บทที่ 3

วิธีดำเนินงาน

โครงงานเรื่อง การวิเคราะห์ข้อมูลเพื่อทำนายลักษณะของผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด ในบทนี้จะเป็นการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล (Data Mining) ซึ่งมีกระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM ที่สำคัญหลากหลายขั้นตอน เมื่อเสร็จสิ้นจากกระบวนการวิเคราะห์ข้อมูลแล้วจะเป็นขั้นตอนการออกแบบเว็บไซต์ และการออกแบบรูปแบบการแสดงผล (Dashboard) และบทสรุปจากวิธีดำเนินงานดังนี้

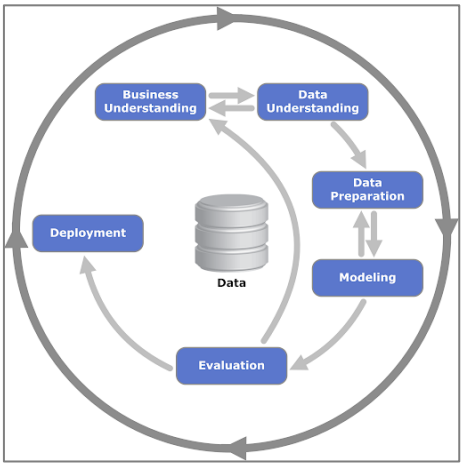
3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM

3.2 การออกแบบเว็บไซต์

3.3 บทสรุป

3.1 การวิเคราะห์ข้อมูลด้วย CRISP-DM

กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM หรือ Cross Industry Standard Process for Data Mining พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือของ 3 บริษัทคือ Daimler Chrysler, SPSS และ NCR ที่มีการพัฒนาเป็น Work flow มาตรฐานสำหรับการทำเหมืองข้อมูล ประกอบด้วย 6 ขั้นตอนหลัก ดังนี้



ภาพที่ 3.1 กระบวนการวิเคราะห์ข้อมูล ด้วย CRISP-DM

(ที่มา: https://n9.cl/qoaaw)

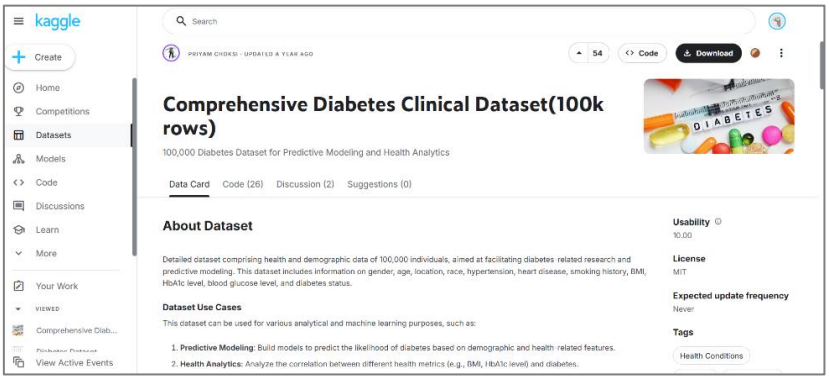
3.1.1 กระบวนการทำความเข้าใจธุรกิจ (Business Understanding)

เป็นขั้นตอนแรกของกระบวนการที่มุ่งเน้นไปที่การทำความเข้าใจกระบวนการทางธุรกิจโดยรวม ผู้วิเคราะห์ข้อมูลทำความเข้าใจกับปัญหาให้อยู่ในรูปของการวิเคราะห์ข้อมูลทาง Data Mining โดยมีเป้าหมายกำหนดวัตถุประสงค์ของโครงการ ซึ่งก็คือการทำนายลักษณะของผู้ป่วยโรคเบาหวานโดยใช้ข้อมูลทุติยภูมิ ระบุปัจจัยเสี่ยงที่มีผลต่อการเกิดโรคเบาหวาน เช่น อายุ BMI ความดันโลหิต ประวัติการสูบบุหรี่ และระดับน้ำตาลในเลือด และนำผลที่ได้ไปช่วยในการวินิจฉัยและให้คำแนะนำกับบุคคลทั่วไปที่มีความเสี่ยงเป็นโรคเบาหวาน โดยใช้ข้อมูลจาก Kaggle ซึ่งเป็นแหล่งข้อมูลทุติยภูมิ วิเคราะห์ข้อมูลโดยใช้การจำแนกประเภท (Classification) ซึ่งมีการใช้เทคนิค Decision Tree, Naive Bayes, Random Forest, Support Vector Machine และ Artificial Neural Network ในการวิเคราะห์ข้อมูล

3.1.2 การทำความเข้าใจข้อมูล (Data Understanding)

ขั้นตอนการจัดเก็บและรวบรวมข้อมูล ตลอดจนการพิจารณาตรวจสอบความถูกต้องของข้อมูลที่ได้รับ โดยเลือกว่าจะใช้ข้อมูลทั้งหมดหรือบางส่วนในการวิเคราะห์ให้สอดรับกับวัตถุประสงค์ที่กำหนดไว้

ผู้วิเคราะห์ข้อมูลทำการรวบรวมข้อมูล เพื่อตรวจสอบรายละเอียด ปริมาณ และ ความน่าเชื่อถือของชุดข้อมูลโรคเบาหวาน ที่ได้จากเว็บไซต์ kaggle.com ซึ่งเป็นเว็บไซต์ที่เก็บรวบรวมชุดข้อมูลต่างๆ เป็นแหล่งรวม Datasets หรือชุดข้อมูลสำหรับฝึกสอน Machine Learning ที่ใหญ่ที่สุดในโลกแห่งหนึ่ง มีข้อมูลทุกประเภท มีทั้ง Datasets ในหมวดหมู่ Finance, Business, Physics, Biology, Sports, News ซึ่งเป็นข้อมูลที่เปิดเผยได้ เพื่อให้ผู้ใช้บริการสามารถนำชุดข้อมูลไปศึกษาหรือวิเคราะห์ให้เกิดประโยชน์ต่อไป



ภาพที่ 3.2 เว็บไซต์ kaggle.com

(ที่มา: https://www.kaggle.com/)

ซึ่งข้อมูลปัญหาโรคเบาหวาน มีจำนวนข้อมูลทั้งหมด 100,000 รายการ ประกอบด้วย 16 แอตทริบิวท์ ดังนี้ ปี (Year) เพศ (Gender) อายุ (Age) สถานที่ (Location) เชื้อชาติ: แอฟริกันอเมริกัน (race:AfricanAmerican) เชื้อชาติ: เอเชีย (race:Asian) เชื้อชาติ: คอเคเซียน (race:Caucasian) เชื้อชาติ: ฮิสแปนิก (race:Hispanic) เชื้อชาติ: อื่นๆ (race:Other) ความดันโลหิตสูง (Hypertension) ประวัติโรคหัวใจ (Heart_disease) ประวัติการสูบบุหรี่ (Smoking_history) ค่าดัชนีมวลกาย (BMI) ระดับน้ำตาลสะสมเฉลี่ยในเลือด (hbA1c_level) ระดับน้ำตาลในเลือด (blood_glucose_level) และการเป็นโรคเบาหวาน (diabetes)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
year	gender	age	location	race:Africa	race:Asian	race:Cauc	race:Hispa	race:Othe	hypertensi	heart_dise	smoking_h	bmi	hbA1c_lev	blood_glu	diabetes
2020	Female	32	Alabama	0	0	0	0	1	0	0	never	27.32	5	100	0
2015	Female	29	Alabama	0	1	0	0	0	0	0	never	19.95	5	90	0
2015	Male	18	Alabama	0	0	0	0	1	0	0	never	23.76	4.8	160	0
2015	Male	41	Alabama	0	0	1	0	0	0	0	never	27.32	4	159	0
2016	Female	52	Alabama	1	0	0	0	0	0	0	never	23.75	6.5	90	0
2016	Male	66	Alabama	0	0	1	0	0	0	0	not curren	27.32	5.7	159	0
2015	Female	49	Alabama	0	0	1	0	0	0	0	current	24.34	5.7	80	0
2016	Female	15	Alabama	0	0	0	0	1	0	0	No Info	20.98	5	155	0
2016	Male	51	Alabama	1	0	0	0	0	0	0	never	38.14	6	100	0
2015	Male	42	Alabama	0	0	1	0	0	0	0	No Info	27.32	5.7	160	0
2016	Male	15	Alabama	1	0	0	0	0	0	0	No Info	19.15	6.6	200	0
2016	Female	53	Alabama	0	0	1	0	0	0	0	never	34.3	6.6	155	0
2015	Female	3	Alabama	1	0	0	0	0	0	0	No Info	20.28	3.5	159	0
2016	Female	40	Alabama	0	0	0	1	0	0	0	never	27.63	6.5	126	0
2016	Female	64	Alabama	0	0	0	0	1	0	0	ever	49.27	8.2	140	1
2015	Female	23	Alabama	0	0	1	0	0	0	0	never	27.32	6	140	0
2015	Female	2	Alabama	0	1	0	0	0	0	0	never	27.32	3.5	155	0
2015	Male	80	Alabama	0	0	0	0	1	0	0	No Info	27.32	6	130	0
2016	Female	12	Alabama	0	0	0	1	0	0	0	never	20.78	6.5	130	0
2015	Female	50	Alabama	0	1	0	0	0	0	0	never	26.39	6.6	158	0
2015	Female	69	Alabama	0	0	0	1	0	0	0	No Info	31.76	6.5	160	0
2016	Female	75	Alabama	0	1	0	0	0	0	0	never	21.3	6	158	0
2016	Male	42	Alabama	1	0	0	0	0	0	0	never	27.32	6	160	0
2016	Male	72	Alabama	0	0	0	0	1	0	0	not curren	20.04	6.2	90	0
2016	Female	46	Alabama	0	1	0	0	0	0	0	current	19.81	5.7	160	0
2016	Female	47	Alabama	1	0	0	0	0	0	0	No Info	22.61	6.6	90	0
2015	Male	14	Alabama	1	0	0	0	0	0	0	never	23.9	4	140	0

ภาพที่ 3.3 ข้อมูลของโรคเบาหวาน

ตาราง 3.1 Data Dictionary อธิบายรายละเอียดในแต่ละตัวแปร

Attribute	Description	Type	Values
year	ปีที่บันทึกข้อมูล	Integer	
gender	เพศของผู้ป่วย	Nominal	Male คือ เพศชาย หรือ Female คือ เพศหญิง หรือOther คือ เพศอื่น ๆ
age	อายุของผู้ป่วย	Real	หน่วยเป็นปี
location	สถานที่หรือภูมิภาคที่ผู้ป่วยอยู่	Nominal	เช่น Alabama California
race:AfricanAmerican	เชื้อชาติแอฟริกันอเมริกัน	Integer	0 = ไม่ใช่ หรือ 1 = ใช่
race:Asian	เชื้อชาติเอเชีย	Integer	0 = ไม่ใช่ หรือ 1 = ใช่
race:Caucasian	เชื้อชาติคอเคเซียน (ผิวขาว)	Integer	0 = ไม่ใช่ หรือ 1 = ใช่
race:Hispanic	เชื้อชาติฮิสแปนิก	Integer	0 = ไม่ใช่ หรือ 1 = ใช่

Attribute	Description	Type	Values
race:Other	เชื้อชาติอื่น ๆ	Integer	0 = ไม่ใช่ หรือ 1 = ใช่
hypertension	ความดันโลหิตสูง	Integer	0 = ไม่มี, 1 = มี
heart_disease	โรคหัวใจ	Integer	0 = ไม่มี, 1 = มี
smoking_history	ประวัติการสูบบุหรี่	Nominal	never, former, current, ever, not current, unknown
bmi	ดัชนีมวลกาย (Body Mass Index) ซึ่งใช้บ่งบอกภาวะอ้วนหรือผอม	Real	ค่า bmi ต่ำกว่า 23 เป็นเกณฑ์ที่น้ำหนักน้อยกว่าปกติ ระหว่าง 23–27.5 เป็นเกณฑ์ที่ปกติ และสูงกว่า 27.5 ถือว่าอยู่ในเกณฑ์ที่น้ำหนักเกิน
hbA1c_level	ระดับฮีโมโกลบิน A1c (%), ใช้วัดระดับน้ำตาลเฉลี่ยในเลือดช่วง 2–3 เดือนที่ผ่านมา	Real	ค่า HbA1c ปกติจะต่ำกว่า 5.7%, ค่าระหว่าง 5.7–6.4% ถือว่าเป็นภาวะเสี่ยงต่อโรคเบาหวาน และค่า 6.5% ขึ้นไป ถือว่าเป็นโรคเบาหวาน
blood_glucose_level	ระดับน้ำตาลในเลือดแบบเฉียบพลัน (mg/dL)	Integer	ต่ำกว่า 70 mg/dL อาจบ่งบอกถึงภาวะน้ำตาลในเลือดต่ำ ปกติอยู่ในช่วง 70–100 mg/dL สูงกว่า 125 mg/dL อาจบ่งบอกถึงภาวะก่อนเบาหวานหรือโรคเบาหวาน
diabetes	การวินิจฉัยโรคเบาหวาน	Integer	0 = ไม่เป็น, 1 = เป็น

3.1.3 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการแปลงข้อมูลที่ได้รวบรวมมา และเลือกไว้ให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปวิเคราะห์ในขั้นตอนต่อไปได้ โดยการทำให้เป็นข้อมูลที่ถูกต้อง (Data cleaning) มักใช้เวลาค่อนข้างมาก

3.1.3.1 ทำการคัดเลือกข้อมูล (Data Selection) คือการคัดเลือกข้อมูลที่เหมาะสมเพื่อนำมาใช้ในการวิเคราะห์ข้อมูล

ผู้วิเคราะห์ข้อมูลทำการคัดเลือกข้อมูล และทำการ Data Cleaning ข้อมูลโรคเบาหวาน โดยตัดส่วนที่ไม่จำเป็นออกให้เหลือเฉพาะข้อมูลที่จำเป็นในการวิเคราะห์ในภาพรวม จำนวน 9 แอตทริบิวท์ ได้แก่ เพศ (Gender) อายุ (Age) ความดันโลหิตสูง (Hypertension) ประวัติโรคหัวใจ (Heart_disease) ประวัติการสูบบุหรี่ (Smoking_history) ค่าดัชนีมวลกาย (BMI) ระดับน้ำตาลสะสมเฉลี่ยในเลือด (hbA1c_level) ระดับน้ำตาลในเลือด (blood_glucose_level) และ การเป็นโรคเบาหวาน (diabetes)

A	B	C	D	E	F	G	H	I
gender	age	hypertension	heart_disease	smoking_history	bmi	hbA1c_level	blood_glucose_level	diabetes
Female	32	0	0	never	27.32	5	100	0
Female	29	0	0	never	19.95	5	90	0
Male	18	0	0	never	23.76	4.8	160	0
Male	41	0	0	never	27.32	4	159	0
Female	52	0	0	never	23.75	6.5	90	0
Male	66	0	0	not current	27.32	5.7	159	0
Female	49	0	0	current	24.34	5.7	80	0
Female	15	0	0	No Info	20.98	5	155	0
Male	51	0	0	never	38.14	6	100	0
Male	42	0	0	No Info	27.32	5.7	160	0
Male	15	0	0	No Info	19.15	6.6	200	0
Female	53	0	0	never	34.3	6.6	155	0
Female	3	0	0	No Info	20.28	3.5	159	0
Female	40	0	0	never	27.63	6.5	126	0
Female	64	0	0	ever	49.27	8.2	140	1
Female	23	0	0	never	27.32	6	140	0
Female	2	0	0	never	27.32	3.5	155	0
Male	80	0	0	No Info	27.32	6	130	0
Female	12	0	0	never	20.78	6.5	130	0
Female	50	0	0	never	26.39	6.6	158	0
Female	69	0	0	No Info	31.76	6.5	160	0

ภาพที่ 3.4 ข้อมูลของโรคเบาหวาน

ในการสร้างโมเดลวิเคราะห์โรคเบาหวาน การเลือกแอตทริบิวต์ที่เหมาะสมมีความสำคัญอย่างยิ่ง เนื่องจากแต่ละโมเดลมีวิธีการเรียนรู้และประมวลผลข้อมูลแตกต่างกัน เช่น Decision Tree และ Random Forest ใช้การแบ่งกลุ่มตามเงื่อนไข, Naïve Bayes ใช้ความน่าจะเป็นของตัวแปรแต่ละตัว, SVM ต้องการตัวแปรที่สามารถใช้แยกกลุ่มข้อมูลได้อย่างชัดเจน ส่วน ANN จะมองหาความสัมพันธ์ซับซ้อนเชิงลึกของข้อมูล ในการนี้จึงได้มีการพิจารณาคัดเลือกแอตทริบิวต์ที่มีความเกี่ยวข้องกับโรคเบาหวาน ดังนี้

1. อายุของผู้ป่วย (age)

อายุเป็นปัจจัยสำคัญที่ส่งผลต่อความเสี่ยงในการเกิดโรคเบาหวาน โดยเฉพาะเบาหวานชนิดที่ 2 ซึ่งมีแนวโน้มพบมากในกลุ่มผู้สูงอายุ เนื่องจากการเปลี่ยนแปลงของระบบเผาผลาญและภาวะดื้อต่ออินซูลินที่เพิ่มขึ้นตามอายุ การใช้ตัวแปรอายุในโมเดล Machine Learning ช่วยให้สามารถทำนายผลลัพธ์ได้แม่นยำมากขึ้น โดยเฉพาะเมื่อใช้กับโมเดลที่รองรับข้อมูลเชิงตัวเลข เช่น Decision Tree และ Neural Network (วนิดา พงษ์สงวน, 2561; เมธาพร ผ่องยิ่ง, 2565; อรุณรักษ์ ตันพานิช และคณะ, 2562; สายชล สินสมบูรณ์ทอง, 2561; Luís Chaves และ Gonçalo Marques, 2021; Maydanchi et al., 2024)

