รายงานสหกิจศึกษาฉบับสมบูรณ์

การสร้างแบบจำลองการหาความผิดปกติของข้อมูลระบบบันทึกข้อมูลพฤติกรรมการขับขี่ โดยการทำเหมืองข้อมูล

The Error Data of In-Vehicle Monitoring System Model by Using Data Mining

นางสาวอังคณา อัครวราวงศ์

ภาควิชาวิศวกรรมคอมพิวเตอร์ สาขาวิชาวิศวกรรมสารสนเทศ
คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2559

รายงานสหกิจศึกษาฉบับสมบูรณ์

การสร้างแบบจำลองการหาความผิดปกติของข้อมูลระบบบันทึกข้อมูลพฤติกรรมการขับขี่ โดยการทำเหมืองข้อมูล

The Error Data of In-Vehicle Monitoring System Model by Using Data Mining

นางสาวอังคณา อัครวราวงศ์

ภาควิชาวิศวกรรมคอมพิวเตอร์ สาขาวิศวกรรมสารสนเทศ

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2559

ชื่อโครงงานสหกิจศึกษา การสร้างแบบจำลองการหาความผิดปกติของข้อมูลระบบบันทึกข้อมูลพฤติกรรมการขับขี่ โดยการทำเหมืองข้อมูล

ชื่อ-สกุล นักศึกษา นางสาวอังคณา อัครวราวงศ์

คณะ วิศวกรรมศาสตร์ **ภาควิชา** วิศวกรรมคอมพิวเตอร์ **สาขา** วิศวกรรมสารสนเทศ

ชื่อ-สกุล อาจารย์นิเทศ ผศ.ดร.สุธีรา พันธุ์ธีรานุรักษ์

ชื่อ-สกุล ผู้นิเทศงาน นายปุญญนันท์ ปิ่นสุข

สถานประกอบการ บริษัท พีทีที ไอซีที โซลูชั่นส์ จำกัด

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่สามารถสนับสนุนการทำงานของระบบภายในองค์กรให้มีความน่าเชื่อถือและนำเสนอการเปรียบเทียบประสิทธิภาพของแบบจำลองในการจำแนกประเภทข้อมูล โดยใช้เทคนิคการทำเหมืองข้อมูล ในการพัฒนาแบบจำลองและทำการเปรียบเทียบประสิทธิภาพ โดยใช้ข้อมูลติดตามสถานะของรถขนส่งผลิตภัณฑ์ (In-Vehicle Monitoring System: IVMS) จากฐานข้อมูลส่วนกลางของระบบนำมาวิเคราะห์ และใช้วิธีการจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีต้นไม้ตัดสินใจ โดยใช้อัลกอริทึมแบบ C4.5 ป่าของต้นไม้ตัดสินใจ และโครงข่ายประสาทเทียมแบบหลายระดับ การวัดประสิทธิภาพสามารถวัดได้จากความถูกต้องของการจำแนกประเภทของข้อมูลโดยนับจากค่าความถูกต้องของการจำแนกประเภทข้อมูลที่วัดได้ ซึ่งการทดสอบแบบจำลองที่ได้จะทำการทดสอบผลบนพื้นฐานวิธี 10 – fold Cross Validation โดยผลการทดลองที่ได้ พบว่า การใช้เทคนิคป่าของต้นไม้ตัดสินใจในการจำแนกข้อมูลนั้นจะมีประสิทธิภาพที่ดีกว่าการใช้แบบจำลองแบบอื่น ๆ ซึ่งผลวิจัยที่ได้สามารถนำไปใช้ในการสนับสนุนการทำงานของระบบภายในองค์กรให้มีความน่าเชื่อถือและความถูกต้องของระบบที่เพิ่มมากขึ้น

คำสำคัญ : เหมืองข้อมูล การจำแนกประเภทข้อมูล ต้นไม้ตัดสินใจ ป่าของต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม

Cooperative Title: The Error Data of In-Vehicle Monitoring System Model by Using Data Mining

Student intern name: Miss Angkana Akarawarawong

Faculty: Engineering **Department:** Computer Engineering **Program:** Information Engineering

Advisor name: Asst.Prof.Dr.Sutheera Puntheeranurak

Mentor name: Mr.Punyanan Pinsuk

Company: PTT ICT Solutions Company Limited

Abstract

This research aims to develop a model which can support the work of the organization are reliable and represented comparing the efficiency of models to classify. Using data mining to develop models and compare the efficiency. We use GPS Tracking data set from In-Vehicle Monitoring System that is in IVMS database. The experiments used the classified algorithm, Decision Tree (C4.5), Random Forest and Neural Network (Multilayer Perceptron: MLP). The accuracy rate of classification used for evaluation efficiency. Moreover, the 10-fold cross-validation is used to testing model. The result of the experiment shows Random Forest for classification that high efficiency more than Decision Tree and Neural Network. The research results can be applied to support the work of the organization with the credibility and accuracy of the system.

Keywords: Data mining, Classification, Decision Tree, Random forest, Neural network, 10-fold cross validation

กิตติกรรมประกาศ

งานวิจัยฉบับนี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอขอบพระคุณอาจารย์ ผศ.ดร.สุธีรา พันธุ์ธีรานุรักษ์ อาจารย์ที่ปรึกษาที่ให้คำแนะนำ และสละเวลาในการตรวจสอบและแก้ไขตลอดการดำเนินงานวิจัยครั้งนี้ และขอขอบพระคุณบริษัท พีทีที ไอซีที โซลูชั่นส์ จำกัด และบุคลากรที่เกี่ยวข้อง ที่ได้ให้โอกาสและความรู้ในการทำงานวิจัยครั้งนี้ นอกจากนี้ขอกราบขอบคุณครู อาจารย์ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัย จนประสบความสำเร็จ

ท้ายนี้ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา ที่ให้การอุปการะอบรมเลี้ยงดู ตลอดจนส่งเสริมการศึกษา และให้กำลังใจเป็นอย่างดี อีกทั้งขอขอบคุณมิตรสหายที่ให้การสนับสนุนและช่วยเหลือด้วยดีเสมอมา และขอขอบคุณเจ้าของเอกสารและงานวิจัยทุกท่าน ที่ผู้ศึกษาค้นคว้าได้นำมาอ้างอิงในการทำวิจัย จนกระทั่งงานวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

นางสาวอังคณา อัครวราวงศ์

ผู้วิจัย

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

ระบบบันทึกข้อมูลพฤติกรรมการขับขี่ (In-Vehicle Monitoring System: IVMS) เป็นระบบติดตามสถานะรถขนส่งน้ำมัน โดยทำการติดตั้งอุปกรณ์ตัวรับส่งสัญญาณระบบกำหนดตำแหน่งบนโลก (Global Positioning System) และกล้องวงจรปิด (Closed Circuit Television) ที่ตัวรถ เพื่อเป็นอุปกรณ์ติดตามและบันทึกข้อมูลการใช้งานรถและพฤติกรรมการขับขี่ของพนักงานขับรถ เช่น การใช้ความเร็ว การเบรกกะทันหัน การขับรถกระชั้นชิด การเหยียบคันเร่งออกรถอย่างกระชาก ลักษณะการขับขี่ก่อนเกิดอุบัติเหตุ เป็นต้น การส่งข้อมูลจากอุปกรณ์ระบบกำหนดตำแหน่งบนโลกและกล้องวงจรปิด มายังฐานข้อมูลส่วนกลางจะส่งโดยผ่านสัญญาณอินเทอร์เน็ต

การมอนิเตอร์สถานะของรถแต่ละคันจะทำการมอนิเตอร์ผ่านเว็บแอปพลิเคชัน (Web Application) โดยดึงข้อมูลจากฐานข้อมูลส่วนกลาง โดยตำแหน่งของรถจะถูกแสดงบนแผนที่ ซึ่งจะมีพนักงานที่ศูนย์ควบคุมการขนส่งผลิตภัณฑ์ปิโตรเลียม (Transportation Control Center: TCC) เป็นผู้ตรวจสอบและติดตามการขับขี่รถขนส่งผลิตภัณฑ์ทั่วประเทศตลอด 24 ชั่วโมง ซึ่งการติดตั้งระบบบันทึกข้อมูลพฤติกรรมการขับขี่นี้ เพื่อเป็นเครื่องมือในการบริหารจัดการรถขนส่งในกลุ่มปิโตเลียมแห่งประเทศไทย ให้มีประสิทธิภาพมากขึ้น และลดความสูญเสียจากอุบัติเหตุที่เกิดจากการขนส่งให้เหลือน้อยที่สุด

เนื่องจากข้อมูลในระบบติดตามรถขนส่งผลิตภัณฑ์ที่ถูกส่งเข้ามาในแต่ละวันมีจำนวนมาก ทำให้ข้อมูลบางส่วนมีความผิดพลาด อาจเกิดจากตัวของอุปกรณ์เอง โดยตัวอย่างข้อมูลที่มีความผิดพลาด เช่น เครื่องยนต์ของรถขนส่งน้ำมันไม่มีการใช้งานหรือไม่ได้สตาร์ทเครื่องยนต์ แต่มีความเร็วมากกว่าศูนย์ กม./ชม. ดังภาพที่ 1.1 ซึ่งในความเป็นจริงไม่สามารถเกิดเหตุการณ์เช่นนี้ได้ ถ้าเกิดข้อมูลในระบบเป็นเช่นนี้มากจะทำให้ระบบไม่มีความน่าเชื่อถือ จึงจำเป็นต้องมีระบบคัดแยกข้อมูลที่ผิดปกติออกจากข้อมูลปกติ โดยนำเทคนิคการทำเหมืองข้อมูล (Data Mining) [1] เข้ามาพัฒนาระบบ โดยนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผล เพื่อสร้างแบบจำลอง ซึ่งเทคนิคการทำเหมืองข้อมูลมีอยู่หลายเทคนิค เช่น โครงข่ายประสาทเทียม (Neural Network) ต้นไม้ตัดสินใจ (Decision Tree) เป็นต้น การเลือกใช้ควรเลือกใช้ให้เหมาะสมกับเป้าหมายหรือปัญหาตามความเหมาะสม โดยทั่วไปประเภทของงานตามลักษณะแบบจำลองที่ใช้ในการทำเหมืองข้อมูลสามารถแบ่งกลุ่มได้เป็น 2 ประเภทใหญ่ ๆ คือ 1) การเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ เทคนิคการเรียนรู้โดยใช้ข้อมูลที่ผ่านมาในอดีตเป็นผู้สอนในการสร้างแบบจำลอง และ 2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คือ เทคนิคหนึ่งของการเรียนรู้โดยการสร้างแบบจำลองที่เหมาะสมกับข้อมูล การเรียนรู้แบบนี้ต่างจากการเรียนรู้แบบมีผู้สอนคือ จะไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน โดยส่วนมากจะเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูล

GPS_ID	GPS_TIME	DB_TIME	SPEED	ENGINE_STAT	DRIVER_LIC_INFO	HDOP	SAT_NO	RSSI	INT_BATT_VDC	EXT_BATT_VDC
[redacted]	28/10/2016 6:00:04.973000	28/10/2016 6:00:08.145000	39	0	[redacted]	7	12	17	4.1	27

ภาพที่ 1.1 ตัวอย่างของข้อมูลที่เครื่องยนต์ไม่ทำงานแต่มีความเร็วเครื่องยนต์มากกว่าศูนย์ กม./ชม.

เทคนิคการจำแนกประเภทข้อมูล (Classification) [2] เป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) หรือเหมืองข้อมูล [1] จุดประสงค์ของการจำแนกประเภทข้อมูลคือ การสร้างแบบจำลองการแยกคุณสมบัติหรือแอตทริบิวท์ (Attribute) หนึ่งโดยขึ้นกับแอตทริบิวท์อื่น ซึ่งแบบจำลองที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้ เทคนิคการจำแนกประเภทข้อมูลนี้สามารถนำไปประยุกต์ใช้ได้กับงานในหลาย ๆ ด้าน

ในงานวิจัยฉบับนี้จะเน้นการนำเสนอวิธีการจำแนกประเภทข้อมูล ซึ่งวิธีการที่เป็นที่นิยมในการนำมาประยุกต์ใช้ในการจำแนกประเภทข้อมูลวิธีที่หนึ่งก็คือ ต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึม C4.5 วิธีที่สองคือ ป่าของต้นไม้ตัดสินใจ และวิธีที่สามคือ โครงข่ายประสาทเทียมชนิดเพอร์เซ็ปตรอนแบบหลายชั้น (Multilayer Perceptron: MLP) เข้ามาเปรียบเทียบประสิทธิภาพ โดยนำข้อมูลระบบติดตามสถานะรถขนส่งน้ำมัน จากฐานข้อมูลมาทำการทดสอบการวิจัยในครั้งนี้ โดยเนื้อหาในบทความได้แบ่งเป็นบทดังนี้ บทที่ 2 กล่าวถึงแนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง บทที่ 3 วิธีดำเนินการวิจัย บทที่ 4 ผลการวิจัย และบทที่ 5 กล่าวถึงการสรุปผลการวิจัยและข้อเสนอแนะ

1.2 วัตถุประสงค์ของการวิจัย

1) เพื่อสร้างแบบจำลองที่สามารถคัดแยกข้อมูลระบบติดตามสถานะรถขนส่งที่ผิดปกติออกจากข้อมูลปกติได้โดยใช้เทคนิคทางเหมืองข้อมูล

2) เพื่อเพิ่มประสิทธิภาพของระบบบันทึกข้อมูลพฤติกรรมการขับขี่ให้มีความถูกต้องและมีความน่าเชื่อถือ

3) เพื่อเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองของแต่ละเทคนิค โดยวัดจากค่าความถูกต้องของการทำนาย

1.3 ขอบเขตของการวิจัย

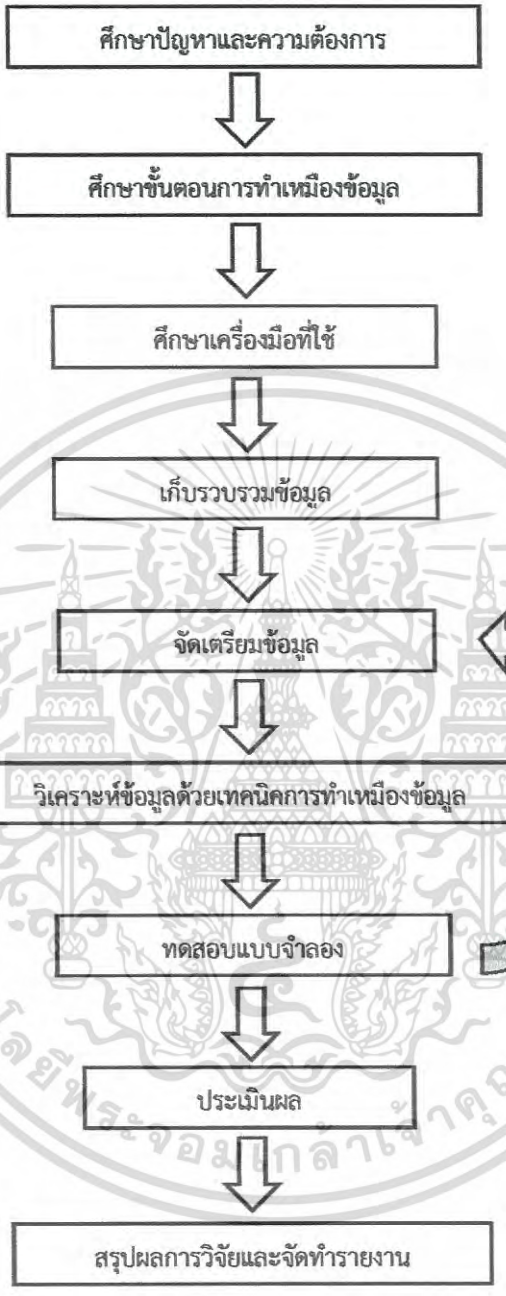
1) วิเคราะห์และหาแบบจำลองที่มีความสามารถคัดแยกข้อมูลที่ไม่ปกติออกจากข้อมูลปกติ โดยให้เหมาะสมแก่ระบบติดตามรถขนส่งน้ำมัน

2) ในการทำวิจัยครั้งนี้ได้ทำการรวบรวมข้อมูลทั้งหมด 402,000 ตัวอย่าง ภายในเดือนตุลาคม พ.ศ.2559 เพื่อใช้เป็นข้อมูลในการทำนายหาแบบจำลองที่ใช้ในการคัดแยกข้อมูลผิดปกติของระบบติดตามรถขนส่งน้ำมัน

3) วิธีที่ใช้ในการทำเหมืองข้อมูลในการวิเคราะห์หาแบบจำลองด้วยเทคนิคการจำแนกประเภท และในการทดสอบใช้ทั้งหมด 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ ป่าของต้นไม้ตัดสินใจ และโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น

4) วัดประสิทธิภาพของแต่ละเทคนิคโดยดูจากค่าความถูกต้องที่พิจารณาจากทุกคลาส (Accuracy) ค่าความแม่นยำ (Precision) ค่าความถูกต้องโดยพิจารณาแยกทีละคลาส (Recall) และ การวัดค่าความแม่นยำ และค่าความถูกต้องโดยพิจารณาแยกทีละคลาสพร้อมกัน (F-measure) โดยนำค่าความถูกต้องที่คำนวณได้จากแต่ละเทคนิคมาเปรียบเทียบกัน

1.4 วิธีการดำเนินการวิจัย



ภาพที่ 1.2 แผนผังวิธีการดำเนินงานวิจัย

จากภาพที่ 1.2 สามารถอธิบายขั้นตอนการดำเนินงานได้ดังนี้

1) ศึกษาปัญหาที่เกิดขึ้นและความต้องการของบริษัท เป็นขั้นตอนแรกของการทำงานวิจัย โดยทำการศึกษาถึงปัญหาที่เกิดขึ้นและความต้องการของบริษัท เพื่อการพัฒนาได้ตรงเป้าหมายที่ต้องการ

2) ศึกษาขั้นตอนการทำเหมืองข้อมูล เป็นขั้นตอนการศึกษาเกี่ยวกับการทำเหมืองข้อมูล เพื่อที่จะนำเทคนิคการทำเหมืองข้อมูลมาใช้ในการวิเคราะห์ข้อมูลของการทำวิจัย

3) ศึกษาเครื่องมือที่ใช้ในการทำเหมืองข้อมูล เป็นวิธีขั้นตอนการศึกษาถึงเครื่องมือ หรือโปรแกรมที่นำมาใช้ในการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูล

4) เก็บรวบรวมข้อมูลเพื่อการวิเคราะห์ เป็นการเก็บรวบรวมข้อมูลที่จะใช้ในการนำไปวิเคราะห์เป็นข้อมูลชุดฝึกสอนในการทำเหมืองข้อมูล

5) จัดเตรียมข้อมูลเพื่อการทำเหมืองข้อมูล เป็นขั้นตอนการจัดเตรียมข้อมูลให้พร้อมก่อนที่จะนำไปวิเคราะห์ เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพ โดยการกำจัดค่าว่าง หรือ ค่าที่ผิดปกติออกก่อน

6) วิเคราะห์ข้อมูลตามวิธีการทำเหมืองข้อมูลที่ได้กำหนดไว้โดยใช้เครื่องมือในการทำเหมืองข้อมูล ในขั้นตอนนี้ เป็นการนำชุดข้อมูลที่พร้อมสำหรับทำการวิเคราะห์ นำไปวิเคราะห์ด้วยโปรแกรม ตามเทคนิคการทำเหมืองข้อมูล

7) ทดสอบแบบจำลองที่ได้ว่ามีความถูกต้องเพียงใด เป็นขั้นตอนการทดสอบแบบจำลองที่ได้ว่ามีความถูกต้องหรือไม่ หรือถูกต้องมากน้อยแค่ไหน ขั้นตอนนี้อาจจะต้องมีการย้อนกลับไปขั้นตอนการเตรียมข้อมูล เพื่อแปรงข้อมูลบางส่วนให้เหมาะสม

8) ประเมินความถูกต้องของผลลัพธ์ที่ได้ ทำการประเมินผลของแบบจำลองที่ได้ ด้วยการทำการทดสอบแบบจำลอง ตามวิธีที่เลือกใช้ในการทดสอบ ว่าแบบจำลองที่ได้มีความน่าเชื่อถือมากน้อยเพียงใด

9) สรุปผลการวิจัยและจัดทำรายงานฉบับสมบูรณ์ เป็นขั้นตอนสุดท้ายของการทำวิจัย เพื่อเป็นการสรุปผลที่ได้จากการทำวิจัยและจัดทำรายงานการทำงานวิจัย

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) เพื่อได้แบบจำลองความสามารถในการคัดแยกข้อมูลที่ผิดปกติออกจากข้อมูลปกติในระบบบันทึกข้อมูลพฤติกรรมการขับขี่ ให้สามารถนำไปใช้ได้กับข้อมูลในอนาคต เพื่อให้เกิดประสิทธิภาพและประสิทธิผลแก่องค์กร

2) ได้เครื่องมือที่สนับสนุนการคัดแยกข้อมูลตามแบบจำลองการคัดแยกข้อมูลผิดปกติออกจากข้อมูลปกติ

3) มีอัลกอริทึมที่เหมาะสมในการสร้างแบบจำลอง

4) สามารถนำแบบจำลองการคัดแยกข้อมูลผิดปกติออกจากข้อมูลปกติ และเครื่องมือสนับสนุนแบบจำลองไปประยุกต์ใช้ในระบบอื่น ๆ ขององค์กร

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 เหมืองข้อมูล (Data Mining)

เหมืองข้อมูล [1] คือ การค้นหาความสัมพันธ์และรูปแบบทั้งหมดซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก การทำเหมืองข้อมูลจะทำการสำรวจและวิเคราะห์อย่างอัตโนมัติในปริมาณข้อมูลจำนวนมากให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่าง ๆ ที่มีประโยชน์ในฐานข้อมูล

กระบวนการค้นพบความรู้จากฐานข้อมูลขนาดใหญ่มาก เป็นกระบวนการสร้างแบบจำลองหรือรูปแบบกฎเกณฑ์จากกลุ่มของข้อมูล ทำให้เกิดความเข้าใจลักษณะรูปแบบความเกี่ยวข้องสัมพันธ์กันของกลุ่มข้อมูล แล้วแนวโน้มเพื่อใช้ในการทำนายข้อมูลนั้น ๆ โดยมีกระบวนการตามทฤษฎี รวม 4 ขั้นตอน ดังนี้

2.1.1 การเตรียมข้อมูล (Data preparation)

