การเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลการเสียชีวิต จากอุบัติเหตุบนโครงข่ายถนน โดยใช้เทคนิคการทำเหมืองข้อมูล
Comparison of Imbalanced Data Classification Efficiency to Predict Road Accident Deaths using Data Mining Techniques

นรินทร์ จิวิตัน (Narin Jiwitan)*

Received: December 22, 2021
Revised: February 8, 2022
Accepted: February 21, 2022

* ผู้นิพนธ์ประสานงาน: นรินทร์ จิวิตัน (Narin Jiwitan) อีเมล: narin@rmutt.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องในการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนน และเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเสียชีวิตบนโครงข่ายถนนของแบบจำลองการพยากรณ์ ประกอบด้วยเทคนิคเกรเดียนท์บูตทรีส์ เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ เทคนิคแรนดอมฟอร์เรสต์ และเทคนิคการเรียนรู้เชิงลึก ซึ่งเป็นเทคนิคในการทำเหมืองข้อมูล กระบวนการวิจัยเป็นไปตามขั้นตอนของ CRISP-DM โดยศึกษาและรวบรวมข้อมูลอุบัติเหตุจากศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงคมนาคม ระหว่างวันที่ 1 มกราคม 2562 ถึง 30 มิถุนายน 2564 จำนวน 51,384 แถว ทำการคัดเลือกข้อมูล การกลั่นกรองข้อมูล การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมก่อนนำไปวิเคราะห์และสร้างแบบจำลอง ผลการวิเคราะห์ปัจจัยที่เกี่ยวข้องในการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนนจากการคำนวณค่าน้ำหนักของข้อมูลที่มีความเกี่ยวข้องหรือสัมพันธ์ของแอตทริบิวต์ด้วยวิธีการ Gain Ratio พบว่าประเภทผู้ใช้รถหรือใช้ถนนมีค่าน้ำหนักมากที่สุด 0.1030 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลของแบบจำลองการพยากรณ์ด้วยการวัดประสิทธิภาพของแบบจำลองการพยากรณ์ด้วยวิธีการ K-Fold Cross Validation โดยแบ่งชุดข้อมูลออกเป็น 30 ส่วน พบว่าเทคนิคที่มีประสิทธิภาพการจำแนกข้อมูลดีที่สุด คือ เทคนิคนาอีฟเบย์มีค่าความถูกต้อง 72.23% ค่าประสิทธิภาพโดยรวม 73.35% และค่าเฉลี่ยเรขาคณิต 72.10% จากการวิจัยนี้สามารถนำไปใช้เป็นข้อมูลป้องกันอุบัติภัยทางถนนในแต่ละพื้นที่ได้อย่างมีประสิทธิภาพ ลดความเสี่ยง ลดผลกระทบ และเตรียมความพร้อมในการป้องกันอุบัติภัยบนโครงข่ายถนนที่อาจจะเกิดขึ้น

คำสำคัญ: อุบัติเหตุ เกรเดียนท์บูตทรีส์ เทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ แรนดอมฟอร์เรสต์ การเรียนรู้เชิงลึก

Abstract

This research has an objective to analyze the factors involved in road accident deaths and to compare the classification efficiency of road fatalities data of forecasting models. Consists of Gradient Boosted Tree, Decision Tree, Naïve Bayes, Random Forest, and Deep Learning which is a data mining technique based on the CRISP-DM methodology. Study and collect accident data 51,384-row from the Information and Communication Technology Center Office of the Permanent Secretary for Transport between 1 January 2019 - 30 June 2021. The data was prepared and cleaned before being analyzed and modeled. The results of the analysis of factors involved in road accident deaths by calculating the weight of relevant data or related attributes using the Gain Ratio method, it was found that the type of vehicle or road user weighted 0.1030. The results of the comparison of the classification efficiency of forecast model data by measuring the performance of the forecast model using the K-Fold Cross Validation method; it was divided into 30 fold, showed that the Naïve Bayes technique had the highest accuracy of 72.23% F1-measure of 73.35% and G-mean of 72.10%. The results of this research can be used as information to effectively prevent road accidents in each area, reduce risks, reduce impacts and prepare for road accident prevention that may occur.

Keywords: Accident, Gradient Boosted Tree, Decision Tree, Naïve Bayes, Random Forest, Deep Learning.

* หลักสูตรระบบสารสนเทศทางธุรกิจ คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยราชมงคลล้านนา
* Business Information Systems department, Faculty of Business Administration and Liberal Arts, Rajamangala University of Technology Lanna.

1. บทนำ

ปัจจุบันอุบัติเหตุบนถนนกลายเป็นสาเหตุที่ทำให้เด็กและวัยรุ่นทั่วโลกเสียชีวิตมากที่สุด องค์การอนามัยโลก (World Health Organization - WHO) [1] ระบุว่าทวีปแอฟริกามีอัตราการเสียชีวิตจากอุบัติเหตุบนถนนสูงที่สุดในโลก จากรายงานสถานการณ์ความปลอดภัยทางถนน ปี 2015 (Global status report) พบว่ามีผู้เสียชีวิตจากอุบัติเหตุทางถนนทั่วโลกสูงถึง 1.25 ล้านคน บาดเจ็บจำนวน 20 – 50 ล้านคนในแต่ละปี ในขณะที่การเสียชีวิตจากอุบัติเหตุทางถนนในภูมิภาคเอเชียใต้และตะวันออกมีอัตราตายอยู่ที่ 17 ต่อประชากรแสนคน ในปี พ.ศ. 2557 ประเทศไทยเป็นประเทศที่มีอัตราการเสียชีวิตจากอุบัติเหตุบนถนนสูงสุดในเอเชียตะวันออกเฉียงใต้ มีผู้เสียชีวิตจากอุบัติเหตุทางถนนจำนวน 24,237 คน เป็นอันดับ 2 ของโลก ด้วยเหตุนี้สมัชชาสหประชาชาติได้ประกาศเจตนารมณ์ในปฏิญญามอสโกให้ปี พ.ศ. 2554 – 2563 เป็นทศวรรษแห่งความปลอดภัยทางถนน (Decade of Action for Road Safety) [2] ประเทศไทยในฐานะประเทศสมาชิกสมัชชาสหประชาชาติ ได้รวมขับเคลื่อนวาระความปลอดภัยทางถนนของโลก ตามมติคณะรัฐมนตรี เมื่อวันที่ 29 มิถุนายน 2553 กำหนดให้ “ปี พ.ศ. 2554 – 2563 เป็นทศวรรษแห่งความปลอดภัยทางถนน” มีเป้าหมายเพื่อลดอัตราการเสียชีวิตจากอุบัติเหตุทางถนนของคนไทยลงครึ่งหนึ่งหรือในอัตราที่ต่ำกว่า 10 คนต่อประชากรหนึ่งแสนคน [3] การเก็บสถิติจากกลุ่มป้องกันการบาดเจ็บจากการจราจรในประเทศไทย พบว่าในปี 2557 เพศชายมีการตายจากอุบัติเหตุทางถนนสูงกว่าเพศหญิงประมาณ 3 เท่า กลุ่มอายุที่เสียชีวิตจากอุบัติเหตุทางถนนมากที่สุด คือ กลุ่มอายุ 15 -19 ปี รถจักรยานยนต์เป็นพาหนะที่มีจำนวนการเสียชีวิตมากที่สุด ช่วงเวลาที่มีจำนวนผู้เสียชีวิตมากที่สุดคือเดือนมกราคม จังหวัดที่มีอัตราการเสียชีวิตต่อประชากรแสนคนสูงที่สุด คือ จังหวัดระยอง ทุกหน่วยงานต่างก็ให้ความสำคัญกับสาเหตุและปัจจัยเสี่ยงที่ก่อให้เกิดอุบัติเหตุบนถนน เช่น ทัศนวิสัยไม่ดี เมาสุรา ขับรถเร็วเกินกำหนด ตัดหน้ากระชั้นชิด หลับใน ฝ่าฝืนสัญญาณไฟจราจร การสวมหมวกนิรภัย การคาดเข็มขัดนิรภัย โทรศัพท์ขณะขับรถ เป็นต้น

