

บทที่ 4

กระบวนการพยากรณ์ผลผลิตสุกร

ในบทนี้ได้กล่าวถึงกระบวนการพยากรณ์ผลผลิตสุกร ซึ่งเป็นส่วนหนึ่งของเทคนิคการทำเหมืองข้อมูล ซึ่งจัดเป็นส่วนที่สำคัญที่สุดส่วนหนึ่งของระบบ โดยเนื้อหาของกระบวนการนี้สามารถแบ่งได้เป็นขั้นตอนได้ดังนี้

- (1) การนิยามปัญหาของการพยากรณ์
- (2) กระบวนการคัดเลือกข้อมูล และเตรียมข้อมูลแซมเปิ้ล
- (3) กระบวนการสร้างโมเดลการพยากรณ์
- (4) การทดสอบกระบวนการพยากรณ์ผลผลิตสุกร

4.1 การนิยามปัญหาของการพยากรณ์

ในกระบวนการผลิตสุกรนั้น การที่สุกรสามารถเจริญเติบโตได้ดีและมีประสิทธิภาพเพียงใดนั้นขึ้นอยู่กับอาหารที่สุกรได้รับซึ่งได้รับการวางแผนและควบคุมการให้อาหารจากผู้ประกอบการ อาหารจำนวนมากไม่ได้หมายความว่าสุกรจะเจริญเติบโตได้ดีหากแต่ขึ้นอยู่กับปริมาณพลังงานสารอาหารและคุณค่าทางโภชนาการที่สุกรได้รับอย่างเหมาะสมด้วย ในธุรกิจการผลิตสุกรนั้นผู้ประกอบการจะให้ความสนใจกับผลผลิตที่มาจากต้นทุนที่ไม่สูงมากนักด้วย ดังนั้นการผลิตสุกรให้ได้ผลผลิตที่มีคุณภาพนั้นจึงขึ้นอยู่กับความเหมาะสมและสมดุลกันระหว่างปริมาณอาหาร คุณค่าทางโภชนาการของอาหาร วิธีการให้อาหาร และราคาค้นทุนการผลิตด้วย ผลผลิตที่มีคุณภาพในที่นี้หมายถึง ผลผลิตสุกรที่ได้มีน้ำหนักที่ดี มีลักษณะการเจริญเติบโตที่ดี และในกระบวนการเลี้ยงสุกรมีการสูญเสียสุกรน้อยที่สุด ซึ่งแต่เดิมทีแล้วการที่จะบอกว่าสุกรกลุ่มใดมีคุณภาพมากน้อยเพียงใดนั้นยังไม่มีกรอบอธิบายเป็นข้อคุณลักษณะตายตัวที่ชัดเจน หากแต่เป็นเชิงการให้คะแนนของผู้ประกอบการเท่านั้นว่าสุกรกลุ่มใดมีคุณภาพเป็นเช่นใด ดังนั้นหากมีการอธิบายเป็นรูปแบบกฎที่ชัดเจนได้ว่าการเลี้ยงแบบใดก่อให้เกิดผลลัพธ์อย่างใดได้จะถือว่าเป็นสิ่งที่มีประโยชน์อย่างยิ่ง อีกปัญหาหนึ่งที่จะสามารถช่วยให้การควบคุมคุณภาพของกระบวนการผลิตดีขึ้นคือปกติแล้วการชั่งน้ำหนักและวัดค่าความกว้างช่วงหลังของสุกรนั้นผู้ประกอบการไม่สามารถจะทำการชั่งทุกวันได้จึงมีการสุ่มชั่งเป็นช่วงๆของการเลี้ยงไป ดังนั้นการมีโมเดลสมการตัวแทนที่สามารถบอกได้ว่าที่ผ่านมามีลักษณะธรรมชาติของข้อมูลการเลี้ยงสุกรโดยรวมเป็นเช่นใด เพื่อ

พยายามควบคุมคุณภาพไม่ให้ตกลงไปจากที่เป็นอยู่ และยังสามารถช่วยให้ผู้ประกอบการได้เห็นมุมมองในการวิเคราะห์เพื่อตัดสินใจเพิ่มมากขึ้นได้ด้วย

จากปัญหาดังกล่าวเหล่านี้เราสามารถแบ่งส่วนกระบวนการพยากรณ์ข้อมูลที่เหมาะสมกับระบบนี้ได้ 2 ส่วนคือ การใช้วิธีแผนผังต้นไม้การตัดสินใจซึ่งเป็นเทคนิคหนึ่งของการจำแนกข้อมูลเพื่อทำการสร้างโมเดลการจำแนกข้อมูลและสร้างกฎการจำแนกข้อมูลเพื่อใช้ในการอธิบายความสัมพันธ์ของกระบวนการผลิตได้ด้วย และส่วนที่สองคือการใช้วิธีรีเกรชันเชิงเส้นตรงในการสร้างสมการตัวแทนเพื่อเป็น โมเดลในการอธิบายธรรมชาติการเกิดขึ้นของข้อมูลจำพวกตัวเลขที่มีการเกิดขึ้นอย่างต่อเนื่องซึ่งมีแนวโน้มและทิศทางเกิดขึ้นของข้อมูลใกล้เคียงกัน โครงสร้างหลักของกระบวนการพยากรณ์ผลผลิตสุกรสามารถแสดงได้ดังรูปที่ 4.1



รูป 4.1 โครงสร้างหลักของกระบวนการพยากรณ์ผลผลิตสุกร

จากรูป 4.1 แสดงโครงสร้างหลักของกระบวนการพยากรณ์ผลผลิตสุกร ซึ่งเป็นส่วนหนึ่งของระบบวิเคราะห์และพยากรณ์ผลผลิตสุกร ด้วยเทคนิคการทำเหมืองข้อมูล โดยมีรายละเอียดดังนี้

