

บทที่ 3

วิธีการดำเนินโครงการ

โครงการเรื่อง การวิเคราะห์ปัจจัยที่ส่งผลต่อความล่าช้าของสายการบิน โดยใช้วิธีการเหมืองข้อมูล ในบทนี้จะเป็นการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้าไมนิ่งซึ่งมีกระบวนการวิเคราะห์ที่สำคัญหลายขั้นตอน เมื่อเสร็จสิ้นจากกระบวนการวิเคราะห์ข้อมูลแล้วจะเป็นการออกแบบเว็บไซต์ และออกแบบรูปแบบการแสดงผลและบทสรุปจากวิธีการดำเนินงาน ดังนี้

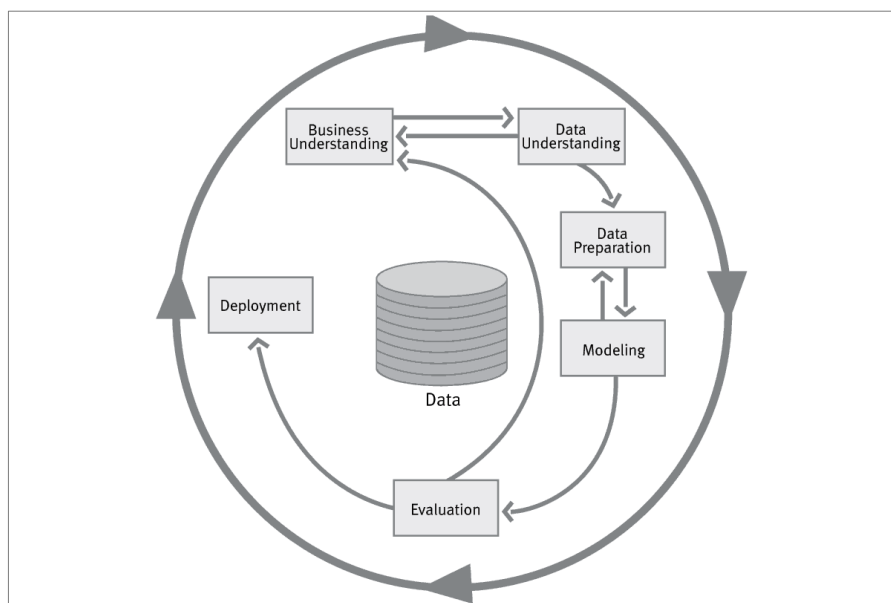
3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM

3.2 การออกแบบเว็บไซต์

3.3 บทสรุป

3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM (Cross Industry Standard Process for Data Mining)

กระบวนการวิเคราะห์ข้อมูล Cross Industry Standard Process for Data Mining หรือ CRISP-DM) พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือของ 3 บริษัทคือ Daimler Chrysler, SPSS และ NCR ซึ่งกระบวนการวิเคราะห์ข้อมูล CRISP-DM จะประกอบด้วย 6 ขั้นตอนหลัก ดังนี้



ภาพที่ 3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM

ที่มา : datasolut.com

3.1.1 การทำความเข้าใจธุรกิจ (Business Understanding) เป็นขั้นตอนแรกของกระบวนการ ที่มุ่งเน้นไปที่การทำความเข้าใจกระบวนการทางธุรกิจ

ผู้วิเคราะห์ข้อมูลทำความเข้าใจกับปัญหา ในบริบทของอุตสาหกรรมการบิน ความตรง

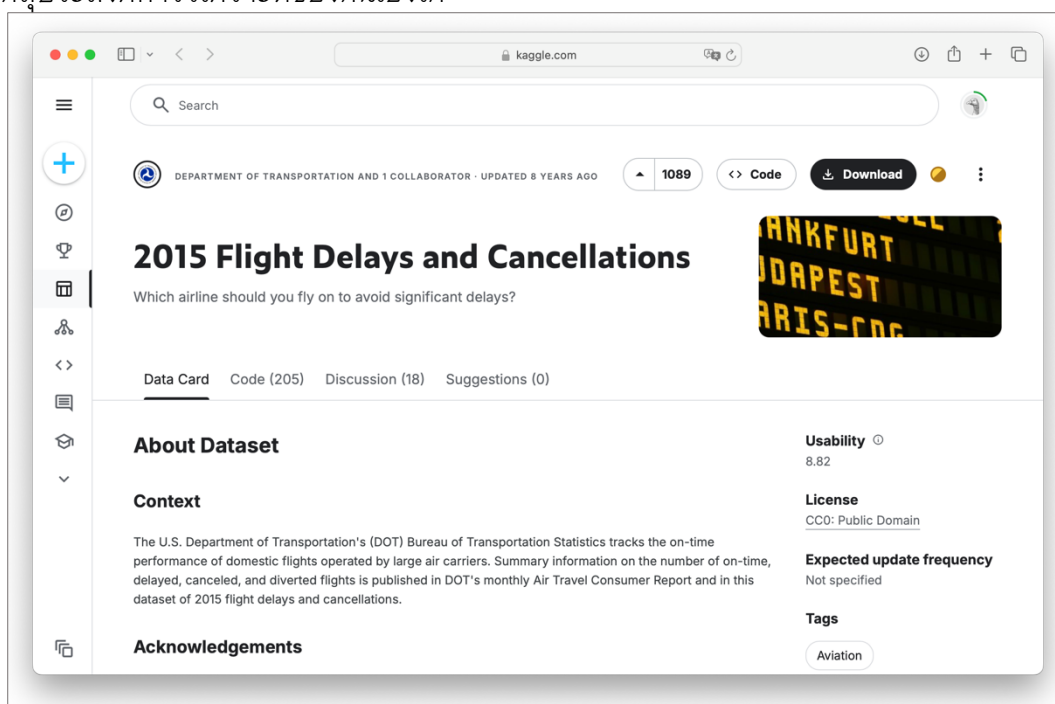
ต่อเวลาเป็นปัจจัยสำคัญที่ส่งผลต่อความพึงพอใจของผู้โดยสารและประสิทธิภาพการดำเนินงานของสายการบิน การวิเคราะห์สาเหตุของความล่าช้าของเที่ยวบินจะช่วยให้สายการบินและหน่วยงานที่เกี่ยวข้องสามารถกำหนดมาตรการปรับปรุงและลดปัญหาความล่าช้าได้

3.1.2 การทำความเข้าใจข้อมูล (Data Understanding) ขั้นตอนการจัดเก็บและรวบรวมข้อมูล รวมถึงการตรวจสอบความถูกต้องของข้อมูลที่ได้รับ ถือเป็นกระบวนการสำคัญในการวิเคราะห์ข้อมูล โดยต้องพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือเพียงบางส่วนเพื่อให้สอดคล้องกับวัตถุประสงค์ที่กำหนดไว้

ผู้วิเคราะห์ข้อมูลเริ่มต้นด้วยการรวบรวมข้อมูลที่เกี่ยวข้อง จากนั้นตรวจสอบความถูกต้องและความน่าเชื่อถือของข้อมูลที่ได้รับ เพื่อลดความคลาดเคลื่อนที่อาจเกิดขึ้น การพิจารณาเลือกใช้ข้อมูลทั้งหมดหรือบางส่วนขึ้นอยู่กับปัจจัยต่างๆ เช่น ความครบถ้วนของข้อมูล คุณภาพของข้อมูล และความสัมพันธ์ของข้อมูลกับปัญหาที่ต้องการวิเคราะห์

ในการวิเคราะห์ข้อมูล นักวิเคราะห์อาจเลือกใช้ข้อมูลบางส่วนจากแหล่งข้อมูลออนไลน์ เช่น Kaggle ซึ่งเป็นแพลตฟอร์มที่รวบรวม Datasets หรือชุดข้อมูลสำหรับการฝึกสอน Machine Learning ที่ใหญ่ที่สุดแห่งหนึ่งของโลก

Kaggle มีชุดข้อมูลหลากหลายประเภท ครอบคลุมหลายอุตสาหกรรม เช่น การแพทย์ การเงิน การตลาด และการบิน ซึ่งช่วยให้นักวิเคราะห์สามารถเลือกใช้ข้อมูลที่เหมาะสมกับวัตถุประสงค์การวิเคราะห์ของตนเองได้



ภาพที่ 3.2 ชุดข้อมูลจาก Kaggle

ที่มา : <https://www.kaggle.com>

ซึ่งชุดข้อมูลของ Flight Delay Data มีข้อมูลทั้งหมด 5,819,079 แถว ประกอบด้วย 31 คอลัมน์ โดยมีข้อมูล ดังนี้

ตารางที่ 3.1 Data Dictionary อธิบายรายละเอียดในแต่ละแอตทริบิวต์

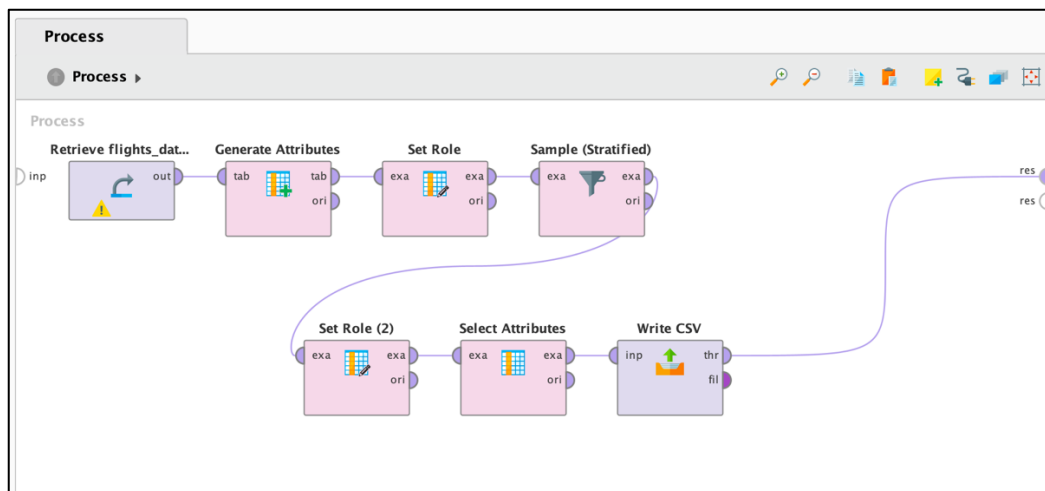
Attribute	ความหมาย	ชนิดข้อมูล
YEAR	ปีที่เที่ยวบินออกเดินทาง	Integer
MONTH	เดือนที่เที่ยวบินออกเดินทาง	Integer
DAY	วันที่ของเที่ยวบิน	Integer
DAY_OF_WEEK	วันในสัปดาห์	Integer
AIRLINE	รหัสสายการบิน	Polynomial
FLIGHT_NUMBER	หมายเลขเที่ยวบิน	Integer
TAIL_NUMBER	รหัสเครื่องบินที่ใช้บินเที่ยวนี้	Polynomial
ORIGIN_AIRPORT	รหัสสนามบินต้นทาง	Polynomial
DESTINATION_AIRPORT	รหัสสนามบินปลายทาง	Polynomial
SCHEDULED_DEPARTURE	เวลาตามกำหนดการออกเดินทาง	Integer
DEPARTURE_TIME	เวลาจริงที่เครื่องบินออกจากสนามบิน	Real
DEPARTURE_DELAY	ความล่าช้าในการออกเดินทาง	Real
TAXI_OUT	เวลาที่เครื่องบินเคลื่อนตัวไปยังรันเวย์ก่อนบิน	Real
WHEELS_OFF	เวลาที่ล้อเครื่องบินออกจากพื้น	Real
SCHEDULED_TIME	เวลาตามกำหนดการของเที่ยวบิน	Integer
ELAPSED_TIME	เวลาทั้งหมดที่ใช้เดินทางจริง	Real
AIR_TIME	เวลาที่เครื่องบินอยู่บนอากาศ	Real
DISTANCE	ระยะทางของเที่ยวบิน	Integer
WHEELS_ON	เวลาที่ล้อเครื่องบินแตะพื้นสนามบินปลายทาง	Real
TAXI_IN	เวลาที่เครื่องบินเคลื่อนตัวไปยังประตูผู้โดยสาร	Real
SCHEDULED_ARRIVAL	เวลาตามกำหนดการถึงปลายทาง	Integer

ARRIVAL_TIME	เวลาจริงที่เครื่องบินถึงสนามบินปลายทาง	Real
ARRIVAL_DELAY	ความล่าช้าในการถึงที่หมาย	Real
DIVERTED	เที่ยวบินถูกเปลี่ยนเส้นทางหรือไม่	Binominal
CANCELLED	เที่ยวบินถูกยกเลิกหรือไม่	Binominal
CANCELLATION_REASON	เหตุผลการยกเลิกเที่ยวบิน	Polynomial
AIR_SYSTEM_DELAY	ความล่าช้าจากระบบควบคุมการบิน	Real
SECURITY_DELAY	ความล่าช้าจากระบบรักษาความปลอดภัย	Real
AIRLINE_DELAY	ความล่าช้าจากปัญหาของสายการบิน	Real
LATE_AIRCRAFT_DELAY	ความล่าช้าจากเที่ยวบินก่อนหน้าที่มาสาย	Real
WEATHER_DELAY	ความล่าช้าจากสภาพอากาศ	Real

