การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุล โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูล การเป็นโรคติดอินเทอร์เน็ต

Improving Decision Tree Technique in Imbalanced Data Sets Using SMOTE for Internet Addiction Disorder Data

ภรัณยา ปาลวิสุทธิ์ (Paranya Palwisut)*

บทคัดย่อ

โรคติดอินเทอร์เน็ตเป็นพฤติกรรมที่ผู้ใช้ทั่วไปที่มีสังคมอยู่ในโลกออนไลน์ให้เวลากับการใช้อินเทอร์เน็ตในปริมาณมากเกินไปจนทำให้เกิดปัญหาขึ้นหลายด้าน ไม่ว่าจะเป็นปัญหาทางการเรียน ความสัมพันธ์กับบุคคลอื่น ในการวิจัยครั้งนี้จึงมีจุดมุ่งหมายเพื่อที่จะพัฒนาตัวแบบสำหรับพยากรณ์การเป็นโรคติดอินเทอร์เน็ตในเยาวชน เนื่องจากสัดส่วนการใช้อินเทอร์เน็ตสูงที่สุดมีอายุระหว่าง 15 ถึง 24 ปี หลายคนขาดการควบคุมตัวเอง และผลจากการวิเคราะห์ข้อมูลพบว่า ข้อมูลมีความไม่สมดุลของกลุ่มข้อมูล โดยมีจำนวนกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งเป็นจำนวนมาก ผู้วิจัยจึงได้ทำการแก้ปัญหาโดยปรับความสมดุลของข้อมูลด้วยวิธีเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-sampling Technique : SMOTE) แล้วพัฒนาตัวแบบด้วยเทคนิค ต้นไม้ตัดสินใจ J48, ID3, LMT, CART และ Random Forest โดยใช้ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ และได้ใช้ค่าความแม่นยำ (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity) ในการวัดประสิทธิภาพการพยากรณ์ของตัวแบบ ผลการทดลองประสิทธิภาพในการพยากรณ์ของตัวแบบ พบว่า เทคนิค Random Forest สามารถพยากรณ์ได้ดีกว่า J48 ID3 LMT และ CART ซึ่งมีค่าความแม่นยำร้อยละ 87.15 ค่าความไว ร้อยละ 85.89 และค่าความจำเพาะ ร้อยละ 87.53 ตามลำดับ

คำสำคัญ: โรคติดอินเทอร์เน็ต ต้นไม้ตัดสินใจ การสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Abstract

Internet Addiction Disorder is behavior of users in social online community that have spent excessive time on internet activities. This leads to many problems such as learning problem, relationship problem. This research aims to development of model for prediction of Internet Addiction Disorder of adolescents because many internet users are between the ages of 15 - 24 and many lack self-control. After analyzing the data were found class imbalanced problem. In order to improve quality of data, SMOTE were used to increase the minority class. Then decision trees J48, Iterative Dichotomiser 3 (ID3), Logistic Model Trees (LMT), Classification and Regression Trees (CART) and Random Forest (RF) were used to build the prediction models. Moreover, 10-fold cross validation ware utilized to split the data into the training and test set. This research has measured performance models with accuracy, sensitivity and the specificity. Experimental result demonstrated that Random Forest superior to J48, ID3, LMT and CART with accuracy 87.15%, sensitivity 85.89% and specificity 87.53% respectively.

Keywords: Internet Addiction Disorder, Decision Tree, SMOTE

1. บทนำ

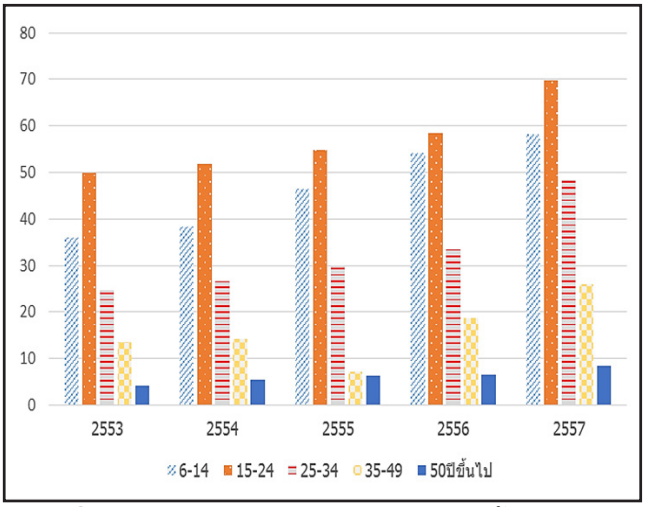
ปัจจุบันอินเทอร์เน็ตได้กลายเป็นส่วนหนึ่งในชีวิตประจำวันของมนุษย์ ต่างนำอินเทอร์เน็ตเข้ามาช่วยในการทำงาน

*สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

หลายๆ ด้าน ไม่ว่าจะเป็นด้านการศึกษา การทำธุรกิจ การหาความรู้ รวมถึงการหาความบันเทิง และจากการสำรวจการมีการใช้เทคโนโลยีสารสนเทศและการสื่อสารในครัวเรือน พ.ศ. 2557 สำนักงานสถิติแห่งชาติ [1] เมื่อพิจารณาการใช้อินเทอร์เน็ตตามกลุ่มอายุต่างๆ พบว่า กลุ่มที่มีสัดส่วนการใช้อินเทอร์เน็ตสูงที่สุดคือ กลุ่มอายุระหว่าง 15-24 ปี มีสัดส่วนการใช้อินเทอร์เน็ตสูงสุด ถึงร้อยละ 69.7 ซึ่งเป็นช่วงของการศึกษาชั้นมัธยมศึกษาตอนปลายและอุดมศึกษา รองลงมาคือ กลุ่มอายุ 6-14 ปีร้อยละ 58.2 กลุ่มอายุ 25-34 ปีร้อยละ 48.5 กลุ่มอายุ 35-49 ปีร้อยละ 25.9 และต่ำสุดในกลุ่มอายุ 50 ปีขึ้นไป ร้อยละ 8.4 ดังตารางที่ 1 และภาพที่ 1

ตารางที่ 1 *ร้อยละของประชากรอายุ 6 ปีขึ้นไปที่ใช้อินเทอร์เน็ต จำแนกตามกลุ่มอายุ พ.ศ. 2553-2557 [1]*

ปี	กลุ่มอายุ (ปี)				
	6-14	15-24	25-34	35-49	50 ปีขึ้นไป
2553	35.9	50.0	24.6	13.6	4.2
2554	38.3	51.9	26.6	14.3	5.5
2555	46.5	54.8	29.7	7.1	6.2
2556	54.1	58.4	33.5	18.7	6.6
2557	58.2	69.7	48.5	25.9	8.4