4.1.1 ส่วนการจำแนกข้อมูลคุณภาพสุกร

ข้อมูลที่นำมาใช้ในการสร้างโมเดลการจำแนกจะเป็นข้อมูลการเลี้ยงสุกรที่มีการระบุคลาสเอาไว้แล้ว ซึ่งคลาสในที่นี้ได้แก่ “คุณภาพของสุกร” ซึ่งมีการระบุเอาไว้โดยความชำนาญและการตัดสินใจของผู้ประกอบการ คลาสคุณภาพสุกรสามารถแบ่งได้ 3 ระดับ หรือ 3 คลาสคือ “High” “Medium” และ “Low” โดยที่ตัวอย่างข้อมูลที่มีความเกี่ยวข้องในการนำมาใช้สร้างโมเดลการจำแนกสามารถแสดงได้ในตารางที่ 4.1 ดังนี้คือ

ตาราง 4.1 แสดงตัวอย่างรูปแบบข้อมูลที่นำมาใช้ในการสร้างโมเดลการจำแนก

รูปแบบการให้อาหาร	น้ำหนักที่เพิ่มขึ้นต่อวัน	ต้นทุนต่อน้ำหนักที่เพิ่มขึ้น	ค่าพลังงานที่ได้ต่อน้ำหนักที่เพิ่มขึ้น	%สุกรสูญเสีย	คุณภาพสุกร
A-01	0.51-0.60	21-30	12K-14K	<2	Medium
A-01	>0.60	31-40	10K-12K	4..6	Medium
B-01	<0.41	>50	<10K	>10	Low
B-02	>0.60	31-40	14K-16K	<2	High
B-01	0.51-0.60	<20	16K-18K	<2	High

จากตารางที่ 4.1 เป็นตัวอย่างรูปแบบข้อมูลแบบเปิดสำหรับนำมาใช้สร้างโมเดลการจำแนกด้วยวิธีแผนผังต้นไม้การตัดสินใจ ซึ่งแอททริบิวต์ “คุณภาพสุกร” ถือเป็นคลาสแอททริบิวต์ซึ่งเป็นตัวบอกลักษณะของสุกรจากเงื่อนไขตามแอททริบิวต์อื่นๆ โดยหลังจากสร้างโมเดลการจำแนกแล้วผลลัพธ์ที่เราจะได้จะอยู่ในรูปแบบกฎการจำแนกซึ่งสามารถใช้อธิบายความสัมพันธ์กันของข้อมูลและสามารถใช้ในการจำแนกข้อมูลใหม่ได้ด้วย

4.1.2 ส่วนการอธิบายธรรมชาติการเกิดข้อมูลตัวเลข

ข้อมูลตัวเลขที่ผู้ประกอบการสนใจซึ่งข้อมูลเหล่านี้มีความสัมพันธ์กันในการผลิตสุกรได้แก่

- (1) อายุวันที่ทำการเลี้ยงสุกร
- (2) ค่าต้นทุนการผลิตเฉลี่ยต่อสุกรสะสม
- (3) ค่าปริมาณอาหารเฉลี่ยต่อสุกรสะสม
- (4) ค่าพลังงานสารอาหารเฉลี่ยต่อสุกรสะสม
- (5) น้ำหนักเฉลี่ยสุกรตัวผู้
- (6) น้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรตัวผู้
- (7) ความกว้างช่วงหลังเฉลี่ยของสุกรตัวผู้
- (8) น้ำหนักเฉลี่ยสุกรตัวเมีย
- (9) น้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรตัวเมีย
- (10) ความกว้างช่วงหลังเฉลี่ยของสุกรตัวเมีย

ซึ่งข้อมูลเหล่านี้จะเป็นข้อมูลของการเลี้ยงสุกรแต่ละกลุ่มที่สนใจนำมาทำการสร้างโมเดล หลังจากมีการสร้างโมเดลการพยากรณ์ออกมาแล้ว เมื่อมีการระบุค่าใดค่าหนึ่งในปัจจัยด้านบนนี้ ซึ่งถือเป็นตัวแปรตั้งต้นแล้ว ระบบการพยากรณ์สามารถคำนวณค่าปัจจัยอื่นๆที่เกิดขึ้นเมื่อค่าตั้งต้นมีค่า

ตามระบุ ออกมาได้ เช่น เมื่อผู้ใช้กรอกค่าวันเป็น 30 หมายความว่าจากข้อมูลเทรนนิ่งที่นำมาสร้างโมเดลการพยากรณ์ด้วยวิธีรีเกรชันเส้นตรงนั้นสามารถทำนายได้ว่า ในวันที่ 30 ของการเลี้ยงสุกรนั้น ข้อมูลปัจจัยอื่นๆมีผลลัพธ์เป็นเช่นใด

4.2 กระบวนการคัดเลือกข้อมูล และเตรียมข้อมูลแชนแนล

ข้อมูลที่ใช้ในระบบวิเคราะห์และพยากรณ์ผลผลิตสุกร ด้วยเทคนิคการทำเหมืองข้อมูลนั้นมีจำนวนมากและค่อนข้างมีความซับซ้อนดังนั้นการคัดเลือกข้อมูลให้เหมาะสมสำหรับใช้ในการนำมาพยากรณ์จึงจัดว่ามีความสำคัญ อีกทั้งข้อมูลสำหรับการนำมาพยากรณ์นั้นส่วนใหญ่เป็นข้อมูลตัวเลข ดังนั้นกระบวนการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์จึงถูกนำมาใช้ในขั้นตอนการเตรียมข้อมูลด้วย โดยมีรายละเอียดดังนี้