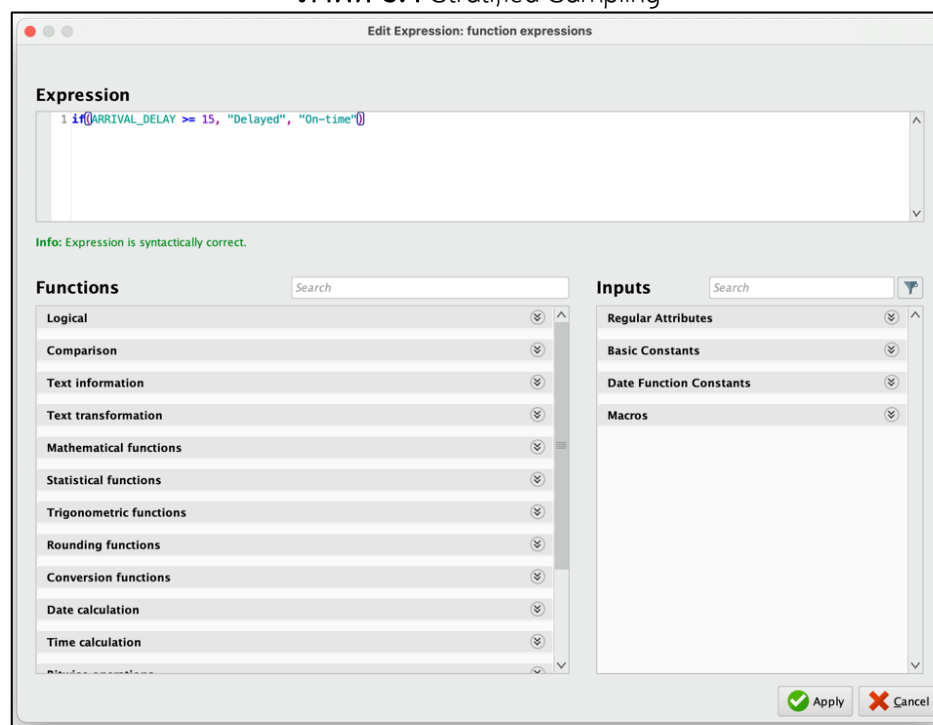
3.1.3 การเตรียมข้อมูล (Data Preparation) ขั้นตอนนี้เป็น การแปลงและจัดเตรียมข้อมูลที่รวบรวมมาให้อยู่ในรูปแบบที่พร้อมสำหรับการวิเคราะห์ โดยใช้กระบวนการ Data Cleaning เพื่อให้ข้อมูลมีความสอดคล้องและเป็นมาตรฐานเดียวกัน รวมถึงการกำจัดข้อมูลที่ไม่น่าเป็นออกไป ขั้นตอนดังนี้

3.1.3.1 จากข้อมูลชุดใหญ่ที่นำมาใช้ในการวิจัยนี้ คือชุดข้อมูล 2015 Flight Delays and Cancellations จากเว็บไซต์ Kaggle ซึ่งประกอบด้วยข้อมูลเที่ยวบินภายในประเทศของสหรัฐอเมริกาในปี พ.ศ. 2558 รวมทั้งสิ้นจำนวน 5,819,079 แถว

3.1.3.2 การสุ่มข้อมูลเพื่อเตรียมชุดข้อมูลเที่ยวบิน การสุ่มตัวอย่างข้อมูล (Sampling) เป็นขั้นตอนสำคัญในกระบวนการเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล โดยเฉพาะอย่างยิ่งเมื่อชุดข้อมูลมีขนาดใหญ่มาก การใช้ข้อมูลทั้งหมดอาจส่งผลกระทบต่อประสิทธิภาพและสิ้นเปลืองทรัพยากรคอมพิวเตอร์ การสุ่มตัวอย่างจึงเป็นวิธีที่ช่วยลดปริมาณข้อมูลลง แต่ยังคงรักษาลักษณะสำคัญและความเป็นตัวแทนของข้อมูลเดิมเอาไว้ การสุ่มตัวอย่างข้อมูลมีหลายรูปแบบ เช่น การสุ่มแบบง่าย (Simple Random Sampling) การสุ่มแบบแบ่งชั้น (Stratified Sampling) และการสุ่มแบบกลุ่ม (Cluster Sampling) สำหรับงานวิจัยนี้ได้เลือกใช้วิธี Stratified Sampling โดยใช้ตัวแปร delay_status เป็นตัวกำหนดชั้น (strata) เพื่อให้ชุดข้อมูลที่สุ่มมาใหม่มีสัดส่วนเที่ยวบินที่ล่าช้าและตรงเวลาใกล้เคียงกับข้อมูลต้นฉบับมากที่สุด ขั้นตอนการสุ่มประกอบด้วย การกำหนดตัวแปร delay_status ให้มีบทบาทเป็น label ในซอฟต์แวร์ RapidMiner จากนั้นใช้เครื่องมือ Sample (Stratified) โดยตั้งค่าอัตราส่วนการสุ่ม (sample ratio) เท่ากับ 0.05 หรือ 5% ของข้อมูลทั้งหมด การตั้งค่านี้ทำให้ได้ข้อมูลย่อยที่มีจำนวนเพียงพอสำหรับการวิเคราะห์ แต่มีขนาดเล็กเพียงพอที่จะประมวลผลได้อย่างรวดเร็วและไม่เกินข้อจำกัดของหน่วยความจำ (Jiawei Han, 2012)



ภาพที่ 3.4 Stratified Sampling



ภาพที่ 3.5 ตัวแปร delay_status

จากการสุ่มครั้งนี้ ข้อมูลถูกลด เหลือ 290,954 แถว โดยยังคงสัดส่วนของคลาส delay_status ให้เหมือนข้อมูลเดิม

Open in Turbo Prep Auto Model Interactive Analysis Filter (290,954 / 290,954 examples):

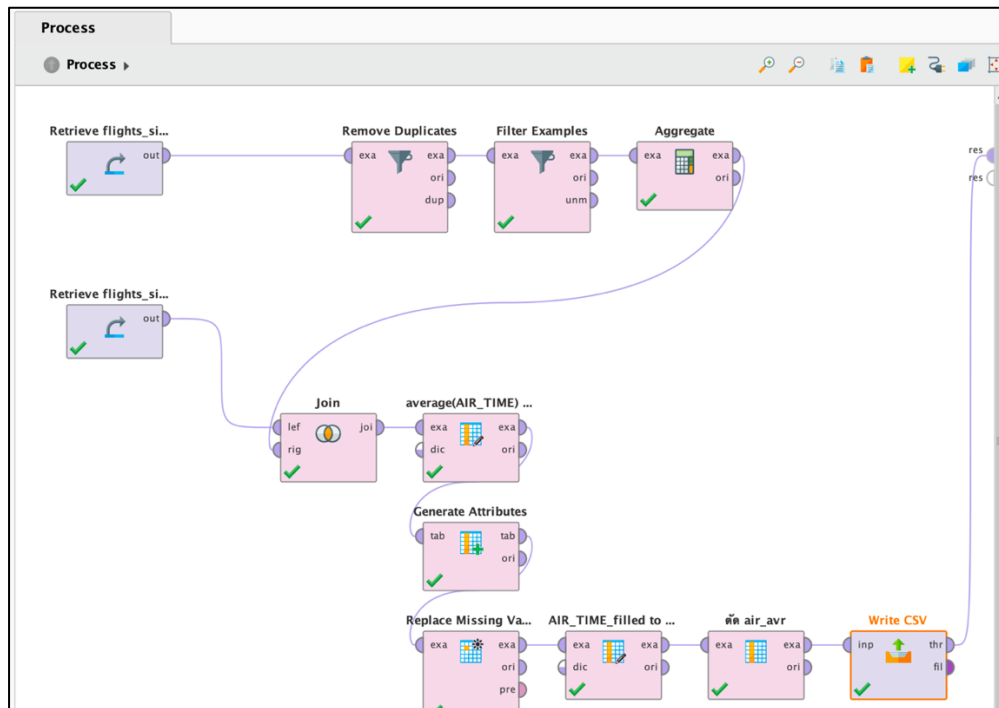
Row No.	YEAR	MONTH	DAY	DAY_OF_W...	AIRLINE	FLIGHT_NU...	TAIL_NUMB...	ORIGIN_AIR...	DESTINA...
1	2015	1	1	4	AA	258	N3HYAA	LAX	MIA
2	2015	1	1	4	NK	168	N629NK	PHX	ORD
3	2015	1	1	4	US	617	N804AW	ANC	PHX
4	2015	1	1	4	B6	730	N621JB	BQN	MCO
5	2015	1	1	4	EV	4685	N26545	BRO	IAH
6	2015	1	1	4	UA	1532	N77066	SFO	DEN
7	2015	1	1	4	DL	824	N3736C	PHX	DTW
8	2015	1	1	4	UA	1237	N65832	SEA	IAH
9	2015	1	1	4	EV	4646	N29917	CHS	EWR
10	2015	1	1	4	MQ	2926	N932MQ	CHS	DFW
11	2015	1	1	4	EV	3941	N16561	LAN	ORD
12	2015	1	1	4	EV	5677	N12563	GSO	EWR
13	2015	1	1	4	US	1941	N708UW	PVD	CLT
14	2015	1	1	4	DL	977	N988DL	ORD	ATL
15	2015	1	1	4	DL	2417	N554NW	LAS	DTW
16	2015	1	1	4	B6	1502	N535JB	FLL	JFK
17	2015	1	1	4	EV	4621	N161EB	CAV	EMD

ภาพที่ 3.6 ผลลัพธ์ Stratified Sampling

3.1.3.3 การทำความสะอาดข้อมูล (Data Cleaning) การทำความสะอาด

สะอาดข้อมูลในงานวิจัยนี้ดำเนินการด้วยโปรแกรม RapidMiner โดยมีขั้นตอนตามลำดับดังนี้ เพื่อให้ได้ชุดข้อมูลที่มีความถูกต้อง ครบถ้วน และพร้อมสำหรับการวิเคราะห์เชิงสถิติและการสร้างแบบจำลองพยากรณ์

- ก. การนำเข้าข้อมูล (Retrieve) เพื่อดึงชุดข้อมูลเที่ยวบินจากแหล่งที่เก็บ (Repository) เข้าสู่สภาพแวดล้อมการประมวลผลของ RapidMiner ซึ่งเป็นจุดเริ่มต้นของกระบวนการทำความสะอาดข้อมูล
- ข. การลบข้อมูลซ้ำ (Remove Duplicates) ระเบียนที่ซ้ำกัน (Duplicate Records) ทำให้การวิเคราะห์เกิดความผิดพลาดและเพิ่มน้ำหนักข้อมูลบางรายการเกินจริง
- ค. การกรองข้อมูล (Filter Examples) ใช้เงื่อนไข 3 ข้อ
 - DEPARTURE_TIME is not missing เก็บเฉพาะเที่ยวบินที่มีเวลาขึ้นบินจริง
 - TAIL_NUMBER is not missing เก็บเฉพาะระเบียนที่มีหมายเลขเครื่องบิน
 - CANCELLED = 0 ตัดเที่ยวบินที่ถูกยกเลิกออกจากชุดข้อมูล เพื่อให้เหลือเฉพาะข้อมูลที่มีความสมบูรณ์และสามารถใช้สร้างโมเดลได้ ข้อมูลที่ถูกยกเลิกหรือขาดข้อมูลสำคัญจะไม่สามารถใช้ในการพยากรณ์การมาถึงล่าช้าได้



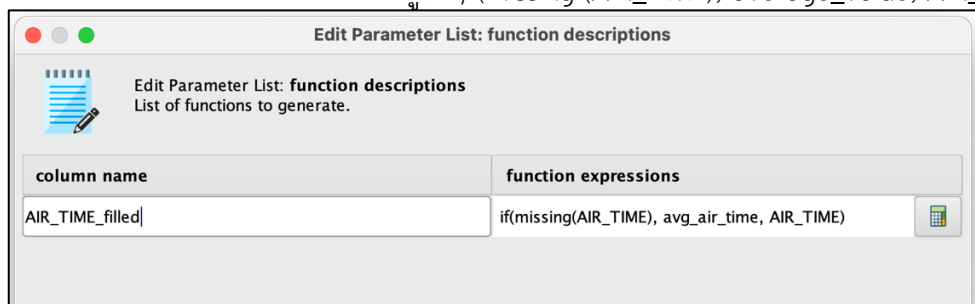
ภาพที่ 3.7 การทำความสะอาดข้อมูล (Data Cleaning)

Field	Condition	Value
DEPARTURE_TIME	is not missing	
TAIL_NUMBER	is not missing	
CANCELLED	=	0

ภาพที่ 3.8 การกรองข้อมูล (Filter Examples)

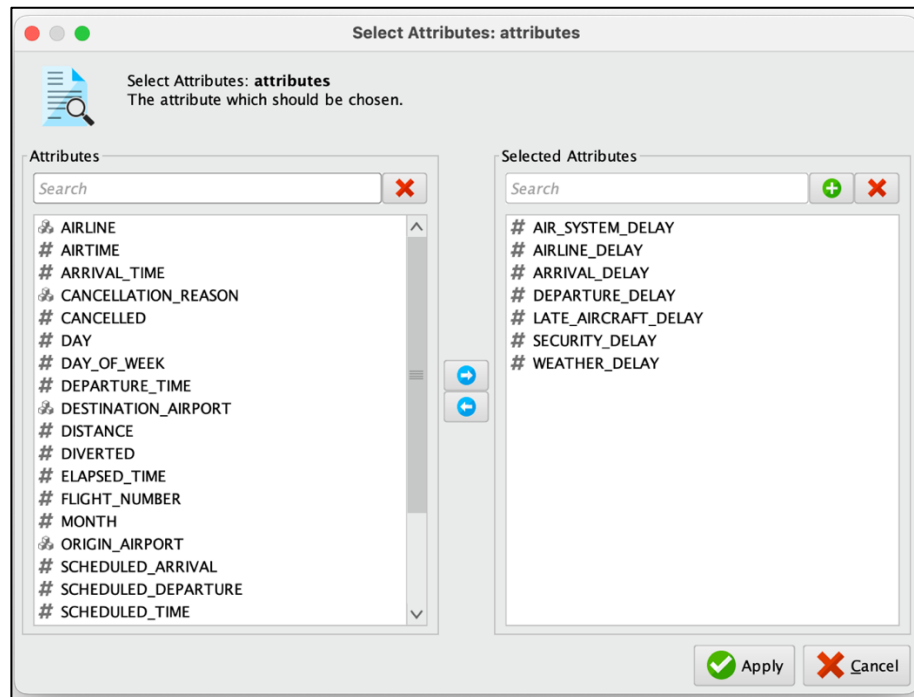
- ง. การสรุปข้อมูล (Aggregate) คำนวณสถิติ ค่าเฉลี่ยเวลาบิน (AIR_TIME) หรือข้อมูลประกอบ ที่จำเป็นต่อการแทนค่าข้อมูลที่ขาดหาย

- จ. การรวมข้อมูล (Join) นำข้อมูลจากตารางที่ผ่านการสรุปค่าเฉลี่ย (AIR_TIME) มารวมเข้ากับข้อมูลเที่ยวบิน เพื่อให้สามารถแทนค่าที่ขาดหายใน AIR_TIME ได้
- ฉ. การคำนวณค่าเฉลี่ยของเวลาบิน (average (AIR_TIME)) หาค่าเฉลี่ยเวลาบินที่ใช้แทนค่าที่ Missing ในตัวแปร AIR_TIME ซึ่งเป็นตัวแปรสำคัญในการคำนวณความล่าช้า
- ช. การสร้างตัวแปรใหม่ (Generate Attributes) สร้างคอลัมน์ใหม่ เช่น AIR_TIME_filled เพื่อเก็บค่าที่แทน Missing โดยใช้สูตร if (missing (AIR_TIME), average_value, AIR_TIME)



