

ระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชน
กรณีศึกษาจังหวัดกาฬสินธุ์

พัฒน์พงษ์ ดลรัตน์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
สิงหาคม 2560
ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

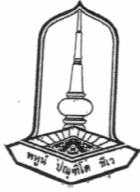


ระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชน
กรณีศึกษาจังหวัดกาฬสินธุ์

พัฒนพงษ์ ดลรัตน์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
สิงหาคม 2560
ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม





คณะกรรมการสอบการศึกษาชั้นคว่ำอิสระ ได้พิจารณาการศึกษาชั้นคว่ำอิสระของ
นายพัฒนพงษ์ ดลรัตน์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบการศึกษาชั้นคว่ำอิสระ

.....
(ผศ.ดร.ภก.กาญจน์ สมประเสริฐศรี) ประธานกรรมการ
(กรรมการบัณฑิตศึกษาประจำคณะ)

.....
(ผศ.ดร.จารี ทองคำ) กรรมการ
(อาจารย์ที่ปรึกษาการศึกษาชั้นคว่ำอิสระ)

.....
(ผศ.รวิชัย ชมศิริ) กรรมการ
(อาจารย์บัณฑิตศึกษาประจำคณะ)

มหาวิทยาลัยยอนมิตีให้การศึกษาค้นคว้าอิสระฉบับนี้ เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัย
มหาสารคาม

.....
(ผศ.ดร.สุจิน บุตระดีสุวรรณ)
คณบดีคณะวิทยาการสารสนเทศ

.....
(ศ.ดร.ประติษฐ์ เทอดทูล)
คณบดีบัณฑิตวิทยาลัย
วันที่ เดือน พ.ศ. 2560



กิตติกรรมประกาศ

การศึกษาค้นคว้าอิสระฉบับนี้ สำเร็จลงได้ด้วยดี เนื่องจากได้รับความกรุณาอย่างสูงจาก ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ อาจารย์ที่ปรึกษาการศึกษาค้นคว้าอิสระ ที่กรุณาให้คำแนะนำ ปรึกษาตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่เป็นอย่างดี ผู้ศึกษาตระหนักถึงความตั้งใจจริงและความทุ่มเทของอาจารย์ และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณ อาชีวศึกษาจังหวัดกาฬสินธุ์ และคณาจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ให้ความอนุเคราะห์ ข้อมูลและคำปรึกษาในการดำเนินการศึกษาค้นคว้าอิสระครั้งนี้จนสำเร็จลุล่วงไปด้วยดี

ขอขอบพระคุณ นายถวิล ดลรัตน์ นางวิไล ดลรัตน์ บิดามารดา ที่ให้กำลังใจตลอดมา ผู้ศึกษามีความหวังว่า การศึกษาค้นคว้าอิสระฉบับนี้จะมีประโยชน์อยู่ไม่น้อย จึงขอมอบ ส่วนดีทั้งหมดนี้ ให้แก่เหล่าคณาจารย์ ที่ได้ประสิทธิ์ประสาทวิชาจนทำให้ผลงานวิจัยเป็นประโยชน์ต่อ ผู้ที่เกี่ยวข้อง และขอมอบความกตัญญูตเวทิตาคุณ แด่บิดา มารดา และผู้มีพระคุณทุกท่าน สำหรับ ข้อบกพร่องต่าง ๆ ที่อาจจะเกิดขึ้นนั้น ผู้ศึกษาก็ขออภัย และยินดีที่จะรับฟังคำแนะนำจากทุกท่านที่ได้ เข้ามาศึกษา เพื่อเป็นประโยชน์ในการพัฒนาต่อไป

พัฒน์พงษ์ ดลรัตน์



ชื่อเรื่อง	ระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษา จังหวัดกาฬสินธุ์		
ผู้ศึกษา	นายพัฒนพงษ์ ดลรัตน์		
ปริญญา	วิทยาศาสตรมหาบัณฑิต	สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ		
มหาวิทยาลัย	มหาวิทยาลัยมหาสารคาม	ปีที่พิมพ์	2560

บทคัดย่อ

งานวิจัยนี้จึงมีจุดประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อการสำเร็จการศึกษสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ สร้างแบบจำลองที่มีประสิทธิภาพในการพยากรณ์และพัฒนาระบบพยากรณ์การสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ ข้อมูลที่ใช้ในงานวิจัยปีการศึกษา 2557-2558 จำนวน 1,266 ข้อมูล 6 เทคนิคที่มีประสิทธิภาพได้ถูกนำมาใช้ในการสร้างแบบจำลอง คือ C4.5, Random Forest, Random Tree, Reduced Error Pruning (REP Tree), k-Nearest Neighbors (k-NN) และ Support Vector Machine (SVM) คณะผู้วิจัยได้ใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ จากการศึกษพบว่าปัจจัยที่มีผลต่อการสำเร็จการศึกษจำนวน 15 ปัจจัย คือ อำเภอที่นักเรียนอยู่ อาชีพบิดา อาชีพมารดา เกรดเฉลี่ยกลุ่มการเรียนอาชีวะและเทคโนโลยี เกรดเฉลี่ยกลุ่มคณิตศาสตร์ เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม เกรดเฉลี่ยกลุ่มศิลปะ เกรดเฉลี่ยกลุ่มภาษาไทย เกรดเฉลี่ยกลุ่มวิทยาศาสตร์ สถานภาพครอบครัว เกรดเฉลี่ยกลุ่มสุขศึกษาพลศึกษา เพศ ประเภทวิชาที่จบ เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ แบบจำลองที่มีประสิทธิภาพในการพยากรณ์สูงที่สุดคือแบบจำลองที่สร้างจากเทคนิค C4.5 ซึ่งมีค่าความถูกต้อง (Accuracy) เท่ากับ 95.09% ค่าความแม่นยำ (Precision) เท่ากับ 95.10% และค่าความระลึก (Recall) เท่ากับ 95.10%

คำสำคัญ: แบบจำลองการพยากรณ์การสำเร็จการศึกษา; เหมืองข้อมูล; ประกาศนียบัตรวิชาชีพ



TITLE Graduation Forecasting System for Private Vocational Education: Case Study of Kalasin Province

AUTHOR Mr. Pattanaphong Donrat

DEGREE Master of Science **MAJOR** Information Technology

ADVISOR Jaree Thongkum, Ph.D.

UNIVERSITY Mahasarakham University **YEAR** 2017

ABSTRACT

The purposes of this Thesis are to study the factors which affect the student's graduation of private vocational education institutes in Kalasin Province, to build effective prediction models and to develop graduation systems for people who are interested in vocational education. The data were collect from private vocational education institutes in Kalasin Province in 2015-2016. The total data are 1,266 records. In order to build effective prediction models, six powerful techniques including C4.5, Random Forest, Random Tree, Reduced Error Pruning (REP Tree), k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM) were utilized. 10-fold cross validation was used to divide the data into training and test sets. The experimental results showed that district, father career, mother career, GPA of career and technology, GPA of mathematics, GPA of social studies religion and culture, GPA of art group, GPA of Thai languages, GPA of science group, family status, GPA of health education group, sex, type of course, GPA of foreign languages are factors for predicting the student's graduation. Moreover, the most efficient model for predicting student's graduation is C4.5 which has accuracy, precision and recall up to 95.09%, 95.10% and 95.10%, respectively.

Key Words: Models Forecasting Graduation; Data Mining; Vocational Education Certificate.



สารบัญ

	หน้า
กิตติกรรมประกาศ	ก
บทคัดย่อภาษาไทย	ข
บทคัดย่อภาษาอังกฤษ	ค
สารบัญ	ง
สารบัญตาราง	ฉ
สารบัญภาพ	ช
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 ความมุ่งหมายของการศึกษา	2
1.3 ความสำคัญของการศึกษา	2
1.4 ขอบเขตของการศึกษา	2
1.5 นิยามศัพท์เฉพาะ	3
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	4
2.1 ทฤษฎีที่เกี่ยวข้อง	4
2.1.1 สถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์	4
2.1.2 สภาวะการขาดแคลนแรงงานฝีมือในประเทศไทย	5
2.1.3 เหมืองข้อมูล	6
2.1.4 การวิเคราะห์ปัจจัย	7
2.1.5 เทคนิคในเหมืองข้อมูล	9
2.1.6 การวัดประสิทธิภาพแบบจำลอง	19
2.2 งานวิจัยที่เกี่ยวข้อง	20
บทที่ 3 วิธีดำเนินการศึกษา	22
3.1 การเตรียมข้อมูล	22
3.2 การทำกระบวนการก่อนการสร้างแบบจำลอง	23
3.3 การสร้างแบบจำลอง	28
3.4 การวัดประสิทธิภาพของแบบจำลอง	28
3.5 การพัฒนาระบบ	28
บทที่ 4 ผลการศึกษา	29
4.1 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการศึกษา	29
4.2 ผลการวัดประสิทธิภาพแบบจำลองทั้ง 6 เทคนิค	32
4.3 ผลการพัฒนาระบบพยากรณ์การสำเร็จการศึกษา	34



