การตรวจจับการฉ้อโกงประกันภัยรถยนต์โดยใช้การวิเคราะห์ข้อความและการเรียนรู้ของเครื่อง

MOTOR INSURANCE FRAUD DETECTION USING TEXT ANALYSIS AND MACHINE LEARNING

ภูริต อำนวยชัย

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2564

การตรวจจับการฉ้อโกงประกันภัยรถยนต์โดยใช้การวิเคราะห์ข้อความและการเรียนรู้ของเครื่อง

ภูริต อำนวยชัย

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2564

MOTOR INSURANCE FRAUD DETECTION USING TEXT ANALYSIS AND MACHINE LEARNING

PHURIT AMNUAYCHAI

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2021

สารนิพนธ์

เรื่อง

การตรวจจับการฉ้อโกงประกันภัยรถยนต์โดยใช้การวิเคราะห์ข้อความและการเรียนรู้ของเครื่อง

ของ

ภูริต อำนวยชัย

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.ศุภร คนธภักดี)

..... ประธาน

(อาจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

ชื่อเรื่อง	การตรวจจับการฉ้อโกงประกันภัยรถยนต์โดยใช้การวิเคราะห์ข้อความและการเรียนรู้ของเครื่อง
ผู้วิจัย	ภูริต อำนวยชัย
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2564
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศุภร คนธภักดี

วัตถุประสงค์ของงานวิจัยเพื่อศึกษาวิเคราะห์ข้อมูลจากข้อความร่วมกันกับการใช้คุณลักษณะอื่นๆมาประกอบร่วมกัน นำมาประยุกต์ใช้กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสร้างแบบจำลองเพื่อทำนายการคาดการความน่าจะเป็นว่าเคลมจะเกิดการทุจริต และเปรียบเทียบประสิทธิภาพของแบบจำลองการแยกประเภท(Classification) ร่วมกับการทดลองกับการจัดการความไม่สมดุลกันของข้อมูล โดยใช้ชุดข้อมูลการเคลมสินไหมรถยนต์ของบริษัทเอเชียประกันภัย 1950 จำกัด(มหาชน) ที่เกิดเคลมในช่วง ม.ค. 2563 ถึง ธ.ค. 2563 โดยรวบรวมข้อมูลการทุจริตเคลมในช่วง ม.ค. 2563 ถึง เม.ย. 2564 จำนวนข้อมูลทั้งหมด 58,579 แถว โดยได้ทำการทดลองด้วย 4 วิธีหลักดังนี้ 1. สร้างแบบจำลองทดลองกับข้อมูลที่มีความไม่สมดุล 2. สร้างแบบจำลองทดลองกับข้อมูลที่จัดการกับความไม่สมดุลด้วยวิธี Random Oversampling 3. สร้างแบบจำลองทดลองกับข้อมูลที่จัดการกับความไม่สมดุลด้วยวิธี SMOTE 4. นำแบบจำลองและวิธีการจัดการความไม่สมดุลของข้อมูลที่เลือกมาทำการปรับจูนพารามิเตอร์ ผู้วิจัยได้ทำการทดลองโดยเปรียบเทียบจากค่า Accuracy, Precision, Recall และ F1-Score ในแต่ละวิธีการที่ทำการวิจัย ซึ่งแบบจำลองที่ให้ค่าผลลัพธ์ที่ดีที่สุดคือ Random Forest และวิธีการจัดการกับความไม่สมดุลกันของข้อมูลคือ SMOTE โดยให้ค่า Accuracy=0.99, Precision=0.803, Recall=0.241, F1-Score=0.371 โดยใช้เวลาเทรนแบบจำลองเพียง 12 นาที จากการทดลองแบบจำลอง Random Forest ร่วมกับการทำ SMOTE สามารถให้ผลลัพธ์ที่ดีกว่าและใช้เวลาในการเทรนที่ไม่มาก ในแง่ของการใช้คุณลักษณะข้อความกับคุณลักษณะที่ไม่ใช่ข้อความพบว่าแบบจำลองยังให้ความสำคัญกับคุณลักษณะที่ไม่ใช่ข้อความมากกว่า

คำสำคัญ : ทุจริตเคลมรถยนต์, การวิเคราะห์ข้อความ, การเรียนรู้ของเครื่อง, ความไม่สมดุลกันของข้อมูล, เทคนิคป่าแบบสุ่ม

Title	MOTOR INSURANCE FRAUD DETECTION USING TEXT ANALYSIS AND MACHINE LEARNING
Author	PHURIT AMNUAYCHAI
Degree	MASTER OF SCIENCE
Academic Year	2021
Thesis Advisor	Dr. Subhorn Khonthapagdee

The purpose of this research was to analyze the data from the text attributes and categorical attributes, in order to generate a model using machine learning techniques. The dataset from motor insurance claims were used and were from the Asia Insurance Company 1950 (Public) and originated in the period from January 2020 to December 2020 and fraudulent claims data from January 2020 to April 2021, which a total of 58,579. The machine learning (ML) algorithms such as Naive Bayes classifier, Logistic regression, Random Forest and support vector machine were applied to the dataset. In this study, two methods were compared to handle an imbalanced dataset: random oversampling and SMOTE. These models were evaluated using Accuracy, Precision, Recall and F1-Score. It was found that Random Forest using SMOTE achieved the best results, with the following values of Accuracy=0.99, Precision=0.803, Recall=0.241, and a F1-Score=0.371.

Keyword : Motor Claim Fraud, Text Analytics, Machine Learning, Imbalance Data, Random Forest Technique

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับความอนุเคราะห์ข้อมูลที่ใช้ในการดำเนินการวิจัยจากบริษัทเอเชียประกันภัย1950 จำกัด(มหาชน) และได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

ภูริต อำนวยชัย

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและความเป็นมาของงานวิจัย

ในอุตสาหกรรมประกันภัยรถยนต์มีการแข่งขันสูง เนื่องจากจำนวนรถที่จดทะเบียนและต่อภาษีกับกรมการขนส่งทางบกทั้งรถเก่าและรถใหม่มีอัตราเพิ่มขึ้นในทุกๆปี โดยจำนวนรถสะสมทั่วประเทศ ณ วันที่ 30 กันยายน 2559 – 2563 มีอัตราเพิ่มขึ้นร้อยละ 2.06 หรือประมาณ 1 ล้านคัน/ปี โดยมีรถตามกฎหมายว่าด้วยรถยนต์มีอัตราเพิ่มขึ้นร้อยละ 2.07 และรถตามกฎหมายว่าด้วยการขนส่งทางบกมีอัตราเพิ่มขึ้นร้อยละ 1.74 จากรายงานสถิติการขนส่งประจำปี 2563(รายงานจดทะเบียนสะสม,รายงานสถิติการขนส่ง ปีงบประมาณ 2559 – 2563) (กองแผนงานกรมการขนส่งทางบก, 2559 - 2563) ส่งผลให้ปริมาณรถยนต์ที่ใช้งานอยู่บนท้องถนนมีการทำประกันภัยรถยนต์ภาคบังคับและภาคสมัครใจเพิ่มขึ้น เป็นผลพลอยได้กับอุตสาหกรรมประกันภัยอย่างมาก จากรายงานสถิติธุรกิจประกันวินาศภัยของสำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (*คปภ.*) ((คปภ.), 2563) จะเห็นได้ว่ามีอัตราจำนวนกรมธรรม์ประกันภัยรถยนต์ในปี 2563 เพิ่มขึ้นร้อยละ 1.34 จากปี 2562 โดยแบ่งออกเป็นกรมธรรม์ภาคบังคับ(พรบ.) เพิ่มขึ้นร้อยละ 0.0188 และ กรมธรรม์ภาคสมัครใจเพิ่มขึ้นร้อยละ 5.5741

การที่บริษัทประกันภัยรับความเสี่ยงในการรับประกันภัยไว้เป็นปริมาณมากขึ้นเท่าไหร่ยิ่งส่งผลให้มีโอกาสเกิดการทุจริต/ฉ้อโกงการเคลมประกันมากขึ้นอย่างหลีกเลี่ยงไม่ได้

การฉ้อโกงและทุจริตการเคลมประกันอาจทำได้ในหลายวิธีการ กล่าวคือ อาจมีการทุจริตจากตัวผู้เอาประกันภัยเอง , ผู้เอาประกันภัยร่วมมือกันกับคู่กรณีร่วมกันทำให้เกิดการทุจริต และการทุจริตอีกวิธีหนึ่งคือเจ้าหน้าที่สำรวจภัยร่วมมือกับผู้เอาประกันภัยดำเนินการทุจริตการเคลม

การทุจริตในรูปแบบเหล่านี้อาจเกิดขึ้นได้หลายวิธีการ เช่น การจัดฉากเพื่อให้ได้มาซึ่งการเคลมประกัน , การตั้งใจหรือจงใจปกปิดข้อมูลการเกิดเหตุที่แท้จริงทำให้บริษัทประกันได้ทราบถึงสาเหตุการเกิดเหตุของเคลมนั้นๆที่ไม่เป็นจริง และที่ร้ายแรงกว่านั้น คือ การที่เจ้าหน้าที่สำรวจภัยชี้แนะให้ผู้เอาประกันภัยให้รายละเอียดความเสียหายที่ไม่ตรงกับความเป็นจริงเพื่อให้ได้มาซึ่งการชดใช้สินไหมจากบริษัทประกัน

สาเหตุทั้งหมดที่กล่าวมาข้างต้นเป็นปัจจัยที่ทำให้บริษัทประกันภัยต้องชดใช้สินไหมเกินความเป็นจริง ทำให้บริษัทประกันภัยต้องสูญเสียจำนวนเงินในการเคลมอย่างมหาศาล ส่งผลกระทบให้ อัตราส่วนค่าสินไหมทดแทน(Loss Ratio) และ อัตราส่วนค่าใช้จ่ายในการจัดการค่าสินไหมทดแทน (Loss Adjustment Expense Ratio) เพิ่มขึ้น ซึ่งทำให้มีผลกระทบต่อการคำนวน

เบี้ยประกันของแต่ละผลิตภัณฑ์รถยนต์ที่จะขายในอนาคต และอาจทำให้สูญเสียลูกค้าในกลุ่มลูกค้าที่ไม่มีการทุจริตอีกด้วยยังมีผลกระทบอื่นๆอีก เมื่ออัตราส่วนค่าสินไหมทดแทน(Loss Ratio) เพิ่มสูงมากขึ้นทำให้กระทบต่อ ความเพียงพอของเงินกองทุน (CAR Ratio) ที่ทางสำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (*คปภ.*) กำหนดไว้ว่า ต้องไม่ต่ำกว่าร้อยละ 140 ถ้าต่ำกว่านี้บริษัทประกันภัยอาจตกอยู่ในความเสี่ยงสูงอาจส่งผลในการให้หยุดการรับประกันภัยในอนาคตได้ ((คปภ.), 2020)

ด้วยสาเหตุนี้การตรวจสอบการฉ้อโกง/การทุจริตในการเคลมสินไหมจำเป็นอย่างยิ่ง ในปัจจุบันการตรวจสอบการฉ้อโกง/การทุจริตการเคลมสินไหมมีการตรวจสอบโดยเจ้าหน้าที่สินไหมซึ่งต้องใช้เวลานานและประสบการณ์จากผู้เชี่ยวชาญเป็นอย่างมาก และส่งผลให้การตรวจสอบการเคลมที่ไม่ทุจริตเกิดการล่าช้าไปด้วย ส่งผลให้ลูกค้าเกิดการไม่พอใจและอาจจะร้องเรียนกับทางสำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (*คปภ.*) เรื่องการชดใช้สินไหมล่าช้าได้ เนื่องจากเป็นการตรวจหาความทุจริตของเคลมโดยเจ้าหน้าที่พิจารณาสินไหมส่งผลให้อาจเกิดการผิดพลาดได้

จากทั้งหมดที่กล่าวมานี้ นำไปสู่การแก้ไขปัญหาที่เกิดขึ้นโดยอาศัยเทคโนโลยีและวิธีการทางด้านการวิเคราะห์ข้อมูลเข้ามาช่วยสนับสนุน เทคนิค Machine Learning เป็นวิธีการหนึ่งที่จะช่วยให้การตรวจสอบการทุจริต/ฉ้อโกง ให้เป็นไปได้อย่างรวดเร็วและถูกต้องมากกว่าการทำงานโดยใช้มนุษย์เป็นผู้กระทำ

ในงานวิจัยนี้เนื่องจากเป็นข้อมูลที่ไม่สมดุลกันมาก ข้อมูลการทุจริต/ฉ้อฉลมีปริมาณที่น้อยกว่าข้อมูลที่ไม่ทุจริตเป็นอย่างมาก จึงต้องใช้วิธีการสุ่มข้อมูลเกิน (Oversampling) เป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก หลังจากนั้นจะนำข้อความจากรายงานสำรวจภัยมาวิเคราะห์ มาประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing Techniques) ในการสกัดคุณลักษณะ (Features Extraction) จากข้อความ ร่วมกับการสร้าง แบบจำลองสำหรับการจำแนกประเภท (Classification Model)โดยเลือกใช้อลักอริทึมการเรียนรู้ของ เครื่อง (Machine Learning Algorithms)

1.2 วัตถุประสงค์

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งจุดมุ่งหมายไว้ดังนี้

1. เพื่อศึกษากระบวนการวิเคราะห์ข้อมูลจากข้อความเพื่อให้เห็นความสำคัญของคำและประโยคที่จะนำไปสู่การทุจริตเคลมและใช้ร่วมกับคุณลักษณะอื่นๆเข้ามาประกอบ นำมาประยุกต์ใช้กับเทคนิคการเรียนรู้ของเครื่อง (Machine learning) นำมาสร้างแบบจำลองเพื่อทำนายการคาดการความน่าจะเป็นว่าเคลมจะเกิดการทุจริตได้
2. เพื่อเปรียบเทียบการทำงานของแบบจำลองตัวแยกประเภท (Classification) และประเมินประสิทธิภาพของแบบจำลองแต่ละชนิดเพื่อให้ได้ความแม่นยำของแบบจำลองที่ใช้ในการทำนายการทุจริต
3. เพื่อทดสอบแบบจำลองการแยกประเภท(Classification) กับความไม่สมดุลกันของข้อมูลเปรียบเทียบประสิทธิภาพของการทำนายในแต่ละแบบจำลอง

1.3 ความสำคัญของการวิจัย

งานวิจัยนี้จะนำรายงานสำรวจภัยของเจ้าหน้าที่สำรวจภัยมาศึกษาการสกัดคำและความสำคัญของคำรวมกับการพิจารณาคุณลักษณะอื่นมาประกอบ เพื่อนำมาทำนายและคาดการณ์ว่าข้อมูลนี้จะเข้าข่ายการทุจริตเคลมหรือไม่โดยการใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน(Supervise Machine learning) เข้ามาช่วยในการทำนายผลซึ่งจะเปรียบเทียบการทำนายในหลายโมเดล โดยจะทำการทดลองกับชุดข้อมูลของบริษัทประกันภัยที่ไม่มีการทุจริตในช่วง ม.ค. 2563 - ธ.ค. 2563 และข้อมูลที่มีการทุจริตในช่วง ม.ค. 2563 – เม.ย. 2564 ข้อมูลก่อนการทำความสะอาดข้อมูล(Preprocessing) ประกอบไปด้วย 18 แอททริบิวต์ และมีจำนวนข้อมูลทั้งหมด 58,579 รายการโดยแบ่งเป็น ข้อมูลที่ไม่ทุจริต 57,890 รายการและข้อมูลที่ทุจริต 689 รายการ หลังจากที่ได้ทำความสะอาดข้อมูลแล้วข้อมูลที่จะนำไปใช้ในการทำวิจัยครั้งนี้คือ จำนวนข้อมูลทั้งหมด 56,459 รายการโดยแบ่งเป็น ข้อมูลที่ไม่ทุจริต 55,817 รายการและข้อมูลที่ทุจริต 678 รายการ

1.4 ขอบเขตของการวิจัย

1.4.1 ประชากรที่ใช้ในการวิจัย

ข้อมูลการเคลมสินไหมรถยนต์ของบริษัทประกันภัยในประเภทไทยแห่งหนึ่งที่เกิดเคลมในช่วง ม.ค. 2563 - ธ.ค. 2563 โดยรวบรวมข้อมูลการทุจริตเคลมในช่วง ม.ค. 2563 - เม.ย. 2564 จำนวนแอททริบิวต์ทั้งหมด 18 แอททริบิวต์ จำนวนข้อมูลทั้งหมด 58,579 แถว

1.4.2 กลุ่มตัวอย่างที่ใช้ในการวิจัย

ข้อมูลการเคลมสินไหมรถยนต์ของบริษัทประกันภัยในประเภทไทยแห่งหนึ่งที่เกิดเคลมในช่วง ม.ค. 2563 - ธ.ค. 2563 โดยรวบรวมข้อมูลการทุจริตเคลมในช่วง ม.ค. 2563 - เม.ย. 2564

1.4.3 คุณลักษณะที่ใช้ศึกษา

คุณลักษณะ มีรายละเอียดดังนี้

- วันที่รับแจ้งอุบิตเหตุ(inform_date)
- เวลาแจ้งอุบัติเหตุ(inform_time)
- วันที่เกิดเหตุ(date_occur)
- เวลาเกิดเหตุ(time_occur)
- วันเริ่มคุ้มครองกรมธรรม์(ins_startdate)
- เพศของผู้ขับขี่(driver_sex)
- รหัสประเภทการเคลม(datacase)
- คำอธิบายประเภทการเคลม(datacase_desc)
- รหัสสาเหตุการเกิดเหตุ(datacasedt)
- คำอธิบายสาเหตุการเกิดเหตุ(datacasedt_desc)
- รายงานผลการสำรวจภัย(comment)
- วันที่ทำกรมธรรม์(tdate)
- รหัสประเภทการใช้รถยนต์(body_type)
- คำอธิบายประเภทการใช้รถยนต์(body_desc)
- รหัสลักษณะการใช้รถยนต์(veh_use)
- คำอธิบายลักษณะการใช้รถยนต์(veh_use_desc)
- อายุรถยนต์(ageofvehicle)
- Label ทุจริต/ไม่ทุจริต(fraudity)

1.5 กรอบแนวคิดในการวิจัย

การพัฒนาแบบจำลองเพื่อทำนายโอกาสการเกิดการทุจริตของการเคลมประกันภัยรถยนต์ของลูกค้าจะแบ่งออกเป็น 4 ขั้นตอน

ขั้นตอนแรกหลังจากการนำข้อมูลเข้าจะทำการเตรียมข้อมูล ทำความสะอาดข้อมูลในส่วนของคุณลักษณะอื่นๆที่ไม่ใช้ข้อความ(Data Preprocessing) โดยหลังจากทำความสะอาด

ข้อมูลจำนวนข้อมูลที่เหลือ 56,459 รายการ โดยข้อมูลที่ไม่ทุจริต 55,817 รายการและข้อมูลที่ทุจริต 678 รายการ หลังจากนั้นจะทำการวิเคราะห์และสำรวจข้อมูลในมุมมองต่างๆ(Exploratory Data Analysis : EDA)

ขั้นตอนที่สองจะมุ่งเน้นไปที่การเตรียมข้อมูลข้อความ การตัดคำ(Tokenization) ทำความสะอาดข้อมูลที่เป็นข้อความ(Text Preprocessing) เปลี่ยนคำให้กลายเป็นเวกเตอร์(Vectorization) เปรียบเทียบความเหมือนของคำสองคำด้วย TF-IDF

ขั้นตอนที่สามจะทำการแก้ไขจัดการปัญหาความไม่สมดุลกันของข้อมูลระหว่างข้อมูลที่ไม่ทุจริตกับข้อมูลที่ทุจริต(Imbalance Data)

ขั้นตอนสุดท้ายจะมุ่งเน้นไปที่การพัฒนาแบบจำลองการทำนายโอกาสการเกิดการทุจริตเคลม ซึ่งเป็นปัญหาแบบ Binary Classification ประกอบไปด้วย 2 ประเภทคือ N และ Y โดยที่ N เป็นรายการที่ไม่ทุจริตเคลม และ Y เป็นรายการที่ทุจริตเคลม โดยใช้เทคนิคการเรียนรู้ของเครื่อง 4 อัลกอริทึมที่แตกต่างกันประกอบไปด้วย อัลกอริทึมการจำแนกประเภทการสุ่มป่าไม้(Random Forest: RF) , อัลกอริทึมการจำแนกประเภทถดถอยโลจิสติก(Logistic Regression: LR) , อัลกอริทึมการจำแนกประเภทนาอีฟเบย์(Naïve Bayes Classifier: NB) และ อัลกอริทึมซัพพอร์ตเวกเตอร์แมทชีน(Support Vector Machine: SVM) โดยจะทำการข้อมูลออกเป็น Training Data 70% และ Testing Data 30% ของข้อมูลทั้งหมด หลังจากนั้นจะนำข้อมูล Training Data มาทำ 10-Folds Cross-Validation เพื่อนำมาทดสอบประสิทธิภาพของโมเดล และในการวัดประสิทธิภาพการทำนายของแบบจำลอง ทดสอบโดยการใช้ Confusion Matrix คำนวณค่า Accuracy , Recall , Precision , F1 Score

1.6 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถนำแบบจำลองที่ได้ไปใช้ช่วยทำนายการทุจริตเคลมจากข้อมูลที่เจ้าหน้าที่สำรวจภัยส่งให้ทางฝ่ายสินไหมตรวจอนุมัติรับผิดชอบเคลม

2. แบบจำลองสามารถช่วยลดระยะเวลาในการพิจารณาตรวจวิเคราะห์หาความผิดปกติจากการเคลมที่ทุจริตได้

3. เพื่อให้สามารถนำผลลัพธ์ของแบบจำลองกับความถูกต้องที่ได้จากผู้ใช้มาปรับปรุงและพัฒนาแบบจำลองให้มีประสิทธิภาพมากยิ่งขึ้น

4. เพื่อทำให้บริษัทลดต้นทุนการจ่ายค่าสินไหมที่ผิดปกติที่เกิดจากการทุจริตเคลม

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยในครั้งนี้ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการพัฒนาแบบจำลองในการทำนายโอกาสเกิดการทุจริตเคลม และได้นำเสนอตามหัวข้อดังต่อไปนี้

1. ทฤษฎีการเตรียมข้อมูล (Data Preprocessing)
2. การประมวลผลข้อความภาษาไทย(PythaiNLP)
3. การจัดการความไม่สมดุลกันของข้อมูล (Imbalance Dataset)
4. ทฤษฎีเกี่ยวกับอัลกอริทึมการจำแนกประเภท (Classification Algorithms)
5. การวัดประสิทธิภาพของแบบจำลอง (Model Evaluation)
6. งานวิจัยที่เกี่ยวข้อง (Related work)

1. **ทฤษฎีการเตรียมข้อมูล** (Data Preprocessing) (Goyal, 2021)

1.1 การเตรียมข้อมูลคุณลักษณะที่ไม่ใช่ข้อความ (Data Preprocessing)

1.1.1 ทำความสะอาดข้อมูล (Data Cleaning)

1.1.1.1 จัดการกับข้อมูลที่ขาดหายไป (Handling missing data)

ชุดข้อมูลขนาดใหญ่มักจะมีข้อมูลหรือค่าที่ขาดหายไป บางทีบุคคลอาจจะลืมป้อนข้อมูลเหล่านั้นหรือแม้กระทั้งระบบจัดเก็บข้อมูลไม่ได้มีการบังคับให้ผู้ใช้งานป้อนข้อมูลในฟิวส์นั้น ข้อมูลที่ขาดหายไปควรได้รับการจัดการก่อนที่จะทำงานกับชุดข้อมูลนั้นๆ การจัดการกับข้อมูลที่ขาดหายไปนั้นเราสามารถทำได้โดยการแทนค่าข้อมูลได้จากข้อมูลที่เราต้องการระบุ, ค่าเฉลี่ย,ข้อมูลก่อนหน้า,ข้อมูลข้างหลัง หรือเลือกที่จะทำการลบข้อมูลในรายการนั้นออกไปเลยก็สามารถทำได้

1.1.1.2 การตรวจสอบค่าที่ผิดปกติที่ไม่ต้องการ (Filtering unwanted outliers)

ค่าที่มีความผิดปกติมันก็จะมีข้อมูลที่จำเป็นต่อข้อมูลของเรา แต่ในขณะเดียวกันมันจะมีผลต่อข้อมูลที่เป็นข้อมูลกลุ่มหลักของเราได้ เราจึงต้องตรวจสอบข้อมูลทั้งที่มีค่าที่ผิดปกติและมีค่าที่ไม่ผิดปกติ ถ้าตรวจพบข้อมูลที่มีค่าที่ผิดปกติไปจากกลุ่มหลักของข้อมูลแล้วนั้น เราต้องหาวิธีการจัดการกับข้อมูลที่ผิดปกติเหล่านั้นเพื่อลดผลกระทบต่อการวิเคราะห์และทำนายของแบบจำลองในอนาคต เราสามารถเลือกที่จะลบข้อมูลรายการที่มีความผิดปกตินั้นทิ้งไปได้ เพื่อให้ได้มาซึ่งความถูกต้องและแม่นย่ำในการทำนายของแบบจำลองของเรา

1.1.1.3 ทำข้อมูลให้เป็นมาตรฐานเดียวกัน (Standardizing data)

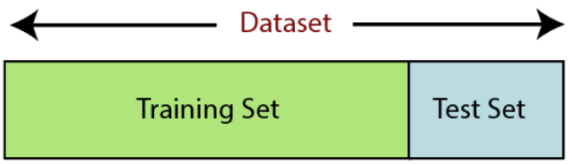
ข้อมูลในตัวแปรคุณสมบัติของเราควรจะเป็นมาตรฐานเดียวกัน ซึ่งจะส่งผลให้การตรวจสอบและการสร้างแบบจำลองของเรา ในกรณีที่ข้อมูลไม่เป็นมาตรฐานเดียวกันเราจำเป็นต้องแปลงข้อมูลให้อยู่ในรูปแบบเดียวกัน เช่น ทำให้เป็นตัวอักษรพิมพ์ใหญ่ ตัวอย่าง TOYOTA กับ toyota แปลงให้เป็นรูปแบบเดียวกันให้เป็น TOYOTA หรือสามารถใช้ function mapping ข้อมูลนำมาใช้ได้ อีกในรูปแบบนึงคือข้อมูลที่เป็นวันที่ก็สามารถเปลี่ยนแปลงให้อยู่ในรูปแบบและฟอร์แมทแบบเดียวกันเพื่อเป็นมาตรฐานก่อนนำไปใช้งาน

1.1.2 การเข้ารหัสข้อมูลหมวดหมู่ (Encoding the categorical data)

การเข้ารหัสข้อมูลเราจะทำกับข้อมูลที่เป็นหมวดหมู่ ข้อมูลที่เป็นหมวดหมู่หมายถึงข้อมูลที่มีหมวดหมู่ภายในชุดข้อมูลที่เรานำมาใช้ โมเดลการเรียนรู้ของเครื่องจะใช้หลักการทางคณิตศาสตร์เป็นหลัก ดังนั้นการจะใช้ข้อมูลที่เป็นหมวดหมู่ข้อความจำเป็นที่จะต้องแปลงให้อยู่ในรูปแบบของตัวเลขเท่านั้น เช่นเราต้องการแปลง Label Yes / No จะถูก Encoding ให้เป็น 0 และ 1

1.1.3 การแยกชุดข้อมูล (Splitting the dataset)

การแยกชุดข้อมูลเป็นขั้นตอนหนึ่งของ Data Pre-processing โดยชุดข้อมูลจะถูกแบ่งออกเป็น 2 ส่วน ส่วนแรกสำหรับนำไปใช้ฝึกโมเดล และ ส่วนที่สองจะถูกแยกเก็บไว้สำหรับทำการทดสอบโมเดล โดยส่วนใหญ่จะแบ่งข้อมูลออกเป็น 80:20 สำหรับฝึกโมเดล:ทดสอบโมเดล



ภาพประกอบ 1แสดงการแบ่งข้อมูลสำหรับการฝึกและทดสอบโมเดล

ที่มา: https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

1.1.4 การเลือกคุณสมบัติ(Feature Selection) (Brownlee, 2014)

การเลือกคุณสมบัติหรือเรียกอีกอย่างว่าการเลือกตัวแปรหรือการเลือกแอตทริบิวต์จะเป็นการเลือกแอตทริบิวต์ที่เกี่ยวข้องและมีผลกับการสร้างแบบจำลองที่กำลังดำเนินการมากที่สุด การเลือกคุณสมบัติจะเลือกจากแอตทริบิวต์ที่มีอยู่ในชุดข้อมูลของเราทั้งหมด โดยจะทำหน้าที่เป็นตัวกรองคุณสมบัติที่ไม่มีประโยชน์จากคุณสมบัติที่มีทั้งหมดออกมาได้

การเลือกคุณสมบัติที่ดีเพื่อนำไปใช้จะส่งผลให้แบบจำลองที่ใช้ในการทำนายมีความแม่นยำมากขึ้นอีกด้วย

อัลกอริทึมของการเลือกคุณสมบัติ(Feature Selection Algorithms) มีทั้งหมดสามประเภททั่วไปดังนี้

- Filter Methods

เป็นการใช้ค่าทางสถิติเพื่อกำหนดคะแนนให้กับแต่ละคุณลักษณะต่างๆ จะจัดอันดับคะแนน และเลือกส่าจะเก็บหรือจะนำออกจากชุดข้อมูล วิธีนี้มักจะไม่แปรผันและพิจารณาคุณลักษณะอย่างอิสระหรือเกี่ยวกับตัวแปรตาม

- Wrapper Methods

จะพิจารณาเลือกชุดคุณลักษณะที่เป็นปัญหาในการค้นหา โดยมีการเตรียม ประเมินและเปรียบเทียบชุดค่าผสมอื่นๆ แบบจำลองคาดการณ์ใช้เพื่อประเมินคุณลักษณะต่างๆร่วมกันและกำหนดคะแนนตามความถูกต้องของแบบจำลอง

- Embedded Methods

จะเรียนรู้ว่าคุณลักษณะใดมีส่วนสนับสนุนความแม่นยำของแบบจำลองได้ดีที่สุดในขณะที่กำลังสร้างแบบจำลอง วิธีการเลือกคุณลักษณะแบบ Embedded ที่ใช้บ่อยที่สุดคือวิธีการ regularization methods

วิธี Regularization methods เรียกอีกอย่างว่า penalization methods จำกัดเพิ่มเติมในการเพิ่มประสิทธิภาพของแบบจำลองในการทำนาย จะทำให้แบบจำลองมีอคติต่อความซับซ้อนที่ต่ำกว่า

1.2 การเตรียมข้อมูลคุณลักษณะที่เป็นข้อความ (Text Preprocessing) (Ganesan, 2019)

ขั้นตอนการประมาลผลข้อความก่อนนำไปใช้เป็นส่วนสำคัญของกระบวนการทำ NLP เนื่องจาก อักขระ คำ ประโยคที่ระบุในขั้นตอนนี้เป็นหน่วยพื้นฐานที่ต้องส่งต่อไปยังขั้นตอนการประมาลผลในขั้นต่อไปทั้งหมดในงานวิจัยนี้ขอพูดถึงวิธีการเตรียมข้อมูลข้อความที่เกี่ยวข้องในงานวิจัยโดยมีขั้นตอนที่เกี่ยวข้องดังนี้

1.2.1 การทำให้เป็นมาตรฐาน (Normalization)

ลบช่องว่างที่ซ้ำกัน สระซ้ำ และตัวอักษรห้อย นอกจากนี้ยังจัดลำดับสระและโทนเสียงใหม่ในระหว่างขั้นตอนการลบสระที่ซ้ำกัน

1.2.2 การตัดคำ (Tokenization)

เป็นกระบวนการที่แบ่งข้อความออกมาเป็นคำ วลี สัญลักษณ์หรือองค์ประกอบอื่นๆ ที่เรียกว่าโทเค็น จุดมุ่งหมายของการสร้างโทเค็นคือการสำรวจคำในประโยค รายการของโทเค็นจะนำไปเป็นอินพุตสำหรับการประมวลผลต่างๆเพิ่มเติม

1.2.3 การลบคำฟุ่มเฟือย (Stopword)

คำในประโยคหรือข้อความที่เรานำมาใช้มักมีคำซ้ำบ่อยมากและเป็นคำที่ไม่มีความหมายอีกด้วย โดยพื้นฐานแล้วคำเหล่านี้ใช้เพื่อเชื่อมคำเข้าด้วยกันในประโยค เป็นที่เข้าใจว่าคำเหล่านี้ไม่ส่งผลต่อบริบทหรือเนื้อหาของข้อความ ดังนั้นเราจึงสามารถลบคำที่เป็นคำฟุ่มเฟือยเหล่านี้ออกไปได้ กระบวนการนี้ยังช่วยลดข้อมูลข้อความและช่วยปรับปรุงประสิทธิภาพของระบบได้อีกด้วย

1.2.4 การลดคำ (Stemming)

เป็นกระบวนการในการลดคำลงในรูปแบบต้นกำเนิด ฐาน หรือราก

Part of Speech Tagging (POS) การแท็กหรือแปะป้ายว่า คำไหนเป็นส่วนไหนของประโยค เช่น คำนาม คำสรรพนาม คำกริยา คำวิเศษ คำสันธาน เป็นต้น

1.2.5 Name Entity Tagging

เป็นการสกัดและแปะป้ายชื่อให้กับ ชื่อบุคคล ชื่อองค์กร ชื่อสถานที่ ตัวเลข จำนวนเงิน วันที่ ในข้อความ

1.2.6 Bag-of-words (BOW)

เป็นโมเดลที่ใช้ในการจัดประเภทข้อความ Text Classification กลุ่มของคำจะถูกอธิบายด้วยกระเป๋าคำ หรือกลุ่มรวมของคำ โดยจะไม่คำนึงถึงหลักไวยกรณ์ ความถี่ที่พบ และลำดับของคำจะนำมาใช้เป็น Feature ในการใช้ในการทดสอบตัวแยกประเภทข้อความ Classifier

1.2.7 N-gram

เป็นโมเดลที่นิยมใช้กันมากที่สุด N-gram คือลำดับของคำต่อเนื่องกันตามจำนวน N คำที่ระบุจากชุดข้อความ เพื่อนำมาเป็นข้อมูลสถิติ เช่นนำมาวิเคราะห์ความถี่ ความสัมพันธ์ระหว่างคำ โอกาสความน่าจะเป็น และสามารถใช้พยากรคำต่อไปได้

1.2.8 Term Frequency- Inverse Document Frequency (TF- IDF) (Prasertsom, 2020)

1.2.8.1 Term-Frequency (TF)

คำคำไหนถูกพูดถึงบ่อยๆในเอกสารนั้นๆ ที่จะมีความเป็นไปได้ว่าคำนั้นมีความเกี่ยวข้องกับใจความสำคัญของเอกสารนั้นๆมาก

ค่าของ Term Frequency เป็นค่าที่บอกความถี่ของคำแต่ละคำที่ปรากฏในเอกสารหนึ่ง โดยคำนวณจากการนำจำนวนครั้งที่คำนั้นๆปรากฏในเอกสารมาหารด้วยจำนวนคำทั้งหมดในเอกสาร

$$TF(\text{ของคำคำหนึ่ง}) = \frac{\text{จำนวนของคำนั้นๆในเอกสาร}}{\text{จำนวนของคำทั้งหมดในเอกสาร}} \quad (1)$$

1.2.8.2 Inverse Document Frequency (IDF)

เป็นการคำนวณค่าน้ำหนัก (weight) ความสำคัญของแต่ละคำโดยคำที่พบเจอได้บ่อยๆจะมีค่า IDF ที่ต่ำซึ่งบอกได้ว่าคำเหล่านั้นจะไม่สามารถดึงเอาจุดเด่นของเอกสารที่คำเหล่านั้นปรากฏอยู่ออกมมาได้ดี โดยค่า IDF สามารถคำนวณได้ด้วยสมการดังนี้

$$IDF(\text{ของคำคำหนึ่ง}) = \log\left(\frac{\text{จำนวนเอกสารทั้งหมดที่ใช้พิจารณา}}{\text{จำนวนเอกสารที่มีคำคำนั้นปรากฏอยู่}}\right) \quad (2)$$

1.2.8.3 คำนวณ TF-IDF

ดังนั้นเมื่อเราเอาค่า TF และ IDF มารวมกันจะคำนวณ TF-IDF ดังนี้

$$TFIDF = TF \times IDF \quad (3)$$

โดยค่าที่ได้จะเป็นการ Weight กันของค่า TF และ IDF

2. **การประมวลผลข้อความภาษาไทย(PythaiNLP)**

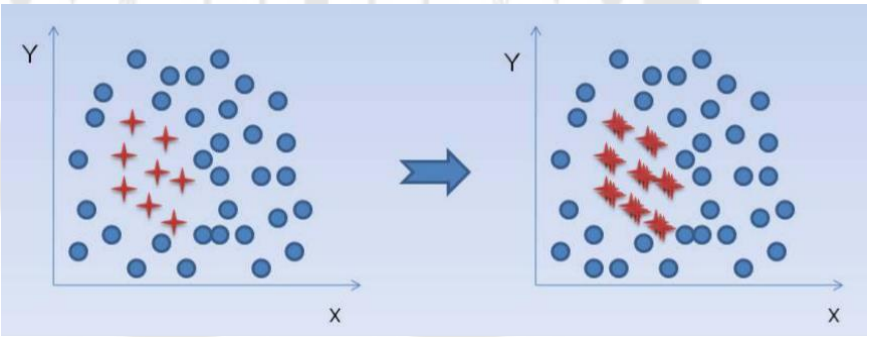
เป็นเครื่องมือที่ใช้สำหรับการประมวลผลข้อความ และการวิเคราะห์ที่ใช้กับภาษาไทย โดยจะมีฟังก์ชั่นการทำงานที่หลากหลาย เช่น ตรวจสอบอักขระ , การเรียงคำ , การตัดคำและแบ่งคำ , การตัดประโยคและแบ่งประโยค , การทำให้เป็นมาตรฐาน , ตรวจการสะกดคำ , การแบ่งชนิดของคำ(Part-of-Speech Tagging) , การจำแนกประเภทของชื่อ(Named Entity Tagging) และการทำ Word Vector

3. **การจัดการความไม่สมดุลกันของข้อมูล** (Imbalance Dataset) (Goswami, 2020)

การที่ข้อมูลของเรามีความไม่สมดุลกันระหว่างคลาส คือข้อมูลของคลาสกลุ่มนึงมีความแตกต่างกับคลาสอีกกลุ่มนึงเป็นอย่างมากส่งผลให้การทำนายของแบบจำลองการแยกประเภทของเราเกิดการลำเอียงไปในทางที่คลาสกลุ่มที่มีสมาชิคอยู่มาก ทำให้เกิดการทำนายที่ไม่มีประสิทธิภาพ ในงานวิจัยของเราจะใช้เทคนิคการสังเคราะห์สุ่มตัวอย่างเพิ่มข้อมูลของคลาสกลุ่มน้อย(Oversampling Technique) เทคนิคสำหรับใช้การแก้ปัญหาการไม่สมดุลกันของข้อมูลที่เกี่ยวข้องในงานวิจัยดังนี้

3.1 Random Oversampling

วิธีการสุ่มตัวอย่างเกินใช้กับข้อมูลของคลาสกลุ่มน้อยเท่านั้น วิธีการนี้จะสุ่มเลือกตัวอย่างจากคลาสกลุ่มน้อยและจะทำซ้ำจนกว่าชุดข้อมูลจะมีความสมดุลกันโดยประมาณ แต่มีข้อเสียคือวิธีนี้อาจจะทำให้เกิดการ Overfitting เนื่องจากข้อมูลที่สร้างขึ้นมาจากการสุ่มจะทำให้ข้อมูลที่ได้จะซ้ำๆกัน



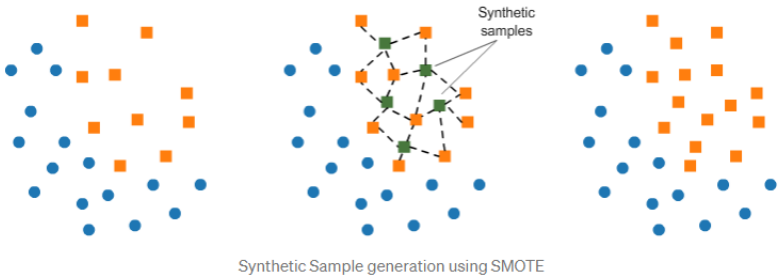
ภาพประกอบ 2 แสดงการสุ่มข้อมูลเพิ่ม(Random Oversampling)

