การจำแนกรุ่นของรถยนต์ประเภทรถกระบะด้วยเทคนิคเหมืองข้อมูล กรณีศึกษาโชว์รูมอีซูซุมหานคร สาขามหาชัย

Classifying pickup truck models using data mining techniques: A Case Study of Isuzu Showroom, Mahachai Branch

ณัฐสิทธิ แพวตะคุ

ภาคนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาตรีวิทยาศาสตรบัณฑิต
รายวิชาโครงการเทคโนโลยีสารสนเทศ
ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยสยาม
พ.ศ.2566

หัวข้อภาคนิพนธ์	: การจำแนกรุ่นของรถยนต์ประเภทรถกระบะด้วยเทคนิคเหมืองข้อมูล กรณีศึกษา โชว์รูมอีซูซุมหานคร สาขามหาชัย
	: Classifying pickup truck models using data mining techniques: A Case Study of Isuzu Showroom, Mahachai Branch.
หน่วยกิตของภาคนิพนธ์	: 3 หน่วยกิต
ผู้จัดทำ	: นาย ณัฐสิทธิ แพวตะคุ 6105100013
อาจารย์ที่ปรึกษา	: อาจารย์ณรงค์ฤทธิ์ สุคนธสิงห์
ระดับการศึกษา	: วิทยาศาสตรบัณฑิต
สาขาวิชา	: เทคโนโลยีสารสนเทศ
ปีการศึกษา	: 2566

อนุมัติให้ภาคนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาตรีวิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะกรรมการสอบภาคนิพนธ์

..... ประธานกรรมการ
(ผศ.ดร.ศักดิ์ชาย ตั้งวรรณวิทย์)

..... กรรมการสอบ
(อาจารย์อรรณพ กางกั้น)

..... อาจารย์ที่ปรึกษา
(อาจารย์ณรงค์ฤทธิ์ สุคนธสิงห์)

หัวข้อภาคนิพนธ์	: การจำแนกรุ่นของรถยนต์ประเภทรถกระบะด้วยเทคนิคเหมืองข้อมูล กรณีศึกษา โชว์รูมอีซูซุมหานครสาขามหาชัย		
หน่วยกิตของภาคนิพนธ์	: 3 หน่วยกิต		
ผู้จัดทำ	: นาย ณัฐสิทธิ	แพวตะคุ	6105100013
อาจารย์ที่ปรึกษา	: อาจารย์ณรงค์ฤทธิ์	สุคนธสิงห์	
ระดับการศึกษา	: วิทยาศาสตรบัณฑิต		
สาขาวิชา	: เทคโนโลยีสารสนเทศ		
ปีการศึกษา	: 2566		

บทคัดย่อ

งานวิจัยนี้ มีวัตถุประสงค์เพื่อสร้างตัวแบบที่สำหรับการจำแนกรุ่นของรถยนต์ประเภทรถกระบะ กรณีศึกษาโชว์รูมรถยนต์ อีซูซุมหานคร สาขามหาชัย และเพื่อเปรียบเทียบประสิทธิภาพที่สุดในการจำแนกรุ่นของรถยนต์ประเภทรถกระบะ โดยใช้อัลกอริทึมดังต่อไปนี้ ได้แก่ ต้นไม้ตัดสินใจ นาอีฟเบย์ เรนดอม ฟลอเรส และเพื่อนบ้านใกล้ที่สุด - เค ข้อมูลที่นำมาสร้างตัวแบบเป็นข้อมูลลูกค้าที่ซื้อรถยนต์อีซูซุ สาขามหาชัย เดือนมกราคม พ.ศ. 2563 ถึง เดือนธันวาคม พ.ศ. 2564 จำนวน 1,500 คน โดยใช้วิธีการแบ่งข้อมูลเพื่อทดสอบจำนวน 5 วิธีได้แก่ 1). การแบ่งข้อมูลเพื่อทดสอบแบบ 70:30 2). การแบ่งข้อมูลเพื่อทดสอบแบบ 75:25 3). การแบ่งข้อมูลเพื่อทดสอบแบบ 80:20 4). การวิเคราะห์ข้อมูลด้วยวิธี 5-Fold Cross Validation และ 5). การวิเคราะห์ข้อมูลด้วยวิธี 10- Fold Cross Validation เพื่อทดสอบประสิทธิภาพของตัวแบบ โดยใช้โปรแกรม Rapid Miner X นำเข้าข้อมูลในการสร้างแบบจำลอง โดยมีทั้งหมด 11 คุณลักษณะ จากการทดสอบพบว่าประสิทธิภาพการทำนายของตัวแบบเรนดอม ฟลอเรส ที่ใช้การวิเคราะห์ข้อมูลด้วยวิธี 10- Fold Cross Validation มีค่าความแม่นยำอยู่ที่ 84.13% ซึ่งให้ค่าความแม่นยำมากที่สุด ลำดับถัดมาเป็นตัวแบบต้นไม้ตัดสินใจ มีค่าความแม่นยำอยู่ที่ 82.73% ลำดับถัดมาเป็นตัวแบบเพื่อนบ้านใกล้ที่สุด – เค ให้ค่าความแม่นยำอยู่ที่ 81.00% และลำดับ

สุดท้ายคือตัวแบบนาอีฟเบย์ โดยให้ค่าความแม่นยำอยู่ที่ 62.20% สรุปได้ว่า ตัวแบบที่ทำนายผลของการจำแนกรุ่นของรถยนต์ประเภทรถกระบะ คือ ตัวแบบเรนดอมฟรอเรส โดยให้ค่าความแม่นยำสูงที่สุดอยู่ที่ 84.13%

คำสำคัญ: การจำแนก, ต้นไม้ตัดสินใจ, นาอีฟเบย์, เพื่อนบ้านใกล้ที่สุด – เค, เรนดอมฟลอเรส

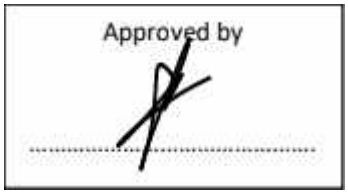
มหาวิทยาลัยสยาม SIAM UNIVERSITY

Project Title	: Classifying pickup truck models using data mining techniques: A Case Study of Isuzu Showroom, Mahachai Branch		
Project Credits	: 3 Credits		
Candidate	: Mr. Natthasitthi	Paewtaku	6105100013
Advisor	: Mr. Narongrit	Sukonthasing	
Program	: Bachelor of Science		
Field of Study	: Information Technology		
Academic Year	: 2023		

Abstract

This research aims to develop a model for classifying pickup truck models using data mining techniques. The case study was conducted at the Isuzu showroom branch in Mahachai. The study also seeks to compare the effectiveness of different algorithms in classifying customer preferences for pickup truck models. The algorithms used in this study included Decision Tree, Naïve Bayes, Random Forest, and K-Nearest Neighbor. The data used to build the model consisted of customer information who purchased Isuzu cars at the Mahachai branch from January 2020 to December 2021, totaling 1,500 individuals. The data was divided into five testing methods, including 70:30, 75:25, and 80:20 data splits, as well as 5-fold and 10-fold Cross Validation. The performance of the models were evaluated using Rapid Miner X, with a total of 11 attributes. The results indicated that the Random Forest model, analyzed with the 10-fold Cross Validation method, achieved the highest accuracy at 84.13%. The Decision Tree model followed with an accuracy of 82.73%, the K-Nearest Neighbor model with 81.00%, and the Naïve Bayes model with 62.20%. In conclusion, the Random Forest model demonstrated the highest accuracy in predicting customer preferences for pickup truck models, reaching 84.13%.

Keywords: classification, Decision Tree, K-Nearest Neighbor, Naïve Bayes, Random Forest



กิตติกรรมประกาศ

งานวิจัยเรื่อง การจำแนกรุ่นของรถยนต์ประเภทรถกระบะด้วยเทคนิคเหมืองข้อมูล กรณีศึกษาโชว์รูมอีซูซุมหานคร สาขามหาชัย สามารถดำเนินได้เสร็จเรียบร้อย เนื่องจากได้รับความกรุณาให้คำแนะนำและการช่วยเหลือจากคณะอาจารย์ในภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยสยาม ให้ผู้วิจัยสามารถดำเนินการวิจัยได้ลุล่วงตามวัตถุประสงค์

ขอขอบพระคุณ อาจารย์ณรงค์ฤทธิ์ สุคนธสิงห์ ที่ให้ความรู้ คำแนะนำและความช่วยเหลือในการจัดทำงานวิจัยฉบับนี้จนเสร็จสมบูรณ์ไปได้ด้วยดี

ขอขอบพระคุณ ท่านผู้จัดการโชว์รูมอีซูซุมหานครสาขามหาชัย และพนักงานที่ปรึกษาการขายที่กรุณาให้ข้อมูลของลูกค้าในการดำเนินการทำงานวิจัยในครั้งนี้

ผู้จัดทำ

สารบัญ

หน้า

บทคัดย่อ ..... ข
กิตติกรรมประกาศ ..... จ
สารบัญ ..... ฉ
สารบัญรูปภาพ ..... ซ
สารบัญตาราง ..... ฑ
บทที่
1 บทนำ
ความเป็นมาและความสำคัญของปัญหา 1
วัตถุประสงค์ของการวิจัย 3
ขอบเขตการวิจัย 3
ขั้นตอนการดำเนินงานวิจัย 3
ประโยชน์ที่คาดว่าจะได้รับ 4
2 งานวิจัยและทฤษฎีที่เกี่ยวข้อง
ทฤษฎีการทำเหมืองข้อมูล 5
ทฤษฎีต้นไม้ตัดสินใจ (Decision Tree) 11
นาอีฟเบย์ (Naïve Bayes) 13
เรนดอม ฟลอเรส (Random Forest) 16
เพื่อนบ้านใกล้ที่สุด - เค (K-Nearest Neighbor) 18
งานวิจัยที่เกี่ยวข้อง 21
บทสรุปของงานวิจัยที่เกี่ยวข้อง 26

สารบัญ (ต่อ)

บทที่ หน้า

3 วิธีดำเนินการวิจัย

ทำความเข้าใจกับปัญหา (Business Understanding) 30

ทำความเข้าใจข้อกับข้อมูล (Data Understanding) 31

การเตรียมข้อมูล (Data Preparation) 31

การสร้างแบบจำลองข้อมูล (Modeling) 33

การประเมินผล (Evaluation) 45

การนำไปใช้ (Deployment) 46

4 ผลการวิจัย

สัดส่วนข้อมูลของการ Train และการ Test ในการประเมินผลประสิทธิภาพตัวแบบ 47

ผลการทดสอบด้วยตัวแบบ Decision Tree 51

ผลการทดสอบด้วยตัวแบบ Naïve Bayes 64

ผลการทดสอบด้วยตัวแบบ K- Nearest Neighbor 65

ผลการทดสอบด้วยตัวแบบ Random Forest 66

5 สรุปผลการวิจัย

สรุปผลการทดลอง 69

ปัญหาและอุปสรรค 71

ข้อเสนอแนะ 71

บรรณานุกรม 73

ภาคผนวก 75

ภาคผนวก ก ขั้นตอนการแสดงผลลัพธ์การทดลองอย่างละเอียด 76

ประวัติผู้จัดทำ 118

สารบัญรูปภาพ

ภาพที่ หน้า

2.1 ขั้นตอนการดำเนินงานตามแบบ CRISP-DM 8
2.2 ตัวอย่างการสร้างต้นไม้ตัดสินใจ 16
2.3 ตัวอย่างการสร้าง K- Nearest Neighbor 18
3.1 ตัวอย่างข้อมูลลูกค้าของโชว์รูม อีซูซุมหานคร สาขา มหาชัย 31
3.2 การกำหนดชนิดข้อมูลที่ต้องการทำนาย 33
3.3 การตั้งค่า Parameter ภายใน Operator Set Role 34
3.4 การแบ่งข้อมูลเป็น Training Data และ Test Data 35
3.5 สร้างตัวแบบ Decision Tree โดยใช้โอเปอร์เรเตอร์ Decision Tree 35
3.6 ขั้นตอนการปรับพารามิเตอร์ Criterion เป็น gini index 35
3.7 การตั้งค่าพารามิเตอร์ Confidence เป็น 0.1 36
3.8 การตั้งค่าพารามิเตอร์ Maximal Depth ความลึกของต้นไม้ตัดสินใจ 36
3.9 โอเปอร์เรเตอร์ Apply Model 36
3.10 โอเปอร์เรเตอร์ Confusion Matrix ที่มีชื่อว่า Performance (Classification) 37
3.11 ตัวแบบต้นไม้ตัดสินใจสำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ 37
3.12 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 38
3.13 ค่า Recall Decision Tree โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 38
3.14 ค่า Precision Decision Tree โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 39
3.15 การปรับพารามิเตอร์ของ Cross validation 39
3.16 Process Cross validation ของตัวแบบ Decision tree 40
3.17 ตัวแบบต้นไม้ตัดสินใจโดยใช้ Cross Validation ในการ Training ข้อมูล 40

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

3.18 สร้างตัวแบบ Naïve Bayes โดยใช้ โอเปอร์เรเตอร์ Naïve Bayes....40

3.19 ตัวแบบ Naïve Bayes สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ 41

3.20 Process Cross validation ของตัวแบบ Naïve Bayes 41

3.21 ตัวแบบ Naïve Bayes โดยใช้ Cross Validation ในการ Training ข้อมูล.... 41

3.22 สร้างตัวแบบ K- Nearest Neighbor โดยใช้ โอเปอร์เรเตอร์ K- Nearest Neighbor 42

3.23 ตัวแบบ K- Nearest Neighbor สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ 42

3.24 Process Cross validation ของตัวแบบ K- Nearest Neighbor.... 42

3.25 ตัวแบบ K- Nearest Neighbor โดยใช้ Cross Validation ในการ Training ข้อมูล.... 43

3.26 ตัวแบบการทำนายโดยใช้เทคนิคของ Random Forest.... 43

3.27 ขั้นตอนการปรับพารามิเตอร์ของตัวแบบ Random Forest.... 43

3.28 สร้างตัวแบบ Random Forest โดยใช้ โอเปอร์เรเตอร์ Random Forest 44

3.29 Process Cross validation ของตัวแบบ Random forest.... 44

3.30 ตัวแบบ Random forest โดยใช้ Cross Validation ในการ Training ข้อมูล 45

4.1 กราฟผลลัพธ์ของตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 5 52

4.2 กราฟผลลัพธ์ของตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 10 53

4.3 สรุปกฎของตัวแบบ Decision Tree.... 53

4.4 กราฟผลลัพธ์ของตัวแบบ Naïve Bayes 64

4.5 กราฟผลลัพธ์ของตัวแบบ K- Nearest Neighbor.... 65

4.6 กราฟผลลัพธ์ของตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 5 66

4.7 กราฟผลลัพธ์ของตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 10 67

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

5.1 กราฟผลการทดสอบตัวแบบที่มีประสิทธิภาพมากที่สุด 70

ก.1 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth 77

ก.2 ค่า Recall Decision Tree โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth 77

ก.3 ค่า Precision Decision Tree โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth..... 78

ก.4 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth 78

ก.5 ค่า Recall Decision Tree โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth 79

ก.6 ค่า Precision Decision Tree โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth..... 79

ก.7 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth 80

ก.8 ค่า Recall Decision Tree โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth 80

ก.9 ค่า Precision Decision Tree โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth..... 81

ก.10 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth 81

ก.11 ค่า Recall Decision Tree โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth 82

ก.12 ค่า Precision Decision Tree โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth..... 82

ก.13 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth 83

ก.14 ค่า Recall Decision Tree โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth 83

ก.15 ค่า Precision Decision Tree โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth..... 84

ก.16 ค่า Accuracy Decision tree โดยใช้ number of folds 5 และค่า Maximal Depth 5 84

ก.17 ค่า Recall โดยใช้ number of folds 5 และค่า Maximal Depth 5..... 85

ก.18 ค่า Precision โดยใช้ number of folds 5 และค่า Maximal Depth 5 85

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

ก.19 ค่า Accuracy Decision tree โดยใช้ number of folds 5 และค่า Maximal Depth 10 86
ก.20 ค่า Recall โดยใช้ number of folds 5 และค่า Maximal Depth 10..... 86
ก.21 ค่า Precision โดยใช้ number of folds 5 และค่า Maximal Depth 10 87
ก.22 ค่า Accuracy Decision tree โดยใช้ number of folds 10 และค่า Maximal Depth 5 87
ก.23 ค่า Recall โดยใช้ number of folds 10 และค่า Maximal Depth 5..... 88
ก.24 ค่า Precision โดยใช้ number of folds 10 และค่า Maximal Depth 5 88
ก.25 ค่า Accuracy Decision tree โดยใช้ number of folds 10 และค่า Maximal Depth 10 89
ก.26 ค่า Recall โดยใช้ number of folds 10 และค่า Maximal Depth 10..... 89
ก.27 ค่า Precision โดยใช้ number of folds 10 และค่า Maximal Depth 10 90
ก.28 ค่า Accuracy Naïve Bayes โดยแบ่งข้อมูล 70:30..... 90
ก.29 ค่า Recall Naïve Bayes โดยแบ่งข้อมูล 70:30..... 91
ก.30 ค่า Precision Naïve Bayes โดยแบ่งข้อมูล 70:30 91
ก.31 ค่า ค่า Accuracy Naïve Bayes โดยแบ่งข้อมูล 75:25 92
ก.32 ค่า Recall Naïve Bayes โดยแบ่งข้อมูล 75:25..... 92
ก.33 ค่า Precision Naïve Bayes โดยแบ่งข้อมูล 75:25 93
ก.34 ค่า ค่า Accuracy Naïve Bayes โดยแบ่งข้อมูล 80:20 93
ก.35 ค่า Recall Naïve Bayes โดยแบ่งข้อมูล 80:20..... 94
ก.36 ค่า Precision Naïve Bayes โดยแบ่งข้อมูล 80:20 94
ก.37 ค่า Accuracy ของตัวแบบ Naïve Bayes พารามิเตอร์ number of folds ที่ 5..... 95
ก.38 ค่า Recall ของตัวแบบ Naïve Bayes พารามิเตอร์ number of folds ที่ 5 95
ก.39 ค่า Precision ของตัวแบบ Naïve Bayes พารามิเตอร์ number of folds ที่ 5 96
ก.40 ค่า Accuracy ของตัวแบบ Naïve Bayes พารามิเตอร์ number of folds ที่ 10..... 96
ก.41 ค่า Recall ของตัวแบบ Naïve Bayes พารามิเตอร์ number of folds ที่ 10 96

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

ก.42 ค่า Precision ของตัวแบบ Naïve Bayes พารามิเตอร์ number of folds ที่ 10 97
ก.43 ค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30 97
ก.44 ค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30 98
ก.45 ค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30..... 98
ก.46 ค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25 99
ก.47 ค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25 99
ก.48 ค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25..... 99
ก.49 ค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20 100
ก.50 ค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20 100
ก.51 ค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20..... 101
ก.52 ค่า Accuracy K- Nearest Neighbor พารามิเตอร์ number of folds ที่ 5..... 101
ก.53 ค่า Recall K- Nearest Neighbor พารามิเตอร์ number of folds ที่ 5..... 102
ก.54 ค่า Precision K- Nearest Neighbor พารามิเตอร์ number of folds ที่ 5 102
ก.55 ค่า Accuracy K- Nearest Neighbor พารามิเตอร์ number of folds ที่ 10..... 102
ก.56 ค่า Recall K- Nearest Neighbor พารามิเตอร์ number of folds ที่ 10..... 103
ก.57 ค่า Precision K- Nearest Neighbor พารามิเตอร์ number of folds ที่10 103
ก.58 ค่า Accuracy Random Forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 104
ก.59 ค่า Recall Random Forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 104
ก.60 ค่า Precision Random Forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 5..... 105
ก.61 ค่า Accuracy Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth 105
ก.62 ค่า Recall Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth 106
ก.63 ค่า Precision Random forest โดยปรับพารามิเตอร์ Maximal depth ที่ 10 106
ก.64 ค่า Accuracy Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth 107
ก.65 ค่า Recall Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth 107

สารบัญรูปภาพ (ต่อ)

ภาพที่ หน้า

ก.66 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth 107
ก.67 ค่า Accuracy Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth 108
ก.68 ค่า Recall Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth 108
ก.69 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth 109
ก.70 ค่า Accuracy Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth 109
ก.71 ค่า Recall Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth 110
ก.72 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth 110
ก.73 ค่า Accuracy Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth 111
ก.74 ค่า Recall Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth 111
ก.75 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth 112
ก.76 ค่า Accuracy Random forest โดยใช้ number of folds 5 Maximal Depth 5 112
ก.77 ค่า Recall Random forest โดยใช้ number of folds 5 Maximal Depth 5 113
ก.78 ค่า Precision Random forest โดยใช้ number of folds 5 Maximal Depth 5 113
ก.79 ค่า Accuracy Random forest โดยใช้ number of folds 5 Maximal Depth 10 114
ก.80 ค่า Recall Random forest โดยใช้ number of folds 5 Maximal Depth 10 114
ก.81 ค่า Precision Random forest โดยใช้ number of folds 5 Maximal Depth 10 115
ก.82 ค่า Accuracy Random forest โดยใช้ number of folds 10 Maximal Depth 5 116
ก.83 ค่า Recall Random forest โดยใช้ number of folds 10 Maximal Depth 5 116
ก.84 ค่า Precision Random forest โดยใช้ number of folds 10 Maximal Depth 5 117
ก.85 ค่า Accuracy Random forest โดยใช้ number of folds 10 Maximal Depth 10 117
ก.86 ค่า Recall Random forest โดยใช้ number of folds 10 Maximal Depth 10 118
ก.87 ค่า Precision Random forest โดยใช้ number of folds 5 Maximal Depth 10 118

สารบัญตาราง

ตารางที่ หน้า

2.1 สรุปงานวิจัยที่เกี่ยวข้อง 26

3.1 การกำหนดตัวแปร 32

3.2 กำหนดประเภทของข้อมูลและกำหนดบทบาทของ Attribute ดังนี้ 34

4.1 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามเพศ 47

4.2 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามอาชีพ 48

4.3 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานเงินเดือน 48

4.4 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานเงินจอง 48

4.5 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาของลูกค้า 49

4.6 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาเงินดาวน์ 49

4.7 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาประวัติเครดิต 49

4.8 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาระบบเครื่องยนต์ 50

4.9 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาระบบเกียร์ 50

4.10 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มารุ่นรถยนต์ 50

4.11 ผลการทดสอบด้วยตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 5 51

4.12 ผลการทดสอบด้วยตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 10 52

4.13 ผลการทดสอบด้วยตัวแบบ Naïve Bayes 64

4.14 ผลการทดสอบด้วยตัวแบบ K- Nearest Neighbor 65

4.15 ผลการทดสอบด้วยตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 5 66

4.16 ผลการทดสอบด้วยตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 10 ……...67

5.1 ตารางเปรียบเทียบผลการทดสอบตัวแบบที่มีประสิทธิภาพมากที่สุดตามลำดับ ……..70

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

อุตสาหกรรมรถยนต์เป็นอุตสาหกรรมที่มีความสำคัญต่อการพัฒนาประเทศในด้านหลายๆ ด้าน เช่น เศรษฐกิจ, การจ้างงาน, การสร้างมูลค่าเพิ่ม, และการพัฒนาด้านเทคโนโลยีรถยนต์ การบริหารประเทศไทยมีนโยบายในการส่งเสริมอุตสาหกรรมรถยนต์เป็นต้นต่อเนื่องตั้งแต่ พ.ศ.2504 โดยนโยบายเหล่านี้เน้นที่การผลิตชิ้นส่วนรถยนต์ในประเทศเพื่อลดการนำเข้า อุตสาหกรรมรถยนต์ในประเทศไทยมีการเติบโตขึ้นเรื่อยๆ จนถึงวิกฤติเศรษฐกิจใน พ.ศ.2540 ที่มีผลกระทบต่ออุตสาหกรรมเช่นเดียวกับอุตสาหกรรมอื่น ๆ แต่อุตสาหกรรมรถยนต์สามารถปรับตัวได้อย่างรวดเร็ว โดยเน้นที่การผลิตเพื่อส่งออก การเติบโตของอุตสาหกรรมรถยนต์ของไทยเป็นอุตสาหกรรมที่มีความสำคัญต่อประเทศ และมียอดขายรถยนต์แบบกระบะและ Pure Pick Up ที่มียอดขายสูง โดยยอดขายรถยนต์ยี่ห้อ Isuzu มียอดขายสูงสุดทั้งในกลุ่มตลาดรถกระบะขนาด 1 ตัน และตลาดรถกระบะ Pure Pick Up โดยยอดขายรถยนต์ Toyota และ Ford ก็มีจำนวนสูง แต่ยี่ห้อ Isuzu ยังคงครองอันดับหนึ่งในทั้งสองกลุ่มตลาดรถกระบะด้วยยอดขายสูงสุดขณะนั้น รวมถึงยอดขายรถกระบะรวมทั้งสิ้นก็สูงมาก นำเสนอให้เห็นว่าอุตสาหกรรมรถยนต์ยังคงเป็นส่วนสำคัญของเศรษฐกิจไทยอย่างต่อเนื่องและมียอดขายสูงในรถกระบะที่นิยมมากที่สุดในประเทศ คือรถยี่ห้อ Isuzu ในกลุ่มรถกระบะขนาด 1 ตันและ Pure Pick Up จากความเร็จสึกข้างต้นจะเห็นได้ว่าอุตสาหกรรมรถยนต์เป็นส่วนสำคัญของเศรษฐกิจของประเทศไทยและมีความสำคัญในการส่งออกสินค้าด้วยความเร็จนี้จึงมีการส่งเสริมการผลิตรถยนต์ในประเทศอย่างต่อเนื่องเพื่อเสริมสร้างเศรษฐกิจและความเจริญของอุตสาหกรรมนี้ในอนาคต (สุรพงษ์ ไพสิฐพัฒนพงษ์ , 2565)

การทำเหมืองข้อมูล (Data Mining) เป็นเทคโนโลยีที่ช่วยในการค้นหารูปแบบที่มีประโยชน์จากข้อมูลขนาดใหญ่ เช่น ฐานข้อมูล, คลังข้อมูล, เว็บไซต์ และแหล่งข้อมูลอื่น ๆ การประยุกต์การทำเหมืองข้อมูลมีความสำคัญในการวิเคราะห์ข้อมูลของลูกค้าเพื่อนำมาวางแผนการตลาด ซึ่งช่วยให้ธุรกิจมีความรวดเร็วและสะดวกสบายในการวิเคราะห์ข้อมูลอีกด้วย การจัดประเภท (Classification)

เป็นกระบวนการสร้างตัวแบบจากข้อมูลที่มีจำนวนมากเพื่อสร้างความเข้าใจเกี่ยวกับข้อมูลที่น่าสนใจ มีหลายอัลกอริทึมที่ใช้ในการทำ Classification เช่น Decision Tree, Naïve Bayes, Random Forest, K-Nearest Neighbor ตัวแบบที่ได้จะสามารถใช้ในการทำนายข้อมูลจากรูปแบบที่ได้เรียนรู้จากชุดข้อมูล ตัวอย่างการประยุกต์ใช้งานทางธุรกิจได้แก่การวิเคราะห์ข้อมูลลูกค้า เพื่อสร้างตัวแบบที่สามารถทำนายพฤติกรรมของลูกค้าในอนาคต แต่ลูกค้านี้อาจมีนิสัยในการซื้อสินค้าหรือบริการต่าง ๆ ในอนาคต หรือการจัดลำดับลูกค้าที่ควรได้รับความสนใจเพิ่มเติมจากบริษัท การทำเหมืองข้อมูลเป็นวิธีการที่มีประสิทธิภาพในการจัดการกับข้อมูลขนาดใหญ่และสร้างความเข้าใจในแนวโน้มและรูปแบบที่อาจไม่เป็นที่รู้จักไว้แล้ว ทั้งนี้เพื่อให้ธุรกิจสามารถดำเนินการในทิศทางที่มีประสิทธิภาพมากขึ้นในธุรกิจสมัยใหม่ที่ความจำเป็นต้องใช้ข้อมูลในการตัดสินใจเพื่อขับเคลื่อนธุรกิจไปสู่ความสำเร็จ

งานวิจัยนี้ผู้วิจัยได้นำเทคนิคเหมืองข้อมูลมาใช้ในการแก้ปัญหา เช่น การใช้เทคนิคเหมืองข้อมูลในการจำแนกรุ่นของรถยนต์ประเภทรถกระบะที่ลูกค้าต้องการ การวิเคราะห์ปัจจัยที่มีผลต่อการเลือกรุ่นของรถยนต์ที่ลูกค้าต้องการ และ การเปรียบเทียบประสิทธิภาพของตัวแบบในการจำแนกรุ่นของรถยนต์ประเภทรถกระบะที่ลูกค้าต้องการ จากปัญหากระบวนการเดิมที่เคยใช้มาการจะทราบถึงรุ่นของรถยนต์ที่ได้รับความนิยมสูงสุดในแต่ละเดือนผู้จัดการจะนำยอดขายของเดือนที่ผ่านมาทำการสรุปผลยอดขายรถยนต์ในแต่ละประเภทรถกระบะไหนที่ลูกค้าทำการซื้อไปเยอะที่สุดก็จะนำรุ่นรถยนต์ที่ขายดีที่สุดมาให้ที่ปรึกษาการขายทำการตลาดส่งเสริมการขายของประเภทรถกระบะที่ได้รับความนิยมสูงที่สุดในแต่ละเดือน ทั้งนี้ผู้วิจัยได้เห็นถึงประโยชน์ของการนำเทคนิคเหมืองข้อมูลที่สามารถนำมาจำแนกประเภทรถกระบะเพื่อให้ที่ปรึกษาการขายสามารถเตรียมประเภทรถกระบะไว้สำหรับความต้องการของลูกค้าได้ทันเวลาที่ลูกค้าต้องการ อีกทั้งได้ตัวแบบที่มีประสิทธิภาพเพื่อนำไปใช้ในการประกอบการตัดสินใจในการดำเนินธุรกิจรถยนต์ต่อไปได้