	หน้า
บทที่ 5 วิจัยรณัผลการทดลอง สรุป และข้อเสนอแนะ	37
5.1 วิจัยรณัผลการทดลอง	37
5.2 สรุป	38
5.3 ข้อเสนอแนะ	38
เอกสารอ้างอิง	39
ภาคผนวก	44
ประวัติย่อผู้ศึกษา	54



สารบัญตาราง

	หน้า
ตารางที่ 2.1 ตัวอย่างชุดข้อมูลการซื้อเครื่องปรับอากาศ	10
ตารางที่ 2.2 แสดงตัวอย่างการทดสอบประสิทธิภาพ เมื่อ $K=5$	19
ตารางที่ 3.1 แสดงจำนวนนักเรียนประจำปีการศึกษา 2557-2558	22
ตารางที่ 3.2 แสดงการจำแนกชนิดของตัวแปร	23
ตารางที่ 3.3 แสดงการแปลงค่าตัวแปร	24
ตารางที่ 4.1 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา	29



สารบัญภาพ

	หน้า
ภาพที่ 2.1 ตัวอย่างการแบ่งกลุ่มข้อมูลที่ แอตทริบิวต์ Age	10
ภาพที่ 2.2 ลักษณะการทำงานของเทคนิค Random Forest	12
ภาพที่ 2.3 การวางตัวของข้อมูลในลักษณะเชิงเส้น	17
ภาพที่ 4.1 การวิเคราะห์ปัจจัยด้วย Gain Ratio	31
ภาพที่ 4.2 ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ทั้ง 6 เทคนิค จากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation	32
ภาพที่ 4.3 ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความแม่นยำ (Precision) ทั้ง 6 เทคนิค จากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation	33
ภาพที่ 4.4 ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความระลึก (Recall) ทั้ง 6 เทคนิค จากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation	34
ภาพที่ 4.5 แสดงหน้าระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชน จังหวัดกาฬสินธุ์	35
ภาพที่ 4.6 แสดงหน้าผลการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชน จังหวัดกาฬสินธุ์	36



บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันประเทศไทยนั้นอยู่สภาวะขาดแคลนแรงงานฝีมือ [1] กระทรวงศึกษาธิการจึงได้มีนโยบายในปีพุทธศักราช 2554 การเปลี่ยนแปลงสัดส่วนของผู้เรียนอาชีวศึกษาต่อผู้เรียนสายสามัญจากเดิม 40:60 เป็น 60:40 ภายในปีพุทธศักราช 2561 ซึ่งเป็นปีสิ้นสุดของการปฏิรูปการศึกษาทศวรรษที่ 2

การเพิ่มจำนวนผู้สำเร็จการศึกษาตามระยะเวลาของหลักสูตรของอาชีวศึกษา ซึ่งหลักสูตรอาชีวศึกษาแบ่งออกได้เป็น 2 หลักสูตรคือ 1) หลักสูตรประกาศนียบัตรวิชาชีพ (ปวช.) มีระยะเวลาการศึกษาตามหลักสูตร 3 ปี 2) หลักสูตรประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) มีระยะเวลาการศึกษาตามหลักสูตร 2 ปี ซึ่งการจัดให้ผู้เรียนสำเร็จการศึกษาตามระยะเวลาของหลักสูตรนั้น เป็นตัวบ่งชี้ตัวหนึ่งที่แสดงว่าสถานศึกษาได้จัดการศึกษาอย่างมีคุณภาพและประสิทธิภาพ ตามตัวบ่งชี้ของการประกันคุณภาพการศึกษาทั้งการประกันคุณภาพภายในและการประกันคุณภาพภายนอก

สถานศึกษาอาชีวศึกษาเอกชน [2] ก็เป็นหน่วยงานหนึ่งในการจัดการศึกษาอาชีวศึกษาที่มีบทบาท สำคัญในการให้บริการการศึกษาทางด้านวิชาชีพเช่นเดียวกับสถานศึกษาอาชีวศึกษารัฐบาล ซึ่งล้วนแต่มีบทบาทและความสำคัญอย่างยิ่งต่อการพัฒนาประเทศเพราะเป็นการศึกษาที่จัดเตรียมบุคคลให้มีอาชีพเป็นหลักในอนาคต และช่วยให้อาชีพที่มีอยู่แล้วมีความก้าวหน้าในอาชีพตน

แต่ในความเป็นจริงนั้นสถานศึกษาอาชีวศึกษาเอกชนนั้นมีข้อเสียเปรียบสถานศึกษาอาชีวศึกษารัฐบาลเช่น งบประมาณในการจัดซื้อสื่อการเรียนการสอนที่ชัดเจนคือประเภทวิชาอุตสาหกรรมเป็นต้น เนื่องจากงบประมาณส่วนใหญ่ของสถานศึกษาอาชีวศึกษาเอกชนนั้นมาจากเงินอุดหนุนรายบุคคลที่รัฐสนับสนุนโดยเงินที่ได้จะมากหรือน้อยนั้นก็ขึ้นอยู่กับจำนวนนักเรียนเป็นหลัก ดังนั้นหากสถานศึกษาอาชีวศึกษาเอกชนมีจำนวนนักเรียนมากก็จะได้งบประมาณมาพัฒนาการเรียนการสอนให้ได้ประสิทธิภาพมากยิ่งขึ้น ซึ่งสถานการณ์ปัจจุบันของสถานศึกษาอาชีวศึกษาในจังหวัดกาฬสินธุ์ ซึ่งมีการรวมเอาสถานศึกษาอาชีวศึกษาทั้งภาครัฐและเอกชนมาไว้ภายใต้หน่วยงานเดียวกันคืออาชีวศึกษาจังหวัดกาฬสินธุ์โดยมีสถานศึกษาทั้งสิ้น 18 สถานศึกษา แบ่งเป็นสถานศึกษารัฐบาลจำนวน 6 แห่ง และสถานศึกษาของเอกชน 12 แห่ง แม้จำนวนสถานศึกษาของเอกชนจะมากกว่าจำนวนสถานศึกษาของรัฐบาลแต่ก็ยังมีจำนวนนักเรียน-นักศึกษา น้อยกว่าสถานศึกษาของรัฐบาลอยู่มาก

เหมือนข้อมูลเป็นกระบวนการในการค้นหาความรู้จากข้อมูล นักวิจัยหลายท่านได้นำเอากระบวนการของเหมือนข้อมูลมาใช้ในการสร้างแบบจำลองเพื่อการพยากรณ์ เช่น Kavitha และ Rangasamy [3] ได้เสนอการทำนายการรอดชีวิตจากมะเร็งเต้านมโดยใช้เทคนิค Naive Bayes และ C4.5 ผลการทดลองพบว่า เทคนิค C4.5 ให้ความถูกต้อง 97.9% ซึ่งสูงกว่า Naive Bayes



เพียงฤทัย หนูสวัสดิ์ [4] ได้เสนอการสร้างโมเดลทำนายอัตราการใช้พลังงานของแบตเตอรี่มือถือ โดยใช้เทคนิคเหมืองข้อมูล 2 เทคนิค คือ Perceptron Neural Network และ Support Vector Machines (SVM) แบบ kernel ผลการวิจัยพบว่าเทคนิค SVM แบบ kernel ให้ประสิทธิภาพความแม่นยำมากที่สุด เทคนิคในเหมืองข้อมูลที่เป็นที่นิยมและมีประสิทธิภาพ [5] เช่น C4.5, The k-means, SVM, The Apriori algorithm, The EM algorithm, PageRank, AdaBoost, kNN: k-nearest neighbor, Naive Bayes, CART, Random Forest, Random Tree และ REP Tree เป็นต้น

ดังนั้นผู้วิจัยจึงได้มีการคิดค้นหาวิธีการแก้ปัญหาโดยการรวบรวมข้อมูลของนักเรียน-นักศึกษาที่สำเร็จการศึกษาจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ มาทำการสร้างแบบจำลองเพื่อการพยากรณ์ด้วยเทคนิคของเหมืองข้อมูล (Data Mining) ประกอบด้วยใช้ 6 เทคนิค คือ C4.5, Random Forest, Random Tree, REP Tree, k-NN และ SVM ซึ่งเป็นเทคนิคที่มีความแม่นยำสูง มาทำการวิเคราะห์เปรียบเทียบทั้ง 6 เทคนิค แล้วจึงนำแบบจำลองที่มีประสิทธิภาพสูงที่สุดนั้นไปสร้างแบบจำลอง เพื่อนำไปพัฒนาระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์ ซึ่งระบบสามารถช่วยส่งเสริมให้ครูแนะแนวและผู้บริหารมีข้อมูลสารสนเทศเพื่อวางแผนในการแนะแนวให้นักเรียน-นักศึกษาเข้าศึกษาต่อในสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

1.2 ความมุ่งหมายของการศึกษา

1.2.1 เพื่อศึกษาปัจจัยที่มีผลต่อการสำเร็จการศึกษาสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

1.2.2 เพื่อสร้างแบบจำลองที่มีประสิทธิภาพในการพยากรณ์

1.2.3 เพื่อพัฒนาระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์

1.3 ความสำคัญของการศึกษา

ระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์สามารถช่วยส่งเสริมให้ครูแนะแนวและผู้บริหารมีข้อมูลสารสนเทศเพื่อวางแผนในการแนะแนวให้นักเรียนเข้าศึกษาต่อในสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

