Vol.2 No.1 (January-June 2020)

การประยุกต์ใช้การทำเหมืองข้อมูลในการทวนสอบปัจจัยในการกระทำผิดซ้ำของเด็กและเยาวชนไทย

เจษฎา ขจรฤทธิ์[1] และจิราภรณ์ ชมบุญ[2]

วันได้รับบทความ: 2 พฤษภาคม 2563 วันแก้ไข: 22 พฤษภาคม 2563 วันยอมรับเผยแพร่: 5 มิถุนายน 2563

บทคัดย่อ

ปัจจุบันการกระทำผิดของเด็กและเยาวชนมีแนวโน้มสูงขึ้น ทำให้หน่วยงานรัฐที่เกี่ยวข้องมีจำเป็นต้องเข้ามาดูแลแก้ไขปัญหานี้ กรมพินิจและคุ้มครองเด็กและเยาวชนมีหน้าที่ต้องฝึกอบรมเด็กและเยาวชนที่กระทำผิดให้มีสภาพจิตใจและแนวคิดที่ถูกต้องก่อนปล่อยคืนเข้าสู่สังคม อย่างไรก็ตามการกระทำผิดซ้ำของเด็กและเยาวชนเหล่านั้นยังคงเกิดขึ้นและเป็นปัญหาหนึ่งที่ต้องได้รับการดูแลอย่างต่อเนื่อง ดังนั้นจึงเกิดการศึกษาวิจัยตัวแบบด้านจิตวิทยาเพื่อหาเหตุปัจจัยของการกระทำผิดซ้ำของเด็กและเยาวชนเหล่านี้ งานวิจัยฉบับนี้เป็นส่วนหนึ่งของการทวนสอบเหตุปัจจัยของตัวแบบดังกล่าว โดยอาศัยข้อมูลที่มีการบันทึกไว้มาเป็นเครื่องมือในการทวนสอบตัวแบบด้วยวิธีการทำเหมืองข้อมูล ข้อมูลเด็กและเยาวชนที่บันทึกไว้มีจำนวน 258 ราย มีการกระทำผิดซ้ำ 99 รายจากทั้งหมด ข้อมูลประกอบด้วย 26 เหตุปัจจัย การทวนสอบนี้แบ่งเป็นสามการทดลอง ได้แก่ (1) การวัดประสิทธิภาพตัวแบบ (2) การวิเคราะห์กลุ่มข้อมูลย่อย และ (3) การคัดเลือกฟีเจอร์ที่มีอิทธิพลต่อการกระผิดซ้ำ ผลการทดลองแรกพบว่าเหตุปัจจัยของการกระทำผิดซ้ำยังไม่สามารถทำนายการกระทำผิดซ้ำของเด็กและเยาวชนได้อย่างมีประสิทธิภาพ สาเหตุที่เป็นไปได้คือเหตุปัจจัยเหล่านั้นไม่ถูกต้องหรือเกิดจากจำนวนของคลาสข้อมูลที่ไม่สมดุลกัน การทดลองที่สองพบว่าการจัดกลุ่มย่อยข้อมูลเพื่อลดความหลากหลายในตัวข้อมูลไม่สามารถเพิ่มประสิทธิภาพในการทำนายได้ ซึ่งหมายความว่าข้อมูลไม่ได้แบ่งออกเป็นกลุ่มย่อยที่ชัดเจน ส่วนการทดลองสุดท้ายแสดงผลการทดลองเชิงคุณภาพคือการคัดเลือกปัจจัยบางตัวที่มีผลต่อการกระทำผิดซ้ำ แบ่งเป็นปัจจัยเดี่ยวที่ควรพิจารณา ปัจจัยเดี่ยวที่ควรตัดออก และปัจจัยกลุ่มที่ควรพิจารณาร่วมกัน

คำสำคัญ: เด็กและเยาวชนไทย การกระทำผิดซ้ำ ตัวแบบทำนาย การทำเหมืองข้อมูล

[1] วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์, มหาวิทยาลัยธุรกิจบัณฑิตย์

[2] กลุ่มเรียนรู้ศาสตร์พระราชาและการพัฒนานวัตกรรม, กรมพินิจและคุ้มครองเด็กและเยาวชน (ผู้ประพันธ์บรรณกิจ)

 ลิขสิทธิ์โดย คณะสังคมศาสตร์ โรงเรียนนายร้อยตำรวจ

Vol.2 No.1 (January-June 2020)

An Application of Data Mining to Validate the Causes of Recidivism of Thai's Children and Juveniles

Jesada Kajornrit[1] & Jiraporn Chomboon[2]

Received: May 2, 2020 Revised: May 22, 2020 Accepted: June 5, 2020

Abstract

Presently, Thai's juvenile delinquency has continuously increased. Many involved state agencies have come to take care and correct this problem. Department of Juvenile Observation and Prevention, one of those involves, has responsibility to rehabilitate those juveniles before returning them to the social. However, the recidivism (habitual criminal activity) of juveniles has still continued and it must be taken care of. Thus, a study about the mental causes of recidivism of those juveniles has been established. This research is a part of whole works to validate the causes of the recidivism prediction model. By using data recorded and data mining techniques, the recidivism prediction model is able to validated, itself. The number of data of 258 occurrences (juveniles) consists of 99 recidivism occurrences. The features (causes) of data are 26. These validation works consist of 3 tests, that is, testing overall performance, testing group performance, and features selections. The first experimental results pointed out that the features (causes) of recidivism prediction model are inefficiently able to predict the recidivism of juveniles. That is because the features, possibly, do not substantially reflect to the recidivism, or the number of classes are imbalanced. In the second experiment, again, clustering do not reduce the heterogeneity in the data and do not improve performance. Finally, causes (features) that affect to recidivism are shown in the third experiment, included interested single features, uninterested single features, and interested grouped features.