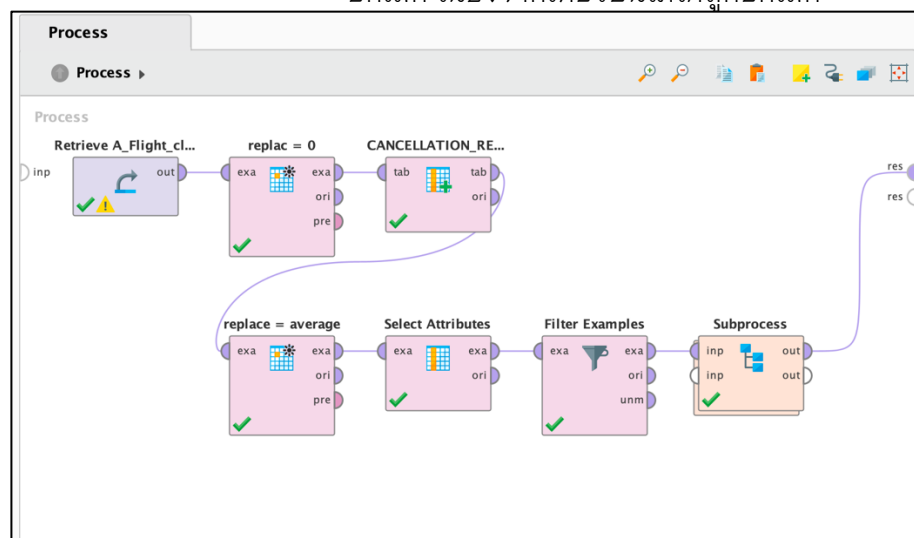
ภาพที่ 3.9 การสร้างตัวแปรใหม่ (Generate Attributes)

- ซ. การแทนค่าที่ขาดหาย (Replace Missing Values) สำหรับตัวแปร DEPARTURE_DELAY, ARRIVAL_DELAY, TAXI_OUT, TAXI_IN ใช้วิธีแทนค่าตามลักษณะตัวแปร เช่น ค่า 0 สำหรับ delay ที่ไม่มี หรือใช้ค่าเฉลี่ยสำหรับตัวแปรเชิงเวลา
- ฅ. การปรับโครงสร้างข้อมูล (Rename / Remove Attributes) เปลี่ยนชื่อคอลัมน์ให้สอดคล้องกับการใช้งาน เปลี่ยน AIR_TIME_filled กลับเป็น AIRTIME และตัดคอลัมน์ที่ไม่จำเป็นออก
- ญ. การแทนค่าที่ขาดหาย (Replace Missing Values) โดยเลือกตัวแปรกลุ่ม Delay ได้แก่ AIR_SYSTEM_DELAY, AIRLINE_DELAY, ARRIVAL_DELAY, DEPARTURE_DELAY, LATE_AIRCRAFT_DELAY, SECURITY_DELAY, และ WEATHER_DELAY เพื่อแทนค่าที่ขาดหาย (Missing) ในกรณีข้อมูลเที่ยวบินมักหมายถึงไม่มีความล่าช้าจากสาเหตุนั้น ๆ การแทนด้วยค่า 0 จึงช่วยให้ความหมายของข้อมูลถูกต้องและป้องกันความคลาดเคลื่อนในการวิเคราะห์



ภาพที่ 3.10 การแทนค่าที่ขาดหายของตัวแปร (Replace Missing Values)

ก. การสร้างตัวแปรสำหรับแทนค่าที่ขาดหาย (Generate Attributes) แทนค่าที่ขาดหายของตัวแปร CANCELLATION_REASON ด้วยรหัส “0” ในกรณีที่ไม่มีข้อมูลเหตุการณ์ยกเลิกเที่ยวบิน (cancellation reason) มักหมายความว่าเที่ยวบินนั้นไม่ได้ถูกยกเลิก การแทนด้วยรหัส 0 ช่วยระบุว่าระเบียบดังกล่าวไม่มีเหตุการณ์ยกเลิก เนื่องจากเที่ยวบินไม่ได้ถูกยกเลิก

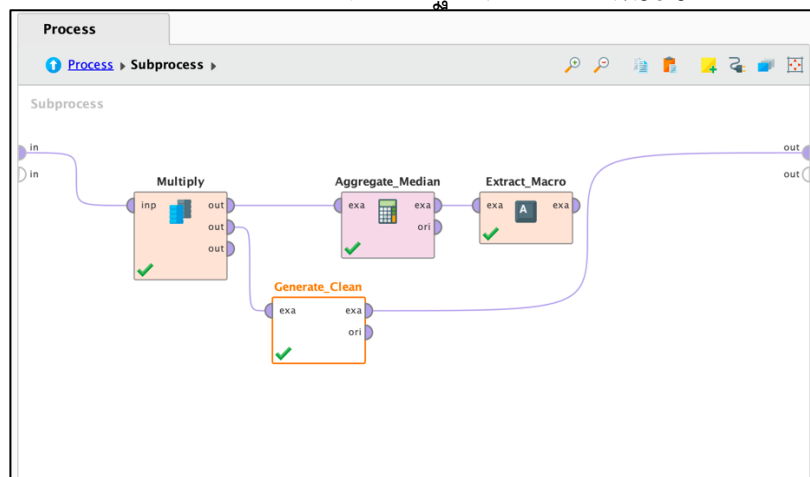


ภาพที่ 3.11 การแทนค่าที่ขาดหายของตัวแปร CANCELLATION_REASON

ก. การแทนค่าที่ขาดหาย (Replace Missing Values) ด้วยค่าเฉลี่ยของตัวแปร เพื่อลดการสูญเสียข้อมูลจากการลบระเบียบที่มีค่าขาดหาย และรักษาขนาดชุดข้อมูลให้

เพียงพอต่อการวิเคราะห์เชิงสถิติและการสร้างแบบจำลองการพยากรณ์ การใช้ค่าเฉลี่ยเหมาะสมกับตัวแปรที่มีการกระจายแบบสมมาตร (Symmetric Distribution) และไม่มีค่าผิดปกติ (Outlier) ที่รุนแรง เนื่องจากค่าเฉลี่ยจะสามารถแทนค่าที่ขาดหายได้โดยไม่บิดเบือนแนวโน้มโดยรวมของข้อมูล

- ฐ. การแทนค่าที่ขาดหายของ ELAPSED_TIME ด้วยค่ามัธยฐาน (Median) ขั้นตอนนี้ใช้ Subprocess ที่ประกอบด้วย โอเปอเรเตอร์ Multiply, Aggregate Median, Extract Macro และ Generate Attributes เพื่อทำการแทนค่าที่ขาดหาย (Missing) ในคอลัมน์ ELAPSED_TIME ด้วยค่ามัธยฐาน เพื่อลดผลกระทบจากค่าผิดปกติ (Outlier) ของเวลาบิน ซึ่งอาจเกิดจากการบันทึกข้อมูลผิดพลาดหรือเที่ยวบินที่มีความผิดปกติด้านเวลา
- Multiply สร้างชุดข้อมูลสำเนาเพื่อให้สามารถประมวลผลการคำนวณและแทนค่าพร้อมกันได้
 - Aggregate Median คำนวณค่ามัธยฐานของ ELAPSED_TIME จากข้อมูลที่มีค่าไม่ Missing
 - Extract Macro เก็บค่ามัธยฐานที่คำนวณได้เป็น Macro เพื่อเรียกใช้ในขั้นตอนสร้างตัวแปรใหม่
 - Generate Attributes สร้างคอลัมน์ โดยแทนค่าที่ Missing ด้วยค่ามัธยฐานที่เก็บไว้ใน Macro



ภาพที่ 3.12 Subprocess

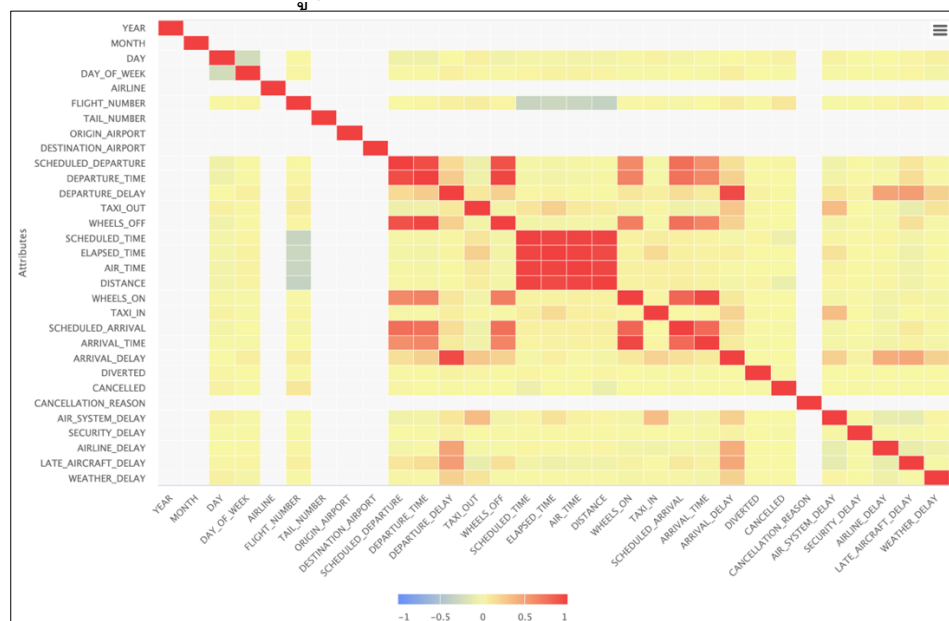
หลังการทำความสะอาดข้อมูลทั้งหมด เหลือข้อมูลที่พร้อมใช้งานจำนวน 290,223 แถว จากข้อมูลดิบทั้งหมด โดยระเบียบที่ถูกกลบออกเป็นระเบียบที่ไม่สามารถใช้ในการวิเคราะห์ได้ เช่น ข้อมูลที่มีค่าขาดหายสำคัญ (Critical Missing Values) หรือเที่ยวบินที่ถูกยกเลิก ซึ่งการลดจำนวนระเบียบดังกล่าวช่วยเพิ่มความถูกต้องของการวิเคราะห์และลดความเอนเอียงของแบบจำลอง (Model Bias)

กำหนดบทบาทของแอตทริบิวต์ ให้เป็นตัวแปรเป้าหมาย (label) โดยใช้โหนด Set Role ซึ่งเป็นขั้นตอนสำคัญในการเตรียมข้อมูลสำหรับการเรียนรู้ของโมเดลในกระบวนการวิเคราะห์ข้อมูล

3.1.4 สร้างแบบจำลอง (Modeling)

3.1.4.1 การสร้างแบบจำลอง (Modeling) เพื่อทำนายความล่าช้าของเที่ยวบิน โดยแบ่งการสร้างโมเดลออกเป็น 2 กลุ่มหลัก) โดยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร ด้วย Correlation Matrix ได้แก่ การจำแนกประเภท (Classification) และการถดถอยเชิงพหุคูณ (Multiple Linear Regression)

- ก. เทคนิคทางสถิติเมทริกซ์สหสัมพันธ์ (correlation matrix) การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรด้วย Correlation Matrix ชุดข้อมูลเที่ยวบิน พบว่ามีตัวแปรจำนวนมากที่มีความสัมพันธ์เชิงเส้นในระดับต่าง ๆ กัน โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson Correlation Coefficient) ซึ่งมีค่าตั้งแต่ -1 ถึง 1 โดยมีความหมายดังนี้
- ค่าสัมประสิทธิ์ใกล้ 1 หรือ -1 หมายถึงมีความสัมพันธ์เชิงเส้นสูง
 - ค่าสัมประสิทธิ์ใกล้ 0 หมายถึงไม่มีความสัมพันธ์เชิงเส้น
 - ในการวิเคราะห์ครั้งนี้ ได้พิจารณาเฉพาะตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์ มากกว่า 0.4 หรือ น้อยกว่า -0.4 เพื่อคัดเลือกเฉพาะตัวแปรที่มีความสัมพันธ์ในระดับปานกลางถึงสูง



ภาพที่ 3.13 Correlation Matrix

โครงการนี้ได้ใช้วิธีการวิเคราะห์ Correlation Matrix เพื่อคัดเลือกตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์ตั้งแต่ 0.4 ขึ้นไป เพื่อนำมาใช้ในการสร้างโมเดล 2 รูปแบบ ได้แก่ การจำแนกประเภท (Classification) และ การถดถอยพหุคูณ (Multiple Linear Regression) โดย

สำหรับโมเดลการจำแนกประเภท จะเลือกใช้เฉพาะตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรตาม ARRIVAL_DELAY ตั้งแต่ 0.4 ขึ้นไป ทั้งนี้ การกำหนดเกณฑ์คัดเลือกที่ระดับ 0.4 มีวัตถุประสงค์เพื่อให้ได้ตัวแปรที่มีความสัมพันธ์ในระดับปานกลางถึงสูง อันสามารถสะท้อนอิทธิพลที่มีนัยสำคัญต่อการเกิดความล่าช้าในการมาถึงของเที่ยวบินได้อย่างเหมาะสม

ตารางที่ 3.2 แสดงค่าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง ARRIVAL_DELAY กับตัวแปรที่ผ่านเกณฑ์การคัดเลือก (≥ 0.4) เพื่อนำมาใช้สร้างโมเดลการจำแนกประเภท (Classification)

ตัวแปร (Attributes)	ค่าสัมประสิทธิ์สหสัมพันธ์กับ ARRIVAL_DELAY
DEPARTURE_DELAY	0.936
LATE_AIRCRAFT_DELAY	0.632
AIRLINE_DELAY	0.617
AIR_SYSTEM_DELAY	0.430

สำหรับการสร้างโมเดล ถดถอยพหุคูณ (Multiple Linear Regression) ได้ใช้วิธีการวิเคราะห์ Correlation Matrix เพื่อตรวจสอบความสัมพันธ์ระหว่างตัวแปรอิสระ (Attributes) กับตัวแปรตาม โดยคัดเลือกเฉพาะตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์ตั้งแต่ 0.4 ขึ้นไป เพื่อให้ได้ชุดตัวแปรที่มีความสัมพันธ์ในระดับปานกลางถึงสูง ซึ่งสามารถสะท้อนอิทธิพลที่มีนัยสำคัญต่อการเกิดความล่าช้าในการมาถึงของเที่ยวบิน และช่วยให้โมเดลสามารถอธิบายความแปรปรวนของข้อมูลได้อย่างเหมาะสม

