public Disaster Prevention and Mitigation, which counted 115,154 single-rider accidents between 2015 and 2020. Toward our study aim, the data was compiled on HR and RR motorcycle accident fatalities, developed a decision tree model, and measured its accuracy. Figure 3.2 displays the steps in the study process, also listed below:

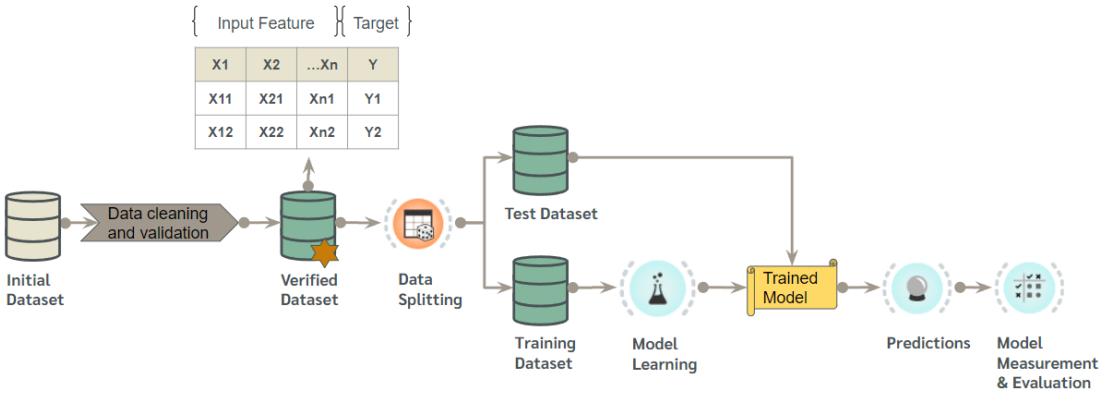


Figure 3.2 The steps in the process for the study.

1) After cleansing the data set, the initial dataset was validated for detecting and correcting missing and incompletely captured data as well as demonstrating the data's quality.
2) Verified Dataset - Set the target to Fatal/non-fatal and partition the data in binary mode both HW and RR data set.
3) Data Separation - Separated test and train data sets.
4) Model Learning enables the model to learn from the test data set and then test with the remaining data set.
5) Prediction and Model Measurement - To assess each model's prediction accuracy.

3.4.1 Data Description

As noted earlier, the 2015–2020 motorcycle accident data from the Thailand Department of Public Disaster Prevention and Mitigation indicated 115,154 single-rider accidents, 61,866 on HWs and 53,288 on RRs (PDPM, 2020). Table 2 presents the categorical and descriptive statistics for the study data, which we divided into four categories: roadway characteristics, external factors involving the environment and weather conditions, internal factors involving driver behavior, and driver details. According to the descriptive data table, most accidents on both HWs and RRs were caused by being a male rider between 15 and 35 years of age exceeding the speed limit; most accidents occurred on dry surfaces and in clear weather, even when the driver stayed on the right side of the road.

Table 3.2 The Categorical Variables and Their Descriptive Statistics

Aciccent Event (Attribute)	HighWay Fatality Yes Count	%	HighWay Fatality No Count	%	Non-HighWay Fatality Yes Count	%	Non-HighWay Fatality No Count	%
RoadWay								
Dry Surface Road	18992	30.7%	40112	64.8%	8063	15.1%	43049	80.8%
Wet Surface	673	1.1%	2089	3.4%	277	0.5%	1899	3.6%
Straight Way	14171	22.9%	29389	47.5%	5619	10.5%	31698	59.5%
Not straight Way (Curve,Slope,Junction, etc)	5494	8.9%	12812	20.7%	2721	5.1%	13250	24.9%
Obstruction	343	0.6%	1301	2.1%	170	0.3%	1847	3.5%
Road condition	194	0.3%	886	1.4%	167	0.3%	1388	2.6%
Vehicle condition	226	0.4%	825	1.3%	126	0.2%	1402	2.6%
External Factor (Envi and Weather Con)								
Day Time (06.00-18.00)	9649	15.6%	24654	39.9%	4021	7.5%	26215	49.2%
Night with Light	5633	9.1%	11310	18.3%	2140	4.0%	9936	18.6%
Night without Light	4383	7.1%	6237	10.1%	2179	4.1%	8797	16.5%
Low visibility	1819	2.9%	5437	8.8%	659	1.2%	5523	10.4%
Clear Weather	17298	28.0%	35951	58.1%	7451	14.0%	39037	73.3%
Not Clear Weather (Rain, fog, etc)	2367	3.8%	6250	10.1%	889	1.7%	5911	11.1%
Internal Factor (Driver Behavior)								
Drunk	2410	3.9%	9521	15.4%	1357	2.5%	11065	20.8%
Over Speed limit	13524	21.9%	20018	32.4%	5888	11.0%	18129	34.0%
Break Through Traffic lights	185	0.3%	283	0.5%	50	0.1%	183	0.3%
Break Through Traffic Signs	289	0.5%	748	1.2%	88	0.2%	573	1.1%
Overtake	526	0.9%	702	1.1%	121	0.2%	576	1.1%
Use Mobile Phone	22	0.0%	171	0.3%	2	0.0%	185	0.3%
Short Cut off	4156	6.7%	9170	14.8%	1248	2.3%	10141	19.0%
Drug	3	0.0%	39	0.1%	3	0.0%	27	0.1%
Drive in opposite direction	385	0.6%	563	0.9%	58	0.1%	268	0.5%
Doze off	260	0.4%	536	0.9%	92	0.2%	369	0.7%
Overweight Carry	11	0.0%	28	0.0%	5	0.0%	37	0.1%
Cannot Conclude	676	1.1%	1686	2.7%	256	0.5%	1772	3.3%
Driver info								
Gender (male)	16921	27.4%	30998	50.1%	7236	13.6%	32500	61.0%
Gender (Female)	2744	4.4%	11203	18.1%	1104	2.1%	12448	23.4%
Youth 15-35	9894	16.0%	23204	37.5%	4010	7.5%	23277	43.7%
Adult 36-60	7111	11.5%	14830	24.0%	3205	6.0%	17126	32.1%
Senior 61-90+	2660	4.3%	4167	6.7%	1125	2.1%	4545	8.5%

**External factors are environment and weather conditions.

3.4.2 The Decision Tree

A decision tree is a predictor, h: X ⟶ Y, of the predecessors of an event x by spanning a tree from its root node to its leaves. For simplicity, we concentrated on the binary classification case, namely, Y = {0, 1}, but decision trees can be used for a range of prediction problems. Based on the division of the input space, the successor child is chosen at each node along the root-to-leaf path. Usually, the splitting is based on one of x's properties or a predefined set of splitting rules as follows:

1) First, set the domain set: X is the accident event that needs to be labeled.

   Set X to be binary {1,0}, and let Y be our possible labels.

   Then, Y = {0, 1}, where 1 and 0 represent the possible options.

2) Training set $S = \big((X1, Y1) \dots (Xn, Yn)\big)$ is a limited number of pairings in $X \times Y$, that is, a list of labeled domain points. This is the information to which the learner has access.

3) For the output, the learner is asked to generate a prediction rule, h: X ⟶ Y. This function is also referred to as a prediction, hypothesis, or classifier. The predictor can forecast new domain elements (Ben-David, 2014).

Decision trees comprise three parts: decision nodes, branches, and leaf nodes. Each decision node in the structure displays the variable, and each branch displays one variable value based on decision rules; the leaf nodes display the expected values of the target variables (Song & Ying, 2015). We used Orange 3.30 software (Demšar et al., 2013) to run the CART decision tree set classification to stop at when majority reach 95% and limit of maximum tree depth is 7. Data was divided into two flows, HW and RR, and extracted 27 binary categorical variables that were most relevant to the 115,154 single-rider motorcycle accidents in Thailand from 2015 to 2020; the variables were set as binary (1 or 0) to facilitate interpretation and classification. Table 3.3 presents the 27 most relevant variables related to single-rider accidents under the following factors: roadway factors, external (environment, weather) and internal (driver behaviors) factors, driver data, and driver status.

Table 3.3 The Measurement Categories for the 27 Identified Motorcycle Accident Variables

Factor and Variables	Measurement
Roadway	
Dry road	1 - Yes, 0-Otherwise
Straight road	1 - Yes, 0-Otherwise
Obstruction	1 - Yes, 0-Otherwise
Road conditions	1 - Yes, 0-Otherwise
Vehicle conditions	1 - Yes, 0-Otherwise
External Factors (Environment and Weather Conditions)	
Day Time (06.00–18.00)	1 - Yes, 0-Otherwise
Night with light	1 - Yes, 0-Otherwise
Night without light	1 - Yes, 0-Otherwise
Low visibility	1 - Yes, 0-Otherwise
Clear weather	1 - Yes, 0-Otherwise
Internal Factors (Driver Behaviors)	
Drunk	1 - Yes, 0-Otherwise
Over speed limit	1 - Yes, 0-Otherwise
Ran a traffic light	1 - Yes, 0-Otherwise
Ran a traffic sign	1 - Yes, 0-Otherwise
Passing (overtaking)	1 - Yes, 0-Otherwise

Table 3.3 The Measurement Categories for the 27 Identified Motorcycle Accident Variables (Continued)

Factor and Variables	Measurement
Used a mobile phone	1 - Yes, 0-Otherwise
Short cutoff	1 - Yes, 0-Otherwise
Used drugs	1 - Yes, 0-Otherwise
Drove in opposite direction	1 - Yes, 0-Otherwise
Dozed off	1 - Yes, 0-Otherwise
Overweight cargo	1 - Yes, 0-Otherwise
Inconclusive	1 - Yes, 0-Otherwise
Driver Data	
Gender	1- Male, 0-Otherwise
Youth 15–35	1 - Yes, 0-Otherwise
Adult 36–60	1 - Yes, 0-Otherwise
Senior 61–90+	1 - Yes, 0-Otherwise
Driver Status	
Fatality	1 – Yes, 0-Otherwise

3.4.3 Performance Measurement

To assess the performance of the supervised machine learning decision tree in this study, we used tests data to validation how well the model performed with a confusion matrix with the following components: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). For example, TP shows the number of

positive values projected to be positive, whereas FP predicts an accident as fatal when it was not; recall and precision were also measured. These strategies are especially useful for unbalanced data sets, in which one answer category accounts for the bulk of the responses.

Precision refers to the accuracy of the classifier findings, expressed as follows and shown in Figure 3.3:

$$Precision = \frac{\mathrm{TP}}{TP+FP} \tag{3.1}$$

Recall, or sensitivity, gives the proportion of the positive class that was correctly classified, expressed as follows:

$$Recall = \frac{\mathrm{TP}}{TP+FN} \tag{3.2}$$

The TN rate, also called specificity, is computed as follows:

$$TNR = \frac{\mathrm{TN}}{\mathrm{TP+FN}} \tag{3.3}$$

The FP rate shows how often the classifier misclassified the negative class and is computed as follows:

$$FPR = \frac{\mathrm{FP}}{TN+FP} = 1 - TNR\ (Specificity) \tag{3.4}$$

The ratio of correct classifications reflects the data accuracy and is calculated as below:

$$Accuracy = \frac{\mathrm{TP+TN}}{TP+TN+FP+FN} \tag{3.5}$$

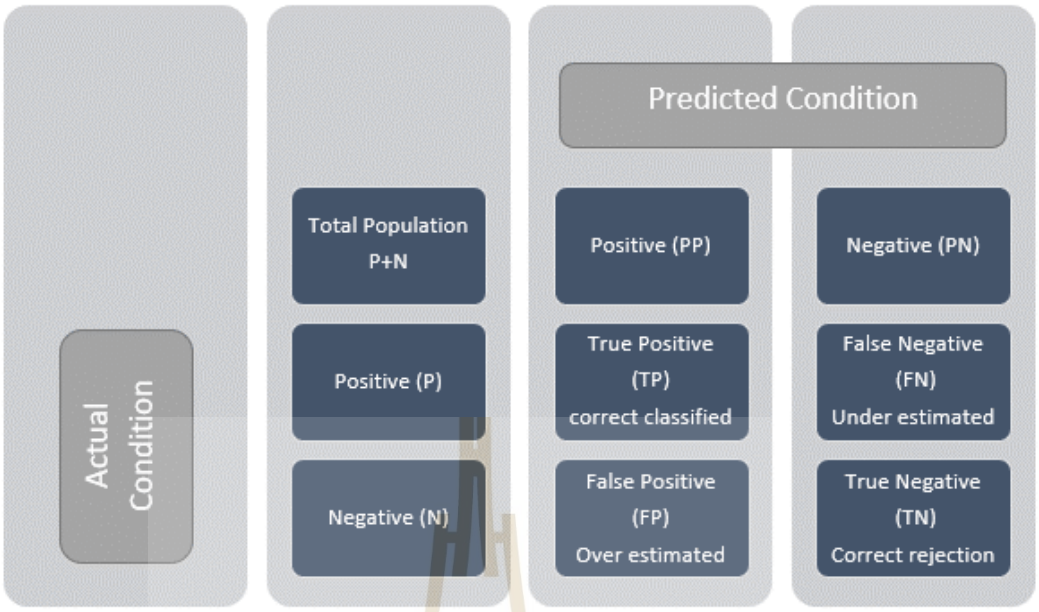


Figure 3.3 Diagram of the confusion matrix.

3.5 Results

Figure 3.4 presents plots of the HW and RR likelihoods of fatalities over 24 hours.

HWs

1 = fatality: μ =13.47, σ = 6.34

0 = nonfatality: μ =13.71, σ = 6.31

RRs

1 = fatality: μ =13.14, σ = 7.11

0 = nonfatality: μ =13.89, σ = 6.29

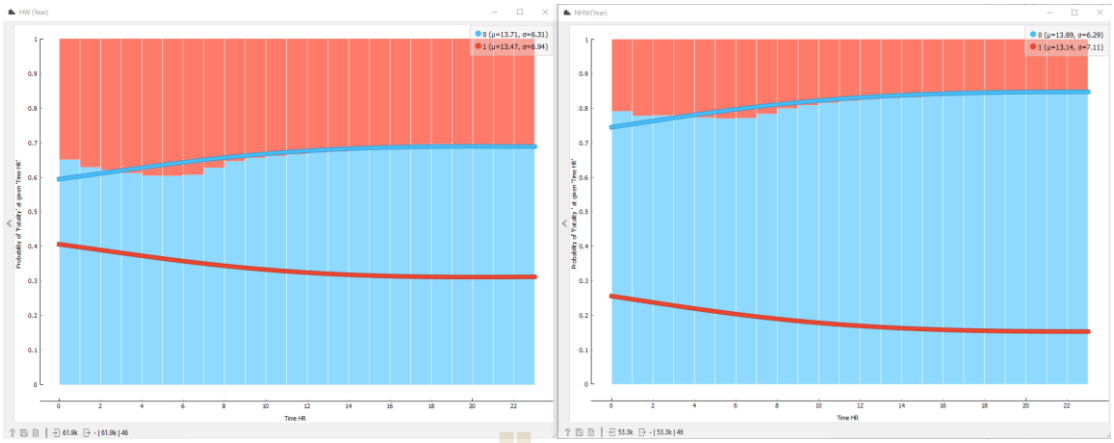


Figure 3.4 HW and RR fatality probabilities at different times of the day.

There is a higher probability of a fatality on a HW than on a RR: HW, 0.3–0.4 and RR, 0.25–0.15. However, both RRs and HWs have higher fatality rates at night (00.00–07.00). Both the HW and the RR decision trees were set to target rider fatalities, and they identified the main causes. The tree node was divided into whether the rider was following or exceeding the speed limit.

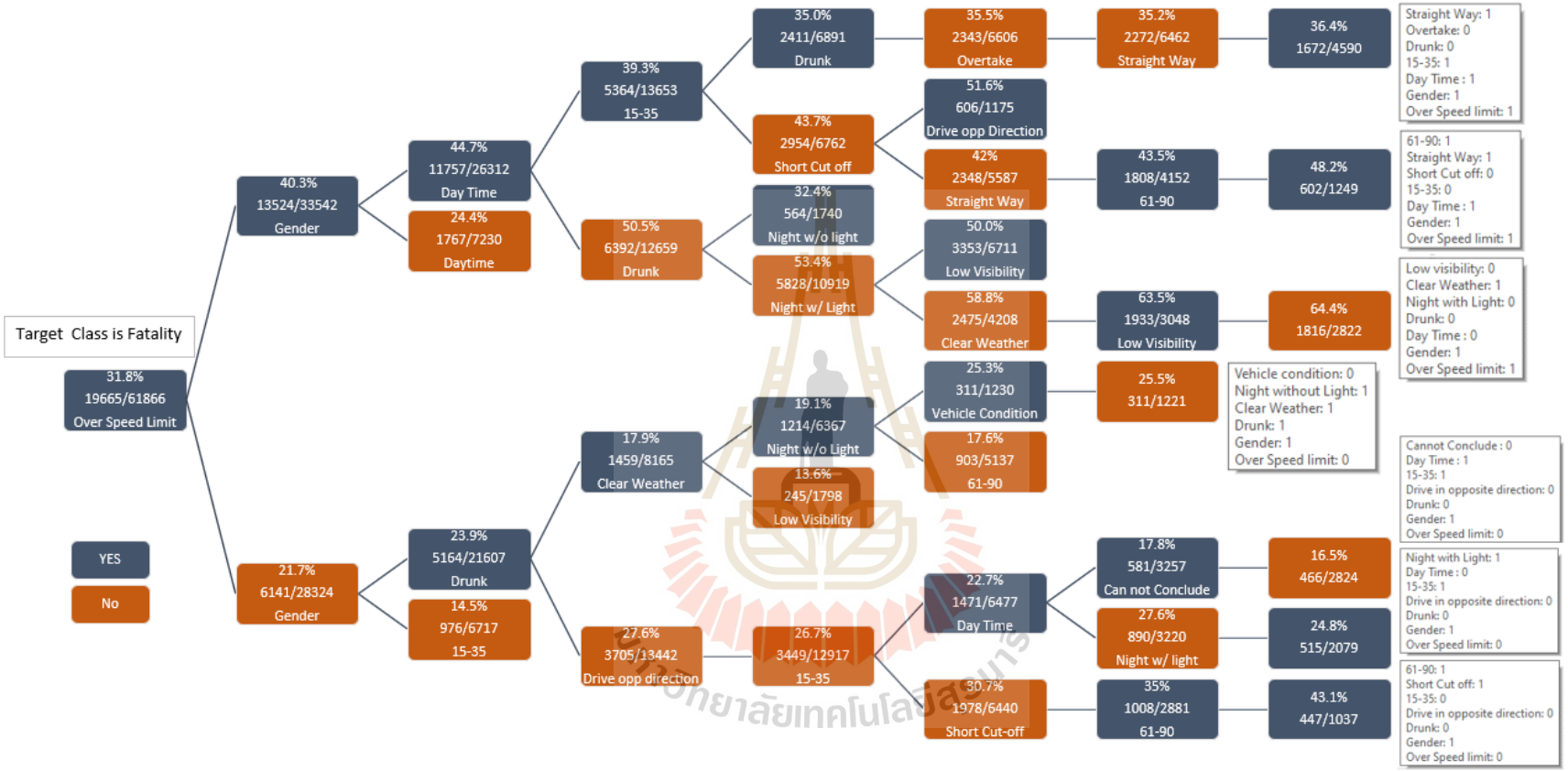


Figure 3.5 The HW tree model.