จากปัญหาและที่มาดังกล่าวผู้วิจัยจึงได้ดำเนินการวิจัยเรื่องการเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนน โดยใช้เทคนิคการทำเหมืองข้อมูล มีจุดประสงค์เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องในการได้รับอุบัติเหตุเสียชีวิตบนโครงข่ายถนน และเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนนของแบบจำลองการพยากรณ์ประกอบด้วยเทคนิคเกรเดียนท์บูตทรีส์ (Gradient Boosted Tree) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคแรนดอมฟอร์เรสต์ (Random Forest) และเทคนิคการเรียนรู้เชิงลึก (Deep Learning) คาดว่างานวิจัยนี้จะเป็นประโยชน์และเป็นข้อมูลป้องกันอุบัติภัยทางถนนในแต่ละพื้นที่ได้อย่างมีประสิทธิภาพ ลดความเสี่ยง ลดผลกระทบ และเตรียมความพร้อมในการป้องกันอุบัติภัยบนโครงข่ายถนนที่อาจจะเกิดขึ้น

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

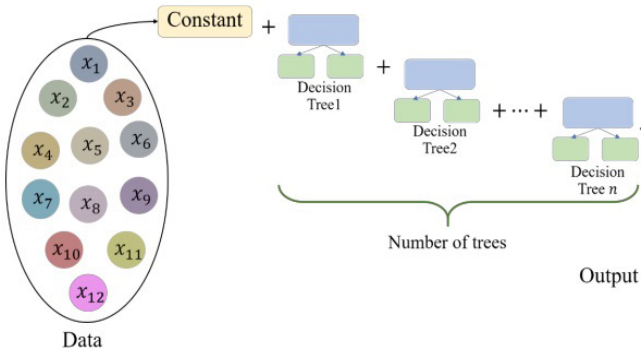
2.1 ทฤษฎีที่เกี่ยวข้อง

ความไม่สมดุลของข้อมูล (Imbalanced data) เป็นข้อมูลที่มีจำนวนข้อมูลในกลุ่มหนึ่งมากกว่าจำนวนข้อมูลของอีกกลุ่มหนึ่งเป็นจำนวนมาก ซึ่งเป็นไปตามธรรมชาติของข้อมูลที่มีความแตกต่างของจำนวนในแต่ละกลุ่ม ข้อมูลที่ไม่สมดุลจะส่งผลให้ไม่สามารถจำแนกข้อมูลของกลุ่มที่มีจำนวนข้อมูลน้อยได้ถูกต้องแม่นยำ ในทางกลับกันจะสามารถจำแนกข้อมูลของกลุ่มที่มีจำนวนมากได้อย่างแม่นยำ ดังนั้น ควรปรับความสมดุลของข้อมูลก่อนการนำไปพยากรณ์ ซึ่งมีอยู่ 4 วิธี คือ วิธีการสุ่มลด (Under-sampling) วิธีการสุ่มเพิ่ม (Over-sampling) วิธีการผสมผสาน (Hybrid method) และ วิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling TEchnique: SMOTE)

การทำเหมืองข้อมูล (Data mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยทำการจำแนกประเภท รูปแบบ เชื่อมโยงข้อมูลที่มีความสัมพันธ์กันและหาความน่าจะเป็นที่จะเกิดขึ้น เพื่อให้ได้องค์ความรู้ใหม่ที่สามารถนำไปใช้ประกอบการตัดสินใจในด้านต่าง ๆ เทคนิคการจำแนกประเภทข้อมูล (Data Classification) ถือเป็นหนึ่งในการทำเหมืองข้อมูลด้วยพื้นฐานของการเรียนรู้แบบมีผู้สอน (Supervised Learning)

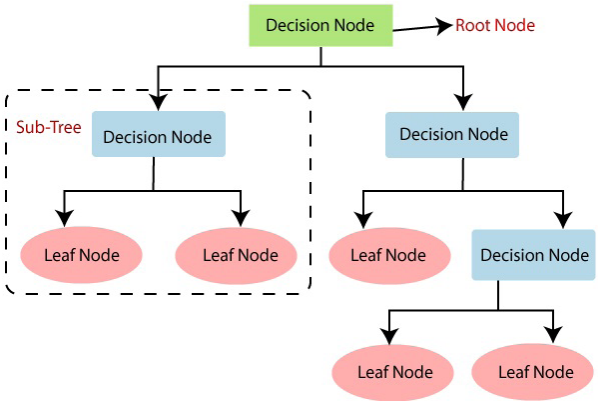
เทคนิคเกรเดียนท์บูตทรี (Gradient Boost tree) [4] เป็นเทคนิคการเรียนรู้ของเครื่อง ใช้แก้ปัญหาการถดถอยและปัญหาการจำแนกประเภท โดยการสุ่มสร้างต้นไม้

การตัดสินใจหลายร้อยแบบจำลองและประเมินผลแต่ละแบบจำลองจนกว่าจะได้แบบจำลองที่เหมาะสมที่สุด ลดความคลาดเคลื่อนจากการเรียนรู้ก่อนหน้า สามารถจัดการกับข้อมูลสูญหาย (Missing Value) มีพารามิเตอร์ที่หลากหลายให้ปรับเพื่อให้เหมาะสมกับแบบจำลอง อาทิเช่น จำนวนต้นไม้ (number of tree) จำนวนชั้นสูงสุดของต้นไม้ (maximal depth) อัตราการเรียนรู้ (learning rate) เป็นต้น ดังภาพที่ 1