Keywords: Thai Children and Juveniles, Recidivism, Prediction Model, Data Mining

[1] College of Innovative Technology and Engineering, Dhurakij Pundit University

[2] The King's Science and Innovation Development Group, Department of Juvenile Observation and Protection (Corresponding author)

RPCA-JCSC

บทนำ (Introduction)

ในปัจจุบันสภาพสังคมไทยมีการเปลี่ยนแปลงไปอย่างมากจากอดีตที่ผ่านมา ปัจจัยที่หลากหลายผลักดันให้สังคมไทยมีการเปลี่ยนแปลงไปอย่างต่อเนื่องและรวดเร็วทั้งในด้านดีและด้านที่ไม่ดี ทุกวันนี้ถ้าสังเกตจากข่าวโทรทัศน์ประจำวันจะพบว่าข่าวอาชญากรรมมีการนำเสนอมากขึ้นจนเห็นได้ชัด และถ้าสังเกตดูให้ละเอียดแล้วจะพบว่าบ่อยครั้งผู้กระทำผิดเหล่านั้นยังเป็นเพียงแค่เด็กและเยาวชน (ไทยโพสต์, 2562) ปัญหาเหล่านี้จำเป็นต้องได้รับการดูแลแก้ไขอย่างต่อเนื่อง ก่อนที่จะกลายเป็นปัญหาใหญ่จนเป็นเหตุให้คนในสังคมดำเนินชีวิตด้วยความหวาดระแวง

กรมพินิจและคุ้มครองเด็กและเยาวชน (กรมพินิจฯ) มีพันธกิจที่จะต้องบำบัดและฝึกอบรมเด็กและเยาวชนที่กระทำผิดให้มีสภาพความคิดและจิตใจที่ถูกต้องก่อนปล่อยคืนเข้าสู่สังคม อย่างไรก็ตามจากการติดตามเด็กและเยาวชนที่ได้รับการปล่อยคืนสู่สังคม พบว่าการกระทำผิดซ้ำของเด็กและเยาวชนเหล่านั้นยังคงเกิดขึ้นต่อเนื่อง ดังนั้นนักวิจัยด้านสังคมจึงเกิดการศึกษาตัวแบบทำนายเชิงจิตวิทยาเพื่อหาเหตุปัจจัยของการกระทำผิดซ้ำของเด็กและเยาวชนเหล่านี้ ตัวอย่างเช่นงานวิจัยของ จารุณี แซ่ตั้ง (2546) และ ณัฐวุฒิ ยอดระบำ (2549)

นักวิจัยของกรมพินิจฯ ได้ทำการเก็บรวบรวมข้อมูลสำหรับตัวแบบนี้ (Mallett et al., 2013, Tafrate & Mitchell, 2014) มาระยะหนึ่ง ข้อสังเกตุที่เกิดขึ้นคือเหตุปัจจัยเหล่านี้สามารถใช้ทำนายการกระทำผิดซ้ำของเด็กและเยาวชนไทยได้หรือไม่ มีหลายวิธีในการทวนสอบประสิทธิภาพในการทำนาย และหนึ่งในนั้นคือการใช้ข้อมูลที่เก็บได้มาเป็นตัวทวนสอบประสิทธิภาพในการทำนายของตัวแบบด้วยวิธีการทำเหมืองข้อมูล

เอกสารฉบับนี้นำเสนอถึงผลที่ได้จากการทวนสอบเหตุปัจจัยเหล่านั้นด้วยวิธีการทำเหมืองข้อมูลจากข้อมูลการกระทำผิดซ้ำของเด็กและเยาวชนภายในระยะเวลา 1 ปี หลังจากถูกปล่อยตัว ข้อมูลที่ใช้เป็นข้อมูลที่เก็บรวบรวมขึ้นจากศูนย์ฝึกอบรมแห่งหนึ่ง หัวข้อต่อจากนี้ประกอบด้วย วัตถุประสงค์ของการวิจัย เครื่องมือและวิธีการดำเนินงานวิจัย สรุปและอธิปรายผลการวิจัย ประเด็นข้อสังเกต และ บทสรุป ตามลำดับ

วัตถุประสงค์การวิจัย (Research Objectives)

การวิจัยนี้มีจุดประสงค์เพื่อทวนสอบประสิทธิภาพของเหตุปัจจัยที่นำไปสู่การกระทำผิดซ้ำของเด็กและเยาวชนไทย วิธีการคือใช้ข้อมูลที่รวบรวมไว้มาทำการทวนสอบความถูกต้องในการทำนายการกระทำผิดซ้ำด้วยวิธีการทำเหมืองข้อมูล ผลลัพธ์ที่ได้จะถูกนำไปใช้เป็นส่วนหนึ่งในประเมินการปรับปรุงตัวแบบทำนายการกระทำผิดซ้ำต่อไป

นอกจากนี้การดำเนินการต่างๆ ในงานวิจัยนี้เป็นนับว่าครั้งแรกสำหรับการทวนสอบตัวแบบทำนาย จึงถือได้ว่าเป็นการสร้างกรอบงานและวิธีการปฏิบัติในครั้งถัดไป ในอนาคตการเก็บข้อมูลจะมีประสิทธิภาพมากขึ้น มีจำนวนข้อมูลมากขึ้น ข้อมูลมีความถูกต้องมากขึ้น ดังนั้นกรอบงานในการทวนสอบจึงเป็นสิ่งที่สำคัญ

RPCA-JCSS

วิธีดำเนินการวิจัย (Research Methods)

1. ข้อมูลในการวิจัย

ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้รับมาจากกรมพินิจและคุ้มครองเด็กและเยาวชน เป็นข้อมูลที่รวบรวมจากศุนย์ฝึกแห่งหนึ่งของกรมพินิจฯ ซึ่งมีการติดตามเด็กและเยาวชนที่ได้รับการปล่อยตัวจากศูนย์ฝึกนี้ภายในระยะเวลา 1 ปี ดังนั้นการกระทำผิดซ้ำในที่นี้คือการกระทำผิดซ้ำในระยะเวลา 1 ปีหลังจากปล่อยตัวเท่านั้น รายละเอียดฟีเจอร์ข้อมูลแสดงในตารางที่ 1