4.2.1 กระบวนการคัดเลือกข้อมูล

จากบทที่ 3 ผู้เขียนได้อธิบายเกี่ยวกับระบบข้อมูลในส่วนที่มีความจำเป็นกับธุรกิจการผลิตสุกร ซึ่งข้อมูลได้ถูกจัดเก็บเป็น 2 ส่วนตามมุมมองคือ ข้อมูลการจัดการอาหารสุกร และข้อมูลการจัดการสุกร โดยทั้งสองส่วนนั้นมีการจัดเก็บแบบทราจเช็คขึ้นเพื่อประโยชน์ในการสืบย้อนข้อมูลในอดีต คือเมื่อมีการกระทำเหตุการณ์ใดกับอาหารหรือสุกรนั้นก็ให้มีการบันทึกรายละเอียดเข้าไปทั้งหมด การออกแบบโครงสร้างตารางสำหรับการเก็บข้อมูลนั้นจึงค่อนข้างมีจำนวนมากและซับซ้อนซึ่งยังไม่เหมาะกับการนำมาใช้ในกระบวนการพยากรณ์ ข้อมูลเกือบทั้งหมดมีความจำเป็นกับระบบธุรกิจและการวิเคราะห์ข้อมูล แต่สำหรับการนำมาพยากรณ์นั้นเราจะทำการคัดเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับการพยากรณ์มาใช้เท่านั้น ซึ่งในการดึงเอาข้อมูลมาใช้วิเคราะห์นั้น เราจึงจำเป็นต้องสร้างตารางชั่วคราวเพิ่มขึ้นมาจากตารางเดิมที่มีอยู่ สำหรับใช้ในการเก็บรวบรวมข้อมูลทั้งหมดที่เหมาะสมจะใช้เป็นแชนแนล และเนื่องจากบริบทการพยากรณ์ข้อมูลถูกแบ่งเป็น 2 ส่วนคือการจำแนกข้อมูลคุณภาพสุกร และการทำนายค่าข้อมูลต่างๆในกระบวนการผลิต ถึงแม้ข้อมูลที่จะนำมาใช้ในการวิเคราะห์จะมีบริบทที่ใกล้เคียงกัน แต่รูปแบบข้อมูลที่ต้องใช้ในกระบวนการก็มีรูปแบบที่ต่างกันตามวิธีการที่ใช้ในการพยากรณ์ โดยที่ข้อมูลที่จะนำมาใช้ในกระบวนการจำแนกด้วยวิธีแผนผังต้นไม้การตัดสินใจนั้นจำเป็นต้องปรับให้อยู่ในลักษณะค่าคงที่และมีขอบเขตเสียก่อน ในขณะที่ข้อมูลที่จะนำมาใช้ในกระบวนการทำนายข้อมูลด้วยวิธีรีเกรชันนั้นจำเป็นต้องใช้ข้อมูลที่มีค่าเป็นตัวเลขต่อเนื่องเพื่อใช้ในการคำนวณค่าที่ใกล้เคียงความเป็นจริงมากที่สุด

ดังนั้นตารางข้อมูลชั่วคราวจะถูกสร้างขึ้นมา 2 ตาราง คือ ตาราง TmpQualityClassification ซึ่งเป็นตารางชั่วคราวสำหรับรวบรวมข้อมูลที่เกี่ยวข้องและใช้ในกระบวนการจำแนกข้อมูลคุณภาพ

สุกรด้วยวิธีแผนผังต้นไม้การตัดสินใจ และตาราง TmpProductPrediction ซึ่งใช้ในการทำนายค่าข้อมูลผลผลิตสุกรด้วยวิธีรีเกรซันเชิงเส้นตรง ซึ่งสามารถอธิบายรายละเอียดได้ดังนี้

(1) ตาราง TmpQualityClassification เป็นตารางที่ใช้ในกระบวนการจำแนกข้อมูลคุณภาพสุกร ข้อมูลทั้งหมดที่ถูกเลือกเป็นแซมเปิ้ลจะถูกนำมารวบรวมไว้ที่ตารางนี้ และเมื่อผู้ใช้ทำการนิยามคุณภาพให้กลุ่มข้อมูลตัวอย่างในตารางนี้แล้ว ข้อมูลคุณภาพที่ได้รับการนิยามจะถูกจัดเก็บไว้ในแอตทริบิวต์ “Quality” รายละเอียดตาราง TmpQualityClassification สามารถแสดงได้ดังตารางที่ 4.2 ดังนี้

ตาราง 4.2 แสดงรายละเอียดของตาราง TmpQualityClassification

ลำดับที่	ชื่อฟิลด์	ชนิดและขนาด	คำอธิบาย	ตัวอย่างข้อมูล
1	GroupID	nvarchar(10)	รหัสกลุ่มสุกร	F01-040844
2	PatternFeed	nvarchar(4)	รหัสรูปแบบการให้อาหาร	A-01
3	AverageWeightGainPerDay	nvarchar(10)	ช่วงค่าน้ำหนักสุกรที่เพิ่มขึ้นเฉลี่ยต่อวัน	0.51-0.60
4	AverageFeedCostPerAverageWeightGain	nvarchar(10)	ช่วงต้นทุนอาหารที่ใช้ต่อน้ำหนักสุกรที่เพิ่มขึ้น	21-30
5	AverageEnergyPerAverageWeightGain	nvarchar(10)	ช่วงพลังงานสารอาหารที่สุกรได้รับต่อน้ำหนักสุกรที่เพิ่มขึ้น	13K-14K
6	PercentSwineLoss	nvarchar(10)	ช่วงค่าจำนวนสุกรที่สูญเสีย	0-3
7	Quality	nvarchar(10)	คลาสที่ระบุคุณภาพของสุกร	Medium

(2) ตาราง TmpProductPrediction เป็นตารางที่ใช้ในกระบวนการทำนายค่าข้อมูลผลผลิตสุกร ข้อมูลทั้งหมดที่ถูกเลือกเป็นแซมเปิ้ลจะถูกนำมารวบรวมไว้ที่ตารางนี้ รายละเอียดตาราง TmpProductPrediction สามารถแสดงได้ดังตารางที่ 4.3 ดังนี้

