การสร้างตัวแบบวิเคราะห์พฤติกรรมและสาเหตุต่อการเป็นโรคติดสุรา ด้วยเทคนิคต้นไม้ตัดสินใจ

A Behavioral Analysis Model and the Cause of Alcohol Dependence with the Decision Tree Technique

ณัฐภัทร ฤทธิ์เดช[*] อดิศร ขาวเอียด[*] และปริพัส ศรีสมบูรณ์[**]

[*] สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

[**] สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

Nattapat Rideach[*], Adisorn Khaoead[*] and Paripas Srisomboon[**]

[*] Department of Information Technology, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

[**] Department of Information Technology, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

บทคัดย่อ

การสร้างตัวแบบวิเคราะห์พฤติกรรมและสาเหตุต่อการเป็นโรคติดสุราด้วยเทคนิคต้นไม้ตัดสินใจ งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและพัฒนาตัวแบบวิเคราะห์ พฤติกรรมและสาเหตุต่อการดื่มเครื่องดื่มแอลกอฮอล์และปัจจัยที่ส่งผลต่อการทำให้เป็นโรคติดสุรา โดยใช้เทคนิคต้นไม้ตัดสินใจ กลุ่มตัวอย่างที่ใช้คือ นักศึกษา มหาวิทยาลัยราชภัฏนครปฐมชั้นปีที่ 1 ถึงชั้นปีที่ 4 อำเภอเมืองนครปฐม จังหวัดนครปฐม จำนวนผู้ทำแบบสำรวจทั้งหมดเป็น 500 คน โดยมีรายละเอียดข้อมูลคือ มีจำนวนคุณลักษณะทั้งหมด 12 คุณลักษณะ และเมื่อได้วิเคราะห์ข้อมูลจะพบว่ามีจำนวนระเบียนที่สมบูรณ์พร้อมใช้งานทั้งหมด 500 ระเบียน และจำแนกได้ทั้งหมด 4 กลุ่ม คือ กลุ่มที่ 1 คือ ผู้ดื่มแบบเสี่ยงต่ำ กลุ่มที่ 2 คือ ผู้ดื่มแบบเสี่ยง กลุ่มที่ 3 คือ ผู้ดื่มแบบอันตราย กลุ่มที่ 4 คือ กลุ่มผู้ดื่มแบบติด ซึ่งใช้เทคนิคการจำแนกข้อมูลประเภทข้อมูลด้วยต้นไม้ตัดสินใจ โดยเปรียบเทียบอัลกอริทึมที่ใช้ในงานวิจัยคือ 4 อัลกอริทึม ได้แก่ อัลกอริทึม J48, อัลกอริทึม LMT, อัลกอริทึม RandomTree และอัลกอริทึม HoeffdingTree และใช้เทคนิคการปรับค่าความสมดุลของข้อมูลด้วยวิธีเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic minority Over-sampling Technique : SMOTE) เพื่อให้ได้ค่าที่เหมาะสมที่สุดและถูกต้องที่สุด นำมาสร้างตัวแบบวิเคราะห์พฤติกรรมและสาเหตุต่อการเป็นโรคติดสุราด้วยเทคนิคต้นไม้ตัดสินใจ โดยผลการวิจัยครั้งนี้ พบว่าอัลกอริทึม J48 มีค่าความถูกต้องและต่าตวามแม่นยำของข้อมูลซึ่งใช้ 8-fold cross validation ในการแบ่งข้อมูล โดยมีค่าความสมดุล SMOTE 800 นั้น มีค่าความถูกต้อง 80.61%, ค่าความแม่นยำ 78.20%, ค่าความระลึก 80.60% และค่าความถ่วงดุล 78.80% ซึ่งมากกว่า LMT, RandomTree และ HoeffdingTree

คำสำคัญ: โรคติดสุรา ต้นไม้ตัดสินใจ เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Abstract

Creating a behavioral model and causes of alcohol consumption by using the decision tree technique. This research aimed to study and develop behavioral analysis models about causes of alcohol use and factors affecting the cause of alcohol dependence. The sample used was students university,

Freshman to Senior, Nakhon Pathom Rajabhat University, Muang District, Nakhon Pathom Province. The total number of survey takers is 500 people and the detailed data is a total of 12 characteristics. When analyzing the data, it was found that there were a total of 500 complete records available and could be classified into 4 groups: Group 1 was low risk drinkers, Group 2 was risk drinkers, Group 3 was dangerous drinkers, and Group 4 was addicted drinkers. Which uses data classification techniques with a decision tree by comparison the algorithms include the J48 algorithm, the LMT algorithm, the RandomTree algorithm, and the HoeffdingTree algorithm. By Using techniques for adjusting the balance of information with the randomly added technique. Minority samples (Synthetic minority Over-sampling Technique: SMOTE) to get the best and most accurate values to create a behavior analysis model and causes of alcohol dependence with decision tree technique. The results of this research revealed that the J48 algorithm has the data validity and precision values, which use 8-fold cross validation to divide the data.The balance SMOTE 800 has 80.61% validity value, 78.20% accuracy value, 78.80% precision value, which is more than LMT, RandomTree and HoeffdingTree.

