การเลือกปัจจัยในการเพิ่มประสิทธิภาพของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์

Factor Selection for Improving Vocational Certificate Course Selection Model Performance using Wrapper Method

ปริวรรต เพียรภายลุน[1*] และจารี ทองคำ[1]

Pariwat Pianpailoon[1*] and Jaree Thongkam[1]

Received: November 8, 2020; Revised: May 5, 2021; Accepted: May 18, 2021

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อคัดเลือกปัจจัยที่มีความเกี่ยวข้องในการเพิ่มประสิทธิภาพของแบบจำลองการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพของนักเรียนระดับชั้นมัธยมศึกษาปีที่ 3 ในงานวิจัยนี้ข้อมูลนักศึกษาที่จบการศึกษาระดับประกาศนียบัตรวิชาชีพจากวิทยาลัยเทคโนโลยีอุตสาหกรรมศรีสงคราม มหาวิทยาลัยนครพนม ตั้งแต่ปีการศึกษา 2555 - 2561 ได้ถูกรวบรวมจำนวน 926 ข้อมูล วิธีแรปเปอร์ได้ถูกนำมาใช้ร่วมกับเทคนิค Random Tree เทคนิค Support Vector Machines และเทคนิค Naive Bayes ในการเลือกปัจจัย เทคนิค Bagging เทคนิค K-Nearest Neighbors เทคนิค Decision Tree C4.5 เทคนิค Random Forest และเทคนิค Artificial Neural ได้ถูกนำมาใช้ในการสร้างแบบจำลองในการวัดประสิทธิภาพการคัดเลือกปัจจัยผู้วิจัยได้ใช้ 10-fold cross validation และวัดด้วยค่าความถูกต้อง ค่าความไว และค่าความจำเพาะ จากการทดลองพบว่าวิธีแรปเปอร์ร่วมกับ Random Tree สามารถเพิ่มค่าความถูกต้องให้กับแบบจำลองการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพสูงถึงร้อยละ 89.13 ค่าความไวสูงถึงร้อยละ 75.86 และค่าความจำเพาะสูงถึงร้อยละ 92.86

คำสำคัญ : การเลือกปัจจัย; วิธีแรปเปอร์; แบบจำลองการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพ; การทำเหมืองข้อมูล

[1] *คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม*

[1] *Faculty of Informatics, Mahasarakham University*

[*] *Corresponding Author E - mail Address: ohmeve74@gmail.com*

Abstract

The objective of this research is to relevant select factors for improving of the vocational certificate program selection models of third secondary education students. In this research paper, data were collected from students who graduated with a vocational certificate from the Sarisongkram Industrial Technology College, Nakhon Phanom University. The data are from 2012 - 2018 academic year, total 926 records. Wrapper with the Random Tree and Support Vector Machines and Naive Bayes and were used to select the relevant factors. Technique Bagging technique, K-Nearest neighbors, Decision Tree C4.5 technique, Random Forest technique and Artificial Neural techniques were used to build the prediction models. In order to measure performance of factors, 10-fold cross validation was employed and measured with Accuracy, Sensitivity and Specificity values. The experimental results showed that the wrapper method in combination with Random Tree was able to increase the accuracy of vocational certificate program selection model up to 89.13 %, sensitivity up to 75.86 % and specificity up to 92.86 %.

Keywords: Factor Selection; Wrapper Method; Vocational Certificate Program Selection Models; Data Mining

บทนำ

การทำเหมืองข้อมูล (Data Mining) [1] เป็นกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภทซึ่งมีนักวิจัยหลายท่านได้นำเทคนิคในการทำเหมืองข้อมูลมาใช้ เช่น [2] ได้ทำการเปรียบเทียบประสิทธิภาพของเทคนิคและวิธีการเลือกคุณสมบัติในการทำนายการเรียนรู้ ผลสัมฤทธิ์ทางการเรียนของนักเรียนอาชีวศึกษา โดยเป็นข้อมูลของนักศึกษาระดับประกาศนียบัตรวิชาชีพ ของสถานศึกษารัฐบาลแห่งหนึ่งในประเทศไทยจำนวน 5,100 ระเบียน ตั้งแต่ปีการศึกษา 2550 – 2559 ประกอบด้วย 9 หลักสูตรวิชา และได้ใช้ 3 เทคนิค Decision Tree C4.5, Naive Bayes และ Rule Induction เพื่อเปรียบเทียบประสิทธิภาพโดยการใช้วิธีการ 10-fold cross validation ซึ่งผลที่ได้เทคนิค Decision Tree C4.5 เหมาะสมที่สุดสำหรับการทำนายผลสัมฤทธิ์ทางการเรียนของนักเรียนอาชีวศึกษา [3] ได้ศึกษาตัวแบบพยากรณ์คุณลักษณะความเหมาะสมสำหรับการเลือกสมัครสาขาวิชาเรียนโดยใช้เทคนิคเหมืองข้อมูล เพื่อเปรียบเทียบประสิทธิภาพตัวแบบพยากรณ์คุณลักษณะความเหมาะสมสำหรับการเลือกสมัครสาขาวิชาเรียนและทำการประเมินตัวแบบพยากรณ์ด้วยเทคนิคต้นไม้ตัดสินใจ และเทคนิคแรนดอมฟอเรส ผลการทดลองพบว่า เทคนิคแรนดอมฟอเรส ให้ค่าความถูกต้องสูงที่สุด ปัญหาของการทำเหมืองข้อมูลนั้นสามารถแก้ไขปัญหาด้วยการคัดเลือกปัจจัยเป็นการวิเคราะห์ข้อมูลเพื่อทำนายการเป็นสมาชิกกลุ่ม โดยอาศัยข้อมูลตัวอย่างคุณลักษณะบางอย่างที่มีประโยชน์ต่อการทำนายและหากนำคุณลักษณะที่มีจำนวนมากไปใช้