ตาราง 4.3 แสดงรายละเอียดของตาราง TmpProductPrediction

ลำดับ ที่	ชื่อฟิลด์	ชนิดและ ขนาด	คำอธิบาย	ตัวอย่างข้อมูล
1	GroupID	nvarchar(10)	รหัสกลุ่มสุกร	F01-040844
2	Age	int	อายุวันที่ทำการเลี้ยงสุกร	81
3	Cumulative AverageCost	float	ค่าต้นทุนการผลิตเฉลี่ยต่อ สุกรสะสม	767.11
4	Cumulative AverageFeed	float	ค่าปริมาณอาหารเฉลี่ยต่อ สุกรสะสม	103.97
5	Cumulative AverageEnergy	float	พลังงานสารอาหารที่ได้รับ สะสม	310,662.21
6	AverageWeight Male	float	น้ำหนักเฉลี่ยสุกรตัวผู้	43.57
7	AverageWeight GainMale	float	น้ำหนักเฉลี่ยที่เพิ่มขึ้นของ สุกรตัวผู้	25.18
8	AverageBackfat Male	float	ความกว้างช่วงหลังสุกรเฉลี่ย ของสุกรตัวผู้	7.42
9	AverageWeight Female	float	น้ำหนักเฉลี่ยสุกรตัวเมีย	44.62
10	AverageWeight GainFemale	float	น้ำหนักเฉลี่ยที่เพิ่มขึ้นของ สุกรตัวเมีย	26.17
11	AverageBackfat Female	float	ความกว้างช่วงหลังสุกรเฉลี่ย ของสุกรตัวเมีย	8.18

4.2.2 กระบวนการเตรียมข้อมูลแซมเปิ้ล

เนื่องจากข้อมูลที่จะนำมาใช้พยากรณ์นั้นเป็นข้อมูลจากฐานข้อมูลซึ่งใช้ในจัดเก็บข้อมูลประวัติทั้งหมดของกระบวนการให้อาหารและการเลี้ยงสุกร ซึ่งมีจำนวนมากและถูกจัดเก็บในลักษณะและรูปแบบที่ค่อนข้างมีการกระจัดกระจายทำให้เกิดความแตกต่างกันหลายรูปแบบ และระหว่างจัดเก็บข้อมูลอาจเกิดความผิดพลาดจากผู้กรอก หรือความผิดปกติของข้อมูล หรือความบกพร่องของเครื่องมือต่างๆได้ ดังนั้นกระบวนการในการจัดเตรียมข้อมูล หรือเราเรียกว่า

กระบวนการ เคต้าพรีโพรเซสซิง เพื่อจัดเตรียมเป็นข้อมูลที่จะนำไปใช้การเป็นข้อมูลเข้าสำหรับการทำเหมืองข้อมูลซึ่งในที่นี้คือการพยากรณ์ข้อมูล กระบวนการนี้จึงมีความจำเป็นและไม่ควรที่จะมองข้ามอย่างยิ่ง เพื่อความถูกต้องของการวิเคราะห์ข้อมูล ซึ่งขั้นตอนในกระบวนการเตรียมข้อมูลมีรายละเอียดดังนี้

(1) เคต้าคลีนนิ่ง

ในส่วนข้อมูลที่ขาดหายไป ข้อมูลที่ไม่น่าเชื่อถือ หรือข้อมูลไม่เข้าพวก สามารถตรวจสอบหรือคัดเลือกข้อมูลได้โดยวิธีการคิวรี (Query) ข้อมูลด้วยภาษา SQL ในส่วนนี้พบว่าข้อมูลบางจุดมีความผิดพลาดและอาจก่อให้เกิดผลผิดพลาดในกระบวนการพยากรณ์ โดยการแก้ปัญหาที่จำเป็นต้องแก้เป็นจุดๆ ไปแล้วแต่บริบทของปัญหาซึ่งสามารถอธิบายรายละเอียดได้ดังนี้

- ข้อมูลการให้อาหารสุกรบางส่วนมีลักษณะขาดหายไปซึ่งอาจเกิดจากการผิดพลาดในการเก็บข้อมูล หรือการกรอกข้อมูลทำให้ในข้อมูลที่ตั้งมามีลักษณะเป็นค่า NULL ยังผลให้การคำนวณค่าบางอย่างเกิดปัญหา เช่น การหาค่าอาหารสะสม ต้นทุนสะสม และค่าพลังงานสะสมเป็นต้น การแก้ปัญหาในส่วนนี้ใช้วิธีการหาค่าเฉลี่ยของข้อมูลในแอททริบิวท์ที่ข้อมูลหายไป โดยค่าข้อมูลที่น่ามาใช้หาค่าเฉลี่ยจะเลือกนำเอาข้อมูลที่มีค่าอายุวันที่เท่ากับข้อมูลที่หายไปมาใช้ในการคำนวณ ซึ่งค่าที่ได้ก็จะมีลักษณะใกล้เคียงกันกับข้อมูลอื่นๆ ที่อยู่ในวันที่เดียวกัน
- ข้อมูลการชั่งน้ำหนักบางกลุ่มสุกรมีไม่ครบซึ่งอาจเกิดจากกระบวนการเก็บข้อมูล ทำให้การคำนวณข้อมูลที่เกี่ยวข้องกับน้ำหนักซึ่งเป็นส่วนที่สำคัญมากกับบริบทผลผลิตมีค่าผิดไปจากความเป็นจริงในส่วนนี้เราจะตัดข้อมูลที่เป็นของกลุ่มสุกรดังกล่าวออก ไม่นำมาใช้ในการพยากรณ์เนื่องจากข้อมูลดังกล่าวถือว่าเป็นข้อมูลที่ไม่มีความน่าเชื่อถือที่จะนำมาใช้พยากรณ์
- ข้อมูลการสูญเสียสุกรบางกลุ่มสุกร มีค่าพุ่งสูงเกินปกติ ข้อมูลส่วนนี้ความเป็นจริงแล้วเกิดไม่บ่อยนัก หากเกิดก็อาจมีผลมาจากปัจจัยอื่นภายนอกเช่น โรคติดต่อ หรืออาหารเป็นพิษ เป็นต้น ดังนั้นข้อมูลกลุ่มสุกรที่เกิดเหตุการณ์อย่างนี้ถึงแม้ว่าอาจเกิดขึ้นจริงแต่ถือว่าเป็นข้อมูลที่ผิดธรรมชาติและมีปัจจัยภายนอกเข้ามาทำให้เกิดความผิดปกติ เราจะตัดข้อมูลกลุ่มสุกรนี้ออกทั้งกลุ่มไม่นำมาใช้ในการพยากรณ์