ตารางที่ 3.3 คู่ตัวแปร (Attributes Pairs) ที่มีค่าสัมประสิทธิ์สหสัมพันธ์ตั้งแต่ 0.4 ขึ้นไป

Attribute 1	Attribute 2	ค่าสัมประสิทธิ์สหสัมพันธ์
SCHEDULED_TIME	DISTANCE	0.984
AIRTIME	SCHEDULED_TIME	0.984
AIRTIME	DISTANCE	0.979
DEPARTURE_TIME	WHEELS_OFF	0.973
WHEELS_ON	ARRIVAL_TIME	0.968
AIRTIME	ELAPSED_TIME	0.963
SCHEDULED_TIME	ELAPSED_TIME	0.960
DEPARTURE_TIME	SCHEDULED_DEPARTURE	0.958

DISTANCE	ELAPSED_TIME	0.950
DEPARTURE_DELAY	ARRIVAL_DELAY	0.936
WHEELS_OFF	SCHEDULED_DEPARTURE	0.933
WHEELS_ON	SCHEDULED_ARRIVAL	0.864
ARRIVAL_TIME	SCHEDULED_ARRIVAL	0.855
WHEELS_OFF	SCHEDULED_ARRIVAL	0.719
SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL	0.708
DEPARTURE_TIME	SCHEDULED_ARRIVAL	0.708
WHEELS_OFF	WHEELS_ON	0.697
DEPARTURE_TIME	WHEELS_ON	0.678
WHEELS_OFF	ARRIVAL_TIME	0.669
DEPARTURE_DELAY	LATE_AIRCRAFT_DELAY	0.656
WHEELS_ON	SCHEDULED_DEPARTURE	0.655
DEPARTURE_TIME	ARRIVAL_TIME	0.650
DEPARTURE_DELAY	AIRLINE_DELAY	0.644
ARRIVAL_DELAY	LATE_AIRCRAFT_DELAY	0.632
ARRIVAL_TIME	SCHEDULED_DEPARTURE	0.627
ARRIVAL_DELAY	AIRLINE_DELAY	0.617
ARRIVAL_DELAY	AIR_SYSTEM_DELAY	0.430

ตารางที่ 3.4 แสดงตัวแปรที่ผ่านเกณฑ์การคัดเลือก เพื่อนำมาใช้ในการสร้างโมเดล Multiple Linear Regression

ลำดับ	Attributes
1	SCHEDULED_TIME
2	DISTANCE

3	AIRTIME
4	ELAPSED_TIME
5	DEPARTURE_TIME
6	SCHEDULED_DEPARTURE
7	ARRIVAL_TIME
8	SCHEDULED_ARRIVAL
9	WHEELS_OFF
10	WHEELS_ON
11	DEPARTURE_DELAY
12	ARRIVAL_DELAY
13	AIRLINE_DELAY
14	LATE_AIRCRAFT_DELAY
15	AIR_SYSTEM_DELAY

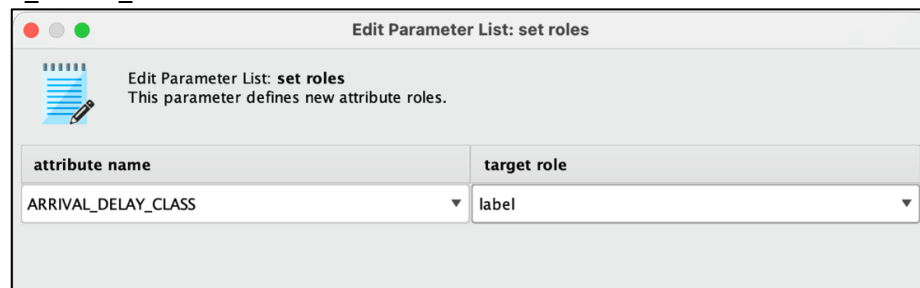
จากตาราง แสดงให้เห็นถึงชุดตัวแปรที่ผ่านเกณฑ์การคัดเลือกด้วยวิธี Correlation Matrix สำหรับใช้สร้างโมเดลทั้งสองรูปแบบ ได้แก่ Classification และ Multiple Linear Regression โดยในส่วนถัดไปจะเป็นการสร้างและประเมินผลของโมเดล Classification ซึ่งมุ่งเน้นการจำแนกว่าเที่ยวบิน โดยใช้ตัวแปรที่มีความสัมพันธ์สูงกับ ARRIVAL_DELAY

- ข. 3.1.4.1.2 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) แบบจำลองการจำแนกประเภท (Classification Model) ในขั้นตอนการสร้างโมเดลการจำแนกประเภท ได้กำหนดตัวแปรเป้าหมาย (Label) คือ ARRIVAL_DELAY โดยทำการสร้างคุณลักษณะใหม่ (Generate Attribute) เพื่อจัดกลุ่มข้อมูลให้อยู่ในรูปแบบ Multiclass ตามเงื่อนไขดังนี้
- ON TIME หมายถึง เมื่อค่า ARRIVAL_DELAY < 5 นาที
 - SHORT DELAY หมายถึง เมื่อค่า ARRIVAL_DELAY อยู่ระหว่าง 5 –14 นาที
 - DELAY หมายถึง เมื่อค่า ARRIVAL_DELAY ≥ 15 นาที



ภาพที่ 3.14 การสร้างตัวแปร ARRIVAL_DELAY_CLASS

จากนั้นได้ทำการกำหนดบทบาทของตัวแปร (Set Role) โดยให้ ARRIVAL_DELAY_CLASS เป็นตัวแปรเป้าหมาย



ภาพที่ 3.15 ขั้นตอนการกำหนดบทบาทของตัวแปร (Set Role)

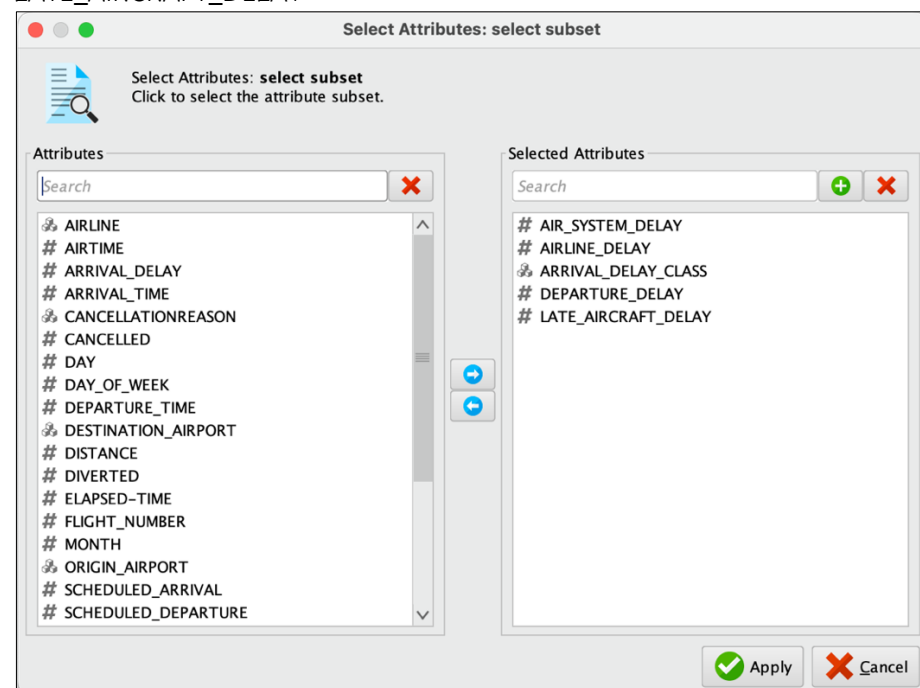
การเลือกตัวแปร (Select Attributes) ขั้นตอนนี้ได้ทำการเลือกเฉพาะตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ตั้งแต่ 0.4 ขึ้นไป เพื่อนำมาใช้ในการสร้างโมเดลการจำแนกประเภท โดยตัวแปรที่ถูกเลือก ได้แก่

AIR_SYSTEM_DELAY

AIRLINE_DELAY

DEPARTURE_DELAY

LATE_AIRCRAFT_DELAY

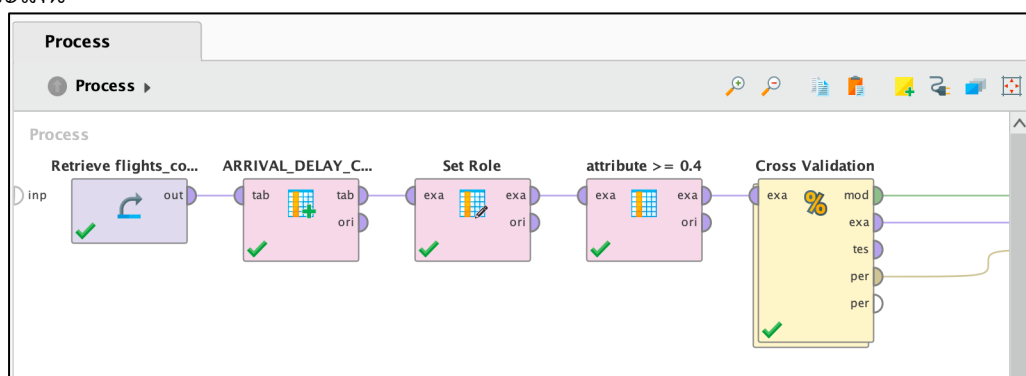


ภาพที่ 3.16 ขั้นตอนการเลือกตัวแปร (Select Attributes)

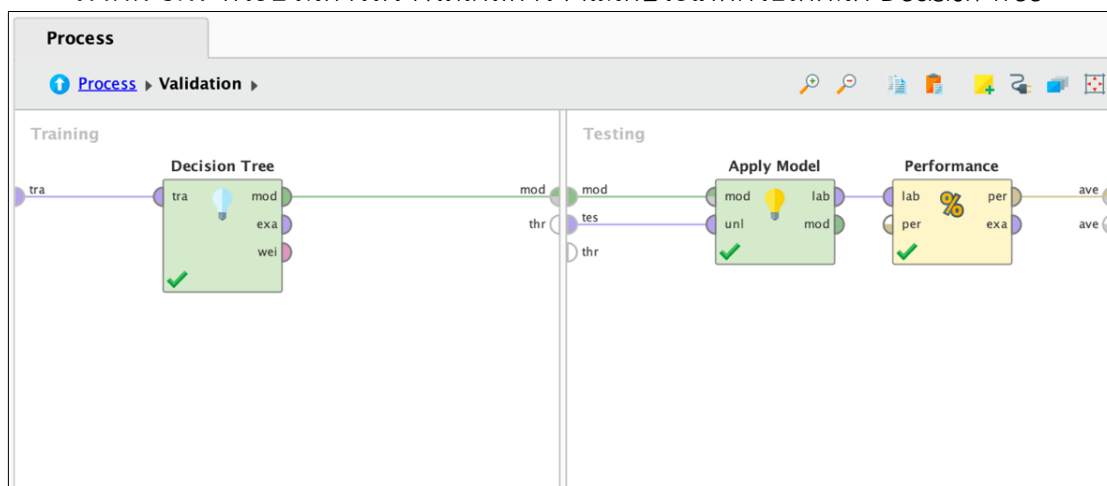
การประเมินโมเดลด้วยวิธี Cross Validation เพื่อประเมินความแม่นยำและประสิทธิภาพของโมเดล โดย Cross Validation ทำหน้าที่แบ่งชุดข้อมูลออกเป็น 2 ส่วนหลัก คือ

Training set สำหรับใช้ในการเรียนรู้ (Train) โมเดล

Testing set สำหรับใช้ทดสอบ (Test) ความสามารถในการทำนายของโมเดล ในโครงการนี้ได้เลือกใช้วิธี k-fold Cross Validation ซึ่งจะสุ่มแบ่งชุดข้อมูลออกเป็น 10 ส่วน และทำการ Train/Test สลับกันจนครบทุกส่วน ผลลัพธ์ที่ได้จะถูกเฉลี่ยเพื่อใช้เป็นตัวแทนประสิทธิภาพโดยรวมของโมเดล ช่วยลดความเอนเอียง (Bias) และเพิ่มความน่าเชื่อถือของการประเมิน



ภาพที่ 3.17 กระบวนการสร้างโมเดลการจำแนกประเภทด้วยเทคนิค Decision Tree



ภาพที่ 3.18 กระบวนการเรียนรู้และประเมินผลโมเดล Decision Tree ด้วย Cross Validation

โมเดล Decision Tree มีค่า Accuracy รวมเท่ากับ $90.01\% \pm 0.27\%$ ซึ่งถือว่าอยู่ในระดับสูงและสะท้อนถึงความสามารถของโมเดลในการจำแนกเที่ยวบินได้อย่างมีประสิทธิภาพ โดยผลการจำแนกเชิงรายละเอียดพบว่า คลาส ON TIME และ DELAY มีค่า Precision และ Recall สูงกว่า 88% ขึ้นไป แสดงให้เห็นว่าโมเดลสามารถทำนายเที่ยวบินที่ตรงเวลาและเที่ยวบินที่ล่าช้าได้อย่างแม่นยำ อย่างไรก็ตาม คลาส SHORT DELAY มีค่า Recall เพียง 11.44% ซึ่งสะท้อนถึงข้อจำกัดของโมเดลในการเรียนรู้และแยกแยะเที่ยวบินที่มีความล่าช้าในระดับสั้น ข้อจำกัดนี้อาจมีสาเหตุมาจากปัญหาความไม่สมดุลของข้อมูล (Class Imbalance) ระหว่างจำนวนตัวอย่างในแต่ละคลาส ส่งผลให้โมเดลไม่สามารถสร้างกฎที่เพียงพอสำหรับการจำแนกคลาส SHORT DELAY ได้อย่างแม่นยำ

accuracy: 90.01% +/- 0.27% (micro average: 90.01%)				
	true ON TIME	true DELAY	true SHORT DELAY	class precision
pred. ON TIME	204848	81	26871	88.37%
pred. DELAY	229	52892	61	99.45%
pred. SHORT DELAY	1564	199	3478	66.36%
class recall	99.13%	99.47%	11.44%	

ภาพที่ 3.19 Confusion Matrix และค่าตัวชี้วัดการจำแนกของโมเดล Decision Tree

PerformanceVector				
PerformanceVector:				
accuracy: 90.01% +/- 0.27% (micro average: 90.01%)				
ConfusionMatrix:				
True:	ON TIME	DELAY	SHORT DELAY	
ON TIME:	204848	81	26871	
DELAY:	229	52892	61	
SHORT DELAY:	1564	199	3478	