ภาพที่ 1 *กราฟเปรียบเทียบสัดส่วนการใช้อินเทอร์เน็ตตามกลุ่มอายุ*

ถึงแม้ว่าอินเทอร์เน็ตจะมีประโยชน์มากแต่หากใช้อินเทอร์เน็ตในทางที่ผิดอาจสร้างปัญหาให้กับสังคมได้อย่างมากมายเช่นกัน เนื่องจาก ผู้ใช้อินเทอร์เน็ตไม่จำเป็นต้องเปิดเผยตัวตนที่แท้จริง สามารถแสดงความรู้สึกแบบที่ไม่สามารถกระทำได้ในชีวิตจริง จนทำให้บางคนเกิดความรู้สึกยึดติดและพึ่งพิงกับการใช้อินเทอร์เน็ต แต่เมื่อใช้อินเทอร์เน็ตในปริมาณมากเกินไปจนทำให้เกิดความเสียหายกับชีวิต การเรียน หน้าที่การงานและครอบครัว จนไม่สามารถควบคุมได้จะเรียกพฤติกรรมการใช้ลักษณะนี้ว่าการเสพติดอินเทอร์เน็ต (Internet Addiction) [2] หรือเรียกว่า โรคติดอินเทอร์เน็ต (Internet Addiction Disorder : IAD) ซึ่งในปัจจุบันพบมากในกลุ่มเยาวชน และในปัจจุบันนี้มีโรคที่เกิดจากการใช้อินเทอร์เน็ตอุบัติขึ้นมาใหม่หลายโรค เช่น โรคติดข่าวสาร (FOMO - Fear to Miss Out) [3] ซึ่งเป็นอาการกลัวการถูกทิ้งจากข้อมูล สังคม และ เพื่อน เป็นต้น ถึงแม้ว่าในปัจจุบันจะมีการศึกษาเรื่องพฤติกรรมการติดอินเทอร์เน็ตอย่างหลากหลายแต่โดยส่วนใหญ่เป็นการศึกษาในเรื่องปัจจัยที่ส่งผลต่อพฤติกรรมการติดอินเทอร์เน็ต และปัญหาที่เกิดขึ้นจากการใช้อินเทอร์เน็ต [4], [5] โดยยังขาดงานวิจัยที่ใช้ในการจำแนกและการพยากรณ์ (Classification and Prediction) เกี่ยวกับการเป็นโรคติดอินเทอร์เน็ต

ดังนั้นผู้วิจัยจึงมีแนวคิดในการสร้างตัวแบบ (Model) เพื่อใช้ในการพยากรณ์การเป็นโรคติดอินเทอร์เน็ตสำหรับกลุ่มประชากรที่มีอายุระหว่าง 15 - 24 ปี ซึ่งปัจจุบันได้มีการนำความรู้เกี่ยวกับการทำเหมืองข้อมูล (Data Mining) มาใช้ในการจำแนกและการพยากรณ์ในงานด้านต่างๆ ไม่ว่าจะเป็นด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ด้านการแพทย์เพื่อจำแนกผู้ป่วยโรคต่างๆ เช่น โรคตับอักเสบ [6] โรคมะเร็งปากมดลูก [7] และโรคติดต่อทางเพศสัมพันธ์ [8] เป็นต้น และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) ก็เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลด้านการจำแนกและการพยากรณ์ที่ได้รับความนิยมจากนักวิจัยหลายๆ ท่าน เช่น ดิษฐพลและลี่ลี [9] ได้ใช้ต้นไม้ตัดสินใจในการวินิจฉัยโรคระบบการหายใจ เป็นต้น

ในการวิจัยครั้งนี้ได้นำข้อมูลที่ได้จากการสำรวจและศึกษาระดับพฤติกรรมการติดอินเทอร์เน็ตของประชากรกลุ่มเยาวชนหรือวัยรุ่น ซึ่งมีอายุระหว่าง 15 - 24 ปี [10] ในเขตอำเภอเมือง จังหวัดนครปฐม มาพัฒนาตัวแบบ จากข้อมูลที่ได้มามีปัญหาข้อมูลไม่สมดุล (Class Imbalance) ซึ่งเป็นปัญหาที่เกิดขึ้นได้กับข้อมูลทุกประเภท โดยทั่วไปจะแบ่งออกเป็น 2 อย่าง คือ ข้อมูลกลุ่มน้อย (Minority Class) จะ

เป็นข้อมูลที่มีอยู่จำนวนน้อยในชุดข้อมูล และข้อมูลกลุ่มมาก (Majority Class) จะเป็นกลุ่มข้อมูลที่มีอยู่จำนวนมากในชุดข้อมูล [11] เมื่อต้องการจำแนกประเภทข้อมูลผลลัพธ์ที่ได้เมื่อจำแนกข้อมูลจะถูกจำแนกไปในกลุ่มข้อมูลกลุ่มมาก ดังนั้น ตัวแบบพยากรณ์การเป็นโรคติดอินเทอร์เน็ตที่นำเสนอในงานวิจัยฉบับนี้ จึงนำเทคนิคต้นไม้ตัดสินใจร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-sampling Technique) หรือที่เรียกว่า SMOTE เพื่อปรับสมดุลข้อมูลและเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจให้จำแนกกลุ่มตัวอย่างข้างน้อยได้ดีขึ้น เมื่อทำการวิจัยเสร็จแล้วสามารถนำค่าที่พยากรณ์ได้มาใช้ประโยชน์ในการวางแผนและการตัดสินใจต่อไป

2. ทฤษฎีที่เกี่ยวข้อง

ผู้วิจัยได้มีการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการวิจัยมีรายละเอียดดังนี้

2.1 โรคติดอินเทอร์เน็ต

โรคติดอินเทอร์เน็ต [12] มีความเป็นมาจากการศึกษาของ ดร.คิมเบอร์ลี ยัง (Kimberly Young) มหาวิทยาลัยการแพทย์พิตซ์เบิร์ก (University of Pittsburgh Medical School) ได้นำเสนองานวิจัยเรื่อง “การเสพติดอินเทอร์เน็ต : การเกิดของโรคชนิดใหม่” (Internet Addiction : The Emergence of a New Disorder) ต่อที่ประชุมประจำปีของสมาคมจิตแพทย์อเมริกัน เมื่อประมาณ 16 ปีที่ผ่านมา ซึ่งปัจจุบันประเทศที่มีผู้ใช้อินเทอร์เน็ตเป็นจำนวนมาก เช่น สหรัฐอเมริกา เกาหลี จีน ต่างยอมรับตรงกันว่าการติดอินเทอร์เน็ตเป็นปัญหาด้านสาธารณสุข

