Application of Data Mining Processes with Machine Learning Techniques to Prevent Corruption in Public Procurement

การประยุกต์ใช้กระบวนการเหมืองข้อมูล (Data mining) ด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine learning) เพื่อการป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ

Weerapun Panich[1] and Punniththa Mrazek[2]

วีระพันธ์ พานิชย์[1] และปุณณิฐฐา มาเชค[2]

[1]Faculty of Education, Burapha University

[1]คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

[2]Faculty of Education, Burapha University

[2]คณะศึกษาศาสตร์ มหาวิทยาลัยบูรพา

nookrai@go.buu.ac.th

Abstract

Corruption in Thai society has a huge negative impact in the country and is a major obstacle to economic, social and political development in all dimensions. The current pattern of corruption has shifted to more complex forms. A lack of transparency in public procurement leads to numerous corruptions. Preventing corruption is a national strategy and agenda because corruption is a major problem in Thailand that must be addressed urgently. This article aims to provide an idea of how to use data mining processes with machine learning techniques to prevent corruption in public procurement. The author studied the principle, theories and research papers from foreign countries, and synthesized information to be guidelines for practical use.

Application of data mining process with machine learning techniques is using algorithms and decision trees to analyze corruption in public procurement by using the completed procurement project data as a training set and a testing set. The steps of the data mining process consist of 1) Selection 2) Preprocessing 3) Transformation 4) Data Mining and 5) Interpretation and Evaluation. These are steps of using data mining for translating into the knowledge and presenting them to relevant parties to support decision-making.

The analysis for predicting corruption in new procurement projects, the computer programs will take the former information to compare with new information. In the case where the program analyzes and predicts that new procurement projects are among the corrupt, the project owner agency can apply this information in the corruption prevention plan. Or this information can be used for relevant agencies to set up surveillance measures and to check each step of the process to ensure fairness and transparency in accordance with the objectives of the use of state budget.

Keywords: data mining, machine learning, decision tree, corruption in public procurement

บทคัดย่อ

การทุจริตในสังคมไทยส่งผลเสียต่อประเทศอย่างมหาศาลและเป็นอุปสรรคสำคัญต่อการพัฒนาเศรษฐกิจ สังคม การเมือง ในทุกมิติ รูปแบบการทุจริตในปัจจุบันได้ปรับเปลี่ยนเป็นการทุจริตที่ซับซ้อนมากขึ้น การจัดซื้อจัดจ้างภาครัฐที่ไม่โปร่งใส จะนำไปสู่การทุจริตคอร์รัปชันต่าง ๆ มากมาย การป้องกันการทุจริตคอร์รัปชันถือเป็นยุทธศาสตร์และวาระแห่งชาติ เนื่องจากการทุจริตคอร์รัปชันเป็นปัญหาที่สำคัญของประเทศไทยที่ต้องได้รับการแก้ไขอย่างเร่งด่วน บทความนี้มีวัตถุประสงค์เพื่อนำเสนอแนวคิดการประยุกต์ใช้กระบวนการเมืองข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง เพื่อการป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ โดยผู้เขียนได้ศึกษาหลักการ ทฤษฎี ตัวอย่างงานงานวิจัยในต่างประเทศ และดำเนินการสังเคราะห์เรียบเรียงขึ้นเพื่อเป็นแนวทางในการนำไปปฏิบัติจริงต่อไป

การประยุกต์ใช้กระบวนการเหมืองข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้อัลกอริทึม ต้นไม้ตัดสินใจ มาวิเคราะห์การทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐโดยใช้ข้อมูลโครงการจัดซื้อจัดจ้างที่ดำเนินการเสร็จสมบูรณ์แล้วมาเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ โดยมีขั้นตอนของกระบวนการเหมืองข้อมูล ประกอบด้วย 1) การเลือกข้อมูล 2) การเตรียมข้อมูล 3) การเปลี่ยนรูปข้อมูล 4) การทำเหมืองข้อมูล และ 5) การแปลผลและการประเมินผล เป็นขั้นตอนที่นำรูปแบบที่ได้จากการทำเหมืองข้อมูล แปลให้อยู่ในรูปของความรู้ และนำเสนอต่อผู้ที่เกี่ยวข้องเพื่อใช้สนับสนุนการตัดสินใจ

การวิเคราะห์เพื่อทำนายการทุจริตโครงการจัดซื้อจัดจ้างใหม่ โปรแกรมคอมพิวเตอร์จะใช้ความรู้จากการเรียนรู้ข้อมูลเก่ามาเปรียบเทียบกับข้อมูลใหม่ ในกรณีที่โปรแกรมวิเคราะห์และทำนายว่าโครงการจัดซื้อจัดจ้างใหม่อยู่ในกลุ่มทุจริต หน่วยงานเจ้าของโครงการสามารถนำมาเป็นข้อมูลในวางแผนป้องกันการทุจริต หรือเป็นข้อมูลให้หน่วยงานที่เกี่ยวข้องวางมาตรการเฝ้าระวังและตรวจสอบการดำเนินการแต่ละขั้นตอนให้เที่ยงตรงโปร่งใสตามเจตนารมณ์ของการใช้งบประมาณแผ่นดินต่อไป

คำสำคัญ: เหมืองข้อมูล การเรียนรู้ของเครื่อง ต้นไม้ตัดสินใจ การทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ

1. บทนำ

การทุจริตในสังคมไทยระหว่างช่วงเวลากว่าทศวรรษ ส่งผลเสียต่อประเทศอย่างมหาศาลและเป็นอุปสรรคสำคัญต่อการพัฒนาเศรษฐกิจ สังคม การเมือง ในทุกมิติ รูปแบบการทุจริตจากเดิมที่เป็นทุจริตทางตรงไม่ซับซ้อน อาทิ การรับสินบน การจัดซื้อจัดจ้าง ในปัจจุบันได้ปรับเปลี่ยนเป็นการทุจริตที่ซับซ้อนมากขึ้น ตัวอย่างเช่น การทุจริตเชิงนโยบาย การทุจริตข้ามแดนข้ามชาติ ซึ่งเชื่อมโยงไปสู่อาชญากรรมอื่น ๆ มากมายและส่งผลกระทบทางลบในวงกว้าง (ศูนย์ปฏิบัติการต่อต้านการทุจริต กระทรวงศึกษาธิการ, 2560) ปริมาณเงินงบประมาณของหน่วยงานราชการในปีหนึ่ง ๆ สูงมาก โดยมีนักวิชาการหลายท่านเคยประเมินไว้ว่าอัตราการทุจริตจากเงินงบประมาณจะมีอัตราตั้งแต่ 10 - 20% (ธราพงศ์ ลิ้มสุทธิวันภูมิ และ ตวงทอง สินชัย, 2563) สภาพปัญหาการจัดซื้อจัดจ้างในหน่วยงานภาครัฐ ที่ก่อให้เกิดปัญหาการทุจริตคอร์รัปชัน ซึ่งประเทศไทยกำลังประสบปัญหาการทุจริตการจัดซื้อจัดจ้างที่รุนแรง โดยรูปแบบการทุจริตส่วนใหญ่จะเป็นในลักษณะของการรับสินบน การฮั้วประมูล การใช้อำนาจหน้าที่กระทำผิดโดยมิชอบ รวมไปถึงเรื่องผลประโยชน์ทับซ้อน และการเอื้อผลประโยชน์ต่อคู่สัญญา สาเหตุส่วนใหญ่มาจากระบบอุปถัมภ์และเป็นระบบที่มีมายาวนานทุกยุคทุกสมัยทำให้ระบบนี้ฝังตัวเป็นรากลึกอยู่ในสังคมไทยมาเป็นระยะเวลานานจนถึงปัจจุบัน (สิรินทรา เสียบไธสง และ สมศักดิ์ อมรสิริพงศ์, 2562) รูปแบบการจัดซื้อจัดจ้างของหน่วยงานภาครัฐต้องปฏิบัติตามพระราชบัญญัติการจัดซื้อจัดจ้างและการบริหารงานพัสดุภาครัฐ พ.ศ. 2560 โดยได้กำหนดกระบวนการจัดซื้อจัดจ้างต้องปฏิบัติ 8 ขั้นตอน คือ 1) การกำหนดความต้องการ 2) การตั้งงบประมาณ 3) การกำหนด Term of Reference (TOR) 4) การออกประกาศประกวดราคา 5) การเสนอราคา 6) การพิจารณาและตัดสิน 7) การส่งมอบและติดตั้ง 8) การจ่ายเงิน (ธราพงศ์ ลิ้มสุทธิวันภูมิ และ ตวงทอง สินชัย, 2563) ซึ่งการจัดซื้อจัดจ้างภาครัฐที่ไม่โปร่งใส จะนำไปสู่การทุจริตคอร์รัปชันต่าง ๆ มากมาย เช่น การติดสินบน การสมรู้ร่วมคิด การฮั้วประมูล การเอื้อผลประโยชน์ ไปจนถึงการมีผลประโยชน์ทับซ้อน ซึ่งเกิดจากผู้มีส่วนเกี่ยวข้องใช้อำนาจหรือตัดสินใจโดยคำนึงถึงผลประโยชน์ส่วนตนมากกว่าผลประโยชน์ส่วนราชการ ไม่ว่าจะเป็นผู้กำหนดนโยบาย เจ้าหน้าที่ของรัฐที่รับผิดชอบในการจัดซื้อจัดจ้าง ผู้เข้าร่วมประมูล คู่สัญญา และผู้มีส่วนร่วมเกี่ยวข้องทั้งหลายในสัญญา

การป้องกันการทุจริตคอร์รัปชันถือเป็นยุทธศาสตร์และวาระแห่งชาติ เนื่องจากการทุจริตคอร์รัปชันเป็นปัญหาที่สำคัญของประเทศไทยที่ต้องได้รับการแก้ไขอย่างเร่งด่วน (กรมสอบสวนคดีพิเศษ, 2562) ซึ่งปัญหาการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐเป็นปัญหาที่เกิดขึ้นอย่างเป็นระบบ และเป็นปัญหาที่เกิดขึ้นซ้ำซาก ดังนั้นในการแก้ปัญหาต้องดำเนินการวิเคราะห์ด้วยวิธีการคิดเชิงระบบ (systems thinking) อย่างละเอียดครอบคลุมถึงพฤติกรรมของปัญหา รูปแบบของปัญหา องค์ประกอบของปัญหา และความสัมพันธ์ขององค์ประกอบของปัญหา การวิเคราะห์ด้วยวิธีการคิดเชิงระบบ เป็นการศึกษาข้อมูลที่เกิดขึ้นในอดีตเพื่อสร้างแบบแผนว่าทำไมถึงเกิดการทุจริตขึ้นได้ ผลการได้แบบแผนการทุจริตที่เกิดขึ้นสามารถนำมาเป็นข้อมูลเพื่อทำนายโครงการจัดซื้อจัดจ้างที่จะเกิดขึ้นใหม่ว่าโครงการมีแนวโน้มเกิดการทุจริต หรือไม่เกิดการทุจริต การที่ทราบว่าโครงการใดมีแนวโน้มจะเกิดการทุจริต จะทำให้หน่วยงานที่เกี่ยวข้องทบทวนโครงการหรือเพิ่มมาตรการตรวจสอบติดตามทุกขั้นตอนส่งผลให้เกิดความโปร่งใสในการจัดซื้อจัดจ้างของหน่วยงานภาครัฐได้ในที่สุด