ที่มา: https://core.ac.uk/download/pdf/42753857.pdf

3.2 SMOTE (Das, 2019)

SMOTE เป็นเทคนิคที่ใช้กันอย่างแพร่หลายในการแก้ปัญหาการไม่สมดุลกันของข้อมูล โดยจะใช้เทคนิคการสุ่มตัวอย่างเพิ่ม (Oversampling Technique) จะทำการแก้ไขชุดข้อมูลให้มีความสมดุลกัน โดยจะสังเคราะห์ข้อมูลของคลาสกลุ่มน้อยให้เพิ่มขึ้นมาให้ใกล้เคียงหรือเท่ากับคลาสกลุ่มที่มีข้อมูลมากกว่า ข้อมูลที่ถูกสังเคราะห์ขึ้นมาจะใช้อัลกอริทึมเพื่อบ้านที่ใกล้ที่สุด

(K-Neighbor algorithm) โดย K เพื่อนบ้านที่ใกล้ที่สุดจะถูกเลือกมาเป็นสมาชิกของคลาสที่เป็นส่วนน้อย ดังรูปตัวอย่าง



ภาพประกอบ 3 แสดงการสังเคราะห์ข้อมูลเพิ่มด้วยวิธี SMOTE

ที่มา: https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35

4. **ทฤษฎีเกี่ยวกับอัลกอริทึมการจำแนกประเภท** (Classification Algorithms)

4.1 การถดถอยโลจิสติก (Logistic Regression (LR))

การถดถอยโลจิสติกเป็นวิธีการทางสถิติในการทำนายคลาสไบนารี ผลลัพธ์หรือตัวแปรเป้าหมายมีเพียงสองคลาสที่เป็นไปได้ ตัวอย่างเช่นสามารถใช้สำหรับปัญหาการพิจารณาเคลมว่าทุจริตหรือไม่ โดยจะคำนวณความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น

การถดถอยเชิงเส้นที่ตัวแปรเป้าหมายมีลักษณะเป็นหมวดหมู่ จะใช้บันทึกของอัตราต่อรองเป็นตัวแปรตาม Logistic Regression ทำนายความน่าจะเป็นของการเกิดเหตุการณ์ไบนารีโดยใช้ฟังก์ชัน logit

สมการถดถอยเชิงเส้น:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn \quad (4)$$

โดยที่ y คือตัวแปรตามและ X1, X2 …และ Xn เป็นตัวแปรอธิบาย

ฟังก์ชัน Sigmoid:

$$p = \frac{1}{1+e^{-(y)}} \quad (5)$$

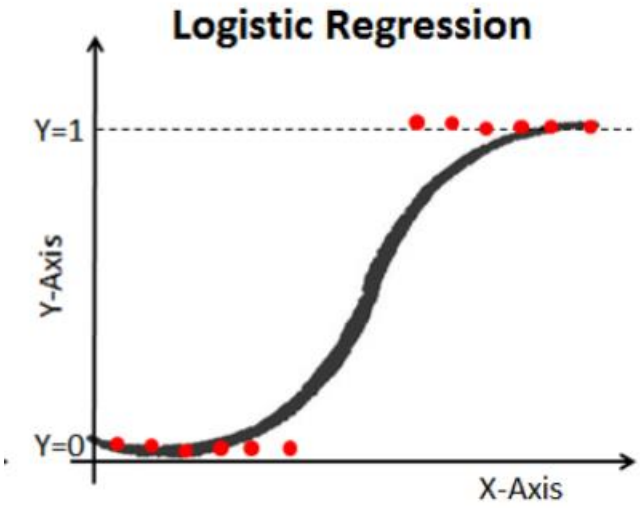
ใช้ฟังก์ชัน Sigmoid กับการถดถอยเชิงเส้น

$$p = \frac{1}{1+e^{-(\beta 0+ \beta 1X1+\beta 2X2+\ldots\ldots+\beta nXn)}} \quad (6)$$

คุณสมบัติของ Logistic Regression:

- ตัวแปรตามในการถดถอยโลจิสติกเป็นไปตาม Bernoulli Distribution
- การประมาณทำได้โดยความเป็นไปได้สูงสุด
- ไม่มี R Square, Model fitness คำนวณผ่าน Concordance, KS-Statistics

การถดถอยโลจิสติกจะให้ผลลัพธ์ที่ต่อเนื่อง ตัวอย่างของผลผลิตต่อเนื่องคือราคาทองและราคาน้ำมัน ส่วนตัวอย่างของผลลัพธ์ที่ไม่ต่อเนื่องคือการทำนายว่าผู้ป่วยเป็นมะเร็งหรือไม่เป็นมะเร็ง การถดถอยโลจิสติกจะประมาณโดยใช้วิธีการประมาณค่าความเป็นไปได้สูงสุด (Maximum likelihood estimation (MLE))



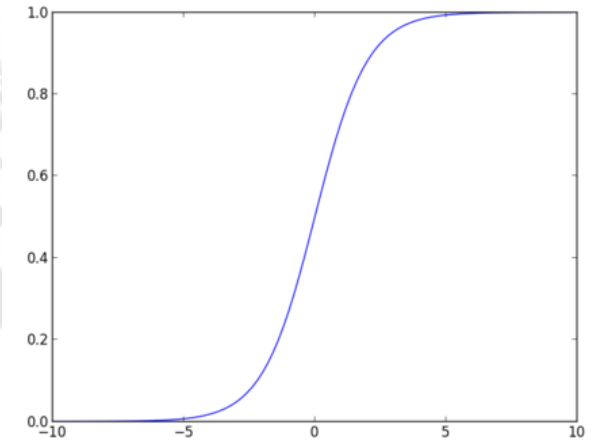
ภาพประกอบ 4 แสดงกราฟ Logistic Regression

ที่มา: https://ichi.pro/th/kar-thakhwam-kheaci-logistic-regression-laea-building-model-ni-python-25643771585192

การประมาณความเป็นไปได้สูงสุด(Maximum likelihood estimation(MLE)) เป็นวิธีการเพิ่มสูงสุดแบบ "ความเป็นไปได้(Possibility)" การเพิ่มฟังก์ชันความน่าจะเป็นสูงสุดจะกำหนดพารามิเตอร์ที่มีแนวโน้มมากที่สุดในการสร้างข้อมูล จากมุมมองทางสถิติ MLE กำหนดค่าเฉลี่ย(Mean) และความแปรปรวน(Variance) เป็นพารามิเตอร์ในการกำหนดค่าพาราเมตริกเฉพาะสำหรับโมเดลที่กำหนด ชุดของพารามิเตอร์นี้สามารถใช้สำหรับการทำนายข้อมูลที่ต้องการในการแจกแจงปกติ(Normal Distribution)

ฟังก์ชัน Sigmoid เรียกอีกอย่างว่าฟังก์ชันโลจิสติกส์ให้เส้นโค้งรูปตัว 'S' ที่สามารถนำจำนวนที่มีมูลค่าจริงใด ๆ มาจับคู่เป็นค่าระหว่าง 0 ถึง 1 ถ้าเส้นโค้งไปที่อินฟินิตี้บวก y ที่คาดการณ์ไว้จะกลายเป็น 1 และถ้า เส้นโค้งไปที่อินฟินิตี้ติดลบ y ที่คาดการณ์ไว้จะกลายเป็น 0 หากเอาต์พุตของฟังก์ชัน sigmoid มากกว่า 0.5 เราสามารถจำแนกผลลัพธ์เป็น 1 หรือ YES และถ้าน้อยกว่า 0.5 เราสามารถจัดประเภทเป็น 0 หรือ NO

$$f(y) = \frac{1}{1+e^{-(y)}} \qquad (7)$$



ภาพประกอบ 5 แสดงกราฟ Function Logistic Regression

ที่มา: https://ichi.pro/th/kar-thakhwam-kheaci-logistic-regression-laea-building-model-ni-python-25643771585192

ประเภทของ Logistic Regression

- การถดถอยโลจิสติกแบบไบนารี เป็นตัวแปรเป้าหมายมีผลลัพธ์ที่เป็นไปได้เพียงสองอย่างเช่น เป็นมะเร็งหรือไม่มีมะเร็ง

- Multinomial Logistic Regression เป็นตัวแปรเป้าหมายมีสามประเภทหรือมากกว่าเล็กน้อยเช่นการทำนายประเภทของภาพยนต์
- การถดถอยลอจิสติกตามลำดับ เป็นตัวแปรเป้าหมายมีสามหมวดหมู่ลำดับขึ้นไปเช่นการให้คะแนนความพึงพอใจตั้งแต่ 0 ถึง 5 ตามลำดับ

4.2 Naïve Bayes (NB)

Naïve Bayes เป็นการจัดหมวดโดยใช้ความน่าจะเป็นเข้ามาช่วยในการคำนวณ โดยมีสมการดังนี้

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (8)$$

สมการจะอธิบายด้วยการแทนแปร 3 ตัวคือ

C คือ Class

X คือ Attribute

P คือ ความน่าจะเป็น

$P(c|x)$ = Posterior Probability คือ ความน่าจะเป็นที่ข้อมูลแอตทริบิวต์เป็น x จะมีคลาส c

$P(x|c)$ = Likelihood คือ ความน่าจะเป็นที่ข้อมูลมีคลาส c และมีแอตทริบิวต์ x

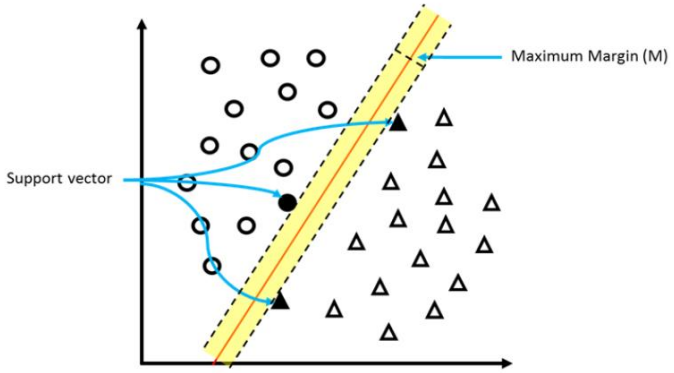
$P(c)$ = Prior Probability คือจำนวนคลาสที่อาจจะเกิดขึ้น / จำนวนคลาสทั้งหมด

$P(x)$ = Predictor Prior Probability คือจำนวนแอตทริบิวต์ทั้งหมด

4.3 Support Vector Machine (SVM)

เป็นตัวจำแนกเชิงเส้น (Linear Classifier) แบบไบนารีคลาสซึ่งมีประสิทธิภาพของการจำแนกมากและเหนือว่าวิธีการจำแนกอื่นๆ ข้อได้เปรียบของ SVM คือมีประสิทธิภาพในการจำแนกข้อมูลที่มีมิติจำนวนมากๆได้ นอกจากนี้การใช้ฟังก์ชันเคอร์เนล (Kernel Function) เพื่อแปลงข้อมูลไปยังมิติที่สูงขึ้นในปริภูมิคุณลักษณะ (Feature Space) สามารถจำแนกข้อมูลที่มี

ความคลุมเครือได้อย่างมีประสิทธิภาพ หลักการของ SVM คือการหาเส้นตรงที่มีมาร์จินที่มากที่สุด (Maximum Margin) ที่สามารถแบ่งข้อมูลออกเป็น 2 คลาส ดังตัวอย่างเป็นข้อมูลขนาด 2 มิติ โดนถูกจำแนกออกเป็น 2 คลาส ได้แก่ + (O) และคลาส – (Δ) โดยเส้นตรงที่ใช้แบ่งข้อมูล มาร์จินเท่ากับ *M=2w* ซึ่ง เป็นความกว้างระหว่างเส้นตรงกับซัพพอร์เวคเตอร์ (Support vector) ของข้อมูลทั้ง 2 คลาส (● และ ▲)



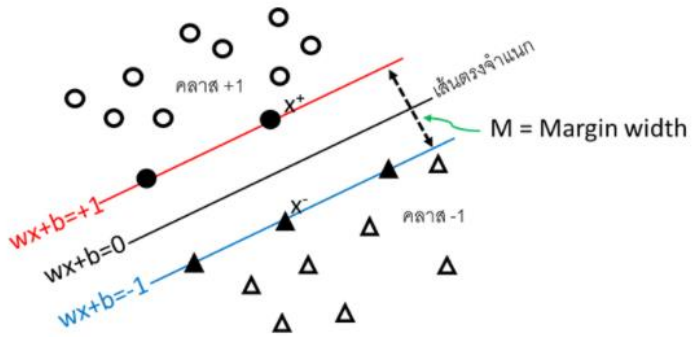
ภาพประกอบ 6 SVM การจำแนกข้อมูล 2 มิติ

ที่มา : https://knowledge.snru.ac.th/ซัพพอร์ตเวกเตอร์แมชชิน/

การจำแนกเชิงเส้นด้วยมาร์จินที่สูงที่สุด(Maximum Margin)

การใช้เส้นตรงสำหรับแบ่งข้อมูลเป็น 2 กลุ่มด้วยมาร์จินที่โตที่สุด (Maximum Margin) เป็นวิธีที่การันตีได้ว่าจะสามารถแยกข้อมูลได้โดยมีความผิดพลาดน้อยที่สุด โดยมี support vector เป็นตัวกำหนดขนาดของ Margin ดังนั้นถ้าข้อมูลมีการเปลี่ยนแปลงใดๆ เส้นตรงจำแนกก็ยังขึ้นอยู่กับ support vector ซึ่งจะยังเป็น Maximum Margin อยู่จากรูปด้านล่างจะเห็นได้ว่าข้อมูล x จะถูกแบ่งเป็นระนาบบวก และระนาบลบ โดยมีสมการคือ w·x+b>=1 สำหรับคลาส+ และ w·x+b<=-1 สำหรับคลาส- ดังนั้นจะสามารจำแนกข้อมูลได้โดย

$$\begin{aligned} +1 \text{ ถ้า } & w \cdot x + b \geq -1 \\ -1 \text{ ถ้า } & w \cdot x + b < +1 \\ \text{ถ้า } & -1 < w \cdot x + b < +1 \end{aligned} \qquad (9)$$



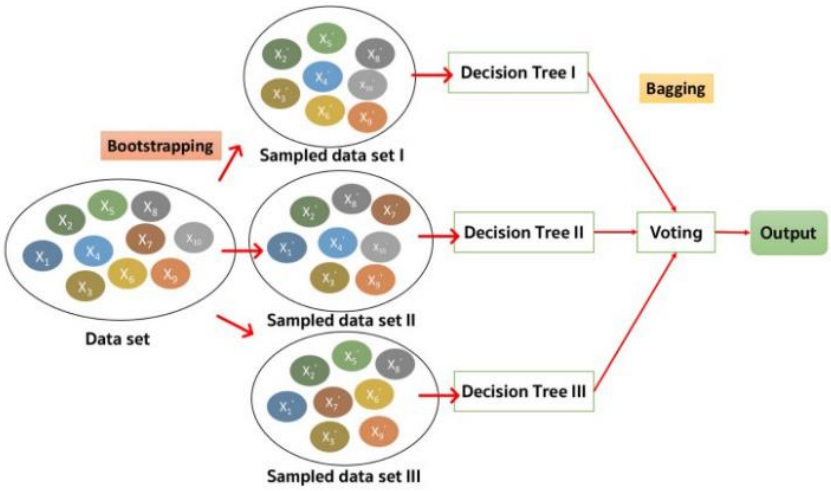
ภาพประกอบ 7 SVM การจำแนกข้อมูล 2 มิติ

ที่มา : https://knowledge.snru.ac.th/ซัพพอร์ตเวกเตอร์แมชชิน/

4.4 Random Forest (RF)

Random Forest จะสร้างแบบจำลองจาก ต้นไม้การตัดสินใจ(Decision Tree) หลายๆ แบบจำลองย่อยๆ (ตั้งแต่ 10 แบบจำลอง ถึง มากกว่า 1000 แบบจำลอง) โดยแต่ละแบบจำลองจะได้รับ ชุดข้อมูล(data set) ที่ไม่เหมือนกัน ซึ่งเป็นชุดข้อมูลย่อยของ ข้อมูลทั้งหมด แล้วให้แบบจำลองทำการ predict แต่ละ ต้นไม้การตัดสินใจ(Decision Tree) ของใครของมัน และคำนวณผลการ predict ด้วยการ vote โดยจะเลือก Decision Tree ที่มีผล vote มากที่สุด กรณีที่เป็นปัญหาของ classification หรือหาค่าเฉลี่ยจากผลลัพธ์ที่ได้ ของแต่ละ ต้นไม้การตัดสินใจ (Decision Tree) กรณีที่เป็นปัญหาของ regression

ต้นไม้การตัดสินใจ(Decision Tree) ที่อยู่ในแบบจำลองของ Random Forest ไม่ได้มีประสิทธิภาพที่เก่งเท่าไหร่นัก แต่เมื่อพอนำเอามา Decision Tree หลายๆ Decision Tree มาทำนายผลร่วมกันจะสามารถสร้างปบบจำลองที่มีประสิทธิภาพ และความแม่นยำมากกว่า Decision Tree แบบต้นเดียว



ภาพประกอบ 8 หลักการทำ Random Forest

ที่มา : https://medium.com/@witchapongdaroontham /@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c

หลักการของ Random Forest คือ

ทำการสุ่มตัวอย่าง(Sample Data) โดยวิธีการ Boostrapping พร้อมการแทนที่ โดยหลักการคือการเพิ่มจำนวนของข้อมูลโดยการสุ่มตัวอย่างไปหลายๆกลุ่ม โดยแต่ละกลุ่มจะมีขนาด 2 ใน 3 ของข้อมูลเดิมและแต่ละกลุ่มจะมีข้อมูลที่ไม่เหมือนกัน

ทำการสร้างแบบจำลองต้นไม้การตัดสินใจ(Decision Tree) สำหรับแต่ละชุดข้อมูล

ทำการรวมผลลัพธ์ที่ได้จากแบบจำลองต้นไม้การตัดสินใจ(Decision Tree) โดยใช้เทคนิคการทำ Bagging(Bootstrap Aggregation) เช่นการ voting ในกรณีของ Classification หรือหาค่าเฉลี่ยในกรณีของ Regression

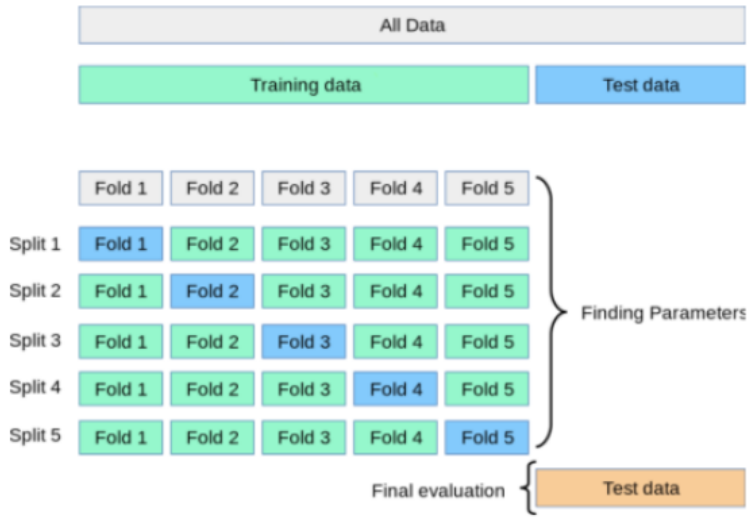
5. **การวัดประสิทธิภาพของแบบจำลอง** (Model Evaluation)

5.1 Cross-Validation

เป็นวิธีการทางสถิติที่ใช้ในการประเมินทักษะของแบบจำลองการเรียนรู้ของเครื่อง การวัดประสิทธิภาพด้วย Cross-Validation นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน โดยแต่ละส่วนจะมีข้อมูลเท่ากัน จากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของแบบจำลอง

5.1.1 K-Fold Cross-Validation (Kumar, 2020)

โดยส่วนมากมักจะแสดงด้วยค่า K เพื่อบ่งบอกจำนวณของการแบ่งข้อมูลออกเป็นส่วนตามจำนวณ K เท่ากัน เช่น K = 5 (5-flod cross-validation) จะแบ่งข้อมูลออกเป็น 5 ส่วนเท่ากันโดยที่ ข้อมูล 1 ส่วนจะถูกใช้เป็นตัวทดสอบประสิทธิภาพของแบบจำลอง และจะทำวนไปจนกว่าจะครบจำนวณ K รอบที่ระบุ



ภาพประกอบ 9 แสดงการทำ Cross-Validation K-Fold

ที่มา: https://vitalflux.com/k-fold-cross-validation-python-example/

5.2 Confusion Metrix (chengz, 2019)

Confusion Matrix คือตารางสำคัญในการวัดความสามารถของการเรียนรู้ของเครื่องในการแก้ปัญหา การจำแนกประเภท(classification)

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

ภาพประกอบ 10 แสดง Confusion Matrix

ที่มา: https://computersciencesource.wordpress.com/2010/01/07/year-2-machine-learning-confusion-matrix/

โดยที่ค่า TP , TN , FP , FN อธิบายได้ดังนี้

True Positive (TP) คือ สิ่งที่โปรแกรมทำนายว่า “จริง” และมีค่าเป็น “จริง”

True Negative (TN) คือสิ่งที่โปรแกรมทำนายว่า “ไม่จริง” และมีค่า “ไม่จริง”

False Positive (FP) คือ สิ่งที่โปรแกรมทำนายว่า “จริง” แต่ มีค่าเป็น “ไม่จริง”

False Negative (FN) คือ สิ่งที่โปรแกรมทำนายว่า “ไม่จริง” แต่ มีค่าเป็น “จริง”

5.2.1 Accuracy (ความถูกต้องที่เราทายได้ตรงกับสิ่งที่เกิดขึ้นจริง) เป็นการวัดความถูกต้องของแบบจำลองโดยพิจารณารวมทุกคลาส

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

5.2.2 Precision (ค่าความแม่นยำ) เป็นการวัดความแม่นยำของข้อมูล โดยพิจารณาแยกทีละคลาส เป็นการเปรียบเทียบ การทำนายที่ถูกต้องว่า” จริง” และก็เกิดขึ้น” จริง” (TP) กับ การทำนายว่า “จริง” แต่สิ่งที่เกิดขึ้น คือ “ไม่จริง” (FP)

$$\frac{TP}{TP+FP} \tag{11}$$

5.2.3 Recall (Sensitivity) เป็นการวัดความถูกต้องของ Model โดยพิจารณาแยกทีละคลาส จะวัดความถูกต้องของการทำนายว่าจะเป็น “จริง” เทียบกับ จำนวนครั้งของเหตุการณ์ทั้งทำนาย และ เกิดขึ้น ว่า “เป็นจริง”

$$\frac{TP}{TP+FN} \tag{12}$$

5.2.4 F1-Score เป็นค่าเฉลี่ยแบบ harmonic mean ระหว่าง precision และ recall จุดประสงค์ของการสร้าง F1 ขึ้นมา คือ เพื่อเป็น single metric ที่วัดความสามารถของโมเดล

$$2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision + Recall}} \quad (13)$$

6. งานวิจัยที่เกี่ยวข้อง (Related work)

การทบทวนวรรณกรรมของงานวิจัยนี้ได้ทำการศึกษาค้นคว้างานวิจัยที่เกี่ยวกับการทำนายเกิดการทุจริตการเคลมประกันภัยรถยนต์ โดยงานวิจัยที่เกี่ยวข้องมีรายละเอียดดังนี้

(1) บทความวิจัยเรื่อง Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique (Harjai, Khatri, และ Singh, 2019)

งานวิจัยนี้ได้สาธิตแนวทางในการสร้างเครื่องตรวจจับการฉ้อโกงประกันภัยรถยนต์ที่ใช้การเรียนรู้ด้วยเครื่อง ซึ่งจะคาดการณ์การเรียกร้องค่าสินไหมทดแทนจากการประกันภัยที่ฉ้อฉลจากชุดข้อมูลบันทึกการเคลมรถยนต์กว่า 15,420 รายการ แบบจำลองที่เสนอสร้างขึ้นโดยใช้เทคนิคการสุ่มตัวอย่างเกินจริงของชนกลุ่มน้อยสังเคราะห์ (SMOTE) ซึ่งขจัดความไม่สมดุลของคลาสของชุดข้อมูล และใช้วิธีการจัดประเภท Random Forest เพื่อจัดประเภท

หลังจากการสุ่มตัวอย่างเกินในคลาสของชนกลุ่มน้อยโดยใช้ SMOTE จำนวนอินสแตนซ์เพิ่มขึ้นเป็น 16,343 โดยที่อินสแตนซ์ของคลาสส่วนน้อยเพิ่มขึ้นเป็น 1,846 ทำให้จำนวนอินสแตนซ์ของคลาสเชิงบวกเท่ากับ 14,497

การวิเคราะห์ข้อมูลเชิงสำรวจ (EDA) ของชุดข้อมูลพบว่า 82% ของคดีที่กลายเป็นการฉ้อโกงเกี่ยวข้องกับรถยนต์ที่มีอายุรถ 6 ถึง 8 ปี นั่นพิสูจน์ให้เห็นว่ารถเก่ามีแนวโน้มที่จะมีส่วนร่วมในการฉ้อโกงมากขึ้น

การเรียกร้องที่เป็นการฉ้อโกง 99.6% ไม่มีพยานในขณะที่การเรียกร้องที่ไม่เป็นการฉ้อโกง 83% ของพวกเขามีพยาน และเมื่อใช้เทคนิคการจำแนกประเภทกับชุดข้อมูลที่สมดุลซึ่งมี 16,343 อินสแตนซ์ พบว่ามีการจัดประเภทอินสแตนซ์อย่างถูกต้อง 15,318 รายการ ซึ่งคิดเป็น 93.72% ของชุดข้อมูลทั้งหมด ความแม่นยำโดยละเอียดของแบบจำลองที่เสนอด้วยค่าเฉลี่ยถ่วงน้ำหนักที่สอดคล้องกัน (Weighted Averages) แนวทางที่เสนอทำให้มีความแม่นยำ(Precision)

และ มูลค่าการเรียกคืน(Recall) 99.9% โดยใช้เวลา 1.43 วินาทีในการสร้างแบบจำลอง ซึ่งพิสูจน์แล้วว่าดีกว่าระบบตรวจจับการฉ้อโกงที่มีอยู่หลายระบบ

(2) บทความวิจัยเรื่อง Thai Defamatory Text Classification on Social Media (Arreerard และ Senivongse, 2018)

งานวิจัยนี้ได้นำข้อความที่โพสต์ข้อมูล แสดงความคิดเห็นจากโซเชียลมีเดียที่อาจทำให้ส่งผลทางลบต่อบุคคลที่กล่าวถึงในโพสต์ และบุคคลนั้นอาจกลายเป็นเป้าหมายของการหมิ่นประมาทได้ ในประเทศไทย แม้ว่าการหมิ่นประมาทใครบางคนบนโซเชียลมีเดียจะผิดกฎหมาย แต่ผู้ใช้โซเชียลมีเดียส่วนใหญ่ไม่ทราบ เพื่อสร้างความตระหนัก งานวิจัยนี้ได้เสนอการจัดหมวดหมู่ข้อความหมิ่นประมาทเป็นภาษาไทย การแบ่งประเภทข้อความใช้วิธีการหลายวิธีในการวิเคราะห์ความคิดเห็นที่เป็นข้อความต่อข่าวการเมืองและบทความบน Facebook รวมถึงคำ n-grams อักขระ n-grams คำศัพท์เฉพาะ แบบจำลองการจัดประเภทข้อความได้รับการพัฒนาโดยใช้การจัดประเภท Naïve Bayes และ Support Vector Machine (SVM) โดยมีวิธีการหลายอย่างรวมกัน วัตถุประสงค์ของการจัดหมวดหมู่ไม่ใช่เพื่อระบุว่าผู้หมิ่นประมาทจะถูกตั้งข้อหาแสดงความคิดเห็นหมิ่นประมาท

การเก็บรวบรวมข้อมูล จะแบ่งออกเป็น 2 ทางคือ 1.) พจนานุกรมคำพิพากษา: ดึงคำศัพท์มาสร้างพจนานุกรมคำพิพากษาจากข้อความที่รวบรวมจากบันทึกของศาลฎีกาเกี่ยวกับคดีตามประมวลกฎหมายอาญามาตรา 326 หมิ่นประมาท และ 393 ดูหมิ่นในที่สาธารณะ 2.) ข้อมูลการทดลอง: จะรวบรวมความคิดเห็นต่อโพสต์บน Facebook ในหน้าการเมืองและข่าวสาธารณะ เป้าหมายที่กล่าวถึงในแต่ละโพสต์เป็นที่รู้จักกันดี เช่น ข้าราชการ ตำรวจ ครู เป็นต้น รวม 1,034 ประโยค โดยมีทนายความ ผู้ช่วยผู้พิพากษา และศาสตราจารย์ด้านกฎหมาย ติดป้าย (Label) ให้ว่าหมิ่นประมาทหรือไม่

การจัดประเภทข้อความหมิ่นประมาท อันดับแรก เครื่องหมายวรรคตอนจะถูกเอกออกแต่จุด (.) ที่ระบุตัวย่อจะถูกเก็บไว้ ประการที่สอง การแปลงโทเค็น(Tokenize) จะถูกประมวลผลโดยโครงข่ายประสาทเทียมที่ได้รับการฝึกฝนโดยใช้คลังข้อมูลที่ดีที่สุดของเนคเทค และ สุดท้ายนี้สำหรับส่วนหนึ่งของการแท็กคำพูด เราใช้ RDR-PoS-Tagger หลังจากประมวลผลล่วงหน้า (Preprocessing) จะแยกคุณลักษณะ(Features extraction) ของแต่ละความคิดเห็นสำหรับแนวทางที่นำมาใช้ทั้ง 5 วิธี 1.) ขั้วของความรู้สึก (Sentiment polarity) บ่งชี้ว่าความคิดเห็นของผู้ทำให้เสื่อมเสียต่อเป้าหมายในแง่บวก เป็นกลาง หรือเชิงลบ 2.) ข้อกำหนดเฉพาะ คือ คุณสมบัติไบนารีที่บ่งชี้ว่ามีเงื่อนไขที่น่าสนใจ ดังกล่าวมี 7 ประเภทคือ (ชื่อหน่วยงาน,กริยา,ดูหมิ่น

,สรรพนามบุรุษที่หนึ่ง,สรรพนามบุรุษที่สองหรือสาม,การรวมกันของสรรพนามบุรุษที่หนึ่งและสองหรือบุคคลที่สาม,การรวมกันของชื่อหน่วยงานและกริยา) 3.)โครงสร้างการพึ่งพา จะพิจารณาความสัมพันธ์แบบพึ่งพา 2 แบบคือความสัมพันธ์ของวัตถุโดยตรงและความสัมพันธ์แบบปกติ ความสัมพันธ์ของวัตถุโดยตรงระบุว่าวัตถุใดถูกกระทำโดยการกระทำในขณะที่ความสัมพันธ์ของวัตถุปกติคือการกระทำที่เป้าหมายทำ 4.) Word n-gram: คุณลักษณะของ Word n-gram เป็นคุณสมบัติไบนารีที่ระบุว่ามีคำที่ประกอบด้วย unigrams, bigrams, trigrams. คุณลักษณะของ Word n-grams จึงมีการรวม unigrams, bigrams และ trigrams 5.) Character n-gram: คุณลักษณะของอักขระ n-gram เป็นคุณลักษณะไบนารีที่ระบุการมีอยู่ของอักขระที่ประกอบด้วย bigrams, trigram และ fourgrams อักขระ n-gram ถูกดึงออกมาในลักษณะเดียวกับ Word n-gram แต่จะพิจารณาอักขระแทน ดังนั้นคุณลักษณะของอักขระ n-gram จึงให้การรวมของ trigrams และ fourgrams

ด้วยข้อมูลที่รวบรวมจากบันทึกของศาลฎีกาและ Facebook และป้ายกำกับโดยผู้เชี่ยวชาญในกฎหมายหมิ่นประมาท ข้อมูลการฝึกอบรมถูกสร้างขึ้นสำหรับห้าแนวทางที่เราใช้ในการฝึกอบรมแบบจำลองการจัดหมวดหมู่(classification models) การใช้ Scikit-learn มาทดลองกับการจัดประเภท Naïve Bayes และ linear SVM แล้วเปรียบเทียบประสิทธิภาพ

การทดลองการจัดประเภทข้อความหมิ่นประมาทในภาษาไทยโดยใช้ Naïve Bayes และ linear SVM ดำเนินการใช้ Ten-fold cross-validation และทดสอบประสิทธิภาพโดยใช้ ความแม่นยำ-Accuracy (Acc), ความแม่นยำ- Precision (P), การเรียกคืน- Recall (R) และคะแนน F-F score (F) คำนวณโดยการหาค่าเฉลี่ยจากประสิทธิภาพ

สำหรับ linear SVM การรวมกันของวิธีการที่มีคำ Word n-gram และอักขระ Character n-grams ให้ความแม่นยำ(accuracy) สูงสุดที่ 0.74 ในขณะที่การรวมกันของแนวทางกับอักขระ Character n-grams ให้ F-score ที่ดีที่สุดที่ 0.64 แม้ว่าฟีเจอร์ของ word n-grams จะให้ความแม่นยำ(precision)สูงสุด แต่การผสมผสานของวิธีการที่ไม่มีคำ Word n-grams จะให้การเรียกคืน(recall) สูงสุด นอกจากนี้โครงสร้างการพึ่งพา(dependency structure), คำเฉพาะ(specific terms), และวิธีการขั้วความรู้สึก(sentiment polarity) ยังทำงานได้ดีด้วยความแม่นยำ(precision) 0.65 และความแม่นยำ(accuracy) 0.66 แต่อัตราการเรียกคืน(recall) ที่ต่ำที่ 0.35 ทำให้คะแนน F ลดลงเหลือ 0.45

พบว่าแม้ว่าข้อความที่หมิ่นประมาทจะมีโครงสร้าง แต่การแยกคุณลักษณะของข้อกำหนดเฉพาะไม่สามารถจดจำชื่อหน่วยงานภายในข้อความได้ เนื่องจากเมื่อมีการกล่าวถึง

เป้าหมาย ผู้ทำลายชื่อเสียงอาจใช้คำที่คล้ายคลึงกันแทนชื่อของเป้าหมาย นอกเหนือจากคำพูด วลีสามารถใช้เพื่อทำให้เสียชื่อเสียงได้เช่นกัน อย่างไรก็ตาม คำศัพท์ในพจนานุกรมเป็นคำศัพท์เดี่ยว ไม่ใช่วลี ดังนั้นโครงสร้างการพึ่งพาจึงไม่รับรู้ถึงความสัมพันธ์ใดๆ ระหว่างวลีนี้กับเป้าหมาย

ในพจนานุกรมคำพิพากษามีคำนาม 120 คำและคำกริยา 127 คำ แม้ว่าจะมีการจัดการคำพ้องความหมาย แต่พจนานุกรมก็ยังขาดคำศัพท์บางคำ ดังนั้น สำหรับโครงสร้างการพึ่งพา การค้นหาแผนผังย่อยจึงไม่พบแผนผังย่อยของ NVO และ PVO นอกจากนี้ การค้นหาแผนผังย่อยยังขึ้นอยู่กับโครงสร้างการพึ่งพาและการแบ่งส่วนคำ

เกี่ยวกับขั้วของความรู้สึก เราคิดว่าข้อความหมิ่นประมาทควรแสดงความคิดเห็นเชิงลบ ตามค่าเฉลี่ยของขั้วของข้อความโดยรวม ข้อความบางข้อความไม่จำเป็นต้องเป็นความคิดเห็นเชิงลบในขณะที่เนื้อหานั้นหมิ่นประมาท

แม้ว่าเราจะพิจารณาโครงสร้างการพึ่งพาและการใช้คำศัพท์เฉพาะที่สมเหตุสมผลสำหรับการจำแนกประเภทการหมิ่นประมาท เนื่องจากรูปแบบต่างๆ ของภาษา คำเหล่านี้ยังคงมีประสิทธิภาพเหนือกว่าด้วยคำ n-gram และอักขระ n-grams นอกจากนี้เรายังทราบด้วยว่ามีคำที่มักพบในความคิดเห็นหมิ่นประมาท เช่น “โกง” (โกง), “โจร” (โจร) และ “คอรัปชั่น” (ทุจริต) ด้วยเหตุนี้ คำว่า n-grams และอักขระ n-grams ซึ่งอาศัยการมีอยู่ของคำดังกล่าวจึงสามารถมีประสิทธิภาพที่ดีได้

(3) บทความวิจัยเรื่อง DETECTING INSURANCE CLAIMS FRAUD USING MACHINE LEARNING TECHNIQUES (Prasasti, Dhini, และ Laoh, 2020)

การกล่าวอ้างที่เป็นการฉ้อโกงรถยนต์นำไปสู่ผลที่ตามมาหลายประการสำหรับบริษัทและผู้ถือกรมธรรม์ ระบบตรวจจับปัจจุบันมีราคาแพงและไม่มีประสิทธิภาพ งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบแบบจำลองการคาดการณ์ในการตรวจจับการฉ้อโกงประกันภัยรถยนต์โดยใช้วิธีการเรียนรู้ของเครื่อง การศึกษานี้ใช้ข้อมูลในโลกแห่งความเป็นจริงของบริษัทประกันภัยรถยนต์แห่งหนึ่งในอินโดนีเซีย ชุดข้อมูลมีการกระจายที่ไม่สมดุลระหว่างข้อมูลของผู้ถือกรมธรรม์ที่กระทำการฉ้อโกงและข้อมูลที่ถูกต้องตามกฎหมาย งานวิจัยนี้จัดการปัญหาชุดข้อมูลที่ไม่สมดุลโดยใช้เทคนิค Synthetic Minority Oversampling Technique (SMOTE) และวิธีการสุ่มตัวอย่างต่ำ ตัวแยกประเภทภายใต้การดูแลที่เสนอ ได้แก่ Multilayer Perceptron (MLP), Decision Tree C4.5 และ Random Forest(RF) ประสิทธิภาพของแบบจำลองจะได้รับการประเมินผ่าน Confusion Matrix , ROC Curve และพารามิเตอร์ต่างๆ เช่น Recall(Sensitivity)

ชุดข้อมูลที่เก็บมาเป็นการรวบรวมการเรียกร้องค่าสินไหมทดแทนจากบริษัทประกันรัฐวิสาหกิจแห่งหนึ่งในอินโดนีเซียตั้งแต่ปี 2559 ถึง 2561 ประกอบด้วย 1881 อินสแตนซ์ที่มีการฉ้อโกง 32 รายการและถูกกฎหมาย 1849 รายการ นี่หมายความว่ามีเพียง 1.7% ของคดีฉ้อโกงในชุดข้อมูล

คุณลักษณะที่ใช้ ได้แก่ รายงานของตำรวจ,ค่าเสียหายส่วนแรก,เดือนที่เกิดเหตุ,ค่าสินไหมทดแทน,อายุรถยนต์,ราคาของรถยนต์,ประเภทรถยนต์,รุ่นของรถยนต์,เพศของผู้ถือกรมธรรม์ และ Label ของชุดข้อมูล

งานวิจัยมีวัตถุประสงค์เพื่อตรวจหาการฉ้อโกงประกันภัยรถยนต์โดยใช้วิธีการลักษณนามภายใต้การดูแลที่เสนอ และเพื่อเปรียบเทียบประสิทธิภาพของตัวแยกประเภท มุ่งเน้นไปที่การจัดการชุดข้อมูลที่ไม่สมดุลและสร้างการตรวจจับการฉ้อโกงด้วยความแม่นยำสูง ใช้เทคนิค Synthetic Minority Oversampling Technique (SMOTE) กับ Multilayer Perceptron (MLP), Support Vector Machine (SVM) และ K-nearest Neighbors (KNN) เพื่อตรวจจับการเคลมประกันรถยนต์ที่เป็นการฉ้อโกง แสดงให้เห็นว่าการใช้ SMOTE ให้ความไว(Sensitivity) ความจำเพาะ(Specificity) และความแม่นยำ(Accuracy) ที่ดีกว่าชุดข้อมูลที่ไม่สมดุล อย่างไรก็ตาม ข้อมูลที่สร้างจาก SMOTE อาจไม่เหมือนกับชุดข้อมูลดั้งเดิมในบางกรณี การสุ่มตัวอย่างน้อยเกินไปสามารถช่วยให้ได้ผลลัพธ์ที่ดีขึ้นเมื่อเปอร์เซ็นต์ของชุดข้อมูลไม่สมดุลอย่างมากเนื่องจากมีกรณีการฉ้อโกงจำนวนน้อย งานวิจัยนี้จึงเสนอการสุ่มตัวอย่างแบบผสมโดยใช้ SMOTE และการสุ่มตัวอย่างต่ำ(Undersampling)เพื่อจัดการกับชุดข้อมูลที่ไม่สมดุล งานวิจัยนี้ยังเสนอการตรวจจับการฉ้อโกงโดยใช้ MLP, Decision Tree และ Random Forest ด้วยความแม่นยำสูง(High Accuracy) โดย MLP มีความแม่นยำสูงสุดเมื่อเทียบกับ KNN และ SVM ในอีกด้านหนึ่ง MLP มีความไว(Sensitivity) และค่าความจำเพาะ(Specificity) ต่ำสุด งานวิจัยนี้พิสูจน์ให้เห็นว่า MLP สามารถบรรลุความแม่นยำสูงด้วยความไวสูงและความจำเพาะสูง Decision Tree ใช้สำหรับความสามารถในการเพิ่มความแม่นยำของแบบจำลองโดยใช้กระบวนการ pruning process. Random Forest พิสูจน์แล้วว่าประสบความสำเร็จในการแก้ปัญหาการจำแนกประเภทจำนวนมากด้วยความแม่นยำสูงในการศึกษานี้

ผลลัพธ์แสดงให้เห็นว่า Multilayer Perceptron, Decision Tree และ Random Forest สร้างความแม่นยำสูง อย่างไรก็ตาม Random Forest มีประสิทธิภาพสูงสุดด้วยความแม่นยำ (Accuracy) 98.5% ความไว(Sensitivity) 100% และความจำเพาะ(Specificity) 98.5% ของแบบจำลอง

(4) บทความวิจัยเรื่อง Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy (Aninditya, Hasibuan, และ Sutoyo, 2019)

การระบุระดับของคำถามตามระดับความรู้ความเข้าใจของอนุกรมวิธานของ Bloom ทำได้ด้วยตนเอง กระบวนการจัดประเภทคำถามด้วยตนเองต้องใช้เวลามากสำหรับข้อมูลขนาดใหญ่ นอกจากนี้ ความแตกต่างในการรับรู้ในการจำแนกคำถามส่งผลให้กระบวนการจัดหมวดหมู่ด้วยตนเองมีความหลากหลาย เพื่อแก้ปัญหาเหล่านี้ กระบวนการจำแนกอัตโนมัติสามารถทำได้โดยใช้การประมวลผลภาษาธรรมชาติ

การศึกษานี้มีวัตถุประสงค์เพื่อกำหนดประสิทธิภาพในการจำแนกคำถามตามระดับความรู้ความเข้าใจของอนุกรมวิธานของ Bloom โดยใช้ตัวจำแนก Naive Bayes นอกจากนี้เพื่อเป็นแนวทางแก้ปัญหาของอาจารย์ในการตั้งคำถามซึ่งจะเป็นเกณฑ์มาตรฐานในการประเมินความเข้าใจของนักเรียนเกี่ยวกับเนื้อหาที่ให้ตามวัตถุประสงค์การเรียนรู้

การศึกษานี้ใช้ชุดข้อมูลจริงที่รวบรวมจากคำถามกลางภาคและข้อสอบปลายภาคที่นำมาจากภาควิชาระบบสารสนเทศ มหาวิทยาลัยเทลคอม ตั้งแต่ปีการศึกษา 2555/2556 ถึง 2561/2562 โดยเฉพาะอย่างยิ่ง เราได้ตรวจสอบ Words, Characters และ N-gram โดยเข้าสู่กระบวนการเตรียมการล่วงหน้า เช่น เทคนิคในการทำ tokenizing, stemming, filtering และการแยกคุณลักษณะต่างๆ ของกระบวนการแล้วตั้งคำถามข้อมูลที่มีอยู่จะถูกแปลงเป็นข้อมูลตัวเลขที่เรียกว่าคุณลักษณะเวกเตอร์ หลังจากนั้นจะถูกป้อนลงในแบบจำลองอาคาร Naive Bayes ผลการศึกษาจะเป็นแบบอย่างในการจำแนกคำถามตามระดับความรู้ความเข้าใจของอนุกรมวิธานของบลู

หลังจากรวบรวมข้อมูลคำถามแล้ว ให้กำหนดป้ายกำกับ (Label) ของแต่ละคำถาม การติดฉลาก (Label) ตามระดับความรู้ความเข้าใจของอนุกรมวิธานของ Bloom ซึ่งแบ่งออกเป็นสองส่วนคือ ลำดับที่ต่ำกว่า (LO) และลำดับสูง (HO) หลังจากเสร็จสิ้นการติดฉลากข้อมูล การเปรียบเทียบหมายเลขป้ายกำกับระดับสูงและระดับต่ำ ระดับสูงที่ 143 คำถามและระดับต่ำสุดที่ 157 คำถาม. จะแบ่งข้อมูล Trans 80% และ ข้อมูล Test 20% และใช้การกระจายข้อมูลโดยทำ 10-Fold Cross-Validation Score. ขั้นตอนการฝึกใช้ Data Training ซึ่งจะทำการวิเคราะห์เอกสารตัวอย่างในรูปแบบของการเลือกคำศัพท์ ผลจาก 10-Fold Cross Validation ที่ใช้ในฟีเจอร์ Words TF-IDF ซึ่งให้ความแม่นยำโดยเฉลี่ย 78% , ผลของ 10 Fold-Cross Validation โดยใช้คุณลักษณะ N-Gram TF-IDF ซึ่งให้ความแม่นยำเฉลี่ยต่ำสุดของคุณลักษณะระดับ TF-IDF เพียง 74% ,

ผลลัพธ์ของ 10 Fold-Cross Validation โดยใช้คุณลักษณะ TF-IDF ของอักขระ อัตราความแม่นยำเฉลี่ยทำได้เพียง 76%. การทดสอบ Naive Bayes Classifier แสดงความสอดคล้องตามผลลัพธ์ของการจำแนกประเภทโดยใช้คุณสมบัติ N-Gram TF-IDF สามารถวิเคราะห์ได้ว่าการใช้คุณสมบัติ TF-IDF ในการสร้างแบบจำลองสมมติมีอิทธิพลอย่างมากต่อผลลัพธ์ . จากผลการจัดหมวดหมู่แสดงให้เห็นว่าประสิทธิภาพของแบบจำลองการทำนายยังได้รับอิทธิพลจากประเภทของคำถามที่ใช้. ความเหมาะสมของ Naive Bayes Classification เป็นแบบอย่างการทำนายคำถามและเป็นตัวแยกประเภทที่เหมาะสมที่สุดผลที่ได้เมื่อใช้ Words TF-IDF ความแม่นยำ 82% การเรียกคืน คือ 82% ในขณะที่ผลลัพธ์ที่ได้รับสำหรับ N-Gram TF-IDF ความแม่นยำคือ 85% และการเรียกคืนคือ 80% และผลลัพธ์ที่ได้จากอักขระ TF-IDF ความแม่นยำคือ 78% และการเรียกคืนคือ 77%.