2. เพศ (Gender)

เพศของผู้ป่วยเป็นอีกปัจจัยหนึ่งที่อาจมีผลต่อโอกาสในการเกิดโรคเบาหวาน เนื่องจากความแตกต่างทางชีวภาพและฮอร์โมนระหว่างชายและหญิง เช่น ผู้หญิงอาจเผชิญกับภาวะเบาหวานขณะตั้งครรภ์ ขณะที่ผู้ชายบางกลุ่มอาจมีพฤติกรรมเสี่ยงสูงกว่า เช่น การสูบบุหรี่หรือบริโภคอาหารไขมันสูง แม้ว่าจะไม่ใช่ปัจจัยชี้ขาดโดยตรง แต่เพศสามารถช่วยเสริมการวิเคราะห์และจำแนกกลุ่มได้ โดยเฉพาะในโมเดลที่ต้องการข้อมูลจำแนกประเภท เช่น Naïve Bayes และโมเดลที่เรียนรู้ความสัมพันธ์เชิงซ้อน เช่น ANN (วนิดา พงษ์สงวน, 2561; Sidong Wei et al., 2018; เมธาพร ผ่องยิ่ง 2565; อรุณรักษ์ ตันพานิช และคณะ, 2562; สายชล สินสมบูรณ์ทอง, 2561; Luís Chaves และ Gonçalo Marques, 2021; Maydanchi et al., 2024; Chun-Yang Chou, Ding-Yang Hsu and Chun-Hung Chou, 2023; Wenhui Zhou, Xiaomin Liu and Hongtao Bai, Lili He, 2024)

3. ความดันโลหิตสูง (Hypertension)

ความดันโลหิตสูงเป็นภาวะที่พบร่วมกับโรคเบาหวานบ่อยครั้ง และมักเป็นผลลัพธ์ของการเปลี่ยนแปลงในระบบเผาผลาญของร่างกาย ความดันสูงยังบ่งชี้ถึงภาวะความเสื่อมของหลอดเลือด ซึ่งสัมพันธ์กับการทำงานของอินซูลินที่ผิดปกติ ดังนั้นการมีภาวะความดันสูงจึงเป็นสัญญาณเตือนล่วงหน้าที่สำคัญในการพยากรณ์โรคเบาหวาน และเหมาะแก่การนำมาใช้กับโมเดลทุกประเภท ทั้งในรูปแบบตัวแปรเชิงตรรกะใน Decision Tree และการคำนวณโอกาสร่วมใน Naïve Bayes (วนิดา พงษ์สงวน, 2561; Sidong Wei et al., 2018; เมธาพร ผ่องยิ่ง, 2565; อรุณรักษ์ ตันพานิช และคณะ, 2562; สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณ์ทอง, 2560; Alain Hennebelle, Huned Materwala and Leila Ismail, 2023; Maydanchi et al.

2024; Chun-Yang Chou, Ding-Yang Hsu and Chun-Hung Chou, 2023; Wenhui Zhou, Xiaomin Liu, Hongtao Bai and Lili He, 2024)

4. โรคหัวใจ (Heart_disease)

โรคหัวใจและเบาหวานมักมีความเกี่ยวข้องกันอย่างใกล้ชิด ทั้งในฐานะโรคเรื้อรังที่มีปัจจัยเสี่ยงร่วมกัน และในฐานะผลสืบเนื่องซึ่งกันและกัน การมีประวัติโรคหัวใจอาจสะท้อนถึงระดับความเสี่ยงที่สูงกว่าปกติต่อภาวะเบาหวาน การบันทึกข้อมูลตัวแปรนี้ในโมเดลช่วยให้ระบบสามารถจับความสัมพันธ์ระหว่างความเสื่อมของระบบไหลเวียนโลหิตกับการควบคุมน้ำตาลในเลือดได้อย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งในโมเดล ANN ที่สามารถเรียนรู้ความเชื่อมโยงซับซ้อนได้ดี (Maydanchi et al., 2024)

5. ประวัติการสูบบุหรี่ (Smoking_history)

การสูบบุหรี่เป็นพฤติกรรมเสี่ยงที่ได้รับการยืนยันว่ามีผลต่อการทำงานของอินซูลิน และเพิ่มโอกาสในการเกิดโรคเบาหวานอย่างมีนัยสำคัญ การบันทึกข้อมูลว่าผู้ป่วยมีประวัติการสูบบุหรี่ในรูปแบบใด เช่น สูบในอดีต ปัจจุบัน หรือไม่เคยสูบ ช่วยเพิ่มมิติการวิเคราะห์ให้ครอบคลุมทั้งด้านชีวภาพและพฤติกรรม โดยเฉพาะในโมเดล Naïve Bayes ที่สามารถใช้ข้อมูลเชิงหมวดหมู่ได้ดี รวมทั้งใน ANN ที่สามารถเรียนรู้ผลกระทบระยะยาวจากพฤติกรรมได้ (วนิดา พงษ์สงวน, 2561; Maydanchi et al., 2024; นพรัตน์ นนท์ศิริ, พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา, 2565)

6. ดัชนีมวลกาย (BMI)

ดัชนีมวลกายเป็นตัวชี้วัดภาวะอ้วนหรือผอม ซึ่งสัมพันธ์อย่างยิ่งกับความเสี่ยงในการเกิดโรคเบาหวาน โดยเฉพาะในกรณีที่ BMI เกิน 27.5 ซึ่งถือว่าอยู่ในเกณฑ์อ้วนและมีความเสี่ยงสูง การนำตัวแปรนี้มาใช้ในโมเดลสามารถช่วยแยกแยะกลุ่มผู้ป่วยที่มีแนวโน้มเป็นโรคได้อย่างแม่นยำ โดยเฉพาะใน SVM ที่ต้องอาศัยการวางเส้นแบ่งข้อมูล และใน ANN ที่เรียนรู้ความสัมพันธ์ระหว่างค่าต่อเนื่องกับผลลัพธ์ได้ดี (วนิดา พงษ์สงวน, 2561; Talha Mahboob Alam et al., 2019; Sidong Wei et al., 2018; เมธาพร ผ่องยิ่ง, 2565; อรุณรักษ์ ตันพานิช และคณะ, 2562; สายชล สินสมบูรณ์ทอง, 2561; Alain Hennebelle, Huned Materwala, Leila Ismail 2023; Maydanchi et al., 2024; Chun-Yang Chou, Ding-Yang Hsu and Chun-Hung Chou, 2023; นพรัตน์ นนท์ศิริ, พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา, 2565; Wenhui Zhou, Xiaomin Liu, Hongtao Bai, Lili He, 2024)

7. ระดับน้ำตาลเฉลี่ยในเลือดช่วง 2–3 เดือน (hbA1c_level)

ระดับ HbA1c ถือเป็นหนึ่งในเกณฑ์มาตรฐานทางคลินิกที่ใช้วินิจฉัยโรคเบาหวาน เพราะสามารถสะท้อนระดับน้ำตาลในเลือดได้ในระยะเวลา 2–3 เดือนที่ผ่านมา ค่านี้จึงมีความสัมพันธ์โดยตรงกับการเกิดโรคเบาหวานและมีบทบาทสำคัญในการทำนายผลลัพธ์ โมเดลอย่าง Decision Tree หรือ Random Forest มักเลือกแอตทริบิวต์นี้เป็นตัวแบ่งกลุ่มหลัก และใน ANN หรือ SVM ก็ใช้เพื่อจับรูปแบบของความผิดปกติได้อย่างละเอียด (Talha Mahboob Alam et al., 2019; Sidong Wei et al., 2018; สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณ์ทอง, 2560; Alain Hennebelle, Huned Materwala, Leila Ismail, 2023; Maydanchi et al., 2024; Chun-Yang Chou, Ding-Yang Hsu and Chun-Hung Chou, 2023; นพรัตน์ นนท์ศิริ, พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา, 2565; Wenhui Zhou, Xiaomin Liu, Hongtao Bai and Lili He, 2024)

8. ระดับน้ำตาลในเลือดเฉียบพลัน (blood_glucose_level)

ค่าน้ำตาลในเลือดที่วัดแบบเฉียบพลันมักใช้ประกอบการวินิจฉัยหรือเฝ้าระวังภาวะเบาหวานและภาวะก่อนเบาหวาน การที่ค่าดังกล่าวสูงผิดปกติเป็นสัญญาณที่บ่งชี้ถึงการทำงานของระบบควบคุมน้ำตาลในร่างกายที่บกพร่อง และสามารถเป็นตัวแปรสนับสนุนสำคัญในโมเดลการวิเคราะห์ได้ โดยเฉพาะอย่างยิ่งใน ANN ซึ่งสามารถเรียนรู้ความสัมพันธ์ระหว่างค่าทางคลินิกหลายค่าได้ในเชิงลึก (Talha Mahboob Alam et al., 2019; Sidong Wei et al., 2018; สุรวัชร ศรีเปารยะ และ สายชล สินสมบูรณ์ทอง, 2560; Alain Hennebelle, Huned Materwala, Leila Ismail, 2023; Maydanchi et al., 2024; Chun-Yang Chou, Ding-Yang Hsu and Chun-Hung Chou, 2023; นพรัตน์ นนท์ศิริ, พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา, 2565; Wenhui Zhou, Xiaomin Liu, Hongtao Bai and Lili He, 2024)

9. เชื้อชาติ (race)

จากข้อมูลทางระบาดวิทยาพบว่า เชื้อชาติบางกลุ่ม เช่น ชาวแอฟริกันอเมริกันและฮิสแปนิก มีอัตราการเกิดโรคเบาหวานสูงกว่ากลุ่มอื่น ซึ่งอาจเกิดจากปัจจัยทางพันธุกรรม วิถีชีวิต และการเข้าถึงบริการทางการแพทย์ การใส่ข้อมูลเชื้อชาติจึงมีความสำคัญในการช่วยให้โมเดลเรียนรู้โครงสร้างความเสี่ยงเชิงประชากร โดยเฉพาะใน Naïve Bayes และ Decision Tree ที่ใช้ข้อมูลเชิงกลุ่มเพื่อจำแนกความเสี่ยง (สายชล สินสมบูรณ์ทอง, 2561; นพรัตน์ นนท์ศิริ, พิศณุ ชัยจิตวณิชกุล และ กริช สมกันธา, 2565)

3.1.3.2 การจัดการข้อมูลหมวดหมู่เพื่อเพิ่มประสิทธิภาพในการทำนาย

1. การเข้ารหัสข้อมูล (Label endcoding)

เนื่องจากข้อมูลที่นำมาใช้ มีทั้งรูปแบบของตัวอักษร (Nominal) ตัวเลขจำนวนเต็ม (Integer) และตัวเลขทศนิยม (Real) จึงทำการแปลงข้อมูลจากตัวอักษรให้อยู่ในรูปแบบของตัวเลขและทศนิยม เพื่อให้ตัวแบบทำนายสามารถทำงานได้ กระบวนการแปลงค่าข้อมูลประเภทหมวดหมู่ (Categorical Data) ที่อยู่ในรูปของข้อความ เช่น ชาย หญิง ให้เป็น ค่าตัวเลขจำนวนเต็ม (Integer) เพื่อให้สามารถนำข้อมูลไปใช้กับอัลกอริธึมของการเรียนรู้ของเครื่อง (Machine Learning) ได้แก่ แอตทริบิวต์ gender ดังตารางที่ 3.2 และแอตทริบิวต์ smoking_history ดังตารางที่ 3.3

ตารางที่ 3.2 การแปลงข้อมูล Label endcoding ของ gender

gender	Label endcoding	ความหมาย
Male	000	ชาย
Female	001	หญิง

ตารางที่ 3.3 การแปลงข้อมูล Label endcoding ของ smoking_history

smoking_history	Label endcoding	ความหมาย
No Info	100	ไม่มีข้อมูลว่าผู้ป่วยเคยสูบบุหรี่หรือไม่
never	101	ไม่เคยสูบบุหรี่
former	102	เคยสูบบุหรี่ในอดีต แต่เลิกแล้ว
not current	102	ไม่ได้สูบบุหรี่ในปัจจุบัน
ever	102	เคยสูบบุหรี่
current	103	ยังสูบบุหรี่อยู่ในปัจจุบัน

2. การตัดทิ้งข้อมูลที่ไม่เกี่ยวข้องกับการทำนาย

ในกระบวนการเตรียมข้อมูลสำหรับการวิเคราะห์ข้อมูลเพื่อทำนายลักษณะผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด **ตัวแปร location** ซึ่งบ่งชี้ถึงชื่อรัฐหรือดินแดนในสหรัฐอเมริกาได้รับการพิจารณาว่าไม่มีความสัมพันธ์โดยตรงกับการทำนายผลการเกิดโรคเบาหวาน จึงได้มีการตัดออกจากการวิเคราะห์ในขั้นตอนการเตรียมข้อมูลเพื่อลด

ความซับซ้อนของโมเดลและเพิ่มประสิทธิภาพในการเรียนรู้ของเครื่อง (Rajkomar et al., 2018) และ**ตัวแปร year** เป็นข้อมูลที่สะท้อนช่วงเวลาการเก็บข้อมูล แต่ไม่สะท้อนลักษณะหรือพฤติกรรมของผู้ป่วยโดยตรง ตัวแปรนี้จึงไม่มีความเชื่อมโยงโดยตรงกับการเกิดโรคเบาหวานในแต่ละบุคคล นอกจากนี้ year ยังอาจเป็นตัวแปรที่มีจำนวนค่าที่หลากหลาย (high cardinality) ซึ่งไม่เหมาะสมกับการป้อนเข้าสู่โมเดลประเภททั่วไป โดยเฉพาะเมื่อไม่มีการเข้ารหัสหรือจัดการอย่างเหมาะสม (Chakraborty et al., 2020) ส่วน**ตัวแปร race** แม้ว่าข้อมูลเชื้อชาติมักถูกเก็บเพื่อการวิจัยทางระบาดวิทยา แต่ในการใช้งานโมเดลเรียนรู้ของเครื่อง การใช้เชื้อชาติเป็นคุณลักษณะอาจนำไปสู่ความลำเอียงหรืออคติที่เกิดขึ้นจากการทำงานของระบบอัลกอริธึมหรือโมเดลปัญญาประดิษฐ์ (algorithmic bias) ได้ (Obermeyer et al., 2019)

3.1.3.3 ทำการกลั่นกรองข้อมูล (Data Cleaning) คือการทำความสะอาดข้อมูลเป็น กระบวนการตรวจสอบและการแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล ตารางหรือฐานข้อมูล ซึ่งเป็นหลักสำคัญของฐานข้อมูล ทางผู้วิเคราะห์ข้อมูลได้ดำเนินการดังนี้

1) ข้อมูลผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด จากการวิเคราะห์พบว่า ข้อมูลเชื้อชาติที่ได้มาเป็นตัวหนังสือซึ่งทำให้ไม่รองรับต่อการนำไปวิเคราะห์ ดังนั้นผู้วิเคราะห์ข้อมูลจึงได้ทำการแทนที่จากข้อมูลที่เป็นตัวหนังสือให้เป็นตัวเลขใน excel

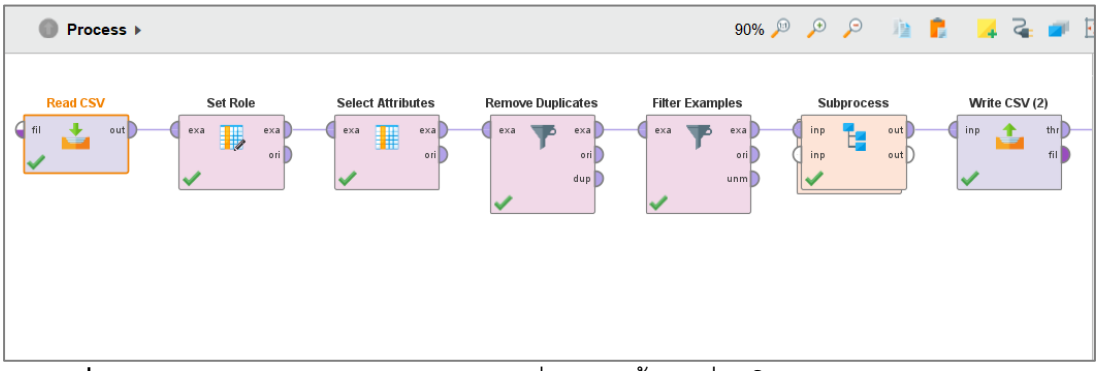
– ข้อมูลประวัติการสูบบุหรี่ จากตัวหนังสือแทนเป็นตัวเลขทั้งหมด คือ No Info = 100, never = 101, former = 102, not current = 102, ever = 102 และ current = 103

smoking_h	bmi	hbA1c_lev	blood_glu	diabetes	
never	27.32	5	100	0	=IF(L2="current","103",IF(L2="ever","102",IF(L2="not current","102",IF(L2
never	19.95	5	90	0	="former","102",IF(L2="never","101","100")))))
never	23.76	4.8	160	0	
never	27.32	4	159	0	101 สูตรที่ใช้แปลงข้อมูล
never	23.75	6.5	90	0	101 ตัวหนังสือเป็นข้อมูลตัวเลข
not curren	27.32	5.7	159	0	102
current	24.34	5.7	80	0	103
No Info	20.98	5	155	0	100
never	38.14	6	100	0	101
No Info	27.32	5.7	160	0	100
No Info	19.15	6.6	200	0	100
never	34.3	6.6	155	0	101