ในกระบวนการวิเคราะห์ข้อมูลเพื่อหารูปแบบการทุจริตจัดซื้อจัดจ้าง ในปัจจุบันสามารถนำเทคโนโลยีเข้ามาเป็นเครื่องมือเพื่อช่วยวิเคราะห์ทำให้มีผลการวิเคราะห์มีความถูกต้องแม่นยำมากขึ้น เทคโนโลยีดังกล่าว ได้แก่ เทคโนโลยีเหมืองข้อมูล (Data mining) ด้วยความแม่นยำของระบบประมวลผลของคอมพิวเตอร์ที่สามารถวิเคราะห์ข้อมูลเชิงลึกที่ให้ผลลัพธ์การวิเคราะห์มีความถูกต้องและเชื่อถือได้ ทำให้เทคโนโลยีเหมืองข้อมูล

ถูกนำไปใช้อย่างกว้างขวางในหลาย ๆ หน่วยงานทั้งภาครัฐและเอกชน ซึ่งกระบวนเหมืองข้อมูล มีเทคนิคที่นำมาใช้วิเคราะห์หลายวิธีการ ในปัจจุบันเทคนิคการเรียนรู้ของเครื่อง (Machine learning) เป็นเทคนิคที่ได้รับการยอมรับว่ามีประสิทธิภาพในวิเคราะห์การจำแนกข้อมูล และสามารถทำนายผลจากข้อมูลที่มีอยู่ถูกต้องในอัตราสูง เนื่องจากเทคนิคการเรียนรู้ของเครื่อง เรียนรู้จากข้อมูลที่เกิดขึ้นในอดีต และสามารถเรียนรู้จากการรวบรวมข้อมูลเข้าสู่ฐานข้อมูลขนาดใหญ่ (Big data) สำหรับการวิเคราะห์และประมวลผลอย่างต่อเนื่องทำให้การเรียนรู้ของเครื่องวิเคราะห์ได้ถูกต้องมากขึ้น ส่งผลให้สามารถทำนายพฤติการณ์ต่าง ๆ นำไปสู่การตัดสินใจการดำเนินงานตามภารกิจของหน่วยงานอย่างมีประสิทธิภาพ

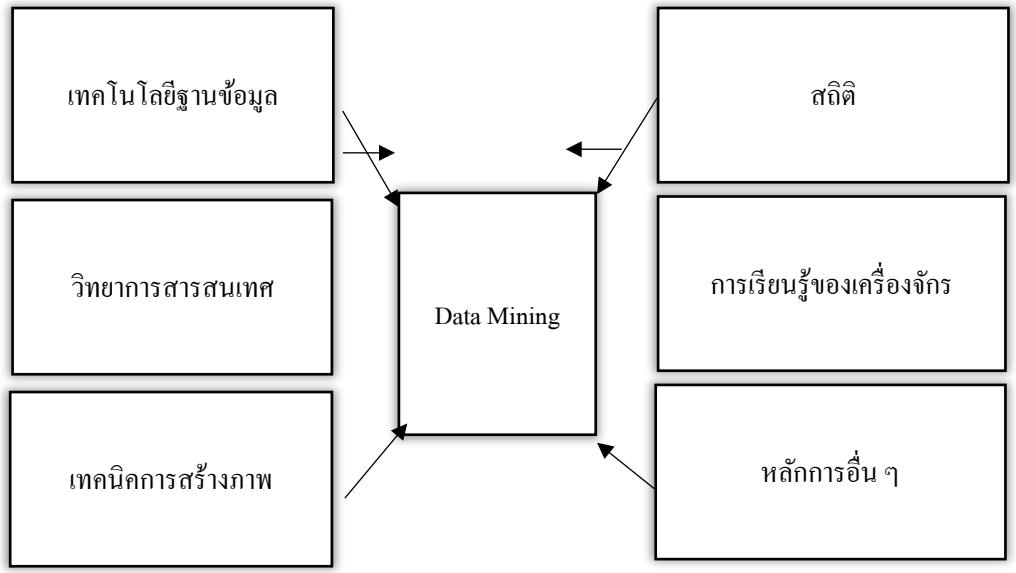
สำหรับบทความนี้มีวัตถุประสงค์เพื่อนำเสนอแนวคิดการป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐโดยการประยุกต์ใช้กระบวนการเมืองข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง เพื่อเป็นข้อมูลนำไปใช้ในการวางแผนเพื่อการป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ เช่น ทำนายแนวโน้มการทุจริตการจัดซื้อจัดจ้างโครงใหม่ที่จะเกิดขึ้น การหาความสัมพันธ์ขององค์ประกอบการทุจริตที่เกิดขึ้น เป็นต้น

การดำเนินการศึกษา ผู้เขียนได้ศึกษาหลักการ ทฤษฎี ประเด็นที่เกี่ยวข้อง ได้แก่ ทฤษฎีเหมืองข้อมูล ทฤษฎีการเรียนรู้ของเครื่อง ทฤษฎีต้นไม้ตัดสินใจ โปรแกรมคอมพิวเตอร์สำหรับการทำเหมืองข้อมูล และตัวอย่างงานวิจัยในต่างประเทศ พร้อมทั้งได้เสนอตัวอย่างการประยุกต์ใช้กระบวนการเหมืองข้อมูล ด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อการป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐเพื่อเป็นแนวทางในการนำไปปฏิบัติจริงต่อไป

2. ทฤษฎีเหมืองข้อมูล (Data Mining)

Data Mining หมายถึง การสกัดหรือวิเคราะห์ค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้หรือกล่าวอีกนัยหนึ่ง คือ ชุดซอฟต์แวร์ (Software) วิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนความต้องการของผู้ใช้ในการค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้ (ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงศึกษาธิการ, 2564) การทำเหมืองข้อมูล (Han & Kamber, 2006) คือ กระบวนการวิเคราะห์ข้อมูลจากมุมมองที่แตกต่างกันและสามารถสรุปผลเป็นข้อมูลที่มีประโยชน์ กระบวนการทำเหมืองข้อมูลใช้หลายหลักการเช่น เทคนิคการเรียนรู้ของเครื่อง สถิติ (Statistics) เทคนิคการสร้างภาพ (Visualization techniques) วิทยาการสารสนเทศ (Information science) เทคโนโลยีฐานข้อมูล (Database technology) และหลักการอื่น ๆ เพื่อค้นพบและนำเสนอความรู้ในรูปแบบที่เข้าใจได้ง่าย โดยอีกความหมายหนึ่งคือกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อสกัดสารสนเทศรวมถึงรูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลขนาดใหญ่นั้น ซึ่งการทำเหมืองข้อมูลเป็นการบรรจบกันของหลายหลักการ ดังภาพที่ 1

ภาพที่ 1 หลักการที่ใช้ในการทำเหมืองข้อมูล

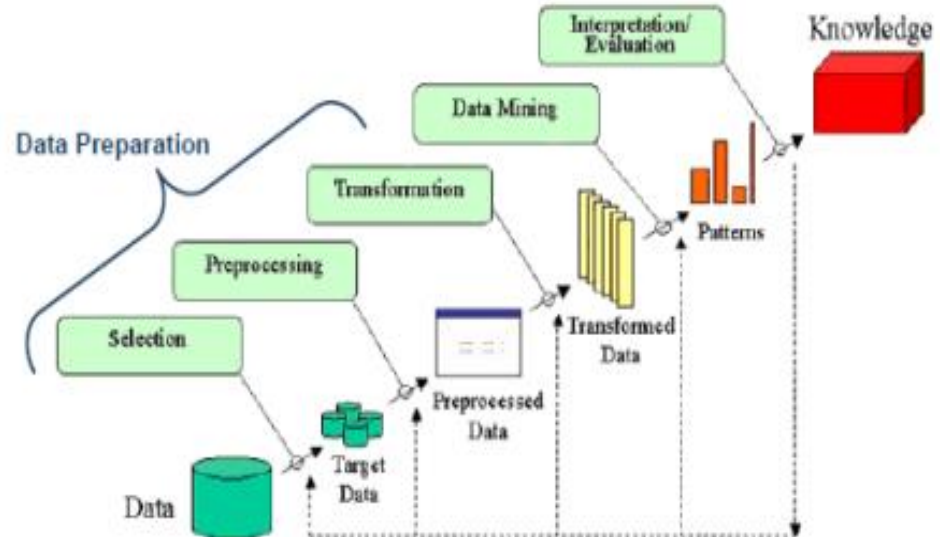


ที่มา: From *Data mining: Concepts and techniques (2nd ed.)* (pp. 29), by โดย J. Han, & M. Kamber, 2006, San Francisco: Morgan Kaufmann.

กระบวนการทำงานของ Knowledge Discovery in Database (KDD) หรือ Data Mining

กระบวนการ Knowledge Discovery in Database (KDD) (Fayyad et al., 1996) หมายถึงกระบวนการในการค้นหาความรู้สารสนเทศจากกลุ่มข้อมูลที่มีจำนวนมาก ซึ่งมีขั้นตอนการทำเหมืองข้อมูลเป็นกระบวนการหลักเพื่อค้นหาลักษณะที่น่าสนใจของข้อมูลเหล่านี้ ทำให้ได้มาซึ่งรูปแบบข้อมูลที่เป็นเหตุเป็นผล เป็นประโยชน์และเข้าใจได้ง่าย ซึ่งการนำเอาเทคนิคการทำเหมืองข้อมูลไปใช้นั้นจะมีวิธีการต่างกัน ขึ้นอยู่กับวัตถุประสงค์ที่จะนำไปใช้งานว่าใช้เพื่อทำอะไร สิ่งที่ต้องการรู้คืออะไร ดังนั้นจึงมีการนำเสนอวิธีการที่หลากหลายสำหรับเป้าหมายที่ต่างกันเพื่อให้ได้ผลลัพธ์ที่เหมาะสมต่อความต้องการหลังจากนำไปใช้งาน โดยขั้นตอนการทำเหมืองข้อมูลสามารถแสดงเป็นรูปได้ดังนี้

ภาพที่ 2 กระบวนการทำเหมืองข้อมูล (Fayyad et al., 1996)



ที่มา: From From Data Mining to Knowledge Discovery in Databases (pp. 41), by U. Fayyad, G. Piatetsky-Shapiro, & P. Shapiro, *AI Magazine*, 17(3).