ภาพที่ 3.20 ผลการประเมินโมเดล Decision Tree ด้วย Confusion Matrix

ตารางที่ 3.5 Confusion Matrix ของผลการจำแนกของโมเดล Decision Tree

	ทำนายเป็น ON TIME	ทำนายเป็น DELAY	ทำนายเป็น SHORT DELAY
จริง = ON TIME	204,848	229	1,564
จริง = DELAY	81	52,892	199
จริง = SHORT DELAY	26,871	61	3,478

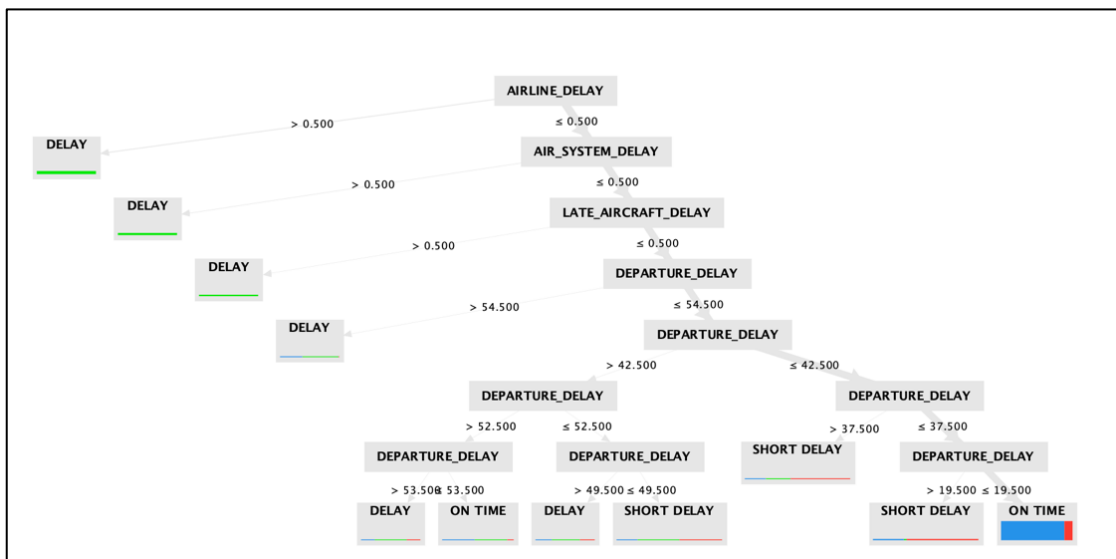
ตารางที่ 3.6 ค่าตัวชี้วัดประสิทธิภาพของโมเดล Decision Tree

ตัวชี้วัด (Metrics)	ค่า (Value)
Accuracy	90.01% ± 0.27%
Precision (ON TIME)	88.37%
Precision (DELAY)	99.45%
Precision (SHORT DELAY)	66.36%
Recall (ON TIME)	99.13%

Recall (DELAY)	99.47%
----------------	--------

Recall (SHORT DELAY)	11.44%
----------------------	--------

โครงสร้างต้นไม้ตัดสินใจพบว่าโมเดลเลือกใช้ตัวแปร AIRLINE_DELAY, AIR_SYSTEM_DELAY, LATE_AIRCRAFT_DELAY และ DEPARTURE_DELAY เป็นเงื่อนไขหลักในการจำแนกเที่ยวบิน โดยแต่ละโหนด (Node) แทนการแบ่งข้อมูลตามค่าของตัวแปร (Threshold) ผลการจำแนกที่ใบไม้ (Leaf Node) จะสรุปผลว่าเที่ยวบินนั้นอยู่ในกลุ่มใด ได้แก่ ON TIME, SHORT DELAY หรือ DELAY โดยแถบสีใต้แต่ละโหนดแสดงจำนวนตัวอย่างที่ถูกจัดให้อยู่ในแต่ละคลาส



ภาพที่ 3.21 โครงสร้างต้นไม้ตัดสินใจ (Decision Tree)

Tree

```

AIRLINE_DELAY > 0.500: DELAY {ON TIME=0, DELAY=28480, SHORT DELAY=0}
AIRLINE_DELAY ≤ 0.500
|   AIR_SYSTEM_DELAY > 0.500: DELAY {ON TIME=0, DELAY=16639, SHORT DELAY=0}
|   AIR_SYSTEM_DELAY ≤ 0.500
|   |   LATE_AIRCRAFT_DELAY > 0.500: DELAY {ON TIME=0, DELAY=7404, SHORT DELAY=0}
|   |   LATE_AIRCRAFT_DELAY ≤ 0.500
|   |   |   DEPARTURE_DELAY > 54.500: DELAY {ON TIME=191, DELAY=307, SHORT DELAY=6}
|   |   |   DEPARTURE_DELAY ≤ 54.500
|   |   |   |   DEPARTURE_DELAY > 42.500
|   |   |   |   |   DEPARTURE_DELAY > 52.500
|   |   |   |   |   |   DEPARTURE_DELAY > 53.500: DELAY {ON TIME=3, DELAY=7, SHORT DELAY=3}
|   |   |   |   |   |   DEPARTURE_DELAY ≤ 53.500: ON TIME {ON TIME=5, DELAY=5, SHORT DELAY=1}
|   |   |   |   |   |   DEPARTURE_DELAY ≤ 52.500
|   |   |   |   |   |   |   DEPARTURE_DELAY > 49.500: DELAY {ON TIME=11, DELAY=19, SHORT DELAY=10}
|   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 49.500: SHORT DELAY {ON TIME=32, DELAY=64, SHORT DELAY=65}
|   |   |   |   |   |   DEPARTURE_DELAY ≤ 42.500
|   |   |   |   |   |   |   DEPARTURE_DELAY > 37.500: SHORT DELAY {ON TIME=45, DELAY=53, SHORT DELAY=128}
|   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 37.500
|   |   |   |   |   |   |   |   DEPARTURE_DELAY > 19.500: SHORT DELAY {ON TIME=1535, DELAY=133, SHORT DELAY=3551}
|   |   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 19.500: ON TIME {ON TIME=204819, DELAY=61, SHORT DELAY=26646}

```

ภาพที่ 3.22 ผลการสร้างกฎจำแนกของโมเดล Decision Tree

กฎที่ได้จากโมเดล Decision Tree สามารถแจกแจงได้ดังนี้

กฎข้อที่ 1

IF AIRLINE_DELAY > 0.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากเกิดความล่าช้าที่เกี่ยวข้องกับสายการบินมากกว่า 0.5 นาที เที่ยวบินจะถูกจัดอยู่ในกลุ่ม “ล่าช้า”

กฎข้อที่ 2

IF AIRLINE_DELAY ≤ 0.5 AND AIR_SYSTEM_DELAY > 0.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากความล่าช้าของสายการบินไม่มาก แต่ระบบการบิน (Air System) มีความล่าช้ามากกว่า 0.5 นาที เที่ยวบินจะถูกจัดว่า “ล่าช้า”

กฎข้อที่ 3

IF AIRLINE_DELAY ≤ 0.5 AND AIR_SYSTEM_DELAY ≤ 0.5 AND

LATE_AIRCRAFT_DELAY > 0.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากไม่มีความล่าช้าจากสายการบินและระบบการบิน แต่เครื่องบินมาถึงล่าช้า จะส่งผลให้เที่ยวบินถูกจัดว่า “ล่าช้า”

กฎข้อที่ 4

IF LATE_AIRCRAFT_DELAY ≤ 0.5 AND DEPARTURE_DELAY > 54.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากไม่มีความล่าช้าจากเครื่องบินก่อนหน้า แต่เที่ยวบินออกช้ากว่า 54.5 นาที เที่ยวบินจะถูกจัดว่า “ล่าช้า”

กฎข้อที่ 5

IF LATE_AIRCRAFT_DELAY ≤ 0.5 AND DEPARTURE_DELAY ≤ 19.5

THEN ผลลัพธ์ = ON TIME หมายความว่า หากเที่ยวบินไม่มีความล่าช้าจากเครื่องบินก่อนหน้า และออกช้าหรือเร็วไม่เกิน 19.5 นาที จะถือว่าเที่ยวบินยังคง “ตรงเวลา”

กฎข้อที่ 6

IF LATE_AIRCRAFT_DELAY ≤ 0.5 AND DEPARTURE_DELAY อยู่ในช่วง 19.5 – 49.5 นาที

THEN ผลลัพธ์ = SHORT DELAY หมายความว่า หากเที่ยวบินออกช้าระหว่าง 20–50 นาที จะถูกจัดให้อยู่ในกลุ่ม “ล่าช้าสั้น”

กฎข้อที่ 7

IF LATE_AIRCRAFT_DELAY ≤ 0.5 AND DEPARTURE_DELAY > 49.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากเที่ยวบินออกช้ามากกว่า 50 นาทีขึ้นไป จะถูกจัดให้อยู่ในกลุ่ม “ล่าช้า”

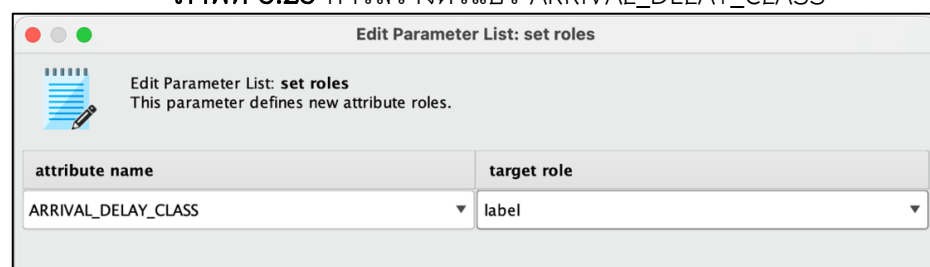
ค. เทคนิคป่าสุ่ม (Random Forest) การสร้างโมเดลการจำแนกประเภทด้วยเทคนิค Random Forest ได้ใช้ตัวแปรเป้าหมาย (Label) เช่นเดียวกับ Decision Tree คือ ARRIVAL_DELAY ซึ่งได้ถูกสร้างคุณลักษณะใหม่ (Generate Attribute) เพื่อจัดกลุ่มข้อมูลให้อยู่ในรูปแบบ Multiclass Classification โดยแบ่งเป็น 3 กลุ่ม ได้แก่

- ON TIME หมายถึง เมื่อค่า ARRIVAL_DELAY < 5 นาที

- SHORT DELAY หมายถึง เมื่อค่า ARRIVAL_DELAY อยู่ระหว่าง 5-14 นาที
- DELAY หมายถึง เมื่อค่า ARRIVAL_DELAY ≥ 15 นาที



ภาพที่ 3.23 การสร้างตัวแปร ARRIVAL_DELAY_CLASS



ภาพที่ 3.24 ขั้นตอนการกำหนดบทบาทของตัวแปร (Set Role)

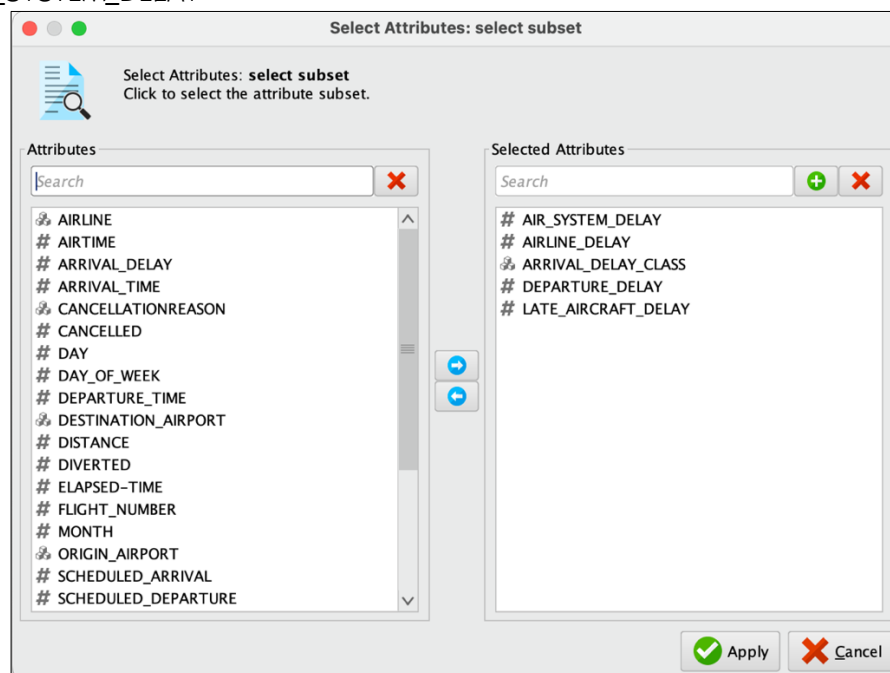
จากนั้นได้ทำการกำหนดบทบาทของตัวแปร (Set Role) โดยให้ ARRIVAL_DELAY_CLASS เป็นตัวแปรเป้าหมาย และเลือกใช้ตัวแปรอิสระ (Independent Variables) ที่ผ่านเกณฑ์การคัดเลือกด้วยค่า Correlation ≥ 0.4 ได้แก่