(2) เคต้าอินทิเกรชัน

ในส่วนของการรวมข้อมูลจากหลายแหล่งเข้ามาเก็บรวมกันเพื่อให้ต่อเนื่องและมีความสัมพันธ์กันนั้น ผู้พัฒนาระบบไม่ได้มีการจัดทำในส่วนนี้ เนื่องจากข้อมูลที่น่ามาใช้ในการวิเคราะห์มีเพียง 1 ฐานข้อมูลเท่านั้น

(3) เคตารีคักชัน

เนื่องจากข้อมูลดิบที่มีการจัดเก็บไว้ในฐานข้อมูลของระบบการจัดการอาหารและสุกรนั้น การเก็บข้อมูลจะเก็บแบบทรานแซคชัน ข้อมูลจะเก็บทุกรายละเอียดการกระทำ ดังนั้นข้อมูลจะมีปริมาณมากและซับซ้อน อยู่ในรูปแบบที่ยังไม่เหมาะจะนำไปใช้ในกระบวนการพยากรณ์ กระบวนการเคตารีคักชันซึ่งจะเป็นการลดปริมาณข้อมูลลง จะมีหน้าที่กวีรีเลือกเอาข้อมูลเฉพาะส่วนที่จำเป็นกับการวิเคราะห์และนำมาทำการสรุปรวมข้อมูลตามมุมมองหรือการทำการสรุปปิ้ง เพื่อนำมาใส่ในตารางข้อมูลชั่วคราวสำหรับใช้ในกระบวนการพยากรณ์ ซึ่งส่วนใหญ่แล้วกระบวนการนี้จะเกิดขึ้นกับ ตาราง TmpQualityClassification มากที่สุด ซึ่งแอททริบิวท์ที่เกิดจากการสรุปรวมข้อมูลและคำนวณออกมาเพื่อให้อยู่ในรูปที่เหมาะสมกับการนำไปใช้ ได้แก่

- ค่าน้ำหนักสุกรที่เพิ่มขึ้นเฉลี่ยต่อวัน
- ต้นทุนอาหารที่ใช้ต่อน้ำหนักสุกรที่เพิ่มขึ้น
- พลังงานสารอาหารที่สุกรได้รับต่อน้ำหนักสุกรที่เพิ่มขึ้น
- ค่าจำนวนสุกรที่สูญเสีย

ซึ่งเดิมทีข้อมูลเหล่านี้อยู่ในรูปแบบรายละเอียดของทรานแซคชัน ซึ่งมีโครงสร้างการเก็บข้อมูลที่กระจายตารางกันออกไปเพื่อประโยชน์ในการเก็บรายละเอียดให้ได้มากที่สุด

(4) เคต่าทรานสฟอร์มชัน

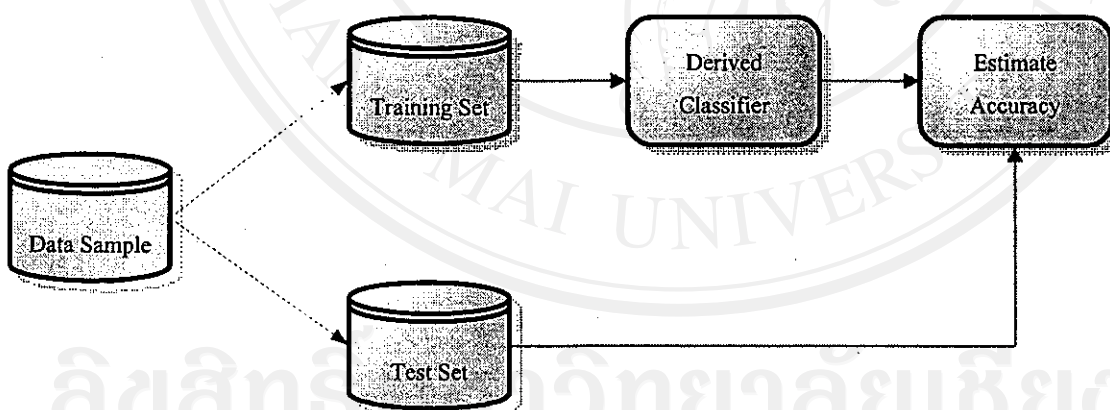
ในการเปลี่ยนแปลงรูปแบบของข้อมูลให้อยู่ในรูปแบบที่สามารถทำการวิเคราะห์ได้นั้น สำหรับกระบวนการพยากรณ์ข้อมูลแล้วถือได้ว่ามีความสำคัญมาก เนื่องจากการพยากรณ์ข้อมูลด้วยทฤษฎีต่างๆ นั้น จะมีรูปแบบของข้อมูลเข้า (Input Data) ที่แตกต่างกันไป สำหรับกระบวนการพยากรณ์ผลผลิตสุกรนั้น ส่วนที่มีความจำเป็นต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการใช้งานนั้นคือส่วนกระบวนการจำแนกข้อมูลด้วยวิธีแผนผังต้นไม้การตัดสินใจ ซึ่งเหมาะกับข้อมูลประเภทมีค่าคงที่และขอบเขตจำกัดมากกว่าที่จะเป็นข้อมูลตัวเลข ซึ่งการแปลงข้อมูลในส่วนดังกล่าวมีรายละเอียดดังนี้

- ช่วงค่าน้ำหนักสุกรที่เพิ่มขึ้นเฉลี่ยต่อวัน ซึ่งจากเดิมจะมีค่าเป็นค่าตัวเลข เช่น 0.54 กิโลกรัม หรือ 0.29 กิโลกรัม เป็นต้น ในส่วนข้อมูลนี้เราจะแปลงให้อยู่ในรูปแบบของช่วงข้อมูลทั้งหมด 4 ช่วงคือ “<0.41” “0.41-0.50” “0.51-0.60” และ “>0.60” ซึ่งช่วงข้อมูลดังกล่าวถือเป็นข้อมูลประเภทค่าคงที่และมีขอบเขตชัดเจน
- ช่วงต้นทุนอาหารที่ใช้ต่อน้ำหนักสุกรที่เพิ่มขึ้น ข้อมูลเดิมจะมีค่าตัวเลข เช่น 20.95 บาท หรือ 34.25 บาท เป็นต้น แปลงให้อยู่ในรูปแบบของช่วงข้อมูลทั้งหมด 5 ช่วงข้อมูลคือ “<21” “21-30” “31-40” “41-50” และ “>50”