การเตรียมข้อมูลเป็นขั้นตอนสำคัญ และใช้เวลามากที่สุด เนื่องจากบางครั้งมีการเลือกข้อมูลมาไม่เหมาะสมและไม่ถูกต้อง หรือการนำข้อมูลมาจากหลายแหล่งที่มารวมเข้าด้วยกันเพื่อพิจารณาความสัมพันธ์ของข้อมูล ซึ่งส่งผลให้เกิดความผิดพลาด ดังนั้นขั้นตอนการเตรียมข้อมูลจึงถือเป็นส่วนสำคัญของงาน การเตรียมข้อมูลสามารถแบ่งออกได้เป็น 3 ขั้นตอนย่อย คือ

1) การคัดเลือกข้อมูล (Data Selection) จุดประสงค์หลัก คือ การระบุลักษณะข้อมูลที่ต้องการ แล้ว ทำการคัดเลือกข้อมูลที่ต้องการ ซึ่งข้อมูลที่ได้จะแตกต่างกันไปตามจุดประสงค์ของแต่ละธุรกิจ

2) การกลั่นกรองข้อมูล (Data Cleaning) จุดประสงค์เพื่อมั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นถูกต้องและเหมาะสม เนื่องจากข้อมูลที่ถูกเลือกมาจากกระบวนการเลือกข้อมูลนั้นอาจมีข้อมูลไม่ถูกต้อง

3) การแปลงรูปข้อมูล (Data Transformation) จุดประสงค์เพื่อแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมจะนำไปวิเคราะห์ตามหลักอัลกอริทึมของการทำเหมืองข้อมูลที่เลือกใช้ เช่น การแบ่งช่วงอายุให้เป็นกลุ่ม ๆ หรือ กำหนดตัวเลขให้กับแต่ละกลุ่มของข้อมูล

2.1.2 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล เป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้ ในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนการเตรียมข้อมูล โดยเมื่อทำขั้นตอนนี้แล้วอาจต้องย้อนกลับไปทำขั้นตอนการเตรียมข้อมูลใหม่

2.1.3 การประเมินรูปแบบ (Pattern evaluation)

การประเมินรูปแบบ เป็นขั้นตอนการวิเคราะห์และประเมินผลของรูปแบบหรือกฎเกณฑ์ที่หาได้จากขั้นตอนการหาความรู้จากข้อมูล การทำงานในส่วนนี้จำเป็นต้องใช้ทักษะในการวิเคราะห์ข้อมูลทางธุรกิจเข้าช่วย

2.1.4 การนำเสนอความรู้ (Knowledge presentation)

การนำเสนอความรู้ เป็นการนำความรู้ที่ค้นพบไปประยุกต์ใช้ต่อไป

2.2 เทคนิคในการทำเหมืองข้อมูล (Data Mining Techniques)

เทคนิคที่ใช้ในการทำเหมืองข้อมูลอาศัยหลักการทางสถิติในการวิเคราะห์ข้อมูล โดยมี 3 เทคนิคสำคัญที่เป็นที่แพร่หลายในปัจจุบันดังต่อไปนี้

2.2.1 กฎความสัมพันธ์ (Association Rule)

กฎความสัมพันธ์ เป็นการค้นหากฎความสัมพันธ์ของข้อมูล โดยค้นหาความสัมพันธ์ของข้อมูลทั้งสองชุดหรือมากกว่าสองชุดขึ้นไปไว้ด้วยกัน ความสำคัญของกฎทำการวัดโดยใช้ข้อมูลสองตัวด้วยกันคือค่าสนับสนุน (Support) ซึ่งเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎสามารถนำไปใช้ หรือเป็นเปอร์เซ็นต์ของการดำเนินการที่กฎที่ใช้มีความถูกต้อง และข้อมูลตัวที่สองที่นำมาใช้วัดคือค่าความมั่นใจ (Confidence) ซึ่งเป็นจำนวนของกรณีที่กฎถูกต้องโดยสัมพันธ์กับจำนวนของกรณีที่กฎสามารถนำไปใช้ได้ ในการหากฎความสัมพันธ์นั้นจะมีขั้นตอนวิธีการหาหลายวิธีด้วยกัน แต่ขั้นตอนวิธีที่เป็นที่รู้จักและใช้อย่างแพร่หลายคือ อัลกอริทึมอะพลิโอริ (Apriori algorithm)

2.2.2 การจัดกลุ่มข้อมูล (Clustering)

การจัดกลุ่มข้อมูล เป็นการจัดกลุ่มข้อมูลซึ่งมีลักษณะคล้ายกับการแบ่งประเภทแต่จะไม่เหมือนกัน โดยการแบ่งประเภทจะวิเคราะห์ข้อมูลตามต้นแบบ แต่สำหรับการแบ่งกลุ่มเป็นการวิเคราะห์โดยไม่พิจารณาจัดกลุ่มตามประเภทที่มีหรือที่รู้จัก แต่จะใช้ขั้นตอนวิธีการจัดกลุ่มเพื่อค้นหากลุ่มที่สามารถยอมรับได้เพื่อจัดเข้ากลุ่ม กล่าวคือ กลุ่มของวัตถุมีการสร้างขึ้นโดยเปรียบเทียบวัตถุที่มีความเหมือนกันจัดเข้ากลุ่มเดียวกัน

2.2.3 การจำแนกประเภทของข้อมูล (Classification)

การจำแนกประเภทของข้อมูล เป็นการจัดแบ่งประเภทของข้อมูล โดยหาชุดต้นแบบหรือชุดของการทำงานที่อธิบายและแบ่งประเภทข้อมูล วัตถุประสงค์เพื่อให้สามารถใช้เป็นต้นแบบทำนายประเภทของวัตถุหรือข้อมูลที่ไม่มีการระบุประเภทหรือชนิดของข้อมูล ซึ่งต้นแบบสร้างจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (Training Data) โดยอาจจะเป็นกลุ่มข้อมูลที่มีการระบุประเภทหรือกลุ่มเรียบร้อยแล้ว รูปแบบของต้นแบบแสดงได้หลายแบบเช่น กฎการจัดแบ่งประเภทของข้อมูล ต้นไม้ตัดสินใจ หรือ โครงข่ายประสาทเทียม เป็นต้น

2.3 เทคนิคการจำแนกประเภทข้อมูล (Classification)

การจำแนกประเภทข้อมูล [2] คือ วิธีการแยกข้อมูลออกเป็นแต่ละประเภทตามลักษณะของข้อมูลเป้าหมาย หรือเรียกว่า คลาส ข้อมูลที่นำมาใช้ในการจำแนกประกอบด้วย 2 ส่วน คือ ชุดข้อมูลที่ใช้ในการฝึกฝน (Training set) และชุดข้อมูลที่ใช้ในการทดสอบ (Test set) ผลลัพธ์ที่ได้คือแบบจำลองที่ใช้ในการจำแนกข้อมูล จุดประสงค์ของการจัดจำแนก คือ สามารถนำแบบจำลองที่สร้างขึ้นมาทำนายข้อมูลที่ไม่เคยพบมาก่อน หรือข้อมูลในอนาคต แล้วได้ผลลัพธ์ถูกต้องแม่นยำเป็นที่น่าพอใจ

การจำแนกประเภทข้อมูลเป็นกระบวนการสร้างแบบจำลองจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดไว้ โดยการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่ เพื่อใช้ทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น โดยเทคนิคการจำแนกประเภทข้อมูลที่ใช้ในงานวิจัยมีทั้งหมด 3 เทคนิค ได้แก่

2.3.1 ต้นไม้ตัดสินใจ (Decision Tree)

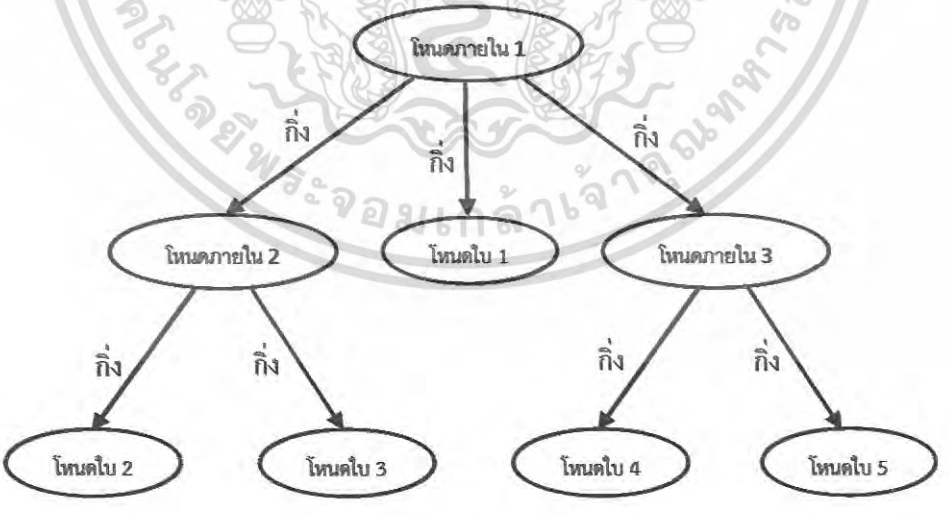
ต้นไม้ตัดสินใจ [3] เป็นการนำข้อมูลมาสร้างแบบจำลอง มีลักษณะเป็นผังงาน (flowchart) เหมือนโครงสร้างต้นไม้ เป็นการเรียนรู้แบบมีผู้สอน (supervised learning) คือ จะสร้างแบบจำลองขึ้นมาจากข้อมูลที่นำมาใช้เรียนรู้ โดยต้นไม้ตัดสินใจสามารถนำมาใช้ในการทำนายค่าต่าง ๆ ได้ โดยผลการทำนายจะขึ้นอยู่กับตัวแปรต้น รูปแบบของต้นไม้ตัดสินใจประกอบด้วยโหนดแรกสุดที่เรียกว่า โหนดราก (root node) จากโหนดรากจะแตกออกเป็นโหนดลูกที่มีกิ่งในการเชื่อมระหว่างโหนด และที่โหนดลูกก็จะมีลูกของตัวเอง ซึ่งที่โหนดระดับสุดท้ายจะเรียกว่า โหนดใบ (leaf node) แต่ละโหนดของโหนดรากและโหนดลูกจะแสดงค่าคุณลักษณะ (attribute) ที่ใช้ทดสอบข้อมูล ส่วนโหนดใบจะแสดงกลุ่ม (class) ที่กำหนดไว้ เมื่อมีข้อมูลที่ต้องการทำนาย การทำงานจะเริ่มต้นจากโหนดราก ซึ่งจะนำค่าคุณลักษณะต่าง ๆ ของข้อมูลนั้นไปเปรียบเทียบกับคุณลักษณะของโหนด และทำการตัดสินใจว่าจะเดินทางไปตามกิ่งใด หลังจากนั้นจะเดินทางผ่านโหนดลูก และทำการเปรียบเทียบคุณลักษณะไปเรื่อย ๆ จนกระทั่งสุดท้ายไปถึงโหนดใบ ก็จะได้กลุ่มที่ถูกกำหนด

1) ส่วนประกอบของต้นไม้ตัดสินใจ

1.1) โหนดภายใน (Internal node) คือ คุณลักษณะต่าง ๆ ของข้อมูลใด ๆ ตกลงมาที่โหนด จะใช้คุณลักษณะนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก

1.2) กิ่ง (Branch) เป็นค่าคุณลักษณะในโหนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณลักษณะของโหนดภายในนั้น

1.3) โหนดใบ (Leaf node) คือ กลุ่มต่าง ๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล



ภาพที่ 2.1 โครงสร้างต้นไม้ตัดสินใจ

จากภาพที่ 2.1 สามารถอธิบายการทำงานของต้นไม้ตัดสินใจได้ คือ ที่โหนดภายใน 1 หรือโหนดราก จะสร้างเส้นเชื่อมหรือแตกกิ่งออกเป็นโหนดภายใน 2 โหนดภายใน 3 และโหนดใบ 1 ที่โหนดภายใน 2 ก็จะแตกกิ่งของตัวเองเป็นโหนดใบ 2 และโหนดใบ 3 ที่โหนดภายใน 3 จะแตกกิ่งออกเป็นโหนด

ใบ 4 และโหนดใบ 5 ซึ่งแต่ละโหนดของโหนดภายในจะแสดงค่าคุณลักษณะ และโหนดใบจะแสดงกลุ่มของผลลัพธ์ในการจำแนกข้อมูล

2) ขั้นตอนการสร้างต้นไม้ตัดสินใจ

2.1) หาคุณลักษณะที่สำคัญที่สุดมาแบ่งข้อมูลโดยคุณลักษณะนี้จะถูกตั้งให้เป็นโหนดราก

2.2) จากโหนดรากจะสร้างเส้นทางเชื่อมหรือกิ่งไปยังโหนดลูก โดยจำนวนเส้นทางเชื่อมจะเท่ากับจำนวนค่าที่เป็นไปได้ของคุณลักษณะของโหนดราก

2.3) ถ้าโหนดลูกเป็นกลุ่มของข้อมูลที่อยู่ในกลุ่มเดียวกันทั้งหมดให้หยุดการสร้างต้นไม้ แต่ถ้าโหนดลูกมีข้อมูลของหลายกลุ่มปะปนกัน ต้องสร้างโหนดลูกเพื่อจำแนกข้อมูลต่อไป โดยวนกลับไปทำขั้นตอนที่ 1 ซ้ำเพื่อเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นตัวแบ่งข้อมูลต่อไป

สำหรับความสำคัญของคุณลักษณะสามารถหาได้จากการคำนวณค่าการเพิ่มสารสนเทศ (Information gain)

3) ค่าการเพิ่มสารสนเทศ (Information Gain)

ค่าการเพิ่มสารสนเทศ ถูกนำมาใช้ในการเลือกคุณลักษณะในแต่ละโหนดของต้นไม้ตัดสินใจ โดยคุณลักษณะที่มีค่าการเพิ่มสารสนเทศสูงสุดจะถูกเลือกให้เป็นโหนดราก

$$Gain(S,A) = E(S) - \sum_{v=value(A)} \frac{|S_v|}{|S|} E(S_v) \quad (1)$$

เมื่อ

ตัวแปร S คือ ตัวอย่างที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลาย ๆ กรณี

ตัวแปร E คือ เอนโทรปีของตัวอย่าง

ตัวแปร A คือตัวแปรต้นที่พิจารณา

ตัวแปร $Value\ (A)$ คือ เซตของค่าของ A ที่เป็นไปได้

ตัวแปร S_v คือตัวอย่างที่ A มีค่า v ทั้งหมด

4) เอนโทรปี (Entropy)

เอนโทรปี คือ การคำนวณหาค่าความยุ่งเหยิงของข้อมูลกลุ่มหนึ่ง

$$Entropy(t_i) = -\sum_{i=0}^{n} p(t_i) \log_2 p(t_i) \quad (2)$$

2.3.2 ป่าของต้นไม้ตัดสินใจ (Random Forest)

ป่าของต้นไม้ตัดสินใจ [4] จะประกอบไปด้วยต้นไม้ตัดสินใจจำนวนหลายต้น แต่ละต้นเป็นอิสระต่อกัน ในการสร้างต้นไม้ตัดสินใจแต่ละต้นจะทำการสุ่มค่าคุณลักษณะ และแถวข้อมูล โดยต้นไม้ตัดสินใจแต่ละต้นจะทำการจำแนกประเภทหรือทำนายเอาต์พุตออกมา หลังจากนั้นเอาต์พุตสุดท้ายจะได้มาจากการโหวตของต้นไม้ตัดสินใจแต่ละต้น โดยเลือกค่าที่ได้รับการโหวตมากที่สุด

1) อัลกอริทึมในการสร้างป่าของต้นไม้ตัดสินใจ

1.1) จำนวนแถวข้อมูลที่ใช้ในการเรียนรู้จำนวน n แถว และจำนวนคุณลักษณะคือ M

1.2) เลือกแถวข้อมูลสำหรับเรียนรู้ โดยใช้วิธีสุ่มแบบใส่คืน (Sampling with replacement) จำนวน 2 ใน 3 จากแถวข้อมูลทั้งหมด เพื่อนำไปใช้ในการสร้างต้นไม้ตัดสินใจ ส่วนแถวข้อมูลที่เหลืออีก 1 ใน 3 (ข้อมูล out-of-bag) จะใช้ในการประมาณความผิดพลาดของต้นไม้โดยนำมาใช้ทำนายเพื่อดูความถูกต้อง

1.3) m คือจำนวนคุณลักษณะที่ถูกเลือกแบบสุ่มมาจาก M และนำมาใช้ในการสร้างโหนดของต้นไม้ตัดสินใจ โดยแต่ละโหนดของต้นไม้ตัดสินได้มาจากการหาคุณลักษณะที่สามารถแยกกลุ่มได้ดีที่สุด และค่า m นี้จะคงที่ตลอดระหว่างการสร้างป่าของต้นไม้ตัดสินใจ ยกตัวอย่างเช่น ถ้ามีค่าคุณลักษณะทั้งหมด 10 ตัว ถ้าค่า m คือ 5 ให้ทำการสุ่มเลือกคุณลักษณะมา 5 ตัว และนำคุณลักษณะ 5 ตัวนี้มาทำการหาว่าค่าคุณลักษณะตัวใดสามารถทำการจำแนกประเภทได้ดีที่สุดก็จะใช้โหนดนั้นเป็นโหนดของต้นไม้ตัดสินใจ จากนั้นในการหาโหนดถัดไปก็จะทำการสุ่มคุณลักษณะขึ้นมาอีก 5 ตัวจากคุณลักษณะที่เหลือและหาคุณลักษณะที่สามารถทำการจำแนกได้ดีที่สุด ทำจนกระทั่งครบคุณลักษณะทุกตัว

2) อัตราความผิดพลาด (Error Rate) ของป่าของต้นไม้ตัดสินใจ

2.1) สหสัมพันธ์ (Correlation) ระหว่างต้นไม้ตัดสินใจสองต้นใด ๆ ในป่าของต้นไม้ตัดสินใจ โดยการเพิ่มขึ้นของสหสัมพันธ์จะทำให้อัตราความผิดพลาดของป่าของต้นไม้ตัดสินใจเพิ่มขึ้นด้วย

2.2) ความแข็งแรง (Strength) ของต้นไม้ตัดสินใจแต่ละต้น ต้นไม้ตัดสินใจที่มีอัตราความผิดพลาดน้อยถือเป็นตัวจำแนกประเภทที่เข้มแข็ง การเพิ่มขึ้นของความเข้มแข็งของต้นไม้ตัดสินใจแต่ละต้นจะช่วยลดอัตราความผิดพลาดของป่าของต้นไม้ตัดสินใจ

การลดค่า m จะลดทั้งสหสัมพันธ์และความเข้มแข็งของต้นไม้ตัดสินใจ ในขณะที่การเพิ่มค่า m ก็จะเพิ่มค่าทั้งสองเช่นเดียวกัน ดังนั้นเราจึงจำเป็นต้องหาค่า m ที่เหมาะสม

3) การทำงานของป่าของต้นไม้ตัดสินใจ

เมื่อข้อมูลสำหรับเรียนรู้ถูกเลือกด้วยวิธีสุ่มแบบใส่คืนจากชุดข้อมูลทั้งหมด ต้นไม้ตัดสินใจจะถูกสร้างขึ้นด้วยข้อมูลชุดนี้ ส่วนข้อมูลที่เหลืออีกประมาณ 1 ใน 3 จะถูกนำมาใช้ในการประมาณความผิดพลาดของต้นไม้ตัดสินใจที่ถูกสร้างขึ้นมา นอกจากนี้ยังถูกใช้เพื่อหาความสำคัญของคุณลักษณะอีกด้วย โดยการทำงานของป่าของต้นไม้ตัดสินใจ มีวิธีดังนี้

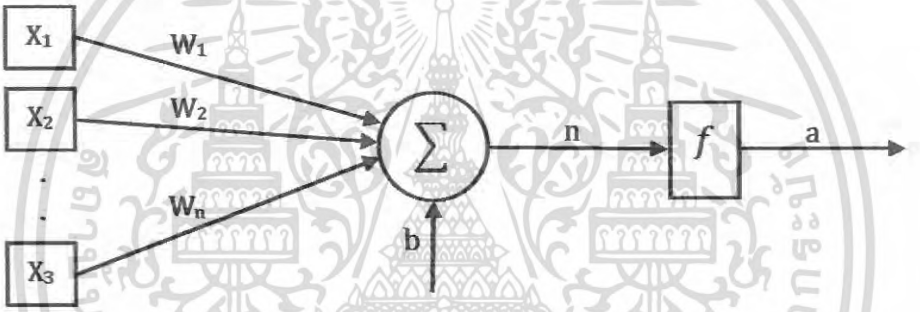
3.1) การประมาณความผิดพลาดจากข้อมูลที่เหลือจากการสุ่ม ต้นไม้ตัดสินใจแต่ละต้นจะถูกสร้างขึ้นมาจากข้อมูลที่มาจากการสุ่มแบบใส่คืน ข้อมูลที่เหลืออีกประมาณ 1 ใน 3 จะไม่นำมาใช้ในการสร้างต้นไม้ดังกล่าว แต่จะถูกนำมาใช้ในการประมาณความผิดพลาด โดยจะถูกใช้เป็นชุดทดสอบความถูกต้อง โดยค่าประมาณความผิดพลาดจะได้มาจากค่าเฉลี่ยของอัตราส่วนในการทำนายผิดของต้นไม้ตัดสินใจทุกต้นในป่าของต้นไม้ตัดสินใจ

3.2) ความสำคัญของคุณลักษณะ สำหรับต้นไม้ตัดสินใจทุก ๆ ต้นในป่าของต้นไม้ตัดสินใจ ให้ทำการนำข้อมูลที่เหลือจากการสุ่ม ไปทำการทำนายและนับจำนวนครั้งที่ทำนายถูกต้อง จากนั้นทำการเปลี่ยนคุณลักษณะ m แบบสุ่มและนำชุดข้อมูลใหม่นี้ไปทำนายอีกครั้งหนึ่ง จากนั้นนำจำนวน

ครั้งที่ทายถูกของข้อมูลที่มีการเปลี่ยนคุณลักษณะ m ไปลบออกจากจำนวนครั้งที่ทายถูกของข้อมูลที่ไม่ได้สับเปลี่ยนคุณลักษณะ เมื่อนำค่าที่ได้นี้จากต้นไม้ทุกต้นมาหาค่าเฉลี่ยก็จะได้ค่าคะแนนที่แสดงถึงความสำคัญของคุณลักษณะ m