DEPARTURE_DELAY

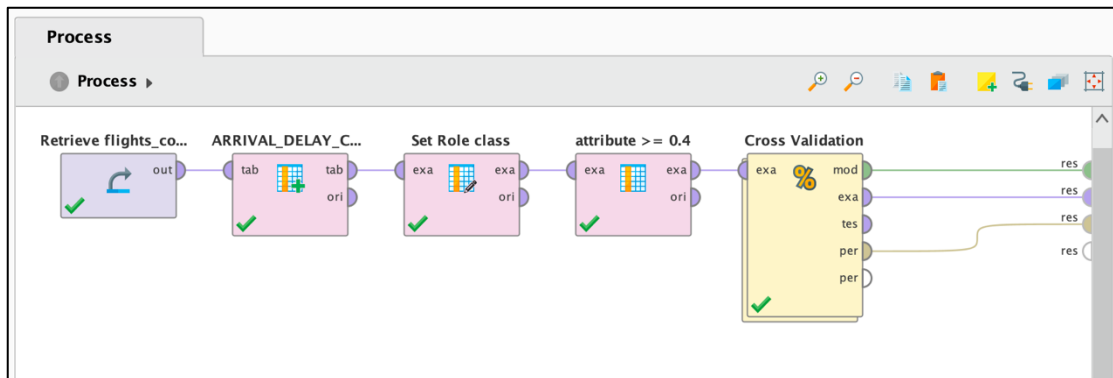
LATE_AIRCRAFT_DELAY

AIRLINE_DELAY

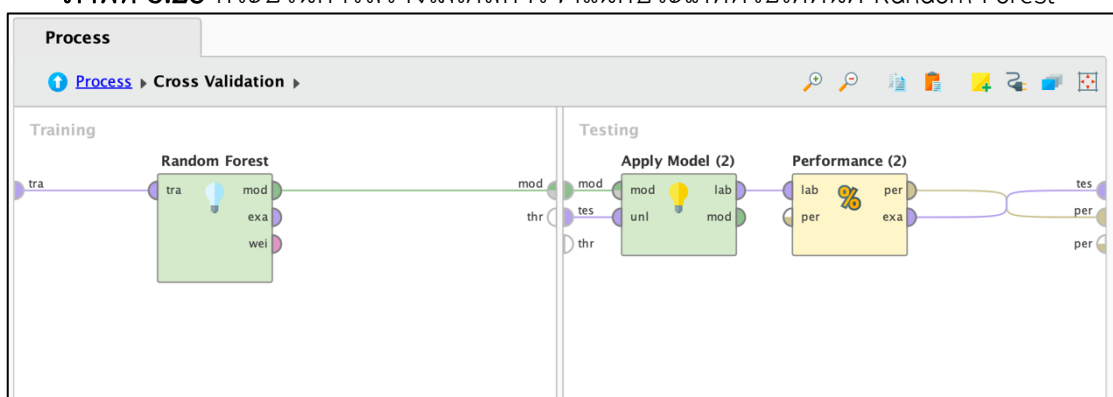
AIR_SYSTEM_DELAY



ภาพที่ 3.25 ขั้นตอนการเลือกตัวแปร (Select Attributes)



ภาพที่ 3.26 กระบวนการสร้างโมเดลการจำแนกประเภทด้วยเทคนิค Random Forest



ภาพที่ 3.27 กระบวนการเรียนรู้และประเมินผลโมเดล Random Forest ด้วย Cross Validation

โมเดล Random Forest แสดงค่า Accuracy รวมเท่ากับ $89.30\% \pm 0.03\%$ ซึ่งสะท้อนถึงประสิทธิภาพในการจำแนกเที่ยวบินได้ในระดับสูงและมีค่าความคลาดเคลื่อนมาตรฐานต่ำ แสดงให้เห็นถึงความเสถียรของโมเดลเมื่อทำการทดสอบซ้ำหลายครั้ง โดยผลการจำแนกในแต่ละคลาสพบว่า คลาส ON TIME และ DELAY มีค่า Precision และ Recall อยู่ในระดับสูง โดยเฉพาะคลาส DELAY ที่ให้ค่า Precision และ Recall ใกล้เคียง 100% แสดงถึงความสามารถของโมเดลในการทำนายเที่ยวบินที่ล่าช้าได้อย่างถูกต้องและแม่นยำ คลาส SHORT DELAY มีค่า Precision และ Recall เท่ากับ 0% แสดงให้เห็นว่าโมเดลไม่สามารถจำแนกเที่ยวบินที่มีความล่าช้าในระดับสั้นได้เลย ปัญหาดังกล่าวอาจเกิดจากความไม่สมดุลของข้อมูล (Class Imbalance) ระหว่างกลุ่ม ON TIME, DELAY และ SHORT DELAY ส่งผลให้การเรียนรู้ของโมเดลไม่สามารถสร้างกฎที่ชัดเจนสำหรับการจำแนกกลุ่ม SHORT DELAY ได้

accuracy: 89.30% +/- 0.03% (micro average: 89.30%)				
	true ON TIME	true DELAY	true SHORT DELAY	class precision
pred. ON TIME	206641	649	30410	86.93%
pred. DELAY	0	52523	0	100.00%
pred. SHORT DELAY	0	0	0	0.00%
class recall	100.00%	98.78%	0.00%	

ภาพที่ 3.28 Confusion Matrix และค่าตัวชี้วัดการจำแนกของโมเดล Random Forest

PerformanceVector

PerformanceVector:

accuracy: 89.30% +/- 0.03% (micro average: 89.30%)

ConfusionMatrix:

True: ON TIME DELAY SHORT DELAY
 ON TIME: 206620 629 30410
 DELAY: 21 52543 0
 SHORT DELAY: 0 0 0

ภาพที่ 3.29 ผลการประเมินโมเดล Random Forest ด้วย Confusion Matrix

ตารางที่ 3.7 Confusion Matrix ของผลการจำแนกของโมเดล Random Forest

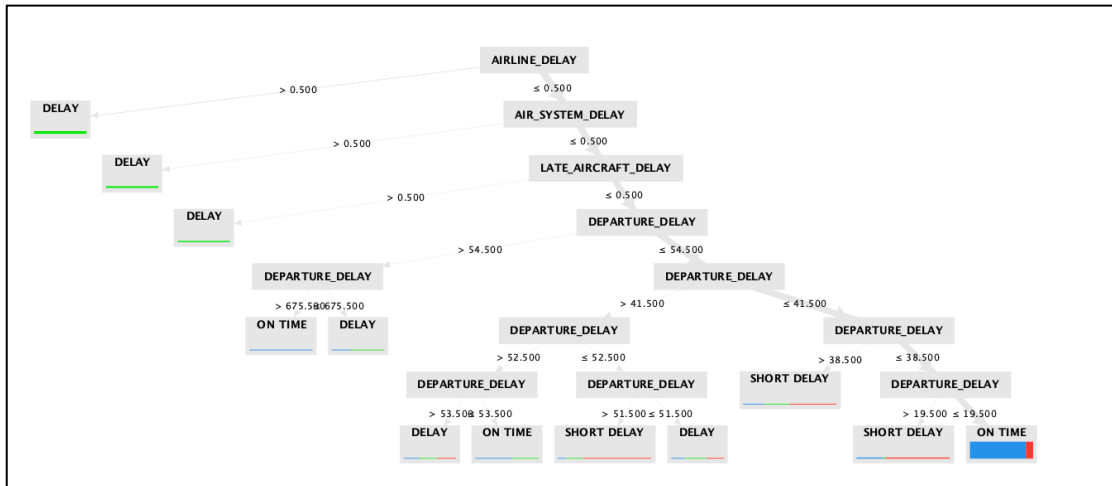
	ทำนายเป็น ON TIME	ทำนายเป็น DELAY	ทำนายเป็น SHORT DELAY
จริง = ON TIME	206,641	0	0
จริง = DELAY	649	52,523	0
จริง = SHORT DELAY	30,410	0	0

ตารางที่ 3.8 ค่าตัวชี้วัดประสิทธิภาพของโมเดล Random Forest

ตัวชี้วัด (Metrics)	ค่า (Value)
Accuracy	89.30% ± 0.03%
Precision (ON TIME)	86.93%
Precision (DELAY)	100.00%
Precision (SHORT DELAY)	0.00%
Recall (ON TIME)	100.00%
Recall (DELAY)	98.78%
Recall (SHORT DELAY)	0.00%

โครงสร้างต้นไม้ตัดสินใจที่เป็นส่วนหนึ่งของโมเดล Random Forest พบว่า โมเดล เลือกใช้ตัวแปร AIRLINE_DELAY, AIR_SYSTEM_DELAY, LATE_AIRCRAFT_DELAY และ DEPARTURE_DELAY เป็นเงื่อนไขหลักในการจำแนกเที่ยวบิน โดยในแต่ละโหนด (Node) แทน การแบ่งข้อมูลตามค่าเกณฑ์ (Threshold) ของตัวแปรที่เลือก เช่น DEPARTURE_DELAY ที่ มากกว่าหรือน้อยกว่าค่าที่กำหนด เพื่อแตกแขนงไปยังโหนดถัดไป

ผลการจำแนกปรากฏที่ใบไม้ (Leaf Node) โดยสรุปว่าเที่ยวบินนั้นจัดอยู่ในกลุ่ม ON TIME, SHORT DELAY หรือ DELAY ตามเงื่อนไขที่กำหนด แถบสีใต้แต่ละโหนดแสดงสัดส่วนจำนวนตัวอย่างที่ถูกจัดให้อยู่ในแต่ละคลาส ซึ่งสะท้อนถึงการกระจายของข้อมูลภายใต้กฎที่สร้างขึ้น



ภาพที่ 3.30 ตัวอย่างโครงสร้างต้นไม้ตัดสินใจในโมเดล Random Forest

Tree

```

AIRLINE_DELAY > 0.500: DELAY {ON TIME=0, DELAY=28160, SHORT DELAY=0}
AIRLINE_DELAY ≤ 0.500
|
|   AIR_SYSTEM_DELAY > 0.500: DELAY {ON TIME=0, DELAY=16514, SHORT DELAY=0}
|   AIR_SYSTEM_DELAY ≤ 0.500
|   |
|   |   LATE_AIRCRAFT_DELAY > 0.500: DELAY {ON TIME=0, DELAY=7452, SHORT DELAY=0}
|   |   LATE_AIRCRAFT_DELAY ≤ 0.500
|   |   |
|   |   |   DEPARTURE_DELAY > 54.500
|   |   |   |
|   |   |   |   DEPARTURE_DELAY > 675.500: ON TIME {ON TIME=1, DELAY=0, SHORT DELAY=0}
|   |   |   |   DEPARTURE_DELAY ≤ 675.500: DELAY {ON TIME=187, DELAY=314, SHORT DELAY=5}
|   |   |   |
|   |   |   |   DEPARTURE_DELAY ≤ 54.500
|   |   |   |   |
|   |   |   |   |   DEPARTURE_DELAY > 41.500
|   |   |   |   |   |
|   |   |   |   |   |   DEPARTURE_DELAY > 52.500
|   |   |   |   |   |   |
|   |   |   |   |   |   |   DEPARTURE_DELAY > 53.500: DELAY {ON TIME=5, DELAY=6, SHORT DELAY=6}
|   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 53.500: ON TIME {ON TIME=7, DELAY=5, SHORT DELAY=0}
|   |   |   |   |   |   |
|   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 52.500
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   DEPARTURE_DELAY > 51.500: SHORT DELAY {ON TIME=1, DELAY=2, SHORT DELAY=8}
|   |   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 51.500: DELAY {ON TIME=55, DELAY=88, SHORT DELAY=70}
|   |   |   |   |   |   |
|   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 41.500
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   DEPARTURE_DELAY > 38.500: SHORT DELAY {ON TIME=31, DELAY=36, SHORT DELAY=69}
|   |   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 38.500
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   DEPARTURE_DELAY > 19.500: SHORT DELAY {ON TIME=1526, DELAY=119, SHORT DELAY=3701}
|   |   |   |   |   |   |   |   |   DEPARTURE_DELAY ≤ 19.500: ON TIME {ON TIME=205136, DELAY=49, SHORT DELAY=26670}

```

ภาพที่ 3.31 ผลการสร้างกฎจำแนกของโมเดล Random Forest

กฎที่ได้จากโมเดล Random Forest สามารถแจกแจงได้ดังนี้

กฎข้อที่ 1

IF AIRLINE_DELAY > 0.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากเที่ยวบินมีความล่าช้าจากสายการบินมากกว่า 0.5 นาที เที่ยวบินจะถูกจัดอยู่ในกลุ่ม “ล่าช้า”

กฎข้อที่ 2

IF AIRLINE_DELAY ≤ 0.5 AND AIR_SYSTEM_DELAY > 0.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากสายการบินไม่ล่าช้า แต่ระบบการบินล่าช้ามากกว่า 0.5 นาที เที่ยวบินจะถูกจัดว่า “ล่าช้า”

กฎข้อที่ 3

IF AIRLINE_DELAY $\leq$ 0.5 AND AIR_SYSTEM_DELAY $\leq$ 0.5 AND
LATE_AIRCRAFT_DELAY $>$ 0.5

THEN ผลลัพธ์ = DELAY หมายความว่า หากไม่มีความล่าช้าจากสายการบินและระบบการ
บิน แต่เครื่องบินเข้ามีความล่าช้ามากกว่า 0.5 นาที เที่ยวบินจะถูกจัดว่า “ล่าช้า”

กฎข้อที่ 4

IF LATE_AIRCRAFT_DELAY $\leq$ 0.5 AND DEPARTURE_DELAY $\leq$ 19.5

THEN ผลลัพธ์ = ON TIME หมายความว่า หากไม่มีความล่าช้าจากเครื่องบินเข้า และ
เวลาออกเดินทางล่าช้าหรือเร็วไม่เกิน 19.5 นาที เที่ยวบินจะถูกจัดว่า “ตรงเวลา”

กฎข้อที่ 5

IF LATE_AIRCRAFT_DELAY $\leq$ 0.5 AND DEPARTURE_DELAY อยู่ระหว่าง 19.5 – 38.5
นาที

THEN ผลลัพธ์ = SHORT DELAY หมายความว่า หากเที่ยวบินออกช้าในช่วง 20–40 นาที
จะถูกจัดให้อยู่ในกลุ่ม “ล่าช้าสั้น”

กฎข้อที่ 6

IF LATE_AIRCRAFT_DELAY $\leq$ 0.5 AND DEPARTURE_DELAY อยู่ระหว่าง 38.5 – 52.5
นาที

THEN ผลลัพธ์ = SHORT DELAY หมายความว่า หากเที่ยวบินออกช้าในช่วง 40–50 นาที
จะถูกจัดให้อยู่ในกลุ่ม “ล่าช้าสั้น”