(5) บทความวิจัยเรื่อง Gender Classification of Thai Facebook Usernames (Yuenyong และ Sinthupinyo, 2020)

บทความนี้จะนำเสนอแอปพลิเคชันการเรียนรู้ของเครื่องเพื่อจำแนกเพศของผู้ใช้ Facebook ตามชื่อผู้ใช้เพียงอย่างเดียว ข้อมูลโปรไฟล์ผู้ใช้บนโซเชียลเน็ตเวิร์กมีความสำคัญในการศึกษาจำนวนมาก แต่ในบางครั้งจะไม่มีข้อมูลเปิดเผยต่อสาธารณะทางออนไลน์ เช่น อายุหรือเพศ การศึกษาส่วนใหญ่ใช้เฉพาะข้อมูลที่เป็นข้อความจากหน้าเว็บเท่านั้น แต่เราเลือกที่จะศึกษาการจำแนกเพศตามชื่อผู้ใช้แทน ซึ่งเพศนั้นอนุมานจากชื่อจริงของผู้ใช้และชื่อนามแฝง เราเน้นเฉพาะชื่อภาษาไทยซึ่งอาจมีรูปแบบบางอย่างที่บ่งบอกเพศของเจ้าของได้ มีการเสนอรูปแบบต่างๆ ร่วมกันเพื่อจำแนกเพศตามชื่อผู้ใช้ Facebook ของไทย แต่ละรุ่นได้รับการฝึกฝนโดยใช้วิธีการเรียนรู้ภายใต้การดูแล นอกจากนี้ ผลการจัดหมวดหมู่ทั้งหมดยังถูกรวมเข้าเป็นแบบจำลองขั้นสุดท้าย

ชุดข้อมูลจะเน้นไปที่ชื่อผู้ใช้ Facebook ของไทยที่รวบรวมระหว่างเดือนมกราคมถึงมีนาคม 2019 โดยใช้ห้องสมุด Selenium ผู้ใช้ได้รับการคัดเลือกที่มีชื่อผู้ใช้เป็นตัวอักษรไทยเท่านั้นและมีโปรไฟล์เพศที่เปิดอยู่ หลังจากรวบรวมข้อมูลทั้งหมดแล้ว ชื่อผู้ใช้จะถูกติดป้ายกำกับ(Label)ด้วยตนเองเป็นสองประเภท ได้แก่ ชื่อจริงและชื่อแทน จากชื่อผู้ใช้ 4317 ที่รวบรวม 2047 ถูกจัดประเภทเป็นผู้หญิง (47.42%) และ 2270 เป็นชาย (52.58%) ยิ่งกว่านั้น 1961 ของพวกเขาเป็นชื่อแรก(First name) ในขณะที่ 2356 เป็นชื่อแทน(Alias name)

การตัดคำ(Tokenization) ในภาษาไทย คำต่างๆจะไม่คั่นด้วยการเว้นวรรค และช่องว่างจะใช้แทนประโยคใหม่แทน นอกจากนี้ ชื่อนามแฝงส่วนใหญ่สามารถแยกเป็นคำได้ในงานวิจัยจะใช้ Library PythaiNLP โดยใช้วิธี Maximal Matching เพื่อแบ่งกลุ่มชื่อภาษาไทย

ลักษณะการจำแนกเสียงพูดภาษาไทย ในภาษาไทย ชื่อมักมีคำนามแต่ไม่รวมคำบางประเภท ตัวอย่างเช่น ชื่อภาษาไทยไม่มีคำสันธาน เช่น "และ" "หรือ" และ "แต่" คำปฏิเสธ เช่น "ไม่" และ "ห้าม") แบ่งส่วนของคำพูดภาษาไทยออกเป็น 14 ประเภท (คำนาม สรรพนาม กริยา กริยาช่วย กำหนด วิเศษณ์ ลักษณนาม คำสันธาน คำบุพบท การฉีด คำนำหน้า ตอนจบ ปฏิเสธ และเครื่องหมายวรรคตอน) และแบ่งออกเป็น 47 หมวดหมู่ย่อย ในการศึกษานี้ PyThaiNLP library ถูกใช้เพื่อแยกส่วนของคำพูดด้วยการแปลงคำ

คุณสมบัติความถี่ตัวอักษรไทยในภาษาไทยมีอักขระหลักสามประเภทคือ พยัญชนะ; สระ; และโทน เหล่านี้อยู่ในระดับบน กลาง และล่าง ความถี่ของอักขระแต่ละตัวจะนับเฉพาะจากชื่อเท่านั้น ในการวัดความไม่เป็นระเบียบของในอักขระ จะคำนวณเอนโทรปีของเพศของอักขระแต่ละอักขระ โดยให้ผลลัพธ์สูงแสดงว่าอักขระนั้นมีเพศที่หลากหลายมาก และผลลัพธ์ต่ำชี้ไปที่ที่ใช้ในอักขระ เพศเดียว โดยเฉพาะอย่างยิ่ง เอนโทรปีถูกกำหนดเป็นผลรวมของความน่าจะเป็นของแต่ละเพศ เวลาโดยความน่าจะเป็นของบันทึกของเพศเดียวกันนั้น เอนโทรปีอักขระขั้นต่ำ 10 ตัวในชื่อแรก ได้แก่ 'ซ', 'ฝ', 'ฟ', ' ', 'ใ', '้', 'เฒ', 'แ', 'โ' และ 'ค'

คุณลักษณะถูกสร้างขึ้นจากสตริงย่อยในชื่อ มีการใช้คุณลักษณะหกประเภทในการทดลองนี้: อักขระสองตัวแรก; อักขระสามตัวแรก อักขระสี่ตัวแรก อักขระสองตัวสุดท้าย อักขระสามตัวสุดท้าย และสี่ตัวอักษรสุดท้าย

ในการศึกษานี้ ใช้กลุ่มคุณลักษณะต่างๆ สี่กลุ่มดังนี้ 1) คุณสมบัติการสร้างโทเค็นของ Word; 2) คุณสมบัติการจำแนกส่วนของคำพูด; 3) คุณสมบัติความถี่ของตัวละอักขระ; 4) คุณสมบัติอักขระสตริงย่อย; เหล่านี้ถูกใช้เป็นอินพุตเดียวกันสำหรับแบบจำลองสี่แบบแยกกันซึ่งมีจุดประสงค์ต่างกัน ตัวอย่างเช่น ก) การจำแนกเพศจากชื่อเท่านั้น B) การจำแนกเพศจากชื่อนามแฝงเท่านั้น ค) จำแนกชื่อผู้ใช้เป็นชื่อจริงหรือชื่อแทน; และ D) การจำแนกเพศจากชื่อผู้ใช้ โมเดลสุดท้ายจะรวมเอาท์พุตของโมเดลทั้งสี่เพื่อสร้างโมเดลสุดท้าย

ใช้ K-Nearest Neighbor, Support Vector Machine, Random Forest, Multinomial Naïve Bayes และ Neural Network ในรุ่น A, B, C และ D ในขณะที่ Neural Network ใช้เพื่อรวมผลลัพธ์จากตัวแยกประเภททั้งหมด

ผลการทดลองแสดงให้เห็นว่าการใช้ word tokenization สำหรับชื่อผู้ใช้ทั้งหมดมีระดับความแม่นยำอยู่ที่ 65.81% แต่รูปแบบที่รวมกันนั้นมีประสิทธิภาพที่ดีขึ้นโดยมีความแม่นยำระดับ 91.75

(6) บทความวิจัยเรื่อง Leveraging deep learning with LDA-base text analytics to detect automobile insurance fraud (Wang และ Xu, 2018)

การฉ้อโกงประกันภัยรถยนต์แสดงถึงเปอร์เซ็นต์ที่สำคัญของต้นทุนของบริษัทประกันภัยทรัพย์สินและส่งผลกระทบต่อกลยุทธ์การกำหนดราคาของบริษัทและผลประโยชน์ทางเศรษฐกิจของสังคมในระยะยาว

การตรวจจับการฉ้อโกงประกันภัยรถยนต์ได้กลายเป็นส่วนสำคัญอย่างยิ่งในการลดต้นทุนของบริษัทประกันภัยการศึกษาก่อนหน้านี้เกี่ยวกับการตรวจจับการฉ้อโกงประกันภัยรถยนต์ได้ตรวจสอบปัจจัยตัวเลขต่างๆ เช่นเวลาเคลมและยี่ห้อรถที่เอาประกันภัย อย่างไรก็ตามข้อมูลที่เป็นข้อความในการอ้างสิทธิ์นั้นแทบไม่มีใครได้ทำการศึกษาวิเคราะห์การทุจริตการประกันภัยบทความนี้ขอเสนอรูปแบบการเรียนรู้เชิงลึกแบบใหม่สำหรับรถยนต์การตรวจจับการฉ้อโกงประกันภัยที่ใช้การวิเคราะห์ข้อความตาม Latent Dirichlet Allocation (LDA) ในข้อเสนอของเราวิธี LDAถูกใช้ครั้งแรกเพื่อแยกคุณสมบัติข้อความที่ซ่อนอยู่ในคำอธิบายข้อความของอุบัติเหตุที่ปรากฏในคำกล่าวอ้างและโครงข่ายประสาทเทียมได้รับการฝึกอบรมเกี่ยวกับข้อมูลซึ่งรวมถึงข้อความฟีเจอร์และฟีเจอร์ตัวเลขแบบดั้งเดิมสำหรับตรวจจับการอ้างสิทธิ์ที่เป็นการฉ้อโกงบนพื้นฐานของการประกันภัยในโลกแห่งความเป็นจริงชุดข้อมูลการฉ้อโกงผลการทดลองเปิดเผยว่ากรอบการทำงานตามการวิเคราะห์ข้อความที่นำเสนอมีประสิทธิภาพดีกว่าแบบดั้งเดิมนอกจากนี้ผลการทดลองแสดงให้เห็นว่าโครงข่ายประสาทส่วนลึกมีประสิทธิภาพดีกว่าโมเดลแมชชีนเลิร์นนิงที่ใช้กันอย่างแพร่หลายเช่น Random Forest และSupport Vector Machine ดังนั้นกรอบงานที่เสนอซึ่งรวมโครงข่ายประสาทเทียมและ LDA เข้าด้วยกันจึงมีศักยภาพที่เหมาะสมสำหรับการตรวจจับการฉ้อโกงประกันภัยรถยนต์

การประมวลผลข้อมูลล่วงหน้า ข้อมูลการฉ้อโกงประกันภัยรถยนต์แบ่งออกเป็นข้อมูลที่มีโครงสร้างและข้อมูลที่ไม่มีโครงสร้าง ข้อมูลที่มีโครงสร้างประกอบด้วยข้อมูลตัวเลขและข้อมูลตามหมวดหมู่ เช่น จำนวนการอ้างสิทธิ์ที่ผ่านมา ข้อมูลที่ไม่มีโครงสร้างหมายถึงข้อมูลที่เป็นข้อความ เช่น คำอธิบายของอุบัติเหตุ ในกระบวนการประมวลผลข้อมูลล่วงหน้า ข้อมูลทั้งสามประเภทจะถูกทำความสะอาดข้อมูลตามลักษณะเฉพาะเพื่ออำนวยความสะดวกในการประมวลผลข้อมูล

การขุดข้อความ(text mining) ประสบการณ์ของผู้เชี่ยวชาญที่เป็นมนุษย์มีอยู่ในคำอธิบายข้อความของอุบัติเหตุ ซึ่งคอมพิวเตอร์ไม่สามารถเข้าใจได้โดยตรง ดังนั้นจึงแนะนำการทำเหมืองข้อความเพื่อช่วยดึงข้อมูล ค่าสูงที่ฝังอยู่ในคำอธิบายข้อความ การทำเหมืองข้อความเป็นเทคโนโลยีที่ค่อนข้างใหม่ในด้านการตรวจจับการฉ้อโกงประกันภัย ดังนั้นจะมีการแนะนำวิธีการขุดข้อความบางวิธีพร้อมกับวิธีการทางเทคนิคอื่นๆ เพื่อสนับสนุนการวิจัย การขุดข้อความสามารถอธิบายได้ว่าเป็นวิธีการทางเทคนิคที่ดึงข้อมูลที่เป็นประโยชน์จากข้อมูลที่ไม่มีโครงสร้าง (ข้อมูลข้อความ) เนื่องจากความซับซ้อนของข้อมูลข้อความ จึงมีการนำทฤษฎีและเทคโนโลยีต่างๆ มาใช้ในการดำเนินการขุดข้อความในบรรดาวิธีการและทฤษฎีต่างๆ มีการใช้การประมวลผลภาษาธรรมชาติ (NLP) และทฤษฎีความน่าจะเป็น ในบทความนี้ จะใช้การแบ่งส่วนคำภาษาจีนและ LDA เพื่อแก้ปัญหาความไม่ลงรอยกันระหว่างข้อมูลข้อความและอัลกอริธึมการทำเหมืองข้อมูล การแบ่งกลุ่มคำภาษาจีนใช้เพื่อประมวลผลข้อมูลข้อความล่วงหน้า และแบบจำลองหัวข้อ LDA ใช้เพื่อแยกหัวข้อ ซึ่งมีประสบการณ์ของผู้เชี่ยวชาญที่เป็นมนุษย์ในข้อมูลข้อความที่ประมวลผล

โมเดลการเรียนรู้เชิงลึก คุณสมบัติหมวดหมู่ ตัวเลข และเฉพาะทั้งหมดจะถูกส่งไปยังเลเยอร์อินพุตของ Deep Neural Network (DNN) เพื่อเริ่มกระบวนการฝึกอบรม จากนั้น เลเยอร์อินพุตจะจับคู่คุณสมบัติกับเลเยอร์แรกที่ซ่อนอยู่ และกระบวนการจะดำเนินต่อไป แต่ละชั้นที่ซ่อนอยู่ประกอบด้วยโหนดจำนวนหนึ่งสำหรับการประมวลผลข้อมูลที่ป้อนเข้าของชั้นและส่งผลลัพธ์ไปยังชั้นถัดไป activation function ของแต่ละเลเยอร์สามารถเพิ่มการแมปแบบไม่เชิงเส้น(nonlinear) ให้กับกระบวนการแมปเพื่อรับประกันว่าความสามารถในการสร้างนามธรรมของ DNN จะมีประสิทธิภาพมากขึ้น ในเวลาเดียวกัน เพื่อหลีกเลี่ยงไม่ให้เกิดข้อผิดพลาด gradient vanish ในกระบวนการ back propagation บทความนี้จึงใช้ ReLU แทน Sigmoid เป็น activation function หลังจากกระบวนการวนซ้ำของการเพิ่มประสิทธิภาพไฮเปอร์พารามิเตอร์ โมเดล DNN จะแสดงผลลัพธ์การตรวจจับและพิจารณาว่าการอ้างสิทธิ์นั้นเป็นการฉ้อโกงหรือไม่

ข้อมูลที่ใช้ในบทความนี้เป็นข้อมูลในโลกแห่งความเป็นจริงซึ่งได้มาจากบริษัทประกันภัยรถยนต์ และการติดฉลากการฉ้อโกงจะได้รับการยืนยันจากแผนกมืออาชีพของบริษัทประกันภัย ในที่สุด ข้อมูลทั้งหมด 37,082 รายการที่มีอยู่ในชุดข้อมูล และแต่ละรายการแสดงถึงการเคลมประกันรถยนต์ โดยรวมแล้ว มีการกล่าวอ้างที่เป็นการฉ้อโกง 415 ครั้ง และการเรียกร้องที่ไม่เป็นการฉ้อโกง 36,667 ครั้ง ในชุดข้อมูล อัตราส่วนของการอ้างสิทธิ์ที่เป็นการฉ้อโกงต่อการเรียกร้องที่ถูกต้องตามกฎหมายนั้นใกล้เคียงกับ 88:1 ซึ่งแสดงถึงข้อมูลที่ไม่สมดุล ข้อมูลที่ไม่สมดุลอาจส่งผล

กระทบอย่างมากต่อประสิทธิภาพของอัลกอริธึมการจัดหมวดหมู่ ดังนั้นจึงใช้วิธีสุ่มตัวอย่างเพื่อแก้ปัญหาความไม่สมดุลของข้อมูล เนื่องจากมีความแตกต่างกันมากในจำนวนข้อมูลระหว่างคลาสของการอ้างสิทธิ์ เราทั้งคู่จึงมองข้ามการอ้างสิทธิ์ที่ถูกต้องตามกฎหมาย (คลาสส่วนใหญ่) และสุ่มตัวอย่างการอ้างสิทธิ์ที่ฉ้อโกงมากเกินไป (คลาสชนกลุ่มน้อย) เพื่อสร้างสมดุลให้กับชุดข้อมูล ใช้ SMOTE เพื่อสุ่มตัวอย่างการอ้างสิทธิ์ที่เป็นการฉ้อโกงและสุ่มตัวอย่างการอ้างสิทธิ์ที่ถูกต้องตามกฎหมายเพื่อรับข้อมูลจำนวนเท่ากันจากคลาสส่วนใหญ่เพื่อสร้างชุดข้อมูลที่สมดุลสุดท้าย ชุดข้อมูลประกอบด้วยการอ้างสิทธิ์ที่ถูกต้องตามกฎหมาย 1660 และการอ้างสิทธิ์ที่เป็นการฉ้อโกง 1660 รายการ การอ้างสิทธิ์แต่ละรายการประกอบด้วยแอตทริบิวต์ 16 รายการและป้ายกำกับหลอกลวง 1 รายการซึ่งระบุว่าการอ้างสิทธิ์เป็นการอ้างสิทธิ์ที่ฉ้อโกงหรือไม่ คุณลักษณะสามารถแบ่งออกเป็นแอตทริบิวต์หมวดหมู่ 10 รายการ คุณลักษณะตัวเลข 5 รายการ และแอตทริบิวต์ข้อความ 1 รายการ คำอธิบายของข้อมูลมีอยู่ในตารางที่ 1 และสถิติสรุปของข้อมูลที่เป็นตัวเลขแสดงไว้ในตารางที่ 2 ประเภทของรถยนต์ที่เอาประกันภัย ได้แก่ รถเก๋ง รถยนต์โดยสาร รถยนต์เอนกประสงค์ รถสปอร์ต รถตู้ และอื่นๆ รถเอาประกันภัยส่วนใหญ่เป็นรถเก๋ง คิดเป็น 59.9% รถยนต์นั่งส่วนบุคคลคิดเป็น 10.1% อันดับสอง สาเหตุของอุบัติเหตุที่รายงานและตรวจสอบแล้ว ได้แก่ การชน ไฟไหม้ การขโมย รอยขีดข่วน ภัยธรรมชาติ และอื่นๆ ในบรรดาการเรียกร้องนั้น อุบัติเหตุ 86.1% เกิดจากการชนและการเกาคิดเป็น 9.3% ในแง่ของภูมิภาคที่เกิดอุบัติเหตุ พื้นที่เมืองคิดเป็น 41.9% ชนบทคิดเป็น 33.9% และส่วนที่เหลือเป็นภูมิภาคอื่น ข้อมูลหมวดหมู่และข้อมูลตัวเลขสามารถใช้เพื่อสร้างคุณสมบัติของประเภทที่สอดคล้องกัน ข้อมูลข้อความใช้เพื่อแยกคุณลักษณะหัวข้อเชิงลึกที่ประกอบด้วยประสบการณ์ของผู้เชี่ยวชาญ คุณลักษณะเฉพาะที่แยกออกมาจะนำเสนอในรายละเอียดในภายหลัง

การทดลองเปรียบเทียบ เพื่อแสดงการมีส่วนร่วมของ LDA ในแบบจำลองการตรวจจับของเรา เราลบเอาต์พุต LDA ในชุดข้อมูล อัลกอริทึมทั้งสามคือ DNN, RF และ SVM ใช้สำหรับการสร้างแบบจำลอง ประการแรก ในแง่ของอัตรา TP การวิเคราะห์ LDA ช่วยให้ DNN และ RF ทำงานได้ดีขึ้น ในขณะเดียวกัน อัตรา TP ของ DNN นั้นต่ำกว่า SVM ที่ 3.9% โดยไม่มีการวิเคราะห์ LDA แต่ LDA ช่วยให้ DNN แซงหน้า SVM ได้ 22.8% ในอัตรา TP ประการที่สอง อัตรา FP ของอัลกอริธึมทั้งสามลดลงด้วยความช่วยเหลือของ LDA นอกจากนี้ ความแม่นยำและความแม่นยำของ RF และ DNN ยังเพิ่มขึ้นหลังจากการวิเคราะห์ LDA ถูกเพิ่มเข้าไปในวิธีการที่เรานำเสนอ จากการวิเคราะห์ผลการทดลองเปรียบเทียบ เราพบว่าการปรับปรุง LDA ในแบบจำลองนั้นชัดเจน โดยเฉพาะอย่างยิ่งสำหรับ DNN อย่างไรก็ตาม การเพิ่มการวิเคราะห์ LDA จะลดอัตรา TP ความ

แม่นยำ และ F1 ของ SVM การทดสอบ t บ่งชี้ว่า LDA ปรับปรุงประสิทธิภาพของ DNN อย่างมีนัยสำคัญ ความแม่นยำของ DNN กับ LDA (ค่าเฉลี่ย = 0.914) นั้นสูงกว่าความแม่นยำของ DNN ที่ไม่มี LDA อย่างมีนัยสำคัญ (ค่าเฉลี่ย = 0.846) ($t = 13.7877$, $p < 0.0001$) ในขณะเดียวกัน มีความแตกต่างที่สำคัญระหว่างความแม่นยำของ DNN กับ LDA (ค่าเฉลี่ย = 0.917) และความแม่นยำของ DNN ที่ไม่มี LDA (ค่าเฉลี่ย = 0.869) ($t = 11.9275$, $p < 0.0001$)

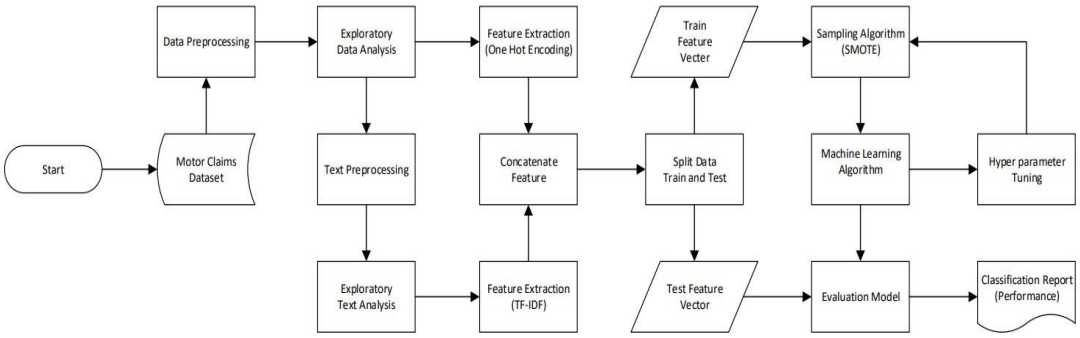
การทดลองเปรียบเทียบแสดงให้เห็นว่าวิธีการที่เสนอในบทความนี้สามารถปรับปรุงความถูกต้องของการตรวจจับการฉ้อโกงประกันภัยรถยนต์ได้ อันดับแรก บทความนี้จะทดสอบ RF และ SVM กับชุดข้อมูลที่ไม่มีคุณลักษณะเฉพาะของ LDA เพื่ออธิบายความสำคัญของการทำเหมืองข้อความในการตรวจจับการฉ้อโกง LDA ถูกใช้เพื่อแยกหัวข้อเพื่อให้อัลกอริทึมสามารถวิเคราะห์ข้อมูลข้อความ และผลการทดลองแสดงให้เห็นว่าการทำเหมืองข้อความมีความสำคัญต่อการปรับปรุงความถูกต้องของการตรวจจับการฉ้อโกง นอกจากนี้ ความแม่นยำของ DNN กับ LDA เพิ่มขึ้น 6.8%, ความแม่นยำเพิ่มขึ้น 4.8%, อัตราการเรียกคืน/TP เพิ่มขึ้น 9.6% และค่า F1 เพิ่มขึ้น 7.2% นอกจากนี้ เพื่อขยายข้อดีของ LDA ให้มากขึ้น บทความนี้จะทดสอบ DNN กับข้อมูลที่ขยายด้วยคุณลักษณะเฉพาะ และ RF และ SVM ถูกใช้เป็นความเปรียบต่าง DNN มีส่วนสนับสนุนอย่างมากต่อประสิทธิภาพที่ดีของตัวแบบ ค่า F1 ของ DNN สูงกว่า RF 9.9% ซึ่งมีความสำคัญในอัลกอริธึมการจำแนกประเภทแบบดั้งเดิมสองแบบ โครงสร้างแบบหลายชั้นช่วยให้ DNN สามารถดึงคุณลักษณะจากข้อมูลได้ดีขึ้น ในขณะที่ LDA ให้วัตถุดิบที่ดีกว่า

บทที่ 3
การดำเนินงานวิจัย

ในการวิจัยในครั้งนี้ ผู้วิจัยได้ดำเนินการวิจัยตามขั้นตอนดังนี้

1. กระบวนการทำงานของแบบจำลอง(Workflows Process of Model)
2. การเก็บรวบรวมข้อมูล(Data Acquisition)
3. การเตรียมข้อมูลคุณลักษณะที่ไม่ใช่ข้อความ(Data Pre-preprocessing)
4. การสำรวจข้อมูลคุณลักษณะที่ไม่ใช่ข้อความ(Exploratory Data Analysis)
5. สร้างคุณลักษณะของข้อมูล(Feature Extraction)
6. การเตรียมข้อมูลคุณลักษณะที่เป็นข้อความ(Text Pre-preprocessing) และการสำรวจข้อมูลคุณลักษณะที่เป็นข้อความ(Exploratory Text Analysis)
7. สร้างคุณลักษณะของข้อมูลข้อความ(Feature Extraction Text)
8. รวมคุณลักษณะที่ไม่ใช่ข้อความ และคุณลักษณะของข้อความเข้าด้วยกัน (Concatenate)
9. การแบ่งข้อมูลสำหรับการเทรนและทดสอบ(Train/Test)
10. ทำการ Scale ข้อมูล(Feature Scaling)
11. อัลกอริทึมและแบบจำลองที่ใช้ทำนาย ทดลองกับข้อมูลที่มีความไม่สมดุลกัน (Model Algorithm with Imbalance Data)
12. แก้ปัญหาความไม่สมดุลกันของข้อมูลด้วยการสุ่มตัวอย่างข้อมูล(Sampling Algorithm)
13. อัลกอริทึมและแบบจำลองที่ใช้ทำนาย ทดลองกับข้อมูลที่มีความสมดุลกัน (Model Algorithm with Balance Data)
14. การวัดประสิทธิภาพและประเมิณผลการทดลองของแบบจำลอง (Model Evaluation)
15. การปรับจูนพารามิเตอร์กับแบบจำลองที่เลือก(Model Parameter Tuning)

1. **กระบวนการทำงานของแบบจำลอง** (Workflows Process of Model)



ภาพประกอบ 11 แสดง Flowchart กระบวนการสร้างแบบจำลอง

จากรูป Flowchart กระบวนการทำงานวิจัยสร้างแบบจำลองและเปรียบเทียบสมรรถนะของอัลกอริทึมที่เราเลือกมา โดยขั้นตอนจะเริ่มจากนำข้อมูลเข้าโดยข้อมูลที่ได้มาจะอยู่ในรูปแบบของ Excel ต่อมาจะทำขั้นตอนการทำความสะอาดข้อมูล(Data Preprocessing) และทำการสำรวจข้อมูล(Exploratory Data Analysis :EDA) โดยจะวิเคราะห์ดูความสัมพันธ์ของข้อมูลจากการเกิดเหตุ เช่นลักษณะการเกิดเหตุ , อายุรถ , เวลาที่เกิดเหตุ , รวมถึงเพศของผู้ขับขี่โดยจะแสดงผลออกมาเป็น Visualization ขั้นตอนต่อมาจะทำความสะอาดข้อความ(Text Preprocessing) และสำรวจข้อความ(Exploratory Text Analysis) ดูการทำความสะอาดข้อมูลข้อความ คำที่ตัดออกมา และ ดูความถี่ของคำที่เกิดขึ้นแสดงออกมาในรูปแบบ Visualization ขั้นตอนถัดมาจะทำการเปลี่ยนขอมูลที่เป็น Category ต่างๆในส่วนของคุณลักษณะที่ไม่ใช่ข้อความให้เป็นตัวเลข(Feature Extraction) และทำการเปลี่ยนข้อมูลข้อความหลังจากที่ได้ทำความสะอาดและตัดคำออกมาแล้วเปลี่ยนให้เป็นตัวเลขให้นำหนักคำโดยใช้ (TFIDF) เพื่อให้ Machine Learning สามารถประมวลผลวิเคราะห์ได้ หลังจากนั้นจะนำเอาข้อมูลที่ทำ Feature Extraction ทั้งข้อมูลที่คุณลักษณะที่ไม่ใช่ข้อความ และข้อมูลที่เป็นข้อความมารวมกัน ขั้นตอนถัดไปจะทำการแบ่งข้อมูล(Split Data) เพื่อใช้ในการฝึกสอน(Training Data) และทดสอบ(Test Data) ในอัตราส่วน 70:30 ขั้นตอนต่อมาจะนำข้อมูลฝึกสอน และข้อมูลทดสอบมาทำการ Scaling ให้อยู่ในมาตรฐานเดียวกัน จากนั้นจะนำข้อมูลที่ใช้ในการฝึกสอน(Training Data) นำมาแก้ปัญหาในการไม่สมดุลกันของข้อมูลก่อนโดยการทำการเพิ่มข้อมูล(Sampling Data) ในประเภทคลาสส่วนน้อย(ทุจริตเคลม) ด้วยวิธีการ Oversampling Technique

หลังจากนั้นจะนำข้อมูลที่ผ่านการทำ Sampling Data แล้วมาทำการสร้างแบบจำลองในการทำนาย โดยใช้เทคนิคการเรียนรู้ของเครื่อง 4 อัลกอริทึมที่แตกต่างกัน ประกอบด้วย Naïve Bayes , Logistic Regression , Support Vector Machine(SVM) และ Random Forest

ขั้นตอนสุดท้ายนำแบบจำลองที่สร้างมาทำการทดสอบประสิทธิภาพและเปรียบเทียบของแต่ละอัลกอริทึมโดยใช้ Confusion Matrix , Accuracy , Precision , Recall , F-measure โดยนำข้อมูลทดสอบ(Test Data) มาทำการทดสอบและประเมินผลประสิทธิภาพของแบบจำลองการทำนายโอกาสเกิดการทุจริตการเคลมสินไหมรถยนต์จากข้อมูลรายงานสำรวจภัยจากเจ้าหน้าที่สำรวจภัย

2. **การเก็บรวบรวมข้อมูล**(Data Acquisition)

ในการวิจัยครั้งนี้ใช้ข้อมูลการเคลมประกันรถยนต์ทั้งที่มีการทุจริตเคลมทั้งหมดในช่วงปี 2562 – เดือนเมษายน 2564 และไม่ทุจริตเคลมในปี 2563 โดยได้รับความอนุเคราะห์ข้อมูลจากบริษัทเอเซียประกันภัย1950 จำกัด(มหาชน) การระบุประเภทว่าเป็นการทุจริตและไม่ทุจริตเคลมนั้นระบุโดยผู้เชี่ยวชาญของฝ่ายสินไหมรถยนต์ ข้อมูลประกอบไปด้วย

- วันที่รับแจ้งอุบัติเหตุ(inform_date)
- เวลาแจ้งอุบัติเหตุ(inform_time)
- วันที่เกิดเหตุ(date_occur)
- เวลาเกิดเหตุ(time_occur)
- วันเริ่มคุ้มครองกรมธรรม์(ins_startdate)
- เพศของผู้ขับขี่(driver_sex)
- รหัสประเภทการเคลม(datacase)
- คำอธิบายประเภทการเคลม(datacase_desc)
- รหัสสาเหตุการเกิดเหตุ(datacasedt)
- คำอธิบายสาเหตุการเกิดเหตุ(datacasedt_desc)
- รายงานผลการสำรวจภัย(comment)
- วันที่ทำกรมธรรม์(tdate)
- รหัสประเภทการใช้รถยนต์(body_type)
- คำอธิบายประเภทการใช้รถยนต์(body_desc)
- รหัสลักษณะการใช้รถยนต์(veh_use)

- คำอธิบายลักษณะการใช้รถยนต์(veh_use_desc)
- อายุรถยนต์(ageofvehicle)
- Label ทุจริต/ไม่ทุจริต(fraudity)

ตาราง 1 รายละเอียดของชุดข้อมูลการเคลมประกันภัยรถยนต์ ปี 2563

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)
1	วันที่รับแจ้งอุบัติเหตุ (inform_date)	วันที่รับแจ้งอุบัติเหตุ
2	เวลาแจ้งอุบัติเหตุ (inform_time)	เวลาแจ้งอุบัติเหตุ
3	วันที่เกิดเหตุ (date_occur)	วันที่เกิดเหตุ
4	เวลาเกิดเหตุ (time_occur)	เวลาเกิดเหตุ
5	วันเริ่มคุ้มครองกรมธรรม์ (ins_startdate)	วันเริ่มคุ้มครองกรมธรรม์
6	เพศของผู้ขับขี่ (driver_sex)	เพศของผู้ขับขี่มีรายละเอียดดังนี้ F = หญิง , M = ชาย
7	รหัสประเภทการเคลม (datacase)	รหัสประเภทการเคลม 100 ฝ่ายผิด 200 ฝ่ายถูก 300 ยังไม่ทราบผลคดี 400 ไม่มีคู่กรณี 428 ประมาทร่วม 988 ยกเลิกรับแจ้ง 989 ยกเลิกเคลม

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
8	คำอธิบายประเภทการเคลม	คำอธิบายประเภทการเคลม	
	(datacase_desc)	100	ฝ่ายผิด
		200	ฝ่ายถูก
		300	ยังไม่ทราบผลคดี
		400	ไม่มีคู่กรณี
		428	ประมาทร่วม
		988	ยกเลิกรับแจ้ง
		989	ยกเลิกเคลม
9	รหัสสาเหตุการเกิดเหตุ	รหัสสาเหตุการเกิดเหตุ	
	(datacasedt)	101	ชนท้ายคู่กรณี
		102	เฉี่ยวชน/เบียดคู่กรณี
		103	ถอยชน/ไหลชนคู่กรณี
		104	ออกจากซอยตัดหน้าคู่กรณี
		105	เลี้ยวเข้าซอยตัดหน้าคู่กรณี
		106	กลับรถตัดหน้าคู่กรณี
		107	เฉี่ยวชนคู่กรณีในวงเวียน
		108	เฉี่ยวชนคู่กรณีในบริเวณสามแยก
		109	เฉียวชนคู่กรณีในสี่แยก
		110	เฉี่ยวชนคู่กรณีที่แยก
		111	เปลี่ยนช่องทางเฉี่ยวชนคู่กรณี
		112	เปลี่ยนช่องทางตัดหน้าคู่กรณี
		113	ชนรถคู่กรณี
		114	ชนคู่กรณีในช่องทางรถสวน
		115	เปิดปรตูชนกับรถคู่กรณี
		116	ชนกันหลายคัน เป็นฝ่ายเสียเปรียบ
		117	ชนคู่กรณีแล้วหลบหนี