2.3.3 โครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

โครงข่ายประสาทเทียม [5] มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ ด้วยโปรแกรมคอมพิวเตอร์ จุดมุ่งหมายของโครงข่ายประสาทเทียมคือต้องการให้คอมพิวเตอร์มีความชาญฉลาดในการเรียนรู้เหมือนที่มนุษย์มีการเรียนรู้ สามารถฝึกฝนได้ และสามารถนำความรู้และทักษะ รวมทั้งสามารถนำไปประยุกต์ใช้ได้ดีกับปัญหาการจำแนกประเภท การถดถอย และการจัดแบ่งกลุ่มข้อมูล เทคนิคนี้มักถูกเรียกว่า "black box" เนื่องจากการทำงานมีความซับซ้อนมากกว่าเทคนิคอื่น ๆ ค่อนข้างมาก การเรียนรู้ของโครงข่ายประสาทเทียมทำได้โดยการส่งข้อมูลเข้ามายังส่วนที่เรียกว่าเพอร์เซ็ปตรอน (perceptron) สามารถเทียบได้กับเซลล์สมองของมนุษย์ โดยที่เพอร์เซ็ปตรอนทำการรับข้อมูลที่อยู่ในรูปของเมทริกซ์ซึ่งเป็นตัวเลข เข้ามาคำนวณ ดังภาพที่ 2.2



ภาพที่ 2.2 โครงข่ายประสาทเทียมเพอร์เซ็ปตรอน

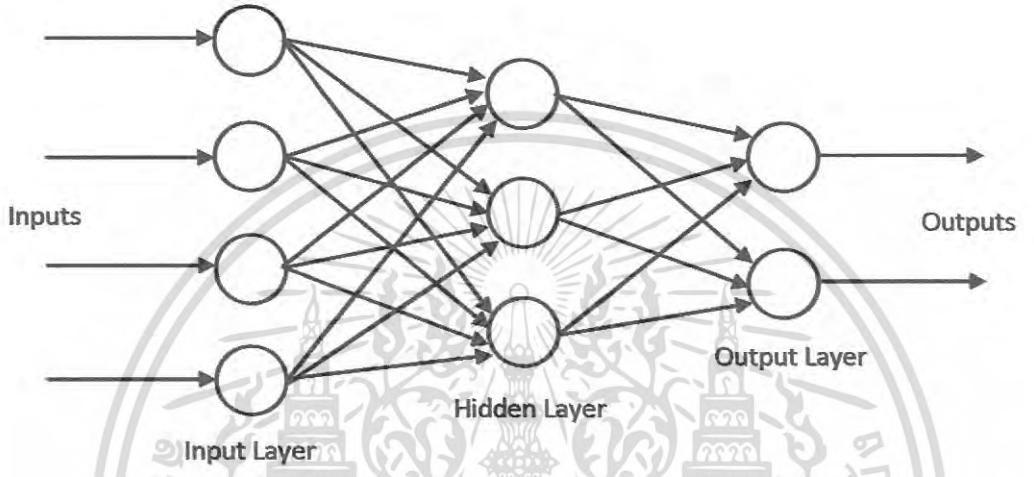
ฟังก์ชันผลรวม (Summation Function)

$$n = \sum_{i=1}^{z} x_i w_i + b \qquad (3)$$

โดยที่ ตัวแปร n คือ ผลรวมที่ได้จากฟังก์ชันผลรวม
ตัวแปร x_i คือ ค่าข้อมูลเข้าตัวที่ i
ตัวแปร w_i คือ ค่าน้ำหนักของโครงข่ายประสาทเทียมตัวที่ i
ตัวแปร z คือ จำนวนโครงข่ายประสาทเทียมชั้นข้อมูลเข้า
ตัวแปร b คือ ค่าความโน้มเอียง
ตัวแปร i มีค่าตั้งแต่ 1 ถึง z

โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น (Multilayer Perceptron: MLP) [6] เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีการฝึกฝนแบบมีผู้สอน และใช้ขั้นตอนการส่งค่าย้อนกลับ (Back-Propagation) สำหรับการฝึกฝน กระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า

(Forward Pass) และการส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นของข้อมูลเข้าและจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ไขข้อผิดพลาด (error-correction) คือ ผลต่างของผลตอบที่แท้จริง (actual response) กับผลตอบเป้าหมาย (target response) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ ค่าน้ำหนักการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย



ภาพที่ 2.3 โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น

1) อัลกอริทึมแบคพรอพาเกชัน (Back-propagation Algorithm) เป็นอัลกอริทึมที่ใช้ในการเรียนรู้ของโครงข่ายประสาทเทียมชนิดหนึ่งที่นิยมใช้ในโครงข่ายประสาทเทียมแบบหลายชั้น ข้อมูลจากชั้นอินพุตจะถูกคำนวณและส่งผ่านฟังก์ชันจากชั้นแฝงไปยังชั้นเอาต์พุต ซึ่งหลักการสำคัญของการเรียนรู้คือการเปลี่ยนแปลงค่าน้ำหนักของแต่ละเส้นเชื่อมระหว่างโหนด โดยในการปรับแก้ค่าน้ำหนักจะขึ้น อยู่กับความแตกต่างระหว่างค่าเอาต์พุตที่คำนวณได้กับค่าเอาต์พุตที่ต้องการ สำหรับขั้นตอนในการปรับแก้ค่าน้ำหนักมีขั้นตอนดังต่อไปนี้

1.1) กำหนดค่าอัตราการเรียนรู้ และค่าโมเมนตัม

1.2) ปรับแก้ค่าน้ำหนักของแต่ละเส้นเชื่อมระหว่างโหนด

สมการการปรับน้ำหนักทำได้ตามสมการที่ 4 ดังนี้

$$\Delta w_{ji}(n+1) = \eta\delta(n) \cdot y_j(n) + \alpha\Delta w_{ji}(n) \qquad (4)$$

เมื่อ

ตัวแปร x_i คือ ค่าข้อมูลด้านเข้าที่ node i

ตัวแปร w_i คือ ค่าถ่วงน้ำหนักที่ node i

ตัวแปร Δw_{ji} คือ ค่าปรับแก้ค่าถ่วงน้ำหนักระหว่าง node i และ j

ตัวแปร η คือ ค่าอัตราการเรียนรู้

ตัวแปร α คือ ค่าโมเมนตัม

ตัวแปร δ_j คือ ค่าผลต่างระหว่างค่าจริงกับค่าที่ได้จากการคำนวณในรูปของอนุพันธ์ของ Transfer function ของ node j

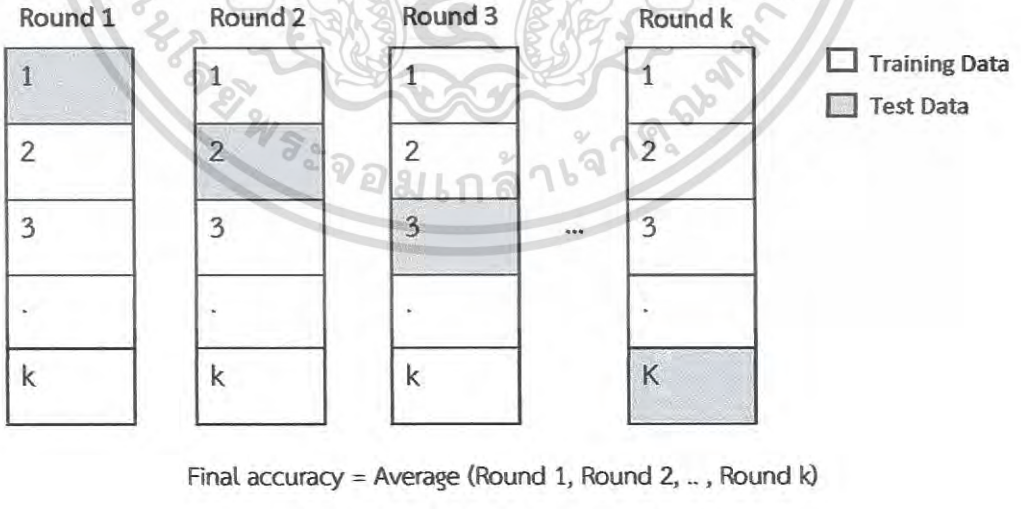
ตัวแปร y_i คือ ค่าผลลัพธ์ของแบบจำลองที่ node j และ n

ตัวแปร $n+1$ คือ ค่าที่แสดงถึงรอบของการปรับแก้ที่ n หรือ $n+1$

2.4 วิธีการทดสอบแบบการตรวจสอบไขว้กัน (K-fold cross-validation)

ผู้วิจัยได้ทำการทดสอบแบบจำลองโดยวิธีการทดสอบแบบการตรวจสอบไขว้กัน [7] เป็นวิธีการตรวจสอบค่าความผิดพลาดในการคาดการณ์ของแบบจำลอง โดยพื้นฐานของวิธีการตรวจสอบไขว้กันคือการสุ่มตัวอย่าง (resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำการตรวจสอบไขว้กันมักถูกใช้เป็นตัวเลือกในการกำหนดแบบจำลอง

การทำการตรวจสอบไขว้กัน จะแบ่งข้อมูลออกเป็น K ชุดเท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุด จะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K-1 ชุด จะถูกเป็นข้อมูลสำหรับการเรียนรู้ เช่น K=10 ก็หมายความว่ามีการแบ่งข้อมูลออกเป็น 10 รอบ โดยรอบที่ 1 คือ การเอาข้อมูลกลุ่มที่ 1 ออกเพื่อใช้ในการทดสอบ และใช้กลุ่มข้อมูลที่ 2-10 ในการสอน ส่วนรอบที่ 2 คือ การเอาข้อมูลกลุ่มที่ 2 ออกเพื่อใช้ในการทดสอบ และใช้กลุ่มข้อมูลที่ 1, 3-10 ในการสอน ดังนั้น รอบที่ 10 คือ การเอาข้อมูลกลุ่มที่ 10 ออกเพื่อใช้ในการทดสอบ และใช้กลุ่มข้อมูลที่ 1-9 ในการสอน ซึ่งวิธีนี้เป็นวิธีที่นิยมใช้ในการทดสอบประสิทธิภาพของแบบจำลองเนื่องจากผลที่ได้มีความน่าเชื่อถือ



ภาพที่ 2.4 โครงสร้าง k-fold Cross-Validation

2.5 ตารางเมทริกซ์ความสับสน (Confusion Matrix)

ตารางเมทริกซ์ความสับสน คือ ตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมน์และเท่ากับจำนวนคลาส เช่น ในงานวิจัยนี้มีคลาสที่เป็นคำตอบอยู่ 2 ค่า คือ Normal และ Abnormal ดังนั้น

ตาราง Confusion Matrix นี้จะสร้างได้เป็นตารางขนาด 2x2 ดังในตารางที่ 2.1 โดยข้อมูลด้านคอลัมน์คือคลาสที่อยู่ในข้อมูลฝึกสอน (Actual) และข้อมูลในแนวแถว คือ คลาสที่แบบจำลองทำนายมาได้ (Predicted)

ตารางที่ 2.1 ตารางเมทริกซ์ความสับสน

Actual \ Predicted	a	b
a	TP	FP
b	FN	TN

จากในตารางที่ 2.1 ค่าที่แสดงในช่องต่าง ๆ ของตารางประกอบด้วย

- ผลบวกจริง (True Positive: TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่
- ผลลบจริง (True Negative: TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่
- ผลบวกลวง (False Positive: FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งกำลังสนใจอยู่
- ผลลบลวง (False Negative: FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งไม่ได้สนใจอยู่

ตัววัดประสิทธิภาพจากตาราง Confusion Matrix

1) ค่าความแม่นยำ (Precision) เป็นการวัดความแม่นยำของแบบจำลอง โดยพิจารณาแยกทีละคลาส

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

2) ค่าความระลึก (Recall) เป็นการวัดความถูกต้องของแบบจำลอง โดยพิจารณาแยกทีละคลาส

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

3) ค่าวัดประสิทธิภาพ (F-measure) เป็นการวัดค่า Precision และ Recall พร้อมกันของแบบจำลอง โดยพิจารณาแยกทีละคลาส

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

4) ค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้องของแบบจำลอง โดยพิจารณารวมทุกคลาส คือ จำนวน True Positive ของทุกคลาสรวมกัน

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

2.6 ฐานข้อมูล (Database)

2.6.1 ความรู้ทั่วไปเกี่ยวกับระบบฐานข้อมูล

ฐานข้อมูล [8] เป็นการจัดเก็บข้อมูลอย่างเป็นระบบ ทำให้ผู้ใช้สามารถใช้ข้อมูลที่เกี่ยวข้องในระบบงานต่าง ๆ ร่วมกันได้ โดยที่จะไม่เกิดความซ้ำซ้อนของข้อมูล และยังสามารถหลีกเลี่ยงความขัดแย้งของข้อมูลด้วย อีกทั้งข้อมูลในระบบก็จะถูกต้องเชื่อถือได้ และเป็นมาตรฐานเดียวกัน โดยจะมีการกำหนดระบบความปลอดภัยของข้อมูลขึ้น

ปัจจุบันข้อมูลสารสนเทศเป็นข้อมูลที่ผ่านการกลั่นกรองอย่างเหมาะสม สามารถนำมาใช้ประโยชน์อย่างมาก เช่น นำมาใช้งานด้านธุรกิจ การบริหาร และกิจการอื่น ๆ ซึ่งองค์กรที่มีข้อมูลปริมาณมาก ๆ จะพบความยุ่งยากลำบากในการจัดเก็บข้อมูล และการนำข้อมูลที่ต้องการออกมาใช้ให้ทันต่อเหตุการณ์ ดังนั้นคอมพิวเตอร์จึงถูกนำมาใช้เป็นเครื่องมือช่วยในการจัดเก็บข้อมูลและประมวลผลข้อมูล ซึ่งทำให้ระบบการจัดเก็บข้อมูลเป็นไปได้สะดวก ฐานข้อมูลจึงเข้ามามีบทบาทสำคัญอย่างมาก โดยเฉพาะระบบงานต่าง ๆ ที่ใช้คอมพิวเตอร์ การออกแบบและพัฒนาระบบฐานข้อมูลจึงต้องคำนึงถึงการควบคุมและการจัดการความถูกต้อง ตลอดจนประสิทธิภาพในการเรียกใช้ข้อมูล จากการจัดเก็บข้อมูลรวมเป็นฐานข้อมูลจะก่อให้เกิดประโยชน์ดังนี้

1) สามารถลดความซ้ำซ้อนของข้อมูล

การเก็บข้อมูลชนิดเดียวกันไว้หลาย ๆ ที่ จะทำให้เกิดความซ้ำซ้อน ดังนั้นการนำข้อมูลมารวมเก็บไว้ในฐานข้อมูล จะช่วยลดปัญหาการเกิดความซ้ำซ้อนของข้อมูลได้ โดยระบบจัดการฐานข้อมูล (Database Management System: DBMS) จะช่วยควบคุมความซ้ำซ้อนได้ เนื่องจากระบบจัดการฐานข้อมูลจะทราบได้ตลอดเวลาว่ามีข้อมูลซ้ำซ้อนกันอยู่ที่ใดบ้าง

2) หลีกเหลี่ยงความขัดแย้งของข้อมูล

หากมีการเก็บข้อมูลชนิดเดียวกันไว้หลาย ๆ ที่ และมีการปรับปรุงข้อมูลเดียวกันนี้แต่ปรับปรุงไม่ครบทุกที่ที่มีข้อมูลเก็บอยู่ ก็จะทำให้เกิดปัญหาข้อมูลชนิดเดียวกันอาจมีค่าไม่เหมือนกันในแต่ละที่ที่เก็บข้อมูลอยู่ จึงก่อให้เกิดความขัดแย้งของข้อมูลขึ้น (Inconsistency)

3) สามารถใช้ข้อมูลร่วมกันได้

ฐานข้อมูลจะเป็นการจัดเก็บข้อมูลรวมไว้ด้วยกัน ดังนั้นหากผู้ใช้ต้องการใช้ข้อมูลในฐานข้อมูลที่มาจากแฟ้มข้อมูลต่าง ๆ ก็จะทำได้โดยง่าย

4) สามารถรักษาความถูกต้องเชื่อถือได้ของข้อมูล

บางครั้งพบว่าการจัดเก็บข้อมูลในฐานข้อมูลอาจมีข้อผิดพลาดเกิดขึ้น เช่น จากการที่ผู้ป้อนข้อมูลป้อนข้อมูลผิดพลาดคือป้อนจากตัวเลขหนึ่งไปเป็นอีกตัวเลขหนึ่ง โดยเฉพาะกรณีมีผู้ใช้หลายคนต้องใช้ข้อมูลจากฐานข้อมูลร่วมกัน หากผู้ใช้คนใดคนหนึ่งแก้ไขข้อมูลผิดพลาดก็ทำให้ผู้อื่นได้รับ

ผลกระทบตามไปด้วย ในระบบจัดการฐานข้อมูล (DBMS) จะสามารถใส่กฎเกณฑ์เพื่อควบคุมความผิดพลาดที่เกิดขึ้น

5) สามารถกำหนดความเป็นมาตรฐานเดียวกันของข้อมูล

การเก็บข้อมูลร่วมกันไว้ในฐานข้อมูลจะทำให้สามารถกำหนดมาตรฐานของข้อมูลได้ รวมทั้งมาตรฐานต่าง ๆ ในการจัดเก็บข้อมูลให้เป็นไปในลักษณะเดียวกัน เช่น การกำหนดรูปแบบการเขียนวันที่ ในลักษณะ วัน/เดือน/ปี หรือ ปี/เดือน/วัน ทั้งนี้จะมีผู้ที่คอยบริหารฐานข้อมูลที่เรียกว่า ผู้บริหารฐานข้อมูล (Database Administrator: DBA) เป็นผู้กำหนดมาตรฐานต่าง ๆ

6) สามารถกำหนดระบบความปลอดภัยของข้อมูล

ระบบความปลอดภัยในที่นี้ เป็นการป้องกันไม่ให้ผู้ใช้ที่ไม่มีสิทธิมาใช้ หรือมาเห็นข้อมูลบางอย่างในระบบ ผู้บริหารฐานข้อมูลจะสามารถกำหนดระดับการเรียกใช้ข้อมูลของผู้ใช้แต่ละคนได้ตามความเหมาะสม

7) เกิดความเป็นอิสระของข้อมูล

ในระบบฐานข้อมูลจะมีตัวจัดการฐานข้อมูลที่ทำหน้าที่เป็นตัวเชื่อมโยงกับฐานข้อมูลโปรแกรมต่าง ๆ อาจไม่จำเป็นต้องมีโครงสร้างข้อมูลทุกครั้ง ดังนั้นการแก้ไขข้อมูลบางครั้งจึงอาจกระทำเฉพาะกับโปรแกรมที่เรียกใช้ข้อมูลที่เปลี่ยนแปลงเท่านั้น ส่วนโปรแกรมที่ไม่ได้เรียกใช้ข้อมูลดังกล่าวก็จะเป็นอิสระจากการเปลี่ยนแปลง

2.6.2 รูปแบบของระบบฐานข้อมูล

รูปแบบของระบบฐานข้อมูล มีอยู่ด้วยกัน 3 ประเภท คือ

1) ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) เป็นการเก็บข้อมูลในรูปแบบที่เป็นตาราง หรือเรียกว่า รีเลชั่น มีลักษณะเป็น 2 มิติ คือเป็นแถว และเป็นคอลัมน์ การเชื่อมโยงข้อมูลระหว่างตารางจะเชื่อมโยงโดยใช้คุณลักษณะ หรือคอลัมน์ที่เหมือนกันทั้งสองตารางเป็นตัวเชื่อมโยงข้อมูล ฐานข้อมูลเชิงสัมพันธ์นี้จะเป็นรูปแบบของฐานข้อมูลที่นิยมใช้ในปัจจุบัน

2) ฐานข้อมูลแบบเครือข่าย (Network Database) จะเป็นการรวมระเบียนต่าง ๆ และความสัมพันธ์ระหว่างระเบียน แต่จะต่างกับฐานข้อมูลเชิงสัมพันธ์คือ ในฐานข้อมูลเชิงสัมพันธ์จะแฝงความสัมพันธ์เอาไว้ โดยระเบียนที่มีความสัมพันธ์กันจะต้องมีค่าของข้อมูลในแอตทริบิวต์หนึ่งเหมือนกัน แต่ฐานข้อมูลแบบเครือข่ายจะแสดงความสัมพันธ์อย่างชัดเจน

3) ฐานข้อมูลแบบลำดับชั้น (Hierarchical Database) เป็นโครงสร้างที่จัดเก็บข้อมูลในลักษณะความสัมพันธ์แบบพ่อ-ลูก (Parent-Child Relationship Type: PCR Type) หรือเป็นโครงสร้างรูปแบบต้นไม้ ข้อมูลที่จัดเก็บในที่นี้คือ ระเบียน ซึ่งประกอบด้วยค่าของเขตข้อมูล (Field) ของเอนทิตี้หนึ่ง ๆ ฐานข้อมูลแบบลำดับชั้นนี้คล้ายคลึงกับฐานข้อมูลแบบเครือข่าย แต่ต่างกันที่ฐานข้อมูลแบบลำดับชั้นมีกฎเพิ่มขึ้นมาอีกหนึ่งประการ คือ ในแต่ละกรอบจะมีลูกศรวิ่งเข้าหาได้ไม่เกิน 1 หัวลูกศร

2.7 เอสคิวแอล (SQL)

SQL ย่อมาจาก Structured Query Language [10] คือภาษาที่ใช้ในการเขียนโปรแกรมเพื่อจัดการกับฐานข้อมูลโดยเฉพาะ เป็นภาษามาตรฐานบนระบบฐานข้อมูลเชิงสัมพันธ์ และเป็นระบบเปิด (open system) สามารถแบ่งการทำงานได้เป็น 4 ประเภท ดังนี้

1) Select query ใช้ในการดึงข้อมูลในฐานข้อมูล จะมีการค้นหารายการจากตารางในฐานข้อมูล ตั้งแต่หนึ่งตารางขึ้นไป ตามเงื่อนไขที่สั่ง ผลลัพธ์ที่ได้จะเป็นเซตของข้อมูลที่สามารถสร้างเป็นตารางใหม่

2) Update query ใช้สำหรับการแก้ไขข้อมูลในตาราง โดยแก้ในคอลัมน์ที่มีค่าตรงตามเงื่อนไข โดยมีรูปแบบดังนี้

Update ชื่อตาราง Set (ชื่อคอลัมน์ = ค่าที่จะใส่เข้าไปในคอลัมน์นั้น ๆ)