ภาพที่ 1 เทคนิคเกรเดียนท์บูตทรี

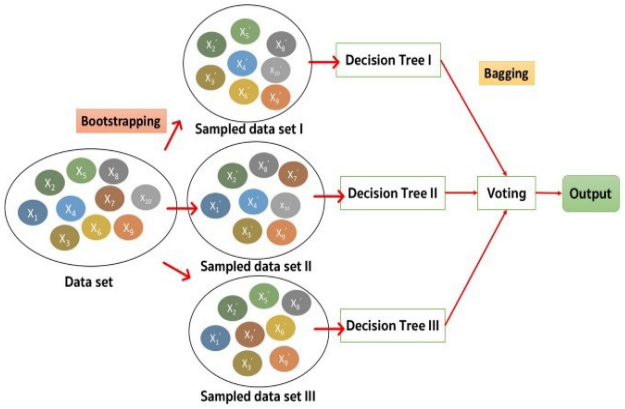
เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นแบบจำลองทางคณิตศาสตร์ที่ใช้จำแนกประเภทข้อมูล ประสิทธิภาพของแบบจำลองขึ้นอยู่กับคุณลักษณะของข้อมูลที่นำมาใช้ ค่าของคุณลักษณะจะอยู่ในรูปของค่าที่ไม่ต่อเนื่อง (Discrete Value) เป็นโครงสร้างข้อมูลชนิดเป็นลำดับชั้น (Hierarchy) มีโหนดราก (RootNode) อยู่ด้านบนสุดและโหนดใบอยู่ล่างสุดของต้นไม้ ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ (Attribute) เป็นตัวทดสอบ กิ่งของต้นไม้ (Branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (Leaf) อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (Class) หรือผลลัพธ์ที่ได้จากการทำนาย



ภาพที่ 2 เทคนิคต้นไม้ตัดสินใจ

เทคนิคนาอีฟเบย์ (Naïve Bayes) เป็นขั้นตอนวิธีในการจำแนกข้อมูลที่อาศัยหลักการความน่าจะเป็น (Probability) ตามทฤษฎีของเบย์ (Bayes's theorem) โดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ มีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับการกระจายความน่าจะเป็น (Probability Distribution) เป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน

เทคนิคแรนดอมฟอร์เรสต์ (Random Forest) เป็นการสร้างแบบจำลองด้วยวิธีการต้นไม้ตัดสินใจขึ้นมาหลาย ๆ แบบจำลอง โดยวิธีการสุ่มตัวแปร แล้วนำผลที่ได้แต่ละแบบจำลองมารวมกัน พร้อมนับจำนวนผลที่มีจำนวนซ้ำกันมากที่สุด สกัดออกมาเป็นผลลัพธ์สุดท้าย



ภาพที่ 3 เทคนิคแรนดอมฟอร์เรสต์

เทคนิคการเรียนรู้เชิงลึก (Deep Learning) [5] เป็นวิธีการเรียนรู้แบบอัตโนมัติด้วยการเลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ (Neurons) โดยนำระบบโครงข่ายประสาท (Neural Network) มาซ้อนกัน หลายชั้น (Layer) และทำการเรียนรู้ข้อมูลตัวอย่าง ข้อมูลดังกล่าวจะถูกนำไปใช้ในการจัดหมวดหมู่ข้อมูล (Classify the Data)

คอนฟิวชันเมทริกซ์ (Confusion Matrix) เป็นตารางประเมินผลลัพธ์การทำนายของแบบจำลอง สามารถวัดจากผลลัพธ์การจำแนกข้อมูล (Classification) รูปแบบของคอนฟิวชันเมทแสดงดังตารางที่ 1

ตารางที่ 1 *แสดงเมตริกซ์วัดประสิทธิภาพ*

	Actual Positive	Actual Negative
Predict Positive	True Positive (TP)	False Positive (FP)
Predict Negative	False Negative (FN)	True Negative (TN)

ค่าที่ได้จากการทำนาย (Prediction) อธิบายได้ ดังนี้

ค่า TP คือจำนวนข้อมูลทดสอบที่เป็นคลาส Positive และแบบจำลองจำแนกได้ถูกต้องว่าเป็น Positive

ค่า FP คือจำนวนข้อมูลทดสอบที่ไม่ใช่คลาส Positive แต่แบบจำลองทำนายผิดว่าเป็นคลาส Positive

ค่า TN คือจำนวนข้อมูลทดสอบที่เป็นคลาส Negative และแบบจำลองทำนายได้ถูกต้องว่าเป็นคลาส Negative

ค่า FN คือจำนวนข้อมูลทดสอบที่เป็นคลาส Positive แต่แบบจำลองทำนายผิดว่าเป็นคลาส Negative

2.2 งานวิจัยที่เกี่ยวข้อง

จตุภพ ดิษผล และ อัญสุรีย์ ศิริโสภณ [6] ได้ศึกษาปัจจัยที่ส่งผลต่อการป้องกันอุบัติเหตุทางถนนของประชาชนในเขตอำเภอสว่างอารมณ์ จังหวัดอุทัยธานี: การวิเคราะห์ MIMIC model มีวัตถุประสงค์เพื่อพัฒนาและตรวจสอบความสอดคล้องของแบบจำลองเชิงสาเหตุที่ส่งผลต่อการป้องกันอุบัติเหตุทางถนนของประชาชนในเขตอำเภอสว่างอารมณ์ จังหวัดอุทัยธานี กับข้อมูลเชิงประจักษ์ ผลการวิจัยปรากฏว่าแบบจำลองเชิงสาเหตุที่พัฒนาขึ้นตามสมมติฐานมีความสอดคล้องกับข้อมูลเชิงประจักษ์ โดยมีแนวทางการป้องกันอุบัติเหตุทางถนนในด้านการบริหารจัดการความปลอดภัยทางถนน ด้านถนนและการสัญจรอย่างปลอดภัย ด้านยานพาหนะที่ปลอดภัย และด้านผู้ใช้รถใช้ถนนอย่างปลอดภัย และมีปัจจัยที่ส่งผลต่อการป้องกันอุบัติเหตุทางถนนอย่างมีนัยสำคัญทางสถิติ $p<.05$ ได้แก่ ยานพาหนะ ผู้ขับขี่ ถนน และสิ่งแวดล้อมรอบข้าง มีขนาดอิทธิพลเท่ากับ 0.40, 0.32, 0.24 และ 0.17 ตามลำดับ โดยที่ปัจจัยทั้ง 4 นี้ สามารถร่วมกันอธิบายความแปรปรวนในการป้องกันอุบัติเหตุทางถนนได้ร้อยละ 58

รัชพล กลัดชื่น และ จรัญ แสนราช [7] ได้ศึกษาการเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษาโดยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา โดยใช้เทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ Decision Tree: J48graft, Naïve Bayes และ Rule Induction ทำการเปรียบเทียบประสิทธิภาพแบบจำลองการทำนาย ผลการศึกษาพบว่าการใช้เทคนิค Decision Tree: J48graft ด้วยการเลือกคุณลักษณะแบบ Forward Selection และ การเลือกคุณลักษณะทั้งหมดมีค่าความถูกต้องเท่ากับ 83.08% และ 81.71% ตามลำดับ