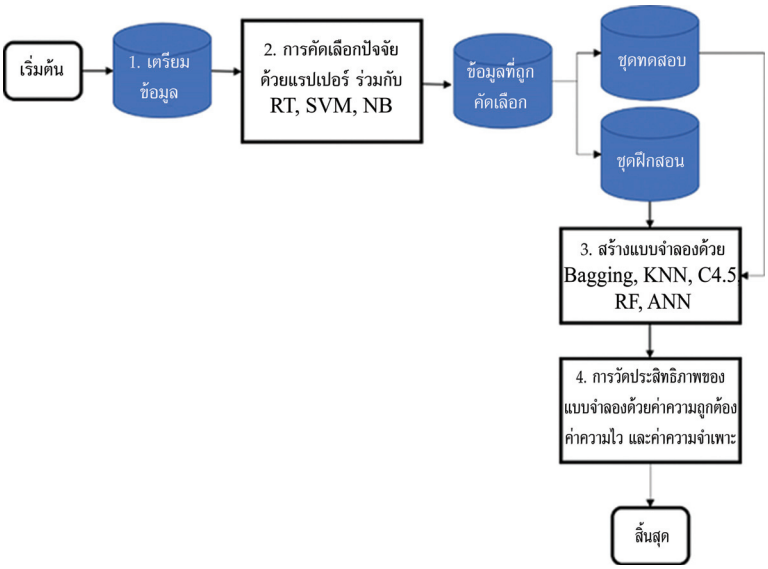
ในการจำแนกอาจจะส่งผลต่อความแม่นยำได้ กระบวนการคัดเลือกปัจจัย [4] การทำเหมืองข้อมูลเหล่านี้จึงเป็นกระบวนการวิเคราะห์เพื่อจำแนกประเภท กำจัดข้อมูลที่ไม่เกี่ยวข้อง และข้อมูลที่ซ้ำซ้อนทำให้เพิ่มความแม่นยำในการเรียนรู้เพิ่มความเข้าใจต่อตัวแบบที่ได้ และยังลดเวลาในการเรียนรู้ข้อมูลกระบวนการนี้จึงเรียกว่า การคัดเลือกปัจจัย

การคัดเลือกปัจจัย [5] เป็นขั้นตอนก่อนการสร้างแบบจำลอง ใช้ในการค้นหาตัวแปรที่มีผลต่อประสิทธิภาพของแบบจำลองในการจำแนกข้อมูลอัตโนมัตินี้ให้แบบจำลองมีประสิทธิภาพเพิ่มมากขึ้น ในงานวิจัยนี้ผู้วิจัยได้ใช้วิธีแรปเปอร์ (Wrapper) ซึ่งเป็นหลักการเรียนรู้เพื่อหาเป้าหมายเพื่อคาดคะเนมูลค่าของกลุ่มปัจจัย แล้วนำกลุ่มปัจจัยที่ได้มาคำนวณหาค่าความถูกต้องหลังจากคัดเลือกปัจจัย

ดังนั้นผู้วิจัยได้มุ่งเน้นที่จะพัฒนาการเลือกปัจจัยในการเพิ่มประสิทธิภาพของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์ร่วมกับเทคนิค Random Tree เทคนิค Support Vector Machines เทคนิค Naive Bayes และทำการวัดประสิทธิภาพของแบบจำลองด้วยเทคนิค Bagging เทคนิค K-Nearest Neighbors เทคนิค Decision Tree C4.5 เทคนิค Random Forest (RF) เทคนิค Artificial Neural Networks ทำการวัดประสิทธิภาพของแบบจำลองด้วยหลักการ 10-fold cross validation เพื่อหาค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity)

วิธีดำเนินการวิจัย

วิธีดำเนินการวิจัยนี้มี 4 ขั้นตอน คือ การเตรียมข้อมูล การคัดเลือกปัจจัย การสร้างแบบจำลอง และการวัดประสิทธิภาพแบบจำลอง ดังรูปที่ 1



รูปที่ 1 ขั้นตอนวิธีดำเนินการวิจัย

1. การเตรียมข้อมูล

การเตรียมข้อมูลเป็นการเก็บรวบรวมข้อมูลนักศึกษาที่สำเร็จการศึกษาจากทั้ง 5 หลักสูตร ได้แก่ หลักสูตรวิชาช่างยนต์ หลักสูตรวิชาช่างอิเล็กทรอนิกส์ หลักสูตรวิชาช่างไฟฟ้า หลักสูตรวิชาการบัญชี และหลักสูตรวิชาคอมพิวเตอร์ธุรกิจ ซึ่งเป็นข้อมูลที่ได้จากวิทยาลัยเทคโนโลยีอุตสาหกรรมศรีสงคราม มหาวิทยาลัยนครพนม จากปีการศึกษา 2555 - 2561 มีจำนวนนักศึกษา 1,350 คน สามารถแสดงจำนวนนักศึกษาได้ดังตารางที่ 1

ตารางที่ 1 จำนวนนักศึกษา

ลำดับ	หลักสูตร	ปีการศึกษา							จำนวนนักศึกษา
		2555	2556	2557	2558	2559	2560	2561	
1	คอมพิวเตอร์ธุรกิจ	21	29	46	44	40	48	62	290
2	ช่างไฟฟ้า	33	41	40	47	50	48	35	294
3	ช่างยนต์	22	18	16	19	44	48	44	211
4	ช่างอิเล็กทรอนิกส์	36	34	33	40	39	37	36	255
5	บัญชี	44	35	38	39	41	42	61	300
								รวม	1,350