วัตถุประสงค์ของการวิจัย

1.เพื่อสร้างตัวแบบที่สำหรับการจำแนกรุ่นของรถยนต์ประเภทรถกระบะ กรณีศึกษาโชว์รูมรถยนต์ อีซูซุมหานคร สาขามหาชัย

2.เพื่อเปรียบเทียบประสิทธิภาพที่สุดในการจำแนกรุ่นรถรถยนต์ประเภทรถกระบะโดยใช้อัลกอริทึมดังต่อไปนี้ ได้แก่ อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) อัลกอริทึมเรนดอม ฟลอเรส (Random forest) อัลกอริทึมเพื่อนบ้านใกล้ที่สุด - เค (K-Nearest Neighbor)

ขอบเขตการวิจัย

1. ขอบเขตด้านข้อมูล - ข้อมูลที่นำมาสร้างตัวแบบเป็นข้อมูลลูกค้าที่ซื้อรถยนต์อีซูซุ สาขามหาชัย เดือนมกราคม พ.ศ. 2563 ถึง เดือนธันวาคม พ.ศ. 2564 จำนวน 1,500 คน

2. ขอบเขตด้านการสร้างตัวแบบ ผู้วิจัยเลือกเทคนิคในการสร้างตัวแบบโดยวิธีการดังนี้ ได้แก่ อัลกอริทึม Decision Tree อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) อัลกอริทึมเรนดอม ฟลอเรส (Random forest) อัลกอริทึมเพื่อนบ้านใกล้ที่สุด - เค (K-Nearest Neighbor)

ขั้นตอนการดำเนินงานวิจัย

1. ทำความเข้าใจกับปัญหาและวัตุประสงค์ของการวิจัยจากธุรกิจ โดยแปลงปัญหาให้อยู่ในรูปแบบของโจทย์สำหรับการวิเคราะห์และวางแผนการดำเนินงานเบื้องต้น

2. ทำความเข้าใจข้อมูลโดยเริ่มจากการรวบรวมข้อมูลของลูกค้าจำนวน 1,500 คน จากผู้จัดการโชว์รูมอีซูซุมหานคร สาขามหาชัย โดยเป็นข้อมูลตั้งแต่ เดือนมกราคม พ.ศ. 2563 ถึง เดือนธันวาคม พ.ศ. 2564 จากนั้นตรวจสอบคุณภาพของข้อมูล และทำการเลือกข้อมูลว่าจะใช้ข้อมูลตัวใดบ้างในการวิเคราะห์

3. เตรียมข้อมูลเพื่อทำให้ข้อมูลที่รวบรวมมาเป็นข้อมูลที่สมบูรณ์พร้อมสู่ขั้นตอนที่ 4 เช่นการลบข้อมูลของลูกค้าที่ไม่ต้องการ และแปลงข้อมูลของลูกค้าให้อยู่ในรูปแบบที่ต้องการ

4. การสร้างตัวแบบ โดยทำการสร้างตัวแบบมาทั้งหมด 4 ตัวแบบ ได้แก่ ตัวแบบ Decision Tree ตัวแบบนาอีฟเบย์ (Naïve Bayes) ตัวแบบเรนดอม ฟลอเรส (Random forest) และตัวแบบเพื่อนบ้าน

ใกล้ที่สุด - เค (K-Nearest Neighbor) เพื่อทดลองตัวแบบที่จะสามารถแก้ไขปัญหาที่ต้องการ เพื่อหาตัวแบบที่เหมาะสมที่สุดในการแก้ปัญหา

5. การวัดประสิทธิภาพของตัวแบบ โดยทำการวัดประสิทธิภาพของตัวแบบที่ได้จากขั้นตอนการสร้างตัวแบบ เพื่อวัดประสิทธิภาพของตัวแบบทั้ง 4 ตัวแบบต่อการนำไปใช้งาน ซึ่งตัวแบบแต่ละประเภทจะมีประสิทธิภาพที่แตกต่างกันออกไป

6. การนำตัวแบบไปใช้งานจริง จะเป็นการนำตัวแบบที่มีความเหมาะสมที่สุดไปใช้ในการจำแนกประเภทรถกระบะที่ได้รับความนิยม และนำไปใช้ในการประกอบการตัดสินใจในการดำเนินธุรกิจรถยนต์

ประโยชน์ที่คาดว่าจะได้รับ

1.ทราบประสิทธิภาพของตัวแบบแต่ละตัวแบบที่ใช้ในการจำแนกประเภทรถกระบะที่ได้รับความนิยมด้วยเทคนิคเหมืองข้อมูล

2.ได้ตัวแบบสำหรับจำแนกรุ่นของรถยนต์โดยตัวแบบสามารถนำไปใช้ในการจำแนกประเภทรถกระบะยี่ห้ออื่นๆได้

บทที่2
งานวิจัยและทฤษฎีที่เกี่ยวข้อง

การวิจัยการใช้เทคนิคเหมืองข้อมูลเพื่อการจำแนกประเภทรถกระบะที่ได้รับความนิยม ผู้วิจัยได้ศึกษาการทำเหมืองข้อมูลด้วย อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) และอัลกอริทึมเรนดอม ฟลอเรส (Random forest) อัลกอริทึม เพื่อนบ้านใกล้ที่สุด - เค (K-Nearest Neighbor) และงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดดังนี้

2.1 ทฤษฎีการทำเหมืองข้อมูล (Data Mining)

2.2 ทฤษฎีต้นไม้ตัดสินใจ Decision Tree

2.3 ทฤษฎีนาอีฟเบย์ (Naïve Bayes)

2.4 ทฤษฎีเรนดอม ฟลอเรส (Random forest)

2.5 ทฤษฎีเพื่อนบ้านใกล้ที่สุด - เค (K-Nearest Neighbor)

2.6 งานวิจัยที่เกี่ยวข้อง

ทฤษฎีการทำเหมืองข้อมูล

การทำเหมืองข้อมูลคือกระบวนการการสืบค้นความรู้ที่มีประโยชน์และน่าสนใจจากฐานข้อมูลขนาดใหญ่ หรือที่เรียกว่า Knowledge Discovery from Very Large Database (KDD) ซึ่งเป็นเทคนิคที่ใช้ในการจัดการข้อมูลขนาดใหญ่ โดยการวิเคราะห์ข้อมูลที่มีอยู่และการดึงข้อมูลที่สำคัญออกมาเพื่อใช้ในการทำนายหรือวิเคราะห์เหตุการณ์ต่าง ๆ ที่อาจจะเกิดขึ้นในอนาคต กระบวนการนี้ช่วยในการค้นหาความรู้และความจริงที่อาจซ่อนอยู่ในข้อมูล โดยการขุดค้นสิ่งที่น่าสนใจจากข้อมูลที่มีปริมาณมาก ซึ่งความแตกต่างหลักคือการทำเหมืองข้อมูลไม่ต้องกำหนดคำสั่งการค้นหาข้อมูลเหมือนในระบบฐานข้อมูลทั่วไป การทำเหมืองข้อมูลมีความสำคัญในการช่วยให้ทราบถึงแนวโน้มและความสัมพันธ์ในข้อมูล ทำให้สามารถทำนายเหตุการณ์ต่าง ๆ ได้ ซึ่งมีประโยชน์ในหลายด้าน เช่น การวิเคราะห์ธุรกิจ, การทำนายทรัพยากร, หรือการค้นหาข้อมูลทางการแพทย์ เพื่อช่วยลดความเสี่ยงหรือเพิ่มประสิทธิภาพในการตัดสินใจ

วิวัฒนาการของเหมืองข้อมูล

การทำเหมืองข้อมูลไม่ได้มีเป้าหมายที่จะค้นพบเพียงองค์ความรู้ที่เคยรู้มาแล้ว แต่ยังเน้นการค้นพบองค์ความรู้ใหม่ที่ไม่เคยรู้จักมาก่อน วิวัฒนาการของการทำเหมืองข้อมูลได้เกิดขึ้นตามนี้

1. Data Collection (ปี ค.ศ. 1960) การนำข้อมูลมาจัดเก็บในรูปแบบที่เหมาะสมและปลอดภัยมีความสำคัญ เพื่อป้องกันการสูญหายข้อมูลได้เป็นอย่างดี
2. Data Access (ปี ค.ศ. 1980) การสร้างความสัมพันธ์ระหว่างข้อมูลที่ถูกจัดเก็บมีความสำคัญ โดยเน้นการนำข้อมูลไปใช้ประโยชน์ในการวิเคราะห์และการตัดสินใจที่มีคุณภาพ
3. Data Warehouse & Decision Support (ปี ค.ศ. 1990) การสร้างระบบ Data Warehouse เพื่อรวบรวมข้อมูลในองค์กรในฐานข้อมูลขนาดใหญ่ เพื่อสนับสนุนการตัดสินใจในทุกด้าน
4. Data Mining (ปี ค.ศ. 2000) การนำข้อมูลมาวิเคราะห์และประมวลผลโดยใช้เทคนิค Data Mining เพื่อสร้างแบบจำลองและค้นพบความสัมพันธ์ทางสถิติที่ไม่เคยรู้มาก่อน

วิวัฒนาการของการทำเหมืองข้อมูลมีการพัฒนาที่ต่อเนื่องเพื่อค้นพบความรู้ที่ไม่เคยรู้มาก่อนและให้ข้อมูลมีประโยชน์ในการตัดสินใจและวิเคราะห์ทางธุรกิจ

ประเภทของการทำเหมืองข้อมูล

1.การจำแนกประเภทข้อมูล (Classification) คือกระบวนการทำเหมืองข้อมูลที่ใช้สร้างตัวแบบเพื่อจำแนกข้อมูลจากชุดข้อมูลการเรียนรู้ โดยการพิจารณาจุดเด่นที่เด่นชัดเจนและกำหนดค่าของจุดเด่นนั้นเพื่อใช้ในการแบ่งหมวดหมู่ข้อมูล การจำแนกประเภทข้อมูลนี้ทำการอธิบายและให้น้ำหนักกับจุดเด่นที่เป็นที่รู้จักดีในการแยกประเภทข้อมูล และใช้ชุดข้อมูลการเรียนรู้ที่สำรวจจุดเด่นเพื่อสร้างตัวแบบ ตัวอย่างของการจำแนกประเภทข้อมูลเป็นชนิดหนึ่งคือการใช้ตัวแบบที่ไม่สามารถจำแนกข้อมูลได้ให้สามารถแบ่งหมวดหมู่ได้

2. กฎความสัมพันธ์ (Association Rule) เป็นทฤษฎีที่ใช้ในการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์ระหว่างรายการในชุดข้อมูล โดยมีสองคำสำคัญคือ "สนับสนุน" (Support) และ "ความ

คงเหลือ" (Confidence) สนับสนุน (Support) เป็นค่าที่บ่งบอกถึงความถี่ของกฎความสัมพันธ์ในชุดข้อมูล ค่าสนับสนุนสูงแสดงว่ากฎมีความถี่มาก ความคงเหลือ (Confidence) เป็นค่าที่บ่งบอกถึงความน่าจะเป็นที่จะเกิดเหตุการณ์ที่ถูกทำนายโดยกฎความสัมพันธ์ ค่าความคงเหลือสูงแสดงว่ากฎมีความแม่นยำมาก กฎความสัมพันธ์มีรูปแบบที่ดีที่จะช่วยในการเข้าใจและวิเคราะห์ความสัมพันธ์ระหว่างรายการในชุดข้อมูล การปรับค่าสนับสนุนและความคงเหลือช่วยกำหนดกฎที่มีความสำคัญและน่าสนใจที่สุดสำหรับการวิเคราะห์

3. การจับกลุ่มข้อมูล (Clustering) เป็นกระบวนการทำเหมืองข้อมูลที่ทำให้ข้อมูลถูกแบ่งเป็นกลุ่มตามความคล้ายคลึงของลักษณะ ข้อมูลที่อยู่ในกลุ่มเดียวกันมีความคล้ายคลึงกันมาก ในขณะที่ข้อมูลที่อยู่ในกลุ่มต่างกันมีความแตกต่างออกไป กระบวนการนี้ใช้วัดความคล้ายและความแตกต่างเพื่อจัดกลุ่มข้อมูลให้อยู่ในลักษณะที่เหมาะสม การจับกลุ่มข้อมูลมีประโยชน์ในการวิเคราะห์ลูกค้า การจัดการข้อมูลทางการแพทย์ หรือการวิเคราะห์ภาพทางดาวเทียม เพื่อช่วยในการตัดสินใจและทำนายแนวโน้ม

4. การค้นหาข้อมูลที่มีความผิดปกติ (Anomaly Detection) เป็นกระบวนการทำเหมืองข้อมูลที่เน้นการค้นหาข้อมูลที่มีความแปลกต่างหรือผิดปกติ เหตุการณ์เหล่านี้มักจะมีจำนวนน้อยและต่างออกไปจากลักษณะทั่วไปของข้อมูล การตรวจจับข้อมูลที่มีความผิดปกติช่วยในการประเมินความเสี่ยงหรือการตรวจสอบข้อผิดปกติในระบบที่มีข้อมูลมากมาย

5. การคาดคะเน (Estimation) และการจำแนกประเภทข้อมูล (Classification) มีความคล้ายคลึงในการสร้างตัวแบบเพื่อทำนายผลลัพธ์ แต่มีความแตกต่างในลักษณะของตัวแปรที่เป็นเป้าหมาย การคาดคะเน (Estimation) มุ่งหวังที่จะคาดคะเนค่าของตัวแปรต่อเนื่อง เช่น การคาดคะเนยอดขายสินค้า, การคาดคะเนราคาหุ้น การจำแนกประเภทข้อมูล (Classification) มุ่งหวังที่จะจำแนกข้อมูลเป็นกลุ่มตามลักษณะของข้อมูล เช่น การจำแนกอีเมลเป็นสแปมหรือไม่สแปม การจำแนกภาพถ่ายเป็นแมวหรือสุนัข

Cross-Industry Standard Process: CRISP-DM

ภาพที่ 2.1 ขั้นตอนการดำเนินงานตามแบบ CRISP-DM

ที่มา https://datacubeth.ai/crisp-dm/

CRISP-DM (Cross-Industry Standard Process for Data Mining) เป็นกระบวนการมาตรฐานที่ใช้ในการทำเหมืองข้อมูล (Data Mining) ซึ่งได้รับความนิยมในวงการและอุตสาหกรรมต่าง ๆ เน้นการดำเนินการขั้นตอนต่าง ๆ ในกระบวนการทำเหมืองข้อมูลอย่างมีระเบียบและเป็นระบบ กระบวนการ CRISP-DM ประกอบด้วยขั้นตอนหลัก 6 ขั้นตอน (อาทิตยาพร โรจรัตน์, 2566)

CRISP-DM ประกอบด้วย 6 ขั้นตอน

1.ขั้นตอนการทำความเข้าใจการวิจัยและธุรกิจ (Business/Research Understanding Phase)

ในขั้นตอนแรกของกระบวนการทำเหมืองข้อมูลมาตรฐาน CRISP-DM เราเริ่มจากการทำความเข้าใจเกี่ยวกับการวิจัยและธุรกิจที่เราจะทำ การสร้างความเข้าใจนี้จะเน้นที่ปัญหาและวัตถุประสงค์ของโครงการที่เกิดขึ้นจากธุรกิจ หลังจากนั้น เราจะทำการแปลงปัญหาที่ได้ให้อยู่ในรูปแบบของโจทย์ที่เหมาะสมสำหรับการวิเคราะห์ข้อมูล และวางแผนการดำเนินงานเบื้องต้น ในขณะนี้, หัวหน้าโครงการหรือที่ปรึกษาด้านการวางระบบวิเคราะห์ข้อมูล จะต้องไปสัมภาษณ์หรือรับฟังปัญหาและความต้องการจากผู้บริหารองค์กรและหน่วยงานที่เกี่ยวข้อง ความต้องการทั้งหมดจะถูก

จัดลำดับความสำคัญเพื่อกำหนดและกำหนดวัตถุประสงค์ของโครงการ การวิเคราะห์ข้อมูลนำไปสู่รูปแบบการวิเคราะห์ข้อมูลขององค์กรจะถูกกำหนดขึ้นต้นในขั้นตอนนี้

2.ขั้นตอนการทำความเข้าข้อมูล (Data Understanding Phase)

ทำความเข้าใจเพื่อการวิเคราะห์ข้อมูลเริ่มต้นด้วยการรวบรวมข้อมูลที่เกี่ยวข้อง โดยการสำรวจแหล่งข้อมูลที่เป็นที่เชื่อถือและเหมาะสม เพื่อเข้าใจลักษณะของข้อมูลที่จะนำมาใช้ การประเมินค่าของประโยชน์ที่จะได้รับจากการนำข้อมูลมาใช้เป็นสิ่งสำคัญ ดังนั้นการเลือกแหล่งข้อมูลที่มีความเหมาะสมและเชื่อถือได้เป็นสิ่งสำคัญ เช่น ข้อมูลการซื้อสินค้าของลูกค้าที่เกี่ยวข้อง หรือข้อมูลการลงทะเบียนและผลการศึกษาของนักเรียนในระบบการศึกษา การรวบรวมข้อมูลที่เกี่ยวข้องจากแหล่งข้อมูลเหล่านี้จะช่วยให้เราเข้าใจข้อมูลที่สำคัญและเตรียมพร้อมสำหรับขั้นตอนถัดไปในการวิเคราะห์ข้อมูลได้ดียิ่งขึ้น โดยเน้นไปที่ความสำคัญของการรวบรวมข้อมูลจากแหล่งที่เชื่อถือได้และเหมาะสมสำหรับการวิเคราะห์ที่เราต้องการทำในภายหลัง

3. ขั้นตอนการเตรียมข้อมูล (Data Preparation Phase)

ขั้นตอนการเตรียมข้อมูลคือขั้นตอนที่ทำให้ข้อมูลที่เรามีเป็นข้อมูลที่สมบูรณ์และพร้อมสำหรับการนำเข้าสู่ตัวแบบในขั้นตอนถัดไป รวมถึงการสร้างตาราง การลบข้อมูลที่ไม่จำเป็น, และการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม การแปลงข้อมูลที่ได้มาจากการรวบรวม (raw data) เป็นข้อมูลที่สามารถนำไปวิเคราะห์ได้ในขั้นถัดไป เป็นขั้นตอนที่ซับซ้อนและทำให้เราต้องใช้เวลามากที่สุดในกระบวนการ CRISP-DM ในขั้นตอนนี้ เราต้องการให้ข้อมูลเป็นอย่างมากเพื่อให้มีคุณภาพที่ดีในการนำไปใช้ในการวิเคราะห์ การทำความสะอาดข้อมูล (data cleaning) เป็นส่วนสำคัญที่อาจรวมถึงการแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกันการเติมข้อมูลที่ขาดหายไป หรือการปรับแต่งข้อมูลให้เป็นรูปแบบที่เหมาะสม ขั้นตอนนี้เป็นการเตรียมพร้อมข้อมูลเพื่อให้มีความเสถียรและเต็มที่สำหรับการวิเคราะห์ในขั้นตอนถัดไป

4. ขั้นตอนการสร้างตัวแบบ (Modeling Phase)

ขั้นตอนการสร้างตัวแบบนั้นมีหลายขั้นตอนที่เน้นการทดลองและปรับแต่งเพื่อหาตัวแบบที่เหมาะสมที่สุดสำหรับการแก้ไขปัญหาที่กำหนด การเลือกและทดลองสร้างตัวแบบหลายแบบจะเป็น

การวิเคราะห์ข้อมูลโดยใช้เทคนิคเหมืองข้อมูล เช่น การจำแนกประเภทข้อมูล การแบ่งกลุ่มข้อมูลหรือการหากฎความสัมพันธ์ ในขั้นตอนนี้ เราจะทดลองสร้างตัวแบบที่น่าสนใจและสามารถแก้ไขปัญหาได้จากนั้น เราจะปรับค่าพารามิเตอร์ในตัวแบบเหล่านั้นเพื่อให้ได้ตัวแบบที่มีประสิทธิภาพมากที่สุด การวิเคราะห์ข้อมูลในขั้นตอนนี้จะใช้เทคนิคเหมืองข้อมูล เพื่อนำเสนอคำตอบที่ดีที่สุด หลายเทคนิคเหมืองข้อมูลอาจถูกนำมาใช้ เช่น การจำแนกประเภทข้อมูล, การแบ่งกลุ่มข้อมูล, และการหากฎความสัมพันธ์ เพื่อปรับข้อมูลให้เหมาะสมกับแต่ละเทคนิค

5. ขั้นตอนการวัดประสิทธิภาพของตัวแบบ (Evaluation Phase)

ขั้นตอนการวัดประสิทธิภาพของตัวแบบเป็นขั้นตอนสำคัญที่ต้องทำเพื่อให้ทราบว่าตัวแบบที่ได้สร้างขึ้นมีประสิทธิภาพตรงตามวัตถุประสงค์หรือไม่ โดยขั้นตอนนี้จะใช้ตัววัดประสิทธิภาพที่เหมาะสมกับประเภทของตัวแบบ เช่น ความแม่นยำ (Accuracy) ค่าความสามารถในการทำนาย (Precision, Recall) หรือ F1 Score การวัดประสิทธิภาพจะช่วยให้ทราบถึงคุณภาพของผลลัพธ์ที่ตัวแบบสร้างขึ้นได้ ว่ามีประสิทธิภาพมากน้อยเพียงใด และว่าได้รับประสิทธิภาพตามที่คาดหวังหรือไม่ หากผลลัพธ์ไม่ตรงตามที่ต้องการ อาจจะต้องทำการปรับแก้และทดลองสร้างตัวแบบใหม่การวัดประสิทธิภาพเป็นขั้นตอนที่สำคัญในการปรับปรุงและพัฒนาตัวแบบเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด

6. ขั้นการนำตัวแบบไปใช้งานจริง (Deployment Phase)

ขั้นตอนการนำตัวแบบไปใช้งานจริงคือการนำตัวแบบที่เหมาะสมที่สุดไปใช้งานจริงเพื่อวิเคราะห์และแก้ปัญหาตามที่ต้องการ ขั้นนตอนนี้รวมถึงการนำผลลัพธ์ที่ได้ไปใช้ทางธุรกิจ เช่น การสร้างรายงานหรือแผนภาพที่สามารถนำไปใช้ในการวางแผน กำหนดกลยุทธ์ และดำเนินการต่างๆ การนำตัวแบบไปใช้งานจริงอาจประกอบด้วยการเตรียมข้อมูล การทดสอบการทำงานในสภาพแวดล้อมจริง และการให้คำแนะนำหรือแนวทางการดำเนินการ ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของกระบวนการ CRISP-DM และเป็นขั้นตอนที่สำคัญในการให้ผลลัพธ์แก่องค์กรหรือธุรกิจได้

ทฤษฎีต้นไม้ตัดสินใจ (Decision Tree)

การสร้างต้นไม้ตัดสินใจเป็นวิธีการที่นิยมในการวิเคราะห์และจำแนกข้อมูล โครงสร้างของต้นไม้ตัดสินใจประกอบด้วยโหนดที่ใช้ในการตัดสินใจและโหนดใบ แต่ละโหนดตัดสินใจสร้างชุดคำสั่งที่ใช้สำหรับทดสอบเงื่อนไข ข้อมูลที่ป้อนเข้ามาถูกทดสอบตามเงื่อนไขเรียงตามลำดับจนกว่าจะถึงโหนดใบเพื่อให้ได้คำตอบที่สุดท้าย

1. การสร้างต้นไม้เริ่มต้นโดยมีโหนดเดียวแสดงถึงชุดข้อมูลฝึก (Training Set)
2. ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกัน ให้โหนดนั้นเป็นใบและตั้งชื่อตามกลุ่มของข้อมูล
3. ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ ทำการวัดค่า Gain ของแต่ละ Attribute เพื่อใช้เป็นเกณฑ์ในการคัดเลือก ที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆได้ดีที่สุด
4. กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่างๆที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่างๆที่สร้างขึ้น
5. ทำการวนซ้ำเพื่อหา Attribute ที่มีค่า Gain มากที่สุดสำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำ Attribute มาสร้างเป็นโหนดตัดสินใจต่อไป
6. ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งเป็นจริง (Han, Kamber, 2001)

หลักการทำงานของต้นไม้ตัดสินใจมีขั้นตอนดังนี้

1. การเลือกคุณลักษณะ (Feature Selection): การเริ่มต้นด้วยการเลือกคุณลักษณะที่เหมาะสมที่สุดสำหรับการแบ่งแยกข้อมูลในขั้นตอนถัดไป คุณลักษณะที่ถูกเลือกจะกำหนดโหนดตัดสินใจแรกของต้นไม้

2. การแบ่งแยก (Splitting): ใช้คุณลักษณะที่เลือกมาเพื่อแบ่งข้อมูลในโหนดปัจจุบันเป็นกลุ่มย่อยๆ โดยใช้กฎหรือเงื่อนไข ข้อมูลจะถูกแบ่งออกไปในโหนดย่อยที่เหมาะสมตามกฎที่กำหนด โหนดย่อยจะถูกสร้างขึ้นและเป็นข้อมูลประเภทใบ หรือจะเริ่มกระบวนการแบ่งแยกอีกครั้งหากมีการแบ่งย่อยต่อไป

3. การสร้างต้นไม้แบบเรียง (Recursion) กระบวนการแบ่งแยกและสร้างโหนดย่อยจะทำซ้ำในข้อมูลย่อยของแต่ละโหนดไปเรื่อยๆ จนกว่าจะพบเงื่อนไขที่ใช้ในการสร้างโหนดใบ โหนดใบคือโหนดสุดท้ายในต้นไม้แทนคำตอบหรือค่าที่ต้องการทำนาย

4. การสร้างโหนดใบ (Leaf Node) เมื่อต้นไม้ตัดสินใจถึงโหนดที่ไม่สามารถแบ่งแยกข้อมูลเพิ่มเติมได้ (ย่อยออกแล้วไม่เหลือข้อมูลที่แตกต่าง) โหนดนั้นจะกลายเป็นโหนดใบและจะถูกตั้งชื่อตามคำตอบหรือค่าที่ควรทำนายในส่วนนี้

5. การจบการสร้าง (Termination) กระบวนการสร้างต้นไม้จะสิ้นสุดเมื่อมีเงื่อนไขหยุดการแบ่งแยกที่กำหนดไว้ เช่นจำนวนโหนดใบสูงสุด ความลึกของต้นไม้ หรือเงื่อนไขอื่น ๆ

(Hastie, Tibshirani และ Friedman, 2009)

การคำนวณหาค่า Information Gain ของต้นไม้ตัดสินใจ

Information Gain (IG) เป็นหลักการในการสร้างต้นไม้ตัดสินใจ (Decision Tree) ในงาน Machine Learning ซึ่งช่วยในกระบวนการตัดสินใจว่า Attribute ใดควรถูกเลือกเพื่อแบ่งกลุ่มข้อมูลอย่างมีประสิทธิภาพในแต่ละโหนดของต้นไม้ เพื่อให้ได้กลุ่มที่มีความแม่นยำสูง Information Gain ใช้ค่าของ Entropy เพื่อวัดความสับสนในข้อมูลก่อนและหลังการแบ่ง Attribute ที่กำลังพิจารณา ค่า Entropy มากจะหมายถึงความสับสนมาก ค่า Entropy น้อยจะหมายถึงความสับสนน้อยลง

การหาค่า Information Gain สามารถคำนวณได้จากสมการดังนี้

$$I(s_1, s_2, \dots, s_n) = -\sum_{i=1}^{n} \frac{s_1}{s} \log_2 \frac{s_1}{s} \qquad [1]$$

s เป็นเซตข้อมูลซึ่งประกอบด้วยข้อมูล s เรคคอร์ด

n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

c_i แทนกลุ่มในลำดับ ที่ i โดยที่ i มีค่าระหว่าง 1 ถึง n

s_i แทนจำนวนข้อมูลที่เป็นสามาชิกของ s และอยู่ในกลุ่ม c_i

S_{ij} แทนจำนวนข้อมูลที่เป็นสามาชิกของ s ในกลุ่ม c จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของคุณลักษณะ A

J ค่าระหว่าง 1 ถึง v

ค่า Entropy ของคุณลักษณะ A ซึ่งมีค่าคุณลักษณะเป็น ($a1, a2, a3, ..., a_v$)

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\cdots+s_{nj}}{s} I(s_{1j}, s_{2j}, \ldots, s_{nj}) \quad [2]$$

ดังนั้นจะสามารถพิจารณาค่ามาตรฐานเกนได้ตามสมการที่ 3 ดังนี้

$$Gain(A) = I(s_{1j}, s_{2j}, \ldots, s_{nj}) - E(A) \quad [3]$$

นาอีฟเบย์ (Naïve Bayes)

ในกระบวนการจำแนกประเภทโดยใช้หลักสถิติเพื่อพยากรณ์ความน่าจะเป็น ทฤษฎีของเบย์ถูกนำมาปรับใช้ในการเรียนรู้โดยการปรับเปลี่ยนแจกแจง เปลี่ยนผลลัพธ์ของการเรียนรู้ด้วยตัวอย่างใหม่ นี้ทำให้ตัวแบบเรียนรู้ได้ตัวอย่างที่มีการเพิ่มหรือลดความน่าจะเป็น โดยมีการปรับตัวแบบไปตามตัวอย่างใหม่ และผนวกความรู้เดิม รวมทั้งทำนายค่าคลาสเป้าหมายของตัวอย่างโดยใช้ความน่าจะเป็นมากที่สุดจากทุกสมมติฐาน ตามทฤษฎีของเบย์ สามารถคำนวณความน่าจะเป็นของสมมติฐานต่าง ๆ โดยใช้สมการในทฤษฎีของเบย์ ทำให้เป็นกระบวนการที่มีการปรับตัวแบบไปตามข้อมูลที่เข้ามาและความน่าจะเป็นของแต่ละสมมติฐานที่เป็นไปได้ ทำให้มีความยืดหยุ่นในการเรียนรู้และปรับตัวแบบให้เหมาะสมกับข้อมูลใหม่ที่เข้ามาในขั้นตอนต่อไปดังนี้

$$P(h|D) = \frac{P(D|h) * P(h)}{P(D)} \quad [1]$$

โดย D แทนข้อมูลที่นำมามาใช้ในการคำนวณการแจกแจงความน่าจะเป็น Posteriori Probability ของสมมติฐาน h คือ P(h|D) ตามทฤษฎี