สายชล สินสมบูรณ์ทอง [8] ได้ศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล วิธีการจำแนกที่นำมาเปรียบเทียบมี 7 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้อัลกอริทึมชนิด IBk วิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึมชนิด J48 วิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล วิธีฐานกฎโดยใช้อัลกอริทึม decision table วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ ผลการศึกษาพบว่าชุดข้อมูล fertility วิธีการถดถอยลอจิสติกทวิภาคที่ random seed = 10, 20 และ 30 มีค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.0000, 1.0000 และ 0.00000 ตามลำดับ ส่วนชุดข้อมูล vertibral volumn วิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 มีค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.0000, 1.0000 และ 0.00024 ตามลำดับ และชุดข้อมูล diabetes วิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed =10, 20 และ 30 มีค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.0000, 1.0000 และ 0.00004 ตามลำดับ และเมื่อพิจารณาข้อมูลทั้ง 3 ชุด ร่วมกัน พบว่าวิธีที่ดีที่สุดในการทำนายผล คือ วิธีเพื่อนบ้านใกล้สุด k ตัว

3. วิธีดำเนินการวิจัย

กระบวนการวิจัยมีขั้นตอนที่เป็นไปตามกรอบแนวคิดของ CRISP – DM จำนวน 6 ขั้นตอน ดังภาพที่ 4

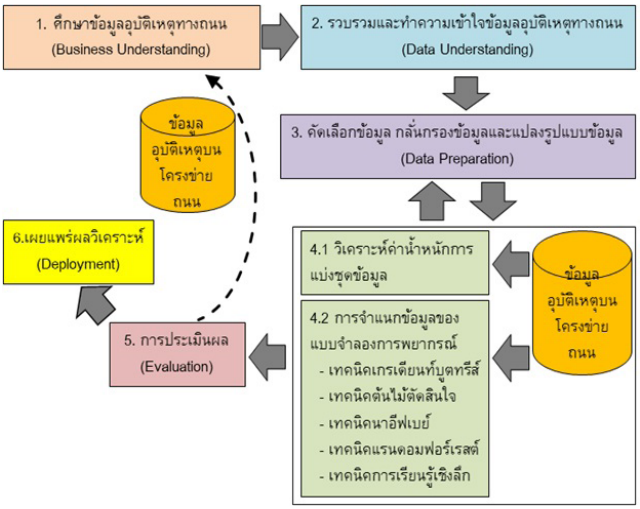
3.1 ศึกษาข้อมูลอุบัติเหตุบนโครงข่ายถนน (Business Understanding)

ทำความเข้าใจสาเหตุของอุบัติเหตุทางถนนก่อนทำความเข้าใจข้อมูลที่จะใช้ในการวิเคราะห์ จากการศึกษาพบว่า

ข้อมูลอุบัติเหตุจรจรทางถนนมาจาก 4 สาเหตุหลัก ได้แก่ 1) สาเหตุจากบุคคล เช่น ขับรถขณะมึนเมา ขับรถเร็วเกินกว่ากำหนด คนขับรถใช้โทรศัพท์ขณะขับรถ ขับรถโดยประมาท ตัดหน้ารถกระชั้นชิด คนเดินถนนหรือคนข้ามถนน เป็นต้น 2) สาเหตุจากยานพาหนะ เช่น อุปกรณ์บกพร่อง กระจกส่องหลังที่ปัดน้ำฝน เบรก ไฟสัญชำรุด เป็นต้น 3) สาเหตุจากเครื่องหมายสัญญาณและทาง เช่น เครื่องหมายสัญญาณชำรุด ทางแยก ทางโค้ง ทางชำรุด เป็นต้น และ 4) สาเหตุจากธรรมชาติ เช่น หมอกลงจัด ฝนตก ควัน เป็นต้น

ประเภททางหลวงในไทยมี 5 ประเภท ประกอบด้วย ทางหลวงพิเศษ ทางหลวงแผ่นดิน ทางหลวงชนบท ทางหลวงท้องถิ่น และทางหลวงสัมปทาน

ระดับความรุนแรงจากอุบัติเหตุจรจรทางบก [9] คือ มีผู้เสียชีวิต มีผู้บาดเจ็บสาหัส มีผู้บาดเจ็บและทรัพย์สินเสียหาย



ภาพที่ 4 *กรอบแนวคิดการวิจัย*

3.2 รวบรวมและทำความเข้าใจข้อมูลอุบัติเหตุบนโครงข่ายถนน (Data Understanding)

ขั้นตอนการจัดเก็บ รวบรวมข้อมูล ตลอดจนการพิจารณาตรวจสอบความถูกต้องของข้อมูลที่ได้รับ โดยเลือกว่าจะใช้ข้อมูลทั้งหมดหรือบางส่วนในการวิเคราะห์ให้สอดรับกับวัตถุประสงค์ที่กำหนดไว้

แหล่งข้อมูลการวิจัยนี้ได้จากแฟ้มข้อมูลอุบัติเหตุบนโครงข่ายถนนของกระทรวงคมนาคม ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงคมนาคม ระหว่างวันที่ 1 มกราคม 2562 - 30 มิถุนายน 2564 จำนวน 51,384 แถว ดังตารางที่ 2

ตารางที่ 2 *ข้อมูลแอตทริบิวท์รายงานการเกิดอุบัติเหตุ*

ชื่อแอตทริบิวท์	ชนิดข้อมูล	ตัวอย่างข้อมูล
วันที่เกิดเหตุ	วันที่	1/1/2019
ช่วงเวลาที่เกิดเหตุ	เวลา	13:00
วันที่รายงาน	วันที่	1/2/2019
เวลาที่รายงาน	เวลา	6:11
ACC_CODE	ข้อความ	571905
ประเภททางหลวงที่เกิดอุบัติเหตุ	ข้อความ	กรมทางหลวงชนบท
สายทาง	ข้อความ	แยกทางหลวงหมายเลข 21 (กม.ที่ 31+000) - บ้านวังเพลิง
ก.ม.	จำนวนจริง	4.00
จังหวัดที่เกิดเหตุ	ข้อความ	ลพบุรี
ประเภทยานพาหนะที่เกิดเหตุ	ข้อความ	รถจักรยานยนต์
บริเวณที่เกิดเหตุ	ข้อความ	ทางตรง+ไม่มีความลาดชัน
สาเหตุการเกิดอุบัติเหตุ	ข้อความ	เมาสุรา
ลักษณะการชน	ข้อความ	พลิกคว่ำ/ตกถนนในทางตรง
จำนวนรถที่เกิดเหตุ	จำนวนเต็ม	1
จำนวนผู้บาดเจ็บ	จำนวนเต็ม	2
ระดับความรุนแรง หรือจำนวนผู้เสียชีวิต	จำนวนเต็ม	1
สภาพอากาศขณะเกิดเหตุ	ข้อความ	แจ่มใส
LATITUDE	จำนวนจริง	14.959105
LONGITUDE	จำนวนจริง	100.873463

3.3 คัดเลือกข้อมูล กลั่นกรองข้อมูลและแปลงรูปแบบข้อมูล (Data Preparation)