จากภาพที่ 2 แสดงกระบวนการทำเหมืองข้อมูล 5 ขั้นตอน ดังนี้

1. การเลือกข้อมูล (Selection) เป็นการเลือกหรือแบ่งข้อมูลตามเงื่อนไขที่กำหนด เช่น เลือกพนักงานที่มีรถยนต์เป็นของตนเอง

2. การเตรียมข้อมูล (Preprocessing) เป็นขั้นตอนในการเตรียมข้อมูลโดยการแยกข้อมูลที่ไม่มีค่าข้อมูลที่ทำการบันทึกผิด ข้อมูลที่มีความซ้ำซ้อนหรือไม่สอดคล้องกันออกไปและทำการรวบรวมข้อมูลที่ต้องการซึ่งได้มาจากหลายๆ ฐานข้อมูลเข้าไว้ด้วยกัน

3. การเปลี่ยนรูปข้อมูล (Transformation) เป็นขั้นตอนการเปลี่ยนรูปหรือรวบรวมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการทำเหมืองข้อมูล เช่น ในรูปผลรวม หรือ ผลสรุป

4. การทำเหมืองข้อมูล (Data mining) เป็นขั้นตอนที่สำคัญเพื่อทำการกลั่นกรองรูปแบบของข้อมูล (Data pattern) จากข้อมูลดิบที่มี ในขั้นตอนนี้มีการนำเทคนิคต่าง ๆ เพื่อเข้ามาช่วยในการดึงรูปแบบ (Pattern) ที่ซ่อนอยู่ในข้อมูลออกมา

5. การแปรผลและการประเมินผล (Interpretation evaluation) เป็นขั้นตอนที่นำรูปแบบ (Pattern) ที่ได้จากการทำเหมืองข้อมูล แปลให้อยู่ในรูปของความรู้ (Knowledge) และนำเสนอต่อผู้ใช้ เพื่อใช้สนับสนุนการตัดสินใจ

3. ทฤษฎีการเรียนรู้ของเครื่อง (Machine learning)

Machine learning หรือ การเรียนรู้ของเครื่อง เป็นสาขาหนึ่งของปัญญาประดิษฐ์ที่พัฒนามาจากการศึกษาการรู้จำแบบเกี่ยวข้องกับการศึกษาและการสร้างอัลกอริทึมที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ อัลกอริทึมนั้นจะทำงานโดยอาศัยโมเดลที่สร้างมาจากชุดข้อมูลตัวอย่างเพื่อการทำนายหรือตัดสินใจโดยไม่ต้องเขียนโปรแกรมไว้อย่างชัดแจ้ง การเรียนรู้ของเครื่องต้องอาศัยวิธีการทางสถิติศาสตร์เป็นอย่างมาก โดยเชื่อมโยงกับองค์ความรู้อื่นที่ต้องการนำไปประยุกต์ เช่น การแพทย์ ชีวสารสนเทศศาสตร์ เคมีสารสนเทศศาสตร์ (ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงศึกษาธิการ, 2564) การเรียนรู้ของเครื่องหมายถึง การนำข้อมูลต่าง ๆ มาฝึกให้เครื่องหรือคอมพิวเตอร์ได้เรียนรู้หาคำตอบ การเรียนรู้ หมายถึง การเรียนรู้วิธีหาคำตอบจากข้อมูลฝึกหัด จากนั้นจึงนำวิธีการหาคำตอบที่ได้ไปใช้พยากรณ์คาดการณ์สิ่งที่จะเกิดขึ้นจากข้อมูลชุดใหม่ ต่างจากการทำเหมืองข้อมูลเพราะการทำเหมืองข้อมูลเป็นการวิเคราะห์สืบค้นหาความรู้ที่เป็นประโยชน์และมีความสำคัญที่แฝงอยู่ในฐานข้อมูลขนาดใหญ่ที่ยังไม่ทราบคำตอบหรือความสัมพันธ์ของปัจจัยต่าง ๆ ที่แฝงอยู่ (Kaur et al., 2015) ซึ่งสอดคล้องกับ ปริญญา สงวนสัตย์ (2562, น. 3) ที่ได้กล่าวว่า “Machine learning” หรือ “การเรียนรู้ของเครื่อง” เป็นเครื่องมือหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence) หรือ AI ที่มุ่งเน้นในการใช้ตัวอย่างหรือประสบการณ์เพื่อการเรียนรู้งาน โดยมนุษย์มีส่วนร่วมเพียงการออกแบบระบบเท่านั้น หลังจากนั้นระบบจะสกัดสาระสำคัญจากตัวอย่างเหล่านี้เอง หลักจากเรียนรู้เสร็จสิ้นด้วยตัวอย่างจำนวนหนึ่งอย่างเพียงพอ เครื่องหรือระบบที่เรียนรู้แล้วนี้สามารถนำไปใช้ในการประมวลผลของตัวอย่างใหม่ที่ไม่เคยพบมาก่อนได้อย่างมีประสิทธิภาพ โดยในการฝึกให้คอมพิวเตอร์เรียนรู้จะใช้ข้อมูลในอดีตจำนวนมากนำมาวิเคราะห์ด้วยเทคนิคของเหมืองข้อมูล เพื่อสร้างวิธีการเรียนรู้ที่เป็นต้นแบบใช้พยากรณ์หรือทำนายสิ่งที่อาจจะเกิดในอนาคต ปัจจุบัน Machine learning ถูกแบ่งอย่างกว้างๆ ออกเป็น 3 ประเภท โดย Shalev-Shwartz and Ben-David (2009) ได้อธิบายตามลักษณะของข้อมูลนำเข้า หรือข้อมูลฝึก ได้แก่

การเรียนรู้จากข้อมูลแบบมีโครงสร้าง (Supervised learning) เป็นการเรียนรู้แบบมีผู้สอน หรือจากข้อมูลตัวอย่างในอดีตที่เฉลยผลลัพธ์ที่ควรจะเป็นแสดงเป็นลาเบล (Label) ไว้ นำมาสอน Machine ให้ค้นหาความสัมพันธ์ และสร้างกฎทั่วไปไว้ เพื่อทำนายว่าข้อมูลนำเข้าแบบนี้แล้วจะทำให้ได้ข้อมูลส่งออกแบบใด สำหรับเทคนิควิธีการจำแนกกลุ่ม (Classification) เช่น การวิเคราะห์เครือข่ายประสาทเทียม (Neuron networks) และการวิเคราะห์ต้นไม้ตัดสินใจ

การเรียนรู้จากข้อมูลแบบไม่มีโครงสร้าง (Unsupervised learning) เป็นการเรียนรู้แบบไม่มีผู้สอน หรือไม่มีข้อมูลเฉลยผลลัพธ์ใด ๆ ให้ Machine ได้เรียนรู้ ซึ่ง Machine ต้องหาโครงสร้างของข้อมูลนำเข้าเอง โดยบอกแค่ความต้องการข้อมูลแบบใด เช่นต้องการแบ่งกลุ่มข้อมูล (Clustering) แต่ไม่สามารถระบุผลลัพธ์ที่จะได้ Machine จะทำการแบ่งกลุ่มข้อมูลตามเงื่อนไขที่เราระบุเท่านั้น สำหรับเทคนิควิธีที่ใช้ในการจัดกลุ่มข้อมูล เช่น Hierarchical clustering algorithms (Single-link, Complete-link, Average-link) และPartitional clustering algorithms (K-means, K-medoids)

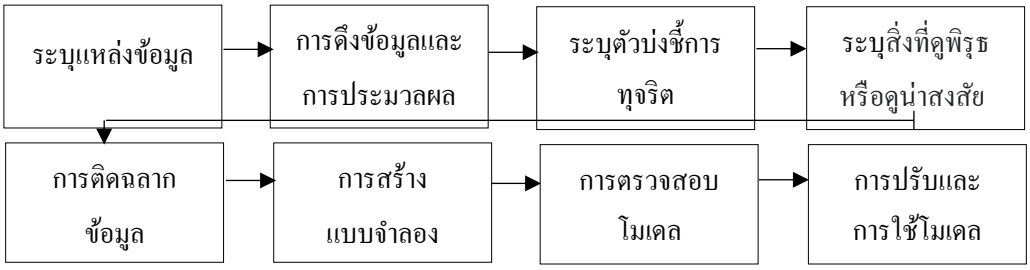
การเรียนรู้ในรูปแบบที่ใกล้เคียงกับการเรียนรู้ของมนุษย์มากที่สุด การเรียนรู้แบบเสริมแรง (Reinforcement learning) คือ Machine มีปฏิสัมพันธ์กับสิ่งแวดล้อมที่เปลี่ยนแปลงตลอดเวลา โดย Machine จะต้องทำงานบางอย่างเอง โดยที่ไม่มี “ผู้สอน” คอยบอกว่าวิธีการแบบใดจะทำให้ถึงเป้าหมาย หรือวิธีการที่ทำอยู่เข้าใกล้เป้าหมายแล้วหรือไม่

บทความนี้ในการนำเสนอตัวอย่างการวิเคราะห์ ผู้เขียนใช้การวิเคราะห์ด้วยประเภทการเรียนรู้จากข้อมูลแบบมีโครงสร้าง เป็นการเรียนรู้แบบมีผู้สอน โดยใช้ข้อมูลหรือจากข้อมูลตัวอย่างในอดีตที่มีผลการสอบสวนสิ้นสุดแล้วนำเข้าการเรียนรู้ของเครื่อง ให้ค้นหาความสัมพันธ์ และสร้างกฎทั่วไปไว้เพื่อทำนายว่าจากข้อมูลที่นำเข้าแล้วจะทำให้ได้ข้อมูลส่งออกแบบใด สำหรับเทคนิควิธีการจำแนกกลุ่มที่นำเสนอคือ ต้นไม้ตัดสินใจ

4. การใช้การเรียนรู้ของเครื่องสร้างโมเดลการทุจริตคอร์รัปชัน

การใช้การเรียนรู้ของเครื่องสร้างโมเดลวิเคราะห์การทุจริตคอร์รัปชัน สามารถใช้วิธีการวิเคราะห์ด้วยวิธีโครงข่ายประสาทเทียม (Artificial Neuron Network) หรือวิธีนาอีฟเบย์ (Naive bayes) หรือวิธี Support Vector Machine หรือ วิธีต้นไม้ตัดสินใจ (Decision tree) เนื่องจากแต่ละวิธีใช้สำหรับการจัดกลุ่มข้อมูล กล่าวคือ โมเดลต่าง ๆ จะใช้ข้อมูลในอดีตและเพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต สามารถให้ข้อมูลเชิงลึกแก่หน่วยงานกำกับดูแลเกี่ยวกับความเสี่ยงที่จะเกิดขึ้น (Gallego et al., 2018) หรือสามารถระบุความสัมพันธ์ขององค์ประกอบพื้นฐานที่นำไปสู่ความทุจริต จากผลการวิเคราะห์กฎความสัมพันธ์ของวิธีต้นไม้ตัดสินใจ (Velasco et al., 2021) ได้ศึกษาโมเดลที่มีคุณสมบัติการทำนายที่ดีที่สุด โดยวิเคราะห์ตัวชี้วัดการทุจริตสูง โดยการรวมฐานข้อมูลที่แตกต่างกัน การรู้จำรูปแบบ และความรู้ในการดึงข้อมูลเป็นส่วนหนึ่งของการค้นพบความรู้ จากการศึกษาได้เสนอแนะว่าการประยุกต์ใช้การเข้าถึงบิ๊กดาต้า (Big data) และวิธีการทำเหมืองข้อมูล จะทำให้ได้ผลลัพธ์ที่ดีขึ้น โดยได้ปรับข้อมูลนำเข้าด้วยฐานข้อมูลประเภทต่าง ๆ (ภาพที่ 3) เพื่อสร้างที่เก็บส่วนกลาง (Data lake) หรือชุดข้อมูลแบบรวมศูนย์ที่สามารถรองรับรูปแบบของพฤติกรรมที่น่าสงสัยหรือทุจริตในกระบวนการจัดซื้อจัดจ้าง ซึ่งคุณภาพการวิเคราะห์สิ่งผิดปกตินั้นขึ้นอยู่กับคุณภาพของสิ่งที่ดูพิรุธ หรือดูน่าสงสัย (Red Flags)

ภาพที่ 3 โมเดลการตรวจจับการฉ้อโกง



ที่มา: From Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies (pp. 3), by N. Modrušan, L. Mršić, & K. Rabuzin, *International Journal of Advanced Computer Science and Applications, 12*(2).