Where เงื่อนไข

3) Insert query ใช้ในการเพิ่มเติมข้อมูลใหม่ ๆ เข้าไปในฐานข้อมูล มีรูปแบบดังนี้

Insert Into ชื่อตาราง (ชื่อคอลัมน์1,2)

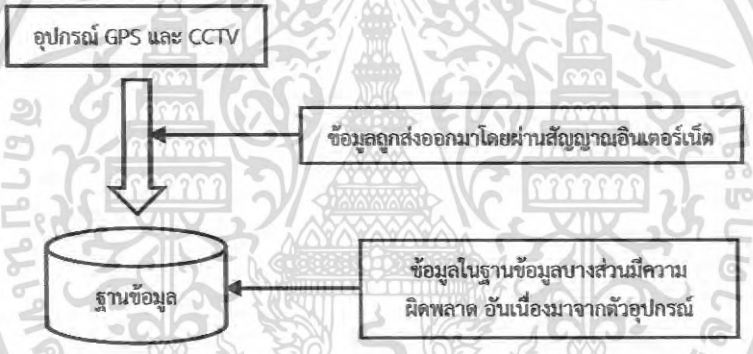
Values (ค่าที่จะใส่ลงในคอลัมน์1,2)

บทที่ 3
วิธีการดำเนินการวิจัย

งานวิจัยนี้ต้องการนำเสนอแบบจำลองหาความผิดปกติของระบบบันทึกข้อมูลพฤติกรรมการขับขี่เพื่อคัดแยกข้อมูลที่ผิดปกติออกจากข้อมูลปกติในระบบ โดยใช้เทคนิคการจำแนกประเภทข้อมูล วิเคราะห์ด้วยโปรแกรมเวกา (Weka) [9] เป็นเครื่องมือในการทำเหมืองข้อมูล โปรแกรมโทดฟอออราเคิล (Toad for Oracle) เป็นเครื่องมือในการจัดการข้อมูลในฐานข้อมูล และโปรแกรมไมโครซอฟท์เอกซ์เซล (Microsoft Excel) เป็นเครื่องมือในการเตรียมข้อมูลสำหรับการดำเนินการวิจัยและมีวิธีดำเนินการวิจัยออกเป็นขั้นตอนดังต่อไปนี้

3.1 ศึกษาปัญหาและความต้องการของระบบ

ในขั้นตอนของงานวิจัยนี้ได้ศึกษากระบวนการทำงานของระบบบันทึกข้อมูลพฤติกรรมการขับขี่รวมถึงข้อมูลของระบบในฐานข้อมูล เพื่อให้ได้เข้าใจถึงปัญหาที่เกิดขึ้นและแนวทางในการพัฒนา เพื่อนำมาเป็นข้อมูลในการวิเคราะห์และออกแบบพัฒนาในขั้นต่อไป

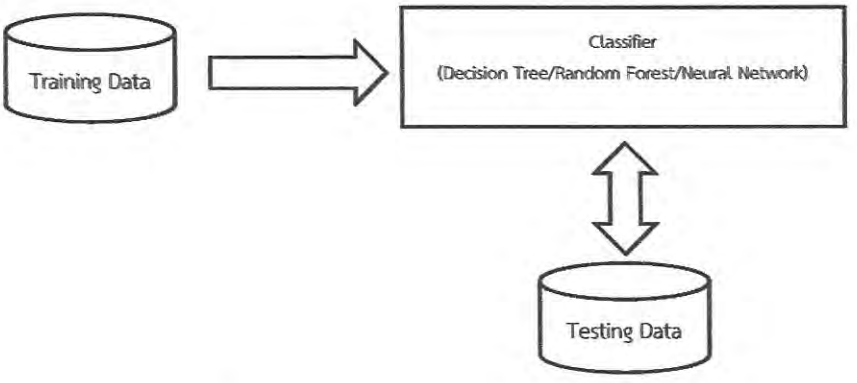


ภาพที่ 3.1 โครงสร้างระบบบันทึกข้อมูลพฤติกรรมการขับขี่

จากภาพที่ 3.1 สามารถอธิบายการทำงานได้ คือ เมื่ออุปกรณ์ระบบกำหนดตำแหน่งบนโลกและกล้องวงจรปิดทำการส่งข้อมูลออกมาโดยผ่านสัญญาณอินเทอร์เน็ต ข้อมูลจะถูกส่งมาเก็บในฐานข้อมูลส่วนกลาง ซึ่งข้อมูลที่ถูกส่งมาบางส่วนนั้นอาจมีข้อมูลที่มีความผิดพลาด งานวิจัยนี้จึงจำเป็นต้องสร้างแบบจำลองในการคัดแยกข้อมูล

3.2 ศึกษาขั้นตอนการทำเหมืองข้อมูล

งานวิจัยนี้ได้ศึกษาแบบจำลองที่เหมาะสมกับเป้าหมายและความต้องการในการพัฒนาระบบเพื่อให้เข้าใจถึงทฤษฎีและแนวคิดสำหรับการนำไปประยุกต์ใช้ในการวิเคราะห์และออกแบบแบบจำลองความสามารถคัดแยกข้อมูลผิดปกติออกจากข้อมูลปกติ ผู้พัฒนาจึงได้ทำการวิเคราะห์รูปแบบข้อมูลโดยใช้เทคนิคการจำแนกประเภทข้อมูล [2] โดยมีอัลกอริทึมของต้นไม้ [3] แบบ C4.5 ป่าของต้นไม้ตัดสินใจ [4] และโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น [6] มาทำการเปรียบเทียบหาประสิทธิภาพเพื่อความแม่นยำในการทำนายค่าโดยวิเคราะห์ลักษณะของชุดข้อมูล



ภาพที่ 3.2 การสร้างแบบจำลองการจำแนกประเภท

จากภาพที่ 3.2 อธิบายขั้นตอนการสร้างแบบจำลองด้วยวิธีการจำแนกประเภท โดยการเตรียมชุดข้อมูลฝึกสอนที่ใช้ในการวิเคราะห์ พร้อมกับชุดข้อมูลสำหรับทดสอบ

3.3 เก็บรวบรวมข้อมูลที่ใช้ในการวิเคราะห์

ในขั้นตอนนี้เป็นการเก็บรวบรวมข้อมูล เพื่อที่จะใช้เป็นชุดข้อมูลสำหรับฝึกสอนในการสร้างแบบจำลอง โดยการเก็บรวบรวมข้อมูลนั้นจะต้องใช้เครื่องมือในการจัดการกับข้อมูลที่อยู่ในฐานข้อมูล ผู้วิจัยได้ใช้โปรแกรม Toad for Oracle เป็นเครื่องมือในการเชื่อมต่อกับฐานข้อมูล [8] โดยการจัดการกับข้อมูลที่อยู่ในฐานข้อมูล เพื่อที่จะเรียกใช้ข้อมูลนั้น จะทำได้โดยใช้ภาษาโปรแกรมเพื่อจัดการฐานข้อมูล หรือ ภาษา SQL [10] เพื่อเรียกข้อมูลนั้น ๆ และทำการส่งออกมาแสดงผลที่โปรแกรม Microsoft Excel เพื่อเป็นการจัดเตรียมข้อมูลให้พร้อมนำไปวิเคราะห์ในขั้นตอนต่อไป

3.4 การจัดเตรียมข้อมูล

การจัดเตรียมข้อมูล ผู้วิจัยได้ทำการจัดเตรียมข้อมูลผ่านโปรแกรม Microsoft Excel โดยทำการคัดเลือกข้อมูล การกลั่นกรองข้อมูล และการแปรงรูปข้อมูล เพื่อให้ชุดข้อมูลมีความเหมาะสมและถูกต้องที่จะนำไปใช้ในการสร้างแบบจำลอง โดยเมื่อเตรียมข้อมูลเสร็จแล้ว ก่อนจะนำไปวิเคราะห์ด้วยโปรแกรม Weka ผู้วิจัยได้ทำการบันทึกข้อมูลเป็นนามสกุลไฟล์ซีเอสวี (csv) เพราะเป็นข้อกำหนดหนึ่งของโปรแกรม Weka

ข้อมูลสำหรับทำการวิจัยนี้ได้มีการจัดเตรียมข้อมูลโดยใช้ข้อมูลติดตามสถานะรถขนส่งผลิตภัณฑ์ในระบบบันทึกข้อมูลพฤติกรรมการขับขี่ที่อยู่ในฐานข้อมูล ซึ่งเก็บรวบรวมข้อมูลเกี่ยวกับตำแหน่งความเร็ว สถานะเครื่องยนต์ ใบอนุญาตคนขับ สัญญาณอินเทอร์เน็ต จำนวนดาวเทียมที่ใช้ในการส่งตำแหน่ง แบตเตอรี่ที่ตัวรถ และแบตเตอรี่ที่ตัวอุปกรณ์ โดยข้อมูลสถานะของรถแต่ละคันจะถูกส่งเข้ามาในฐานข้อมูลทุก ๆ 30 วินาที ซึ่งข้อมูลที่นำมาวิเคราะห์เป็นตัวอย่างข้อมูลในเดือนตุลาคม พ.ศ.2559 จำนวน 402,000 ตัวอย่าง โดยมีรายละเอียดของข้อมูลคือ มีแอตทริบิวต์ทั่วไปที่ใช้ทั้งหมด 8 แอตทริบิวต์ และ แอตทริบิวต์ที่เป็นตัวทำนาย 1 แอตทริบิวท์ โดยแบ่งเป็น 2 Classes เพื่อเป็นตัวบอกสถานะของข้อมูลแต่ละตัวอย่าง ซึ่งประกอบด้วยคลาส Normal และ Abnormal โดยรวมทั้งหมดแล้วชุดข้อมูลที่ใช้มีแอตทริบิวท์ทั้งหมด 9 แอตทริบิวท์

ตารางที่ 3.1 อธิบายความหมายของแต่ละคุณลักษณะ (Attributes)

คุณลักษณะ	ความหมาย
SPEED	ความเร็วของรถ ณ เวลาที่ข้อมูลถูกส่งออกมา
ENGINE_STAT	สถานะของเครื่องยนต์
DRIVER_LIC_INFO	ใบอนุญาตของคนขับก่อนขับขี่
HDOP	ค่าความถูกต้องของตำแหน่งทางราบ
SAT_NO	จำนวนดาวเทียมที่ใช้ระบุตำแหน่ง
RSSI	ค่าความแรงของสัญญาณอินเทอร์เน็ต
INT_BATT_VDC	แบตเตอรี่ที่อุปกรณ์ GPS
EXT_BATT_VDC	แบตเตอรี่รถขนส่ง
STATUS	สถานะของข้อมูลแต่ละตัวอย่าง

ตารางที่ 3.2 อธิบายความหมายของค่าแต่ละคุณลักษณะ (Field)

คุณลักษณะ	ค่าของคุณลักษณะ	ความหมาย
SPEED	>= 0	ค่าความเร็วของรถ มีค่าตั้งแต่ 0 ขึ้นไป
ENGINE_STAT	0	เครื่องยนต์ดับ
	1	เครื่องยนต์ทำงาน
DRIVER_LIC_INFO	0	ไม่มีข้อมูลใบอนุญาตก่อนขับ
	1	มีข้อมูลใบอนุญาตก่อนขับ
HDOP	>= 0	ค่าความถูกต้องของตำแหน่งทางราบ ถ้ามีค่ามากจะมีความถูกต้องมาก
SAT_NO	0 - 14	จำนวนดาวเทียมที่ใช้ในการระบุตำแหน่ง หากมีจำนวนดาวเทียมมาก จะระบุตำแหน่งได้แม่นยำ มีค่า 0 ถึง 14
INT_BATT_VDC	3.0 – 4.3	แบตเตอรี่อุปกรณ์ (Volt)
EXT_BATT_VDC	0 - 54	แบตเตอรี่รถ (Volt)
STATUS	Normal	ตัวอย่างข้อมูลปกติ
	Abnormal	ตัวอย่างข้อมูลที่ผิดปกติ

SPEED	ENGINE_STAT	DRIVER_LIC_INFO	HDOP	SAT_NO	RSSI	INT_BATT_VDC	EXT_BATT_VDC	STATUS
58	1	1	7	12	10	4	28	Normal
63	1	0	7	13	12	4	28	Normal
55	1	1	7	13	21	4	28	Normal
54	1	1	7	13	19	4	28	Normal
49	1	1	8	11	17	4	28	Normal
59	1	1	8	11	15	4	28	Normal
0	1	0	7	12	17	4	26	Normal
0	1	1	6	13	8	4	28	Normal
0	1	1	7	13	31	4	28	Normal
0	1	0	7	11	14	4	27	Normal
57	1	1	7	12	17	4	28	Normal
56	1	1	7	12	12	4	28	Normal
64	1	0	7	13	12	4	28	Normal
63	1	1	7	12	14	4	28	Normal
63	1	1	7	12	20	4	27	Normal
55	1	1	6	13	15	4	28	Normal
56	1	1	7	12	20	4	28	Normal
56	1	1	8	10	9	4	28	Normal
57	1	1	7	13	21	4	29	Normal
71	1	0	7	12	14	3	29	Normal
58	1	1	8	11	10	4	28	Normal
57	1	1	7	12	20	4	28	Normal
0	1	1	7	13	21	4	27	Normal
59	1	0	7	13	10	4	28	Normal
55	1	1	7	12	20	4	28	Normal
57	1	1	7	12	19	4	27	Normal
0	1	1	7	11	21	4	27	Normal
60	1	1	7	11	19	4	28	Normal
63	1	1	8	11	12	4	27	Normal
0	1	1	8	10	10	4	29	Normal
6	1	1	8	10	20	4	28	Normal
53	1	1	7	12	12	4	28	Normal
0	1	1	7	11	21	4	27	Normal
0	1	0	6	14	15	4	27	Normal
60	1	1	7	12	15	4	27	Normal
0	1	1	7	12	12	4	27	Normal
52	1	1	7	12	16	4	28	Normal
60	1	1	7	12	12	4	28	Normal
50	1	1	8	11	22	4	28	Normal
56	1	1	7	13	19	4	27	Normal
56	1	1	7	12	17	4	27	Normal
56	1	1	7	12	17	4	28	Normal
51	1	1	7	12	21	4	28	Normal
62	1	1	7	12	23	4	28	Normal
0	1	1	7	13	21	4	27	Normal
0	1	0	7	12	14	4	26	Normal
55	1	1	6	13	24	4	28	Normal
57	1	1	7	12	13	4	27	Normal
81	1	0	7	12	22	4	29	Normal
57	1	1	6	14	21	4	28	Normal
63	1	1	7	12	20	4	27	Normal
53	1	1	9	10	22	4	28	Normal
57	1	1	7	12	13	4	27	Normal
0	1	1	6	13	8	4	28	Normal
52	1	1	7	12	16	4	28	Normal
0	1	1	7	13	15	4	28	Normal
60	1	0	6	14	21	4	28	Normal
0	1	0	7	11	17	4	27	Normal
31	1	1	7	12	18	4	27	Normal
44	1	1	7	12	20	4	29	Normal
49	1	1	6	13	17	4	27	Normal
56	1	1	7	12	16	4	28	Normal
9	1	1	8	11	14	4	29	Normal
63	1	1	7	12	9	4	27	Normal
0	1	1	7	13	21	4	27	Normal
56	1	1	7	11	11	4	28	Normal
50	1	1	8	11	19	4	28	Normal
47	1	1	6	13	20	4	27	Normal
58	1	1	6	13	14	4	28	Normal
0	1	0	7	11	17	4	27	Normal
56	1	1	7	12	14	4	28	Normal
76	1	0	7	12	22	4	29	Normal
54	1	1	7	12	11	4	28	Normal
54	1	1	7	12	20	4	28	Normal
58	1	1	6	13	20	4	28	Normal
56	1	1	7	13	13	4	28	Normal
54	1	1	7	12	16	4	28	Normal
0	1	0	8	10	14	4	28	Normal
54	1	1	7	12	20	4	28	Normal
0	1	0	7	11	17	4	28	Normal
59	1	1	7	11	25	4	27	Normal
0	1	1	9	6	10	4	29	Normal
51	1	1	7	11	11	4	28	Normal

ภาพที่ 3.3 ตัวอย่างของข้อมูลปกติที่ใช้ในการวิเคราะห์

SPEED	ENGINE_STAT	DRIVER_LIC_INFO	HDOP	SAT_NO	RSSI	INT_BATT_VOC	EXT_BATT_VOC	STATUS
0	0	1	8	10	9	4	25	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	9	6	17	4	27	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	10	31	4	27	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	7	12	13	4	27	Abnormal
0	0	1	8	10	30	4	27	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	7	12	14	4	27	Abnormal
0	0	1	7	13	16	4	26	Abnormal
0	0	1	8	11	9	4	25	Abnormal
0	0	1	8	10	9	4	25	Abnormal
0	0	1	8	10	9	4	25	Abnormal
0	0	1	8	11	17	4	27	Abnormal
0	0	1	6	12	19	4	27	Abnormal
0	0	1	8	10	9	4	25	Abnormal
0	0	1	8	10	9	4	25	Abnormal
0	0	1	8	11	7	4	25	Abnormal
0	0	1	8	11	7	4	25	Abnormal
0	0	1	8	11	7	4	25	Abnormal
0	0	1	8	11	7	4	25	Abnormal
21	0	0	8	10	11	4	26	Abnormal
11	0	0	8	11	20	4	25	Abnormal
53	0	0	8	11	19	4	28	Abnormal
52	0	0	8	11	19	4	28	Abnormal
52	0	0	8	11	19	4	28	Abnormal
9	0	0	9	10	12	4	24	Abnormal
10	0	0	8	11	20	4	25	Abnormal
53	0	0	8	11	19	4	28	Abnormal
53	0	0	8	11	19	4	28	Abnormal
13	0	0	8	10	11	4	26	Abnormal
16	0	0	8	11	20	4	25	Abnormal
57	0	0	8	10	15	4	28	Abnormal
58	0	0	8	11	15	4	28	Abnormal
17	0	0	8	10	11	4	26	Abnormal
55	0	0	7	11	11	4	28	Abnormal
15	0	0	8	11	20	4	25	Abnormal
53	0	0	8	11	12	4	28	Abnormal
54	0	0	8	11	12	4	28	Abnormal
7	0	0	8	10	11	4	26	Abnormal
55	0	0	8	11	12	4	28	Abnormal
54	0	0	8	11	12	4	28	Abnormal
55	0	0	8	11	12	4	28	Abnormal
17	0	0	8	10	11	4	26	Abnormal
8	0	0	9	10	12	4	24	Abnormal
53	0	0	8	11	12	4	28	Abnormal
51	0	0	8	11	12	4	28	Abnormal
51	0	0	8	11	12	4	28	Abnormal
6	0	0	8	10	11	4	26	Abnormal
58	0	0	8	10	14	4	28	Abnormal
54	0	0	8	10	16	4	28	Abnormal
18	0	0	9	10	13	4	28	Abnormal
76	0	0	7	11	25	4	27	Abnormal
83	0	0	7	11	25	4	27	Abnormal
27	0	0	7	12	9	4	28	Abnormal
49	0	0	7	11	19	4	13	Abnormal
72	0	0	7	11	13	4	28	Abnormal
10	0	0	8	11	23	4	28	Abnormal
29	0	0	9	10	11	4	28	Abnormal
26	0	0	9	10	11	4	28	Abnormal
62	0	0	8	10	19	4	13	Abnormal
17	0	0	9	10	11	4	27	Abnormal
52	0	0	8	10	11	4	28	Abnormal
55	0	0	8	10	13	4	28	Abnormal
9	0	0	9	10	11	4	28	Abnormal
68	0	0	8	10	27	4	13	Abnormal
64	0	0	7	11	15	4	27	Abnormal
24	0	0	6	13	14	4	27	Abnormal
72	0	0	8	10	27	4	13	Abnormal
8	0	0	9	10	15	4	28	Abnormal
6	0	0	8	10	15	4	27	Abnormal
66	0	0	7	11	9	4	28	Abnormal
30	0	0	8	10	18	4	28	Abnormal
60	0	0	9	6	23	4	13	Abnormal
60	0	0	9	6	23	4	13	Abnormal
61	0	0	9	6	23	4	13	Abnormal
63	0	0	8	10	11	4	28	Abnormal

ภาพที่ 3.4 ตัวอย่างของข้อมูลผิดปกติที่ใช้ในการวิเคราะห์

3.5 การทำงานของโปรแกรมที่ใช้วิเคราะห์แบบจำลอง

โปรแกรมที่ใช้ในการวิเคราะห์และสร้างแบบจำลองคือ โปรแกรม Weka [9] ขั้นตอนการทำงานของโปรแกรมที่ใช้วิเคราะห์แบบจำลองเพื่อการจำแนกประเภทข้อมูล สามารถอธิบายได้ดังต่อไปนี้

3.5.1 การเตรียมข้อมูลสำหรับวิเคราะห์แบบจำลอง

ข้อมูลตัวอย่างที่ใช้เป็นข้อมูลติดตามสถานะรถขนส่งผลิตภัณฑ์ในระบบบันทึกข้อมูลพฤติกรรมการขับขี่ที่อยู่ในฐานข้อมูล โดยข้อมูลที่นำมาวิเคราะห์เป็นตัวอย่างข้อมูลในเดือนตุลาคม พ.ศ.2559 จำนวน 402,000 ตัวอย่าง โดยรายละเอียดของข้อมูลมีทั้งหมด 9 แอตทริบิวต์ โดยทำการบันทึกข้อมูลในโปรแกรม Microsoft Excel เป็นนามสกุลไฟล์ csv เพื่อที่สามารถนำไปวิเคราะห์ในโปรแกรม Weka ได้

SPEED	ENGINE_STA	DRIVER_LIC_INFO	HDOP	SAT_NO	RSST	INT_BATT_VDC	EXT_BATT_VDC	STATUS
0	0	0	7	12	15	4.3	24	Normal
0	0	0	7	12	12	4	25	Normal
0	0	0	7	12	15	4.1	25	Normal
0	0	0	9	10	14	4.1	26	Normal
0	0	0	8	11	10	4.1	25	Normal
0	0	0	8	11	24	4	26	Normal
0	0	0	8	11	20	4.3	26	Normal
0	0	0	8	11	21	4.1	25	Normal
0	0	0	9	10	16	4	25	Normal
0	0	0	8	11	20	4.1	25	Normal
0	0	0	8	11	20	4	25	Normal
0	0	0	7	12	15	4	25	Normal
0	0	0	9	10	21	4.3	25	Normal
0	0	0	7	11	16	4.2	25	Normal
0	0	0	8	11	18	4.2	26	Normal
0	0	0	8	11	21	4.3	25	Normal
0	0	0	6	13	14	4.1	25	Normal
0	0	0	7	12	9	4.1	26	Normal
0	0	0	7	12	19	4.1	25	Normal
0	0	0	8	10	22	4.3	25	Normal
0	0	0	8	11	16	4.1	26	Normal
0	0	0	6	13	19	4.2	25	Normal