3.3.1 การคัดเลือกข้อมูล (Data Selection) เป็นการคัดเลือกข้อมูลที่เหมาะสมเพื่อนำมาใช้ในการวิเคราะห์ข้อมูลให้เหลือเฉพาะแอตทริบิวท์ที่จำเป็นสำหรับการวิเคราะห์จำนวน 8 แอตทริบิวท์ ได้แก่ 1) ช่วงเวลาที่เกิดเหตุ 2) ประเภททางหลวงที่เกิดอุบัติเหตุ 3) จังหวัดที่เกิดเหตุ 4) ประเภทยานพาหนะที่เกิดเหตุ

5) บริเวณที่เกิดเหตุ 6) สาเหตุการเกิดอุบัติเหตุ 7) ลักษณะการชน และ 8) สภาพอากาศขณะเกิดเหตุ สำหรับแอตทริบิวท์ระดับความรุนแรงหรือจำนวนผู้เสียชีวิตใช้เป็นหมวดหมู่ของข้อมูลพยากรณ์

3.3.2 การกลั่นกรองข้อมูล (Data Cleaning) เป็นการทำความสะอาดข้อมูล ตรวจสอบ แก้ไขหรือลบรายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล ตารางหรือฐานข้อมูล ซึ่งได้ดำเนินการ ดังนี้

กำจัดแถวรายการที่มีข้อมูลไม่ถูกต้อง (Outlier) ได้แก่ ประเภทยานพาหนะที่เกิดเหตุ มีข้อมูลที่เป็นค่าว่าง (Missing Value) จำนวน 248 แถว สาเหตุการเกิดอุบัติเหตุมีข้อมูลที่เป็นค่าว่างจำนวน 279 แถว ลักษณะการชนมีข้อมูลที่เป็นค่าว่างจำนวน 741 รายการ จังหวัดที่เกิดเหตุมีข้อมูลที่เป็นค่าว่าง 28 แถว

แก้ไขข้อมูลแถวรายการบริเวณที่เกิดเหตุมีข้อมูลเดิมเป็นค่าว่าง เปลี่ยนเป็น “ไม่ระบุ” จำนวน 4,234 แถว

จัดกลุ่มข้อมูลเวลาเกิดเหตุเป็น 2 ประเภท [10] คือ ตั้งแต่ 06.00-17.59 น. เป็น “กลางวัน” และ 18.00-05.59 น. เป็น “กลางคืน”

จัดกลุ่มข้อมูลระดับความรุนแรงหรือจำนวนผู้เสียชีวิตเป็น 2 ประเภท คือ ถ้าจำนวนผู้เสียชีวิตเท่ากับ 0 คือ “ไม่เสียชีวิต” และถ้าจำนวนผู้เสียชีวิตมากกว่า 0 คือ “เสียชีวิต”

จัดกลุ่มข้อมูลบริเวณที่เกิดเหตุ [9] ตัวอย่างเช่น บริเวณลักษณะถนนที่เกิดหตุ เช่น สามแยก วนเวียน ห้าแยก ทางร่วม จุดเข้า/ออก ทางหลัก จุดเข้าออกสถานที่ คือ ทางแยก บริเวณลักษณะถนนที่เกิดหตุเป็นสะพาน ทางลอด จุดตัดทางรถไฟ จุดเปลี่ยนความกว้างถนน ทางจักรยาน ทางเดินเท้า ทางม้าลาย คือ บริเวณเฉพาะ เป็นต้น

จัดกลุ่มข้อมูลสาเหตุการเกิดอุบัติเหตุ เช่น สาเหตุของอุบัติเหตจากเบรก ไฟสัญญาณ กระจกส่องหลัง ที่ปัดน้ำฝน คือ สาเหตุที่เกิดจากอุปกรณ์ชำรุด สาเหตุของอุบัติเหตุจากไม่มีป้ายจราจร ป้ายจราจรชำรุด เครื่องหมายสัญญาณชำรุด คือ สาเหตุที่เกิดจากป้ายเครื่องหมาย เป็นต้น

จัดกลุ่มข้อมูลจังหวัดที่เกิดอุบัติเหตุ 77 จังหวัด [11] โดยจำแนกเป็นภูมิภาคตามลักษณะภูมิประเทศแทน ประกอบด้วย 6 ภูมิภาค ได้แก่ ภาคเหนือ ภาคกลาง ภาคตะวันออกเฉียงเหนือ ภาคตะวันออก ภาคตะวันตก ภาคใต้

หลังจากคัดเลือกข้อมูลและการกลั่นกรองข้อมูลเรียบร้อยแล้ว คงเหลือข้อมูลจำนวน 51,000 แถว

พบว่ามีแอตทริบิวท์ระดับความรุนแรงหรือจำนวนผู้เสียชีวิตจำนวน 1 แอตทริบิวท์ เกิดปัญหาความไม่สมดุลของข้อมูล โดยมีอัตราส่วนข้อมูล “เสียชีวิต” : “ไม่เสียชีวิต” คิดเป็นร้อยละ 13.34 : 86.11 ดังนั้น จึงได้ทำการปรับความสมดุลของข้อมูลเพื่อให้ข้อมูลใกล้เคียงกันด้วยวิธีการสุ่มลด (Under-sampling) ทำให้อัตราส่วนข้อมูล “เสียชีวิต” : “ไม่เสียชีวิต” คิดเป็นร้อยละ 49.83 : 50.17 คงเหลือข้อมูลสำหรับนำไปวิเคราะห์จำนวน 13,653 แถว แปลงข้อมูลด้วยวิธีการนอร์มอลไลซ์ปรับการกระจายของข้อมูล และค่าเบี่ยงเบนมาตรฐาน (Z-Score Normalization) ก่อนนำไปวิเคราะห์

3.4 การวิเคราะห์ข้อมูลและสร้างแบบจำลอง

3.4.1 วิเคราะห์ค่าน้ำหนักการแบ่งชุดข้อมูลโดยใช้ Gain Ratio [12] ซึ่งเป็นตัวชี้วัดการแบ่งชุดข้อมูลออกเป็นชุดย่อยที่พัฒนามาจาก Information Gain เนื่องจากการใช้ Information Gain ในการแบ่งชุดข้อมูลมีโอกาสทำให้เกิดความเอนเอียงขึ้น เมื่อคุณลักษณะที่ทำการพิจารณาได้ค่า Gain ที่สูงเป็นจำนวนมาก จะส่งผลให้คุณลักษณะที่ถูกคัดเลือกไม่ถูกต้อง

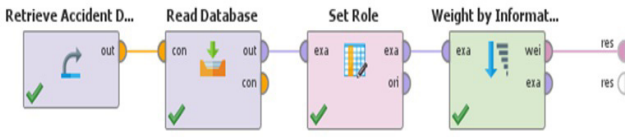
การหาค่า Gain Ratio คำนวณได้จากสมการที่ 1

$$\text{GainRatio}\ (A) = \frac{\text{Gain}\ (\text{A})}{\text{SplitInfo}_{\text{A}}\ (D)} \qquad (1)$$