ตัวอย่างงานวิจัยต่างประเทศการทำเหมืองข้อมูลวิเคราะห์การทุจริต ดังตารางที่ 1 (Modrušan et al., 2021)

ตารางที่ 1 งานวิจัยการทำเหมืองข้อมูลวิเคราะห์การทุจริต

ผู้วิจัย	เป้าหมาย	รูปแบบ/วิธีการ	ตัวแปรเป้าหมาย
Wensink and Maarten de Vet (2006)	การประมาณความน่าจะเป็นของการทุจริต	Probit -Linear Regression	การทุจริต, กรณีโปร่งใส
Mencıa et al. (2013)	การทำนายจำนวนผู้ประมูล	k-NN, LibSVM, LibLinear Ensemble, Neural network	“น่าสงสัย” การประกวดราคาเดียว
Wang (2016)	แบบจำลองการคาดการณ์ความเสี่ยงจากการฉ้อโกงในสัญญา	One-Class Support Vector Machine, Logistic Regression	"เสี่ยง": ไม่รวมผู้รับเหมาเนื่องจากการฉ้อโกง, การทุจริตคอร์รัปชัน, การละเมิดกฎหมายป้องกันการผูกขาด
Gallego et al. (2018)	การคาดคะเนการกระทำผิดตามสัญญา	Lasso Logistic Regression, Conditional Inference Tree, Gradient Boosting machine	"น่าสงสัย": การขยายสัญญา, ผู้รับเหมาที่ถูกลงโทษ, ผู้รับเหมาบัญชีดำ
Carvalho et al. (2014)	การแยกซื้อ	Tree Augmented Network, Bayesian Networks	"น่าสงสัย" : สถาบันเดียวกันในที่เดียวกัน เดือนและปีที่รวมกันมากกว่า 8,000 ยูโร
Sales and Carvalho (2016)	แบบจำลองการคาดการณ์ความเสี่ยงจากการทุจริตในสัญญา	Naive Bayes, Tree-Augmented Naive Bayes score-based learning algorithms	"เสี่ยง": ระงับการประมูลชั่วคราว, ประกาศความไม่น่าเชื่อถือ, อุปสรรคเพื่อประมูลและจ้าง
Sales (2013)	แบบจำลองการคาดการณ์ความเสี่ยงจากการทุจริตในสัญญา	Logistic Regression, Decision Tree	"เสี่ยง": ข้อผิดพลาดร้ายแรงของซัพพลายเออร์ในการดำเนินการตามสัญญาใด ๆ
Sun and Sales (2018)	แบบจำลองการคาดการณ์ความผิดปกติของการจัดซื้อจัดจ้างสาธารณะที่ออกแบบมาสำหรับการคัดกรองเบื้องต้นของผู้รับเหมา	A neural network, Deep Neural Network, Logistic Regression, Discriminant Function Analysis	"เสี่ยง" : บริษัทรับประมูลได้อย่างน้อย หนึ่งบทลงโทษที่รุนแรงเนื่องจากเกิดความผิดปกติอย่างร้ายแรง
Fazekas and Kocsis (2020)	สัมประสิทธิ์ที่แสดงถึงความแข็งแกร่งของความสัมพันธ์ระหว่างการป้อนข้อมูลที่อาจก่อให้เกิดการทุจริตและผลการทุจริตที่อาจเกิดขึ้น	Logistic Regression, Linear Regression	"น่าสงสัย": ส่วนแบ่งผู้ชนะของผู้ออกหุ้นสัญญา, ผู้เสนอราคารายเดียว, ยกเว้นทั้งหมดแต่ผู้เสนอราคารายเดียว

ผู้วิจัย	เป้าหมาย	รูปแบบ/วิธีการ	ตัวแปรเป้าหมาย
Ralha and Silva (2012)	การตรวจจับพันธมิตร	Clustering, association rules, multi-agent approach	ความสัมพันธ์ระหว่างองค์การ
Baldomir et al. (2018)	การตรวจจับพันธมิตร	Association rules – A-priori algorithm	ความสัมพันธ์ระหว่างองค์การ
Domingos et al. (2016)	การตรวจจับความผิดปกติ	Deep Learning Auto-encoder algorithm	ความผิดปกติ
Tas (2017)	การเปิดเผยโครงสร้างของพฤติกรรมสมรู้ร่วมคิด	The reduced form of linear regression enriched KRLS method with the CF approach	ความสัมพันธ์ระหว่างองค์การ
Van Erven et al. (2017) และ Carneiro et al. (2020)	ระบุความสัมพันธ์ระหว่างบริษัท	Graph databases, Decision support system, Rule-based	หน่วยงานที่เกี่ยวข้องในกระบวนการ
Arief et al. (2016)	การตรวจจับการฉ้อโกงกระบวนการจัดซื้อจัดจ้างสาธารณะ	Naive Bayes, Bayesian networks, Decision tree and neural network.	"น่าสงสัย": คำตัดสินของศาล , คณะกรรมการปราบปรามการทุจริต (KPK) และความคิดเห็นของประชาชน
Modrušan et al. (2020)	การตรวจจับกระบวนการจัดซื้อจัดจ้างสาธารณะที่น่าสงสัย	Data mining, Linear regression, Support vector machines, Naive Bayes, Process mining	ประกวดราคาเดียว -ประมูลครั้งเดียว
Velasco et al. (2021)	การสมรู้ร่วมคิดระหว่างผู้เสนอราคา, ความขัดแย้งทางผลประโยชน์ทับซ้อนและเจ้าของบริษัทผู้เป็นคนบิดเบือนที่มีศักยภาพเพื่ออำพรางเจ้าของที่แท้จริง	Graph theory, Clustering, and Regression analysis with advanced data science methods	รูปแบบความเสี่ยงของการสมรู้ร่วมคิด, รูปแบบความเสี่ยงระดับองค์การ, รูปแบบความเสี่ยงระดับบุคคล

จากตารางที่ 1 แสดงให้เห็นถึงการนำกระบวนการทำเหมืองข้อมูล มาวิเคราะห์การทุจริตในการจัดซื้อจัดจ้าง งานวิจัยได้ใช้เทคนิคการเรียนรู้ของเครื่องโดยใช้วิธีการที่หลากหลาย (ช่องคอลัมน์ รูปแบบ/วิธีการ) และตัวแปรเป้าหมายการวิเคราะห์ ที่แตกต่างกันตามพฤติกรรมของการทุจริตที่เกิดขึ้น

5. ทฤษฎีต้นไม้ตัดสินใจ (Decision tree)

บทความนี้ผู้เขียนขอเสนอการประยุกต์ใช้กระบวนการเมืองข้อมูล ด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้วิธีการต้นไม้ตัดสินใจ วิเคราะห์การทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ การใช้วิธีการต้นไม้ตัดสินใจ จะทำให้ทราบความสัมพันธ์ของส่วนประกอบต่าง ๆ ของการของการทุจริตจัดซื้อจัดจ้างของหน่วยงานภาครัฐ ด้วยภาพที่สามารถอธิบายการเชื่อมโยงความสัมพันธ์ได้อย่างชัดเจน แนวคิดวิธีการต้นไม้ตัดสินใจ มีดังนี้

ต้นไม้ตัดสินใจ คือ แบบจำลองทางคณิตศาสตร์ เพื่อการหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน สามารถสร้างแบบจำลองการจัดหมวดหมู่ (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ล่วงหน้า ได้โดยอัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย

โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” (พยูน พาณิชย์กุล, 2548) เช่น

“If Income = High and Married = No THEN Risk = Poor”

“If Income = High and Married = Yes THEN Risk = Good”

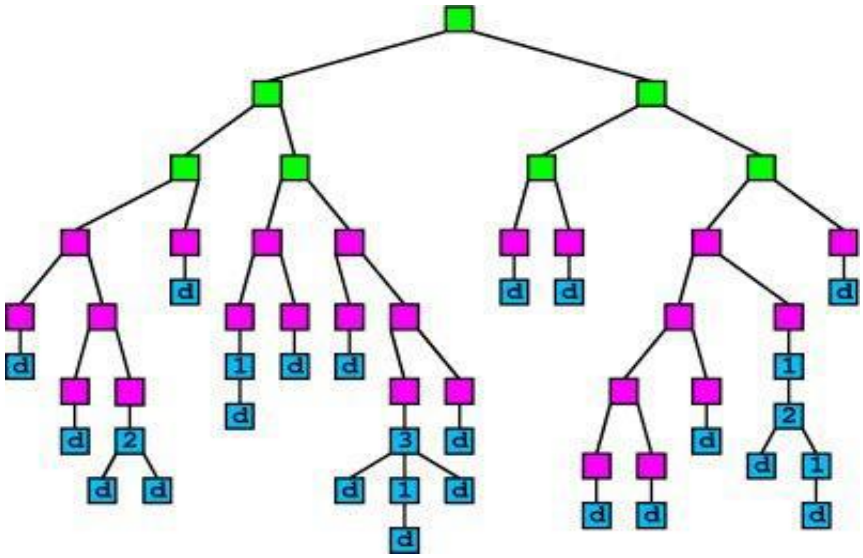
5.1 ส่วนประกอบของต้นไม้ตัดสินใจ ประกอบด้วย (จิตตภู พูลวัน, 2550)

(1) โหนด (Node) คือ คุณสมบัติต่าง ๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root node)

(2) กิ่ง (Branch) คือ คุณสมบัติของคุณสมบัติในโหนดที่แตกออกมา โดยจำนวนของกิ่งจะเท่ากับคุณสมบัติของโหนด

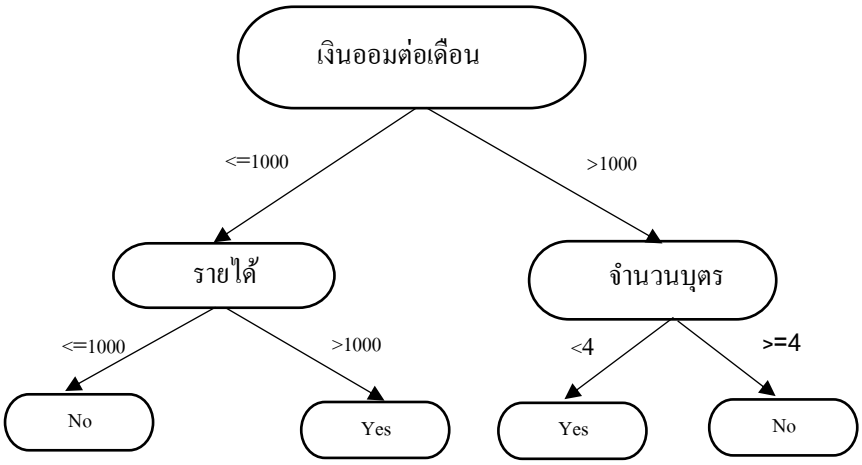
(3) ใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล โดยสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจ ดังภาพที่ 4 และ 5

ภาพที่ 4 แสดงส่วนประกอบของต้นไม้ตัดสินใจ



ที่มา: จาก *ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ* (รายงานวิจัย) (น. 68), โดย รุจิรา ธรรมสมบัติ, 2554, นนทบุรี: สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์.