กฎข้อที่ 7

IF LATE_AIRCRAFT_DELAY $\leq$ 0.5 AND DEPARTURE_DELAY $>$ 52.5

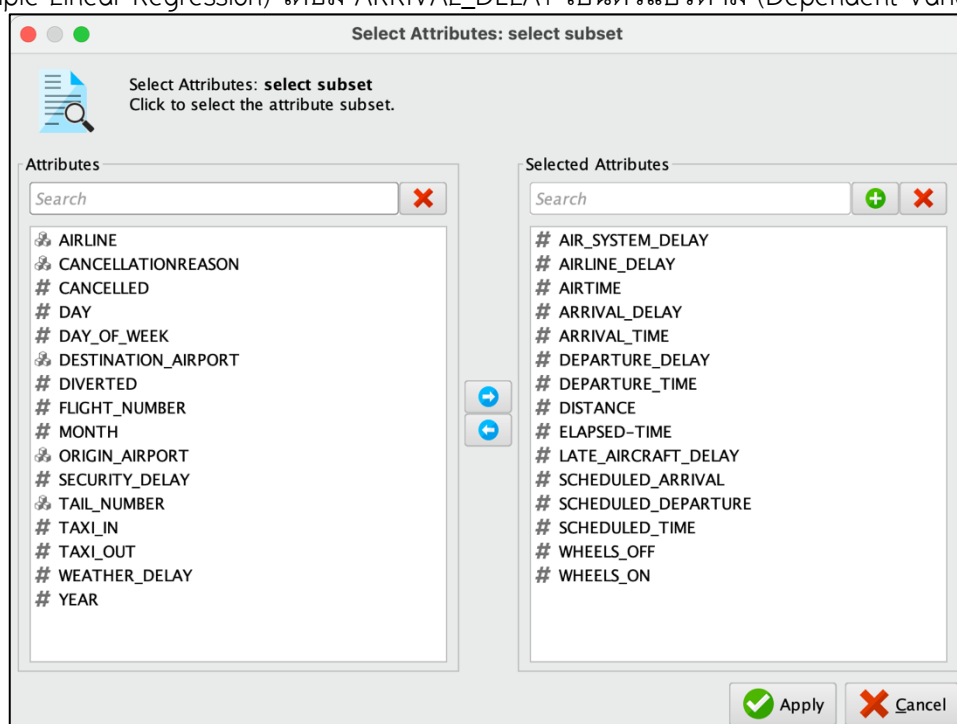
THEN ผลลัพธ์ = DELAY หมายความว่า หากเที่ยวบินออกช้าเกินกว่า 52.5 นาทีขึ้นไป
เที่ยวบินจะถูกจัดว่า “ล่าช้า”

- ง. เทคนิคการวิเคราะห์การถดถอยเชิงเส้นหลายตัวแปร (Multiple Linear Regression) เทคนิคการถดถอยเชิงเส้นหลายตัวแปร (Multiple Linear Regression: MLR) เป็นวิธีการวิเคราะห์ทางสถิติที่ใช้ในการสร้างแบบจำลองความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) กับตัวแปรอิสระ (Independent Variables) มากกว่าหนึ่งตัว โดยมีวัตถุประสงค์เพื่อตรวจสอบและทำนายค่าของตัวแปรตามจากการเปลี่ยนแปลงของตัวแปรอิสระ ผู้จัดทำได้ทำการเลือกใช้ตัวแปรอิสระที่มีค่าสหสัมพันธ์ (Correlation Coefficient) ตั้งแต่ 0.4 ขึ้นไป ได้แก่

DEPARTURE_DELAY
LATE_AIRCRAFT_DELAY
AIRLINE_DELAY
AIR_SYSTEM_DELAY
SCHEDULED_TIME
DISTANCE
AIRTIME

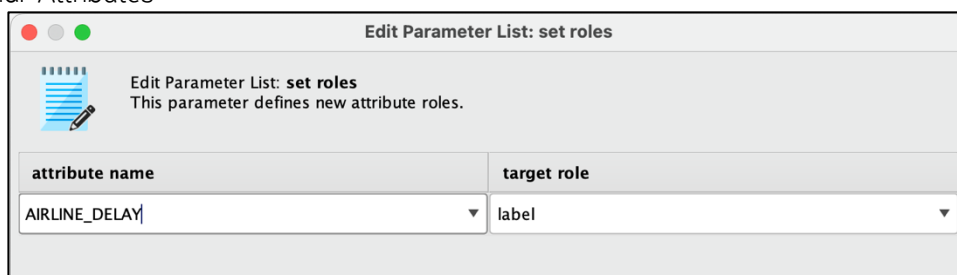
ELAPSED_TIME
 DEPARTURE_TIME
 SCHEDULED_DEPARTURE
 ARRIVAL_TIME
 SCHEDULED_ARRIVAL
 WHEELS_OFF
 WHEELS_ON

ตัวแปรดังกล่าวถูกคัดเลือกจากการวิเคราะห์ Correlation Matrix และถูกนำมาใช้เป็นตัวแปรอิสระ (Independent Variables) สำหรับสร้างแบบจำลองการถดถอยเชิงเส้นหลายตัวแปร (Multiple Linear Regression) โดยมี ARRIVAL_DELAY เป็นตัวแปรตาม (Dependent Variable)

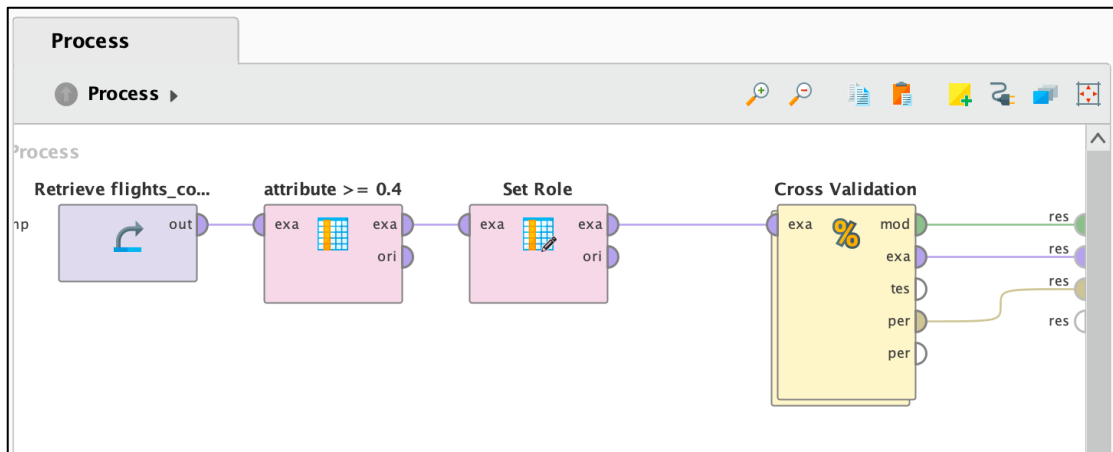


ภาพที่ 3.32 การเลือกตัวแปร (Select Attributes) สำหรับโมเดล Multiple Linear Regression

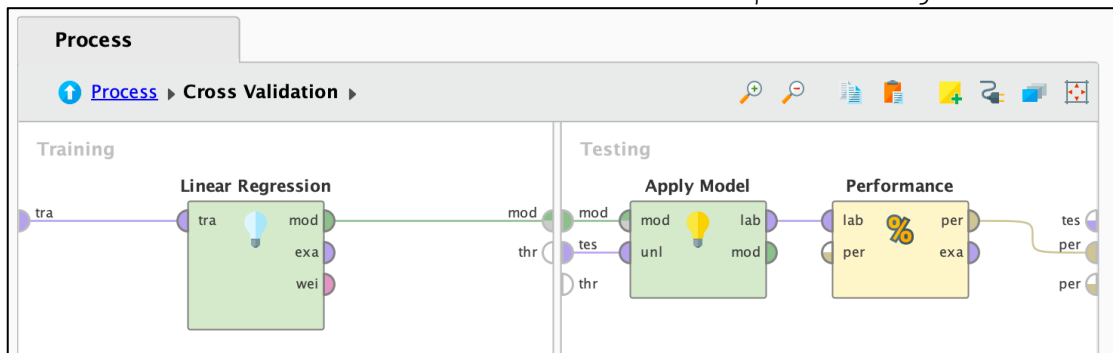
การกำหนดบทบาทของตัวแปรเป้าหมาย (Target Variable) โดยใช้ Operator Set Role เพื่อกำหนดให้ ARRIVAL_DELAY ทำหน้าที่เป็น Label ซึ่งเป็นตัวแปรตาม (Dependent Variable) ที่ต้องการทำนายในโมเดล Multiple Linear Regression ส่วนตัวแปรอื่น ๆ ถูกกำหนดเป็น Regular Attributes



ภาพที่ 3.33 การกำหนดบทบาทของตัวแปร (Set Role) ใน RapidMiner



ภาพที่ 3.34 กระบวนการสร้างและประเมินโมเดล Multiple Linear Regression



ภาพที่ 3.35 กระบวนการเรียนรู้และประเมินผลโมเดล ด้วย Cross Validation

ผลการประเมินโมเดล Multiple Linear Regression พบว่าค่า Root Mean Squared Error (RMSE) มีค่าเท่ากับ 9.875 ± 0.592 โดยมีค่าเฉลี่ยแบบ micro average เท่ากับ 9.891 ± 0.000 ผลลัพธ์ดังกล่าวสะท้อนว่าโมเดลสามารถทำนายค่าความล่าช้าในการมาถึงของเที่ยวบิน (ARRIVAL_DELAY) ได้ค่อนข้างใกล้เคียงกับค่าจริง โดยมีความคลาดเคลื่อนเฉลี่ยอยู่ที่ประมาณ 10 นาที ซึ่งถือว่าอยู่ในระดับที่ยอมรับได้สำหรับการสร้างแบบจำลองเชิงทำนายจากข้อมูลเที่ยวบินจริง ค่าของ RMSE แสดงให้เห็นถึงประสิทธิภาพของโมเดลในเชิงการทำนายที่มีความเสถียร เนื่องจากค่าความคลาดเคลื่อนมาตรฐาน (± 0.592) อยู่ในระดับต่ำ สะท้อนว่าโมเดลสามารถสร้างผลลัพธ์ที่มีความสม่ำเสมอเมื่อทำการทดสอบซ้ำหลายครั้ง อย่างไรก็ตาม แม้โมเดลจะมีความสามารถในการทำนายเชิงตัวเลขได้ดี แต่การตีความค่าที่ได้ควรสัมพันธ์กับลักษณะของข้อมูล โดยเฉพาะการกระจายตัวของค่าความล่าช้าในชุดข้อมูลเที่ยวบินที่มีช่วงเวลาแตกต่างกัน

root_mean_squared_error

root_mean_squared_error: 9.875 +/- 0.592 (micro average: 9.891 +/- 0.000)

ภาพที่ 3.36 ผลการประเมินโมเดล Multiple Linear Regression ด้วยค่า Root Mean Squared Error (RMSE)

PerformanceVector

PerformanceVector:

root_mean_squared_error: 9.875 +/- 0.592 (micro average: 9.891 +/- 0.000)

ภาพที่ 3.37 ผลการประเมินโมเดล Multiple Linear Regression ในรูปแบบ Performance Vector

3.1.5 การวัดประสิทธิภาพของโมเดล (Evaluation)

การวัดประสิทธิภาพของโมเดลเป็นขั้นตอนสำคัญในการประเมินความสามารถของอัลกอริทึมที่นำมาใช้ในการจำแนกหรือทำนายผลลัพธ์ โดยมีวัตถุประสงค์เพื่อเปรียบเทียบความแม่นยำ ความสม่ำเสมอ และความเหมาะสมของโมเดลแต่ละประเภทที่นำมาใช้ในโครงการ ทั้งนี้วิธีการวัดประสิทธิภาพจะแตกต่างกันไปตามลักษณะของปัญหา กล่าวคือ หากเป็นปัญหาการจำแนกประเภท (Classification) จะใช้ตัวชี้วัดที่แตกต่างจากปัญหาการถดถอย (Regression)

1. การวัดประสิทธิภาพของโมเดลจำแนกประเภท (Classification Models) สำหรับโมเดลที่ใช้จำแนกเที่ยวบินออกเป็นกลุ่ม ON TIME, SHORT DELAY และ DELAY ได้แก่ Decision Tree และ Random Forest การประเมินประสิทธิภาพดำเนินการโดยใช้ Confusion Matrix และตัวชี้วัดดังนี้

- Accuracy วัดสัดส่วนของการทำนายที่ถูกต้องทั้งหมด
- Precision วัดความถูกต้องของการทำนายแต่ละคลาสเมื่อโมเดลทำนายว่าเป็นคลาสนั้น
- Recall (Sensitivity) วัดความสามารถของโมเดลในการจับข้อมูลของคลาสนั้น ๆ ได้ครบถ้วน
- F1-score ค่ากลางเชิงฮาร์โมนิกของ Precision และ Recall ใช้ประเมินประสิทธิภาพเมื่อข้อมูลไม่สมดุล

ผลลัพธ์จากโมเดล Decision Tree และ Random Forest ได้ถูกนำเสนอในรูปของ Confusion Matrix และตารางสรุปค่าชี้วัด เพื่อเปรียบเทียบความสามารถของแต่ละโมเดล

ตารางที่ 3.9 การวัดประสิทธิภาพของโมเดล Classification

โมเดล (Models)	Accuracy	Precision	Recall	F1-score
Decision Tree				
Random Forest				

2. การวัดประสิทธิภาพของโมเดลถดถอย (Regression Model) สำหรับโมเดล Multiple Linear Regression ที่ใช้ในการทำนายค่าความล่าช้าในการมาถึงของเที่ยวบิน (ARRIVAL_DELAY) ซึ่งเป็นข้อมูลเชิงตัวเลข การวัดประสิทธิภาพจะใช้ตัวชี้วัดทางสถิติดังนี้

- Root Mean Squared Error (RMSE) ค่าความคลาดเคลื่อนเฉลี่ยที่ถ่วงน้ำหนักด้วยกำลังสอง แสดงถึงความแม่นยำของโมเดลในการทำนายค่าเชิงปริมาณ

- Mean Absolute Error (MAE) ค่าความคลาดเคลื่อนเฉลี่ยเชิงสัมบูรณ์ ใช้วัดความผิดพลาดโดยไม่คำนึงถึงทิศทาง
- Coefficient of Determination (R^2) แสดงสัดส่วนของความแปรปรวนในตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแปรอิสระในโมเดล

ตารางที่ 3.10 การวัดประสิทธิภาพของโมเดล Regression

โมเดล (Models)	RMSE	MAE	R^2
Multiple Linear Regression			

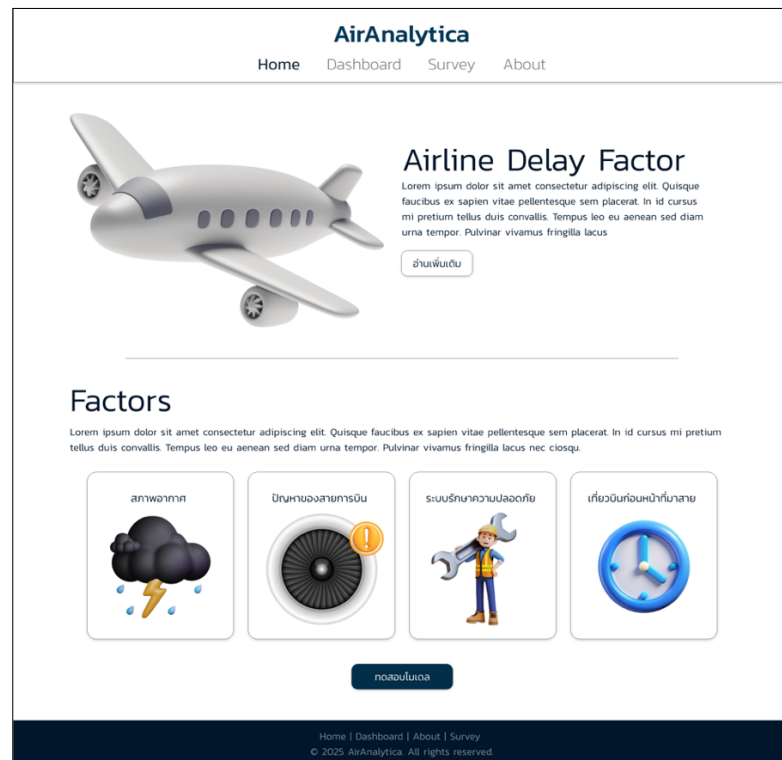
3.1.6 การนำโมเดลไปใช้งานจริง (Deployment)