Keywords: alcohol dependence, Decision Tree, Synthetic minority Over-sampling Technique

บทนำ

รายงาน “สถานการณ์การบริโภคเครื่องดื่มแอลกอฮอล์ในสังคมไทย ประจำ ปี พ.ศ. 2560 ผลการสำรวจ พบว่า จากประชากรไทยอายุ 15 ปีขึ้นไปจำนวน 55.9 ล้านคนทั่วประเทศ มีผู้ไม่เคยดื่มสุราเลยร้อยละ 57.18 ผู้เคยดื่มแต่เลิกแล้วร้อยละ 14.41 และมีนักดื่มปัจจุบันร้อยละ 28.41 โดยมีความแตกต่างระหว่างเพศชัดเจน ประชากรเพศหญิงร้อยละ 81.05 ไม่เคยดื่มสุราเลยในชีวิต และเป็นนักดื่มปัจจุบันร้อยละ 10.62 ส่วนในประชากรเพศชายมีผู้ไม่เคยดื่มสุราเลยเพียงร้อยละ 31.62และเป็นนักดื่มปัจจุบันร้อยละ 47.46 สัดส่วนนักดื่มปัจจุบันสูงที่สุดในภาคเหนือ ร้อยละ 35.40 และต่ำสุดในภาคใต้ร้อยละ 16.05 ไม่พบความแตกต่างระหว่างพฤติกรรมการดื่มในเขตเทศบาลและนอกเขตเทศบาล นักดื่มประจำ ร้อยละ 56.06 ดื่มเป็นครั้งคราว (น้อยกว่าสัปดาห์ละ 1 ครั้ง) ส่วนร้อยละ 43.94 ดื่มเป็นประจำ (Wichaidit, McNally, Saenam, & Asanangkornchai, 2017)

เครื่องดื่มแอลกอฮอล์ หมายถึงเครื่องดื่มที่มีเอทิลแอลกอฮอล์ผสมอยู่ ได้แก่ สุรา และเมรัย แอลกอฮอล์มีฤทธิ์กดระบบประสาทส่วนกลาง ผู้ที่ดื่มสุราในปริมาณไม่มาก จะรู้สึกผ่อนคลาย เนื่องจากแอลกอฮอล์ไปกดจิตใต้สำนึกที่คอยควบคุมตนเองอยู่ แต่เมื่อกินมากขึ้นก็จะกดสมองบริเวณอื่นๆ ทำให้เสียการทรงตัว พูดไม่ชัด จนแม้กระทั่งหมดสติในที่สุด เครื่องดื่มชนิดนี้ผลิตจากวัตถุดิบที่มีส่วนประกอบของน้ำตาลมาหมัก และเติมยีสต์ลงไปเพื่อให้ยีสต์กินน้ำตาลที่อยู่ในวัตถุดิบ และเปลี่ยนให้กลายเป็นแอลกอฮอล์ (Alcoholic beverages, 2020) โดยทั่วไปสุราจะมีอยู่ 2 ประเภทคือ สุรากลั่น และสุราไม่กลั่น สุรากลั่น คือ สุราที่เกิดจากการเอาน้ำกากส่า แป้งหรือน้ำตาล หมักด้วยยีสต์ จนเกิดแอลกอฮอล์แล้วนำมามากลั่น อาจจะมีการเติมสีปรุงกลิ่น แต่งรส ด้วยสารปรุงแต่ง อื่น ๆ

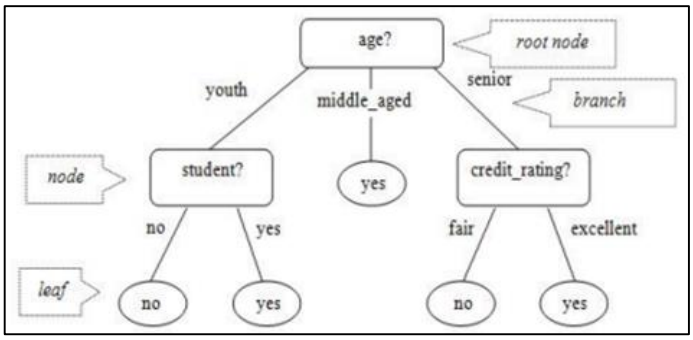
แล้วเก็บบ่มต่อไป สุราประเภทนี้ได้แก่ สุราขาว สุราผสมสุราผสมพิเศษ และสุราปรุงพิเศษ เป็นต้น สุราไม่กลั่น คือ สุราที่ได้จากการหมักส่าให้เกิดเป็นน้ำเมา แต่ไม่นำมากลั่น มักจะหมักจากเมล็ดธัญพืช ผลไม้และน้ำตาลจากพืช เช่น เบียร์ ไวน์ กระแช่ น้ำขาว น้ำตาลเมา เป็นต้น (Cartoon, A. 2012)

โรคติดสุรา หรือพิษสุราเรื้อรัง เป็นคำกว้าง ๆ ที่ใช้ในการวินิจฉัยผู้ติดสุรา ที่ทำให้เกิดปัญหาหน้าที่การงาน หรือสุขภาพ มีลักษณะการดำเนินโรคอย่างเรื้อรัง ก้าวหน้า โดยมีลักษณะที่สูญเสียความสามารถในการควบคุมการดื่มสุรา และมีผลเสียตามมาในด้านต่าง ๆ เช่น ด้านสังคม กฎหมาย สุขภาพจิต และสุขภาพทางร่างกาย และผู้ป่วยมักไม่ตระหนักถึงปัญหาเหล่านี้ว่าเกิดจากการใช้สุรา การดื่มสุรามาเป็นเวลานาน ทำให้เกิดความเปลี่ยนแปลงทางความรู้สึก แลทำให้ผู้ใช้อยากใช้อีกแม้ว่าจะเกิดผลเสียมากมาย เป็นการเสพติดทางใจและการเสพติดทางร่างกายทำให้เกิดการถอนสุราหรือไม่สบาย หากไม่ได้ใช้ (Faculty of medicine, Chulalongkorn University, 2012)