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		121	เข้าเกียร์ค้างไว้ สตาร์ทรถพุ่งชนคู่กรณี
		123	เฉี่ยวชน ท/ส คู่กรณีเสียหาย
		124	เฉี่ยวชนสิ่งมีชีวิต(คน/สัตว์)ได้รับบาดเจ็บ
		125	เฉี่ยวชนสิ่งมีชีวิต(คน/สัตว์)เสียชีวิต
		126	ชนสิ่งมีชีวิต(คน/สัตว์)แล้วหลบหนี
		127	ทับ ท/ส คู่กรณีเสียหาย
		128	ทับสิ่งมีชีวิต(คน/สัตว์)ได้รับบาดเจ็บ
		129	ทับสิ่งมีชีวิต(คน/สัตว์)เสียชีวิต
		130	ท/ส บนรถประกันหล่นใส่คู่กรณี
		131	อุปกรณ์ส่วนควบรถประกันหลุดไปชนคู่กรณี
		132	เหยียบหิน/ไม้ กระเด็นไปถูกคู่กรณี
		133	ชนเสาไฟฟ้า/เสาโทรศัพท์
		134	ชนการีดเลนส์หรือหลัก กม.
		135	ยกดั้มพ์เกี่ยวสายไฟฟ้า/สายโทรศัพท์
		136	พุ่งชนคู่กรณี
		137	ฝ่าสัญญาณไฟจราจรชนคู่กรณี
		138	ตกข้างทางแล้วเสียหลักชนรถคู่กรณี/ทส.คู่กรณี
		139	ผู้โดยสารหล่นจากรถ
		140	เฉี่ยวชนและหลบหนี
		141	รถประกันแซง ชนรถคู่กรณี
		142	รถประกันย้อนศร ชนคู่กรณี
		144	หินกระเด่นใส่
		201	ถูกคู่กรณีชนท้าย
		202	ถูกคู่กรณีถอยชน/ไหลชน

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		203	ถูกคู่กรณีเฉี่ยวชน/เบียด
		204	ถูกคู่กรณีกลับรถตัดหน้า
		205	ถูกคู่กรณีกลับรถตัดหน้า
		206	จอดไว้ถูกชนเสียหาย
		208	ถูกคู่กรณีเฉี่ยวชนบริเวณสามแยก
		209	ถูกคู่กรณีเฉี่ยวชนบริเวณสี่แยก
		211	ถูกคู่กรณีเปลี่ยนช่องทางเฉี่ยวชน
		212	ถูกคู่กรณีเปลี่ยนช่องทางตัดหน้า
		213	ถูกคู่กรณีเปิดประตูชน
		214	ถูก ท/ส คู่กรณีหล่นใส่
		215	ถูกคู่กรณีส่วนควบรถคู่กรณีหลุดมาชน
		217	ถูกคู่กรณีล้ำช่องทางเฉี่ยวชน
		218	ถูกคู่กรณีเสียหลักมาชน
		220	รับหลักฐานชดใช้แล้ว
		223	ถูกบุคคลทำให้รถประกันเสียหาย
		224	ถูกคู่กรณีเฉี่ยวชนและหลบหนี
		225	คู่กรณีหลุดโค้งมาชน
		226	ถูกคู่กรณีฝ่าสัญญาณไฟจราจรมาชน
		227	ถูกคู่กรณีแซง ชนรถประกัน
		228	คู่กรณีย้อนศร ชนรถคู่กรณี
		300	ยังไม่ทราบผลคดี
		301	เฉี่ยวชนกับคู่กรณี
		302	สวนชนกับคู่กรณี
		304	เฉี่ยวชนกับคู่กรณีในวงเวียน
		307	เฉี่ยวชนกับคู่กรณีในบริเวณที่แยก
		309	ตกลงกันไม่ได้

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		311	ถอยชนกับคู่กรณี
		400	ไม่มีคู่กรณี
		401	จอดไว้รถหาย
		402	จอดไว้เสียหายไม่ทราบคู่กรณี
		403	จอดไว้ถูกงัด
		404	จอดไว้ถูกลักทรัพย์/ส่วนรวม
		405	เสียหลักตกหลุม/ตกข้างทาง
		406	เสียหลักพลิกคว่ำ
		407	หินกระเด็นใส่
		408	ยางระเบิด
		410	กระแทก/ครูดก้อนหิน
		411	ชนเกาะกลางถนน/ชนฟุตบาท
		412	ชนเสาไฟฟ้า/เสาโทรศัพท์
		413	ชนการ์ดเลน/หลัก กม.
		414	ชนป้ายโฆษณา/ป้ายจราจร
		415	ชนต้นไม้/กระถางต้นไม้
		416	ชนราวสะพาน
		417	ชนกำแพง/รั้ว
		418	ชนขอบทางด่วน/โทล์เวย์
		419	ชนเสาบ้าน/ประตูบ้าน
		420	น้ำท่วม
		421	ไฟไหม้/ไฟฟ้าลัดวงจร
		422	ตกน้ำ/แม่น้ำลำคลอง
		423	กระจกบังลมหน้าแตก
		424	กระจกบังลมหลังแตก
		425	กระจกประตูแตก

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		427	วัตถุหล่นใส่
		428	ประมาทร่วม
		429	เสียหลักพลิกคว่ำไฟไหม้
		430	ชนวัสดุ
		431	ชนกระจก
		432	ชนท่อ
		433	ความเสียหายเกิดจากสัตว์เลี้ยง
		434	เคลมรอบคัน
		435	ความเสียหายเกิดจากการเปิด - ปิดประตู
		436	ชนเหล็กกั้น
		437	ชนโต๊ะ/เก้าอี้
		438	รีโมทกุญแจสูญหาย
		439	เคลมสีรอบคัน
		440	เคลมสีบางส่วน
		441	คู่กรณีเฉี่ยวชนและหลบหนี(ไม่ทราบทะเบียน)
		988	ยกเลิกรับแจ้ง
		989	ยกเลิกเคลม
10	คำอธิบายสาเหตุการเกิดเหตุ (datacasedt_desc)	คำอธิบายสาเหตุการเกิดเหตุ	
		101	ชนท้ายคู่กรณี
		102	เฉี่ยวชน/เบียดคู่กรณี
		103	ถอยชน/ไหลชนคู่กรณี
		104	ออกจากซอยตัดหน้าคู่กรณี
		105	เลี้ยวเข้าซอยตัดหน้าคู่กรณี
		106	กลับรถตัดหน้าคู่กรณี
		107	เฉี่ยวชนคู่กรณีในวงเวียน

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		108	เฉี่ยวชนคู่กรณีในบริเวณสามแยก
		109	เฉียวชนคู่กรณีในสี่แยก
		110	เฉี่ยวชนคู่กรณีที่แยก
		111	เปลี่ยนช่องทางเฉี่ยวชนคู่กรณี
		112	เปลี่ยนช่องทางตัดหน้าคู่กรณี
		113	ชนรถคู่กรณี
		114	ชนคู่กรณีในช่องทางรถสวน
		115	เปิดประตูชนกับรถคู่กรณี
		116	ชนกันหลายคัน เป็นฝ่ายเสียเปรียบ
		117	ชนคู่กรณีแล้วหลบหนี
		121	เข้าเกียร์ค้างไว้ สตาร์ทรถพุ่งชนคู่กรณี
		123	เฉี่ยวชน ท/ส คู่กรณีเสียหาย
		124	เฉี่ยวชนสิ่งมีชีวิต(คน/สัตว์)ได้รับบาดเจ็บ
		125	เฉี่ยวชนสิ่งมีชีวิต(คน/สัตว์)เสียชีวิต
		126	ชนสิ่งมีชีวิต(คน/สัตว์)แล้วหลบหนี
		127	ทับ ท/ส คู่กรณีเสียหาย
		128	ทับสิ่งมีชีวิต(คน/สัตว์)ได้รับบาดเจ็บ
		129	ทับสิ่งมีชีวิต(คน/สัตว์)เสียชีวิต
		130	ท/ส บนรถประกันหล่นใส่คู่กรณี
		131	อุปกรณ์ส่วนควบรถประกันหลุดไปชนคู่กรณี
		132	เหยียบหิน/ไม้ กระเด็นไปถูกคู่กรณี
		133	ชนเสาไฟฟ้า/เสาโทรศัพท์
		134	ชนการ์ดเลนส์หรือหลัก กม.
		135	ยกดั๊มพ์เกี่ยวสายไฟฟ้า/สายโทรศัพท์
		136	พุ่งชนคู่กรณี

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		137	ฝ่าสัญญาณไฟจราจรชนคู่กรณี
		138	ตกข้างทางแล้วเสียหลักชนรถคู่กรณี/ทส. คู่กรณี
		139	ผู้โดยสารหล่นจากรถ
		140	เฉี่ยวชนและหลบหนี
		141	รถประกันแซง ชนรถคู่กรณี
		142	รถประกันย้อนศร ชนคู่กรณี
		144	หินกระเด่นใส่
		201	ถูกคู่กรณีชนท้าย
		202	ถูกคู่กรณีถอยชน/ไหลชน
		203	ถูกคู่กรณีเฉี่ยวชน/เบียด
		204	ถูกคู่กรณีกลับรถตัดหน้า
		205	ถูกคู่กรณีกลับรถตัดหน้า
		206	จอดไว้ถูกชนเสียหาย
		208	ถูกคู่กรณีเฉี่ยวชนบริเวณสามแยก
		209	ถูกคู่กรณีเฉี่ยวชนบริเวณสี่แยก
		211	ถูกคู่กรณีเปลี่ยนช่องทางเฉี่ยวชน
		212	ถูกคู่กรณีเปลี่ยนช่องทางตัดหน้า
		213	ถูกคู่กรณีเปิดประตูชน
		214	ถูก ท/ส คู่กรณีหล่นใส่
		215	ถูกคู่กรณีส่วนควบรถคู่กรณีหลุดมาชน
		217	ถูกคู่กรณีล้ำช่องทางเฉี่ยวชน
		218	ถูกคู่กรณีเสียหลักมาชน
		220	รับหลักฐานชดใช้แล้ว
		223	ถูกบุคคลทำให้รถประกันเสียหาย

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		224	ถูกคู่กรณีเฉี่ยวชนและหลบหนี
		225	คู่กรณีหลุดโค้งมาชน
		226	ถูกคู่กรณีฝ่าสัญญาณไฟจราจรมาชน
		227	ถูกคู่กรณีแซง ชนรถประกัน
		228	คู่กรณีย้อนศร ชนรถคู่กรณี
		300	ยังไม่ทราบผลคดี
		301	เฉี่ยวชนกับคู่กรณี
		302	สวนชนกับคู่กรณี
		304	เฉี่ยวชนกับคู่กรณีในวงเวียน
		307	เฉี่ยวชนกับคู่กรณีในบริเวณที่แยก
		309	ตกลงกันไม่ได้
		311	ถอยชนกับคู่กรณี
		400	ไม่มีคู่กรณี
		401	จอดไว้รถหาย
		402	จอดไว้เสียหายไม่ทราบคู่กรณี
		403	จอดไว้ถูกงัด
		404	จอดไว้ถูกลักทรัพย์/ส่วนรวม
		405	เสียหลักตกหลุม/ตกข้างทาง
		406	เสียหลักพลิกคว่ำ
		407	หินกระเด็นใส่
		408	ยางระเบิด
		410	กระแทก/ครูดก้อนหิน
		411	ชนเกาะกลางถนน/ชนฟุตบาท
		412	ชนเสาไฟฟ้า/เสาโทรศัพท์
		413	ชนการ์ดเลน/หลัก กม.
		414	ชนป้ายโฆษณา/ป้ายจราจร

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		415	ชนต้นไม้/กระถางต้นไม้
		416	ชนราวสะพาน
		417	ชนกำแพง/รั้ว
		418	ชนขอบทางด่วน/โทล์เวย์
		419	ชนเสาบ้าน/ประตูบ้าน
		420	น้ำท่วม
		421	ไฟไหม้/ไฟฟ้าลัดวงจร
		422	ตกน้ำ/แม่น้ำลำคลอง
		423	กระจกบังลมหน้าแตก
		424	กระจกบังลมหลังแตก
		425	กระจกประตูแตก
		427	วัตถุหล่นใส่
		428	ประมาทร่วม
		429	เสียหลักพลิกคว่ำไฟไหม้
		430	ชนวัสดุ
		431	ชนกระจก
		432	ชนท่อ
		433	ความเสียหายเกิดจากสัตว์เลี้ยง
		434	เคลมรอบคัน
		435	ความเสียหายเกิดจากการเปิด - ปิดประตู
		436	ชนเหล็กกั้น
		437	ชนโต๊ะ/เก้าอี้
		438	รีโมทกุญแจสูญหาย
		439	เคลมสีรอบคัน
		440	เคลมสีบางส่วน

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)
		441 คู่กรณีเฉี่ยวชนและหลบหนี(ไม่ทราบทะเบียน)
		988 ยกเลิกรับแจ้ง
		989 ยกเลิกเคลม
11	รายงานผลการสำรวจภัย (comment)	รายละเอียดที่ทางเต้าหน้าที่สำรวจภัยบันทึกรายงานผลนำส่งรายงานให้ฝ่ายสินไหมรถยนต์เพื่อพิจารณาตรวจอนุมัติรับผิดชอบเคลม
12	วันที่ทำกรมธรรม์ (tdate)	วันที่ทำกรมธรรม์ หรือ วันที่บันทึกออกกรมธรรม์
13	รหัสประเภทการใช้รถยนต์ (body_type)	รหัสประเภทการใช้รถยนต์
		0 0
		01 เก๋ง2ตอน
		01-1 นั่งสามตอน
		01-2 นั่งสองตอนท้ายบรรทุก
		01-3 เก๋งตอนเดียว
		02 รถตู้
		021 ตู้บรรทุก
		02-1 นั่งสองตอน
		02-2 นั่งสองตอนสองแถว
		02-3 นั่งสองตอนแวน
		03 รถบรรทุก
		031 รถบรรทุก 6 ล้อ
		032 รถบรรทุก10ล้อ
		04 กระบะบรรทุก
		05 จักรยานยนต์
		07 รถโดยสาร

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)	
		08	ปิคอัพแวน
		10	ดั๊มพ์สิบล้อ
		14	รถพ่วง
		15	รถลากจูง
		20	รถขนขยะ
		21	รถแท็กซี่
		29	รถสามล้อเครื่อง
		29-1	สามล้อไฟฟ้า
		99	อื่นๆ
		MLHJA1407K5146125 0	
		จยย.	0
		รถจักรยานยนต์ 0	
14	คำอธิบายประเภทการใช้รถยนต์ (body_desc)	คำอธิบายประเภทการใช้รถยนต์	
		0	0
		01	เก๋ง2ตอน
		01-1	นั่งสามตอน
		01-2	นั่งสองตอนท้ายบรรทุก
		01-3	เก๋งตอนเดียว
		02	รถตู้
		021	ตู้บรรทุก
		02-1	นั่งสองตอน
		02-2	นั่งสองตอนสองแถว
		02-3	นั่งสองตอนแวน
		03	รถบรรทุก
		031	รถบรรทุก 6 ล้อ
		032	รถบรรทุก10ล้อ

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)
		04 กระบะบรรทุก
		05 จักรยานยนต์
		07 รถโดยสาร
		08 ปิคอัพแวน
		10 ดั๊มพ์สิบล้อ
		14 รถพ่วง
		15 รถลากจูง
		20 รถขนขยะ
		21 รถแท็กซี่
		29 รถสามล้อเครื่อง
		29-1 สามล้อไฟฟ้า
		99 อื่นๆ
		MLHJA1407K5146125 0
		จยย. 0
		รถจักรยานยนต์ 0
15	อายุรถยนต์ (ageofvehicle)	อายุของรถรถยนต์คำนวณจากวันที่เกิดเหตุ กับปีผลิตรถยนต์
16	รหัสลักษณะการใช้รถยนต์ (veh_use)	รหัสลักษณะการใช้รถยนต์
		0 0
		110 ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
		120 ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ
		210 ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
		220 ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ
		230 ใช้รับจ้างสาธารณะ

ตาราง 1 (ต่อ)

ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)
		320 ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุกและขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส และ ไม่ใช้ลากจูงรถพ่วง
		420 ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุกและขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส
		520 ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุกและขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส
		610 ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
		620 ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ
		730 ใช้รับจ้างสาธารณะ
		802 รถพยาบาล
17	คำอธิบายลักษณะการใช้รถยนต์ (veh_use_desc)	คำอธิบายลักษณะการใช้รถยนต์
		0 0
		110 ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
		120 ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ
		210 ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
		220 ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ
		230 ใช้รับจ้างสาธารณะ
		320 ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุกและขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส และ ไม่ใช้ลากจูงรถพ่วง
		420 ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุกและขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส

ตาราง 1 (ต่อ)

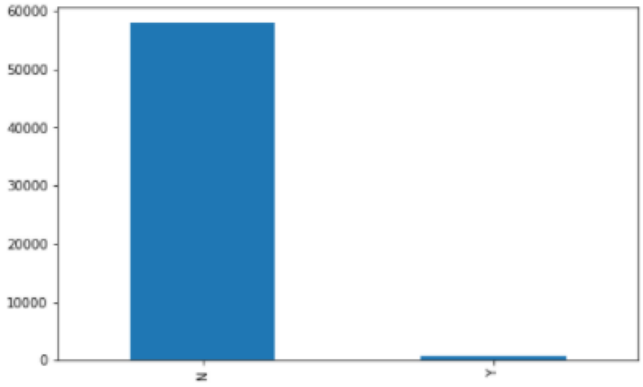
ลำดับที่	ข้อมูลตัวแปร(Variable)	คำอธิบายข้อมูล(Description)
		520 ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุกและขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส
		610 ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
		620 ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ
		730 ใช้รับจ้างสาธารณะ
		802 รถพยาบาล
18	Label ทุจริต/ไม่ทุจริต (fraudity)	Label ว่าเป็นเคลมทุจริตหรือไม่ทุจริต N = ไม่ทุจริต, Y = ทุจริต

โดยข้อมูลมีจำนวนทั้งหมด 58,579 รายการก่อนทำ Data Pre-processing โดยแบ่งออกเป็น ข้อมูลที่ไม่มีการทุจริตเคลมจำนวน 57,890 รายการ และ ข้อมูลที่มีการทุจริตเคลมจำนวน 689 รายการ

inform_date	inform_time	date_occur	time_occur	ins_startdate	driver_sex	datacase	datacase_desc	datacased	datacasedt_desc	comment	tdate	body_typ	body_desc	veh_use	veh_use_desc	ageofvehicle	fraudity
2018-12-17	08:44:30.960	2018-12-17	08:44:3	2018-12-17	08:44:3	2018-12-17	08:3	2018-06-30	00:00	F	28	ประมาทรถ	28	ประมาทรถ	AA6112/1818 ร้นที่ 19/12/61 ส	2018-06-29	02	รถยนต์	220	ใช้เพื่อการพาณิชย์	11	Y
2018-12-17	08:44:30.960	2018-12-17	08:44:3	2018-12-17	08:44:3	2018-12-17	08:3	2018-06-30	00:00	F	28	ประมาทรถ	28	ประมาทรถ	****ลูกใน**** เพิ่มเติม	2018-06-29	02	รถยนต์	220	ใช้เพื่อการพาณิชย์	11	Y
2019-05-11	02:59:38.523	2019-05-11	02:59:4	2019-05-11	02:59:3	2019-05-11	01:5	2019-02-27	00:00	M	300	ยังไม่ทราบผลคดี	301	เฉี่ยวชนกับบุคคลนี้	รับแจ้ง62052466 AA6305/00	2019-02-23	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	10	Y
2019-05-11	02:59:38.523	2019-05-11	02:59:4	2019-05-11	02:59:3	2019-05-11	01:5	2019-02-27	00:00	M	300	ยังไม่ทราบผลคดี	301	เฉี่ยวชนกับบุคคลนี้	AA6205/1021 รับ11/05/2562	2019-02-23	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	10	Y
2019-07-17	17:52:41.537	2019-07-17	17:52:4	2019-07-17	17:52:4	2019-07-17	17:5	2018-12-25	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	ขายและซ่อมทางด้านยนต์ทั่วไป ทาง	2018-12-20	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2019-07-17	17:52:41.537	2019-07-17	17:52:4	2019-07-17	17:52:4	2019-07-17	17:5	2018-12-25	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	สรุปรายงานการตรวจสอบ รอบ	2018-12-20	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2019-07-17	17:52:41.537	2019-07-17	17:52:4	2019-07-17	17:52:4	2019-07-17	17:5	2018-12-25	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	ความเสียหายของรถคู่เอาประกัน :	2018-12-20	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2019-07-17	17:52:41.537	2019-07-17	17:52:4	2019-07-17	17:52:4	2019-07-17	17:5	2018-12-25	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	รายการประเมินราคา ที่ : ชื่อ-นามสกุล	2018-12-20	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2019-07-17	17:52:41.537	2019-07-17	17:52:4	2019-07-17	17:52:4	2019-07-17	17:5	2018-12-25	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	AN6207/0988 รร.620	2018-12-20	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2019-08-25	19:24:52.230	2019-08-25	19:24:5	2019-08-25	19:24:5	2019-08-25	19:2	2019-02-11	00:00	M	100	ฝ่ายผิด	103	ออกชน/ไหล่ชนถูกรถนี้	62085539 AA6208/2521 สต/เก่า	2019-02-11	02	รถตู้	210	ใช้สาธารณะ ไม่ใช้	2	Y
2019-10-30	17:48:30.130	2019-10-30	17:48:3	2019-10-29	17:48:3	2019-10-30	15:0	2019-06-01	00:00	F	989	ยกเลิกเคลม	989	ยกเลิกเคลม	An6210/1620 ร	2019-05-22	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	5	Y
2019-11-04	19:07:18.510	2019-11-04	19:07:1	2019-11-04	19:07:1	2019-11-04	18:5	2018-11-13	00:00	M	100	ฝ่ายผิด	121	ไฟไหม้ไฟฟ้าดับทรัพย์	เมื่อวันที่ 04/11/2562เวลา 19.15 น	2018-11-02	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	6	Y
2019-12-26	08:56:55.187	2019-12-26	08:56:5	2019-12-26	08:56:5	2019-12-26	08:5	2019-01-28	00:00	M	989	ยกเลิกเคลม	989	ยกเลิกเคลม	พบรอยประทับและรอยถูกรถนี้พร้อมมีรูป	2018-12-25	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	13	Y
2019-12-28	17:41:37.683	2019-12-28	17:41:3	2019-12-28	17:41:3	2019-12-28	08:4	2019-06-15	00:00	M	100	ฝ่ายผิด	405	เสียหลักลอกลอย/ลอยข้างทาง	62125976, AN6301/0273 รับที่	2019-05-30	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	2	Y
2019-12-28	17:41:37.683	2019-12-28	17:41:3	2019-12-28	17:41:3	2019-12-28	08:4	2019-06-15	00:00	M	100	ฝ่ายผิด	405	เสียหลักลอกลอย/ลอยข้างทาง	62125976, AN6211/1439 รับที่	2019-05-30	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	2	Y
2019-12-31	02:04:35.893	2019-12-31	02:04:3	2019-12-31	02:04:3	2019-12-31	02:0	2019-06-04	00:00	M	100	ฝ่ายผิด	103	ออกชน/ไหล่ชนถูกรถนี้	AA6212/2876 รับที่ 31/12/2562	2019-05-25	21	รถแท็กซี่	730	ใช้รับจ้างสาธารณะ	5	Y
2019-12-31	14:38:59.743	2019-12-31	14:38:5	2019-12-31	14:38:5	2019-12-31	12:0	2019-03-26	00:00	F	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	เรียน ผู้จัดการฝ่ายสินไหมรถยนต์	2019-03-21	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	15	Y
2019-12-31	14:38:59.743	2019-12-31	14:38:5	2019-12-31	14:38:5	2019-12-31	12:0	2019-03-26	00:00	F	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	AN 6301/0418 (ตามวาง1 สต.เมื่อ	2019-03-21	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	15	Y
2019-12-31	15:16:40.300	2019-12-31	15:16:4	2019-12-31	15:16:4	2019-12-31	15:1	2019-02-18	00:00	F	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	**พบขยาย** เมื่อช่วงวันที่ระดั	2019-02-18	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	20	Y
2019-12-31	15:16:40.300	2019-12-31	15:16:4	2019-12-31	15:16:4	2019-12-31	15:1	2019-02-18	00:00	F	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	AA6212/2911 สต ทำเคลมวัน	2019-02-18	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	20	Y
2020-01-01	01:46:02.660	2020-01-01	01:46:0	2020-01-01	01:46:0	2020-01-01	01:4	2019-04-17	00:00	F	989	ยกเลิกเคลม	989	ยกเลิกเคลม	++เพิ่มเติม++ -ภายหลัง อ/ส เอก	2019-03-27	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	16	Y
2020-01-01	01:46:02.660	2020-01-01	01:46:0	2020-01-01	01:46:0	2020-01-01	01:4	2019-04-17	00:00	F	989	ยกเลิกเคลม	989	ยกเลิกเคลม	An6301/0001 น้ำ 1 1/1/63 ร	2019-03-27	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	16	Y
2020-01-01	11:38:05.817	2020-01-01	11:38:0	2020-01-01	11:38:0	2020-01-01	11:3	2019-03-14	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	ออกตรวจสอบอุบัติเหตุที่ น้ำ ม.ล.	2019-03-14	05	รัถยานยนต์	510	ใช้สาธารณะ ไม่ใช้	8	Y
2020-01-02	11:44:38.440	2020-01-02	11:44:3	2020-01-02	11:44:3	2020-01-02	11:2	2019-01-19	00:00	F	100	ฝ่ายผิด	412	ชนเสาไฟฟ้า/เสารั้วทรัพย์ทาง	AN 6301 / 0067 รร 6301020	2018-12-17	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	8	Y
2020-01-02	11:44:38.440	2020-01-02	11:44:3	2020-01-02	11:44:3	2020-01-02	11:2	2019-01-19	00:00	F	100	ฝ่ายผิด	412	ชนเสาไฟฟ้า/เสารั้วทรัพย์ทาง	AN 6301 / 0269 รร 6301020	2018-12-17	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	8	Y
2020-01-02	19:47:01.870	2020-01-02	19:47:0	2020-01-02	19:47:0	2020-01-02	19:4	2019-04-28	00:00	M	100	ฝ่ายผิด	111	เปลี่ยนช่องทางเรื่องชนถูกรถนี้	AA6301/0095 สต ทำเคลมวัน	2019-04-15	01	เก๋ง2ตอน	110	ใช้รับจ้างสาธารณะ	5	Y
2020-01-03	00:05:21.420	2020-01-03	00:05:2	2020-01-03	00:05:2	2020-01-03	00:0	2019-11-17	00:00	M	100	ฝ่ายผิด	405	เสียหลักลอกลอย/ลอยข้างทาง	เลขรับแจ้ง 63010337/เลขเคลีส 2	2019-11-15	04	กระบะบรรทุก	320	ใช้สาธารณะ ไม่ใช้	2	Y
2020-01-03	18:37:28.170	2020-01-03	18:37:2	2020-01-03	18:37:2	2020-01-03	18:3	2019-12-24	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	AA6301-08539 ตามนัดหมาย รับ	2019-12-24	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2020-01-03	18:37:28.170	2020-01-03	18:37:2	2020-01-03	18:37:2	2020-01-03	18:3	2019-12-24	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	mp80/0163 ผู้เห็นเหตุประกันแจ้งว่า	2019-12-24	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	12	Y
2020-01-06	20:07:03.150	2020-01-06	20:07:0	2020-01-06	20:07:0	2020-01-06	20:0	2019-07-04	00:00	F	100	ฝ่ายผิด	101	ชนท้ายกลอยนี้	พยักงานรับแจ้งอุบัติเหตุรับวัน	2019-07-04	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	16	Y
2020-01-06	21:52:49.160	2020-01-06	21:52:4	2020-01-06	21:52:4	2020-01-06	21:5	2019-07-04	00:00	F	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	เลขรับแจ้ง 63011124 เลขเคลีส /2	2019-07-04	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	18	Y
2020-01-13	17:12:18.000	2020-01-13	17:12:1	2020-01-13	16:50:1	2020-01-13	16:5	2019-08-14	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	เลขรับแจ้ง 630113162 เลขเคลีส 2	2019-05-10	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	21	Y
2020-01-13	18:50:02.000	2020-01-13	18:50:0	2020-01-13	18:20:0	2020-01-13	18:2	2019-02-17	00:00	M	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	เรียน ผู้จัดการฝ่ายสินไหมฯ ทราบ	2019-02-17	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	15	Y
2020-01-10	04:20:40.423	2020-01-10	04:20:4	2020-01-10	04:20:4	2020-01-10	04:1	2019-05-24	00:00	M	100	ฝ่ายผิด	101	ชนท้ายกลอยนี้	รอประกันแสมาตามสนามหลวงสิน ทา	2019-05-24	01	เก๋ง2ตอน	730	ใช้รับจ้างสาธารณะ	1	Y
2020-01-18	16:58:58.000	2020-01-18	16:58:5	2020-01-18	16:50:5	2020-01-18	16:5	2020-01-18	00:00	M	989	ยกเลิกเคลม	989	ยกเลิกเคลม	630118117 AN6301/0843	2020-01-18	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	9	Y
2020-01-19	11:46:26.000	2020-01-19	11:46:2	2020-01-19	11:40:2	2020-01-19	11:4	2019-11-04	00:00	M	100	ฝ่ายผิด	415	ชนต้นไม้/ขอบทางต้นไม้	รช630119038 AN6301/0875 (แร	2019-11-04	04	กระบะบรรทุก	320	ใช้เพื่อการพาณิชย์	5	Y
2020-01-19	14:34:30.000	2020-01-19	14:34:3	2020-01-19	14:30:3	2020-01-19	14:3	2019-03-22	00:00	F	100	ฝ่ายผิด	429	เสียหลักหลบหลีกการไฟไหม้	เรียน ทุกท่านด้วยสินไหมรถยนต์ทาง	2019-03-04	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	7	Y
2020-01-11	16:56:34.330	2020-01-11	16:56:3	2020-01-11	16:56:3	2020-01-11	16:4	2019-08-03	00:00	M	100	ฝ่ายผิด	103	ออกชน/ไหล่ชนถูกรถนี้	เมื่อวันที่ 11 มกราคม 2563 เวลา 17	2019-08-03	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	26	Y
2020-01-11	18:31:24.100	2020-01-11	18:31:2	2020-01-11	18:31:2	2020-01-11	18:2	2019-04-08	00:00	F	100	ฝ่ายผิด	102	เฉี่ยวชน/ปิดถนนนี้	AA6301/0853 แจกน้ำใบหล วันที่	2019-03-30	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	17	Y
2020-01-21	03:10:21.000	2020-01-21	03:10:2	2020-01-21	03:09:2	2020-01-21	03:0	2019-08-30	00:00	M	100	ฝ่ายผิด	123	เฉี่ยวชน ท/ส ถูกรถเสียหาย	รอประกันเรื่องรถท่านเอง ขับเคลื่อน	2019-08-02	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	7	Y
2020-01-21	03:10:21.000	2020-01-21	03:10:2	2020-01-21	03:09:2	2020-01-21	03:0	2019-08-30	00:00	M	100	ฝ่ายผิด	123	เฉี่ยวชน ท/ส ถูกรถเสียหาย	หลักงานเส้นทางออกตรวจสอบว	2019-08-02	01	เก๋ง2ตอน	110	ใช้สาธารณะ ไม่ใช้	7	Y

ภาพประกอบ 12 ตัวอย่างข้อมูลที่ใช้สร้างแบบจำลอง

```
N    57890
Y      689
Name: fraudity, dtype: int64
```



ภาพประกอบ 13 แสดงจำนวนข้อมูลที่ทุจริต และ ไม่ทุจริต

3. **การเตรียมข้อมูลคุณลักษณะที่ไม่ใช่ข้อความ**(Data Pre-processing)

ขั้นตอนในการเตรียมข้อมูลคุณลักษณะที่ไม่ใช้ข้อความเริ่มจากการสำรวจ ประเภทข้อมูล (Data type) ข้อมูลในแต่ละคุณลักษณะว่าพร้อมที่จะนำไปใช้งานหรือไม่ ปรากฏว่าประเภทของข้อมูลยังไม่ตรง

```
In [3]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 58579 entries, 0 to 58578
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   inform_date      58579 non-null  object
 1   inform_time      58558 non-null  object
 2   date_occur       58579 non-null  object
 3   time_occur       58190 non-null  object
 4   ins_startdate    58225 non-null  object
 5   driver_sex       57635 non-null  object
 6   datacase         58579 non-null  int64
 7   datacase_desc    58579 non-null  object
 8   datacasedt       58579 non-null  int64
 9   datacasedt_desc  58579 non-null  object
 10  comment          58516 non-null  object
 11  fraudity         58579 non-null  object
 12  tdate            57220 non-null  object
 13  body_type        57220 non-null  object
 14  body_desc        57220 non-null  object
 15  veh_use          57219 non-null  float64
 16  veh_use_desc     57219 non-null  object
 17  ageofvehicle     58413 non-null  float64
dtypes: float64(2), int64(2), object(14)
memory usage: 8.0+ MB
```

ภาพประกอบ 14 แสดงประเภทของข้อมูล

จากนั้นจะทำการตรวจสอบค่าNull Value จะพบว่าข้อมูลมีค่าNull Value ของ Feature tdate , body_type , body_desc , veh_use , veh_use_desc อยู่จำนวน 1,359 รายการจึงทำการเลือก Drop Null ของ tdate ทิ้งเนื่องจากให้ความสำคัญกับ feature นี้มากที่สุด

```
Out[6]: inform_date          0
        inform_time         21
        date_occur           0
        time_occur         389
        ins_startdate      354
        driver_sex         944
        datacase             0
        datacase_desc        0
        datacasedt           0
        datacasedt_desc      0
        comment             63
        fraudity             0
        tdate             1359
        body_type         1359
        body_desc         1359
        veh_use           1360
        veh_use_desc      1360
        ageofvehicle       166
        dtype: int64
```

```
Out[10]: inform_date          0
         inform_time          6
         date_occur           0
         time_occur         277
         ins_startdate        7
         driver_sex         943
         datacase             0
         datacase_desc        0
         datacasedt           0
         datacasedt_desc      0
         comment             58
         fraudity             0
         tdate                0
         body_type            0
         body_desc            0
         veh_use              1
         veh_use_desc         1
         ageofvehicle         0
         dtype: int64
```

ภาพประกอบ 15 แสดง Drop Null Value ของ feature tdate

จากการตรวจสอบ data type เบื้องต้นเราจะทำการเปลี่ยน data type ให้กับ feature ที่เกี่ยวข้องเพื่อใช้ในการคำนวณต่อไปดังนี้ veh_use ให้เป็น string , inform_date ให้เป็น datatime และ date_occur ให้เป็น datatime

```
In [11]: df['veh_use'] = df['veh_use'].astype(str)
         df['inform_date'] = pd.to_datetime(df['inform_date'])
         df['date_occur'] = pd.to_datetime(df['date_occur'])

In [12]: df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 57220 entries, 0 to 58578
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   inform_date      57220 non-null  datetime64[ns]
 1   inform_time      57214 non-null  object
 2   date_occur       57220 non-null  datetime64[ns]
 3   time_occur       56943 non-null  object
 4   ins_startdate    57213 non-null  object
 5   driver_sex       56277 non-null  object
 6   datacase         57220 non-null  int64
 7   datacase_desc    57220 non-null  object
 8   datacasedt       57220 non-null  int64
 9   datacasedt_desc  57220 non-null  object
 10  comment          57162 non-null  object
 11  fraudity         57220 non-null  object
 12  tdate            57220 non-null  object
 13  body_type        57220 non-null  object
 14  body_desc        57220 non-null  object
 15  veh_use          57220 non-null  object
 16  veh_use_desc     57219 non-null  object
 17  ageofvehicle     57220 non-null  float64
dtypes: datetime64[ns](2), float64(1), int64(2), object(13)
memory usage: 8.3+ MB
```

ภาพประกอบ 16 แสดงการเปลี่ยน data type ของ feature ที่เราสนใจ

หลังจากนั้นจะทำการแทนที่ข้อมูลที่เป็น NaN ของ feature inform_time และ time_occur โดยทำให้เป็นเวลา 00:00:00

ทำการสร้าง Feature date_inform , time_inform , date_acci , time_acci , date_start , tdate สำหรับการเปลี่ยน data type ให้ถูกต้อง และเพื่อรองรับในการคำนวณ จากนั้นทำการสร้าง Feature date diff ระหว่าง วันที่รับแจ้งอุบัติเหตุกับวันที่ทำกรมธรรม์ให้เป็น tday_apart และ วันที่รับแจ้งอุบัติเหตุกับวันที่เริ่มคุ้มครอง เป็น eday_apart

Extract Data - Time and calculate date diff

```
In [16]: df['date_inform'] = pd.to_datetime(df['inform_date']).dt.date
         df['time_inform'] = pd.to_datetime(df['inform_date']).dt.time
         df['date_acci'] = pd.to_datetime(df['date_occur']).dt.date
         df['time_acci'] = pd.to_datetime(df['date_occur']).dt.time
         df['date_start'] = pd.to_datetime(df['ins_startdate']).dt.date
         df['tdate'] = pd.to_datetime(df['tdate']).dt.date
         df['tday_apart'] = df['date_inform'] - df['tdate'] #diff days between informcliam to date policy
         df['eday_apart'] = df['date_inform'] - df['date_start'] #diff days between informcliam to date policy
```

ภาพประกอบ 17 แสดงการสร้าง Feature tday_apart

จากนั้นจะทำการ Group ข้อมูลสาเหตุการเกิดเหตุ(datacasedt) ให้ลดน้อยลง โดยการนำ Table Mapping ข้อมูลของ Code datacasedt เข้ามาทำการ Mearge เพื่อให้ได้ Group ของสาเหตุการเกิดเหตุ

Out[18]:

	datacasedt	groupcasedt	groupcasedtdesc
0	dt_code	GroupCode	GroupDesc
1	121	1	เข้าเกียร์ค้างไว้ สตาร์ทรถพุ่งชนคู่กรณี
2	434	2	เคลมรอบคัน
3	440	3	เคลมสี
4	439	3	เคลมสี

```
In [22]: dfclaim.head(100)
```

esc	datacasedt	datacasedt_desc	...	ageofvehicle	date_inform	time_inform	date_acci	time_acci	date_start	tday_apart	eday_apart	groupcasedt	groupcasedtdesc
ม ...	428	ประมาทร่วม ...	...	11.0	2018-12-17	08:44:30.960000	2018-12-17	08:44:30.960000	2018-06-30	171 days	170 days	36	ประมาทร่วม
ม ...	428	ประมาทร่วม ...	...	11.0	2018-12-17	08:44:30.960000	2018-12-17	08:44:30.960000	2018-06-30	171 days	170 days	36	ประมาทร่วม
ลคดี ...	301	เฉี่ยวชนกับคู่กรณี ...	...	10.0	2019-05-11	02:59:38.523000	2019-05-11	02:59:38.523000	2019-02-27	77 days	73 days	4	เฉี่ยวชนคู่กรณี
ลคดี ...	301	เฉี่ยวชนกับคู่กรณี ...	...	10.0	2019-05-11	02:59:38.523000	2019-05-11	02:59:38.523000	2019-02-27	77 days	73 days	4	เฉี่ยวชนคู่กรณี
ค ...	102	เฉี่ยวชน/เบียดคู่กรณี ...	...	12.0	2019-07-17	17:52:41.537000	2019-07-17	17:52:41.537000	2018-12-25	209 days	204 days	4	เฉี่ยวชนคู่กรณี
...	...	...	...	...	...	...	...	...	...	...	...	...	...
ลคดี ...	301	เฉี่ยวชนกับคู่กรณี ...	...	13.0	2020-02-20	22:39:55	2020-02-20	21:30:55	2019-09-05	175 days	168 days	4	เฉี่ยวชนคู่กรณี