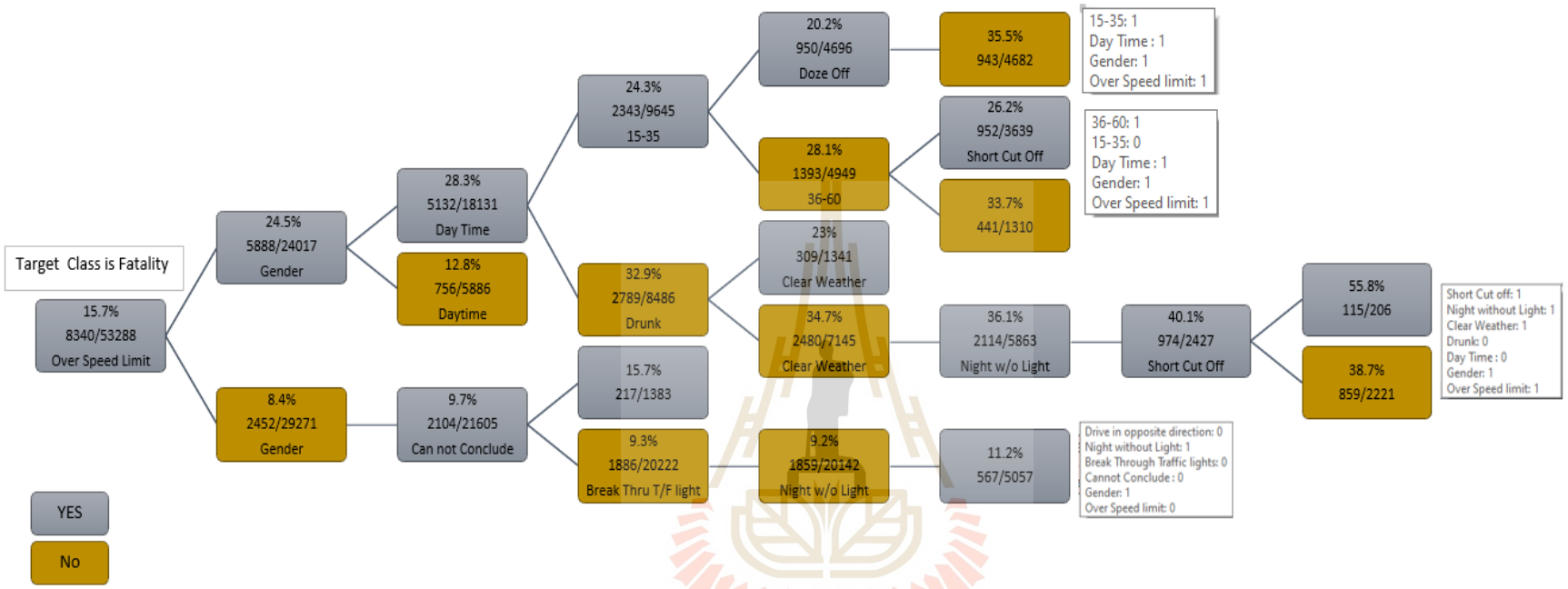


Figure 3.6 The RR tree model.

HW Fatalities (Figure 3.5):

1) Rider exceeds the speed limit (40.3%: 13,524/33,542): male (44.7%: 11,757/26,312), age 15–35 years (35%: 2,411/6,891), age 61–90 years (48.2%: 1,808/4,152), straight road (43.5%: 1,808/4,152), daytime (39.3%: 5,364/13,653), clear weather (63.5%: 1,933/3,048).
2) Rider does not exceed the speed limit (21.7%: 6,141/28,324): male (23.9%: 5,164/21,607), age 15–35 years (22.7%: 1,471/6,477), age 61–90 years (43.1%: 447/1,037), drunk (17.9%: 1,459/8,165), daytime (17.8%: 581/3,257), night w/o light (25.3%: 311/1,230), clear weather (19.1%: 1,214/6,367), short cutoff (35%: 1,008/2,881). On HWs, fatalities occurred most among male riders who were between the ages of 15 and 35 years and were exceeding the speed limit on a straight road in clear weather during the day. For riders aged 61 to 90 years, the fatalities occurred most often at night, without lights, under rider intoxication, and with short cutoffs.

RR Fatalities (Figure 3.6):

1) Rider exceeds the speed limit (24.5%: 5,888/24,017): male (28.3%: 5,132/18,131), age 15–35 years (20.2%: 950/4,696), age 36–60 years (26.2%: 952/3,639), daytime (24.3%: 2,343/9,645), night w/o light (40.1%: 974/2,427), short cutoff (55.8%: 115/206), clear weather (36.1%: 2,114/5,863).
2) Rider does not exceed the speed limit: male (9.7%: 2,104/21,605), night w/o light (11.2%: 567/5,057).

Table 3.4 presents the final-level leaf sets for HW and RR motorcycle accident fatalities according to whether the driver was exceeding the speed limit. For instance, for RR fatalities, the common factors in both age groups were being male and riding over the speed limit during the day; at night on RRs, male riders who died were most often speeding and making short cutoffs at night with no light. When drivers were not speeding, most fatalities occurred among males at night with no light.

Short cutoffs were among the most common causes of fatalities on both HWs and RRs, but excess speed was also a factor only on RRs. Gender was the most

significant variable in fatalities: Most fatalities were among male riders on both HWs and RRs irrespective of the rider speed limit, consistent with earlier findings that men who ride motorcycles are more likely to be involved in serious accidents Ospina-Mateus et al. (2019).

Table 3.4 The Final HW and RR Sets by Rider Speed

	HW Fatalities (Figure 4):	RR Fatalities (figure 5):
Rider exceeds the speed limit	Set 1: (15-35, Day time, Male, Straightway) 36.4% (1,672/4,590)	Set 1: (15-35, Day time, Male) 35.5% (943/4,682)
	Set 2: (61-90, Day time, Male, Straightway) 48.2% (602/1,249)	Set 2: (36-60, Day time, Male) 26.2% (952/3,639)
	Set 3 (Clear Weather, Male) 64.4% (1,816/2,822)	Set 3: (Clear weather, short cutoff, Night without light, Male) 55.8% (115/206)
Rider *does not exceed* the speed limit	Set 1: (Clear Weather, Male, Drunk) 25.5% (311/1,221)	Set 1: (Night without light, Male) 11.2% (567/5,057)
	Set 2: (Daytime,15-35, Male) 16.5% (466/2,824)	
	Set 3: (Night with light,15-35, Male) 24.8% (515/2,079)	
	Set 4: (Short cutoff,61-90, Male) 43.1% (477/1,037)	

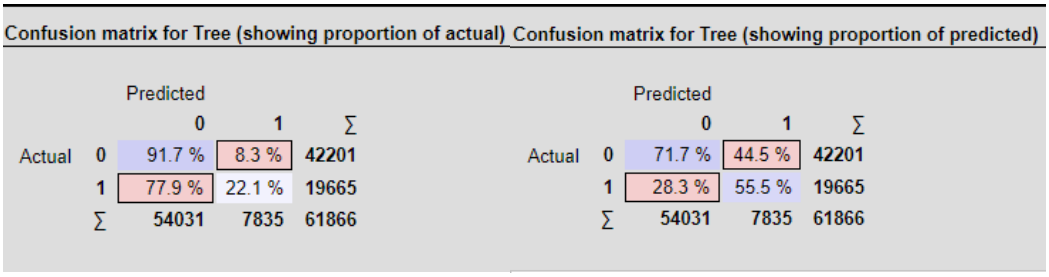
3.5.1 Evaluation Results

Table 3.5 presents the cross-validated data results with 20 folds. The table presents the area under the receiver-operating curve (AUC) and the classification accuracy (CA; Equation 5), recall (Equation 2), and precision (Equation 1). CA is high for HWs and RRs. When 0.5 < AUC < 1, there is a good possibility that the classifier will be able to differentiate between positive and negative class values. This is because the classifier is better able to recognize TP and TN (Equation 3) than FN and FP (Equation 4).

Table 3.5 The Model Evaluation Results

Model	Road	Target Class	AUC	CA	Precision	Recall
Tree	HW	Avg over	0.685	0.696	0.665	0.696
		Fatality	0.686	0.696	0.555	0.221
		Nonfatality	0.686	0.696	0.717	0.917
	RR	Avg Over	0.706	0.842	0.776	0.842
		Fatality	0.703	0.842	0.400	0.021
		Nonfatality	0.703	0.842	0.845	0.994

The HW confusion matrix in Figure 3.7 shows misclassifications in 77.9% of actual fatalities to nonfatality accidents and 8.3% of nonfatalities to fatalities. For the predicted values, we identified a misclassification of 28.3% for accident fatalities to nonfatalities and 44.5% for nonfatalities to fatalities. Figure 3.8 shows the RR confusion matrix, reflecting misclassifications of 97.9% (fatality to nonfatality) and 0.6% (nonfatality to fatality) for the actual cases and of 15.5% (fatality to nonfatality) and 60% (nonfatality to fatality). That is, the decision tree for this study shows significant misclassifications of fatalities for both HWs and RRs but much better ability with nonfatalities.



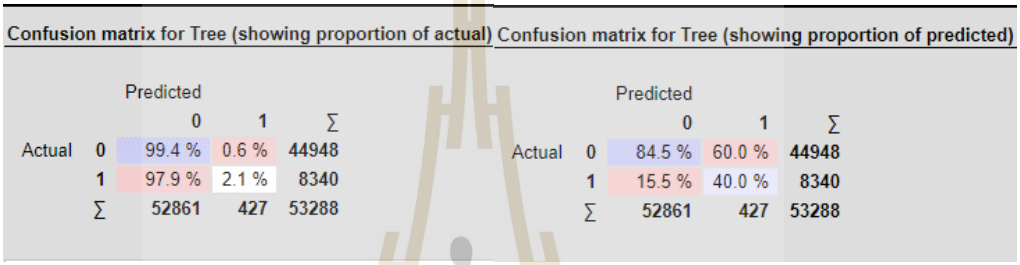
Confusion matrix for Tree (showing proportion of actual)

		Predicted 0	Predicted 1	∑
Actual	0	91.7 %	8.3 %	42201
	1	77.9 %	22.1 %	19665
	∑	54031	7835	61866

Confusion matrix for Tree (showing proportion of predicted)

		Predicted 0	Predicted 1	∑
Actual	0	71.7 %	44.5 %	42201
	1	28.3 %	55.5 %	19665
	∑	54031	7835	61866

Figure 3.7 The confusion matrix actual and predicted results for HWs.



Confusion matrix for Tree (showing proportion of actual)

		Predicted 0	Predicted 1	∑
Actual	0	99.4 %	0.6 %	44948
	1	97.9 %	2.1 %	8340
	∑	52861	427	53288

Confusion matrix for Tree (showing proportion of predicted)

		Predicted 0	Predicted 1	∑
Actual	0	84.5 %	60.0 %	44948
	1	15.5 %	40.0 %	8340
	∑	52861	427	53288

Figure 3.8 The confusion matrix actual and predicted results for RRs.

3.6 Conclusion and Discussion

The aim of this research was to design a decision tree to identify individual contributors to motorcycle accident fatalities among riders in Thailand, with a focus on single-rider crashes. Contributing variables included roadway features along with external and internal (driver-related) factors. In addition, using accident data from 2015 to 2020, we performed a nonparametric analysis to determine the importance of factors that influence target variables, such as road and weather conditions, speeding, being on a straightaway (with no intersections), gender, and substance use.

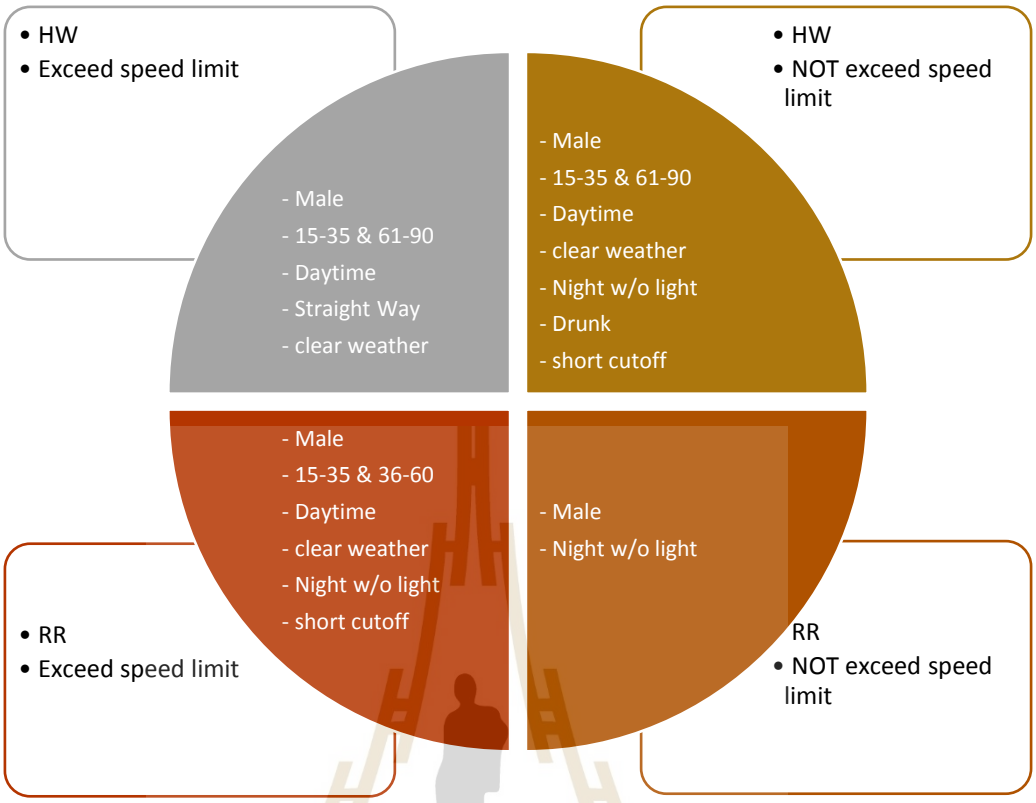


Figure 3.9 Key accident factors: HWs versus RRs.

The decision tree concluded that the most common causes of fatality on both HWs and RRs were being a male rider and exceeding the speed limit, with the other variables showing differing levels of importance (Figure 3.9). HWs have more contributors to fatalities than RRs; for instance, accidents were common on HWs when riders had been drinking, especially at night with no light. HWs also have much heavier traffic and a wider variety of vehicles traveling at higher speeds than RRs, facilitating accidents that can cause serious injury and death. Age was another significant contributor to motorcycle accident fatalities on both types of roads, although notably, only HW fatalities extended to riders up to the age of 90 years; the highest age in RR fatalities was 60 years. One interesting observation here was that RR fatalities generally occurred at night with no street lighting whether a rider was speeding. Road accidents have many contributing factors, but speeding is a key area of concern for the severity of road accident injuries (M. Yu et al., 2020), (Osman et al., 2018), (Krull et al., 2000). Thus, we propose that Thailand must properly post speed restrictions and support the

enforcement of compliance within these limits. Short-cutoff riding was another key predictor of motorcycle fatalities Bahiru et al. (2018). Although male gender was a primary factor in HW motorcycle fatalities in this study, we found less influence from age, accident location, and vehicle type. However, substance use was a factor in many accidents, and we propose more education on the dangers of riding while intoxicated in addition to tighter enforcement of penalties for infractions.

In Thailand, persons of all ages ride motorcycles, although most riders are between the ages of 15 (bike capacity of no more than 110cc) and 35 years. Looking at all four accident scenarios in the research, three featured young people, consistent with Zhang and Fan (2013), in which accidents were more likely among young people (25 years old), who are less disciplined and unfamiliar with traffic laws and have less driving experience. Policymakers might consider raising the minimum age for obtaining a motorcycle license to at least 18 years or imposing further restrictions on engine size dependent on rider age. We also identified road lighting as a considerable factor in motorcycle accidents, particularly on RRs but in fact in all accident categories except for speeding-related deaths on HWs. Therefore, we propose that better lighting be installed wherever possible on Thai roadways, particularly in rural areas.