P(h) คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h

P(D) คือ ความน่าจะเป็นก่อนหน้าของชุดข้อมูลตัวอย่าง D

P(h|D) คือ ความน่าจะเป็นของ h เมื่อรู้ D

ตัวจำแนกแบบนาอีฟ เบย์อย่างง่าย (Naïve Bayesian Classifier) เป็นตัวจำแนกประเภทข้อมูลที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของทฤษฎีของเบย์ (Bayes' Theorem) และสมมติฐานว่าการเกิดของเหตุการณ์ต่าง ๆ เป็นอิสระต่อกัน (Independence) แนวคิดของทฤษฎีเบย์ช่วยในการคำนวณความน่าจะเป็นของเหตุการณ์ A เมื่อทราบข้อมูล B โดยสามารถแสดงได้ในรูปแบบของสมการเบย์ (Peachapong Poolpol ,2021)

หลักการทำงานของ Naive Bayes มีดังนี้

1. เริ่มต้นต้องมีชุดข้อมูลฝึก (training data) ที่รวบรวมข้อมูล ซึ่งใช้สำหรับการเรียนรู้ของ Naive Bayes ชุดข้อมูลฝึกประกอบด้วยตัวอย่างที่มีคุณลักษณะ (features) และคลาส (class) เพื่อเป็นเฉลย เช่น ข้อมูลเกี่ยวกับการจำแนกอีเมลว่าเป็นอีเมลขยะหรือไม่

2.การคำนวณความน่าจะเป็น (Probability Calculation) Naive Bayes ใช้ทฤษฎีเบยส์ในการคำนวณความน่าจะเป็นของเหตุการณ์ต่าง ๆ โดยพิจารณาคุณลักษณะที่เกี่ยวข้องกับคลาสของเหตุการณ์ อัลกอริทึมนี้หาความน่าจะเป็นที่เหตุการณ์จะเป็นคลาสที่ต่างกันตามคุณลักษณะของเหตุการณ์ นั่นหมายความว่า Naive Bayes สมมติว่าไม่มีสหสัมพันธ์ระหว่างคุณลักษณะแต่ละอันและคลาสที่ถูกจำแนกในแต่ละคุณลักษณะ

3. Naive Assumption Naive Bayes ใช้สมมติฐาน "ความไม่เหมาะสม" หรือ "Naive" ซึ่งสมมติว่าคุณลักษณะที่เกี่ยวข้องกับเหตุการณ์นี้เป็นอิสระต่อกัน นั่นหมายความว่า Naive Bayes สมมติว่าไม่มีสัมพันธ์ระหว่างคุณลักษณะแต่ละอันและคลาสที่ถูกจำแนกในแต่ละคุณลักษณะ

4. เมื่อได้รับข้อมูลทดสอบ (test data) Naive Bayes จะคำนวณความน่าจะเป็นของเหตุการณ์ที่เป็นไปได้ในแต่ละคลาสของคลาสที่สามารถจำแนกได้ จากนั้นเลือกคลาสที่มีความน่าจะเป็นสูงสุดเป็นคลาสที่จะกำหนดให้กับข้อมูล

5.การใช้เป็นเครื่องมือจำแนก (Classification Tool) Naive Bayes สามารถนำไปใช้ในงานจำแนกหรือการจำแนกความน่าจะเป็นของเหตุการณ์ในหลายสถานการณ์ เช่น การจำแนกอีเมลเป็นอีเมลขยะหรือไม่ การจำแนกข้อมูลการสื่อสารละเลยและไม่ละเลย การจำแนกความน่าจะเป็นของโรคในการแพทย์ และอื่นๆ (Russell, Norvig, 2009)

การนำตัวแบบ Naïve Bayes ไปใช้

1. ความสมมาตร (Bayes' Theorem) เป็นหลักการทางคณิตศาสตร์ที่ใช้ในการคำนวณความน่าจะเป็นของเหตุการณ์ๆ โดยใช้ข้อมูลที่เรามีเป็นหลัก หลักการนี้มีความสำคัญในสถิติ วิทยาการคอมพิวเตอร์ การเรียนรู้ของเครื่อง และหลายด้านอื่นๆ ด้วยความเฉพาะกลุ่มที่ใช้ใน Naive Bayes สำหรับการจำแนกหรือทำนาย Bayes' Theorem สามารถเขียนแทนได้ดังนี้

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad [2]$$

P(A|B) คือความน่าจะเป็นของเหตุการณ์ A เมื่อทราบเหตุการณ์ B (Posterior Probability)

P(B|A) คือความน่าจะเป็นของเหตุการณ์ B เมื่อทราบเหตุการณ์ A (Likelihood)

P(A) คือความน่าจะเป็นของเหตุการณ์ A (Prior Probability)

P(B) คือความน่าจะเป็นของเหตุการณ์ B (Evidence or Marginal Likelihood)

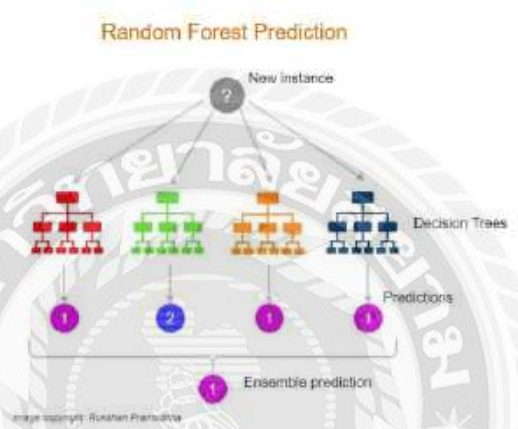
2. สมมติฐาน Naive (Naive Assumption) คือ สมมติฐานหรือความเชื่อที่เรียบง่ายหรือตรงไปตรงมาเกินไป โดยมักไม่ได้คำนึงถึงปัจจัยหรือเงื่อนไขอื่นๆ ที่อาจส่งผลต่อผลลัพธ์หรือการตัดสินใจ สมมติฐาน Naive มักเกิดขึ้นจากความไม่รู้ ความเข้าใจผิด หรืออคติของผู้สร้างสมมติฐาน

3. การคำนวณค่าความน่าจะเป็น (Probability) โดยใช้ข้อมูลการฝึก (training data) Naive Bayes จะคำนวณความน่าจะเป็นของคลาสเริ่มต้นและความน่าจะเป็นของคุณลักษณะที่ระบุสิ่งนั้น

4. การทำนาย (Prediction) การทำนายของ Naive Bayes เป็นเทคนิคที่ใช้หลักการความน่าจะเป็นแบบมีเงื่อนไขของเบส์เพื่อคำนวณความน่าจะเป็นที่ตัวอย่างหนึ่งๆ จะอยู่ในคลาสใดคลาสหนึ่ง การทำนายของ Naive Bayes มีข้อดีคือง่ายต่อการคำนวณและเข้าใจ มีประสิทธิภาพและรวดเร็ว นอกจากนี้ยังสามารถใช้กับข้อมูลที่มีขนาดใหญ่และซับซ้อนได้ อย่างไรก็ตาม สมมติฐาน Naive Assumption อาจไม่เป็นจริงเสมอไป โมเดลอาจทำงานผิดพลาดได้หากข้อมูลไม่ (Sarah Guido, 2017)

เรนดอม ฟลอเรส (Random Forest)

Random Forest เป็นอัลกอริธึมการเรียนรู้ของเครื่องประเภท ensemble learning ที่ใช้ decision tree หลายต้นเพื่อสร้างโมเดลที่มีประสิทธิภาพสูง การทำงานของ Random Forest คือการสร้าง decision tree แต่ละต้นจากชุดข้อมูลย่อยที่แตกต่างกันและรวมผลลัพธ์จากทุก decision tree เพื่อให้ได้คำตอบสุดท้าย การสุ่มตัวอย่างข้อมูลและคุณสมบัติช่วยลดความเสี่ยงของ overfitting และเพิ่มความหลากหลายของโมเดล Random Forest สามารถนำไปใช้ในการจำแนกที่มีประสิทธิภาพและความเสถียร (Andreas C. Müller และ Sarah Guido, 2017)



ภาพที่ 2.2 ตัวอย่างการสร้างต้นไม้ตัดสินใจ

ที่มา : https://www.almabetter.com/bytes/tutorials/data-science/random-forest

การทำงานของ Random Forest

Random Forest เป็นอัลกอริธึมการเรียนรู้ ประเภท ensemble learning ที่สร้าง decision tree หลายต้นจากชุดข้อมูลย่อยที่แตกต่างกัน การสร้าง decision tree แต่ละต้นเป็นการสุ่มตัวอย่างข้อมูลต้นฉบับและคุณสมบัติบางส่วน การลดความเสี่ยงของ overfitting ใน Random Forest ทำได้โดยการลดขนาดของชุดข้อมูล การลดจำนวนคุณสมบัติที่ใช้, และการลดความแปรปรวนของ decision tree แต่ละต้น หลังจากสร้าง decision tree หลายต้น, Random Forest จะรวมการตัดสินใจของทุก decision tree เพื่อสร้างการคาดการณ์สุดท้ายวิธีการรวมการตัดสินใจของ decision tree แต่ละต้นสามารถทำได้

หลายวิธี เช่น การโหวตเสียงข้างมาก การหาค่าเฉลี่ย หรือวิธีอื่นๆ การรวมการตัดสินใจนี้ช่วยลดความแปรปรวนของโมเดลโดยรวม และทำให้โมเดลมีความแม่นยำมากขึ้น

หลักกการเลือกตัวแปรเพื่อแบ่งการตัดสินใจ

การเลือกตัวแปรเพื่อใช้ในการแบ่งการตัดสินใจในตัวอัลกอริทึม Random Forest มีหลายวิธี ซึ่งส่วนใหญ่ขึ้นอยู่กับการสร้าง Random Forest แต่ละครั้ง วิธีที่มักจะใช้เป็นการสร้าง Random Forest จากการสุ่มตัวแปรในแต่ละโหนดในต้นไม้ตัดสินใจแล้ววัดความสำคัญของแต่ละตัวแปรด้วยวิธีต่าง ๆ ซึ่งสามารถแบ่งเป็นสองวิธีหลักคือการใช้ Variable Importance (VIMP) และค่า Minimal Depth

Variable Importance (VIMP)

Variable Importance (VIMP) เป็นวิธีที่ใช้ในการวัดความสำคัญของตัวแปรในการสร้างโมเดล Random Forest หรือต้นไม้ตัดสินใจอื่นๆ VIMP ช่วยให้เราทราบถึงความสำคัญของแต่ละตัวแปรต่อการตัดสินใจของโมเดลโดยวิเคราะห์โหนดและการแบ่งข้อมูลในแต่ละโหนดของต้นไม้ วิธีการคำนวณ VIMP สามารถทำได้หลายวิธีตามอัลกอริทึมที่ใช้ในการสร้าง Random Forest หรือต้นไม้ตัดสินใจ

1. Mean Decrease Accuracy (MDA): ใช้การประมาณค่าความถูกต้องของโมเดล (Accuracy) เมื่อตัวแปรนั้นถูกเอาออกจากโมเดล MDA คำนวณจากความแตกต่างของค่า Accuracy ระหว่างโมเดลที่ใช้ตัวแปรนั้นและโมเดลที่ไม่ใช้ตัวแปรนั้น
2. Mean Decrease Gini (MDG): ใช้ Gini Impurity ในการวัดความสำคัญของแต่ละตัวแปร MDG คำนวณจากความแตกต่างของค่า Gini Impurity ระหว่างโมเดลที่ใช้ตัวแปรนั้นและโมเดลที่ไม่ใช้ตัวแปรนั้น

การคำนวณ VIMP ต่างๆ อาจมีค่าความเป็นไปได้เชิงสถิติ และสามารถสรุปผลลัพธ์ได้หลายวิธีตามวิธีการและอัลกอริทึมที่ใช้ในการสร้างโมเดล เพื่อความถูกต้องและเข้าใจดีขึ้น ควรอ่านเอกสารที่เกี่ยวข้องกับอัลกอริทึมหรือเครื่องมือที่ใช้ในการสร้าง Random Forest หรือต้นไม้ตัดสินใจเพื่อเรียนรู้เพิ่มเติมเกี่ยวกับวิธีการคำนวณ VIMP และการใช้ VIMP ในการวิเคราะห์ข้อมูล

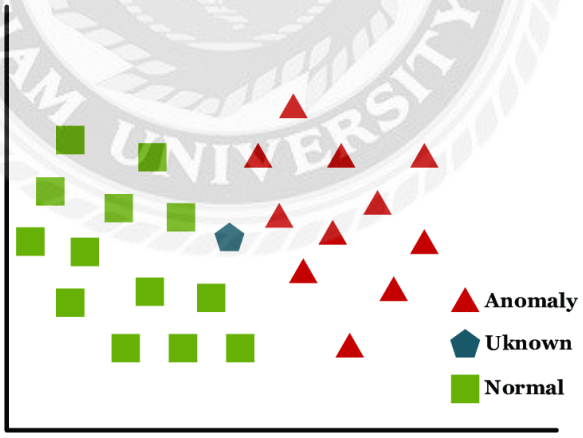
Minimal Depth

เป็นวิธีที่ใช้ในการวัดความสำคัญของตัวแปรจากค่าความลึกของโหนดที่ใช้ในการแบ่งข้อมูลในต้นไม้ตัดสินใจ ค่า Minimal Depth หมายถึงค่าที่บ่งบอกถึงว่าตัวแปรใดมีความสำคัญมากในการ

แบ่งข้อมูลในแต่ละโหนดของต้นไม้ถ้าตัวแปรใดมีค่า Minimal Depth น้อยมาก แสดงว่าการแบ่งด้วยตัวแปรนั้นมีผลมากในการสร้างตัวแบบ การใช้ Minimal Depth เป็นวิธีในการเลือกตัวแปรที่มีความสำคัญมากที่สุดในแต่ละโหนดของต้นไม้ตัดสินใจ การเลือกตัวแปรที่มีค่า Minimal Depth น้อยที่สุดช่วยให้เราทราบถึงตัวแปรที่มีผลมากในการแบ่งข้อมูล ซึ่งสามารถใช้ในการทำนายหรือวิเคราะห์ผลลัพธ์ของต้นไม้ (Adele Cutler และ Matthias Wiener , 2002)

เพื่อนบ้านใกล้ที่สุด - เค (K-Nearest Neighbor)

K-Nearest Neighbor (KNN) เป็นเทคนิคที่นิยมใช้ในการจำแนกหรือทำนายข้อมูลโดยการค้นหาข้อมูลที่อยู่ใกล้ที่สุดกับข้อมูลเป้าหมายจากชุดข้อมูลที่มีอยู่ โดยอาศัยระยะห่างระหว่างข้อมูลในการค้นหา เทคนิคนี้มีข้อดีในความเรียบง่ายและความสามารถในการใช้กับหลายประเภทของข้อมูล แต่ก็มีความซับซ้อนในกระบวนการในกรณีที่มีข้อมูลมาก และการเลือกค่า K ที่เหมาะสมมีความสำคัญ เนื่องจากมีผลต่อผลลัพธ์การใช้เทคนิค K-Nearest Neighbor เน้นการคำนวณระยะห่างระหว่างข้อมูล เพื่อหาข้อมูลที่ใกล้ที่สุด ข้อมูลที่ใกล้ที่สุดนี้จะถูกใช้ในการตัดสินใจเกี่ยวกับประเภทหรือค่าที่ต้องการทำนายความยืดหยุ่นและการใช้กับข้อมูลหลายประเภททำให้ KNN เป็นเครื่องมือที่นิยมในการแก้ปัญหาทั้งทางด้านการจำแนกและการทำนาย



ภาพที่ 2.3 ตัวอย่างการสร้าง K- Nearest Neighbor

ที่มา : https://cutt.ly/40semGG

วิธีการทำงานของเค (K-Nearest Neighbor) ได้แก่

1. การคำนวณระยะห่าง ในขั้นตอนแรก จะทำการคำนวณระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลในชุดข้อมูลทั้งหมด โดยมักใช้วิธีการคำนวณระยะยูคลิเดียน (Euclidean distance) หรือวิธีการคำนวณระยะห่างอื่น ๆ ตามความเหมาะสมของงานและข้อมูล

2. เลือกข้อมูล จากนั้นเรียงลำดับข้อมูลในชุดข้อมูลตามระยะห่างที่คำนวณได้ และเลือก K ข้อมูลที่อยู่ใกล้ที่สุด เรียกข้อมูลเหล่านี้ว่า "เพื่อนบ้าน" (neighbors)

3. การตัดสินใจ ข้อมูลที่ต้องการทำนายจะถูกตัดสินใจโดยการ Vote หรือคิดคะแนนจากความเหมือนหรือความใกล้เคียงกับเพื่อนบ้าน ตัวแบบที่ใช้ในการ Vote จะขึ้นอยู่กับประเภทของการจำแนก (classification) หรือการประมวลผลที่ทำ (regression)

4. การคำนวณผลลัพธ์ ผลลัพธ์จากข้อมูลที่เพื่อนบ้านเลือกจะถูกใช้ในการจำแนกหรือทำนายข้อมูลตัวเป้าหมาย และเป็นผลลัพธ์สุดท้ายของการทำงาน

K-Nearest Neighbor มีการปรับค่า K และการใช้วิธีการคำนวณระยะห่างต่าง ๆ ตามความเหมาะสมของงานและข้อมูลที่ใช้งาน และมีการพัฒนาอัลกอริทึมเพื่อปรับปรุงประสิทธิภาพในการคำนวณ (Christopher M. Bishop, 2006)

ฟังก์ชันความเหมือน (Similarity Function)

การวัดความเหมือนเป็นวิธีการวัดค่าความคล้ายคลึงของวัตถุ 2 ตัว โดยทั่วไปจะมีความหมายตรงกันข้ามกับการวัดระยะห่าง (Distance Measure) ได้นิยามฟังก์ชันความเหมือนดังสมการดังนี้

$$s = \sqrt{(X_i - X_k)^2}$$

x_i คือ ค่าของข้อมูลตัว i

x_k คือ ค่าของข้อมูลตัว k

ระยะห่างยูคลิเลียน (Euclidean Distant)

เป็นการวัดระยะห่างปกติระหว่างจุด 2 จุด ในแนวเส้นตรง ที่วัดได้ด้วยไม้บรรทัด ที่ได้จากทฤษฎีพีทาโกรัส ระยะห่างยูคลิเดียนระหว่างจุด p และ จุด q แสดงด้วย d(p,q) ดังสมการนี้

$$d(p,q) = \sqrt{(p_i - q_i)^2 + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad [2]$$

โดยที่ p = {p1 , p2 , p3 , … , pn} และ q = {q1 , q2 , q3 , … , qn} คือ จุด 2 จุดที่ต้องการคำนวณระยะห่าง

ค่า d(p , q) น้อยแสดงว่า 2 จุด p และ q มีความใกล้เคียงกันมาก (หากมีค่าเป็นศูนย์หมายถึงทั้ง 2 จุด คือจุดเดียวกัน) หากมีค่ามากแสดงว่า 2 จุดนี้ มีความห่างกัน หรือแตกต่างกันมาก

ระยะห่างแมนฮัตตัน (Manhattan Distant)

เป็นการวัดวัดระยะทางระหว่างจุดสองจุดตามแกนวัดมุมขวา ชื่อของสมการนั้นมาจากเค้าโครงของถนนในเมืองแมนฮัตตัน ซึ่งสามารถใช้เส้นทางที่สั้นที่สุดระหว่างจุดสองจุดในเมือง คำนวณดังสมการนี้

$$d = \sum_{i}^{n} \left| X_i - Y_i \right|$$

โดยที่ x_i และ Y_i คือ จุด 2 จุด ที่ต้องการคำนวณระยะห่าง

สหสัมพันธ์โคไซน์ (Cosine Coefficient)

ความคล้ายคลึงโคไซน์ (Cosine Similarity) เป็นการวัดความคล้ายระหว่าง 2 เวกเตอร์ โดยวัดมุมโคไซน์ของเวกเตอร์ทั้ง 2 คำนวณได้ดังสมการนี้

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \qquad [1]$$

$$= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum^{n} (A_i)^2} \times \sqrt{\sum^{n} (B_i)^2}} \qquad [2]$$

โดยที่ A = {a1 , a2 , a3 , … , an} และ B = {b1 , b2 , b3 , … , bn} 2 เวกเตอร์ที่ต้องการเปรียบเทียบ ค่าสหสัมพันธ์โคไซน์จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยมีความหมายดังนี้

ถ้าค่าเข้าใกล้ 1 หมายถึง ทั้ง 2 เวกเตอร์มีค่าความสัมพันธ์กันมากไปในทางเดียวกัน

ถ้าค่าเข้าใกล้ -1 หมายถึง ทั้ง 2 เวกเตอร์มีค่าความสัมพันธ์ไปในทิศทางตรงกันข้าม

ถ้าค่าเข้าใกล้ 0 หมายถึง ทั้ง 2 เวกเตอร์ไม่มีความสัมพันธ์กัน

สหสัมพันธ์เพียร์สัน (Pearson Coefficient)

เป็นวิธีที่ใช้วัดความสัมพันธ์ระหว่างตัวแปร หรือข้อมูล 2 ชุด โดยที่ตัวแปรข้อมูล 2 ชุดนั้นอยู่ในรูปข้อมูลในระบบมาตราอันตรภาคหรืออัตราส่วน (Interval or Ratioscale) คำนวณดังสมการนี้

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left(N\sum X^2 - (\sum X)^2\right)\left(N\sum Y^2 - (\sum Y)^2\right)}}$$

โดยที่ r_{XY} เป็นค่าสัมประสิทธิ์สหพันธ์แบบเพียร์สัน

Σ_x เป็นผลรวมของข้อมูลที่วัดได้จากตัวแปรที่ 1 (X)

Σ_y เป็นผลรวมของข้อมูลที่วัดได้จากตัวแปรที่ 2 (Y)

Σ_{xy} เป็นผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2

$\sum x^2$ เป็นผลรวมกำลังสองของข้อมูลที่ได้จากตัวแปรที่ 1

งานวิจัยที่เกี่ยวข้อง

ธนะวัฒน์ วรรณประภา (2564) การใช้เทคนิคเหมืองข้อมูลเพื่อคัดกรองบุคคลที่มีแนวโน้มสัมฤทธิผลในการศึกษาระดับปริญญาตรี สาขาวิชาเทคโนโลยีการศึกษา งานวิจัยนี้ใช้เทคนิคเหมืองข้อมูลเพื่อคัดกรองบุคคลที่มีแนวโน้มสัมฤทธิผลในการศึกษาระดับปริญญาตรี สาขาเทคโนโลยีการศึกษา ด้วย Decision Tree, Support Vector Machine, และ Deep Learning. ผลลัพธ์แสดงว่า Deep Learning ให้ความถูกต้องสูงสุดที่ 85.94% ในการคัดกรองและทำนายผลการศึกษา ด้วยการวิเคราะห์

แบบ Stepwise Selection และให้ค่าความถูกต้องดีกว่า Decision Tree และ Support Vector Machine ที่มีความถูกต้องตามลำดับคือ 84.46% และ 80.81%

สุรวัชร ศรีเปารยะ สายชล สินสมบูรณ์ทอง (2560) การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย งานวิจัยนี้มีวัตถุประสงค์เปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มที่ใช้วิธีความใกล้เคียง ต่าง ๆ ในการทำนายผลโรคไตเรื้อรัง โดยใช้ข้อมูลผู้ป่วยจากโรงพยาบาลอพอลโล ประเทศอินเดีย แบ่งข้อมูลเป็นชุดสร้างแบบจำลองและชุดทดสอบตัวแบบในอัตราส่วน 70:30 ตามลำดับ จากผลการเปรียบเทียบประสิทธิภาพ พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องที่ 100% และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.0059 ซึ่งเป็นวิธีการที่มีประสิทธิภาพการจำแนกดีที่สุดในการทำนายโรคไตเรื้อรัง

บีสุตา ดาวเรือง จรัญ แสนราช และ อนิราช มิ่งขวัญ (2564) การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อทำนายแขนงวิชาเรียนของนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศและการจัดการอุตสาหกรรม การวิจัยนี้มุ่งเน้นเปรียบเทียบประสิทธิภาพของโมเดลการจำแนกในการทำนายแขนงวิชาเรียน โดยใช้เทคนิคเหมืองข้อมูล 3 วิธีคือ ต้นไม้ตัดสินใจ, แบบเบย์, และฐานกฎ โดยใช้ข้อมูลนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม ในช่วงปีการศึกษา 2548-2559 จำนวน 759 รายการ ผลลัพธ์ของการวิเคราะห์บนพื้นฐานของวิธี 10-Fold Cross Validation แสดงให้เห็นว่า วิธีฐานกฎมีประสิทธิภาพสูงสุดด้วยความถูกต้องที่ 78.66% ตามมาด้วยวิธีต้นไม้ตัดสินใจที่ 78.26% และวิธีแบบเบย์ที่ 78.12% ผลลัพธ์นี้เสนอแนะให้นำวิธีฐานกฎไปใช้ในการทำนายแขนงวิชาเพื่อช่วยนักศึกษาในการเลือกเรียนและประสบความสำเร็จในการศึกษาตามหลักสูตรที่เหมาะสม

ประเทือง วงษ์ทอง วาฤทธิ์ กันแก้ว อภิชชญา ขวัญแก้ว และ ยุวดี โฉมแดง (2563) การประยุกต์เทคนิคเหมืองข้อมูลค้นหาลักษณะนิสัยของผู้ประกอบอาชีพด้านคอมพิวเตอร์ งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างและเปรียบเทียบประสิทธิภาพของตัวแบบความสัมพันธ์ของข้อมูลจากเทคนิค Neural Networks, Naive Bayes, และ Decision Tree เพื่อค้นหาลักษณะนิสัยของผู้ประกอบอาชีพด้านคอมพิวเตอร์ตามทฤษฎีการเลือกอาชีพของฮอลแลนด์ การเก็บรวบรวมข้อมูลทำจากผู้ประกอบอาชีพด้านคอมพิวเตอร์ ทั้งหมด 395 คนที่สำเร็จการศึกษาระดับปริญญาตรีในหลักสูตรวิศวกรรม

คอมพิวเตอร์และวิทยาการคอมพิวเตอร์ ผลลัพธ์ของงานวิจัยพบว่า ตัวแบบที่มีประสิทธิภาพที่สุดคือ ตัวแบบจากเทคนิค Naïve Bayes ซึ่งมีค่าความถูกต้องเท่ากับ 79.64% ตามลำดับต่อมาคือ Neural Network ที่มีค่าเท่ากับ 73.60% และเทคนิค Decision Tree ที่มีค่าเท่ากับ 54.69% ผลลัพธ์นี้จึงช่วยให้เข้าใจลักษณะนิสัยของผู้ประกอบอาชีพด้านคอมพิวเตอร์ตามทฤษฎีการเลือกอาชีพของฮอลแลนด์โดยมีการเปรียบเทียบความสามารถของตัวแบบที่ใช้ในการจำแนกข้อมูล

วนิดา พงษ์สงวน ทิพยา ถินสูงเนิน และ มาโนช ถินสูงเนิน (2560) การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ โครงงานนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคเหมืองข้อมูลแบบต้นไม้ตัดสินใจ เพื่อช่วยในการวิเคราะห์หาแบบจำลองของปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน ใช้ข้อมูลผู้เข้ารับบริการที่โรงพยาบาลด่านขุนทด จังหวัด นครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402 ราย แบ่งข้อมูลสำหรับฝึกและทดสอบแบบจำลองออกเป็นร้อยละ 90:10 พัฒนาแบบจำลองด้วยอัลกอริทึม J48 ประเมินประสิทธิภาพของ แบบจำลองด้วยค่าความแม่น ผลการวิจัยพบว่าแบบจำลองที่พัฒนาให้ประสิทธิภาพที่มีค่าความแม่นยำ 76.14 % และสามารถจำแนกต้นไม้ตัดสินใจทั้งหมด 97 กฎ

จุฑาภรณ์ สุคตะ ศรัญญู สวนทอง วิภาวรรณ บัวทอง และ สมใจ จิตคำนึงสุข (2564) การใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์โรคของผู้สูงอายุที่เกิดขึ้นพร้อมกัน งานวิจัยนี้มุ่งหวังที่จะแสดงความสัมพันธ์ของโรคที่พบบ่อยในผู้สูงอายุที่มีโรคประจำตัว ผ่านการวิเคราะห์ข้อมูลจาก 549 ผู้สูงอายุที่มีโรคประจำตัว โดยใช้เทคนิคการทำเหมืองข้อมูลแบบการสร้างกฎความสัมพันธ์ โครงสร้างการทำเหมืองข้อมูลแบบ Decision Tree และ J48 ได้แบ่งกฎความสัมพันธ์ออกเป็น 17 กฎที่มีความถูกต้องถึง 96.95% ในการทดสอบโมเดล ทำให้งานวิจัยเป็นประโยชน์ในการวิเคราะห์ความสัมพันธ์ของโรคในกลุ่มผู้สูงอายุที่มีโรคประจำตัวและสามารถนำข้อมูลเหล่านี้ไปใช้ในการปรับปรุงการดูแลรักษาและเตรียมการให้บริการด้านสุขภาพสำหรับกลุ่มผู้สูงอายุได้ในอนาคต