โรคติดอินเทอร์เน็ต เป็นกลุ่มอาการทางจิตอย่างหนึ่ง ซึ่งเกิดจากการใช้อินเทอร์เน็ตในการเสพข้อมูลข่าวสารมากเกินไป แตกต่างกับการติดสื่ออื่น ๆ คือผู้ที่ใช้อินเทอร์เน็ตจะสามารถโต้ตอบกับผู้ที่เข้ามาใช้คนอื่น ๆ ได้ทันที ซึ่งทำให้โลกของอินเทอร์เน็ตเป็นเหมือนอีกโลกหนึ่งที่ทำให้ผู้ใช้รู้สึกว่าตัวเองมีตัวตนในโลกนั้นได้โดยปราศจากกฎเกณฑ์ไร้ขอบเขต

ในการเดินทาง และสร้างตัวตนตามที่ต้องการได้ บริการเครือข่ายสังคมออนไลน์ในอินเทอร์เน็ตที่มีลักษณะการสร้างเป็นสังคมเสมือน (Virtual Community) เช่น เฟซบุ๊ก ทวิตเตอร์ โปรแกรมแชท เว็บบอร์ด หรือแม้กระทั่งเกมออนไลน์ ถ้าหากผู้ใช้ยึดติดกับสังคมในโลกของอินเทอร์เน็ตจนแยกไม่ออกระหว่างโลกของความจริงและโลกเสมือน จะนำมาซึ่งสาเหตุของโรคติดอินเทอร์เน็ตได้ ก่อให้เกิดผลเสียกับระบบร่างกาย กระทบต่อการเรียน สภาพสังคม ทำให้เสียเวลา เสียความสัมพันธ์กับคนในครอบครัว เพื่อน และหน้าที่การงาน

2.2 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล [13] คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม

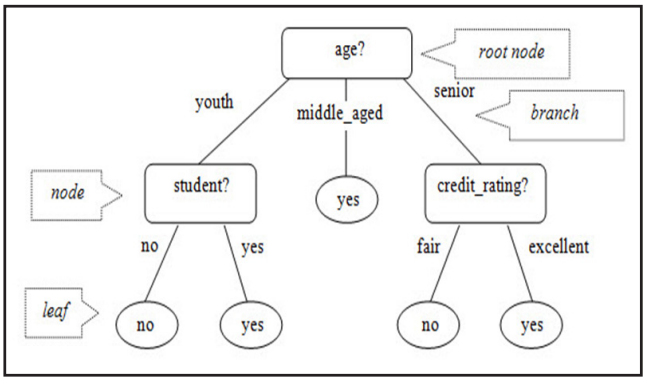
การทำเหมืองข้อมูลเปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล ซึ่งจุดประสงค์ของการทำเหมืองข้อมูลสามารถใช้ค้นข้อมูลสำคัญที่ปะปนกับข้อมูลอื่นๆ ในฐานข้อมูลที่ไม่ใช่แค่การสุ่มหา เรียกว่า KDD (Knowledge Discovery in Database) หรือการค้นหาข้อมูลด้วยความรู้ มีด้วยกัน 5 รูปแบบ คือ 1) การค้นหากฎความสัมพันธ์ 2) การจำแนกประเภทและการพยากรณ์ 3) การจัดกลุ่มข้อมูล 4) การหาค่าผิดปกติที่เกิดขึ้น 5) การวิเคราะห์แนวโน้ม

ซึ่งในงานวิจัยนี้ได้ใช้เทคนิคการจำแนกประเภทและการพยากรณ์ [14] ซึ่งเป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่จุดประสงค์ คือการสร้างตัวแบบการแยกคุณลักษณะ(Attribute) หนึ่งโดยขึ้นกับคุณลักษณะอื่น ตัวแบบที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้ โดยงานวิจัยฉบับนี้เลือกใช้เทคนิคต้นไม้ตัดสินใจ

2.3 เทคนิคต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ [15] เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ ซึ่งเมื่อมีข้อมูลที่ต้องการจัดกลุ่มก็จะนำคุณลักษณะต่างๆ ของข้อมูลนั้นไปเทียบกับเส้นทางในต้นไม้จนกระทั่งคลาสปลายทางซึ่งก็คือกลุ่มของข้อมูลที่เหมือนกัน ภายในต้นไม้จะประกอบไปด้วยโหนด (node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ เป็นตัวทดสอบ กิ่งของต้นไม้

(branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (class) ก็คือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก (root node) โครงสร้างของต้นไม้ตัดสินใจแสดงดังภาพที่ 2



ภาพที่ 2 *โครงสร้างต้นไม้ตัดสินใจ [6]*

ต้นไม้ตัดสินใจมีค่าวัดความสามารถของการจัดกลุ่มของแต่ละคุณลักษณะหรือปัจจัย ดังต่อไปนี้

Gini Index ค่าที่บ่งบอกว่าคุณลักษณะหรือปัจจัยใดควรนำมาใช้เป็นคุณลักษณะในการแบ่งกลุ่มของอัลกอรึธึม J48 และ CART

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

Entropy ค่าคาดคะเนของข้อมูลเป็นค่าที่แยกโดยใช้ลักษณะประจำของอัลกอริทึม ID3

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

โดยที่

t_i คือ คุณลักษณะที่นำมาวัดค่า Entropy

$P(t_i)$ คือ สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

ซึ่งแต่ละอัลกอริทึมจะให้ผลของโครงสร้างต้นไม้ตัดสินใจที่แตกต่างกันไป เทคนิคต้นไม้ตัดสินใจที่นำมาใช้ในงานวิจัยมีดังนี้

2.3.1 J48 หรืออัลกอริทึมของ C4.5 เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจจากกลุ่มของข้อมูลฝึกสอนโดยใช้ความถูกต้องของแต่ละคุณลักษณะของข้อมูล เพื่อใช้เป็นการตัดสินใจแบ่งกลุ่มข้อมูลกลุ่มย่อยๆ โดยพิจารณาจากค่าความแตกต่างใน Entropy ผลลัพธ์จากการเลือกคุณลักษณะสำหรับแบ่งกลุ่มข้อมูล ด้วยค่า Normalized information gain ที่สูงสุดนั้นคือการสร้างการตัดสินใจ รายละเอียดศึกษาได้จากเอกสารอ้างอิง [16]

2.3.2 Iterative Dichotomiser 3 (ID3) เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจโดยใช้หลักการทฤษฎีสารสนเทศ ค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการแบ่งข้อมูลโดยวิธีกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นการเลือกข้อมูลตามลำดับของตัวชี้วัดหรือค่าเกน (Gain) สูงที่สุดเป็นข้อมูลเริ่มต้นและข้อมูลถัดไปมีค่าลดหลั่นกันตามลำดับ รายละเอียดศึกษาได้จากเอกสารอ้างอิง [17]

2.3.3 Logistic Model Trees (LMT) เป็นการรวมกันของเทคนิคต้นไม้ (Trees) และการถดถอยโลจิสติก (Logistic Regression) รายละเอียดศึกษาได้จากเอกสารอ้างอิง [18]

