บทที่ 2
แนวคิดและทฤษฎีที่เกี่ยวข้อง

การวิจัยด้านการวิเคราะห์ปัจจัยที่ส่งผลต่อการขอตำแหน่งทางวิชาการจำเป็นต้องอาศัยพื้นฐานด้านทฤษฎี แนวคิด และงานวิจัยที่เกี่ยวข้อง เพื่อสร้างความเข้าใจที่ถูกต้องและเป็นกรอบอ้างอิงสำหรับการวิเคราะห์ในขั้นตอนต่อไป ดังนั้น บทนี้จะมุ่งนำเสนอแนวคิดและทฤษฎีที่เกี่ยวข้องกับการทำเหมืองข้อมูล เทคนิคการวิเคราะห์ที่เลือกใช้ ตลอดจนวรรณกรรมที่สนับสนุนแนวทางการดำเนินงาน เพื่อเป็นรากฐานสำคัญในการออกแบบกระบวนการวิจัยและการพัฒนาแบบจำลองในบทถัดไป

2.1 แนวคิดที่เกี่ยวข้อง

2.1.1 การทำเหมืองข้อมูล (Data Mining)

คือกระบวนการที่ดำเนินการกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

การทำเหมืองข้อมูล (Data Mining) เปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย ๆ มาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้ จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล โดยสามารถแบ่งออกเป็นข้อ ๆ ได้ดังนี้

กระบวนการหรือการเรียงลำดับของการค้นหาข้อมูลจำนวนมากและเก็บข้อมูลที่เกี่ยวข้อง การนำมาใช้โดยหน่วยงานทางธุรกิจและนักวิเคราะห์ทางการเงิน หรือการนำมาใช้ในด้านวิทยาศาสตร์ เพื่อวิเคราะห์ข้อมูลขนาดใหญ่ที่สร้างโดยวิธีการทดลองและการสังเกตการณ์ที่ทันสมัย การสกัดหรือแยกข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่หรือฐานข้อมูลการวางแผนทรัพยากรขององค์กรโดยสามารถวิเคราะห์ข้อมูลทางสถิติและตรรกะของข้อมูลขนาดใหญ่ เพื่อค้นหารูปแบบที่สามารถช่วยการตัดสินใจได้ วิวัฒนาการของการทำเหมืองข้อมูล

- ปี 1960 Data Collection คือ การนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือและป้องกันการสูญหายได้อย่างดี

- ปี 1980 Data Access คือ การนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ต่อกันในข้อมูล เพื่อประโยชน์ในการนำไปวิเคราะห์และการตัดสินใจอย่างมีคุณภาพ
- ปี 1990 Data Warehouse & Decision Support คือ การรวบรวมข้อมูลมาจัดเก็บลงในฐานข้อมูลขนาดใหญ่ โดยครอบคลุมทุกด้านขององค์กร เพื่อช่วยสนับสนุนการตัดสินใจ

2.2 ทฤษฎีที่เกี่ยวข้อง

ผู้จัดทำจึงได้รวบรวมข้อมูลที่มีความสำคัญและเกี่ยวข้องกับการพัฒนาโครงงาน โดยประกอบไปด้วย ทฤษฎี ที่มีความเกี่ยวเนื่องดังนี้

2.2.1 ความสัมพันธ์เชิงสหสัมพันธ์ (Correlation)

Correlation เป็นเทคนิคทางสถิติที่ใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรสองตัว โดยมุ่งเน้นที่การวัดขอบเขตและทิศทางของความสัมพันธ์เหล่านั้น ผลลัพธ์ของการคำนวณ correlation จะมีค่าอยู่ในช่วงระหว่าง –1 ถึง 1 โดยที่

- ค่า +1 หมายถึง ความสัมพันธ์เชิงบวกที่สมบูรณ์แบบ ตัวแปรทั้งสองเพิ่มขึ้นพร้อมกันในทิศทางเดียวกัน
- ค่า –1 หมายถึง ความสัมพันธ์เชิงลบที่สมบูรณ์แบบ เมื่อหนึ่งตัวแปรเพิ่มขึ้นอีกตัวแปรจะลดลงในทิศทางตรงข้าม
- ค่า 0 หมายถึง ไม่มีความสัมพันธ์ระหว่างตัวแปรทั้งสอง