1.4 ขอบเขตของการศึกษา

การศึกษารั้งนี้มุ่งเน้นการสร้างแบบจำลองการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์ซึ่งผู้ศึกษาได้กำหนดขอบเขตการศึกษาดังนี้



1.4.1 เตรียมข้อมูลนักเรียนระดับประกาศนียบัตรวิชาชีพ ปีการศึกษา 2557-2558 ของสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

1.4.2 สร้างแบบจำลองเพื่อทำการพยากรณ์ของ 6 เทคนิค คือ C4.5, Random Forest, Random Tree, REP Tree, k-NN และ SVM

1.4.3 วัดประสิทธิภาพของทั้ง 6 เทคนิค ด้วย k-fold cross validation และวิเคราะห์ประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

1.4.4 สร้างระบบการพยากรณ์การสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์

1.5 นิยามศัพท์เฉพาะ

1.5.1 นักเรียน หมายถึง ผู้เรียนที่เรียนในระดับประกาศนียบัตรวิชาชีพ (ปวช.) ในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์

1.5.2 ปัจจัยที่มีอิทธิพลต่อการสำเร็จการศึกษา หมายถึง ตัวแปรที่ทำให้ผลต่อการสำเร็จการศึกษาสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์



บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษารังนี้ ผู้ศึกษาได้ศึกษาทฤษฎี ตำราและงานวิจัยที่เกี่ยวข้องกับการพัฒนาระบบพยากรณ์ การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์ ผู้ศึกษาได้เสนอทฤษฎี และงานวิจัยที่เกี่ยวข้อง ดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 สถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

สถานศึกษาอาชีวศึกษาเอกชนสังกัดสำนักงานคณะกรรมการส่งเสริมการศึกษาเอกชน (สช.) และมีหน่วยงานต้นสังกัดคือสำนักงานเขตพื้นที่การศึกษาประถมศึกษากาฬสินธุ์ เขต1 เขต2 และ เขต3 ตามพื้นที่ที่สถานศึกษาดำเนินอยู่ แต่ใช้หลักสูตรของสำนักงานคณะกรรมการการอาชีวศึกษา (สอศ.) จนมาถึงวันที่ 12 กุมภาพันธ์ 2559 เว็บไซต์ราชกิจจานุเบกษาได้เผยแพร่คำสั่งหัวหน้าคณะรักษาความสงบแห่งชาติที่ 8/2559 เรื่องการบริหารจัดการรวมสถานศึกษาอาชีวศึกษาภาครัฐและภาคเอกชน โดยให้รวมสถานศึกษาทั้งภาครัฐและเอกชนไว้ด้วยกันภายใต้สังกัดอาชีวศึกษาจังหวัดกาฬสินธุ์ (อศจ.กาฬสินธุ์)

หลักสูตรประกาศนียบัตรวิชาชีพเป็นหลักสูตรหลังมัธยมศึกษาตอนต้นหรือเทียบเท่าด้านวิชาชีพที่สอดคล้องกับแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ แผนการศึกษาแห่งชาติ และประชาคมอาเซียนเพื่อผลิตและพัฒนากำลังคนระดับฝีมือให้มีสมรรถนะ มีคุณธรรม จริยธรรม และจรรยาบรรณ วิชาชีพสามารถประกอบอาชีพได้ตรงตามความต้องการของสถานประกอบการและการประกอบอาชีพ อิสระเป็นหลักสูตรที่เปิดโอกาสให้เลือกเรียนได้อย่างกว้างขวาง เน้นสมรรถนะเฉพาะด้านด้วยการปฏิบัติจริงสามารถเลือกวิธีการเรียนตามศักยภาพและโอกาสของผู้เรียน เปิดโอกาสให้ผู้เรียนสามารถเทียบโอนผลการเรียน สะสมผลการเรียน เทียบความรู้และประสบการณ์จากแหล่งวิทยาการ สถานประกอบการ และสถานประกอบอาชีพอิสระ เป็นหลักสูตรที่สนับสนุนการประสานความร่วมมือในการจัดการศึกษา ร่วมกันระหว่างหน่วยงานและองค์กรที่เกี่ยวข้อง ทั้งภาครัฐและเอกชน เป็นหลักสูตรที่เปิดโอกาสให้ สถานศึกษา สถานประกอบการ ชุมชนและท้องถิ่น มีส่วนร่วมในการ พัฒนาหลักสูตรให้ตรงตามความต้องการและสอดคล้องกับสภาพยุทธศาสตร์ของภูมิภาค เพื่อเพิ่มขีดความสามารถในการแข่งขันของประเทศ

โครงสร้างของหลักสูตรประกาศนียบัตรวิชาชีพ พุทธศักราช 2556 แบ่งเป็น 3 หมวดวิชา และกิจกรรมเสริมหลักสูตร ดังนี้

1. หมวดวิชาทักษะชีวิต
 - 1.1 กลุ่มวิชาภาษาไทย
 - 1.2 กลุ่มวิชาภาษาต่างประเทศ
 - 1.3 กลุ่มวิชาวิทยาศาสตร์
 - 1.4 กลุ่มวิชาคณิตศาสตร์



- 1.5 กลุ่มวิชาสังคมศึกษา
- 1.6 กลุ่มวิชาสุขศึกษาและพลศึกษา
2. หมวดวิชาทักษะวิชาชีพ
 - 2.1 กลุ่มทักษะวิชาชีพพื้นฐาน
 - 2.2 กลุ่มทักษะวิชาชีพเฉพาะ
 - 2.3 กลุ่มทักษะวิชาชีพเลือก
 - 2.4 ฝึกประสบการณ์ทักษะวิชาชีพ
 - 2.5 โครงการพัฒนาทักษะวิชาชีพ
3. หมวดวิชาเลือกเสรี
4. กิจกรรมเสริมหลักสูตร
 - 2.1.2 สภาวะการขาดแคลนแรงงานฝีมือในประเทศไทย

จงจิตต์ ฤทธิรงค์, รินา ต๊ะดี [6] กล่าวว่าความต้องการแรงงานฝีมือในอุตสาหกรรมที่สำคัญ 3 ประเภท คือ อุตสาหกรรมยานยนต์ การผลิตอาหารและการท่องเที่ยว นั้นยังมีมากโดยเฉพาะแรงงานที่อยู่ในระดับปฏิบัติงานคือกลุ่มผู้ที่สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพและประกาศนียบัตรวิชาชีพชั้นสูง เนื่องจากบุคลากรกลุ่มนี้มีทักษะที่จำเป็นต่อการปฏิบัติงาน จากการฝึกประสบการณ์วิชาชีพซึ่งมีความร่วมมือระหว่างสถานบันอาชีวศึกษาและภาคเอกชน ในการฝึกทักษะให้แก่ผู้เรียนมีความสามารถและทักษะตรงตามความต้องการของตลาดแรงงาน แต่อย่างไรก็ตาม การผลิตแรงงานฝีมือยังไม่เพียงพอและมีแนวโน้มว่าจะขาดแคลนอันเนื่องมาจากค่านิยมที่ให้ความสำคัญกับปริญญาบัตรมากกว่าความสามารถและทักษะในการทำงาน ดังสะท้อนออกมาเป็นอัตราค่าจ้างที่แปรผันตามระดับวุฒิการศึกษา แนวทางการแก้ไขคือ สร้างความเข้าใจให้กับผู้ประกอบการเปลี่ยนแนวคิดที่จะให้เรียนสูงๆ แต่เมื่อจบออกมาแล้วไม่มีงานทำเนื่องจากตำแหน่งงานน้อย และส่งเสริมเพิ่มโอกาสให้กับผู้เรียนทางด้านหลักสูตรทวิภาคี โดยให้ทุนการศึกษาแก่นักเรียนที่เรียนในหลักสูตรดังกล่าวเพื่อเพิ่มแรงจูงใจสำหรับนักเรียนที่มีข้อจำกัดด้านการเงิน รวมทั้งให้ข้อมูลเกี่ยวกับแนวทางการศึกษาในระดับที่สูงขึ้นและการทำงานสายอาชีพอย่างตรงไปตรงมา และอีกอย่างคืออัตราค่าจ้างแบบทางเลือกโดยพิจารณาทักษะและความสามารถในการปฏิบัติงานโดยไม่ยึดติดกับวุฒิการศึกษาจะช่วยลดจำนวนผู้จบการศึกษาระดับปริญญาแต่ขาดทักษะและประสบการณ์ได้

พิเชษฐ์ สุขเสกสรร [7] ได้ทำการสำรวจประสบการณ์ของหน่วยงานกลุ่มรับเหมาก่อสร้างในการเผชิญปัญหาการขาดแคลนแรงงานก่อสร้างกลุ่มช่างฝีมือหรือช่างเทคนิค คือผู้ที่สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพและประกาศนียบัตรวิชาชีพชั้นสูงโดยมีผลการสำรวจดังนี้ 1) หน่วยงานของท่านเคยประสบปัญหาการขาดแคลนแรงงานกลุ่มนี้คิดเป็นร้อยละ 89.7 2) หน่วยงานของท่านกำลังประสบปัญหาการขาดแคลนแรงงานกลุ่มนี้คิดเป็นร้อยละ 60.3 3) หน่วยงานของท่านได้เคยประเมินสถานการณ์หรือได้ประเมินสถานการณ์การขาดแคลนแรงงานกลุ่มนี้ในอนาคต คิดเป็นร้อยละ 75.6