2.3.4 Classification and Regression Trees (CART) เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจแบบ Binary ซึ่งประกอบด้วย กิ่งหรือแขนง 2 กิ่งสำหรับแต่ละโหนด เทคนิคนี้จะทำการแบ่งระเบียนในชุดข้อมูลฝึกสอนออกเป็น ระเบียนย่อยที่ให้ค่าเป้าหมายที่เหมือนกัน รายละเอียดศึกษาได้จากเอกสารอ้างอิง [19]

2.3.5 Random Forest (RF) เป็นชุดของการจำแนกประเภทแบบไม่ตัดแต่งกิ่ง (unpruned) หรือต้นไม้ถดถอย (Regression Trees) ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูลและคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจ ซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก เรียกข้อมูลส่วนนี้ว่า Out-of-Bag (OOB) จะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ รายละเอียดศึกษาได้จากเอกสารอ้างอิง [20]

2.4 เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

Synthetic Minority Over-sampling Technique: SMOTE [21] เป็นเทคนิคที่ใช้ในการแก้ปัญหาที่ต้องการจำแนกข้อมูลไม่สมดุล ซึ่งข้อมูลมีจำนวนตัวอย่างแตกต่างกันมากในแต่ละคลาส เมื่อทำการจำแนกประเภท จะทำให้มีการเรียนรู้แต่ข้อมูลกลุ่มที่มาก ผลที่ได้ก็จะจำแนกไปในข้อมูลกลุ่มมาก วิธี SMOTE เป็นวิธีการเพิ่มจำนวนข้อมูลประเภทที่มีข้อมูลน้อย ให้เพิ่มปริมาณข้อมูลใกล้เคียงกับประเภทที่มีมากที่สุด โดยสุ่มค่าขึ้นมาหนึ่งค่า และหาค่าระยะห่างระหว่างค่าที่เลือกกับทุกๆ ค่า เลือกค่าที่ใกล้เคียงที่สุด เช่น กำหนดไว้ 5 ค่า สุ่มค่าจากที่เลือก 1 ใน 5 หาค่าอยู่ระหว่างค่าที่เลือกตอนแรกและค่าที่สุ่มมาตอนหลัง เพื่อนำค่าที่ได้มาเพิ่มจำนวนข้อมูล 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

2.5 การวัดประสิทธิภาพ

ในการคำนวณประสิทธิภาพของตัวแบบได้ใช้ Confusion Matrix คือ ตารางสรุปจำนวนข้อมูลที่ตัวแบบมีการจำแนกได้อย่างถูกต้องและไม่ถูกต้อง ดังแสดงในภาพที่ 3

Know Class. d	0	1	2	⋯	*j*
0	TP	FN	FN	FN	FN
1	FP	TN	FN	FN	FN
2	FP	FN	TN	FN	FN
⋮	FP	FN	FN	TN	FN
j	FP	FN	FN	FN	TN

ภาพที่ 3 *ภาพประกอบ TP, FP, TN, และ FN สำหรับ class 0 โดยใช้ multi-class confusion [22]*

สำหรับวิธีการวัดประสิทธิภาพของวิธีการที่นำเสนอมีรายละเอียดดังนี้ [23]

1) ค่าความแม่นยำ (Accuracy) เป็นการวัดประสิทธิภาพการพยากรณ์ของตัวแบบโดยรวม ดังสมการที่ 4

2) ค่าความไว (Sensitivity) หรือ True-Positive Rate เป็นค่าความน่าจะเป็นหรืออัตราส่วนการพยากรณ์ถูกต้องในกลุ่มที่สนใจ ดังสมการที่ 5

3) ค่าความจำเพาะ (Specificity) หรือ True-Negative Rate เป็นค่าความน่าจะเป็นหรืออัตราส่วนการพยากรณ์ถูกต้องในกลุ่มอื่นๆ ดังสมการที่ 6

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \quad (4)$$

$$Sensitivity = \frac{TP}{TP+FP} \times 100 \quad (5)$$

$$Specificity = \frac{TN}{TN+FN} \times 100 \quad (6)$$

โดยที่

TP คือ ค่าที่พยากรณ์ถูกต้องในกลุ่มที่สนใจ

FP คือ ค่าที่พยากรณ์ผิดในกลุ่มที่สนใจ

TN คือ ค่าที่พยากรณ์ถูกต้องในกลุ่มอื่นๆ

FP คือ ค่าที่พยากรณ์ผิดในกลุ่มอื่นๆ

2.6 วิธีการวิเคราะห์ความแม่นตรงของตัวแบบ K-fold cross-validation

การตรวจสอบไขว้กัน (Cross Validation) [24] เป็นวิธีการตรวจสอบค่าความผิดพลาดในการคาดการณ์ของตัวแบบ โดยพื้นฐานของวิธีการการตรวจสอบไขว้กันคือ การสุ่มตัวอย่างโดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำการตรวจสอบไขว้กันมักถูกใช้เป็นตัวเลือกในการกำหนดตัวแบบ ในกรณีการทำ K - fold cross-validation จะแบ่งข้อมูลออกเป็น K ชุดเท่าๆ กัน เช่น 5 ชุด จะทำการคำนวณค่าความผิดพลาด 5 รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล 5 ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก 4 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ ตัวอย่างดังภาพที่ 4

lteration 1: train on	2	3	4	5	test on	1
lteration 1: train on	1	3	4	5	test on	2
lteration 1: train on	1	2	4	5	test on	3
lteration 1: train on	1	2	3	5	test on	4
lteration 1: train on	1	2	3	4	test on	5