ปัจจุบันในทางการแพทย์ยังไม่สามารถสรุปสาเหตุของโรคติดสุราได้อย่างแน่ชัด แต่คาดว่าการดื่มแอลกอฮอล์ปริมาณมากส่งผลให้สารเคมีในสมองเปลี่ยนแปลงและทำให้เกิดความรู้สึกสุขสมตามมา จนอาจทำให้ความรู้สึกดังกล่าวส่งผลต่อผู้บริโภคให้ต้องการดื่มแอลกอฮอล์ปริมาณมากขึ้นเรื่อย ๆ และเกิดการเสพติดในที่สุด ซึ่งหากหยุดดื่มอาจมีอาการขาดสุราตามมาจึงทำให้เลิกดื่มได้ยากไปด้วย (PobPad, 2017)

การทำเหมืองข้อมูล(Data mining) คือ กระบวนการค้นหาสารสนเทศหรือข้อความรู้ที่อยู่ในฐานข้อมูลขนาดใหญ่ที่ซับซ้อนเพื่อนำข้อความรู้ที่ได้ไปใช้ประโยชน์ในการตัดสินใจ สารสนเทศที่ได้อาจนำมาสร้างการพยากรณ์หรือสร้างตัวแบบสำหรับการจำแนกหน่วยหรือกลุ่มหรือแสดงความสัมพันธ์ระหว่างหน่วยต่าง ๆ หรือให้ข้อสรุปของสาระในฐานข้อมูล โดยการทำเหมืองข้อมูลมีขั้นตอนการดำเนินงานหลายขั้นตอน ซึ่งต้องอาศัยเทคนิคหรือวิธีการต่าง ๆ เช่นวิธีการ จัดกลุ่มการค้นหาความสัมพันธ์ การพยากรณ์ เป็นต้น การดำเนินงานมักอยู่ในลักษณะของการสร้างตัวแบบที่อธิบายความเป็นไปหรือสภาพการหนึ่งที่เกิดขึ้นแล้วหรือที่ทราบคำตอบแล้วนำตัวแบบนี้มาใช้อธิบายสถานการณ์ที่ยังไม่เกิดขึ้นหรือที่ยังไม่ทราบคำตอบ ตัวแบบเหล่านี้อาจเป็นตัวแบบที่เรียบง่ายไปจนถึงตัวแบบที่ยุ่งยากซับซ้อนและอาจใช้การผสมผสานแนวคิดหรือเครื่องมือต่าง ๆ เข้าด้วยกันเพื่อที่จะสามารถสกัดข้อความรู้ที่อยู่ในข้อมูลขนาดใหญ่ได้โดยใช้เทคโนโลยีครั้งข้อมูลเข้ามาช่วยในการจัดการข้อมูลเพื่อเพิ่มประสิทธิภาพของการทำเหมืองข้อมูลดังนั้นถ้ามีฐานข้อมูลขนาดใหญ่ที่มีข้อมูลคุณภาพดีเทคโนโลยีการทำเหมืองข้อมูลจะช่วยในการค้นหรือแสวงหาโอกาสทางธุรกิจใหม่โดยการทำเหมืองข้อมูลจะก่อให้เกิดกระบวนการอัตโนมัติในการค้นพบสารสนเทศหรือข้อความรู้ในฐานข้อมูลขนาดใหญ่ด้วยการใช้วิธีการเช่น การพยากรณ์แนวโน้มและพฤติกรรม การบริโภคแบบอัตโนมัติ หรือเกิดกระบวนการอัตโนมัติในการค้นพบรูปแบบที่ไม่เคยรู้จักมาก่อนด้วยการใช้วิธีการค้นหาเข้าไปในรายละเอียดของฐานข้อมูลเพื่อหารูปแบบที่ซ่อนอยู่ในฐานข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภททั้งด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหารในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการในการจัดเก็บและตีความหมายข้อมูลจากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายง่ายมาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูลซึ่งจุดประสงค์ของการทำเหมืองข้อมูลสามารถใช้ค้นข้อมูลสำคัญที่ปะปนกับข้อมูลอื่น ๆ ในฐานข้อมูลที่ไม่ใช่แค่การสุ่มเรียกว่า KDD(Knowledge Discovery in Database) หรือการค้นหาข้อมูลด้วยความรู้มีด้วยกัน 5 รูปแบบคือ 1) การค้นหากฎความสัมพันธ์ 2) การจำแนกประเภทและการพยากรณ์ 3) การจัดกลุ่มข้อมูล 4) การหาค่าผิดปกติที่เกิดขึ้น 5) การวิเคราะห์แนวโน้ม (Boonserm, 2003)

ต้นไม้ตัดสินใจ(Decision Tree) เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ ซึ่งเมื่อมีข้อมูลที่ต้องการจัดกลุ่มก็จะนำคุณลักษณะ(Attribute) ต่าง ๆ ของข้อมูลนั้นไปเทียบกับเส้นทางในต้นไม้จนกระทั่งคลาสปลายทางซึ่งก็คือกลุ่มของข้อมูลที่เหมือนกัน ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ เป็นตัวทดสอบกิ่งของต้นไม้(Branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ(Leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (Class) ก็คือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก(root node) โครงสร้างของต้นไม้ตัดสินใจแสดงดังรูปที่ 1