ภาพที่ 3.5 ตัวอย่างข้อมูลของชุดข้อมูลเป็นไฟล์ซีเอสวี

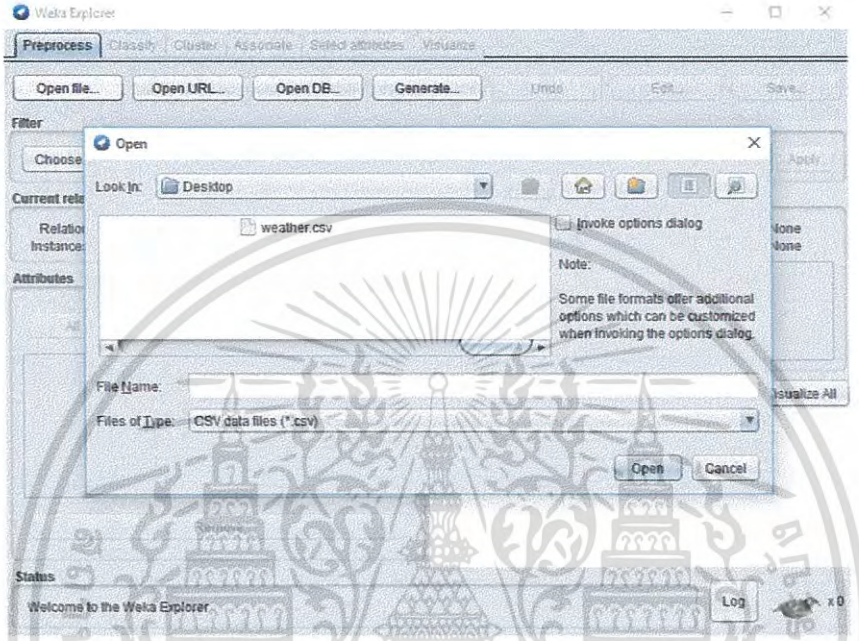
3.5.2 การเรียกใช้โปรแกรมในการสร้างแบบจำลอง

โปรแกรมที่ใช้ในการสร้างแบบจำลองคือโปรแกรม Weka 3.8.0 [9] โดยมีวิธีการเรียกใช้โปรแกรม ดังขั้นตอนต่อไปนี้

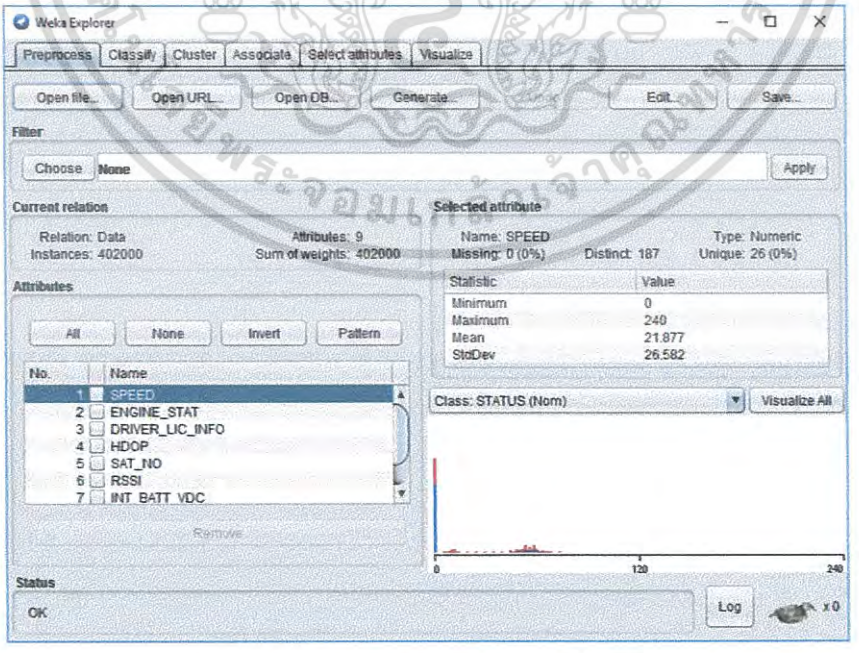
ขั้นตอนที่ 1 เริ่มต้นการทำงานของโปรแกรม เริ่มจากเปิดโปรแกรม และแสดงหน้าโปรแกรมดังภาพที่ 3.6

ภาพที่ 3.6 การเข้าหน้าโปรแกรม Weka [9]

ขั้นตอนที่ 2 นำเข้าข้อมูลที่จัดเตรียมไว้ โดยการเลือก Applications >> Explorer >> Open file เลือกไฟล์ข้อมูลที่ต้องการนำเข้า ตามภาพที่ 3.7 และหลังจากนั้น จะแสดงหน้าจอข้อมูลตามภาพที่ 3.8 ซึ่งหน้าจอนี้จะทำการวิเคราะห์หาจำนวนของข้อมูลที่ใช้ในการวิจัย เพื่อให้ทราบจำนวนเปอร์เซ็นต์ต่อข้อมูลทั้งหมด และค่าเฉลี่ยของข้อมูลในแต่ละแอตทริบิวต์

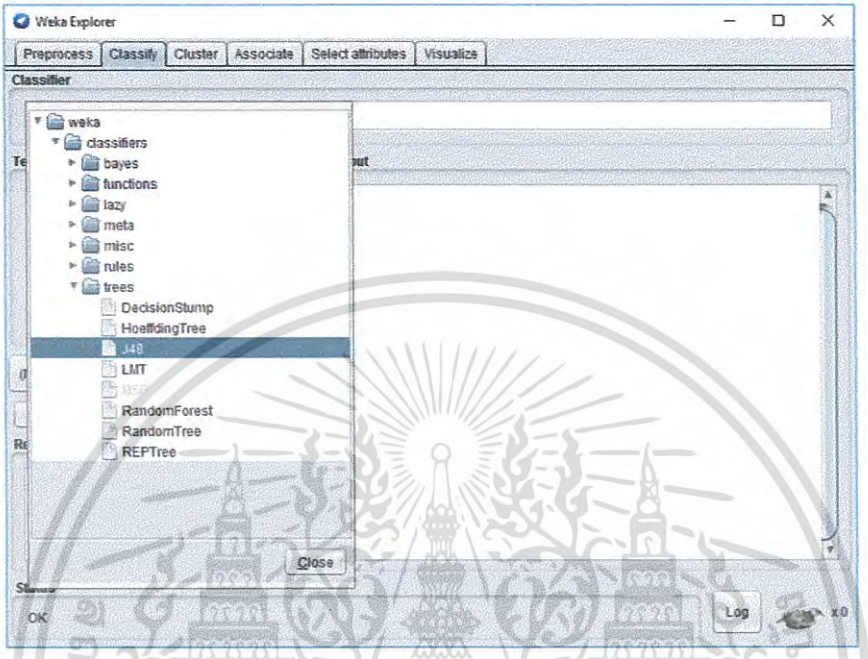


ภาพที่ 3.7 การนำเข้าข้อมูลในโปรแกรม Weka

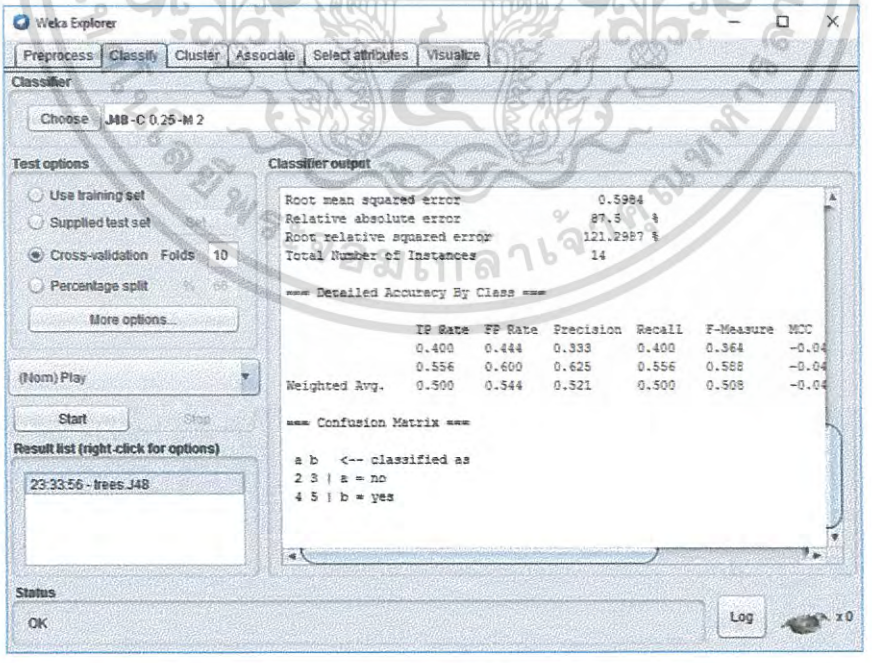


ภาพที่ 3.8 หน้าจอการแสดงข้อมูลที่นำเข้า

ขั้นตอนที่ 3 เลือกเทคนิคการจำแนกประเภทข้อมูล Classify >> Choose >> trees >> เลือกเทคนิคการจำแนกข้อมูล เช่น J48 ตามภาพที่ 3.9 จากนั้นกดปุ่ม Start จะได้ผลลัพธ์และค่าความถูกต้องของการสร้างแบบจำลองออกมา ภาพที่ 3.10



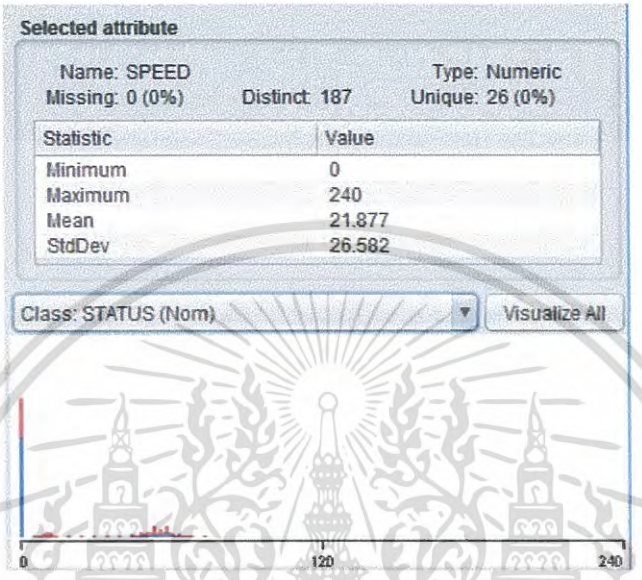
ภาพที่ 3.9 การเลือกเทคนิคที่ใช้ในการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ



ภาพที่ 3.10 หน้าจอแสดงผลลัพธ์ที่ได้ด้วยเทคนิคต้นไม้ตัดสินใจ

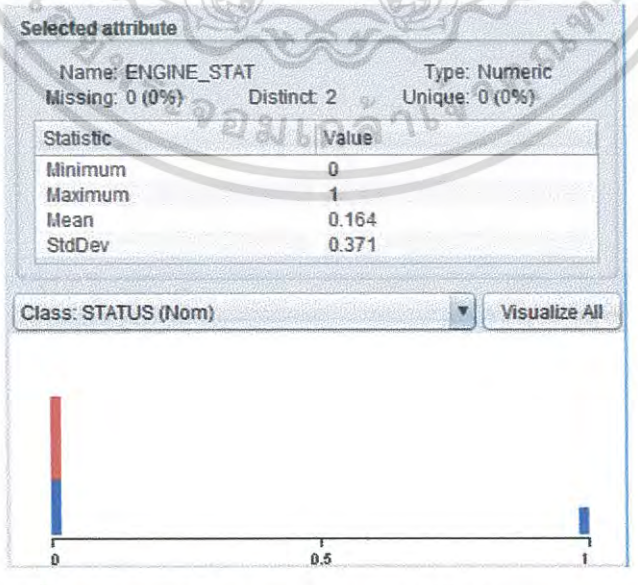
3.5.3 รายละเอียดของข้อมูลแต่ละแอตทริบิวต์

ภายในแต่ละแอตทริบิวต์จะประกอบไปด้วยข้อมูลต่าง ๆ ภายในแอตทริบิวต์ ซึ่งโปรแกรม Weka สามารถคำนวณหาค่าสถิติของข้อมูล และบอกจำนวนข้อมูลที่ผิดพลาดภายในแอตทริบิวต์ ซึ่งมีกราฟแสดงปริมาณข้อมูลตามค่าของข้อมูลนั้น ๆ ในแอตทริบิวต์



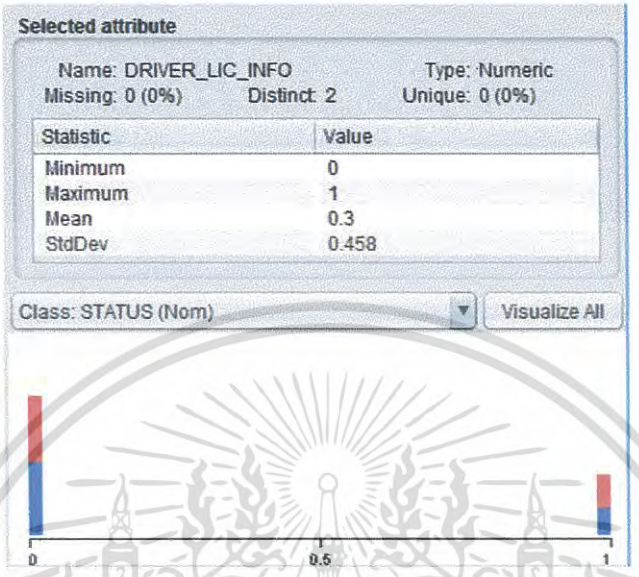
ภาพที่ 3.11 รายละเอียดของข้อมูลแอตทริบิวต์ SPEED

จากภาพที่ 3.11 ที่แอตทริบิวต์ SPEED มีค่า Minimum อยู่ที่ 0 และมีค่า Maximum สูงสุด 240 คำนวณค่า Mean ของความเร็วได้ความเร็วที่ 21.877 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 26.582



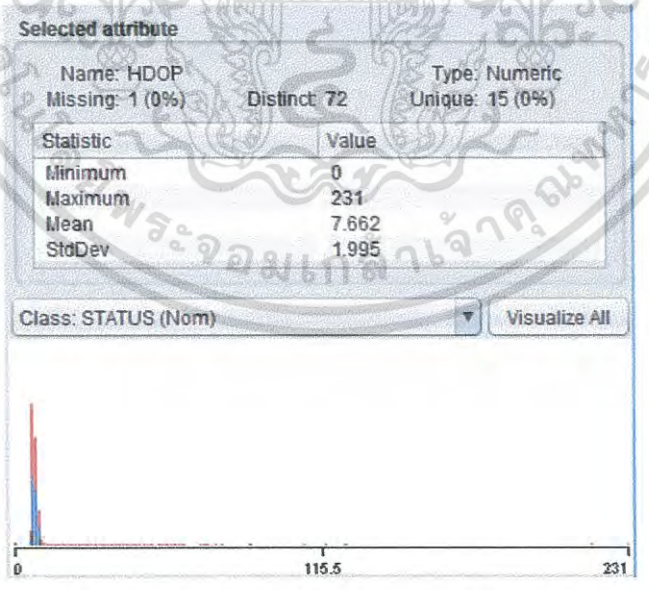
ภาพที่ 3.12 รายละเอียดของข้อมูลแอตทริบิวต์ ENGINE_STAT

จากภาพที่ 3.12 ที่แอตทริบิวต์ ENGINE_STAT มีค่าของข้อมูลอยู่ 2 ค่า คือ 0 และ 1 โดยค่า Mean อยู่ที่ 0.164 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 0.371



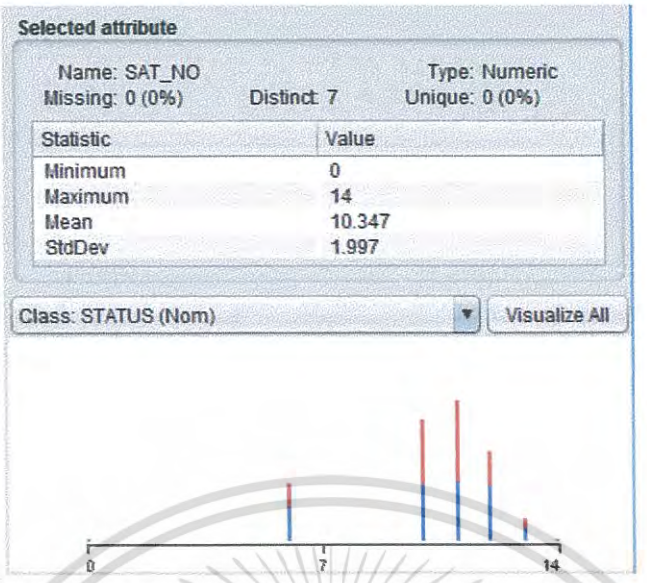
ภาพที่ 3.13 รายละเอียดของข้อมูลแอตทริบิวต์ DRIVER_LIC_INFO

จากภาพที่ 3.13 ที่แอตทริบิวต์ DRIVER_LIC_INFO มีค่าของข้อมูลอยู่ 2 ค่า คือ 0 และ 1 โดยค่า Mean อยู่ที่ 0.3 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 0.458



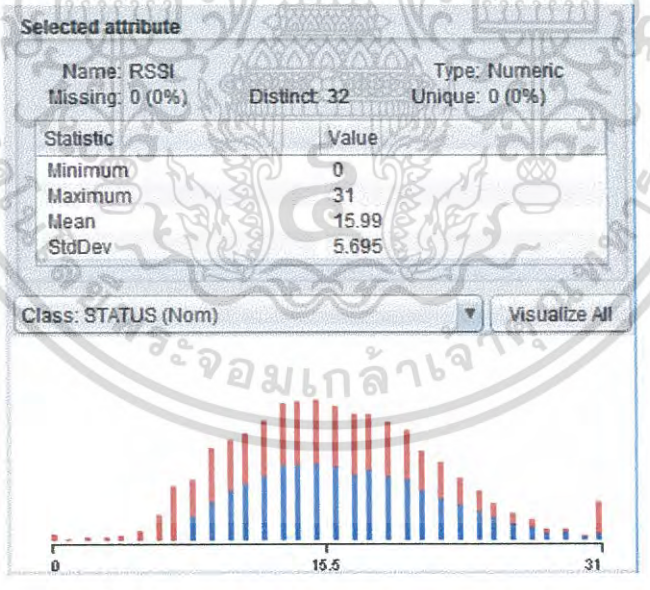
ภาพที่ 3.14 รายละเอียดของข้อมูลแอตทริบิวต์ HDOP

จากภาพที่ 3.14 ที่แอตทริบิวต์ HDOP มีค่า Minimum อยู่ที่ 0 และมีค่า Maximum สูงสุดที่ 231 ซึ่งคำนวณค่า Mean ได้ค่า Mean อยู่ที่ 7.662 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 1.995 ซึ่งมีจำนวนของข้อมูลที่ผิดพลาดในแอตทริบิวต์อยู่ 1 จำนวน



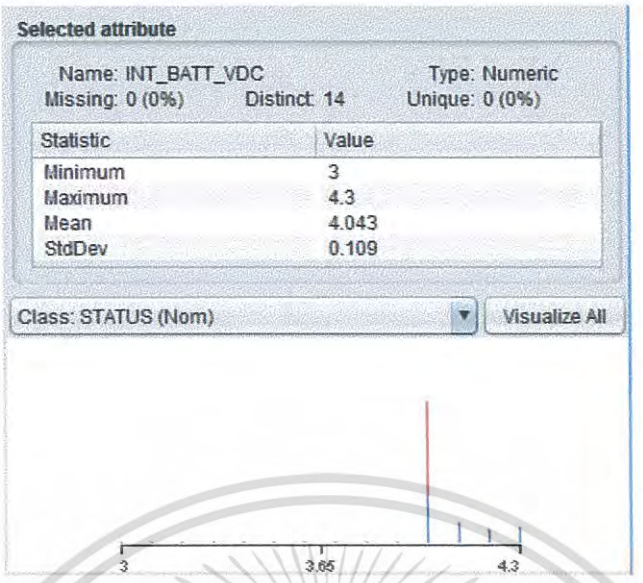
ภาพที่ 3.15 รายละเอียดของข้อมูลแอตทริบิวต์ SAT_NO

จากภาพที่ 3.15 ที่แอตทริบิวต์ SAT_NO มีจำนวน 7 ค่า มีค่า Minimum อยู่ที่ 0 และมีค่า Maximum 14 ซึ่งคำนวณค่า Mean ได้ค่า Mean อยู่ที่ 10.347 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 1.997



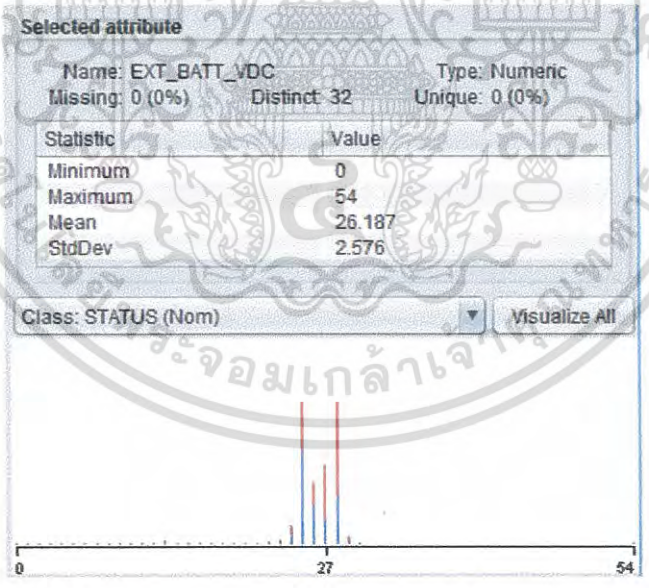
ภาพที่ 3.16 รายละเอียดของข้อมูลแอตทริบิวต์ RSSI