ค่า $\text{SplitInfo}_{\text{A}}(\text{D})$ หมายถึง ปริมาณข้อมูลที่ถูกพิจารณาโดยการแบ่งชุดข้อมูลในชุดข้อมูล D ออกเป็น m ชุด ข้อมูลย่อยตามคุณลักษณะ A ดังสมการที่ 2

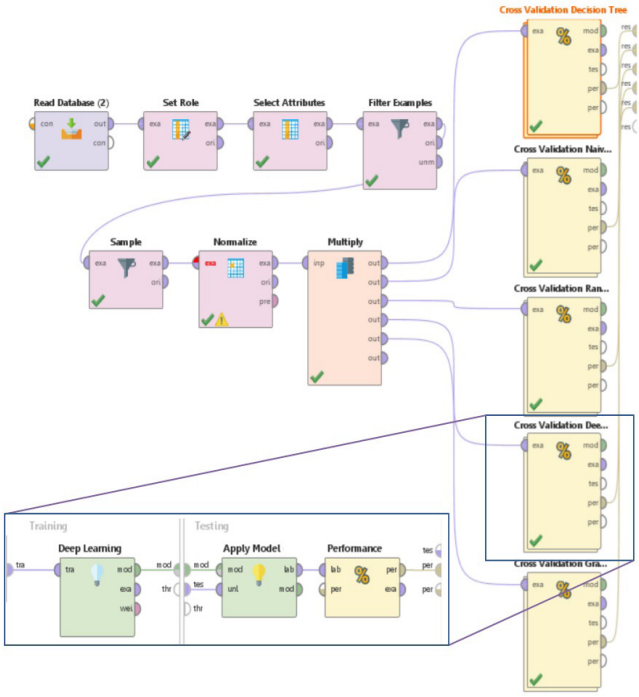
$$SplitInfo_{\text{A}}\ (D) = -\sum_{\text{j}=1}^{m} \frac{|\text{D}_\text{j}|}{|D|}\ X\ log_2 \frac{|\text{D}_\text{j}|}{|D|} \qquad (2)$$

การวิเคราะห์ค่าน้ำหนักการแบ่งชุดข้อมูล Gain Ratio จำนวน 8 แอตทริบิวต์ จากสมการข้างต้นและใช้โปรแกรม Rapidminer Studio สนับสนุนการประมวลผลดังภาพที่ 5



ภาพที่ 5 *การวิเคราะห์ค่าน้ำหนักการแบ่งชุดข้อมูล*

3.4.2 สร้างแบบจำลอง (Modeling) เลือกใช้เทคนิคการจัดทำแบบจำลองที่เป็นมาตรฐานเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด โดยออกแบบและสร้างแบบจำลองการพยากรณ์จำนวน 5 แบบจำลอง ประกอบด้วย เทคนิคเกรเดียนท์บูตทรีส์ เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ เทคนิคแรนดอมฟอร์เรสต์ และเทคนิคการเรียนรู้เชิงลึก ซึ่งใช้สมการทางคณิตศาสตร์และสถิติในการคำนวณ และใช้โปรแกรม Rapidminer Studio สนับสนุนการประมวลผล ดังภาพที่ 6



ภาพที่ 6 *ออกแบบและสร้างแบบจำลองการพยากรณ์*

3.5 การประเมินผล (Evaluation)

การประเมินผลเป็นการวัดประสิทธิภาพของแบบจำลองการพยากรณ์มีนัยสำคัญหรือความน่าเชื่อถือมากน้อยเพียงใด ในกรณีที่ความน่าเชื่อถือยังไม่เพียงพอจะกลับไปทบทวนหัวข้อที่ 3.2 – 3.4 ซ้ำอีกครั้ง

การทดสอบประสิทธิภาพของแบบจำลองได้ดำเนินการวัดค่าความแม่นยำ ค่าเรียกกลับ ค่าความจำเพาะ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต โดยวิธีการ K-Fold Cross Validation Test แบ่งชุดข้อมูลเป็น 10 ส่วน 20 ส่วน และ 30 ส่วน ในแต่ละส่วนมาจากการสุ่มเพื่อให้ข้อมูลกระจายเท่า ๆ กัน เพื่อสร้างแบบจำลองการพยากรณ์จากการเรียนรู้และทดสอบแบบจำลองการพยากรณ์

ค่าความถูกต้อง (Accuracy) คำนวณจากสมการที่ 3

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \quad (3)$$

ค่าความแม่นยำ (Precision) คำนวณจากสมการที่ 4

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

ค่าเรียกกลับ (Recall หรือ Sensitivity) คำนวณจากสมการที่ 5

$$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (5)$$

ค่าความจำเพาะ (Specificity) คำนวณจากสมการที่ 6

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (6)$$

ค่าประสิทธิภาพโดยรวม (F-measure) คำนวณจากสมการที่ 7

$$\text{F1} - \text{measure} = 2\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (7)$$

ค่าเฉลี่ยเรขาคณิต (G-mean) คำนวณจากสมการที่ 8

$$\text{G} - \text{mean} = \sqrt{\text{Sensitivity} * \text{Specificity}} \quad (8)$$

3.6 เผยแพร่ผลวิเคราะห์ (Deployment)

เผยแพร่ผลวิเคราะห์หรือการนำผลลัพธ์ที่ได้ไปใช้งานเพื่อที่หน่วยงานต่าง ๆ จะได้นำไปใช้ประโยชน์และเป็นข้อมูลป้องกันอุบัติภัยทางถนนในแต่ละพื้นที่ได้อย่างมีประสิทธิภาพ ลดความเสี่ยง ลดผลกระทบ และเตรียมความพร้อมในการป้องกันอุบัติภัยบนโครงข่ายถนนที่อาจจะเกิดขึ้น

4. สรุปผลการวิจัย

4.1 ผลการคำนวณปัจจัยที่เกี่ยวข้องในการได้รับอุบัติเหตุเสียชีวิต

ผลการคำนวณค่าน้ำหนักการแบ่งชุดข้อมูลด้วยวิธีการ Gain Ratio จำนวน 8 แอตทริบิวต์ แสดงดังตารางที่ 3

จากตารางที่ 3 ปัจจัยที่เกี่ยวข้องกับการได้รับอุบัติเหตุบนโครงข่ายถนนของกระทรวงคมนาคม ด้วยการคำนวณหาค่าน้ำหนักการแบ่งชุดข้อมูลด้วยวิธี Gain Ratio พบว่าประเภทยานพาหนะที่เกิดเหตุมีค่าน้ำหนัก 0.1030 มากที่สุด รองลงมาเป็นลักษณะการชนมีค่าน้ำหนัก 0.0893