การนำโมเดลไปใช้งานจริง (Deployment) ในงานวิจัยนี้ เป็นขั้นตอนสุดท้ายหลังจากทำการวิเคราะห์ เปรียบเทียบ และเลือกโมเดลที่มีประสิทธิภาพดีที่สุด โดยอาศัยข้อมูลเที่ยวบินที่ผ่านการประมวลผลในโปรแกรม RapidMiner จากนั้นนำผลลัพธ์การทำนายและการวิเคราะห์มาจัดทำเป็น แดชบอร์ด (Dashboard) ด้วยโปรแกรม Tableau เพื่อแสดงผลข้อมูลเชิงภาพ (Data Visualization) ให้เข้าใจง่ายและนำไปใช้ประกอบการตัดสินใจได้อย่างมีประสิทธิภาพ และได้พัฒนา เว็บไซต์แสดงผลการวิเคราะห์ โดยเชื่อมโยงกราฟและแดชบอร์ดที่สร้างจาก Tableau เพื่อนำเสนอข้อมูลเชิงโต้ตอบ (Interactive Dashboard)

3.2 การออกแบบเว็บไซต์

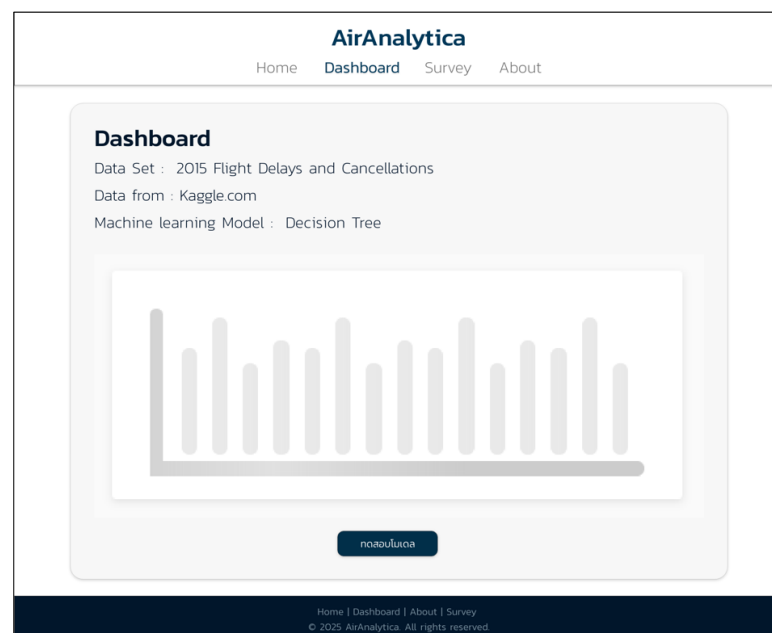
การออกแบบเว็บไซต์ (Website Design) ในงานวิจัยนี้ มีวัตถุประสงค์เพื่อสร้างเครื่องมือในการนำเสนอผลการวิเคราะห์ข้อมูลการล่าช้าเที่ยวบินให้สามารถเข้าถึงได้ง่าย สะดวก และมีความน่าสนใจ โดยเว็บไซต์จะทำหน้าที่เป็นช่องทางเผยแพร่ผลลัพธ์จากการประมวลผลใน RapidMiner และการสร้าง Dashboard เชิงโต้ตอบจาก Tableau เพื่อให้ผู้ใช้งานสามารถสำรวจและทำความเข้าใจกับข้อมูลได้ด้วยตนเอง

3.2.1 หน้าแรก (Index Page) แสดงข้อมูลภาพรวมของโครงการในหัวข้อ Airline Delay Factor พร้อมปัจจัยที่อาจส่งผลกระทบต่อความล่าช้าของเที่ยวบิน เช่น สภาพอากาศ ปัญหาของสายการบิน ระบบรักษาความปลอดภัย และเที่ยวบินก่อนหน้าที่ล่าช้า โดยออกแบบให้ผู้ใช้งานสามารถรับรู้ปัจจัยเบื้องต้นที่เกี่ยวข้องกับการล่าช้าได้ทันทีที่เข้าสู่เว็บไซต์



ภาพที่ 3.38 หน้าแรก (Index Page) ของเว็บไซต์ AirAnalytica

3.2.2 หน้าแสดง Dashboard แสดงผลการวิเคราะห์ข้อมูลเที่ยวบินโดยใช้ชุดข้อมูล 2015 Flight Delays and Cancellations จาก Kaggle ผ่าน Tableau Dashboard ในรูปแบบกราฟเชิงโต้ตอบ เพื่อให้ผู้ใช้งานเข้าใจข้อมูลได้ง่ายและสามารถตรวจสอบแนวโน้มการล่าช้าได้ด้วยตนเอง



ภาพที่ 3.39 หน้าแสดง Dashboard ของเว็บไซต์ AirAnalytica

3.2.3 หน้าแบบสอบถาม (Survey) สำหรับเก็บข้อมูลจากผู้ใช้งานผ่านแบบสอบถาม โดยผู้ใช้งานสามารถเลือกหรือตอบข้อมูลตามที่ระบบกำหนดไว้ เช่น ปัจจัยที่อาจมีผลต่อการ

ล่าช้าของเที่ยวบิน ข้อมูลที่ได้จะถูกนำไปใช้ในการประมวลผลและแสดงผลลัพธ์สรุปในหน้าถัดไป

AirAnalytica

[Home](#)
[Dashboard](#)
[Survey](#)
[About](#)

แบบสอบถาม

Survey

1. เที่ยวบินของคุณออกเดินทางในช่วงเวลาใด?

00:00–06:00 (กลางคืน)

06:01–12:00 (เช้า)

12:01–18:00 (บ่าย)

18:01–23:59 (เย็น)

2. วันที่เดินทางของคุณตรงกับวันใดในสัปดาห์?

อาทิตย์ (Sunday)

▼

3. สภาพอากาศในวันเดินทางเป็นอย่างไร?

☀️

ปกติ

☁️⚡️

ฝนตก

☁️

หมอก/ควัน

💨

พายุ/ลมแรง

❄️

หิมะตกหนัก

👉

อื่น ๆ

4. เที่ยวบินของคุณเคยมีความล่าช้าในการออกเดินทางหรือไม่?

ไม่ล่าช้า

ล่าช้าเล็กน้อย (ไม่เกิน 15 นาที)

ล่าช้าปานกลาง (15–60 นาที)

ล่าช้ามาก (มากกว่า 60 นาที)

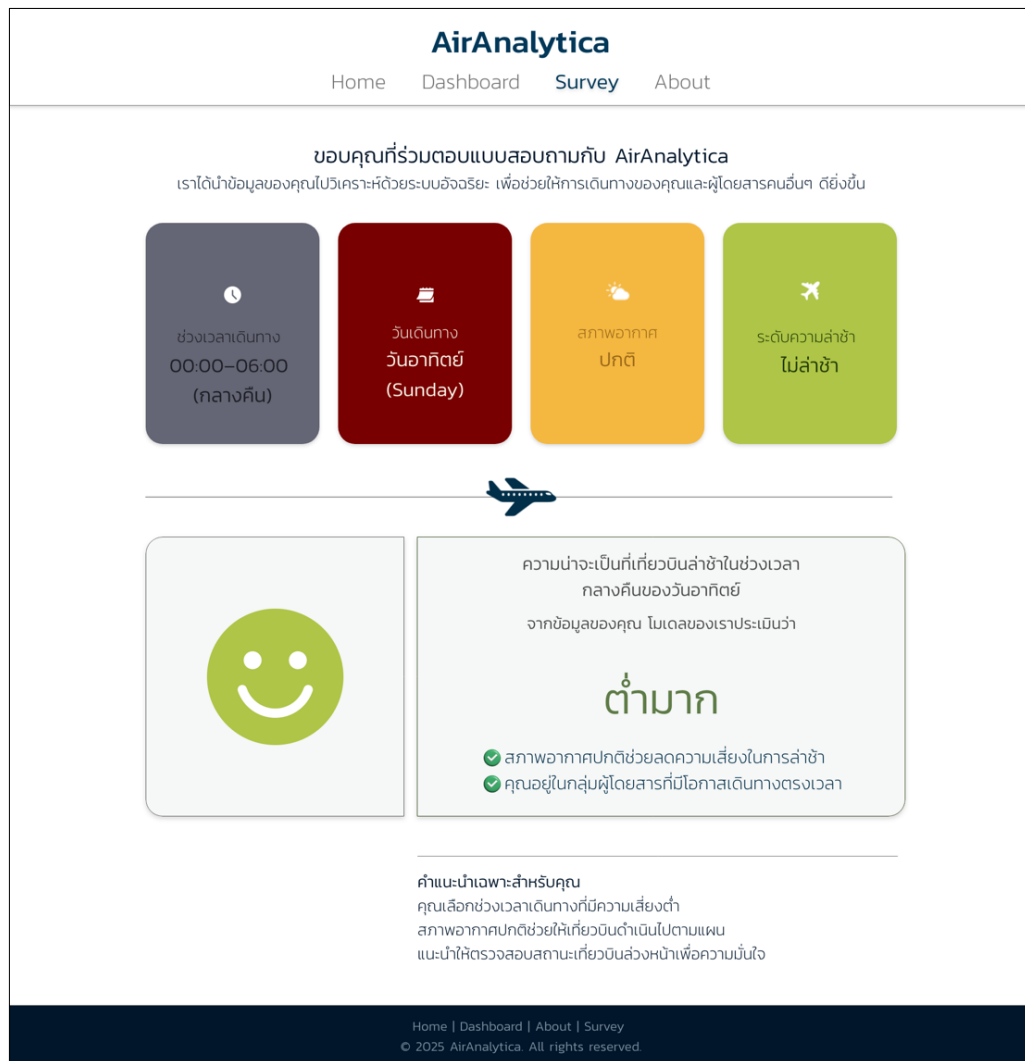
Submit

Home | Dashboard | About | Survey

© 2025 AirAnalytica. All rights reserved.

ภาพที่ 3.40 หน้าแบบสอบถาม (Survey) ของเว็บไซต์ AirAnalytica

3.2.4 หน้าสรุปผล (Result) สรุปผลหลังผู้ใช้งานตอบแบบสอบถาม โดยระบบนำข้อมูลที่กรอกมาไปประมวลผลและแสดงผลลัพธ์ในลักษณะภาพรวม



ภาพที่ 3.41 หน้าสรุปผล (Result) ของเว็บไซต์ AirAnalytica

3.2.5 หน้าเกี่ยวกับฉัน (About) ให้ข้อมูลเกี่ยวกับวัตถุประสงค์ของเว็บไซต์ แนวคิดในการพัฒนา และผู้จัดทำ พร้อมช่องทางการติดต่อสำหรับผู้สนใจหรือมีข้อเสนอแนะ โดยออกแบบให้เรียบง่ายและชัดเจน เพื่อสร้างความน่าเชื่อถือและให้ผู้ใช้งานสามารถเข้าถึงข้อมูลได้สะดวก



ภาพที่ 3.42 หน้าเกี่ยวกับฉัน (About) ของเว็บไซต์ AirAnalytica สำหรับแสดงข้อมูลโครงการ
ผู้พัฒนา และช่องทางการติดต่อ

3.3 บทสรุป

จากกระบวนการดำเนินโครงการที่กล่าวมาในบทนี้ สามารถสรุปได้ว่า การวิเคราะห์ข้อมูลความล่าช้าของเที่ยวบินโดยใช้เทคนิคเหมืองข้อมูล (Data Mining) ตามกระบวนการ CRISP-DM ได้ถูกดำเนินการอย่างเป็นระบบ เริ่มตั้งแต่ขั้นตอนการทำความเข้าใจปัญหาเชิงธุรกิจและข้อมูล (Business and Data Understanding) การเตรียมข้อมูล (Data Preparation) การสร้างโมเดล (Modeling) การประเมินประสิทธิภาพ (Evaluation) จนถึงการนำโมเดลไปใช้งานจริง (Deployment) ในการสร้างแบบจำลอง (Modeling) ได้เลือกใช้สองแนวทางหลัก ได้แก่ โมเดลการจำแนกประเภท (Classification Models) คือ Decision Tree และ Random Forest ซึ่งมีเป้าหมายเพื่อทำนายสถานะเที่ยวบินว่าอยู่ในกลุ่ม ON TIME, SHORT DELAY หรือ DELAY และโมเดลการถดถอยเชิงเส้นหลายตัวแปร (Multiple Linear Regression) เพื่อทำนายค่าความล่าช้าในการมาถึงของเที่ยวบินในเชิงปริมาณ ผลการประเมินโมเดลด้วยวิธี Cross Validation พบว่า Decision Tree และ Random Forest สามารถจำแนกเที่ยวบินในคลาส ON TIME และ DELAY ได้อย่างแม่นยำ แต่ยังมีข้อจำกัดในการจำแนกคลาส SHORT DELAY ขณะที่ Multiple Linear Regression สามารถทำนายค่าความล่าช้าเชิงตัวเลขได้ใกล้เคียงกับค่าจริง โดยมีค่า RMSE ประมาณ 10 นาที ซึ่งถือว่าอยู่ในระดับที่ยอมรับได้สำหรับการวิเคราะห์เชิงพยากรณ์ ในส่วนของการนำโมเดลไปใช้งานจริง (Deployment) ผลลัพธ์จากการวิเคราะห์และการทำนายถูกนำเสนอผ่าน แดชบอร์ด (Dashboard) บน Tableau และ เว็บไซต์ AirAnalytica ซึ่งออกแบบมาเพื่อแสดงข้อมูลเชิงภาพและผลการวิเคราะห์ในรูปแบบที่เข้าใจง่าย และสามารถโต้ตอบได้ (Interactive)