จากภาพที่ 3.16 ที่แอตทริบิวต์ RSSI มีค่า Minimum อยู่ที่ 0 และมีค่า Maximum ที่ 31 ซึ่งคำนวณค่า Mean ได้ค่า Mean อยู่ที่ 15.99 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 5.695 ซึ่งสีน้ำเงิน หมายถึงข้อมูลปกติ และสีแดง หมายถึงข้อมูลที่ผิดปกติ



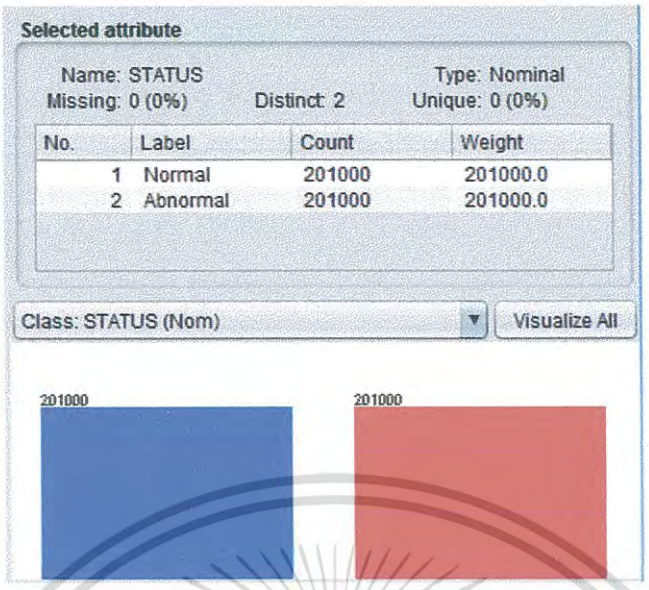
ภาพที่ 3.17 รายละเอียดของข้อมูลแอตทริบิวต์ INT_BATT_VDC

จากภาพที่ 3.17 ที่แอตทริบิวต์ INT_BATT_VDC มีค่า Minimum อยู่ที่ 3 และมีค่า Maximum อยู่ที่ 4.3 ซึ่งคำนวณค่า Mean ได้ค่า Mean อยู่ที่ 4.043 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 0.109



ภาพที่ 3.18 รายละเอียดของข้อมูลแอตทริบิวต์ EXT_BATT_VDC

จากภาพที่ 3.18 ที่แอตทริบิวต์ EXT_BATT_VDC มีค่า Minimum อยู่ที่ 0 และมีค่า Maximum อยู่ที่ 54 ซึ่งคำนวณค่า Mean ได้ค่า Mean อยู่ที่ 26.187 และมีค่า Standard Deviation หรือค่าเบี่ยงเบนมาตรฐานอยู่ที่ 2.576 โดยดูจากคลาส STATUS ซึ่งสีน้ำเงิน หมายถึงข้อมูลปกติ และสีแดง หมายถึงข้อมูลที่ผิดปกติ



ภาพที่ 3.19 รายละเอียดของข้อมูลแอตทริบิวต์ STATUS

จากภาพที่ 3.19 ที่แอตทริบิวต์ STATUS มีค่า 2 จำนวน คือ Normal และ Abnormal ประกอบด้วย Normal 201,000 จำนวน และ Abnormal 201,000 จำนวน

ภาพที่ 3.20 รายละเอียดของชุดข้อมูลทุกแอตทริบิวต์

จากภาพที่ 3.20 แสดงกราฟข้อมูลของแต่ละแอตทริบิวต์ของชุดข้อมูล โดยดูจากคลาส STATUS ซึ่งสีน้ำเงิน หมายถึงข้อมูลปกติ และสีแดง หมายถึงข้อมูลที่ผิดปกติ

3.6 ขั้นตอนการทดลอง

ในการทำงานวิจัยนี้ได้มีการทดลองเพื่อหาแบบจำลองที่ใช้ในการคัดแยกข้อมูลที่ผิดปกติออกจากข้อมูลปกติในระบบบันทึกข้อมูลพฤติกรรมการขับขี่ โดยมีขั้นตอนการทดลอง แสดงดังภาพที่ 3.21

ภาพที่ 3.21 แผนภาพขั้นตอนการทดลอง

จากภาพที่ 3.21 สามารถอธิบายขั้นตอนการทดลองได้ดังนี้

3.6.1 การเตรียมข้อมูล

การเตรียมข้อมูล เป็นการนำข้อมูลของระบบติดตามรถขนส่งผลิตภัณฑ์ที่มีอยู่ในฐานข้อมูลนำมาจัดเตรียมเพื่อนำไปวิเคราะห์และพัฒนาแบบจำลอง ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากแบบจำลองที่ได้จากการทำเหมืองข้อมูลจะให้ผลลัพธ์ที่ถูกต้องหรือแม่นยำนั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ ถ้าข้อมูลที่ใช้ไม่ถูกต้องหรือมีความผิดพลาด จะส่งผลให้ผลลัพธ์ที่ได้คลาดเคลื่อน โดยการเตรียมข้อมูลนั้น สามารถแบ่งออกได้เป็น 3 ขั้นตอน คือ

1) การคัดเลือกข้อมูล เป็นการเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่จะทำการวิเคราะห์ เพื่อลดขนาดของข้อมูลที่ใช้ทดสอบและเพิ่มความถูกต้องให้แก่แบบจำลอง เพราะข้อมูลบางชนิดไม่มีความสัมพันธ์กับข้อมูลชนิดอื่น ซึ่งไม่มีความจำเป็นในการที่จะนำข้อมูลนี้ไปวิเคราะห์ จึงทำการตัดข้อมูลในส่วนนี้ออก

2) การกลั่นกรองข้อมูล เป็นเทคนิคที่ใช้ในการกำจัดข้อมูลที่ไม่ปกติ (Noisy) เพื่อยกระดับคุณภาพของข้อมูล และเพิ่มความแม่นยำของข้อมูล ในบางครั้งอาจพบข้อมูลที่ไม่ถูกต้องเนื่องมาจากปัญหาในระหว่างการจัดเก็บข้อมูล เช่น การกรอกข้อมูลไม่ครบ กรอกข้อมูลซ้ำซ้อน ในขั้นตอนนี้จะทำการกรองข้อมูลที่ไม่ถูกต้องหรือซ้ำซ้อนออก หรือทำการซ่อมข้อมูลที่ขาดหายไป เช่น แทนที่ด้วยค่าศูนย์ (0) เพื่อให้สามารถนำข้อมูลไปวิเคราะห์ต่อได้

3) การแปรงรูปข้อมูล เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมและพร้อมนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของเหมืองข้อมูลที่เลือกใช้

3.6.2 การพัฒนาแบบจำลอง

การพัฒนาแบบจำลอง เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล ได้แก่การสร้างตัวทำนาย (Prediction Model) และมีการนำเทคนิคการทำเหมืองข้อมูลหลายเทคนิคมาใช้ในการวิเคราะห์ข้อมูล ซึ่งจะต้องเลือกเทคนิคที่มีความเหมาะสมกับประเภทของงาน ข้อจำกัด และการแก้ปัญหา เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ดังนั้นเมื่อทำขั้นตอนนี้แล้ว อาจมีการย้อนกลับไปที่ขั้นตอนการเตรียมข้อมูล เพื่อแปรงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิค งานวิจัยนี้ผู้วิจัยได้เลือกใช้เทคนิคการจำแนกประเภทข้อมูล เพื่อใช้ในการจำแนกข้อมูลระหว่างข้อมูลปกติและข้อมูลที่ผิดปกติ โดยนำเทคนิคในการทำเหมืองข้อมูลมาทั้งหมด 3 เทคนิคที่ใช้ในการวิเคราะห์ ได้แก่

1) ต้นไม้ตัดสินใจ อัลกอริทึม C4.5

2) ป่าของต้นไม้ตัดสินใจ

3) โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น

3.6.3 ทดสอบความถูกต้องของแบบจำลอง

การทดสอบและตรวจสอบความถูกต้องของแบบจำลองที่สร้างขึ้น จะเปรียบเทียบกับผลลัพธ์ที่ได้จากการใช้เครื่องมือของ Weka [9]

1) การตรวจสอบผลการทำนาย

นำผลหรือแบบจำลองที่ได้จากการวิเคราะห์ข้อมูลในการจำแนกประเภทข้อมูลนำไปทดสอบกับข้อมูลชุดใหม่

2) ประเมินผลจากการตรวจสอบการทำนาย

นำผลที่ได้จากการตรวจสอบมาประเมินผล โดยวัดจากประสิทธิภาพของการทำนาย เพื่อเป็นตัวบ่งชี้ความน่าเชื่อถือของแบบจำลองที่ได้ โดยผู้วิจัยได้เลือกใช้วิธีทดสอบแบบ 10-fold Cross-Validation เป็นตัวทดสอบแบบจำลอง

3.6.4 การเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค

การวัดประสิทธิภาพของความถูกต้องของข้อมูล วัดได้จากค่าความถูกต้องของการจัดกลุ่มของข้อมูล ซึ่งการทดสอบประสิทธิภาพจะแบ่งออกเป็น 3 ส่วน คือ ส่วนแรกเป็นการทดสอบเทคนิคต้นไม้ตัดสินใจ โดยใช้อัลกอริทึม C4.5 ส่วนที่สองคือป่าของต้นไม้ตัดสินใจ และส่วนที่สามจะใช้อัลกอริทึมเพอร์เซ็ปตรอนแบบหลายชั้นของเทคนิคโครงข่ายประสาทเทียม

3.6.5 สรุปผลการวิจัย

การสรุปผล เป็นขั้นตอนสุดท้ายของการดำเนินการวิจัย เพื่อสรุปผลการทำเหมืองข้อมูลในการเลือกใช้เทคนิคการทำเหมืองข้อมูล เพื่อให้ตรงกับเป้าหมายและความต้องการของการพัฒนาระบบ และทำการเปรียบเทียบประสิทธิภาพและตรวจสอบความถูกต้องของแต่ละแบบจำลองที่ได้จากแต่ละเทคนิค เพื่อสามารถนำไปประยุกต์ใช้งานกับระบบได้

3.7 ตัวแปรที่ใช้ในงานวิจัย

ชุดข้อมูลที่ผู้วิจัยนำมาใช้ในการวิเคราะห์ เป็นข้อมูลที่ได้จากระบบบันทึกข้อมูลพฤติกรรมการขับขี่ที่อยู่ในฐานข้อมูล โดยมีรายละเอียดของชุดข้อมูลคือ จำนวนข้อมูลที่ใช้ในการวิเคราะห์ทั้งหมด 402,000 ตัวอย่าง มีแอตทริบิวต์ทั่วไปที่ใช้ทั้งหมด 8 แอตทริบิวต์ และ แอตทริบิวต์ที่ใช้เป็นตัวทำนาย 1 แอตทริบิวต์

1) เปลี่ยนข้อมูลในแอตทริบิวต์ของ DRIVER_LIC_INFO เป็น 2 ค่า คือ ค่า 0 และค่า 1 โดยที่ค่า 0 เป็นค่าที่บ่งบอกว่ามีไม่มีข้อมูลใบอนุญาตขับขี่ของพนักงานคนขับ และค่า 1 เป็นค่าที่บ่งบอกว่ามีข้อมูลใบอนุญาตขับขี่ของพนักงานคนขับ

2) เพิ่มแอตทริบิวต์ STATUS เพื่อบอกสถานะของข้อมูลแต่ละตัวอย่าง แบ่งเป็น 2 classes คือ Normal และ Abnormal โดยคลาส Normal เป็นคลาสที่มีข้อมูลเป็นปกติ และคลาส Abnormal เป็นคลาสที่มีข้อมูลผิดปกติ

3) ข้อมูลจำนวน 402,000 ตัวอย่าง แบ่งเป็นข้อมูลที่มีสถานะ Normal 201,000 ตัวอย่าง และ Abnormal 201,000 ตัวอย่าง

4) วิธีการที่ใช้ในการทดสอบแบบจำลอง ผู้วิจัยได้เลือกใช้วิธีการตรวจสอบไขว้กัน [7]

3.7.4.1 K - fold Cross-Validation โดย K=10 โดยแบ่งชุดข้อมูลออกเป็น 10 ชุด ชุดละเท่า ๆ กัน

5) การวิเคราะห์แบบจำลองด้วยเทคนิคโครงข่ายประสาทเทียม มีการปรับค่าดังนี้

3.7.5.1 ค่าอัตราการเรียนรู้ เท่ากับ 0.33

3.7.5.2 ค่าโมเมนตัม เท่ากับ 0.2

3.8 เครื่องมือที่นำมาใช้ในการวิจัย

1) ระบบปฏิบัติการ Windows 10

2) โปรแกรม Toad for Oracle 11.6 ใช้ในการจัดการเกี่ยวกับฐานข้อมูล

3) โปรแกรม Weka 3.8.0 [9] ใช้ในการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล

4) โปรแกรม Microsoft Excel 2015 เป็นโปรแกรมจัดการข้อมูลที่ใช้ในการคัดเลือกข้อมูลและจัดเตรียมข้อมูล สำหรับนำเข้ามาทดสอบในโปรแกรม Weka

5) โปรแกรม Microsoft Word 2015 เป็นโปรแกรมที่ใช้ในการทำรายงานวิจัย

บทที่ 4
ผลการวิจัย

จากการวัดประสิทธิภาพแบบจำลองของแต่ละเทคนิคกับข้อมูลติดตามสถานะรถขนส่งผลิตภัณฑ์ของระบบบันทึกข้อมูลพฤติกรรมการขับขี่ ตามขั้นตอนการดำเนินงานวิจัย สามารถอธิบายผลการดำเนินงานได้ดังนี้

4.1 ผลการวิเคราะห์ข้อมูลของแบบจำลอง

ในการทำนายแบบจำลองโดยใช้โปรแกรม Weka เวอร์ชั่น 3.8.0 [9] ข้อมูลที่ใช้ในการทดสอบมีทั้งหมด 402,000 ตัวอย่าง และ 9 แอตทริบิวต์ ใช้การจำแนกประเภทข้อมูล โดยใช้เทคนิคทั้งหมด 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ [3] ป่าของต้นไม้ตัดสินใจ [4] และเพอร์เซ็ปตรอนแบบหลายชั้น [6] ซึ่งวิธีที่ใช้ในการทดสอบแบบจำลองคือวิธี 10-folds Cross-Validation [7] ได้แก่การแบ่งกลุ่มข้อมูลออกเป็น 10 กลุ่ม ในแต่ละรอบจะนำกลุ่มข้อมูลจำนวน 9 กลุ่ม เป็นกลุ่มศึกษา และกลุ่มข้อมูลที่เหลือเป็นกลุ่มทดสอบ โดยทำซ้ำเป็นจำนวน 10 รอบ เพื่อเปลี่ยนกลุ่มทดสอบให้ครบทุกกลุ่ม และนำผลลัพธ์ที่ได้มาหาค่าเฉลี่ย

4.1.1 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ [3] แบบ C4.5 ซึ่งแต่ละกิ่งแทนคุณลักษณะต่าง ๆ ของข้อมูลระบบติดตามรถขนส่งน้ำมัน และโหนดใบแสดงสถานะของข้อมูลหรือคลาสที่กำหนดไว้ 2 คลาส คือ ข้อมูลปกติและข้อมูลที่ไม่ปกติ ผลการวิเคราะห์ข้อมูลการเรียนรู้ต้นไม้ตัดสินใจจากชุดข้อมูลได้ความถูกต้องและความผิดพลาด แสดงดังในภาพที่ 4.1

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      401914               99.9786 %
Incorrectly Classified Instances        86                0.0214 %
Kappa statistic                          0.9996
Mean absolute error                      0.0004
Root mean squared error                  0.0145
Relative absolute error                  0.0746 %
Root relative squared error              2.8919 %
Total Number of Instances           402000
```

ภาพที่ 4.1 สรุปผลของการทำนายด้วยเทคนิคต้นไม้ตัดสินใจ

จากภาพที่ 4.1 อธิบายได้ว่า สำหรับข้อมูลติดตามสถานะรถขนส่งน้ำมัน จำนวน 402,000 ตัวอย่าง เมื่อนำมาวิเคราะห์แบบจำลอง พบว่ามีความถูกต้องคิดเป็นร้อยละ 99.9786% และมีความผิดพลาดร้อยละ 0.0214% ถือว่ามีเปอร์เซ็นต์ความถูกต้องที่สูง

จากการทดสอบนี้ได้จำนวนโหนดใบเท่ากับ 28 ใบ และขนาดต้นไม้ 55 โหนด

```
=== Confusion Matrix ===

      a      b   <-- classified as
 200987     13 |      a = Normal
     73 200927 |      b = Abnormal
```

ภาพที่ 4.2 ผลลัพธ์ค่าเมทริกซ์ความสับสนที่ได้ของต้นไม้ตัดสินใจ

จากภาพที่ 4.2 แสดงค่าผลลัพธ์ของตารางเมทริกซ์ความสับสนที่ได้จากแบบจำลองของต้นไม้ตัดสินใจ โดย a แทนข้อมูลที่มีแอตทริบิวต์ Status มีค่าเป็น Normal และ b แทนข้อมูลที่มีแอตทริบิวต์ Status มีค่าเป็น Abnormal เป็นการเปรียบเทียบระหว่างข้อมูลจริงและข้อมูลที่แบบจำลองทำนายได้ โดยค่าผิดพลาดที่ได้มีค่าผลลบลวงเท่ากับ 13 และผลบวกลวงเท่ากับ 73

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 1.000    0.000    1.000      1.000    1.000
                 1.000    0.000    1.000      1.000    1.000
Weighted Avg.    1.000    0.000    1.000      1.000    1.000
```

ภาพที่ 4.3 ผลลัพธ์ค่าความถูกต้องโดยคำนวณเป็นคลาสของต้นไม้ตัดสินใจ

จากภาพที่ 4.3 แสดงผลลัพธ์ที่ได้จากแบบจำลองของต้นไม้ตัดสินใจ โดยคำนวณค่าความถูกต้องเป็นคลาส ซึ่งผลลัพธ์บรรทัดแรกเป็นการคำนวณของคลาส Normal โดยการคำนวณได้ค่าความแม่นยำ ค่าความระลึก และค่าการวัดประสิทธิภาพ เท่ากับ 1 หรือ 100% และผลลัพธ์บรรทัดที่สองเป็นการคำนวณของคลาส Abnormal โดยการคำนวณได้ค่าความแม่นยำ ค่าความระลึก และค่าการวัดประสิทธิภาพ เท่ากับ 1 หรือ 100%

Classifier output

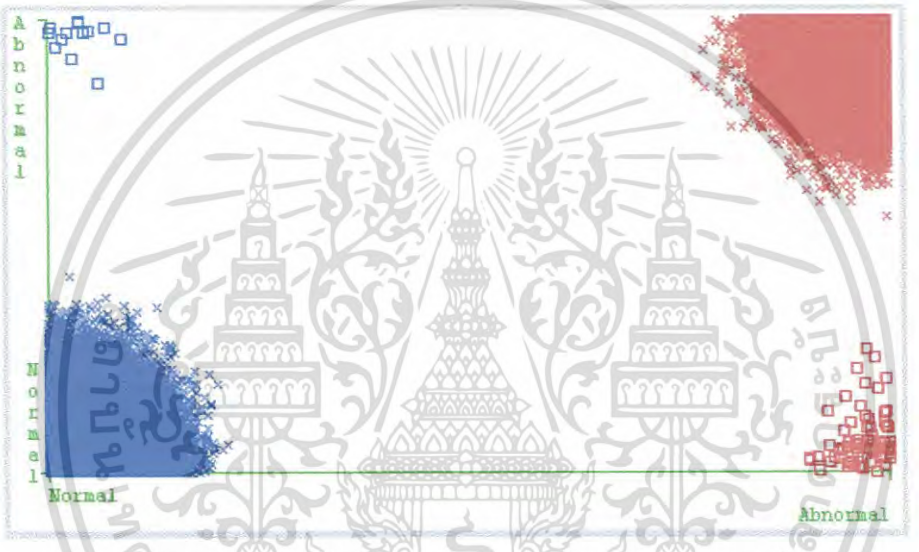
```
J48 pruned tree
------------------

INI_BATT_VDC <= 4
|   ENGINE_STAT <= 0
|   |   SPEED <= 0
|   |   |   DRIVER_LIC_INFO <= 0
|   |   |   |   RSSI <= 7
|   |   |   |   |   HDOP <= 8
|   |   |   |   |   |   INI_BATT_VDC <= 3.9: Normal (3.0/1.0)
|   |   |   |   |   |   INI_BATT_VDC > 3.9: Abnormal (19.0/2.0)
|   |   |   |   |   HDOP > 8: Normal (6.0/2.0)
|   |   |   |   RSSI > 7: Normal (41143.0/7.0)
|   |   |   DRIVER_LIC_INFO > 0: Abnormal (57672.0/1.0)
|   |   SPEED > 0: Abnormal (139321.0)
|   ENGINE_STAT > 0
|   |   SPEED <= 100
|   |   |   RSSI <= 8
|   |   |   |   DRIVER_LIC_INFO <= 0: Normal (213.0/1.0)
```

ภาพที่ 4.4 ผลลัพธ์บางส่วนจากการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ

จากการทดสอบ ผลลัพธ์ที่ได้จะแสดงเงื่อนไขของกฎ โดยค่าของข้อมูลของแอตทริบิวต์ที่มีค่าต่อเนื่องจะใช้เครื่องหมาย "<" ">" และ "<=" "=>" และในการพิจารณาส่วนข้อมูลที่เป็นค่าไม่ต่อเนื่องจะใช้เครื่องหมาย "=" และใช้เงื่อนไข "AND" ในการเชื่อมโยงระหว่างแอตทริบิวต์ โดยจะเชื่อมโยงไปจนถึงแอตทริบิวต์ที่สามารถแบ่งประเภทได้ และจะมีข้อมูลความถูกต้องของกฎ

SPEED > 0: Abnormal (139321.0) ภายใต้การแตกกิ่งของ ENGINE_STAT <= 0 และ INT_BATT_VDC <= 4 หมายถึง ถ้าความเร็วของรถ มีค่ามากกว่า 0 แต่สถานะของเครื่องยนต์เป็น 0 หรือเครื่องยนต์ดับอยู่ และภายใต้ของค่าแบตเตอรี่รถน้อยกว่าหรือเท่ากับ 4 จะเป็นข้อมูลที่ผิดปกติ และมีข้อมูลที่สอดคล้องหรือให้ผลลัพธ์ตรงกันจำนวน 139,321 ตัวอย่าง และให้ผลลัพธ์ที่ไม่ตรงกัน 0 ตัวอย่าง