ทิพย์ธิดา วงศ์พิพันธ์ (2560) การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ กรณีศึกษากรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน) งานวิจัยนี้เน้นการใช้ความรู้ที่ได้จากการทำเหมืองข้อมูลเพื่อช่วยในการตัดสินใจในการให้สินเชื่อที่มีประสิทธิภาพ โดยใช้เทคนิคต้นไม้ตัดสินใจและ Naïve Bayes เพื่อจำแนกกลุ่มลูกค้าเป็นกลุ่มที่ดีและไม่ดี ผลลัพธ์ของการทดลองแสดงว่าเทคนิคต้นไม้ตัดสินใจให้ผลลัพธ์ที่ดีที่สุด ด้วยความถูกต้องถึง 90.47% โดยมีความชัดเจนในการแบ่งกลุ่มลูกค้า

ผลลัพธ์ที่ได้จากงานวิจัยนี้สามารถนำมาปรับปรุงกระบวนการตัดสินใจการให้สินเชื่อเพื่อลดปัญหาหนี้สูญในองค์กรได้ นอกจากนี้ยังเป็นแนวทางที่สามารถนำไปใช้ในการพัฒนาระบบการตัดสินใจการให้สินเชื่อในสายธุรกิจการเงินหรือสินเชื่ออื่น ๆ อย่างมีประสิทธิภาพ

สำราญ วานนท์ ธรัช อารีราษฎร์ และจรัญ แสนราช (2561) การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีสาขาคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล งานวิจัยนี้มุ่งสู่การศึกษาเทคนิคการพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีในสาขาคอมพิวเตอร์ โดยใช้เทคนิคเหมืองข้อมูล เบื้องต้นนำเข้าข้อมูลระเบียนประวัติของนิสิตในช่วง 5 ปีย้อนหลัง (2555-2559) จำนวน 65,335 รายการ เพื่อทดลองวัดความแม่นยำของเทคนิคต่าง ๆ ได้แก่ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคแรนดอมฟอร์เรส (Random Forest) และเทคนิคแบ็กกิง (Bagging) ผลการวิจัยพบว่าเทคนิคแรนดอมฟอร์เรสมีความแม่นยำสูงที่สุดที่เท่ากับ 84.29% เมื่อเทียบกับเทคนิคต้นไม้ตัดสินใจ (81.91%) และเทคนิคแบ็กกิง (81.71%) ทำให้เห็นถึงความเหมาะสมของเทคนิคแรนดอมฟอร์เรสในการทำนายอาชีพของนักศึกษาระดับปริญญาตรีในสาขาคอมพิวเตอร์ ซึ่งสามารถนำไปใช้ในการให้คำแนะนำหรือสนับสนุนการตัดสินใจทางอาชีพของนักศึกษาได้อย่างมีประสิทธิภาพ

ภาภรณ์ เหล่าพิลัย และจรัญ แสนราช (2563) การวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่เกี่ยวข้องกับการลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิค Rule Induction, K-Nearest Neighbor, Decision Tree และ Naive Bayes ด้วยข้อมูลจากฐานข้อมูลงานทะเบียนของมหาวิทยาลัย เทคโนโลยีราชมงคลอีสาน นักศึกษาระดับปริญญาตรีในระหว่างปีการศึกษา 2557-2561 จำนวน 14 แอททริบิวต์และ 10,151 ชุดข้อมูลผลการวิจัยพบว่าโมเดลที่สร้างด้วยเทคนิควิธี Rule Induction มีประสิทธิภาพสูงสุด มีค่าเฉลี่ยความถูกต้องที่ 94.70% และปัจจัยที่เกี่ยวข้องสูงสุด 5 อันดับได้แก่ เกรดเฉลี่ย ปีการศึกษา โรงเรียนเดิมสาขาวิชา และอาชีพของบิดา ทำให้เห็นถึงความสำคัญของปัจจัยเหล่านี้ในการทำนายพฤติกรรมการลาออกของนักศึกษา

นภาพร สุทธินาคสมบัติ (2558) การจำแนกข้อมูลเพื่อพยากรณ์ราคาเครื่องสำอางด้วยเทคนิคเหมืองข้อมูล งานวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกข้อมูลและพยากรณ์ราคาเครื่องสำอางโดยใช้เทคนิคเหมืองข้อมูลสำหรับการสร้างตัวแบบในการจำแนกประเภทข้อมูล จำนวน 3 เทคนิคได้แก่ นาอีฟเบย์

ต้นไม้ตัดสินใจ และเพื่อนบ้านใกล้ที่สุด โดยใช้ข้อมูลจาก Vanilla.in.th และ Cosmenet.in.th จำนวนรวม 6,000 รายการ ผลการทดสอบจากข้อมูลที่มีความเหมาะสมในการพยากรณ์ (4,323 รายการ) แสดงให้เห็นว่าเทคนิคที่เหมาะสมที่สุดในการจำแนกประเภทข้อมูลคือ "เพื่อนบ้านใกล้ที่สุด" ซึ่งให้ค่าความถูกต้องที่ร้อยละ 99.47% นั้นหมายถึงความแม่นยำและประสิทธิภาพสูงสุดของโมเดลที่ได้รับการทดสอบในงานวิจัยนี้

บทสรุปของงานวิจัยที่เกี่ยวของ

จากที่ได้ศึกษางานวิจัยที่เกี่ยวข้องของนักวิจัยหลายๆท่าน สามารถนำมาสรุปได้ดังนี้

ตารางที่ 2.1 สรุปงานวิจัยที่เกี่ยวข้อง

ผู้แต่ง ปี	ชื่อเรื่อง	ข้อมูล	อัลกอริทึม	สรุปผล
ธนะวัฒน์ วรรณประภา (2564)	การใช้เทคนิคเหมืองข้อมูลเพื่อคัดกรองบุคคลที่มีแนวโน้มสัมฤทธิผลในการศึกษาระดับปริญญาตรี สาขาวิชาเทคโนโลยีการศึกษา	ข้อมูลจากบัณฑิตที่สำเร็จการศึกษา ปีการศึกษา 2554-2558 จำนวน 738 คน	-Decision Tree -Support Vector Machine - Deep Learning	Deep Learning ให้ค่าความถูกต้อง เท่ากับ 85.94 %
สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง (2560)	การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย	กลุ่ม ผู้ป่วยโรคไตเรื้อรัง ของโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย จำนวน 120 ระเบียน	- Decision Tree - Neural Network - Support Vector machine - Naïve Bayes	Decision Tree ให้ค่าความถูกต้อง คือ 94.17%

ผู้แต่ง ปี	ชื่อเรื่อง	ข้อมูล	อัลกอริทึม	สรุปผล
บีสุตา ดาวเรือง จรัญ แสนราช และ อนิราช มิ่งขวัญ (2564)	การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อทำนายแขนงวิชาเรียนของนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศและการจัดการอุตสาหกรรม	ข้อมูลของนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยี และการจัดการอุตสาหกรรม ตั้งแต่ปีการศึกษา 2548-2559 จำนวน 759 ระเบียน	-Rule-Based Classification - Decision Tree - Naïve Bayes	Rule-Based Classificationให้ประสิทธิภาพสูงสุด ให้ค่าความถูกต้องเท่ากับ 78.66%
ประเทือง วงษ์ทอง วาฤทธิ์ กันแก้ว อภิชชญา ขวัญแก้ว และ ยุวดี โฉมแดง (2563)	การประยุกต์เทคนิคเหมืองข้อมูลค้นหาลักษณะนิสัยของผู้ประกอบอาชีพด้านคอมพิวเตอร์	ข้อมูลจากผู้ประกอบอาชีพด้านคอมพิวเตอร์ กลุ่มตัวอย่าง 395 คน	-Decision Tree -Neural Networks -Naive Bayes	Naïve Bayes มีค่าความถูกต้องเท่ากับ79.64%
วนิดา พงษ์สงวน ทิพยา ถินสูงเนิน และ มาโนช ถินสูงเนิน (2560)	การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ	ข้อมูลโรงพยาบาลด่านขุนทด จังหวัดนครราชสีมา ระหว่างปี 2550 – 2555 จำนวนทั้งสิ้น 4,402 ราย	-Decision Tree	Decision Tree มีค่าความแม่นยำ 76.14 %

ผู้แต่ง ปี	ชื่อเรื่อง	ข้อมูล	อัลกอริทึม	สรุป
จุฑาภรณ์ สุคตะ ศรัญญู สวนทอง วิภาวรรณ บัวทอง และ สมใจ จิตคำนึงสุข (2564)	การใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์โรคของผู้สูงอายุที่เกิดขึ้นพร้อมกัน	ข้อมูลผู้สูงอายุที่มีโรคประจำตัว จำนวน 549 คน	-Decision Tree	Decision Tree มีความถูกต้อง 96.95%
ทิพย์ธิดา วงศ์พิพันธ์ (2560)	การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อกรณีศึกษากรุงไทยคาร์เร้นท์ แอนด์ ลีส	ข้อมูลลูกค้าที่ผ่านการอนุมัติตั้งแต่ ปี2551-2553 จำนวน 441 ราย	- Decision Tree -Naïve Bayes	Decision Tree ผลลัพธ์มีความถูกต้อง 90.47 %
สำราญ วานนท์ ธรัช อารีราษฎร์ และจรัญ แสนราช (2561)	การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีสาขาคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล	ข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษาย้อนหลัง 5 ปี คือ ปี 2555-2559 จำนวน 65,335 ระเบียน	-Decision Tree -Random Forest - Bagging	Random Forest มีค่าความแม่นยำเท่ากับ 84.29%

ผู้แต่ง ปี	ชื่อเรื่อง	ข้อมูล	อัลกอริทึม	สรุปผล
ภาภรณ์ เหล่าพิลัย และจรัญ แสนราช (2563)	การวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรีโดยใช้เทคนิควิธีการทำเหมืองข้อมูล	ข้อมูลจากฐานข้อมูลมหาวิทยาลัยเทคโนโลยีราชมงคลอีสานของนักศึกษาระดับปริญญาตรีระหว่างปีการศึกษา 2557-2561 มีจำนวน 14 แอททริ บิวต์และ 10,151 ชุดข้อมูล	-Rule Induction -KNearest Neighbor -Decision Tree -Naive Bayes	Rule Induction มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้อง 94.70 %
นภาพร สุทธินาคสมบัติ (2558)	การจำแนกข้อมูลเพื่อพยากรณ์ราคาเครื่องสำอางด้วยเทคนิคเหมืองข้อมูล	ข้อมูลเครื่องสำอางรวมทั้งสิ้น 6,000 รายการ ข้อมูลที่มีความเหมาะสมในการพยากรณ์มีจำนวน 4,323 รายการ	-KNearest Neighbor -Decision Tree -Naive Bayes	KNearest Neighbor ให้ค่าความถูกต้องอยู่ที่ร้อยละ 99.47 %

บทที่3

วิธีดำเนินการวิจัย

งานวิจัยนี้ได้นำเทคนิคเหมืองข้อมูลมาใช้เพื่อการจำแนกรุ่นรถยนต์ประเภทรถกระบะ กรณีศึกษา โชว์รูมอีซูซุมหานคร สาขามหาชัย โดยมีขั้นตอนการดำเนินการวิจัยด้วยกระบวนการ CRIS-DM ดังต่อไปนี้

3.1 ทำความเข้าใจกับปัญหา (Business Understanding)

3.2 ทำความเข้าใจกับข้อมูล (Data Understanding)

3.3 การเตรียมข้อมูล (Data Preparation)

3.4 การสร้างแบบจำลองข้อมูล (Modeling)

3.5 การประเมินผล (Evaluation)

3.6 การนำไปใช้ (Deployment)

ทำความเข้าใจกับปัญหา (Business Understanding)

กระบวนการทำงานในการจำหน่ายรถกระบะให้แก่ลูกค้าของ โชว์รูม อีซูซุมหานคร สาขา มหาชัย จะเริ่มจากการให้ลูกค้านั้นเลือกซื้อรถก่อนแล้วจึงทำการสั่งผลิตรถ ทำให้เกิดความล่าช้าในการจำหน่ายรถกระบะเป็นอย่างมาก ผู้จัดทำจึงคิดหาวิธีแก้ปัญหา โดยการทำเหมืองข้อมูล เพื่อพยากรณ์รุ่นของรถกระบะที่ลูกค้าต้องการ เพื่อทางผู้จัดการของ โชว์รูม อีซูซุมหานคร สาขา มหาชัย จะนำข้อมูลที่ได้จากการพยากรณ์มาทำการวางแผนในการสั่งผลิตรถกระบะเพื่อรอพร้อมจำหน่ายให้แก่ลูกค้า โดยที่ลูกค้าไม่ต้องรอรถเป็นเวลานาน และยังสามารถส่งเสริมการขายให้กับรถกระบะรุ่นอื่นๆ ให้ตรงกับกลุ่มเป้าหมายได้

ทำความเข้าใจกับข้อมูล (Data Understanding)

ข้อมูลที่นำมาพยากรณ์จะเป็นข้อมูลจากแฟ้มข้อมูลลูกค้าจากการจำหน่ายรถกระบะของ โชว์รูม อีซูซุมหานคร สาขา มหาชัย ตั้งแต่เดือนมกราคม พ.ศ. 2563 ถึง เดือนธันวาคม พ.ศ. 2564จำนวน 1,500 คน โดยจะนำข้อมูลมาใช้ดังนี้ 1. เพศ 2. อาชีพ 3. ฐานเงินเดือน 4. เงินจองสำหรับทำสัญญาจองรถยนต์ 5. เงินดาวน์ 6. ที่มาของลูกค้า 7. ประวัติเครดิต 8. ระบบเครื่องยนต์ 9. ระบบเกียร์ 10.ประเภทของรถยนต์ เพื่อนำมาวิเคราะห์ในลำดับต่อไป

การเตรียมข้อมูล (Data Preparation)

หลังจากได้ทำความเข้าใจกับข้อมูล และทำการรวบรวมข้อมูล ทำให้ทราบถึงคุณลักษณะที่จะนำมาใช้ในการพิจารณาในการจำแนกประเภทของรถกระบะ ซึ่งก่อนที่จะนำข้อมูลที่ได้มาเข้าสู่กระบวนการ การสร้างตัวแบบจะต้องทำการจัดเตรียมข้อมูลเพื่อให้มีความเหมาะสมกันที่จะนำมาทำการพยากรณ์ และเป็นข้อมูลที่มีลักษณะการจัดเก็บเหมือนกัน โดยมีคุณลักษณะที่นำมาพิจารณาดังนี้

	A	B	C	D	E	F	G	H	I	J
1	Gender	Career	Salary	Booking_Money	Cus_Source	Down_Payment	Engine_System	Gear_System	Credit_history	Car_Model
2	Female	Employee	High	High	Facebook	High	1.9 M/T		Bad	SPC
3	Male	Merchant	Normal	High	Walk	Normal	3 M/T		Good	SPARK
4	Female	Employee	High	Low	Old_Cus	Normal	1.9 M/T		Good	SPC
5	Male	Self Busine	Normal	High	Walk	Normal	1.9 M/T		Good	CAB4
6	Female	Merchant	Normal	Low	Facebook	Normal	3 M/T		Good	SPC
7	Female	Governme	High	High	Rec_Cus	High	3 A/T		Good	CAB4 HR
8	Male	Merchant	High	High	Facebook	Normal	3 M/T		Good	SPARK
9	Female	Employee	High	High	Rec_Cus	Normal	1.9 A/T		Good	CAB4 HR
10	Female	Employee	Normal	High	Walk	Normal	1.9 M/T		Good	SPC HR
11	Male	Self Busine	High	Low	Walk	Normal	1.9 M/T		Good	CAB4
12	Female	Employee	High	Low	Old_Cus	Normal	1.9 A/T		Bad	CAB4 HR
13	Female	Contractor	High	High	Walk	High	1.9 M/T		Good	SPC HR
14	Male	Governme	High	High	Facebook	Normal	1.9 A/T		Good	CAB4 HR
15	Female	Merchant	High	High	Facebook	Normal	1.9 M/T		Good	SPARK
16	Male	Employee	High	High	Rec_Cus	Normal	1.9 M/T		Good	SPC
17	Male	Employee	High	High	Walk	High	1.9 A/T		Good	CAB4 HR
18	Male	Farmer	High	High	Rec_Cus	High	3 A/T		Good	CAB4 HR
19	Male	Contractor	High	High	Walk	High	3 M/T		Good	SPARK
20	Female	Employee	High	High	Facebook	High	1.9 M/T		Good	SPC HR
21	Female	Merchant	Normal	High	Rec_Cus	High	1.9 M/T		Good	CAB4 HR
22	Female	Self Busine	High	High	Facebook	High	1.9 M/T		Good	CAB4
23	Female	Contractor	High	High	Facebook	High	1.9 M/T		Bad	CAB4 HR
24	Female	Self Busine	High	High	Old_Cus	High	1.9 M/T		Bad	CAB4 HR
25	Female	Employee	Normal	Low	Facebook	Normal	1.9 M/T		Good	SPC HR
26	Male	Merchant	High	High	Rec_Cus	Normal	1.9 M/T		Bad	SPC HR
27	Female	Farmer	High	High	Walk	High	3 A/T		Good	CAB4 HR
28	Female	Merchant	High	High	Facebook	High	1.9 A/T		Good	CAB4 HR
29	Female	Contractor	Normal	High	Facebook	High	1.9 M/T		Bad	CAB4
30	Male	Employee	High	High	Facebook	High	1.9 M/T		Good	SPC HR
31	Male	Employee	Normal	Low	Walk	Normal	1.9 M/T		Bad	CAB4
32	Female	Contractor	High	High	Old_Cus	High	1.9 M/T		Good	SPARK
33	Female	Merchant	High	Low	Facebook	Normal	3 M/T		Bad	SPC
34	Female	Merchant	High	High	Facebook	High	1.9 A/T		Good	CAB4 HR
35	Female	Self Busine	High	High	Walk	High	1.9 M/T		Good	SPC HR
36	Female	Employee	Normal	High	Walk	Normal	1.9 M/T		Good	SPC HR
37	Female	Employee	High	High	Rec_Cus	Normal	1.9 A/T		Good	CAB4 HR
38	Female	Employee	Normal	High	Walk	Normal	1.9 M/T		Good	SPC HR
39	Female	Farmer	High	High	Walk	High	1.9 M/T		Good	CAB4 HR

ภาพที่ 3.1 ตัวอย่างข้อมูลลูกค้าของโชว์รูม อีซูซุมหานคร สาขา มหาชัย

ตารางที่ 3.1 การกำหนดตัวแปร

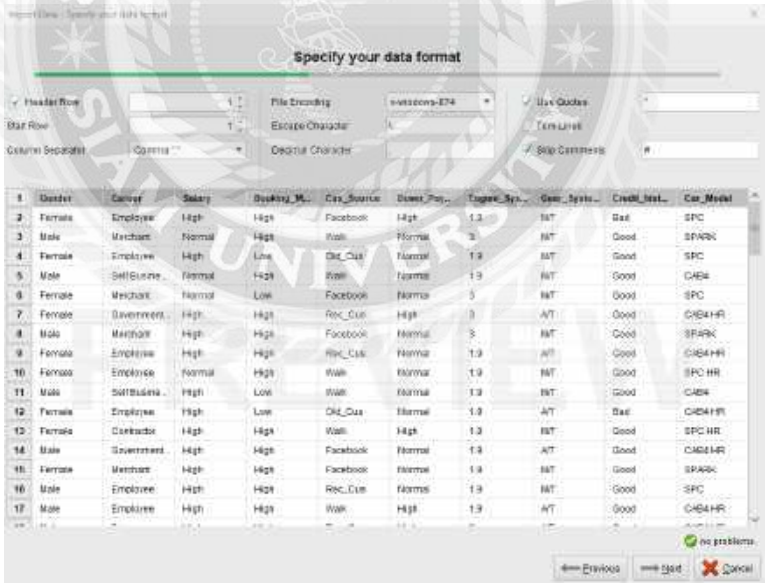
Attribute	คำอธิบายค่า	ความหมาย
Gender	Male = เพศชาย Female = เพศหญิง	เพศ
Career	Employee = พนักงานบริษัท Farmer = เกษตรกร Merchant = ค้าขาย Self Business = ธุรกิจส่วนตัว Government Officer = ข้าราชการ Contractor = รับเหมาก่อสร้าง Lawyer = ทนายความ	อาชีพ
Salary	Normal = ฐานเงินเดือน 15,000 – 35,999 บาท High = ฐานเงินเดือน 36,000 – 55,000 บาท	ฐานเงินเดือน
Booking_Money	Normal = ใช้เงินจอง 1,000 ถึง 2,999 บาท High = ใช้เงินจอง 3,000 ถึง 5,000 บาท	เงินจอง
Down_Payment	Normal = ใช้เงินดาวน์ 10% ถึง 19% High = ใช้เงินดาวน์ 20% ถึง 30%	เงินดาวน์
Cus_Source	Facebook =ลูกค้ามากจากเฟสบุ๊ค Walk = ลูกค้ามาจากโชว์รูม Rec_Cus = ลูกค้าแนะนำ Old_Cus = ลูกค้าเก่า Call = ลูกค้าโทรเข้ามา	ที่มาของลูกค้า
Credit_history	Good = เครดิตดี Bad = เครดิตแย่	ประวัติเครดิต
Engine_System	1.9 = เครื่องยนต์ 1.9 3.0 = เครื่องยนต์ 3.0	ระบบเครื่องยนต์
Gear_System	A/T = เกียร์อัตโนมัติ M/T = เกียร์กระปุก	ระบบเกียร์

Attribute	คำอธิบายค่า	ความหมาย
Car_Model	SPARK = กระบะหัวเดียว SPC = กระบะแค็บ 2 ประตู SPC HR = กระบะแค็บ 2 ประตู ยกสูง CAB4 = กระบะ 4 ประตู CAB4 HR = กระบะ 4 ประตูยกสูง	ประเภทของรถยนต์

การสร้างแบบจำลองข้อมูล (Modeling)

เมื่อทำการ Cleaning Data เรียบร้อยแล้ว สร้างตัวแบบการพยากรณ์ให้ครบ 4 เทคนิค ในโปรแกรม RapidMiner Studio ได้แก่ Decision Tree, Naïve Bayes, Random forest และ K-Nearest Neighbor เพื่อนำมาเปรียบเทียบประสิทธิภาพในการพยากรณ์ จึงได้กระบวนการสร้างตัวแบบดังนี้

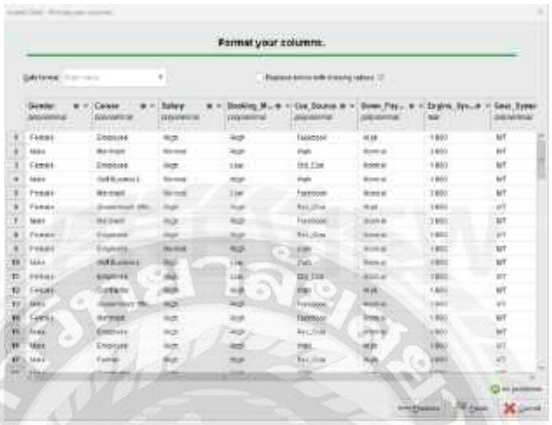
1. ทำการนำเข้าข้อมูลเข้าสู่โปรแกรม RapidMiner X เพื่อวิเคราะห์และคำนวณค่าต่างๆที่ใช้ในการทำนาย



ภาพที่ 3.2 การกำหนดชนิดข้อมูลที่ต้องการทำนาย

จากการสำรวจข้อมูลประกอบด้วย 1500 แถว 10 ตัวแปร ซึ่งแต่ละตัวแปรจะแสดงให้เห็นถึงคุณลักษณะเฉพาะของข้อมูล และสามารถนำมาสร้างตัวแบบในการทำนายได้

2. การสร้างตัวแบบในการจำแนกรุ่นรถกระบะที่ลูกค้าต้องการ โดยนำเข้าชุดข้อมูล IsuzuCustomerData2563To2564.csv ผ่านโอเปอร์เรเตอร์ ReadCSV จึงกำหนดบทบาทของคุณลักษณะ โดยใช้โอเปอร์เรเตอร์ Set Role ระบุตัวแปร Car_Model เป็น label

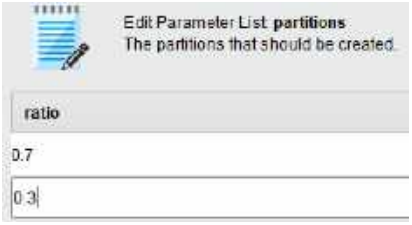


ภาพที่ 3.3 การตั้งค่าภายใน Operator Set Role

ตารางที่ 3.2 กำหนดประเภทของข้อมูลและกำหนดบทบาทของ Attribute ดังนี้

Attribute	**Change Type**	**Change Role**
Gender	Binominal	-
Career	Polynominal	-
Salary	Binominal	-
Booking_Money	Binominal	-
Cus_Source	Polynominal	-
Down_Payment	Binominal	-
Credit_History	Binominal	-
Engine_System	Binominal	-
Gear_System	Binominal	-
Car_Model	Polynominal	Label

3. ทำการแบ่งชุดข้อมูลออกเป็นจำนวน 2 กลุ่ม ผ่านโอเปอร์เรเตอร์ Split Data คือกลุ่มข้อมูลที่ใช้ในการสอน (Training Data) โดยผู้วิจัยกำหนดค่าสำหรับการ Train ตั้งแต่ 70:30, 75:25 และ 80:20 ตามลำดับ เพื่อให้ได้จำนวนกลุ่มตัวอย่างก่อนการเชื่อมต่อกับเทคนิคการทำนายต่างๆโดยใช้เทคนิคเหมืองข้อมูลดังต่อไปนี้



ภาพที่ 3.4 การแบ่งข้อมูลเป็น Training Data และ Test Data

4. การสร้างตัวแบบ Decision Tree

ภาพที่ 3.5 สร้างตัวแบบ Decision Tree โดยใช้โอเปอร์เรเตอร์ Decision Tree

4.1 กำหนดค่า Parameters ในตัวเลือก Criterion ตั้งค่าเป็น Gini index เป็นค่าที่วัดความบริสุทธิ์ของข้อมูล สามารถใช้เพื่อเลือกคุณสมบัติที่ดีที่สุดในการแยกข้อมูลออกเป็นกลุ่มต่างๆ หรือเพื่อเปรียบเทียบประสิทธิภาพของคุณสมบัติต่างๆ ในการแยกข้อมูลออกเป็นกลุ่ม

ภาพที่ 3.6 ขั้นตอนการปรับพารามิเตอร์ Criterion เป็น gini index

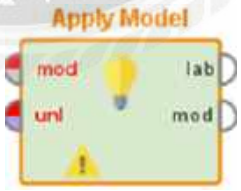
4.2 Confidence ตั้งค่าเป็น 0.1 เพื่อกำหนดค่าความเชื่อมั่นที่ใช้ในการคำนวณ Error ในการทำนายผลของอัลกอริทึมต้นไม้ตัดสินใจ

ภาพที่ 3.7 การตั้งค่าพารามิเตอร์ Confidence เป็น 0.1

4.3 Maximal Depth ตั้งค่าตามลำดับตั้งแต่ 5 ขั้น และ 10 ขั้น เพื่อกำหนดความลึกของต้นไม้ โดยเปรียบเทียบค่า Maximal Depth ที่ให้ผลการทำนายที่ดีที่สุด

ภาพที่ 3.8 การตั้งค่าพารามิเตอร์ Maximal Depth ความลึกของต้นไม้ตัดสินใจ

4.4 ขั้นตอนต่อไปคือการเลือกใช้โอเปอร์เรเตอร์ Apply Model เพื่อใช้ในการนำชุดข้อมูลที่ได้มาทำการประมวลผลลัพธ์สำหรับการทำนาย ดังภาพต่อไปนี้



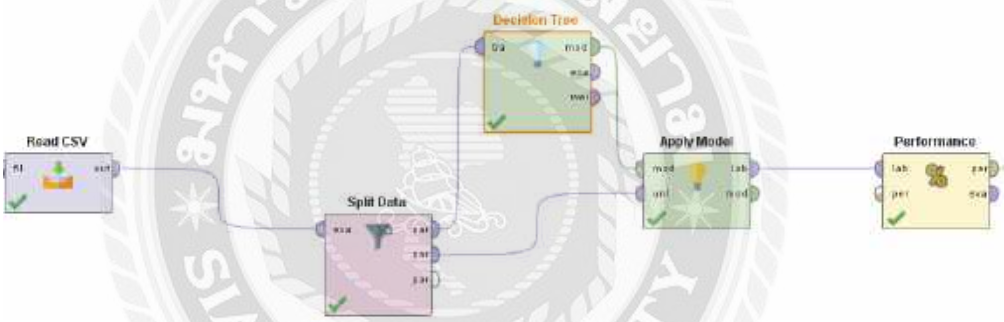
ภาพที่ 3.9 โอเปอร์เรเตอร์ Apply Model

4.5 ทำการเชื่อมต่อชุดข้อมูลสำหรับเชื่อมต่อเข้ากับตัวแบบต้นไม้ตัดสินใจสำหรับการทดสอบเข้ากับ Apply Model

4.6 ขั้นตอนต่อมาทำการเลือกใช้โอเปอร์เรเตอร์ของ Confusion Matrix ที่มีชื่อว่า Performance (Classification) ในการหาผลลัพธ์ความแม่นยำในการทำนายในรูปแบบของตารางแสดงค่าความถูกต้อง

ภาพที่ 3.10 โอเปอร์เรเตอร์ Confusion Matrix ที่มีชื่อว่า Performance (Classification)

4.7 เมื่อทำการเชื่อมต่อโอเปอร์เรเตอร์ ที่มีชื่อว่า Performance (Classification) เข้ากับ Apply Model แล้ว ก็จะได้ขั้นตอนการสร้างตัวแบบของตัวแบบต้นไม้ตัดสินใจสำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ



ภาพที่ 3.11 ตัวแบบต้นไม้ตัดสินใจสำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ

โดยในการทดสอบผลลัพธ์จะทำการวัดค่า Accuracy, Recall และ Precision เนื่องจากค่าเหล่านี้จะช่วยให้ผู้อ่านเข้าใจประสิทธิภาพของแบบจำลองได้อย่างครอบคลุม

Accuracy คือค่าวัดประสิทธิภาพแบบจำลองจำแนกประเภท (classification model) โดยพิจารณาจากตัวอย่างทั้งหมด

Recall คือการวัดความไวของแบบจำลองโดยพิจารณาจากตัวอย่างที่ถูกจำแนกว่าเป็นคลาสที่เป็นบวก ช่วยให้เข้าใจว่าแบบจำลองสามารถตรวจจับตัวอย่างที่เป็นคลาสที่เป็นบวกได้ถูกต้องมากน้อยเพียงใด

Precision คือการวัดความแม่นยำของแบบจำลองโดยพิจารณาจากตัวอย่างที่ถูกจำแนกว่าเป็นคลาสที่เป็นบวก ช่วยให้ผู้อ่านเข้าใจว่าแบบจำลองสามารถจำแนกตัวอย่างว่าเป็นคลาสที่เป็นบวกได้อย่างแม่นยำมากน้อยเพียงใด

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

```
PerformanceVector:
accuracy: 77.78%
ConfusionMatrix:
True:   SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:        146     19      23      20      0
SPARK:  5       38      3       4       1
CAB4 1.9:       11      2       138     2       3
SPC 3.0:        0       0       0       2       0
CAB4 3.0:       0       1       6       0       26
```

ภาพที่ 3.12 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth

ผลของการทำนายค่าความถูกต้องของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 77.78%

```
weighted_mean_recall: 65.69%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:        146     19      23      20      0
SPARK:  5       38      3       4       1
CAB4 1.9:       11      2       138     2       3
SPC 3.0:        0       0       0       2       0
CAB4 3.0:       0       1       6       0       26
```

ภาพที่ 3.13 ค่า Recall Decision Tree โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 65.96%

```
weighted_mean_precision: 82.39%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:        146     19      23      20      0
SPARK:  5       38      3       4       1
CAB4 1.9:       11      2       138     2       3
SPC 3.0:        0       0       0       2       0
CAB4 3.0:       0       1       6       0       26
```

ภาพที่ 3.14 ค่า Precision Decision Tree โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth

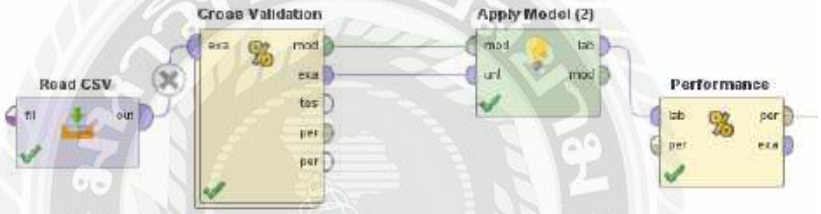
ผลของการทำนายค่า Precision ของของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผล ค่า Precision อยู่ที่ 82.39%

4.8 การกำหนด Cross validation คือวิธีการทดสอบประสิทธิภาพแบบจำลองโดยแบ่งชุดข้อมูลออกเป็นชุดย่อยๆ หลายชุด จากนั้นใช้ชุดย่อยหนึ่งชุดสำหรับทดสอบประสิทธิภาพของแบบจำลอง และชุดย่อยที่เหลือสำหรับฝึกแบบจำลอง โดยในการปรับพารามิเตอร์จะมีการปรับค่าเป็น 2 แบบ ดังนี้ ปรับค่า number of folds เป็น 5 , 10 เพื่อเปรียบเทียบประสิทธิภาพของการทำนาย

ภาพที่ 3.15 การปรับพารามิเตอร์ของ Cross validation

ภาพที่ 3.16 Process Cross validation ของตัวแบบ Decision tree

4.9 โดยภายใน Process Cross validation ของตัวแบบ Decision tree ได้มีการปรับพารามิเตอร์ของ Maximal depth 2 ค่า ดังนี้ 5 , 10 เพื่อกำหนดความลึกของต้นไม้ โดยเปรียบเทียบค่า Maximal Depth ที่ให้ผลการทำนายที่ดีที่สุด



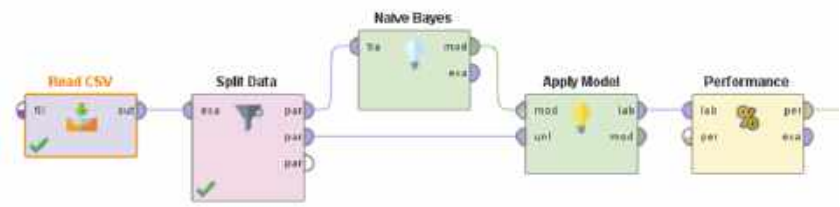
ภาพที่ 3.17 ตัวแบบต้นไม้ตัดสินใจโดยใช้ Cross Validation ในการ Training ข้อมูล

4.10 ทำการเชื่อมต่อ Operator แต่ละอันเพื่อทำการทดสอบประสิทธิภาพการทำนายของตัวแบบ Decision tree โดยเชื่อมต่อเข้ากับ Operator Apply Model (2) และเชื่อมต่อเข้ากับ Operator Performance

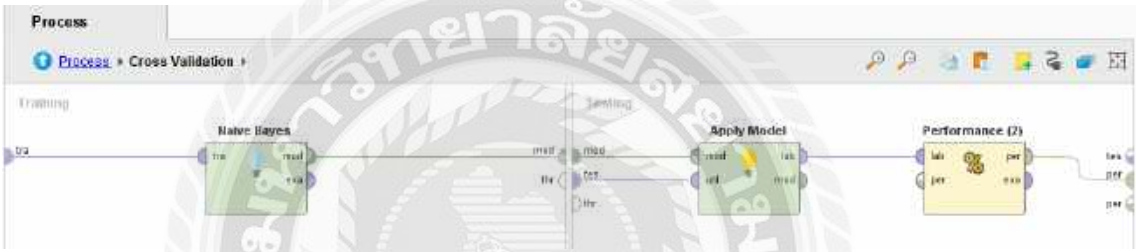
5. การสร้างตัวแบบ Naïve Bayes

ภาพที่ 3.18 สร้างตัวแบบ Naïve Bayes โดยใช้ โอเปอร์เรเตอร์ Naïve Bayes

5.1 เชื่อมต่อโอเปอร์เรเตอร์ มีชื่อว่า Performance (Classification) เข้ากับ Apply Model แล้ว ก็จะได้ขั้นตอนการสร้างตัวแบบของตัวแบบ Naïve Bayes สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ

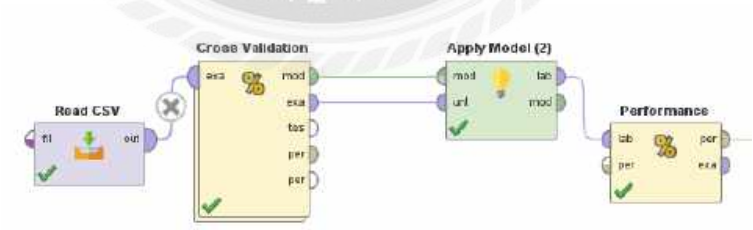


ภาพที่ 3.19 ตัวแบบ Naïve Bayes สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ



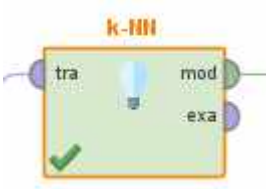
ภาพที่ 3.20 Process Cross validation ของตัวแบบ Naïve Bayes

5.2 โดยภายใน Process Cross validation ของตัวแบบ Naïve Bayes ได้มีการปรับพารามิเตอร์ของ number of folds 2 ค่า ดังนี้ 5 , 10 โดยเปรียบเทียบค่าให้ผลการทำนายที่ดีที่สุด



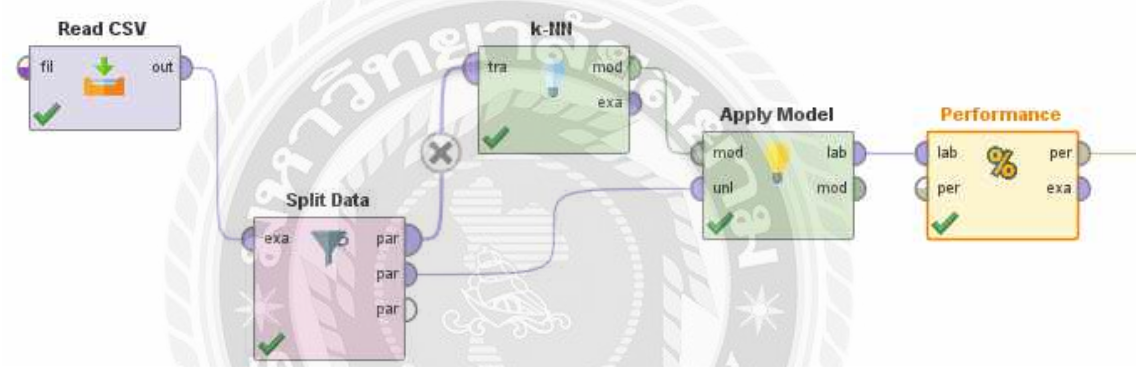
ภาพที่ 3.21 ตัวแบบ Naïve Bayes โดยใช้ Cross Validation ในการ Training ข้อมูล

6. การสร้างตัวแบบ K- Nearest Neighbor



ภาพที่ 3.22 สร้างตัวแบบ K- Nearest Neighbor โดยใช้ โอเปอร์เรเตอร์ K- Nearest Neighbor

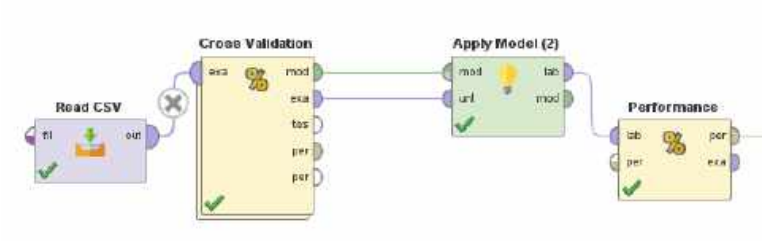
6.1 เชื่อมต่อโอเปอร์เรเตอร์ มีชื่อว่า Performance (Classification) เข้ากับ Apply Model แล้ว ก็จะได้ขั้นตอนการสร้างตัวแบบของตัวแบบ K- Nearest Neighbor สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ



ภาพที่ 3.23 ตัวแบบ K- Nearest Neighbor สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ

ภาพที่ 3.24 Process Cross validation ของตัวแบบ K- Nearest Neighbor

6.2 โดยภายใน Process Cross validation ของตัวแบบ K- Nearest Neighbor ได้มีการปรับพารามิเตอร์ของ number of folds 2 ค่า ดังนี้ 5 , 10 โดยเปรียบเทียบค่าให้ผลการทำนายที่ดีที่สุด



ภาพที่ 3.25 ตัวแบบ K- Nearest Neighbor โดยใช้ Cross Validation ในการ Training ข้อมูล

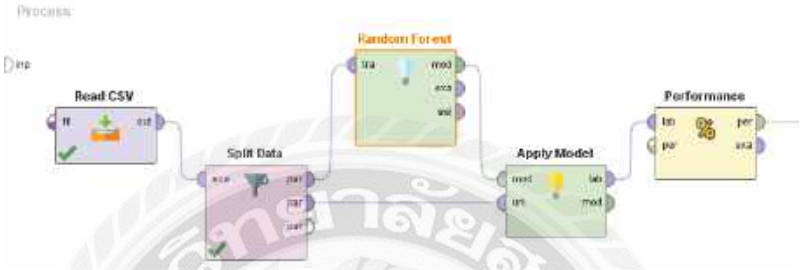
6.3 ทำการเชื่อมต่อ Operator แต่ละอันเพื่อทำการทดสอบประสิทธิภาพการทำนายของตัวแบบ K- Nearest Neighbor โดยเชื่อมต่อเข้ากับ Operator Apply Model (2) และเชื่อมต่อเข้ากับ Operator Performance

7. การสร้างตัวแบบ Random Forest

ภาพที่ 3.26 ตัวแบบการทำนายโดยใช้เทคนิคของ Random Forest

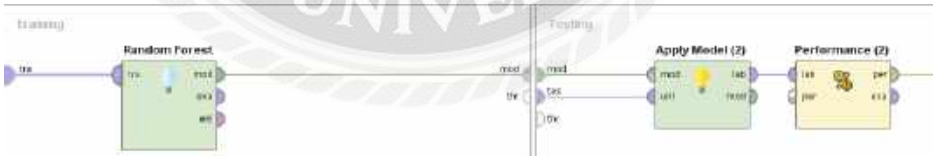
ภาพที่ 3.27 ขั้นตอนการปรับพารามิเตอร์ของตัวแบบ Random Forest

7.1 Criterion ตั้งค่าเป็น gini_index เป็นค่าที่บอกถึงคุณลักษณะที่ควรนำมาใช้งานเป็นคุณลักษณะที่แบ่งกลุ่มของอัลกอริทึม โดยคำนวณให้ได้ค่า gini ที่น้อยที่สุดเป็นโหนดราก Number of trees เพื่อกำหนดจำนวนต้นไม้ที่นำมาสร้างให้เป็น Random Forest โดยตั้งค่าจำนวนป่าเท่ากับ 100 ต้น Maximal Depth เพื่อกำหนดค่าความลึกของต้นไม้แต่ละต้น โดยการเปรียบเทียบค่า Maximal Depth ให้ได้ผลการทำนายที่ดีที่สุดของ Random Forest โดยกำหนดค่า Maximal Depth = 5,10



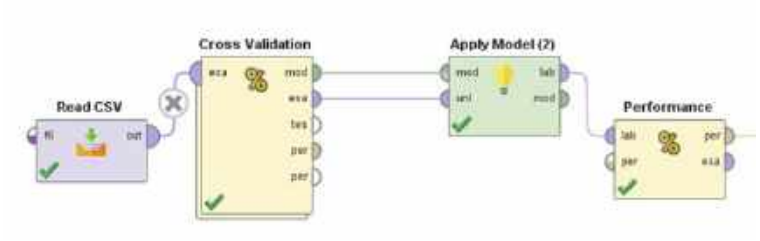
ภาพที่ 3.28 สร้างตัวแบบ Random Forest โดยใช้ โอเปอร์เรเตอร์ Random Forest

7.2 เชื่อมต่อโอเปอร์เรเตอร์ มีชื่อว่า Performance (Classification) เข้ากับ Apply Model แล้ว ก็จะได้ขั้นตอนการสร้างตัวแบบของตัวแบบ Random Forest สำหรับการจำแนกรุ่นของรถกระบะที่ลูกค้าต้องการ



ภาพที่ 3.29 Process Cross validation ของตัวแบบ Random forest

7.3 โดยภายใน Process Cross validation ของตัวแบบ Random forest ได้มีการปรับพารามิเตอร์ของ number of folds 2 ค่า ดังนี้ 5 , 10 โดยเปรียบเทียบค่าให้ผลการทำนายที่ดีที่สุด



ภาพที่ 3.30 ตัวแบบ Random forest โดยใช้ Cross Validation ในการ Training ข้อมูล

7.4 ทำการเชื่อมต่อ Operator แต่ละอันเพื่อทำการทดสอบประสิทธิภาพการทำนายของตัวแบบ Random forest โดยเชื่อมต่อเข้ากับ Operator Apply Model (2) และเชื่อมต่อเข้ากับ Operator Performance

การประเมินผล (Evaluation)

ในการแบ่งข้อมูลมูล 70:30 75:25 80:20 นั้น เป็นการแบ่งข้อมูลออกเป็นชุดฝึกและชุดทดสอบ โดยกำหนดสัดส่วนระหว่างชุดฝึกและชุดทดสอบดังนี้

1. 70:30 หมายถึง ชุดฝึก 70% และชุดทดสอบ 30%

2. 75:25 หมายถึง ชุดฝึก 75% และชุดทดสอบ 25%

3. 80:20 หมายถึง ชุดฝึก 80% และชุดทดสอบ 20%

สำหรับการแบ่งข้อมูลแบบ Cross validation นั้นเป็นการแบ่งข้อมูลออกเป็นหลายชุด โดยแต่ละชุดจะใช้สำหรับทดสอบตัวแบบหนึ่งครั้ง จากนั้นนำผลลัพธ์ที่ได้จากการทดสอบแต่ละชุดมาเฉลี่ยกันเพื่อประเมินประสิทธิภาพของโมเดล

กำหนดค่า number of folds ไว้ 2 ค่า คือ 5 และ 10 หมายถึงการแบ่งข้อมูลออกเป็น 5 ชุด และ 10 ชุด ตามลำดับ การแบ่งข้อมูลแบบ Cross validation จะช่วยให้ได้ผลการวัดที่เชื่อถือได้มากกว่าการแบ่งข้อมูลแบบ Split data

ดังนั้น ในการเปรียบเทียบประสิทธิภาพของตัวแบบในการจำแนกรุ่นของรถยนต์ประเภทรถกระบะที่ลูกค้าต้องการ กรณีศึกษาโชว์รูมรถยนต์ อีซูซุมหานคร สาขามหาชัย โดยใช้อัลกอริทึม

Decision Tree, Naïve Bayes, Random forest และ K-Nearest Neighbor ควรทำการประเมินผลโดยใช้ทั้งการแบ่งข้อมูลแบบ Split data และแบบ Cross validation โดยพิจารณาจากค่าเกณฑ์การประเมินผลทั้งหมด ได้แก่ ความแม่นยำ (Accuracy) ความครอบคลุม (Recall) และค่าความถูกต้อง (Precision)

การนำไปใช้ (Deployment)

การนำตัวแบบไปใช้งานโดยผ่านกระบวนการเปรียบเทียบประสิทธิภาพของตัวแบบที่ได้ค่าความถูกต้องสูงที่สุด หมายถึง การนำตัวแบบที่ผ่านการทดสอบและประเมินผลแล้วมาเปรียบเทียบกัน โดยพิจารณาจากเกณฑ์การประเมินผลต่างๆ เช่น ความแม่นยำ ความครอบคลุม และค่าความถูกต้อง ในกรณีศึกษานี้ ตัวแบบจำแนกรุ่นรถกระบะที่มีค่าความถูกต้องสูงที่สุดจะถูกนำไปใช้ในการจำแนกรุ่นรถกระบะที่ลูกค้าต้องการ โดยนำข้อมูลของลูกค้า เช่น เพศ รายได้ อาชีพ เป็นต้น มาป้อนให้กับตัวแบบ ตัวแบบจะทำการวิเคราะห์ข้อมูลและทำการจำแนกรุ่นรถกระบะที่ลูกค้าต้องการ ข้อมูลจำแนกรุ่นรถกระบะของลูกค้าจะถูกนำไปใช้ในการวางแผนในการสั่งผลิตรถกระบะเพื่อรอพร้อมจำหน่ายให้แก่ลูกค้า โดยที่ลูกค้าไม่ต้องรอรถเป็นเวลานาน ตัวอย่างเช่น หากตัวแบบพบว่าลูกค้าส่วนใหญ่เป็นเพศชาย มีรายได้ปานกลาง ทำงานเป็นพนักงานออฟฟิศ โชว์รูมรถกระบะก็จะสั่งผลิตรถกระบะอเนกประสงค์และรถกระบะขับเคลื่อนสี่ล้อไว้ล่วงหน้า เพื่อให้เพียงพอกับความต้องการของลูกค้ากลุ่มดังกล่าว การนำตัวแบบจำแนกรุ่นรถกระบะไปใช้งานจริง จะช่วยให้ธุรกิจสามารถตอบสนองความต้องการของลูกค้าได้อย่างตรงจุด และสามารถเพิ่มยอดขายได้อย่างมีประสิทธิภาพ

บทที่4

ผลการวิจัย

งานวิจัยนี้ได้นำเทคนิคเหมืองข้อมูล เพื่อสร้างตัวแบบสำหรับการจำแนกรุ่นของรถยนต์ประเภทรถกระบะด้วยเทคนิคเหมืองข้อมูล กรณีศึกษา โชว์รูมอีซูซุ มหานครสาขามหาชัย เพื่อเปรียบเทียบประสิทธิภาพของตัวแบบจำแนกตามร้อยละ จำนวนการฝึกสอนและจำนวนข้อมูลที่แตกต่างกันด้วย เทคนิค Decision Tree , เทคนิค Naïve Bayes , เทคนิค K- Nearest Neighbor และเทคนิค Random Forest

สัดส่วนข้อมูลของการ Train และการ Test ในการประเมินผลประสิทธิภาพตัวแบบ

จำนวนกลุ่มตัวอย่างจำนวน 1500 คน นำมาแบ่งสัดส่วนการ Train และการ Test เพื่อทดสอบตัวแบบตามเทคนิคเหมืองข้อมูลทั้ง 4 เทคนิค คือ Decision Tree, Naïve Bayes, K- Nearest Neighbor , Random Forest โดยแบ่งกลุ่มการฝึกสอน (Training set) ดังสัดส่วนต่อไปนี้ 70%, 75%, และ 80% ตามลำดับ สำหรับการแบ่งข้อมูลแบบ Cross validation นั้นเป็นการแบ่งข้อมูลออกเป็นหลายชุด โดยแต่ละชุดจะใช้สำหรับทดสอบตัวแบบหนึ่งครั้ง จากนั้นนำผลลัพธ์ที่ได้จากการทดสอบแต่ละชุดมาเฉลี่ยกันเพื่อประเมินประสิทธิภาพของโมเดล กำหนดค่า number of folds ไว้ 2 ค่า คือ 5 และ 10 หมายถึงการแบ่งข้อมูลออกเป็น 5 ชุด และ 10 ชุด ตามลำดับ

ตารางที่ 4.1จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามเพศ

เพศ	จำนวน	ร้อยละ
ชาย	507	33.8
หญิง	993	66.2
รวม	1,500	100

ตารางที่ 4.2 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามอาชีพ

อาชีพ	จำนวน	ร้อยละ
รับเหมาก่อสร้าง	110	7.3
พนักงานบริษัท	689	45.9
เกษตกร	102	6.8
ข้าราชการ	34	2.3
ทนายความ	9	0.6
ค้าขาย	355	23.6
ธุรกิจส่วนตัว	201	13.5
รวม	1,500	100

ตารางที่ 4.3 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานเงินเดือน

ฐานเงินเดือน	จำนวน	ร้อยละ
ฐานเงินเดือนสูง	898	59.8
ฐานเงินเดือนปกติ	602	40.2
รวม	1,500	100

ตารางที่ 4.4 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานเงินจอง

เงินจอง	จำนวน	ร้อยละ
ฐานเงินจองสูง	886	59.1
ฐานเงินจองปกติ	614	40.9
รวม	1,500	100

ตารางที่ 4.5 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาของลูกค้า

ที่มาของลูกค้า	จำนวน	ร้อยละ
โทร	39	2.6
เฟสบุ๊ค	610	40.7
ลูกค้าเก่า	106	7.1
ลูกค้าแนะนำ	257	17.1
ลูกค้ามาจากโชว์รูม	488	32.5
รวม	1,500	100

ตารางที่ 4.6 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาเงินดาวน์

เงินดาวน์	จำนวน	ร้อยละ
เงินดาวน์สูง	547	36.5
เงินดาวน์ปกติ	953	63.5
รวม	1,500	100

ตารางที่ 4.7 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาประวัติเครดิต

ประวัติเครดิต	จำนวน	ร้อยละ
ประวัติเครดิตดี	1,105	74
ประวัติเครดิตไม่ดี	395	26
รวม	1,500	100

ตารางที่ 4.8 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาระบบเครื่องยนต์

ระบบเครื่องยนต์	จำนวน	ร้อยละ
เครื่องยนต์ขนาด 1.9	1,173	78.2
เครื่องยนต์ขนาด 3.0	327	21.8
รวม	1,500	100

ตารางที่ 4.9 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาระบบเกียร์

ระบบเกียร์	จำนวน	ร้อยละ
ระบบเกียร์กระปุก	1,228	82
ระบบเกียร์อัตโนมัติ	272	18
รวม	1,500	100

ตารางที่ 4.10 จำนวนกลุ่มตัวอย่างสำหรับการเก็บข้อมูล จำแนกตามฐานที่มาประเภทรถยนต์

ประเภทรถยนต์	จำนวน	ร้อยละ
SPARK	200	13
SPC	346	23
SPC HR	286	19.5
CAB4	158	10.5
CAB4 HR	510	34
รวม	1,500	100

งานวิจัยเรื่อง การจำแนกรุ่นของรถยนต์ประเภทรถกระบะด้วยเทคนิคเหมืองข้อมูล กรณีศึกษาโชว์รูมอีซูซุ มหานครสาขามหาชัย ข้อมูลเพศส่วนใหญ่เป็นเพศ หญิง คิดเป็นร้อยละ 66.2% ข้อมูลอาชีพส่วนใหญ่เป็นอาชีพ พนักงานบริษัท คิดเป็นร้อยละ 45.9% ข้อมูลฐานเงินเดือนส่วนใหญ่เป็นฐานเงินเดือนสูง คิดเป็นร้อยละ 59.8% ข้อมูลฐานเงินจองส่วนใหญ่เป็น เงินจองสูง คิดเป็นร้อยละ

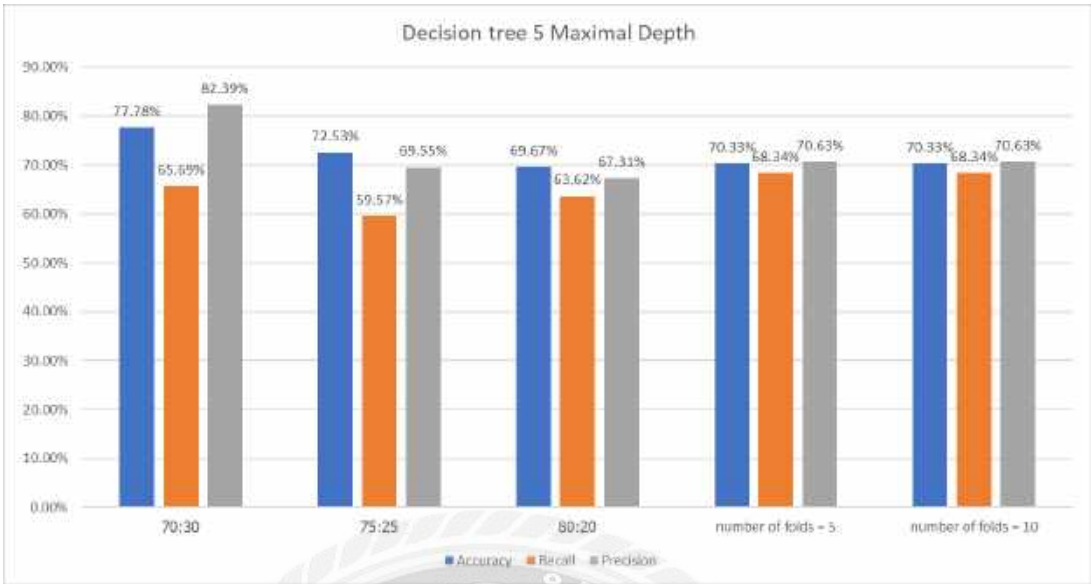
59.1% ข้อมูลที่มาของลูกค้าส่วนใหญ่เป็น เฟสบุ๊ค คิดเป็นร้อยละ 40.7% ข้อมูลเงินดาวน์ส่วนใหญ่เป็น เงินดาวน์ปกติ คิดเป็นร้อยละ 63.5% ข้อมูลประวัติเครดิตส่วนใหญ่เป็น ประวัติเครดิตดี คิดเป็นร้อยละ 74% ข้อมูลระบบเครื่องยนต์ส่วนใหญ่เป็น เครื่องยนต์ขนาด 1.9 คิดเป็นร้อยละ 78.2% ข้อมูลระบบเกียร์ส่วนใหญ่เป็น ระบบเกียร์กระปุก คิดเป็นร้อยละ 82% และข้อมูลประเภทรถยนต์ส่วนใหญ่เป็น CAB4 HR คิดเป็นร้อยละ 34%

ผลการทดสอบด้วยตัวแบบ Decision Tree

ตารางที่ 4.11 ผลการทดสอบด้วยตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 5

Data Training	Accuracy	Recall	Precision
70:30	77.78%	65.69%	82.39%
75:25	72.53%	59.57%	69.55%
80:20	69.67%	63.62%	67.31%
number of folds = 5	70.33%	68.34%	70.63%
number of folds = 10	70.33%	68.34%	70.63%

จากผลการทดสอบตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 5 ทดสอบบนชุดข้อมูลแบบ 70:30 พบว่ามีค่า Accuracy อยู่ที่ 77.78% ค่า Recall ทดสอบบนชุดข้อมูล number of folds = 5 และ number of folds = 10 พบว่ามีค่า Recall อยู่ที่ 68.34% เท่ากัน และค่า Precision ทดสอบบนชุดข้อมูลแบบ 70:30 พบว่ามีค่า Precision อยู่ที่ 82.39%