3.7 Limitations and Future Studies

This study's model showed acceptable (above 50%) accuracy, but there is room for improvement; adjusting the parameters in a future study could increase the accuracy. Additionally, we used accident data from 2015 to 2020, but during the last two years, 2019 and 2020, the circumstances in Thailand as well as around the world changed drastically overnight because of the COVID-19 pandemic. Governments worldwide locked down and ordered people to stay indoors, and Thailand limited travel between provinces, particularly between the hours of 22.00 and 04.00. Because mobility was so limited during 2019 and 2020, the overall findings for those years might not accurately reflect what would have been the country's true numbers of accidents and fatalities.

3.8 References

Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. *Expert Systems with Applications*, *40*(15), 6047-6054. https://doi.org/10.1016/j.eswa.2013.05.027

Al Mamlook, R. E., Ali, A., Hasan, R. A., & Mohamed Kazim, H. A. (2019). Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation. Proceedings of the IEEE National Aerospace Electronics Conference, NAECON,

Anvari, M. B., Tavakoli Kashani, A., & Rabieyan, R. (2017). Identifying the Most Important Factors in the At-Fault Probability of Motorcyclists by Data Mining, Based on Classification Tree Models. *International Journal of Civil Engineering*, *15*(4), 653-662. https://doi.org/10.1007/s40999-017-0180-0

Bahiru, T. K., Kumar Singh, D., & Tessfaw, E. A. (2018). Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018,

Ben-David, S. S.-S. a. S. (2014). <understanding-machine-learning-theory-algorithms.pdf>. *Cambridge University Press.* http://www.cs.huji.ac.il/~shais/UnderstandingMachine Learning

Bhavsar, R., Amin, A., & Zala, L. (2021). Development of Model for Road Crashes and Identification of Accident Spots [Article]. *International Journal of Intelligent Transportation Systems Research*, *19*(1), 99-111. https://doi.org/10.1007/s13177-020-00228-z

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., & Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia,

Champahom, T., Jomnonkwao, S., Chatpattananan, V., Karoonsoontawong, A., & Ratanavaraha, V. (2019). Analysis of Rear-End Crash on Thai Highway: Decision Tree Approach. *Journal of Advanced Transportation*, *2019*, 1-13. https://doi.org/10.1155/2019/2568978

Chen, M.-Y. (2012). Comparing Traditional Statistics, Decision Tree Classification And Support Vector Machine Techniques For Financial Bankruptcy Prediction. *Intelligent Automation & Soft Computing*, *18*(1), 65-73. https://doi.org/10.1080/10798587.2012.10643227

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., . Zupan, B. (2013). Orange: Data mining toolbox in python [Article]. *Journal of Machine Learning Research*, *14*, 2349-2353. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885599052&partnerID=40&md5=75d2df52a0c46b5ab58ab08e1576114e

DLT. (2021). Department of Land Transportation. *https://www.dlt.go.th/th/public-news/view.php?_did=2806*.

El Abdallaoui, H. E. A., El Fazziki, A., Ennaji, F. Z., & Sadgal, M. (2018). Decision Support System for the Analysis of Traffic Accident Big Data. Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018,

Feng, M., Zheng, J., Ren, J., & Xi, Y. (2020). Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. In *Advances in Brain Inspired Cognitive Systems* (pp. 520-529). https://doi.org/10.1007/978-3-030-39431-8_50

Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests [Article]. *Accident Analysis and Prevention*, *41*(1), 98-107. https://doi.org/10.1016/j.aap.2008.09.009

Helen, W. R., Almelu, N., & Nivethitha, S. (2019). Mining Road Accident Data Based on Diverted Attention of Drivers. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018,

Jafari Anarkooli, A., Hosseinpour, M., & Kardar, A. (2017). Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. *Accident Analysis & Prevention*, *106*, 399-410. https://doi.org/10.1016/j.aap.2017.07.008

Jomnonkwao, S., Uttra, S., & Ratanavaraha, V. (2020). Forecasting Road Traffic Deaths in Thailand: Applications of Time-Series, Curve Estimation, Multiple Linear Regression, and Path Analysis Models. *Sustainability*, *12*(1). https://doi.org/10.3390/su12010395

Jou, R. C., Yeh, T. H., & Chen, R. S. (2012). Risk factors in motorcyclist fatalities in Taiwan. *Traffic Inj Prev*, *13*(2), 155-162. https://doi.org/10.1080/15389588.2011.641166

Khorashadi, A., Niemeier, D., Shankar, V., & Mannering, F. (2005). Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. *Accid Anal Prev*, *37*(5), 910-921. https://doi.org/10.1016/j.aap.2005.04.009

Kim, J.-K., Ulfarsson, G. F., Kim, S., & Shankar, V. N. (2013). Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention*, *50*, 1073-1081. https://doi.org/https://doi.org/10.1016/j.aap.2012.08.011

Krull, K. A., Khattak, A. J., & Council, F. M. (2000). Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes. *Transportation Research Record*, *1717*(1), 46-54. https://doi.org/10.3141/1717-07

Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations [Article]. *Journal of Modern Transportation*, *24*(1), 62-72. https://doi.org/10.1007/s40534-016-0095-5

Kuşkapan, E., Çodur, M. Y., & Atalay, A. (2021). Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms [Article]. *Accident Analysis and Prevention*, *155*, Article 106098. https://doi.org/10.1016/j.aap.2021.106098

Mafi, S., AbdelRazig, Y., & Doczy, R. (2018). Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. In *Transportation Research Record* (Vol. 2672, pp. 171-183).

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accid Anal Prev*, *122*, 181-188. https://doi.org/10.1016/j.aap.2018.10.014

Mphela, T. (2020). Causes of road accidents in botswana: An econometric model [Article]. *Journal of Transport and Supply Chain Management*, *14*, 1-8, Article a509. https://doi.org/10.4102/jtscm.v14i0.509

Osman, M., Mishra, S., & Paleti, R. (2018). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*. https://doi.org/10.1016/j.aap.2018.05.004

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists [Article]. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051-10072. https://doi.org/10.1007/s12652-020-02759-5

Ospina-Mateus, H., Quintana Jiménez, L. A., López-Valdés, F. J., Morales-Londoño, N., & Salas-Navarro, K. (2019). Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia. In *Communications in Computer and Information Science* (Vol. 1052, pp. 309-320).

Pakgohar, A., Tabrizi, R. S., Khalili, M., & Esmaeili, A. (2011). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. *Procedia Computer Science*, *3*, 764-769. https:// doi.org/10.1016/j.procs.2010.12.126

PDPM. (2020). Thailand Department of Public Disaster Prevention and Mitigation. *https://www.disaster.go.th/en/*

Recal, F., & Demirel, T. (2021). Comparison of machine learning methods in predicting binary and multi-class occupational accident severity [Article]. *Journal of Intelligent and Fuzzy Systems*, *40*(6), 10981-10998. https://doi.org/10.3233/JIFS-202099

Rezapour, M., Mehrara Molan, A., & Ksaibati, K. (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. *International Journal of Transportation Science and Technology*, *9*(2), 89-99. https://doi.org/10.1016/j.ijtst.2019.10.002

RSC, T. (2019). Thailand Accident Research Center *Thailand Accident Research Center* https://www.thairsc.com/

Se, C., Champahom, T., Jomnonkwao, S., Chaimuang, P., & Ratanavaraha, V. (2021). Empirical comparison of the effects of urban and rural crashes on motorcyclist injury severities: A correlated random parameters ordered probit approach with heterogeneity in means. *Accid Anal Prev*, *161*, 106352. https://doi.org/10.1016/j.aap.2021.106352

Shaheed, M. S., Gkritza, K., Zhang, W., & Hans, Z. (2013). A mixed logit analysis of two-vehicle crash seventies involving a motorcycle. *Accident; analysis and prevention*, *61*. https://doi.org/10.1016/j.aap.2013.05.028

Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., & Hanks, H. (2011). Risk factors for fatal crashes in rural Australia. *Accident Analysis & Prevention*, *43*(3), 1082-1088. https://doi.org/https://doi.org/10.1016/j.aap.2010.12.016

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

WHO. (2018). World Health Organization: Global status report on road safety 2018. . *https://extranet.who.int/roadsafety/death-on-the-roads/*.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Xie, Y., & Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. *Accident; analysis and prevention*, *47*, 36-44. https://doi.org/10.1016/j.aap.2011.12.012

Yu, M., Zheng, C., & Ma, C. (2020). Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *27*, 100126. https://doi.org/https://doi.org/10.1016/j.amar.2020.100126

Zhang, X. F., & Fan, L. (2013). A decision tree approach for traffic accident analysis of Saskatchewan highways. Canadian Conference on Electrical and Computer Engineering,

Zhou, M., & Chin, H. C. (2019). Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. *Accident Analysis & Prevention*, *124*, 104-112. https://doi.org/10.1016/j.aap.2019.01.009

CHAPTER 4
COMPARISON OF MACHINE LEARNING PREDICTABILITY PERFORMANCE: THE CASE OF MOTORCYCLE ACCIDENT IN THAILAND

4.1 Abstract

Every year in Thailand and other place around the globe, traffic accidents kill, injure, and kill, causing millions of deaths, injuries, and fatalities as well as billions of dollars in economic damages. Accurate models for predicting the severity of traffic accidents are essential for transportation systems. This inquiry attempt provides methods for selecting a collection of influential criteria and developing a model for categorizing the severity of injuries. Various supervised machine learning methods approaches are used such as, Decision Tree (DT), Support Vector Machine (SVM), Random Forests (RF), K-Nearest Neighbors (kNN), Neural Network (NN), Naive Bayes (NB) Logistic Regression (LR), and Gradient Boosting (GB).

The researcher included information about the incidence of motorcycle accidents in Thailand for 5 years; a total of 112,837 events. The data includes several factors causing accidents and focuses solely on the motorcycle rider who were the cause of accidents excluding the passengers.

Random Forest (RF) outperformed the other seven ML algorithms in predicting road accidents in 2-Class classification giving an average over class AUC of 0.768, CA of 0.777, Precision of 0.752, and recall of 0.777. Which top 5 features giving highest info gain from RF consist of Highway, Riding over speed limit, Day time, Night w/o light and gender. The ML is an effective tool for predicting accidents; the model performs well in terms of non-fatality prediction, but there is still room for improvement in fatality prediction, which can be interpreted to mean that the factors that cause fatal accidents may cause an accident but do not always result in a fatality.

4.1.1 Highlights

1) Random Forest (RF) outperformed the other seven ML algorithms in predicting motorcycle accidents

2) All ML models for this study found significant misclassifications of fatalities but a much better ability to predict nonfatalities

3) Most accidents occur on highway, Riding over speed limit, Day time, Night w/o light and gender

4.2 Introduction

Thailand has the highest number of road fatalities, ranking in the top ten worldwide. Thais die in traffic crashes at a rate of 32.7 for each 100,000 people. (WHO, 2018). Motorcycles comprise half of all 41 million registered vehicles. (DLT, 2021), potentially contributing to the highest number of fatalities from major accidents. As Jomnonkwao et al. (2020) observed, motorcyclists are responsible for most road fatalities., while prior studies showed different types of cars in difference country, such as rollover SUV/vans (Jafari Anarkooli et al., 2017), large truck (Huo et al., 2020; Jafari Anarkooli et al., 2017; Li et al., 2018), and pick-ups (Li et al., 2018). Despite government efforts to increase law enforcement for drink driving, speed control, engineering solutions for road safety, and so on, the number of road fatalities has remained consistent at 32-35 percent from 2015 to 2020 (PDPM, 2020). The current policy for reducing road fatalities appears to be ineffective. Machine learning is one of the new solutions for predicting and minimizing accidents that should be implemented.

Machine learning (ML) is the learning of computer algorithms that change automatically through education. It is pictured as the subset of artificial power which algorithms develop a science framework using sample information, called ‘preparation information’, in order to make prediction or conclusion without being explicitly programmed to do so. Why is ML comparison required for road accident prediction study? Because the ML model itself has different advantages and disadvantages, as shown in table 4.1 (Géron, 2019; Guido, 2017; Sarker, 2021; Wu et al., 2007). Understanding and analyzing the model is essential before deciding on a machine

learning algorithm. Before settling on a single machine learning algorithm, compare their accuracies on training and test sets.

Table 4.1 Comparison of the advantages and disadvantages of ML models

Model	PROs	CONs
SVM	- Performs admirably in higher dimensions - Outliers have less influence	- Does not perform well when classes overlap - Selecting appropriate hyperparameters is important
GB	- Running Quickly & interpret - High dimension of data are well handled	- Tuning parameters are difficult to find - If the parameter is not well tuned, it's easy to overfit
LR	- Simple and effective when the dataset can be divided linearly - Feature scaling is not required	- Inadequate performance on non-linear data - More commonly used for classification not regression - When used with high-dimensional datasets, it has the potential to overfit.
NN	- Can build extremely complex model for large datasets - Excellent for large data set w/ high dimension data - High prediction accuracy	- Scaling data and parameter selection are critical considerations - Time required for training

Table 4.1 Comparison of the advantages and disadvantages of ML models (Continued)

Model	PROs	CONs
Naïve Bayes	- Good performance with high-dimensional data - It only requires a small amount of training data to quickly estimate the required parameters.	- Training data should accurately represent the population
kNN	- It can be used for both classification and regression. - Can deal with multi-class problems - Very resistant to noisy training data	- Sensitive to outlier - Not handle well with missing data - The accuracy is determined by the quality of the data
Tree	- Missing value handling - Simple to explain and visualize - Can used for both the classification and regression tasks	- Prone to overfitting - Sensitive to data
RF	- Fits both categorical and continuous values well - Less overfitting - Useful to extract feature importance	- Not suitable for extremely high-dimensional sparse data

The primary objective of this research is to evaluate the performance of machine learning methods in classifying the severity of road accidents. (Fatality and Non-Fatality) and attempting to identify major factors in forecasting the severity of motorcycle accidents. The proposed study is unique in that it compared eight machine learning models in the same data set that only included motorcycle accidents involving the rider alone (no passenger or victim involved) in attempt to choose the model that predicts future accidents the most accurately.

4.3 Literature Review

According to international research, earlier studies on traffic accidents have been divided by group of factors that perceived to be related in traffic crashes.

4.3.1 Traffic Study Using Machine Learning

According to Table 4.2 shows a previous study that used several machine learning models to investigate traffic accidents; The studies compared prediction accuracy of ML models and discovered feature importance as it relates to road accidents.

Table 4.2 Machine learning models in traffic accident study

Author	Methodology											
	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic	Neural Network	Naïve Bayes	Random Forest	Support Vector Machine
Ospina-Mateus et al. (2021)	-	-	-	✓	-	-	✓	-	✓	✓	✓	✓
Kumar and Toshniwal (2016)	✓	-	-	-	-	✓	-	-	-	-	-	-
Harb et al. (2009)	-	-	-	✓	-	-	-	-	-	-	✓	-
Mafi et al. (2018)	-	-	-	-	-	-	-	-	-	-	✓	-

Table 4.2 Machine learning models in traffic accident study (Continued)

Author	Methodology											
	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic	Neural Network	Naïve Bayes	Random Forest	Support Vector Machine
Kuşkapan et al. (2021)	-	-	-	-	-	-	✓	-	-	✓	-	✓
Al Mamlook et al. (2019)	-	✓	✓	✓	-	-	✓	-	-	✓	✓	✓
Recal and Demirel (2021)	-	-	-	✓	✓	-	-	✓	✓	-	-	✓
Bahiru et al. (2018)	-	-	-	✓	-	-	-	-	-	✓	-	-
Ospina-Mateus et al. (2021)	-	-	-	✓	-	-	✓	-	✓	✓	✓	✓
Feng et al. (2020)	✓	-	-	-	-	-	-	✓	-	-	-	-
Helen et al. (2019)	✓	-	-	-	-	✓	-	-	-	-	-	-
(Santos et al., 2021)	-	-	-	✓	-	-	-	-	-	✓	✓	-
(Kim et al., 2021)	-	-	-	-	✓	-	-	-	✓	-	✓	-

4.3.2 Driver/Rider information

When the rider was not at wrong, Thailand cases discovered that the rider's age was the most significant element. (Champahom et al., 2019). Riders between the ages of 18 and 24 are lacking the driving experience, such as adjusting the speed to accommodate different roadways. (Bucsuházy et al., 2020). Motorcyclists aged 20-39 are more likely to engage in serious crashes, however when no motorbike or bicycle is engaged, the magnitude is likely to be mild, and men are involved in serious crashes than women. (Ospina-Mateus et al., 2019).

4.3.3 Roadway

Poor road conditions increase the likelihood of an accident, particularly on highways. (Malin et al., 2019). The characteristic of road, brightness, vehicle speed, and road conditions all have an impact on the frequency of accidents. (Feng et al., 2020). Roads that were dark or dim also played important roles in road accidents. (Shweta et al., 2021). Highway intersections have also been recognized as the ones most risky for all types of accidents. (Kumar & Toshniwal, 2016). Highways had an impact on injury severity outcomes in rural motorcycle crashed (Geedipally et al., 2011). Around 92.5% of the crashes on motorcycle occurred on dry roadway surfaces. (Shaheed et al., 2013)

4.3.4 Internal Factor (Driver/Rider Behavior)

Rider characteristics increase the risk of serious and fatal injury in motorcycle accidents (Cunto & Ferreira, 2017). Intoxicated drivers have a higher accident rate than other drivers. (Helen et al., 2019) and the most important factor of an injury is speeding. (Al Mamlook et al., 2019). Most alcohol-related accidents are caused by young people (35 years old) late at night. (John & Shaiba, 2022). Motorcycles are more dangerous in rural areas. Male riders, exceeding the posted speed limit, overtaking, and exhaustion are all contributing factors to serious and fatal injuries. (Mohamad et al., 2022)

4.3.5 External Factor (Environment & Weather Condition)

More than two-thirds of the two-vehicle motorcycle crashes reported occurred in clear weather (because clear weather encourages motorcycle riding), while one-quarter occurred in cloudy or partly cloudy conditions. Approximately 80% of the

crashes occurred during the day, while nearly one-fifth of the crashes occurred at night. These findings are most likely due to the increased exposure of motorcycles in daylight versus at night, as well as the higher associated crash risk. (Shaheed et al., 2013). Weather conditions, such as poor visibility, have a greater impact on traffic accidents than internal factors such as the driver. Sonal and Suman (2018). Driving at night increases the likelihood of a car accident. (Mphela, 2020). Traveling at night increases the likelihood of a car accident. (Mphela, 2020).