ภาพที่ 4 *5 - fold Cross Validation [24]*

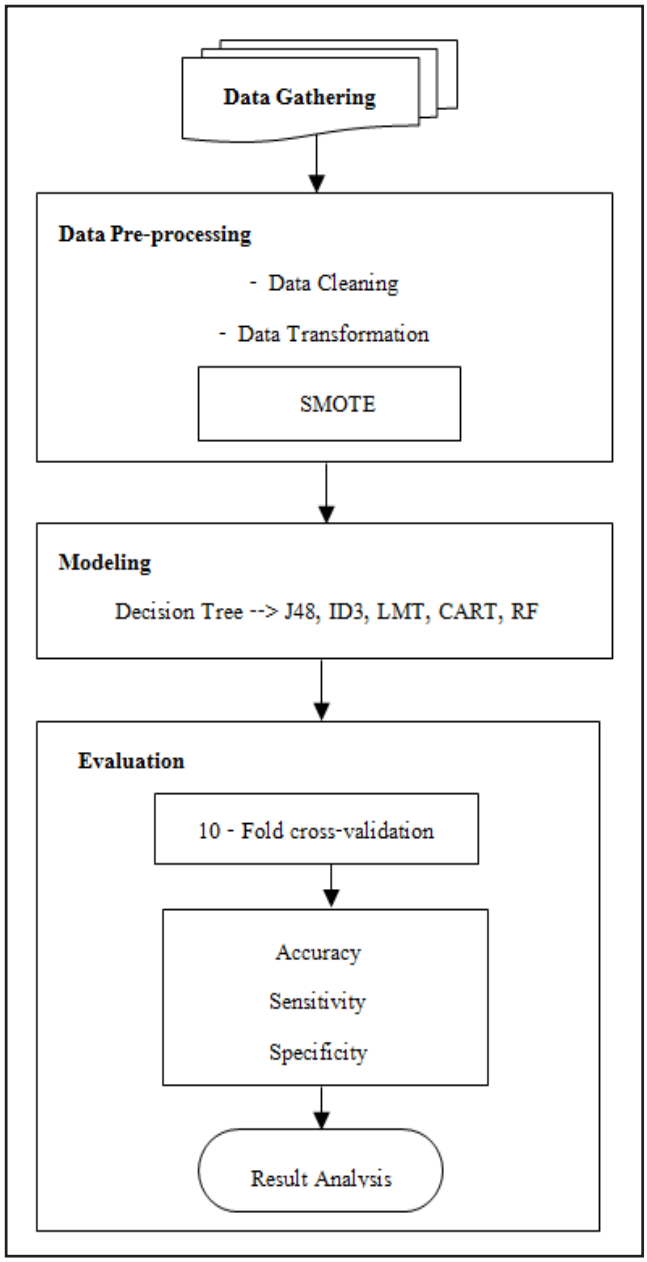
3. วิธีการดำเนินการวิจัย

ในการศึกษาวิจัยครั้งนี้ นำเสนอตัวแบบการพยากรณ์การเป็นโรคติดอินเทอร์เน็ต ซึ่งใช้เทคนิคต้นไม้ตัดสินใจร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย โดยได้มีดำเนินการ 4 ขั้นตอนดังต่อไปนี้ 1) รวบรวมข้อมูลจากแบบสอบถาม 2) การเตรียมข้อมูลสำหรับทำเหมืองข้อมูล 3) พัฒนาตัวแบบพยากรณ์ และ 4) การตรวจสอบประสิทธิภาพของตัวแบบ

ภาพการทำงานโดยรวมแสดงดังภาพที่ 5 โดยมีรายละเอียดดังนี้

3.1 การรวบรวมข้อมูลจากแบบสอบถาม

จากการสำรวจการการใช้อินเทอร์เน็ต พบว่า กลุ่มที่ใช้อินเทอร์เน็ตมากที่สุด คือ กลุ่มเยาวชนอายุระหว่าง 15 - 24 ปี โดยมีสัดส่วนการใช้อินเทอร์เน็ตสูงสุด ซึ่งสูงถึงร้อยละ 69.7 และปัญหาการเสพติดอินเทอร์เน็ต ในปัจจุบันพบมากขึ้นตลอดในกลุ่มเยาวชน ดังนั้น หากมีการนำเอาข้อมูลมาสร้างตัวแบบในการพยากรณ์ความเสี่ยงต่อการเป็นโรคติดอินเทอร์เน็ต จะทำให้สามารถใช้พยากรณ์ความเสี่ยงและเฝ้าระวังการการเสพติดอินเทอร์เน็ตในเด็กและเยาวชนได้



***ภาพที่ 5** ขั้นตอนการดำเนินงานวิจัย*

โดยกลุ่มตัวอย่างที่ใช้ในงานวิจัยเป็นกลุ่มเยาวชน ซึ่งมีอายุระหว่าง 15 - 24 ปี ในเขต อำเภอเมือง จังหวัดนครปฐม จากการเก็บรวบรวมข้อมูล ระหว่างวันที่ 3 - 25 สิงหาคม พ.ศ. 2558 และการเก็บรวบรวมข้อมูลได้จากการสำรวจโดยการใช้แบบประเมินอาการติดอินเทอร์เน็ต (Internet Addiction Test) [25] ของศูนย์โรคติดอินเทอร์เน็ต (The Center for Internet Addiction)

3.2 การเตรียมข้อมูลสำหรับทำเหมืองข้อมูล

ชุดข้อมูลที่ใช้เป็นข้อมูลที่ได้จาก การเก็บรวบรวมข้อมูลประเมินการติดอินเทอร์เน็ต ของกลุ่มเยาวชน ซึ่งมีอายุระหว่าง 15 - 24 ปี ในเขต อำเภอเมือง จังหวัดนครปฐม ในการเก็บรวบรวมข้อมูลได้ผลการประเมินการติดอินเทอร์เน็ตแล้วนำส่วนของข้อมูลส่วนตัวของผู้ตอบแบบสอบถามมาสร้างตัวแบบในการพยากรณ์ความเสี่ยงต่อการเป็นโรคติดอินเทอร์เน็ต โดยมีรายละเอียดข้อมูล คือ มีจำนวนคุณลักษณะทั้งหมด 15 คุณลักษณะ และเมื่อได้วิเคราะห์ข้อมูล จะพบว่ามีจำนวนระเบียนที่สมบูรณ์ทั้งหมด 880 ระเบียน จาก 892 ระเบียน ซึ่งระเบียนที่มีข้อมูลที่ไม่สมบูรณ์และมีค่าที่หายไปมีจำนวน 12 ระเบียนจึงได้ใช้วิธีตัดระเบียน (Ignore the tuple) เหล่านั้นทิ้ง

ซึ่งคุณลักษณะข้อมูลที่ใช้ปรากฏดังตารางที่ 2

***ตารางที่ 2** รายละเอียดคุณลักษณะ ที่นำมาสร้างตัวแบบ*

ลำดับ	ชื่อคุณลักษณะ	คำอธิบาย
1	Gender	เพศนักศึกษา
2	GPA	ผลการเรียน
3	Status	สถานภาพครอบครัว
4	Sibling	จำนวนพี่น้อง
5	Income	รายได้เฉลี่ยต่อเดือน
6	Occ_fa	อาชีพบิดา
7	Occ_Ma	อาชีพมารดา
8	Living	บุคคลที่พักอาศัยอยู่ด้วย
9	Habitat	ลักษณะของที่พักอาศัย
10	Net_Cafe	ร้านอินเทอร์เน็ตในแหล่งที่อยู่
11	Time	ระยะเวลาที่ใช้อินเทอร์เน็ตถึงปัจจุบัน
12	Hobbies	งานอดิเรก
13	cigarette	การสูบบุหรี่
14	Drink	ดื่มแอลกอฮอล์
15	UseNet	แหล่งที่ใช้อินเทอร์เน็ตบ่อยที่สุด
16	Class	1 = ไม่เสี่ยงเป็น IAD 2 = เสี่ยงเป็น IAD 3 = เป็น IAD

จากตารางที่ 2 คุณลักษณะที่ 16 เป็นคุณลักษณะที่เป็นกลุ่มของข้อมูลที่ใช้ในการจำแนกข้อมูล ซึ่งกลุ่มที่แบ่งได้จากผลการประเมินจากแบบประเมินอาการติดอินเทอร์เน็ต ของศูนย์โรคติดอินเทอร์เน็ต โดยแบ่งออกได้ 3 กลุ่มได้แก่