ภาพประกอบ 18 แสดงการจัด Group ข้อมูลสาเหตุการเกิดเหตุ

หลังจากนั้นจะทำการสร้าง feature นับจำนวนตัวอักษรในรายงานสำรวจภัย(comment) ว่ามีปริมาณตัวอักษรมากน้อยแค่ไหน โดยจะทำการตัดช่องว่างหน้าหลังของประโยครายงานสำรวจภัย และทำการนับคำเก็บไว้เป็น feature CommentLength เพื่อดูว่าปริมาณตัวอักษรในแต่ละรายงานสำรวจภัยมากน้อยแค่ไหน และ ลองดูค่าทางสถิติของข้อมูลที่ทำ Preprocessing ทั้งหมด ก่อนที่จะเลือก Feature ไปใช้งาน

	datacase	ageofvehicle	tday_apart	eday_apart	CommentLength
count	57220.000000	57220.000000	57220	57213	57162.000000
mean	239.935687	10.221059	203 days 06:49:05.934987	266 days 17:27:35.309108	978.608341
std	288.901425	7.857184	216 days 22:31:16.678487	1807 days 22:01:06.762088	431.392881
min	100.000000	1.000000	-12 days +00:00:00	-31 days +00:00:00	87.000000
25%	100.000000	4.000000	103 days 00:00:00	94 days 00:00:00	691.000000
50%	100.000000	8.000000	206 days 00:00:00	197 days 00:00:00	921.000000
75%	200.000000	14.000000	301 days 00:00:00	291 days 00:00:00	1192.000000
max	989.000000	78.000000	44074 days 00:00:00	44183 days 00:00:00	5103.000000

ภาพประกอบ 19 แสดงค่าทางสถิติของข้อมูลที่ทำ Pre-processing ก่อนเลือก Feature ไปใช้งาน

จากนั้นจะทำการ สร้าง Feature ชั่วโมงที่เกิดเหตุ(acci_hours) สำหรับไว้ดูปริมาณช่วงเวลาที่เกิดอุบัติเหตุบ่อยที่สุด

จากนั้นจะเลือก Feature มาใช้งานทั้งหมด 10 Feature จาก 29 Feature ที่เรามี โดยจะเลือกมาดังนี้

- จำนวนวันที่แจ้งเคลมหลังจากทำกรมธรรม์(tday_apart)
- คำอธิบายประเภทการเคลม(datacase_desc)
- สาเหตุการเกิดเหตุที่จัดกลุ่มมาแล้ว(groupcasedtdesc)
- ประเภทตัวถังรถยนต์(body_desc)
- เพศของผู้ขับขี่(driver_sex)
- อายุรถยนต์(ageofvehicle)
- ชั่วโมงที่เกิดอุบัติเหตุ(acci_hours)
- ประเภทการใช้รถยนต์(veh_use_desc)
- รายงานสำรวจภัยจาก Surveyor(comment)
- Label ทุจริต/ไม่ทุจริต(fraudity)

4. **การสำรวจข้อมูลคุณลักษณะที่ไม่ใช่ข้อความ**(Exploratory Data Analysis)

ทำการสำรวจข้อมูลอายุรถยนต์ที่เกิน 30 ปี พบว่ามีปริมาณรายการ 615 รายการ

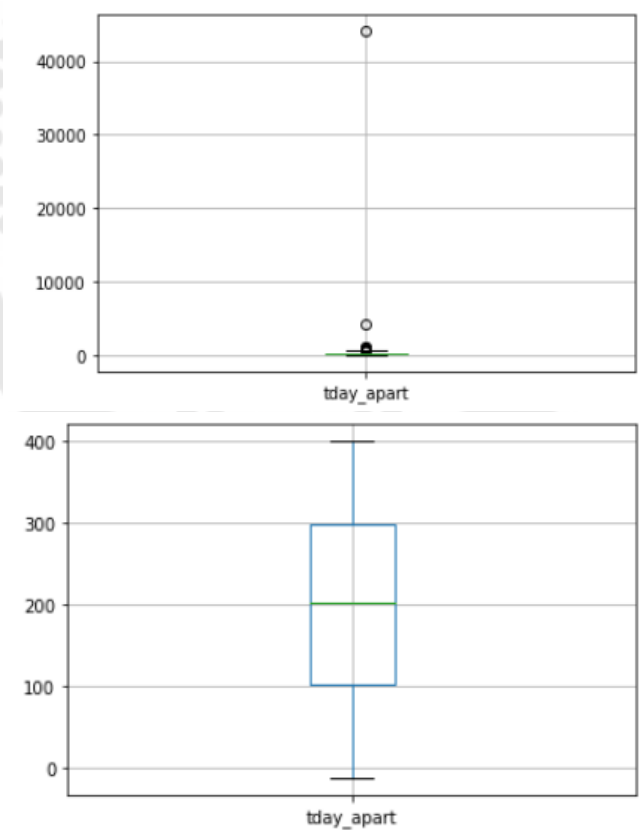
```
In [38]: condition1 = dfclaim['ageofvehicle'] > 30
         dfclaim[condition1]
```

56293	353 days	ฝ่ายผิด ...	ถอยชนคู่กรณี	รถโดยสาร	M	34.0	17	ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า	พนักงานเดินทางออกตรวจสอบสถานที่เกิดเหตุพบทุกฝ...	N
56436	337 days	ฝ่ายถูก ...	คู่กรณีถอยชน	เก๋ง2ตอน	M	32.0	09	ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า	MP3967/0463 - ออกตรวจสอบที่เกิดเหตุธนาคารกสิ...	N
56772	258 days	ฝ่ายผิด ...	ถอยชนคู่กรณี	กระบะบรรทุก	M	64.0	09	ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส...	- ออกตรวจสอบที่เกิดเหตุตามรับแจ้งบริเวณ หมู่บ...	N
56882	322 days	ยกเลิกเคลม ...	ยกเลิกรับแจ้ง/ยกเลิกเคลม	ปิคอัพแวน	F	36.0	16	ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส...	MP2619/0263 เมื่อวันที่ 26 กุมภาพันธ์ 2563 เว...	N
56948	247 days	ฝ่ายผิด ...	เฉี่ยวชนคู่กรณี	กระบะบรรทุก	M	32.0	17	ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส...	เมื่อวันที่ 23 มกราคม 2561 เวลา 18.01 น ได้รับ...	N

615 rows × 10 columns

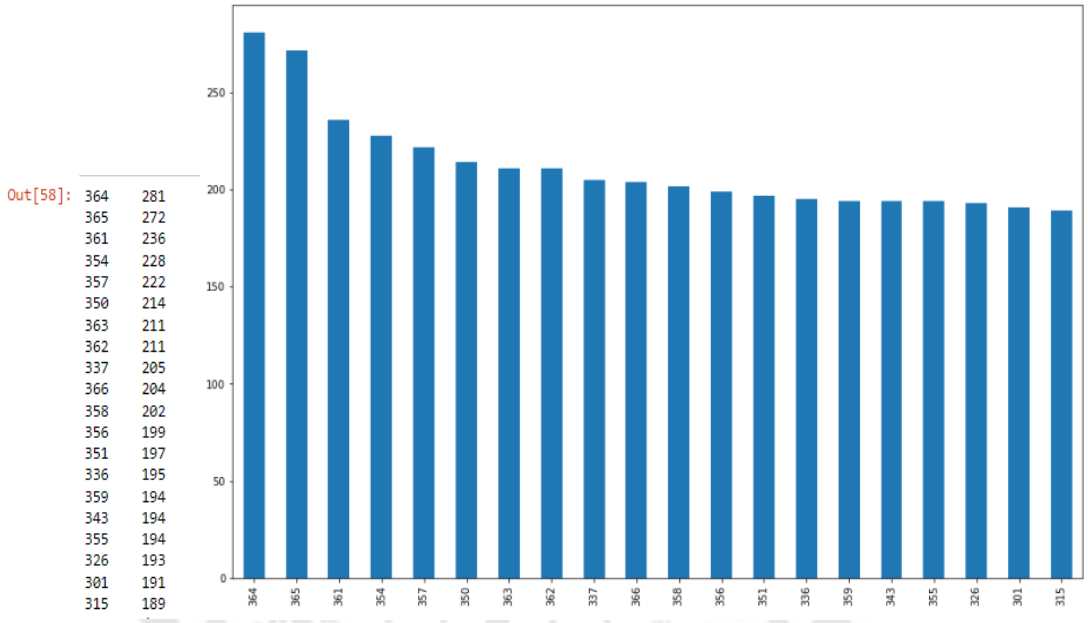
ภาพประกอบ 20 แสดงข้อมูลเคลมที่รถมีอายุ > 30 ปี

ทำการสำรวจข้อมูล วันที่ลูกค้าแจ้งเคลมหลังจากวันที่ทำกรมธรรม์(tday_apart) จะเห็นว่าข้อมูลมี Outliner อยู่อาจส่งผลให้การทำนายค่าของ Model ผิดเพี้ยนไปได้ เราเลยทำการตัดข้อมูลที่เป็น Outliner มากทิ้งไป โดยการเลือกข้อมูลที่ tday_apart < 400 วันมาทำ



ภาพประกอบ 21 แสดงข้อมูล tday_apart ที่มี Outliner และหลังกำจัด Outliner

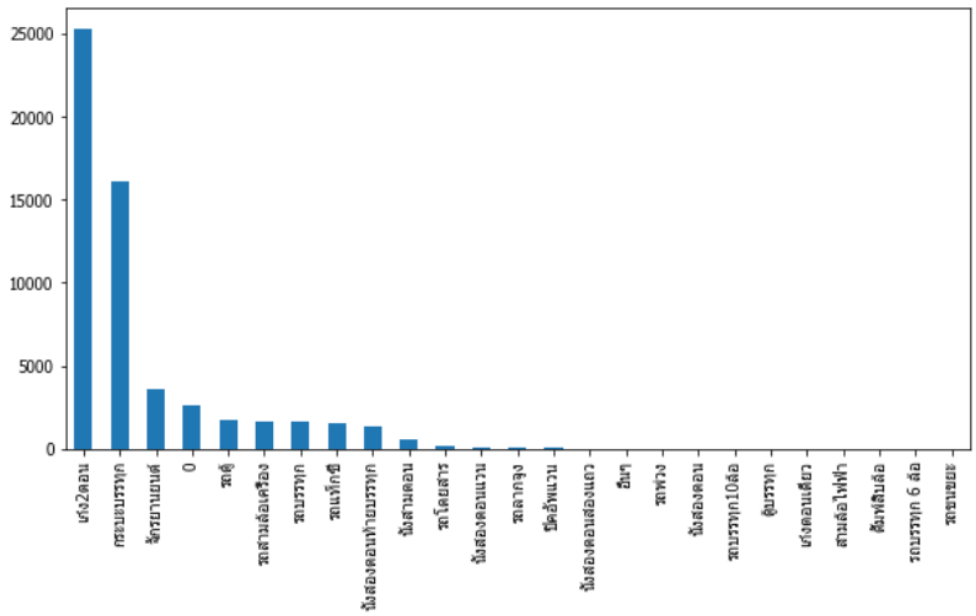
Plot ข้อมูลการวันที่ลูกค้าแจ้งเคลมหลังจากวันที่ทำกรมธรรม์(tday_apart) ออกมาดูปริมาณการแจ้งเคลมอยู่ในช่วงไหนมากที่สุด ดูจากกราฟจะเห็นได้ว่ามีการแจ้งเคลมก่อนจะหมดอายุกรมธรรม์เยอะที่สุดเนื่องจากลูกค้าส่วนใหญ่จะเก็บความเสียหายไว้ก่อนแล้วนำมาแจ้งทำเคลมใกล้ๆกับวันหมดอายุกรมธรรม์ สืบเนื่องมากจากลูกค้าอาจจะหลีกเลี่ยงค่าเบี้ยประกันปีต่ออายุที่อาจจะโดนปรับเพิ่มขึ้นได้



ภาพประกอบ 22 แสดงข้อมูลระยะเวลาที่ลูกค้ามาทำเคลมหลังจากทำกรมธรรม์

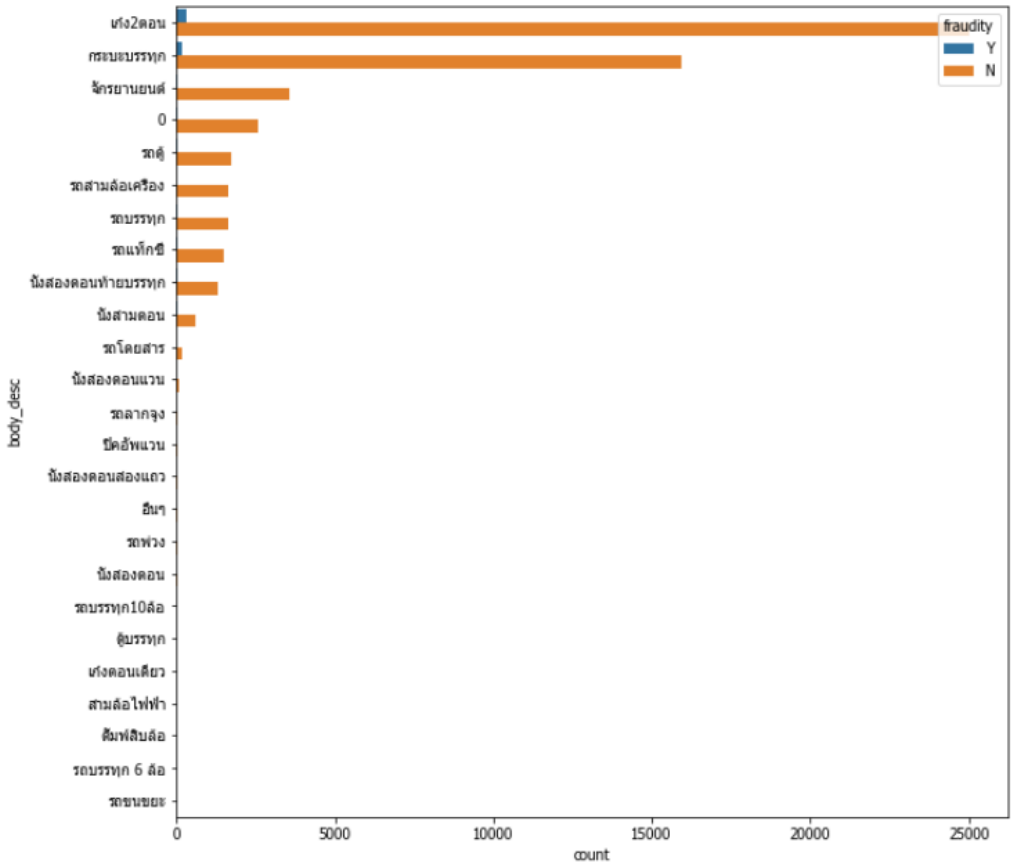
ดูข้อมูลตัวถังรถว่าเป็นรถประเภทไหน และนับจำนวนเพื่อวิเคราะห์ได้ว่ารถประเภทไหนมีการทำเคลมมากน้อยต่างกันทำให้เราวางแผนการรับประกันหรือออกผลิตภัณฑ์ให้สอดคล้องได้ เราจะเห็นได้ว่าในปริมาณเคลมที่เกิดขึ้นส่วน 3 อันดับแรกจะเป็น รถเก๋ง , รถกระบะ และ รถมอเตอร์ไซค์

```
Out[56]: เก่ง2ตอน                25317
         กระบะบรรทุก             16128
         จักรยานยนต์              3606
         0                         2624
         รถตู้                    1757
         รถสามล้อเครื่อง          1645
         รถบรรทุก                 1633
         รถแท็กซี่                1524
         นั่งสองตอนท้ายบรรทุก     1299
         นั่งสามตอน                584
         รถโดยสาร                  167
         นั่งสองตอนแวน              58
         รถลากจูง                   44
         ปิคอัพแวน                  30
         นั่งสองตอนสองแถว           15
         อื่นๆ                      13
         รถพ่วง                     10
         นั่งสองตอน                 10
         รถบรรทุก10ล้อ               8
         ตู้บรรทุก                   7
         เก่งตอนเตียว                7
         สามล้อไฟฟ้า                 3
         ดั๊มพ์สิบล้อ                2
         รถบรรทุก 6 ล้อ              2
         รถขนขยะ                     2
         Name: body_desc, dtype: int64
```



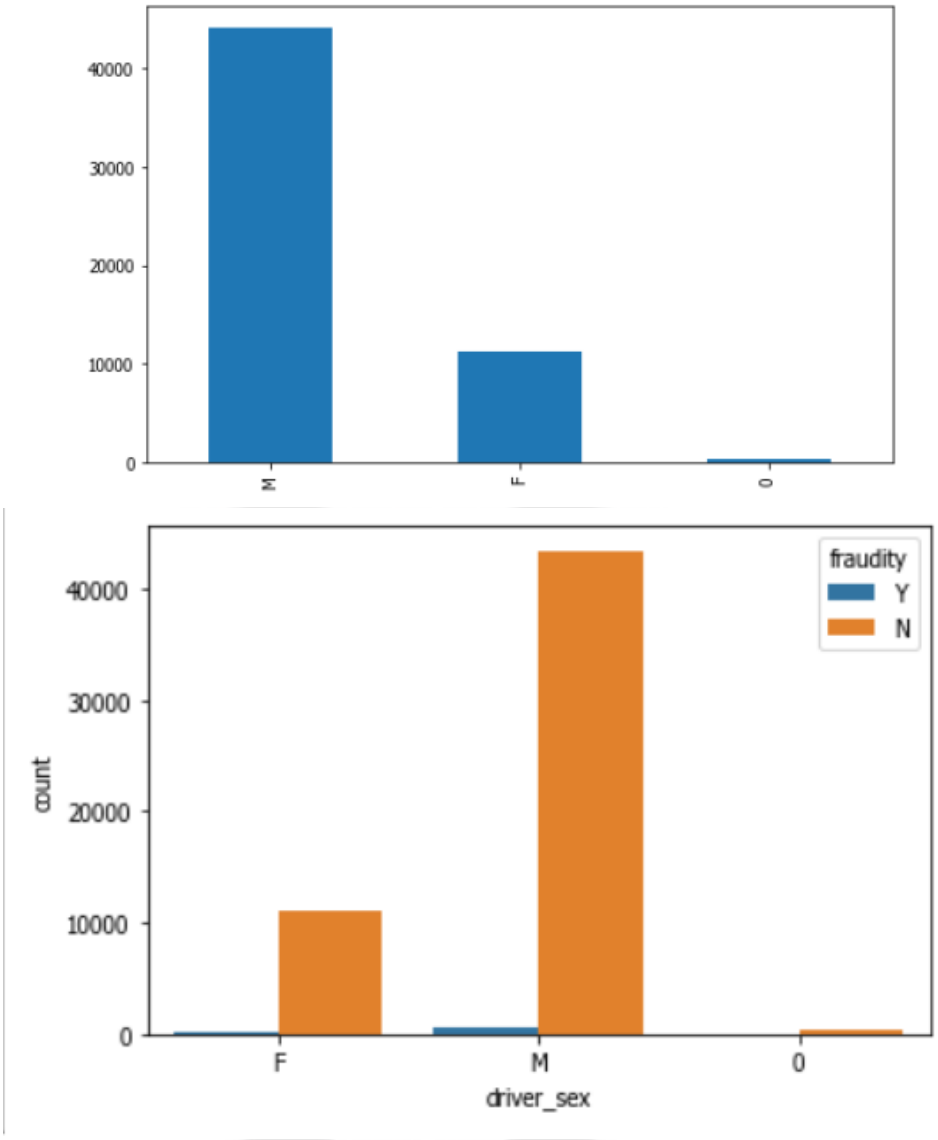
ภาพประกอบ 23 แสดงข้อมูลปริมาณเคลมแยกตามประเภทรถยนต์

มองให้ลึกลงไปในประเภทรถยนต์ โดยดูว่าประเภทไหนเกิดการทุจริตเคลมเยอะที่สุด ดูจากสัดส่วนแล้วรถยนต์ประเภท เก๋ง2ตอน และกระบะบรรทุก มีการทุจริตมากเป็นอันดับแรก



ภาพประกอบ 24 แสดงข้อมูลปริมาณเคลมที่มีการทุจริตและไม่ทุจริตแยกตามประเภทรถยนต์

นับจำนวนการเกิดเคลมของผู้ขับขี่ที่เป็นเพศ ชาย , หญิง , อื่นๆ ปริมาณการเกิดเคลมโดยผู้ขับขี่ที่เป็นผู้ชายมีปริมาณมากกว่าผู้หญิง และ ผู้ขับขี่ที่เป็นผู้ชายมีการทุจริตเคลมมากกว่าผู้หญิงแต่ถ้าเทียบสัดส่วนแล้วจะเห็นว่าผู้ขับขี่ที่เป็นผู้หญิงจะมีโอกาสที่เกิดการทุจริตเคลมมากกว่าผู้ชาย



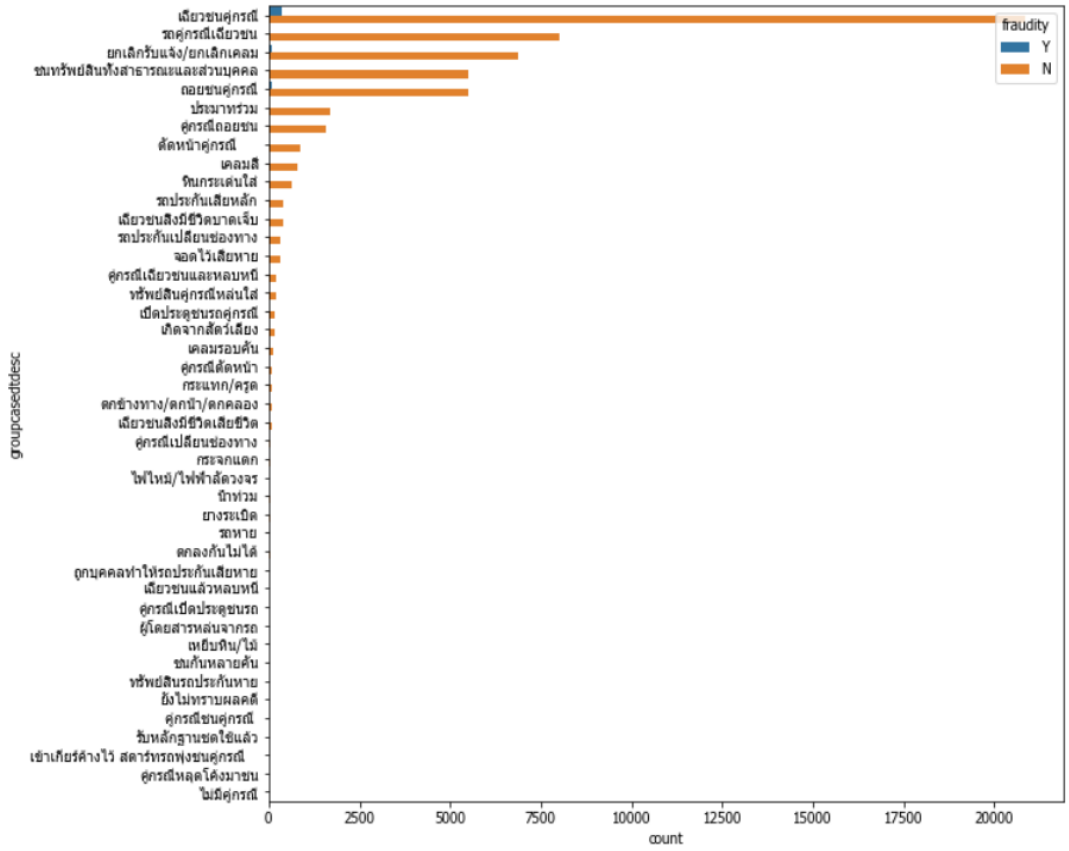
ภาพประกอบ 25 แสดงข้อมูลปริมาณเคลมที่มีการทุจริตและไม่ทุจริตแยกตามเพศของผู้ขับขี่

จากนั้นลองมาดูการเกิดอุบัติเหตุจากสาเหตุการเกิดอุบัติเหตุ จะสังเกตเห็นว่าการเกิดอุบัติเหตุส่วนใหญ่เป็นการเฉี่ยวชนคู่กรณี , รถคู่กรณีเฉี่ยวชน , ยกเลิกรับแจ้ง/ยกเลิกเคลม , ถอยชนคู่กรณี , ชนทรัพย์สินทั้งสาธารณะและคู่กรณี เป็นอันดับต้นๆ

```
เฉี่ยวชนคู่กรณี                                21226
รถคู่กรณีเฉี่ยวชน                               8009
ยกเลิกรับแจ้ง/ยกเลิกเคลม                        6947
ถอยชนคู่กรณี                                    5599
ชนทรัพย์สินทั้งสาธารณะและส่วนบุคคล              5531
ประมาทร่วม                                      1696
คู่กรณีถอยชน                                    1572
ตัดหน้าคู่กรณี                                  894
เคลมสี                                           786
หินกระเด็นใส่                                   654
รถประกันเสียหลัก                                418
เฉี่ยวชนสิ่งมีชีวิตบาดเจ็บ                      400
รถประกันเปลี่ยนช่องทาง                          336
จอดไว้เสียหาย                                   328
คู่กรณีเฉี่ยวชนและหลบหนี                        210
ทรัพย์สินคู่กรณีหล่นใส่                         199
เปิดประตูชนรถคู่กรณี                            168
เกิดจากสัตว์เลี้ยง                              151
เคลมรอบคัน                                      134
คู่กรณีตัดหน้า                                  101
กระแทก/ครูด                                     83
ตกข้างทาง/ตกน้ำ/ตกคลอง                          77
เฉี่ยวชนสิ่งมีชีวิตเสียชีวิต                    67
คู่กรณีเปลี่ยนช่องทาง                           65
กระจกแตก                                        56
ไฟไหม้/ไฟฟ้าลัดวงจร                             38
น้ำท่วม                                         38
ยางระเบิด                                       36
รถหาย                                           34
ตกลงกันไม่ได้                                   29
ถูกบุคคลทำให้รถประกันเสียหาย                    20
คู่กรณีเปิดประตูชนรถ                            20
เฉี่ยวชนแล้วหลบหนี                              18
ผู้โดยสารหล่นจากรถ                              10
เหยียบหิน/ไม้                                   9
ชนกันหลายคัน                                    7
ทรัพย์สินรถประกันหาย                            3
ยังไม่ทราบผลคดี                                 3
คู่กรณีชนคู่กรณี                                3
รับหลักฐานชดใช้แล้ว                             2
เข้าเกียร์ค้างไว้ สตาร์ทรถพุ่งชนคู่กรณี          1
คู่กรณีหลุดโค้งมาชน                             1
ไม่มีคู่กรณี                                    1
Name: groupcasedtdesc, dtype: int64
```

ภาพประกอบ 26 แสดงข้อมูลปริมาณเคลมที่แยกตามสาเหตุการเกิดเหตุ

โดยพอลองมาดูต่อโดยการแยกประเภทว่าสาเหตุการเกิดเหตุไหนมีการทุจริตเคลมบ้าง เราจะเห็นว่าเฉี่ยวชนคู่กรณี , ยกเลิกรับแจ้ง/ยกเลิกเคลม , ถอยชนคู่กรณี จะมีการทุจริตการเคลมในสาเหตุการเกิดเหตุเหล่านี้

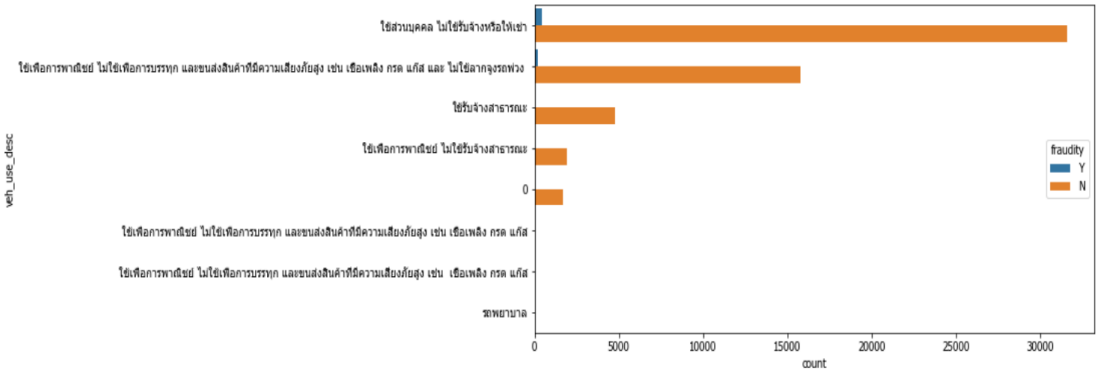


ภาพประกอบ 27 แสดงข้อมูลปริมาณเคลมที่มีการทุจริตและไม่ทุจริตแยกตามสาเหตุการเกิดเหตุ

จากนั้นมาดูการเคลมของประเภทการใช้รถยนต์ว่าจำนวนเคลมที่มากที่สุดเกิดจากประเภทการใช้รถยนต์แบบไหนจะเห็นว่าประเภทใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า และ ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส และ ไม่ใช้ลากจูงรถพ่วง มีการเคลมสูงเป็นอันดับต้นๆของการเคลมทั้งหมด

โดย ประเภทใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า และ ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส และ ไม่ใช้ลากจูงรถพ่วง มีการเคลมสูง จะมีปริมาณการทุจริตเคลมมากที่สุด

ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า	32009
ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส และ ไม่ใช้ลากจูงรถพ่วง	15950
ใช้รับจ้างสาธารณะ	4798
ใช้เพื่อการพาณิชย์ ไม่ใช้รับจ้างสาธารณะ	1973
0	1714
ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส	31
ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส	17
รถพยาบาล	2

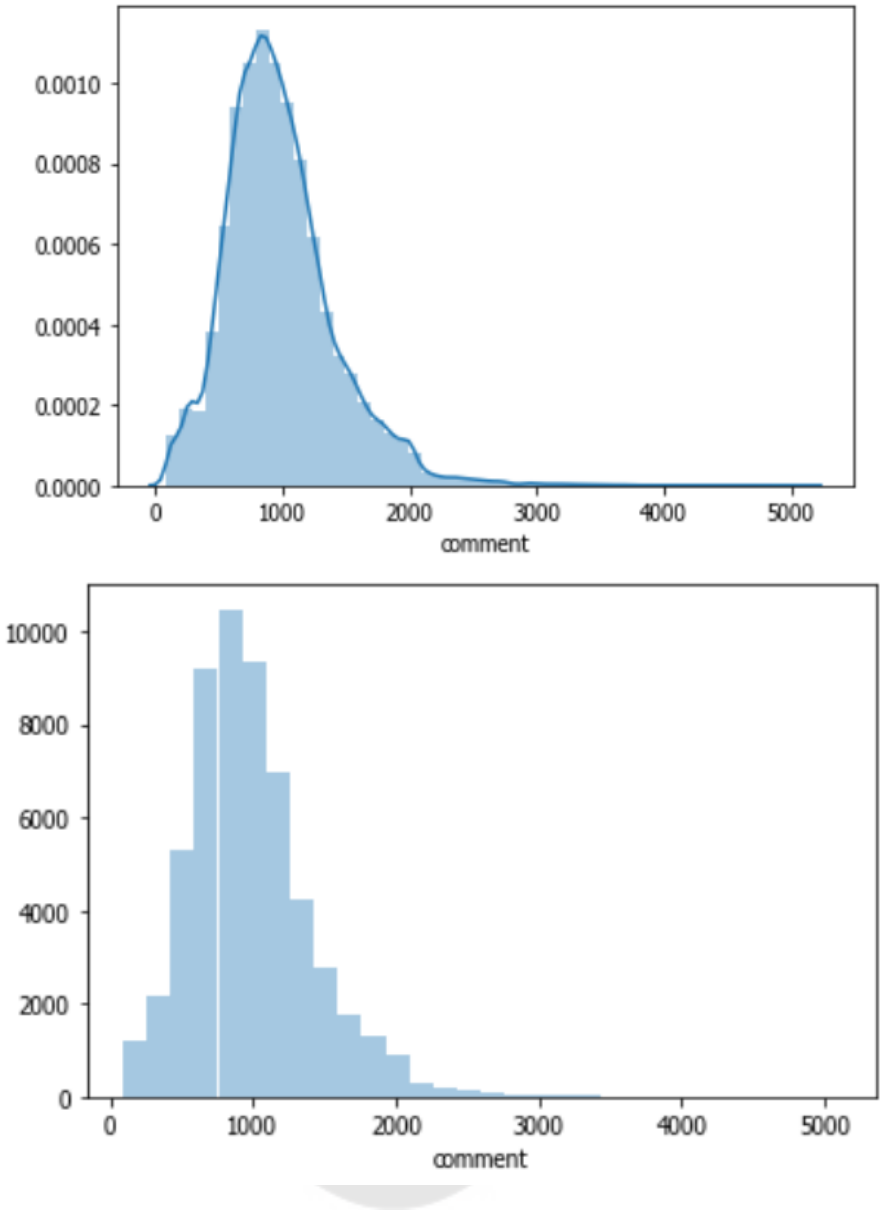


ภาพประกอบ 28 แสดงข้อมูลปริมาณเคลมที่มีการทุจริตและไม่ทุจริตแยกตามประเภทการใช้รถยนต์

มาดูรายงานสำรวจภัย(comment) ของเจ้าหน้าที่สำรวจภัยรายงานความเสียหายและรายงานผลการเกิดอุบุติเหตุว่ามีกี่ตัวอักษร และปริมาณตัวอักษรอยู่ที่เท่าไหร่บ้าง ปรากฏว่าจำนวนตัวอักษรของรายงานสำรวจภัย(comment) จะเห็นได้ว่าค่าเฉลี่ย(mean) ของตัวอักษรในรายงานสำรวจภัยอยู่ที่ 978 ตัวอักษร

	datacase	ageofvehicle	tday_apart	eday_apart	CommentLength
count	57220.000000	57220.000000	57220	57213	57162.000000
mean	239.935687	10.221059	203 days 06:49:05.934987	266 days 17:27:35.309108	978.608341
std	288.901425	7.857184	216 days 22:31:16.678487	1807 days 22:01:06.762088	431.392881
min	100.000000	1.000000	-12 days +00:00:00	-31 days +00:00:00	87.000000
25%	100.000000	4.000000	103 days 00:00:00	94 days 00:00:00	691.000000
50%	100.000000	8.000000	206 days 00:00:00	197 days 00:00:00	921.000000
75%	200.000000	14.000000	301 days 00:00:00	291 days 00:00:00	1192.000000
max	989.000000	78.000000	44074 days 00:00:00	44183 days 00:00:00	5103.000000

ภาพประกอบ 29 แสดงข้อมูลจำนวนตัวอักษรเป็นค่าทางสถิติ



ภาพประกอบ 30 แสดงข้อมูลจำนวนตัวอักษรของรายงานสำรวจภัย

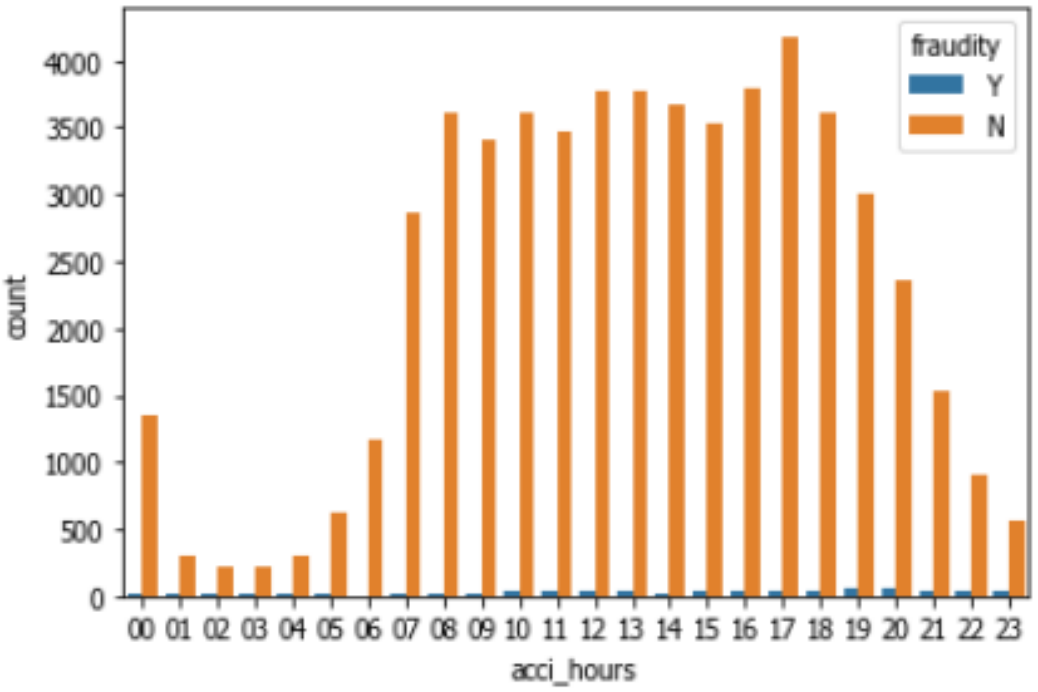
มาสำรวจข้อมูลการเกิดอุบัติเหตุตามชั่วโมงในแต่ละวันดูว่าความถี่และปริมาณการเกิดเหตุในช่วงเวลาไหนที่มีการเกิดเหตุมากที่สุด

จะเห็นว่าปริมาณการเกิดอุบัติเหตุอยู่ในช่วงเวลาตั้งแต่ 08:00 – 20:00 น. ซึ่งเป็นเวลาที่คนเดินทางและใช้รถยนต์กันเป็นปริมาณมาก แต่จะมีปริมาณเยอะจริงๆอยู่ในช่วงเวลาเร่งด่วน 08:00 – 10:00 น. และ 16:00 – 19:00 น.

```
00    1375
01     307
02     225
03     238
04     310
05     628
06    1174
07    2866
08    3630
09    3430
10    3651
11    3503
12    3805
13    3821
14    3702
15    3579
16    3821
17    4227
18    3650
19    3058
20    2413
21    1567
22     931
23     584
Name: acci_hours, dtype: int64
```

```
Y
```

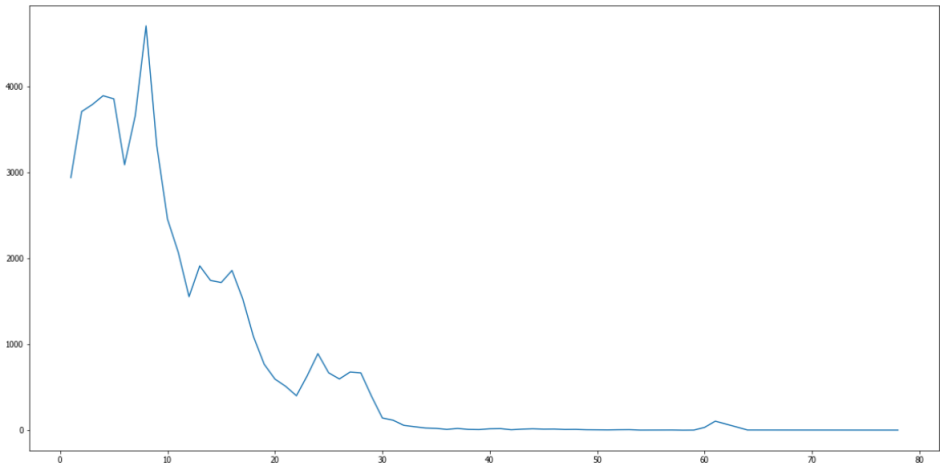
```
00    26
01    12
02    12
03    22
04    11
05    15
06     4
07     8
08    18
09    26
10    34
11    42
12    30
13    43
14    25
15    40
16    28
17    46
18    41
19    49
20    50
21    42
22    31
23    27
```

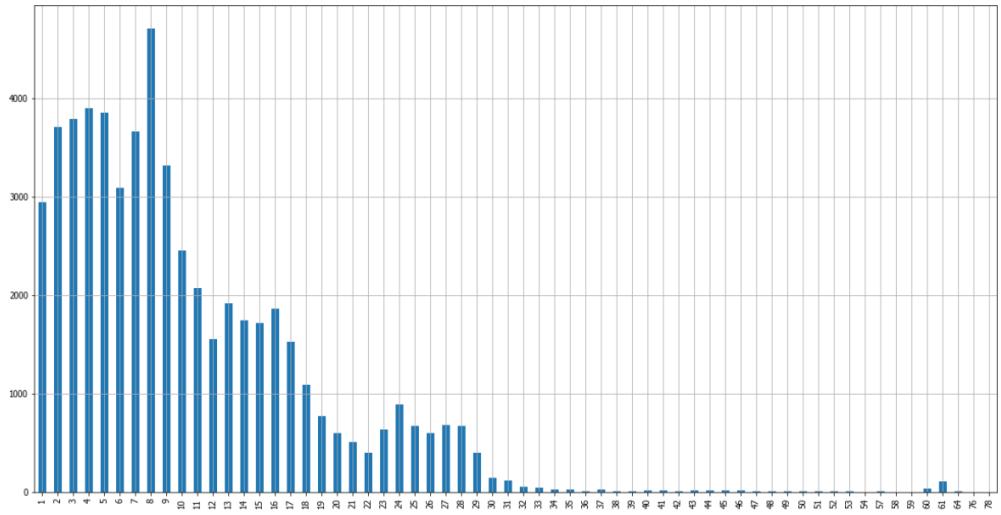


ภาพประกอบ 31 แสดงข้อมูลการเกิดอุบัติเหตุตามชั่วโมง

ข้อมูลการเคลมประกันรถยนต์โดยดูจากอายุรถยนต์ของการเกิดเหตุของรถประกัน สังเกตได้ว่ารถยนต์ที่มีอายุรถอยู่ในช่วง 1-9 ปีมีการเคลมที่ค่อนข้างสูงยิ่งเป็นช่วงอายุรถอยู่ที่ 8-9 ปีจะมีปริมาณการเคลมสูงที่สุด

```
1     2942
2     3711
3     3793
4     3896
5     3859
6     3092
7     3668
8     4709
9     3314
10    2458
11    2072
12    1556
13    1914
14    1745
15    1720
16    1861
17    1530
18    1091
19     769
20     596
21     511
22     402
23     633
24     892
25     669
26     597
27     678
28     668
29     394
30     143
Name: ageofvehicle, dtype: int64
```





ภาพประกอบ 32 แสดงข้อมูลการเกิดอุบัติเหตุแบ่งตามอายุรถยนต์

หลังจากที่ผ่านขั้นตอนการทำ Pre-processing ของคุณลักษณะที่ไม่ใช้ข้อความเรียบร้อยแล้วข้อมูลรายการเคลมประกันจะเหลือทั้งหมด 56,495 รายการโดยแบ่งเป็นข้อมูลที่ไม่ทุจริตจำนวน 55,817 รายการ และ ข้อมูลที่ทุจริต 678 รายการ

5. **สร้างคุณลักษณะของข้อมูลข้อความ**(Feature Extraction)

ทำการเปลี่ยนคุณลักษณะของข้อมูลที่ไม่ใช่ข้อความให้อยู่ในรูปแบบที่สามารถนำไปใช้งานกับ Machine Learning โดยการทำ One-Hot Encoding เปลี่ยนข้อมูลจาก Categories ให้กลายเป็น Feature ที่เป็นตัวเลขโดยจะทำกับ Feature body_desc , driver_sex , groupcasedtdesc , datacase_desc , veh_use_desc เป็นการเสร็จขั้นตอนของการเตรียมข้อมูลในส่วนของ Feature ที่ไม่ใช่ข้อความ แล้วทำการ Join ข้อมูลที่ทำ One-Hot Encoding เข้ากับข้อมูลที่ใช้ในการทดลอง จากนั้นจะทำการ Drop Column ที่นำไปทำ One-Hot Encoding แล้วออกไป

	body_desc_0	body_desc_กระบะบรรทุก	body_desc_จักรยานยนต์	body_desc_ดั๊มพ์สิบล้อ	body_desc_ตู้บรรทุก	body_desc_นั่งสองตอน	body_desc_นั่งสองตอนท้ายบรรทุก	body_desc_นั่งสองตอนสองแถว	body_desc_นั่งสองตอนแวน	b
0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	
2	0	1	0	0	0	0	0	0	0	
3	0	1	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	
56490	0	0	0	0	0	0	0	0	0	
56491	0	0	0	0	0	0	0	0	0	
56492	0	0	0	0	0	0	0	0	0	
56493	0	0	0	0	0	0	0	0	0	
56494	0	0	0	0	0	0	0	0	0	

56495 rows × 86 columns

	tday_apart	ageofvehicle	acci_hours	comment	fraudity	body_desc_0	body_desc_กระบะบรรทุก	body_desc_จักรยานยนต์	body_desc_ดั๊มพ์สิบล้อ	body_desc_ตู้บรรทุก	...
0	171	11	08	AA6112/1818 วันที่ 19/12/61 นัดหมาย เรียน หัวห...	Y	0	0	0	0	0	...
1	171	11	08	****แก้ไข**** เพิ่มเติม ...	Y	0	0	0	0	0	...
2	77	10	02	รับแจ้ง62052466 AA6305/0072 วันที่02/05/...	Y	0	1	0	0	0	...
3	77	10	02	AA6205/1021 วันที่11/05/2562 เวลา08.00น.นัดหม...	Y	0	1	0	0	0	...
4	209	12	17	รายละเอียดการเกิดอุบัติเหตุ ทางบริษัทฯได้ร...	Y	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...
56490	119	1	08	AN 6312/0567 (เคลมพรบ.นัดหมายสภ.ศรีมโหสถ วันที่...	N	0	0	0	0	0	...
56491	138	1	15	AN6306/0653 63060729 วันที่ 20/6/63 เคลมน...	N	0	0	0	0	0	...
56492	299	1	13	เมื่อวันที่ 11 มกราคม 2563 เวลา 14.05น. ทางบริ...	N	0	0	0	0	0	...
56493	105	1	10	วันที่ 15 เดือนตุลาคม 2563 เวลา 10.10 น. ทางมร...	N	0	0	0	0	0	...
56494	324	1	11	ออกตรวจสอบรถตามรับแจ้งพนักงานเดินทางไปถึงที่สถ...	N	0	0	0	0	0	...