ตารางที่ 1 แสดงรายละเอียดของข้อมูล

ลำดับ	สัญลักษณ์	ความหมายของปัจจัย	ประเภท
1	RC	กระทำผิดซ้ำ	Nominal (0,1)
2	IQ_LEVEL	รหัส IQ	Nominal (1,2,3,4,5,6,7)
3	IRC	รหัส IRC	Nominal (0,1)
4	R1	มีประวัติต่อต้านสังคม	Nominal (0,1)
5	R2	มีรูปแบบพฤติกรรมต่อต้านสังคม	Nominal (0,1)
6	R3	เพื่อนต่อต้านสังคม	Nominal (0,1)
7	R4	การศึกษาอาชีพ	Nominal (0,1)
8	R5	สารเสพติด	Nominal (0,1)
9	R6	ครอบครัว ชีวิตคู่	Nominal (0,1)
10	R7	กิจกรรมยามว่าง	Nominal (0,1)
11	R8	ความคิดแบบผู้กระทำผิด	Nominal (0,1)
12	CT1	เชื่อว่าตนเองเหนือกว่าผู้อื่น	Nominal (0,1)
13	CT2	เป็นพวกเดียวกับเพื่อนฯ	Nominal (0,1)
14	CT3	ไม่ใส่ใจผู้อื่น	Nominal (0,1)
15	CT4	เป็นศัตรูกับผู้บังคับใช้กฎหมาย	Nominal (0,1)
16	CT5	ไม่แสดงความรู้สึก	Nominal (0,1)
17	CT6	แสวงหาอำนาจและการควบคุมผู้อื่น	Nominal (0,1)
18	CT7	ประเมินสถานการณ์ต่ำเกินไป	Nominal (0,1)

RPCA-JCSS

ลำดับ	สัญลักษณ์	ความหมายของปัจจัย	ประเภท
19	CT8	ไม่ทำอะไรปล่อยให้มันเป็นไป	Nominal (0,1)
20	CT9	ไม่สามารถจัดการปัญหาได้	Nominal (0,1)
21	CT10	เอาเปรียบผู้อื่น	Nominal (0,1)
22	CT11	ไม่เคารพกฎหมาย	Nominal (0,1)
23	CT12	ต้องการความตื่นเต้นเร้าใจ	Nominal (0,1)
24	CT13	ยกเหตุผลข้ออ้าง	Nominal (0,1)
25	ABCD	รหัส ABCD	Nominal (A,B,C,D)
26	HML	รหัส HML	Nominal (H,M,L)

ข้อมูลมีจำนวน 258 เรคคอร์ด ข้อมูลที่ใช้ในการวิจัยนี้ได้ทำการลบ ชื่อ นามสกุล รหัสบัตรประชนชน และข้อมูลใดๆ ที่ใช้สืบย้อนไปหาตัวเด็กและเยาวชนออก ดังนั้นข้อมูลนี้จึงไม่สามารถใช้สืบย้อนถึงตัวเด็กและเยาวชนผู้นั้นได้ ไม่ว่ากรณีใดๆ

ข้อมูลมีลักษณะเป็น sparse matrix กล่าวคือในแต่ละเรคคอร์ดจะมีค่า 0 (ไม่ใช่) มากกว่าค่า 1 (ใช่) รายละเอียดทางสถิติแสดงในตารางที่ 2 ข้อมูลแบ่งออกเป็นสองคลาส (binary classes) คือ คลาส 0 (ไม่กระทำผิดซ้ำ) และ คลาส 1 (กระทำผิดซ้ำ) จำนวนข้อมูลของคลาส 1 คือ 99 เรคคอร์ด ดังนั้นคลาสของข้อมูลนี้จึงไม่สมดุล

ตารางที่ 2 แสดงรายละเอียดขทางสถิติของข้อมูล

ลำดับ	สัญลักษณ์	จำนวนที่เป็น 0	จำนวนที่เป็น 1	ลำดับ	สัญลักษณ์	จำนวนที่เป็น 0	จำนวนที่เป็น 1
1	RC	159	99	14	CT3	205	53
2	IQ_LEVEL	L3(4), L4(191), L5(45), L6(17), L7(3)		15	CT4	258	0
3	IRC	214	44	16	CT5	257	1
4	R1	247	11	17	CT6	252	6
5	R2	156	102	18	CT7	174	84
6	R3	20	238	19	CT8	253	5
7	R4	174	84	20	CT9	257	1
8	R5	42	216	21	CT10	257	1
9	R6	13	245	22	CT11	257	1
10	R7	52	206	23	CT12	182	76
11	R8	4	254	24	CT13	10	248
12	CT1	253	5	25	ABCD	A(1), B(7), C(13), D(237)	
13	CT2	144	114	26	HML	H(21), M(122), L(115)	

RPCA-JCSC

2. ซอฟต์แวร์ที่ใช้ในการวิจัย

การวิจัยนี้ใช้โปรแกรมประยุกต์ RapidMiner Studio ซึ่งหาได้จาก https://rapidminer.com/ เป็นเครื่องมือหลัก ซอฟต์แวรดังกล่าวเป็นที่รู้จักและใช้งานกันอย่างกว้างขวางเกี่ยวกับงานด้านเหมืองข้อมูลอัน เนื่องมาจากเสถียรภาพและประสิทธิภาพในการทำงาน เครื่องมือต่างๆ ที่อยู่ในซอฟต์แวร์สามารถช่วยลด ขั้นตอนการเขียนโปรแกรมและอำนวยความสะดวกให้กับผู้ดำเนินการวิจัยได้อย่างมาก กลุ่มชุดเครื่องมือหลัก ที่ใช้ในการวิจัย ได้แก่ กลุ่มเครื่องมือตัวแบบทำนาย (predictive) กลุ่มเครื่องมือการจัดกลุ่มข้อมูล (segmentation) และ กลุ่มเครื่องมือคัดเลือกฟีเจอร์ (optimization) ส่วนชุดเครื่องมือย่อยจะกล่าวต่อไปในส่วนของผลการทดลอง

3. วิธีดำเนินการวิจัย

การวิจัยนี้ประกอบด้วยการทดลอง 3 เรื่อง ได้แก่ การวัดประสิทธิภาพของตัวแบบ (models evaluation) การวิเคราะห์กลุ่มข้อมูลย่อย (clusters analysis) และ การคัดเลือกฟีเจอร์ (features selection) มีรายละเอียดดังนี้