- ช่วงพลังงานสารอาหารที่สุกรได้รับต่อน้ำหนักสุกรที่เพิ่มขึ้น ข้อมูลเดิมจะมีค่าตัวเลข เช่น 11,561.15 กิโลแคลลอรี่ หรือ 52,439.08 กิโลแคลลอรี่เป็นต้น แปลงให้อยู่ในรูปแบบของช่วงข้อมูลทั้งหมด 6 ช่วงคือ “<10K” “10K-12K” “12K-14K” “14K-16K” “16K-18K” และ “>18K”
- ช่วงค่าจำนวนสุกรที่สูญเสีย ข้อมูลเดิมจะมีค่าตัวเลข เช่น 6.26 % หรือ 2.54 % เป็นต้น แปลงให้อยู่ในรูปแบบของช่วงข้อมูลทั้งหมด 6 ช่วงคือ “<2” “2-4” “4-6” “6-8” “8-10” และ “>10”

4.3 กระบวนการสร้างโมเดลการพยากรณ์

หลังจากเราเตรียมข้อมูลสำหรับเป็นข้อมูลแซมเปิ้ลเสร็จเรียบร้อยแล้ว กระบวนการต่อมาคือการเตรียมสร้างโมเดลการพยากรณ์ ในกระบวนการสร้างโมเดลพยากรณ์นั้นเราจำเป็นต้องแบ่งข้อมูลแซมเปิ้ลเป็น 2 ส่วน ส่วนแรกเราจะเรียกว่าข้อมูลเทรนนิ่ง ซึ่งข้อมูลส่วนนี้จะใช้เป็นข้อมูลในการสร้างโมเดล ข้อมูลส่วนที่สองซึ่งเป็นข้อมูลที่เหลือจากข้อมูลส่วนแรก เราเรียกว่าข้อมูลทดสอบ เราจะใช้ข้อมูลส่วนนี้เป็นข้อมูลในการตรวจสอบความถูกต้องแม่นยำของโมเดลการพยากรณ์ โครงสร้างการแบ่งข้อมูลแซมเปิ้ลแสดงได้ในรูปที่ 4.2



รูป 4.2 โครงสร้างการแบ่งข้อมูลแซมเปิ้ล

จากรูปที่ 4.2 ข้อมูลส่วนเทรนนิ่งจะถูกนำไปสร้างโมเดลการจำแนกหรือพยากรณ์ และใช้ข้อมูลทดสอบเป็นตัวทดสอบความแม่นยำ ในการแบ่งข้อมูลทั้งสองส่วนนั้นไม่มีหลักเกณฑ์ตายตัวว่าต้องแบ่งกันด้วยสัดส่วนเท่าไร ขึ้นอยู่กับผู้ออกแบบและบริบทของการพยากรณ์

ในส่วนของการสร้างโมเดลการพยากรณ์ของระบบวิเคราะห์และพยากรณ์ผลผลิตสุกร ด้วยเทคนิคการทำเหมืองข้อมูลนั้น เราแบ่งการสร้างโมเดลพยากรณ์ออกเป็น 2 ส่วนดังนี้

4.3.1 โมเดลการจำแนกข้อมูลคุณภาพสุกร ด้วยวิธีแผนผังต้นไม้การตัดสินใจ

ข้อดีของการจำแนกข้อมูลด้วยวิธีแผนผังต้นไม้การตัดสินใจนั้น คือเป็นวิธีที่ค่อนข้างประหยัดเวลาในการคำนวณ และยังสามารถแปลงผลลัพธ์ที่ได้ออกมาอยู่ในรูปแบบของกฎการจำแนกซึ่งง่ายต่อการทำความเข้าใจ ผู้ใช้ระบบสามารถทำความเข้าใจกับกฎดังกล่าวได้ทันที ในบทที่ 2 ผู้จัดทำได้อธิบายวิธีการและอัลกอริทึมของการสร้างแผนผังต้นไม้การตัดสินใจแล้ว โดยกำหนดให้มีข้อมูลรับเข้าเป็นกลุ่มข้อมูลสุกรที่ได้เตรียมไว้แล้วจากตาราง TmpQualityClassification ซึ่งแต่ละแอททริบิวต์ได้มีการเตรียมข้อมูลโดยการแปลงข้อมูลเชิงปริมาณให้อยู่ในรูปแบบค่าคงที่ซึ่งมีขอบเขตชัดเจนไว้แล้ว เมื่อผ่านการสร้างแผนผังต้นไม้การตัดสินใจตามวิธีการในอัลกอริทึมแล้ว ผลลัพธ์ที่ได้จะเป็นโครงสร้างแผนผังต้นไม้การตัดสินใจ ซึ่งโครงสร้างดังกล่าวจะถือเป็นโมเดลการจำแนกข้อมูลคุณภาพสุกร โดยในการนำไปพัฒนานั้นผู้พัฒนาได้ทำการแปลงให้อยู่ในรูปแบบของกฎการจำแนก เพื่อการง่ายต่อการแสดงผลลัพธ์และการทำความเข้าใจ ตัวอย่างเช่น

IF(PercentSwineLoss= “2-4”)AND(AverageFeedCostPerAverageWeightGain= “41-50”)

THEN(Quality= “Low”)

โครงสร้างเหล่านี้จะสามารถนำไปใช้ในการพยากรณ์ข้อมูลสุกรกลุ่มใหม่ได้ว่าจัดอยู่ในกลุ่มคุณภาพแบบใด อีกทั้งผู้ประกอบการสามารถนำกฎการจำแนกเหล่านี้ไปใช้เป็นเครื่องมือช่วยในการวิเคราะห์ธรรมชาติการเกิดข้อมูลการเลี้ยงสุกรของตนได้อีกด้วย