กลุ่ม 1 คือ กลุ่มไม่เสี่ยงเป็น IAD

กลุ่ม 2 คือ กลุ่มเสี่ยงเป็น IAD

กลุ่ม 3 คือ กลุ่มเป็น IAD

เมื่อพิจารณาข้อมูลที่ได้จากแบบประเมินแล้วนำข้อมูลมาทำการศึกษา จะพบว่าข้อมูลที่ได้มีปัญหาข้อมูลไม่สมดุล ดังนี้

กลุ่มไม่เสี่ยงเป็น IAD จำนวน 200 ระเบียน

กลุ่มเสี่ยงเป็น IAD จำนวน 623 ระเบียน

กลุ่มเป็น IAD จำนวน 57 ระเบียน

ดังนั้น ผู้วิจัยจึงนำวิธี SMOTE มาใช้เพื่อปรับสมดุลข้อมูล คือเพิ่มจำนวนข้อมูลกลุ่มที่มีข้อมูลน้อย ก่อนที่นำเข้าสู่กระบวนการต่อไป และทำการปรับค่าพารามิเตอร์ให้เหมาะสม ผลการทดลองพบว่า ขนาดชุดข้อมูลที่สามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีที่สุด คือ ร้อยละ 600 เพิ่มข้อมูลขึ้นจำนวน 342 ระเบียน รวมเป็น 1,222 ระเบียน ดังตารางที่ 3 ทั้งนี้จะทำการทดลองเปรียบเทียบประสิทธิภาพกับข้อมูลที่ไม่ผ่านวิธี SMOTE

ตารางที่ 3 *ข้อมูลที่ผ่านวิธี SMOTE แก้ปัญหาข้อมูลไม่สมดุล*

กลุ่ม	Original	SMOTE
กลุ่มไม่เสี่ยงเป็น IAD	200	200
กลุ่มเสี่ยงเป็น IAD	623	623
กลุ่มเป็น IAD	57	399
จำนวนทั้งหมด	880	1,222

3.3 การพัฒนาตัวแบบพยากรณ์

ในการพัฒนาตัวแบบได้ทำการเปรียบเทียบหาเทคนิคและตัวแบบที่สามารถพยากรณ์ได้แม่นยำและมีประสิทธิภาพสูงที่สุดโดยวัดจากค่าความถูกต้องของแต่ละตัวแบบ ซึ่งเทคนิคต้นไม้ตัดสินใจที่นำมาใช้เปรียบเทียบประสิทธิภาพในงานวิจัยได้แก่ J48, ID3, LMT, CART และ RF

3.4 การวัดประสิทธิภาพของตัวแบบ

ในงานวิจัยนี้ใช้การทดสอบความแม่นตรงของตัวแบบแบบ 10-fold Cross-Validation เพื่อให้ข้อมูลทุกตัวมีโอกาสเป็นชุดทดสอบและชุดสอน โดยแบ่งข้อมูลออกเป็นชุดข้อมูลสอน (Training Data) และข้อมูลทดสอบ (Testing Data) โดยแบ่งข้อมูลออกเป็น 10 ส่วนเท่าๆ กัน ใช้ 9 ส่วนเป็นชุดข้อมูลสอน และอีก 1 ส่วนเป็นชุดข้อมูลทดสอบ ซึ่งจะทำสลับกันจนครบทั้งหมด 10 รอบ ในส่วนการวัดประสิทธิภาพของตัวแบบในงานวิจัยนี้ ได้วัดจากค่าความแม่นยำ ค่าความไว และค่าความจำเพาะ

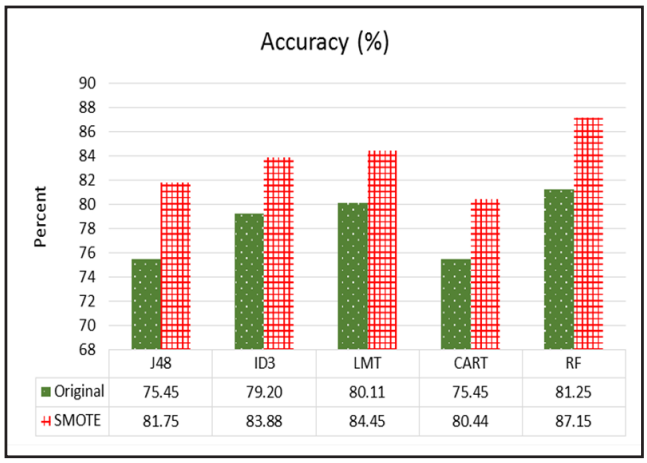
4. ผลการดำเนินงาน

ในงานวิจัยนี้ได้นำข้อมูลมาปรับสมดุลด้วยวิธีการ SMOTE และนำข้อมูลเข้าสู่กระบวนการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจได้แก่ J48, ID3, LMT, CART และ RF เพื่อพยากรณ์การเป็นโรคติดอินเทอร์เน็ตในกลุ่มเยาวชน การดำเนินงานได้ทำการเปรียบเทียบระหว่างข้อมูลที่ไม่ได้ปรับสมดุลกับข้อมูลที่ผ่านการปรับสมดุล โดยใช้ค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ เพื่อวัดประสิทธิภาพของตัวแบบ ได้ผลการดำเนินงาน ดังนี้

1) ค่าความแม่นยำผลการเปรียบเทียบค่าความแม่นยำสามารถสรุปผลได้ดังตารางที่ 4 และภาพที่ 6

ตารางที่ 4 *ค่าความแม่นยำ*

Algorithms	Original	SMOTE
J48	75.45	81.75
ID3	79.20	83.88
LMT	80.11	84.45
CART	75.45	80.44
RF	81.25	87.15



ภาพที่ 6 *ค่าความแม่นยำ*

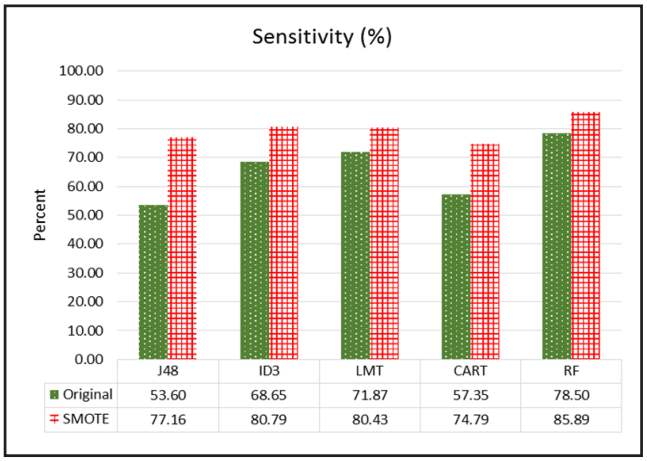
จากตารางที่ 4 และ ภาพที่ 6 การวัดค่าความแม่นยำซึ่งถือว่าเป็นประสิทธิภาพโดยรวมของตัวแบบในการพยากรณ์ ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี SMOTE ทำให้ค่าความแม่นยำของตัวแบบสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค RF ให้ค่าความแม่นยำสูงสุดร้อยละ 87.15

ตามด้วยเทคนิค LMT ให้ค่าความแม่นยำร้อยละ 84.45 เทคนิค ID3 ให้ค่าความแม่นยำร้อยละ 83.88 เทคนิค J48 ให้ค่าความแม่นยำร้อยละ 81.75 และตัวแบบที่สร้างด้วยเทคนิค CART ให้ค่าความแม่นยำน้อยที่สุดที่ร้อยละ 80.44

2) ค่าความไว ผลการเปรียบเทียบค่าความไว สามารถสรุปผลได้ดังตารางที่ 5 และภาพที่ 7

ตารางที่ 5 *ค่าความไว*

Algorithms	Original	SMOTE
J48	53.60	77.16
ID3	68.65	80.79
LMT	71.87	80.43
CART	57.35	74.79
RF	78.50	85.89



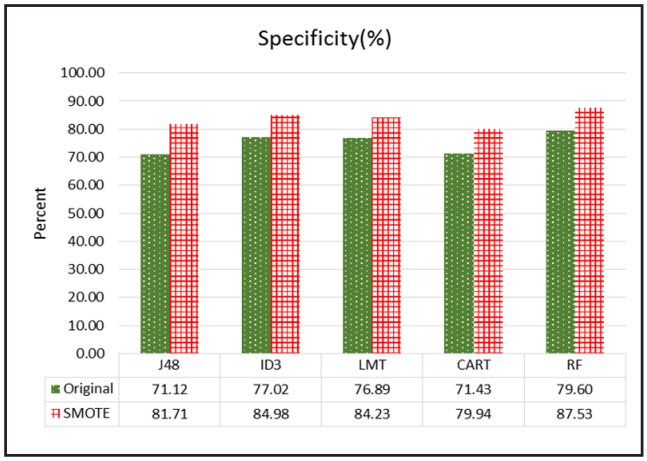
ภาพที่ 7 *ค่าความไว*

จากตารางที่ 5 และ ภาพที่ 7 การวัดค่าความไว เป็นการบอกถึงอัตราส่วนของการพยากรณ์ถูกต้องในกลุ่มที่สนใจ เช่น กลุ่มที่สนใจ คือ กลุ่มเป็น IAD หมายความว่าโอกาสที่ผู้ที่เป็นโรคติดอินเทอร์เน็ตจะได้รับผลการวินิจฉัยว่าเป็นโรคติดอินเทอร์เน็ต ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี SMOTE ทำให้ค่าความไวของตัวแบบสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค RF ให้ค่าความไวสูงสุดร้อยละ 85.89 ตามด้วยเทคนิค ID3 ให้ค่าความไวร้อยละ 80.79 เทคนิค LMT ให้ค่าความไวร้อยละ 80.43 เทคนิค J48 ให้ค่าความไวร้อยละ 77.16 และตัวแบบที่สร้างด้วย เทคนิค CART ให้ค่าความไวน้อยที่สุดที่ร้อยละ 74.79

3) ค่าความจำเพาะ ผลการเปรียบเทียบค่าความถูกต้องสามารถสรุปผลได้ดังตารางที่ 6 และภาพที่ 8

ตารางที่ 6 *ค่าความจำเพาะ*

Algorithms	Original	SMOTE
J48	71.12	81.71
ID3	77.02	84.98
LMT	76.89	84.23
CART	71.43	79.94
RF	79.60	87.53



ภาพที่ 8 *ผลการทดลอง*

จากตารางที่ 6 และ ภาพที่ 8 การวัดค่าความจำเพาะเป็นการบอกถึงอัตราส่วนการพยากรณ์ถูกต้องในกลุ่มอื่นๆ เช่นกลุ่มอื่นๆ คือ กลุ่มไม่เสี่ยงเป็น IAD และกลุ่มเสี่ยงเป็น IAD หมายความว่า โอกาสที่ผู้ที่ไม่ได้เป็นโรคติดอินเทอร์เน็ตจะได้รับผลการวินิจฉัยว่าไม่เป็นโรคติดอินเทอร์เน็ต ผลการทดลองพบว่า ข้อมูลที่ผ่านการปรับสมดุลด้วยวิธี SMOTE ทำให้ค่าความจำเพาะของตัวแบบสูงขึ้นทุกเทคนิค โดยเฉพาะเทคนิค RF ให้ค่าความจำเพาะสูงสุดร้อยละ 87.53 ตามด้วยเทคนิค ID3ให้ค่าความจำเพาะร้อยละ 84.94 เทคนิค LMT ให้ค่าความจำเพาะร้อยละ 84.23 เทคนิค J48 ให้ค่าความจำเพาะร้อยละ 81.71 และตัวแบบที่สร้างด้วยเทคนิค CART ให้ค่าความจำเพาะน้อยที่สุดที่ร้อยละ 79.94

5. สรุป

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อนำเสนอตัวแบบการพยากรณ์การเป็นโรคติดอินเทอร์เน็ตในกลุ่มเยาวชนหรือ

วัยรุ่น ซึ่งมีอายุระหว่าง 15 - 24 ปี ในเขต อำเภอเมือง จังหวัดนครปฐม ซึ่งเทคนิคการพยากรณ์ในงานวิจัยฉบับนี้ใช้เทคนิคต้นไม้ตัดสินใจและแก้ปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE จากการทดลอง พบว่า

1) การแก้ปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE สามารถเพิ่มประสิทธิภาพให้ตัวแบบเพิ่มขึ้นในทุกเทคนิคที่ได้นำมาเปรียบเทียบ เมื่อพิจารณาเป็นด้าน สามารถสรุปได้ว่า ค่าความแม่นยำเพิ่มขึ้นเฉลี่ยร้อยละ 5.24 ค่าความไวเพิ่มขึ้นเฉลี่ยร้อยละ 13.82 และค่าความจำเพาะเพิ่มขึ้นเฉลี่ยร้อยละ 8.47

2) ตัวแบบการพยากรณ์จากเทคนิค RF เป็นตัวแบบที่มีประสิทธิภาพในการพยากรณ์สูงที่สุด ซึ่งเทคนิค RF เป็นเทคนิคที่สร้างตัวแบบต้นไม้ตัดสินใจหลายๆ ตัวแบบมาช่วยในการหาคำตอบ สร้างตัวแบบโดยใช้การสุ่มข้อมูลจากชุดข้อมูลสอนออกมาเป็นหลายๆ ชุด ร่วมกับการสุ่มคุณลักษณะที่มีมาใช้ในการสร้างตัวแบบทำให้ตัวแบบที่สร้างขึ้นมามีลักษณะต้นไม้ตัดสินใจที่หลากหลาย หลังจากได้ตัวแบบมาชุดหนึ่งแล้วถึงนำไปพยากรณ์ข้อมูลที่ยังไม่รู้คำตอบ ซึ่งแต่ละตัวแบบก็จะให้คำตอบออกมา และในขั้นตอนสุดท้ายจะนำคำตอบเหล่านี้มารวมกันเพื่อดูว่าคำตอบไหนเหมาะสมที่สุด โดยอาจจะใช้วิธีการโหวต (vote) เพื่อเลือกคำตอบที่ตอบตรงกันมากที่สุด จากวิธีการสร้างตัวแบบเช่นนี้ทำให้เทคนิค RF มีประสิทธิภาพในการพยากรณ์ที่ดี และเมื่อได้ทำการปรับสมดุลของข้อมูลด้วยวิธี SMOTE แล้วทำให้ประสิทธิภาพการพยากรณ์เพิ่มขึ้นอีก โดยมีค่าความแม่นยำเท่ากับร้อยละ 87.15 ค่าความไวเท่ากับร้อยละ 85.89 และค่าความจำเพาะเท่ากับร้อยละ 87.53 จึงเป็นตัวแบบที่เหมาะสมต่อการนำไปประยุกต์ใช้สำหรับการพยากรณ์การเป็นโรคติดอินเทอร์เน็ตต่อไป

อย่างไรก็ตามตัวแบบการพยากรณ์ที่ผู้วิจัยได้นำเสนอนั้นเป็นเพียงการใช้เทคนิคหนึ่งของการทำเหมืองข้อมูลจากหลากหลายเทคนิคที่มีอยู่ และยังมีข้อจำกัดด้านกลุ่มประชากรตัวอย่างที่เก็บในพื้นที่จังหวัดเดียวเท่านั้น ซึ่งในอนาคตหากมีการพัฒนางานวิจัยให้มีประสิทธิภาพมากขึ้นอาจจะต้องเพิ่มปริมาณข้อมูลที่นำมาใช้ในการสร้างตัวแบบหรือขยายบริเวณพื้นที่ในการเก็บข้อมูลจากกลุ่มประชากรตัวอย่างเพิ่มจากจังหวัดในภาคอื่นๆ เช่น ภาคเหนือ ภาคตะวันออกเฉียงเหนือ และภาคใต้ เพื่อให้เกิดความหลากหลายของข้อมูลมากขึ้น และเพื่อให้ตัวแบบสามารถนำไปใช้ได้อย่างมีประสิทธิภาพสูงสุด

6. เอกสารอ้างอิง

[1] National Statistical Office. “The Household Survey on Information and Communication Technology in 2014.” *National Statistical Office*, Bangkok, Thailand, pp. 4, 2014.

[2] O. Wongkaewpotong. “Internet Addiction of Online Community.” *Executive Journal*, Vol. 31, No. 3, pp. 39-42, July - September 2011.

[3] D. Cioban. *Social Media FOMO– Fear of Missing Out*. Retrieved August 5, 2015, Available online at http://www.brandingmagazine.com /2012/07/24/social-media-fomo-fear-of-missing-out/.

[4] N. Kannika. “Online Gaming Habits of Children in Bangkok Metropolis and the Suburb in 2005.” *ABAC Poll*, Assumption University, Bangkok, Thailand, 2005.

[5] L. Ghassemzadeh, M. Shahraray, and A. Moradi. “Prevalence of internet addiction and comparison of internet addicts and non-addicts in Iranian high schools.” *CyberPsychology & Behavior*, Vol. 11, No. 6, pp. 731-733, December, 2008.

[6] S. Uhmn, D. H. Kim and J. Kim. “Chronic Hepatitis Classification using SNP data and Data Mining Techniques.” *IEEE computer society*, pp. 81–86, 2007.

[7] A. Madaruk. “Pap-smear Classification Using Efficient Second Order Neural Network Training Algorithm.” seminar papers in Computer Science, Khon Kaen University, Thailand, 2006.

[8] S. Boonlue, T. Kammanat and K. Kawsan. “Diagnosis preliminary for leukorrhea system using back-propagation neural network.” *The 5th National Conference on Computing and Information Technology (NCCIT’09)*, Thailand, pp. 163, 2009.

[9] D. Muntham and L. Ingsrisawang, “An Application of Decision Tree Algorithms for Diagnosis of the Respiratory System: A Case Study of Pranakorn

Sri Ayudthaya Hospital." *Journal of Health Systems Research*, Vol. 4, No.1, pp. 73-81, January - March 2010.

[10] S. Toonkul. *Health Promotion in Community Nursing and First Aid, Aroonprinting*. Bangkok, Thailand, 2002.

[11] K. Boonchuay, K. Sinapiromsaran and C. Lursinsap. "Minority Split and Gain Ratio for a Class Imbalance." *Eighth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 2060 - 2064, 26-28 July 2011.

[12] K. Loylert. *Internet Addiction Disorder*. Retrieved January 4, 2013, Available online at http://www.thaihealth.or.th.

[13] B. Kijsirikul. *Data Mining Algorithm*. Department of Computer Engineering,Faculty of Engineering, Chulalongkorn University, Thailand, 2002.

[14] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[15] C. Kaewchinporn. *Data Classification with Decision Tree and Clustering Techniques*. Thesis in Computer Science, King Mongkut's Institute of Technology Ladkrabang, Thailand, 2010.

[16] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[17] R. Quinlan. "Induction of decision trees." Machine Learning, Vpl. 1, No. 1, pp. 81-106, 1986.

[18] N. Landwehr, M. Hall and E. Frank. "Logistic Model Trees." *Machine Learning*, Vol. 95, No. 1-2, pp. 161-205, 2005.

[19] L. Breiman, J. H. Friedman, R. Olshen and C. J. Stone. *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.

[20] L. Breiman and R. Forests. *Machine Learning*, Vol. 45, No. 1, pp.5-32, 2001.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmayer. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligent Research*, pp.321-357, 2002.

[22] P. R. Perea, J. C. Ruiz, J. A. Perez Venzor, D. G. Chaparro, J. G. Rosiles. "LP-SVR Model Selection Using an Inexact Globalized Quasi-Newton Strategy." *Journal of Intelligent Learning Systems and Applications*, pp. 19-28, 2013.

[23] J. Thongkam, G. Xu, Y. Zhang and F. Huang. "Toward breast cancer survivability prediction models through improving training space." Expert Systems with Applications, pp.12200–9, 2009.

[24] Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Vol. 2, No. 12, pp. 1137–1143, 1995.

[25] The Centre for Internet Addiction. *Internet Addiction Test (IAT)*. Retrieved June 12, 2015, Available online at http://netaddiction.com/internet-addiction-test/.