ภาพที่ 1 โครงสร้างต้นไม้ตัดสินใจ (Atcharaporn, J. 2016)

ต้นไม้ตัดสินใจมีค่าวัดความสามารถของการจัดกลุ่มของแต่ละคุณลักษณะหรือปัจจัย คือ Gini Index ค่าที่บ่งบอกว่าคุณลักษณะหรือปัจจัยใดควรนำมาใช้เป็นคุณลักษณะในการแบ่งกลุ่มของอัลกอริทึม โดยมีสูตรในการคำนวณดังนี้

$$Gini(ti) = 1 - \sum[P(ti)]2 \quad (1)$$

Entropy ค่าคาดคะเนของข้อมูลเป็นค่าที่แยกโดยใช้ ลักษณะประจำ ของอัลกอริทึม ID3

$$Entropy(ti) = 1 - \sum_{i=1}^{N}[p(t_i)]log_2 p(t_i) \quad (2)$$

โดยที่

t_i คือ คุณลักษณะที่นำ มาวัดค่า Entropy

$p(t_i)$ คือสัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

ซึ่งแต่ละอัลกอริทึมจะให้ผลของโครงสร้างต้นไม้ตัดสินใจที่แตกต่างกันไป (Atcharaporn, 2016) เทคนิคต้นไม้ตัดสินใจที่นำมาใช้ในงานวิจัย มีดังนี้ J48 หรืออัลกอริทึมของ C4.5 เป็นอัลกอริทึม ในการสร้างต้นไม้ตัดสินใจจากกลุ่มของข้อมูลฝึกสอนโดยใช้ ความถูกต้องของแต่ละคุณลักษณะของข้อมูล เพื่อใช้เป็นการตัดสินใจแบ่งกลุ่มข้อมูลกลุ่มย่อยๆ โดยพิจารณาจากค่า ความแตกต่างใน Entropy ผลลัพธ์จากการเลือกคุณลักษณะ สำ หรับแบ่งกลุ่มข้อมูล ด้วยค่า Normalized information gain ที่สูงสุดนั้นคือการสร้างการตัดสินใจ Logistic Model Trees (LMT) เป็นการรวมกัน ของเทคนิคต้นไม้(Trees)และการถดถอยโลจิสติก (Logistic Regression) (Paranya, P. 2016) Random Tree เป็นเทคนิคที่ใช้ใน การจำแนกหมวดหมู่เช่นเดียวกับ C4.5 โดยมีหลักการสร้าง Tree จากการสุ่ม Tree หลายๆแบบ ในแต่ละโหนดแล้วเลือกมาประมวลผลโดยไม่ใช้การ Prune และเนื่องจากจำนวนของ Tree เพิ่มขึ้นอย่างรวดเร็ว ซึ่งยากแก่การแก้ปัญหา การสร้าง Tree ที่เป็นไปได้ทั้งหมด ส่วนประกอบสามารถสร้างชุดสุ่มของ Tree ออกมาจากการกระจายชุดต่างๆของ Tree Random Tree เป็นการสุ่มวาดที่สุ่มจากชุดของ Tree ที่เป็นไปได้ (Pattanaphong & Jaree, 2021) และอัลกอริทึม Hoeffding Tree

SMOTE เป็นเทคนิคที่ใช้ในการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุลเนื่องจากข้อมูลในแต่ละคลาสมีจำนวนแตกต่างกันมาก ทำให้ผลลัพธ์ของการจำแนกอยู่ ในข้อมูลกลุ่มมาก ดังนั้นวิธี SMOTE เป็นการเพิ่มจำนวน ข้อมูลกลุ่มน้อย ให้มีจำนวนเพิ่มขึ้นโดยการเพิ่มข้อมูลใน กลุ่มน้อยนั้นทำให้การกระจายของกลุ่มข้อมูลมีความสมดุลมากขึ้น โดยทำการสุ่มค่าข้อมูลที่อยู่ในกลุ่มข้อมูล น้อยขึ้นมา 1 ค่าหลังจากนั้นพิจารณาค่าข้อมูลใกล้เคียง อีกจำนวน K ค่า(K-nearest neighbor) แล้วคำนวณค่า ระยะทาง(Euclidean distance) ระหว่างค่าที่สุ่มกับค่า ข้อมูลใกล้เคียงแต่ละค่า เพื่อหาค่าระยะทางที่น้อยที่สุด ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียง จากนั้นจึงสร้าง ข้อมูลเทียมระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัว ที่ให้ค่าระยะทางที่น้อยที่สุด เพื่อนำค่าที่ได้มาเพิ่มจำนวน ข้อมูล ดังสมการที่ 3

$$X_{new} = X_i + (\hat{X}_i - X_i) \times \delta \quad (3)$$

โดยที่

X_{new} คือ ข้อมูลใหม่

X_i คือ ข้อมูลที่สุ่มในตอนแรก

$\hat{X}_i$ คือ ข้อมูลที่สุ่มมาอีก เช่น สุ่มมาอีก 5 จุด

δ คือ ค่าสุ่มตั้งแต่ 0-1

การคำนวณประสิทธิภาพของตัวแบบได้ใช้ Confusion Matrix คือตารางสรุปจำนวนข้อมูลที่ตัวแบบมีรายละเอียดดังรูปที่ 2 (Chawla, Bowyer, Hall, & Kegelmayer, 2002)