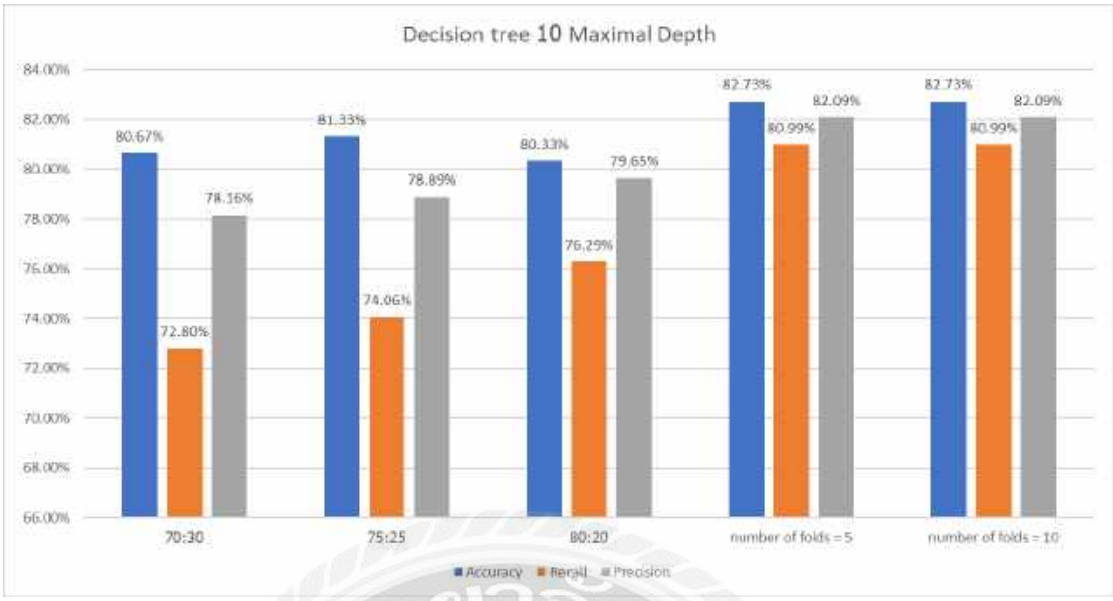


ภาพที่ 4.1 กราฟผลลัพธ์ของตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 5

ตารางที่ 4.12 ผลการทดสอบด้วยตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 10

Data Training	Accuracy	Recall	Precision
70:30	80.67%	72.80%	78.16%
75:25	81.33%	74.06%	78.89%
80:20	80.33%	76.29%	79.65%
number of folds = 5	82.73%	80.99%	82.09%
number of folds = 10	82.73%	80.99%	82.09%

จากผลการทดสอบตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 10 ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Accuracy อยู่ที่ 82.73% ค่า Recall ทดสอบบนชุดข้อมูล number of folds = 5 และ number of folds = 10 พบว่ามีค่า Recall อยู่ที่ 80.99% และค่า Precision ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Precision อยู่ที่ 82.09%



ภาพที่ 4.2 กราฟผลลัพธ์ของตัวแบบ Decision Tree พารามิเตอร์ Maximal Depth ที่ 10

```
Tree
Gear_System = A/T: CAB4 HR {SPC=0, SPARK=0, CAB4=0, CAB4 HR=226, SPC HR=0}
Gear_System = M/T
|   Career = Contractor
|   |   Booking_Money = High
|   |   |   Down_Payment = High
|   |   |   |   Engine_System = 1.9
|   |   |   |   |   Gender = Female
|   |   |   |   |   |   Salary = High
|   |   |   |   |   |   |   Credit_history = Bad: SPARK {SPC=0, SPARK=6, CAB4=0, CAB4 HR=1, SPC HR=3}
|   |   |   |   |   |   |   Credit_history = Good
|   |   |   |   |   |   |   |   Cus_Source = Facebook: SPARK {SPC=0, SPARK=2, CAB4=2, CAB4 HR=0, SPC HR=0}
|   |   |   |   |   |   |   |   Cus_Source = Old_Cus: SPARK {SPC=0, SPARK=6, CAB4=0, CAB4 HR=0, SPC HR=0}
|   |   |   |   |   |   |   |   Cus_Source = Rec_Cus: SPARK {SPC=0, SPARK=2, CAB4=0, CAB4 HR=0, SPC HR=0}
|   |   |   |   |   |   |   |   Cus_Source = Walk: SPC HR {SPC=0, SPARK=0, CAB4=2, CAB4 HR=2, SPC HR=3}
|   |   |   |   |   |   Salary = Normal: CAB4 {SPC=0, SPARK=0, CAB4=1, CAB4 HR=1, SPC HR=0}
|   |   |   |   |   Gender = Male
|   |   |   |   |   |   Credit_history = Bad: CAB4 {SPC=0, SPARK=0, CAB4=3, CAB4 HR=0, SPC HR=0}
|   |   |   |   |   |   Credit_history = Good: SPC {SPC=1, SPARK=0, CAB4=1, CAB4 HR=0, SPC HR=1}
|   |   |   |   Engine_System = 3
|   |   |   |   |   Credit_history = Bad: SPC {SPC=1, SPARK=1, CAB4=0, CAB4 HR=0, SPC HR=0}
|   |   |   |   |   Credit_history = Good: SPARK {SPC=0, SPARK=7, CAB4=0, CAB4 HR=2, SPC HR=0}
|   |   |   Down_Payment = Normal
|   |   |   |   Cus_Source = Rec_Cus: SPC {SPC=2, SPARK=0, CAB4=0, CAB4 HR=0, SPC HR=0}
|   |   |   |   Cus_Source = Walk
|   |   |   |   |   Salary = High
```

ภาพที่ 4.3 สรุปกฎของตัวแบบ Decision Tree

ผลลัพธ์การทำนายของตัวแบบ Decision tree ได้ผลลัพธ์จำนวนทั้งหมด กฎ ดังนี้

กฎที่ 1 คือ ถ้าระบบเกียร์ A/T จะได้รถกระบะประเภท CAB4 HR

กฎที่ 2 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ หญิง ฐานเงินเดือน สูง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPARK

กฎที่ 3 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ หญิง ฐานเงินเดือน สูง ประวัติเครดิต ดี ที่มาของลูกค้า Facebook จะได้รถกระบะประเภท SPARK

กฎที่ 4 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ หญิง ฐานเงินเดือน สูง ประวัติเครดิต ดี ที่มาของลูกค้า ลูกค้าเก่า จะได้รถกระบะประเภท SPARK

กฎที่ 5 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ หญิง ฐานเงินเดือน สูง ประวัติเครดิต ดี ที่มาของลูกค้า ลูกค้าแนะนำ จะได้รถกระบะประเภท SPARK

กฎที่ 6 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ หญิง ฐานเงินเดือน สูง ประวัติเครดิต ดี ที่มาของลูกค้า Walk จะได้รถกระบะประเภท SPC HR

กฎที่ 7 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ หญิง ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท CAB4

กฎที่ 8 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ ชาย ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท CAB4

กฎที่ 9 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 1.9 เพศ ชาย ประวัติเครดิต ดี จะได้รถกระบะประเภท SPC

กฎที่ 10 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 3.0 ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPC

กฎที่ 11 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ สูง ระบบเครื่องยนต์ 3.0 ประวัติเครดิต ดี จะได้รถกระบะประเภท SPARK

กฎที่ 12 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ ปกติ ที่มาของลูกค้า ลูกค้าแนะนำ จะได้รถกระบะประเภท SPC

กฎที่ 13 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ ปกติ ที่มาของลูกค้า Walk ฐานเงินเดือน สูง เพศ หญิง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท CAB4

กฎที่ 14 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ ปกติ ที่มาของลูกค้า Walk ฐานเงินเดือน สูง เพศ หญิง ประวัติเครดิต ดี จะได้รถกระบะประเภท CAB4 HR

กฎที่ 15 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ ปกติ ที่มาของลูกค้า Walk ฐานเงินเดือน สูง เพศ ชาย จะได้รถกระบะประเภท CAB4

กฎที่ 16 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง สูง เงินดาวน์ ปกติ ที่มาของลูกค้า Walk ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท CAB4

กฎที่ 17 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง ต่ำ เงินดาวน์ สูง จะได้รถกระบะประเภท SPC HR

กฎที่ 18 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง ต่ำ เงินดาวน์ ปกติ ฐานเงินเดือน สูง จะได้รถกระบะประเภท SPC

กฎที่ 19 คือ ถ้าระบบเกียร์ M/T อาชีพ รับเหมาก่อสร้าง เงินจอง ต่ำ เงินดาวน์ ปกติ ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท CAB4

กฎที่ 20 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Call เพศ หญิง จะได้รถกระบะประเภท SPC HR

กฎที่ 21 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Call เพศ ชาย จะได้รถกระบะประเภท CAB4

กฎที่ 22 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ สูง จะได้รถกระบะประเภท SPC HR

กฎที่ 23 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ หญิง เงินจอง สูง ประวัติเตรดิต ไม่ดี จะได้รถกระบะประเภท SPC

กฎที่ 24 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ หญิง เงินจอง สูง ประวัติเตรดิต ดี ฐานเงินเดือน สูง จะได้รถกระบะประเภท SPC HR

กฎที่ 25 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ หญิง เงินจอง สูง ประวัติเตรดิต ดี ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 26 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ หญิง เงินจอง ต่ำ ฐานเงินเดือน สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 27 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ หญิง เงินจอง ต่ำ ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC HR

กฎที่ 28 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ ชาย เงินจอง สูง ฐานเงินเดือน สูง ประวัติเตรดิต ไม่ดี จะได้รถกระบะประเภท SPC HR

กฎที่ 29 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ ชาย เงินจอง สูง ฐานเงินเดือน สูง ประวัติเตรดิต ดี จะได้รถกระบะประเภท CAB4

กฎที่ 30 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ ชาย เงินจอง ต่ำ ฐานเงินเดือน สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 31 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ เพศ ชาย เงินจอง ต่ำ ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท CAB4

กฎที่ 32 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า ลูกค้าเก่า เงินจอง สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 33 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า ลูกค้าเก่า เงินจอง ต่ำ จะได้รถกระบะประเภท SPC

กฎที่ 34 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง สูง ฐานเงินเดือน สูง เพศ หญิง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 35 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง สูง ฐานเงินเดือน สูง เพศ ชาย จะได้รถกระบะประเภท SPC

กฎที่ 36 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง สูง ฐานเงินเดือน ปกติ เพศ หญิง ประวัติเตรดิต ไม่ดี จะได้รถกระบะประเภท SPARK

กฎที่ 37 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง สูง ฐานเงินเดือน ปกติ เพศ หญิง ประวัติเตรดิต ดี จะได้รถกระบะประเภท CAB4 HR

กฎที่ 38 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง สูง ฐานเงินเดือน ปกติ เพศ ชาย จะได้รถกระบะประเภท SPC

กฎที่ 39 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง ต่ำ เพศ หญิง จะได้รถกระบะประเภท SPC

กฎที่ 40 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง ต่ำ เพศ ชาย ประวัติเตรดิต ไม่ดี จะได้รถกระบะประเภท SPC HR

กฎที่ 41 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า แนะนำ เงินจอง ต่ำ เพศ ชาย ประวัติเตรดิต ดี จะได้รถกระบะประเภท CAB4

กฎที่ 42 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Walk แนะนำ เงินจอง สูง เพศ หญิง จะได้รถกระบะประเภท SPC HR

กฎที่ 43 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Walk แนะนำ เงินจอง สูง เพศ ชาย ประวัติเตรดิต ไม่ดี จะได้รถกระบะประเภท CAB4 HR

กฎที่ 44 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Walk แนะนำ เงินจอง สูง เพศ ชาย ประวัติเตรดิต ดี จะได้รถกระบะประเภท SPC HR

กฎที่ 45 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Walk แนะนำ เงินจอง ต่ำ เพศ หญิง จะได้รถกระบะประเภท SPC HR

กฎที่ 46 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 1.9 ที่มาของลูกค้า Walk แนะนำ เงินจอง ต่ำ เพศ ชาย จะได้รถกระบะประเภท CAB4

กฎที่ 47 คือ ถ้าระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง สูง ที่มาของลูกค้า Call เพศ หญิง จะได้รถกระบะประเภท SPC

กฎที่ 48 คือ ระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง สูง ที่มาของลูกค้า Call เพศ ชาย จะได้รถกระบะประเภท SPARK

กฎที่ 49 คือ ระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง สูง ที่มาของลูกค้า Facebook จะได้รถกระบะประเภท SPARK

กฎที่ 50 คือ ระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง สูง ที่มาของลูกค้า แนะนำ จะได้รถกระบะประเภท SPARK

กฎที่ 51 คือ ระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง สูง ที่มาของลูกค้า Walk เพศ หญิง จะได้รถกระบะประเภท SPC

กฎที่ 52 คือ ระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง สูง ที่มาของลูกค้า Walk เพศ ชาย จะได้รถกระบะประเภท SPARK

กฎที่ 53 คือ ระบบเกียร์ M/T อาชีพ พนักงานบริษัท ระบบเครื่องยนต์ 3.0 เงินจอง ต่ำ จะได้รถกระบะประเภท SPC

กฎที่ 54 คือ ระบบเกียร์ M/T อาชีพ เกษตกร ที่มาของลูกค้า Facebook เงินจอง สูง ระบบเครื่องยนต์ 1.9 จะได้รถกระบะประเภท CAB4 HR

กฎที่ 55 คือ ระบบเกียร์ M/T อาชีพ เกษตกร ที่มาของลูกค้า Facebook เงินจอง สูง ระบบเครื่องยนต์ 3.0 จะได้รถกระบะประเภท SPARK

กฎที่ 56 คือ ระบบเกียร์ M/T อาชีพ เกษตกร ที่มาของลูกค้า Facebook เงินจอง ต่ำ จะได้รถกระบะประเภท CAB4

กฎที่ 57 คือ ระบบเกียร์ M/T อาชีพ เกษตกร ที่มาของลูกค้า ลูกค้าเก่า จะได้รถกระบะประเภท SPARK

กฎที่ 58 คือ ระบบเกียร์ M/T อาชีพ เกษตกร ที่มาของลูกค้า ลูกค้าแนะนำ จะได้รถกระบะประเภท CAB4 HR

กฎที่ 59 คือ ระบบเกียร์ M/T อาชีพ เกษตกร ที่มาของลูกค้า Walk จะได้รถกระบะประเภท CAB4 HR

กฎที่ 60 คือ ระบบเกียร์ M/T อาชีพ ข้าราชการ จะได้รถกระบะประเภท CAB4 HR

กฎที่ 61 คือ ระบบเกียร์ M/T อาชีพ ทนายความ จะได้รถกระบะประเภท SPC HR

กฎที่ 62 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ หญิง เงินจอง สูง ฐานเงินเดือน สูง เงินดาวน์ สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 63 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ หญิง เงินจอง สูง ฐานเงินเดือน สูง เงินดาวน์ ปกติ จะได้รถกระบะประเภท SPARK

กฎที่ 64 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ หญิง เงินจอง สูง ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 65 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ หญิง เงินจอง ต่ำ จะได้รถกระบะประเภท SPC

กฎที่ 66 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ ชาย เงินจอง สูง เงินดาวน์ สูง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPC

กฎที่ 67 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ ชาย เงินจอง สูง เงินดาวน์ สูง ประวัติเครดิต ดี จะได้รถกระบะประเภท CAB4 HR

กฎที่ 68 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ ชาย เงินจอง สูง เงินดาวน์ ปกติ จะได้รถกระบะประเภท SPARK

กฎที่ 69 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ ชาย เงินจอง ต่ำ เงินดาวน์ สูง จะได้รถกระบะประเภท SPARK

กฎที่ 70 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ ชาย เงินจอง ต่ำ เงินดาวน์ ปกติ ประวัติดครดิต ไม่ดี จะได้รถกระบะประเภท SPC

กฎที่ 71 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Facebook เพศ ชาย เงินจอง ต่ำ เงินดาวน์ ปกติ ประวัติดครดิต ดี จะได้รถกระบะประเภท SPARK

กฎที่ 72 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า ลูกค้าเก่า ฐานเงินเดือน สูง จะได้รถกระบะประเภท SPARK

กฎที่ 73 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า ลูกค้าเก่า ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 74 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 1.9 เงินดาวน์ สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 75 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ ฐานเงินเดือน สูง เพศ หญิง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPC

กฎที่ 76 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ ฐานเงินเดือน สูง เพศ หญิง ประวัติเครดิต ดี จะได้รถกระบะประเภท CAB4 HR

กฎที่ 77 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ ฐานเงินเดือน สูง เพศ ชาย จะได้รถกระบะประเภท CAB4 HR

กฎที่ 78 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 79 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 3.0 เงินดาวน์ สูง จะได้รถกระบะประเภท SPARK

กฎที่ 80 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง สูง ระบบเครื่องยนต์ 3.0 เงินดาวน์ ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 81 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง ต่ำ เงินดาวน์ สูง จะได้รถกระบะประเภท SPC

กฎที่ 82 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง ต่ำ เงินดาวน์ ปกติ ฐานเงินเดือน สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 83 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า แนะนำ เงินจอง ต่ำ เงินดาวน์ ปกติ ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC HR

กฎที่ 84 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 1.9 เงินดาวน์ สูง จะได้รถกระบะประเภท CAB4 HR

กฎที่ 85 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ เพศ หญิง จะได้รถกระบะประเภท SPC

กฎที่ 86 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ เพศ ชาย เงินจอง สูง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPARK

กฎที่ 87 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ เพศ ชาย เงินจอง สูง ประวัติเครดิต ดี จะได้รถกระบะประเภท SPC

กฎที่ 88 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 1.9 เงินดาวน์ ปกติ เพศ ชาย เงินจอง ต่ำ จะได้รถกระบะประเภท CAB4

กฎที่ 89 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 3.0 ฐานเงินเดือน สูง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPARK

กฎที่ 90 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 3.0 ฐานเงินเดือน สูง ประวัติเครดิต ดี จะได้รถกระบะประเภท CAB4 HR

กฎที่ 91 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 3.0 ฐานเงินเดือน ปกติ เพศ หญิง ประวัติเครดิต ไม่ดี จะได้รถกระบะประเภท SPC

กฎที่ 92 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 3.0 ฐานเงินเดือน ปกติ เพศ หญิง ประวัติเครดิต ดี จะได้รถกระบะประเภท SPARK

กฎที่ 93 คือ ระบบเกียร์ M/T อาชีพ ค้าขาย ที่มาของลูกค้า Walk ระบบเครื่องยนต์ 3.0 ฐานเงินเดือน ปกติ เพศ ชาย จะได้รถกระบะประเภท SPARK

กฎที่ 94 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Call จะได้รถกระบะประเภท SPARK

กฎที่ 95 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Facebook เงินดาวน์ สูง เงินจอง สูง ฐานเงินเดือน สูง เพศ หญิง จะได้รถกระบะประเภท CAB4

กฎที่ 96 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Facebook เงินดาวน์ สูง เงินจอง สูง ฐานเงินเดือน สูง เพศ ชาย จะได้รถกระบะประเภท SPARK

กฎที่ 97 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Facebook เงินดาวน์ สูง เงินจอง สูง ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท CAB4

กฎที่ 98 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Facebook เงินดาวน์ สูง เงินจอง ต่ำ จะได้รถกระบะประเภท CAB4 HR

กฎที่ 99 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Facebook เงินดาวน์ ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 100 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า ลูกค้าเก่า จะได้รถกระบะประเภท CAB4

กฎที่ 101 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า ลูกค้าแนะนำ เงินดาวน์ สูง เพศ หญิง จะได้รถกระบะประเภท SPC HR

กฎที่ 102 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า ลูกค้าแนะนำ เงินดาวน์ สูง เพศ ชาย จะได้รถกระบะประเภท CAB4 HR

กฎที่ 103 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า ลูกค้าแนะนำ เงินดาวน์ ปกติ ฐานเงินเดือน สูง จะได้รถกระบะประเภท SPARK

กฎที่ 104 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า ลูกค้าแนะนำ เงินดาวน์ ปกติ ฐานเงินเดือน ปกติ จะได้รถกระบะประเภท SPC

กฎที่ 105 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Walk เพศ หญิง เงินจอง สูง เงินดาวน์ สูง จะได้รถกระบะประเภท SPC HR

กฎที่ 106 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Walk เพศ หญิง เงินจอง สูง เงินดาวน์ ปกติ จะได้รถกระบะประเภท CAB4 HR

กฎที่ 107 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Walk เพศ หญิง เงินจอง ต่ำ จะได้รถกระบะประเภท SPC HR

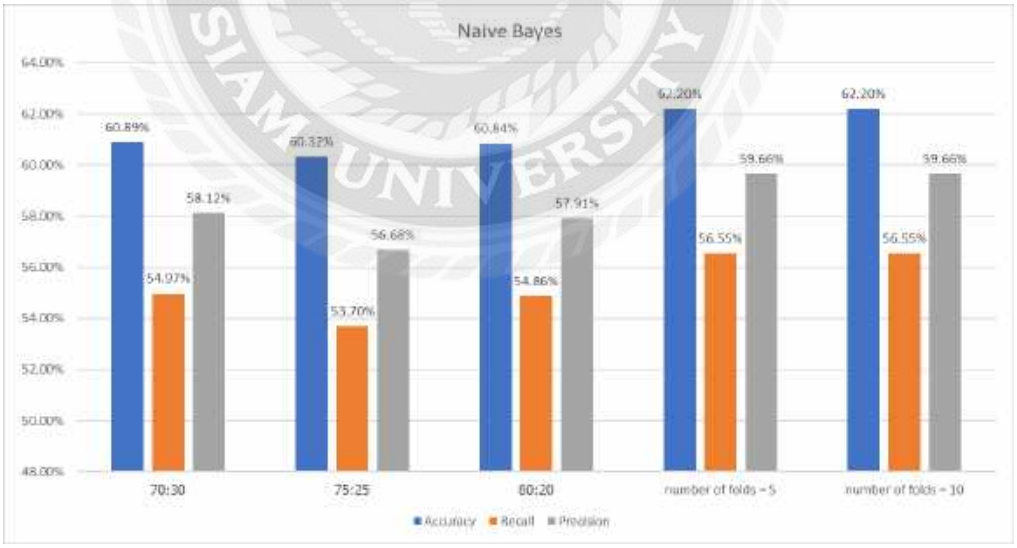
กฎที่ 108 คือ ระบบเกียร์ M/T อาชีพ ธุรกิจส่วนตัว ที่มาของลูกค้า Walk เพศ ชาย จะได้รถกระบะประเภท CAB4

ผลการทดสอบด้วยตัวแบบ Naïve Bayes

ตารางที่ 4.13 ผลการทดสอบด้วยตัวแบบ Naïve Bayes

Data Training	Accuracy	Recall	Precision
70:30	60.89%	54.97%	58.12%
75:25	60.32%	53.70%	56.68%
80:20	60.84%	54.86%	57.91%
number of folds = 5	62.20%	56.55%	59.66%
number of folds = 10	62.20%	56.55%	59.66%

จากผลการทดสอบโมเดล Naïve Bayes ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Accuracy อยู่ที่ 62.20% ค่า Recall ทดสอบบนชุดข้อมูล number of folds = 5 และ number of folds = 10 พบว่ามีค่า Recall อยู่ที่ 56.55% และค่า Precision ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Precision อยู่ที่ 59.66%



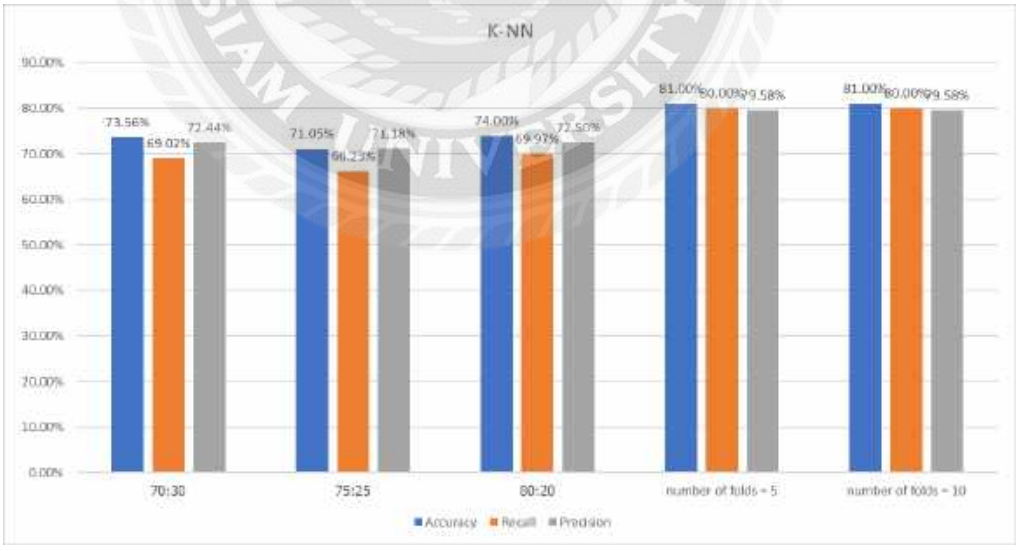
ภาพที่ 4.4 กราฟผลลัพธ์ของตัวแบบ Naïve Bayes

ผลการทดสอบด้วยตัวแบบ K- Nearest Neighbor

ตารางที่ 4.14 ผลการทดสอบด้วยตัวแบบ K- Nearest Neighbor

Data Training	Accuracy	Recall	Precision
70:30	73.56%	69.02%	72.44%
75:25	71.05%	66.23%	71.18%
80:20	74.00%	69.97%	72.50%
number of folds = 5	81.00%	80.00%	79.58%
number of folds = 10	81.00%	80.00%	79.58%

จากผลจากผลการทดสอบตัวแบบ K- Nearest Neighbor ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Accuracy อยู่ที่ 81.00% ค่า Recall ทดสอบบนชุดข้อมูล number of folds = 5 และ number of folds = 10 พบว่ามีค่า Recall อยู่ที่ 80.00%และค่า Precision ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Precision อยู่ที่ 79.58%



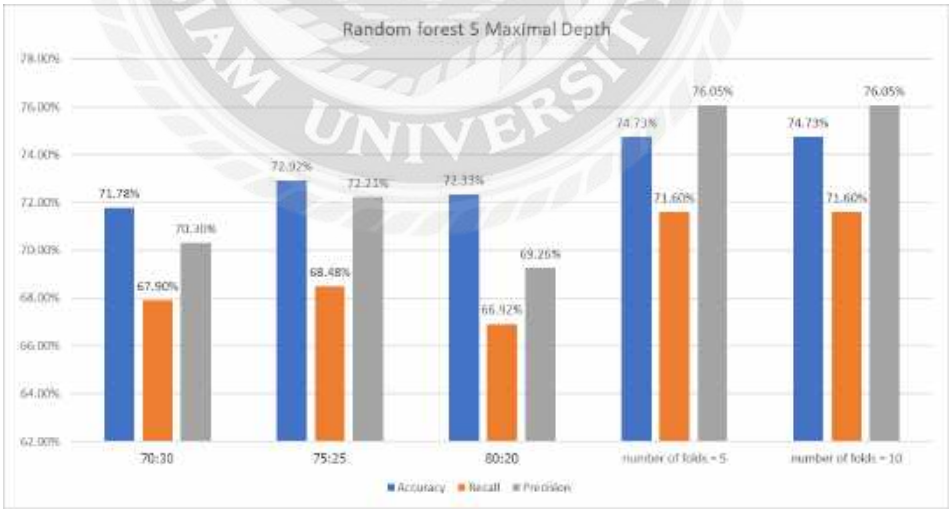
ภาพที่ 4.5 กราฟผลลัพธ์ของตัวแบบ K- Nearest Neighbor

ผลการทดสอบด้วยตัวแบบ Random Forest

ตารางที่ 4.15 ผลการทดสอบด้วยตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 5

Data Training	Accuracy	Recall	Precision
70:30	71.78%	67.90%	70.30%
75:25	72.92%	68.48%	72.21%
80:20	72.33%	66.92%	69.26%
number of folds = 5	74.73%	71.60%	76.05%
number of folds = 10	74.73%	71.60%	76.05%

จากผลการทดสอบโมเดล Random Forest พารามิเตอร์ Maximal Depth ที่ 5 ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Accuracy อยู่ที่ 74.73% ค่า Recall ทดสอบบนชุดข้อมูล number of folds = 5 และ number of folds = 10 พบว่ามีค่า Recall อยู่ที่ 71.60% เท่ากัน และค่า Precision ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Precision อยู่ที่ 76.05%

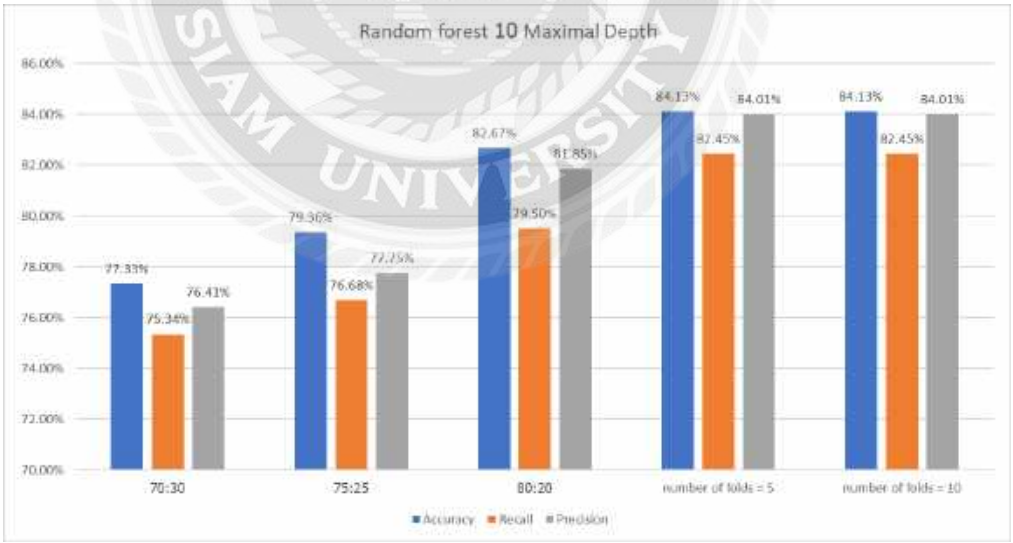


ภาพที่ 4.6 กราฟผลลัพธ์ของตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 5

ตารางที่ 4.16 ผลการทดสอบด้วยตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 10

Data Training	Accuracy	Recall	Precision
70:30	77.33%	75.34%	76.41%
75:25	79.36%	76.68%	77.75%
80:20	82.67%	79.50%	81.85%
number of folds = 5	84.13%	82.45%	84.01%
number of folds = 10	84.13%	82.45%	84.01%