ภาพที่ 3.5 ข้อมูลประวัติการสูบบุหรี่ จากตัวหนังสือแทนเป็นตัวเลขทั้งหมด

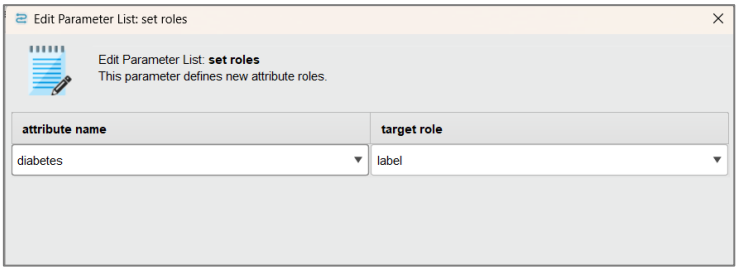
2. การแสดงกระบวนการกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio โดยใช้รูปแบบการทำงานแบบกระบวนการที่สามารถปรับแต่งและจัดการได้อย่างเป็น

ระบบ ช่วยให้ผู้ใช้งานสามารถดำเนินการตามลำดับขั้นตอน ได้แก่ การอ่านข้อมูล การเลือกตัวแปร การกำหนดบทบาทของข้อมูล การกรองค่าที่ไม่จำเป็น รวมถึงการลบข้อมูลซ้ำซ้อน เพื่อให้ได้ชุดข้อมูลที่มีคุณภาพสูงและเหมาะสมต่อการวิเคราะห์เชิงลึก ดังแสดงในภาพที่ 3.6



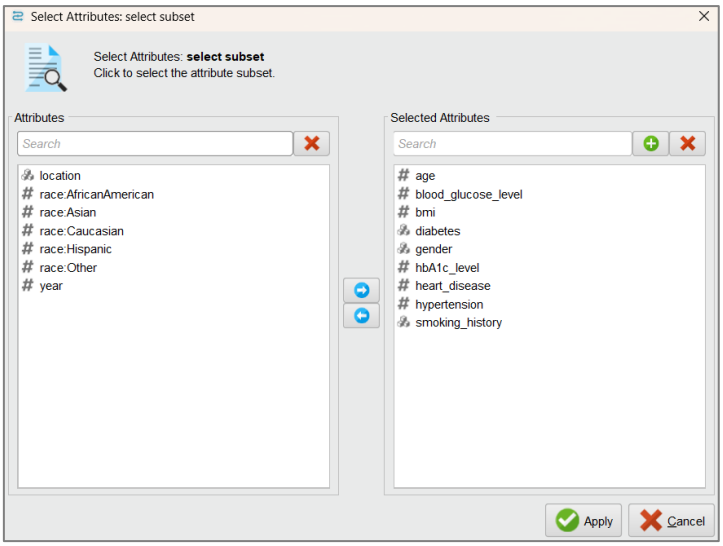
ภาพที่ 3.6 แสดงกระบวนการประมวลผลกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio

จากภาพที่ 3.6 แสดงถึงกระบวนการเตรียมและประมวลผลข้อมูลเบื้องต้นในโปรแกรม RapidMiner Studio โดยเริ่มจากการนำเข้าข้อมูลด้วยโหนด Read CSV และกำหนดบทบาทของแอตทริบิวต์ผ่านโหนด Set Role จากนั้นใช้โหนด Select Attributes เพื่อเลือกเฉพาะคุณลักษณะที่จำเป็น และโหนด Remove Duplicates เพื่อลบข้อมูลซ้ำ หลังจากนั้นใช้โหนด Filter Examples สำหรับคัดกรองแถวข้อมูลเฉพาะที่ต้องการใช้งาน ต่อมาคือโหนด Subprocess เป็นกระบวนการย่อยที่จัดการข้อมูลเพิ่มเติม เช่น ปรับสมดุลคลาสหรือเติมข้อมูลที่ขาดหาย และจบด้วยการบันทึกข้อมูลใหม่ผ่าน Write CSV เพื่อเตรียมไว้สำหรับขั้นตอนการวิเคราะห์หรือสร้างโมเดลในลำดับถัดไป



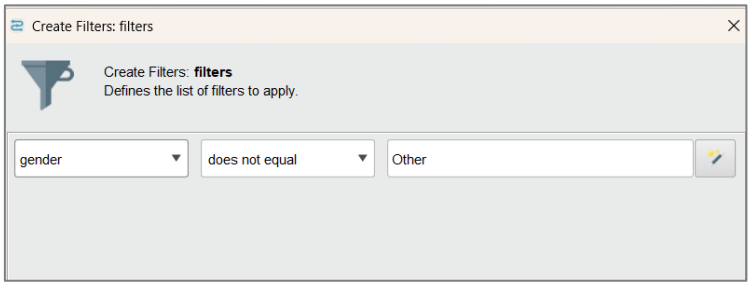
ภาพที่ 3.7 แสดงการกำหนดบทบาทของแอตทริบิวต์

จากภาพที่ 3.7 กำหนดบทบาทของแอตทริบิวต์ diabetes ให้เป็นตัวแปรเป้าหมาย (label) โดยใช้โหนด Set Role ซึ่งเป็นขั้นตอนสำคัญในการเตรียมข้อมูลสำหรับการเรียนรู้ของโมเดลในกระบวนการวิเคราะห์ข้อมูล



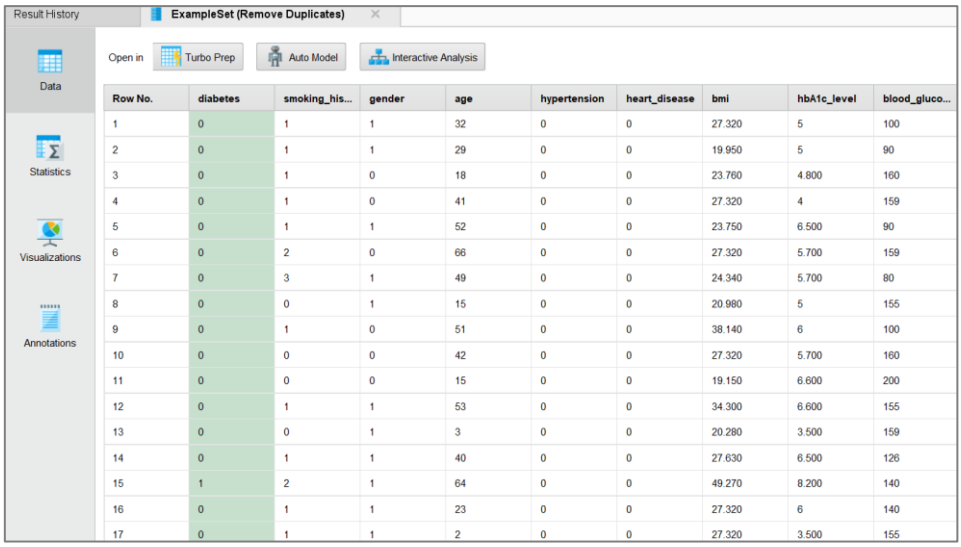
ภาพที่ 3.8 แสดงขั้นตอนการคัดเลือกแอตทริบิวต์ที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล

จากภาพที่ 3.8 คัดเลือกแอตทริบิวต์ที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลผ่านโหนด Select Attributes โดยเลือกเฉพาะคุณลักษณะที่ต้องการใช้ เช่น อายุ เพศ โรคประจำตัว และค่าทางสุขภาพ เพื่อเพิ่มประสิทธิภาพในการประมวลผลและลดความซับซ้อนของข้อมูล



ภาพที่ 3.9 แสดงการใช้โหนด Filter Examples เพื่อกรองข้อมูล

จากภาพที่ 3.9 ใช้โหนด Filter Examples เพื่อกรองข้อมูล โดยตั้งเงื่อนไขให้แสดงเฉพาะแถวที่มีค่าเพศ (gender) ไม่เท่ากับ Other ซึ่งช่วยลดความคลาดเคลื่อนของข้อมูล และเพิ่มความแม่นยำในการวิเคราะห์



Row No.	diabetes	smoking_his...	gender	age	hypertension	heart_disease	bmi	hbA1c_level	blood_gluco...
1	0	1	1	32	0	0	27.320	5	100
2	0	1	1	29	0	0	19.950	5	90
3	0	1	0	18	0	0	23.760	4.800	160
4	0	1	0	41	0	0	27.320	4	159
5	0	1	1	52	0	0	23.750	6.500	90
6	0	2	0	66	0	0	27.320	5.700	159
7	0	3	1	49	0	0	24.340	5.700	80
8	0	0	1	15	0	0	20.980	5	155
9	0	1	0	51	0	0	38.140	6	100
10	0	0	0	42	0	0	27.320	5.700	160
11	0	0	0	15	0	0	19.150	6.600	200
12	0	1	1	53	0	0	34.300	6.600	155
13	0	0	1	3	0	0	20.280	3.500	159
14	0	1	1	40	0	0	27.630	6.500	126
15	1	2	1	64	0	0	49.270	8.200	140
16	0	1	1	23	0	0	27.320	6	140
17	0	1	1	2	0	0	27.320	3.500	155

ภาพที่ 3.10 แสดงผลลัพธ์ข้อมูลชุดโรคเบาหวาน

จากภาพที่ 3.10 แสดงผลลัพธ์ข้อมูลหลังผ่านขั้นตอนการลบข้อมูลซ้ำ (Remove Duplicates) ซึ่งทำให้ได้ชุดข้อมูลที่สะอาดและปราศจากข้อมูลซ้ำซ้อน พร้อมสำหรับการนำไปใช้ในการสร้างโมเดลต่อไป

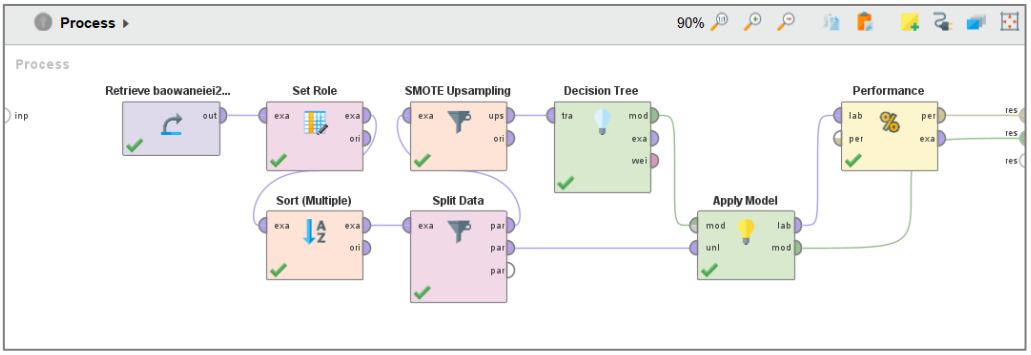
3.1.4 การสร้างแบบจำลอง (Modeling)

ขั้นตอนการสร้างตัวแบบทางคณิตศาสตร์ และ สถิติเพื่อการวิเคราะห์ข้อมูล โดยสามารถใช้เทคนิควิธีการต่างๆ อาทิ การจำแนก (Classification) และการสร้างความสัมพันธ์ (Association rule) โดยใช้โปรแกรม RapidMiner Studio

โดยเทคนิคทั้ง 5 ที่ผู้วิจัยได้เลือกออกมาทำเป็นโมเดลนั้นคือเครื่องมือสำคัญในกระบวนการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งเป็นการสร้างแบบจำลองจากข้อมูลที่มีป้ายกำกับ (Label) อยู่แล้ว เพื่อใช้ในการพยากรณ์หรือจำแนกประเภทของข้อมูลใหม่ในอนาคต โดยในกระบวนการนี้ระบบจะเรียนรู้ความสัมพันธ์ระหว่างข้อมูลอินพุต (Input Features) และคำตอบหรือผลลัพธ์ที่ต้องการ (Target Output) เพื่อให้สามารถคาดเดาคำตอบของข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างถูกต้อง โดยเทคนิคแรกคือ (1) ต้นไม้ตัดสินใจ (Decision Tree) ซึ่งทำงานโดยการแบ่งข้อมูลตามคุณลักษณะต่าง ๆ ด้วยเกณฑ์เงื่อนไขแบบ "ถ้า–แล้ว" (if–then rules) ไปทีละชั้นจาก root node ไปยัง leaf node จนได้คำตอบ เช่น ถ้าอายุ < 45 และมีค่า BMI > 30 อาจมีแนวโน้มเป็นเบาหวาน โดย Decision Tree ต้องอาศัยชุดข้อมูล

ที่รู้ผลลัพธ์อยู่แล้วเพื่อให้สามารถเรียนรู้ว่าเงื่อนไขใดนำไปสู่ผลลัพธ์ใด อีกทั้งยังสามารถใช้ได้ทั้งในงานจำแนก (Classification) และพยากรณ์เชิงปริมาณ (Regression) ต่อมา (2) เทคนิคนาอีฟเบย์ (Naïve Bayes) เป็นอัลกอริธึมเชิงความน่าจะเป็นที่ใช้กฎของเบย์ (Bayes' Theorem) โดยสมมติว่าตัวแปรอิสระต่อกันทั้งหมด แม้ว่าความจริงอาจไม่เป็นเช่นนั้น โมเดลนี้ต้องการข้อมูลที่มี label ชัดเจน เพื่อคำนวณความน่าจะเป็นแบบมีเงื่อนไข เช่น ความน่าจะเป็นที่จะเป็นเบาหวานเมื่อรู้ว่าเพศชาย อายุ 50 ปี และมีประวัติสูบบุหรี่ Naïve Bayes เหมาะกับงานที่มีข้อมูลจำนวนมาก และต้องการโมเดลที่ประมวลผลเร็ว เช่น การคัดกรองสแปม หรือการวิเคราะห์ความรู้สึกในข้อความ ส่วน (3) เทคนิคป่าสุ่ม (Random Forest) เป็นการรวมหลาย Decision Trees เข้าด้วยกันโดยใช้วิธีสุ่มเลือกข้อมูลและคุณลักษณะ เพื่อสร้างต้นไม้หลายต้นแล้วรวมผลลัพธ์ด้วยวิธีการโหวต (สำหรับงานจำแนก) หรือเฉลี่ย (สำหรับงานพยากรณ์) โดยทุกต้นไม้จะฝึกด้วยข้อมูลที่มี label เหมือนกัน จึงถือว่าอยู่ภายใต้ Supervised Learning เช่นเดียวกับ Decision Tree แต่มีความแม่นยำสูงกว่าและลดปัญหา overfitting ได้ดีมาก ส่วนเทคนิคที่สี่คือ (4) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine หรือ SVM) ซึ่งมุ่งหาเส้นแบ่งข้อมูล (Hyperplane) ที่ดีที่สุดในการแยกกลุ่มข้อมูลให้ออกจากกัน โดยพยายามเพิ่ม "ระยะห่าง" หรือ margin ระหว่างกลุ่มให้กว้างที่สุด SVM ต้องฝึกจากข้อมูลที่มี label เช่นเดียวกัน โดยอาศัยข้อมูลตัวอย่างที่มีอยู่เพื่อเรียนรู้ว่าแต่ละคลาสควรแบ่งอย่างไรให้ชัดเจนที่สุด และสามารถประยุกต์กับข้อมูลที่ไม่สามารถแยกด้วยเส้นตรงได้ผ่านเทคนิค Kernel Trick สุดท้ายคือ (5) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่ได้แรงบันดาลใจจากสมองมนุษย์ โดยมีโครงสร้างเป็นชั้น ๆ ของโหนดที่เชื่อมโยงกัน ข้อมูลจะไหลผ่านและถูกคำนวณโดยแต่ละโหนด จากนั้นระบบจะปรับน้ำหนักภายในเพื่อให้ผลลัพธ์แม่นยำขึ้นเรื่อย ๆ ผ่านกระบวนการที่เรียกว่า Backpropagation เทคนิคนี้เหมาะกับข้อมูลที่มีความซับซ้อนสูง เช่น รูปภาพ เสียง หรือข้อมูลทางการแพทย์ ทั้งหมดนี้ล้วนเป็นเทคนิคที่ต้องอาศัยข้อมูลที่มีคำตอบเพื่อเรียนรู้ จึงจัดอยู่ในกลุ่มของ Supervised Learning ทั้งสิ้น

3.1.4.1 เทคนิคต้นไม้ตัดสินใจ (Decision tree)



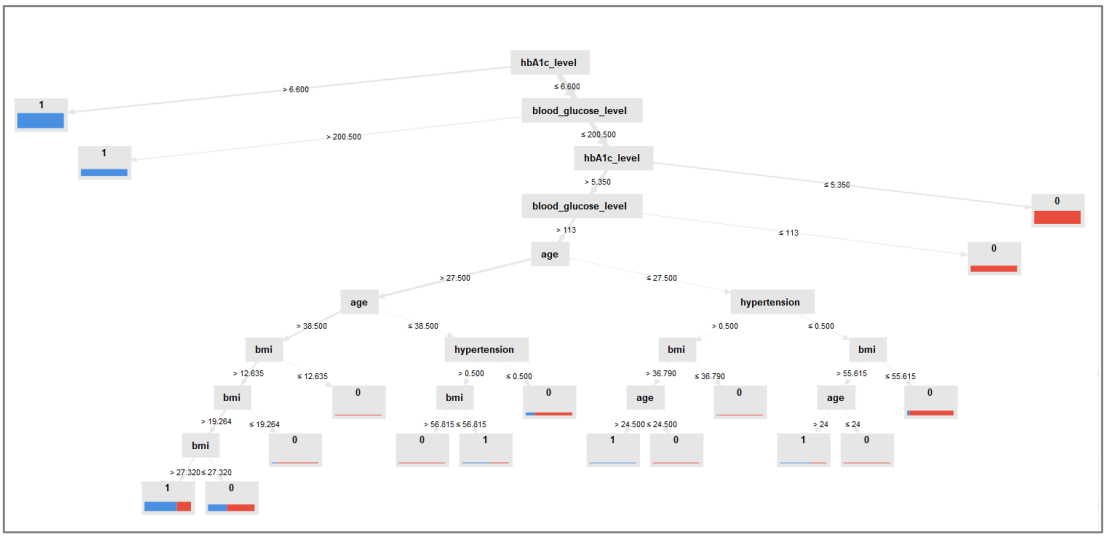
ภาพที่ 3.11 แสดงกระบวนการประมวลผลกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio

จากภาพที่ 3.11 เริ่มต้นจากการเรียกใช้ชุดข้อมูลเบาหวานผ่านโหนด Retrieve และใช้โหนด Set Role กำหนดบทบาทของแอตทริบิวต์ diabetes ให้เป็นตัวแปรเป้าหมาย (label) จากนั้นใช้โหนด Sort Multiple เรียงลำดับข้อมูลตามแอตทริบิวต์ diabetes ในลำดับแบบ descending หรือเรียงจากค่ามากไปน้อย จากนั้นข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data โดยกำหนดอัตราส่วนเป็น 0.7 : 0.3 ซึ่งหมายความว่า 70% ของข้อมูลจะถูกนำไปใช้สำหรับการฝึกโมเดล อีก 30% ใช้สำหรับการประเมินผลการทำงานของโมเดล เพื่อป้องกันการเกิด overfitting และช่วยให้สามารถประเมินประสิทธิภาพของโมเดลกับข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างเหมาะสม พร้อมทั้งเลือกวิธีการสุ่มตัวอย่างแบบ linear sampling เพื่อจัดเรียงข้อมูลตามลำดับที่กำหนด ซึ่งชุดฝึกผ่านการปรับสมดุลด้วย SMOTE Upsampling แล้วนำไปสร้างโมเดลด้วย Decision Tree ต่อไป

accuracy: 90.69%

	true 0	true 1	class precision
pred. 0	26116	0	100.00%
pred. 1	2680	0	0.00%
class recall	90.69%	0.00%	

ภาพที่ 3.12 แสดงค่าผลลัพธ์โมเดล Decision Tree



ภาพที่ 3.13 แสดงโมเดล Graph Decision Tree ในโปรแกรม Rapid Miner Studio

Tree

```
hbA1c_level > 6.600: 1 {1=28385, 0=0}
hbA1c_level ≤ 6.600
|   blood_glucose_level > 200.500: 1 {1=12296, 0=0}
|   blood_glucose_level ≤ 200.500
|   |   hbA1c_level > 5.350
|   |   |   blood_glucose_level > 113
|   |   |   |   age > 27.500
|   |   |   |   |   age > 38.500
|   |   |   |   |   |   bmi > 12.635
|   |   |   |   |   |   |   bmi > 19.264
|   |   |   |   |   |   |   |   bmi > 27.320: 1 {1=11889, 0=5192}
|   |   |   |   |   |   |   |   bmi ≤ 27.320: 0 {1=4826, 0=7032}
|   |   |   |   |   |   |   bmi ≤ 19.264: 0 {1=34, 0=234}
|   |   |   |   |   |   bmi ≤ 12.635: 0 {1=0, 0=2}
|   |   |   |   |   age ≤ 38.500
|   |   |   |   |   |   hypertension > 0.500
|   |   |   |   |   |   |   bmi > 56.815: 0 {1=0, 0=2}
|   |   |   |   |   |   |   bmi ≤ 56.815: 1 {1=117, 0=81}
|   |   |   |   |   |   hypertension ≤ 0.500: 0 {1=812, 0=3496}
|   |   |   |   age ≤ 27.500
|   |   |   |   |   hypertension > 0.500
|   |   |   |   |   |   bmi > 36.790
|   |   |   |   |   |   |   age > 24.500: 1 {1=7, 0=0}
|   |   |   |   |   |   |   age ≤ 24.500: 0 {1=0, 0=3}
|   |   |   |   |   |   bmi ≤ 36.790: 0 {1=1, 0=18}
|   |   |   |   |   hypertension ≤ 0.500
|   |   |   |   |   |   bmi > 55.615
|   |   |   |   |   |   |   age > 24: 1 {1=2, 0=1}
|   |   |   |   |   |   |   age ≤ 24: 0 {1=0, 0=4}
|   |   |   |   |   |   bmi ≤ 55.615: 0 {1=393, 0=7648}
|   |   |   blood_glucose_level ≤ 113: 0 {1=0, 0=10658}
|   |   hbA1c_level ≤ 5.350: 0 {1=0, 0=24391}
```

ภาพที่ 3.14 แสดงคำบรรยายลักษณะงาน Decision Tree ในโปรแกรม Rapid Miner

เมื่อได้โมเดล Decision tree แล้วจากนั้น จึงนำโมเดลมาใช้ในการเขียน กฎ Decision Tree เพื่อในในการตัดสินใจ โดยสามารถจำแนกกฎได้ ดังนี้

กฎข้อที่ 1 IF hbA1c_level > 6.600 ผลลัพธ์ = 1 หมายความว่า หากค่า hbA1c_level สูงกว่า 6.600 จะได้ผลลัพธ์เป็น 1 ถือว่าเข้าสู่ภาวะเบาหวานแน่นอน

กฎข้อที่ 2 IF hbA1c_level ≤ 6.600 AND blood_glucose_level > 200.500 ผลลัพธ์ = 1 หมายความว่า ถ้าค่า hbA1c_level ไม่เกิน 6.600 แต่กลูโคสในเลือดสูงมาก (>200.500) จะได้ผลลัพธ์เป็น 1 ซึ่งแสดงถึงระดับน้ำตาลในเลือดที่สูงมาก

กฎข้อที่ 3 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND age > 38.5 AND bmi > 27.320 ผลลัพธ์ = 1 หมายความว่า กลุ่มอายุ 38 ปีขึ้นไปที่น้ำหนักเกิน (BMI สูงมาก) และมีค่าน้ำตาลในเลือดมากกว่า 113 mg/dL มีแนวโน้มเป็นเบาหวานสูง

กฎข้อที่ 4 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND age > 38.5 AND 19.264 < bmi ≤ 27.320 ผลลัพธ์ = 0 หมายความว่า อายุเกิน 38 ปีแต่น้ำหนักแค่เกินเล็กน้อย–ปานกลาง ส่วนใหญ่ยังไม่เป็นเบาหวาน

กฎข้อที่ 5 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND age > 38.5 AND bmi ≤ 12.635 ผลลัพธ์ = 0 หมายความว่า ผู้ที่มีค่า BMI ต่ำมาก (ผอมมาก) แม้อายุมากแต่น้ำตาลในเลือดไม่สูงจัด มักไม่เป็นเบาหวาน

กฎข้อที่ 6 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND 27.5 < age ≤ 38.5 AND hypertension > 0.5 ผลลัพธ์ = 1 หมายความว่า กลุ่มวัยกลาง (~30–38 ปี) ที่มีความดันโลหิตสูงร่วมด้วย มักเข้าสู่ภาวะเบาหวาน

กฎข้อที่ 7 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND 27.5 < age ≤ 38.5 AND hypertension ≤ 0.5 ผลลัพธ์ = 0 หมายความว่า อายุไม่เกิน 38 ปี และไม่มีความดันโลหิตสูง มีแนวโน้มว่ายังไม่เป็นเบาหวาน

กฎข้อที่ 8 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND age ≤ 27.5 AND hypertension > 0.5 AND bmi > 36.790 ผลลัพธ์ = 1 หมายความว่า กลุ่มอายุน้อยกว่า 27 ปีที่มีทั้งความดันโลหิตสูงและน้ำหนักมาก (อ้วนจัด) มีความเสี่ยงสูงเป็นเบาหวาน

กฎข้อที่ 9 IF hbA1c_level ≤ 6.600 AND blood_glucose_level ≤ 200.500 AND hba1c_level > 5.350 AND blood_glucose_level > 113 AND age ≤ 27.5 AND hypertension ≤ 0.5 ผลลัพธ์ = 0 หมายความว่า ผู้วัยหนุ่มสาว ไม่มีความดันสูง และค่าน้ำตาลเพียงสูงเล็กน้อย ยังไม่เป็นเบาหวาน

กฎข้อที่ 10 IF hbA1c_level ≤ 5.350 ผลลัพธ์ = 0 หมายความว่า ค่าฮีโมโกลบิน (HbA1c) ต่ำกว่า 5.35 จัดว่าอยู่ในเกณฑ์ปกติ ไม่เป็นเบาหวาน

จากผลการประเมินพบว่า โมเดลมีความแม่นยำโดยรวมเท่ากับ 90.69% แต่เมื่อพิจารณาแล้วเห็นว่าโมเดลมีปัญหาใหญ่ เพราะทำนายได้เพียงคลาส 0 เท่านั้น

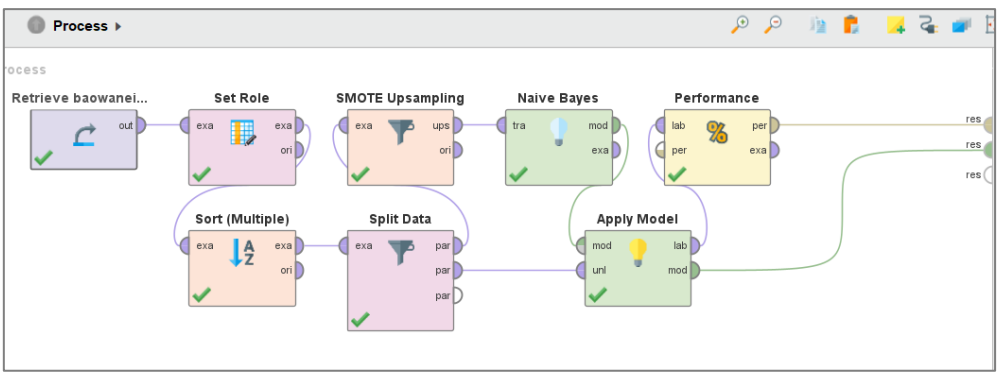
- ข้อมูลจริงที่เป็น คลาส 0 ⟶ โมเดลทำนายถูกทั้งหมด 26,116 ตัวอย่าง
- ข้อมูลจริงที่เป็น คลาส 1 ⟶ โมเดลทำนายผิดทั้งหมด 2,680 ตัวอย่าง โดยทำนายเป็นคลาส 0 หมด

ผลที่ตามมาคือ

- สำหรับ คลาส 0 ⟶ โมเดลทำงานได้ดีมาก ทั้ง precision และ recall สูง
- สำหรับ คลาส 1 ⟶ โมเดลไม่สามารถจับค่าได้ (precision และ recall = 0%)

แม้ Accuracy จะสูง เพราะข้อมูลส่วนใหญ่เป็นคลาส 0 แต่โมเดลไม่สามารถทำนายคลาส 1 ได้เลย สะท้อนให้เห็นถึงความไม่สมดุลของข้อมูล (class imbalance) ที่ทำให้โมเดลเรียนรู้ไปทางคลาสที่มีจำนวนมากกว่าเป็นหลัก ในส่วนโครงสร้างต้นไม้ตัดสินใจสะท้อนว่า HbA1c_level และ blood_glucose_level เป็นตัวแปรที่มีอิทธิพลสูงสุด ต่อการจำแนกผลลัพธ์ ขณะที่ตัวแปร age, BMI และ hypertension ทำหน้าที่เสริมในลำดับขั้นถัดมา เพื่อเพิ่มความแม่นยำในการทำนาย

3.1.4.2 เทคนิคนาอีฟเบย์ (Naïve Bayes)



ภาพที่ 3.15 แสดงกระบวนการประมวลผลกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio

จากภาพที่ 3.15 เป็นกระบวนการสร้างโมเดล Naive Bayes เริ่มจากการเรียกใช้ชุดข้อมูลเบาหวานผ่านโหนด Retrieve และใช้โหนด Set Role กำหนดบทบาทของแอตทริบิวต์ diabetes ให้เป็นตัวแปรเป้าหมาย (label) จากนั้นใช้โหนด Sort Multiple เรียงลำดับข้อมูลตามแอตทริบิวต์ diabetes ในลำดับแบบ descending หรือเรียงจากค่ามากไปน้อย จากนั้นข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data ทำหน้าที่แบ่งข้อมูลออกเป็นสองส่วน คือ 70% สำหรับชุดฝึก (training set) และ 30% สำหรับชุดทดสอบ (test set) ตามค่า ratio ที่กำหนดไว้ ส่วนการเลือก linear sampling หมายถึงการแบ่งข้อมูลตามลำดับที่จัดเรียงไว้ ไม่ได้สุ่ม ทำให้ข้อมูลถูกแบ่งแบบต่อเนื่องตามลำดับ พร้อมทั้งเลือกวิธีการสุ่มตัวอย่างแบบ linear sampling เพื่อจัดเรียงข้อมูลตามลำดับที่กำหนด ซึ่งชุดฝึกผ่านการปรับสมดุลด้วย SMOTE Upsampling แล้วนำไปสร้างโมเดลด้วย Naive Bayes ต่อไป

accuracy: 89.16%

	true 0	true 1	class precision
pred. 0	25675	0	100.00%
pred. 1	3121	0	0.00%
class recall	89.16%	0.00%	

ภาพที่ 3.16 ผลลัพธ์ของโมเดล Naïve Bayes โดยใช้ SMOTE

PerformanceVector

```
PerformanceVector:
accuracy: 89.16%
ConfusionMatrix:
True:    0       1
0:       25675   0
1:       3121    0
```

ภาพที่ 3.17 แสดงคำบรรยายลักษณะงาน Naïve Bayes ในโปรแกรม Rapid Miner

จากภาพที่ 3.16 แสดงผลการประเมินประสิทธิภาพของโมเดล Naïve Bayes โมเดลมีค่า Accuracy = 89.16% ซึ่งดูเหมือนสูง แต่เมื่อดูรายละเอียดใน Confusion Matrix จะเห็นปัญหาชัดเจน คือ

- คลาส 0 (ไม่เป็นโรค/ปกติ) : โมเดลทำนายถูกทั้งหมด 25,675 ตัวอย่าง

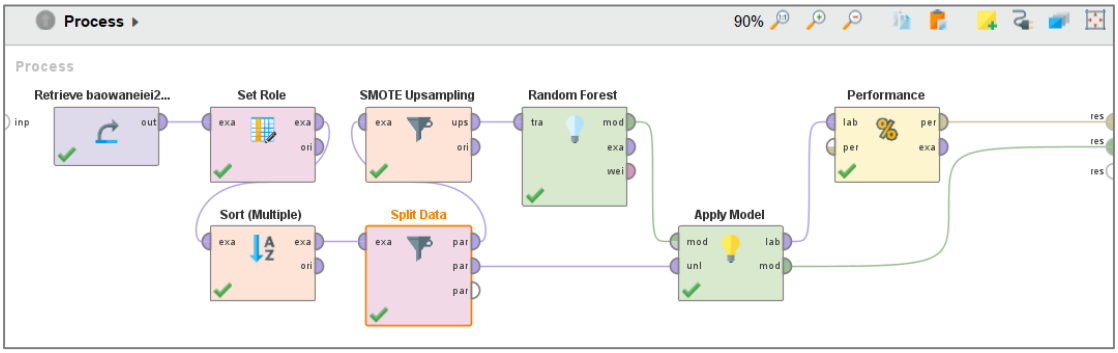
- คลาส 1 (เป็นโรค/เสี่ยง) : โมเดลทำนายผิดทั้งหมด 3,121 ตัวอย่าง เพราะระบบทำนายเป็นคลาส 0 หมด
- ไม่มีตัวอย่างใดเลยที่ถูกทำนายเป็นคลาส 1

ดังนั้น ค่าที่ได้สรุปคือ

- Precision ของคลาส 0 = 100%, แต่ Recall ของคลาส 1 = 0%
- หมายความว่าโมเดลสามารถทำนายคลาส 0 ได้ดีมาก แต่ไม่สามารถแยกแยะหรือทำนายคลาส 1 ได้

แม้โมเดล Naive Bayes จะได้ Accuracy สูงถึง 89.136 แต่ประสิทธิภาพแท้จริงไม่ดี เพราะโมเดลล้มเหลวในการจำแนกคลาส 1 ทั้งหมด สาเหตุเกิดจากข้อมูลไม่สมดุล (class imbalance) ทำให้โมเดลเรียนรู้ไปทางคลาสที่มีจำนวนมากกว่าและทำนายเป็นคลาสนั้นเสมอ

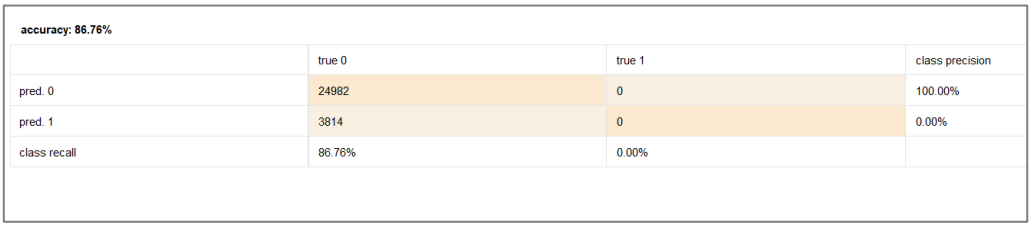
3.1.4.3 เทคนิคป่าสุ่ม (Random Forest)



ภาพที่ 3.18 แสดงกระบวนการประมวลผลกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio

จากภาพที่ 3.18 เป็นการสร้างกระบวนการจำแนกข้อมูลด้วยอัลกอริทึม Random Forest เริ่มจากการนำเข้าข้อมูล และใช้โหนด Set Role กำหนดบทบาทของแอตทริบิวต์ diabetes ให้เป็นตัวแปรเป้าหมาย (label) จากนั้นใช้โหนด Sort (Multiple) เรียงลำดับข้อมูลตามแอตทริบิวต์ diabetes ในลำดับแบบ descending เพื่อควบคุมลำดับ จากนั้นข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data โดยกำหนดอัตราส่วนเป็น 0.7 : 0.3 ซึ่งหมายความว่า 70% ของข้อมูลจะถูกนำไปใช้สำหรับการฝึกโมเดล อีก 30% ใช้สำหรับการประเมินผลการทำงานของโมเดล เพื่อป้องกันการเกิด overfitting และช่วยให้สามารถประเมินประสิทธิภาพของโมเดลกับข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างเหมาะสม พร้อมทั้งเลือกวิธีการสุ่มตัวอย่างแบบ linear sampling เพื่อจัดเรียงข้อมูลตามลำดับที่กำหนด ซึ่งชุดฝึกผ่านการปรับสมดุลด้วย SMOTE Upsampling แล้วนำไปสร้างโมเดล Random Forest ซึ่งเป็นอัลกอริทึมที่รวมผลของ

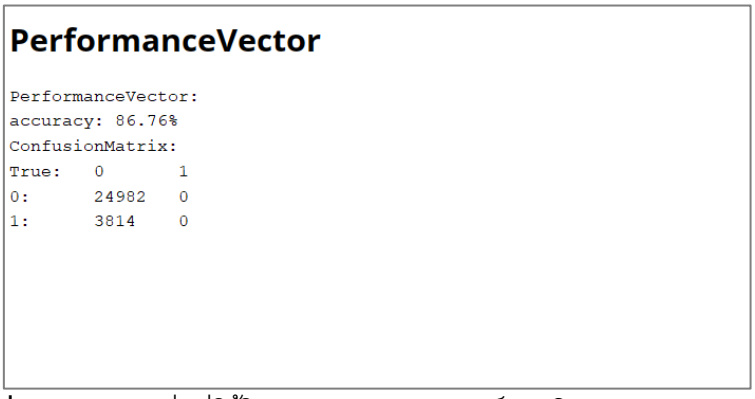
หลายๆ ต้นไม้ตัดสินใจเข้าด้วยกัน (ensemble learning) เพื่อเพิ่มความแม่นยำและลดความเอนเอียงของการทำนาย



accuracy: 86.76%

	true 0	true 1	class precision
pred. 0	24982	0	100.00%
pred. 1	3814	0	0.00%
class recall	86.76%	0.00%	

ภาพที่ 3.19 ผลลัพธ์ของโมเดล Random Forest



PerformanceVector

```
PerformanceVector:
accuracy: 86.76%
ConfusionMatrix:
True:    0       1
0:       24982   0
1:       3814    0
```

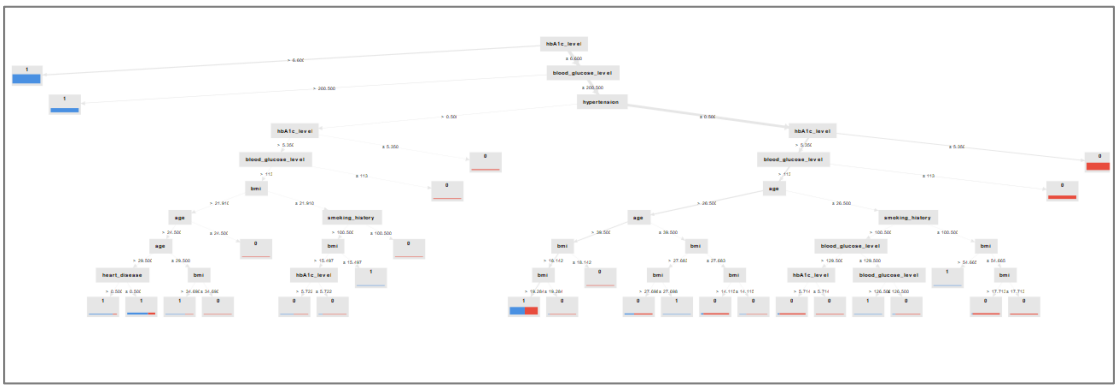
ภาพที่ 3.20 แสดงค่าที่ใช้ในการประเมินผลลัพธ์ของโมเดล Random Forest

จากภาพที่ 3.19 และ 3.20 เป็นผลการประเมินโมเดล Random Forest มีค่า Accuracy = 86.76% ซึ่งถือว่าสูง แต่เมื่อตรวจสอบรายละเอียดพบว่าโมเดลทำนายได้เฉพาะคลาส 0 เท่านั้น

- ข้อมูลจริงเป็น คลาส 0 ทำนายถูกทั้งหมด 24,982 ตัวอย่าง
- ข้อมูลจริงเป็น คลาส 1 ทำนายผิดทั้งหมด 3,814 ตัวอย่าง (ถูกทำนายเป็นคลาส 0 หมด)
- ไม่มีตัวอย่างใดเลยที่ถูกทำนายเป็นคลาส 1

ดังนั้น ถึงแม้ Accuracy จะยังสูง แต่โมเดลไม่สามารถแยกแยะคลาส 1 ทำให้ค่า Precision และ Recall ของคลาส 1 เท่ากับ 0% ในขณะที่คลาส 0 มี Precision 100% และ Recall 86.76% แต่ผลลัพธ์นี้สะท้อนถึงโมเดลที่มีความมั่นคงและปลอดภัยในการใช้งาน โดยเฉพาะใน

บริบทที่ต้องการลดความผิดพลาดจากการแจ้งเตือนเกินหรือการคัดกรองผิดกลุ่ม ซึ่งถือเป็นจุดแข็งที่สามารถนำไปต่อยอดได้อย่างมีประสิทธิภาพ



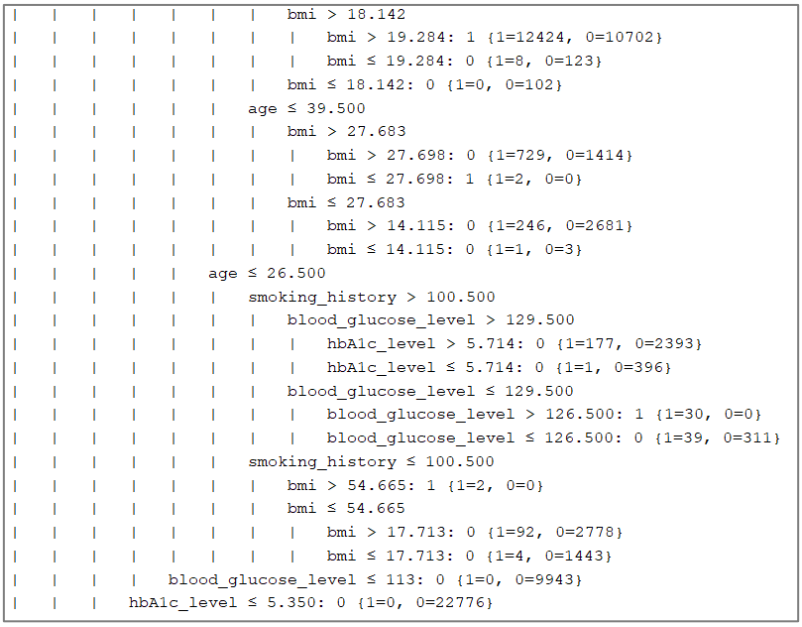
ภาพที่ 3.21 แสดงโมเดล Graph Random Forest ในโปรแกรม Rapid Miner Studio

โมเดล Random Forest ในแผนภาพนี้แสดงการสร้างต้นไม้ตัดสินใจ เห็นว่าตัวแปรสำคัญในการจำแนกความเสี่ยงโรคเบาหวาน ได้แก่ hbA1c_level และ blood_glucose_level ซึ่งถูกใช้เป็นเกณฑ์การแบ่งข้อมูลในลำดับต้นๆ ส่วนตัวแปรอื่นๆ เช่น hypertension, bmi, age, smoking_history และ heart_disease ทำหน้าที่เป็นปัจจัยเสริมเพื่อเพิ่มความแม่นยำในการจำแนก โหนดปลายจะแสดงผลลัพธ์ว่าเป็นกลุ่ม 0 (ปกติ) หรือ 1 (เสี่ยง/ป่วย) โดยสีของโหนดบอกสัดส่วนข้อมูลในแต่ละคลาส

Tree

```
hbA1c_level > 6.600: 1 {1=28402, 0=0}
hbA1c_level ≤ 6.600
|   blood_glucose_level > 200.500: 1 {1=12393, 0=0}
|   blood_glucose_level ≤ 200.500
|   |   hypertension > 0.500
|   |   |   hbA1c_level > 5.350
|   |   |   |   blood_glucose_level > 113
|   |   |   |   |   bmi > 21.910
|   |   |   |   |   |   age > 24.500
|   |   |   |   |   |   |   age > 29.500
|   |   |   |   |   |   |   |   heart_disease > 0.500: 1 {1=737, 0=128}
|   |   |   |   |   |   |   |   heart_disease ≤ 0.500: 1 {1=3635, 0=1239}
|   |   |   |   |   |   |   age ≤ 29.500
|   |   |   |   |   |   |   |   bmi > 34.690: 1 {1=7, 0=3}
|   |   |   |   |   |   |   |   bmi ≤ 34.690: 0 {1=0, 0=10}
|   |   |   |   |   |   age ≤ 24.500: 0 {1=0, 0=10}
|   |   |   |   |   bmi ≤ 21.910
|   |   |   |   |   |   smoking_history > 100.500
|   |   |   |   |   |   |   bmi > 15.497
|   |   |   |   |   |   |   |   hbA1c_level > 5.722: 0 {1=44, 0=47}
|   |   |   |   |   |   |   |   hbA1c_level ≤ 5.722: 0 {1=3, 0=17}
|   |   |   |   |   |   |   bmi ≤ 15.497: 1 {1=4, 0=0}
|   |   |   |   |   |   smoking_history ≤ 100.500: 0 {1=0, 0=5}
|   |   |   |   blood_glucose_level ≤ 113: 0 {1=0, 0=667}
|   |   |   hbA1c_level ≤ 5.350: 0 {1=0, 0=1353}
|   |   hypertension ≤ 0.500
|   |   |   hbA1c_level > 5.350
|   |   |   |   blood_glucose_level > 113
|   |   |   |   |   age > 26.500
|   |   |   |   |   |   age > 39.500
|   |   |   |   |   |   |   bmi > 18.142
```

ภาพที่ 3.22 แสดงคำบรรยายลักษณะงาน Random Forest ในโปรแกรม Rapid Miner



ภาพที่ 3.23 แสดงคำบรรยายลักษณะงาน Random Forest ในโปรแกรม Rapid Miner

ผลลัพธ์ของโมเดล Random Forest ที่สร้างขึ้นในภาพนี้ถูกใช้ในการจำแนกผู้ป่วยออกเป็น กลุ่มเสี่ยง และ ไม่เสี่ยง ต่อโรค โดยใช้ข้อมูลทางสุขภาพหลายตัวแปร เช่น อายุ (age), ระดับ HbA1c (HbA1c_level), ระดับน้ำตาลในเลือด (blood_glucose_level), ดัชนีมวลกาย (BMI) และ ภาวะความดันโลหิตสูง (hypertension) เป็นตัวแปรหลักในการตัดสินใจ โครงสร้างต้นไม้แสดงเงื่อนไขที่ใช้ในการแบ่งกลุ่มข้อมูลตามลำดับ และสามารถทำนายผลลัพธ์ว่าเป็นคลาส 0 (ไม่เสี่ยง) หรือ คลาส 1 (เสี่ยง) ตามเงื่อนไขที่กำหนดในแต่ละโหนด

กฎข้อที่ 1: IF hbA1c_level > 6.600 ผลลัพธ์ = 1 หมายความว่า HbA1c สูงกว่า 6.6 เข้าข่ายภาวะเบาหวานตามเกณฑ์ อาจมีภาวะเบาหวานระดับเริ่มต้น

กฎข้อที่ 2: IF hbA1c_level ≤ 6.600 AND blood_glucose_level > 200.500 ผลลัพธ์ = 1 หมายความว่า แม้ HbA1c ยังไม่เกิน 6.6 แต่ระดับน้ำตาลในเลือดขณะตรวจสูงมาก (>200.5 mg/dL) แสดงถึงภาวะเบาหวานที่ควรได้รับการดูแล

กฎข้อที่ 3: IF hbA1c_level > 5.350 AND blood_glucose_level > 113 AND bmi > 21.910 AND age > 24.500 AND age > 29.500 AND heart_disease > 0.500 ผลลัพธ์ = 1 หมายความ ค่า HbA1c มากกว่า 5.35% และน้ำตาลในเลือดมากกว่า 113 mg/dL และค่า

BMI มากกว่า 21.91 kg/m² ประกอบกับ อายุเกิน 30 ปี และมีภาวะโรคหัวใจ (Heart Disease = 1) จะถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 4: IF hbA1c_level > 5.350 AND blood_glucose_level > 113 AND bmi > 21.910 AND age > 24.500 AND age > 29.500 AND heart_disease ≤ 0.500 ผลลัพธ์ = 1 หมายความว่า ค่า HbA1c มากกว่า 5.35% และน้ำตาลในเลือดมากกว่า 113 mg/dL และค่าBMI มากกว่า 21.91 kg/m² ประกอบกับ อายุเกิน 30 ปี และมีภาวะโรคหัวใจ (Heart Disease = 1) จะถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 5: IF age ≤ 29.500 AND bmi > 34.690 ผลลัพธ์ = 1 หมายความว่า ถ้าอายุ <30 แต่ BMI สูงมาก (อ้วนมาก) ร่วมกับน้ำตาลสูง ถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 6: IF age ≤ 29.500 AND bmi ≤ 34.690 ผลลัพธ์ = 0 หมายความว่า ถ้าอายุ <30 แต่ BMI ต่ำกว่า 34.69 ยังไม่จัดเป็นกลุ่มเสี่ยง

กฎข้อที่ 7: IF age ≤ 24.500 ผลลัพธ์ = 0 หมายความว่า ถ้าอายุไม่เกิน 25 ไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 8: IF bmi ≤ 21.910 AND smoking_history > 100.500 AND bmi > 15.497 AND hbA1c_level > 5.722 ผลลัพธ์ = 0 หมายความว่า ถ้าผอม/น้ำหนักน้อย สูบบุหรี่ และค่า HbA1c สูงกว่า 5.722 แต่ยังจัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 9: IF bmi ≤ 21.910 AND smoking_history > 100.500 AND bmi > 15.497 AND hbA1c_level ≤ 5.722 ผลลัพธ์ = 0 หมายความว่า ถ้าผอม/น้ำหนักน้อย สูบบุหรี่ และค่า HbA1c ไม่เกิน 5.722 แต่ยังจัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 10: bmi ≤ 21.910 AND smoking_history > 100.500 AND bmi ≤ 15.497 ผลลัพธ์ = 1 หมายความว่า ถ้าค่า BMI ไม่เกิน 21.910 และมีประวัติสูบบุหรี่ ถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 11: IF bmi ≤ 21.910 AND smoking_history ≤ 100.500 ผลลัพธ์ = 0 หมายความว่า ผอมและไม่มีสัญญาณสูบบุหรี่ จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 12: IF blood_glucose_level ≤ 113 ผลลัพธ์ = 0 หมายความว่า ถ้าปริมาณของน้ำตาลกลูโคสที่อยู่ในกระแสเลือด ไม่เกิน 113 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 13: IF hbA1c_level ≤ 5.350 ผลลัพธ์ = 0 หมายความว่า หากค่า HbA1c ต่ำกว่าหรือเท่ากับ 5.35 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 14: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age > 39.500 AND bmi > 18.142 AND bmi > 19.284 ผลลัพธ์ = 1 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI > 19.284 ถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 15: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age > 39.500 AND bmi > 18.142 AND bmi ≤ 19.284 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI ไม่เกิน 19.284 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 16: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age > 39.500 AND bmi > 18.142 AND bmi ≤ 18.142 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI ไม่เกิน 18.142 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 17: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age ≤ 39.500 AND bmi > 27.683 AND bmi > 27.698 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI มากกว่า 27.698 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 18: : IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age ≤ 39.500 AND bmi > 27.683 AND

bmi ≤ 27.698 ผลลัพธ์ = 1 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI ไม่เกิน 27.698 ถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 19: : IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age ≤ 39.500 bmi ≤ 27.683 AND bmi > 14.115 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI ไม่เกิน 27.683 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 20: : IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 AND age > 26.500 AND age ≤ 39.500 AND bmi > 27.683 AND bmi ≤ 14.115 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และอายุ > 40 ปีและมีค่า BMI ไม่เกิน 27.698 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 21: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND blood_glucose_level > 129.500 AND hbA1c_level > 5.714 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.714% และน้ำตาลในเลือด > 129.5 และมีประวัติสูบบุหรี่ จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 22: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND blood_glucose_level > 129.500 AND hbA1c_level ≤ 5.714 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c ระหว่าง 5.35%–5.714% และน้ำตาลในเลือด > 129.5 และมีประวัติสูบบุหรี่ จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 23: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND blood_glucose_level ≤ 129.500 AND blood_glucose_level > 126.500 ผลลัพธ์ = 1 หมายความว่า หากมีความดันโลหิต ไม่เกิน

0.5 และมีค่า HbA1c มากกว่า 5.35% และน้ำตาลในเลือดอยู่ระหว่าง 126–129.5 และมีประวัติสูบบุหรี่ ถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