3.1 การวัดประสิทธิภาพของตัวแบบ

ในการวัดประสิทธิภาพของตัวแบบ คือการใช้ข้อมูลทั้งหมดที่ได้จากขั้นตอนการเตรียมข้อมูล มาทำการสร้างตัวแบบและวัดประสิทธิภาพในการทำนาย ผู้วิจัยเลือกใช้กระบวนการ n-fold cross validation ในการวัดประสิทธิภาพของตัวแบบโดยกำหนดให้ n = 10 และผู้วิจัยเลือกตัวแบบทำนายประเภทต้นไม้ตัดสินใจ (decision tree) เป็นตัวแบบพื้นฐาน เนื่องจากตัวแบบสามารถเข้าใจได้ สอดคล้องกับการอ้างเหตุผลของมนุษย์ และทำให้ผู้วิจัยสามารถทำความเข้าใจถึงการได้มาของคำตอบได้

ตัวแบบต้นไม้ตัดสินใจในการทดลองนี้ประกอบด้วย decision tree, ID3, decision stump, random tree นอกจากนี้ผู้วิจัยได้เพิ่ม random forest, gradient boosted tree และ support vector machine (Tan et al. 2018) ไว้สำหรับเปรียบเทียบค่าอีกด้วย โดยสองตัวแบบแรกเป็นต้นไม้ตัดสินใจแบบกลุ่ม (ensemble) ส่วนตัวหลังเป็นตัวจำแนกที่นิยมใช้ในปัจจุบัน นอกจากนี้ผู้วิจัยยังได้ทดลองกับอัลกอริทึม 3 ตัวสำหรับสร้างตัวแบบ ได้แก่ gain ratio, information gain และ gini index

3.2 การวิเคราะห์กลุ่มข้อมูลย่อย

การทดลองนี้แบ่งข้อมูลออกเป็นกลุ่มข้อมูลย่อยโดยคาดหวังว่าจะเพิ่มประสิทธิภาพในการทำนายของตัวแบบ อ้างอิงจากสมมติฐานที่ว่า “เนื่องจากข้อมูลมีลักษณะที่หลากหลาย (heterogeneous) อาจเป็นผลให้ตัวแบบทำนายที่สร้างขึ้นมีประสิทธิภาพต่ำ จึงทำการแบ่งข้อมูลออกเป็นกลุ่มข้อมูลย่อยเพื่อลดความหลากหลายเหล่านั้น (homogeneous) ซึ่งอาจทำให้ประสิทธิภาพของตัวแบบทำนายในแต่ละกลุ่มเพิ่มขึ้น” เทคนิคที่ใช้ในการแบ่งกลุ่มข้อมูลคือ k-mean clustering (Tan et al. 2018)

RPCA-JCSC

k-mean clustering จะแบ่งข้อมูลทั้งหมดออกเป็น k กลุ่ม ซึ่งผู้วิจัยจะต้องเป็นผู้กำหนดพารามิเตอร์ k เอง ดังนั้นจึงต้องมีกระบวนการพิจารณาค่า k ที่เหมาะสม โดยกระบวนการดังกล่าวทำโดยการทดลองแบ่งกลุ่มข้อมูลเป็น 1, 2, 3, ..., n กลุ่ม จากนั้นพิจารณาค่าเฉลี่ยของระยะทางภายในกลุ่ม (average distance within clusters) ค่าที่เหมาะสมคือค่าที่เริ่มคงที่ หรือกล่าวได้ว่าแบ่งกลุ่มให้มากขึ้นก็ไม่ได้ลดค่าเฉลี่ยดังกล่าวลงอย่างมีนัยเมื่อทำการแบ่งกลุ่มย่อยแล้ว ข้อมูลแต่ละกลุ่มจะถูกนำไปทดสอบด้วยวิธีการเดียวกับการทดลองวัดประสิทธิภาพของตัวแบบอีกครั้ง

3.3 การคัดเลือกฟีเจอร์

การคัดเลือกฟีเจอร์คือการตรวจสอบว่าฟีเจอร์ใดบ้างที่มีความสัมพันธ์กับการกระทำผิดซ้ำ ในการทดลองนี้จะใช้วิธี wrapper approach ในการตรวจสอบ กล่าวคือตัวคัดเลือก (selection algorithms) จะทำการคัดเลือกฟีเจอร์ตามอัลกอริทึมทีละชุดแล้วลองไปสร้างตัวแบบทำนายและทดสอบประสิทธิภาพ ตัวคัดเลือกจะทำการปรับเปลี่ยนฟีเจอร์เพื่อให้ค่าประสิทธิภาพในการทำนายสูงที่สุด ในการทดลองนี้ใช้วิธีการคัดเลือกฟีเจอร์สามแบบได้แก่ forward selection, backward elimination และ optimized selection

ผลการวิจัยและอภิปรายผล (Results and Discussion)

1. ผลการทดลองที่ 1

รายละเอียดผลการทดลองแสดงในตารางที่ 3 ผลการทดลองแยกตามตัวแบบทำนายและแยกตามอัลกอริทึมแสดงในตารางที่ 4 และ 5 ตามลำดับ ในภาพรวมจะเห็นว่าค่าความถูกต้องเฉลี่ยอยู่ที่ 61.61% มีค่าความแม่นยำเฉลี่ยของคลาส 1 ที่ 49.07% และมีค่าความครบถ้วนของคลาส 1 ที่ 28.81% จากตัวเลขทั้งสามยืนยันได้ว่าเหตุปัจจัยทั้งหมดของการทำนายกระทำผิดซ้ำ (26 ฟีเจอร์) ยังไม่สามารถใช้ในการทำนายได้อย่างมีประสิทธิภาพภายใต้ข้อมูลชุดนี้อันเนื่องจากค่าผลลัพธ์ในการจำแนกที่น้อย โดยเฉพาะค่าความครบถ้วนของคลาส 1 ซึ่งมีค่าต่ำมาก