ในงานวิจัยนี้ผู้วิจัยได้นำข้อมูลที่ได้รวบรวมจากแฟ้มข้อมูลมาใส่ในไฟล์ Excel แล้วทำการคัดกรองข้อมูลที่สามารถนำมาใช้และทำความสะอาดข้อมูลโดยการตัดข้อมูลที่มีค่าไม่ครบถ้วน จากนั้นทำการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมนำเอาไปใช้ในการวิเคราะห์จึงได้ทำการแทนค่าข้อมูลให้อยู่ในรูปแบบที่สามารถนำมาวิเคราะห์ได้ เช่น เพศชาย แปลงเป็น 1, เพศหญิงแปลงเป็น 2, รายได้บิดา 150,000 - 300,000/ปี แปลงเป็น 1, น้อยกว่า 150,000/ปี แปลงเป็น 2, มากกว่า 300,000/ปี แปลงเป็น 3, ไม่มีรายได้แปลงเป็น 4, รายได้มารดา 150,000 - 300,000/ปี แปลงเป็น 1, น้อยกว่า 150,000/ปี แปลงเป็น 2, มากกว่า 300,000/ปี แปลงเป็น 3, ไม่มีรายได้แปลงเป็น 4, หลักสูตรคอมพิวเตอร์ธุรกิจ แปลงเป็น 1, หลักสูตรช่างไฟฟ้า แปลงเป็น 2, หลักสูตรช่างยนต์ แปลงเป็น 3, หลักสูตรช่างอิเล็กทรอนิกส์ แปลงเป็น 4, หลักสูตรการบัญชี แปลงเป็น 5 เป็นต้น และทำการเลือกตัวแปรต้นที่เป็นอิสระต่อกันแต่มีความสัมพันธ์กับตัวแปรตาม (คลาส) ได้จำนวนนักศึกษาที่จบการการศึกษาและนำไปประมวลผลมีจำนวนทั้งสิ้น 926 ข้อมูล ปัจจัยในงานวิจัยสามารถแสดงผลได้ดังตารางที่ 2 และ 3

ตารางที่ 2 ปัจจัยแบบ Nominal

ลำดับที่	ชื่อปัจจัย	คำอธิบายปัจจัย
1	Sex	เพศ
2	Fa_occ	อาชีพบิดา
3	Ma_occ	อาชีพมารดา

ตารางที่ 2 ปัจจัยแบบ Nominal (ต่อ)

ลำดับที่	ชื่อปัจจัย	คำอธิบายปัจจัย
4	Fa_income	รายได้บิดา
5	Ma_income	รายได้มารดา
6	GPA	เกรดเฉลี่ยจบ (ปวช.)
7	Class	หลักสูตร

ตารางที่ 3 ปัจจัยแบบ Numeric

ลำดับที่	ชื่อปัจจัย	คำอธิบายปัจจัย	ค่าต่ำสุด (Min)	ค่าสูงสุด (Max)	ค่าเฉลี่ย (Mean)	ค่าเบี่ยงเบนมาตรฐาน (SD.)
1	Thai	เกรดวิชาภาษาไทย	1.08	40	2.5	0.63
2	English	เกรดวิชาภาษาอังกฤษ	1	4.03	2.09	0.75
3	Science	เกรดวิชาวิทยาศาสตร์	1	4	2.5	0.63
4	Mathematics	เกรดวิชาคณิตศาสตร์	1	6.06	2.22	0.74
5	Social	เกรดวิชาสังคมศึกษา	1	23.8	2.74	1.16
6	Gymnastics	เกรดวิชาพละศึกษา	1.75	4	3.31	0.40

2. การคัดเลือกปัจจัยด้วยวิธีแรปเปอร์

การคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ (Wrapper) [6] เป็นวิธีการลดปัจจัยให้น้อยลงเพื่อที่จะทำให้ประสิทธิภาพที่ดีขึ้น เริ่มต้นด้วยชุดข้อมูลที่มีคุณลักษณะแบบเต็มและทำการลบคุณลักษณะที่ดีที่สุดโดยลองเลือกความเป็นไปได้ทั้งหมดและดำเนินการต่อไป ในการเลือกคุณลักษณะของโปรแกรมกระบวนการทำงานของแรปเปอร์นั้นจะรวมตัวแยกประเภทไว้ในลูปการตรวจสอบความถูกต้องโดยจะค้นหาผ่านช่องว่างของคุณลักษณะและใช้ลักษณะนามเพื่อค้นหาชุดคุณลักษณะที่ดีในการค้นหาสามารถไปข้างหน้าถอยหลังหรือแบบสองทิศทาง เริ่มต้นจากส่วนย่อยของตัวนั้น ๆ แรปเปอร์ [7] จะค้นหาคุณลักษณะ X_i จะเรียกว่ามีความเกี่ยวข้องกับตัวแปรกลุ่มหรือตัวแปรเป้าหมายอย่างแข็ง ก็ต่อเมื่อการแจกแจงความน่าจะเป็นอย่างมีเงื่อนไขของกลุ่มโดยกำหนดคุณลักษณะทั้งหมดที่เปลี่ยนไปตาม ถ้า X_i ถูกกำจัด และคุณลักษณะ X_i จะเรียกว่ามีความเกี่ยวข้องกับกลุ่มตัวแปรอย่างอ่อน ถ้าการแจกแจงความน่าจะเป็นอย่างมีเงื่อนไขของกลุ่มเมื่อกำหนดเซตย่อยของคุณลักษณะ S (ซึ่งรวม X_i) นั้นเปลี่ยนไปถ้า X_i ถูกกำจัด คุณลักษณะใดที่ไม่ได้มีความเกี่ยวข้องอย่างแข็งหรืออ่อน จะเรียกว่าเป็นคุณลักษณะที่ไม่มีความเกี่ยวข้อง ในการทดลองครั้งนี้ผู้วิจัยได้ใช้วิธีแรปเปอร์ร่วมกับเทคนิค Random Tree เทคนิค Support Vector Machines เทคนิค Naive Bayes เข้ามาทำการคัดเลือกปัจจัย

เทคนิค Random Tree (RT) [8] เป็นขั้นตอนวิธีในการสร้างต้นไม้ตัดสินใจโดยมีหลักการสร้างต้นไม้จากการสุ่มต้นไม้หลาย ๆ แบบ ในแต่ละโหนดแล้วเลือกมาประมวลผล โดยไม่ใช้การตัดแต่งกิ่ง Random Tree ทำงานคล้ายกับ C4.5 หรือ CART แต่ Random Tree จะสุ่มเลือกชุดย่อยของแอตทริบิวต์ก่อนที่จะนำไปใช้ ซึ่งขนาดของชุดย่อยถูกระบุโดยพารามิเตอร์อัตราส่วนชุดย่อย