กฎข้อที่ 24: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND blood_glucose_level ≤ 129.500 AND blood_glucose_level ≤ 126.500 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c มากกว่า 5.35% และน้ำตาลในเลือดอยู่ระหว่าง 113–129.5 และมีประวัติสูบบุหรี่ จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 25: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND bmi > 54.665 ผลลัพธ์ = 1 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และมีประวัติสูบบุหรี่ และมีค่า BMI สูงกว่า 54.665 ถูกจัดให้อยู่ในกลุ่มที่มีความเสี่ยงสูงต่อการเกิดโรคเบาหวาน

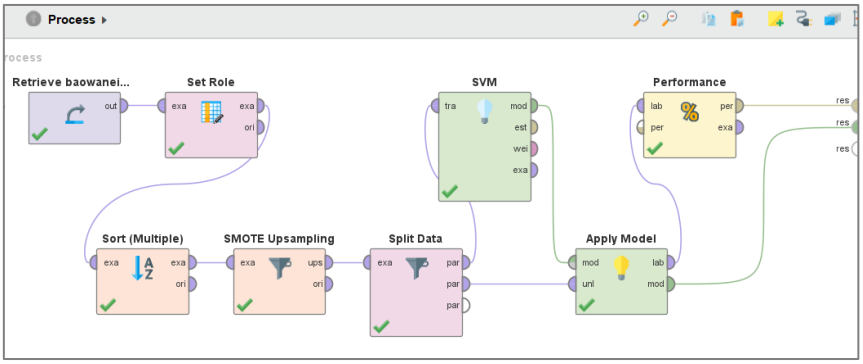
กฎข้อที่ 26: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND bmi ≤ 54.665 AND bmi > 17.713 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และมีประวัติสูบบุหรี่ และมีค่า BMI ระหว่าง 17.714–54.665 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 27: IF hypertension ≤ 0.500 AND hbA1c_level > 5.350 AND blood_glucose_level > 113 smoking_history > 100.500 AND bmi ≤ 54.665 AND bmi ≤ 17.713 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 และมีค่า HbA1c > 5.35% และน้ำตาลในเลือด > 113 และมีประวัติสูบบุหรี่ และมีค่า BMI ไม่เกิน 54.665 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 28: IF hypertension ≤ 0.500 AND blood_glucose_level ≤ 113 ผลลัพธ์ = 0 หมายความว่า หากมีความดันโลหิต ไม่เกิน 0.5 จัดว่าไม่ใช่กลุ่มเสี่ยง

กฎข้อที่ 29: IF hbA1c_level ≤ 5.350 ผลลัพธ์ = 0 หมายความว่า หากมีระดับน้ำตาลเฉลี่ยในเลือดในช่วง 2–3 เดือน ไม่เกิน 5.35% จัดว่าไม่ใช่กลุ่มเสี่ยง

3.1.4.4 เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)



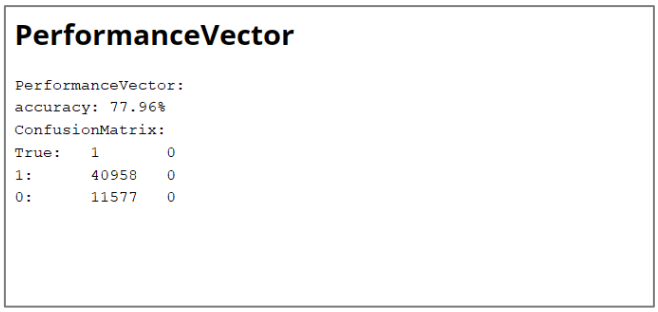
ภาพที่ 3.24 แสดงกระบวนการประมวลผลกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio

จากภาพที่ 3.24 เป็นการสร้างกระบวนการจำแนกข้อมูลด้วยอัลกอริทึม Support Vector Machine เริ่มจากการนำเข้าข้อมูล และใช้โหนด Set Role กำหนดบทบาทของแอตทริบิวต์ diabetes ให้เป็นตัวแปรเป้าหมาย (label) จากนั้นใช้โหนด Sort (Multiple) เรียงลำดับข้อมูลตามแอตทริบิวต์ diabetes ในลำดับแบบ descending เพื่อควบคุมลำดับ จากนั้นนำข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data จากนั้นข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data โดยกำหนดอัตราส่วนเป็น 70 : 30 ซึ่งหมายความว่า 70% ของข้อมูลจะถูกนำไปใช้สำหรับการฝึกโมเดล อีก 30% ใช้สำหรับการประเมินผลการทำงานของโมเดล พร้อมทั้งเลือกวิธีการสุ่มตัวอย่างแบบ linear sampling เพื่อจัดเรียงข้อมูลตามลำดับที่กำหนด ซึ่งชุดฝึกผ่านการปรับสมดุลด้วย SMOTE Upsampling เพื่อให้ชุดข้อมูลมีความสมดุลมากขึ้น แล้วนำไปสร้างโมเดล Support Vector Machine ต่อไป

accuracy: 77.96%

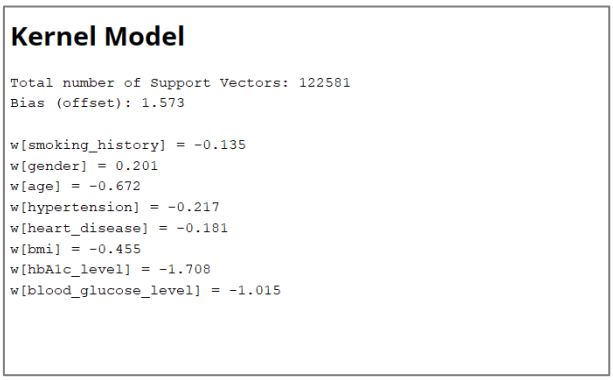
	true 1	true 0	class precision
pred. 1	40958	0	100.00%
pred. 0	11577	0	0.00%
class recall	77.96%	0.00%	

ภาพที่ 3.25 ผลลัพธ์ของโมเดล Support Vector Machine



ภาพที่ 3.26 แสดงค่าที่ใช้ในการประเมินผลลัพธ์ของโมเดล Support Vector Machine

โมเดลมีความแม่นยำโดยรวมสูงถึง 77.96% ซึ่งแบ่งเป็น Class 0 จำนวน 40,958 รายการ และ Class 1 จำนวน 11,577 รายการ และมีจุดเด่นสำคัญคือ สามารถจำแนก Class 0 ได้อย่างแม่นยำสมบูรณ์แบบ 100% ไม่มีข้อผิดพลาดเลย ซึ่งแสดงให้เห็นถึงประสิทธิภาพที่ยอดเยี่ยมในการระบุข้อมูลส่วนใหญ่ของชุดข้อมูล อย่างไรก็ตาม โมเดลกลับมีข้อบกพร่องที่ร้ายแรงคือ ล้มเหลวโดยสิ้นเชิงในการจำแนก Class 1 โดยไม่สามารถระบุข้อมูลในกลุ่มนี้ได้เลยแม้แต่รายการเดียว ทำให้โมเดลนี้ไม่เหมาะกับการนำไปใช้งานจริงหากต้องการจำแนกข้อมูลทั้งสองประเภทอย่างเท่าเทียม



ภาพที่ 3.27 แสดงคำบรรยายลักษณะงาน Support Vector Machine

ผลลัพธ์นี้แสดงให้เราเห็นถึงโครงสร้างและปัจจัยสำคัญที่โมเดลใช้ในการตัดสินใจ โดยโมเดลนี้มีจุดข้อมูลสำคัญที่เรียกว่า Support Vectors ถึง 122,581 จุด ซึ่งเป็นตัวกำหนดเส้นแบ่งสำหรับแยกแยะข้อมูล นอกจากนี้ยังมีค่า Bias หรือค่าคงที่ในการทำนายอยู่ที่ 1.573

Attribute	Weight
smoking_history	-0.135
gender	0.201
age	-0.672
hypertension	-0.217
heart_disease	-0.181
bmi	-0.455
hbA1c_level	-1.708
blood_glucose_level	-1.015

ภาพที่ 3.28 ผลลัพธ์ของโมเดล Support Vector Machine

counter	label	function...	alpha	abs(alp...	support...	smokin...	gender	age	hyperte...	heart_di...	bmi	hbA1c_l...	blood_g...
0	0	-0.470	0.138	0.138	support v...	0.914	0	0.359	0	0	1.004	0.627	0.150
1	0	-0.440	0.138	0.138	support v...	-0.131	0.867	1.389	0	0	0.021	0.627	0.189
2	0	0.269	0.138	0.138	support v...	1.959	0	0.941	0	0	-0.560	0.627	-0.401
3	1	-1.004	-0	0	no suppor...	-1.176	0	0.538	0	3.840	1.441	0.545	0.091
4	0	-0.492	0.138	0.138	support v...	0.914	0	0.359	0	0	0.964	0.627	0.189
5	0	2.263	0	0	no suppor...	-0.131	0.867	0.001	0	0	-0.529	-0.030	-0.204
6	0	-0.023	0.138	0.138	support v...	-1.176	0	0.717	0	0	0.243	0.627	0.091
7	0	0.073	0.138	0.138	support v...	0.914	0.867	1.523	0	0	-0.955	0.627	-0.106
8	0	-0.319	0.138	0.138	support v...	-0.131	0.867	1.523	0	3.840	-0.080	0.134	0.170
9	1	-2.821	0	0	no suppor...	-0.131	0	1.120	2.778	0	-0.146	0.548	2.155
10	0	2.301	0	0	no suppor...	-0.131	0.867	-0.088	0	0	-0.043	-0.030	-0.401
11	0	2.704	0	0	no suppor...	0.914	0.867	0.359	0	3.840	-0.786	-0.850	-0.204
12	1	0.248	-0.138	0.138	support v...	-0.131	0.867	-1.297	0	0	-0.024	0.825	0.975
13	0	0.283	0.138	0.138	support v...	-1.176	0.867	1.523	0	0	-0.139	0.298	0.150
14	0	-0.029	0.138	0.138	support v...	0.914	0.867	1.478	0	0	-0.972	0.545	0.170
15	0	-0.230	0.138	0.138	support v...	-0.131	0	1.344	0	0	-0.405	0.545	0.170
16	0	0.117	0.138	0.138	support v...	-1.176	0.867	1.075	0	0	0.534	0.545	-0.106
17	0	2.661	0	0	no suppor...	0.914	0.867	0.314	0	3.840	1.169	-0.686	-1.285

ภาพที่ 3.29 ผลลัพธ์ของโมเดล Support Vector Machine

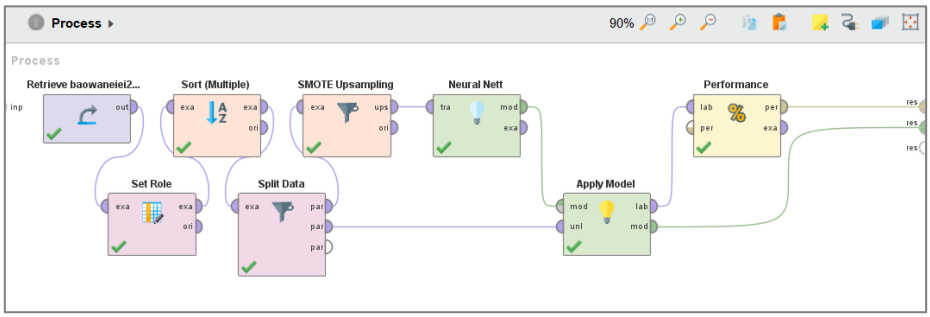
จากตัวเลขในตาราง สามารถเรียงลำดับความสำคัญของปัจจัยจากมากไปน้อยได้ดังนี้

1. hbA1c_level (–1.708): มีอิทธิพลมากที่สุดในเชิงลบต่อการตัดสินใจของโมเดล
2. blood_glucose_level (–1.015): รองลงมา เป็นปัจจัยสำคัญในการทำนายเช่นกันage (–0.823)
3. age (–0.672), bmi (–0.455), hypertension (–0.217), heart_disease (–0.181), smoking_history (–0.135): มีอิทธิพลเชิงลบน้อยลงตามลำดับ

โดยรวมผลลัพธ์แสดงให้เห็นว่า HbA1c level และ Blood glucose level เป็นปัจจัยหลักที่โมเดลใช้ในการจำแนก ซึ่งสะท้อนตรงกับเกณฑ์มาตรฐานทางการแพทย์ ขณะที่ตัวแปรด้านประชากรและสุขภาพอื่น ๆ เช่น Age, BMI, Hypertension, Heart disease, Smoking history

มีบทบาทเสริมในการเพิ่มความแม่นยำ ส่วน Gender แม้มีค่าน้ำหนักเชิงบวก แต่มีอิทธิพลน้อยกว่าปัจจัยทางชีวเคมี

3.1.4.5 เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN)



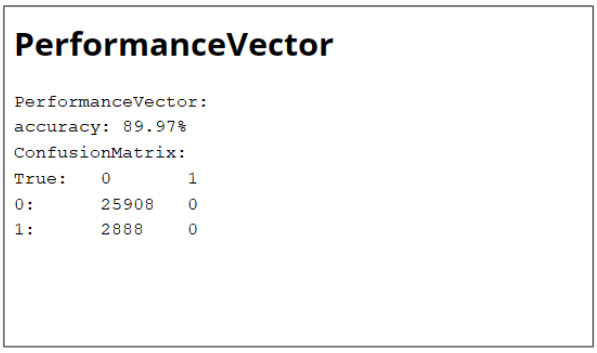
ภาพที่ 3.30 แสดงกระบวนการประมวลผลกลั่นกรองข้อมูลผ่านโปรแกรม RapidMiner Studio

จากภาพที่ 3.21 เป็นการสร้างกระบวนการจำแนกข้อมูลด้วยอัลกอริทึม Artificial Neural Network เริ่มจากการนำเข้าข้อมูล และใช้โหนด Set Role กำหนดบทบาทของแอตทริบิวต์ diabetes ให้เป็นตัวแปรเป้าหมาย (label) จากนั้นใช้โหนด Sort (Multiple) เรียงลำดับข้อมูลตามแอตทริบิวต์ diabetes ในลำดับแบบ descending เพื่อควบคุมลำดับ จากนั้นนำข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data จากนั้นข้อมูลไปแบ่งเป็นชุดฝึกและทดสอบด้วย Split Data โดยกำหนดอัตราส่วนเป็น 70 : 30 ซึ่งหมายความว่า 70% ของข้อมูลจะถูกนำไปใช้สำหรับการฝึกโมเดล อีก 30% ใช้สำหรับการประเมินผลการทำงานของโมเดล พร้อมทั้งเลือกวิธีการสุ่มตัวอย่างแบบ linear sampling เพื่อจัดเรียงข้อมูลตามลำดับที่กำหนด ซึ่งชุดฝึกผ่านการปรับสมดุลด้วย SMOTE Upsampling เพื่อให้ชุดข้อมูลมีความสมดุลมากขึ้น แล้วนำไปสร้างโมเดล Artificial Neural Network ต่อไป

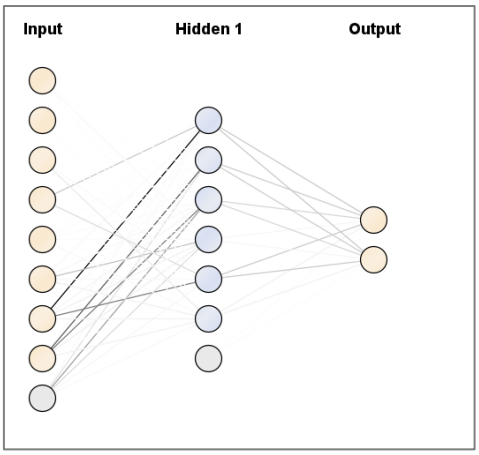
accuracy: 89.97%

	true 0	true 1	class precision
pred. 0	25908	0	100.00%
pred. 1	2888	0	0.00%
class recall	89.97%	0.00%	

ภาพที่ 3.31 ผลลัพธ์ของโมเดล Artificial Neural Network



ภาพที่ 3.32 แสดงค่าที่ใช้ในการประเมินผลลัพธ์ของโมเดล Artificial Neural Network

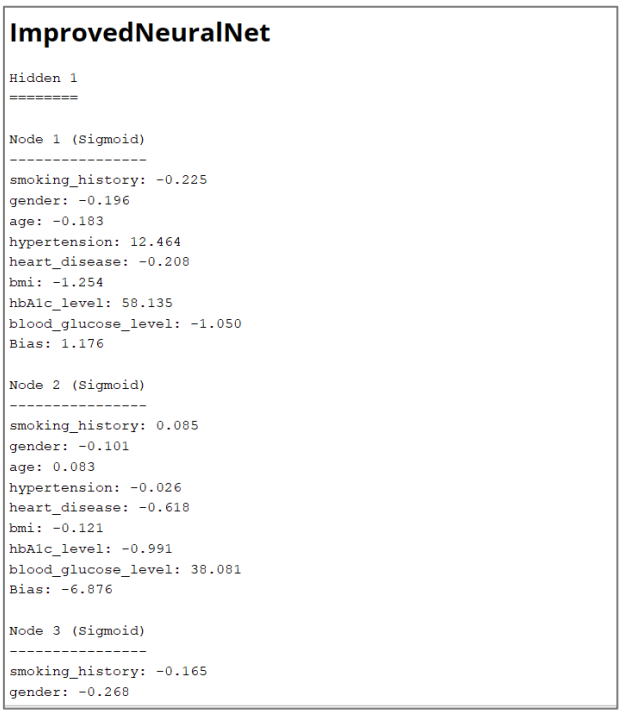


ภาพที่ 3.33 ผลลัพธ์ของโมเดล Artificial Neural Network

จากภาพที่ 3.31 แสดงผลลัพธ์ของโมเดล Artificial Neural Network (ANN) ที่สร้างขึ้นมีความแม่นยำรวมสูงถึง 89.97% แต่เมื่อพิจารณาจาก Confusion Matrix พบว่าโมเดลสามารถทำนายเฉพาะกลุ่ม class 0 ได้ถูกต้องทั้งหมด ขณะที่ class 1 ไม่สามารถทำนายได้เลย ทำให้ recall ของ class 1 เท่ากับศูนย์ ปัญหานี้เกิดจากความไม่สมดุลของข้อมูลที่ class 0 มีจำนวนมากกว่ามาก ส่งผลให้โมเดลเรียนรู้ไปทำนายเพียงกลุ่มใหญ่เพื่อคงค่า accuracy ไว้สูง