ประเภททางหลวงที่เกิดอุบัติเหตุมีค่าน้ำหนัก 0.0215 สาเหตุการเกิดอุบัติเหตุมีค่าน้ำหนัก 0.0156 ภูมิภาคที่เกิดอุบัติเหตุมีค่าน้ำหนัก 0.0135 สภาพอากาศขณะเกิดเหตุมีค่าน้ำหนัก 0.0125 ช่วงเวลาที่เกิดเหตุมีค่าน้ำหนัก 0.0111 และบริเวณที่เกิดเหตุมีค่าน้ำหนักน้อยที่สุด 0.0059

ตารางที่ 3 *ผลการวิเคราะห์ค่าน้ำหนักการแบ่งชุดข้อมูล*

แอตทริบิวต์	Gain Ratio
ประเภทยานพาหนะที่เกิดเหตุ	0.1030
ลักษณะการชน	0.0893
ประเภททางหลวงที่เกิดอุบัติเหตุ	0.0215
สาเหตุการเกิดอุบัติเหตุ	0.0156
ภูมิภาคที่เกิดอุบัติเหตุ	0.0135
สภาพอากาศขณะเกิดเหตุ	0.0125
ช่วงเวลาที่เกิดเหตุ	0.0111
บริเวณที่เกิดเหตุ	0.0059

4.2 ผลการคำนวณประสิทธิภาพการจำแนกข้อมูลของแบบจำลองการพยากรณ์

ผลการทดสอบประสิทธิภาพการจำแนกข้อมูลของแบบจำลองพยากรณ์ด้วยวิธี Cross-validation ได้แบ่งชุดข้อมูลออกเป็น 10 ส่วน 20 ส่วน และ 30 ส่วน แต่ละส่วนมีจำนวนข้อมูลเท่ากัน ใช้ฟังก์ชัน use local random seed สำหรับการสุ่มตัวอย่างของชุดข้อมูลย่อย เพื่อให้เกิดชุดย่อยเหมือนกัน ผลการคำนวณค่าความแม่นยำ ค่าเรียกกลับ และค่าความจำเพาะ แสดงดังตารางที่ 4 ผลการคำนวณค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต แสดงดังตารางที่ 5

จากตารางที่ 5 พบว่าผลการคำนวณค่าความถูกต้อง กรณีที่มีการแบ่งชุดข้อมูลออกเป็น 30 ส่วน เพื่อดำเนินการเรียนรู้และทดสอบ เทคนิคนาอีฟเบย์มีค่าความถูกต้องสูงสุด 72.23% รองลงมาเทคนิคเกรเดียนท์บูตทรีส์มีค่า 72.17% เทคนิคต้นไม้ตัดสินใจมีค่าความถูกต้อง 71.98% เทคนิคการเรียนรู้เชิงลึกมีค่า 71.95% และเทคนิคแรนดอมฟอร์เรสต์มีค่าน้อยสุด 71.23%

ผลการคำนวณประสิทธิภาพโดยรวม กรณีที่มีการแบ่งชุดข้อมูลออกเป็น 30 ส่วน เพื่อดำเนินการเรียนรู้และทดสอบ เทคนิคนาอีฟเบย์มีค่าประสิทธิภาพโดยรวมมากที่สุด 73.35% รองลงมาเทคนิคต้นไม้ตัดสินใจมีค่า 73.07% เทคนิคแรนดอมฟอร์เรสต์มีค่า 72.59% เทคนิคการเรียนรู้เชิงลึกมีค่า 72.50% และเทคนิคเกรเดียนท์บูตทรีส์มีค่าน้อยที่สุด 70.99%

ผลการคำนวณค่าเฉลี่ยเรขาคณิต กรณีที่มีการแบ่งชุดข้อมูลออกเป็น 30 ส่วน เพื่อดำเนินการเรียนรู้และทดสอบ เทคนิคนาอีฟเบย์มีค่าเฉลี่ยเรขาคณิตมากที่สุด 72.10% รองลงมาคือเทคนิคเกรเดียนท์บูตทรีส์มีค่า 72.06% เทคนิคต้นไม้ตัดสินใจมีค่า 71.86% เทคนิคการเรียนรู้เชิงลึกมีค่า 71.68% และเทคนิคแรนดอมฟอร์เรสต์มีค่าน้อยที่สุด 71.08%

5. อภิปลายผลการวิจัยและข้อเสนอแนะ

5.1 อภิปลายผลการวิจัย

การวิจัยการเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนน มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องในการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนน และเพื่อเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเสียชีวิตบนโครงข่ายถนน โดยใช้เทคนิคการทำเหมืองข้อมูล มีกระบวนการวิจัยเป็นไปตามขั้นตอนของ CRISP- DM ศึกษาข้อมูล รวบรวม ทำความเข้าใจข้อมูลอุบัติเหตุบนโครงข่ายถนน ดำเนินการคัดเลือกข้อมูล กลั่นกรองข้อมูล แปลงข้อมูลด้วยวิธีการนอร์มอลไลซ์ ทำการแปลงค่าข้อมูล ปรับการกระจายของข้อมูลค่าเบี่ยงเบนมาตรฐาน (Z-Score Normalization) ปรับปรุงข้อมูลกลุ่มตัวอย่างด้วยวิธีการสุ่มลด เพื่อให้ข้อมูลมีจำนวนใกล้เคียงกันก่อนนำไปวิเคราะห์ ผลการศึกษาพบว่า

5.1.1 ปัจจัยที่เกี่ยวข้องในการเสียชีวิตบนโครงข่ายถนนจากข้อมูลของกระทรวงคมนาคมที่ทำการเปรียบเทียบทั้งหมด 8 แอตทริบิวต์ เป็นข้อมูลที่ได้จากการวิเคราะห์ค่าน้ำหนักการแบ่งชุดข้อมูลโดยใช้ Gain Ratio เรียงลำดับจากมากที่สุดไปน้อยที่สุด ดังนี้ ประเภทยานพาหนะที่เกิดเหตุ ลักษณะการชน ประเภททางหลวงที่เกิดอุบัติเหตุ สาเหตุการเกิดอุบัติเหตุ ภูมิภาคที่เกิดอุบัติเหตุ สภาพอากาศขณะเกิดเหตุ ช่วงเวลาที่เกิดเหตุ บริเวณที่เกิดเหตุ แสดงว่าประเภทยานพาหนะที่เกิดเหตุมีค่าน้ำหนักในการจำแนกข้อมูลที่ก่อให้เกิดผู้เสียชีวิตและไม่เสียชีวิตสูงสุด ซึ่งสอดคล้องกับงานวิจัยของ นันทพร กลิ่นจันทร์

ตารางที่ 4 *ผลการคำนวณค่าความแม่นยำ ค่าเรียกกลับ และค่าความจำเพาะ*

Algorithm	Precision			Recall			Specificity		
	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน
Decision Tree	0.7081	0.7086	0.7055	0.7523	0.7553	0.7577	0.6878	0.6872	0.6816
Naïve Bayes	0.7060	0.7068	0.7072	0.7618	0.7612	0.7618	0.6806	0.6821	0.6825
Random Forest	0.6986	0.6977	0.6965	0.7527	0.7556	0.7561	0.6729	0.6703	0.6682
Deep Learning	0.6986	0.6977	0.6965	0.7527	0.7556	0.7561	0.6729	0.6703	0.6682
Gradient Boosted Tree	0.7472	0.7471	0.7443	0.6765	0.6759	0.6785	0.7695	0.7697	0.7653

ตารางที่ 5 *ผลการคำนวณค่าความถูกต้อง ค่าประสิทธิภาพโดยรวม และค่าเฉลี่ยเรขาคณิต*

Algorithm	Accuracy			F1-measure			G-mean		
	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน	10 ส่วน	20 ส่วน	30 ส่วน
Decision Tree	0.7201	0.7214	0.7198	0.7295	0.7312	0.7307	0.7193	0.7205	0.7186
Naïve Bayes	0.7213	0.7217	0.7223	0.7328	0.7330	0.1335	0.7200	0.7205	0.7210
Random Forest	0.7130	0.7131	0.7123	0.7245	0.7264	0.7259	0.7117	0.7117	0.7108
Deep Learning	0.7209	0.7165	0.7195	0.7246	0.7255	0.7250	0.7190	0.7130	0.7168
Gradient Boosted Tree	0.7228	0.7226	0.7217	0.7101	0.7097	0.7099	0.7215	0.7213	0.7206

นันทนา และคณะ [13] ที่ได้ศึกษาการหาสาเหตุการบาดเจ็บจากรถตู้โดยสารประจำทางชนต้นไม้ จังหวัดพัทลุง พบว่าอุบัติเหตุรถโดยสารประจำทางส่วนใหญ่ทำให้มีการบาดเจ็บและเสียชีวิตจำนวนมาก จำนวนผู้บาดเจ็บและเสียชีวิตขึ้นอยู่กับชนิดของยานพาหนะ ลักษณะการชน การเกิดอุบัติเหตุ และพฤติกรรมความปลอดภัยของผู้ขับขี่และผู้โดยสาร

5.1.2 ประสิทธิภาพการทำนายผลความไม่สมดุลของข้อมูลการเสียชีวิตจากอุบัติเหตุบนโครงข่ายถนนของแบบจำลองการพยากรณ์ เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเสียชีวิตบนโครงข่ายถนนของแบบจำลองการพยากรณ์ ประกอบด้วยเทคนิคเกรเดียนท์บูตทรีส์ เทคนิคต้นไม้ตัดสินใจ เทคนิคนาอีฟเบย์ เทคนิคแรนดอมฟอร์เรสต์ และเทคนิคการเรียนรู้เชิงลึก โดยคำนวณค่าความแม่นยำ ค่าเรียกกลับ ค่าความจำเพาะ ค่าความถูกต้อง ค่าประสิทธิภาพโดยรวมและค่าเฉลี่ยเรขาคณิต จากการทดสอบประสิทธิภาพของแบบจำลองพยากรณ์ กรณีที่มีการแบ่งชุดข้อมูลออกเป็น 30 ส่วน พบว่าเทคนิคที่มีประสิทธิภาพการจำแนกข้อมูลดีที่สุด คือ เทคนิคนาอีฟเบย์มีค่าความถูกต้อง 72.23% ค่าประสิทธิภาพโดยรวม 73.35% และค่าเฉลี่ยเรขาคณิต 72.10%

5.2 ข้อเสนอแนะ

เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่ครอบคลุมมากขึ้น อาจเปรียบเทียบประสิทธิภาพโดยใช้อัลกอริทึมประเภทอื่นเพิ่มเติม เช่น วิธีเคเนียเรสเนเบอร์ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน เป็นต้น มีการเปรียบเทียบกับการใช้เครื่องมือสำเร็จรูปในการช่วยวิเคราะห์อื่น เช่น Microsoft Power BI, Tableau, Pentaho Big Data Integration

and Analytics เป็นต้น และพัฒนาระบบจากผลการวิเคราะห์ที่ได้ไปใช้งาน เพื่อที่หน่วยงานต่าง ๆ จะได้นำผลที่เกิดขึ้นไปใช้ให้ก่อเกิดประโยชน์สูงสุด

6. เอกสารอ้างอิง

[1] World Health Organization. *Global Situation of Road Safety Report 2015*. Global Situation of Road Safety Report. World Health Organization Publishers (WHO Press), 2015.

[2] Road Safety Operation Center. *Thailand Road Safety Master Plan 2018-2021.* Thailand Road Safety Master Plan. 2019.

[3] Traffic Injury Prevention Group Bureau of Non-Communicable Diseases. *Road accident situation in 2014.* 2015.

[4] S. Sinlapasorn. Modeling used to predict osteoarthritis scores WOMAC of Post-Knee Replacement Surgery Patients with Characteristic Engineering and Machine Learning Techniques. *Proceedings of the 25th Annual Meeting in Mathematics (AMM 2021)*. Department of Mathematics, Faculty of Science King Mongkut's Institute of Technology Ladkrabang, 2021.

[5] ABB Corporate Research Center. *What is Deep Learning?.* Available Online at www.new.abb.com/news/detail/58004/deep-learning , accessed on 1 November 2021.

[6] J. Disapon, and A. Sirisophon. "Factors affecting the prevention of road traffic accident among population in Sawang arom district, Uthai thani province: MIMIC model analysis". *Journal of Nurses Association of Thailand Northern Region*, Vol. 27, pp. 101 – 112, 2021

[7] R. Kladchuen, and C. Saenrat. "An Efficiency Comparison of Algorithms and Feature Selection Methods to Predict the Learning Achievement of Vocational Students". *Research Journal Rajamangala University of Technology Thanyaburi*, ISSN 1686-8420, Vol 17, No. 1, pp.1-10, 2018.

[8] S. Sinsomboonthong. "An Efficiency Comparison in Prediction of Imbalanced Data Classification with Data Mining Techniques". *Thai Science and Technology Journal (TSTJ),* Vol. 28 No. 3, pp.383 - 393, 2020.

[9] A. Leelakajonjit, and P. Panudulkitti. Severity level from road traffic accidents. Road accident data management. *Education Research and Development Center and Academic Center for Road Safety,* Aksornthai Press (fahmuang Thai Newspaper) Limited Partnership, Thailand, 2015.

[10] Bureau of Highway Safety. Time of incident. *New Year Festival 2020 (During 27 Dec. 19 – 2 Jan. 20).* 2020.

[11] S. Dechapromphan. *Geography of Thailand : Geographic Perspectives.* Geoinformatics Faculty, Burapha University, Thailand, 2017.

[12] W. Nuipian, and P. Meesut. "Comparison of filtering and consolidation techniques of text mining for text classification". *The Journal of Industrial Technology.* Vol. 9, No. 3, pp.118 - 129, September – December 2013.

[13] N. Klinjun, and N. Supphasri, "Asip Useng and Akeon Sawangnipun. A Case Study on Public Transportation and Road Safety: Why so Many Passengers' Injuries from Van Crashes onto Trees?". *The Southern College Network Journal of Nursing and Public Health*, .pp. 26 – 38, Vol.8, No.2, May-August 2021.