เทคนิค Support Vector Machine (SVM) [9] เป็นขั้นตอนวิธีการที่มีความรวดเร็วและเป็นอัลกอริทึมที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูลใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูลโดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้โดยเน้นไปยัง เส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุดแนวความคิดของเทคนิควิธี SVM นั้นเกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature Space) ในลักษณะเชิงเส้นจากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกันโดยจะทำการสร้างเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่า เส้นตรงที่แบ่งกลุ่มสองกลุ่มออกจากกันนั้น เส้นใดเป็นเส้นที่ดีที่สุด

เทคนิค Naive Bayes (NB) [10] เป็นขั้นตอนวิธีที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (Class) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อนเป็นขั้นตอนวิธีในการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่เป็นการจำแนกข้อมูลโดยใช้ความน่าจะเป็นและคำนวณการแจกแจงความน่าจะเป็นตามสมมติฐานหลักการของนาอีฟเบส์ใช้การคำนวณหาความน่าจะเป็นซึ่งถูกใช้ในการทำนายผลเป็นวิธีในการแก้ปัญหา แบบการจำแนกที่สามารถคาดการณ์ผลลัพธ์และวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์นาอีฟเบส์เป็นวิธีจำแนกข้อมูลที่มีประสิทธิภาพ มีอัลกอริทึมในการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน

จากการนำวิธีการแรปเปอร์ร่วมกับเทคนิค Random Tree เทคนิค Support Vector Machines เทคนิค Naive Bayes เมื่อทำการเลือกข้อมูลแล้วสามารถแสดงปัจจัยหลังจากการคัดเลือกได้ดังตารางที่ 4

ตารางที่ 4 ปัจจัยหลังจากใช้วิธีแรปเปอร์

Feature Select with Wrapper	Random Tree	Support Vector Machines	Naive Bayes
Wrapper Subset Evaluation	เพศ อาชีพบิดา เกรดวิชาภาษาอังกฤษ เกรดวิชาวิทยาศาสตร์ เกรดวิชาสังคม เกรดวิชาพละศึกษา รายได้บิดา รายได้มารดา หลักสูตร	เพศ อาชีพบิดา อาชีพมารดา เกรดวิชาวิทยาศาสตร์ รายได้มารดา หลักสูตร	เพศ อาชีพบิดา อาชีพมารดา เกรดเฉลี่ยจบ(ปวช.) หลักสูตร

3. **ขั้นตอนการสร้างแบบจำลอง**

ในการสร้างแบบจำลองการเลือกปัจจัยในการเพิ่มประสิทธิภาพของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์ จากการศึกษาปัจจัยการเลือกหลักสูตรของนักศึกษาระดับประกาศนียบัตรวิชาชีพ วิทยาลัยเทคโนโลยีอุตสาหกรรมศรีสงคราม ผู้วิจัยได้นำเทคนิคในการทำเหมืองข้อมูลจำนวน 5 เทคนิคมาสร้างแบบจำลอง ดังนี้

เทคนิค Bagging [11] เป็นการสร้างแบบจำลองออกมาหลาย ๆ แบบจำลองโดยใช้การสุ่มข้อมูลตัวอย่างจากข้อมูลฝึกสอนออกมาเป็นหลาย ๆ ชุดสำหรับวิธีการสุ่มข้อมูลออกมาเป็นวิธีสุ่มแบบแทนที่ (Random with Replacement) หมายความว่าข้อมูลที่เรามียังคงเดิมไม่ได้ลดลงหลังจากการสุ่ม ซึ่งสามารถสุ่มข้อมูลหลาย ๆรอบเพื่อให้ได้ classifier หลาย ๆ ตัวเวลาทำนายจะใช้ classifier ทุกตัวที่สร้างขึ้นมาเพื่อทำนายชุดข้อมูลใหม่ที่พบ

เทคนิค K-Nearest Neighbors (KNN) [12] เป็นวิธีการที่จะจำแนกโดยดูว่าคลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน (“K” ตัว) ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไขสำหรับแต่ละคลาสและกำหนดเงื่อนไขใหม่ ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด การนำเทคนิค K-Nearest neighbors ไปใช้นั้น เป็นการหาระยะห่างระหว่างแต่ละตัวแปร (Attribute) ในข้อมูลแล้ว หาผลรวมของระยะห่างทั้งหมดแล้วพิจารณาเลือกคำตอบจากจำนวน K ตัว

เทคนิค Decision Tree C4.5 (C4.5) [13] เป็นวิธีหนึ่งที่จะประมาณฟังก์ชันที่มีค่าที่ไม่ต่อเนื่อง (Discrete-Value Function) ด้วยแผนผังต้นไม้อาจประกอบด้วยเซตของกฎต่าง ๆ แบบถ้า-แล้ว (if-then) เพื่อให้มนุษย์สามารถอ่านแล้วเข้าใจการตัดสินใจของต้นไม้ได้ ข้อดีอย่างหนึ่งของการใช้ Decision Tree คือเลือกตัวแปรที่มีความสำคัญที่ช่วยแบ่งแยกข้อมูลออกมาได้

เทคนิค Random Forest (RF) [14] เป็นการสร้างตัวแบบจากต้นไม้ตัดสินใจหลาย ๆ ตัวแบบย่อย ๆ โดยแต่ละตัวแบบจะได้รับชุดข้อมูลที่ไม่เหมือนกัน (Random Sampling) โดยข้อมูลย่อยจะถูกแบ่งออกมาเป็น n ชุด ที่ไม่เหมือนกันตามจำนวนต้นไม้ตัดสินใจ และจะสร้างตัวแบบเพื่อทำการพยากรณ์คำตอบของตนเองด้วยการโหวต