ภาพโครงสร้าง Input – Hidden – Output ที่เห็นนั้นคือการทำงานของโครงข่ายประสาทเทียม โดยข้อมูลจากตัวแปรนำเข้าแต่ละตัวจะไหลเข้าสู่ชั้น Input แล้วถูกส่งต่อไปยังโหนดในชั้น Hidden ซึ่งทำหน้าที่ประมวลผลด้วยการคูณค่าน้ำหนักและผ่านฟังก์ชันคณิตศาสตร์เพื่อดึงความสัมพันธ์เชิงซับซ้อนออกมา จากนั้นผลลัพธ์ที่ได้จะถูกส่งไปยังชั้น Output เพื่อแสดงการทำนายขั้นสุดท้าย ในที่นี้คือการตัดสินใจว่าเป็น class 0 หรือ class 1 ซึ่ง

โครงสร้างนี้เปรียบเสมือนสมองที่มีเซลล์ประสาทเชื่อมโยงกันหลายชั้น เพื่อให้โมเดลสามารถเรียนรู้และสรุปผลจากข้อมูลที่ซับซ้อนได้อย่างมีประสิทธิภาพ



ImprovedNeuralNet

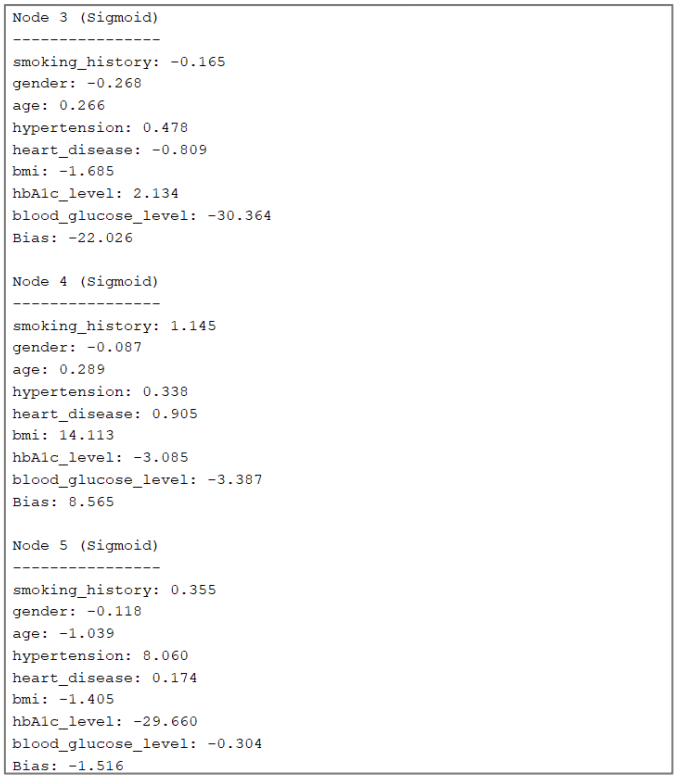
```
Hidden 1
========

Node 1 (Sigmoid)
----------------
smoking_history: -0.225
gender: -0.196
age: -0.183
hypertension: 12.464
heart_disease: -0.208
bmi: -1.254
hbA1c_level: 58.135
blood_glucose_level: -1.050
Bias: 1.176

Node 2 (Sigmoid)
----------------
smoking_history: 0.085
gender: -0.101
age: 0.083
hypertension: -0.026
heart_disease: -0.618
bmi: -0.121
hbA1c_level: -0.991
blood_glucose_level: 38.081
Bias: -6.876

Node 3 (Sigmoid)
----------------
smoking_history: -0.165
gender: -0.268
```

ภาพที่ 3.34 แสดงคำบรรยายลักษณะงาน Artificial Neural Network

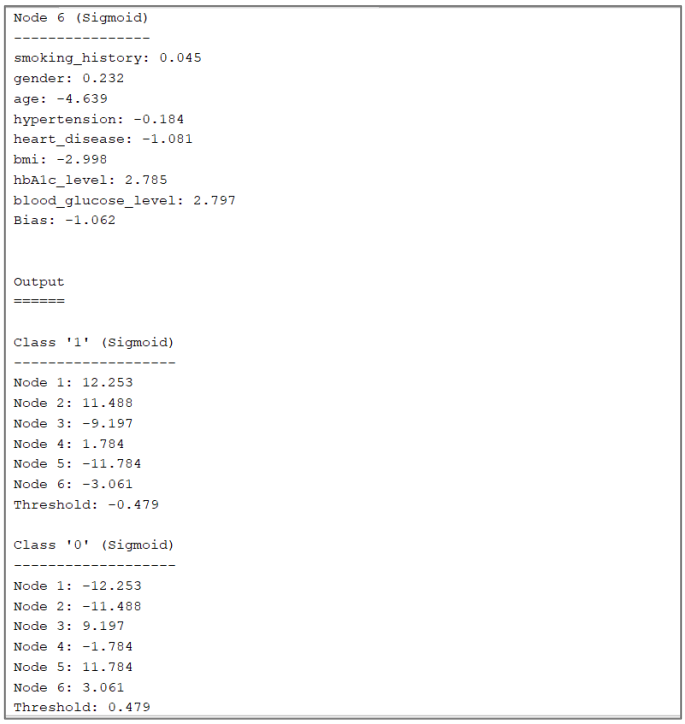


```
Node 3 (Sigmoid)
----------------
smoking_history: -0.165
gender: -0.268
age: 0.266
hypertension: 0.478
heart_disease: -0.809
bmi: -1.685
hbA1c_level: 2.134
blood_glucose_level: -30.364
Bias: -22.026

Node 4 (Sigmoid)
----------------
smoking_history: 1.145
gender: -0.087
age: 0.289
hypertension: 0.338
heart_disease: 0.905
bmi: 14.113
hbA1c_level: -3.085
blood_glucose_level: -3.387
Bias: 8.565

Node 5 (Sigmoid)
----------------
smoking_history: 0.355
gender: -0.118
age: -1.039
hypertension: 8.060
heart_disease: 0.174
bmi: -1.405
hbA1c_level: -29.660
blood_glucose_level: -0.304
Bias: -1.516
```

ภาพที่ 3.35 แสดงคำบรรยายลักษณะงาน Artificial Neural Network



```
Node 6 (Sigmoid)
----------------
smoking_history: 0.045
gender: 0.232
age: -4.639
hypertension: -0.184
heart_disease: -1.081
bmi: -2.998
hbA1c_level: 2.785
blood_glucose_level: 2.797
Bias: -1.062


Output
======

Class '1' (Sigmoid)
-------------------
Node 1: 12.253
Node 2: 11.488
Node 3: -9.197
Node 4: 1.784
Node 5: -11.784
Node 6: -3.061
Threshold: -0.479

Class '0' (Sigmoid)
-------------------
Node 1: -12.253
Node 2: -11.488
Node 3: 9.197
Node 4: -1.784
Node 5: 11.784
Node 6: 3.061
Threshold: 0.479
```

ภาพที่ 3.36 แสดงคำบรรยายลักษณะงาน Artificial Neural Network

โหนดที่ 1: โหนดนี้เน้น HbA1c สูงมาก ร่วมกับ ความดันโลหิตสูง เป็นตัวผลักดันหลัก แม้ BMI, น้ำตาล และปัจจัยอื่น ๆ จะกดสัญญาณเล็กน้อย แต่ผลรวมยังคงทำให้โหนดถูกกระตุ้นแรง ถือเป็น ตัวจับ HbA1c + ความดันโลหิต

โหนดที่ 2: ถูกขับเคลื่อนด้วย น้ำตาลในเลือดสูง เป็นหลัก เสริมด้วยอายุและประวัติการสูบบุหรี่ แต่ถูกกดด้วย HbA1c และโรคหัวใจ พร้อมค่า Bias ลบที่รุนแรง ทำให้ต้องใช้น้ำตาลในเลือดสูงมากจึงจะกระตุ้นได้แรง เป็น โหนดเน้นกลูโคสเฉียบพลัน

โหนดที่ 3: โหนดนี้ถูกกดสัญญาณแรงจาก กลูโคสสูงมาก และ Bias ลบ แต่ยังได้รับแรงเสริมเล็กน้อยจากอายุ ความดัน และ HbA1c จึงเป็น โหนดกดกลูโคส พร้อมถ่วงสมดุลจากปัจจัยเสริม

โหนดที่ 4: เน้นปัจจัยพฤติกรรมและโรคร่วม เช่น การสูบบุหรี่, BMI สูง, โรคหัวใจ ที่ดันสัญญาณขึ้นแรง แม้ HbA1c และกลูโคสจะกดสัญญาณลง แต่ผลรวมยังเป็นบวกชัดเจน ถือเป็น โหนดจับปัจจัยพฤติกรรม + โรคร่วม

โหนดที่ 5: ให้ความสำคัญกับ ความดันโลหิตสูง และการสูบบุหรี่ที่ดันสัญญาณขึ้น แต่จะถูกกดแรงทันทีหาก HbA1c สูงมาก แสดงว่าเป็น โหนดเสริมความดัน แต่ถูกควบคุมด้วย HbA1c

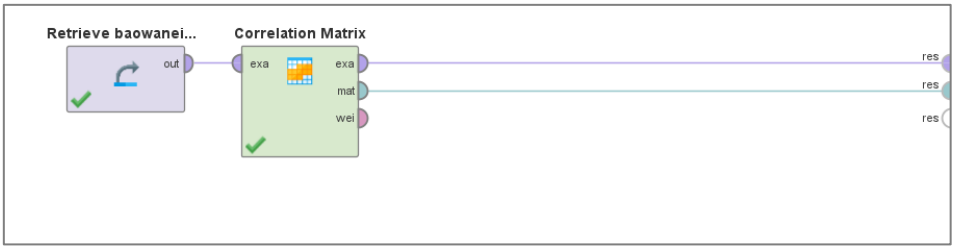
โหนดที่ 6: โหนดนี้ทำหน้าที่สมดุล โดยให้อายุสูง, BMI สูง และโรคหัวใจกดสัญญาณลง ขณะที่ HbA1c และกลูโคสดันสัญญาณขึ้น เป็น โหนดถ่วงสมดุลระหว่างปัจจัยประชากรกับตัวแปรชีวเคมี

โดยสรุป การทำงานร่วมกันของโหนดทั้งหกสะท้อนให้เห็นว่าโมเดลให้ความสำคัญอย่างยิ่งกับ ค่า HbA1c และกลูโคส ในฐานะตัวแปรหลัก ขณะเดียวกันยังผสานปัจจัยด้านร่างกายและพฤติกรรมเพื่อสร้างการจำแนกที่แม่นยำยิ่งขึ้น

ผลการวิเคราะห์ชั้นเอาต์พุตของโมเดล Improved NeuralNet พบว่า การจำแนกขึ้นกับสัญญาณจากโหนดซ่อนทั้งหก โดยคลาส “1” (เสี่ยงเบาหวาน) ได้แรงสนับสนุนหลักจากโหนดที่เน้น **HbA1c** และ **Blood Glucose** ส่วนคลาส “0” (ไม่เสี่ยง) ได้แรงหนุนจากโหนดที่ทำหน้าที่กดหรือถ่วงสมดุลสัญญาณเหล่านี้ โดยเฉพาะโหนดที่เกี่ยวข้องกับปัจจัยประชากรและโรคร่วม แสดงให้เห็นว่าโมเดลผสมผสานทั้งตัวแปรทางชีวเคมีและปัจจัยเสริมเพื่อเพิ่มความแม่นยำในการทำนาย

3.1.4.6 การวิเคราะห์สหสัมพันธ์ (Correlation Analysis)

เป็นการแสดงการสร้างกระบวนการในโปรแกรม RapidMiner โดยใช้โมเดล Correlation Matrix สำหรับคำนวณหาความสัมพันธ์เชิงเส้นระหว่างตัวแปรต่างๆ ภายในชุดข้อมูลที่เกี่ยวข้องกับโรคเบาหวาน โดยเริ่มจากการดึงข้อมูลด้วยโมดูล Retrieve จากนั้นจึงส่งข้อมูลเข้าสู่โมดูล Correlation Matrix เพื่อสร้างตารางค่าความสัมพันธ์ระหว่างตัวแปรทุกคู่ และส่งผลลัพธ์ออกมาในรูปแบบที่สามารถนำไปวิเคราะห์ต่อได้



ภาพที่ 3.37 ขั้นตอนการสร้างกระบวนการเพื่อวิเคราะห์ความสัมพันธ์

ผลลัพธ์ที่ได้เป็นตาราง Correlation Matrix ซึ่งประกอบด้วยตัวแปรสำคัญ เช่น smoking, gender, age, hypertension, heart_disease, bmi, hbA1c_level, blood_glucose, และ diabetes โดยค่าที่แสดงมีช่วงระหว่าง –1 ถึง 1 หมายถึงระดับความสัมพันธ์เชิงเส้นระหว่างตัวแปรแต่ละคู่ หากค่าใกล้ +1 หมายถึงมีความสัมพันธ์เชิงบวกสูง ค่าใกล้ –1 หมายถึงมีความสัมพันธ์เชิงลบสูง และค่าใกล้ 0 หมายถึงแทบไม่มีความสัมพันธ์เชิงเส้น ดังภาพที่ 3.38

Attributes	smokin...	gender	age	hyperte...	heart_di...	bmi	hbA1c_l...	blood_g...	diabetes
smoking_...	1	-0.003	0.266	0.093	0.066	0.199	0.041	0.048	0.106
gender	-0.003	1	0.029	-0.014	-0.079	0.024	-0.020	-0.018	-0.038
age	0.266	0.029	1	0.258	0.239	0.345	0.106	0.114	0.264
hypertens...	0.093	-0.014	0.258	1	0.120	0.148	0.082	0.085	0.197
heart_dis...	0.066	-0.079	0.239	0.120	1	0.061	0.068	0.071	0.172
bmi	0.199	0.024	0.345	0.148	0.061	1	0.084	0.093	0.216
hbA1c_le...	0.041	-0.020	0.106	0.082	0.068	0.084	1	0.172	0.407
blood_glu...	0.048	-0.018	0.114	0.085	0.071	0.093	0.172	1	0.425
diabetes	0.106	-0.038	0.264	0.197	0.172	0.216	0.407	0.425	1

ภาพที่ 3.38 ผลลัพธ์ที่แสดงตารางสหสัมพันธ์

การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) ถูกนำมาใช้เพื่อประเมินความเชื่อมโยงระหว่างตัวแปรต่างๆ ในชุดข้อมูลเบาหวาน โดยผู้วิจัยสร้างเมทริกซ์สหสัมพันธ์เพื่อวัดค่าความสัมพันธ์เชิงตัวเลขระหว่างแต่ละคู่ของตัวแปรอย่างเป็นระบบ ผลการวิเคราะห์แสดงให้เห็นว่า ตัวแปร ระดับ HbA1c, ระดับน้ำตาลกลูโคส และอายุ มีความสัมพันธ์กับการเกิดโรคเบาหวานในระดับที่ชัดเจนกว่าตัวแปรอื่น ขณะที่ปัจจัยอื่นๆ เช่น ความดันโลหิต โรคหัวใจ ดัชนีมวลกาย หรือประวัติการสูบบุหรี่ มีค่าสหสัมพันธ์ค่อนข้างต่ำ สะท้อนให้เห็นว่าไม่มีตัวแปรใดที่มีความสัมพันธ์สูงจนก่อให้เกิดปัญหา multicollinearity ภายในชุดข้อมูล ซึ่งช่วยยืนยันได้ว่าสามารถนำตัวแปรทั้งหมดไปใช้ร่วมกันในกระบวนการสร้างแบบจำลองได้อย่างเหมาะสม

อย่างไรก็ตาม การที่ตัวแปรหนึ่งมีค่าสหสัมพันธ์ต่ำ ไม่ได้หมายความว่าตัวแปรนั้นไม่มีความสำคัญต่อการทำนายโรคเบาหวาน แต่เป็นข้อบ่งชี้ว่าตัวแปรดังกล่าวอาจไม่เด่นชัดเมื่อพิจารณาเพียงลำพัง ทั้งนี้ในการสร้างแบบจำลองด้วยวิธีการเรียนรู้ของเครื่อง ตัวแปรที่มีความสัมพันธ์ต่ำก็ยังคงมีบทบาท เนื่องจากอาจมีอิทธิพลร่วมกับตัวแปรอื่นๆ หรือสะท้อนความสัมพันธ์ที่ไม่เป็นเชิงเส้น ซึ่งโมเดลสามารถเรียนรู้และนำไปใช้ปรับปรุงประสิทธิภาพในการพยากรณ์ได้ ดังนั้น ผลจากการวิเคราะห์สหสัมพันธ์จึงยืนยันได้ว่าตัวแปรทั้งหมดในชุดข้อมูลมี