ภาพที่ 4.5 Classifier errors ของแบบจำลองต้นไม้ตัดสินใจ

เมื่อ แกน X คือ STATUS ของข้อมูลจริง
แกน Y คือ predicted STATUS

และ สีน้ำเงิน คือ คลาส Normal
สีแดง คือ คลาส Abnormal

จากภาพที่ 4.5 แสดงการเปรียบเทียบข้อมูลที่ได้จากการทำนายของแบบจำลองต้นไม้ตัดสินใจกับชุดข้อมูลจริงที่ใช้ในการฝึกสอน โดยที่แกน X คือแอตทริบิวต์ Status ของชุดข้อมูลฝึกสอน และแกน Y คือ Predict Status หรือข้อมูลที่แบบจำลองทำนายได้ ซึ่งสีน้ำเงินหมายถึงข้อมูลที่มีสถานะปกติ และสีแดงหมายถึงข้อมูลที่มีสถานะผิดปกติ โดยที่รูปสี่เหลี่ยมหมายถึงค่าผิดพลาดของการทำนายที่แบบจำลองทำนายได้ต่างจากข้อมูลจริง ซึ่งค่าความผิดพลาดที่ได้มีค่าผลลบลวงเท่ากับ 13 คือรูปสี่เหลี่ยมสีน้ำเงิน และผลบวกลวงเท่ากับ 73 คือรูปสี่เหลี่ยมสีแดง โดยผลบวกจริงเท่ากับ 200987 และผลลบจริงเท่ากับ 200927

ผลการรันฉบับเต็มของ Decision Tree (C4.5)

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Data

Instances: 402000

Attributes: 9

SPEED

ENGINE_STAT

DRIVER_LIC_INFO

HDOP

SAT_NO

RSSI

INT_BATT_VDC

EXT_BATT_VDC

STATUS

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

```
INT_BATT_VDC <= 4
|  ENGINE_STAT <= 0
|  |  SPEED <= 0
|  |  |  DRIVER_LIC_INFO <= 0
|  |  |  |  RSSI <= 7
|  |  |  |  |  HDOP <= 8
|  |  |  |  |  |  INT_BATT_VDC <= 3.9: Normal (3.0/1.0)
|  |  |  |  |  |  INT_BATT_VDC > 3.9: Abnormal (19.0/2.0)
|  |  |  |  |  HDOP > 8: Normal (6.0/2.0)
|  |  |  |  RSSI > 7: Normal (41143.0/7.0)
|  |  |  DRIVER_LIC_INFO > 0: Abnormal (57672.0/1.0)
|  |  SPEED > 0: Abnormal (139321.0)
|  ENGINE_STAT > 0
|  |  SPEED <= 100
|  |  |  RSSI <= 8
|  |  |  |  DRIVER_LIC_INFO <= 0: Normal (213.0/1.0)
|  |  |  |  DRIVER_LIC_INFO > 0
|  |  |  |  |  EXT_BATT_VDC <= 26
|  |  |  |  |  |  RSSI <= 7: Abnormal (5.0)
|  |  |  |  |  |  RSSI > 7: Normal (48.0)
|  |  |  |  |  EXT_BATT_VDC > 26
|  |  |  |  |  |  HDOP <= 6
|  |  |  |  |  |  |  HDOP <= 0: Abnormal (4.0)
|  |  |  |  |  |  |  HDOP > 0: Normal (56.0)
|  |  |  |  |  |  HDOP > 6: Normal (799.0/10.0)
```

```
|  |  |  RSSI > 8: Normal (63222.0/2.0)
|  |  SPEED > 100: Abnormal (260.0)
INT_BATT_VDC > 4
|  SPEED <= 0
|  |  RSSI <= 7
|  |  |  SAT_NO <= 10
|  |  |  |  INT_BATT_VDC <= 4.2
|  |  |  |  |  RSSI <= 6: Normal (57.0/5.0)
|  |  |  |  |  RSSI > 6: Abnormal (3.0/1.0)
|  |  |  |  INT_BATT_VDC > 4.2: Abnormal (7.0)
|  |  |  SAT_NO > 10
|  |  |  |  EXT_BATT_VDC <= 27: Abnormal (33.0)
|  |  |  |  EXT_BATT_VDC > 27
|  |  |  |  |  ENGINE_STAT <= 0: Normal (2.0)
|  |  |  |  |  ENGINE_STAT > 0: Abnormal (4.0)
|  |  RSSI > 7
|  |  |  DRIVER_LIC_INFO <= 0: Normal (94167.0/26.0)
|  |  |  DRIVER_LIC_INFO > 0
|  |  |  |  ENGINE_STAT <= 0: Abnormal (35.0)
|  |  |  |  ENGINE_STAT > 0
|  |  |  |  |  SAT_NO <= 10
|  |  |  |  |  |  SAT_NO <= 0: Abnormal (7.0)
|  |  |  |  |  |  SAT_NO > 0: Normal (47.0)
|  |  |  |  SAT_NO > 10: Normal (230.0/1.0)
|  SPEED > 0
```

| | ENGINE_STAT <= 0: Abnormal (3534.0)

| | ENGINE_STAT > 0

| | | RSSI <= 7: Abnormal (45.0/4.0)

| | | RSSI > 7: Normal (1058.0/4.0)

Number of Leaves : 28

Size of the tree : 55

Time taken to build model: 13.97 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	401914	99.9786 %
Incorrectly Classified Instances	86	0.0214 %
Kappa statistic	0.9996	
Mean absolute error	0.0004	
Root mean squared error	0.0145	
Relative absolute error	0.0746 %	
Root relative squared error	2.8919 %	
Total Number of Instances	402000	

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area
PRC Area  Class
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000    1.000
Normal
                 1.000    0.000    1.000      1.000   1.000      1.000    1.000    1.000
Abnormal
Weighted Avg.    1.000    0.000    1.000      1.000   1.000      1.000    1.000    1.000

=== Confusion Matrix ===

     a      b   <-- classified as
 200987     13 |     a = Normal
     73 200927 |     b = Abnormal
```

ภาพที่ 4.6 ผลการรันฉบับเต็มของต้นไม้ตัดสินใจ

จากภาพที่ 4.6 แสดงผลลัพธ์ฉบับเต็มที่ได้จากการวิเคราะห์ด้วยเทคนิคต้นไม้ตัดสินใจ สามารถอธิบายได้คือ ที่โหนดรากได้แก่แอตทริบิวต์ INT_BATT_VDC สามารถแตกกิ่งหรือเส้นเชื่อม 2 เส้น ซึ่งเส้นเชื่อมแรกคือ ค่าของข้อมูลที่ INT_BATT_VDC <= 4 และเส้นเชื่อมที่สองมีค่า INT_BATT_VDC > 4 และจากโหนดรากหรือจากกิ่ง INT_BATT_VDC <= 4 สามารถสร้างโหนดลูกได้แก่ ENGINE_STAT และจากกิ่ง INT_BATT_VDC > 4 สามารถสร้างโหนดลูกได้แก่ SPEED ซึ่งที่โหนด ENGINE_STAT สามารถแตกกิ่งได้เป็น 2 เส้นคือ ENGINE_STAT เป็น 0 และ ENGINE_STAT เป็น 1 และภายใต้การแตกกิ่งของ ENGINE_STAT เท่ากับ 0 สามารถสร้างโหนดลูกได้คือ SPEED จากโหนด SPEED สามารถแตกกิ่งออกเป็น SPEED <= 0 และ SPEED > 0: Abnormal โดยที่ภายใต้ SPEED มากกว่า 0 แล้วจะเป็นข้อมูลที่มีสถานะผิดปกติ โดยค่าความถูกต้องของแบบจำลองที่ได้จากการวิเคราะห์ของต้นไม้ตัดสินใจเท่ากับ 99.9786% และค่าความผิดพลาดของแบบจำลองเท่ากับ 0.0214%

4.1.2 ป่าของต้นไม้ตัดสินใจ

ป่าของต้นไม้ตัดสินใจ [4] ประกอบไปด้วยต้นไม้ตัดสินใจจำนวนหลายต้น แต่ละต้นเป็นอิสระต่อกัน ในการทำนายจะเริ่มต้นจากนำข้อมูลไปทำนายในต้นไม้ตัดสินใจแต่ละต้น โดยต้นไม้แต่ละต้นจะทำการจำแนกประเภทหรือทำนายเอาต์พุตออกมา หลังจากนั้นเอาต์พุตสุดท้ายจะได้มาจากการโหวตของต้นไม้ตัดสินใจแต่ละต้น โดยเลือกค่าที่ได้รับการโหวตมากที่สุด และโหนดใบของต้นไม้ตัดสินใจจะแสดงสถานะของข้อมูลหรือคลาสที่กำหนดไว้ 2 คลาส คือ ข้อมูลปกติ และข้อมูลที่ไม่ปกติ ผลการวิเคราะห์ข้อมูลการเรียนรู้ป่าของต้นไม้ตัดสินใจจากชุดข้อมูลได้ความถูกต้องและความผิดพลาด แสดงดังในภาพที่ 4.7

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      401939               99.9848 %
Incorrectly Classified Instances        61                0.0152 %
Kappa statistic                          0.9997
Mean absolute error                      0.0003
Root mean squared error                  0.0115
Relative absolute error                  0.0516 %
Root relative squared error              2.2945 %
Total Number of Instances           402000
```

ภาพที่ 4.7 สรุปผลของการทำนายด้วยเทคนิคป่าของต้นไม้ตัดสินใจ

จากภาพที่ 4.7 อธิบายได้ว่า สำหรับข้อมูลติดตามสถานะรถขนส่งน้ำมัน จำนวน 402,000 ตัวอย่าง พบว่ามีความถูกต้องคิดเป็นร้อยละ 99.9848% และมีความผิดพลาดร้อยละ 0.0152% ซึ่งแบบจำลองที่ได้จากการใช้เทคนิคป่าของต้นไม้ตัดสินใจมีเปอร์เซ็นต์ความถูกต้องสูงกว่าเทคนิคต้นไม้ตัดสินใจอยู่ 0.0062%

```
=== Confusion Matrix ===

      a      b   <-- classified as
 200980     20 |     a = Normal
     41 200959 |     b = Abnormal
```

ภาพที่ 4.8 ผลลัพธ์ค่าเมทริกซ์ความสับสนที่ได้ของป่าของต้นไม้ตัดสินใจ

จากภาพที่ 4.8 แสดงค่าผลลัพธ์ของตารางเมทริกซ์ความสับสนที่ได้จากแบบจำลองของป่าของต้นไม้ตัดสินใจ โดย a จากตารางแทนข้อมูลที่มีแอตทริบิวต์ Status มีค่าเป็น Normal และ b แทนข้อมูลที่มีแอตทริบิวต์ Status มีค่าเป็น Abnormal เป็นการเปรียบเทียบระหว่างข้อมูลจริงและข้อมูลที่แบบจำลองทำนายได้ โดยค่าความผิดพลาดที่ได้มีค่าผลลบลวงเท่ากับ 20 และผลบวกลวงเท่ากับ 41

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure
                 1.000     0.000     1.000       1.000    1.000
                 1.000     0.000     1.000       1.000    1.000
Weighted Avg.    1.000     0.000     1.000       1.000    1.000
```

ภาพที่ 4.9 ผลลัพธ์ค่าความถูกต้องโดยคำนวณเป็นคลาสของป่าของต้นไม้ตัดสินใจ

จากภาพที่ 4.9 แสดงผลลัพธ์ที่ได้จากแบบจำลองของป่าของต้นไม้ตัดสินใจ โดยคำนวณค่าความถูกต้องเป็นคลาส ซึ่งผลลัพธ์บรรทัดแรกเป็นการคำนวณของคลาส Normal โดยการคำนวณได้ค่าความแม่นยำ ค่าความระลึก และค่าการวัดประสิทธิภาพ เท่ากับ 1 หรือ 100% และผลลัพธ์บรรทัดที่สองเป็นการคำนวณของคลาส Abnormal โดยการคำนวณได้ค่าความแม่นยำ ค่าความระลึก และค่าการวัดประสิทธิภาพ เท่ากับ 1 หรือ 100%

ภาพที่ 4.10 Classifier errors ของแบบจำลองป่าของต้นไม้ตัดสินใจ

เมื่อ แกน X คือ STATUS ของข้อมูลจริง
แกน Y คือ predicted STATUS

และ สีน้ำเงิน คือ คลาส Normal
สีแดง คือ คลาส Abnormal

จากภาพที่ 4.10 แสดงการเปรียบเทียบข้อมูลที่ได้จากการทำนายของแบบจำลองเทียบกับข้อมูลจริง ซึ่งสีน้ำเงินหมายถึงข้อมูลที่มีสถานะปกติ และสีแดงหมายถึงข้อมูลที่มีสถานะผิดปกติ โดยที่สี่เหลี่ยมหมายถึงค่าผิดพลาดของการทำนาย ซึ่งค่าความผิดพลาดที่ได้มีค่าผลลบลวงเท่ากับ 20 คือรูปสี่เหลี่ยมสีน้ำเงิน และผลบวกลวงเท่ากับ 41 คือรูปสี่เหลี่ยมสีแดง

ผลการรันฉบับเต็มของ Random Forest

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: Data

Instances: 402000

Attributes: 9

SPEED

ENGINE_STAT

DRIVER_LIC_INFO

HDOP

SAT_NO

RSSI

INT_BATT_VDC

EXT_BATT_VDC

STATUS

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 246.36 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	401939	99.9848 %
Incorrectly Classified Instances	61	0.0152 %
Kappa statistic	0.9997	
Mean absolute error	0.0003	
Root mean squared error	0.0115	
Relative absolute error	0.0516 %	
Root relative squared error	2.2945 %	
Total Number of Instances	402000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Normal
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Abnormal
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

```
=== Confusion Matrix ===

     a      b   <-- classified as
200980     20 |     a = Normal
    41 200959 |     b = Abnormal
```

ภาพที่ 4.11 ผลการรันฉบับเต็มของป่าของต้นไม้ตัดสินใจ

จากภาพที่ 4.11 แสดงผลลัพธ์ฉบับเต็มของเทคนิคป่าของต้นไม้ตัดสินใจ ซึ่งได้ค่าความถูกต้องของแบบจำลองเท่ากับ 99.9848% และจากตารางเมทริกซ์ความสับสน มีค่าผลบวกจริงเท่ากับ 200980 ผลลบลวงเท่ากับ 20 ผลลบจริงเท่ากับ 200959 และผลบวกลวงเท่ากับ 41

4.1.3 โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น

โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น [6] เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบชั้น และใช้ขั้นตอนการส่งค่าย้อนกลับ สำหรับการฝึกฝน ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นของข้อมูลเข้าและจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ผลการวิเคราะห์ข้อมูลการเรียนรู้โครงข่ายประสาทเทียมแบบหลายชั้นจากชุดข้อมูลได้ความถูกต้องและความผิดพลาด แสดงดังในภาพที่ 4.12

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      401697               99.9246 %
Incorrectly Classified Instances       303                0.0754 %
Kappa statistic                          0.9985
Mean absolute error                      0.0009
Root mean squared error                  0.0265
Relative absolute error                  0.1795 %
Root relative squared error              5.3047 %
Total Number of Instances           402000
```

ภาพที่ 4.12 สรุปผลของการทำนายด้วยเทคนิคเพอร์เซ็ปตรอนแบบหลายชั้น

จากภาพที่ 4.12 อธิบายได้ว่า สำหรับข้อมูลติดตามสถานะรถขนส่งน้ำมัน จำนวน 402,000 ตัวอย่าง พบว่ามีความถูกต้องคิดเป็นร้อยละ 99.9246% และมีความผิดพลาดร้อยละ 0.0754% ซึ่งค่าความถูกต้องของเทคนิคโครงข่ายประสาทเทียมมีค่าน้อยกว่าทั้ง 2 เทคนิคข้างต้น คือ น้อยกว่าเทคนิคต้นไม้ตัดสินใจ และ ป่าของต้นไม้ตัดสินใจ

```
=== Confusion Matrix ===

     a      b   <-- classified as
200993      7 |      a = Normal
   296 200704 |      b = Abnormal
```

ภาพที่ 4.13 ผลลัพธ์ค่าเมทริกซ์ความสับสนที่ได้ของเพอร์เซ็ปตรอนแบบหลายชั้น

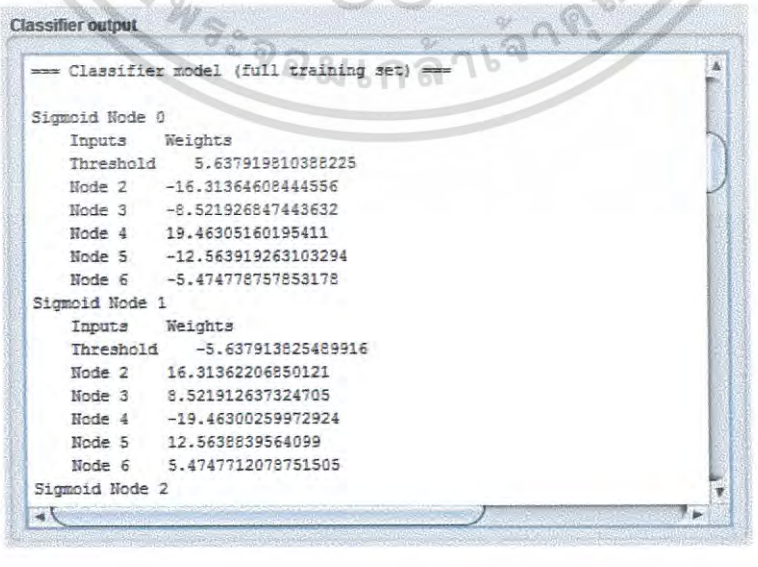
จากภาพที่ 4.13 แสดงค่าผลลัพธ์ของตารางเมทริกซ์ความสับสนที่ได้จากแบบจำลองของเพอร์เซ็ปตรอนแบบหลายชั้น โดย a แทนข้อมูลที่มีแอตทริบิวต์ Status มีค่าเป็น Normal และ b แทนข้อมูลที่มีแอตทริบิวต์ Status มีค่าเป็น Abnormal เป็นการเปรียบเทียบระหว่างข้อมูลจริง และข้อมูลที่แบบจำลองทำนายได้ โดยค่าความผิดพลาดที่ได้มีค่าผลลบลวงเท่ากับ 7 และผลบวกลวงเท่ากับ 296