เทคนิค Artificial Neural Networks (ANN) [15] เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์ (Artificial Intelligence: AI) ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ หลักการสำคัญของโครงข่ายประสาทเทียม คือ ความพยายามที่จะลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์เพื่อทำงานได้อย่างมีประสิทธิภาพ ลักษณะทั่วไปของโครงข่ายประสาทเทียม คือ การที่โหนด (Node) ต่าง ๆ จำลองมาจากไซแนปส์ (Synapse) ของเซลล์ประสาทระหว่างเดนไดรต์ (Dendrite) และแอกซอน (Axon) โดยมีฟังก์ชันเป็นตัวกำหนดสัญญาณส่งออก (Activation Function or Transfer Function) ลักษณะของโครงข่ายประสาทเทียม สามารถแบ่งได้ 2 แบบ คือ 1) โครงข่ายประสาทเทียมแบบชั้นเดียว (Single Layer) ซึ่งจะมีเพียงชั้นสัญญาณประสาทขาเข้าและชั้นสัญญาณประสาทขาออกเท่านั้น เช่น โครงข่ายเพอเซบตรอนอย่างง่าย (Simple Perceptron) และโครงข่ายโฮปฟิลด์ (Hopfield Networks) เป็นต้น และ 2) โครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer) ซึ่งมีลักษณะเช่นเดียวกับโครงข่ายประสาทเทียมแบบชั้นเดียวแต่จะมีชั้นแอบแฝง (Hidden) เพิ่มขึ้นโดยอยู่ส่วนกลางระหว่างชั้นนำข้อมูลป้อนเข้าและชั้นส่งข้อมูลออกทั้งนี้ชั้นแอบแฝงอาจมีมากกว่า 1 ชั้น

4. การวัดประสิทธิภาพของแบบจำลอง

เมื่อทำการสร้างแบบจำลองเสร็จแล้วนำแบบจำลองมาทดสอบประเมินประสิทธิภาพด้วยวิธีการของ 10-fold cross validation [16] โดยการแบ่งข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กัน จากนั้นจะทำการทดสอบทั้งหมด 10 รอบ โดยในแต่ละรอบจะใช้ข้อมูล 1 ชุดเป็นชุดทดสอบและอีก 9 ชุดที่เหลือเป็นชุดฝึกสอนในรอบต่อไปก็ใช้ชุดข้อมูลถัดไปเป็นชุดทดสอบจนครบทั้ง 10 ชุดข้อมูล ซึ่งข้อมูลทั้งหมด

จำนวน 926 ข้อมูล แบ่งออกเป็น 10 ชุดข้อมูล ซึ่งคิดเป็นอัตราข้อมูลทดสอบต่อข้อมูลฝึกเป็นอัตราส่วน 10:90 แล้วทำการเปรียบเทียบค่าด้วยการจำแนกกลุ่มข้อมูล คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Sensitivity) และค่าความจำเพาะ (Specificity) [17]

4.1 ค่าความถูกต้อง (Accuracy) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพได้อย่างถูกต้องต่อข้อมูลดังสมการที่ (1)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

4.2 ค่าความไว (Sensitivity) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพได้อย่างถูกต้องต่อการเลือกหลักสูตรที่เกิดขึ้นจริงดังสมการที่ (2)

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{2}$$

4.3 ค่าความจำเพาะ (Specificity) คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพที่ไม่ได้เลือกหลักสูตรได้อย่างถูกต้องต่อหลักสูตรที่พยากรณ์ว่ามีการเลือกหลักสูตรดังสมการที่ (3)

$$\text{Specificity} = \frac{TP}{TP+FN} \tag{3}$$

TP คือ จำนวนข้อมูลที่พยากรณ์ถูกต้องอยู่ในคลาสนั้น
TN คือ จำนวนข้อมูลที่พยากรณ์ถูกต้องอยู่ในคลาสอื่น
FP คือ จำนวนข้อมูลที่พยากรณ์ผิดแต่เป็นข้อมูลที่อยู่ในคลาสนั้น
FP คือ จำนวนข้อมูลที่พยากรณ์ผิดแต่ไม่เป็นข้อมูลที่อยู่ในคลาสนั้น

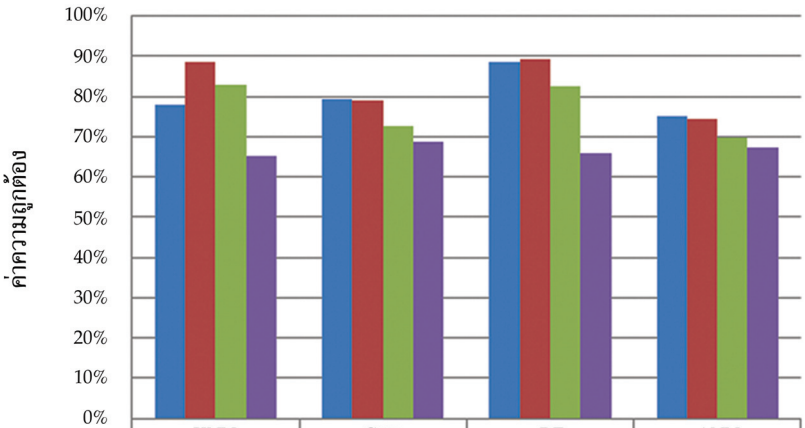
ผลการวิจัย

ผลการวิจัยจากการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ร่วมกับเทคนิค Random Tree (RT) เทคนิค Support Vector Machines (SVM) และเทคนิค Naive Bayes (NB) จากการทดลองโดยทำการแยกข้อมูลออกเป็น 10 ชุดเท่า ๆ กันนำชุดข้อมูล 9 ชุดมาเป็นชุดฝึกสอน ส่วนชุดที่เหลือเป็นชุดทดสอบ และทำการทดสอบจำนวน 10 รอบโดยแต่ละรอบชุดทดสอบจะทดลองไม่ซ้ำกัน ทำให้การทดลองใช้ข้อมูลที่ไม่อยู่ในชุดฝึกสอน ในการนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค K-Nearest neighbors (KNN) เทคนิค Decision Tree C4.5 (C4.5) เทคนิค Random Forest (RF) เทคนิค Artificial Neural Networks (ANN) ซึ่งในการทดลองนี้ได้นำเอาโปรแกรม Weka Version 3.8.4