คุณค่าและสามารถนำมาสร้างโมเดลทำนายได้ โดยไม่ก่อให้เกิดอคติหรือผลกระทบเชิงลบต่อสมรรถนะของโมเดล (HuiZhi Shao1,2, Xiang Liu2 , DaShuai Zong2 and QingJun Song, 2024)

3.1.5 การวัดประสิทธิภาพของโมเดล (Evaluation)

การทดสอบประสิทธิภาพของแบบจำลองได้ดำเนินการวัดค่าความแม่นยำ ค่าเรียกกลับ และค่าความถูกต้อง เพื่อสร้าง แบบจำลองการพยากรณ์จากการเรียนรู้และทดสอบแบบจำลองการพยากรณ์โดยค่าความถูกต้อง (Accuracy)

ตารางที่ 3.4 เปรียบเทียบผลลัพธ์ของโมเดล

Model	Accuracy	Precision	Recall
Decision Tree	90.69%	class0=**100%** class1=0**%**	class0=**90.69%** class1=0**%**
Naïve Bayes	89.16%	class0=**100%** class1=0**%**	class0=89**.16%** class1=0**%**
Random Forest	86.76%	class0=**100%** class1=0**%**	class0=86.76**%** class1=0**%**
SVM	77.96%	class0=0**%** class1=77.96**%**	class0=0**%** class1=100**%**
ANN	89.97%	class0=**100%** class1=0**%**	class0=89.97**%** class1=0**%**

ตารางที่ 3.4 แสดงการทดสอบของโมเดลการเรียนรู้ของเครื่อง (Machine Learning) 5 ตัว ได้แก่ Decision Tree, Naïve Bayes, Random Forest, SVM (Support Vector Machine) และ ANN (Artificial Neural Network) โดยการประเมินผลใช้ 3 เมตริกหลัก ได้แก่ Accuracy, Precision และ Recall สำหรับทั้งสองกลุ่ม class 0 (ไม่เป็นโรคเบาหวาน) และ class 1 (เป็นโรคเบาหวาน) ซึ่งสะท้อนถึงประสิทธิภาพของแต่ละโมเดลในการจำแนกประเภทผู้ป่วยโรคเบาหวาน

เริ่มต้นที่ Decision Tree ซึ่งมี Accuracy สูงถึง 90.69% โดยสามารถทำนาย class 0 (ไม่เป็นโรคเบาหวาน) ได้อย่างถูกต้อง 100% แต่ไม่สามารถทำนาย class 1 (เป็นโรคเบาหวาน) ได้เลย ซึ่งหมายความว่าโมเดลนี้ไม่สามารถตรวจจับผู้ที่เป็นโรคเบาหวานได้เลย แม้ว่าจะสามารถทำนายผู้ที่ไม่เป็นโรคเบาหวานได้ดี (Recall 90.69%) การที่โมเดลไม่สามารถทำนาย class 1 ได้ทำให้ Recall สำหรับ class 1 เป็น 0%

ในกรณีของ Naïve Bayes มี Accuracy ที่ 89.16% ซึ่งสามารถทำนาย class 0 ได้ถูกต้อง 100% แต่ไม่สามารถทำนาย class 1 ได้เลย แม้ว่าจะมี Recall สำหรับ class 1 ที่เท่ากับ Accuracy (89.16%) ซึ่งแสดงว่าโมเดลนี้สามารถตรวจจับผู้ป่วยโรคเบาหวานได้ดี แต่ไม่ได้ตรวจจับผู้ที่ไม่เป็นโรคเบาหวาน

Random Forest แสดงผลคล้ายกับ Decision Tree โดยมีการระบุ class 0 ได้อย่างถูกต้อง 100% (Precision 100%) แต่ไม่สามารถจำแนก class 1 ได้เลย ซึ่งทำให้ Precision และ Recall ของ class 1 เป็น 0%

ในส่วนของ SVM มี Accuracy ที่ต่ำที่สุดที่ 77.96% โดยสามารถทำนาย class 1 ได้ดี (Recall 100%) แต่ไม่สามารถทำนาย class 0 ได้เลย ซึ่งทำให้ SVM จึงเป็นโมเดลที่เหมาะสมในการทำนาย class 1 เท่านั้น แต่ไม่สามารถทำนาย class 0 ได้

สุดท้าย ANN มี Accuracy สูงที่ 89.97% และสามารถทำนาย class 0 ได้ 100% แต่ไม่สามารถทำนาย class 1 ได้เลย นอกจากนี้ Recall สำหรับ class 1 ก็สูงถึง 89.97% ซึ่งแสดงว่าโมเดลนี้สามารถตรวจจับผู้ป่วยโรคเบาหวานได้ดี แต่ไม่สามารถตรวจจับผู้ที่ไม่เป็นโรคเบาหวาน

การวิเคราะห์นี้แสดงให้เห็นว่าโมเดลหลายตัวมีปัญหาในการจัดการกับการจำแนกทั้งสองกลุ่มอย่างสมดุล โดยบางโมเดลสามารถระบุ class 0 ได้ดีแต่ไม่สามารถระบุ class 1 ได้เลย ขณะที่บางโมเดลสามารถระบุ class 1 ได้ดี แต่ล้มเหลวในการจำแนก class 0 ซึ่งสะท้อนถึงการไม่สมดุลในการแยกแยะทั้งสองกลุ่มและทำให้โมเดลไม่สามารถใช้งานได้ดีในสถานการณ์ที่มีการจำแนกทั้งสองกลุ่มอย่างเท่าเทียม อย่างไรก็ตาม โมเดล Decision Tree มีข้อดีที่เด่นชัดหลายประการ เนื่องจากมีโครงสร้างที่เข้าใจง่าย ตีความได้ชัดเจน สามารถแสดงกฎการตัดสินใจออกมาในรูปแบบ if–then ที่โปร่งใส และระบุความสำคัญของปัจจัยที่มีผลต่อการทำนายได้อย่างตรงไปตรงมา โดยไม่จำเป็นต้องพึ่งพาโมเดลที่ซับซ้อนหรือการตั้งค่าพารามิเตอร์ที่ยุ่งยาก เหมาะสำหรับการทำความเข้าใจและการอธิบายผลการทำนายให้ผู้ใช้

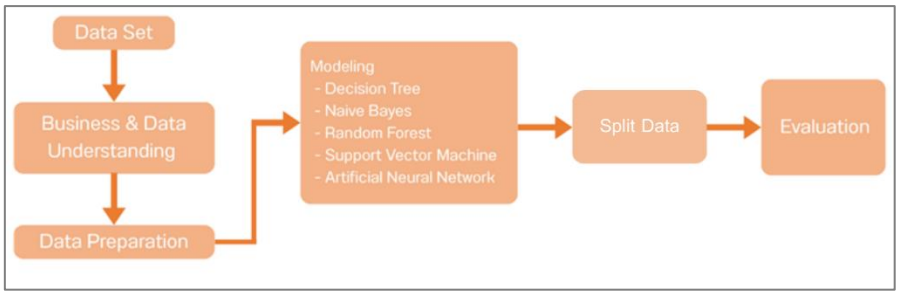
ที่ไม่เชี่ยวชาญทางเทคนิคเข้าใจได้ง่าย สามารถจัดการกับข้อมูลที่มีทั้งประเภทเชิงปริมาณและเชิงคุณภาพได้อย่างดี เหมาะกับงานที่ข้อมูลมีความหลากหลาย โดยไม่จำเป็นต้องแปลงข้อมูลให้เป็นรูปแบบเดียวกันเหมือนกับโมเดลอื่นๆ ที่ซับซ้อนกว่า นอกจากนี้ Decision Tree ยังสามารถจัดการกับข้อมูลที่มีลักษณะไม่สมดุลหรือมีค่า missing ได้ดี ซึ่งช่วยให้มันสามารถใช้งานได้ในงานที่มีข้อมูลไม่สมบูรณ์หรือมีความแตกต่างหลากหลาย

3.1.6 การนำโมเดลไปใช้งานจริง (Deployment)

นำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูล ได้ทำการนำเสนอข้อมูลแบบ Visualization ด้วยการแสดงแผนภาพสรุปข้อมูลและเผยแพร่ ข้อมูลสารสนเทศนี้บนเว็บบราวเซอร์โดยการใช้ภาษา HTML, CSS และ PHP ร่วมกับการ นำเสนอข้อมูลแบบ Visualization ด้วยการแสดงผลข้อมูลในรูปแบบของภาพโดยใช้โปรแกรม Power BI ซึ่งทางผู้วิเคราะห์ข้อมูลได้ยกตัวอย่างการจัดทำเป็นรูปแบบของรายงาน (Report) หรือแผนภาพ (Dashboard) ในการพัฒนาหน้าเว็บไซต์สำหรับการเปิดเผย และเผยแพร่ข้อมูลการวิเคราะห์ของชุดข้อมูลโรคเบาหวาน เพื่อให้ผู้ใช้งานทราบถึงเกี่ยวกับการวิเคราะห์โรคเบาหวานที่ได้ผลสรุปที่ถูกต้อง สะดวกรวดเร็วและเข้าใจง่ายต่อเวลาที่ต้องการศึกษาเพิ่มเติม

3.2 การออกแบบเว็บไซต์

ในการดำเนินโครงงานในครั้งนี้เรื่อง การวิเคราะห์ข้อมูลเพื่อทำนายลักษณะของผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด โดยวิธีการทำเหมืองข้อมูลโดยใช้โปรแกรม RapidMiner studio 9.5 ซึ่งมีกระบวนการตามมาตรฐานการทำเหมืองข้อมูลแบบ CRISP-DM ขั้นตอนการดำเนินงานวิจัยประกอบด้วยการทำความเข้าใจกับข้อมูลเพื่อให้รู้จักโครงสร้าง และรูปแบบของข้อมูลเมื่อเข้าใจข้อมูลและรู้รูปแบบ หรือปัญหาของข้อมูลแล้ว จะนำข้อมูลไปแปลงให้อยู่ในรูปแบบที่สามารถนำไปใช้สร้างตัวแบบทำนายได้พร้อมทั้งคัดเลือกเฉพาะข้อมูลที่สำคัญ และนำข้อมูลเข้าสู่โมเดล โดยจะมีการแบ่งข้อมูลออกเป็นข้อมูลที่ใช้ในการสอน และข้อมูลที่ใช้ทดสอบ เพื่อตรวจสอบประสิทธิภาพของโมเดล และประเมินผลโมเดล เมื่อเสร็จสิ้นจากกระบวนการตามมาตรฐานการทำเหมืองข้อมูลแบบ CRISP-DM แล้วจะเป็นขั้นตอนการออกแบบเว็บไซต์ และการออกแบบรูปแบบการแสดงผล (Dashboard) และบทสรุปจากวิธีดำเนินงานดังนี้



ภาพที่ 3.39 แผนภาพการทำงานของกระบวนการวิเคราะห์ข้อมูลและโมเดลเทคนิคเหมืองข้อมูล

จากภาพที่ 3.39 แสดงขั้นตอนการวิเคราะห์ข้อมูลเพื่อสร้างโมเดลการเรียนรู้ของเครื่อง (Machine Learning) โดยเริ่มจากกล่องด้านบนสุดคือ Data set ซึ่งเป็นแหล่งข้อมูลต้นทาง Kaggle จากนั้นเข้าสู่ขั้นตอน Business & Data Understanding เพื่อทำความเข้าใจข้อมูลและปัญหาที่ต้องการแก้ไข ต่อด้วยขั้นตอน Data Preparation ซึ่งเป็นการจัดการข้อมูลให้สะอาดและพร้อมใช้งาน เช่น การลบค่าที่ขาดหาย การแปลงค่าประเภทต่าง ๆ และการเลือกคุณลักษณะที่สำคัญในการวิเคราะห์ข้อมูลเพื่อทำนายลักษณะของผู้ป่วยโรคเบาหวานจากชุดข้อมูลทุติยภูมิแบบเปิด จากนั้นข้อมูลจะถูกส่งเข้าสู่ขั้นตอน Modeling ซึ่งรวมเทคนิคการสร้างโมเดลต่าง ๆ เช่น Decision Tree, Naive Bayes, Random Forest, Support Vector Machine และ Artificial Neural Network หลังจากได้โมเดลแล้ว จะทำการประเมินด้วยกระบวนการ Split Data ซึ่งเป็นขั้นตอนที่ใช้แบ่งชุดข้อมูลออกเป็นสองส่วน ได้แก่ชุดฝึก (Training set) สำหรับการเรียนรู้ของโมเดล และชุดทดสอบ (Test set) สำหรับตรวจสอบประสิทธิภาพของโมเดลกับข้อมูลที่ไม่เคยเห็นมาก่อน ซึ่งช่วยให้สามารถวัดความแม่นยำและความน่าเชื่อถือของผลการทำนายได้ สุดท้ายผลลัพธ์ทั้งหมดจะถูกสรุปในขั้นตอน Evaluation เพื่อวิเคราะห์ว่าโมเดลมีความแม่นยำและเหมาะสมต่อการใช้งานจริงเพียงใด โดยแต่ละขั้นตอนเชื่อมโยงกันด้วยลูกศรอย่างชัดเจน แสดงให้เห็นถึงลำดับการทำงานและความสัมพันธ์ของแต่ละขั้นตอนในกระบวนการวิเคราะห์ข้อมูล

ภาพที่ 3.40 หน้าแรกของเว็บไซต์ ให้ข้อมูลความรู้เกี่ยวกับโรคเบาหวาน

ภาพที่ 3.41 หน้าทดสอบความเสี่ยงโรคเบาหวาน

หน้าเว็บไซต์ที่ใช้สำหรับแบบทดสอบความเสี่ยงโรคเบาหวานนี้ โดยมีจุดประสงค์หลักเพื่อค้นหาคำแนะนำที่ช่วยในการจัดการและลดความเสี่ยงของการเกิดโรคเบาหวาน ซึ่งในหน้าเว็บไซต์นี้ ผู้ใช้จะต้องตอบคำถาม 4 ข้อ เช่น อายุ ดัชนีมวลกาย การออกกำลังกาย ประวัติคนในครอบครัวเป็นโรคเบาหวาน เป็นต้น เมื่อผู้ใช้กรอกข้อมูลครบถ้วนแล้ว จะสามารถกดปุ่ม "ส่งแบบทดสอบ" เพื่อรับผลการประเมินและคำแนะนำเพิ่มเติมได้

ภาพที่ 3.42 แดชบอร์ดแสดงข้อมูลที่เกี่ยวข้องกับโรคเบาหวานจากโปรแกรม Power BI

หน้าเว็บนี้เป็นหน้าแดชบอร์ดที่รวบรวมข้อมูลและตัวชี้วัดที่สำคัญไว้ในที่เดียว เพื่อช่วยให้ผู้ใช้เห็นภาพรวมและวิเคราะห์ข้อมูลต่างๆ ได้ง่ายขึ้น

ภาพที่ 3.43 คำแนะนำเกี่ยวกับโรคเบาหวาน

3.3 บทสรุป

การดำเนินงานในบทนี้เริ่มจากการนิยามปัญหาและวัตถุประสงค์เชิงธุรกิจสำหรับการทำนายลักษณะผู้ป่วยโรคเบาหวาน จากนั้นจึงเข้าสู่กระบวนการ CRISP-DM อย่างเป็นระบบตลอดจนการเข้ารหัสตัวแปรเชิงหมวดหมู่และการจัดการความไม่สมดุลของข้อมูลในชุดฝึก เพื่อให้ข้อมูลอยู่ในสภาพพร้อมสำหรับการสร้างแบบจำลองและการประเมินผล

ในขั้นตอนการสร้างแบบจำลอง ได้ทดสอบอัลกอริทึมแบบมีผู้สอน 5 เทคนิค ได้แก่ Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine (SVM) และ Artificial Neural Network (ANN) โดยวัดผลด้วย Accuracy, Precision, Recall และ F1-score บนชุดทดสอบ ผลพบว่า Decision Tree, Naïve Bayes, Random Forest และ ANN ให้ Accuracy ค่อนข้างสูง แต่ตรวจจับคลาส 1 ไม่ได้ สะท้อนผลกระทบของความไม่สมดุลของข้อมูล ขณะที่ SVM แม้ Accuracy ต่ำกว่า แต่สามารถตรวจจับคลาส 1 ได้ครบ และให้ F1-score ของคลาส 1 สูงที่สุด อย่างไรก็ดี เมื่อพิจารณามิติการตีความและอธิบายผล ซึ่งสำคัญต่อบริบทสุขภาพ โมเดล Decision Tree ยังมีความโดดเด่นจากความโปร่งใสของกฎ if-then และความสามารถในการชี้ตัวแปรสำคัญอย่างชัดเจน

สุดท้าย นำผลการวิเคราะห์ไปออกแบบสื่อสารในรูปแบบเว็บไซต์และแดชบอร์ด เพื่อให้ผู้ใช้ปลายทางเข้าถึงข้อมูลและผลทำนายได้สะดวก รวดเร็ว และเข้าใจง่าย บทนี้จึงสรุปได้ว่าการผสานกระบวนการเตรียมข้อมูลที่เหมาะสม เข้ากับการทดลองหลายโมเดลและการประเมินผลหลายตัวชี้วัด ช่วยให้เห็นทั้งศักยภาพและข้อจำกัดของแต่ละเทคนิค โดยเฉพาะอย่างยิ่งเหตุผลเชิงวิชาการในการเลือก Decision Tree เพื่อนำไปวิจัยต่อ ได้แก่ ความสามารถในการตีความกฎตัดสินใจ ความสอดคล้องกับเกณฑ์ทางคลินิกของตัวแปรสำคัญ (เช่น HbA1c และระดับน้ำตาลในเลือด) และความพร้อมต่อการนำไปใช้จริงร่วมกับเครื่องมือแสดงผลบนเว็บและแดชบอร์ดในสภาพแวดล้อมการใช้งานจริง