4.4 Methodology and Data

The framework of the article is known as data mining and ML techniques. (Figure 1), and it measures the performance of ML model predictability for fatalities and non-fatalities on road accident as following steps.

Initial Dataset – endorsed for identifying and fixing incomplete and imprecisely collected data, in addition to demonstrating data integrity after the data set has been purified.

Verified Dataset - Data partitioning to binary mode and set Fatal/non-fatal as target.

Data Splitting – Separated test and train data set to ratio 75:25.

Model Learning -Allows the machine to learn with 75% of the test dataset and later test with the leftover 25%.

Model Prediction and Measurement – To check the prediction accuracy of each model.

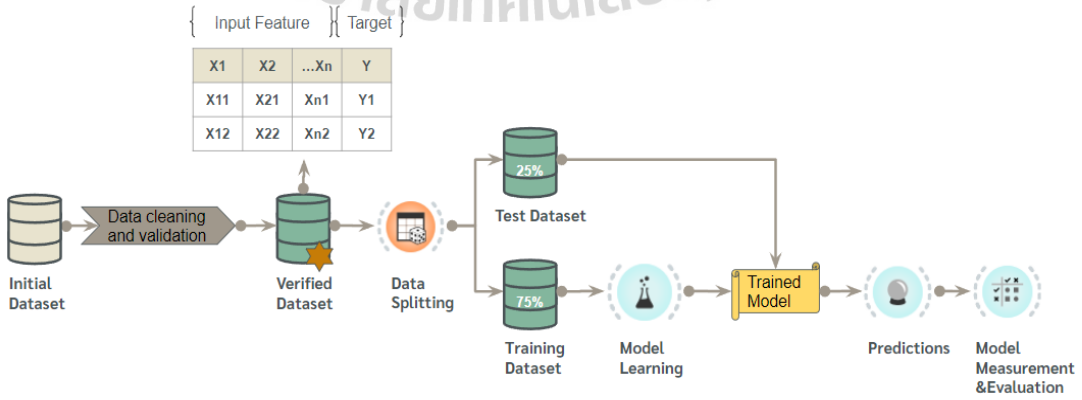


Figure 4.1 Machine learning Process flow

4.4.1 Data Description

The study begins with data on road accidents combined all kind of road, which totaled 112,837 incidents between 2015 and 2020 (PDPM, 2020), with the focus solely on rider who were primary caused in an accident (Table 4.3).

Table 4.3 Categorical Attribute and descriptive statistics.

Accident Event (Attribute)	Fatality			
	Yes		No	
	Count	%	Count	%
Roadway				
Dry Surface Road	26609	23.6%	81459	72.2%
Wet Surface	920	0.8%	3849	3.4%
Straight Way	19460	17.2%	59907	53.1%
Not straight Way (Curve, Slope, Junction, etc.)	8069	7.2%	25401	22.5%
Obstruction	504	0.4%	3080	2.7%
Road condition	374	0.3%	2001	1.8%
Vehicle condition	298	0.3%	2116	1.9%
Highway	19325	17.1%	41403	36.7%
Non-Highway	8204	7.3%	43905	38.9%
External Factor (Environment & Weather Condition)				
Day Time (06.00-18.00)	13439	11.9%	49845	44.2%
Night with Light	7631	6.8%	20824	18.5%
Night without Light	6459	5.7%	14639	13.0%
Low visibility	2427	2.2%	10534	9.3%
Clear Weather	24362	21.6%	73456	65.1%
Not Clear Weather (Rain, fog, etc.)	3167	2.8%	11852	10.5%

Table 4.3 Categorical Attribute and descriptive statistics. (Continued)

Accident Event (Attribute)	Fatality			
	Yes		No	
	Count	%	Count	%
Internal Factor (Driver Behavior)				
Drunk	3670	3.3%	19928	17.7%
Over Speed limit	19121	16.9%	37338	33.1%
Break Through Traffic lights	228	0.2%	456	0.4%
Break Through Traffic Signs	373	0.3%	1299	1.2%
Overtake	641	0.6%	1238	1.1%
Use Mobile Phone	23	0.0%	348	0.3%
Short Cut off	5315	4.7%	19059	16.9%
Drug	6	0.0%	65	0.1%
Drive in opposite direction	432	0.4%	816	0.7%
Doze off	347	0.3%	881	0.8%
Overweight Carry	16	0.0%	64	0.1%
Cannot Conclude	932	0.8%	3458	3.1%
Driver info				
Gender (male)	23744	21.0%	62152	55.1%
Gender (Female)	3785	3.4%	23152	20.5%
Youth 15-35	13671	12.1%	45605	40.4%
Adult 36-60	10141	9.0%	31171	27.6%
Senior 61-90+	3717	3.3%	8532	7.6%

4.4.2 Methodology

To predict accident severity, eight ML algorithms were separately implemented on binary classification types (fatal/non-fatal) using Orange3.30 python base software (Demšar et al., 2013) to run. To determine the best-performing algorithm, the performance of predictive models was compared, and the most important variables were extracted from the eight models listed below.

4.4.2.1 Decision Tree (DT)_is a non-parametric supervised algorithm for learning that can be used for classification as well as regression tasks. Decision trees are a popular and powerful way of modeling decisions in machine learning. They are a type of tree data structure that begins with a root node and divides into two branches at each node. For classification and regression tasks, decision trees are used. They can be used to estimate values for a target variable or to predict the probability of an event based on features.

4.4.2.2 K-Nearest Neighbors- is an algorithm for non-parametric classification and pattern recognition (Sarker, 2021). It is frequently used to divide data into two or more classes. This algorithm begins by assigning each training instance in the dataset to the training set's most similar training instance. The distance between two instances is calculated using some measure of similarity, and it can then be any number, such as the Euclidean distance. The procedure is repeated until all data points are assigned to a class.

4.4.2.3 Support Vector Machine (SVM) - SVMs (Cortes & Vapnik, 1995) are supervised learning algorithms that can classify both continuous and categorical data. The SVM's goal is to find the best hyperplane in the input space that separates two classes of data. The hyperplane should minimize the distance between each class's nearest point.

4.4.2.4 Random Forests (RF) - Random Forest (Breiman, 2001) is a machine learning technique that is used for classification and regression. It can be used for both supervised and unsupervised learning. It has been shown to be more accurate than other algorithms in some cases. The models will then vote to determine which class is the most popular.

4.4.2.5 Neural Network (NN), - Neural network is a type of artificial intelligence that mimics the human brain. They enable computers to learn from examples and make predictions based on data patterns. A neural network is composed of many interconnected layers, each of which contains a number of nodes linked to the next layer. Weighted or unweighted links can exist between nodes in adjacent layers. In general, a node will have an output value for each input it receives from its connections with other nodes in the preceding layer, which will then be used as an input for its connections with other nodes in the subsequent layer. (Dongare et al., 2012; Géron, 2019)

4.4.2.6 Naive Bayes (NB) - It is a classification which employs probability principles to aid in estimation or predict likelihood of an event. (Webb, 2010)

4.4.2.7 Logistic Regression (LR) - is a technique in statistics and machine learning that predicts the probability of an event occurring. It is mainly used in predicting binary outcomes Logistic regression is often used for classification tasks where the dependent variable can take on any value from a discrete set of values. (Tolles & Meurer, 2016)

4.4.2.8 Gradient Boosting (GB) - Gradient Boosting choosing an optimization method by attempting to obtain each new Classifier instance. It improves its accuracy by learning from the cumulative error generated by the previous instance's prediction. (Géron, 2019)

According to table 4.4, the overall 112,837 incidents include 27 attributes from data collection that span coverage Roadway, environment, weather condition, driving behavior, driver data, and driver status. When two classification types were compared, binary models (Fatal & Non-fatal) outperformed the 3-Class model (Fatal, Major, Minor), which can be explained by the inability to effectively separate major and minor accidents in contrast, fatal accidents behaved similarly in both classification types. (Recal & Demirel, 2021).

Table 4.4 Total 28 Attributes with setting description

Attribute Name	Attribute Description
Roadway	
Highway	1 - Yes, 0-Otherwise
Dry Surface Road	1 - Yes, 0-Otherwise
Straight Way	1 - Yes, 0-Otherwise
Obstruction	1 - Yes, 0-Otherwise
Road condition	1 - Yes, 0-Otherwise
Vehicle condition	1 - Yes, 0-Otherwise
External Factor (Environment and Weather Condition)	
Day Time (06.00-18.00)	1 - Yes, 0-Otherwise
Night with Light	1 - Yes, 0-Otherwise
Night without Light	1 - Yes, 0-Otherwise
Low visibility	1 - Yes, 0-Otherwise
Clear Weather	1 - Yes, 0-Otherwise
Internal Factor (Driver Behavior)	
Drunk	1 - Yes, 0-Otherwise
Over Speed limit	1 - Yes, 0-Otherwise
Break Through Traffic lights	1 - Yes, 0-Otherwise
Break Through Traffic Signs	1 - Yes, 0-Otherwise

Table 4.4 Total 28 Attributes with setting description (Continued)

Attribute Name	Attribute Description
Overtake	1 - Yes, 0-Otherwise
Use Mobile Phone	1 - Yes, 0-Otherwise
Short Cut off	1 - Yes, 0-Otherwise
Drug	1 - Yes, 0-Otherwise
Drive in opposite direction	1 - Yes, 0-Otherwise
Doze off	1 - Yes, 0-Otherwise
Overweight Carry	1 - Yes, 0-Otherwise
Cannot Conclude	1 - Yes, 0-Otherwise
Driver info	
Gender	1- Male, 0-Otherwise
Youth 15-35	1 - Yes, 0-Otherwise
Adult 36-60	1 - Yes, 0-Otherwise
Senior 61-90+	1 - Yes, 0-Otherwise
Driver Status	
Fatality (Death)	1 – Yes, 0-Otherwise

4.4.3 Performance Measurement

A confusion matrix could be used to evaluate the performance of the machine learning method. The measurements are true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

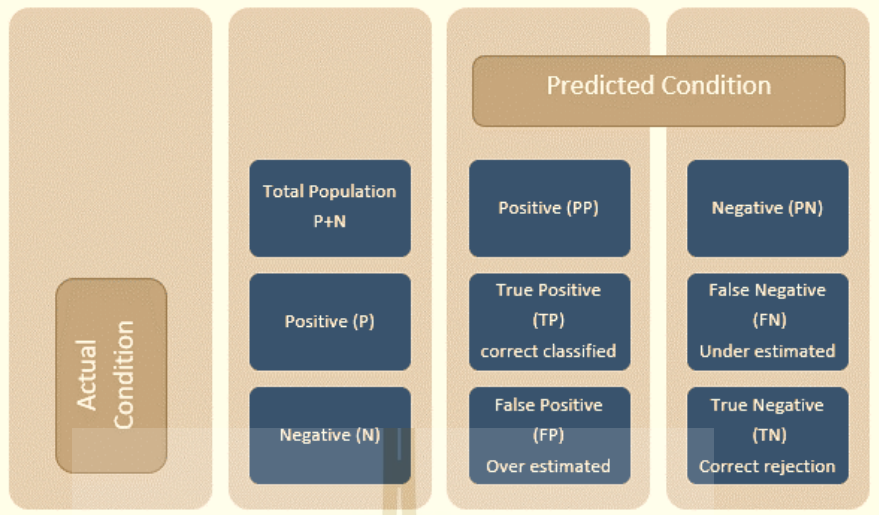


Figure 4.2 Confusion matrix diagram

Precision is a statistic that evaluates how accurate the classifier findings are. This statistic may be expressed as follows and figure 2:

$$Precision = \frac{\text{TP}}{TP+FP} \tag{4.1}$$

Recall: Sensitivity (recall) Sensitivity (recall) represents the proportion of the positive class that was correctly identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4.2}$$

True Negative rate also called specificity:

$$\text{TNR} = \frac{\text{TN}}{\text{TP+FN}} \tag{4.3}$$

False Positive Rate shows us how much of the negative class was misclassified by the classifier.

$$\text{FPR} = \frac{\text{FP}}{\text{TN+FP}} = 1 - \text{ TNR (Specificity)} \tag{4.4}$$

Accuracy: The ratio of correctly classification

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (4.5)$$

Information Gain is a measure of the entropy value.

$$Information\ Gain = \ Entropy\ (Initial) - \sum_{i=1}^{N} Pi\ \log_2 Pi$$

Where P is possibility of event of N (4.6)

4.5 Results

4.5.1 Model Result

According to table 4.5, the most common feature that causes motorcycle accidents for all eight models is the Highway and Gender which is ranked by top score (8 out of 8), Day time (7 out of 8) Over Speed limit (7 out 8), Night w/o light (5 out of 8), Drunk (2 out of 8), Night with light/Break through traffic light signal and Straight way (1 out of 8)

Table 4.5 Info. Gain Ranking by model

Model	Ranking		info. Gain
SVM	1	Break through Traffic light signal	0.009
	2	Highway	0.001
	3	Straight way	0.001
	4	Night with light	0.001
	5	Gender	0.000
Gradient Boosting	1	Highway	0.042
	2	Over Speed limit	0.038
	3	Day	0.037
	4	Night w/o light	0.029
	5	Gender	0.017

Table 4.5 Info. Gain Ranking by model (Continued)

Model	Ranking		info. Gain
Logistic Regression	1	Over Speed limit	0.058
	2	Highway	0.049
	3	Night w/o light	0.024
	4	Gender	0.023
	5	Day	0.019
Neural Network	1	Highway	0.046
	2	Over Speed limit	0.036
	3	Day	0.029
	4	Night w/o light	0.027
	5	Gender	0.021
Naïve Bayes	1	Over Speed limit	0.089
	2	Highway	0.075
	3	Day	0.042
	4	Gender	0.034
	5	Drunk	0.027
kNN	1	Over Speed limit	0.067
	2	Highway	0.047
	3	Gender	0.031
	4	Day	0.028
	5	Drunk	0.021
Tree	1	Night w/o light	0.043
	2	Highway	0.032
	3	Over Speed limit	0.030
	4	Day	0.013
	5	Gender	0.008

Table 4.5 Info. Gain Ranking by model (Continued)

Model	Ranking		info. Gain
Random Forest	1	Highway	0.041
	2	Over Speed limit	0.035
	3	Day	0.031
	4	Night w/o light	0.027
	5	Gender	0.015

4.5.2 Evaluation Results

Evaluation result from models in Table 4.6 & performance measurement avg over class in Figure 4.3 using sampling 75:25 (test data: train data) has high classification accuracy (Eq 5), Recall (Eq 2) and precision (Eq 1) to predict non-Fatality but low for fatality on recall.

Table 4.6 evaluation result from models

Model	Target Class	AUC	CA	Precision	Recall
SVM	Avg over	0.602	0.756	0.695	0.756
	Fatality	0.602	0.756	0.505	0.004
	Non-Fatality	0.602	0.756	0.756	0.999
kNN	Avg over	0.700	0.749	0.728	0.749
	Fatality	0.700	0.749	0.481	0.350
	Non-Fatality	0.700	0.749	0.807	0.878
Naïve Bayes	Avg over	0.724	0.762	0.726	0.762
	Fatality	0.724	0.762	0.526	0.232
	Non-Fatality	0.724	0.762	0.790	0.933
Logistic Regression	Avg over	0.729	0.764	0.725	0.764
	Fatality	0.729	0.764	0.556	0.162
	Non-Fatality	0.729	0.764	0.780	0.958

Table 4.6 evaluation result from models (Continued)

Model	Target Class	AUC	CA	Precision	Recall
Tree	Avg over	0.743	0.769	0.738	0.769
	Fatality	0.743	0.769	0.624	0.135
	Non-Fatality	0.743	0.769	0.777	0.974
Neural Network	Avg over	0.745	0.770	0.738	0.770
	Fatality	0.745	0.770	0.603	0.168
	Non-Fatality	0.745	0.770	0.782	0.964
Gradient Boosting	Avg over	0.751	0.771	0.742	0.771
	Fatality	0.751	0.771	0.616	0.168
	Non-Fatality	0.751	0.771	0.783	0.966
Random Forest	Avg over	0.768	0.777	0.752	0.777
	Fatality	0.768	0.777	0.640	0.198
	Non-Fatality	0.768	0.777	0.788	0.964

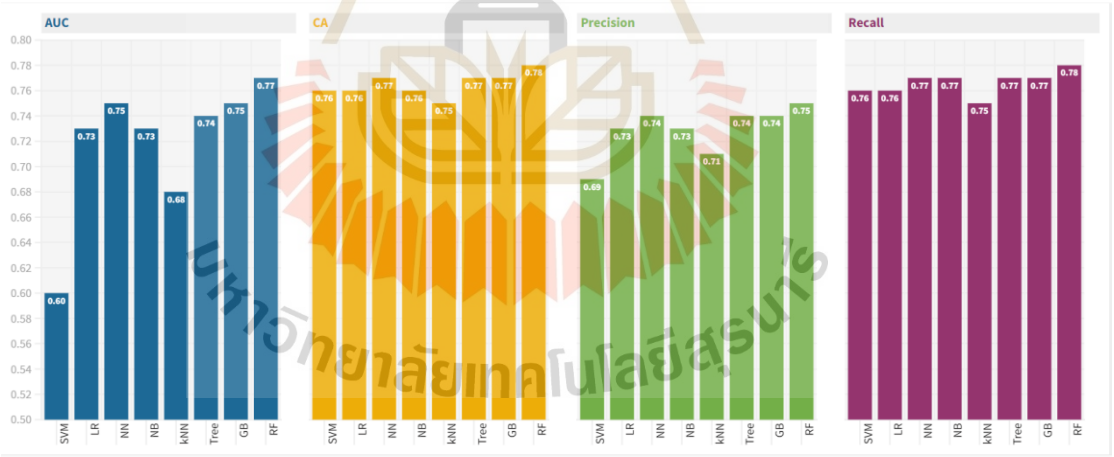


Figure 4.3 Performance Measurement models

Once the AUC (Area under the curve) of the ROC (Receiver Characteristic Operator) is between 0.5 and 1, there is a good chance that the classifier will be able to distinguish between positive and negative class values as Figure 4.4 since the classifier recognizes TF and TN (Eq 3) more than FN and FP (Eq 4).