56495 rows × 91 columns

ภาพประกอบ 33 แสดงข้อมูลหลังจากการทำ One-Hot Encoding ของ Feature ที่ไม่ใช่ข้อความ

6. **การเตรียมข้อมูลคุณลักษณะที่เป็นข้อความ** (Text Pre-preprocessing) **และ การสำรวจข้อมูลคุณลักษณะที่เป็นข้อความ** (Exploratory Text Analysis)

การเตรียมข้อมูลของข้อความเราจะใช้ Library ของ PythaiNLP และ NLTK มาช่วยจัดการกับข้อมูลที่เป็นข้อความโดยมีวิธีการดังนี้

ในงานวิจัยนี้เราจำทำการสำรวจข้อมูลโดยแบ่งออกเป็นข้อมูลออกเป็นสองชุดคือข้อมูลที่มีการทุจริต และ ข้อมูลไม่มีการทุจริต โดยจะทดลองทำความสะอาดข้อมูลในแต่ละชุดข้อมูลเพื่อจะทำการสำรวจข้อมูลแยกกัน

- เริ่มจากการเลือกข้อมูลที่เป็นการทุจริตเคลม fraudity = “Y” โดยมีข้อมูลทั้งหมด 678 รายการข้อความ หลังจากนั้นจะทำการทำความสะอาดข้อมูลโดยจะมีขั้นตอนดังนี้
- เราจะใช้ Corpus thai_stopwords() ของ PythaiNLP มาใช้และทำการเพิ่มเติม stopwords บางส่วนเข้าในขั้นตอนการทำ stopwords removal
- จากนั้นจะทำการลบข้อมูลตัวอักษรที่เป็นภาษาอังกฤษ , ตัวเลข , อักขระต่างๆที่ไม่ได้ใช้ออกไป จนเหลือแต่ตัวอักษรที่เป็นคำภาษาไทย
- จากนั้นจะทำการจัดการกับการพิมพ์ข้อความที่เรียงผิดหรือใช้ผิดอักษร (Normalize)
- ทำการตัดคำด้วย word_tokenize engine = “newmm”
- และทำ Stopwords Removal เพื่อลดคำที่ไม่ได้ใช้ออกไป
- ทำการเลือกคำที่ตัดออกมาแล้วมีตัวอักษรที่มากกว่า 2 ตัวอักษรมาดำเนินการ

หลังจากที่ผ่านการทำ Pre-processing ของข้อความที่เป็นการจุจริตเคลมแล้วจะมานับคำทั้งหมดเพื่อดูว่ามีปริมาณคำทั้งหมดเท่าไหร่ โดยเราจะตรวจสอบดูคำทั้งหมดและคำที่ผ่านการทำการ Unique ของคำมาแล้วได้จำนวนคือ คำที่ผ่านการทำ Pre-processing ก่อนทำการ Unique คำเท่ากับ 87,701 คำ และ คำที่ทำการ Unique แล้วจะเหลือคำทั้งสิ้น 4,486 คำ

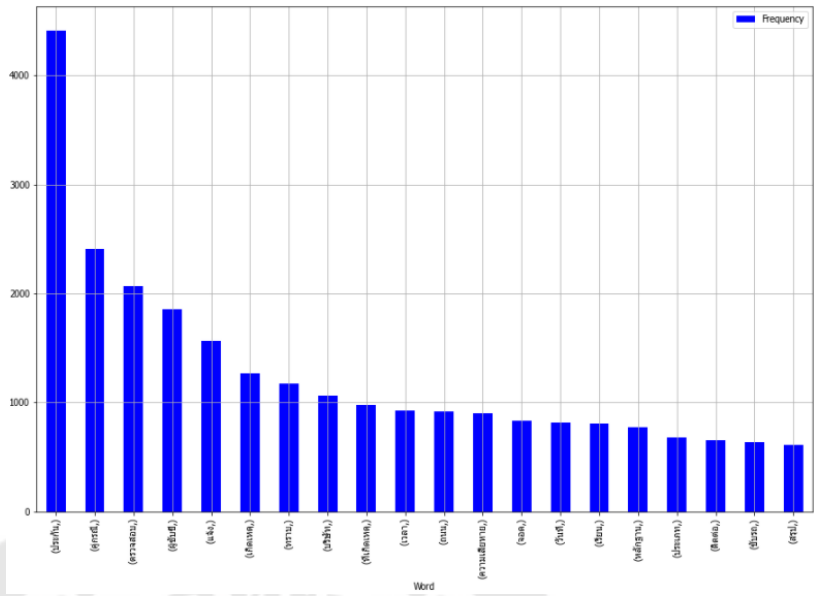
```
1 print("All word in comment: " , len(all_comment)) #จำนวนคำทั้งหมดของ Comment ที่เป็น fraud
2 print("Unique word in comment: " , len(wordfraud_unique))
```

```
All word in comment:  87701
Unique word in comment:  4486
```

ภาพประกอบ 34 แสดงจำนวนคำของชุดข้อมูลทุจริตเคลม

จากนั้นจะทำการทดสอบการตัดคำแบบ N-gram ของ NLTK Library โดยเริ่มจากการทดสอบตัดคำแบบ Uni-gram(1 คำ) และทำการหาความถี่ของคำที่มากที่สุด(Frequency) ผลลัพธ์ที่ได้จะเห็นว่าคำที่มีความถี่มากที่สุดของข้อความที่มีการทุจริตทั้งหมด คือ “ประกัน” เกิดขึ้นทั้งหมด 4,412 ครั้ง รองลงมาคือคำว่า “คู่กรณี” เกิดขึ้นทั้งหมด 2,407 ครั้ง

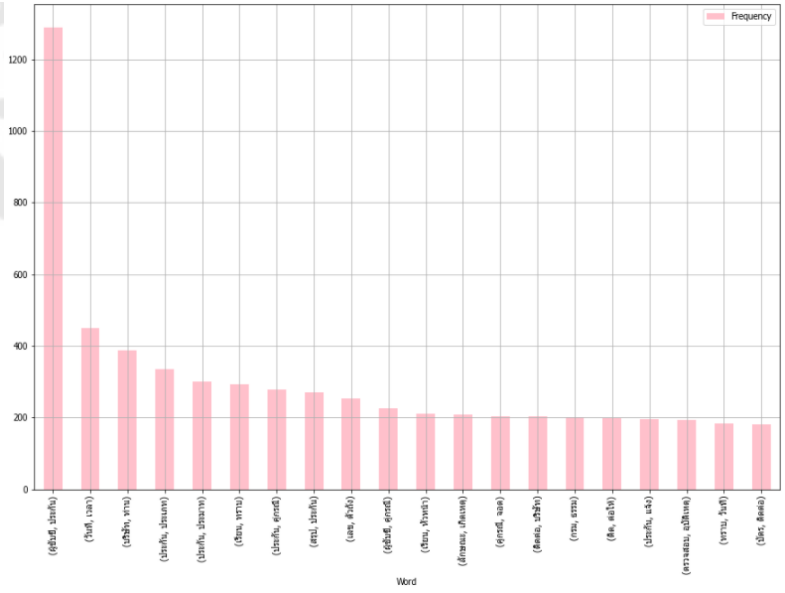
	Word	Frequency
0	(ประกัน,)	4412
1	(คู่กรณี,)	2407
2	(ตรวจสอบ,)	2068
3	(ผู้ขับขี่,)	1853
4	(แจ้ง,)	1564
5	(เกิดเหตุ,)	1270
6	(ทราบ,)	1173
7	(บริษัท,)	1062
8	(ที่เกิดเหตุ,)	973
9	(เวลา,)	921
10	(ถนน,)	919



ภาพประกอบ 35 แสดงความถี่ของคำ Uni-gram ของข้อมูลที่มีการทุจริต

จากนั้นจะทำการทดสอบการตัดคำแบบ Bi-gram(2 คำ) และทำการหาความถี่ของคำที่มากที่สุด(Frequency) ผลลัพธ์ที่ได้จะเห็นว่าคำที่มีความถี่มากที่สุดของข้อความที่มีการทุจริตทั้งหมด คือ “ผู้ขับขี่, ประกัน” เกิดขึ้น 1,289 ครั้ง รองลงมาเป็นคำว่า “วันที่, เวลา” เกิดขึ้น 450 ครั้ง

	Word	Frequency
0	(ผู้ขับขี่, ประกัน)	1289
1	(วันที่, เวลา)	450
2	(บริษัท, ท่าน)	388
3	(ประกัน, ประเภท)	336
4	(ประกัน, ประมาท)	301
5	(เรียน, ทราบ)	292
6	(ประกัน, คู่กรณี)	277
7	(สรุป, ประกัน)	270
8	(เลข, ตัวถัง)	252
9	(ผู้ขับขี่, คู่กรณี)	225
10	(เรียน, หัวหน้า)	212



ภาพประกอบ 36 แสดงความถี่ของคำ Bi-gram ของข้อมูลที่มีการทุจริต

จากนั้นจะเลือกข้อมูลที่ไม่เป็นการทุจริตเคลม fraudity = “N” โดยมีข้อมูลทั้งหมด 55,817 รายการข้อความ หลังจากนั้นจะทำการทำความสะอาดข้อมูลโดยจะมีขั้นตอนดังนี้

- เราจะใช้ Corpus thai_stopwords() ของ PythaiNLP มาใช้และทำการเพิ่มเติม stopwords บางส่วนเข้าในขั้นตอนการทำ stopwords removal
- จากนั้นจะทำการลบข้อมูลตัวอักษรที่เป็นภาษาอังกฤษ , ตัวเลข , อักขระต่างๆที่ไม่ได้ใช้ออกไป จนเหลือแต่ตัวอักษรที่เป็นคำภาษาไทย
- จากนั้นจะทำการจัดการกับการพิมพ์ข้อความที่เรียงผิดหรือใช้ผิดอักษร (Normalize)
- ทำการตัดคำด้วย word_tokenize engine = “newmm”
- และทำ Stopwords Removal เพื่อลดคำที่ไม่ได้ใช้ออกไป
- ทำการเลือกคำที่ตัดออกมาแล้วมีตัวอักษรที่มากกว่า 2 ตัวอักษรมาดำเนินการ

หลังจากที่ผ่านการทำ Pre-processing ของข้อความที่ไม่เป็นการจุจริตเคลมแล้วจะมานับคำทั้งหมดเพื่อดูว่ามีปริมาณคำทั้งหมดเท่าไหร่ โดยเราจะตรวจสอบดูคำทั้งหมดและคำที่ผ่านการทำการ Unique ของคำ จำนวนคำที่ผ่านการทำ Pre-processing ก่อนทำการ Unique คำเท่ากับ 5,070,310 คำ และ คำที่ทำการ Unique แล้วจะเหลือคำทั้งสิ้น 20,874 คำ

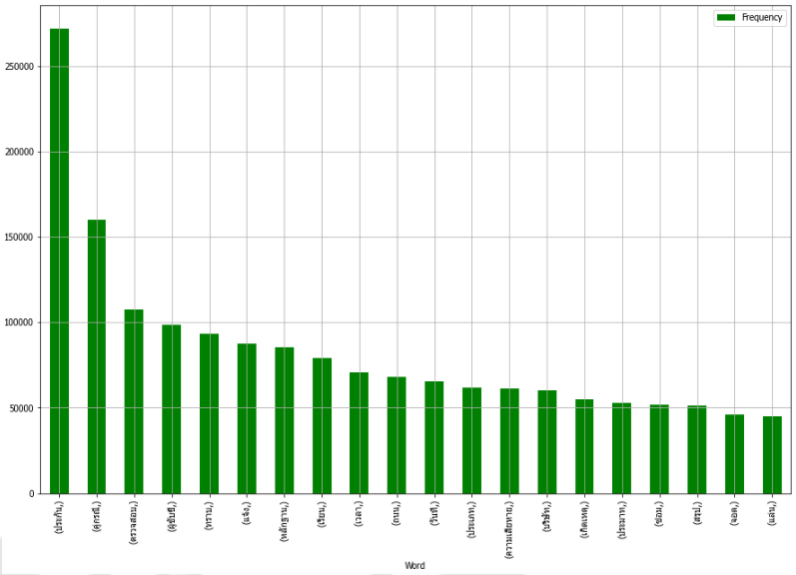
```
1 print("All word in comment: " , len(all_comment)) #จำนวนคำทั้งหมดของ Comment ที่ไม่เป็น fraud
2 print("Unique word in comment: " , len(wordfraud_unique))
```

```
All word in comment:  5070310
Unique word in comment:  20874
```

ภาพประกอบ 37 แสดงจำนวนคำของชุดข้อมูลที่ไม่ทุจริตเคลม

จากนั้นจะทำการทดสอบการตัดคำแบบ N-gram ของ NLTK Library โดยเริ่มจากการทดสอบตัดคำแบบ Uni-gram(1 คำ) และทำการหาความถี่ของคำที่มากที่สุด(Frequency) ผลลัพธ์ที่ได้จะเห็นว่าคำที่มีความถี่มากที่สุดของข้อความที่มีการทุจริตทั้งหมด คือ “ประกัน” เกิดขึ้นทั้งหมด 271,922 ครั้ง รองลงมาคือคำว่า “คู่กรณี” เกิดขึ้นทั้งหมด 159,787 ครั้ง

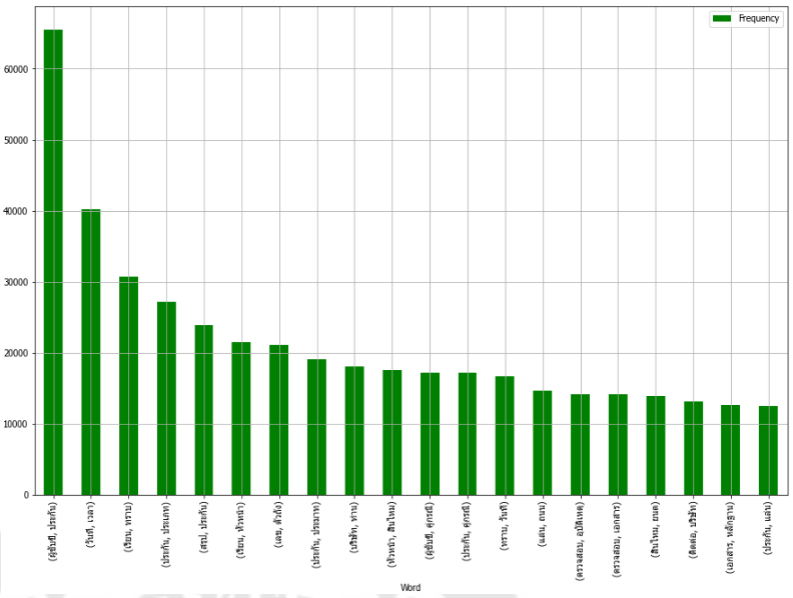
	Word	Frequency
0	(ประกัน,)	271922
1	(คุ้กรณี,)	159787
2	(ตรวจสอบ,)	107387
3	(ผู้ขับขี่,)	98481
4	(ทราบ,)	93394
5	(แจ้ง,)	87309
6	(หลักฐาน,)	85365
7	(เรียน,)	78931
8	(เวลา,)	70470
9	(ถนน,)	67939
10	(วันที่,)	65407



ภาพประกอบ 38 แสดงความถี่ของคำ Uni-gram ของข้อมูลที่มีไม่มีการทุจริต

จากนั้นจะทำการทดสอบการตัดคำแบบ Bi-gram(2 คำ) และทำการหาความถี่ของคำที่มากที่สุด(Frequency) ผลลัพธ์ที่ได้จะเห็นว่าคำที่มีความถี่มากที่สุดของข้อความที่ไม่มีการทุจริตทั้งหมด คือ “ผู้ขับขี่, ประกัน” เกิดขึ้น 65,508 ครั้ง รองลงมาเป็นคำว่า “วันที่, เวลา” เกิดขึ้น 40,251 ครั้ง

	Word	Frequency
0	(ผู้ขับขี่, ประกัน)	65508
1	(วันที่, เวลา)	40251
2	(เรียน, ทราบ)	30767
3	(ประกัน, ประเภท)	27219
4	(สรุป, ประกัน)	23862
5	(เรียน, หัวหน้า)	21440
6	(เลข, ตัวถัง)	21053
7	(ประกัน, ประมาท)	19084
8	(บริษัท, ท่าน)	18003
9	(หัวหน้า, สินไหม)	17486
10	(ผู้ขับขี่, คู่กรณี)	17171



ภาพประกอบ 39 แสดงความถี่ของคำ Bi-gram ของข้อมูลที่ไม่มีการทุจริต

จากนั้นจะทำการทดสอบโดยนำข้อความทั้งหมดมารวมกันเพื่อสร้าง Bag of word(Dictionary) ของทุกๆ รายงานสำรวจภัย(comment) จำนวนทั้งสิ้น 56,495 รายการข้อความ หลังจากนั้นจะทำการทำความสะอาดข้อมูลโดยจะมีขั้นตอนดังนี้

- เราจะใช้ Corpus thai_stopwords() ของ PythaiNLP มาใช้และทำการเพิ่มเติม stopwords บางส่วนเข้าในขั้นตอนการทำ stopwords removal
- จากนั้นจะทำการลบข้อมูลตัวอักษรที่เป็นภาษาอังกฤษ , ตัวเลข , อักขระต่างๆที่ไม่ได้ใช้ออกไป จนเหลือแต่ตัวอักษรที่เป็นคำภาษาไทย
- จากนั้นจะทำการจัดการกับการพิมพ์ข้อความที่เรียงผิดหรือใช้ผิดอักษร (Normalize)
- ทำการตัดคำด้วย word_tokenize engine = “newmm”
- และทำ Stopwords Removal เพื่อลดคำที่ไม่ได้ใช้ออกไป
- ทำการเลือกคำที่ตัดออกมาแล้วมีตัวอักษรที่มากกว่า 2 ตัวอักษรมาดำเนินการ

หลังจากที่ผ่านการทำ Pre-processing ของข้อความทั้งหมดแล้วจะมานับคำทั้งหมดเพื่อดูว่ามีปริมาณคำทั้งหมดเท่าไหร่ โดยเราจะตรวจสอบดูคำทั้งหมดและคำที่ผ่านการทำการ Unique ของคำมาแล้วได้จำนวนคือ คำที่ผ่านการทำ Pre-processing ก่อนทำการ Unique คำเท่ากับ 5,158,011 คำ และ คำที่ทำการ Unique แล้วจะเหลือคำทั้งสิ้น 21,182 คำ

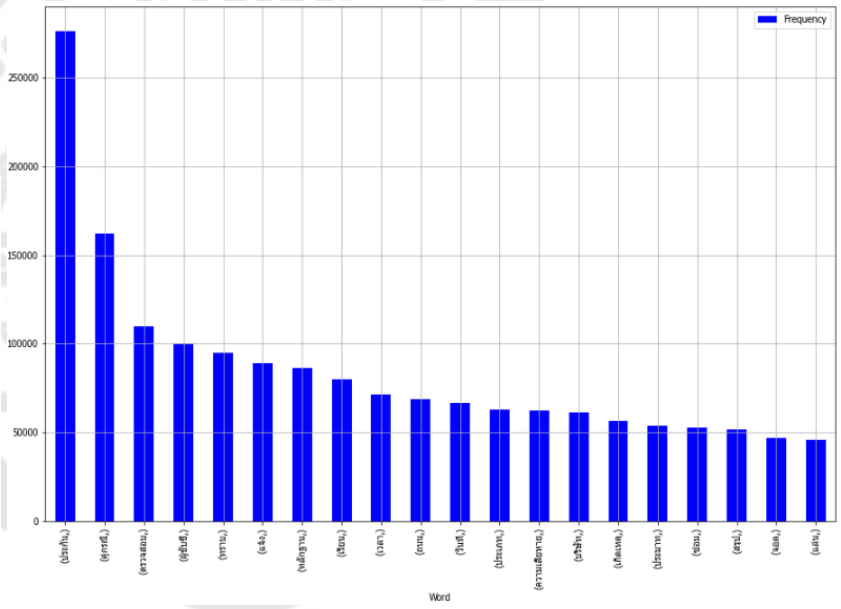
```
print("All word in comment: " , len(all_comment)) #จำนวนคำทั้งหมดของ Comment
print("Unique word in comment: " , len(wordfraud_unique)) # Unique ของคำ Unigram
```

```
All word in comment:  5158011
Unique word in comment:  21182
```

ภาพประกอบ 40 แสดงจำนวนคำของชุดข้อมูลทั้งหมด

จากนั้นจะทำการทดสอบการตัดคำแบบ N-gram ของ NLTK Library โดยเริ่มจากการทดสอบตัดคำแบบ Uni-gram(1 คำ) และทำการหาความถี่ของคำที่มากที่สุด(Frequency) ผลลัพธ์ที่ได้จะเห็นว่าคำที่มีความถี่มากที่สุดของข้อความทั้งหมด คือ “ประกัน” เกิดขึ้นทั้งหมด 276,334 ครั้ง รองลงมาคือคำว่า “คู่กรณี” เกิดขึ้นทั้งหมด 162,194 ครั้ง

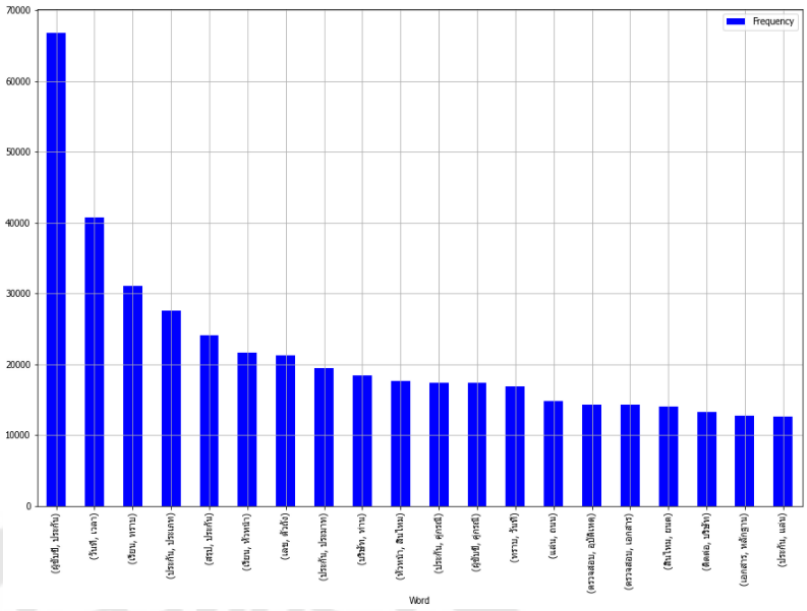
	Word	Frequency
0	(ประกัน,)	276334
1	(คู่กรณี,)	162194
2	(ตรวจสอบ,)	109455
3	(ผู้ขับขี่,)	100334
4	(ทราบ,)	94567
5	(แจ้ง,)	88873
6	(หลักฐาน,)	86136
7	(เรียน,)	79740
8	(เวลา,)	71391
9	(ถนน,)	68858
10	(วันที่,)	66221



ภาพประกอบ 41 แสดงความถี่ของคำ Uni-gram ของข้อมูลทั้งหมด

จากนั้นจะทำการทดสอบการตัดคำแบบ Bi-gram(2 คำ) และทำการหาความถี่ของคำที่มากที่สุด(Frequency) ผลลัพธ์ที่ได้จะเห็นว่าคำที่มีความถี่มากที่สุดของข้อความทั้งหมด คือ “ผู้ขับขี่, ประกัน” เกิดขึ้น 66,797 ครั้ง รองลงมาเป็นคำว่า “วันที่, เวลา” เกิดขึ้น 40,701 ครั้ง

	Word	Frequency
0	(ผู้ขับขี่, ประกัน)	66797
1	(วันที่, เวลา)	40701
2	(เรียน, ทราบ)	31059
3	(ประกัน, ประเภท)	27555
4	(สรุป, ประกัน)	24132
5	(เรียน, หัวหน้า)	21652
6	(เลข, ตัวถัง)	21305
7	(ประกัน, ประมาท)	19385
8	(บริษัท, ท่าน)	18391
9	(หัวหน้า, สินไหม)	17643
10	(ประกัน, คู่กรณี)	17426



ภาพประกอบ 42 แสดงความถี่ของคำ Bi-gram ของข้อมูลทั้งหมด

จับกลุ่มคำโดยเรียงจากคำที่มีมากที่สุดไปน้อยที่สุดโดยใช้ Word Cloud เป็นประโยชน์ในการทำรายงานข้อความ เพื่อให้มองเห็นคำที่ถูกใช้มากที่สุดได้ง่าย

ภาพประกอบ 43 แสดงการจับกลุ่มคำของข้อมูลทั้งหมดโดยเรียงจากคำที่มีมากที่สุดไปน้อยที่สุด โดยใช้ Word Cloud

ภาพประกอบ 44 แสดงการจับกลุ่มคำของข้อมูลที่เป็นการทุจริตเคลมโดยเรียงจากคำที่มีมากที่สุดไปน้อยที่สุด โดยใช้ Word Cloud

ภาพประกอบ 45 แสดงการจับกลุ่มคำของข้อมูลที่ไม่เป็นการทุจริตเคลมโดยเรียงจาก คำที่มีมากที่สุดไปน้อยที่สุด โดยใช้ Word Cloud

จะนำคำทั้งหมดมาสร้าง Bag of Word(BOW) Dictionary ของคำโดยจะได้คำที่ Unique เก็บไว้จำนวนทั้งหมด 21,1 82 คำที่ไม่ซ้ำกัน และจะทำการ write file เป็น csv ไว้เพื่อเรียกใช้ในครั้งต่อไปได้

```
print('Count BOW before remove dupplicate: ' , len(bows_before))
print('Count BOW: ' , len(bows))
```

```
Count BOW before remove dupplicate:  5158011
Count BOW:  21182
```

ภาพประกอบ 46 แสดงจำนวนคำของชุดข้อมูลทั้งหมดที่นำไปสร้าง BOW

7. **สร้างคุณลักษณะของข้อมูลข้อความ**(Feature Extraction Text)

จากนั้นเราจะทำการสร้างคุณลักษณะของข้อความที่ตัดคำมาแล้วให้เป็นตัวเลขโดยเราจะใช้วิธีการ TFIDF เพื่อจะเตรียมข้อมูลก่อนจะนำไปเข้าแบบจำลองเพื่อฝึกสอน

โดยเราจะทดลองกำหนดค่าพารามิเตอร์ TfidfVectorizer เพื่อทดสอบประสิทธิภาพของแบบจำลองโดยเราใช้พารามิเตอร์ use_idf = true , norm = l2 , max_features=1000 ข้อมูลหลังจากทำ TfidfVectorizer ออกมาแล้วเรามาดูค่าที่ vectorizer.fit_transform และ vectorizer.get_feature_names() ออกมา และจากนั้นจะนำ word vector ที่ได้มาทำให้อยู่ในรูปแบบ data frame ดังรูปด้านล่างนี้

```
(0, 975)	0.12158808616505906
(0, 392)	0.12178000291466491
(0, 345)	0.05784792071464038
(0, 112)	0.07276564505820968
(0, 542)	0.08627513745360597
(0, 386)	0.029256218073967544
(0, 79)	0.07746517327204844
(0, 660)	0.02641279447733461
(0, 895)	0.028568759637599275
(0, 375)	0.08832315168767363
(0, 68)	0.13640521973455288
(0, 589)	0.12407713389588629
(0, 740)	0.07949170136508207
(0, 334)	0.08783105044152946
(0, 189)	0.07600597853489933
(0, 962)	0.08505245835963374
```

```
['กทม',
 'กรกฎาคม',
 'กรณี',
 'กรม',
 'กรวย',
 'กระ',
 'กระจก',
 'กระทันหัน',
 'กระทุ่ม',
 'กระบะ',
 'กระเด็น',
 'กระแทก',
 'กระโปรง',
 'กริช',
 'กรุง',
 'กรุงเทพ',
 'กรุงเทพมหานคร',
 'กรุงไทย',
 'กลับบ้าน',
 'กลับมา',
 'กลับรถ',
 'กลาก',
 'กลาง',
 'กลิ่น',
 'กล้อง',
 'กล้องวงจรปิด',
```

ภาพประกอบ 47 แสดงผลลัพธ์ของการทำ TFIDF

	กทม	กรกฎาคม	กรณี	กรม	กรวย	กระ	กระจก	กระทันหัน	กระทุ่ม	กระบะ	...	ไม่	ไฟท้าย	ไฟหน้า	ไฟแดง	ไม้	ไร่	ไลน	ไหม	ไหล	ไหล่ทาง
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.121429	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 1000 columns

ภาพประกอบ 48 แสดงData frame ของ word vector

หลักการที่ใช้ในการพิจารณาเลือกค่าพารามิเตอร์ max_feature = 1000 นั้นจะพิจารณาจากค่า TFIDF ของคำที่มีความสำคัญที่มีค่ามากที่สุดเป็นจำนวณ 1000 คำนำมาใช้เป็นคุณลักษณะในส่วนของข้อความ โดยที่เกณฑ์(threshold) คือเลือกผลรวมค่า TFIDF ของคำแต่ละคำที่มีค่ามากกว่าเท่ากับ 60 ตามที่แสดงในรูปที่ 49 – 50

```
In [23]:
# select numeric columns and calculate the sums
sums = tfidf_sklearn.select_dtypes(pd.np.number).sum().rename("total")
tfidf_sklearn.append(sums)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: FutureWarning: The pandas.np module is deprecated and will be r
emoved from pandas in a future version. Import numpy directly instead
```

Out[23]:

	กกิต	กขกท	กขข	กขนพ	กขสค	กขสด	กขนภ	กงจักร	กจก	กจกท	กจง	กจจ	กจชร	กจด	กจนสส	
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
56491	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
56492	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
56493	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
56494	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000	0.000
total	0.182259	0.332674	0.512415	0.432151	0.288335	0.707768	0.24989	0.600909	0.205067	0.737577	0.163371	0.271284	0.253906	0.1887	0.453152	0.147

56496 rows × 21184 columns

ภาพประกอบ 49 แสดงการรวมผลค่า TFIDF ของแต่ละคำ

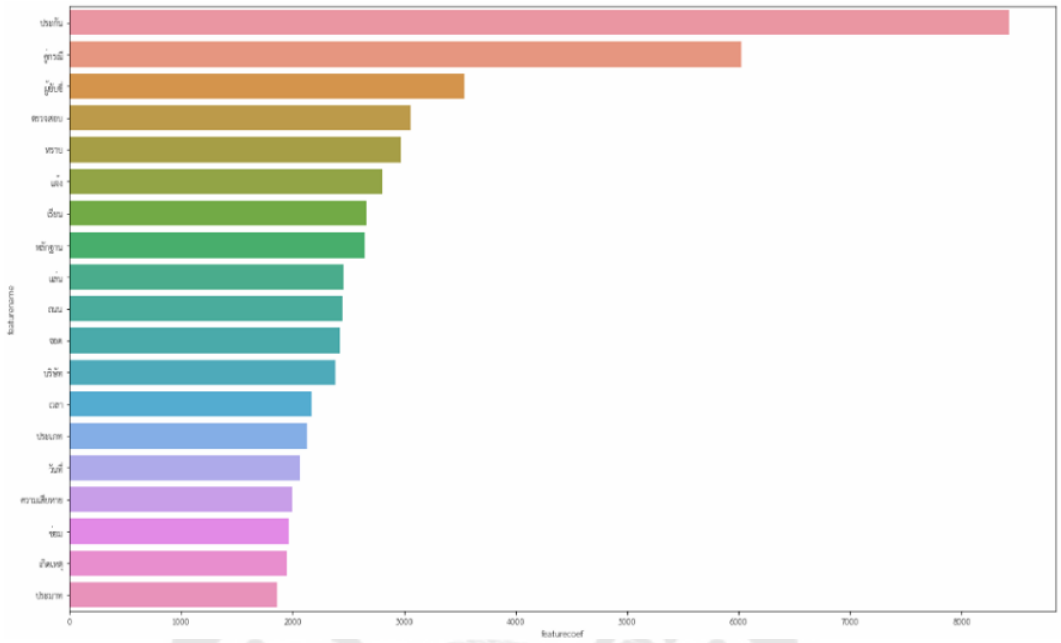
```
In [38]:
condition1 = sorted_df['featurecoef'] >= 60
dfimportance = sorted_df[condition1]
dfimportance
```

Out[38]:

	featurename	featureindex	featurecoef
8100	ประกัน	8100	8425.252930
2456	คู่กรณี	2456	6025.056152
9063	ผู้ขับขี่	9063	3540.320068
4700	ตรวจสอบ	4700	3058.428467
5665	ทราบ	5665	2968.153564
...	...	...	...
3579	ช่าง	3579	60.333607
6189	ทำข้าม	6189	60.265343
5573	ถ่ายภาพ	5573	60.258705
11542	รถติด	11542	60.226601
9521	พระนครศรีอยุธยา	9521	60.168167

959 rows × 3 columns

ภาพประกอบ 50 แสดงคุณลักษณะคำที่มีค่าผลรวม TFIDF ตามเกณฑ์ที่เลือกใช้



ภาพประกอบ 51 แสดงคุณลักษณะคำที่มีค่าผลรวม TFIDF
ตามเกณฑ์ที่เลือกในรูปของ bar chart

8. **รวมคุณลักษณะที่ไม่ใช่ข้อความ และคุณลักษณะของข้อความเข้าด้วยกัน (Concatenate)**

หลังจากที่ทำ Pre-processing ทั้งคุณลักษณะที่ไม่ใช่ข้อความ และคุณลักษณะที่เป็นข้อ และผ่านการทำ One-hot encoding และ word vector เรียบร้อยแล้วเราจะนำคุณลักษณะทั้งหมดมาทำการรวมกันโดยการนำมาต่อกัน(Concatenate) จะได้ข้อมูลที่พร้อมจะทำในขั้นตอนการทำ Feature Scaling และการทำการสุ่มตัวอย่างข้อมูลต่อไป

	tday_apart	ageofvehicle	acci_hours	fraudity	body_desc_0	body_desc_กระบะบรรทุก	body_desc_จักรยานยนต์	body_desc_ดั๊มพ์ล้อ	body_desc_ตู้บรรทุก	body_desc_นั่งสองตอน	...
0	171	11	8	Y	0	0	0	0	0	0	
1	171	11	8	Y	0	0	0	0	0	0	
2	77	10	2	Y	0	1	0	0	0	0	
3	77	10	2	Y	0	1	0	0	0	0	
4	209	12	17	Y	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	
56490	119	1	8	N	0	0	0	0	0	0	
56491	138	1	15	N	0	0	0	0	0	0	
56492	299	1	13	N	0	0	0	0	0	0	
56493	105	1	10	N	0	0	0	0	0	0	
56494	324	1	11	N	0	0	0	0	0	0	

56495 rows × 1091 columns

y_desc_0	body_desc_กระบะบรรทุก	body_desc_จักรยานยนต์	body_desc_ดั๊มพ์สิบล้อ	body_desc_ตู้บรรทุก	body_desc_นั่งสองตอน	...	ไผ่	ไฟท้าย	ไฟหน้า	ไฟแดง	ไม้	ไร่	ไลน	ไหม	ไหล	ไหล่ทาง
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	1	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	1	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	...	0.0	0.0	0.0	0.058254	0.0	0.0	0.0	0.0	0.0	0.0

ภาพประกอบ 52 แสดงการนำข้อมูลที่ไม่ใช่ข้อความและข้อมูลข้อความมารวมกัน

9. **การแบ่งข้อมูลสำหรับการเทรนและทดสอบ** (Train/Test)

จากนั้นจะทำการแบ่งข้อมูลหลังจากที่ได้ทำการรวมข้อมูลคุณลักษณะที่ไม่ใช่ข้อความและคุณลักษณะที่เป็นข้อความเข้าด้วยกันแล้วเพื่อใช้สำหรับการเทรนแบบจำลองและทดสอบแบบจำลองโดยจะแบ่งข้อมูลเทรนออกเป็น 70% และ ทดสอบออกเป็น 30%

10. **ทำการ Scale ข้อมูล**(Feature Scaling)