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

รูปที่ 2 TP, FP, TN และFN สำหรับ class Predicted

สำหรับวิธีการวัดประสิทธิภาพของวิธีการที่นำเสนอมีรายละเอียดดังนี้

Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

Recall เป็นการวัดความระลึกของโมเดล โดยพิจารณาแยกทีละคลาส

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

F-Measure เป็นการวัดค่า Precision และRecall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

การวัดค่าความถูกต้องของโมเดล โดยพิจารณารวมทุก คลาส คือ จำนวน *True Positive* ของทุกคลาสรวมกัน

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (7)$$

วัตถุประสงค์ของงานวิจัย

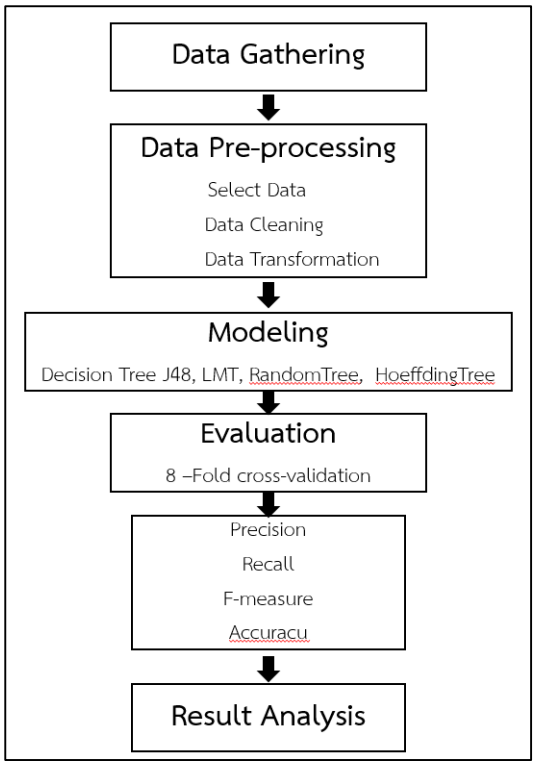
เพื่อพัฒนาตัวแบบวิเคราะห์พฤติกรรมและสาเหตุต่อการดื่มเครื่องดื่มแอลกอฮอล์และปัจจัยที่ส่งผลต่อการทำให้เป็นโรคติดสุรา

ประชากรและกลุ่มตัวอย่าง

ใช้ข้อมูลจากแบบสำรวจผู้ที่ดื่มเครื่องดื่มแอลกอฮอล์ในกลุ่มนักศึกษา มหาวิทยาลัยราชภัฏนครปฐมชั้นปีที่ 1 ถึง 4 จากระบบบริการออนไลน์ Google forms ที่จัดทำไว้มีจำนวน 3 ส่วน ได้แก่ ส่วนที่ 1. สถานภาพทั่วไป 2. แบบสำรวจปัญหาการดื่มสุรา (ซึ่งได้มาจากคณะแพทยศาสตร์จุฬาลงกรณ์มหาวิทยาลัย) 3. พฤติกรรมการดื่มเครื่องดื่มแอลกอฮอล์ในปัจจุบัน จำนวน 25 ข้อ และมีเกณฑ์ประเมินจำนวน 4 ระดับ คือ 1. ผู้ดื่มแบบเสี่ยงต่ำ 2. ผู้ดื่มแบบเสี่ยง 3. ผู้ดื่มแบบอันตราย 4. ผู้ดื่มแบบติด ทั้งหมด 500 ชุด

วิธีดำเนินการวิจัย

ในการทำวิจัยเรื่อง การสร้างตัวแบบวิเคราะห์พฤติกรรมและสาเหตุต่อการเป็นโรคติดสุราด้วยเทคนิคต้นไม้ตัดสินใจ ผู้วิจัยต้องค้นคว้าข้อมูลและความรู้ที่จะนำมาใช้ในงานวิจัย ดังนี้ 1) รวบรวมข้อมูลจากแบบสำรวจผู้ที่ดื่มเครื่องดื่มแอลกอฮอล์ในกลุ่มนักศึกษา มหาวิทยาลัยราช ภัฏนครปฐม อำเภอเมืองนครปฐม จังหวัดนครปฐม 2) การเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล 3) พัฒนาตัวแบบการประเมิน 4) การตรวจสอบประสิทธิภาพของตัวแบบ



รูปที่ 4 ขั้นตอนการดำเนินงานวิจัย

โดยกลุ่มตัวอย่างที่ใช้ในงานวิจัยเป็นกลุ่มนักศึกษา มหาวิทยาลัยราชภัฏนครปฐมชั้นปีที่ 1 ถึง 4 อำเภอเมืองนครปฐม จังหวัดนครปฐม ในการเก็บรวบรวมข้อมูลได้ผลการประเมินความเสี่ยงต่อการเป็นโรคติดสุราแล้วนำส่วนของข้อมูลส่วนตัวและพฤติกรรมการดื่มเครื่องดื่มแอลกอฮอล์ในปัจจุบันของผู้ตอบแบบสำรวจมาสร้างตัวแบบในการพยากรณ์ความเสี่ยงต่อการเป็นโรคติดสุรา โดยมีรายละเอียดข้อมูลคือ มีจำนวนคุณลักษณะทั้งหมด 12 คุณลักษณะ และเมื่อได้วิเคราะห์ข้อมูลจะพบว่ามีจำนวนระเบียนที่สมบูรณ์พร้อมใช้งานทั้งหมด 500 ระเบียน ซึ่งคุณลักษณะข้อมูลที่ใช้ปรากฏดังตารางที่ 1