เทคนิค correlation มีการใช้อย่างแพร่หลายในการวิเคราะห์ข้อมูลเพื่อประเมินว่าตัวแปรสองตัวมีความสัมพันธ์กันอย่างไร ไม่ว่าจะในด้านธุรกิจ เศรษฐกิจ หรือวิทยาศาสตร์ ตัวอย่างเช่น ในการวิเคราะห์การตลาด ผู้วิเคราะห์อาจต้องการตรวจสอบความสัมพันธ์ระหว่างค่าใช้จ่ายในการโฆษณากับยอดขายสินค้า หรือในด้านการแพทย์ อาจต้องการตรวจสอบความสัมพันธ์ระหว่างการบริโภคอาหารบางชนิดกับความเสี่ยงในการเกิดโรค การคำนวณ correlation ส่วนใหญ่จะใช้สูตร Pearson Correlation Coefficient (r) ซึ่งนิยมใช้สำหรับตัวแปรที่มีการกระจายเชิงเส้น (linear relationship) และสามารถใช้กับข้อมูลที่เป็นสเกล interval หรือ ratio ได้การวิเคราะห์ correlation ยังสามารถนำไปใช้ในการทำนายพฤติกรรมของข้อมูล ซึ่งถ้าพบว่ามีความสัมพันธ์สูง อาจนำไปสู่การสร้างโมเดลทำนายที่แม่นยำขึ้น

วิธีการวัดค่า Correlation ในการคำนวณความสัมพันธ์ระหว่างตัวแปรสองตัว คือ (Pearson Correlation Coefficient) (r) ซึ่งจะใช้เมื่อตัวแปรทั้งสองมีการกระจายแบบเชิงเส้น

(linear relationship) และเป็นข้อมูลเชิงปริมาณ (interval หรือ ratio) สูตรการคำนวณ Pearson's คือ

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (1)$$

โดยที่

n = จำนวนข้อมูลคู่

x และ y = ตัวแปรสองตัวที่ต้องการวัดความสัมพันธ์

ผลลัพธ์ที่ได้จากค่า Pearson's r จะอยู่ในช่วง –1 ถึง 1 โดย

- r=1 หมายถึงความสัมพันธ์เชิงบวกที่สมบูรณ์แบบ
- r=–1หมายถึงความสัมพันธ์เชิงลบที่สมบูรณ์แบบ
- r=0 หมายถึงไม่มีความสัมพันธ์ระหว่างตัวแปรทั้งสอง

แม้ว่าการวิเคราะห์ correlation จะเป็นเครื่องมือที่มีประโยชน์ในการตรวจสอบความสัมพันธ์ แต่ควรระมัดระวังการตีความผลลัพธ์ เนื่องจาก correlation ไม่ได้หมายถึงเหตุและผล (Causation) เช่น หากพบว่ามีความสัมพันธ์ระหว่างจำนวนห้างสรรพสินค้ากับอัตราการเกิดอาชญากรรม ไม่ได้หมายความว่าห้างสรรพสินค้าเป็นสาเหตุของการเกิดอาชญากรรม แต่อาจมีตัวแปรที่สามที่ส่งผลกระทบต่อทั้งสอง เช่น การเพิ่มขึ้นของประชากรในพื้นที่

2.2.2 ทฤษฎีการจำแนกประเภท (Classification)

1) เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

เป็นวิธีการวิเคราะห์ข้อมูลที่อยู่ในกลุ่มของ การเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยมุ่งใช้เพื่อการจำแนกประเภท (Classification) และการพยากรณ์ค่าตัวเลข (Regression) ลักษณะของแบบจำลองอยู่ในรูปโครงสร้างต้นไม้ ซึ่งประกอบด้วย โหนดราก (Root Node), โหนดภายใน (Internal Nodes) และ โหนดใบ (Leaf Nodes) วิธีการทำงานของ Decision Tree

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (2)$$

โดยที่

- S = ชุดข้อมูล
- c = จำนวนคลาสทั้งหมด
- pi = ความน่าจะเป็นที่ตัวอย่างใน S อยู่ในคลาสที่ i

เริ่มต้นจากการเลือกตัวแปรหรือคุณลักษณะ (Feature) ที่เหมาะสมที่สุดในการแบ่งแยกข้อมูล โดยใช้เกณฑ์วัดความบริสุทธิ์ (Impurity Measures) เช่น Gini Index, Entropy (Information Gain) หรือ Variance Reduction จากนั้นจึงทำการแตกแขนงข้อมูลออกเป็นกิ่งก้าน (Branches) จนกระทั่งได้โหนดใบที่แสดงถึงผลลัพธ์การจำแนกประเภทหรือค่าที่ทำนายได้ ข้อเด่นของเทคนิคต้นไม้ตัดสินใจคือ ความสามารถในการตีความผลลัพธ์ได้ง่ายและแสดงเหตุผลของการตัดสินใจในเชิงตรรกะอย่างชัดเจน อีกทั้งยังสามารถประยุกต์ใช้ได้กับข้อมูลเชิงคุณภาพ (Categorical) และข้อมูลเชิงปริมาณ (Numerical) อย่างไรก็ตาม ข้อจำกัดที่ควรระมัดระวังคือ การเกิดปัญหา Overfitting หากต้นไม้มีความซับซ้อนมากเกินไป ดังนั้นจึงมักมีการใช้กระบวนการ Pruning (การตัดกิ่ง) เพื่อลดความซับซ้อนและเพิ่มความสามารถในการทำนายกับข้อมูลใหม่

2) เทคนิคป่าสุ่ม (Random Forest)

เป็นเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ที่อยู่ในกลุ่มการเรียนรู้แบบมีผู้สอน (Supervised Learning) และถูกใช้สำหรับงานจำแนกประเภท (Classification) และการทำนายค่าตัวเลข (Regression) เช่นเดียวกับ Decision Tree แต่ Random Forest มีการนำจุดแข็งของหลาย ๆ Decision Tree มารวมกันเพื่อเพิ่มความแม่นยำและลดปัญหาของการใช้ Decision Tree เพียงต้นเดียว เช่น การเกิดปัญหา Overfitting Random Forest

$$\hat{y} = \arg\max_{c \in C} \sum_{b=1}^{B} I(h_b(x) = c) \quad (3)$$

โดยที่

- B = จำนวนต้นไม้ทั้งหมด
- hb(x) = ผลลัพธ์จากต้นไม้ที่ b
- I = Indicator Function (ถ้าเป็นจริง = 1, ถ้าไม่ใช่ = 0)

ทำงานโดยการสร้าง "ป่า" ของต้นไม้หลาย ๆ ต้น ซึ่งแต่ละต้นในป่านี้จะถูกสร้างจากการสุ่มข้อมูลตัวอย่างย่อย ๆ (Bootstrap Sampling) และการสุ่มคุณสมบัติ (Features) บางส่วนสำหรับการสร้างต้นไม้ในแต่ละโหนด จากนั้นเมื่อมีการทำนายข้อมูลใหม่ ระบบจะทำนายผลลัพธ์จากแต่ละต้นไม้ และใช้เสียงข้างมาก (Majority Voting) สำหรับงาน Classification หรือค่าเฉลี่ย (Averaging) สำหรับงาน Regression ในการหาผลลัพธ์สุดท้าย

3) เทคนิค K–Nearest Neighbors (K–NN)

เป็นวิธีการจำแนกประเภท (Classification) ที่อยู่ในกลุ่มของ การเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยอาศัยหลักการเปรียบเทียบความใกล้เคียงระหว่างข้อมูลใหม่กับข้อมูลตัวอย่างที่มีป้ายกำกับ (Label) อยู่แล้ว

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} \quad (4)$$

โดยที่

- dist : ค่าระยะทางระหว่างจุดข้อมูลสองจุด (Point) คือ p และ q
- pk: ค่าของจุดข้อมูล p ในมิติที่ k
- qk ค่าของจุดข้อมูล q ในมิติที่
- n : จำนวนมิติ (Dimension) ของข้อมูล

วิธีการทำงานของ K–NN เริ่มจากการคำนวณ ระยะทาง (Distance Metric) ระหว่างข้อมูลใหม่กับข้อมูลทั้งหมดในชุดฝึก (Training Set) โดยทั่วไปนิยมใช้ระยะทางแบบยูคลิด (Euclidean Distance) จากนั้นทำการเลือกข้อมูลจำนวน K ตัวอย่างที่อยู่ใกล้ที่สุด (Nearest Neighbors) แล้วใช้วิธีการตัดสินใจตาม เสียงข้างมาก (Majority Voting) ของกลุ่มเพื่อนบ้านนั้น หากส่วนใหญ่เป็นคลาสใด ข้อมูลใหม่ก็จะถูกจัดให้อยู่ในคลาสนั้น ข้อเด่นของ K–NN คือความเรียบง่าย ไม่จำเป็นต้องสร้างแบบจำลองเชิงสมการที่ซับซ้อน และ สามารถประยุกต์ใช้ได้ทั้งกับข้อมูลเชิงคุณภาพและเชิงปริมาณ อย่างไรก็ตาม ข้อจำกัดที่สำคัญคือการใช้เวลาในการคำนวณสูงเมื่อต้องประมวลผลข้อมูลจำนวนมาก และประสิทธิภาพของแบบจำลองขึ้นอยู่กับการเลือกค่าพารามิเตอร์ K ที่เหมาะสม รวมทั้งการเลือกมาตรวัดระยะทางที่สอดคล้องกับลักษณะของข้อมูล

2.2.3 เทคนิคการจัดกลุ่ม (Clustering)

Clustering เป็นเทคนิคการวิเคราะห์ข้อมูลที่มุ่งเน้นการจัดกลุ่ม (กลุ่มคลัสเตอร์) ข้อมูลที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน โดยไม่จำเป็นต้องอ้างอิงถึงข้อมูลที่มีการแบ่งประเภทไว้ล่วงหน้า การทำ Clustering มักใช้กับข้อมูลที่ไม่มีการระบุหมวดหมู่ หรือมีลักษณะกระจัดกระจาย การแบ่งกลุ่มนี้ช่วยให้นักวิเคราะห์สามารถทำความเข้าใจโครงสร้าง

ของข้อมูลและค้นพบรูปแบบที่ซ่อนอยู่ การทำ (Clustering) ทำงานโดยการจัดกลุ่มข้อมูลให้มีความคล้ายคลึงกันสูงสุดภายในกลุ่มเดียวกัน และความแตกต่างสูงสุดระหว่างกลุ่มที่ต่างกัน นั่นหมายความว่าข้อมูลในกลุ่มเดียวกันจะมีคุณสมบัติที่คล้ายคลึงกันมากกว่ากลุ่มอื่น ๆ

2.2.4 การวัดประสิทธิภาพ คอนฟิวชันเมททริกซ์ (Confusion Matrix)

(Confusion Matrix) เป็นวิธีหนึ่งที่ใช้วัดประสิทธิภาพของโมเดลการเรียนรู้ของเครื่อง (Machine Learning) โดยเฉพาะในงานจำแนกประเภท (Classification) (Confusion Matrix) เป็นตารางที่ใช้ในการแสดงผลลัพธ์ของโมเดลเปรียบเทียบกับข้อมูลจริง เพื่อให้เห็นภาพรวมของการทำนายได้อย่างชัดเจนโครงสร้างของ Confusion Matrix โดยประกอบด้วยค่า 4 ค่าหลัก ดังนี้

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

ภาพที่ 2.1 ตาราง Confusion Matrix

- True Positive (TP) จำนวนตัวอย่างที่โมเดลทำนายว่ามี Positive และจริง ๆ แล้วก็เป็น Positive
- True Negative (TN) จำนวนตัวอย่างที่โมเดลทำนายว่าเป็น Negative และจริง ๆ แล้วก็เป็น Negative
- False Positive (FP) จำนวนตัวอย่างที่โมเดลทำนายว่าเป็น Positive แต่จริง ๆ แล้วเป็น Negative (เรียกว่า Type I Error หรือ False Alarm)
- False Negative (FN) จำนวนตัวอย่างที่โมเดลทำนายว่าเป็น Negative แต่จริง ๆ แล้วเป็น Positive (เรียกว่า Type II Error หรือ Missed Detection)

ค่าที่ได้จาก Confusion Matrix

- จากค่าในตาราง Confusion Matrix สามารถคำนวณตัวชี้วัดต่าง ๆ ได้ เช่น
- Accuracy (ความถูกต้อง) = (TP + TN) / (TP + TN + FP + FN)
- Precision (ความแม่นยำ) = TP / (TP + FP)
- Recall (Sensitivity หรือ TPR) = TP / (TP + FN)
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
- Specificity (TNR) = TN / (TN + FP)

การใช้ Confusion Matrix ช่วยให้สามารถประเมินผลการทำงานของโมเดลได้ละเอียดขึ้น ไม่ได้พิจารณาแค่ความถูกต้องโดยรวม (Accuracy) แต่ยังสามารถวิเคราะห์ข้อผิดพลาดในรูปแบบต่าง ๆ เพื่อนำไปปรับปรุงโมเดลให้ดีขึ้น

2.3 เครื่องมือในการออกแบบและวิเคราะห์ระบบ

2.3.1 โปรแกรม (Microsoft excel)

ไมโครซอฟท์ เอ็กซ์เซล (อังกฤษ Microsoft Excel) เป็นโปรแกรมประเภทตารางการคำนวณ (สเปรดชีต) พัฒนาโดยบริษัทไมโครซอฟท์ และเป็นโปรแกรมหนึ่งในชุดไมโครซอฟท์ ออฟฟิศ สำหรับจัดการและคำนวณข้อมูลในรูปแบบตาราง อีกทั้งสามารถจัดทำกราฟ แผนภูมิเพื่อแสดงผลข้อมูลได้ โดยเวอร์ชันล่าสุดคือ ไมโครซอฟท์ เอกซ์เซล 2016 (Microsoft Excel 2016) ไมโครซอฟท์ เอกซ์เซล เป็นโปรแกรมที่ได้รับความนิยมในด้านการการคำนวณทางคณิตศาสตร์โดยใช้ฟังก์ชันพื้นฐาน บวก ลบ คูณ หาร ยกกำลัง รวมถึงฟังก์ชันทางคณิตศาสตร์ระดับสูง เช่น Modulo, ตรีโกณมิติ (Sin Cos Tan) ฟังก์ชันทางสถิติ เช่น ค่าเบี่ยงเบนมาตรฐาน ฟังก์ชันทางการเงิน เช่น การคิดค่าเสื่อมราคา, การคำนวณค่าปัจจุบัน ฟังก์ชันในการตัดต่อคำ เช่น Concatenate ฟังก์ชันในการค้นหาข้อมูล เช่น Lookup, vlookup และ hlookup สำหรับส่วนที่ถือว่าเป็นสิ่งที่เยี่ยมยอดของ ไมโครซอฟท์ เอกซ์เซล คือ การใช้งานในรูปแบบของฐานข้อมูล ซึ่งสามารถจัดการฐานข้อมูลที่มีขนาดไม่ใหญ่มาก คือมีประมาณไม่เกิน 65,000 ตาราง ไม่ว่าจะเป็น ตัวกรอง, การเรียงลำดับข้อมูล (Sort) , คำนวณยอดรวม (Subtotal) และตารางไพวอต (Pivot Table) เป็นคำสั่งสำหรับสรุปข้อมูลให้อยู่ในรูปแบบที่ดูได้ง่าย สามารถหมุนเปลี่ยนตามต้องการ นอกจากนี้ยังสามารถทำกราฟในแบบต่าง ๆ เช่น เส้นตรง วงกลม กราฟรูปแท่ง กราฟแท่งเทียนที่ใช้กับการวิเคราะห์หุ้นก็ทำได้ กราฟพื้นที่ สามารถทำกราฟต่าง ๆ ให้อยู่ในรูปแบบ 2 มิติ หรือ 3 มิติได้ด้วย รวมถึงทำกราฟ 2 ชนิดในรูปเดียวกันได้ด้วย

2.3.2 โปรแกรม (RapidMiner Studio)

(Rapidminer) คือซอฟต์แวร์ (Data Science) ใช้สำหรับการเตรียมข้อมูล การเรียนรู้เครื่อง การเรียนรู้ลึก การทำเหมืองข้อความ และการวิเคราะห์การทำนาย (Predictive analysis) เป็นซอฟต์แวร์ที่ช่วยในการจัดส่งข้อมูล และลดข้อผิดพลาดจนแทบจะไม่จำเป็นต้องเขียนโค้ดเพิ่ม แต่ที่ทำให้เป็นเครื่องมือที่เหล่า Data Scientist นิยมเลือกใช้เป็นเพราะว่าตัว (Rapidminer) มีขั้นตอนพร้อมสำหรับการทำ (Data mining) ขุดข้อมูล และ (Machine learning)

ซึ่งรวมไปถึงการโหลดและการแปลงข้อมูล (ETL) การประมวลผลล่วงหน้าและการวาดภาพจากข้อมูล การวิเคราะห์เชิงพยากรณ์และการสร้างแบบจำลองทางสถิติ การประเมินผลและการปรับใช้ ต่างๆ ล้วนเป็นสิ่งที่เหล่า (Data Scientist) จำเป็นต้องทำในการเข้าใจข้อมูลมากขึ้น

2.4 วรรณกรรมที่เกี่ยวข้อง

วรการ ใจดี และ นรินทร์ จิวิตัน ได้ศึกษาปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรี คณะบริหารธุรกิจและศิลปศาสตร์ รวมถึงการเปรียบเทียบประสิทธิภาพเทคนิคการทำเหมืองข้อมูล โดยใช้ข้อมูลพื้นฐานการรับสมัครนักศึกษาจากงานส่งเสริมวิชาการและงานทะเบียน มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่ ในช่วงปีการศึกษา 2563–2565 โดยมีข้อมูลผู้สมัครทั้งหมด 2,509 รายการ คณะผู้วิจัยได้นำข้อมูลมาวิเคราะห์ปัจจัยโดยใช้เทคนิคการคัดเลือกคุณลักษณะและเทคนิคการทำเหมืองข้อมูลเทคนิคการคัดเลือกคุณลักษณะที่ใช้ในการศึกษาประกอบด้วย 3 เทคนิค ได้แก่ เทคนิคการหาค่าสถิติไคสแควร์ เทคนิคการหาค่าสัมประสิทธิ์สหสัมพันธ์ และเทคนิคการหาค่าเกนความรู้ ส่วนเทคนิคการทำเหมืองข้อมูลที่ใช้ประกอบด้วย 6 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคป่าสุ่ม เทคนิคเกรเดียนท์บูสทรีส์ เทคนิคนาอีฟเบย์ เทคนิคการถดถอยโลจิสติก และเทคนิคการโหวตผล การศึกษาพบว่า ปัจจัยที่ส่งผลต่อการตัดสินใจเข้าศึกษาต่อในระดับปริญญาตรีมากที่สุด 3 อันดับแรกจากทุกเทคนิคของการคัดเลือกคุณลักษณะ ได้แก่ สาขาวิชาหรือหลักสูตรที่เลือก สายการเรียนเดิม และวุฒิการศึกษาเดิม นอกจากนี้ เทคนิคการทำเหมืองข้อมูลที่มีค่าความถูกต้องสูงสุดในการพยากรณ์คือเทคนิคการโหวต ซึ่งมีค่าความถูกต้องเท่ากับ 73.44% โดยมีค่าความถูกต้องสูงกว่าเทคนิคป่าสุ่ม (71.45%) เทคนิคการถดถอยโลจิสติก (71.05%) เทคนิคต้นไม้ตัดสินใจ (68.26%) เทคนิคนาอีฟเบย์ (65.60%) และเทคนิคเกรเดียนท์บูสทรีส์ (64.81%) ตามลำดับ ผลการศึกษาพบว่า กลุ่มตัวอย่างส่วนใหญ่เป็นเพศหญิง มีอายุระหว่าง 31 – 40 ปี มีสถานภาพโสด มีวุฒิการศึกษาระดับปริญญาตรี ส่วนใหญ่ประกอบอาชีพเป็นพนักงานของรัฐ/พนักงานมหาวิทยาลัย มีรายได้เฉลี่ย ไม่น้อยกว่า 50,000 บาท และมีประสบการณ์ในการทำงานด้านการศึกษาวิทยาศาสตร์สุขภาพ ในช่วง 1 – 3 ปี โดยผู้ตอบแบบสอบถามให้ความสำคัญกับปัจจัยที่มีผลต่อการตัดสินใจเข้าศึกษาต่อระดับปริญญาโท สาขาวิชาการศึกษาวิทยาศาสตร์สุขภาพ อยู่ในระดับมาก และให้ความสำคัญกับปัจจัยทางด้านสถาบันการศึกษา

มากที่สุด รองลงมาคือ ปัจจัยทางด้านหลักสูตร ปัจจัยทางด้านสังคม และ ปัจจัยทางด้านค่าใช้จ่ายในการศึกษา ตามลำดับ

จีระนันต์ เจริญรัตน์ ได้ศึกษาปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาที่มีผลการเรียนปกติ โดยพบว่าโมเดลที่สร้างขึ้นมีขนาดต้นไม้ที่ใหญ่และมีโหนดจำนวนมาก ทำให้โมเดลมีความซับซ้อน การทำความเข้าใจและการแปลผลทำได้ยาก ซึ่งสอดคล้องกับแนวคิดของนิตยา เกิดประสบ ที่กล่าวว่าโมเดลหรือความสัมพันธ์ที่สร้างขึ้นจำเป็นต้องได้รับการทดสอบอัตราความผิดพลาดและวิเคราะห์ความซับซ้อนของรูปแบบโมเดล หากอัตราความผิดพลาดยังคงสูงเกินไป อาจจำเป็นต้องย้อนกลับไปที่ขั้นตอนการค้นหาโมเดลหรือขั้นตอนการคัดเลือกข้อมูลเพื่อปรับปรุงโมเดลให้ถูกต้องยิ่งขึ้น ในทำนองเดียวกัน หากโมเดลที่ได้มามีรูปแบบที่ซับซ้อนเกินไปจนยากต่อการทำความเข้าใจ อาจต้องย้อนกลับไปที่ขั้นตอนการสร้างและค้นหาโมเดลเพื่อหาโมเดลใหม่ที่มีความถูกต้องเท่าเดิมแต่มีรูปแบบที่ซับซ้อนน้อยลง จากผลการวิจัยนี้สามารถนำกฎการจำแนกข้อมูลที่ได้มาใช้ในการพัฒนาระบบพยากรณ์การพ้นสภาพของนักศึกษาและผู้บริหารสามารถใช้เป็นแนวทางในการบริหารจัดการหรืออาจารย์ที่ปรึกษาสามารถวางแผนการเรียน ดูแลการลงทะเบียนเรียนของนักศึกษา ให้ความช่วยเหลือ และส่งเสริมนักศึกษาได้อย่างเหมาะสม

เยาวลักษณ์ เกิดปั้น และ สุวรรณี อัศวกุลชัย ได้ศึกษาการเรียนรู้ของเครื่องในการพยากรณ์ผลผลิตกล้วยไม้ในประเทศไทยในการศึกษาครั้งนี้ใช้กระบวนการวิเคราะห์ข้อมูลมาตรฐาน เรียกว่า Cross-industry standard process (CRISP) ประกอบด้วย 6 ขั้นตอน ได้แก่ เข้าใจปัญหาของธุรกิจ เข้าใจข้อมูล เตรียมข้อมูล พัฒนาแบบจำลอง การประเมิน และการนำไปใช้จริงข้อมูลที่ใช้รวบรวมข้อมูลพื้นที่เพาะปลูก จำนวนครัวเรือน และผลผลิตกล้วยไม้จาก www.oae.go.th และ www.ditp.go.th ในปี 2559-2563 จำนวน 20 จังหวัด การเตรียมและคัดเลือกข้อมูลให้สมบูรณ์เหมาะสำหรับการทำเหมืองข้อมูลด้วยกระบวนการ Extract transform load(ETL)จากนั้นทำการโอนย้ายข้อมูล การลดขนาดของข้อมูลและการทำความสะอาดข้อมูลใช้โปรแกรม (Knime) เป็นเครื่องมือ วิเคราะห์หาความสัมพันธ์และใช้อัลกอริทึมพยากรณ์ผลผลิตกล้วยไม้ด้วย 3 แบบได้แก่ Simple regression tree, (Gradientboosted) trees และ Random forest การทดสอบประสิทธิภาพอัลกอริทึมด้วยการหาR2เพื่อหาอัลกอริทึมที่เหมาะสมที่สุดและวัดผล ซึ่งผลการศึกษา พบว่า การพยากรณ์ผลผลิตกล้วยไม้ในประเทศไทยจากอัลกอริทึม Gradient boosted tree มีค่าถูกต้องสูงสุด คิดเป็นร้อยละ 97.50 เมื่อเทียบกับ