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 1.000    0.001    0.999      1.000    0.999
                 0.999    0.000    1.000      0.999    0.999
Weighted Avg.    0.999    0.001    0.999      0.999    0.999
```

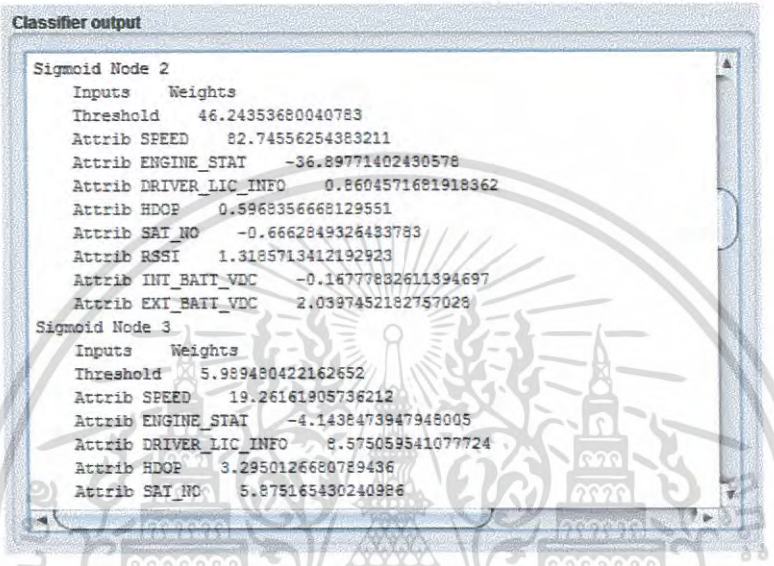
ภาพที่ 4.14 ผลลัพธ์ค่าความถูกต้องโดยคำนวณเป็นคลาสเพอร์เซ็ปตรอนแบบหลายชั้น

จากภาพที่ 4.14 แสดงผลลัพธ์ที่ได้จากแบบจำลองของของโครงข่ายประสาทเทียมแบบหลายชั้น โดยคำนวณค่าความถูกต้องเป็นคลาส ซึ่งผลลัพธ์บรรทัดแรกเป็นการคำนวณของคลาส Normal โดยการคำนวณได้ค่าความแม่นยำ และค่าการวัดประสิทธิภาพ เท่ากับ 0.999 หรือ 99.9% คำนวณค่าความระลึกได้เท่ากับ 1 หรือ 100% และผลลัพธ์บรรทัดที่สองเป็นการคำนวณของคลาส Abnormal โดยการคำนวณได้ค่าความแม่นยำเท่ากับ 1 หรือ 100% คำนวณค่าความระลึก และค่าการวัดประสิทธิภาพได้เท่ากับ 0.999 หรือ 99.9%



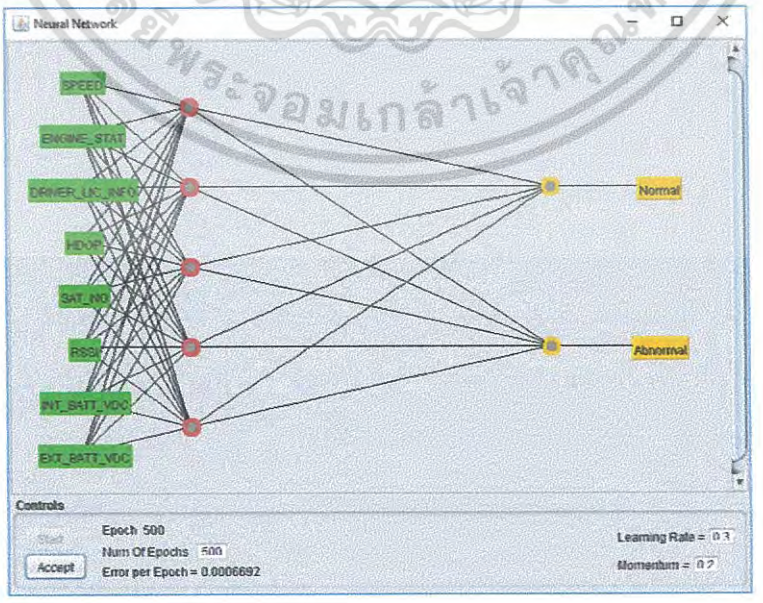
ภาพที่ 4.15 ผลลัพธ์บางส่วนจากการจำแนกข้อมูลแบบเพอร์เซ็ปตรอนแบบหลายชั้น

จากการทดสอบ ผลลัพธ์ที่ได้จะแสดงค่าน้ำหนักของแต่ละเส้นข้อมูลนำเข้าภายใน Sigmoid Node นั้น ๆ เช่น Sigmoid Node 0 มีค่าน้ำหนักของ Threshold เท่ากับ 5.637919810388225 และมีค่าน้ำหนักของโหนดนำเข้าจาก Node 2 เท่ากับ -16.31364608444556 ค่าน้ำหนักของโหนดนำเข้าจาก Node 3 เท่ากับ -8.521926847443632 เป็นต้น โดยที่คลาส Normal มี Input คือ Node 0 และคลาส Abnormal มี Input คือ Node 1



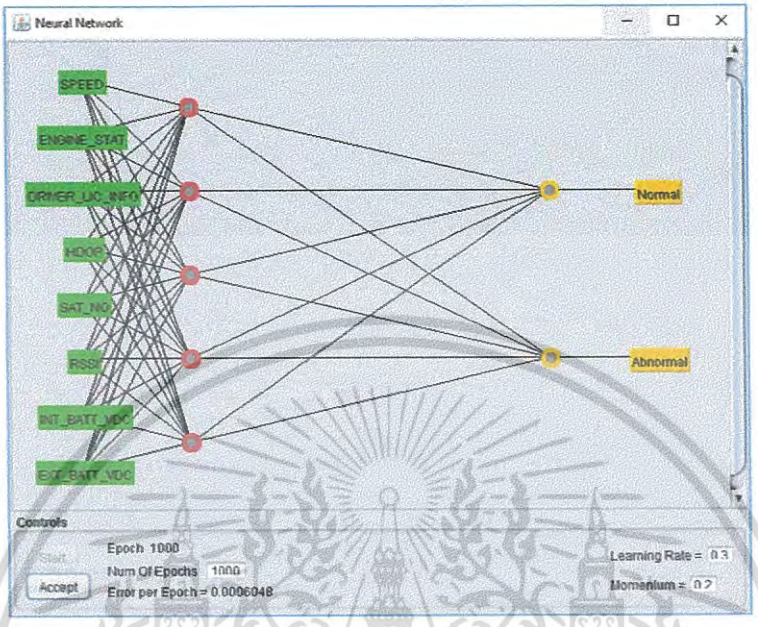
ภาพที่ 4.16 ผลลัพธ์บางส่วนจากการจำแนกข้อมูลแบบเพอร์เซ็ปตรอนแบบหลายชั้น

จากภาพที่ 4.16 ที่ Node 2 มีเส้นข้อมูลนำเข้าแต่ละเส้นมาจาก Attribute ทั้ง 8 Attributes โดยน้ำหนักของแต่ละเส้นข้อมูลนำเข้าจะแตกต่างกันไป และที่ Node 2 มีค่า Threshold เท่ากับ 46.243536800407



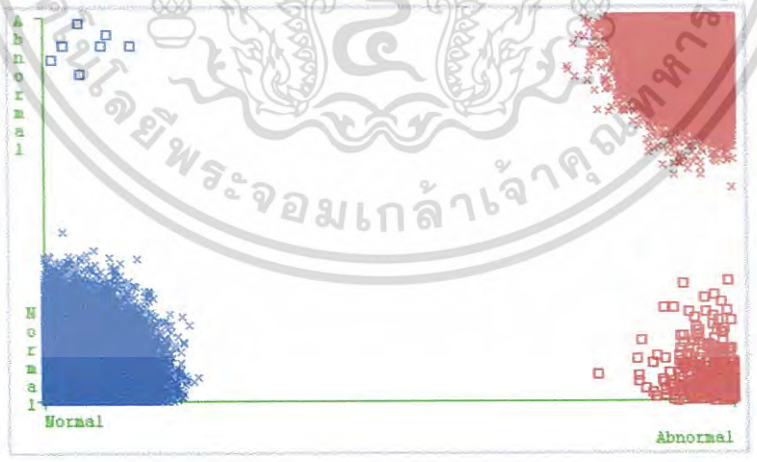
ภาพที่ 4.17 โครงสร้างแบบจำลองโครงข่ายประสาทเทียม 500 รอบ

จากผลลัพธ์ที่ได้ ในการรันแบบจำลอง 500 รอบ โดยมีค่าอัตราการเรียนรู้อยู่ที่ 0.3 และค่าโมเมนตัม เท่ากับ 0.2 ได้ค่า error ต่อรอบอยู่ที่ 0.0006692



ภาพที่ 4.18 โครงสร้างแบบจำลองโครงข่ายประสาทเทียม 1000 รอบ

จากผลลัพธ์ที่ได้ ในการรันแบบจำลอง 1000 รอบ โดยมีค่าอัตราการเรียนรู้อยู่ที่ 0.3 และค่าโมเมนตัม เท่ากับ 0.2 ได้ค่า error ต่อรอบอยู่ที่ 0.0006048 ซึ่งมีค่า error น้อยกว่าการรัน 500 รอบอยู่ 0.0000644



ภาพที่ 4.19 Classifier errors ของแบบจำลอง Multilayer Perceptron

เมื่อ แกน X คือ STATUS ของข้อมูลจริง
แกน Y คือ predicted STATUS

และ สีน้ำเงิน คือ คลาส Normal
สีแดง คือ คลาส Abnormal

จากภาพที่ 4.19 แสดงการเปรียบเทียบข้อมูลที่ได้จากการทำนายของแบบจำลองเทียบกับข้อมูลจริง ซึ่งสีน้ำเงินหมายถึงข้อมูลที่ปกติ และสีแดงหมายถึงข้อมูลที่ผิดปกติ โดยที่สี่เหลี่ยมหมายถึงค่าผิดพลาดของการทำนาย ซึ่งค่าความผิดพลาดที่ได้มีค่าผลลบลวงเท่ากับ 7 คือรูปสี่เหลี่ยมสีน้ำเงิน และผลบวกลวงเท่ากับ 296 คือรูปสี่เหลี่ยมสีแดง

ผลการรันฉบับเต็มของ Multilayer Perceptron

```
=== Run information ===

Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V
0 -S 0 -E 20 -H a
Relation:     final1
Instances:    402000
Attributes:   9
              SPEED
              ENGINE_STAT
              DRIVER_LIC_INFO
              HDOP
              SAT_NO
              RSSI
              INT_BATT_VDC
              EXT_BATT_VDC
              STATUS
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Sigmoid Node 0
```

Inputs Weights

Threshold 5.637919810388225

Node 2 -16.31364608444556

Node 3 -8.521926847443632

Node 4 19.46305160195411

Node 5 -12.563919263103294

Node 6 -5.474778757853178

Sigmoid Node 1

Inputs Weights

Threshold -5.637913825489916

Node 2 16.31362206850121

Node 3 8.521912637324705

Node 4 -19.46300259972924

Node 5 12.5638839564099

Node 6 5.4747712078751505

Sigmoid Node 2

Inputs Weights

Threshold 46.24353680040783

Attrib SPEED 82.74556254383211

Attrib ENGINE_STAT -36.89771402430578

Attrib DRIVER_LIC_INFO 0.8604571681918362

Attrib HDOP 0.5968356668129551

Attrib SAT_NO -0.6662849326433783

Attrib RSSI 1.3185713412192923

Attrib INT_BATT_VDC -0.16777832611394697

Attrib EXT_BATT_VDC 2.0397452182757028

Sigmoid Node 3

Inputs Weights

Threshold 5.989480422162652

Attrib SPEED 19.26161905736212

Attrib ENGINE_STAT -4.14384739479480005

Attrib DRIVER_LIC_INFO 8.575059541077724

Attrib HDOP 3.2950126680789436

Attrib SAT_NO 5.875165430240986

Attrib RSSI -10.586337503658301

Attrib INT_BATT_VDC 8.904824056088795

Attrib EXT_BATT_VDC -21.339675209496196

Sigmoid Node 4

Inputs Weights

Threshold -52.247354163021754

Attrib SPEED -87.80684377042232

Attrib ENGINE_STAT 39.80680880748771

Attrib DRIVER_LIC_INFO -4.416784162949497

Attrib HDOP -0.21284143066290478

Attrib SAT_NO -1.65034083623704

Attrib RSSI 6.293551136363971

Attrib INT_BATT_VDC 0.7392445577386461

Attrib EXT_BATT_VDC -5.364516944131885

Sigmoid Node 5

Inputs Weights

Threshold 20.229926200261723

Attrib SPEED 38.14958338011225

Attrib ENGINE_STAT -14.628416687503039

Attrib DRIVER_LIC_INFO -2.2076067017297656

Attrib HDOP 0.4923669276301948

Attrib SAT_NO -6.60688442992643

Attrib RSSI 16.466160641799632

Attrib INT_BATT_VDC 2.3474447789633848

Attrib EXT_BATT_VDC -3.795592242557512

Sigmoid Node 6

Inputs Weights

Threshold 2.4583081907263455

Attrib SPEED 9.914689130214903

Attrib ENGINE_STAT -1.2172596071005377

Attrib DRIVER_LIC_INFO 2.9572810697595124

Attrib HDOP 3.335107808387892

Attrib SAT_NO -1.0477696630251825

Attrib RSSI -26.670701300912818

Attrib INT_BATT_VDC -2.852207112315339

Attrib EXT_BATT_VDC -21.948941449626815

Class Normal

Input

Node 0

Class Abnormal

Input

Node 1

Time taken to build model: 428.23 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	401697	99.9246 %
Incorrectly Classified Instances	303	0.0754 %
Kappa statistic	0.9985	
Mean absolute error	0.0009	
Root mean squared error	0.0265	
Relative absolute error	0.1795 %	
Root relative squared error	5.3047 %	
Total Number of Instances	402000	

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area
PRC Area  Class
                 1.000    0.001    0.999      1.000    0.999      0.998    1.000
0.999     Normal
                 0.999    0.000    1.000      0.999    0.999      0.998    1.000
1.000     Abnormal
Weighted Avg.    0.999    0.001    0.999      0.999    0.999      0.998    1.000
0.999

=== Confusion Matrix ===

     a      b   <-- classified as
 200993      7 |     a = Normal
    296 200704 |     b = Abnormal
```

ภาพที่ 4.20 ผลการรันฉบับเต็มของเพอร์เซ็ปตรอนแบบหลายชั้น

จากภาพที่ 4.20 แสดงผลลัพธ์ฉบับเต็มที่ได้จากเทคนิคเพอร์เซ็ปตรอนแบบหลายชั้น จะแสดงค่าน้ำหนักของแต่ละเส้นข้อมูลนำเข้าภายใน Sigmoid Node นั้น ๆ เช่น Sigmoid Node 0 มีค่าน้ำหนักของ Threshold เท่ากับ 5.637919810388225 และมีค่าน้ำหนักของโหนดนำเข้าจาก Node 2 เท่ากับ -16.31364608444556 ค่าน้ำหนักของโหนดนำเข้าจาก Node 3 เท่ากับ -8.521926847443632 เป็นต้น โดยที่คลาส Normal มี Input คือ Node 0 และคลาส Abnormal มี Input คือ Node 1 ซึ่งค่าความถูกต้องของแบบจำลองที่ได้ของเพอร์เซ็ปตรอนแบบหลายชั้น เท่ากับ 99.9246% และค่าความผิดพลาดของแบบจำลองเท่ากับ 0.0754%

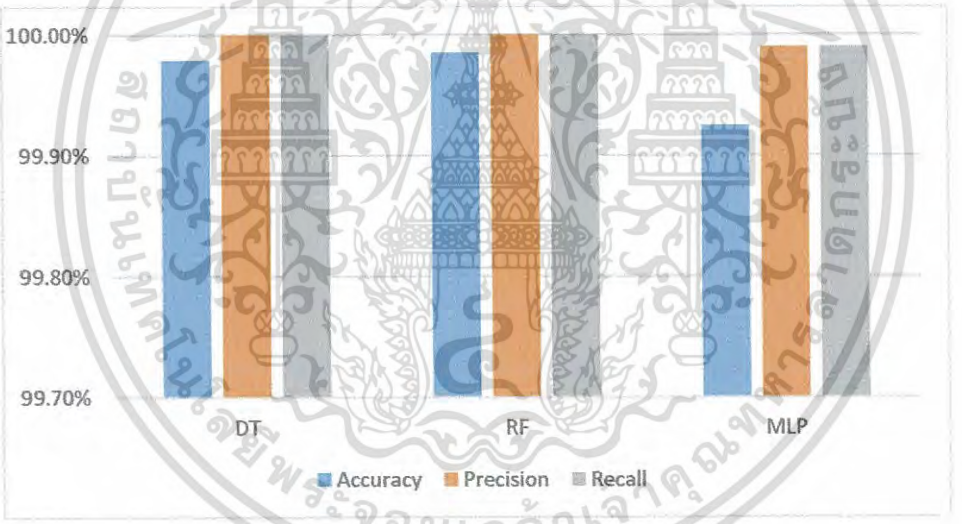
4.2 เปรียบเทียบประสิทธิภาพของแต่ละเทคนิค

จากการทดสอบแบบจำลองทั้ง 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ ป่าของต้นไม้ตัดสินใจ และโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น จะได้ค่าความถูกต้องของแต่ละแบบจำลอง รวมถึงเวลาที่ใช้ในการสร้างแบบจำลอง โดยจะทำการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค เพื่อหาแบบจำลองที่ให้ผลลัพธ์ที่ดีที่สุด และเหมาะสมที่สุดในการนำไปประยุกต์ใช้งาน

ตารางที่ 4.1 เปรียบเทียบประสิทธิภาพแต่ละเทคนิค

Classifier	Instances	Correctly	Incorrectly	Time to build model
Decision Tree	402,000	99.9786%	0.0214%	13.97 seconds
Random Forest	402,000	99.9848%	0.0152%	246.36 seconds
Multilayer Perceptron	402,000	99.9246%	0.0754%	428.23 seconds

จากตารางที่ 4.1 จะเห็นได้ว่า วิธีป่าของต้นไม้ตัดสินใจให้ค่าความถูกต้องที่สูงที่สุด รองลงมาคือ วิธีต้นไม้ตัดสินใจ และวิธีที่ให้ค่าความถูกต้องน้อยที่สุดคือ วิธีโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น และวิธีโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้นนั้นยังใช้เวลาในการสร้างแบบจำลองนานที่สุดอีกด้วย



ภาพที่ 4.21 แผนภูมิแสดงการเปรียบเทียบประสิทธิภาพของแต่ละเทคนิค

โดยที่ DT หมายถึง Decision Tree
RF หมายถึง Random Forest
MLP หมายถึง Multilayer Perceptron

จากภาพที่ 4.21 เป็นกราฟแสดงค่าความถูกต้องของการทำนายของแต่ละเทคนิค จากกราฟจะพบว่าเทคนิคต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ จะให้ค่าความถูกต้องที่ใกล้เคียงกัน โดยเทคนิคป่าของต้นไม้ตัดสินใจ จะมีความถูกต้องมากกว่าเทคนิคต้นไม้ตัดสินใจ อยู่ 0.0062% ส่วนเทคนิคโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้นจะมีค่าความถูกต้องน้อยที่สุด ซึ่งน้อยกว่าเทคนิคต้นไม้ตัดสินใจอยู่ 0.054% และน้อยกว่าเทคนิคป่าของต้นไม้ตัดสินใจอยู่ 0.0602%

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การวิจัยเรื่อง การสร้างแบบจำลองการหาความผิดปกติของข้อมูลระบบบันทึกข้อมูลพฤติกรรมการขับขี่โดยการทำเหมืองข้อมูล มีวัตถุประสงค์เพื่อคัดแยกข้อมูลที่ผิดปกติออกจากข้อมูลปกติ เพื่อเพิ่มความถูกต้องและความน่าเชื่อถือของระบบ โดยมีขั้นตอนหลัก ๆ 3 ส่วน ได้แก่

ส่วนที่ 1 ศึกษาและเก็บรวบรวมข้อมูลที่จะนำไปใช้ในการวิเคราะห์ ซึ่งเป็นข้อมูลระบบติดตามสถานะรถขนส่งผลิตภัณฑ์ ภายในเดือน ตุลาคม พ.ศ.2559 จำนวน 402,000 ตัวอย่าง

ส่วนที่ 2 เตรียมข้อมูลให้เป็นรูปแบบมาตรฐานเดียวกัน โดยทำการคัดเลือกเฉพาะข้อมูลที่เกี่ยวข้อง กลั่นกรองข้อมูลด้วยการกำจัดข้อมูลที่ผิดปกติ และแปรงรูปข้อมูลให้เหมาะสมในการนำไปวิเคราะห์ โดยใช้โปรแกรม Weka 3.8.0 [9] ในการทำเหมืองข้อมูล เพื่อหาแบบจำลองที่ใช้ในการคัดแยกข้อมูลที่ผิดปกติ

ส่วนที่ 3 การทดสอบข้อมูลด้วยการทำเหมืองข้อมูล และเลือกเทคนิคที่เหมาะสม ซึ่งในการศึกษาวิจัยครั้งนี้ ได้ทดสอบเทคนิคการจำแนกประเภท จำนวน 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ ป่าของต้นไม้ตัดสินใจ และโครงข่ายประสาทเทียมเพอร์เซ็ปตรอนแบบหลายชั้น จากการทดสอบการจำแนกประเภทข้อมูลทั้ง 3 เทคนิค พบว่า การจำแนกประเภทแบบ ป่าของต้นไม้ตัดสินใจ ให้ผลลัพธ์ที่ดีที่สุด เนื่องจากมีความถูกต้องมากที่สุด โดยผลลัพธ์ที่ได้มีความถูกต้อง 99.9848% ซึ่งมากกว่าเทคนิคต้นไม้ตัดสินใจ อยู่ 0.0062% และมีความถูกต้องมากกว่าเทคนิคโครงข่ายประสาทเทียมแบบหลายชั้น อยู่ 0.0602% โดยแบบจำลองที่ได้สามารถนำไปประยุกต์ใช้กับระบบ เพื่อคัดแยกข้อมูลที่ไม่ปกติออกจากข้อมูลปกติได้

5.2 ข้อเสนอแนะ

งานวิจัยชิ้นนี้ อาจมีข้อบกพร่องและข้อผิดพลาดอยู่บ้าง อันเนื่องมาจากปัจจัยต่าง ๆ ดังนั้น ผู้วิจัยจึงขอเสนอแนะแนวทางในการปรับปรุงงานวิจัยนี้ในครั้งต่อไป

1) การทำเหมืองข้อมูล มีอยู่หลายประเภทและหลายเทคนิค ซึ่งแต่ละเทคนิคมีความเหมาะสมกับความต้องการในการนำแบบจำลองไปใช้ที่ต่างกัน ดังนั้นหากต้องการทำเหมืองข้อมูลจึงต้องทำความเข้าใจถึงหลักการและประโยชน์ของแต่ละเทคนิค และเลือกเทคนิคที่ตรงกับความต้องการให้มากที่สุด

2) การศึกษางานวิจัยครั้งนี้ ผู้วิจัยได้นำข้อมูลระบบติดตามสถานะรถขนส่งผลิตภัณฑ์มาบางส่วนที่ใช้ในการวิเคราะห์และสร้างแบบจำลอง หากจะนำแบบจำลองที่ได้ไปใช้งานจริง ควรมีการนำข้อมูลชุดใหม่มาวิเคราะห์ และประเมินผลอีกครั้ง เพื่อให้ได้แบบจำลองที่ถูกต้อง และแม่นยำมากที่สุด

3) การทำเหมืองข้อมูลเป็นเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล ซึ่งจะช่วยหารูปแบบและความสัมพันธ์ของข้อมูล โดยแบบจำลองที่ได้จากการวิเคราะห์จะใช้ในการทำนายหรือคาดการณ์สำหรับข้อมูลในอนาคต ดังนั้นหากต้องการนำแบบจำลองไปใช้ ผู้ใช้ควรนำแบบจำลองที่ได้มาวิเคราะห์ และตัดสินใจอีกครั้งหนึ่ง ก่อนนำไปใช้งานจริง

เอกสารอ้างอิง

[1] M.Kantardzic, **Data Mining Concepts, Model, Methods, and Algorithm**, John Wiley & Sons, Inc., 2003.

[2] Jiawei Han and Micheline Kamber, **Data Mining Concepts and Techniques**, Morgan Kaufmann Publishers, 2001.

[3] ปกรณ์ จารุตระกูลชัย, **การทำนายปริมาณสารส้มที่ใช้ในกระบวนการผลิตน้ำประปาด้วยวิธีป่าของต้นไม้ตัดสินใจและโปรแกรมเชิงพันธุกรรม**, สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์, พ.ศ.2550.

[4] ปกรณ์ จารุตระกูลชัย, **การทำนายปริมาณสารส้มที่ใช้ในกระบวนการผลิตน้ำประปาด้วยวิธีป่าของต้นไม้ตัดสินใจและโปรแกรมเชิงพันธุกรรม**, สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์, พ.ศ.2550.

[5] พยุง มีสัจ, **ระบบฟัซซีและโครงข่ายประสาทเทียม**, เอกสารประกอบการสอน, คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, พ.ศ.2551.

[6] เอกรินทร์ แซ่เฮ้ง, **โครงข่ายประสาทเทียมกับการประยุกต์ใช้งาน** (ตอนที่ 1 รู้จักกับโครงข่ายประสาทเทียม), แผนกสารสนเทศ สำนักวิชาการ วิทยาลัยนอร์ท.

[7] Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," **Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence**, vol. 2, no. 12, pp. 1137-1143, 1995.

[8] ทิพย์ธิดา วงศ์พิพันธ์, **การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ กรณีศึกษา: บริษัท กรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน)**, สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์, พ.ศ.2556.

[9] WEKA, Machine Learning Group at the University of Waikato.

[10] ทิพย์ธิดา วงศ์พิพันธ์, **การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ กรณีศึกษา: บริษัท กรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน)**, สาขาวิชาเทคโนโลยีคอมพิวเตอร์และการสื่อสาร คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์, พ.ศ.2556.

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวอังคณา อัครวราวงศ์
เกิดเมื่อ	วันที่ 15 เดือน มิถุนายน พ.ศ.2538
การศึกษา	ระดับปริญญาตรี ชั้นปีที่ 4 ภาควิชาวิศวกรรมคอมพิวเตอร์ สาขาวิศวกรรมสารสนเทศ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง