

การพัฒนาแบบจำลองทำนายผลฟุตบอลพรีเมียร์ลีกด้วยเทคนิคดาต้าไมนิ่ง

The developing of prediction model Premier League football results with datamining technique

เบญจกัฏ จงหมื่นไวย^{1*} และ เกรียงศักดิ์ ภูมิรินทร์²

Benjapuk Jongmuanwai^{1*} and Kreangsak Poommarin²

หลักสูตรระบบสารสนเทศเพื่อการจัดการ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา^{1,2}

E-Mail “benjapuk.j@nrru.ac.th”, “6140214105@nrru.ac.th”

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อศึกษาและพัฒนาแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก โดยใช้เทคนิคเหมืองข้อมูล 2) เพื่อประเมินประสิทธิภาพการทำนายผลและวางแผนสำหรับนำไปใช้ในการตัดสินใจการแข่งขันฟุตบอลพรีเมียร์ลีก กลุ่มเป้าหมาย คือ ผลฟุตบอลพรีเมียร์ลีก ระหว่างปี ค.ศ. 2012-2022 ได้มาจากการเลือกแบบเจาะจง จำนวน 3,992 เรคคอร์ด ขั้นตอนการดำเนินการวิจัยแบบ CRISP-DM มีจำนวน 6 ขั้นตอน และใช้เทคนิคในการสร้างแบบจำลองข้อมูล 2 เทคนิค คือ การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจและการค้นหาความสัมพันธ์ของข้อมูลแบบเบย์ สถิติที่ใช้สำหรับการวิจัย ได้แก่ ค่าร้อยละ ค่าความถูกต้อง

ผลการวิจัยพบว่า 1) ผลการพัฒนาแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก แผนภูมิต้นไม้ที่มีผลต่อการทำนายผลฟุตบอลพรีเมียร์ลีก ซึ่งแอททริบิวต์ที่มีผลสูงสุด ได้แก่ วันที่มากกว่า 77.80 ครั้ง จะเกี่ยวข้องกับผู้ตัดสินผลฟุตบอลที่ชื่อว่า M Dean และ R East วันที่น้อยกว่าเท่ากับ 77.80 จะมีผลต่อประตูทีมเยือนครั้งแรกหากทีม PSCA มากกว่า 1.72 ส่งผลให้เจ้าบ้านทำผิดกติกา (HF) และถ้าน้อยกว่าหรือเท่ากับ 1.72 ส่งผลให้ชนะเสมอ 2) ประสิทธิภาพของแบบจำลองร้อยละความถูกต้องของ 2 เทคนิค ได้แก่ Decision Tree, Naïve Bayes ค่าความถูกต้องเฉลี่ยเท่ากับร้อยละ 80.67 อยู่ในระดับมาก และเทคนิคที่มีค่าความถูกต้องสูงสุด คือ Decision Tree ร้อยละ 98.58 รองลงมา คือ เทคนิค Naïve Bayes ร้อยละ 62.75

คำสำคัญ: ทำนายผลฟุตบอล, เทคนิคเหมืองข้อมูล, ต้นไม้ตัดสินใจ, กฎหาความสัมพันธ์แบบเบย์

ABSTRACT

The purposes of the research were to 1) study and development of prediction model of Premier League football results with a data mining technique, 2) assess the efficiency of prediction and planning for decision support Premier League football. The participants all were by purposive samplings and 3992 records. The process of research CRISP-DM is six steps and two technique data models in the Decision Tree and Naïve Bayes. The statically is percentage and accuracy.

The research findings showed that the developing of a prediction model for Premier League football results in a Decision Tree which a high factor of date of more than 77.80, with the referee football name being M Dean and R East on a date less than equal to 77.80 the result Half Time Away Team Goals (HTAG) then if PSCA team more than 1.72 to result in Home Team Fouls Committed (HF) and If less than or equal to 1.72, results in a draw 2) the performance of the model percentage accuracy of 2 techniques: Decision Tree, Naïve Bayes. The average accuracy was 80.67% at a high level. Furthermore, the technique with the highest accuracy was Decision Tree at 98.58%, followed by the Naïve Bayes technique at 62.75%.

Keywords: Prediction model Premier League football, Data Mining, Decision Tree, Naïve Bayes

บทนำ

กีฬาฟุตบอล ช่วยเพิ่มความแข็งแรงของกล้ามเนื้อขา เพราะกีฬาฟุตบอลเป็นกีฬาที่ต้องเดินและวิ่งอยู่ตลอดเวลา นอกจากนี้ กีฬาฟุตบอลยังเป็นกีฬาที่มีส่วนช่วยในเรื่องความเชื่อมั่นในตนเอง มีความกล้าที่จะตัดสินใจ เสริมสร้างทางไหวพริบและความคิด เรียนรู้คำว่า รู้แพ้-รู้ชนะ-รู้ภัย ซึ่งฟุตบอลเป็นกีฬาประเภทหนึ่งที่มีประโยชน์ต่อผู้เล่น ผู้เล่นต้องเคลื่อนไหวร่างกายตลอดเวลา รวมทั้งต้องมีไหวพริบ มีอารมณ์มั่นคง มีสมาธิ มีความเชื่อมั่นด้วยตนเอง สามารถแก้ปัญหาต่าง ๆ และตัดสินใจที่รวดเร็วและถูกต้อง ย่อมก่อให้เกิดประโยชน์แก่ผู้เล่นทั้งทางตรงและทางอ้อมเสมอ ดังนั้น ฟุตบอลจึงเป็นกีฬาอีกประเภทหนึ่งที่ช่วยฝึกฝนให้ผู้เล่นมีไหวพริบที่ชาญฉลาดและแก้ปัญหาอย่างฉับพลันได้ดี ฟุตบอลมีการเล่นที่มีกฎ กติกา ผู้เล่นต้องเคารพและปฏิบัติตามกฎกติกาการเล่น ซึ่งจะทำให้ผู้เล่นรู้จักความยุติธรรมและปฏิบัติตนให้อยู่ในขอบเขตที่ควรกระทำ สามารถนำไปใช้ชีวิตประจำวัน ให้รู้จักเคารพสิทธิของผู้อื่น มีความอดกลั้น อดทน ยอมรับฟังความคิดเห็นของผู้อื่น รู้แพ้ รู้ชนะ รู้ภัย รู้จักการเป็นผู้นำ และผู้ตามที่ดีตลอดจนรู้จักปฏิบัติหน้าที่อันถูกต้อง

ปัจจุบันวงการฟุตบอลนิยมอย่างมากในหมู่กีฬาทั้งหลาย ทำให้เกิดการขยายผลเพื่อความสนุกและนำมาทายผลในรายการโทรทัศน์ในช่องกีฬาต่าง ๆ ทำให้ทายผิดและทายถูกผสมกันไปทำให้ไม่แน่นอน โดยโครงการการพัฒนาทางระบบสารสนเทศเพื่อการจัดการครั้งนี้ เป็นการนำหลักการวิเคราะห์ข้อมูลทางด้านเทคโนโลยีสารสนเทศที่มีการเรียนรู้ เพื่อสร้างอัลกอริทึมที่คาดการณ์ผลลัพธ์ของการแข่งขันได้อย่างแม่นยำ นอกเหนือจากการให้สถิติเชิงลึกเกี่ยวกับทีมผู้เล่นและการแข่งขันโดยทั่วไปแล้ว ยังมีการให้คำแนะนำที่สมบูรณ์เกี่ยวกับการเดิมพันแบบอัจฉริยะ ซึ่งจะช่วยให้ผู้ทายผลบอลมือใหม่ได้ตัดสินใจอย่างมีความรู้ เน้นการนำเสนอเคล็ดลับฟรี และยังมีคำแนะนำแบ่งปันกลยุทธ์และวิเคราะห์สถิติเพื่อให้ผู้ที่เกี่ยวข้องมีความเข้าใจกับโลกออนไลน์ของการทำนายผลฟุตบอล

ดังนั้น ทางผู้พัฒนาโครงการได้ศึกษาและรวบรวมข้อมูลการทำนายผลฟุตบอลพรีเมียร์ลีก เพื่อนำเข้าข้อมูลประมวลผลและสร้างแบบจำลองเพื่อนำไปใช้สำหรับการตัดสินใจในด้านต่าง ๆ ได้ง่าย สะดวกรวดเร็ว และแม่นยำเพิ่มมากขึ้น ประสิทธิภาพของแบบจำลองร้อยละความถูกต้องของ 2 เทคนิค ได้แก่ Decision Tree, Naïve Bayes ค่าความถูกต้องเฉลี่ยเท่ากับร้อยละ 80.67 อยู่ในระดับมาก และเทคนิคที่มีค่าความถูกต้องสูงสุด คือ Decision Tree ร้อยละ 98.58 รองลงมา คือ เทคนิค Naïve Bayes ร้อยละ 62.75

1. วัตถุประสงค์การวิจัย

- 1.1 เพื่อศึกษาและพัฒนาแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก โดยใช้เทคนิคเหมืองข้อมูล
- 1.2 เพื่อประเมินประสิทธิภาพการทำนายผลและวางแผนสำหรับนำไปใช้ในการตัดสินใจการแข่งขันฟุตบอลพรีเมียร์ลีก

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

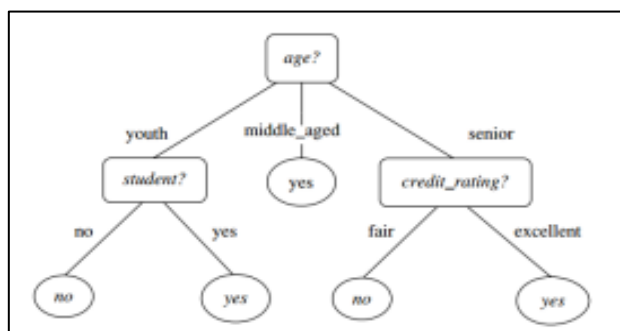
กลุ่มข้อมูลสำหรับการเรียนรู้ (Training Data) หมายถึง การนำเอากลุ่มข้อมูลตัวอย่าง (Training set) มาทำการวิเคราะห์โดยใช้เทคนิคของการจัดกลุ่มข้อมูล เพื่อทำการเรียนรู้ และทำการสร้างแบบจำลองที่สามารถอธิบายถึงลักษณะความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ภายในฐานข้อมูล ซึ่งแบบจำลองนี้มีลักษณะของข้อมูลถูกแจกแจงออกเป็นคลาสต่าง ๆ ด้วยการจัดกลุ่มข้อมูลแบบใช้กฎ และคลาส แต่ละคลาสจะมีลักษณะเฉพาะกลุ่มที่สามารถสรุปออกมาเป็นรูปแบบความสัมพันธ์ได้

กลุ่มข้อมูลสำหรับการทดสอบ (Testing Data) หมายถึง การนำข้อมูลทดสอบมาทดสอบแบบจำลองเพื่อดูความถูกต้องของ กฎการจัดกลุ่มข้อมูลที่ถูกสร้างขึ้นตอนการเรียนรู้ เพื่อพิจารณาว่ากฎการจัดกลุ่มที่ถูกสร้างขึ้นมีความเหมาะสมที่จะนำไปใช้กับกลุ่มข้อมูลใหม่ ๆ หรือไม่

การจัดกลุ่มข้อมูล หมายถึง กระบวนการสร้างแบบจำลองเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด เป็นการสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ล่วงหน้า และสามารถพยากรณ์

กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ แบบจำลองที่ได้อาจอยู่ในรูปแบบการตัดสินใจแบบต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ หมายถึง ต้นไม้ที่ใช้ในการสนับสนุนการตัดสินใจ ซึ่งมีลักษณะเป็นโครงสร้างต้นไม้หัวกลับที่มีรากอยู่ด้านบนและใบอยู่ด้านล่างสุด โดยที่ภายในต้นไม้จะประกอบด้วยโหนด (Node) ซึ่งแต่ละโหนดจะแสดงถึงการตัดสินใจบนข้อมูลของคุณสมบัติต่าง ๆ กิ่งของต้นไม้ แสดงถึงค่าหรือผลลัพธ์ที่ได้จากการทดสอบ และใบซึ่งอยู่ด้านล่างสุดของต้นไม้ตัดสินใจจะแสดงถึงกลุ่มของข้อมูล (Class) หรือผลลัพธ์โหนดที่อยู่บนสุดเรียกว่าโหนดราก (Root Node) ดังแสดงภาพที่ 1



ภาพที่ 1 ตัวอย่างต้นไม้ตัดสินใจ [5]

จากภาพที่ 1 แทนกฎการจำแนกประเภท ดังนี้

IF < age >= youth and <student>= no THEN no

IF < age >= youth and <student>= yes THEN yes

IF < age >= middle_aged THEN yes

IF <age>=senior and <credit_rating>= fair

THEN no

IF <age>= senior and credit_rating>= excellent THEN yes

RapidMiner เป็นโปรแกรมที่ออกแบบมาสำหรับวิเคราะห์ข้อมูล การทำเหมืองข้อมูล การเรียนรู้ของเครื่อง การทำเหมืองข้อความ การวิเคราะห์เชิงพยากรณ์ การวิเคราะห์เชิงธุรกิจ ซึ่งเป็นกระบวนการซึ่งทำกับข้อมูลขนาดใหญ่ เพื่อหาแนวทาง รูปแบบ และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ เพื่อให้ได้สารสนเทศที่ผู้ใช้ไม่รู้ออกมา โดยสารสนเทศที่ได้จะมีเหตุผล และสามารถนำไปใช้ประโยชน์ได้ RapidMiner สร้างขึ้นมาเพื่อใช้สำหรับธุรกิจ ช่วยให้ผู้ใช้เรียกดูและสร้างแบบจำลองเพื่อระบุแนวโน้มของข้อมูลได้อย่างง่ายดาย เมื่อผู้ใช้งานต้องการจัดการกับฐานข้อมูลขนาดใหญ่ [4]

รัชพล กลัดชื่น และจรัญ แสนราช [5] กล่าวว่า อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) นาอีฟเบย์เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็นตามทฤษฎีของเบย์ ซึ่งมีอัลกอริทึมไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ หลักการของนาอีฟเบย์ใช้หลักการของความน่าจะเป็น โดยมีสมมติฐานว่าปริมาณของความสนใจขึ้นอยู่กับความน่าจะเป็น ซึ่งเป็นเทคนิคในการแก้ปัญหาแบบจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณลักษณะของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับสมการ

จิราภรณ์ เจริญยิ่ง [1] ได้พัฒนางานวิจัยเกี่ยวกับการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูล โดยใช้ Rapid Miner ซึ่งเป็นการหาความสัมพันธ์ของตัวแปร และเปรียบเทียบประสิทธิภาพของอัลกอริทึม ซึ่ง

เป็นการศึกษาด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิคป่าแห่งการทำนายการเรียนรู้แบบ และ K-NN โดยใช้ข้อมูลของนักเรียนระดับมัธยมศึกษาในโรงเรียนโปรตุเกส ประกอบด้วยข้อมูลด้านผลการเรียน ด้านความเป็นอยู่ และความเชื่อมโยงทางสังคม จำนวน 649 รายการ 31 แอททริบิวต์ ผลการวิจัย พบว่า เทคนิคให้ค่าความถูกต้องมากที่สุด คือ เทคนิคป่าไม้ตัดสินใจเท่ากับ 80.74 โดยแบ่งกลุ่ม 2 กลุ่ม คือ 80:20 ซึ่งมีการปรับตัวแปรที่เกี่ยวข้องเพื่อให้แบบทำนายมีผลลัพธ์ที่ดีที่สุด และองค์ประกอบหลักของข้อมูลส่งผลต่อค่าความถูกต้องของตัวแบบทำนาย บางเทคนิคอาจเหมาะสมกับการใช้ตัวแปรที่มีจำนวนมาก และมีความซับซ้อน จึงส่งผลต่อค่าความถูกต้องให้ลดลงด้วย

สุกัสนรา สมเจตนา และจารี ทองคำ [9] ได้เปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์โฟนของบุตร มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในเหมืองข้อมูลสร้างแบบจำลองจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์โฟนของบุตร โดยใช้ 6 เทคนิค ได้แก่ เทคนิคปรบเปอร์ เทคนิคต้นไม้ตัดสินใจแบบซี 4.5 เทคนิคนาอียูเบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคเคเนียร์เนสเบอร์ และเทคนิคต้นไม้ป่าสุ่ม มาสร้างแบบจำลองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์โฟนของบุตร โดยข้อมูลถูกรวบรวมจากเครือข่ายสังคมออนไลน์ผ่านเว็บไซต์พันทิป และเฟซบุ๊ก จำนวน 1,925 ข้อความ ผลการทดลอง พบว่า เทคนิคป่าสุ่ม เป็นเทคนิคที่ดีที่สุดในการวิเคราะห์ความคิดเห็นสำหรับข้อมูลชุดนี้ โดยให้ค่าความแม่นยำ 89.62% ค่าความระลึก 78.38% และค่าความถ่วงดุล 83.55% เพราะเทคนิคนี้ได้ทำนายผลอย่างแม่นยำ โดยไม่มีปัญหาเรื่อง overfitting

สมศักดิ์ ศรีสุวรรณ และสมัย ศรีสว [7] ได้วิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ ซึ่งแยกความคิดเห็นเชิงบวก เชิงลบ และได้ทำการเปรียบเทียบประสิทธิภาพจากค่าความถูกต้องด้วยอัลกอริทึมนาอียูเบย์ อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจ ผลการศึกษา พบว่า อัลกอริทึมนาอียูเบย์ ให้ค่าความถูกต้อง 87.97% อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด ให้ค่าความถูกต้อง 83.80% และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจ ให้ค่าความถูกต้อง 79.89% ผลการวิจัย พบว่า ระบบสามารถสกัดคำแยกความคิดเห็นเชิงบวกและเชิงลบตามเวลาจริงทำงานได้อย่างมีประสิทธิภาพ ผลการเปรียบเทียบพบว่า เทคนิควิธีนาอียูเบย์ มีประสิทธิภาพสูงสุด ค่าเฉลี่ยความถูกต้อง ร้อยละ 87.97 และโมเดลเทคนิควิธีนาอียูเบย์ วิธีการหาเพื่อนบ้านใกล้ที่สุด และวิธีการเรียนรู้ของต้นไม้ตัดสินใจ ผลที่ได้โมเดลที่สร้างด้วยเทคนิควิธีนาอียูเบย์ มีประสิทธิภาพสูงสุดค่าเฉลี่ยความถูกต้องร้อยละ 82.97

สัณญา พันธุ์แพง [8] ได้ศึกษาประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลสำหรับการพยากรณ์การศึกษาต่อ นักศึกษาใหม่ระดับปริญญาตรี โดยใช้เทคนิคการทำเหมืองข้อมูล ใช้เทคนิคการทำเหมืองข้อมูลแบบจำลองการหาความสัมพันธ์ ด้วยวิธี Apriori เพื่อใช้ในการค้นหากฎความสัมพันธ์ของพฤติกรรมทางเลือกสมัคร ในแต่ละสาขาวิชาของผู้สมัครในแต่ละกลุ่มทั้งแบบทั่วไปและโควตา ผลการวิเคราะห์เป็นการประเมินผลโมเดลที่เป็นกฎความสัมพันธ์ พิจารณาค่าสนับสนุน ค่าความเชื่อมั่น และค่าความสอดคล้อง จากข้อมูลกฎความสัมพันธ์ที่มีทั้งหมด 340 กฎ พิจารณาจากค่าสนับสนุนเท่ากับ 0.01 ค่าความเชื่อมั่น 95% ค่าความสอดคล้องที่มีมากกว่า 1 ได้ผลตามกฎความสัมพันธ์ 89 กฎ

พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตย์กุล [3] ได้ศึกษาเทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ จำนวนข้อมูล 370 เรคคอร์ด และข้อมูล 4 คลาส ได้แก่ คลาส 0 ยังไม่พบอาการผิดปกติ 200 เรคคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 115 เรคคอร์ด คลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง 39 เรคคอร์ด และคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 16 เรคคอร์ด สำหรับการวินิจฉัยการแพทย์ แบ่งชุดข้อมูลการสอนและชุดข้อมูลทดสอบ จำแนกข้อมูลด้วย multi-class imbalanced data classification ด้วยวิธี one-vs-one และ one-vs-all และเทคนิค Gentleboost ผลการทดสอบประสิทธิภาพ พบว่า ค่าความถูกต้อง 97.31% และทดสอบตัวแบบกับชุดข้อมูลจริงที่ไม่สมดุลจากโรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหัวคู อำเภอนาทวี จ.สงขลา จังหวัดนครศรีธรรมราช

ศุภามณ จันทรสกุล [6] ได้ศึกษาเทคนิคเหมืองข้อมูลในการวิเคราะห์ข้อมูลทางการแพทย์ โดย การวิเคราะห์ข้อมูลแบ่งสองกลุ่มใหญ่ คือ 1) การวิเคราะห์เหมืองข้อมูลเพื่อการทำนายมีการเรียนรู้แบบมีการสอน

และ 2) การวิเคราะห์เหมืองข้อมูลเพื่อการอธิบายการเรียนรู้แบบไม่มีการสอน เทคนิคเหมืองข้อมูล 3 เทคนิค ที่ได้รับความนิยม ได้แก่ การจำแนกประเภท (ต้นไม้ตัดสินใจ และเครือข่ายประสาท) การจัดกลุ่มข้อมูล และการค้นหาความสัมพันธ์ของข้อมูล ผลการวิเคราะห์ที่ได้จากเหมืองข้อมูลสร้างองค์ความรู้ และเป็นข้อค้นพบที่เป็นประโยชน์ในการนำไปใช้เพื่อการตัดสินใจ

วิธีดำเนินการวิจัย

1. ขั้นตอนการดำเนินการวิจัยประกอบด้วย 6 ขั้นตอน ดังนี้

กระบวนการ CRISP-DM ประกอบด้วย 6 ขั้นตอน ดังนี้

1.1 การทำความเข้าใจทางธุรกิจ (Business Understanding) เป็นขั้นตอนแรกในกระบวนการที่ต้องทำความเข้าใจปัญหาและแปลงข้อมูลให้อยู่ในรูปโจทย์การวิเคราะห์ข้อมูลทางตาต้าไมนิ่ง และวางแผนดำเนินการโดยใช้เทคนิคในการวิเคราะห์ข้อมูล เพื่อทำการคัดเลือก และจัดกลุ่มข้อมูล

1.2 การทำความเข้าใจข้อมูล (Data Understanding) ทำการเก็บรวบรวมข้อมูล และตรวจสอบข้อมูลที่ได้รวบรวมเพื่อดูความถูกต้องและเลือกข้อมูลหรือแอททริบิวต์ที่สำคัญมาวิเคราะห์ข้อมูล

1.3 การเตรียมข้อมูล (Data Preparation) แปลงข้อมูลที่เก็บรวบรวมมา ให้กลายเป็นข้อมูลที่นำไปวิเคราะห์ และแปลงข้อมูลให้อยู่ในช่วงสเกลเดียวกัน หรือเติมข้อมูลขนาดที่ขาดหายไป เมื่อได้ข้อมูลแล้วทำการเตรียมข้อมูลเพื่อที่จะนำไปคัดกรองภายใต้หลักการทำงานของเหมืองข้อมูล ซึ่งแบ่งเป็นขั้นตอนต่าง ๆ คือ การทำข้อมูลให้สมบูรณ์ การคัดเลือกข้อมูล การปรับเปลี่ยนรูปแบบข้อมูล ให้มีความเหมาะสม

ขั้นตอนการคัดกรองแอททริบิวต์ แยกประเภทข้อมูลในการทำงาน คือ การใช้เทคนิคการจำแนกข้อมูลและสร้างความสัมพันธ์ของแอททริบิวต์ โดยมีข้อมูลสำหรับทดสอบ ดังตารางที่ 1

ตารางที่ 1 แอททริบิวต์ที่ใช้ในการทดสอบ

ที่	แอททริบิวต์	ชนิดข้อมูล
1	Date	วันที่
2	HomeTeam	ข้อความ
3	AwayTeam	ข้อความ
4	FTHG, FTAG, HTHG, HTAG	ตัวเลข
5	FTR, HTR	ข้อความ
6	Referee	ข้อความ
7	HS, AS, HST, AST, HF, AF, HC, AC, HY, AY, HR, AR	ตัวเลข
8	B365H, B365D, B365A, BWH, BWD, BWA, GBH, GBD, GBA, IWH, IWD, IWA, LBH, LBD, LBA, PSH, PSD, PSA, WHH, WHD, WHA, SJH, SJD, SJA, VCH, VCD, VCA, BSH, BSD, BSA	ตัวเลข
9	Bb1X2, BbMxH, BbAvH, BbMxD, BbAvD, BbMxA, BbAvA, BbOU, BbMx>2.5, BbAv>2.5, BbMx<2.5, BbAv<2.5, BbAH, BbAHh, BbMxAHH, BbAvAHH, BbMxAHA, BbAvAHA, PSCH, PSCD, PSCA	ตัวเลข

จากตารางที่ 1 สามารถอธิบายรายละเอียดของแอททริบิวต์ ได้ดังนี้

Data displays several attributes

Div = League Division (การแบ่งทีมลีก)

Date = Match Date (dd/mm/yy) (วันที่แข่งขัน)

HomeTeam = Home Team (ทีมเจ้าบ้าน)

AwayTeam = Away Team (ทุกทีม)

FTHG = Full Time Home Team Goals (ประตูทีมเหย้าเต็มเวลา)

FTAG = Full Time Away Team Goals (ประตูทีมเยือนเต็มเวลา)

FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win) (รายงานผลเต็มเวลา)

HTHG = Half Time Home Team Goals (ประตูทีมเจ้าบ้านครึ่งแรก)

HTAG = Half Time Away Team Goals (ประตูของทีมเยือนครึ่งแรก)

HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win) (ผลครึ่งแรก)

Match Statistics (where available)

HS = Home Team Shots (ช็อตทีมเหย้า)

AS = Away Team Shots (ช็อตทีมเยือน)

HST = Home Team Shots on Target (ทีมเหย้ายิงตรงเป้าหมาย)

AST = Away Team Shots on Target (ทีมเยือนยิงเข้ากรอบ)

HC = Home Team Corners (ลูกเตะมุม)

AC = Away Team Corners (ลูกเตะมุม)

HF = Home Team Fouls Committed (เจ้าบ้านทำผิดกติกา)

AF = Away Team Fouls Committed (การฟาล์วของทีมเยือน)

HY = Home Team Yellow Cards (เจ้าบ้านใบเหลือง)

AY = Away Team Yellow Cards (ใบเหลืองทีมเยือน)

HR = Home Team Red Cards (เจ้าบ้านใบแดง)

AR = Away Team Red Cards (ใบแดงทีมเยือน)

1.4 แบบจำลองข้อมูล (Modeling) เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางดาต้าไมนิง เช่น การจำแนกประเภทข้อมูลหรือการแบ่งกลุ่มข้อมูล โดยใช้ขั้นตอนการหากฎจำแนก (Classification Rules) การจำแนกประเภทข้อมูล เป็นเทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่าง ๆ โดยทำการสำรวจรายการในฐานข้อมูล เพื่อแยกแยะให้อยู่ในหมวดที่ได้กำหนดไว้ล่วงหน้า เทคนิคที่ใช้สำหรับจำแนกประเภทข้อมูล ได้แก่ ต้นไม้ตัดสินใจ ใช้โปรแกรม RapidMiner สำหรับเชื่อมความสัมพันธ์ของข้อมูลและใช้อัลกอริทึมในการจำแนกกฎข้อมูล

1.5 การประเมินผล (Evaluation) วิเคราะห์ข้อมูลด้วยเทคนิคดาต้าไมนิงแบบจัดกลุ่มข้อมูลและทำนายผล โดยทดสอบประสิทธิภาพของโมเดล คือ Cross-validation test

1.6 การนำไปใช้งาน (Deployment) การนำองค์ความรู้ไปใช้จริง เช่น การสร้างรายงานเพื่อดูผลให้อ่านแล้วเข้าใจได้รวดเร็ว เป็นต้น

2. เครื่องมือการวิจัย

แบบบันทึกรายการข้อมูลผลฟุตบอลพรีเมียร์ลีกจำนวน 3992 เรคคอร์ด

3. กลุ่มตัวอย่าง

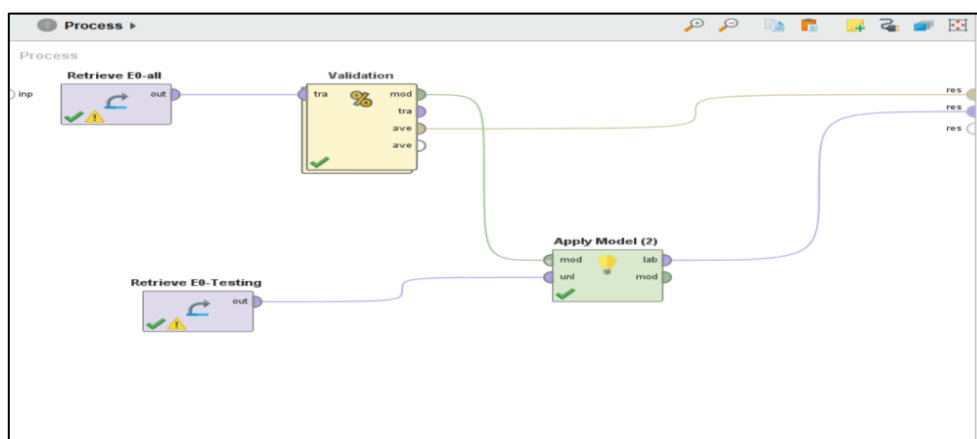
กลุ่มตัวอย่าง คือ ผลฟุตบอลพรีเมียร์ลีก ระหว่างปี ค.ศ. 2012-2022 ได้มาจากการเลือกแบบเจาะจง (Purposive Sampling) จำนวน 3992 เรคคอร์ด ข้อมูลสำหรับชุดฝึกสอน (Training) และ ชุดทดสอบ (Testing) แบ่งอัตราส่วนร้อยละ 70 : 30

4. สถิติที่ใช้ในการวิจัย ได้แก่ การหาค่าร้อยละ ค่าเฉลี่ย โดยนำผลที่ได้จากการนำเข้าสู่ข้อมูลโปรแกรม RapidMiner ที่มีการหาความสัมพันธ์ของข้อมูลฟุตบอลพรีเมียร์ลีก

ผลการวิจัย

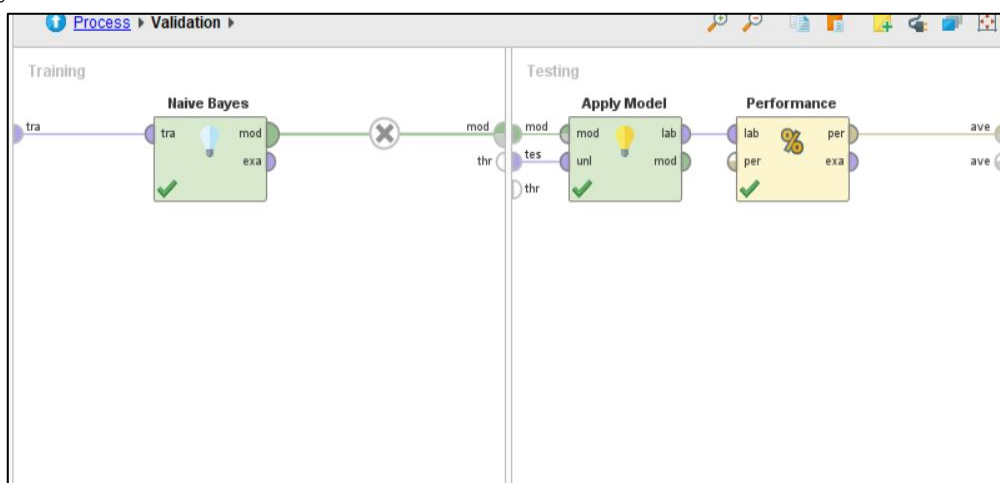
1. ผลการพัฒนาแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก

จากการนำข้อมูลไปวิเคราะห์และทำนายผลฟุตบอลพรีเมียร์ลีก โดยใช้เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคจำแนกประเภทข้อมูลแบบ Naïve bays มีผลการวิจัย ดังนี้

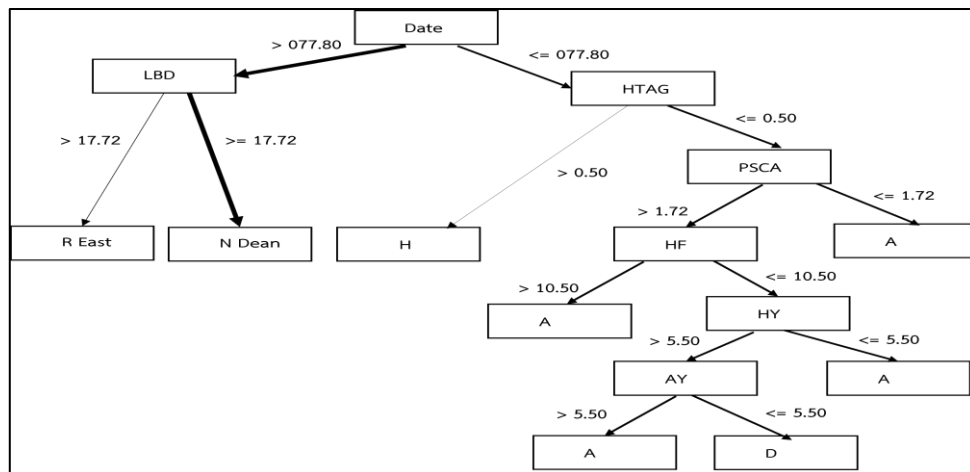


ภาพที่ 2 ผลการพัฒนาแบบต้นไม้ตัดสินใจ

จากภาพที่ 2 เป็นการเชื่อมต่อโดยใช้โปรแกรม RapidMiner โดยนำเข้าสู่ข้อมูลจากไฟล์ Excel และใช้วิธีการตรวจสอบค่าความผิดพลาดด้วย Validation เชื่อมต่อกับ Apply Model โดยใช้ประเภทอัลกอริทึมแบบ Naive bays



ภาพที่ 3 ผลการพัฒนาแบบจำลองการจำแนกประเภทข้อมูล (Naïve bays)



ภาพที่ 4 ผลการทำนายข้อมูลแบบต้นไม้ตัดสินใจ

จากภาพที่ 2-3 เป็นการเชื่อมโยงแผนผังข้อมูลด้วยโปรแกรม RapidMiner studio และภาพที่ 4 แสดงแผนภูมิต้นไม้ที่มีผลต่อการทำนายผลฟุตบอลพรีเมียร์ลีก ซึ่งแอททริบิวต์ที่มีผลสูงสุด ได้แก่ Date ถ้าวันที่มากกว่า 77.80 ครั้ง จะเกี่ยวข้องกับผู้เล่นฟุตบอลที่ชื่อว่า M Dean และ R East วันที่น้อยกว่าเท่ากับ 77.80 จะมีผลต่อประตูทีมเยือนครั้งแรก หากทีม PSCA มากกว่า 1.72 ส่งผลให้เจ้าบ้านทำผิดกติกา (HF) และถ้าน้อยกว่าหรือเท่ากับ 1.72 ส่งผลให้ชนะเสมอ หากมากกว่า 10.50 ส่งผลให้ชนะเสมอ (A) และเจ้าบ้านทำผิดกติกา (HF) น้อยกว่าเท่ากับ 10.50 ส่งผลให้เจ้าบ้านโดนใบเหลือง เมื่อโดนใบเหลืองมากกว่า 5.50 ก็จะถูกใบเหลืองตลอด (AY) หากน้อยกว่าเท่ากับ 5.50 จะถูกถอดถอนจากครั้งแรก ดังนั้นวันที่จัดฟุตบอลมีผลต่อช่วงระยะเวลาในการแข่งขัน รวมถึงผู้ตัดสินและผลแพ้ชนะในแต่ละครั้ง

2. ผลการประเมินประสิทธิภาพแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีกด้วยเทคนิคดาต้าไมนิง

ผลการประเมินประสิทธิภาพแบบจำลอง 2 เทคนิค คือ Decision Tree, Naïve Bayes พบว่า การหาค่าร้อยละ ค่าความถูกต้อง ค่าทำนายถูก ค่าทำนายผิด ดังตารางที่ 2

ตารางที่ 2 ผลการหาค่าประสิทธิภาพแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก

เทคนิค	precision	recall	ค่าความถูกต้อง
Decision Tree	98.88	98.27	98.58
Naïve Bayes	62.32	63.18	62.75
โดยรวม	80.60	80.73	80.67

จากตารางที่ 2 ผลการหาค่าร้อยละความถูกต้องของ 2 เทคนิค ได้แก่ Decision Tree, Naïve Bayes ค่าความถูกต้องเฉลี่ยเท่ากับร้อยละ 80.67 อยู่ในระดับมาก และเทคนิคที่มีค่าความถูกต้องสูงสุด คือ Decision Tree ร้อยละ 98.58 รองลงมา คือ เทคนิค Naïve Bayes ค่าความถูกต้องร้อยละ 62.75 จากผลการประเมินประสิทธิภาพแบบจำลองสามารถให้ผู้ใช้งานนำไปใช้ตัดสินใจสำหรับการแข่งขันฟุตบอลพรีเมียร์ลีกได้