หลังจากที่ได้ทำการแบ่งข้อมูลที่ใช้ในการฝึกฝนและข้อมูลที่ใช้ในการทดสอบเรียบร้อยแล้ว เราต้องการทำให้ค่าในแต่ละคุณลักษณะอยู่ใน Scale มาตรฐานเดียวกัน จะส่งผลให้แบบจำลองได้ค่าความแม่นยำที่ดีขึ้น ในงานวิจัยนี้เราใช้สูตร StandardScaler() ของ Scikit-Learn โดยเราจะทำการ Scale ข้อมูลที่เป็นคุณลักษณะที่ไม่ใช่ Class Label โดยข้อมูลคุณลักษณะ(Feature) ทั้งหมดที่นำมาใช้จะเหลือเพียง 1,089 คุณลักษณะ กับ 1 คุณลักษณะที่เป็น Class Label โดยที่เราจะทำการ Scale ข้อมูลแยกกันระหว่างข้อมูลที่ใช้ฝึกฝน กับข้อมูลที่ใช้ในการทดสอบ

	tday_apart	ageofvehicle	acci_hours	body_desc_0	body_desc_กระบะ บรรทุก	body_desc_จักรยานยนต์	body_desc_ดั๊มพ์ สิบล้อ	body_desc_ตู้ บรรทุก	body_desc_นั่ง สองตอน
0	-0.253241	0.097194	-1.018141	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
1	-0.253241	0.097194	-1.018141	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
2	-1.098040	-0.029906	-2.213673	-0.220701	1.582060	-0.261114	-0.00595	-0.011132	-0.013306
3	-1.098040	-0.029906	-2.213673	-0.220701	1.582060	-0.261114	-0.00595	-0.011132	-0.013306
4	0.088273	0.224294	0.775158	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
...	...	...	...	...	...	...	...	...	...
56490	-0.720576	-1.173806	-1.018141	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
56491	-0.549819	-1.173806	0.376647	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
56492	0.897123	-1.173806	-0.021864	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
56493	-0.846398	-1.173806	-0.619630	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306
56494	1.121803	-1.173806	-0.420374	-0.220701	-0.632087	-0.261114	-0.00595	-0.011132	-0.013306

56495 rows × 1089 columns

body_desc_ตู้ บรรทุก	body_desc_นั่ง สองตอน	body_desc_นั่ง สองตอนท้าย บรรทุก	...	ไผ่	ไฟท้าย	ไฟหน้า	ไฟแดง	ไม้	ไร่	ไลน	ไหม	ไหล	ไหล่ทาง
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
...	...	...	...	...	...	...	...	...	...	...	...	...	...
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	-0.137258	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715
-0.011132	-0.013306	-0.153409	...	-0.067965	-0.129675	-0.074466	1.867167	-0.101972	-0.070979	-0.139529	-0.14448	-0.157765	-0.093715

ภาพประกอบ 53 แสดงข้อมูลตัวอย่างที่ได้หลังจากทำ Feature Scaling

11. อัลกอริทึมและแบบจำลองที่ใช้ทำนาย ทดลองกับข้อมูลที่มีความไม่สมดุลกัน (Model Algorithm with Imbalance Data)

หลังจากที่เราทำการแบ่งข้อมูลที่ใช้ในการฝึกฝน(Train Data) และใช้ในการทดสอบ(Test Data) เรียบร้อยแล้วมาทำการทดลองกับโมเดลแบบจำลองในแต่ละแบบจำลองที่เราเลือกมาทำวิจัย ร่วมกับการทำ 10-Folds Cross Validation โดยแบบจำลองการจำแนกประเภททั้ง 4 แบบจำลองที่เราใช้ในงานวิจัยมีดังนี้

1. Naïve Bayes
2. Logistic Regression
3. Support Vector Machine
4. Random Forest

โดยเราจะใช้ค่ามาตรฐาน (Default) ของแบบจำลองแต่ละประเภทใช้ในการทดลอง หลังจากนั้นจะนำมาทดสอบกับข้อมูลที่ใช้ทดสอบกับแบบจำลองการจำแนกประเภททั้ง 4 และนำมาเปรียบเทียบประสิทธิภาพของแต่ละวิธีการ

12. **แก้ปัญหาความไม่สมดุลกันของข้อมูลด้วยการสุ่มตัวอย่างข้อมูล(Sampling Algorithm)**

หลังจากที่เราทำการแบ่งข้อมูลที่ใช้ในการสอน(Train Data) และใช้ในการทดสอบ(Test Data) เรียบร้อยแล้วจะทำการนำข้อมูลในส่วนที่ใช้ทำการสอนมาทำการทำกระบวนการเพิ่มข้อมูลในคลาสกลุ่มน้อย(Oversampling Technique) เพื่อจัดการกับความไม่สมดุลกันของข้อมูล (Imbalance Data) โดยในงานวิจัยเราจะใช้วิธีการเพิ่มข้อมูลในคลาสกลุ่มน้อยด้วยกัน 2 วิธีคือ Random Oversampling และ Synthetic Minority Oversampling Technique(SMOTE) เพื่อลดความไม่สมดุลกันของข้อมูล

13. **อัลกอริทึมและแบบจำลองที่ใช้ทำนาย ทดลองกับข้อมูลที่มีความสมดุลกัน(Model Algorithm with Balance Data)**

เมื่อผ่านขั้นตอนในการทำ Sampling Algorithm เรียบร้อยแล้วเราจะนำข้อมูลที่ได้หลังจากการทำ Sampling Algorithm มาทดลองกับแบบจำลองการจำแนนประเภททั้ง 4 แบบจำลองที่เราใช้ในงานวิจัยร่วมกับการทำ 10-Folds Cross Validation โดยแบบจำลองการจำแนกประเภททั้ง 4 แบบจำลองที่เราใช้ในงานวิจัยมีดังนี้

1. Naïve Bayes
2. Logistic Regression
3. Support Vector Machine
4. Random Forest

โดยจะใช้ค่ามาตรฐาน (Default) ของแบบจำลองแต่ละประเภทตามที่แต่ละแบบจำลองกำหนดใช้ในการทดลอง หลังจากนั้นจะนำมาทดสอบกับข้อมูลที่ใช้ทดสอบกับแบบจำลองการจำแนกประเภททั้ง 4 และนำมาเปรียบเทียบประสิทธิภาพของแต่ละวิธีการ

14. **การวัดประสิทธิภาพและประเมิณผลการทดลองของแบบจำลอง (Model Evaluation)**

การทดลองของงานวิจัยเราจะวัดประสิมธิภาพและประเมิณผลของแบบจำลองโดยการใช้ Confusion Matrix ร่วมกันกับ ค่าAccuracy , Precision , Recall และ F1-Score เป็นตัว

พิจารณาความแม่นยำของแบบจำลอง และพิจารณาเลือกแบบจำลองที่มีประสิทธิภาพที่ดีที่สุดเพื่อใช้ในการทดสอบกับข้อมูลเพื่อทำนายค่าความน่าจะเป็นต่อไป

15. การปรับจูนพารามิเตอร์กับแบบจำลองที่เลือก

หลังจากที่ได้ทำการทดลองกับแบบจำลองจำแนกประเภทร่วมกับวิธีการจัดการกับข้อมูลที่ไม่สมดุลกันเป็นที่เรียบร้อยแล้ว ทางผู้วิจัยจะเลือกแบบจำลองที่มีประสิทธิภาพโดยการดูค่า Accuracy, Recall , Precision และ F1-Score เพื่อพิจารณาเลือกแบบจำลองนำมาทดลองเพิ่มเติมโดยการปรับจูนพารามิเตอร์ของแบบจำลองที่เลือกมาโดยใช้ GridSearchCV เป็นตัวหาพารามิเตอร์ที่ดีเหมาะสมและดีที่สุดเพื่อนำมาใช้งาน แบบจำลองที่ผู้วิจัยเลือกมาทดลองเพิ่มเติมนั้นคือ Random Forest นำมาทดลองปรับจูนพารามิเตอร์โดยใช้ GridSearchCV กับข้อมูลที่ได้ผ่านการแก้ปัญหาการไม่สมดุลกันของข้อมูลโดยใช้เทคนิค SMOTE พารามิเตอร์ที่ทางผู้วิจัยใช้ในการปรับจูนในครั้งนี้คือ n_estimators และ max_features โดยค่าพารามิเตอร์ที่ปรับจูนโดยใช้ GridSearchCV ในการหาพารามิเตอร์ที่ดีที่สุดผู้วิจัยจะใช้ค่าดังนี้ 'n_estimators': [50, 100, 200, 300, 400, 500, 600, 700, 800] และ 'max_features': ['auto', 'sqrt', 'log2'] โดยจะใช้รวมกับการทำ 10-Fold Cross Validation

บทที่ 4
ผลการดำเนินงานวิจัย

ในการวิจัยเพื่อศึกษาวิธีการวิเคราะห์ข้อมูลข้อความร่วมกันกับข้อมูลที่ไม่ใช้ข้อความ ซึ่งใช้ข้อมูลการเคลมประกันของ บมจ.เอเชียประกันภัย 1950 โดยใช้เทคนิคการเรียนรู้ของเครื่องร่วมกับการวิเคราะห์ข้อความ เพื่อทำการจำแนกประเภทของเคลมนั้นว่าเป็นการทุจริตเคลมหรือไม่ทุจริตเคลม ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามขบวนการและขั้นตอนการวิจัยต่างๆ ตลอดจนการวัดประสิทธิภาพ เพื่อให้บรรลุจุดประสงค์ของการวิจัยที่ได้กำหนดไว้ ดังนี้

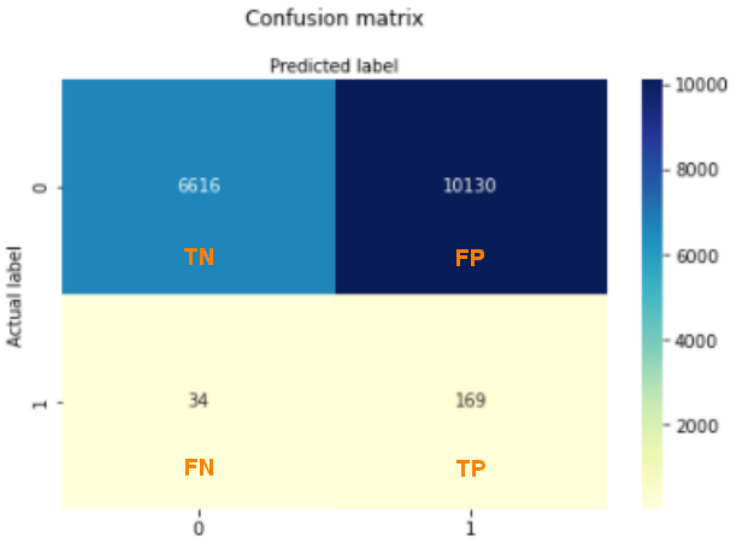
1. ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ไม่สมดุลกัน
2. ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ทำ Random Oversampling
3. ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ทำ SMOTE
4. ผลลัพธ์ของการทดลองปรับจูนพารามิเตอร์ของแบบจำลองที่เลือก
5. เปรียบเทียบผลลัพธ์ของการทดลองของแบบจำลองที่เลือก Random Forest
6. ผลลัพธ์ของการทดลองกับข้อมูลใหม่ที่ไม่ได้ใช้ในการฝึกฝนและทดสอบ

1. ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ไม่สมดุลกัน

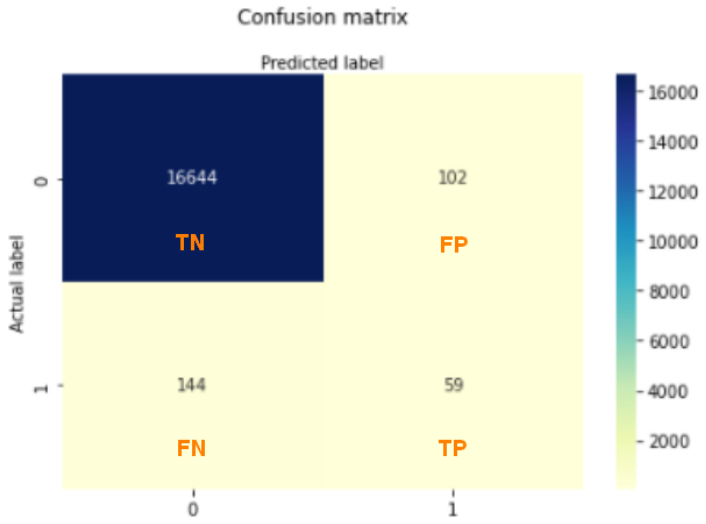
ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูลนำข้อมูลที่ได้ทั้งคุณลักษณะที่เป็นข้อความกับคุณลักษณะที่ไม่ใช่ข้อความโดยข้อมูลที่ได้มีความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับคลาสที่ไม่เป็นการทุจริตเคลม โดยผลลัพธ์ที่ได้จากการทดลอง ดังตารางที่ 2 ถึง 3 และ ภาพประกอบที่ 54 ถึง 59

ตาราง 2 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ไม่สมดุลกัน

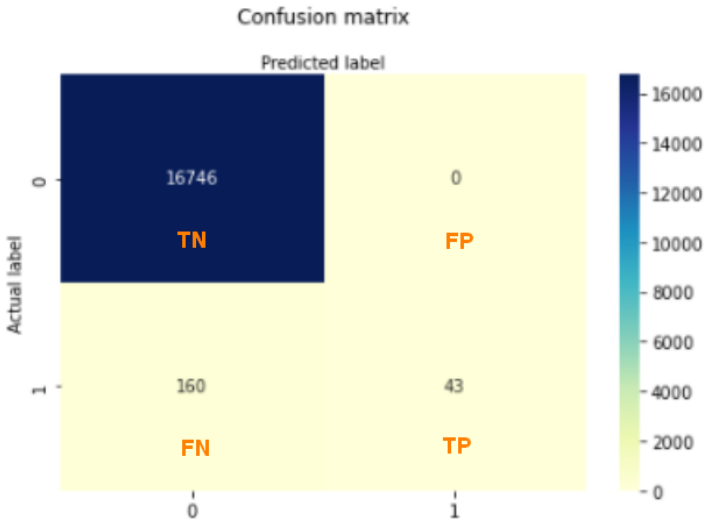
Imbalance Data				
Model	NB	LR	RF	SVM
Accuracy	0.401	0.986	0.991	0.984
Precision	0.017	0.367	1	0.304
Recall	0.833	0.291	0.212	0.321
F1-Score	0.033	0.325	0.35	0.312
Train Duration	0:00:38	0:01:04	0:07:11	2:59:56



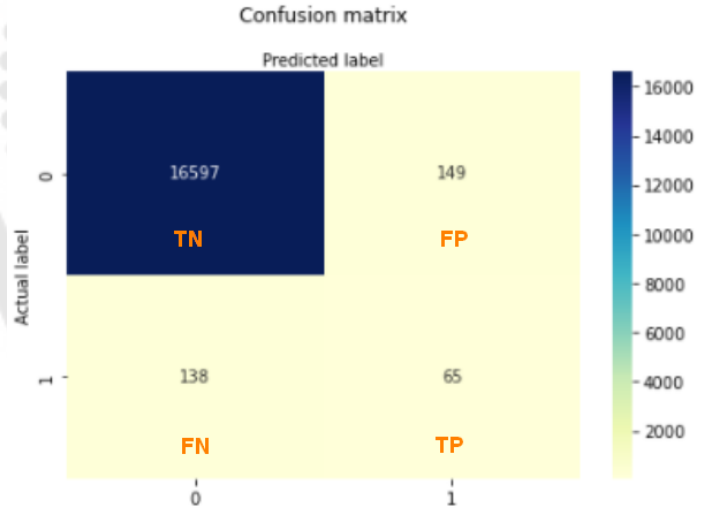
ภาพประกอบ 54 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Naïve Bayes ร่วมกับข้อมูลที่ไม่สมดุลกัน



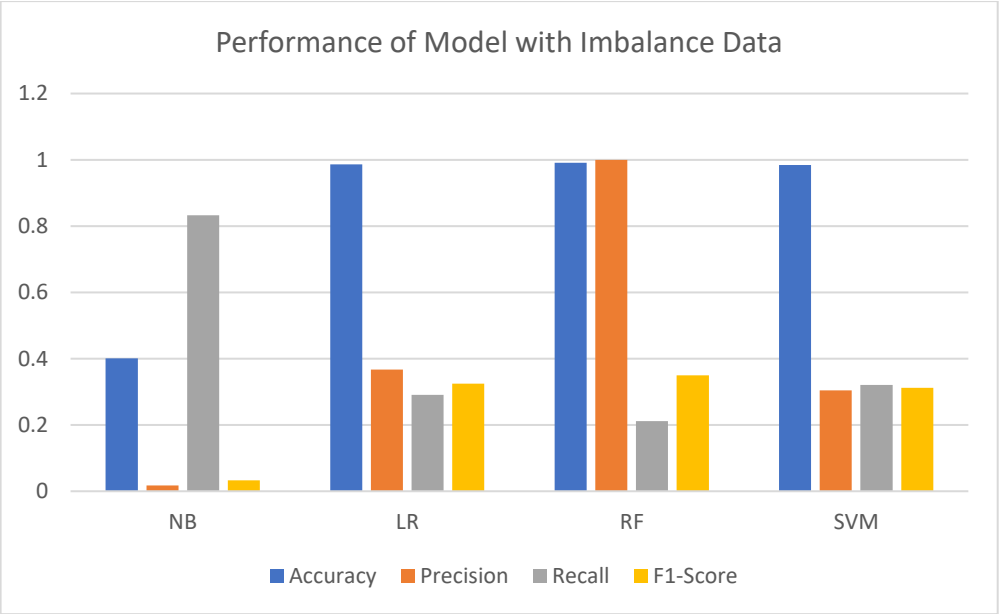
ภาพประกอบ 55 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Logistic Regression ร่วมกับข้อมูลที่ไม่สมดุลกัน



ภาพประกอบ 56 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Random Forest ร่วมกับข้อมูลที่ไม่สมดุลกัน



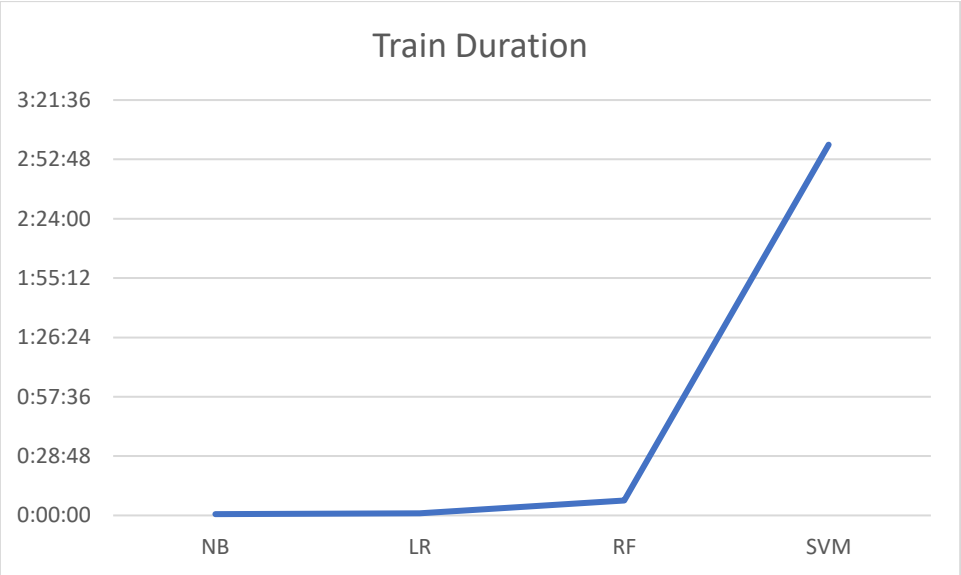
ภาพประกอบ 57 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง SVM ร่วมกับข้อมูลที่ไม่สมดุลกัน



ภาพประกอบ 58 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ไม่สมดุลกัน

ตาราง 3 ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ไม่สมดุลกัน

Model	Train Duration
NB	0:00:38
LR	0:01:04
RF	0:07:11
SVM	2:59:56



ภาพประกอบ 59 แสดงระยะเวลาที่ใช้ในการฝึกสอนแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ไม่สมดุลกัน

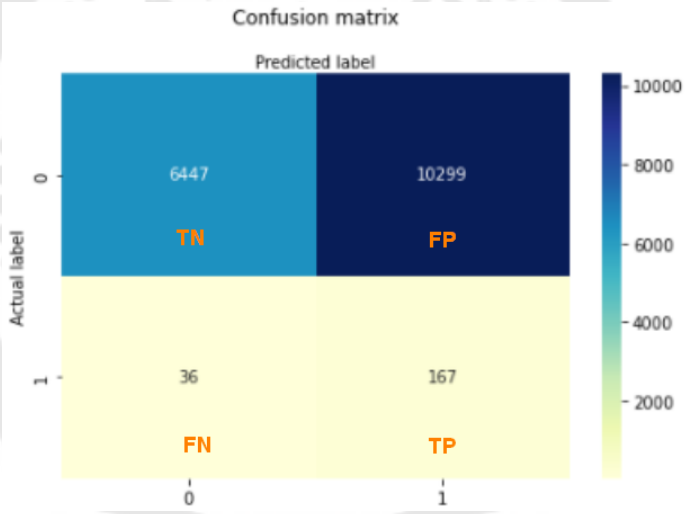
จากผลลัพธ์จะเห็นได้ว่าแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy = 0.991 , ค่า Precision = 1.0 , ค่า Recall = 0.212 และ ค่า F1-Score = 0.35 ภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น โดยที่ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 07:11 นาที ถ้าเปรียบเทียบกับแบบจำลอง Naïve Bayes และ Logistic Regression แล้วใช้เวลาน้อยกว่าแต่ได้ประสิทธิภาพที่น้อยกว่าแบบจำลอง Random Forest

2. **ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ทำ** Random Oversampling

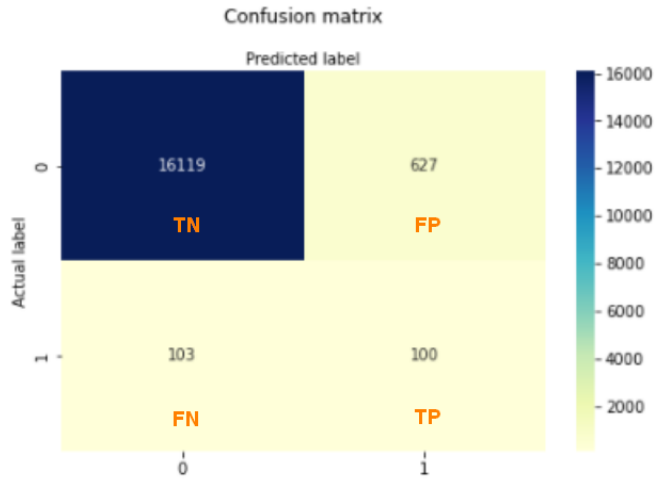
ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูลนำข้อมูลที่ได้ทั้งคุณลักษณะที่เป็นข้อความกับคุณลักษณะที่ไม่ใช่ข้อความโดยข้อมูลที่ได้ถูกแก้ไขความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับคลาสที่ไม่เป็นการทุจริตเคลมโดยใช้วิธีการ Random Oversampling โดยผลลัพธ์ที่ได้จากการทดลอง ดังตารางที่ 4 ถึง 5 และภาพประกอบที่ 60 ถึง 65

ตาราง 4 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ Random Oversampling

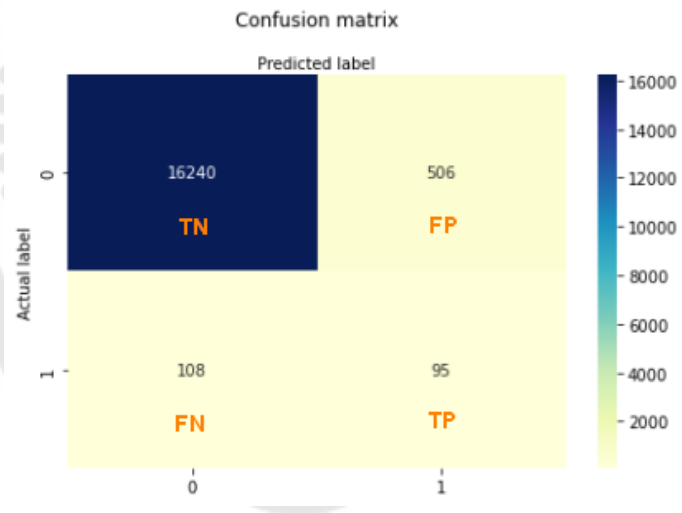
Random Oversampling				
Model	NB	LR	RF	SVM
Accuracy	0.391	0.957	0.991	0.964
Precision	0.016	0.138	0.977	0.159
Recall	0.823	0.493	0.207	0.468
F1-Score	0.032	0.216	0.342	0.237
Train Duration	0:01:18	0:03:18	0:21:01	5 days, 12:04:14



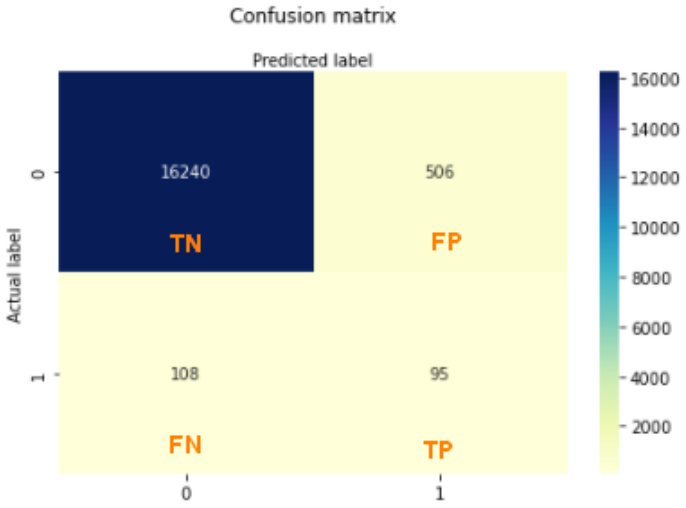
ภาพประกอบ 60 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Naïve Bayes ร่วมกับข้อมูลที่ทำ Random Oversampling



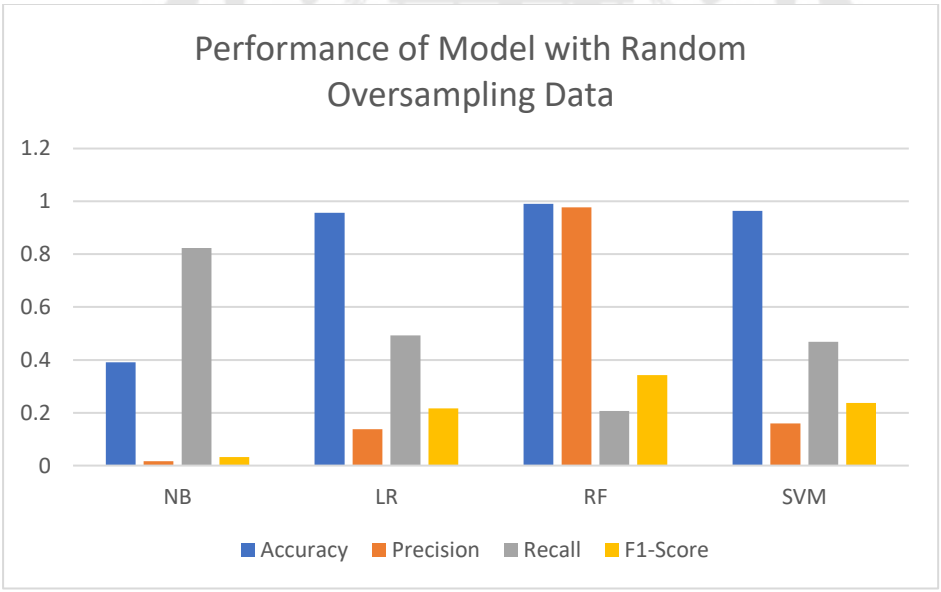
ภาพประกอบ 61 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Logistic Regression ร่วมกับข้อมูลที่ทำ Random Oversampling



ภาพประกอบ 62 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Random Forest ร่วมกับข้อมูลที่ทำ Random Oversampling



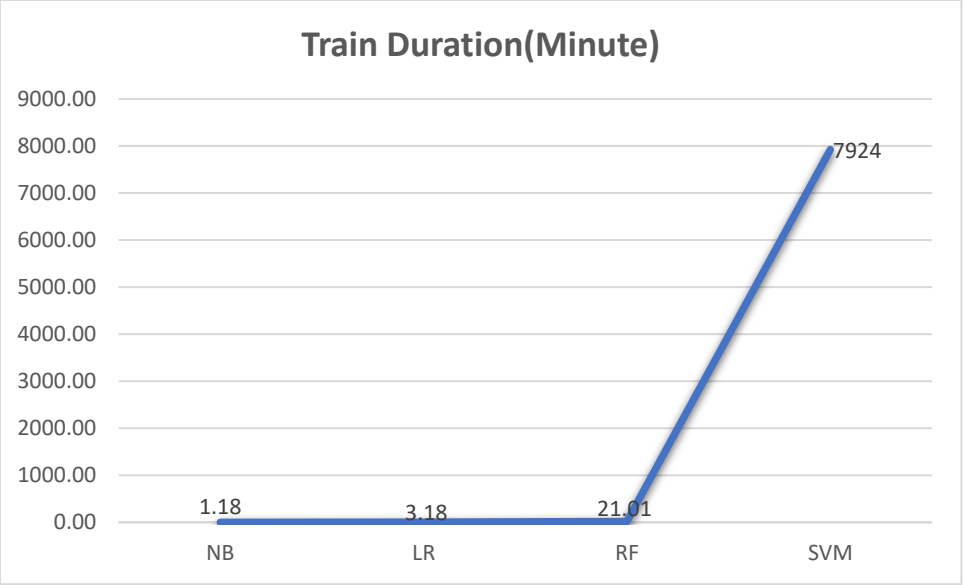
ภาพประกอบ 63 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง SVM ร่วมกับข้อมูลที่ทำ Random Oversampling



ภาพประกอบ 64 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ Random Oversampling

ตาราง 5 ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ Random Oversampling

Model	Train Duration (Min)
NB	1.18
LR	3.18
RF	21.01
SVM	7924



ภาพประกอบ 65 แสดงระยะเวลาที่ใช้ในการฝึกสอนแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ Random Oversampling

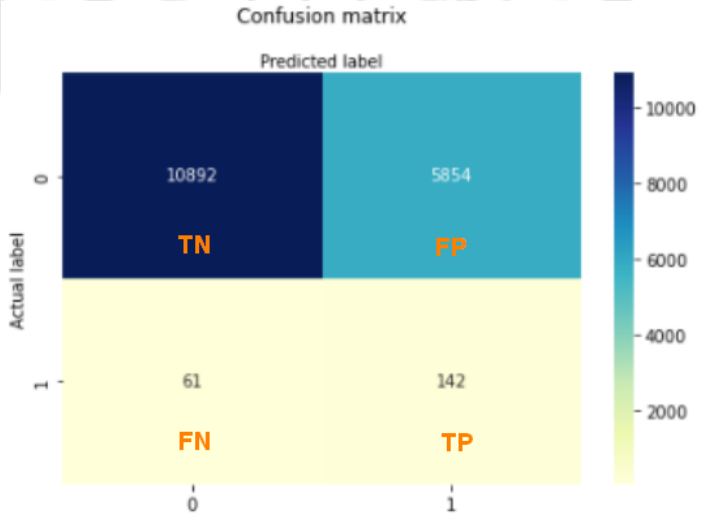
จากผลลัพธ์จะเห็นได้ว่าแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy = 0.991 , ค่า Precision = 0.977 , ค่า Recall = 0.207 และ ค่า F1-Score = 0.342 ภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น โดยที่ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 21:01 นาที ถ้าเปรียบเทียบกับแบบจำลอง Naïve Bayes และ Logistic Regression แล้วใช้เวลาน้อยกว่าแต่ได้ประสิทธิภาพที่น้อยกว่าแบบจำลอง Random Forest

3. **ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ทำ** SMOTE

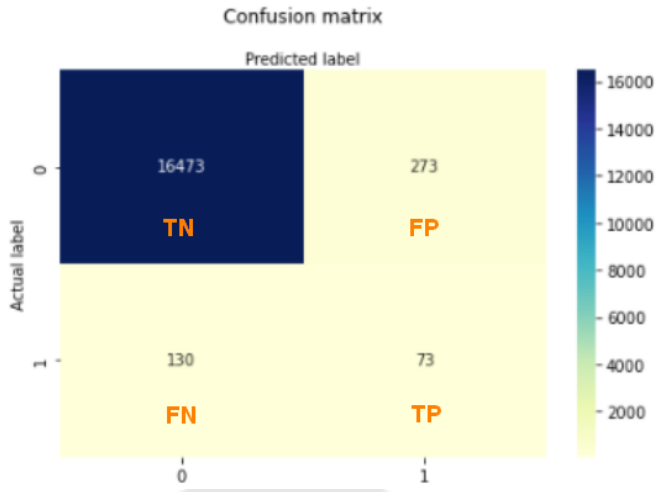
ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูลนำข้อมูลที่ได้ทั้งคุณลักษณะที่เป็นข้อความกับคุณลักษณะที่ไม่ใช่ข้อความโดยข้อมูลที่ได้ถูกแก้ไขความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับคลาสที่ไม่เป็นการทุจริตเคลมโดยใช้วิธีการ SMOTE โดยผลลัพธ์ที่ได้จากการทดลอง ดังตารางที่ 6 ถึง 7 และ ภาพประกอบที่ 66 ถึง 71

ตาราง 6 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ SMOTE

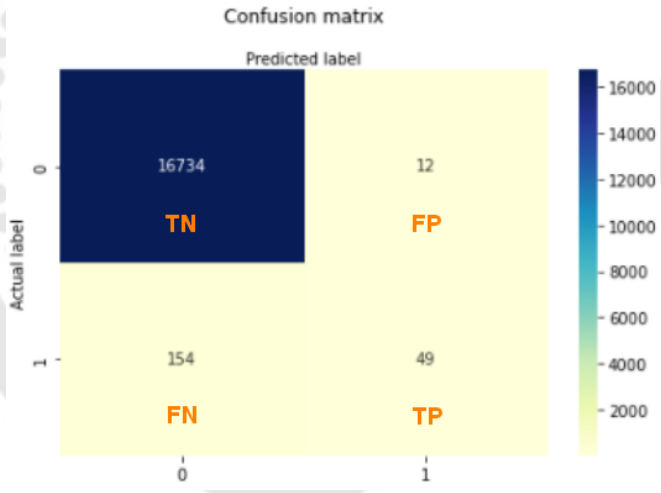
	SMOTE			
Model	NB	LR	RF	SVM
Accuracy	0.652	0.977	0.99	0.977
Precision	0.024	0.211	0.803	0.207
Recall	0.7	0.36	0.241	0.335
F1-Score	0.046	0.266	0.371	0.256
Train Duration	0:02:54	0:04:34	0:12:00	1 day, 11:12:16



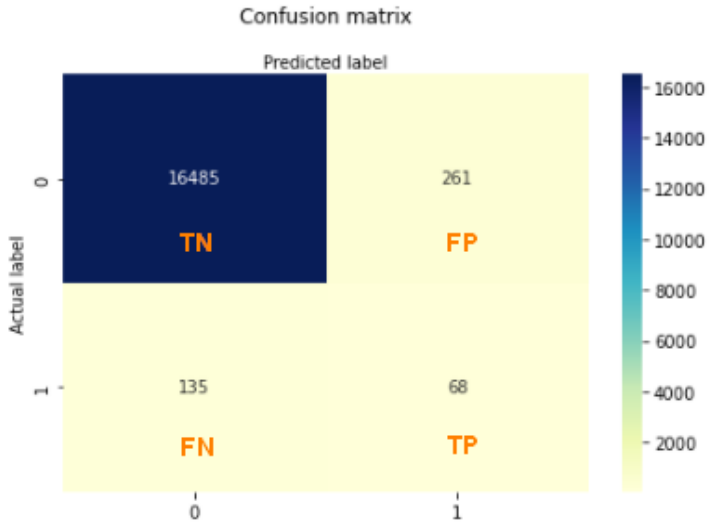
ภาพประกอบ 66 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Naïve Bayes ร่วมกับข้อมูลที่ทำ SMOTE



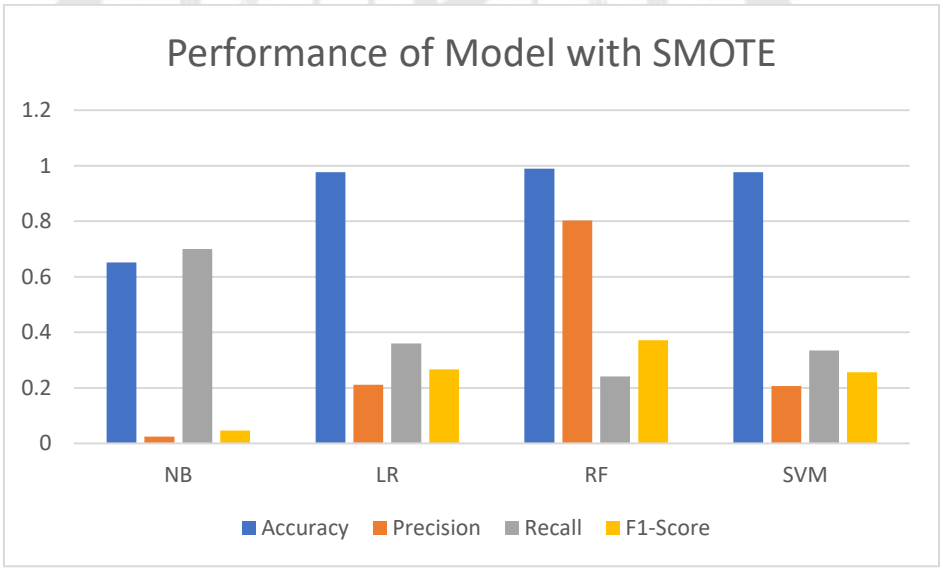
ภาพประกอบ 67 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Logistic Regression ร่วมกับข้อมูลที่ทำ SMOTE



ภาพประกอบ 68 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง Random Forest ร่วมกับข้อมูลที่ทำ SMOTE



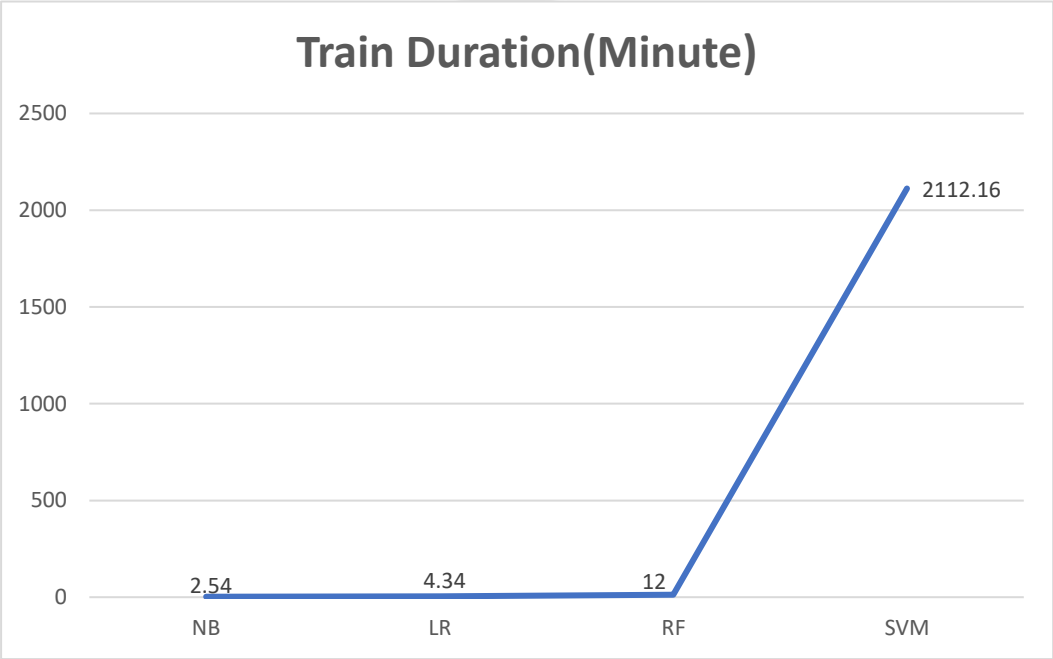
ภาพประกอบ 69 Confusion Matrix ที่ได้จากการทดลองแบบจำลอง SVM ร่วมกับข้อมูลที่ทำ SMOTE