ส่วนเศรษฐกิจภาค ธนาคารแห่งประเทศไทย สำนักงานภาคตะวันออกเฉียงเหนือ [8] ได้ทำการศึกษาเรื่อง “ความไม่สมดุลของตลาดแรงงานไทย นโยบายการขาดแคลนแรงงาน” ผลการศึกษาพบว่ากลุ่มธุรกิจขนาดใหญ่ เช่น กลุ่มสิ่งทอ/เครื่องแต่งกายและเครื่องหนัง กลุ่มอิเล็กทรอนิกส์และเครื่องใช้ไฟฟ้า กลุ่มผลิตภัณฑ์อาหารและเครื่องดื่ม กลุ่มผลิตภัณฑ์ยางและ



พลาสติกและกลุ่มก่อสร้าง ซึ่งกลุ่มดังกล่าวส่วนใหญ่เป็นกลุ่มผู้ผลิตเพื่อการส่งออกและมีบทบาทสำคัญต่อการขับเคลื่อนเศรษฐกิจไทยมาโดยตลอดแต่กลับประสบปัญหาการแคลนแรงงานสูงถึงร้อยละ 76.3 ซึ่งสาเหตุของการเกิดปัญหาขาดแคลนแรงงานมี 4 ประการหลัก คือ 1) การเปลี่ยนแปลงโครงสร้างประชากร ซึ่งเป็นผลมาจากอัตราเจริญพันธุ์ที่ต่ำกว่าระดับทดแทน และการก้าวเข้าสู่สังคมผู้สูงอายุอย่างรวดเร็ว ส่งผลให้กำลังแรงงานใหม่ที่จะเข้าสู่ตลาดแรงงานลดลงอย่างต่อเนื่องไม่เพียงพอสำหรับการทดแทนแรงงานที่จะเกษียณได้ทัน 2) เกิดจากความไม่สอดคล้องกันระหว่างคุณลักษณะของแรงงานที่เข้าสู่ตลาดกับความต้องการของตลาดแรงงานใน 2 ประเด็น คือ ประเด็นที่ 1 ด้านอายุ ภาคธุรกิจส่วนใหญ่ต้องการแรงงานที่มีอายุระหว่าง 20-39 ปี เป็นจำนวนมาก แต่แรงงานกลุ่มดังกล่าวกลับมีจำนวนลดลงประกอบกับแรงงานต่างมั่งเรียนในระดับที่สูงขึ้น โดยเฉพาะในระดับอุดมศึกษา ซึ่งเป็นผลมาจากค่านิยมของสังคมไทยที่มุ่งให้ความสำคัญกับวุฒิการศึกษามากกว่าความรู้และประสบการณ์ ประเด็นที่ 2 ด้านโครงสร้างการศึกษาที่ผู้เรียนส่วนใหญ่นอกจากจะเลือกเรียนสายสามัญมากกว่าสายอาชีวศึกษาแล้วในระดับอาชีวศึกษาเองยังเลือกเรียนสาขาที่เป็นที่ต้องการของตลาดแรงงานในระดับค่อนข้างน้อย และในระดับอุดมศึกษาส่วนใหญ่เลือกเรียนสายสังคมศาสตร์มากกว่าสายวิทยาศาสตร์ ในขณะที่ตลาดแรงงานต้องการแรงงานด้านสายวิทยาศาสตร์มากกว่าส่งผลให้ปริมาณแรงงานไม่สอดคล้องกับความต้องการของตลาดแรงงาน 3) ภาคการผลิตไทยมีผลิตภาพการผลิตที่ต่ำเนื่องจากที่ผ่านมามีความได้เปรียบในการแข่งขันจากการใช้แรงงานพื้นฐานที่มีค่าแรงต่ำเป็นหลัก ในขณะที่ปัจจุบันที่หลายประเทศได้แก่อินโดนีเซีย เวียดนาม และฟิลิปปินส์ มีค่าจ้างต่ำกว่าไทยและมีกำลังแรงงานที่มากกว่า 4) แรงงานไทยที่เดินทางไปทำงานในต่างประเทศยังคงสูง มีสาเหตุมาจากความแตกต่างของระดับค่าจ้าง กล่าวคือแรงงานที่ไปทำงานในต่างประเทศได้รับค่าจ้างสูงกว่าค่าจ้างแรงงานในประเทศเฉลี่ย 3.2 เท่า และยิ่งสูงกว่าแรงงานนอกระบบเฉลี่ย 4.5 เท่า

จากที่ผู้ศึกษาได้ศึกษามานี้ธุรกิจและอุตสาหกรรมในภาคต่างๆ นั้น ยังคงมีความต้องการแรงงานที่สำเร็จการศึกษาในระดับประกาศนียบัตรวิชาชีพและประกาศนียบัตรวิชาชีพชั้นสูงอยู่เป็นจำนวนมากจึงมีความจำเป็นอย่างยิ่งที่จะต้องแนะแนวและส่งเสริมให้ผู้เรียนสายอาชีวศึกษาเพิ่มขึ้น ซึ่งตรงกับระบบที่ผู้ศึกษาจะพัฒนาขึ้นเพื่อมุ่งที่จะเพิ่มจำนวนผู้เรียนในสายอาชีวศึกษาซึ่งสอดคล้องกับข้อมูลดังกล่าว

2.1.3 เหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) [9] คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม

ขั้นตอนการทำเหมืองข้อมูล

1. Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป
2. Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูล

ชุดเดียวกัน

3. Data Selection เป็นขั้นตอนการดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกไว้
4. Data Transformation เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน
5. Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่



6. Pattern Evaluation เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล

7. Knowledge Representation เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ

2.1.4 การวิเคราะห์ปัจจัย

การวิเคราะห์ปัจจัย [10] เป็นกระบวนการก่อนการทำเหมืองข้อมูลเพื่อทำการคัดเลือกเอาเฉพาะแอตทริบิวต์ที่เป็นปัจจัยมีผลต่อการพยากรณ์เพื่อลดระยะเวลาการสร้างแบบจำลองและลดการใช้ทรัพยากรคอมพิวเตอร์แต่ยังคงความถูกต้องของการพยากรณ์ไว้เช่นเดิมซึ่งมีเทคนิคดังนี้

2.1.4.1 เทคนิค Cfs Subset Evaluation [11] เป็นการหาปัจจัยที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์ โดยปัจจัยที่ถูกคัดเลือกใช้สำหรับจำแนกประเภทของข้อมูลสามารถจัดการกับปัจจัยที่ไม่เกี่ยวข้องกันที่จะให้ประสิทธิภาพในการจำแนกประเภทข้อมูลที่ต่ำคำนวณได้จากสมการที่ 1

$$Merit_s = \frac{\overline{r_{cf}}}{\sqrt{k + k(k-1)r_{ff}}} \quad (1)$$

เมื่อ $Merit_s$ คือ Heuristic ที่เป็นกลุ่มของแอตทริบิวต์ S ที่ประกอบด้วย K แอตทริบิวต์ที่ถูกคัดเลือก

$\overline{r_{cf}}$ คือ ค่าเฉลี่ยของกลุ่มของ Attribute ที่ถูกคัดเลือกของความสัมพันธ์ที่เกี่ยวข้องกับประเภทของข้อมูล

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยของกลุ่ม Attribute ที่ถูกคัดเลือกจากความสัมพันธ์ที่เกี่ยวข้องกันภายในกลุ่มแอตทริบิวต์ที่ถูกคัดเลือก

มีนักวิจัยจำนวนมากได้นำเทคนิค Cfs Subset Evaluation มาใช้ในการวิเคราะห์ปัจจัย เช่น นิเวศ จิระวิจิตชัย [12] ได้เสนอ การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ ทำการทดลองกับ 7 เทคนิค ซึ่งประกอบด้วย Naïve Bayes, Multilayer Perceptron, Radial Basis Function Network, SVM, k-NN, Decision Tree, Ripper และทำการศึกษาเปรียบเทียบวิธีลดคุณลักษณะที่เหมาะสมด้วยวิธี Correlation-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) ผลการทดลองพบว่าการลดคุณลักษณะด้วยวิธี Correlation-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) ส่งผลให้ประสิทธิภาพความถูกต้อง (Accuracy) ในการจำแนกประเภทของโรคใกล้เคียงกับการไม่ลดคุณลักษณะ แต่การลดมิติของข้อมูลดังกล่าวทำให้ประหยัดทรัพยากรของระบบคอมพิวเตอร์และระยะเวลาในการเรียนรู้เพื่อสร้างโมเดลได้เป็นอย่างดี



พฤทธิพงศ์ เฟื่องศิริ และคณะ [13] ได้เสนอการลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์ตโฟน ด้วยวิธีการจำแนกข้อมูลแบบ ANN และการลดจำนวนมิติข้อมูลด้วยเทคนิค PAA, Cfs, Information Gain และ Relief ผลการทดลองพบว่าเมื่อใช้ ANN ร่วมกับ PAA ให้ค่าความถูกต้องที่ 84.1%, เมื่อใช้ ANN ร่วมกับ Cfs ให้ค่าความถูกต้องที่ 84.48%, เมื่อใช้ ANN ร่วมกับ Information Gain ให้ค่าความถูกต้องที่ 80.46% และเมื่อใช้ ANN ร่วมกับ ReliefF ให้ค่าความถูกต้องที่ 82.57% จะเห็นได้ว่า ANN เมื่อใช้ร่วมกับ Cfs จะให้ค่าความถูกต้องสูงที่สุด