แต่เมื่อทำการพิจารณาแยกตามตัวแบบและอัลกอริทึมจะเห็นว่าตัวแบบ decision tree ให้ผลลัทธ์ในการทำนายในระดับที่ดีกว่าค่าเฉลี่ยโดยรวม ตัวแบบ decision tree ให้ค่าความถูกต้องเฉลี่ยที่ 60.61% ค่าความแม่นยำเฉลี่ยของคลาส 1 ที่ 48.54% และมีค่าความครบถ้วนของคลาส 1 ที่ 44.78% (ตารางที่ 4) ซึ่งค่าความครบถ้วนของคลาส 1 นี้เป็นค่าที่สำคัญที่สุดในการพิจารณาเพราะเป็นค่าที่บ่งบอกถึงการแยกคลาส 1 ออกจากคลาส 0 บนข้อมูลที่มีความไม่สมดุล จะเห็นได้ว่าค่าความครบถ้วนของ decision tree สูงกว่า ID3 และสูงกว่า decision stump และ random tree อย่างมาก อันที่จริงแล้ว decision stump และ random tree ไม่สามารถแยกแยะคลาส 1 ออกจากคลาส 0 ได้ เมื่อพิจรณาจากค่าความครบถ้วนที่ได้ (2.02% และ 10.10%)

เมื่อพิจารณาแยกตามอัลกอริทึมจะเห็นว่า gini index ให้ผลลัพธ์ที่ดีที่สุด ให้ค่าความถูกต้องเฉลี่ยที่ 62.42% ค่าความแม่นยำเฉลี่ยของคลาส 1 ที่ 50.07% และมีค่าความครบถ้วนของคลาส 1 ที่ 27.27% (ตารางที่ 5) เมื่อพิจารณาจากข้อมูลทั้งหมดแล้วการเลือกตัวแบบ decision tree คู่กับอัลกอริทึม gini index

RPCA-JCSS

จะให้ผลลัพธ์ที่ดีที่สุดบนข้อมูลชุดนี้ดังที่ได้กล่าวไว้ตอนต้น อย่างไรก็ตาม ที่ค่าความถูกต้อง 60.86% ค่าความแม่นยำของคลาส 1 ที่ 48.91% และค่าความคบถ้วนของคลาส 1 ที่ 45.45% ก็ยังไม่มากพอที่จะสรุปได้ว่าเหตุปัจจัยของตัวแบบนี้สามารถใช้ทำนายการกระทำผิดซ้ำได้

2. ผลการทดลองที่ 2

ข้อมูลทั้งหมดถูกนำมาแบ่งออกเป็นกลุ่มย่อยด้วยวิธีการจัดกลุ่ม k-means ดังนี้แล้วการวิเคราะห์เพื่อหาจำนวนคสัสเตอร์ (k) ที่เหมาะสมจึงเป็นสิ่งที่ต้องพิจารณาเป็นอย่างแรก ผู้วิจัยทดลองปรับค่า k และพิจารณาดูค่าเฉลี่ยนระยะทางภายในกลุ่ม (Avg. within centroid distance) ดังแสดงในตารางที่ 6 และภาพที่ 1

ตารางที่ 3 แสดงผลการทดลองที่ 1 (รวม)

ลำดับ	ตัวแบบทำนาย	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1
1	Decision Tree: gain_ratio	60.49%	67.07%	48.35%	70.44%	44.44%
2	Decision Tree: information_gain	60.48%	67.07%	48.35%	70.44%	44.44%
3	Decision Tree: gini_index	60.86%	67.47%	48.91%	70.44%	45.45%
4	ID3: gain_ratio	54.71%	63.29%	41.00%	62.89%	41.41%
5	ID3: information_gain	59.71%	66.87%	47.37%	68.55%	45.45%
6	ID3: gini_index	61.63%	66.48%	50.00%	76.10%	38.38%
7	Decision stump: grain_ratio	61.26%	61.66%	40.00%	98.11%	2.02%
8	Decision stump: information_gain	61.26%	61.66%	40.00%	98.11%	2.02%
9	Decision stump: gini_index	61.26%	61.66%	40.00%	98.11%	2.02%
10	Random tree: gain_ratio	62.03%	62.76%	52.63%	94.34%	10.10%
11	Random tree: information_gain	62.03%	62.76%	52.63%	94.34%	10.10%
12	Random tree: gini_index	62.03%	62.76%	52.63%	94.34%	10.10%
13	Random Forest: gain_ratio	65.91%	67.49%	60.00%	86.16%	33.33%
14	Random Forest: information_gain	64.00%	67.01%	54.69%	81.76%	35.35%
15	Random Forest: gini_index	66.34%	68.95%	58.82%	82.39%	40.40%
16	Gradient Boosted Tree	58.54%	69.70%	46.83%	57.86%	59.60%
17	Support Vector Machine	64.77%	64.90%	52.00%	84.91%	26.26%
	Average	61.61%	65.27%	49.07%	81.72%	28.87%

ตารางที่ 4 แสดงผลการทดลองที่ 1 (แยกตามตัวแบบ)

ลำดับ	ตัวแบบทำนาย	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1
1	**Decision Tree**	60.61%	67.20%	48.54%	70.44%	44.78%
2	ID3	58.68%	65.55%	46.12%	69.18%	41.75%
3	Decision Stump	61.26%	61.66%	40.00%	98.11%	2.02%
4	Random tree	62.03%	62.76%	52.63%	94.34%	10.10%
5	Random forest	65.42%	67.82%	57.84%	83.44%	36.36%
6	Gradient Boosted Tree	58.54%	69.70%	46.83%	57.86%	59.60%
7	Support Vector Machine	64.77%	64.90%	52.00%	84.91%	26.26%