ตารางที่ 1 รายละเอียดคุณลักษณะ ที่นำมาสร้างตัวแบบ

ลำดับ	ชื่อคุณลักษณะ	คำอธิบาย
1	Sex	เพศ
2	Age	อายุ
3	Classroom	ระดับชั้นการศึกษา
4	Salary	รายได้เฉลี่ย
5	Shelter	ที่พักอาศัย
7	Drinking time	ช่วงวันที่ดื่ม
8	Place to drink	สถานที่ดื่ม
9	Often to drink	ความถี่ที่ดื่ม
10	Beverage type	ประเภทเครื่องดื่ม
11	Motivation	แรงจูงใจ
12	Class	1 = ผู้ดื่มแบบเสี่ยงต่ำ 2 = ผู้ดื่มแบบเสี่ยง 3 = ผู้ดื่มแบบอันตราย 4 = ผู้ดื่มแบบติด

จากตารางที่ 1 คุณลักษณะที่ 15 เป็นคุณลักษณะที่กลุ่มของข้อมูลที่ใช้ในการจำแนกข้อมูลซึ่งกลุ่มที่แบ่งได้จากผลการประเมินจากแบบสำรวจผู้ที่ดื่มสุราของกรมสุขภาพจิตโดยแบ่งออกได้ 4 กลุ่มได้แก่

กลุ่มที่ 1 คือ ผู้ดื่มแบบเสี่ยงต่ำ

กลุ่มที่ 2 คือ ผู้ดื่มแบบเสี่ยง

กลุ่มที่ 3 คือ ผู้ดื่มแบอันตราย

กลุ่มที่ 4 คือ ผู้ดื่มแบบติด

เมื่อพิจารณาข้อมูลที่ได้จากแบบสำรวจแล้วนำข้อมูลมาทำการศึกษา จะพบว่าข้อมูลที่ได้มีปัญหาข้อมูลไม่สมดุล ดังนี้

กลุ่มผู้ดื่มแบบเสี่ยงต่ำ จำนวน 241 ระเบียน

กลุ่มผู้ดื่มแบบเสี่ยง จำนวน 78 ระเบียน

กลุ่มผู้ดื่มแบอันตราย จำนวน 31 ระเบียน

กลุ่มผู้ดื่มแบบติด จำนวน 150 ระเบียน

ดังนั้นผู้วิจัยจึงนำวิธี SMOTE มาใช้เพื่อปรับสมดุลข้อมูล คือเพิ่มจำนวนข้อมูลกลุ่มที่มีข้อมูลน้อยก่อนที่จะนำเข้าสู่กระบวนการต่อไป และทำการปรับค่าพารามิเตอร์ให้เหมาะสมผลการทดสอบพบว่า ขนาดชุดข้อมูลที่สามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีที่สุด คือ ร้อยละ 49.60 เพิ่มข้อมูลขึ้น 248 ระเบียน รวมเป็น 748 ระเบียน ดังตารางที่ 2 ทั้งนี้จะทำการทดลองเปรียบเทียบกับข้อมูลที่ไม่ผ่านวิธี SMOTE

ตารางที่ 2 ข้อมูลที่ผ่านวิธี SMOTE แก้ไขปัญหาข้อมูลไม่สมดุล

กลุ่ม	Original	SMOTE
ผู้ดื่มแบบเสี่ยงต่ำ	241	241
ผู้ดื่มแบบเสี่ยง	78	78
ผู้ดื่มแบบอันตราย	31	279
กลุ่มผู้ดื่มแบบติด	150	150
จำนวนทั้งหมด	500	878

หลังรวบรวมข้อมูลและการเตรียมข้อมูลเสร็จสิ้นจะได้ข้อมูลซึ่งอยู่ในรูปแบบฐานข้อมูล จากนั้นจะเข้าสู่ขั้นตอนการทดสอบหาอัลกอริทึมคือ ขั้นตอนที่ 3 Classification การจำแนกประเภทข้อมูล เป็นกระบวนการทำเหมืองข้อมูล ซึ่งเป็นเทคนิคในการจำแนกข้อมูล ตามที่กำหนดด้วยต้นไม้ช่วยการตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยต้นไม้ช่วยการตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ที่แต่ละโหนดแสดงคุณลักษณะแต่ละกิ่งแสดงเงื่อนไขใน การทดสอบ และโหนดปลายแสดงกลุ่มที่กำหนดไว้ โดยจะเลือกเทคนิคในต้นไม้ตัดสินใจ ดังต่อไปนี้ J48, LMT, RandomTree และ HoeffdingTree โดยในงานวิจัยนี้ใช้การทดสอบความถูกต้องของตัวแบบแบบ Cross-Validation