4.3.2 โมเดลการทำนายข้อมูล ด้วยวิธีรีเกรชันเชิงเส้นตรง

เป้าหมายของวิธีการรีเกรชันแบบเส้นตรงคือการพยายามสร้างสมการเส้นตรงเพื่อเป็นโมเดลในการทำนายค่าข้อมูล Y ซึ่งเราจะเรียกค่าใดๆในกลุ่มข้อมูล Y ซึ่งเป็นผลจากการทำนายว่าตัวแปรตาม หรือค่าผล และค่าใดๆในกลุ่มข้อมูล X ซึ่งเป็นค่าที่เป็นปัจจัยในการทำนายว่า ตัวแปรต้น หรือตัวแปรการทำนาย โดยมีผลลัพธ์ของโมเดลการทำนายอยู่ในรูปของสมการ $Y = \alpha + \beta X$ ซึ่งถือเป็นสมการตัวแทนของข้อมูลทั้งหมดที่เป็นข้อมูลเทรนนิ่ง ซึ่งใช้เป็นสมการในการทำนายค่าบางค่าที่เราต้องการจะรู้

ดังนั้นในบริบทของการทำนายค่าข้อมูลในกระบวนการผลิตสุกรนั้น เรามีปัจจัยในการวิเคราะห์ผลผลิตทั้งหมด 10 ปัจจัย โดยเมื่อกำหนดให้ปัจจัยใดปัจจัยหนึ่งเป็น ตัวแปรต้นแล้ว เราสามารถคำนวณหาอีก 9 ปัจจัยที่เหลือได้ทันที เนื่องจากปัจจัยทั้ง 10 ต่างมีความสัมพันธ์ซึ่งกันและกันอยู่ เช่น หลังจากที่เราได้โมเดลการทำนายมาแล้ว เมื่อเรากำหนดให้ตัวแปรต้นคือ วันที่ ซึ่งมีค่าเป็น 35 นั้นหมายความว่า ระบบสามารถคำนวณหาตัวปัจจัยอีก 9 ตัวที่เหลือออกมาได้ว่า ในวันที่ 35 นั้น ค่าต่างๆมีค่าเท่าใด ซึ่งรายละเอียดของปัจจัยข้อมูลที่มีความสัมพันธ์กัน ซึ่งนำไปใช้ในการคำนวณหาตัวอื่นๆสามารถอธิบายได้ดังตาราง 4.4 เมื่อกำหนดค่าต่อไปนี้แทนปัจจัยต่างๆ

- (1) A-1 หมายถึง อายุวันที่ทำการเลี้ยงสุกร
- (2) A-2 หมายถึง ค่าต้นทุนการผลิตเฉลี่ยต่อสุกรสะสม
- (3) A-3 หมายถึง ค่าปริมาณอาหารเฉลี่ยต่อสุกรสะสม
- (4) A-4 หมายถึง ค่าพลังงานสารอาหารเฉลี่ยต่อสุกรสะสม
- (5) A-5 หมายถึง น้ำหนักเฉลี่ยสุกรตัวผู้
- (6) A-6 หมายถึง น้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรตัวผู้
- (7) A-7 หมายถึง ความกว้างช่วงหลังสุกรเฉลี่ยของสุกรตัวผู้
- (8) A-8 หมายถึง น้ำหนักเฉลี่ยสุกรตัวเมีย
- (9) A-9 หมายถึง น้ำหนักเฉลี่ยที่เพิ่มขึ้นของสุกรตัวเมีย
- (10) A-10 หมายถึง ความกว้างช่วงหลังสุกรเฉลี่ยของสุกรตัวเมีย

ตาราง 4.4 แสดงรายละเอียดของปัจจัยที่มีความสัมพันธ์กัน

ลำดับที่	X	Y
1	A-1	A-2
2	A-1	A-3
3	A-1	A-4
4	A-4	A-5
5	A-4	A-6
6	A-4	A-8
7	A-4	A-9
8	A-5	A-7
9	A-8	A-10

ค่าตัวแปรตามที่ได้ อาจถูกนำไปใช้เป็นตัวแปรต้นเพื่อหาค่าอื่นๆต่อไปได้ ตัวอย่างเช่น เมื่อตัวแปรต้นเป็น อายุวัน เมื่อหาค่าตัวแปรตามที่ได้คือค่าพลังงานสะสมได้แล้ว ค่าพลังงานสะสมจึงถูกนำไปใช้เป็นตัวแปรต้นสำหรับหาค่าน้ำหนักสุกรต่อไปเป็นต้น

4.4 การทดลองกระบวนการพยากรณ์ผลผลิตสุกร

หลังจากกระบวนการพัฒนาระบบพยากรณ์ผลผลิตสุกรเสร็จสิ้นลง ผู้จัดทำได้ทำการทดลองกระบวนการพยากรณ์ผลผลิตสุกรที่ได้ทำการพัฒนาขึ้นในลักษณะต่างๆ โดยใช้ข้อมูลตัวอย่างกลุ่มเดียวกัน เพื่อสรุปผลการทดลองและอภิปรายผลที่เกิดขึ้นจากกระบวนการ

4.4.1 การออกแบบการทดลอง

ในการทดลองระบบพยากรณ์ผลผลิตสุกร ผู้จัดทำได้เลือกทำการทดลองเฉพาะส่วนการจำแนกข้อมูลคุณภาพสุกรด้วยวิธีแผนผังต้นไม้การตัดสินใจ ซึ่งสิ่งที่สำคัญที่สุดของกระบวนการคือค่าความแม่นยำของการพยากรณ์ว่าความถูกต้องเมื่อนำข้อมูลทดสอบมาทดสอบกับโมเดลการพยากรณ์แล้วมีเปอร์เซ็นต์ความถูกต้องเป็นเช่นใด