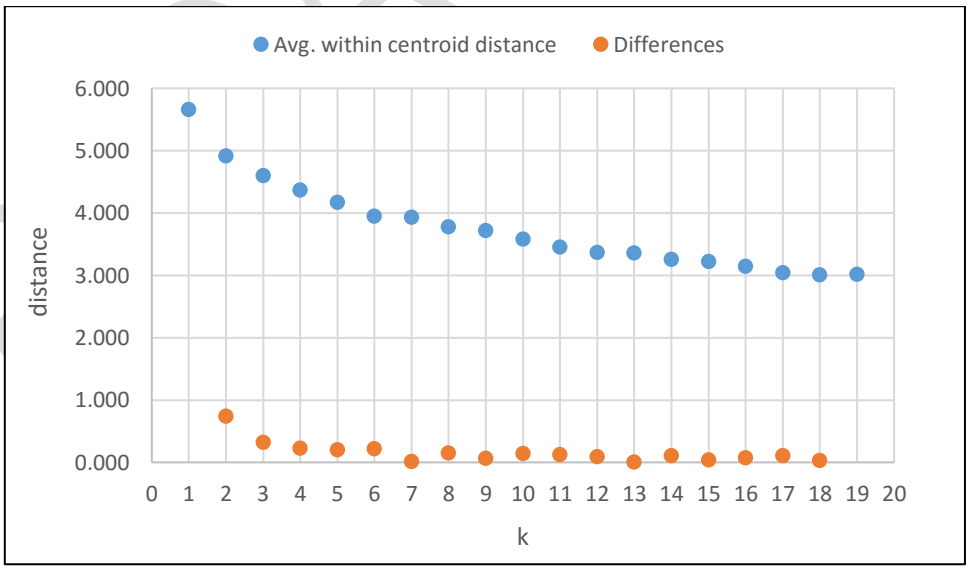
RPCA-JCSS

ตารางที่ 5 แสดงผลการทดลองที่ 1 (แยกตามอัลกอริทึม)

ลำดับ	ตัวแบบทำนาย	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1
1	gain_ratio	60.88%	64.45%	48.40%	82.39%	26.26%
2	information_gain	61.50%	65.07%	48.61%	82.64%	27.47%
3	**gini_index**	62.42%	65.46%	50.07%	84.28%	27.27%
4	Support Vector Machine	64.77%	64.90%	52.00%	84.91%	26.26%

ตารางที่ 6 แสดงผลการวิเคราะห์คลัสเตอร์

k	ค่าเฉลี่ยนระยะทางในกลุ่ม (Avg. within centroid distance)	Difference	k	ค่าเฉลี่ยนระยะทางในกลุ่ม (Avg. within centroid distance)	Difference
1	5.661		11	3.453	0.125
2	4.918	0.743	12	3.364	0.090
3	4.596	0.322	13	3.361	0.003
4	4.366	0.230	14	3.256	0.105
5	4.168	0.197	15	3.222	0.034
6	3.948	0.220	16	3.146	0.077
7	3.932	0.016	17	3.039	0.106
8	3.781	0.151	18	3.008	0.032
9	3.716	0.066	19	3.012	-0.004
10	3.579	0.137	20	2.876	0.136



ภาพที่ 1 แสดงผลการทดลองพิจารณาค่า k

ในตารางที่ 6 ค่า Difference คือค่าความแตกต่างระหว่างค่าเฉลี่ยนระยะทางในกลุ่มของ k-1 กับ k จากภาพที่ 1 เห็นได้ว่าหลังจาก k = 3 ขึ้นไปค่า Difference มีการเปลี่ยนแปลงเชิงเส้น ดังนั้นการเลือก k =

2 จึงน่าจะเหมาะสมที่สุดเมื่อพิจารณาจากค่า Difference นี้ รวมถึงจำนวนข้อมูลที่มี (258 เรคคอร์ด) ผลการแบ่งกลุ่มได้ว่า Cluster 0: 130 เรคคอร์ด และ Cluster 1: 128 เรคคอร์ด

กลุ่มข้อมูลแต่ละกลุ่มถูกนำไปทดสอบค่าความถูกต้องด้วยตัวแบบ decision tree ที่ใช้อัลกอริทึม gini index ผลการทดลองแสดงในตารางที่ 7 จากตารางจะเห็นได้ว่า Cluster 0 มีค่าความถูกต้องมากขึ้น มีค่าความแม่นยำและค่าความครบถ้านของคลาส 0 เพิ่มขึ้น แต่มีค่าความแม่นยำและค่าความครบถ้านของคลาส 1 ลดลง ในทางตรงกันข้าม Cluster 1 มีค่าความถูกต้องลดลง มีค่าความแม่นยำและค่าความครบถ้านของคลาส 0 ลดลง แต่มีค่าความแม่นยำและค่าความครบถ้านของคลาส 1 เพิ่มขึ้น

ลักษณะเช่นนี้ชี้ให้เห็นว่า ข้อมูลใน Cluster 0 มีข้อมูล คลาส 0 เป็นส่วนใหญ่มีการปนกับข้อมูลคลาส 1 บ้าง แต่ตัวแบบไม่สามารถแยกข้อมูลคลาส 1 เหล่านั้นออกมาได้เลย เมื่อพิจารณาจากค่าความแม่นยำและค่าความครบถ้วนของคลาส 1 ส่วนใน Cluster 1 ข้อมูลคลาส 1 ไปจับกลุ่มกันจริงแต่มีการปนกับข้อมูลคลาส 0 ที่มีปัจจัยเหมือนกับคลาส 1 ในสัดส่วนที่ใกล้ๆกัน ซึ่งทำให้ได้ประสิทธิภาพในการจำแนกต่ำ ดังนั้นสรุปได้ว่าการจัดกลุ่มข้อมูลไม่สามารถเพิ่มความสามารถในการทำนายของปัจจัยการกระทำผิดซ้ำนี้ได้ภายใต้ข้อมูลชุดนี้

ตารางที่ 7 แสดงผลการวัดค่าความถูกต้องของแต่ละคลัสเตอร์

ลำดับ	กลุ่มข้อมูล	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1
1	Cluster 0	75.38%	82.57%	38.1%	87.38%	29.63%
2	Cluster 1	49.74%	41.30%	54.88%	33.93%	62.50%

3. ผลการทดลองที่ 3

ผลการทดลองคัดเลือกฟีเจอร์โดยใช้อัลกอริทึม forward selection แสดงในตารางที่ 8 วิธีการทดลองคือคัดเลือกเพิ่มฟีเจอร์ทีละ 1 ฟีเจอร์จนกว่าอัลกอริทึมจะไม่คัดเลือกเพิ่ม การทดลองพบว่าอัลกอริทึมได้คัดเลือกฟีเจอร์เพียง 5 ตัวที่มีความสำคัญตามลำดับดังนี้ R1 > CT1 > IQ_LEVEL > R4 > CT7 และไม่มีการคัดเลือกต่อภายใต้พารามิเตอร์ที่กำหนดไว้