ภาพที่ 5 ตัวอย่างต้นไม้ตัดสินใจ



ที่มา: จาก *การออกแบบและพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการโดยใช้เทคนิคต้นไม้ตัดสินใจ กรณีศึกษามูลนิธิที่อยู่อาศัยเพื่อมนุษยชาติ*, โดย ศุภชัย ประคองศิลป์, 2551, ปัญหาพิเศษปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

จากภาพที่ 5 ต้นไม้ตัดสินใจแสดงความสัมพันธ์ของปัจจัยที่มีผลต่อการออมเงิน > 1000 บาท หรือ ออมเงิน <= 1000 บาท โดยโหนดราก (Root node) เป็นเงินออมต่อเดือน โดยมี 2 กิ่ง (Branch) คือ กิ่งรายได้ และ กิ่งจำนวนบุตร และมีใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล คือ Yes และ No

5.2 ขั้นตอนการสร้างต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจ หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ เป็นการสร้างในลักษณะจากบนลงล่าง (Top-Down) คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้ (Han & Kamber, 2001)

1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก (Training set)

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าเกน (Gain) ของแต่ละ Attribute เพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือก Attribute ที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด โดย Attribute ที่มีค่าเกนมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือ Attribute ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ ถูกสร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อ Attribute ที่มีค่าเกนมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำ Attribute นี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่ Attribute ที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีก สำหรับโหนดในระดับต่อ ๆ ไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้เป็นจริง

หลังจากทำการสร้างต้นไม้ตัดสินใจแล้ว สามารถใช้ต้นไม้ตัดสินใจในการจำแนกข้อมูลใหม่ที่เข้ามาได้ โดยจะทำการจำแนกหมวดหมู่ของข้อมูลใหม่ (ที่ประกอบไปด้วย Attribute ต่าง ๆที่ไม่ทราบหมวดหมู่ข้อมูลนั้น) การเปรียบเทียบ Attribute ที่อยู่ในโหนดรากกับค่าของ Attribute ในข้อมูลใหม่ที่พิจารณา จะทำการเปรียบเทียบจากโหนดรากไปจนถึงโหนดใบ เมื่อเราทราบถึงโหนดใบจะทำให้เราทราบถึงหมวดหมู่ของข้อมูลใหม่

6. โปรแกรมคอมพิวเตอร์สำหรับวิเคราะห์เหมืองข้อมูล

ในการสร้างโมเดลเพื่อทำนาย ปัจจุบันมีโปรแกรมสำเร็จรูปที่สามารถนำมาสร้างโมเดลการเรียนรู้ของเครื่อง โปรแกรมที่ได้รับความนิยม ได้แก่ โปรแกรม Weka, โปรแกรม RapidMiner Studio หรือเขียนด้วยภาษา Python โดยใช้ไลบรารี Scikit-learn ในบทความนี้ผู้เขียนนำเสนอการใช้โปรแกรม RapidMine Studio สร้างโมเดลวิเคราะห์การทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ

โปรแกรม RapidMiner Studio แรกเริ่มพัฒนาขึ้นจากบริษัทที่ชื่อว่า Rapid-I ในประเทศเยอรมนี และปี พ.ศ. 2556 ได้รับทุนจากนักลงทุนในประเทศสหรัฐอเมริกาจึงเปลี่ยนชื่อบริษัทจาก Rapid-I เป็น RapidMiner แทนและย้ายสำนักงานใหญ่มาอยู่ประเทศสหรัฐอเมริกา ความสามารถและการทำงานของ RapidMiner Studio version Education (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557) เป็น Software ที่สามารถใช้งานฟรีทำให้สามารถประหยัดค่าใช้จ่าย กรณีต้องการใช้งานเพื่อทดสอบ การวิเคราะห์ข้อมูล การทำความสะอาดข้อมูล (Data Cleansing) และการรายงานข้อมูลในรูปแบบต่าง ๆ เช่น ตาราง แผนภูมิ เป็นต้น ความสามารถและการทำงานของ Software RapidMiner Studio ดังนี้

1. สามารถนำเข้าข้อมูลได้หลายลักษณะ เช่น การเชื่อมโยงจากฐานข้อมูลโดยตรง ไฟล์ Excel และไฟล์ CSV เป็นต้น
2. สามารถเขียนไฟล์ให้อยู่ในรูปแบบของ Excel และ CSV
3. สามารถแสดงข้อมูลในกราฟแบบต่าง ๆ เช่น Scatter Plot และ Time Series
4. สามารถแปลงข้อมูลจากฐานข้อมูล Relation Database ให้เป็นฐานข้อมูล Transaction Database
5. สามารถหากฎความสัมพันธ์ (Association Rules) การสร้างจาก Item ที่เกิดขึ้นบ่อย ๆ โดยเทคนิคการหากฎความสัมพันธ์ ได้แก่ Apriori และ FP Growth
6. สามารถแบ่งกลุ่มข้อมูล (Clustering) การแบ่งกลุ่มข้อมูล โดยข้อมูลที่มีลักษณะคล้าย ๆ กันอยู่กลุ่มเดียวกัน และข้อมูลที่อยู่คนละกลุ่มจะมีลักษณะที่แตกต่างกันมาก ๆ เทคนิค การแบ่งกลุ่มข้อมูล ได้แก่ K-Means, Agglomerative Clustering และ DBScan
7. สามารถจำแนกประเภทข้อมูล (Classification) การนำข้อมูลเดิมที่มีคำตอบที่สนใจ หรือคลาส (Class) มาสร้างเป็นโมเดล (Model) เพื่อหาคำตอบให้กับข้อมูลใหม่ (Unseen Data) โดยคลาสคำตอบเป็นประเภท (Nominal) เช่น ฝนตกหรือไม่ตก หรือ Spam Email หรือ Normal Email เป็นต้น เทคนิคการจำแนกประเภทข้อมูล ได้แก่ Linear Regression, Naïve Bayes, Decision Tree, K-Nearest Neighbors, Neural Networks และ Support Vector Machines
8. สามารถเปรียบเทียบประสิทธิภาพของโมเดลจากเทคนิค Classification ต่าง ๆ และ T-test
9. สามารถทำ Text Mining และ Image Mining, Selection และ Visualization

7. ตัวอย่าง การประยุกต์ใช้กระบวนการเหมืองข้อมูล (Data mining) ด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine learning) เพื่อการป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐโดยใช้โปรแกรม RapidMiner Studio *(ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลสมมติขึ้นเพื่อการศึกษาเท่านั้น)*

ขั้นตอนการวิเคราะห์โดยใช้วิธีการต้นไม้ตัดสินใจประกอบด้วย 5 ขั้นตอน ดังตารางที่ 2 ดังนี้

ตารางที่ 2 แสดงขั้นตอนการดำเนินงาน

ขั้นตอน	การดำเนินการ
ขั้นตอนที่ 1 การเลือกข้อมูล (Selection)	- ศึกษาคุณลักษณะของข้อมูลที่เกี่ยวข้องกับกระบวนการจัดซื้อจัดจ้างโครงการที่ดำเนินการเสร็จสมบูรณ์ - การกำหนดคุณลักษณะข้อมูลนำเข้า (Attribute) - การกำหนดข้อมูลนำออก (Label) - ดำเนินเก็บรวบรวมข้อมูลจากหน่วยงานภาครัฐต่าง ๆ ที่มีการดำเนินโครงการจัดซื้อจัดจ้างเสร็จสมบูรณ์แล้ว - ดำเนินเก็บรวบรวมข้อมูลจากหน่วยงานภาครัฐต่าง ๆ ที่มีการร้องเรียนการทุจริตการจัดซื้อจัดจ้าง และคดีถึงที่สุดแล้ว
ขั้นตอนที่ 2 การเตรียมข้อมูล (Preprocessing)	- คัดแยกข้อมูลที่ไม่สมบูรณ์ค่าข้อมูล ข้อมูลซ้ำซ้อน หรือทำการบันทึกผิดพลาดออกจากข้อมูลสมบูรณ์ - รวบรวมข้อมูลจากการเก็บจากวิธีการต่าง ๆ เข้าไว้ด้วยกันในรูปแบบไฟล์ดิจิทัล
ขั้นตอนที่ 3 การเปลี่ยนรูปข้อมูล (Transformation)	- เปลี่ยนรูปข้อมูลให้อยู่ในรูปแบบที่ใช้กับเทคนิคการเรียนรู้ของเครื่อง โดยใช้วิธีการต้นไม้ตัดสินใจได้ - กรณีที่ต้องการคัดกรองจำนวนคุณลักษณะของข้อมูลนำเข้า (Attribute) สามารถใช้การวิเคราะห์ความถดถอยพหุคูณ (Multiple Regression) วิเคราะห์ความสัมพันธ์ในเบื้องต้น
ขั้นตอนที่ 4 การทำเหมืองข้อมูล (Data mining)	- สร้างโมเดลทำนายการทุจริตการจัดซื้อจัดจ้างภาครัฐด้วยโปรแกรม Rapidminer Studio สร้างโมเดลด้วยโดยใช้วิธีการต้นไม้ตัดสินใจได้
ขั้นตอนที่ 5 การแปลผลและการประเมินผล (Interpretation evaluation)	- พิจารณาประสิทธิภาพการทำนายการทุจริตโครงการจัดซื้อจัดจ้างโดยดูจากค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความลึก (Recall) - แปลผลอยู่ในรูปของความรู้ (Knowledge) - แสดงกฎความสัมพันธ์ของระหว่างองค์ประกอบต่าง ๆ ของโครงการจัดซื้อจัดจ้างกับการทุจริตหรือไม่ทุจริต

จากตารางที่ 2 กระบวนการเมืองข้อมูลด้วยเทคนิคการเรียนรู้ของเครื่อง โดยใช้วิธีการต้นไม้ตัดสินใจเพื่อวิเคราะห์การทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ

ขั้นตอนที่ 1 การเลือกข้อมูล (Selection)

รวบรวมข้อมูลย้อนหลัง 5 ปี หรือมากกว่า การกำหนดประเภทของข้อมูล (Attribute) ที่เก็บรวบรวม เช่น 1) ประเภทโครงการ 2) สังกัดหน่วยงาน 3) จำนวนงบประมาณ 4) ระยะเวลาของโครงการ 5) พื้นที่โครงการ 6) อายุหัวหน้าสูงสุดของหน่วยงาน 7) โครงการ ทุจริต/ไม่ทุจริต (กำหนดเป็น Label)

ขั้นตอนที่ 2 การเตรียมข้อมูล (Preprocessing)

รวบรวมข้อมูลจากการเก็บจากวิธีการต่าง ๆ เข้าไว้ด้วยกันในรูปแบบไฟล์ Excel ดังภาพที่ 6

ภาพที่ 6 แสดงตัวอย่างการเตรียมข้อมูล

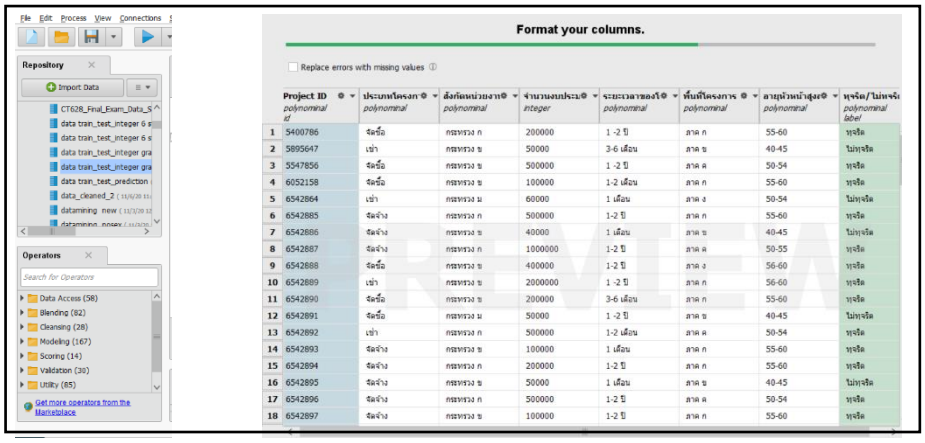
	A	B	C	D	E	F	G	H
1	Project ID	ประเภทโครงการ	สังกัดหน่วยงาน	จำนวนงบประมาณ	ระยะเวลาของโครงการ	พื้นที่โครงการ	อายุหัวหน้าสูงสุดของหน่วยงาน	ทุจริต/ไม่ทุจริต
2	5400786	จัดซื้อ	กระทรวง ก	200,000	1 -2 ปี	ภาค ก	55-60	ทุจริต
3	5895647	เช่า	กระทรวง ข	50,000	3-6 เดือน	ภาค ข	40-45	ไม่ทุจริต
4	5547856	จัดซื้อ	กระทรวง ข	500,000	1 -2 ปี	ภาค ค	50-54	ทุจริต
5	6052158	จัดซื้อ	กระทรวง ข	100,000	1-2 เดือน	ภาค ก	55-60	ทุจริต
6	6542864	เช่า	กระทรวง ม	60,000	1 เดือน	ภาค ง	50-54	ไม่ทุจริต
7	6542885	จัดจ้าง	กระทรวง ก	500,000	1-2 ปี	ภาค ก	55-60	ทุจริต
8	6542886	จัดจ้าง	กระทรวง ข	40,000	1 เดือน	ภาค ข	40-45	ไม่ทุจริต
9	6542887	จัดจ้าง	กระทรวง ก	1,000,000	1-2 ปี	ภาค ค	50-55	ทุจริต
10	6542888	จัดซื้อ	กระทรวง ข	400,000	1-2 ปี	ภาค ง	56-60	ทุจริต
11	6542889	เช่า	กระทรวง ข	2,000,000	1 -2 ปี	ภาค ก	56-60	ทุจริต
12	6542890	จัดซื้อ	กระทรวง ข	200,000	3-6 เดือน	ภาค ก	55-60	ทุจริต

ขั้นตอนที่ 3 การเปลี่ยนรูปข้อมูล (Transformation)

เนื่องจากข้อมูลตามตัวอย่างสามารถนำเข้าโปรแกรม RapidMiner Studio และสามารถวิเคราะห์ด้วยวิธีต้นไม้ตัดสินใจ แต่ต้องมีการปรับรูปข้อมูล 2 Attribute คือ Project ID ตั้งค่าเป็น id และ Attribute ทุจริต/ไม่ทุจริต ตั้งค่าเป็น Label ขั้นตอนการตั้งค่า ดำเนินการในขั้นตอนนำข้อมูลเข้าโปรแกรม RapidMiner Studio ดังภาพที่ 7

ภาพที่ 7 การ Import data และการตั้งค่า Attribute

โปรแกรม RapidMiner Studio และใช้เมนู Import data เพื่อนำข้อมูลเข้าโปรแกรม

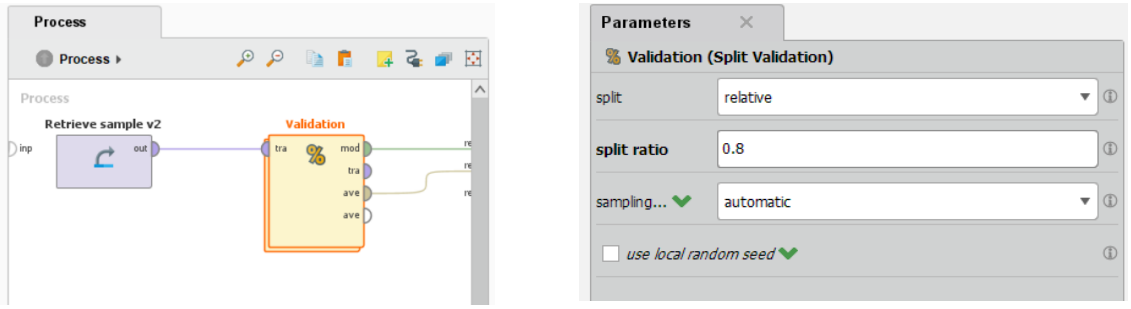


ขั้นตอนที่ 4 การทำเหมืองข้อมูล (Data mining)

ขั้นตอนการสร้างโมเดลด้วยโปรแกรม RapidMiner Studio ดังนี้

ขั้นที่ 1 เลือก Data set และเลือก Operator Split validation จากนั้นลากเส้นเชื่อมโยง Operator ต่าง ๆ เข้าด้วยกันเพื่อทำให้เกิด Process และตั้งค่า Operator Split Validation ดังนี้ split = relative, split ratio = 0.8 (Train set 80 %, Test set 20 %) และเลือกกลุ่มตัวอย่างแบบ Automatic ดังภาพที่ 8

ภาพที่ 8 การตั้งค่าพารามิเตอร์ Operator Validation

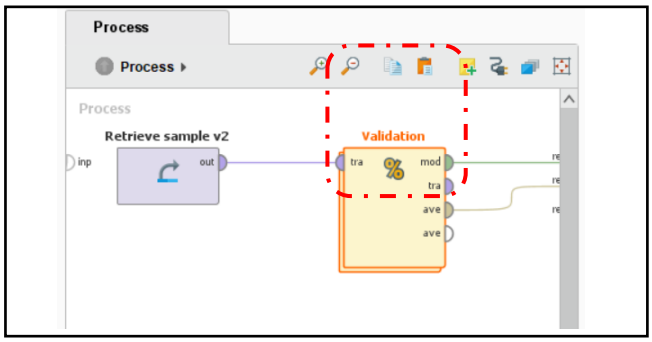


ขั้นที่ 2 เลือก Operator สำหรับสร้างโมเดลทำนาย

การสร้างโมเดล Decision Tree

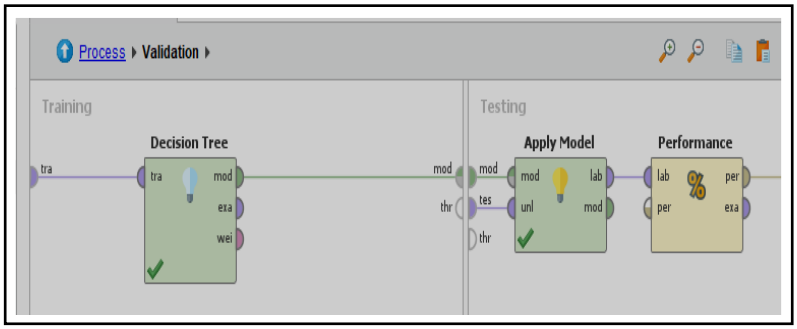
1) เปิด Operator Validation ดังภาพที่ 9

ภาพที่ 9 การเปิด Operator Validation



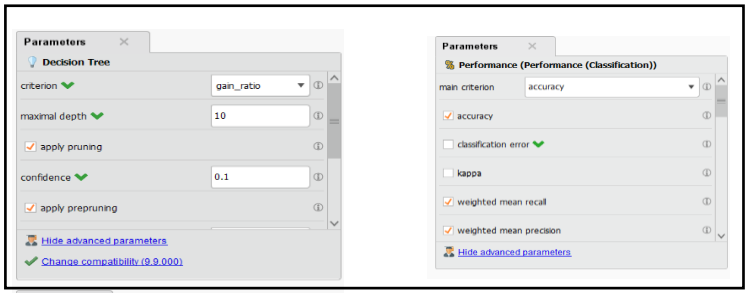
2) หลังจากเปิด Operator Validation จะพบหน้าต่าง Training และ Testing ในหน้าต่าง Training เลือก Operator Decision tree และในหน้าต่าง Testing วาง Operator Apply Model และ Operator Performance จากนั้น ลากเส้นเชื่อมโยง operator ต่าง ๆ เข้าด้วยกันเพื่อทำให้เกิด Process ดังภาพที่ 10

ภาพที่ 10 การสร้างโมเดล Decision Tree



3) การตั้งค่าพารามิเตอร์ Operator Decision tree ค่าพารามิเตอร์ Operator Performance ดังภาพที่ 11

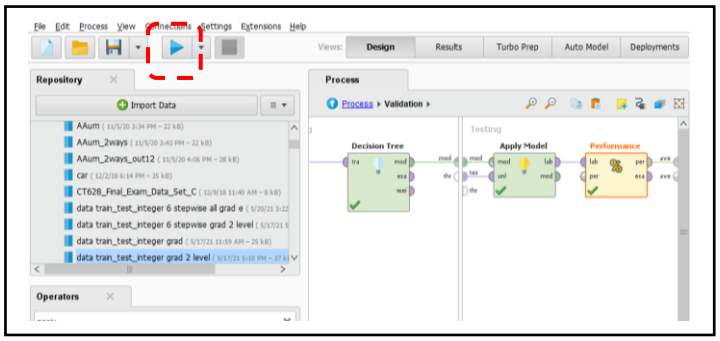
ภาพที่ 11 การตั้งค่าพารามิเตอร์ Operator Decision tree, Operator Performance