พรเทพ คงไชย และรัชฎา คงคะจันทร์ [14] ได้เสนอ การศึกษาเชิงเปรียบเทียบในการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับการทำเหมืองข้อมูลเพื่อพยากรณ์โอกาสการสำเร็จการศึกษาของนักศึกษา ผลการทดลองพบว่าเมื่อใช้เทคนิค SVM ร่วมกับ Cfs ได้ค่าความถูกต้องที่ 90.50% ซึ่งเป็นค่าความถูกต้องที่สูงและยังสามารถลดคุณลักษณะจาก 20 คุณลักษณะให้เหลือเพียง 4 คุณลักษณะ

2.1.4.2 เทคนิค Gain Ratio Attribute Evaluation [15] เป็นกระบวนการหนึ่งในการหาคุณลักษณะข้อมูลในการตัดสินใจ ซึ่งได้มีการพัฒนาต่อจากวิธีการ Information Gain โดยมีการเพิ่มการหา split info เข้ามาแล้วนำไปหารกับค่าที่ได้จาก Information Gain จะได้ค่า Gain Ratio ซึ่งเป็นวิธีการวัดค่าตัวแปรเพื่อนำไปตั้งโหนดเพื่อตัดสินใจของวิธีต้นไม้การตัดสินใจแบบ C4.5 ค่า Gain ratio ได้จากสมการที่ 2 และ 3

$$Gain(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (2)$$

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (3)$$

เมื่อ D = ข้อมูลทั้งหมดของแถวตาม Class
 A = ตัวแปร
 D_j = ข้อมูลของตัวแปรนั้น ณ ตำแหน่ง j

เมื่อได้ค่าจากผลลัพธ์จากการนำข้อมูลไปผ่านสมการนี้ ทำให้สามารถทราบได้ว่าตัวแปรใดมีความสัมพันธ์ และนำคุณลักษณะที่มีผลลัพธ์มากที่สุดไปตั้งเป็นโหนดในการตัดสินใจ มีนักวิจัยจำนวนมากได้นำเทคนิค Gain Ratio Attribute Evaluation มาใช้ในการวิเคราะห์ปัจจัย เช่น วิระ จิริกิจอนุสรณ์ และชุลีรัตน์ จรัสกุลชัย [16] ทำการคาดการณ์ภาษีมูลค่าเพิ่มด้วยเทคนิคของเหมืองข้อมูล ทำการเปรียบเทียบระหว่าง linear regression กับ multilayer perceptron มีการใช้วิธี gain ratio และวิธี information gain เพื่อหาความสัมพันธ์ของตัวแปร และคัดกรองตัวแปร ผลการทดลองพบว่า multilayer perceptron เมื่อใช้ร่วมกับวิธี gain ratio มีการคาดการณ์ที่แม่นยำถึง 98% ซึ่งมากกว่า linear regression



สุโกโชค เรืองศรี [17] ได้เสนอการพัฒนาระบบประมาณการใช้กระดาษภายในแผนกผู้โดยสารด้วยวิธี ANN โดยเรียนรู้ข้อมูล 4 ชุดคือ ชุดที่ 1 ข้อมูลดิบ ชุดที่ 2 วิธีวิเคราะห์ linear regression ชุดที่ 3 ผ่านวิธี Information Gain ได้ 10 ตัวแปร และ ชุดที่ 4 ผ่านวิธี Gain Ratio ได้ 10 ตัวแปร ผลการทดลองพบว่า ข้อมูลชุดที่ 4 ผ่านวิธี Gain Ratio ให้ค่า MSE ต่ำที่สุดคือ ขนาด 4 ชั้น มีฟังก์ชันกระตุ้นแบบ Tansig-Logsig-Tansig-Purelin โดยมีขนาดในแต่ละชั้นตามลำดับ 30-15-7-1 ให้ค่า MSE ต่ำที่สุดคือ 0.001050 และมีค่า R คือ 0.90874

2.1.4.3 เทคนิค Wrapper Subset Evaluation [18] เป็นวิธีการอย่างง่ายซึ่งสร้างเซตของลักษณะเฉพาะใหม่ขึ้นมาโดยการเพิ่มหรือลดคุณลักษณะจากเซตของลักษณะเฉพาะเดิม จากนั้นสร้างตัวจำแนกประเภทจากเซตของลักษณะเฉพาะใหม่ที่ได้แล้วประเมินประสิทธิภาพของตัวจำแนกซึ่งเซตใหม่ที่ได้ผลลัพธ์ที่มีประสิทธิภาพมากที่สุดจะถูกเลือกใช้ มีนักวิจัยจำนวนมากได้นำเทคนิค Wrapper Subset Evaluation มาใช้ในการวิเคราะห์ปัจจัย เช่น สุคนธ์ทิพย์ วงศ์พันธ์ [19] ได้เสนอการเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรม การกระทำความผิดของนักเรียนระดับอาชีวศึกษา ผลการทดลองพบว่า Genetic Search ร่วมกับ Wrapper Subset Evaluator ให้ผลการคัดเลือกคุณลักษณะที่เหมาะสมก่อนนำไปประมวลผลต่อในเทคนิค C4.5 ให้ค่าความถูกต้องสูงที่สุดถึง 83.01% และสูงกว่าอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.1 และสามารถให้ค่า F-Measure ที่สามารถจำแนกได้ครบทุกประเภท

Rodriguez [20] ได้เสนอ การเลือกแอตทริบิวต์ในชุดข้อมูลวิศวกรรมซอฟต์แวร์ สำหรับการตรวจสอบความผิดพลาดของโมดูล โดยเปรียบเทียบกับอัลกอริทึมในการประเมินค่าแบบ Filter approach และ Wrapper approach โดยใช้อัลกอริทึมในการจำแนกประเภทข้อมูล 2 อัลกอริทึมคือ IB1 : an instance-base classifier และ C4.5 ผลการทดลองแสดงให้เห็นว่า เมื่อมีการนำการคัดเลือกคุณลักษณะมาใช้จะทำให้จำนวนของคุณลักษณะลดลงและค่าความถูกต้องในการจำแนกเพิ่มมากขึ้นและวิธีการที่ให้ค่าความถูกต้องสูง คืออัลกอริทึมในการประเมินค่าโดยใช้ Wrapper approach ร่วมกับอัลกอริทึม C4.5 จะให้ค่าความถูกต้องที่ดีกว่าวิธีการอื่นๆ

2.1.5 เทคนิคในเหมืองข้อมูล

2.1.5.1 เทคนิค C4.5 [15] เป็นเทคนิคในการสร้างต้นไม้การตัดสินใจพัฒนาโดย J. Ross Quinlan ในปี 1993 โดยนำเอา ID3 มาปรับปรุงให้มีความสามารถมากขึ้นใช้วิธีการ Information Gain เพิ่มเติมการจัดการกับข้อมูล, ตัวเลข, ข้อมูลที่ขาดไปและไม่สมบูรณ์ และการ Prune ด้วยการแทนกิ่ง (Branch) ที่ไม่ช่วยในการตัดสินใจด้วย Leaf Node ที่ตัดสินใจได้ดีกว่า การแบ่งของ tree ในการทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 คือต้องหา Info และ Gain ออกมาก่อน ตัวอย่างเช่นชุดข้อมูลการซื้อเครื่องปรับอากาศ



ตารางที่ 2.1 ตัวอย่างชุดข้อมูลการซื้อเครื่องปรับอากาศ

Age	Income	Working	Credit	Product (Air condition)
<30	High	No	Fair	No
<30	High	No	Excellent	No
31-41	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	Yes
31-41	Low	Yes	Excellent	Yes
<30	Medium	No	Fair	No
<30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<30	Medium	Yes	Excellent	Yes
31-41	Medium	No	Excellent	Yes
31-41	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

โดยมีคลาสในการซื้อเครื่องปรับอากาศ กำหนดไว้ 2 ค่า คือ Yes และ No สิ่งนำมาพิจารณาคือ อายุรายได้ สถานการณ์ทำงาน และเครดิต ดังรายละเอียดต่อไปนี้

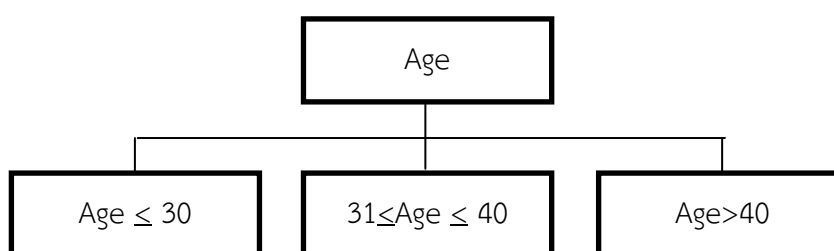
1. อายุ แบ่งออกได้ 3 ค่า คือ อายุต่ำกว่า 30 ปี, อายุระหว่าง 31-41 ปี, และอายุมากกว่า 40 ปี