ผลการทดลองนี้ชี้ให้เห็นว่าว่าถ้านักวิเคราะห์จำเป็นต้องเลือกเหตุปัจจัยที่น่าจะมีผลต่อการกระทำผิดซ้ำไปวิเคราะห์ในเชิงลึกทีละตัวตามลำดับแล้วละก็ ปัจจัย R1 (มีประวัติต่อต้านสังคม) CT1 (เชื่อว่าตนเองเหนือกว่าผู้อื่น) IQ_LEVEL (ระดับเชาว์ปัญญา) R4 (การศึกษาอาชีพ) และ CT7 (ประเมินสถานการณ์ต่ำเกินไป) เป็นเหตุปัจจัยที่ผลการทดลองนี้แนะนำสำหรับใช้เป็นแนวทางในการคัดเลือก

RPCA-JCSS

ตารางที่ 8 แสดงผลการทดลองด้วยวิธี forward selection

No.	Attribute	accuracy	precision 0	precis 1	recall 0	recall 1
1	R1	63.60%	63.16%	72.73%	98.11%	8.08%
2	R1 > CT1	63.95%	63.20%	87.50%	99.37%	7.07%
3	R1 > CT1 > IQ_LEVEL	64.35%	63.45%	88.89%	99.37%	8.08%
4	R1 > CT1 > IQ_LEVEL > R4	64.75%	64.53%	66.67%	94.97%	16.16%
5	R1 > CT1 > IQ_LEVEL > R4 > CT7	67.06%	65.68%	81.82%	97.48%	18.18%

ผลการทดลองคัดเลือกฟีเจอร์โดยใช้อัลกอรึททึม backward elimination เพื่อทำการกำจัดฟีเจอร์ที่ไม่มีผลต่อการกระทำผิดซ้ำออกไปทีละ 1 ตัว ผลการทดลองพบว่าอัลกอริทึมได้กำจัดฟีเจอร์เพียง 5 ตัว ตามลำดับดังนี้ HML > CT13 > CT7 > R7 > R1 และไม่มีการกำจัดต่อภายใต้พารามิเตอร์ที่กำหนดไว้ ค่าความถูกต้องของตัวแบบทำนายแสดงในตารางที่ 9

ผลการทดลองนี้ชี้ให้เห็นว่าถ้านักวิเคราะห์จำเป็นต้องเลือกเหตุปัจจัยที่ไม่น่าจะมีผลต่อการกระทำผิดซ้ำออกไปแล้วละก็ เหตุปัจจัย HML (รหัส HML) CT13 (ยกเหตุผลข้ออ้าง) CT7 (ประเมินสถานการณ์ต่ำเกินไป) และ R7 (กิจกรรมยามว่าง) เป็นเหตุปัจจัยที่ผลการทดลองนี้แนะนำให้พิจารณา

สำหรับ CT7 (ประเมินสถานการณ์ต่ำเกินไป) นั้นเป็นเหตุปัจจัยที่ไม่ควรตัดออกเนื่องจากขัดแย้งกับผลที่ได้จากผลการทดลองด้านบน ที่เป็นเช่นนี้เพราะว่าอาจมีเหตุปัจจัยบางตัวที่ไม่ได้ถูกคัดออกซึ่งเมื่อรวมกันแล้วให้ความสำคัญมากกว่า CT7 เพียงตัวเดียว อัลกอริทึมจึงเลือกที่จะนำ CT7 ออกไป

ตารางที่ 9 แสดงผลการทดลองด้วย backward elimination

No.	Attributes	accuracy	precis 0	precis 1	recall 0	recall 1
1	HML	62.37%	68.24%	51.14%	72.96%	45.45%
2	HML > CT13	66.66%	70.39%	58.23%	79.25%	46.46%
3	HML > CT13 > CT7	66.98%	70.56%	58.97%	79.87%	46.46%
4	HML > CT13 > CT7 > R7	67.48%	71.68%	58.82%	77.99%	50.51%
5	HML > CT13 > CT7 > R7 > R1	69.40%	72.99%	61.90%	79.87%	52.53%

ผลการทดลองคัดเลือกฟีเจอร์โดยใช้อัลกอรึททึม optimize selection ทั้งแบบ forward selection และ backward selection พบว่าอัลกอริทึมได้คัดเลือกฟีเจอร์ภายใต้พารามิเตอร์ที่กำหนดไว้ ดังแสดงในภาพที่ 2 เนื่องจากอัลกอรึทึมทั้งสองนี้ทำงานบนพารามิเตอร์ที่มีการสุ่มเลือก ผลที่ได้ระหว่างสองอัลกอรึทึมจึงมีโอกาสที่จะต่างกัน จากตารางจะเห็นว่ากลุ่มฟีเจอร์ที่ได้จาก forward optimization ให้ค่าความครบถ้วนของ คลาส 0 สูง ในขณะที่กลุ่มฟีเจอร์ที่ได้จาก backward optimization ให้ค่าค่าความครบถ้วนของคลาส 1 สูงกว่าการทดลองที่ผ่านมามาก เมื่อพิจารณาเฉพาะตัวแบบทำนายเดี่ยว (single model)

RPCA-JCSC

ตารางที่ 10 แสดงผลการทดลองด้วย optimization

No.	algorithm	accuracy	Precision 0	precision 1	recall 0	recall 1
1	Forward optimization	69.35%	70.62%	65.62%	86.16%	42.42%
2	Backward optimization	70.98%	74.71%	63.64%	79.87%	56.57%

attributes	forward selection	backward elimination	Forward optimization	backward optimization
IQ_LEVEL	1	1	1	0
IRC	0	1	1	1
R1	1	0	1	0
R2	0	1	0	0
R3	0	1	1	1
R4	1	1	1	1
R5	0	1	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	1	1	1
CT1	1	1	1	0
CT2	0	1	0	1
CT3	0	1	1	1
CT4	0	1	1	0
CT5	0	1	1	1
CT6	0	1	1	0
CT7	1	0	1	0
CT8	0	1	0	0
CT9	0	1	1	1
CT10	0	1	1	1
CT11	0	1	1	1
CT12	0	1	0	1
CT13	0	0	1	0
ABCD	0	1	0	0
HML	0	0	0	0