ภาพประกอบ 70 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ SMOTE

ตาราง 7 ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ SMOTE

Model	Train Duration (Min)
NB	2.54
LR	4.34
RF	12
SVM	2112.16

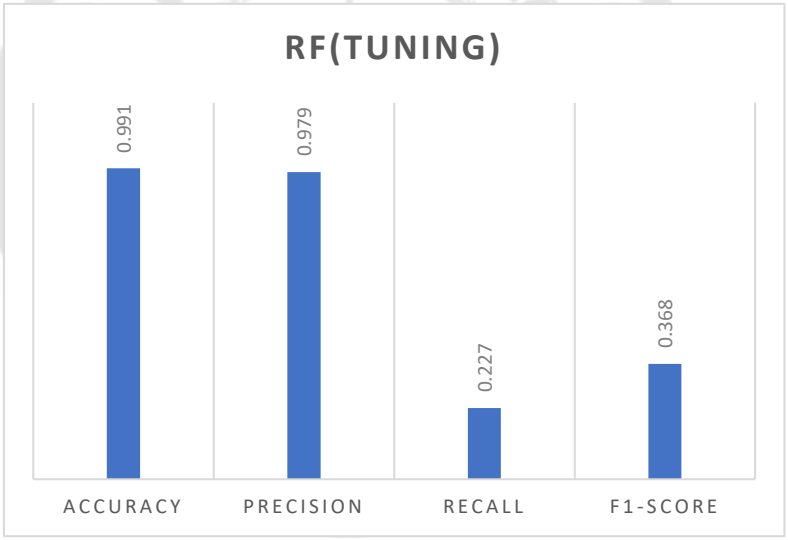


ภาพประกอบ 71 แสดงระยะเวลาที่ใช้ในการฝึกสอนแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ SMOTE

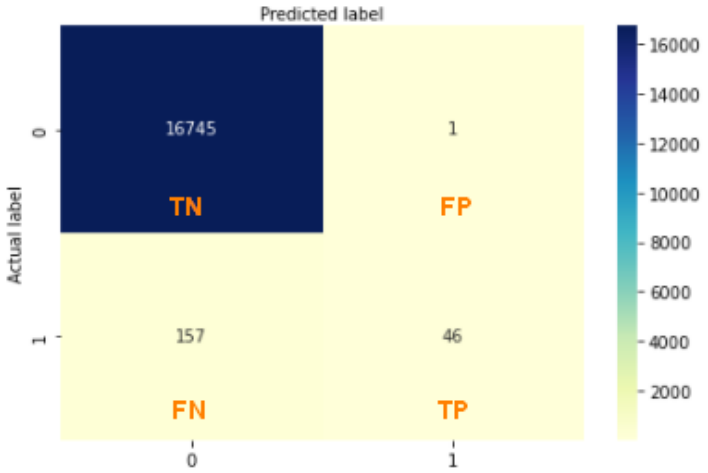
จากผลลัพธ์จะเห็นได้ว่าแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy = 0.99 , ค่า Precision = 0.803 , ค่า Recall = 0.241 และ ค่า F1-Score = 0.371 ภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น โดยที่ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 12 นาที ถ้าเปรียบเทียบกับแบบจำลอง Naïve Bayes และ Logistic Regression แล้วใช้เวลาน้อยกว่าแต่ได้ประสิทธิภาพที่น้อยกว่าแบบจำลอง Random Forest

4. **ผลลัพธ์ของการทดลองปรับจูนพารามิเตอร์ของแบบจำลองที่เลือก**

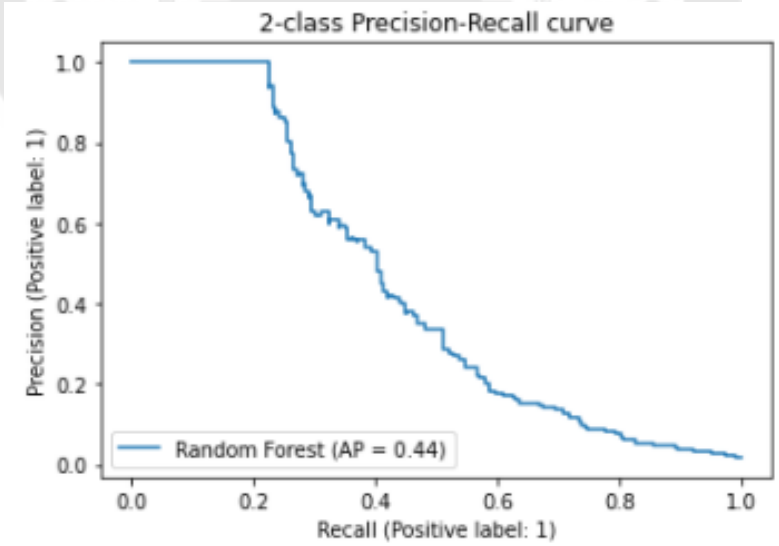
จากการทดลองในการปรับจูนพารามิเตอร์โดยใช้ GridSearchCV ในการเลือกพารามิเตอร์ที่ดีที่สุดของแบบจำลอง Random Forest ที่ทดลองกับข้อมูลที่ผ่านการทำ SMOTE ผลลัพธ์ที่ได้ค่าของพารามิเตอร์ที่เลือกออกมา คือ 'max_features': 'log2', 'n_estimators': 600 โดยใช้เวลาในการทำ GridSearchCV ทั้งสิ้น 4 ชั่วโมง 47 นาทีในการหาพารามิเตอร์ที่ดีที่สุด และเมื่อนำค่าพารามิเตอร์ที่ได้มาทดสอบเพื่อดูความแม่นยำและระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองจะได้ผลลัพธ์ดังนี้ ได้ประสิทธิภาพและความแม่นยำในการทำนายออกมาโดยมีค่า Accuracy = 0.991 , Precision = 0.979 , Recall = 0.227 และ F1-Score = 0.368 โดยได้ค่า TP=46 , FN=157 , TN=16745 และ FP=1 โดยใช้เวลาการฝึกฝนแบบจำลอง 23:37 นาที ดังแสดงในภาพประกอบที่ 72 ถึง 75



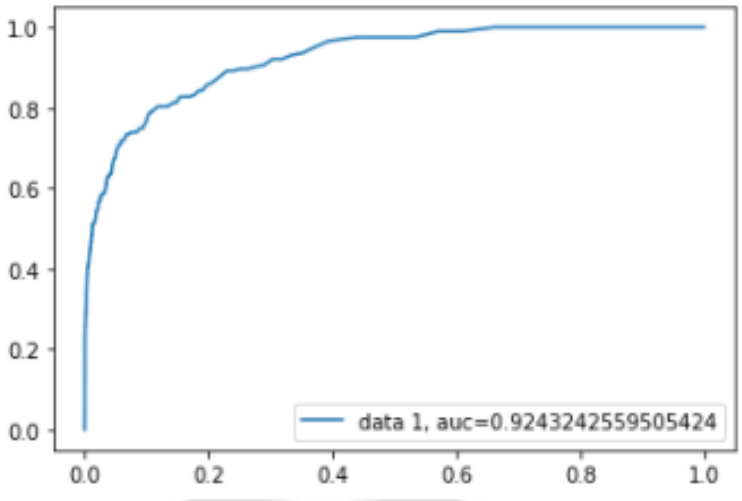
ภาพประกอบ 72 ผลลัพธ์ที่ได้จากการทดลองปรับจูนพารามิเตอร์กับแบบจำลอง Random Forest



ภาพประกอบ 73 Confusion Matrix ของการทดลองปรับจูนพารามิเตอร์กับแบบจำลอง Random Forest



ภาพประกอบ 74 Precision-Recall Curve ของการทดลองปรับจูนพารามิเตอร์กับแบบจำลอง Random Forest



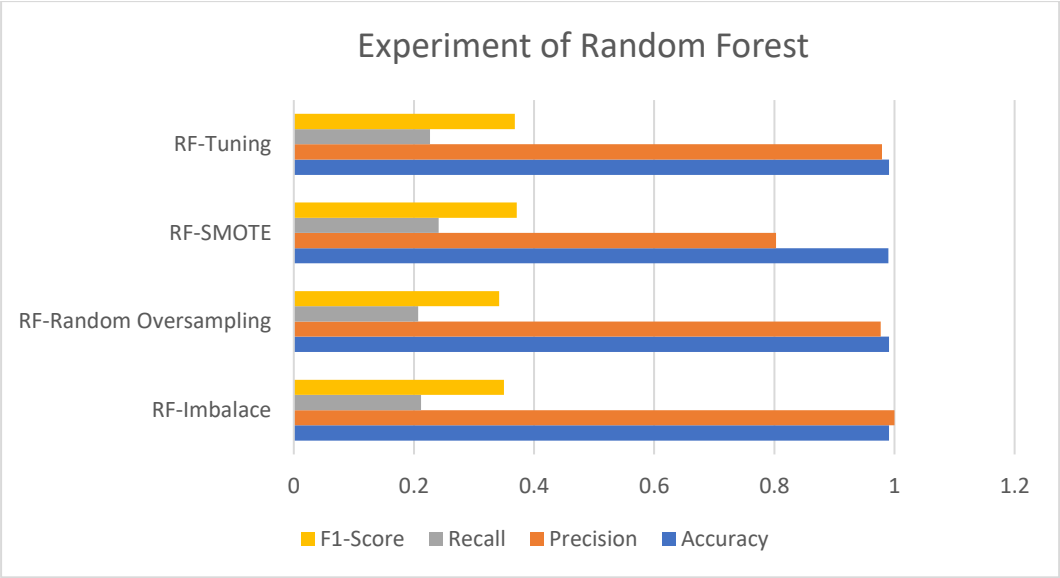
ภาพประกอบ 75 ROC Curve ของการทดลองปรับจูนพารามิเตอร์กับแบบจำลอง Random Forest

5. เปรียบเทียบผลลัพธ์ของการทดลองของแบบจำลองที่เลือก Random Forest

จากการทดลองทั้งหมดเรานำแบบจำลองที่เราเลือกมาทดลองคือ Random Forest ซึ่งเราได้นำการทดลองแต่ละวิธีการมาเปรียบเทียบเพื่อดูประสิทธิภาพ ของแบบจำลอง Random Forest เพื่อให้เห็นว่าวิธีการใดเหมาะสมที่จะนำมาใช้หรือต่อยอดปรับปรุงประสิทธิภาพต่อไปโดยมีรายละเอียดดังคารางเปรียบเทียบตามตามรางที่ 8

ตาราง 8 เปรียบเทียบ Random Forest กับวิธีการทดลองแบบต่างๆ

	RF-Imbalance	RF-Random Oversampling	RF-SMOTE	RF-Tuning
Accuracy	0.991	0.991	0.99	0.991
Precision	1	0.977	0.803	0.979
Recall	0.212	0.207	0.241	0.227
F1-Score	0.35	0.342	0.371	0.368
Train Duration	0:07:11	0:21:01	0:12:00	0:23:37



ภาพประกอบ 76 ผลลัพธ์ที่ได้จากการทดลองแบบจำลอง Random Forest ในแต่ละวิธีการทดลอง

6. **ผลลัพธ์ของการทดลองกับข้อมูลใหม่ที่ไม่ได้ใช้ในการฝึกฝนและทดสอบ**

การนำข้อมูลใหม่ที่ไม่ได้ผ่านการฝึกฝนและทดลองของแบบจำลองมาทำการทดสอบแบบจำลองการทำนายโอกาสเกิดเคลมทุจริต และให้ค่าความน่าจะเป็นของข้อมูลที่ได้นำมาทำนายโดยมีผลการทำนายดัง 4 ตัวอย่างนี้

การทำนายผลที่1 ข้อมูลที่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 77 และผลลัพธ์ที่ได้ดังรูปที่ 78

inform_date	2020-01-21 09:18:34.000
inform_time	2020-01-21 09:18:34.000
date_occur	2020-01-20 00:00:00.000
time_occur	2020-01-20 00:00:00.000
ins_startdate	2019-03-15 00:00:00.000
driver_sex	F
datacase	100
datacase_desc	ฝ่ายผิด
datacasedt	401
datacasedt_desc	จอดไว้รถหาย
comment	เอเอ1637 เรียนหัวหน้าฝ่ายสืนไหมทราบ ในวันที่ 21/1/63 เวลานัดหมาย 10.30 น. เดินทางตามนัดหมาย สภ.ชัยพฤกษ์พบ ป.และ พ.ง.ส. สอบถามทราบว่า ในวันที่ 20/1/63
fraudity	Y
tdate	2019-03-18
body_type	05
body_desc	จักรยานยนต์
veh_use	610
veh_use_desc	ใช้ส่วนบุคคล ไม่ใช่รับจ้างหรือให้เช่า
ageofvehicle	1

ภาพประกอบ 77 การทำนายผลที่1 ข้อมูลที่มีการทุจริตเคลม

```
Predicted Class :  [0]
Actual Lable fraudity(Y/N) =  9    Y
Name: fraudity, dtype: object
Negative Class(0)  0.7
Positive Class(1)  0.3
All Class  1.0
Probability of Non fraud is 70.02%
Probability of fraud is 30.00%
```

ภาพประกอบ 78 การทำนายผลที่1 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่2 ข้อมูลที่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 79 และผลลัพธ์ที่ได้ดังรูปที่ 80

inform_date	2020-01-26 22:46:52.000
inform_time	2020-01-26 22:46:52.000
date_occur	2020-01-26 22:45:52.000
time_occur	2020-01-26 22:45:52.000
ins_startdate	2019-12-23 00:00:00.000
driver_sex	M
datacase	100
datacase_desc	ฝ่ายผิด
datacasedt	103
datacasedt_desc	ถอยชน/ไหลชนคู่กรณี
comment	เจ้าหน้าที่บริษัทท่านมีคำสั่งให้ออกตรวจสอบอุบัติเหตุรถประกันรายนี้ที่ประชาอุทิศ 99 ทางพนักงานเดินทาง
fraudity	Y
tdate	2019-12-23
body_type	04
body_desc	กระบะบรรทุก
veh_use	320
veh_use_desc	ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส
ageofvehicle	2

ภาพประกอบ 79 การทำนายผลที่2 ข้อมูลที่มีการทุจริตเคลม

```
Predicted Class :  [1]
Actual Lable fraudity(Y/N) =  10    Y
Name: fraudity, dtype: object
Negative Class(0)  0.13
Positive Class(1)  0.87
All Class  1.0
Probability of Non fraud is 13.00%
Probability of fraud is 87.01%
```

ภาพประกอบ 80 การทำนายผลที่2 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่3 ข้อมูลที่ไม่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่ไม่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 81 และผลลัพธ์ที่ได้ดังรูปที่ 82

inform_date	11/11/2020
inform_time	11/11/2020
date_occur	11/11/2020
time_occur	11/11/2020
ins_startdate	3/18/2020
driver_sex	M
datacase	200
datacase_desc	ฝ่ายถูก
datacasedt	202
datacasedt_desc	ถูกคู่กรณีถอยชน/ไหลชน
comment	นัดหมายไม่แจกน้ำวันที่ 11/11/2563 เวลา15.30น.ออกตรวจสอบรถประกันตามนัดหมายที่ บิ๊กขีนวนคร
fraudity	N
tdate	2020-03-19
body_type	02
body_desc	รถตู้
veh_use	220
veh_use_desc	ใช้เพื่อการพาณิชย์ ไม่ใช่รับจ้างสาธารณะ
ageofvehicle	15

ภาพประกอบ 81 การทำนายผลที่3 ข้อมูลที่ไม่มีการทุจริตเคลม

```
Predicted Class :  [0]
Actual Lable fraudity(Y/N) =  15    N
Name: fraudity, dtype: object
Negative Class(0)  1.0
Positive Class(1)  0.0
All Class  1.0
Probability of Non fraud is 100.00%
Probability of fraud is 0.00%
```

ภาพประกอบ 82 การทำนายผลที่3 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่4 ข้อมูลที่ไม่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่ไม่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 83 และผลลัพธ์ที่ได้ดังรูปที่ 84

inform_date	11/1/2020
inform_time	11/1/2020
date_occur	11/1/2020
time_occur	11/1/2020
ins_startdate	11/28/2019
driver_sex	M
datacase	100
datacase_desc	ฝ่ายผิด
datacasedt	144
datacasedt_desc	หินกระเด่นใส่
comment	631101054 รายละเอียดการเกิดเหตุ บริษัทได้รับมอบหมายจากคุณถาวร ฯ ให้ออกตรวจสอบอุบัติเหตุที่ถนนเทพารักษ์กม. 11 อำเภอบางพลี
fraudity	N
tdate	2019-11-07
body_type	01
body_desc	เก๋ง2ตอน
veh_use	110
veh_use_desc	ใช้ส่วนบุคคล ไม่ใช่รับจ้างหรือให้เช่า
ageofvehicle	3

ภาพประกอบ 83 การทำนายผลที่4 ข้อมูลที่ไม่มีการทุจริตเคลม

```
Predicted Class :  [0]
Actual Lable fraudity(Y/N) =  20    N
Name: fraudity, dtype: object
Negative Class(0)  1.0
Positive Class(1)  0.0
All Class  1.0
Probability of Non fraud is 100.00%
Probability of fraud is 0.00%
```

ภาพประกอบ 84 การทำนายผลที่4 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่5 ข้อมูลที่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 85 และผลลัพธ์ที่ได้ดังรูปที่ 86

inform_date	2019-12-28 17:41:37.683
inform_time	2019-12-28 17:41:37.763
date_occur	2019-12-28 17:41:37.683
time_occur	2019-12-28 08:40:37.683
ins_startdate	2019-06-15 00:00:00.000
driver_sex	M
datacase	100
datacase_desc	ฝ่ายผิด
datacasedt	405
datacasedt_desc	เสียหลักตกหลุม/ตกข้างทาง
comment	62125976, AN6301/0273 วันที่ 06/01/2563 เรียน ผู้จัดการฝ่ายสินไหม ผ่านหัวหน้าอุบัติเหตุ ออกตรวจสอบอุบัติเหตุตามแจ้งเป็นเคลม ตามคดี1 ที่ สภ.เทพสถิต ตามคำสั่งหัวหน้า
fraudity	Y
tdate	2019-05-30
body_type	01
body_desc	เก๋ง2ตอน
veh_use	110
veh_use_desc	ใช้ส่วนบุคคล ไม่ใช้รับจ้างหรือให้เช่า
ageofvehicle	2

ภาพประกอบ 85 การทำนายผลที่5 ข้อมูลที่มีการทุจริตเคลม

```
Predicted Class :  [0]
Actual Lable fraudity(Y/N) =  3    Y
Name: fraudity, dtype: object
Negative Class(0)  0.96
Positive Class(1)  0.04
All Class  1.0
Probability of Non fraud is 96.00%
Probability of fraud is 4.00%
```

ภาพประกอบ 86 การทำนายผลที่5 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่6 ข้อมูลที่ไม่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่ไม่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 87 และผลลัพธ์ที่ได้ดังรูปที่ 88

inform_date	10/29/2020
inform_time	10/29/2020
date_occur	10/29/2020
time_occur	10/29/2020
ins_startdate	3/9/2020
driver_sex	M
datacase	100
datacase_desc	ฝ่ายผิด
datacasedt	101
datacasedt_desc	ชนท้ายคู่กรณี
comment	เลขรับแจ้ง 631029164 เลขเอพลัส 2010261329 (พนักงานออกตรวจสอบ ได้รับแจ้งเหตุ จากบริษัทฯ ท่าน ให้ออกตรวจสอบ บริเวณ ซ่อย พหลโยธิน 11 แขวง สามเสนใน
fraudity	N
tdate	2020-03-06
body_type	01
body_desc	เก๋ง2ตอน
veh_use	110
veh_use_desc	ใช้ส่วนบุคคล ไม่ใช่รับจ้างหรือให้เช่า
ageofvehicle	18

ภาพประกอบ 87 การทำนายผลที่6 ข้อมูลที่ไม่มีการทุจริตเคลม

```
Predicted Class :  [0]
Actual Lable fraudity(Y/N) =  29    N
Name: fraudity, dtype: object
Negative Class(0)  0.98
Positive Class(1)  0.02
All Class  1.0
Probability of Non fraud is 98.00%
Probability of fraud is 2.00%
```

ภาพประกอบ 88 การทำนายผลที่6 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่7 ข้อมูลที่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 89 และผลลัพธ์ที่ได้ดังรูปที่ 90

inform_date	2020-01-02 19:47:03.870
inform_time	2020-01-02 19:47:03.200
date_occur	2020-01-02 19:47:03.870
time_occur	2020-01-02 19:40:03.870
ins_startdate	2019-04-28 00:00:00.000
driver_sex	M
datacase	100
datacase_desc	ฝ่ายผิด
datacasedt	111
datacasedt_desc	เปลี่ยนช่องทางเฉี่ยวชนคู่กรณี
comment	AA6301/0095 สด ทำเคลมวันที่ 2/12/2563 เรียนหัวหน้าฝ่ายสินไหมทราบ เดินทางถึงที่เกิดเหตุ ก่อนถึงซอยนวลจันทร์ ถนนประดิษฐ์มนูธรรม พบประกันและคู่กรณี จากการ
fraudity	Y
tdate	2019-04-25
body_type	01
body_desc	เก๋ง2ตอน
veh_use	730
veh_use_desc	ใช้รับจ้างสาธารณะ
ageofvehicle	6

ภาพประกอบ 89 การทำนายผลที่7 ข้อมูลที่มีการทุจริตเคลม

```
Predicted Class :  [1]
Actual Lable fraudity(Y/N) =  6    Y
Name: fraudity, dtype: object
Negative Class(0)  0.48
Positive Class(1)  0.52
All Class  1.0
Probability of Non fraud is 48.00%
Probability of fraud is 52.00%
```

ภาพประกอบ 90 การทำนายผลที่7 ผลลัพธ์ที่แบบจำลองทำนาย

การทำนายผลที่8 ข้อมูลที่ไม่มีการทุจริตเคลม

ผู้วิจัยได้เลือกข้อมูลที่ไม่เป็นการทุจริตเคลมนำมาทดสอบแบบจำลองให้แบบจำลองทำนายโอกาสการเกิดการทุจริตเคลมโดยใช้ข้อมูลดังรูปที่ 91 และผลลัพธ์ที่ได้ดังรูปที่ 92

inform_date	11/23/2020
inform_time	11/23/2020
date_occur	11/23/2020
time_occur	11/23/2020
ins_startdate	12/26/2019
driver_sex	M
datacase	200
datacase_desc	ฝ่ายถูก
datacasedt	203
datacasedt_desc	ถูกคู่กรณีเฉี่ยวชน/เบียด
comment	เลขรับแจ้ง 631123010 เลข AN 6311/0844 น้ำหมด ตรวจสอบวันที่ 23/10/2563 เวลา 12.24 น. เรียน หน/อบ ทราบ จากการสอบปากคำผู้ขับขี่รถประกันได้ความว่า รถประกัน
fraudity	N
tdate	2019-12-27
body_type	01
body_desc	เก๋ง2ตอน
veh_use	320
veh_use_desc	ใช้เพื่อการพาณิชย์ ไม่ใช้เพื่อการบรรทุก และขนส่งสินค้าที่มีความเสี่ยงภัยสูง เช่น เชื้อเพลิง กรด แก๊ส และ ไม่ใช้ลากจูงรถพ่วง
ageofvehicle	6

ภาพประกอบ 91 การทำนายผลที่8 ข้อมูลที่ไม่มีการทุจริตเคลม

```
Predicted Class :  [0]
Actual Lable fraudity(Y/N) =  23    N
Name: fraudity, dtype: object
Negative Class(0)  0.95
Positive Class(1)  0.05
All Class  1.0
Probability of Non fraud is 95.02%
Probability of fraud is 5.00%
```

ภาพประกอบ 92 การทำนายผลที่8 ผลลัพธ์ที่แบบจำลองทำนาย

จากการทดสอบข้อมูลใหม่กับแบบจำลองที่ได้ผ่านการฝึกฝนและทดสอบมาแล้วนั้นผลที่ได้ออกมาคือ การทำนายในส่วนของข้อมูลที่ผู้เชียวชาญฝ่ายสินไหมรถยนต์ได้พิจารณาแล้วว่าเป็นเคลมที่มีการทุจริต โดยแบบจำลองทำนายถูก 2 รายการจากตัวอย่าง 4 รายการ จากนั้นได้ทำการทดสอบกับข้อมูลรายการที่ผู้เชียวชาญสินไหมรถยนต์ได้พิจารณาแล้วว่าไม่เป็นการทุจริตเคลม โดยแบบจำลองทำนายถูกทั้ง 4 รายการจากตัวอย่าง 4 รายการ ส่งผลให้เห็นว่าแบบจำลองแม่นยำในการทำนายคลาสที่ไม่เป็นการทุจริตเคลมมากกว่าคลาสที่เป็นการทุจริตเคลมโดยแบบจำลองจะบอกความน่าจะเป็นในการทำนายในแต่ละคลาสออกมา

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยเพื่อศึกษากระบวนการวิเคราะห์ข้อความเพื่อให้เห็นความสำคัญของคำและใช้ร่วมกันกับคุณลักษณะที่ไม่ใช่ข้อความ ที่จะนำไปสู่การทุจริตเคลมเพื่อนำมาสร้างแบบจำลองและเปรียบเทียบแบบจำลองต่างๆ เพื่อทำนายการคาดการความน่าจะเป็นว่าเคลมนั้นจะเกิดการทุจริต ผู้วิจัยได้ประเมินประสิทธิภาพของแบบจำลองแต่ละเทคนิค เพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

1. สรุปผลการวิจัย และ อภิปราผลการวิจัย
2. ข้อเสนอแนะ

สรุปผลการวิจัย และ อภิปรายผลการวิจัย

ในการวิจัยครั้งนี้เป็นการวิจัยเพื่อศึกษากระบวนการวิเคราะห์ข้อความร่วมกันกับคุณลักษณะที่ไม่ใช่ข้อความเพื่อนำไปสู่การตรวจจับการทุจริตเคลม โดยใช้เทคนิคการเรียนรู้ของเครื่อง(Machine Learning) นำมาสร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในแต่ละการทดลอง หลังจากนั้นได้เลือกแบบจำลองที่มีประสิทธิภาพที่สุดกับชุดข้อมูลในงานวิจัยนี้นำมาปรับจูนพารามิเตอร์เพื่อทำการทดลองเพื่อได้ผลลัพธ์ออกมา และทำการทดสอบข้อมูลใหม่กับแบบจำลองที่เราคิดว่าให้ประสิทธิภาพที่ดี

เนื่องจากการวิจัยนี้ผู้วิจัยได้มุ่งเน้นไปที่การนำคุณลักษณะที่ไม่ใช่ข้อความ มาใช้ร่วมกันกับคุณลักษณะที่เป็นข้อความ ในขั้นตอน Preprocess ทั้งคุณลักษณะที่ไม่ใช่ข้อความและคุณลักษณะที่เป็นข้อความค่อนข้างมีความสำคัญเป็นอย่างมาก ถ้าเราทำความสะอาดข้อมูลรวมถึงการตัดคำของคุณลักษณะของข้อความออกมาไม่ดีส่งผลให้จะไม่มีคุณค่าเท่าที่ควรในส่วนของคุณลักษณะที่ไม่ใช่ข้อความเราสามารถจัดกลุ่มรวมประเภทของคุณลักษณะเดียวกันให้ลดน้อยลงได้เพื่อเพิ่มความสำคัญของประเภทย่อยที่อยู่ในคุณลักษระเดียวกันมากขึ้น และทำให้คุณลักษณะที่นำมาใช้ลดน้อยลงไปด้วย ทำให้การฝึกฝนแบบจำลองทำได้เร็วมากขึ้น ในส่วนของคุณลักษณะที่เป็นข้อความการทำ Feature Selection และ Feature Importance ค่อนข้างมีความสำคัญต่อการเลือกคำมาใช้ เนื่องจากถ้าเราไม่เลือกคำที่มีความสำคัญมากๆมาใช้คุณลักษณะข้อความจะมีจำนวนมากส่งผลให้การฝึกฝนแบบจำลองใช้เวลาค่อนข้างมากและใช้ทรัพยากรหน่วยความจำ(Memory) มากทำให้ไม่สามารถฝึกฝนแบบจำลองได้สำเร็จเพราะเกิด

ปัญหาหน่วยความจำไม่เพียงพอ ในงานวิจัยนี้จึงใช้เวลาในขั้นตอน Preprocess ค่อนข้างมาก ผู้วิจัยได้ทำการทดลองตรวจสอบคุณลักษณะที่สำคัญ(Feature Importance) ของคุณลักษณะทั้งหมดร่วมกับแบบจำลอง Random Forest ใน 100 อันดับแรกผลปรากฏว่าคุณลักษณะที่มีความสำคัญที่แบบจำลองเลือกใช้ส่วนใหญ่เป็นคุณลักษณะที่ไม่ใช่ข้อความเช่น ‘เวลาที่เกิดเหตุ(ชั่วโมง)’,’อายุรถ’,’วันที่เกิดเหตุหลังจากวันที่กรมธรรม์เริ่มคุ้มครอง’,’สาเหตุของการเกิดอุบัติเหตุ(เช่น รถประกันเสียหลัก,เฉี่ยวชนคู่กรณี เป็นต้น)’,’ลักษณะของการใช้รถยนต์(เช่น ส่วนบุคคล,เพื่อการพาณิชย์,ใช้รับจ้างสาธารณะ เป็นต้น)’,’ประเภทของตัวรถยนต์(เช่น เก๋ง2ตอน,รถแท็กซี่,กระบะแวน เป็นต้น)’,’ผลคดี(เช่นฝ่ายถูก,ฝ่ายผิด,ประมาทร่วม เป็นต้น)’ ส่วนคุณลักษณะที่เป็นข้อความแบบจำลองที่ได้เลือกมาใช้เป็นคำที่ไม่ได้สื่อความหมายไปในทางที่จะสื่อว่าเป็นการทุจริตเคลม เช่น ‘นุ่น’,’นุ่ง’,’นุ่ม’,’นุด’,’นี้’ แต่มีความสำคัญที่น้อยกว่าคุณลักษณะที่ไม่ใช่ข้อความ ในส่วนของการเลือก คุณลักษณะที่สำคัญ(Feature Importance) ของข้อความโดยผู้วิจัยได้เลือกคำที่มีความสำคัญโดยพิจารณาจากค่า TFIDF ที่มีค่ามากเป็น 1000 คำมาใช้เพื่อเป็นคุณลักษระที่เป็นข้อความ โดยผู้วิจัยได้เลือกค่า Threshold โดยการรวมค่า TFIDF ของแต่ละคำของทุกรายการที่มากกว่าหรือเท่ากับ 60 เพื่อใช้เลือกคำที่สำคัญโดยค่า Threshold ที่เลือกเป็นการทดลองค่าเริ่มต้นเท่านั้น โดยคำส่วนใหญ่ที่มีค่าผลรวมของ TFIDF กระจุกรวมกันอยู่น้อยกว่า 60 ซึ่งคำที่มีค่า IFIDF น้อยจะไม่ได้ให้น้ำหนักของคำมากนักและคำเหล่านั้นไม่ได้ทำให้เกินประโยชน์ต่อแบบจำลองมากนัก

การทดลองส่วนที่เกี่ยวข้องกับความไม่สมดุลกันของข้อมูล ผู้วิจัยได้ลองทดสอบแบบจำลองแต่ละแบบจำลองกับข้อมูลทั้งที่ไม่สมมดุลกัน และข้อมูลที่ได้จัดการให้สมมดุลกัน พบว่าประสิทธิภาพของแบบจำลองที่ได้ทดลองกับข้อมูลที่ผ่านการทำ Oversampling ในวิธีการ SMOTE นั้นมีประสิทธิภาพที่ดีกว่าวิธีการ Random Oversampling สำหรับในชุดข้อมูลที่เราใช้ในงานวิจัยในครั้งนี้

ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้ถูกแก้ไขความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับไม่เป็นการทุจริตเคลม โดยใช้วิธีการ SMOTE จากผลลัพธ์จะเห็นได้ว่าภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy=0.99, Precision=0.803, Recall=0.241, F1-Score=0.371 โดย Naïve Bayes ให้ค่า Recall=0.7, Accuracy=0.652, Precision=0.024, F1-Score=0.0046 โดยที่Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น จากภาพรวมของการวัดประสิทธิภาพ

โดยพิจารณาจากค่า F1-Score(weighted average ระหว่าง Precision และ Recall) ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 12:00 นาที

ปัญหาที่ทางผู้วิจัยพบกับข้อมูลในชุดนี้คือปัญหา Overfitting ซึ่งแบบจำลองที่นำมาใช้ในการทดลองกับการจัดการข้อมูลทั้ง 3 วิธีการก็ยังให้ผลลัพธ์ที่แบบจำลอง Overfitting กับข้อมูลในคลาสที่มีข้อมูลส่วนมาก(คลาสที่ไม่ทุจริตเคลม) อยู่ดีโดยผู้วิจัยได้พิจารณาการวัดประสิทธิภาพจากค่า Accuracy, Precision, Recall, F1-Score ร่วมกับ Confusion Matrix โดยค่า Accuracy ในแต่ละแบบจำลองให้ผลลัพธ์ที่ดีมาก แต่ค่า F1-Score ให้ผลลัพธ์ที่ไม่ค่อยดี เมื่อได้มาดูค่า Confusion Matrix แล้วพบว่าค่า TN(True Negative) ออกมาค่อนข้างแม่นยำ แต่ค่า TP(True Positive) ออกมาค่อนข้างไม่แม่นยำเท่าที่ควร แบบจำลองเลยเอนเอียงไปในทางทำนายคลาสที่ไม่ทุจริตเคลม(คลาส 0) ได้ดีกว่าคลาสที่ทุจริตเคลม(คลาส 1) ที่เราสนใจ

ผู้วิจัยได้เล็งเห็นว่าแบบจำลองที่ผู้วิจัยเลือกมานั้นเป็นแบบจำลองที่ยังไม่สามารถอธิบายวิธีการที่แบบจำลองเลือกคุณลักษณะและเงื่อนไขอะไรนำมาใช้ให้กับผู้ใช้งานทราบได้ ส่งผลให้ผู้ใช้งานอาจจะยังไม่มั่นใจในผลลัพธ์ที่แบบจำลองทำนายผลออกมาและผู้ใช้ต้องการเห็นเงื่อนไขที่แบบจำลองเลือกมาใช้ด้วย เพื่อทำให้เกิดความง่ายและทำให้เกิดความเข้าใจของผู้ใช้เราควรเลือกแบบจำลองประเภท Whitebox ที่สามารถอธิบายให้เห็นถึงวิธีการที่แบบจำลองเลือกคุณลักษณะมาใช้งาน เช่น Decision Tree แต่แบบจำลองประเภทนี้ อาจจะให้ประสิทธิภาพการของการทำนายผลไม่ดีเท่าแบบจำลองประเภท Backbox เช่น Deep Neural Network ซึ่งผู้ใช้งานจะไม่เห็นเงื่อนไขที่แบบจำลองเลือกมาใช้ได้เลย แต่ถ้าเราเลือกแบบจำลองประเภท Whitebox มาใช้ผู้ใช้สามารถอ่านค่าและเห็นถึงสิ่งที่แบบจำลองเลือกมาใช้ประกอบการตัดสินใจเพื่อจะได้ให้น้ำหนักของแบบจำลองของเราได้

ข้อเสนอแนะ

จากปัญหาที่เราพบในการดำเนินการวิจัยในครั้งนี้ทั้งหมด ผู้วิจัยได้เล็งเห็นว่าในขั้นตอนกระบวนการทำ Preprocess และ Text Preprocess มีความสำคัญมาก ในงานวิจัยนี้สามารถที่จะจัดการกับข้อมูลในขั้นตอนนี้เพิ่มเติมได้อีกและเพื่อเพิ่มประสิทธิภาพของคุณลักษณะที่เป็นข้อความผู้วิจัยคิดว่าจำเป็นต้องสกัด Information ออกมาจากคุณลักษณะที่เป็นข้อความและนำมาสร้างให้เป็นคุณลักษณะที่เป็น Categorical Data

เพื่อให้สามารถระบุความสำคัญของคำที่จะนำไปสู่การทุจริตได้ ผู้วิจัยอาจจะต้องขอให้ทางเจ้าหน้าที่สินไหมระบุคำเฉพาะ(Keyword) สำคัญที่ใช้ในการระบุว่าเคลมนั้นเป็นการทุจริตเพื่อนำมาสร้าง Dictionary คำที่มีความทุจริตและนำมาใช้ในการให้น้ำหนักของคำในข้อความด้วย

ในส่วนของการทำ Feature Importance สามารถช่วยให้เราเลือกคุณลักษณะที่มีความสำคัญในงานวิจัยออกมาได้อย่างดีซึ่งอาจจะพิจารณาทำเพิ่มเติมกับคุณลักษณะที่ไม่ใช่ข้อความ

ในส่วนของการลดขนาดของ Feature ลงทางผู้วิจัยได้พิจารณาแล้วว่างานวิจัยสามารถต่อยอดนำเทคนิคการจัดกลุ่มของประโยคและกลุ่มของคำโดยใช้วิธีการ LDA(Latent Dirichlet Allocation) เข้ามาทดลองเพิ่มเติม

ในวิธีการที่ผู้วิจัยคาดว่าจะนำมาทดลองกับการจัดการข้อมูลที่ไม่สมดุลกันของชุดข้อมูลนี้คือการทำ Under sampling Technique ของคลาสที่ไม่เป็นการทุจริตเคลม

ปัญหาอีกอย่างที่ผู้ใช้ทั่วไปอาจจะยังไม่มั่นใจในการทำนายผลของแบบจำลองคือผู้ใช้ไม่รู้เลยว่าผลลัพธ์ที่แบบจำลองทำนายออกมาว่าเป็นการทุจริตนั้นแบบจำลองใช้ข้อมูลอะไรมาตัดสินใจว่าเคลมรายการนี้เป็นการทุจริตเคลมหรือไม่ทุจริตเคลม ผู้วิจัยเลยอยากจะเลือกแบบจำลองที่สามารถอธิบายที่มาของคุณลักษณะและเงื่อนไขที่แบบจำลองได้นำมาใช้สามารถทำให้ผู้ใช้งานเข้าใจในเงื่อนไขที่แบบจำลองเลือกมา โดยจะนำอัลกอริทึม Decision Tree เข้ามาใช้เพื่อทำการทดลองเพิ่มเติมและสามรถอธิบายเงื่อนไขกับคุณลักษณะที่แบบจำลองเลือกมาได้

บรรณานุกรม

(คปภ.), ส. (2020). กรอบการลงทุนตามความเสี่ยง(Risk-Base Capital Framework). สืบค้นจาก https://www.oic.or.th/sites/default/files/3029-9267-2.pdf

(คปภ.), ส. (2563). รายงานภาวะธุรกิจประกันภัยไทย ประจำปี 2563. สืบค้นจาก https://www.oic.or.th/th/industry/statistic/data/39/2

Aninditya, A., Hasibuan, M. A., และ Sutoyo, E. (2019, 5-7 Nov. 2019). *Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy*. Paper presented at the 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS).

Arreerard, R., และ Senivongse, T. (2018, 12-13 July 2018). *Thai Defamatory Text Classification on Social Media*. Paper presented at the 2018 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD).

Brownlee, J. (2014). An Introduction to Feature Selection. Retrieved from https://machinelearningmastery.com/an-introduction-to-feature-selection/

chengz. (2019). วัดประสิทธิภาพ Model จาก Confusion Matrix. สืบค้นจาก https://medium.com/@cheng3374/วัดประสิทธิภาพ-model-จาก-confusion-matrix-69d391bcd48

Das, A. (2019). Oversampling to remove class imbalance using SMOTE. *Transportation Research Part C: Emerging Technologies*. Retrieved from https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35

Ganesan, K. (2019). All you need to know about text preprocessing for NLP and Machine Learning. Retrieved from https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

Goswami, D. S. (2020). Class Imbalance, SMOTE, borderline SMOTE, ADASYN. Retrieved from https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804

Goyal, K. (2021). Data Preprocessing in Machine Learning: 7 Easy Steps To Follow. Retrieve from https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

Harjai, S., Khatri, S. K., และ Singh, G. (2019, 21-22 Nov. 2019). *Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique.* Paper presented at the 2019 4th International Conference on Information Systems and Computer Networks (ISCON).

Kumar, A. (2020). K-Fold Cross Validation – Python Example. Retrieved from https://vitalflux.com/k-fold-cross-validation-python-example/

Prasasti, I. M. N., Dhini, A., และ Laoh, E. (2020, 17-18 Oct. 2020). *Automobile Insurance Fraud Detection using Supervised Classifiers.* Paper presented at the 2020 International Workshop on Big Data and Information Security (IWBIS).

Prasertsom, P. (2020). สกัดใจความสำคัญของข้อความด้วยเทคนิคการประมวลผลทางภาษาเบื้องต้น: TF-IDF. สืบค้นจาก https://bigdata.go.th/big-data-101/tf-idf-1/

Wang, Y., และ Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87-95.

Yuenyong, S., และ Sinthupinyo, S. (2020). Gender classification of thai facebook usernames. *International Journal of Machine Learning and Computing, 10*(5).

กองแผนงานกรมการขนส่งทางบก, ก. (2559 - 2563). รายงานสถิติการขนส่ง ปีงบประมาณ 2559 – 2563. สืบค้นจาก https://web.dlt.go.th/statistics/

ประวัติผู้เขียน

ชื่อ-สกุล	ภูริต อำนวยชัย
วัน เดือน ปี เกิด	16 มิถุนายน 2527
สถานที่เกิด	สมุทรปราการ
วุฒิการศึกษา	วิทยาศาสตร์บัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์