2. รายได้แบ่งออกได้เป็น 3 ค่า คือ High, Medium และ Low

3. การทำงาน แบ่งออกเป็น 2 ค่า คือ Yes และ No

4. เครดิต แบ่งออกเป็น 2 ค่า คือ Fair และ Excellent

ตัวอย่างชุดข้อมูลการซื้อเครื่องปรับอากาศ ถ้าพิจารณาที่ Age จะสามารถแบ่งกลุ่มข้อมูลได้ดังภาพที่ 2.1



ภาพที่ 2.1 ตัวอย่างการแบ่งกลุ่มข้อมูลที่ แอตทริบิวต์ Age



เมื่อนำมาคำนวณหาค่า Information ของแต่ละกลุ่มจะได้ดังนี้

$$\begin{aligned}\text{Age} : \text{Info}([9,5]) &= -\left(\frac{9}{14} \log_2 \frac{9}{14}\right) - \left(\frac{5}{14} \log_2 \frac{5}{14}\right) \\ &= 0.971\end{aligned}$$

$$\begin{aligned}\text{Age} \leq 30 : \text{Info}([2,3]) &= -\left(\frac{2}{5} \log_2 \frac{2}{5}\right) - \left(\frac{3}{5} \log_2 \frac{3}{5}\right) \\ &= 0.940\end{aligned}$$

$$\begin{aligned}31 \leq \text{Age} \leq 40 : \text{Info}([4,0]) &= -\left(\frac{4}{4} \log_2 \frac{4}{4}\right) - \left(\frac{0}{4} \log_2 \frac{0}{4}\right) \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Age} > 40 : \text{Info}([3,2]) &= -\left(\frac{3}{5} \log_2 \frac{3}{5}\right) - \left(\frac{2}{5} \log_2 \frac{2}{5}\right) \\ &= 0.940\end{aligned}$$

ดังนั้นค่าของ information เมื่อพิจารณาที่ แอตทริบิวต์ Age คือ

$$\begin{aligned}\text{Age} : \text{Info}([3,2], [4,0], [2,3]) &= \left(\frac{5}{14} \times 0.971\right) + \left(\frac{4}{14} \times 0\right) + \left(\frac{5}{14} \times 0.971\right) \\ &= 0.693\end{aligned}$$

และค่าของ Gain เมื่อพิจารณาที่ แอตทริบิวต์ Age คือ

$$\begin{aligned}\text{Gain} &= \text{Info}([9,5]) - \text{Info}([3,2], [4,0], [2,3]) \\ &= 0.940 - 0.693 \\ &= 0.247\end{aligned}$$

สิ่งที่ C4.5 ต้องทำเพิ่มเติมคือ การหาค่า Gain ratio ซึ่งจะใช้ตัวนี้เป็นตัวแบ่งซึ่งหาได้

จากสูตร

$$\text{Gain ratio} = \frac{\text{Gain}}{\text{SplitInfo}}$$

ซึ่ง Split info ใช้หลักการเดิมในการหา Information ออกมาเป็น 3 กลุ่ม โดยการนับจำนวนรายการของข้อมูล ไม่ได้นับจำนวน ค่า Yes หรือ No จากข้อมูลดังกล่าวจะทราบว่ากลุ่มอายุน้อยกว่าหรือเท่ากับ 30 มีจำนวน 5 รายการ กลุ่มอายุน้อยกว่าหรือเท่ากับ 30 – 40 มีจำนวน 4 รายการ กลุ่ม



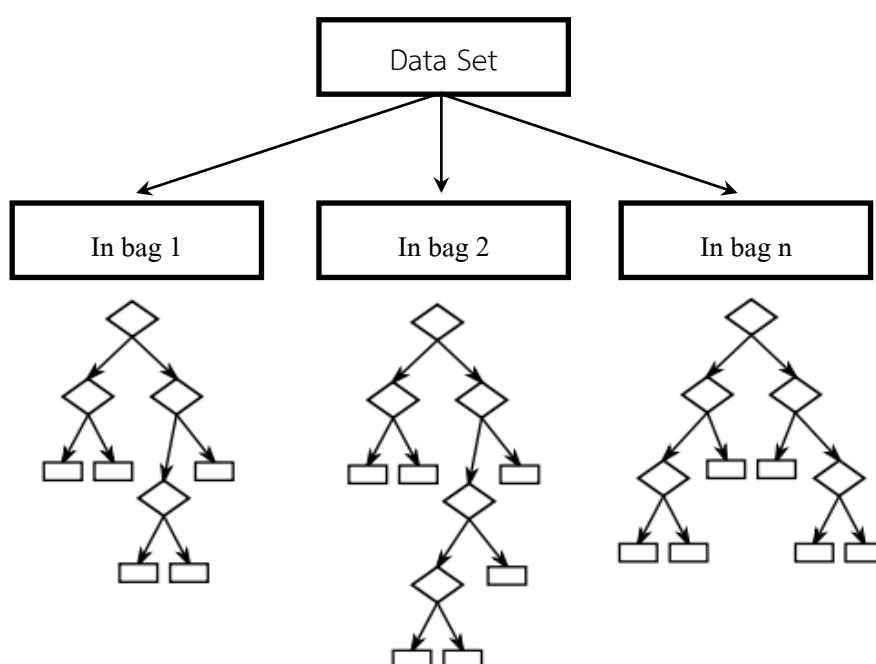
อายุน้อยกว่าหรือเท่ากับ 40 มีจำนวน 5 รายการ จะได้ค่า Split Info คือ

$$\begin{aligned}\text{Split Info}([5,4,5]) &= -\left(\frac{5}{14}\log_2\frac{5}{14}\right) - \left(\frac{4}{14}\log_2\frac{4}{14}\right) - \left(\frac{5}{14}\log_2\frac{5}{14}\right) \\ &= 1.577\end{aligned}$$

$$\begin{aligned}\text{Gain ratio} &= \frac{\text{Gain}}{\text{SplitInfo}} \\ &= \frac{0.247}{1.577} \\ &= 0.156\end{aligned}$$

จะได้ค่า Gain ratio และทำการคำนวณในรายการอื่น ๆ จนครบ ซึ่งจะสามารถแสดงได้ว่าจะแบ่งการทำงานอยู่ค่าของรายการใด

2.1.5.2 เทคนิค Random Forest [21] เป็นเทคนิคการสุ่มเลือกใช้ข้อมูลและคุณลักษณะ Decision Tree ซึ่งถูกสร้างจากการนำข้อมูลไปสุ่มเลือกตัวอย่างแบบเลือกแล้วใส่กลับ (Sampling with Replacement) แล้วนำมาสร้างเป็น Tree ซึ่งจะมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก ซึ่งข้อมูลส่วนนี้เรียกว่า Out-of-Bag (OOB) จะถูกนำมาใช้ในการทดสอบ Decision Tree วิธีการดังกล่าวนี้เรียกว่า Bagging ผลลัพธ์ที่ได้อย่างอิสระจาก Decision Tree ในแต่ละต้นถูกนำมาคิดเป็นผลการโหวต ผลโหวตที่มากที่สุดจะใช้ระบุสถานะของคลาสดังที่แสดงในภาพที่ 2.2 เทคนิค Random Forest ไม่จำเป็นต้องมีข้อมูลทดสอบ เพื่อประมาณความผิดพลาดเพราะข้อมูล OOB นั้นถูก นำมาใช้ทดสอบ Decision Tree นั้นแล้ว



ภาพที่ 2.2 ลักษณะการทำงานของเทคนิค Random Forest



ความสำคัญของตัวแปร (Variable Importance) คือ ดัชนีที่นำไปใช้ในการระบุตัวแปรที่เป็นตัวแปรต้น ซึ่งคำนวณโดยนำข้อมูล OOB ทดสอบกับ Decision Tree จากนั้นสุ่มค่าข้อมูลในข้อมูล OOB ทั้งชุด แล้วนำข้อมูลดังกล่าวทดสอบกับ Decision Tree นับจำนวนครั้งที่โหนดถูกนำจำนวนครั้งที่โหนดถูกก่อนทำการสุ่มค่า ลบจำนวนครั้งที่โหนดถูกหลังการสุ่มค่า แล้วหาค่าเฉลี่ยจะได้คะแนนดิบของความสำคัญของตัวแปร ดังสมการที่ 4

$$I_T(A) = \frac{1}{T} \sum_{j=1}^T \frac{1}{N_j} \sum_{i=1}^N [1(V_j(X_i) = y_i) - 1(V_j(X_i^{(A,j)}) = y_i)] t_{ij} \quad (4)$$

โดยที่ y_i	คือ สถานะจริงของข้อมูล
X_i	คือ ข้อมูลตัวอย่าง
V_j	คือ ผลโหนดของต้นไม้ตัดสินใจแต่ละต้น
N_j	คือ จำนวนข้อมูลใน OOB
T	คือ จำนวนต้นไม้ตัดสินใจ
$X_i^{(A,j)}$	คือ ข้อมูลที่ถูกสุ่มขึ้นมาจาก OOB ของต้นไม้ที่ j
t_{ij}	คือ ค่าแสดงสถานะของข้อมูลว่าอยู่ใน OOB หรือไม่ (เป็น 1 อยู่ใน OOB, เป็น 0 ไม่อยู่ใน OOB)