มาใช้เป็นเครื่องมือในการทดสอบและหาค่าความถูกต้อง ค่าความไว และค่าความจำเพาะของแบบจำลองการเลือกปัจจัยในการเพิ่มประสิทธิภาพของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์

1. **ผลของค่าความถูกต้อง**

ค่าความถูกต้อง คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพได้อย่างถูกต้อง ในการทดลองนี้แบบจำลองการคัดเลือกหลักสูตรจากข้อมูลดั้งเดิม (Raw Data) และจากข้อมูลที่ได้จากการคัดเลือกปัจจัยด้วยแรปเปอร์ร่วมกับเทคนิค RT เทคนิค SVM และเทคนิค NB เมื่อนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค RF เทคนิค ANN จากการทดลองสามารถแสดงผลได้ดังรูปที่ 2



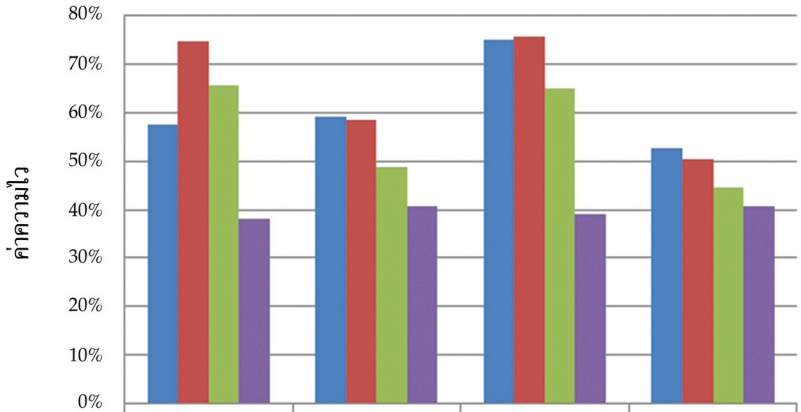
	KNN	C4.5	RF	ANN
Raw Data	78.05%	79.40%	88.66%	74.98%
W+RT	88.35%	78.81%	89.13%	74.33%
W+SVM	82.95%	72.68%	82.44%	69.82%
W+NB	65.10%	68.58%	65.75%	67.42%

รูปที่ 2 ค่าความถูกต้องของแบบจำลอง

จากรูปที่ 2 แสดงผลค่าความถูกต้องของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์ จากข้อมูลดั้งเดิม (Raw Data) และจากข้อมูลที่ได้จากการคัดเลือกปัจจัยด้วยแรปเปอร์ ร่วมกับเทคนิค RT เทคนิค SVM และเทคนิค NB เมื่อนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค RF เทคนิค ANN จากการทดลองพบว่า ค่าความถูกต้องของแบบจำลองหลังการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ ร่วมกับเทคนิค RT สามารถเพิ่มค่าความถูกต้องเฉลี่ยขึ้นสูงที่สุดเมื่อใช้เทคนิค KNN ในการสร้างแบบจำลองสูงขึ้นถึงร้อยละ 10.30 เมื่อวิธีแรปเปอร์ร่วมกับเทคนิค SVM สามารถเพิ่มค่าความถูกต้องเฉลี่ยให้กับแบบจำลองที่สร้างด้วยเทคนิค KNN สูงขึ้นเพียงร้อยละ 4.90 แต่อย่างไรก็ตามค่าความถูกต้องสูงที่สุดของแบบจำลองหลังการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ร่วมกับเทคนิค RT มีค่าความถูกต้องเฉลี่ยอยู่ที่ 89.13 เมื่อใช้เทคนิค RF และค่าความถูกต้องที่ต่ำสุดด้วยวิธีแรปเปอร์ร่วมกับเทคนิค NB มีค่าต่ำสุดเฉลี่ยอยู่ที่ 65.10 เมื่อใช้เทคนิค KNN

2. ผลของค่าความไว

ค่าความไว คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพได้อย่างถูกต้องต่อการเลือกหลักสูตรที่เกิดขึ้นจริง ในการทดลองนี้แบบจำลองการคัดเลือกหลักสูตรจากข้อมูลดั้งเดิม (Raw Data) และจากข้อมูลที่ได้จากการคัดเลือกปัจจัยด้วยแรปเปอร์ ร่วมกับเทคนิค RT เทคนิค SVM และเทคนิค NB เมื่อนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค RF เทคนิค ANN จากการทดลองสามารถแสดงผลได้ดังรูปที่ 3



	KNN	C4.5	RF	ANN
Raw Data	57.44%	59.24%	75.16%	52.71%
W+RT	74.78%	58.63%	75.86%	50.49%
W+SVM	65.82%	48.95%	65.03%	44.46%
W+NB	38.16%	40.68%	39.19%	40.75%