1 = used, 0 = unused

ภาพที่ 2 แสดงฟีเจอร์ที่แต่ละอัลกอริทึมได้คัดเลือกมา

ผลการทดลองชี้ให้เห็นว่ามีความเป็นไปได้สูงที่กลุ่มฟีเจอร์ที่ได้จาก forward optimization จะชี้นำไปสู่การไม่กระทำผิดซ้ำ และ กลุ่มฟีเจอร์ที่ได้จาก backward optimization จะชี้นำไปสู่การกระทำผิดซ้ำ ดังนั้นเพื่อให้เกิดประโยชน์ต่อนักวิจัยที่จะศึกษาในเชิงลึกเรื่องเหตุปัจจัยการกระทำผิดซ้ำ ผลการทดลองนี้แนะนำว่าควรศึกษากลุ่มฟีเจอร์นี้ร่วมกันเนื่องจากมีความสัมพันธ์กันซ่อนอยู่

ผลการทดลองทั้งสามอ้างอิงอยู่บนชุดข้อมูลนี้เท่านั้น ถ้าจำนวนและคุณภาพของข้อมูลนี้เปลี่ยนไป ผลการทดลองย่อมเปลี่ยนตาม ไม่มากก็น้อย อย่างไรก็ตามเพื่อให้เกิดประโยชน์โดยแท้จริงนักวิเคราะห์ควรเก็บข้อมูลต่อไป และนำข้อมูลที่เก็บได้นั้นมาทดลองด้วยวิธีการเช่นนี้อยู่เป็นระยะ เพื่อประเมินคุณภาพของตัวแบบทำนายทางเชิงจิตวิทยาว่าตรงตามสมมุติฐานที่ตั้งไว้หรือไม่

ข้อเสนอแนะ (Recommendations)

โดยสรุป เอกสารฉบับนี้นำเสนอผลการทวนสอบเหตุปัจจัยเชิงจิตวิทยาเกี่ยวกับการกระทำผิดซ้ำของเด็กและเยาวชนไทยด้วยวิธีการทำเหมืองข้อมูล ข้อมูลที่ใช้ในการทวนสอบได้จากการติดตามเด็กและเยาวชนที่ได้รับการปล่อยตัวจากศูนย์ฝึกแห่งหนึ่งของกรมพินิจและคุ้มครองเด็กและเยาวชนเป็นเวลา 1 ปี การทดลองทั้งสามมีประเด็นข้อสังเกตดังนี้

- ตัวแบบทำนายที่นักวิเคราะห์สามารถแปลความหมายได้ (interpretable model) เช่น decision tree (หรือ rule-based model) เป็นสิ่งจำเป็นในการวิจัยเรื่องการกระทำผิดซ้ำ
- ตัวแบบทำนาย decision Tree ร่วมกับอัลกอริทึม gini index ให้ผลการทำนายได้ดีที่สุดเมื่อพิจารณาโดยรวม
- การแบ่งกลุ่มย่อยด้วย k-mean clustering เป็น 2 กลุ่ม ยังไม่สามารถแยกคลาส 1 (กระทำผิดซ้ำ) ออกจากคลาส 0 (ไม่กระทำผิดซ้ำ) ได้
- การคัดเลือกฟีเจอร์ด้วย forward selection ทำให้เราทราบว่าฟีเจอร์ตัวใดที่นักวิเคราะห์ควรให้ความสำคัญก่อน (เลือกทีละตัวตามลำดับ)
- การคัดเลือกฟีเจอร์ด้วย backward elimination ทำให้เราทราบว่าฟีเจอร์ตัวใดที่นักวิเคราะห์ควรให้ความสำคัญน้อย (กำจัดทีละตัวตามลำดับ)
- การคัดเลือกฟีเจอร์ด้วย optimized selection ทำให้เราทราบว่าฟีเจอร์กลุ่มใดที่นักวิเคราะห์ควรให้ความสำคัญร่วมกัน

เอกสารอ้างอิง (References)

ไทยโพสต์. (2562). *เด็กก่อคดีเฉลี่ยวันละ 83 ราย กรมสุขภาพจิตเผยน่าห่วงยิ่งขึ้น 3 ปีพบก่อคดีซ้ำเพิ่ม 7%.* สืบค้นเมื่อจาก https://www.thaipost.net/main/detail/16642.

จารุณี แซ่ตั้ง. (2546). การศึกษาปัจจัยที่มีผลต่อการไม่กระทำผิดซ้ำของเยาวชน กรณีศึกษาศูนย์ฝึกและอบรมเด็กและเยาวชนชายบ้านกรุณา. วิทยานิพนธ์ศิลปศาสตรมหาบัณฑิต สาขาอาชญาวิทยาและงานยุติธรรม บัณฑิตวิทยาลัย, มหาวิทยลัยมหิดล.

ณัฐวุฒิ ยอดระบำ. (2549). มูลเหตุจูงใจในการกระทำผิดซ้ำคดีเกี่ยวกับทรัพย์: ศึกษาเฉพาะกรณีเรือนจำจังหวัดนนทบุรี. สารนิพนธ์ศิลปศาสตรมหาบัณฑิต (การบริหารงานยุติธรรม) คณะสังคมสงเคราะห์ศาสตร์, มหาวิทยาลัยธรรมศาสตร์.

Mallett, C. A., Fukushima, M., Stoddard-Dare, P., & Quinn, L. (2013). Factors Related to Recidivism for Youthful. Cleveland State University.

Tan, P., Steinbach, M., Karpatne, A., Kumar, V. (2018). Introduction to Data Mining (2nd ed). Pearson Education (US).

Tafrate, R. C., & Mitchell, D. (2014). Forensic CBT: A Handbook for Clinical Practice. Chichester, UK: Wiley.

RPCA-JCSC