ผลการศึกษาและอธิปรายผล

นำข้อมูลมาปรับสมดุลด้วยวิธีการ SMOTE และนำข้อมูลเข้าสู่กระบวนการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจได้แก่ J48, LMT, RandomTree & HoeffdingTree เพื่อพยากรณ์การเป็นโรคติดสุราในกลุ่มนักศึกษา มหาวิทยาลัยราชภัฏนครปฐมชั้นปีที่ 1 ถึง 4 การดำเนินงานได้ทำการเปรียบเทียบระหว่างข้อมูลที่ผ่านการปรับสมดุลโดยใช้ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล เพื่อวัดประสิทธิภาพของตัวแบบ ได้ผลการดำเนินงานดังตารางที่ 3

ตารางที่ 3 ตารางผลการดำเนินการทั้ง 4 อัลกอริทึม

Algorithms	Original				CFS : SMOTE			
	Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
J48	74.00	69.80	74.00	70.90	80.61	78.20	80.60	78.80
LMT	70.20	67.10	70.20	68.50	79.54	77.40	79.50	78.20
RandomTree	70.00	68.10	70.00	69.00	78.34	75.20	78.30	76.30
HoeffdingTree	62.20	58.60	62.20	59.30	62.03	58.20	62.00	58.10

จากตารางที่ 3 สามารถสรุปผลได้ดังนี้

การวัดค่าความถูกต้องซึ่งถือได้ว่าเป็นประสิทธิภาพโดยรวมของตัวแบบในการพยากรณ์ ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี CFS : SMOTE ทำให้ค่าความถูกต้องของตัวแบบสูงขึ้นทุกเทคนิค โดยเฉพาะ J48 ให้ค่าความถูกต้องสูงสุดร้อยละ 80.61 ตามด้วยเทคนิค LMT ให้ค่าความถูกต้องร้อยละ 79.54 เทคนิค RandomTree ให้ค่าความถูกต้องร้อยละ 78.34 และตัวแบบที่สร้างด้วยเทคนิค HoeffdingTree ให้ค่าความถูกต้องน้อยที่สุดที่ร้อยละ 62.03

การวัดค่าความแม่นยำเป็นการบ่งบอกถึงอัตราส่วนการพยากรณ์ถูกต้องในแต่ละกลุ่ม ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี CFS : SMOTE ทำให้ค่าความแม่นยำของตัวแบบสูงขึ้นทุกเทคนิคโดยเฉพาะ J48 ให้ค่าความแม่นยำค่าสูงสุดร้อยละ 78.20 ตามด้วยเทคนิค LMT ให้ค่าความแม่นยำร้อยละ 77.40 เทคนิค RandomTree ให้ค่าความแม่นยำร้อยละ 75.20 และตัวแบบที่สร้างด้วยเทคนิค HoeffdingTree ให้ค่าความแม่นยำน้อยที่สุดที่ร้อยละ 58.20

การวัดค่าความระลึกเป็นการบ่งบอกถึงอัตราส่วนการพยากรณ์ถูกต้องในแต่ละกลุ่ม ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี CFS : SMOTE ทำให้ค่าความระลึกของตัวแบบสูงขึ้นทุกเทคนิคโดยเฉพาะ J48 ให้ค่าความระลึกสูงสุดร้อยละ 80.60 ตามด้วยเทคนิค LMT ให้ค่าความระลึกร้อยละ 79.50เทคนิค RandomTree ให้ค่าความระลึกร้อยละ 78.30 และตัวแบบที่สร้างด้วยเทคนิค HoeffdingTree ให้ค่าความระลึกน้อยที่สุดที่ร้อยละ 62.00

การวัดค่าความถ่วงดุลเป็นการบ่งบอกถึงอัตราส่วนการพยากรณ์ถูกต้องในแต่ละกลุ่ม ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี CFS : SMOTE ทำให้ค่าความถ่วงดุลของตัวแบบสูงขึ้นทุกเทคนิคโดยเฉพาะ J48 ให้ค่าความถ่วงดุลสูงสุดร้อยละ 78.80 ตามด้วยเทคนิค LMT ให้ค่าความถ่วงดุลร้อยละ 78.20 เทคนิค RandomTree ให้ค่าความถ่วงดุลร้อยละ 76.30 และตัวแบบที่สร้างด้วยเทคนิค HoeffdingTree ให้ค่าความถ่วงดุลน้อยที่สุดที่ร้อยละ 58.10

จำนวนข้อมูลที่เลือกมาทดสอบทั้งหมด 500 ระเบียน โดยใช้การทดสอบแบบต้นไม้ตัดสินใจในงานวิจัยนี้จะใช้เทคนิคที่กล่าวมานี้มาวิเคราะห์ข้อมูลชุดข้อมูลทั้ง 500 ระเบียน ว่ามีประสิทธิภาพความแตกต่าง และความถูกต้อง เพื่อหาเทคนิคที่เหมาะกับงานวิจัย ศึกษาถึงการวิเคราะห์ข้อมูลเพื่อพัฒนาตัวแบบการประเมินความเสี่ยงต่อการเป็นโรคติสุรา การวัดประสิทธิภาพของความถูกต้องของข้อมูลในงานวิจัยนี้วัดได้จากค่าความถูกต้องของการจัดกลุ่มของข้อมูล

สรุปและอภิปรายผล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อนำเสนอตัวแบบการประเมินความเสี่ยงต่อการเป็นโรคติดสุราของนักศึกษา มหาวิทยาลัยราชภัฏนครปฐมชั้นปีที่ 1 ถึง 4 อำเภอเมืองนครปฐม จังหวัดนครปฐม ซึ่งเทคนิคการพยากรณ์ในงานวิจัยฉบับนี้ใช้เทคนิคต้นไม้ตัดสินใจและแก้ปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE จากการทดลอง พบว่า