มีนักวิจัยจำนวนมากได้นำเทคนิค Random Forest มาใช้ในการจำแนก เช่น Krishnaveni [22] ได้พัฒนาแบบจำลองสำหรับทำนายการบาดเจ็บที่เกิดจากอุบัติเหตุบนท้องถนน โดยใช้เทคนิควิธี Naive, PART Rule, J48 Decision tree และ Random forest พบว่าการสร้างแบบจำลองด้วยเทคนิควิธี Random forest ให้ความถูกต้องคิดเป็นร้อยละ 74.34 ซึ่งดีกว่าวิธีอื่นๆ

ภรณ์ยา ปาลวิสุทธิ [23] ได้เสนอการเพิ่มประสิทธิภาพเทคนิค Decision tree บนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดเชื้อในเทคนิค Decision Tree J48, ID3, LMT, CART และ Random Forest ผลการทดลองประสิทธิภาพในการพยากรณ์ของตัวแบบพบว่าเทคนิค Random Forest มีค่าความแม่นยำร้อยละ 87.15 สามารถพยากรณ์ได้ดีกว่า J48 ID3 LMT และ CART

2.1.5.3 เทคนิค Random Tree [24] คือ เทคนิคที่ใช้ในการจำแนกหมวดหมู่เช่นเดียวกับ C4.5 โดยมีหลักการสร้าง Tree จากการสุ่ม Tree หลายๆแบบ ในแต่ละโหนดแล้วเลือกมาประมวลผลโดยไม่ใช้การ Prune และเนื่องจากจำนวนของ Tree เพิ่มขึ้นอย่างรวดเร็ว ซึ่งยากแก่การแก้ปัญหาการสร้าง Tree ที่เป็นไปได้ทั้งหมด ส่วนประกอบสามารถสร้างชุดสุ่มของ Tree ออกมาจากการกระจายชุดต่างๆของ Tree Random Tree เป็นการสุ่มวาดที่สุ่มจากชุดของ Tree ที่เป็นไปได้ ในบริบทนี้ "สุ่ม" หมายความว่า Tree ในชุดของ Tree มีโอกาสเท่าเทียมกันของการเป็นตัวอย่าง วิธีที่บอกนี้ก็คือว่าการกระจายของ Tree คือ "ชุด" Tree สุ่มแบบต่อเนื่อง (CRT) เป็นแบบสุ่ม Tree จริง T_e และมีระยะในการเดินทางสั้นที่สุด Tree ชุดย่อย ประกอบด้วย Tree ที่มี n จุด แต่ละองค์ประกอบของโครงสร้างที่แท้จริงคือรากมีนักวิจัยจำนวนมากได้นำเทคนิค Random Tree มาใช้ในการจำแนกเช่น ฐิติมา ช่างชัย [25] ได้เสนอการวิเคราะห์หารูปแบบการเรียนรู้โดยใช้เหมือนข้อมูลของนักศึกษาต่อการจัดทำปริญญานิพนธ์ โดยใช้วิธีวิเคราะห์ผลด้วยรูปแบบ Rule Based Classification ด้วยวิธี Decision Table, Jrip



และ PART และรูปแบบ Decision Tree Classification ด้วยวิธี LMT, J48 และ Random Tree จากการวิเคราะห์ผลทั้งหมดพบว่า รูปแบบของ Decision Tree ด้วยวิธีการ Random Tree ให้ค่าความถูกต้องสูงสุด (100%) ส่วนรูปแบบของ Rule Based ด้วยวิธี PART ให้ค่าความถูกต้องสูงสุด (84.12%)

Sushikumar Kalmegh [26] นำเสนอการเปรียบเทียบการวิเคราะห์อัลกอริทึมในโปรแกรม Weka ได้แก่ REPTree, Simple Cart และ Random Tree ในการจำแนกข้าวอินเดียวพบว่า Random Tree มีค่าความถูกต้อง 100% ซึ่งมากกว่า REPTree และ Simple Cart

2.1.5.4 เทคนิค REP Tree [24] คือเทคนิคที่ใช้ regression tree logic และสร้าง Tree หลายๆ ต้นที่แตกต่างกัน หลังจากนั้นก็เลือกที่ดีที่สุดจาก Tree ที่สร้างทั้งหมดมาเป็นตัวแทนของ Tree ทั้งหมด ในการตัดกิ่งใช้ค่า mean square error ในการพยากรณ์เป็นพื้นฐานการวัด Reduced Error Pruning Tree (REP Tree) เป็น Decision Tree ที่มีการเรียนรู้และสร้างแบบจำลองอย่างรวดเร็วบนพื้นฐานของ Information gain หรือ reducing the variance และตัดกิ่งโดยใช้การลดข้อผิดพลาดในการตัดแต่ใช้ได้เฉพาะตัวแปรที่เป็นตัวเลขเท่านั้น มีนักวิจัยจำนวนมากได้นำเทคนิค REP Tree มาใช้ในการจำแนก เช่น Prerna Kapoor และ Reena Rani [27] ได้เสนอประสิทธิภาพการตัดสินใจเทคนิค Decision Tree โดยใช้อัลกอริทึม J48 และ Reduced Error Pruning ผลการทดลองพบว่าทำให้การพยากรณ์มีความแม่นยำมากขึ้น

Kittipol Wisaeng [28] ได้เสนอการเปรียบเทียบอัลกอริทึมของ Decision Tree ในการจำแนก UCI Repository ได้แก่ Nursery, Iris, Anneal, Shuttle_trn, Voting, Waveform และ Sick โดยมีอัลกอริทึมคือ อัลกอริทึม functional tree, อัลกอริทึม logistic model trees, อัลกอริทึม REP Tree และอัลกอริทึม best-first decision tree ผลการทดลองพบว่า REP Tree ให้ค่าความถูกต้องที่ 92.87%

2.1.5.5 เทคนิค K-Nearest Neighbour (k-NN) [29] เป็นขั้นตอนวิธีการในการหาสมาชิกที่ใกล้เคียงที่สุดเป็นอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูล โดยการจัดข้อมูลที่อยู่ใกล้กันให้เป็นกลุ่มเดียวกัน ซึ่งเทคนิคนี้จะทำให้ตัดสินใจได้ว่า คลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง โดยการตรวจสอบจำนวน k ซึ่งถ้าหากเงื่อนไขของการตัดสินใจมีความซับซ้อน วิธีนี้จะสามารถสร้างโมเดลที่มีประสิทธิภาพได้ แต่ขั้นตอนวิธีการหาสมาชิกที่ใกล้เคียงที่สุดจะใช้ระยะเวลาในการคำนวณนาน ถ้าตัวแปร(แอตทริบิวต์) มีจำนวนมากจะเกิดปัญหาในการคำนวณค่า และค่อนข้างที่จะใช้ปริมาณงานในการคำนวณสูงมากบนคอมพิวเตอร์ เพราะเวลาที่ใช้สำหรับการคำนวณจะเพิ่มขึ้นแบบแฟกทอเรียลตามจำนวนจุดทั้งหมด ดังนั้นเพื่อจะเพิ่มความรวดเร็วสำหรับเทคนิคขั้นตอนวิธีการหาสมาชิกที่ใกล้เคียงที่สุดให้มากขึ้น ข้อมูลทั้งหมดที่ใช้บ่อยจะต้องถูกเก็บไว้ในหน่วยความจำ โดยวิธีการเข้าถึงหน่วยความจำพื้นฐานอย่างมีประสิทธิภาพ (Memory-Based Reasoning) ซึ่งจะเป็นวิธีที่นำมาอ้างอิงถึงเป็นประจำในการจัดเก็บกลุ่มคลาสของขั้นตอนวิธีการหาสมาชิกที่ใกล้เคียงที่สุดในหน่วยความจำ และถ้าหากข้อมูลที่ต้องการหาคำตอบมีตัวแปรอิสระเพียงไม่กี่ตัวแล้ว จะทำให้เราสามารถเข้าใจโมเดลขั้นตอนวิธีการหาสมาชิกที่ใกล้เคียงที่สุดได้ง่ายขึ้น ตัวแปรเหล่านี้ยังมีประโยชน์สำหรับนำมาสร้างโมเดลต่าง ๆ ที่เกี่ยวข้องกับชนิดของข้อมูลที่ไม่เป็นมาตรฐาน เช่น ข้อความ เพียงแต่อาจต้องมีมาตรฐานการวัดค่าสำหรับชนิดของข้อมูลดังกล่าวที่เหมาะสมด้วย นอกจากนี้ประสิทธิภาพของขั้นตอนวิธีการหาสมาชิกที่ใกล้เคียงที่สุดนี้ จะขึ้นอยู่กับจำนวนระยะห่าง การอธิบายระหว่างข้อมูลทั้งคู่ที่สามารถแบ่งแยกอย่างมีประสิทธิภาพระหว่างข้อมูลปกติ และ