จากผลการทดสอบโมเดล Random Forest พารามิเตอร์ Maximal Depth ที่ 10 ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Accuracy อยู่ที่ 84.13% ค่า Recall ทดสอบบนชุดข้อมูล number of folds = 5 และ number of folds = 10 พบว่ามีค่า Recall อยู่ที่ 82.45% เท่ากัน และค่า Precision ทดสอบบนชุดข้อมูลแบบ number of folds = 5 และ number of folds = 10 พบว่ามีค่า Precision อยู่ที่ 84.01%



ภาพที่ 4.7 กราฟผลลัพธ์ของตัวแบบ Random Forest พารามิเตอร์ Maximal Depth ที่ 10

การทดสอบการแบ่งข้อมูลแบบ Cross validation เป็นวิธีที่มีประสิทธิภาพมากกว่าการแบ่งข้อมูลแบบ Split data เนื่องจากการใช้ข้อมูลทดสอบที่หลากหลายทำให้ตัวแบบสามารถเรียนรู้ความสัมพันธ์ระหว่างข้อมูลได้แม่นยำมากขึ้น ความแตกต่างนี้เกิดจากการที่ข้อมูลที่ใช้ทดสอบประสิทธิภาพของตัวแบบใน Cross validation มีความหลากหลายมากกว่าข้อมูลที่ใช้ใน Split data ทำให้ตัวแบบสามารถจับความหลากหลายนี้ได้ดีขึ้น จากการทดลองพบว่าตัวแบบที่ถูกสอนให้จำแนกรถกระบะ โดยมีค่า Accuracy ที่สูงถึง 84.13% ค่า Recall ที่สูงถึง 82.45% และค่า Precision ที่สูงถึง 84.01% โดยตัวแบบที่มีประสิทธิภาพสูงสุดคือ Random Forest ที่ใช้ Cross validation แบบ5-fold และ 10-fold สาเหตุที่ Random Forest ทำให้ประสิทธิภาพสูงอาจเกิดจากอัลกอริทึมที่สามารถเรียนรู้ความสัมพันธ์ระหว่างข้อมูลได้แม่นยำ และการแบ่งข้อมูลแบบ Cross validation ที่ใช้ทดสอบประสิทธิภาพของตัวแบบ ทำให้ตัวแบบสามารถเรียนรู้ความสัมพันธ์ระหว่างข้อมูลได้หลากหลาย

บทที่5

สรุปผลการวิจัย

งานวิจัยนี้ได้นำเทคนิคเหมืองข้อมูล เพื่อสร้างตัวแบบสำหรับการจำแนกรุ่นของรถยนต์ประเภทรถกระบะ กรณีศึกษา โชว์รูมอีซูซุมหานครสาขามหาชัย และเปรียบเทียบประสิทธิภาพของตัวแบบที่เหมาะสมในการจำแนกข้อมูลโดยใช้เทคนิคการจำแนกข้อมูลของตัวแบบดังนี้ Decision Tree , Naïve Bayes , K- Nearest Neighbor , Random forest

สรุปผลการทดลอง

งานวิจัยนี้สามารถนำเทคนิคเหมืองข้อมูลมาประยุกต์ใช้ในการจำแนกรุ่นของรถยนต์ประเภทรถกระบะได้อย่างมีประสิทธิภาพ ซึ่งสามารถนำไปใช้ประโยชน์เพื่อปรับปรุงกระบวนการขายและการตลาดของธุรกิจรถยนต์ต่อไปได้

จากผลการทดสอบประสิทธิภาพของตัวแบบ ผู้วิจัยได้เลือกใช้อัลกอริทึมที่เหมาะสมในการสร้างตัวแบบสำหรับการจำแนกรุ่นของรถยนต์ประเภทรถกระบะ กลุ่มตัวอย่างข้อมูลที่ใช้ในการวิจัยในครั้งนี้คือ ข้อมูลลูกค้าที่ซื้อรถยนต์อีซูซุ สาขามหาชัย เดือนมกราคม พ.ศ. 2563 ถึง เดือนธันวาคม พ.ศ. 2564 จำนวน 1,500 คน โดยจะนำข้อมูลมาใช้ดังนี้ 1. เพศ 2. อาชีพ 3. ฐานเงินเดือน 4. เงินจองสำหรับทำสัญญาจองรถยนต์ 5. เงินดาวน์ 6. ที่มาของลูกค้า 7. ประวัติเครดิต 8. ระบบเครื่องยนต์ 9. ระบบเกียร์ 10.ประเภทของรถยนต์ และนำข้อมูลที่ได้มาทำการวิเคราะห์ โดยใช้โปรแกรม RapidMiner X ข้อมูลที่ได้มาทำการสร้างตัวแบบในการจำแนกประเภทรถกระบะ จึงได้ Attribute จำนวนทั้งหมด 10 Attributes โดยแบ่งกลุ่มข้อมูลการฝึกสอน (Training set) ดังสัดส่วนต่อไปนี้ 70:30 75:25 , 80:20 ตามลำดับ และได้ใช้การแบ่งสัดส่วนข้อมูลการฝึกสอนด้วยวิธี Cross validation โดยเป็นการแบ่งข้อมูลออกเป็นหลายๆชุดโดยแต่ละชุดจะใช้สำหรับทดสอบตัวแบบหนึ่งครั้ง จากนั้นนำผลลัพธ์ที่ได้จากการทดสอบแต่ละชุดมาเฉลี่ยกันเพื่อประเมินประสิทธิภาพของตัวแบบ โดยกำหนดค่า Number of folds ไว้ 2 ค่า คือ 5 และ 10 ผู้วิจัยได้เลือกใช้อัลกอริทึมสำหรับการทดสอบการสร้างตัวแบบทั้งหมด 4 อัลกอริทึม ได้แก่ 1. อัลกอริทึม Decision Tree 2. อัลกอริทึม Naïve Bayes 3. อัลกอริทึม K-Nearest Neighbor และ 4. อัลกอริทึม Random Forest จากนั้นพิจารณาค่าเกณฑ์สำหรับการประเมินผลทั้งหมดในการทดสอบ ได้แก่ ค่าความแม่นยำ (Accuracy) ค่าความครอบคลุม (Recall)

และค่าความถูกต้อง (Precision) เมื่อได้ค่าเกณฑ์สำหรับการทดสอบแล้ว จากนั้นจึงทำการทดสอบตัวแบบสำหรับการจำแนกประเภทรถกระบะ จึงได้ผลลัพธ์การทดสอบดังตารางต่อไปนี้

ตารางที่ 5.1 ตารางเปรียบเทียบผลการทดสอบตัวแบบที่มีประสิทธิภาพมากที่สุดตามลำดับ

No.	Model	Accuracy	Recall	Precision
1	Random Forest 5-fold , 10-fold	84.13%	82.45%	84.01%
2	Decision Tree 5-fold , 10-fold	82.73%	80.99%	82.09%
3	K- Nearest Neighbor 5-fold , 10-fold	81.00%	80.00%	79.58%
4	Naïve Bayes 5-fold , 10-fold	62.20%	56.55%	59.66%

สรุปผลการทดสอบของตัวแบบที่มีประสิทธิภาพมากที่สุดตามลำดับแต่ละอัลกอริทึม โดยอันดับที่ 1 คือ อัลกอริทึม Random Forest แบ่งข้อมูลแบบ 5-fold , 10-fold ได้ผลค่า Accuracy ที่ 84.13% ค่า Recall ที่ 82.45% และค่า Precision 84.01% อันดับที่ 2 คือ อัลกอริทึม Decision Tree แบ่งข้อมูลแบบ 5-fold , 10-fold ได้ผลค่า Accuracy ที่ 82.73% ค่า Recall ที่ 80.99% และค่า Precision 82.09% อันดับที่ 3 คือ อัลกอริทึม K- Nearest Neighbor แบ่งข้อมูลแบบ 5-fold , 10-fold ได้ผลค่า Accuracy ที่ 81.00% ค่า Recall ที่ 80.00% และค่า Precision 79.58% และอันดับที่ 4 คือ อัลกอริทึม Naïve Bayes ได้ผลค่า Accuracy ที่ 62.20% ค่า Recall ที่ 56.55% และค่า Precision 59.66%

ภาพที่ 5.1 กราฟผลการทดสอบตัวแบบที่มีประสิทธิภาพมากที่สุด

ปัญหาและอุปสรรค

1. ปัญหาด้านเทคนิคของโครงงานนี้คือ เทคนิคเหมืองข้อมูลบางเทคนิคอาจมีความซับซ้อนและต้องใช้ความรู้เฉพาะทางในการนำไปใช้ ตัวอย่างเช่น เทคนิค Decision Tree เป็นเทคนิคที่ได้รับความนิยมในการจำแนกประเภทข้อมูล แต่เทคนิคนี้มีความซับซ้อนในการกำหนดค่าพารามิเตอร์ต่างๆ เช่น ความลึกของต้นไม้ วิธีการแบ่งข้อมูล เป็นต้น ผู้วิจัยที่ไม่มีความรู้เฉพาะทางอาจไม่เข้าใจถึงหลักการและวิธีการใช้เทคนิคต่างๆ ได้อย่างถูกต้อง ซึ่งอาจส่งผลต่อความแม่นยำของตัวแบบ

2. ปัญหาด้านการใช้งานของโครงงานนี้คือ การเปลี่ยนแปลงของตลาด ความต้องการของลูกค้า เป็นต้น ตัวอย่างเช่น ความต้องการลูกค้าอาจเปลี่ยนแปลงได้ตลอดเวลา ผู้วิจัยจึงควรติดตามการเปลี่ยนแปลงของตลาดและความต้องการของลูกค้าอยู่เสมอ เพื่อปรับปรุงตัวแบบให้เหมาะสมกับสถานการณ์ปัจจุบัน

ข้อเสนอแนะ

1. ปรับปรุงตัวแบบและข้อมูลตัวแบบประสิทธิภาพของผลลัพธ์อาจดีขึ้นหากมีการเพิ่มข้อมูลเพิ่มเติมที่เป็นประโยชน์ เช่น ข้อมูลเกี่ยวกับความพึงพอใจของลูกค้า , ประวัติการบำรุงรักษารถยนต์ หรือข้อมูลอื่น ๆ ที่อาจมีผลต่อการเลือกซื้อรถยนต์

2. ควรเลือกศึกษาประสิทธิภาพของตัวแบบจากเทคนิคอื่นๆ เพื่อให้ผลลัพธ์ที่ได้จากการพยากรณ์มีความแม่นยำมากยิ่งขึ้น

บรรณานุกรม

จุฑาภรณ์ สุคตะ, ศรัญญู สวนทอง, วิภาวรรณ บัวทอง, และ สมใจ จิตคำนึงสุข. (2564). การใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์โรคของผู้สูงอายุที่เกิดขึ้นพร้อมกัน. ใน *การประชุมระดับชาติ มหาวิทยาลัยราชภัฏภูเก็ต ครั้งที่ 13*. https://fund.pkru.ac.th/storage/download/611b2485027439000170f1a4?sector=files2021&bucket=publish_paper&ver=0&sk=244a3d7d5ce95266cbb7bf93b9a63636

ทิพย์ธิดา วงศ์พิพันธ์. (2560). *การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ กรณีศึกษากรุงไทยคาร์เร้นท์ แอนด์ ลีส จำกัด (มหาชน)*. (งานค้นคว้าอิสระปริญญาวิทยาศาสตรมหาบัณฑิต). มหาวิทยาลัยธุรกิจบัณฑิตย์. https://libdoc.dpu.ac.th/thesis/147907.pdf

ธนะวัฒน์ วรรณประภา. (2564). *การใช้เทคนิคเหมืองข้อมูลเพื่อคัดกรองบุคคลที่มีแนวโน้มสัมฤทธิผลในการศึกษาระดับปริญญาตรี สาขาวิชาเทคโนโลยีการศึกษา*. (วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต). มหาวิทยาลัยธุรกิจบัณฑิตย์. https://libdoc.dpu.ac.th/thesis/Thanawat.Wan.pdf

นภาพร สุทธินาคสมบัติ. (2558). *การจำแนกข้อมูลเพื่อพยากรณ์ราคาเครื่องสำอางด้วยเทคนิคเหมืองข้อมูล*. (รายงานการวิจัย). มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ . https://drive.google.com/file/d/1cfamWLQeb9vQ1RphpGTrK7hIKLhLiXJ/view?usp

บีสุตา ดาวเรือง, จรัญ แสนราช, และ อนิราช มิ่งขวัญ. (2564). *การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลเพื่อทำนายแขนงวิชาเรียนของนักศึกษาภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศและการจัดการอุตสาหกรรม*. (รายงานการวิจัย). มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ วิทยาเขตปราจีนบุรี. https://shorturl.at/lqEX3

ประเทือง วงษ์ทอง, วาฤทธิ์ กันแก้ว, อภิชชญา ขวัญแก้ว, และ ยุวดี โฉมแดง. (2563). *การประยุกต์เทคนิคเหมืองข้อมูลค้นหาลักษณะนิสัยของผู้ประกอบอาชีพด้านคอมพิวเตอร์*. (รายงานการวิจัย). มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ. https://rb.gy/z2lbxe

ภาภรณ์ เหล่าพิลัย และ จรัญ แสนราช. (2563). *การวิเคราะห์การลาออกกลางคันของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิควิธีการทำเหมืองข้อมูล. วารสารวิทยาศาสตร์ แห่งมหาวิทยาลัยราชภัฏเพชรบุรี, 16*(2), 61-71. http://sciencejournal.pbru.ac.th/phocadownloadpap/2019_2/2019_2pp61-71.pdf

บรรณานุกรม(ต่อ)

วนิดา พงษ์สงวน, ทิพยา ถินสูงเนิน, และ มาโนช ถินสูงเนิน. (2560). *การพัฒนาแบบจำลองปัจจัยที่มีผลต่อการเป็นโรคเบาหวานด้วยเทคนิคต้นไม้ตัดสินใจ*. (รายงานการวิจัย). มหาวิทยาลัยราชภัฏนครราชสีมา. https://sciencetech.nrru.ac.th/mis-science/pages/journal/file/journal_142705c5a5818b6ab3.pdf

สำราญ วานนท์, ธรัช อารีราษฎร์, และ จรัญ แสนราช. (2561). การศึกษาเทคนิคพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรีสาขาคอมพิวเตอร์โดยใช้เทคนิคเหมืองข้อมูล. *วารสารวิชาการ การจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม, 15*(2). 164-171. https://ph02.tci-thaijo.org/index.php/itm-journal/article/view/140249

สุรพงษ์ ไพสิฐพัฒนพงษ์. (2565). *อุตสาหกรรมรถยนต์ในประเทศไทย*. http://service.nso.go.th publish/census/files/2012_Automotive.pdf

สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณ์ทอง. (2560). *การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย*. (รายงานการวิจัย). สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

อาทิตยาพร โรจรัตน์. (2566). *Datamining กับการพยากรณ์ทางธุรกิจ*. http://olarik.it.msu.ac.th/wp-content/uploads/2023/03/data-mining-%.pdf

Ashraf, J., Moustafa, N., & Khurshid, H. (2020). *A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, Solutions and future directions.* Electronics, 9(7), 1177. https://www.mdpi.com/2079-9292/9/7/1177 =sharing

Bishop, C. M. (2006). *Pattern recognition and machine learning*. https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Cutler, A. (2010). *Random forests for regression and classification*. https://docplayer.net/20757057-Random-forests-for-regression-and-classification-adele-cutler-utah-state-university.html

บรรณานุกรม(ต่อ)

Han, J., Kamber, M., & Pei, J. (2001). *Data mining concepts and techniques* (3rd ed.). https://ia800702.us.archive.org/7/items/datamining_201811/DS-book%20u5.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning* (2nd ed.) https://hastie.su.domains/Papers/ESLII.pdf

Müller, A. C. &Guido, S. (2017). *Introduction to machine learning with python: A guide for data scientists*. https://t.ly/WlwcH

Peachapong Poolpol. (2021). *Naïve bayes classification*. https://peachapong-poolpol.medium.com/na%C3%AFve-bayes-classification-cb6cf905505d

Random Forest Algorithm in Machine Learning. (2023). https://www.almabetter.com/bytes/tutorials/data-science/random-forest

Russell, N. (2009). *Artificial intelligence: A modern approach*. https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

Thapanee Boonchob. (2020). *ขั้นตอนการดำเนินงานตามแบบ CRISP-DM*. https://kamboonchob.medium.com-b0913050198f

ภาคผนวก

ภาคผนวก ก

ขั้นตอนการแสดงผลลัพธ์การทดลองอย่างละเอียด

1. ผลการทดลองของตัวแบบ Decision Tree

1.1 การแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 80.67%
ConfusionMatrix:
True:   SPC 1.9 SPARK   CAB4 1.9         SPC 3.0 CAB4 3.0
SPC 1.9:        138     16      13      15      0
SPARK:  5       39      1       1       1
CAB4 1.9:       11      2       149     0       4
SPC 3.0:        8       2       4       12      0
CAB4 3.0:       0       1       3       0       25
```

ภาพที่ ก.1 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth

ผลของการทำนายค่าความถูกต้องของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 80.67%

```
weighted_mean_recall: 72.80%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC 1.9 SPARK   CAB4 1.9         SPC 3.0 CAB4 3.0
SPC 1.9:        138     16      13      15      0
SPARK:  5       39      1       1       1
CAB4 1.9:       11      2       149     0       4
SPC 3.0:        8       2       4       12      0
CAB4 3.0:       0       1       3       0       25
```

ภาพที่ ก.2 ค่า Recall Decision Tree โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 72.80%

```
weighted_mean_precision: 76.18%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:        138     16      13      15      0
SPARK:  5       39      1       1       1
CAB4 1.9:       11      2       149     0       4
SPC 3.0:        8       2       4       12      0
CAB4 3.0:       0       1       3       0       25
```

ภาพที่ ก.3 ค่า Precision Decision Tree โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth

ผลของการทำนายค่า Precision ของของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผล ค่า Precision อยู่ที่ 78.16%

1.2 การแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

```
PerformanceVector:
accuracy: 72.53%
ConfusionMatrix:
True:   SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:        130     30      28      17      0
SPARK:  2       14      0       1       0
CAB4 1.9:       3       3       103     2       3
SPC 3.0:        0       3       1       3       0
CAB4 3.0:       0       0       10      0       22
```

ภาพที่ ก.4 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth

ผลของการทำนายค่าความถูกต้องของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 72.53%

```
weighted_mean_recall: 59.57%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:         130     30      28       17      0
SPARK:  2        14      0       1        0
CAB4 1.9:        3       3       103      2       3
SPC 3.0:         0       3       1        3       0
CAB4 3.0:        0       0       10       0       22
```

ภาพที่ ก.5 ค่า Recall Decision Tree โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 59.57%

```
weighted_mean_precision: 69.55%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC 1.9 SPARK   CAB4 1.9        SPC 3.0 CAB4 3.0
SPC 1.9:         130     30      28       17      0
SPARK:  2        14      0       1        0
CAB4 1.9:        3       3       103      2       3
SPC 3.0:         0       3       1        3       0
CAB4 3.0:        0       0       10       0       22
```

ภาพที่ ก.6 ค่า Precision Decision Tree โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth

ผลของการทำนายค่า Precision ของของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผล ค่า Precision อยู่ที่ 69.55%

1.3 การแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

PerformanceVector:
accuracy: 81.33%
ConfusionMatrix:

True:	SPC 1.9	SPARK	CAB4 1.9	SPC 3.0	CAB4 3.0
SPC 1.9:	121	15	14	11	0
SPARK: 2	32	1	0	0	
CAB4 1.9:	5	2	120	0	5
SPC 3.0:	7	1	4	12	0
CAB4 3.0:	0	0	3	0	20

ภาพที่ ก.7 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth

ผลของการทำนายค่าความถูกต้องของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 81.33%

weighted_mean_recall: 74.06%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC 1.9	SPARK	CAB4 1.9	SPC 3.0	CAB4 3.0
SPC 1.9:	121	15	14	11	0
SPARK: 2	32	1	0	0	
CAB4 1.9:	5	2	120	0	5
SPC 3.0:	7	1	4	12	0
CAB4 3.0:	0	0	3	0	20

ภาพที่ ก.8 ค่า Recall Decision Tree โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 74.06%

weighted_mean_precision: 78.89%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC 1.9	SPARK	CAB4 1.9	SPC 3.0	CAB4 3.0
SPC 1.9:	121	15	14	11	0
SPARK: 2	32	1	0	0	
CAB4 1.9:	5	2	120	0	5
SPC 3.0:	7	1	4	12	0
CAB4 3.0:	0	0	3	0	20

ภาพที่ ก.9 ค่า Precision Decision Tree โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth

ผลของการทำนายค่า Precision ของของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผล ค่า Precision อยู่ที่ 78.89%

1.4 การแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

PerformanceVector:
accuracy: 69.67%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	47	8	6	6	4	
SPARK:	4	19	0	1	1	
CAB4:	0	7	11	3	1	
CAB4 HR:		4	5	1	82	1
SPC HR:	14	1	14	10	50	

ภาพที่ ก.10 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่าความถูกต้องของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 69.67%

```
weighted_mean_recall: 63.62%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     47      8       6       6       4
SPARK:   4       19      0       1       1
CAB4:    0       7       11      3       1
CAB4 HR:         4       5       1       82      1
SPC HR: 14       1       14      10      50
```

ภาพที่ ก.11 ค่า Recall Decision Tree โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 63.62%

```
weighted_mean_precision: 67.31%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     47      8       6       6       4
SPARK:   4       19      0       1       1
CAB4:    0       7       11      3       1
CAB4 HR:         4       5       1       82      1
SPC HR: 14       1       14      10      50
```

ภาพที่ ก.12 ค่า Precision Decision Tree โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่า Precision ของของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผล ค่า Precision อยู่ที่ 67.31%

1.5 การแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 80.33%
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    54      2       1       0       6
SPARK:  1       24      0       0       1
CAB4:   1       6       23      4       3
CAB4 HR:        6       7       3       96      3
SPC HR: 7       1       5       2       44
```

ภาพที่ ก.13 ค่า Accuracy Decision Tree โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth

ผลของการทำนายค่าความถูกต้องของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 80.33%

```
weighted_mean_recall: 76.29%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    54      2       1       0       6
SPARK:  1       24      0       0       1
CAB4:   1       6       23      4       3
CAB4 HR:        6       7       3       96      3
SPC HR: 7       1       5       2       44
```

ภาพที่ ก.14 ค่า Recall Decision Tree โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 76.29%

```
weighted_mean_precision: 79.65%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     54      2       1       0       6
SPARK:   1       24      0       0       1
CAB4:    1       6       23      4       3
CAB4 HR:         6       7       3       96      3
SPC HR: 7        1       5       2       44
```

ภาพที่ ก.15 ค่า Precision Decision Tree โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth

ผลของการทำนายค่า Precision ของของตัวแบบ Decision Tree โดยการแบ่งข้อมูลแบบ 80:20 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผล ค่า Precision อยู่ที่ 79.65%

1.6 การแบ่งข้อมูลแบบ Cross Validation โดยปรับพารามิเตอร์ number of folds ที่ 5 และ ปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Decision Tree

```
PerformanceVector:
accuracy: 70.33%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     217     18      27      41      21
SPARK:   30      158     9       23      8
CAB4:    0       8       65      16      2
CAB4 HR:         18      9       6       366     6
SPC HR: 81       7       51      64      249
```

ภาพที่ ก.16 ค่า Accuracy Decision tree โดยใช้ number of folds 5 และค่า Maximal Depth 5

ผลของการทำนายค่า Accuracy ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 70.33%

```
weighted_mean_recall: 68.34%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     217     18      27      41      21
SPARK:   30      158     9       23      8
CAB4:    0       8       65      16      2
CAB4 HR:         18      9       6       366     6
SPC HR: 81       7       51      64      249
```

ภาพที่ ก.17 ค่า Recall โดยใช้ number of folds 5 และค่า Maximal Depth 5

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 68.34%

```
weighted_mean_precision: 70.63%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     217     18      27      41      21
SPARK:   30      158     9       23      8
CAB4:    0       8       65      16      2
CAB4 HR:         18      9       6       366     6
SPC HR: 81       7       51      64      249
```

ภาพที่ ก.18 ค่า Precision โดยใช้ number of folds 5 และค่า Maximal Depth 5

ผลของการทำนายค่า Precision ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Precision อยู่ที่ 70.63%

1.7 ปรับพารามิเตอร์ number of folds ที่ 5 และ ปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 82.73%
ConfusionMatrix:
True:    SPC     SPARK    CAB4     CAB4 HR SPC HR
SPC:     265     12       10       17       11
SPARK:   7       163      2        7        4
CAB4:    5       5        114      15       9
CAB4 HR:         19       15       5        454      17
SPC HR: 50       5        27       17       245
```

ภาพที่ ก.19 ค่า Accuracy Decision tree โดยใช้ number of folds 5 และค่า Maximal Depth 10

ผลของการทำนายค่า Accuracy ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 82.73%

```
weighted_mean_recall: 80.99%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK    CAB4     CAB4 HR SPC HR
SPC:     265     12       10       17       11
SPARK:   7       163      2        7        4
CAB4:    5       5        114      15       9
CAB4 HR:         19       15       5        454      17
SPC HR: 50       5        27       17       245
```

ภาพที่ ก.20 ค่า Recall โดยใช้ number of folds 5 และค่า Maximal Depth 10

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 80.99%

```
weighted_mean_precision: 82.09%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     265     12      10      17      11
SPARK:   7       163     2       7       4
CAB4:    5       5       114     15      9
CAB4 HR:         19      15      5       454     17
SPC HR: 50       5       27      17      245
```

ภาพที่ ก.21 ค่า Precision โดยใช้ number of folds 5 และค่า Maximal Depth 10

ผลของการทำนายค่า Precision ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Precision อยู่ที่ 82.09%

1.8 ปรับพารามิเตอร์ number of folds ที่ 10 และ ปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

```
PerformanceVector:
accuracy: 70.33%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     217     18      27      41      21
SPARK:   30      158     9       23      8
CAB4:    0       8       65      16      2
CAB4 HR:         18      9       6       366     6
SPC HR: 81       7       51      64      249
```

ภาพที่ ก.22 ค่า Accuracy Decision tree โดยใช้ number of folds 10 และค่า Maximal Depth 5

ผลของการทำนายค่า Accuracy ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 10 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 70.33%

```
weighted_mean_recall: 68.34%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     217     18      27      41      21
SPARK:   30      158     9       23      8
CAB4:    0       8       65      16      2
CAB4 HR:         18      9       6       366     6
SPC HR: 81       7       51      64      249
```

ภาพที่ ก.23 ค่า Recall โดยใช้ number of folds 10 และค่า Maximal Depth 5

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 10 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 68.34%

```
weighted_mean_precision: 70.63%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     217     18      27      41      21
SPARK:   30      158     9       23      8
CAB4:    0       8       65      16      2
CAB4 HR:         18      9       6       366     6
SPC HR: 81       7       51      64      249
```

ภาพที่ ก.24 ค่า Precision โดยใช้ number of folds 10 และค่า Maximal Depth 5

ผลของการทำนายค่า Precision ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 10 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth ได้ผลค่า Precision อยู่ที่ 70.63%

1.9 ปรับพารามิเตอร์ number of folds ที่ 10 และ ปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

PerformanceVector:
accuracy: 82.73%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	265	12	10	17	11	
SPARK:	7	163	2	7	4	
CAB4:	5	5	114	15	9	
CAB4 HR:		19	15	5	454	17
SPC HR:	50	5	27	17	245	

ภาพที่ ก.25 ค่า Accuracy Decision tree โดยใช้ number of folds 10 และค่า Maximal Depth 10

ผลของการทำนายค่า Accuracy ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 82.73%

weighted_mean_recall: 80.99%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	265	12	10	17	11	
SPARK:	7	163	2	7	4	
CAB4:	5	5	114	15	9	
CAB4 HR:		19	15	5	454	17
SPC HR:	50	5	27	17	245	

ภาพที่ ก.26 ค่า Recall โดยใช้ number of folds 10 และค่า Maximal Depth 10

ผลของการทำนายค่า Recall ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 80.99%

weighted_mean_precision: 82.09%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	265	12	10	17	11	
SPARK:	7	163	2	7	4	
CAB4:	5	5	114	15	9	
CAB4 HR:		19	15	5	454	17
SPC HR:	50	5	27	17	245	

ภาพที่ ก.27 ค่า Precision โดยใช้ number of folds 10 และค่า Maximal Depth 10

ผลของการทำนายค่า Precision ของตัวแบบ Decision Tree โดยใช้ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth ได้ผลค่า Precision อยู่ที่ 82.09%

2. ผลการทดลองของตัวแบบ Naïve Bayes

2.1 การแบ่งข้อมูลแบบ 70:30

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 70:30

PerformanceVector:
accuracy: 60.89%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	59	12	16	16	19	
SPARK:	13	37	0	7	0	
CAB4:	4	2	12	1	6	
CAB4 HR:		12	8	10	122	17
SPC HR:	16	1	9	7	44	

ภาพที่ ก.28 ค่า Accuracy Naïve Bayes โดยแบ่งข้อมูล 70:30

ผลของการทำนายค่าความถูกต้องของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 70:30 ได้ผลค่าความถูกต้องอยู่ที่ 60.89%

```
weighted_mean_recall: 54.97%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     59      12      16      16      19
SPARK:   13      37      0       7       0
CAB4:    4       2       12      1       6
CAB4 HR:         12      8       10      122     17
SPC HR: 16       1       9       7       44
```

ภาพที่ ก.29 ค่า Recall Naïve Bayes โดยแบ่งข้อมูล 70:30

ผลของการทำนายค่า Recall ของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 70:30 ได้ผลค่า Recall อยู่ที่ 54.97%