การแก้ไขปัญหาข้อมูลที่ไม่สมดุลด้วยเทคนิค CFS:SMOTE สามารถเพิ่มประสิทธิภาพให้ตัวแบบเพิ่มขึ้นในทุกเทคนิคที่ได้นำมาเปรียบเทียบ เมื่อพิจารณาเป็นด้าน สามารถสรุปได้ว่า ค่าความถูกต้องเพิ่มขึ้นเฉลี่ยร้อยละ 8.93 ค่าความแม่นยำเพิ่มขึ้นเฉลี่ยร้อยละ 12.03 ค่าความระลึกเพิ่มขึ้นเฉลี่ยร้อยละ 8.91 และค่าความถ่วงดุลเพิ่มขึ้นเฉลี่ยร้อยละ 11.14

ตัวแบบการประเมินจากอัลกอริทึม J48 เป็นตัวแบบที่มีประสิทธิภาพในการพยากรณ์สูงที่สุด ซึ่งอัลกอริทึม J48 เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจโดยใช้หลักการทฤษฎีสารสนเทศค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่งข้อมูล โดยวิธีกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นการเลือกข้อมูลตามลำดับของตัวชี้วัดหรือค่าเกณฑ์ สูงที่สุดเป็นข้อมูลเริ่มต้นและข้อมูลถัดไปมีค่าลดหลั่นกันตามลำดับ (9) และเมื่อได้ปรับความสมดุลของข้อมูลแล้วทำให้ได้จำนวนชุดข้อมูลจากเดิม 500 ระเบียน เพิ่มเป็น 748 ระเบียน และทำให้ประสิทธิภาพการประเมินเพิ่มขึ้นอีก โดยค่าความถูกต้องร้อยละ 80.61 ค่าความแม่นยำร้อยละ 78.20 ค่าความ ระลึกร้อยละ 80.60 และค่าความถ่วงดุลร้อยละ 78.00 จึงเป็นตัวแบบที่เหมาะสมต่อการนำไปประยุกต์ใช้สำหรับการประเมินความเสี่ยงต่อการเป็นโรคติดสุราต่อไป

อย่างไรก็ตามตัวแบบการพยากรณ์ที่ผู้วิจัยได้นำเสนอนั้นเป็นเพียงการใช้เทคนิคหนึ่งของการทำเหมืองข้อมูลจากหลายเทคนิคที่มีอยู่และยังมีข้อจำกัดด้านกลุ่ม ประชากรตัวอย่างที่เก็บในพื้นที่จังหวัดเดียวเท่านั้น ซึ่งในอนาคตหากมีการพัฒนางานวิจัยให้มีประสิทธิภาพมากขึ้น อาจจะต้องเพิ่มปริมาณข้อมูลที่นำ มาใช้ในการสร้างตัวแบบ หรือขยายบริเวณพื้นที่ในการเก็บข้อมูลจากกลุ่มประชากร ตัวอย่างเพิ่มจากจังหวัดในภาคอื่น ๆ เช่น ภาคเหนือ ภาคตะวันออกเฉียงเหนือ และภาคใต้ เพื่อให้เกิดความหลากหลายของข้อมูล และเพื่อให้ตัวแบบสามารถนำไปใช้ได้อย่างมีประสิทธิภาพสูงสุด

References

Wichaidit, W., McNally, E., Saenam, D., & Asanangkornchai, S., (2017). *Alcohol consumption in Thai society in 2017. Center for Alcohol Studies (CAS).* Epidemiology Unit, faculty of Medicine, Prince of Songkhla University. Retrieved from http://cas.or.th/cas/?p=6105. (in Thai)

Alcoholic beverages. (2020). *Thai Encyclopedia for Youth*. Retrieved from https://bit.ly/3sNXDlk. (in Thai)

Cartoon, A. (2012). *Drinker*. Retrieved from http://lovedring.blogspot.com/2012/11/blog-post.html. (in Thai)

Be addicted to alcohol. (2012). *Faculty of medicine*. Chulalongkorn University. Retrieved from https://bit.ly/3niGnmX. (in Thai)

Definition Of Alcoholism. (2017). *PobPad.* Retrieved from https://bit.ly/3sNAS0T. (in Thai)

Boonserm, K. (2003). *Algorithms for Data Mining. Engineering Computer*. Chulalongkorn University. Retrieved from https://www.cp.eng.chula.ac.th/~boonserm/publication/AlgoDataMining.pdf. (in Thai)

Atcharaporn, J. (2016). *Developing Information Systems to Forecast the Number of New Studentsv Using Decision Tree Classification Rules.* Business Administration and Accountancy Roi et Rajabhat University.(2), 267. (in Thai)

Paranya, P. (2016). Improving Decision Tree Technique in Imbalanced Data Sets Using SMOTE for Internet Addiction Disorder Data. *Information Technology Journal. (12)*, 57. (in Thai)

Pattanaphong, D. & Jaree, T., (2021). The Comparison of Performance Models for Predicting Students Success in Vocation Education. *Informatics Faculty Informatics Mahasarakham University. 37*(1), 383. (in Thai)

Chawla N. V., Bowyer K. W., Hall L. O., & Kegelmayer W. P.,(2002). SMOTE: Synthetic Minority Over Sampling Technique. *Journal of Artificial Intelligent Research*. (16). 323.