จากขั้นตอนการสร้างโมเดลการพยากรณ์นั้นข้อมูลแซมเปิลได้ถูกแบ่งเป็น 2 ส่วนคือ ส่วนที่ใช้สำหรับเป็นข้อมูลเทรนนิ่ง และส่วนที่เหลือใช้เป็นข้อมูลทดสอบ ดังนั้นการทดลองของกระบวนการพยากรณ์ผลผลิตสุกรนั้น ผู้พัฒนาได้ทำการทดลองโดยให้สัดส่วนของข้อมูลเทรนนิ่งและข้อมูลทดสอบ มีสัดส่วนที่แตกต่างกันออกไป เพื่อต้องการจะดูผลลัพธ์ค่าความแม่นยำของการพยากรณ์ที่ออกมาว่าจะมีผลลัพธ์เป็นเช่นใด

ปัจจัยสำคัญในการทดลอง มีดังนี้

(1) ข้อมูลที่ใช้ในการทดลอง

สำหรับข้อมูลที่นำมาใช้ทดลองในกระบวนการพยากรณ์ผลผลิตสุกร เป็นข้อมูลจากกลุ่มสุกรทั้งหมดที่มีในฐานข้อมูล ซึ่งเป็นข้อมูลของประวัติการเลี้ยงสุกรตั้งแต่วันที่ 1 มกราคม พ.ศ. 2542 ถึงวันที่ 31 ธันวาคม พ.ศ. 2544 คือเป็นข้อมูลของ 3 ปี ซึ่งผ่านกระบวนการคัดเลือกข้อมูลและเตรียมข้อมูล ตามความเหมาะสมของบริบทการพยากรณ์ หลังจากคัดข้อมูลที่ไม่ได้คุณภาพพอจะนำมาใช้ในการพยากรณ์แล้ว ทำให้มีข้อมูลเหลือสำหรับการใช้ในการทดลองทั้งสิ้น 476 ข้อมูลตัวอย่าง

(2) รูปแบบการพยากรณ์

สำหรับรูปแบบการพยากรณ์ผลผลิตสุกรนั้น ใช้การพยากรณ์ด้วยการจำแนกข้อมูลโดยใช้วิธีแผนผังต้นไม้การตัดสินใจ เพื่อผลลัพธ์ในการระบุชื่อคลาสคุณภาพของสุกร และสังเกตว่ามีเปอร์เซ็นต์ความถูกต้องจากการระบุเป็นเท่าไร

(3) สัดส่วนของข้อมูลเทรนนิ่งและข้อมูลทดสอบ

เนื่องจากข้อมูลเทรนนิ่งเป็นข้อมูลที่ใช้ในการสร้างโมเดลการพยากรณ์ ดังนั้นสัดส่วนของข้อมูลเทรนนิ่งที่ใช้ในกระบวนการพยากรณ์จึงอาจถือได้ว่าน่าจะเป็นปัจจัยในการส่งผลให้ผลลัพธ์โมเดลและความแม่นยำของการพยากรณ์ส่งผลต่างกัน

4.4.2 การทดลองและผลการทดลอง

จากการออกแบบการทดลองเพื่อต้องการทดสอบว่าโมเดลการพยากรณ์ที่ระบบสร้างขึ้นมีศักยภาพในการพยากรณ์ในแง่ของค่าความแม่นยำเป็นเช่นไร เมื่อให้สัดส่วนของข้อมูลเทรนนิ่งและข้อมูลทดสอบมีสัดส่วนที่ต่างกัน ผู้จัดทำได้ทำการทดลองโดยทำการแบ่งสัดส่วนข้อมูลเทรนนิ่งต่อข้อมูลทดสอบเป็น 80:20 70:30 60:40 และ 50:50 โดยข้อมูลที่ถูกนำมาใช้ในการเป็นข้อมูลเทรนนิ่งนั้นเกิดจากวิธีการสุ่ม ผลการทดลองที่ได้ของการจำแนกข้อมูลด้วยวิธีแผนผังต้นไม้การตัดสินใจสามารถแสดงในตาราง 4.5

ตาราง 4.5 แสดงการทดลองและผลการทดลองของการจำแนกข้อมูล

สัดส่วนของข้อมูล	% ความแม่นยำของคลาส Low	% ความแม่นยำของคลาส Medium	% ความแม่นยำของคลาส High	% ความแม่นยำรวม
80:20	79.05%	92.74%	80.13%	86.26%
70:30	57.56%	82.45%	83.95%	74.35%
60:40	81.24%	90.29%	77.65%	81.44%
50:50	78.19%	64.75%	74.09%	71.75%

จากผลการทดลองที่แสดงในตารางที่ 4.5 พบว่าเมื่อนำข้อมูลทดสอบไปทดลองให้กระบวนการพยากรณ์ซึ่งสร้างขึ้นจากข้อมูลเทรนนิ่งทำการพยากรณ์ ค่าความแม่นยำที่ได้ในแต่ละการทดลองมีค่าต่างกันเนื่องจากการทดลองแต่ละครั้งเกิดจากโมเดลที่ถูกสร้างขึ้นมาคนละตัวกัน ซึ่งความแม่นยำโดยรวมของการทดลองอยู่ในระดับค่อนข้างเป็นที่น่าพอใจ โดยค่าความแม่นยำที่ได้ของสัดส่วน 80:20 มีค่าความแม่นยำโดยรวมสูงที่สุด

4.4.3 สรุปผลการทดลอง

จากผลการทดลองเพื่อสังเกตค่าความแม่นยำของการพยากรณ์ผลผลิตสุกร สรุปได้ว่าโมเดลการพยากรณ์ที่กระบวนการพยากรณ์สร้างขึ้น มีความสามารถในการพยากรณ์ระดับคุณภาพสุกรโดยที่โมเดลการพยากรณ์ที่ระบบสร้างขึ้นนั้นจะมีความแม่นยำเพียงใดนั้น อาจขึ้นอยู่กับปัจจัยภายนอกที่ส่งผลกับข้อมูลเทรนนิ่ง เช่น ความหลากหลายของลักษณะข้อมูลที่น่ามาสร้างโมเดล ปริมาณข้อมูลที่น่ามาสร้างโมเดล สัดส่วนของข้อมูลเทรนนิ่งและข้อมูลทดสอบ เป็นต้น