อัลกอริทึม Simple regression treeและ Random forest ถูกต้องร้อยละ 96.40 และ 92.80 ตามลำดับ สามารถนำมาแสดงผลข้อมูลด้วยภาพโดยใช้ Power business intelligence (Power BI)ซึ่งเป็นประโยชน์ในการวางแผนการผลิตเพื่อเพิ่มยอดขายได้เป็นอย่างดี เทคนิคการเรียนรู้ของเครื่องเป็นเทคโนโลยีสมัยใหม่ที่สามารถประยุกต์ใช้ด้านการเกษตรเพื่อการเปลี่ยนถ่ายสู่เกษตรยุค 4.0 เป็นเกษตรแม่นยำ (Precision agriculture)ช่วยให้เกษตรกร และหน่วยงานที่เกี่ยวข้องสามารถพยากรณ์ผลผลิตได้แม่นยำขึ้น สนับสนุนการทำเกษตรกรรม ยกระดับชีวิตเกษตรกร และช่วยเพิ่มขีดความสามารถในการแข่งขันมากยิ่งขึ้นตลอดจนสามารถประยุกต์ใช้ในอุตสาหกรรมของพืชเศรษฐกิจชนิดต่าง ๆ ได้

วชราวรรณ อินทาโยด ชยพล กมยด และ ปุณณ์นมล เติมดี ปัญหาที่ศึกษา ข้อมูลจากการใช้งานเว็บของการศึกษา (educational web usage data) มักมีการแบ่งกลุ่มคลาสแบบ “ผ่าน / ไม่ผ่าน” โดยที่คลาส “ไม่ผ่าน” (fail) เป็นคลาสน้อย (minority class) ซึ่งทำให้โมเดล classification มีแนวโน้มลำเอียงไปยังคลาสใหญ่ (majority class) และประสิทธิภาพในการทำนายคลาสน้อยต่ำ โดยใช้เทคนิค SMOTE และเวอร์ชันปรับปรุง ได้แก่ Borderline–SMOTE1, Borderline–SMOTE2 และ SVM–SMOTE ทำ oversample คลาสน้อยให้มีตัวอย่างมากขึ้น จึงได้ผลลัพธ์ ทั้ง precision, recall และ F1–score ของคลาสน้อยดีขึ้นเมื่อใช้ SMOTE และเวอร์ชันของ SMOTE เหล่านี้ โดยเฉพาะโมเดลจำแนกประเภทที่ใช้ Decision Tree และ K–NN ได้รับประโยชน์ชัดเจน เพราะสามารถทำนายกลุ่มไม่ผ่านได้ดีขึ้นหลังจาก oversampling แล้ว จัดการจำแนก (classification) โดยใช้โมเดล Naive Bayes, Decision Tree, และ K–Nearest Neighborsสรุปได้ว่าโครงสร้างของข้อมูล (features) และวิธีแบ่งชุดทดสอบ–ฝึก (cross validation) มีผลต่อผลลัพธ์และการ SMOTE ช่วยปรับปรุงการทำนายคลาสเล็ก แต่ก็อาจเพิ่มโอกาส overfitting ถ้าไม่ระวัง

พิมพ์ประภา พาลพ่าย และ วชิระ พรหมวงศ์ ได้ทำการศึกษาการใช้เทคนิคเหมืองข้อมูลเพื่อ “จำแนก” ผลการเรียนของนักศึกษาปริญญาโทบนระบบ STOU eLearning โดยเน้นตัวแปรพฤติกรรมการเรียน (log) แล้วสร้างแบบจำลอง Decision Tree เปรียบเทียบเส้นทางกฎการตัดสินใจ 3 เส้นทาง (เช่น เกณฑ์จำนวนครั้งที่เข้าเรียน และการทำแบบทดสอบก่อน–หลังเรียน) ผลพบว่าโมเดลทำนาย “สอบผ่าน” ได้แม่น (≈87.9%) และความแม่นยำรวม ≈84.6% ชี้ให้เห็นศักยภาพของกฎเชิงตีความสำหรับสนับสนุนการตัดสินใจในบริบทการศึกษา—สอดคล้องกับแนวทางในรายงานของคุณที่ใช้ Decision Tree เพื่อสร้างกฎประเมินความพร้อม/ศักยภาพบุคลากร