อภิปรายผลการวิจัย

1. ผลการพัฒนาแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก พบว่า แอททริบิวต์ที่มีผลสูงสุด ได้แก่ Date ถ้าวันทันทีมากกว่า 77.80 ครั้ง จะเกี่ยวข้องกับผู้ตัดสินผลฟุตบอลที่ชื่อว่า M Dean และ R East วันที่น้อยกว่าเท่ากับ 77.80 จะมีผลต่อประตูที่เียนครั้งแรก หากทีม PSCA มากกว่า 1.72 ส่งผลให้เจ้าบ้านทำผิดกติกา (HF) และถ้าน้อยกว่าหรือเท่ากับ 1.72 ส่งผลให้ชนะเสมอ หากมากกว่า 10.50 ส่งผลให้ชนะเสมอ (A) และเจ้าบ้านทำผิดกติกา (HF) น้อยกว่าเท่ากับ 10.50 ส่งผลให้เจ้าบ้านโดนใบเหลือง เมื่อโดนใบเหลืองมากกว่า 5.50 ก็จะโดนใบเหลืองตลอด (AY) หากน้อยกว่าเท่ากับ 5.50 จะถูกถอดถอนจากครั้งแรก ดังนั้นวันที่จัดฟุตบอลมีผลต่อช่วงระยะเวลาในการแข่งขัน รวมถึงผู้ตัดสิน และผลแพ้ชนะในแต่ละครั้ง สอดคล้องกับงานวิจัยของ สุภัสสร่า สมเจตนา และจารี ทองคำ [9] ได้เปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์โฟนของบุตร มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในเหมืองข้อมูลสร้างแบบจำลองจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์โฟนของบุตร โดยใช้ 6 เทคนิค ได้แก่ เทคนิคปรบเปอร์ เทคนิคต้นไม้ตัดสินใจแบบซี 4.5 เทคนิคนาอ์เบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคเคเนียร์สเนเบอร์ และเทคนิคต้นไม้ป่าสุ่ม มาสร้างแบบจำลองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์โฟนของบุตร โดยข้อมูลถูกรวบรวมจากเครือข่ายสังคมออนไลน์ผ่านเว็บไซต์พันทิปและเฟซบุ๊ก จำนวน 1,925 ข้อความ ผลการทดลอง พบว่าเทคนิคป่าสุ่ม เป็นเทคนิคที่ดีที่สุดในการวิเคราะห์ความคิดเห็นสำหรับข้อมูลชุดนี้ โดยให้ค่าความแม่นยำ 89.62% ค่าความระลึก 78.38% และค่าความถ่วงดุล 83.55% เพราะเทคนิคนี้ได้ทำนายผลอย่างแม่นยำ โดยไม่มีปัญหาเรื่อง overfitting

2. ผลการประเมินประสิทธิภาพแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก การประเมินประสิทธิภาพแบบจำลอง 2 เทคนิค คือ Decision Tree, Naïve Bayes พบว่า การหาค่าร้อยละ ค่าความถูกต้อง ค่าทำนายถูก ค่าทำนายผิด การหาค่าร้อยละความถูกต้องของ 2 เทคนิค ได้แก่ Decision Tree, Naïve Bayes ค่าความถูกต้องเฉลี่ยเท่ากับร้อยละ 80.67 อยู่ในระดับมาก และเทคนิคที่มีค่าความถูกต้องสูงสุด คือ Decision Tree ร้อยละ 98.58 รองลงมา คือ เทคนิค Naïve Bayes ค่าความถูกต้องร้อยละ 62.75 สอดคล้องกับงานวิจัยของ ปริญญา ชินจ่อหอ สายสุนีย์ จัปโจร และเบญจกัศ จงหมื่นไวย [2] ได้พัฒนาแบบจำลองเทคโนโลยีสารสนเทศในการคาดการณ์อาชีพอนาคตของบัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏนครราชสีมา พบว่า ผลการพัฒนาแบบจำลองมีประสิทธิภาพในการคาดการณ์ข้อมูลภาพรวมเฉลี่ยร้อยละ 73.86 ซึ่งอยู่ในระดับสูง

ข้อเสนอแนะ

ผลการทดลองและสรุปผลการพัฒนาแบบจำลองการทำนายผลฟุตบอลพรีเมียร์ลีก ควรมีการเก็บรวบรวมข้อมูลเพื่อใช้ในการวิเคราะห์ผลที่มีจำนวนเพิ่มมากขึ้น และให้ได้ผลที่ถูกต้องและมีประสิทธิภาพสูงสุด

เอกสารอ้างอิง

- [1] จิราภรณ์ เจริญยิ่ง. (2563). การพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูลโดยใช้ Rapid Miner. (สารนิพนธ์ตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ). กรุงเทพฯ: บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ.
- [2] ปริญญา ชินจ่อหอ, สายสุนีย์ จัปโจร และเบญจกัศ จงหมื่นไวย. (2564). การพัฒนาแบบจำลองเทคโนโลยีสารสนเทศในการคาดการณ์ อาชีพอนาคตของบัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏนครราชสีมา. *วารสารวิชาการการจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม*, 8(2), 20-32.
- [3] พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสดีกุล. (2562). เทคนิคการจำแนกข้อมูลที่เหมาะสมสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ. *วารสารวิทยาศาสตร์และเทคโนโลยี*, 27(6), 1164-1178.
- [4] มณฑิชา เกตุชนะ. (2563). เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลขององค์กรธุรกิจ. *วารสารธุรกิจและสังคมศาสตร์ มหาวิทยาลัยรามคำแหง*, 3(3), 14-30.
- [5] รัชพล กลัดชื่น และจรัญ แสนราช. (2561). การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสม เพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาในระดับอาชีวศึกษา. *วารสารวิจัยมหาวิทยาลัยเทคโนโลยีธบุรี*, 17(1), 1-10.

- [6] ศุภามณ จันทร์สกุล. (2561). เทคนิคเหมืองข้อมูลในการวิเคราะห์ข้อมูลทางการแพทย์. *วารสารวิชาการ มหาวิทยาลัยอีสเทิร์นเอเชีย ฉบับวิทยาศาสตร์และเทคโนโลยี*, 12(2), 83-96.
- [7] สมศักดิ์ ศรีสุวรรณ และสมัย ศรีสวย. (2563). การวิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ. *วารสารวิชาการประยุกต์ใช้เทคโนโลยีสารสนเทศ*, 6(2), 95-104.
- [8] สันญา พันธุ์แพง. (2563). การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลสำหรับการพยากรณ์การศึกษาค้นคว้าใหม่ ระดับปริญญาตรี ในมหาวิทยาลัยราชภัฏเชียงใหม่. เชียงใหม่: มหาวิทยาลัยราชภัฏเชียงใหม่.
- [9] สุภัทสร สมนัดนา และจาร์ ทองคำ. (2564). การเปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ตโฟนของบุตร. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอุบลราชธานี*, 23(1), 21-30.