จากภาพที่ 11 กรณีตัวอย่างในบทความนี้ ตั้งค่า Decision tree ดังนี้ criterion = gain_ratio, maximal dept = 10, confidence = 0.1 และตั้งค่า Performance ดังนี้ main criterion = accuracy และเลือกผลทำนายประกอบด้วย ค่า accuracy , weighted mean recall, weighted mean precision

ขั้นที่ 3 Run โมเดล เพื่อดูค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) การ Run โมเดลดังภาพที่ 12

ภาพที่ 12 การ Run โมเดล Decision tree



ขั้นตอนที่ 5 การแปลผลและการประเมินผล (Interpretation evaluation)

ผลลัพธ์ประสิทธิภาพการทำนายดังภาพที่ 13

ภาพที่ 13 แสดงผลโมเดลทำนายได้ค่า Accuracy, Precision, Recall 100 % ทุกค่า

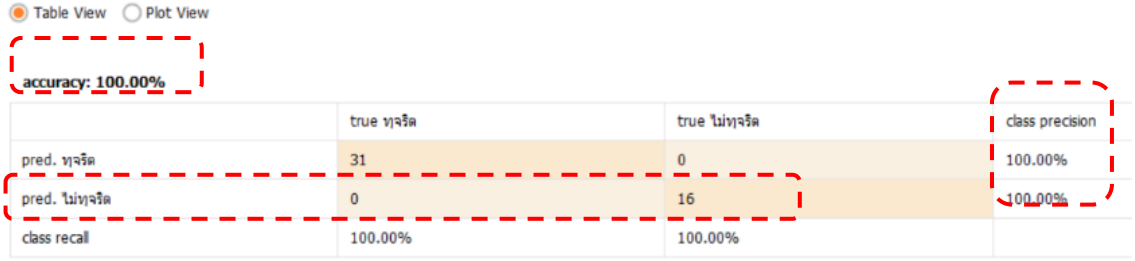
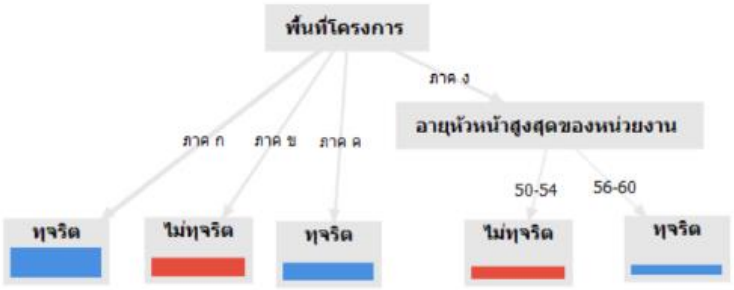


Table View Plot View

accuracy: 100.00%

	true ทุจริต	true ไม่ทุจริต	class precision
pred. ทุจริต	31	0	100.00%
pred. ไม่ทุจริต	0	16	100.00%
class recall	100.00%	100.00%	

ภาพที่ 14 ภาพตัวอย่างต้นไม้ตัดสินใจจากการวิเคราะห์ด้วยโปรแกรม RapidMiner Studio



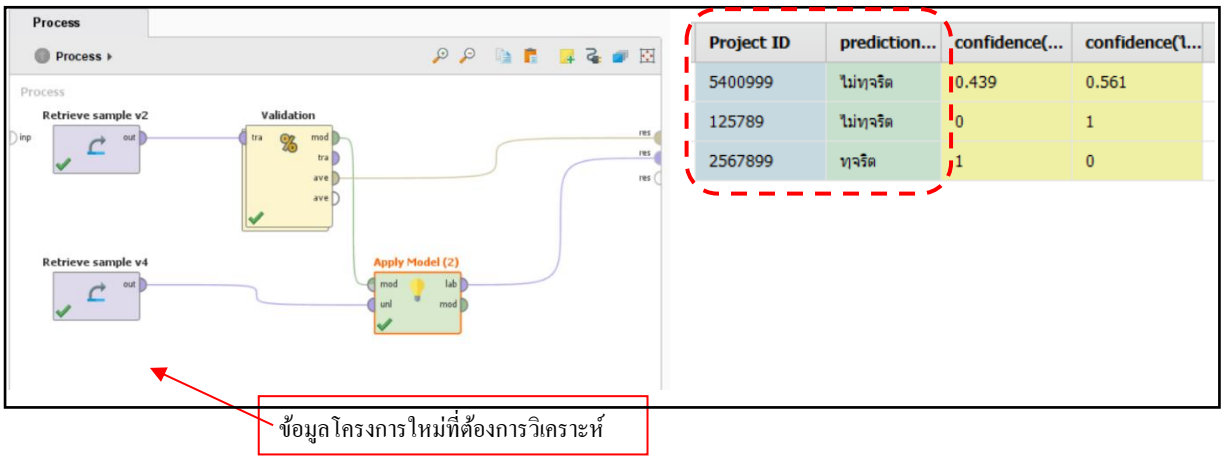
นอกจากการนำเสนอด้วยแผนภาพต้นไม้ตัดสินใจ โปรแกรม RapidMiner Studio ได้นำเสนอกฎความสัมพันธ์ของ Attribute ต่าง ๆ มีความสัมพันธ์กับการทุจริตหรือไม่ทุจริต ลักษณะกฎความสัมพันธ์แสดงดังนี้

if พื้นที่โครงการ = ภาค ก then ทุจริต
if พื้นที่โครงการ = ภาค ข then ไม่ทุจริต
if พื้นที่โครงการ = ภาค ค then ทุจริต
if พื้นที่โครงการ = ภาค ง and อายุหัวหน้าสูงสุดของหน่วยงาน = 50-54 then ไม่ทุจริต
if พื้นที่โครงการ = ภาค ง and อายุหัวหน้าสูงสุดของหน่วยงาน = 56-60 then ทุจริต

หลังจากดำเนินการสร้างโมเดลการเรียนรู้ของเครื่องโดยใช้วิธีการต้นไม้ตัดสินใจได้ ถ้าค่าความถูกต้องไม่ถึง 95 % ควรดำเนินการเพิ่ม Attribute ของข้อมูลเข้าใหม่ หรือเพิ่มจำนวนข้อมูลมากขึ้น เพื่อที่การเรียนรู้ของเครื่องจะได้มีข้อมูลในการเรียนรู้มากขึ้นซึ่งจะส่งผลต่อค่าความถูกต้องของการทำนายมากขึ้นเช่นกัน

ขั้นตอนการทำนายโครงการจัดซื้อจัดจ้างใหม่ สามารถใช้กฎความสัมพันธ์ของ Attribute เขียนเป็นโปรแกรมคอมพิวเตอร์เพื่อวิเคราะห์โครงการที่กำลังจะดำเนินการในอนาคตได้ หรือใช้การวิเคราะห์โครงการจัดซื้อจัดจ้างใหม่ด้วยโปรแกรม RapidMiner Studio โดยสร้างไฟล์ข้อมูลโครงการใหม่ และนำข้อมูลเข้าในแต่ละ Attribute ยกเว้น Attribute ทุจริต/ไม่ทุจริต เพราะโปรแกรมจะวิเคราะห์ และนำผลการวิเคราะห์เพิ่มใน Attribute ทุจริต/ไม่ทุจริต โดยอัตโนมัติ การสร้างโมเดลดังภาพที่ 15 ดังนี้

ภาพที่ 15 แสดงตัวอย่างการสร้างโมเดล และผลการวิเคราะห์โครงการจัดซื้อจัดจ้าง



8. สรุป

การทุจริตในสังคมไทยระหว่างช่วงเวลากว่าทศวรรษ ส่งผลเสียต่อประเทศอย่างมหาศาลและเป็นอุปสรรคสำคัญต่อการพัฒนาเศรษฐกิจ สังคม การเมือง ในทุกมิติ รูปแบบการทุจริตจากเดิมที่เป็นทุจริตทางตรงไม่ซับซ้อน อาทิ การรับสินบน การจัดซื้อจัดจ้าง ในปัจจุบันได้ปรับเปลี่ยนเป็นการทุจริตที่ซับซ้อนมากขึ้น ซึ่งการจัดซื้อจัดจ้างภาครัฐที่ไม่โปร่งใส จะนำไปสู่การทุจริตคอร์รัปชันต่าง ๆ มากมาย เช่น การติดสินบน การสมรู้ร่วมคิด การฮั้วประมูล การเอื้อผลประโยชน์ ไปจนถึงการมีผลประโยชน์ทับซ้อน ซึ่งเกิดจากผู้มีส่วนเกี่ยวข้องใช้อำนาจ หรือตัดสินใจโดยคำนึงถึงผลประโยชน์ส่วนตนมากกว่าผลประโยชน์ส่วนราชการ ไม่ว่าจะเป็นผู้กำหนดนโยบาย เจ้าหน้าที่ของรัฐที่รับผิดชอบในการจัดซื้อจัดจ้าง ผู้เข้าร่วมประมูล คู่สัญญา และผู้มีส่วนร่วมเกี่ยวข้องทั้งหลายในสัญญา การป้องกันการทุจริตคอร์รัปชันถือเป็นยุทธศาสตร์และวาระแห่งชาติ เนื่องจากการทุจริตคอร์รัปชันเป็นปัญหาที่สำคัญของประเทศไทยที่ต้องได้รับการแก้ไขอย่างเร่งด่วน

ปัจจุบันสามารถประยุกต์ใช้กระบวนการเหมืองข้อมูล ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยใช้วิธี ต้นไม้ตัดสินใจ วิเคราะห์การทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐโดยใช้ข้อมูลโครงการจัดซื้อจัดจ้างที่ดำเนินการเสร็จสมบูรณ์แล้วมาเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ ขั้นตอนของกระบวนการเหมืองข้อมูลประกอบด้วย 1) การเลือกข้อมูล 2) การเตรียมข้อมูล 3) การเปลี่ยนรูปข้อมูล 4) การทำเหมืองข้อมูล 5) การแปลผลและการประเมินผล ซึ่งเป็นขั้นตอนที่นำรูปแบบ (Pattern) ที่ได้จากการทำเหมืองข้อมูล แปลให้อยู่ในรูปของภาพต้นไม้ตัดสินใจ และแสดงกฎความความสัมพันธ์ของ Attribute เป็นชุดความรู้เพื่อใช้สนับสนุนการตัดสินใจ

การวิเคราะห์เพื่อทำนายโครงการจัดซื้อจัดจ้างใหม่ โปรแกรมคอมพิวเตอร์จะใช้ความรู้จากการเรียนรู้ข้อมูลเก่ามาเปรียบเทียบกับข้อมูลใหม่ ในกรณีที่โปรแกรมวิเคราะห์และทำนายว่าโครงการจัดซื้อจัดจ้างใหม่อยู่ในกลุ่มทุจริต หน่วยงานเจ้าของโครงการสามารถนำมาเป็นข้อมูลในวางแผนป้องกันทุจริต หรือเป็นข้อมูลให้หน่วยงานที่เกี่ยวข้องวางมาตรการเฝ้าระวังและตรวจสอบการดำเนินการแต่ละขั้นตอนให้เที่ยงตรงโปร่งใสตามเจตนารมณ์ของการใช้งบประมาณแผ่นดินต่อไป