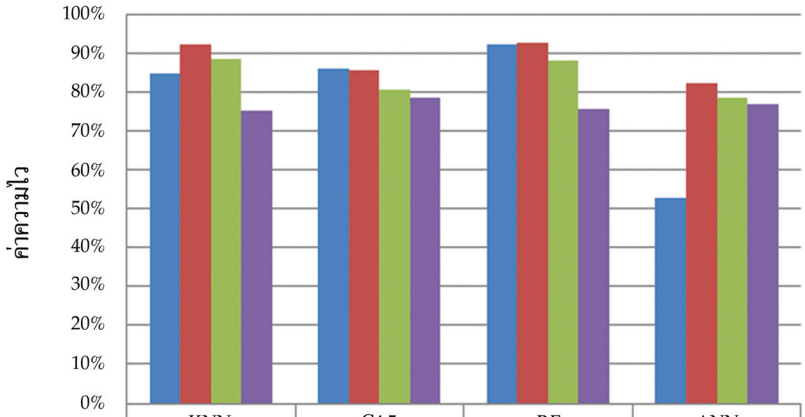
รูปที่ 3 ค่าความไวของแบบจำลอง

จากรูปที่ 3 แสดงผลค่าความไวของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์ จากข้อมูลดั้งเดิม (Raw Data) และจากข้อมูลที่ได้จากการคัดเลือกปัจจัยด้วยแรปเปอร์ร่วมกับเทคนิค RT เทคนิค SVM และเทคนิค NB เมื่อนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค RF เทคนิค ANN จากการทดลองพบว่าค่าความไวของแบบจำลองหลังการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ร่วมกับเทคนิค RT สามารถเพิ่มค่าความไวเฉลี่ยขึ้นสูงที่สุดเมื่อใช้เทคนิค KNN ในการสร้างแบบจำลองสูงขึ้นถึงร้อยละ 15.78 เมื่อวิธีแรปเปอร์ร่วมกับเทคนิค SVM สามารถเพิ่มค่าความไวเฉลี่ยให้กับแบบจำลองที่สร้างด้วยเทคนิค KNN สูงขึ้นเพียงร้อยละ 8.38 แต่อย่างไรก็ตามค่าความไวสูงที่สุดของแบบจำลองหลังการคัดเลือกปัจจัย ด้วยวิธีแรปเปอร์ร่วมกับเทคนิค RT มีค่าความไวเฉลี่ยอยู่ที่ 75.86 เมื่อใช้เทคนิค RF และค่าความไวที่ต่ำสุดด้วยวิธีแรปเปอร์ร่วมกับเทคนิค NB มีค่าต่ำสุดเฉลี่ยอยู่ที่ 38.16 เมื่อใช้เทคนิค KNN

3. ผลของค่าความจำเพาะ

ค่าความจำเพาะ คือ ค่าที่แบบจำลองสามารถพยากรณ์ข้อมูลการเลือกหลักสูตรระดับประกาศนียบัตรวิชาชีพที่ไม่ได้เลือกหลักสูตรได้อย่างถูกต้องต่อหลักสูตรที่พยากรณ์ว่ามีการเลือก

หลักสูตร ในการทดลองนี้แบบจำลองการคัดเลือกหลักสูตรจากข้อมูลดั้งเดิม (Raw Data) และจากข้อมูลที่ได้จากการคัดเลือกปัจจัยด้วยแรปเปอร์ร่วมกับเทคนิค RT เทคนิค SVM และเทคนิค NB เมื่อนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค RF เทคนิค ANN จากการทดลองสามารถแสดงผลได้ดังรูปที่ 4



	KNN	C4.5	RF	ANN
Raw Data	84.99%	86.01%	92.53%	52.71%
W+RT	92.35%	85.66%	92.86%	82.29%
W+SVM	88.54%	80.94%	88.18%	78.78%
W+NB	75.44%	78.79%	75.81%	77.00%

รูปที่ 4 ค่าความจำเพาะของแบบจำลอง

จากรูปที่ 4 แสดงผลค่าความจำเพาะของแบบจำลองการคัดเลือกหลักสูตรประกาศนียบัตรวิชาชีพด้วยวิธีแรปเปอร์ จากข้อมูลดั้งเดิม (Raw Data) และจากข้อมูลที่ได้จากการคัดเลือกปัจจัยด้วยแรปเปอร์ร่วมกับเทคนิค RT เทคนิค SVM และเทคนิค NB เมื่อนำข้อมูลชุดฝึกสอนแต่ละชุดมาสร้างแบบจำลองด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค RF เทคนิค ANN จากการทดลองพบว่า ค่าความจำเพาะของแบบจำลองหลังการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ร่วมกับเทคนิค RT สามารถเพิ่มค่าความจำเพาะเฉลี่ยขึ้นสูงที่สุดเมื่อใช้เทคนิค ANN ในการสร้างแบบจำลองสูงขึ้นถึงร้อยละ 32.58 เมื่อวิธีแรปเปอร์ร่วมกับเทคนิค SVM สามารถเพิ่มค่าความจำเพาะเฉลี่ยให้กับแบบจำลองที่สร้างด้วยเทคนิค 26.07 สูงขึ้นเพียงร้อยละ 8.38 แต่อย่างไรก็ตามค่าความจำเพาะสูงที่สุดของแบบจำลองหลังการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ร่วมกับเทคนิค RT มีค่าความจำเพาะเฉลี่ยอยู่ที่ 92.86 เมื่อใช้เทคนิค RF และค่าความจำเพาะที่ต่ำสุดด้วยวิธีแรปเปอร์ร่วมกับเทคนิค NB มีค่าต่ำสุดเฉลี่ยอยู่ที่ 75.44 เมื่อใช้เทคนิค KNN

สรุปและการอภิปรายผล

ในงานวิจัยฉบับนี้ได้มีการเตรียมข้อมูลเป็นการเก็บรวบรวมข้อมูลนักศึกษาที่สำเร็จการศึกษาจาก 5 หลักสูตร ประกอบด้วย หลักสูตรวิชาช่างยนต์ หลักสูตรวิชาช่างอิเล็กทรอนิกส์ หลักสูตรวิชาช่างไฟฟ้า หลักสูตรวิชาการบัญชี และหลักสูตรวิชาคอมพิวเตอร์ธุรกิจ ของวิทยาลัยเทคโนโลยีอุตสาหกรรมศรีสงคราม มหาวิทยาลัยนครพนม จากปีการศึกษา 2555 - 2561 มีจำนวนนักศึกษา 1,350 คน จากการศึกษาปัจจัยที่ได้จากคัดเลือกหลักสูตรด้วยวิธีแรปเปอร์ร่วมกับ RT เทคนิค SVM และเทคนิค NB พบว่า