```
weighted_mean_precision: 58.12%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     59      12      16      16      19
SPARK:   13      37      0       7       0
CAB4:    4       2       12      1       6
CAB4 HR:         12      8       10      122     17
SPC HR: 16       1       9       7       44
```

ภาพที่ ก.30 ค่า Precision Naïve Bayes โดยแบ่งข้อมูล 70:30

ผลของการทำนายค่า Precision ของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 70:30 ได้ผลค่า Precision อยู่ที่ 58.12%

2.2 การแบ่งข้อมูลแบบ 75:25

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 75:25

```
PerformanceVector:
accuracy: 60.32%
ConfusionMatrix:
True:    SPC     SPARK    CAB4    CAB4 HR SPC HR
SPC:     49      11       15      13      16
SPARK:   10      28       0       5       0
CAB4:    5       2        9       1       6
CAB4 HR:         10       8       9       102     12
SPC HR: 12       1        6       6       37
```

ภาพที่ ก.31 ค่า ค่า Accuracy Naïve Bayes โดยแบ่งข้อมูล 75:25

ผลของการทำนายค่าความถูกต้องของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 75:25 ได้ผลค่าความถูกต้องอยู่ที่ 60.32%

```
weighted_mean_recall: 53.70%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK    CAB4    CAB4 HR SPC HR
SPC:     49      11       15      13      16
SPARK:   10      28       0       5       0
CAB4:    5       2        9       1       6
CAB4 HR:         10       8       9       102     12
SPC HR: 12       1        6       6       37
```

ภาพที่ ก.32 ค่า Recall Naïve Bayes โดยแบ่งข้อมูล 75:25

ผลของการทำนายค่า Recall ของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 75:25 ได้ผลค่า Recall อยู่ที่ 53.70%

```
weighted_mean_precision: 56.68%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     49      11      15      13      16
SPARK:   10      28      0       5       0
CAB4:    5       2       9       1       6
CAB4 HR:         10      8       9       102     12
SPC HR: 12       1       6       6       37
```

ภาพที่ ก.33 ค่า Precision Naïve Bayes โดยแบ่งข้อมูล 75:25

ผลของการทำนายค่า Precision ของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 75:25 ได้ผลค่า Precision อยู่ที่ 56.68%

2.3 การแบ่งข้อมูลแบบ 80:20

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 80:20

```
PerformanceVector:
accuracy: 60.84%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     56      11      18      16      21
SPARK:   12      35      0       7       0
CAB4:    5       2       11      1       6
CAB4 HR:         11      8       11      116     12
SPC HR: 15       1       5       6       43
```

ภาพที่ ก.34 ค่า ค่า Accuracy Naïve Bayes โดยแบ่งข้อมูล 80:20

ผลของการทำนายค่าความถูกต้องของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 80:20 ได้ผลค่าความถูกต้องอยู่ที่ 60.84%

```
weighted_mean_recall: 54.86%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     56      11      18      16      21
SPARK:   12      35      0       7       0
CAB4:    5       2       11      1       6
CAB4 HR:         11      8       11      116     12
SPC HR: 15       1       5       6       43
```

ภาพที่ ก.35 ค่า Recall Naïve Bayes โดยแบ่งข้อมูล 80:20

ผลของการทำนายค่า Recall ของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 80:20 ได้ผลค่า Recall อยู่ที่ 54.86%

```
weighted_mean_precision: 57.91%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     56      11      18      16      21
SPARK:   12      35      0       7       0
CAB4:    5       2       11      1       6
CAB4 HR:         11      8       11      116     12
SPC HR: 15       1       5       6       43
```

ภาพที่ ก.36 ค่า Precision Naïve Bayes โดยแบ่งข้อมูล 80:20

ผลของการทำนายค่า Precision ของตัวแบบ Naïve Bayes โดยการแบ่งข้อมูลแบบ 80:20 ได้ผลค่า Precision อยู่ที่ 57.91%

2.4 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 5

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Naïve Bayes

```
PerformanceVector:
accuracy: 62.20%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     206     43      46      44      74
SPARK:   29      121     0       27      0
CAB4:    27      3       39      5       7
CAB4 HR:         37      27      31      392     30
SPC HR: 47       6       42      42      175
```

ภาพที่ ก.37 ค่า Accuracy ของตัวแบบ Naïve Bayes ปรับพารามิเตอร์ number of folds ที่ 5

ผลของการทำนายค่า Accuracy ของตัวแบบ Naïve Bayes โดยใช้ number of folds ที่ 5 ได้ผลค่าความถูกต้องอยู่ที่ 62.20%

```
weighted_mean_recall: 56.55%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     206     43      46      44      74
SPARK:   29      121     0       27      0
CAB4:    27      3       39      5       7
CAB4 HR:         37      27      31      392     30
SPC HR: 47       6       42      42      175
```

ภาพที่ ก.38 ค่า Recall ของตัวแบบ Naïve Bayes ปรับพารามิเตอร์ number of folds ที่ 5

ผลของการทำนายค่า Recall ของตัวแบบ Naïve Bayes โดยใช้ number of folds ที่ 5 ได้ผลค่า Recall อยู่ที่ 56.55%

weighted_mean_precision: 59.66%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	206	43	46	44	74	
SPARK:	29	121	0	27	0	
CAB4:	27	3	39	5	7	
CAB4 HR:		37	27	31	392	30
SPC HR:	47	6	42	42	175	

ภาพที่ ก.39 ค่า Precision ของตัวแบบ Naïve Bayes ปรับพารามิเตอร์ number of folds ที่ 5

ผลของการทำนายค่า Precision ของตัวแบบ Naïve Bayes โดยใช้ number of folds ที่ 5 ได้ผลค่า Precision อยู่ที่ 59.66%

2.5 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 10

PerformanceVector:
accuracy: 62.20%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	206	43	46	44	74	
SPARK:	29	121	0	27	0	
CAB4:	27	3	39	5	7	
CAB4 HR:		37	27	31	392	30
SPC HR:	47	6	42	42	175	

ภาพที่ ก.40 ค่า Accuracy ของตัวแบบ Naïve Bayes ปรับพารามิเตอร์ number of folds ที่ 10

ผลของการทำนายค่า Accuracy ของตัวแบบ Naïve Bayes โดยใช้ number of folds ที่ 10 ได้ผลค่าความถูกต้องอยู่ที่ 62.20%

weighted_mean_recall: 56.55%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	206	43	46	44	74	
SPARK:	29	121	0	27	0	
CAB4:	27	3	39	5	7	
CAB4 HR:		37	27	31	392	30
SPC HR:	47	6	42	42	175	

ภาพที่ ก.41 ค่า Recall ของตัวแบบ Naïve Bayes ปรับพารามิเตอร์ number of folds ที่ 10

ผลของการทำนายค่า Recall ของตัวแบบ Naïve Bayes โดยใช้ number of folds ที่ 10 ได้ผลค่า Recall อยู่ที่ 56.55%

weighted_mean_precision: 59.66%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	206	43	46	44	74	
SPARK:	29	121	0	27	0	
CAB4:	27	3	39	5	7	
CAB4 HR:		37	27	31	392	30
SPC HR:	47	6	42	42	175	

ภาพที่ ก.42 ค่า Precision ของตัวแบบ Naïve Bayes ปรับพารามิเตอร์ number of folds ที่ 10

ผลของการทำนายค่า Precision ของตัวแบบ Naïve Bayes โดยใช้ number of folds ที่ 10 ได้ผลค่าค่า Precision อยู่ที่ 59.66%

3. ผลการทดลองของตัวแบบ K- Nearest Neighbor

3.1 โดยการแบ่งข้อมูลแบบ 70:30

ผลการทดลอง ค่า Accuracy , ค่า Recall และค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30

PerformanceVector:
accuracy: 73.56%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	70	2	11	7	8	
SPARK:	5	45	2	5	0	
CAB4:	1	3	19	2	3	
CAB4 HR:		8	8	8	131	9
SPC HR:	20	2	7	8	66	

ภาพที่ ก.43 ค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30

ผลของการทำนายค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30 ได้ผลค่าความถูกต้องอยู่ที่ 73.56%

```
weighted_mean_recall: 69.02%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     70      2       11      7       8
SPARK:   5       45      2       5       0
CAB4:    1       3       19      2       3
CAB4 HR:         8       8       8       131     9
SPC HR: 20       2       7       8       66
```

ภาพที่ ก.44 ค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30

ผลของการทำนายค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30 ได้ผลค่า Recall อยู่ที่ 69.02%

```
weighted_mean_precision: 72.44%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     70      2       11      7       8
SPARK:   5       45      2       5       0
CAB4:    1       3       19      2       3
CAB4 HR:         8       8       8       131     9
SPC HR: 20       2       7       8       66
```

ภาพที่ ก.45 ค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30

ผลของการทำนายค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 70:30 ได้ผลค่า Precision อยู่ที่ 72.44%

3.2 การแบ่งข้อมูลแบบ 75:25

ผลการทดลอง ค่า Accuracy , ค่า Recall และค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25

```
PerformanceVector:
accuracy: 71.05%
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    60      2       6       6       6
SPARK:  3       35      2       5       0
CAB4:   1       3       17      2       3
CAB4 HR:        13      10      8       109     18
SPC HR: 9       0       6       5       44
```

ภาพที่ ก.46 ค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25

ผลของการทำนายค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25 ได้ผลค่าความถูกต้องอยู่ที่ 71.05%

```
weighted_mean_recall: 66.23%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    60      2       6       6       6
SPARK:  3       35      2       5       0
CAB4:   1       3       17      2       3
CAB4 HR:        13      10      8       109     18
SPC HR: 9       0       6       5       44
```

ภาพที่ ก.47 ค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25

ผลของการทำนายค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25 ได้ผลค่า Recall อยู่ที่ 66.23%

```
weighted_mean_precision: 71.18%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    60      2       6       6       6
SPARK:  3       35      2       5       0
CAB4:   1       3       17      2       3
CAB4 HR:        13      10      8       109     18
SPC HR: 9       0       6       5       44
```

ภาพที่ ก.48 ค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25

ผลของการทำนายค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 75:25 ได้ผลค่า Precision อยู่ที่ 71.18%

3.3 การแบ่งข้อมูลแบบ 80:20

ผลการทดลอง ค่า Accuracy , ค่า Recall และค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20

```
PerformanceVector:
accuracy: 74.00%
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    50      1       5       4       6
SPARK:  2       28      0       4       0
CAB4:   1       3       15      3       3
CAB4 HR:        7       7       5       85      4
SPC HR: 9       1       7       6       44
```

ภาพที่ ก.49 ค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20

ผลของการทำนายค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20 ได้ผลค่าความถูกต้องอยู่ที่ 74.00%

```
weighted_mean_recall: 69.97%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    50      1       5       4       6
SPARK:  2       28      0       4       0
CAB4:   1       3       15      3       3
CAB4 HR:        7       7       5       85      4
SPC HR: 9       1       7       6       44
```

ภาพที่ ก.50 ค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20

ผลของการทำนายค่า Recall ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20 ได้ผลค่า Recall อยู่ที่ 69.97%

```
weighted_mean_precision: 72.50%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     50      1       5       4       6
SPARK:   2       28      0       4       0
CAB4:    1       3       15      3       3
CAB4 HR:         7       7       5       85      4
SPC HR: 9        1       7       6       44
```

ภาพที่ ก.51 ค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20

ผลของการทำนายค่า Precision ของตัวแบบ K- Nearest Neighbor โดยการแบ่งข้อมูลแบบ 80:20 ได้ผลค่า Precision อยู่ที่ 72.50%

3.4 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 5

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ K- Nearest Neighbor

```
PerformanceVector:
accuracy: 81.00%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     251     10      10      17      19
SPARK:   14      166     2       16      0
CAB4:    5       6       117     19      15
CAB4 HR:         27      10      5       441     12
SPC HR: 49       8       24      17      240
```

ภาพที่ ก.52 ค่า Accuracy K- Nearest Neighbor ปรับพารามิเตอร์ number of folds ที่ 5

ผลของการทำนายค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยใช้ number of folds ที่ 5 ได้ผลค่าความถูกต้องอยู่ที่ 81.00%

weighted_mean_recall: 80.00%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	251	10	10	17	19	
SPARK:	14	166	2	16	0	
CAB4:	5	6	117	19	15	
CAB4 HR:		27	10	5	441	12
SPC HR:	49	8	24	17	240	

ภาพที่ ก.53 ค่า Recall K- Nearest Neighbor ปรับพารามิเตอร์ number of folds ที่ 5

ผลของการทำนายค่า Recall ของตัวแบบ K- Nearest Neighbor โดยใช้ number of folds ที่ 5 ได้ผลค่าค่า Recall อยู่ที่ 80.00%

weighted_mean_precision: 79.58%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	251	10	10	17	19	
SPARK:	14	166	2	16	0	
CAB4:	5	6	117	19	15	
CAB4 HR:		27	10	5	441	12
SPC HR:	49	8	24	17	240	

ภาพที่ ก.54 ค่า Precision K- Nearest Neighbor ปรับพารามิเตอร์ number of folds ที่ 5

ผลของการทำนายค่า Precision ของตัวแบบ K- Nearest Neighbor โดยใช้ number of folds ที่ 5 ได้ผลค่าค่า Precision อยู่ที่ 79.58%

3.5 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 10

PerformanceVector:
accuracy: 81.00%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR	
SPC:	251	10	10	17	19	
SPARK:	14	166	2	16	0	
CAB4:	5	6	117	19	15	
CAB4 HR:		27	10	5	441	12
SPC HR:	49	8	24	17	240	

ภาพที่ ก.55 ค่า Accuracy K- Nearest Neighbor ปรับพารามิเตอร์ number of folds ที่ 10

ผลของการทำนายค่า Accuracy ของตัวแบบ K- Nearest Neighbor โดยใช้ number of folds ที่ 10 ได้ผลค่าความถูกต้องอยู่ที่ 81.00%

```
weighted_mean_recall: 80.00%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     251     10      10      17      19
SPARK:   14      166     2       16      0
CAB4:    5       6       117     19      15
CAB4 HR:         27      10      5       441     12
SPC HR:  49      8       24      17      240
```

ภาพที่ ก.56 ค่า Recall K- Nearest Neighbor ปรับพารามิเตอร์ number of folds ที่ 10

ผลของการทำนายค่า Recall ของตัวแบบ K- Nearest Neighbor โดยใช้ number of folds ที่ 10 ได้ผลค่าค่า Recall อยู่ที่ 80.00%

```
weighted_mean_precision: 79.58%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     251     10      10      17      19
SPARK:   14      166     2       16      0
CAB4:    5       6       117     19      15
CAB4 HR:         27      10      5       441     12
SPC HR:  49      8       24      17      240
```

ภาพที่ ก.57 ค่า Precision K- Nearest Neighbor ปรับพารามิเตอร์ number of folds ที่10

ผลของการทำนายค่า Precision ของตัวแบบ K- Nearest Neighbor โดยใช้ number of folds ที่ 10 ได้ผลค่าค่า Precision อยู่ที่ 79.58%

4. ผลการทดลองของตัวแบบ Random Forest

4.1 การแบ่งข้อมูลแบบ 70:30 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random Forest โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

PerformanceVector:
accuracy: 71.78%
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR
SPC:	67	7	13	14	8
SPARK:	4	44	1	7	1
CAB4:	0	5	18	4	2
CAB4 HR:		6	2	0	122 3
SPC HR:	27	2	15	6	72

ภาพที่ ก.58 ค่า Accuracy Random Forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 71.78%

weighted_mean_recall: 67.90%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:

True:	SPC	SPARK	CAB4	CAB4 HR	SPC HR
SPC:	67	7	13	14	8
SPARK:	4	44	1	7	1
CAB4:	0	5	18	4	2
CAB4 HR:		6	2	0	122 3
SPC HR:	27	2	15	6	72

ภาพที่ ก.59 ค่า Recall Random Forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth ได้ผลค่าค่า Recall อยู่ที่ 67.90%

```
weighted_mean_precision: 70.30%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     67      7       13      14      8
SPARK:   4       44      1       7       1
CAB4:    0       5       18      4       2
CAB4 HR:         6       2       0       122     3
SPC HR: 27       2       15      6       72
```

ภาพที่ ก.60 ค่า Precision Random Forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 5

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยแบ่งข้อมูล 70:30 และ 5 Maximal Depth 5ได้ผลค่า Precision อยู่ที่ 70.30%

4.2 การแบ่งข้อมูลแบบ 70:30 และปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random Forest โดยการแบ่งข้อมูลแบบ 70:30 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 77.33%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     64      1       5       5       2
SPARK:   6       47      1       2       1
CAB4:    3       4       30      5       3
CAB4 HR:         11      6       4       133     6
SPC HR: 20       2       7       8       74
```

ภาพที่ ก.61 ค่า Accuracy Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 77.33%

```
weighted_mean_recall: 75.34%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     64      1       5       5       2
SPARK:   6       47      1       2       1
CAB4:    3       4       30      5       3
CAB4 HR:         11      6       4       133     6
SPC HR: 20       2       7       8       74
```

ภาพที่ ก.62 ค่า Recall Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 75.34%

```
weighted_mean_precision: 76.41%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     64      1       5       5       2
SPARK:   6       47      1       2       1
CAB4:    3       4       30      5       3
CAB4 HR:         11      6       4       133     6
SPC HR: 20       2       7       8       74
```

ภาพที่ ก.63 ค่า Precision Random forest โดยปรับพารามิเตอร์ Maximal depth ที่ 10

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยแบ่งข้อมูล 70:30 และ 10 Maximal Depth ได้ผลค่า Precision อยู่ที่ 76.41%

4.3 การแบ่งข้อมูลแบบ 75:25 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random Forest โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 5 Maximal Depth

```
PerformanceVector:
accuracy: 72.92%
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    55      9       7       7       3
SPARK:  4       33      1       2       1
CAB4:   0       5       16      3       1
CAB4 HR:        4       1       0       105     3
SPC HR: 23      2       15      10      63
```

ภาพที่ ก.64 ค่า Accuracy Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 72.92%

```
weighted_mean_recall: 68.48%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    55      9       7       7       3
SPARK:  4       33      1       2       1
CAB4:   0       5       16      3       1
CAB4 HR:        4       1       0       105     3
SPC HR: 23      2       15      10      63
```

ภาพที่ ก.65 ค่า Recall Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 68.48%

```
weighted_mean_precision: 72.21%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    55      9       7       7       3
SPARK:  4       33      1       2       1
CAB4:   0       5       16      3       1
CAB4 HR:        4       1       0       105     3
SPC HR: 23      2       15      10      63
```

ภาพที่ ก.66 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยแบ่งข้อมูล 75:25 และ 5 Maximal Depth ได้ผลค่า Precision อยู่ที่ 72.21%

4.4 การแบ่งข้อมูลแบบ 75:25 และปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random Forest โดยการแบ่งข้อมูลแบบ 75:25 โดยปรับพารามิเตอร์ Maximal Depth อยู่ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 79.36%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     60      1       5       7       3
SPARK:   2       37      1       2       1
CAB4:    3       4       25      4       2
CAB4 HR:         6       5       2       112     3
SPC HR: 15       3       6       2       62
```

ภาพที่ ก.67 ค่า Accuracy Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 79.36%

```
weighted_mean_recall: 76.68%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     60      1       5       7       3
SPARK:   2       37      1       2       1
CAB4:    3       4       25      4       2
CAB4 HR:         6       5       2       112     3
SPC HR: 15       3       6       2       62
```

ภาพที่ ก.68 ค่า Recall Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 76.68%

```
weighted_mean_precision: 77.75%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     60      1       5       7       3
SPARK:   2       37      1       2       1
CAB4:    3       4       25      4       2
CAB4 HR:         6       5       2       112     3
SPC HR: 15       3       6       2       62
```

ภาพที่ ก.69 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยแบ่งข้อมูล 75:25 และ 10 Maximal Depth ได้ผลค่า Precision อยู่ที่ 77.75%

4.5 การแบ่งข้อมูลแบบ 80:20 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random Forest โดยการแบ่งข้อมูลแบบ 80:20 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

```
PerformanceVector:
accuracy: 72.33%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     48      5       6       4       3
SPARK:   4       24      1       1       1
CAB4:    0       6       12      4       2
CAB4 HR:         4       4       1       85      3
SPC HR: 13       1       12      8       48
```

ภาพที่ ก.70 ค่า Accuracy Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 72.33%

```
weighted_mean_recall: 66.92%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     48      5       6       4       3
SPARK:   4       24      1       1       1
CAB4:    0       6       12      4       2
CAB4 HR:         4       4       1       85      3
SPC HR: 13       1       12      8       48
```

ภาพที่ ก.71 ค่า Recall Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth ได้ผลค่า Recall อยู่ที่ 66.92%

```
weighted_mean_precision: 69.26%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     48      5       6       4       3
SPARK:   4       24      1       1       1
CAB4:    0       6       12      4       2
CAB4 HR:         4       4       1       85      3
SPC HR: 13       1       12      8       48
```

ภาพที่ ก.72 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth ได้ผลค่า Precision อยู่ที่ 69.26%

4.6 การแบ่งข้อมูลแบบ 80:20 และปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random Forest โดยการแบ่งข้อมูลแบบ 80:20 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

```
accuracy: 82.67%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     54      0       1       1       2
SPARK:   1       27      0       1       1
CAB4:    1       5       23      4       2
CAB4 HR:         6       7       2       94      2
SPC HR: 7        1       6       2       50
```

ภาพที่ ก.73 ค่า Accuracy Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth ได้ผลค่าความถูกต้องอยู่ที่ 82.67%

```
weighted_mean_recall: 79.50%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     54      0       1       1       2
SPARK:   1       27      0       1       1
CAB4:    1       5       23      4       2
CAB4 HR:         6       7       2       94      2
SPC HR: 7        1       6       2       50
```

ภาพที่ ก.74 ค่า Recall Random forest โดยแบ่งข้อมูล 80:20 และ 5 Maximal Depth

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth ได้ผลค่า Recall อยู่ที่ 79.50%

```
weighted_mean_precision: 81.85%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     54      0       1       1       2
SPARK:   1       27      0       1       1
CAB4:    1       5       23      4       2
CAB4 HR:         6       7       2       94      2
SPC HR: 7        1       6       2       50
```

ภาพที่ ก.75 ค่า Precision Recall Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยแบ่งข้อมูล 80:20 และ 10 Maximal Depth ได้ผลค่า Precision อยู่ที่ 81.85%

4.7 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

ผลการทดลองค่า Accuracy , ค่า Recall และค่า Precision ของตัวโดยแบบ Random forest

```
PerformanceVector:
accuracy: 74.73%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     232     26      25      16      16
SPARK:   11      151     2       11      0
CAB4:    1       7       72      15      2
CAB4 HR:         21      9       2       410     12
SPC HR: 81       7       57      58      256
```

ภาพที่ ก.76 ค่า Accuracy Random forest โดยใช้ number of folds 5 Maximal Depth 5

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยใช้ number of folds ที่ 5 และปรับค่า Maximal Depth ที่ 5 ได้ผลค่าความถูกต้องอยู่ที่ 74.73%

```
weighted_mean_recall: 71.60%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     232     26      25      16      16
SPARK:   11      151     2       11      0
CAB4:    1       7       72      15      2
CAB4 HR:         21      9       2       410     12
SPC HR: 81       7       57      58      256
```

ภาพที่ ก.77 ค่า Recall Random forest โดยใช้ number of folds 5 Maximal Depth 5

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยใช้ number of folds ที่ 5 และปรับค่า Maximal Depth ที่ 5 ได้ผลค่า Recall อยู่ที่ 71.60%

```
weighted_mean_precision: 76.05%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     232     26      25      16      16
SPARK:   11      151     2       11      0
CAB4:    1       7       72      15      2
CAB4 HR:         21      9       2       410     12
SPC HR: 81       7       57      58      256
```

ภาพที่ ก.78 ค่า Precision Random forest โดยใช้ number of folds 5 Maximal Depth 5

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยใช้ number of folds ที่ 5 และปรับค่า Maximal Depth ที่ 5 ได้ผลค่า Precision อยู่ที่ 76.05%

4.8 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 5 และปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 84.13%
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    265     10      6       14      10
SPARK:  4       167     2       6       0
CAB4:   7       5       115     11      8
CAB4 HR:        21      13      7       460     13
SPC HR: 49      5       28      19      255
```

ภาพที่ ก.79 ค่า Accuracy Random forest โดยใช้ number of folds 5 Maximal Depth 10

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยใช้ number of folds ที่ 5และปรับค่า Maximal Depth ที่ 10 ได้ผลค่าความถูกต้องอยู่ที่ 84.13%

```
weighted_mean_recall: 82.45%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    265     10      6       14      10
SPARK:  4       167     2       6       0
CAB4:   7       5       115     11      8
CAB4 HR:        21      13      7       460     13
SPC HR: 49      5       28      19      255
```

ภาพที่ ก.80 ค่า Recall Random forest โดยใช้ number of folds 5 Maximal Depth 10

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยใช้ number of folds ที่ 5 ได้ผลและปรับค่า Maximal Depth ที่ 10 ค่า Recall อยู่ที่ 82.45%

ภาพที่ ก.81 ค่า Precision Random forest โดยใช้ number of folds 5 Maximal Depth 10

```
weighted_mean_precision: 84.01%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     265     10      6       14      10
SPARK:   4       167     2       6       0
CAB4:    7       5       115     11      8
CAB4 HR:         21      13      7       460     13
SPC HR: 49       5       28      19      255
```

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยใช้ number of folds ที่ 5 และปรับค่า Maximal Depth ที่ 10 ได้ผลค่า Precision อยู่ที่ 84.01%

4.9 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 10 และปรับพารามิเตอร์ Maximal Depth ที่ 5 Maximal Depth

```
PerformanceVector:
accuracy: 74.73%
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     232     26      25      16      16
SPARK:   11      151     2       11      0
CAB4:    1       7       72      15      2
CAB4 HR:         21      9       2       410     12
SPC HR: 81       7       57      58      256
```

ภาพที่ ก.82 ค่า Accuracy Random forest โดยใช้ number of folds 10 Maximal Depth 5

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยใช้ number of folds ที่ 10 และปรับค่า Maximal Depth ที่ 5 ได้ผลค่าความถูกต้องอยู่ที่ 74.73%

```
weighted_mean_recall: 71.60%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     232     26      25      16      16
SPARK:   11      151     2       11      0
CAB4:    1       7       72      15      2
CAB4 HR:         21      9       2       410     12
SPC HR: 81       7       57      58      256
```

ภาพที่ ก.83 ค่า Recall Random forest โดยใช้ number of folds 10 Maximal Depth 5

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยใช้ number of folds ที่ 10 และปรับค่า Maximal Depth ที่ 5 ได้ผลค่า Recall อยู่ที่ 71.60%

```
weighted_mean_precision: 76.05%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    232     26      25      16      16
SPARK:  11      151     2       11      0
CAB4:   1       7       72      15      2
CAB4 HR:        21      9       2       410     12
SPC HR: 81      7       57      58      256
```

ภาพที่ ก.84 ค่า Precision Random forest โดยใช้ number of folds 10 Maximal Depth 5

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยใช้ number of folds ที่ 10 และปรับค่า Maximal Depth ที่ 5 ได้ผลค่า Precision อยู่ที่ 76.05%

4.10 การแบ่งข้อมูลแบบ Cross Validation โดยค่าปรับพารามิเตอร์ number of folds ที่ 10 และปรับพารามิเตอร์ Maximal Depth ที่ 10 Maximal Depth

```
PerformanceVector:
accuracy: 84.13%
ConfusionMatrix:
True:   SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:    265     10      6       14      10
SPARK:  4       167     2       6       0
CAB4:   7       5       115     11      8
CAB4 HR:        21      13      7       460     13
SPC HR: 49      5       28      19      255
```

ภาพที่ ก.85 ค่า Accuracy Random forest โดยใช้ number of folds 10 Maximal Depth 10

ผลของการทำนายค่า Accuracy ของตัวแบบ Random forest โดยใช้ number of folds ที่ 10 และปรับค่า Maximal Depth ที่ 10 ได้ผลค่าความถูกต้องอยู่ที่ 84.13%

```
weighted_mean_recall: 82.45%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     265     10      6       14      10
SPARK:   4       167     2       6       0
CAB4:    7       5       115     11      8
CAB4 HR:         21      13      7       460     13
SPC HR: 49       5       28      19      255
```

ภาพที่ ก.86 ค่า Recall Random forest โดยใช้ number of folds 10 Maximal Depth 10

ผลของการทำนายค่า Recall ของตัวแบบ Random forest โดยใช้ number of folds ที่ 10 ได้ผลและปรับค่า Maximal Depth ที่ 10 ค่า Recall อยู่ที่ 82.45%

```
weighted_mean_precision: 84.01%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:    SPC     SPARK   CAB4    CAB4 HR SPC HR
SPC:     265     10      6       14      10
SPARK:   4       167     2       6       0
CAB4:    7       5       115     11      8
CAB4 HR:         21      13      7       460     13
SPC HR: 49       5       28      19      255
```

ภาพที่ ก.87 ค่า Precision Random forest โดยใช้ number of folds 5 Maximal Depth 10

ผลของการทำนายค่า Precision ของตัวแบบ Random forest โดยใช้ number of folds ที่ 10 และปรับค่า Maximal Depth ที่ 10 ได้ผลค่า Precision อยู่ที่ 84.01%

ประวัติผู้จัดทำ

รหัสนักศึกษา	: 6150100013
ชื่อ-นามสกุล	: นายณัฐสิทธิ แพวตะคุ
ที่อยู่	: 223/13 หมู่ 6 ตำบลอ้อมน้อย อำเภอกระทุ่มแบน จังหวัดสมุทรสาคร 74130
เบอร์โทรศัพท์	: 065-8211997
E-Mail	: nattasitt19@gmail.com
ระดับประกาศนียบัตรวิชาชีพ	: โรงเรียนสารสาสน์วิเทศนครปฐม
ระดับปริญญาตรี	: มหาวิทยาลัยสยาม