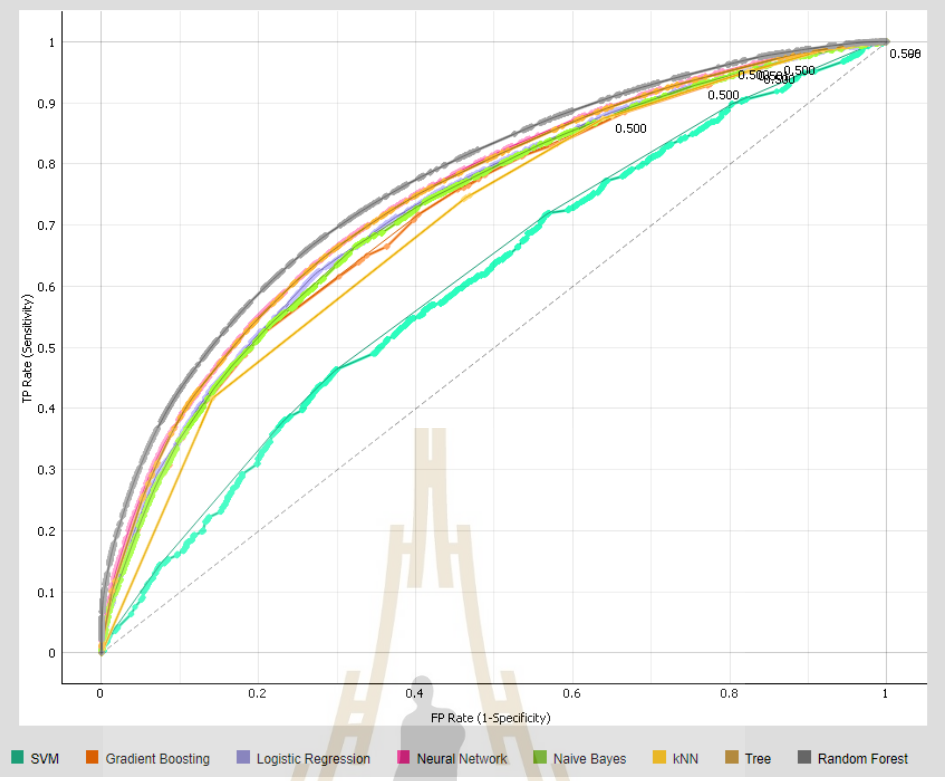


Figure 4.4 Model-specific ROC plot for predicting non-fatality.

From the evaluation model result, all models perform well for non-fatality prediction, with high precision and recall, but not for fatality prediction, with only fair precision and low recall. The best model for prediction is Random Forest, that has an average AUC of 0.768, CA of 0.777, Precision of 0.752, and recall of 0.777 follow by GB avg AUC of 0.751, CA of 0.771, Precision of 0.742, and recall 0. 771.Neural network avg AUC of 0.745, CA of 0.770, Precision of 0.738, and recall 0.770. Tree avg AUC of 0.743, CA of 0.769, Precision of 0.738, and recall 0.769. Logistic Regression avg AUC of 0.729, CA of 0.764, Precision of 0.725, and recall 0.764. Naïve Bayes avg AUC of 0.724, CA of 0.762, Precision of 0.726, and recall 0. 762.kNN avg AUC of 0.700, CA of 0.749, Precision of 0.728, and recall 0. 749. SVM avg AUC of 0.602, CA of 0.756, Precision of 0.695, and recall 0. 756.

Table 4.7 Confusion Metrix for each model

Confusion Metrix			Proportion of actual		Proportion of predicted		
					Predicted		
Model			Non-Fatality	Fatality	Non-Fatality	Fatality	Total
SVM	Actual	Non-Fatality	99.90%	0.10%	75.60%	49.50%	85308
		Fatality	99.60%	0.40%	24.40%	50.50%	27529
		Total	112637	200	112637	200	112837
Gradient Boosting	Actual	Non-Fatality	96.6%	3.4%	78.3%	38.4%	85308
		Fatality	83.2%	16.8%	21.7%	61.1%	27529
		Total	105328	7509	105328	7509	112837
Logistics Regression	Actual	Non-Fatality	95.80%	4.20%	78.00%	44.40%	85308
		Fatality	83.80%	16.20%	22.00%	55.60%	27529
		Total	104838	7999	104838	7999	112837
Neural Network	Actual	Non-Fatality	96.40%	3.60%	78.20%	39.70%	85308
		Fatality	83.20%	16.80%	21.80%	60.30%	27529
		Total	105142	7695	105142	7695	112837
Naïve Bayes	Actual	Non-Fatality	93.90%	6.70%	79.00%	47.40%	85308
		Fatality	76.80%	23.20%	21.00%	52.60%	27529
		Total	100723	12114	100723	12114	112837
kNN	Actual	Non-Fatality	87.80%	12.20%	80.70%	51.90%	85308
		Fatality	65.00%	35.00%	19.30%	48.10%	27529
		Total	92797	20040	92797	20040	112837
Tree	Actual	Non-Fatality	97.40%	2.60%	77.70%	37.60%	85308
		Fatality	86.50%	13.50%	22.30%	62.40%	27529
		Total	106887	5950	106887	5950	112837

Table 4.7 Confusion Metrix for each model (Continued)

Confusion Metrix		Proportion of actual		Proportion of predicted		
		Predicted				
Model		Non-Fatality	Fatality	Non-Fatality	Fatality	Total
Random Forest	Non-Fatality	96.50%	3.50%	78.80%	35.70%	85308
	Fatality	80.60%	19.40%	21.20%	64.30%	27529
	Total	104519	8318	104519	8318	112837

The confusion matrix in table 4.7 Random Forest is best in class, revealing 80.6% classification errors of actual F (fatalities) to NF (nonfatalities) and 3.5% of NF to F. There was a misclassifying of 21.2% for F to NF and 35.7% for NF to F for the predicted values. The worst model (SVM) misclassifies 99.6% of actual F to NF accidents and 0.1% of NF to F accidents. There was a misclassifying of 24.4% for F to NF and 49.5% for NF to F for predicted values. About information gain (Eq 6) from the model's prediction which top 5 attributes (features) from RF model consist of Highway, Over speed limit, Day time, Night w/o light and Gender. According to feature finding interpretation, all those features play a significant role in motorcycle accidents.

4.6 Conclusion & Discussion

This study aimed to focus on motorcycle crashes by focusing on the rider who was involved in the accident as well as the environmental factors that contribute to fatal crashes. To begin, identify the machine learning model that is best suited for predicting road accidents with high accuracy, as well as factors that are contributing to an increase in fatality accidents. A nonparametric analysis was performed on accident data from 2015 to 2020 to assess the significance of other elements that affect target factors such as rider information, road condition, weather condition, and rider behaviors. All eight models discovered significant classification errors of fatalities but a much better ability to predict nonfatalities. Random Forest outperformed the other seven ML algorithms in predicting road accidents in 2-Class classification giving an

average over class AUC of 0.768, CA of 0.777, Precision of 0.752, and recall of 0.777. Fatality AUC of 0.768, CA of 0.777, Precision of 0.640 and recall 0.198. Non-fatality AUC of 0.768, CA of 0.777, Precision of 0.788 and recall of 0.964. Which top 5 features which giving highest info gain consist of Highway, Riding over speed limit, Day time, Night w/o light and gender.

In many accident-related fields, ML models have been used to predict accidents. Kim et al. (2021) used to predict accidents at a Korean Container Port; the best model was chosen by comparing the accuracy, precision, recall, and F1 score of different models. The results show that a deep neural network model and a gradient boosting model outperform all other performance metrics, but the data set for port accidents is limited. While Santos et al. (2021) was discovered that a decision tree can detect the most important factors describing the severity of a road accident. Furthermore, the predictive model results indicate that the RF model could be a useful tool for forecasting accident hotspots, which is consistent with this study's finding that RF is the most accurate one to predict road accidents. For two-class prediction when compared to other methods, SVM and GB are the best in class (Recal & Demirel, 2021) while our research discovered that GB is the second runner after RF. Because of the complexities and wide range of factors involved in traffic accidents. The comparison analysis aids in determining which models outperformed and provided a useful prediction with the least amount of error, and which will be implemented. Different models work better for different data. Naive Bayes works well when features are highly independent. SVM is useful when there are too many features, and the dataset is medium in size. If the dependent and independent variables have a linear relationship, linear regression, logistic regression, and SVM are appropriate. kNN can be used with small data sets where the relationship between the dependent and independent variables is unknown. As a result, before deciding on which ML algorithm to use, the data must first be understood and analyzed or compared their accuracies on training and test sets.

The ML is effective tools in accident predicting, the model performs well in terms of non-fatality prediction, but fatality prediction still has room to improve, which can be interpreted to mean that the factors that cause fatal accidents may cause a

major accident but do not always result in a fatality. Specific feature selection may be required before entering the model to predict fatality since fatality has been a quite random feature involve, such as riding experience, rider health fit, and even the time duration from the accident area to hospital. Over in terms of information gained from the model, speed limit is the key runner for road accident (M. Yu et al., 2020), (Osman et al., 2018), (Krull et al., 2000) and a drunk cyclist is also a potential accident risk. Nonetheless, more education and stricter enforcement for intoxicated motorcycle riders may necessitate more research. For internal factors such as gender, age, accident location, and vehicle type were observed (Bahiru et al., 2018). Those were discovered to have an impact on the severity of road accidents, even though being male is still one of the leading causes of highway fatalities since gender and highway road were seen as a key factor in our study as well same as Ospina-Mateus et al. (2019) observed that men are more likely to be involved in serious accidents. As motorcycle accidents are heavily influenced by factors like as speed limit, age, Highway functional class, and speed compliance. (Rezapour et al., 2020). Ospina-Mateus et al. (2019) observed that men are more likely to be involved in serious accidents. According to previous research, intoxicated drivers have a higher accident rate. (Krull et al., 2000), (Xie & Huynh, 2012), (Kim et al., 2013), (Wu et al., 2016),(Zhou & Chin, 2019), (John & Shaiba, 2019), (Helen et al., 2019) ,(Champahom et al., 2020). The severity of the injury will increase as the rider's age below 25(Behnood & Mannering, 2017) , (Li, Ci, et al., 2019). Policymakers can use the prediction model with the most recent data set to see if there any factors changed or after the laws are implemented. Authorities should consider proposed laws to regulate speed limits and drunk riders more serious than before.

4.7 Limitations and Future Studies

It was predicted in an acceptable level (above 50% accuracy) by the model, and there is still room to improve the model and adjust the parameters in future studies to increase accuracy. The study used accident data from 2015 to 2020, but the covid-19 pandemic has spread so far in the last two years (2019-2020) that the government has ordered the country to close down and prohibit travel between

provinces, especially between the hours of 22.00 and 04.00. People are also hesitant to travel only to the separated zone, implying that they have not traveled far from home. Finally, the figures for 2019-2020 may not accurately reflect the country's true number of accidents and fatalities.

4.8 Reference

Al Mamlook, R. E., Ali, A., Hasan, R. A., & Mohamed Kazim, H. A. (2019). Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation. Proceedings of the IEEE National Aerospace Electronics Conference, NAECON,

Bahiru, T. K., Kumar Singh, D., & Tessfaw, E. A. (2018). Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018,

Behnood, A., & Mannering, F. (2017). The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. *Analytic Methods in Accident Research*, *14*, 41-53. https://doi.org/https://doi.org/10.1016/j.amar.2017.04.001

Breiman, L. (2001). Mach Learn.

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., & Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia,

Champahom, T., Jomnonkwao, S., Chatpattananan, V., Karoonsoontawong, A., & Ratanavaraha, V. (2019). Analysis of Rear-End Crash on Thai Highway: Decision Tree Approach. *Journal of Advanced Transportation*, *2019*, 1-13. https://doi.org/10.1155/2019/2568978

Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., & Ratanavaraha, V. (2020). Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. *Accident Analysis & Prevention*, *141*, 105537. https://doi.org/ https://doi.org/10.1016/j.aap.2020.105537

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cunto, F. J. C., & Ferreira, S. (2017). An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. *Journal of Transportation Safety & Security*, *9*(sup1), 33-46. https://doi.org/10.1080/19439962.2016.1162891

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Zupan, B. (2013). Orange: Data mining toolbox in python [Article]. *Journal of Machine Learning Research*, *14*, 2349-2353. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885599052&partnerID=40&md5=75d2df52a0c46b5ab58ab08e1576114e

DLT. (2021). Department of Land Transportation. *https://www.dlt.go.th/th/public-news/view.php?_did=2806*.

Dongare, A., Kharde, R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

Feng, M., Zheng, J., Ren, J., & Xi, Y. (2020). Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. In *Advances in Brain Inspired Cognitive Systems* (pp. 520-529). https://doi.org/10.1007/978-3-030-39431-8_50

Geedipally, S. R., Turner, P. A., & Patil, S. (2011). Analysis of Motorcycle Crashes in Texas with Multinomial Logit Model. *Transportation Research Record*, *2265*(1), 62-69. https://doi.org/10.3141/2265-07

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. http://oreilly.com/catalog/errata.csp?isbn=9781492032649

Guido, A. C. M. S. (2017). Introduction to machinelearning with python. http://oreilly.com/catalog/errata.csp?isbn=9781449369415 (Third Release) (O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.)

Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests [Article]. *Accident Analysis and Prevention*, *41*(1), 98-107. https://doi.org/10.1016/j.aap.2008.09.009

Helen, W. R., Almelu, N., & Nivethitha, S. (2019). Mining Road Accident Data Based on Diverted Attention of Drivers. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018,

Huo, X., Leng, J., Hou, Q., & Yang, H. (2020). A Correlated Random Parameters Model with Heterogeneity in Means to Account for Unobserved Heterogeneity in Crash Frequency Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*, 036119812092221. https://doi.org/10.1177/0361198120922212

Jafari Anarkooli, A., Hosseinpour, M., & Kardar, A. (2017). Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. *Accident Analysis & Prevention*, *106*, 399-410. https://doi.org/10.1016/j.aap.2017.07.008

John, M., & Shaiba, H. (2019). Apriori-Based Algorithm for Dubai Road Accident Analysis. Procedia Computer Science,

John, M., & Shaiba, H. (2022). Analysis of Road Accidents Using Data Mining Paradigm. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 68, pp. 215-223).

Jomnonkwao, S., Uttra, S., & Ratanavaraha, V. (2020). Forecasting Road Traffic Deaths in Thailand: Applications of Time-Series, Curve Estimation, Multiple Linear Regression, and Path Analysis Models. *Sustainability*, *12*(1). https://doi.org/10.3390/su12010395

Kim, J.-K., Ulfarsson, G. F., Kim, S., & Shankar, V. N. (2013). Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention*, *50*, 1073-1081. https://doi.org/https://doi.org/10.1016/j.aap.2012.08.011

Kim, J. H., Kim, J., Lee, G., & Park, J. (2021). Machine Learning-Based Models for Accident Prediction at a Korean Container Port. *Sustainability*, *13*(16), 9137. https://www.mdpi.com/2071-1050/13/16/9137

Krull, K. A., Khattak, A. J., & Council, F. M. (2000). Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes. *Transportation Research Record*, *1717*(1), 46-54. https://doi.org/10.3141/1717-07

Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations [Article]. *Journal of Modern Transportation*, *24*(1), 62-72. https://doi.org/10.1007/s40534-016-0095-5

Kuşkapan, E., Çodur, M. Y., & Atalay, A. (2021). Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms [Article]. *Accident Analysis and Prevention*, *155*, Article 106098. https://doi.org/10.1016/j.aap.2021.106098

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P. D., & Ma, D. T. (2018). Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. *Analytic Methods in Accident Research*, *20*, 1-14. https://doi.org/https://doi.org/10.1016/j.amar.2018.08.001

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z., Ma, D. T. (2019). Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. *Accident Analysis & Prevention*, *124*, 219-229. https://doi.org/https://doi.org/10.1016/j.aap.2018.12.020

Mafi, S., AbdelRazig, Y., & Doczy, R. (2018). Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. In *Transportation Research Record* (Vol. 2672, pp. 171-183).

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accid Anal Prev*, *122*, 181-188. https://doi.org/10.1016/j.aap.2018.10.014

Mohamad, I., Jomnonkwao, S., & Ratanavaraha, V. (2022). Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand. *Case Studies on Transport Policy*, *10*(4), 2165-2174. https://doi.org/https://doi.org/10.1016/j.cstp.2022.09.016

Mphela, T. (2020). Causes of road accidents in botswana: An econometric model [Article]. *Journal of Transport and Supply Chain Management*, *14*, 1-8, Article a509. https://doi.org/10.4102/jtscm.v14i0.509

Osman, M., Mishra, S., & Paleti, R. (2018). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*. https://doi.org/10.1016/j.aap.2018.05.004

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists [Article]. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051-10072. https://doi.org/10.1007/s12652-020-02759-5

Ospina-Mateus, H., Quintana Jiménez, L. A., López-Valdés, F. J., Morales-Londoño, N., & Salas-Navarro, K. (2019). Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia. In *Communications in Computer and Information Science* (Vol. 1052, pp. 309-320).

PDPM. (2020). Thailand Department of Public Disaster Prevention and Mitigation. *https://www.disaster.go.th/en/*

Recal, F., & Demirel, T. (2021). Comparison of machine learning methods in predicting binary and multi-class occupational accident severity [Article]. *Journal of Intelligent and Fuzzy Systems*, *40*(6), 10981-10998. https://doi.org/10.3233/JIFS-202099

Rezapour, M., Mehrara Molan, A., & Ksaibati, K. (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. *International Journal of Transportation Science and Technology*, *9*(2), 89-99. https://doi.org/10.1016/j.ijtst.2019.10.002

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. *Computers*, *10*(12), 157. https://www.mdpi.com/2073-431X/10/12/157

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shaheed, M. S., Gkritza, K., Zhang, W., & Hans, Z. (2013). A mixed logit analysis of two-vehicle crash seventies involving a motorcycle. *Accident; analysis and prevention*, *61*. https://doi.org/10.1016/j.aap.2013.05.028

Shweta, Yadav, J., Batra, K., & Goel, A. K. (2021). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. Journal of Physics: Conference Series,

Sonal, S., & Suman, S. (2018). A framework for analysis of road accidents. 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research, ICETIETR 2018,

Tolles, J., & Meurer, W. J. (2016). Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, *316*(5), 533-534. https://doi.org/10.1001/jama.2016.7653

Webb, G. I. (2010). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 713-714). Springer US. https://doi.org/10.1007/978-0-387-30164-8_576

WHO. (2018). World Health Organization: Global status report on road safety 2018. . *https://extranet.who.int/roadsafety/death-on-the-roads/*.

Wu, Q., Zhang, G., Zhu, X., Liu, X. C., & Tarefder, R. (2016). Analysis of driver injury severity in single-vehicle crashes on rural and urban roadways. *Accident Analysis & Prevention*, *94*, 35-45. https://doi.org/https://doi.org/10.1016/j.aap.2016.03.026

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Xie, Y., & Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. *Accident; analysis and prevention*, *47*, 36-44. https://doi.org/10.1016/j.aap.2011.12.012

Yu, M., Zheng, C., & Ma, C. (2020). Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *27*, 100126. https://doi.org/ https://doi.org/10.1016/j.amar.2020.100126

Zhou, M., & Chin, H. C. (2019). Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. *Accident Analysis & Prevention*, *124*, 104-112. https://doi.org/10.1016/j.aap.2019.01.009

CHAPTER 5
CONCLUSION AND RECOMMENDATION

The ML is an effective tool for predicting accidents; the model performs well in terms of non-fatality prediction, but there is still room for improvement in fatality prediction, which can be interpreted to mean that the factors that cause fatal accidents may cause a major accident but do not always result in a fatality. Because fatality is such a random feature, specific feature selection may be required before entering the model to predict fatality, such as riding experience, rider health fit, and even the time duration from the accident area to hospital.

In Thailand now, the majority of accidents occur during daytime (08.00–18.00), while peaks occur at 19.00–20.00 and 22.00–23.00 and high fatality rate at night (19.00–07.00). Accidents were more likely among young people (15-35 years old), who are less disciplined, unfamiliar with traffic laws, and have less driving experience, according to the study. Authorities are considering proposed laws to control speed limits on long straightaways by using light signs, warning signs, and cameras that closely monitor driving speeds, especially motorcycles and may consider raising the minimum age for obtaining a motorcycle license to at least 18 years old or imposing additional engine size restrictions based on rider age. Road lighting was also identified as a significant factor in motorcycle accidents, particularly on RRs, but in all accident categories except speeding-related deaths on HWs. As a result, we propose that better lighting be installed on Thai roads wherever possible, particularly in rural areas. It's important to note that predicting road accidents is a complex task and will likely require a combination of machine learning and other techniques, such as analyzing traffic patterns and engineering road designs to make them safer. In term of data, I would like to propose a recommendation to the organization responsible for collecting accident data, particularly regarding the importance of ensuring meticulous attention to data formatting and typing. We have encountered substantial data loss as a consequence of frequent errors and omissions during data entry. To address this issue,

I suggest implementing a drop-down selection feature for individuals involved in data input. This approach has the potential to significantly reduce or even eliminate errors, enhancing the overall accuracy and reliability of the collected accident data. By giving due consideration to these measures, the organization can ensure the utmost precision and integrity in its data collection practices.

Study 1: Males riding motorcycles at over-the-limit speeds on straight roads in clear weather have a higher risk of injury or death in traffic accidents than other conditions, with confidence levels increasing from 0.5x, 06x, and 0.7x to 0.99x if the consequences are motorcycle and dry surface road with high lift. As stated on the finding rule, the greater the number of elements involved, the greater the likelihood of an accident. In addition, the newly discovered straightway contributes significantly, while transportation authority's exercise caution at intersections and on curve roads. With all of the rules discovered in this study, policymakers may be able to eliminate some of the factors that contribute to highway traffic accidents. It should, at the very least, raise awareness of risky driving behaviors. Authorities are considering proposed legislation to control speed limits on long straightaways through the use of light signs, warning signs, and cameras that closely monitor driving speeds, particularly those of motorcycles.

Study 2: HWs have more fatality contributors than RRs; for example, accidents on HWs were common when riders had been drinking, especially at night with no light. HWs also have significantly more traffic and a wider variety of vehicles traveling at higher speeds than RRs, facilitating accidents that can result in serious injury or death. Age was another significant factor in motorcycle accident fatalities on both types of roads, though only HW fatalities included riders as old as 90 years; the highest age in RR fatalities was 60 years. One interesting observation here was that whether or not a rider was speeding, RR fatalities generally occurred at night with no street lighting. As a result, we propose that Thailand properly post speed limits and support enforcement of compliance within these limits. Another important predictor of motorcycle fatalities was short-cutting. Although male gender was the most common cause of HW motorcycle fatalities in this study, age, accident location, and vehicle type had less of an impact. Substance use, on the other hand, was a factor in many

accidents, and we propose more education on the dangers of riding while intoxicated, as well as stricter enforcement of penalties for infractions.

Study 3: According to our findings, GB is the second runner after RF. Because of the complexities and wide range of factors that contribute to traffic accidents. The comparison analysis helps to determine which models outperformed and provided the most accurate prediction with the least amount of error, and which will be implemented. Different models are better suited to different types of data. When features are highly independent, Naive Bayes performs well. When there are too many features and the dataset is medium in size, SVM is useful. Linear regression, logistic regression, and SVM are appropriate when the dependent and independent variables have a linear relationship. kNN can be applied to small data sets with an unknown relationship between the dependent and independent variables. As a result, before making a decision.

APPENDIX A

LIST OF PUBLICATIONS

List of Publications

Mohamad, I., Jomnonkwao, S., & Ratanavaraha, V. (2022). Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand. *Case Studies on Transport Policy*, *10*(4), 2165-2174. https://doi.org/https://doi.org/10.1016/j.cstp.2022.09.016 (IF : 3.043, Q1 in Scopus)

Case Studies on Transport Policy 10 (2022) 2165–2174

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand

Ittirit Mohamad [a], Sajjakaj Jomnonkwao [b,*], Vatanavongs Ratanavaraha [c]

[a] Program of Energy and Logistics Management Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand
[b] School of Transportation Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand
[c] School of Transportation Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand

ARTICLE INFO

Keywords:
Comparative analysis
Decision tree
Accidents
Big data
Transportation
Machine learning

ABSTRACT

Thailand ranks first in Asia and ninth in the world in term of road accident. As of 2020, the number of vehicles registered in Thailand was over 41 million, with motorcycles accounting for half of all vehicles. This study aimed to determine the cause of fatalities to reduce motorcycle accidents. The research entailed separating the accidents and fatalities into those occurring on highways (HWs) versus those occurring on rural roadways (RRs) and focused solely on rider at fault accidents to involve any confounding factors related to passengers or others involved. In Thailand, HWs have higher speed limits and allow more vehicle types than some RRs. Thailand's Department of Public Disaster Prevention and Mitigation recorded 115,154 motorcycle accidents from 2015 to 2020. Decision trees allow for processing large amounts of data to drill down into associations between the individual variables in a large data set; in this study, the tree also separated accidents into whether or not the driver was exceeding the speed limit. The model's performance for HWs, predicted misclassifications were found to be 28.3% (fatality to nonfatality) and a 44.5% (nonfatality to fatality) while predicted misclassification for RRs were 15.5% (fatality to nonfatality) and 60% (nonfatality to fatality). At all ages, the most fatalities were among male riders on dry straightaways in clear daytime weather; notably, however, on RRs, even when the rider was driving responsibly, fatalities were high at night on roads with no light. Following the presentation of the study findings, suggestions are made for ways the Thai government can improve the motorcycle accident and fatality statistics, including increasing the age limit for a motorcycle license, with engine size limits further divided according to age; proper enforcement of the existing rules will also improve the country's accident statistics. It will also be highly effective to improve road lighting, particularly on RRs.

1. Introduction

Thailand is one of the countries with a high rate of fatalities from road accidents, ranking first in Asia and ninth in the world. Thais are killed in traffic accidents at a rate of 32.7 per 100,000 persons (WHO, 2018). As of 2020, the total number of vehicles registered in Thailand was over 41 million, with motorcycles accounting for half of all vehicles (DLT, 2021).

Fig. 1 shows that the number of motorcycles registered in Thailand grew continually from 2015 to 2020, when there were 41,471,135 vehicles registered in the country, of which 21,396,980 were motorcycles; these were followed by lightweight 4-wheeled drive vehicles (10,446,505), mini-trucks (6,878,050), and others (2,749,600) (DLT, 2021). However, although motorcycles only account for half the vehicles in Thailand, motorcyclists account for the majority of road accident fatalities (Jomnonkwao et al., 2020). According to the Thailand Accident Research Center, in 2019, there were 4802 motorcycle fatalities, an average of 13.15 people per day, with most occurring among people aged 20 years and up, also known as the working age group. In terms of causes, 54 % of accidents were the fault of the motorcyclists, drivers were at fault in 40 % of the accidents, and the road and vehicle accounted for 4 % and 2 %, respectively (RSC, 2019). Worldwide research has identified elements that appear to be common to every accident. Thailand has six road types:

- A motorway is a HW designed for high mobility with low access based on limited entrances and exits to designated points, and no

* Corresponding author.
E mail address: sajjakaj@g.sut.ac.th (S. Jomnonkwao).

https://doi.org/10.1016/j.cstp.2022.09.016
Received 29 June 2022; Received in revised form 19 August 2022; Accepted 27 September 2022
Available online 30 September 2022

two-wheeled vehicles are permitted. Motorways are supervised by the Department of Highways.

- The national HWs link regions, provinces, and districts; they emphasize mobility, but access is not limited, and rules in general are less strict than those for motorways are. It is difficult to travel through cities, and the HWs were designed to bypass the cities; they are also supervised by the Department of Highways.
- Rural roadways (RRs) are located outside of municipalities and connect with the national HWs. They are supervised by the Department of Rural Roads.
- Municipal roadways provide the streets in municipalities and are maintained by the municipalities.
- Subdistrict roadways serve as the streets for those areas, and they are supervised by subdistrict organizations.
- Concession roadways are privately owned; the government grants concessions to private entities that are then responsible for supervising the roads.

For this study, all roads managed by the Department of Highways are designated HWs and rural and subdistrict roads are RRs. The motorcycle speed limits are lower on RRs—80 km/hr. for engines larger than 400 cc and 60 km/hr. for smaller cycles; for HWs, the limits are 90 km/hr. for engines larger than 400 cc and 70 km/hr. otherwise (DLT, 2021) and the study was distinctive in that it separated accidents and fatalities into those that occurred on highways (HWs) and those that occurred on rural roadways (RR), and it focused solely on rider-at-fault accidents in order to eliminate any confounding factors related to passengers or others involved.

In statistical side, Discriminant analysis (DA) is a technique commonly used in Statistical Algorithms to classify a set of observations into predefined classes and LDA (linear Discriminant analysis) is a diagnostic method for detecting potentially influential observations. The usual assumptions relevant to discriminant analysis are linearity, normality, and homoscedasticity of within-group variances of independent variables. However, due to violations of these assumptions, discriminant analysis has been supplanted by LR (logistic Regression), which requires fewer assumptions, produces more robust results, and is easier to use and comprehend than discriminant analysis (Chen, 2012). A regression-type model is a CART model that predicts the value of continuous variables using a set of continuous or categorical predictor variables. For this study, we selected a decision tree (CART regression Tree) to drill down into a set of big data to identify the relevant variables and analyze the relationships among them. Decision tree mining is among the most popular machine learning techniques (Wu et al., 2007) for its comprehensibility and ease of interpretation. One of the primary advantages of a decision tree is the ability to derive decision rules; these rules can aid in identifying safety issues and developing performance metrics (Abellán et al., 2013).

2. Literature review

Previous studies on road accidents have been classified by group components that are suspected to be involved in every accident, according to international research.

2.1. Age and gender

Motorcycle accidents are more likely to occur among young people because they are less disciplined, are unfamiliar with traffic laws, and have less driving experience (Zhang and Fan, 2013). Men and women aged 20–39 years who ride motorcycles are more likely to be involved in major accidents, whereas when no motorcycle or cyclist is involved in the incident, the severity is likely to be minor(Ospina-Mateus et al., 2019). Jou et al. (2012) found that being older, male, and unlicensed; not wearing a helmet; riding after drinking; and driving heavy motorcycles (above 550 cc) were linked to higher motorcycle fatality rates. Additionally, rider age was the most important factor when the rider was not at fault (Champahom et al., 2019). Pakgohar et al. (2011) found that the majority of fatalities were among young persons who were in good health before the accident. Riders between the ages of 18 and 24 years have insufficient experience to make adjustments while driving including adjusting their speed to the road conditions (Bucsuházy et al., 2020).

2.2. Weather and road conditions

Research has established that external conditions such as fog, rain, and snow have a greater influence on road accidents than rider-related internal factors and that the drivers/riders are more likely than passengers to be injured or killed in an accident (El Abdallaoui et al., 2018). According to the findings, the most important and influential road accident variables are speed limit; weather conditions; road factors such as type, surface, and number of lanes; lighting conditions; and time of the accident. Factors that had less influence on accidents were gender, age, accident site, and vehicle type (Feng et al., 2020). Highway (HW) intersections have been identified as the most dangerous for all accidents (Kumar and Toshniwal, 2016),Malin et al. (2019). As noted above, however, there are still significant accidents on straightaways with no intersections, in part because riders disobey the speed limit and in part because of poor road conditions.

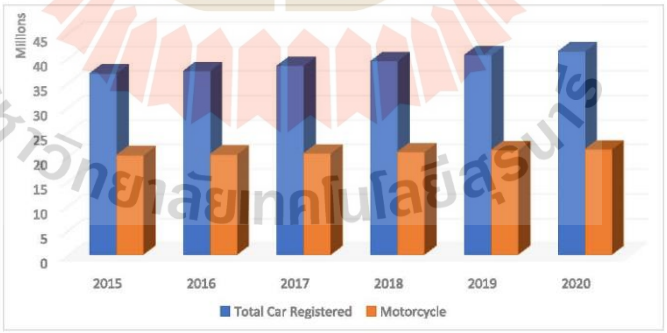


Fig. 1. Total number of vehicles and motorcycles registered in Thailand from 2015 to 2020.

2.3. Other important factors

Vehicle speed is the most critical determinant of an accident's severity (Al Mamlook et al., 2019), followed by factors such as speed limit, age, and road type (Rezapour et al., 2020). Travel at night increases the risk of an accident (Mphela, 2020) and increases the severity of any injuries, particularly when there is no light Shaheed et al. (2013), (Kim et al., 2013), (Jafari Anarkooli et al., 2017) and after midnight (Zhou and Chin, 2019). Xie and Huynh (2012) determined that the severity of injuries from accidents on dark roads decreases when riders are more cautious. Motorcycles are riskier in rural areas. Male riders, pillion riders, speeding, improper overtaking, and fatigue are all important factors that influence severe and fatal injuries (Se et al., 2021).

Additionally, the risk of motorcycle death increases for single-vehicle accidents that occur on nonurban roads at night, and the major factors that affect rear-end crashes are passenger characteristics and the rider's age, whereas side collisions are most commonly the result of lighting conditions and landscape (Anvari et al., 2017; Siskind et al., 2011). Focusing on driver factors, researchers discovered that high-speed driving, driving while intoxicated. And traffic violations all contributed to high rates of fatalities on RRs (Khorashadi et al., 2005). Researchers have used many tools in accident analysis, including measuring the accuracy between models or methods (Table 1), but few have studied the same model or method to compare two road types.

3. Methodology

The research begins with motorcycle accident data from Thailand's Department of Public Disaster Prevention and Mitigation, which counted 115,154 single-rider accidents between 2015 and 2020. Toward our study aim, the data was compiled on HR and RR motorcycle accident fatalities, developed a decision tree model, and measured its accuracy. Fig. 2 displays the steps in the study process, also listed below:

- After cleansing the data set, the initial dataset was validated for detecting and correcting missing and incompletely captured data as well as demonstrating the data's quality.
- Verified Dataset – Set the target to Fatal/non-fatal and partition the data in binary mode both HW and RR data set.
- Data Separation – Separated test and train data sets.
- Model Learning enables the model to learn from the test data set and then test with the remaining data set.
- Prediction and Model Measurement - To assess each model's prediction accuracy.

3.1. Data description

As noted earlier, the 2015–2020 motorcycle accident data from the Thailand Department of Public Disaster Prevention and Mitigation indicated 115,154 single-rider accidents, 61,866 on HWs and 53,288 on RRs (PDPM, 2020). Table 2 presents the categorical and descriptive statistics for the study data, which we divided into four categories: roadway characteristics, external factors involving the environment and weather conditions, internal factors involving driver behavior, and driver details. According to the descriptive data table, most accidents on both HWs and RRs were caused by being a male rider between 15 and 35 years of age exceeding the speed limit; most accidents occurred on dry surfaces and in clear weather, even when the driver stayed on the right side of the road.

Table 1
The Machine Learning Models Used in Extant Traffic Accident Studies.

Author	Methodology Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-Nearest Neighbor	K-Means	Multinomial Logistic Regression	Neural Network	Naïve Bayes	Random Forest	Support Vector Machine
Ospina-Mateus et al. (2021)	–	–	–	✓	–	✓	–	–	✓	✓	✓	✓
Harb et al. (2009)	–	–	–	✓	–	–	–	–	–	–	✓	–
Kuşkapan et al. (2021)	–	–	–	–	–	✓	–	–	–	✓	–	✓
Abellán et al. (2013)	–	–	–	✓	–	–	–	–	–	–	–	–
Mafi et al. (2018)	–	–	–	–	–	–	–	–	–	–	✓	–
Al Mamlook et al. (2019)	–	✓	✓	✓	–	✓	–	–	–	✓	✓	✓
Reeal and Demirel (2021)	–	–	–	✓	✓	–	–	✓	✓	–	–	✓
Kumar and Toshniwal (2016)	✓	–	–	–	–	–	✓	–	–	–	–	–
Helen et al. (2019)	✓	–	–	–	–	–	✓	–	–	–	–	–
Feng et al. (2020)	✓	–	–	–	–	–	–	–	✓	–	–	–
Bhavsar et al. (2021)	✓	–	–	–	–	–	–	–	–	–	–	–
Bahiru et al. (2018)	–	–	–	✓	–	–		–	–	✓	–	–

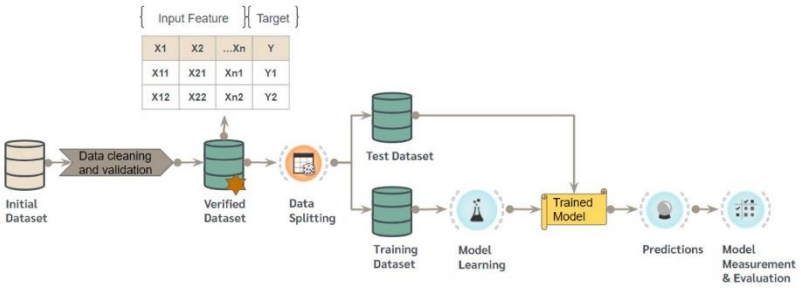


Fig. 2. The steps in the process for the study.

Table 2
The categorical variables and their descriptive statistics.

	HighWay				Non-HighWay			
	Fatality				Fatality			
	Yes		No		Yes		No	
Aciccent Event (Attribute)	Count	%	Count	%	Count	%	Count	%
RoadWay								
Dry Surface Road	18992	30.7%	40112	64.8%	8063	15.1%	43049	80.8%
Wet Surface	673	1.1%	2089	3.4%	277	0.5%	1899	3.6%
Straight Way	14171	22.9%	29389	47.5%	5619	10.5%	31698	59.5%
Not straight Way (Curve, Slope, Junction, etc)	5494	8.9%	12812	20.7%	2721	5.1%	13250	24.9%
Obstruction	343	0.6%	1301	2.1%	170	0.3%	1847	3.5%
Road condition	194	0.3%	886	1.4%	167	0.3%	1388	2.6%
Vehicle condition	226	0.4%	825	1.3%	126	0.2%	1402	2.6%
External Factor (Envi and Weather Con)								
Day Time (06.00-18.00)	9649	15.6%	24654	39.9%	4021	7.5%	26215	49.2%
Night with Light	5633	9.1%	11310	18.3%	2140	4.0%	9936	18.6%
Night without Light	4383	7.1%	6237	10.1%	2179	4.1%	8797	16.5%
Low visibility	1819	2.9%	5437	8.8%	659	1.2%	5523	10.4%
Clear Weather	17298	28.0%	35951	58.1%	7451	14.0%	39037	73.3%
Not Clear Weather (Rain, fog, etc)	2367	3.8%	6250	10.1%	889	1.7%	5911	11.1%
Internal Factor (Driver Behavior)								
Drunk	2410	3.9%	9521	15.4%	1357	2.5%	11065	20.8%
Over Speed limit	13524	21.9%	20018	32.4%	5888	11.0%	18129	34.0%
Break Through Traffic lights	185	0.3%	283	0.5%	50	0.1%	183	0.3%
Break Through Traffic Signs	289	0.5%	748	1.2%	88	0.2%	573	1.1%
Overtake	526	0.9%	702	1.1%	121	0.2%	576	1.1%
Use Mobile Phone	22	0.0%	171	0.3%	2	0.0%	185	0.3%
Short Cut off	4156	6.7%	9170	14.8%	1248	2.3%	10141	19.0%
Drug	3	0.0%	39	0.1%	3	0.0%	27	0.1%
Drive in opposite direction	385	0.6%	563	0.9%	58	0.1%	268	0.5%
Doze off	260	0.4%	536	0.9%	92	0.2%	369	0.7%
Overweight Carry	11	0.0%	28	0.0%	5	0.0%	37	0.1%
Cannot Conclude	676	1.1%	1686	2.7%	256	0.5%	1772	3.3%
Driver info								
Gender (male)	16921	27.4%	30998	50.1%	7236	13.6%	32500	61.0%
Gender (Female)	2744	4.4%	11203	18.1%	1104	2.1%	12448	23.4%
Youth 15-35	9894	16.0%	23204	37.5%	4010	7.5%	23277	43.7%
Adult 36-60	7111	11.5%	14830	24.0%	3205	6.0%	17126	32.1%
Senior 61-90+	2660	4.3%	4167	6.7%	1125	2.1%	4545	8.5%

**External factors are environment and weather conditions.

3.2. The decision tree

A decision tree is a predictor, h: X → Y, of the predecessors of an event x by spanning a tree from its root node to its leaves. For simplicity, we concentrated on the binary classification case, namely, Y = {0, 1}, but decision trees can be used for a range of prediction problems. Based on the division of the input space, the successor child is chosen at each node along the root-to-leaf path. Usually, the splitting is based on one of x's properties or a predefined set of splitting rules as follows:

- First, set the domain set: X is the accident event that needs to be labeled.

Set X to be binary {1,0}, and let Y be our possible labels.

Then, Y = {0, 1}, where 1 and 0 represent the possible options.

- Training set $S = ((X1, Y1)\cdots(Xn, Yn))$ is a limited number of pairings in $X \times Y$, that is, a list of labeled domain points. This is the information to which the learner has access.
- For the output, the learner is asked to generate a prediction rule, h: X → Y. This function is also referred to as a prediction, hypothesis, or classifier. The predictor can forecast new domain elements (Ben-David, 2014).

Decision trees comprise three parts: decision nodes, branches, and leaf nodes. Each decision node in the structure displays the variable, and each branch displays one variable value based on decision rules; the leaf nodes display the expected values of the target variables (Song and Ying, 2015). We used Orange 3.30 software (Demšar et al., 2013) to run the CART decision tree set classification to stop at when majority reach 95 % and limit of maximum tree depth is 7. Data was divided into two flows, HW and RR, and extracted 27 binary categorical variables that were most relevant to the 115,154 single-rider motorcycle accidents in Thailand from 2015 to 2020; the variables were set as binary (1 or 0) to facilitate interpretation and classification. Table 3 presents the 27 most relevant variables related to single-rider accidents under the following factors: roadway factors, external (environment, weather) and internal (driver behaviors) factors, driver data, and driver status.

3.3. *Performance measurement*

To assess the performance of the supervised machine learning decision tree in this study, we used tests data to validation how well the model performed with a confusion matrix with the following components: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). For example, TP shows the number of positive values projected to be positive, whereas FP predicts an accident as fatal when it was not; recall and precision were also measured. These strategies are especially useful for unbalanced data sets, in which one answer category accounts for the bulk of the responses. Precision refers to the accuracy of the classifier findings, expressed as eq (1) and shown in Fig. 3:

$$Precision = \frac{\mathrm{TP}}{TP + FP} \tag{1}$$

Recall, or sensitivity, gives the proportion of the positive class that was correctly classified, expressed as eq (2) :

$$Recall = \frac{\mathrm{TP}}{TP + FN} \tag{2}$$

The TN rate (TNR), also called specificity, is computed as eq (3) :

$$TNR = \frac{\mathrm{TN}}{TP + FN} \tag{3}$$

The FP rate (FPR) shows how often the classifier misclassified the negative class and is computed as eq (4):

$$FPR = \frac{\mathrm{FP}}{TN + FP} = 1 - TNR(Specificity) \tag{4}$$

The ratio of correct classifications reflects the data accuracy and is calculated as eq (5):

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + TN + FP + FN} \tag{5}$$

Table 3
The Measurement Categories for the 27 Identified Motorcycle Accident Variables.

Factor and Variables	Measurement
Roadway	
Dry road	1 – Yes, 0-Otherwise
Straight road	1 – Yes, 0-Otherwise
Obstruction	1 – Yes, 0-Otherwise
Road conditions	1 – Yes, 0-Otherwise
Vehicle conditions	1 – Yes, 0-Otherwise
External Factors (Environment and Weather Conditions)	
Day Time (06.00–18.00)	1 – Yes, 0-Otherwise
Night with light	1 – Yes, 0-Otherwise
Night without light	1 – Yes, 0-Otherwise
Low visibility	1 – Yes, 0-Otherwise
Clear weather	1 – Yes, 0-Otherwise
Internal Factors (Driver Behaviors)	
Drunk	1 – Yes, 0-Otherwise
Over speed limit	1 – Yes, 0-Otherwise
Ran a traffic light	1 – Yes, 0-Otherwise
Ran a traffic sign	1 – Yes, 0-Otherwise
Passing (overtaking)	1 – Yes, 0-Otherwise
Used a mobile phone	1 – Yes, 0-Otherwise
Short cutoff	1 – Yes, 0-Otherwise
Used drugs	1 – Yes, 0-Otherwise
Drove in opposite direction	1 – Yes, 0-Otherwise
Dozed off	1 – Yes, 0-Otherwise
Overweight cargo	1 – Yes, 0-Otherwise
Inconclusive	1 – Yes, 0-Otherwise
Driver Data	
Gender	1- Male, 0-Otherwise
Youth 15–35	1 – Yes, 0-Otherwise
Adult 36–60	1 – Yes, 0-Otherwise
Senior 61–90+	1 – Yes, 0-Otherwise
Driver Status	
Fatality	1 – Yes, 0-Otherwise

4. Results

Fig. 4 presents plots of the HW and RR likelihoods of fatalities over 24 h.

HWs

1 = fatality: μ =13.47, σ = 6.34
0 = nonfatality: μ =13.71, σ = 6.31

RRs

1 = fatality: μ =13.14, σ = 7.11
0 = nonfatality: μ =13.89, σ = 6.29

There is a higher probability of a fatality on a HW than on a RR: HW, 0.3–0.4 and RR, 0.25–0.15. However, both RRs and HWs have higher fatality rates at night (00.00–07.00). Both the HW and the RR decision trees were set to target rider fatalities, and they identified the main causes. The tree node was divided into whether the rider was following or exceeding the speed limit.

- HW Fatalities (Fig. 5):
  - o Rider exceeds the speed limit (40.3 %: 13,524/33,542): male (44.7 %: 11,757/26,312), age 15–35 years (35 %: 2,411/6,891), age 61–90 years (48.2 %: 1,808/4,152), straight road (43.5 %: 1,808/4,152), daytime (39.3 %: 5,364/13,653), clear weather (63.5 %: 1,933/3,048).
  - o Rider does not exceed the speed limit (21.7 %: 6,141/28,324): male (23.9 %: 5,164/21,607), age 15–35 years (22.7 %: 1,471/6,477), age 61–90 years (43.1 %: 447/1,037), drunk (17.9 %: 1,459/8,165), daytime (17.8 %: 581/3,257), night w/o light (25.3 %: 311/1,230), clear weather (19.1 %: 1,214/6,367), short cutoff (35 %: 1,008/2,881). On HWs, fatalities occurred most commonly among male riders who were between the ages of 15 and 35 years and were exceeding the speed limit on a straight road in clear

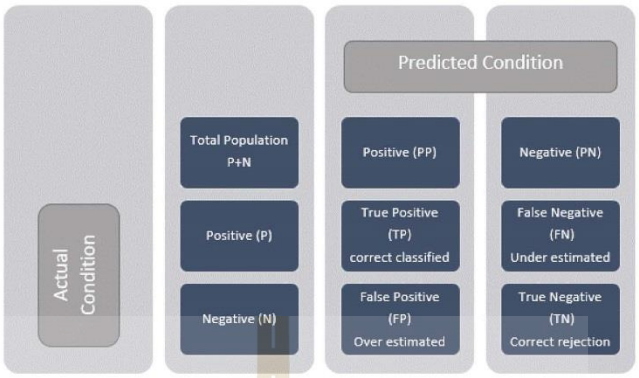


Fig. 3. Diagram of the confusion matrix.

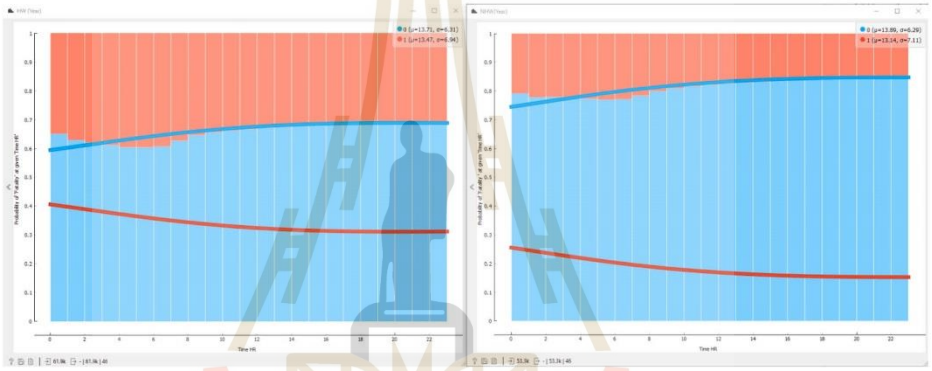


Fig. 4. HW and RR fatality probabilities at different times of the day.

weather during the day. For riders aged 61 to 90 years, the fatalities occurred most often at night, without lights, under rider intoxication, and with short cutoffs.

- RR Fatalities (Fig. 6):
  - o Rider exceeds the speed limit (24.5 %: 5,888/24,017): male (28.3 %: 5,132/18,131), age 15–35 years (20.2 %: 950/4,696), age 36–60 years (26.2 %: 952/3,639), daytime (24.3 %: 2,343/9,645), night w/o light (40.1 %: 974/2,427), short cutoff (55.8 %: 115/206), clear weather (36.1 %: 2,114/5,863).
  - o Rider does not exceed the speed limit: male (9.7 %: 2,104/21,605), night w/o light (11.2 %: 567/5,057).

Table 4 presents the final-level leaf sets for HW and RR motorcycle accident fatalities according to whether or not the driver was exceeding the speed limit. For instance, for RR fatalities, the common factors in both age groups were being male and riding over the speed limit during the day; at night on RRs, male riders who died were most often speeding and making short cutoffs at night with no light. When drivers were not speeding, most fatalities occurred among males at night with no light.

Short cutoffs were among the most common causes of fatalities on both HWs and RRs, but excess speed was also a factor only on RRs. Gender was the most significant variable in fatalities: Most fatalities were among male riders on both HWs and RRs irrespective of the rider speed limit, consistent with earlier findings that men who ride motorcycles are more likely to be involved in serious accidents Ospina-Mateus et al. (2019).

4.1. Evaluation results

Table 5 presents the cross-validated data results with 20 folds. The table presents the area under the receiver-operating curve (AUC) and the classification accuracy (CA; Equation (5)), recall (Equation (2)), and precision (Equation 1). CA is high for HWs and RRs. When $0.5 < AUC < 1$, there is a good possibility that the classifier will be able to differentiate between positive and negative class values. This is because the classifier is better able to recognize TP and TN (Equation (3)) than FN and FP (Eq. (4)).

The HW confusion matrix in Fig. 7 shows misclassifications in 77.9 % of actual fatalities to nonfatality accidents and 8.3 % of nonfatalities to fatalities. For the predicted values, we identified a misclassification of 28.3 % for accident fatalities to nonfatalities and 44.5 % for nonfatalities to fatalities. Fig. 8 shows the RR confusion matrix, reflecting

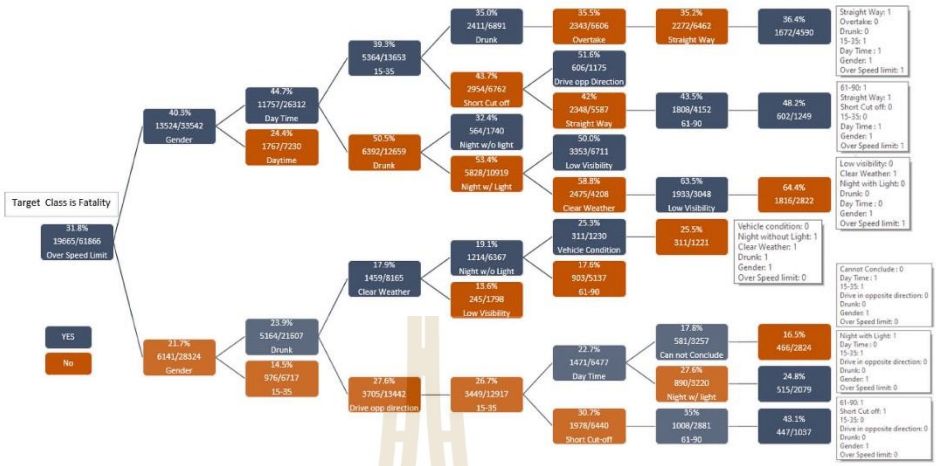


Fig. 5. The HW tree model.

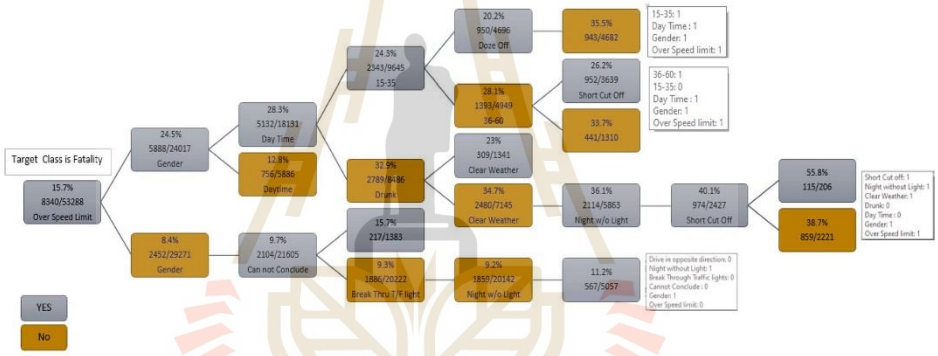


Fig. 6. The RR tree model.

misclassifications of 97.9 % (fatality to nonfatality) and 0.6 % (nonfatality to fatality) for the actual cases and of 15.5 % (fatality to nonfatality) and 60 % (nonfatality to fatality). That is, the decision tree for this study shows significant misclassifications of fatalities for both HWs and RRs but much better ability with nonfatalities.

5. Conclusion and discussion

The aim of this research was to design a decision tree to identify individual contributors to motorcycle accident fatalities among riders in Thailand, with a focus on single-rider crashes. Contributing variables included roadway features along with external and internal (driver-related) factors. In addition, using accident data from 2015 to 2020, we performed a nonparametric analysis to determine the importance of factors that influence target variables, such as road and weather conditions, speeding, being on a straightaway (with no intersections), gender, and substance use.

The decision tree concluded that the most common causes of fatality on both HWs and RRs were being a male rider and exceeding the speed limit, with the other variables showing differing levels of importance (Fig. 9). HWs have more contributors to fatalities than RRs; for instance, accidents were common on HWs when riders had been drinking, especially at night with no light. HWs also have much heavier traffic and a wider variety of vehicles traveling at higher speeds than RRs, facilitating accidents that can cause serious injury and death. Age was another significant contributor to motorcycle accident fatalities on both types of roads, although notably, only HW fatalities extended to riders up to the age of 90 years; the highest age in RR fatalities was 60 years. One interesting observation here was that RR fatalities generally occurred at night with no street lighting whether or not a rider was speeding. Road accidents have many contributing factors, but speeding is a key area of concern for the severity of road accident injuries (Yu et al., 2020; Osman et al., 2018; Krull et al., 2000). Thus, we propose that Thailand must properly post speed restrictions and support the enforcement of compliance within these limits. Short-cutoff riding was another key predictor of motorcycle fatalities Bahiru et al. (2018). Although male

Table 4
The Final HW and RR Sets by Rider Speed.

	HW Fatalities (Fig. 4):	RR Fatalities (Fig. 5):
Rider exceeds the speed limit	Set 1: (15–35, Day time, Male, Straight way) 36.4 % (1,672/4,590) Set 2: (61–90, Day time, Male, Straight way) 48.2 % (602/1,249) Set 3 (Clear Weather, Male) 64.4 % (1,816/2,822)	Set 1: (15–35, Day time, Male) 35.5 % (943/4,682) Set 2: (36–60, Day time, Male) 26.2 % (952/3,639) Set 3: (Clear weather, short cutoff, Night without light, Male) 55.8 % (115/206)
Rider *does not exceed* the speed limit	Set 1: (Clear Weather, Male, Drunk) 25.5 % (311/1,221) Set 2: (Daytime,15–35, Male) 16.5 % (466/2,824) Set 3: (Night with light,15–35, Male) 24.8 % (515/2,079) Set 4: (Short cutoff,61–90, Male) 43.1 % (477/1,037)	Set 1: (Night without light, Male) 11.2 % (567/5,057)

Table 5
The Model Evaluation Results.

Model	Road	Target Class	AUC	CA	Precision	Recall
Tree	HW	Avg over	0.685	0.696	0.665	0.696
		Fatality	0.686	0.696	0.555	0.221
		Nonfatality	0.686	0.696	0.717	0.917
	RR	Avg Over	0.706	0.842	0.776	0.842
		Fatality	0.703	0.842	0.400	0.021
		Nonfatality	0.703	0.842	0.845	0.994

gender was a primary factor in HW motorcycle fatalities in this study, we found less influence from age, accident location, and vehicle type. However, substance use was a factor in many accidents, and we propose more education on the dangers of riding while intoxicated in addition to tighter enforcement of penalties for infractions.

In Thailand, persons of all ages ride motorcycles, although the majority of riders are between the ages of 15 (bike capacity of no more than 110 cc) and 35 years. Looking at all four accident scenarios in the research, three featured young people, consistent with Zhang and Fan (2013), in which accidents were more likely among young people (25 years old), who are less disciplined and unfamiliar with traffic laws and have less driving experience. Policymakers might consider raising the minimum age for obtaining a motorcycle license to at least 18 years or imposing further restrictions on engine size dependent on rider age. We also identified road lighting as a considerable factor in motorcycle accidents, particularly on RRs but in fact in all accident categories except for speeding-related deaths on HWs. Therefore, we propose that better lighting be installed wherever possible on Thai roadways, particularly in rural areas.

6. Limitations and future studies

This study's model showed acceptable (above 50 %) accuracy, but there is room for improvement; adjusting the parameters in a future study could increase the accuracy. Additionally, we used accident data from 2015 to 2020, but during the last two years, 2019 and 2020, the circumstances in Thailand as well as around the world changed drastically overnight because of the COVID-19 pandemic. Governments worldwide locked down and ordered people to stay indoors, and Thailand limited travel between provinces, particularly between the hours of 22.00 and 04.00. Because mobility was so limited during 2019 and 2020, the overall findings for those years might not accurately reflect what would have been the country's true numbers of accidents and fatalities.

CRediT authorship contribution statement

Ittirit Mohamad: Conceptualization, Data curation, Formal analysis, Methodology, Software. **Sajjakaj JomnonKwao:** Validation, Writing – review & editing. **Vatanavongs Ratanavaraha:** Visualization, Supervision.

Confusion matrix for Tree (showing proportion of actual)

		Predicted 0	Predicted 1	Σ
Actual	0	91.7 %	8.3 %	42201
	1	77.9 %	22.1 %	19665
	Σ	54031	7835	61866

Confusion matrix for Tree (showing proportion of predicted)

		Predicted 0	Predicted 1	Σ
Actual	0	71.7 %	44.5 %	42201
	1	28.3 %	55.5 %	19665
	Σ	54031	7835	61866

Fig. 7. The confusion matrix actual and predicted results for HWs.

Confusion matrix for Tree (showing proportion of actual)

		Predicted 0	Predicted 1	Σ
Actual	0	99.4 %	0.6 %	44948
	1	97.9 %	2.1 %	8340
	Σ	52861	427	53288

Confusion matrix for Tree (showing proportion of predicted)

		Predicted 0	Predicted 1	Σ
Actual	0	84.5 %	60.0 %	44948
	1	15.5 %	40.0 %	8340
	Σ	52861	427	53288

Fig. 8. The confusion matrix actual and predicted results for RRs.

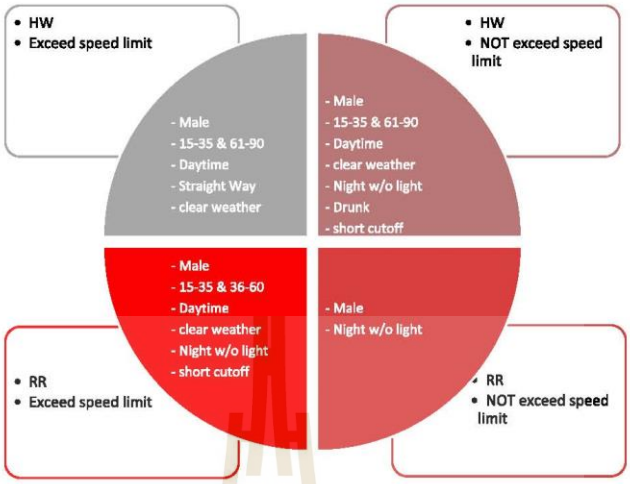


Fig. 9. Key accident factors: HWs versus RRs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abellán, J., López, G., de Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl. 40 (15), 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027.

Al Mamlook, R.E., Ali, A., Hasan, R.A., Mohamed Kazim, H.A., 2019. Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation. Proceedings of the IEEE National Aerospace Electronics Conference.

Anvari, M.B., Tavakoli Kashani, A., Rabieyan, R., 2017. Identifying the most important factors in the at-fault probability of motorcyclists by data mining, based on classification tree models. Int. J. Civil Eng. 15 (4), 653–662. https://doi.org/10.1007/s40999-017-0180-0.

Bahiru, T.K., Kumar Singh, D., Tessfaw, E.A., 2018. Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018.

Ben-David, S.S.-S.A.S. (2014). <understanding-machine-learning-theory-algorithms.pdf>. Cambridge University Press. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

Bhavsar, R., Amin, A., Zala, L., 2021. Development of model for road crashes and identification of accident spots [Article]. Int. J. Intell. Transp. Syst. Res. 19 (1), 99–111. https://doi.org/10.1007/s13177-020-00228-z.

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., Mikulec, R., 2020. Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia.

Champahom, T., Jomnonkwao, S., Chatpattananan, V., Karoonsoontawong, A., Ratanavaraha, V., 2019. Analysis of rear-end crash on Thai highway: decision tree approach. J. Adv. Transp. 2019, 1–13. https://doi.org/10.1155/2019/2568978.

Chen, M.-Y., 2012. Comparing traditional statistics, decision tree classification and support vector machine techniques for financial bankruptcy prediction. Intell. Autom. Soft Comput. 18 (1), 65–73. https://doi.org/10.1080/10798587.2012.10643227.

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Zupan, B., 2013. Orange: Data mining toolbox in python [Article]. J. Mach. Learn. Res. 14, 2349–2353. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885599052&partnerID=40&md5=75d2df52a0c46b5ab58ab08e1576114e.

DLT, 2021. Department of Land Transportation. https://www.dlt.go.th/th/public-news/view.php?_did=2806.

El Abdallaoui, H.E.A., El Fazziki, A., Ennaji, F.Z., Sadgal, M., 2018. Decision Support System for the Analysis of Traffic Accident Big Data. Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018.

Feng, M., Zheng, J., Ren, J., Xi, Y., 2020. Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. In Advances in Brain Inspired Cognitive Systems (pp. 520-529). 10.1007/978-3-030-39431-8_50.

Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests [Article]. Accid. Anal. Prev. 41 (1), 98–107. https://doi.org/10.1016/j.aap.2008.09.009.

Helen, W.R., Almetu, N., Nivethitha, S., 2019. Mining Road Accident Data Based on Diverted Attention of Drivers. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018.

Jafari Anarkooli, A., Hosseinpour, M., Kardar, A., 2017. Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. Accid. Anal. Prev. 106, 399–410. https://doi.org/10.1016/j.aap.2017.07.008.

Jomnonkwao, S., Uttra, S., Ratanavaraha, V., 2020. Forecasting road traffic deaths in Thailand: applications of time-series, curve estimation, multiple linear regression and path analysis models. *Sustainability* 12 (1). https://doi.org/10.3390/su12010395.

Jou, R.C., Yeh, T.H., Chen, R.S., 2012. Risk factors in motorcyclist fatalities in Taiwan. Traffic Inj Prev 13 (2), 155–162. https://doi.org/10.1080/15389588.2011.641166.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accid. Anal. Prev. 37 (5), 910–921. https://doi.org/10.1016/j.aap.2005.04.009.

Kim, J.-K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. Accid. Anal. Prev. 50, 1073–1081. https://doi.org/10.1016/j.aap.2012.08.011.

Krull, K.A., Khattak, A.J., Council, F.M., 2000. Injury effects of rollovers and events sequence in single-vehicle crashes. Transp. Res. Rec. 1717 (1), 46–54. https://doi.org/10.3141/1717-07.

Kumar, S., Toshniwal, D., 2016. A data mining approach to characterize road accident locations [Article]. J. Modern Transp. 24 (1), 62–72. https://doi.org/10.1007/s40534-016-0095-5.

Kuşkapan, E., Çodur, M.Y., Atalay, A., 2021. Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms [Article]. Accid. Anal. Prev. 155, 106098 https://doi.org/10.1016/j.aap.2021.106098.

Mafi, S., AbdelRazig, Y., Doczy, R., 2018. Machine learning methods to analyze injury severity of drivers from different age and gender groups. Transp. Res. Rec. 2672, 171–183.

Malin, F., Norros, I., Innamaa, S., 2019. Accident risk of road and weather conditions on different road types. Accid. Anal. Prev. 122, 181–188. https://doi.org/10.1016/j.aap.2018.10.014.

Mphela, T., 2020. Causes of road accidents in botswana: An econometric model [Article]. J. Transp. Supply Chain Manage., 14, 1-8, Article a509. 10.4102/jtscm.v14i0.509.

Osman, M., Mishra, S., Paleti, R., 2018. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: accounting for unobserved heterogeneity and age group differences. Accid. Anal. Prev. 118 https://doi.org/10.1016/j.aap.2018.05.004.

Ospina-Mateus, H., Quintana Jiménez, L. A., López-Valdés, F. J., Morales-Londoño, N., Salas-Navarro, K., 2019. Using Data-Mining Techniques for the Prediction of the

Severity of Road Crashes in Cartagena, Colombia. In Communications in Computer and Information Science (Vol. 1052, pp. 309-320).
Ospina-Mateus, H., Quintana Jiménez, L.A., Lopez-Valdes, F.J., Berrio Garcia, S., Barrero, L.H., Sana, S.S., 2021. Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists [Article]. J. Ambient Intell. Hum. Comput. 12 (11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5.
Pakgohar, A., Tabrizi, R.S., Khalili, M., Esmaeili, A., 2011. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Comput. Sci. 3, 764–769. https://doi.org/10.1016/j.procs.2010.12.126.
PDPM, 2020. Thailand Department of Public Disaster Prevention and Mitigation. https://www.disaster.go.th/en/.
Recal, F., Demirel, T., 2021. Comparison of machine learning methods in predicting binary and multi-class occupational accident severity [Article]. J. Intell. Fuzzy Syst. 40 (6), 10981–10998. https://doi.org/10.3233/JIFS-202099.
Rezapour, M., Mehrara Molan, A., Ksaibati, K., 2020. Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. Int. J. Transp. Sci. Technol. 9 (2), 89–99. https://doi.org/10.1016/j.ijtst.2019.10.002.
RSC, T., 2019. Thailand Accident Research Center Thailand Accident Research Center https://www.thairsc.com/.
Se, C., Champahom, T., Jomnonkwao, S., Chaimuang, P., Ratanavaraha, V., 2021. Empirical comparison of the effects of urban and rural crashes on motorcyclist injury severities: A correlated random parameters ordered probit approach with heterogeneity in means. Accid. Anal. Prev. 161, 106352 https://doi.org/10.1016/j.aap.2021.106352.
Shaheed, M.S., Gkritza, K., Zhang, W., Hans, Z., 2013. A mixed logit analysis of two-vehicle crash seventies involving a motorcycle. Accid. Anal. Prev. 61 https://doi.org/10.1016/j.aap.2013.05.028.
Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., Hanks, H., 2011. Risk factors for fatal crashes in rural Australia. Accid. Anal. Prev. 43 (3), 1082–1088. https://doi.org/10.1016/j.aap.2010.12.016.
Song, Y.-Y., Ying, L., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27 (2), 130.
WHO. (2018). World Health Organization: Global status report on road safety 2018. https://extranet.who.int/roadsafety/death-on-the-roads/.
Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D., 2007. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14 (1), 1–37. https://doi.org/10.1007/s10115-007-0114-2.
Xie, Y., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accid. Anal. Prevent. 47, 36–44. https://doi.org/10.1016/j.aap.2011.12.012.
Yu, M., Zheng, C., Ma, C., 2020. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. Anal. Methods Accid. Res. 27, 100126 https://doi.org/10.1016/j.amar.2020.100126.
Zhang, X.F., Fan, L., 2013. A decision tree approach for traffic accident analysis of Saskatchewan highways. Canadian Conference on Electrical and Computer Engineering.
Zhou, M., Chin, H.C., 2019. Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. Accid. Anal. Prev. 124, 104–112. https://doi.org/10.1016/j.aap.2019.01.009.

BIOGRAPHY

Mr. Ittirit Mohamad received his Bachelor's degree in Industrial Engineering from the Sirindhorn International Institute of Technology (SIIT), Thammasat University, in 2007, and a Master's degree in Public Administration with a concentration in Public and Private Organization Management from the National Institute of Development Administration (NIDA) in 2010.

From 2008 to 2018, he worked as an Engineering Manager for Operation Quality and Data Analytics at Seagate Technology (Thailand) Co., Ltd. He later obtained a Commercial Pilot License (CPL) from Bangkok Aviation Center (BAC) in 2019.

Since 2021, Mr. Mohamad has been a Project and Data Analysis Consultant at BRANDED, a partner of top-performing Amazon sellers, and is currently pursuing a Doctor of Philosophy degree in Energy and Logistics Management Engineering at the School of Mechanical Engineering, Suranaree University of Technology. His research interests include machine learning applications in engineering, big data analytics, transportation systems, accident analysis, and logistics management.