1. วิธีการแรปเปอร์ร่วมกับเทคนิค RT คัดเลือกปัจจัยได้สูงสุด 9 ปัจจัย (เพศ อาชีพบิดา เกรดวิชาภาษาอังกฤษ เกรดวิชาวิทยาศาสตร์ เกรดวิชาสังคม เกรดวิชาพละศึกษา รายได้บิดา รายได้มารดา และหลักสูตร) ค่าความถูกต้องของแบบจำลองหลังการคัดเลือกปัจจัยด้วยวิธีแรปเปอร์ร่วมกับเทคนิค RT สามารถเพิ่มค่าความถูกต้องเฉลี่ยขึ้นสูงที่สุดเมื่อใช้เทคนิค KNN ในการสร้างแบบจำลองสูงขึ้นถึงร้อยละ 10.30 และเพิ่มค่าความไวสูงขึ้นร้อยละ 15.78

2. ค่าความถูกต้องสูงที่สุดของแบบจำลองที่สร้างด้วยเทคนิค RF มีค่าความถูกต้องเฉลี่ยสูงสุดที่ 89.13 ค่าความไวสูงถึงร้อยละ 75.86 และค่าความจำเพาะสูงถึงร้อยละ 92.86 ซึ่งดีกว่าด้วยเทคนิค Bagging เทคนิค KNN เทคนิค C4.5 เทคนิค ANN เช่นเดียวกับงานวิจัยของ [18] ที่ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์อาชีพสำหรับนักศึกษาระดับปริญญาตรี สร้างด้วยเทคนิค RF ให้ค่าความถูกต้องสูงที่สุดร้อยละ 84.29

กิตติกรรมประกาศ

ขอขอบคุณวิทยาลัยเทคโนโลยีอุตสาหกรรมศรีสงคราม มหาวิทยาลัยนครพนมที่ให้ข้อมูลนักศึกษาที่จบการศึกษาระดับประกาศนียบัตรวิชาชีพ ที่ให้ความอนุเคราะห์ในการเก็บรวบรวมข้อมูลมาใช้ในงานวิจัยครั้งนี้

References

[1] Hand, D. J., and Adams, N. M. (2014). Data Mining. **Wiley StatsRef: Statistics Reference Online**. pp. 1-7. DOI: 10.1002/9781118445112.stat06466.pub2

[2] Kladchuen, R. and Sanrach, C. (2018). An Efficiency Comparison of Algorithms and Feature Selection Methods to Predict the Learning Achievement of Vocational Students. **RMUTT Research Journal Rajamangala University of Technology Thanyaburi**. Vol. 17, Issue 1, pp. 1-10

[3] Wanon, S. M., and Muangsan, R. (2020). A Study and Development of Forecasting Model for the Suitability Characteristics on the Applying Major Selection by Using Data Mining Techniques. **Journal of Management Sciences Suratthani Rajabhat University**. Vol. 7, No. 1, pp. 135-152

[4] Kira, K. and Rendell, L. A. (1992). A Practical Approach to Feature Selection. **Machine Learning Proceedings 1992**. pp. 249-256. DOI: 10.1016/B978-1-55860-247-2.50037-1

[5] Akarachantachote, N. (2016). Feature Selection for High-Dimensional Data in Classification. **Research Methodology and Cognitive Science**. Vol. 8, No. 2, October 2010 - March 2011. pp. 1-13

[6] Witten, I. H. (2014). **More Data Mining with Weka**. MOOC Course Material, Department of Computer Science University of Waikato New Zealand

[7] John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. **Machine Learning Proceedings 1994**. pp. 121-129. DOI: 10.1016/B978-1-55860-335-6.50023-4

[8] Aldous, D. (1991). The Continuum Random Tree. I. **The Annals of Probability**. Vol. 19, No. 1, pp. 1-28

[9] Pinmuang, N. and Thongkam, J. (2018). Classifying Thai Opinions on Online Media Using Text Mining. **Journal of Science and Technology Mahasarakham University**. Vol. 37, No. 3, pp. 372-379

[10] Larose, D. T. and Larose, D. T. (2006). **Data Mining Methods and Models**. John Wiley & Sons, Inc., Hoboken, New Jersey

[11] Breiman, L. (1996). Bagging Predictors. **Machine Learning**. Vol. 24, No. 2, pp. 123-140. DOI: 10.1023/A:1018054314350

[12 Ertugrul, O. F. and Tagluk, M. E. (2017). A Novel Version of k Nearest Neighbor: Dependent Nearest Neighbor. **Applied Soft Computing**. Vol. 55, pp. 480-490. DOI: 10.1016/j.asoc.2017.02.020

[13] Jankowski, D. and Jackowski, K. (2014). Evolutionary Algorithm for Decision Tree Induction. In: Saeed K., Snášel V. (eds). **Computer Information Systems and Industrial Management**. CISIM 2015. Lecture Notes in Computer Science, Vol. 8838. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-662-45237-0_4

[14] Breiman, L. (2001). Random Forests. **Machine learning**. Vol. 45, No. 1, pp. 5-32. DOI: 10.1023/A:1010933404324

[15] Yegnanarayana, B. (2009). **Artificial Neural Networks**. PHI Learning Pvt. Ltd.

[16] Donrat, P. and Thongkam, J. (2018). The Comparison of Performance Models for Predicting Students Success in Vocation Education. **Journal of Science and Technology Mahasarakham University**. Vol. 37, No. 3, pp. 380-388

[17] Baratloo, A., Hosseini, M., Negida, A., and1 El Ashal, G. (2015). Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. **Emergency (Tehran, Iran)**. Vol. 3, No. 2, pp. 48-49

[18] Wanon, S., Arreerard, T., and Sanrach, C. (2018). A Study of Techniques in Predicting Career Counseling for Undergraduate Students of the Computer Program by Using Data Mining Technique. **Journal of Information Technology Management and Innovation**. Vol. 5, No. 1, pp. 164-171