9. ข้อเสนอแนะ

9.1 เพื่อให้การป้องกันการทุจริตการจัดซื้อจัดจ้างของหน่วยงานภาครัฐโดยการประยุกต์ใช้กระบวนการเหมืองข้อมูล (Data mining) ด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine learning) มีประสิทธิภาพมากขึ้น หน่วยงานภาครัฐควรออกแบบโครงสร้างฐานข้อมูลการจัดซื้อจัดจ้างให้เป็นรูปแบบเดียวกันทุกกระทรวง และหน่วยงานที่ดำเนินการจัดซื้อจัดจ้างต้องป้อนข้อมูลตามคุณลักษณะที่ออกแบบไว้เข้าระบบฐานข้อมูลกลาง (Data Center) ทุกครั้ง เพื่อจัดเก็บเป็นฐานข้อมูลขนาดใหญ่ (Big data) จะทำให้ข้อมูลมีการเพิ่มเข้าระบบ (Update) อย่างต่อเนื่อง ส่งผลให้การเรียนรู้ของเครื่อง (Machine learning) มีข้อมูลในการเรียนรู้มากขึ้น ค่าความถูกต้องของการวิเคราะห์ผลจะถูกต้อง แม่นยำมากขึ้น

9.2 ควรพัฒนาโปรแกรมคอมพิวเตอร์เพื่อวิเคราะห์แนวโน้มการทุจริตโครงการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ โดยใช้เงื่อนไขการการวิเคราะห์จากกฎความสัมพันธ์ที่ได้จากโมเดล อัลกอริทึม (Algorithm) ต้นไม้ตัดสินใจ (Decision tree)

9.3 ควรพัฒนาระบบปัญญาประดิษฐ์ (Artificial Intelligence: AI) และระบบอินเตอร์เน็ตของสรรพสิ่ง (Internet of Thing: IOT) มาเป็นเครื่องมือในจัดเก็บข้อมูลและส่งข้อมูลเข้าระบบฐานข้อมูลกลาง (Data center) เนื่องจากโครงการจัดซื้อจัดจ้างแต่ละโครงการมีความแตกต่างกัน เช่น พื้นที่โครงการ ขนาดโครงการ บุคคลที่เกี่ยวข้องในโครงการ หรืองบประมาณ ซึ่งการจัดเก็บข้อมูลโดยระบบปัญญาประดิษฐ์ (Artificial Intelligence: AI) อาจเก็บข้อมูลได้เที่ยงตรงมากว่าเก็บด้วยบุคคล เป็นต้น

เอกสารอ้างอิง

กรมสอบสวนคดีพิเศษ. (2564). *ประกาศกรมสอบสวนคดีพิเศษ เรื่อง มาตรการส่งเสริมความโปร่งใสในการจัดซื้อจัดจ้างของกรมสอบสวนคดีพิเศษ.* สืบค้นเมื่อวันที่ 10 กรกฎาคม 2564, จาก https://www.dsi.go.th/oit/file/o44/มาตรการส่งเสริมความโปร่งใสในการจัดซื้อจัดจ้าง.pdf

จิตตภู พูลวัน. (2550). *ระบบวิเคราะห์โรคทั่วไปโดยใช้การจำแนกข้อมูลบนอินเทอร์เน็ต*. กรุงเทพฯ: ภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ธราพงศ์ ลิ้มสุทธิวันภูมิ และ ตวงทอง สินชัย. (2563). การทุจริตในการจัดซื้อจัดจ้างของหน่วยงานภาครัฐ Procurement Corruption in Government Agencies. *วารสารสหวิทยาการมนุษยศาสตร์และสังคมศาสตร์, 3*(2), 311-323.

ปริญญา สงวนสัตย์. (2562). *Articial Intelligence with Machine Learning, AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง*. นนทบุรี: ไอดีซี พรีเมียร์.

พยูน พาณิชย์กุล. (2548). *การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Decision Tree* [โครงการพัฒนาระบบงานปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศแขนงวิทยาการสารสนเทศสถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง]

รุจิรา ธรรมสมบัติ. (2554). *ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ* (รายงานวิจัย). นนทบุรี: สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์.

ศุภชัย ประคองศิลป์. (2551). *การออกแบบและพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการโดยใช้เทคนิคต้นไม้ตัดสินใจ กรณีศึกษามูลนิธิที่อยู่อาศัยเพื่อมนุษยชาติ* [ปัญหาพิเศษปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ]

ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงศึกษาธิการ. (2564). *ราชบัณฑิตฯ บัญญัติ "Big Data" เป็นภาษาไทยว่า "ข้อมูลมหัต" และ Machine Learning ว่า "การเรียนรู้ของเครื่อง."* สืบค้นเมื่อวันที่ 10 กรกฎาคม 2564, จาก http://www.bict.moe.go.th/2019/index.php?option=com_content&view=article&id=766:big-data-machine-learning-25-2563&catid=43&Itemid=351

ศูนย์ปฏิบัติการต่อต้านการทุจริต กระทรวงศึกษาธิการ. (2560). *แผนปฏิบัติการป้องกันและปราบปรามการทุจริตและประพฤติมิชอบกระทรวงศึกษาธิการระยะ 5 ปี (พ.ศ. 2560 – 2564).* สืบค้นเมื่อวันที่ 10 กรกฎาคม 2564, เข้าถึงได้จาก http://www.reo6.moe.go.th/web/images/stories/ข้อ%2035%20แผนปฏิบัติ การป้องกันการทุจริต.pdf.

สิรินทรา เสียบไธสง และ สมศักดิ์ อมรสิริพงศ์. (2562). สภาพปัญหาการจัดซื้อจัดจ้างในหน่วยงานภาครัฐที่ก่อให้เกิดการทุจริตคอร์รัปชัน State of Problems: Procurement Problems in Government Agencies That Cause Corruption, ใน มหาวิทยาลัยเกษตรศาสตร์, *การประชุมทางวิชาการครั้งที่ 57 มหาวิทยาลัยเกษตรศาสตร์* (น. 437 - 446). กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *Introduction to Data Mining with RapidMiner Studio 7*. กรุงเทพฯ : ห้างหุ้นส่วนสามัญดาต้าคิวบ์.

Arief, H. A., Saptawati, G. A. P., & Asnar, Y. D. W. (2016). Fraud detection based-on data mining on Indonesian E-Procurement System (SPSE). In IEEE, *2016 International Conference on Data and Software Engineering (ICoDSE)* (pp. 1-6). Denpasar: IEEE

Baldomir, R. A., Van Erven, G. C., & Ralha, C. G. (2018). Brazilian Government Procurements: an Approach to Find Fraud Traces in Companies Relationships. In Mauá, D. D., & Naldi, M. (Eds.), *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional* (pp. 752-762). São Paulo: SBC.

Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., & Costa, J. (2020). Network Analysis for Fraud Detection in Portuguese Public Procurement. In Analide, C., Novais, P., Camacho, D., & Yin, H. (Eds.), *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 390-401). Guimaraes: Springer.

Carvalho, R. N., Sales, L. J., Da Rocha, H. A., & Mendes, G. L. (2014). Using Bayesian networks to identify and prevent split purchases in Brazil. In Laskey, K. B., Jones, J., & Almond, R. (Eds.), *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop 1218* (pp. 70-78). Quebec: CEUR-WS.

Domingos, S. L., Carvalho, R. N., Carvalho, R. S., & Ramos, G. N. (2016). Identifying IT purchases anomalies in the Brazilian government procurement system using deep learning. In IEEE, *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 722-727). CA: Conference Publishing Services (CPS).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, *AI Magazine, 17*(3), 37 - 54.

Fazekas, M., & Kocsis, G. (2020). Uncovering high-level corruption: crossnational objective corruption risk indicators using public procurement data. *British Journal of Political Science, 50*(1), 155-164.

Gallego, J., Rivero, G., & Martínez, J. D. (2018). *Preventing rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement* (Documentos de trabajo 016724). Bogotá: Universidad del Rosario.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. CA: Morgan Kaufmann.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Kaur, P., Singh, M., & Josan, GS. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science, 57*, 500-508.

Mencıa, E. L., Holthausen, S., Schulz, A., & Janssen, F. (2013). Using data mining on linked open data for analyzing e-procurement information. In Claudia D'Amato, C., Berka, P., Svátek, V., & Wecel, K. (Eds.), *Proceedings of the 2013 International Conference on Data Mining on Linked Data - Volume 1082* (pp. 50–57). Aachen: CEUR-WS.org.

Modrušan, N., Mršić, L., & Rabuzin, K. (2021). Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies. *International Journal of Advanced Computer Science and Applications, 12*(2), 1-9.

Modrušan, N., Rabuzin, K., & Mršić, L. (2020). Improving Public Sector Efficiency using Advanced Text Mining in the Procurement Process. In Hammoudi, S., Quix, C., & Bernardino, J. (Eds.), *Proceedings of the 9th International Conference on Data Science*, *Technology and Applications - Volume 1*: DATA (pp. 200-206). Coimbra: ISEC.

Ralha, C. G., & Silva, C. V. (2012). A multi-agent data mining system for cartel detection in Brazilian government procurement. *Expert Systems with Applications, 39*(14), 11642-11656.

Sales, L. (2013). Risk prevention of public procurement in the brazilian government using credit scoring. In Húmus, E. (Ed.), *Paper in the International Conference Interdisciplinary Insights on Fraud and Corruption.* OBEGEF-Observatório de Economiae Gestão de Fraude & OBEGEF.

Sales, L. J., & Carvalho, R. N. (2016). Measuring the Risk of Public Contracts Using Bayesian Classifiers. In *Uncertainty in Artificial Intelligence BMA@UAI*. (pp. 7–13). N.p.: BMA@UAI.

Shalev-Shwartz, S., & Ben-David, S. (2009). *Understanding Machine Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019.

Sun, T., Sales, L. J. (2018). Predicting public procurement irregularity: An application of neural Networks. *Journal of Emerging Technologies in Accounting, 15*(1), 141-154.

Tas, B. K. (2017). *Collusion Detection in Public Procurement with Limited Information*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2929222.

Van Erven, G. C. G., Carvalho, R. N., De Holanda, M. T., & Ralha, C. (2017). Graph database: A case study for detecting fraud in acquisition of brazilian government. In IEEE, *2017 12th Iberian Conference on Information Systems and Technologies (CISTI).* (pp. 1-6). Lisbon: IEEE

Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C., & Ribeiro, O. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research, 28*(1), 27-47.

Wang, Y. (2016). *Detecting Fraud in Public Procurement*. [Abstract of the Dissertation, University Libraries on behalf of The Graduate School at Stony Brook University].

Wensink, W., & Maarten de Vet, J. (2006). *Identifying and Reducing Corruption in Public Procurement in the EU*. Woluwedal: PwC and Ecorys.