ข้อมูลผิดปกติ การอธิบายจำนวนระยะห่างระหว่างข้อมูลเป็นความท้าทายอย่างมากเมื่อข้อมูลมีความซับซ้อน อย่างเช่น ข้อมูลกราฟ และข้อมูลแบบลำดับ เป็นต้น

k-NN [30] หลักการของวิธีนี้ จะคำนวณหาสมาชิกที่ใกล้ที่สุด k ตัว รวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัว และเลือกคลาสที่สมาชิกส่วนใหญ่ที่อยู่ในกลุ่ม k ตัว ที่มีอยู่มากให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว จะประกอบไปด้วยแอตทริบิวต์หลายตัวแปร X_i ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ k ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณีที่จะต้องค้นหาในการทำนายกรณีใหม่ เทคนิคแบบ k-NN ได้แก่ 1-NN , 2-NN , 3-NN , ... , k-NN ตัวอย่างเช่น 3-NN หมายถึง เทคนิคนี้จะค้นหา 3 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (3 Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝนมาเรียงลำดับจากน้อยไปหามาก และเลือกสมาชิกที่มีระยะทางใกล้เคียงที่สุด k ตัว โดยใช้การวัดระยะทาง เช่น แบบ Euclidian Distance ซึ่งมีหลักการ คือ การวัดระยะทางระหว่าง 2 จุด ถ้าระยะห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย ถ้าระยะห่างกันน้อยแสดงว่ามีความคล้ายคลึงกันมาก ดังแสดงในสมการที่ 5 และสามารถย่อในรูปแบบของสมการที่ 6 ได้ดังนี้

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \quad (5)$$

$$d(X,Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (6)$$

เมื่อ	X	=	เรคคอร์ดของข้อมูลที่ใช้ระบุ
	Y	=	เรคคอร์ดของข้อมูลทั้งหมดในฐานข้อมูล
	x_i	=	ค่าข้อมูลในแอตทริบิวต์ i ของเรคคอร์ด X
	y_i	=	ค่าข้อมูลในแอตทริบิวต์ i ของเรคคอร์ด Y
	n	=	จำนวนข้อมูลทั้งหมดที่มีอยู่ในระบบ

2.1.5.6 เทคนิค Support Vector Machines (SVM) [9] คือ ขั้นตอนวิธีการที่มีความรวดเร็วและเป็นเทคนิคที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการของการหาสมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกต้องเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด

แนวความคิดของเทคนิควิธี SVM นั้นเกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางลงในพีเจอรส์เปซ จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกัน โดยจะทำการสร้างเส้นแบ่งที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่าเส้นตรงที่แบ่งกลุ่มสองกลุ่มออกจากกันนั้น เส้นใดเป็นเส้นที่ดีที่สุด สำหรับ SVM นั้นเดิมได้มีการนำมาใช้กับข้อมูลที่เป็นเชิงเส้น แต่ในความเป็นจริงแล้วข้อมูลที่นำมาใช้ในระบบการสอนให้ระบบเรียนรู้ส่วนใหญ่มักเป็นข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยการนำ Kernel Function มาใช้ การจำแนกข้อมูลบนระนาบหลายมิติ จะใช้ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า โครงสร้างในการคัดเลือกซึ่งโครงสร้างในการคัดเลือกมาจากข้อมูลที่สอนให้ระบบเรียนรู้ จำนวนเซตของโครงสร้างที่ใช้อธิบายในกรณีหนึ่ง เรียกว่า เวกเตอร์ ดังนั้นจุดมุ่งหมายของตัว



แบบ SVM คือ แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรของเป้าหมายที่อยู่ข้างหนึ่งของระนาบและกรณีของกลุ่มอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า ซัพพอร์ตเวกเตอร์ ซึ่งวิธีการนี้เหมาะสำหรับข้อมูลที่มีมิติของข้อมูลสูง

จากวิธีการของ SVM สามารถนำมาเขียนเป็นสมการเพื่อใช้ในการแก้ปัญหา ซึ่งข้อมูลที่ใช้ในการวิเคราะห์จะถูกจัดลงบนพื้นที่ของคุณลักษณะที่เหมาะสมกันเป็นกลุ่มของข้อมูลที่อยู่ในรูปของเวกเตอร์ ดังสมการที่ 7

$$X = ((x_1, y_1), \dots, (x_i, y_i)) \quad (7)$$

เมื่อ X คือ ชุดค่าลักษณะเด่น

จากนั้นทำการกำหนดสมการเส้นตรงเพื่อนำมาสร้างเป็นเส้นตรงบนเส้นแบ่ง ซึ่งจะเป็นการแบ่งกลุ่มข้อมูลที่มีลักษณะเป็นเชิงเส้นสองกลุ่มออกจากกัน ซึ่งจะมีการกำหนดกลุ่มของข้อมูลทั้งสองฝั่งเป็นเพียงสองค่าที่แทนด้วยค่า y ซึ่งข้อมูลที่เป็นตัวกำหนดความชันและระนาบที่เกิดขึ้นบนเส้นแบ่งเกิดจากคู่ของข้อมูล (w, b) และกำหนดสมการที่เป็นตัวบ่งชี้ข้อมูลในแต่ละกลุ่มว่าข้อมูลนั้นจะอยู่ส่วนใดของเส้นแบ่ง ดังแสดงในสมการที่ 8 และสมการที่ 9 และเมื่อนำเงื่อนไขทั้งหมดมาวิเคราะห์เชิงเรขาคณิตโดยทำการพิจารณาในกรณีที่ข้อมูลได้ถูกแบ่งกลุ่มตามเงื่อนไขของ SVM อย่างสมบูรณ์ ดังในสมการที่ 8

$$W^T x + b \geq y \quad \text{เมื่อกำหนดให้ } y = 1 \quad (8)$$

$$W^T x + b \geq y \quad \text{เมื่อกำหนดให้ } y = -1 \quad (9)$$

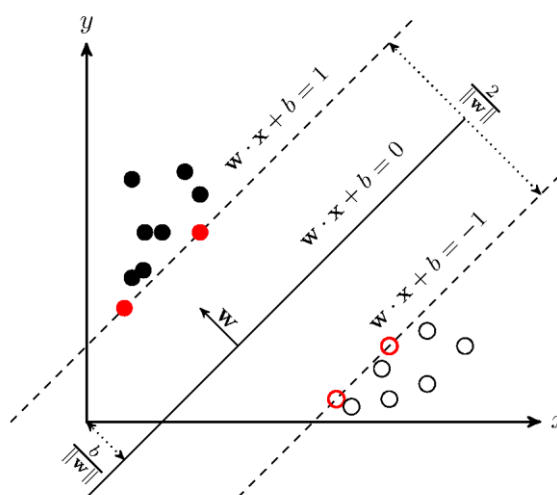
$$y(W^T x + b) - 1 \geq 0 \quad (10)$$

เมื่อ y คือ ค่ากลุ่มของข้อมูล (1,-1)
 w คือ ความชัน
 x คือ ค่าลักษณะเด่น
 b คือ ค่าคงที่ (ค่าตัดแกน y)

และข้อมูลการสอนให้ระบบเรียนรู้จะต้องอยู่ในรูปแบบเชิงเส้น ดังแสดงในภาพที่ 2-3 ซึ่งเวกเตอร์ของข้อมูลที่ถูกป้อนเข้าสู่ระบบการสอน เพื่อให้ระบบเรียนรู้แทนด้วยสมการและข้อมูลทั้งสองด้านแบ่งเป็นบวกและลบ สถานะของข้อมูลจึงแทนด้วย y ซึ่งมีสองค่า คือ $y = 1$ และ $y = -1$ แต่เส้นแบ่งที่ได้นั้นจะยังคงไม่ใช่เส้นแบ่งที่ดีที่สุด วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดนั้นคือ การเพิ่มเส้นขอบให้กับเส้นแบ่งทั้งสองข้าง จะทำให้ได้เส้นแบ่งใหม่ที่น่าเชื่อถือและเป็นเส้นขอบของข้อมูลแต่ละฝั่งอีกด้วย เส้นขอบ (Margin) ที่เป็นเส้นขอบของเส้นแบ่งนั้น จะเป็นเส้นที่สัมผัสกับค่าข้อมูลในฟีเจอร์สเปซ (Feature Space) ที่ใกล้ที่สุด เส้นขอบของทั้งสองเส้นที่เพิ่มขึ้นมานี้จะถูกแทนด้วยสมการ $W^T x + b \geq y \geq 1$ ถ้าอยู่ด้าน $y = 1$ และ $W^T x + b \geq y \geq -1$ เมื่อ $y = -1$ และหากเส้นแบ่งขอบใดๆ มีค่าความกว้างมากที่สุด แสดงให้เห็นว่าข้อมูลสองชุดมีการแยกจากกันอย่างชัดเจนมากที่สุด ดังนั้นเส้น



บางที่มีเส้นขอบกว้างที่สุดจึงถือได้ว่าเป็นเส้นแบ่งที่ดีที่สุด และเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบว่า เวกเตอร์สนับสนุน (Support Vector) โดยสามารถเขียนสมการการคำนวณความกว้างของเส้นขอบจำเป็นต้องทำการคำนวณพจน์ให้อยู่ในรูปปกติ (Normalization) ซึ่งคำนวณได้จากการแทนค่าลงไปในสมการที่ 8 และ 9



ที่มา: <http://blog.pengyifan.com/wp-content/uploads/2013/09/svm.png>

ภาพที่ 2.3 การวางตัวของข้อมูลในลักษณะเชิงเส้น

ซึ่งสามารถหาสมการความกว้างของเส้นขอบได้จากสมการที่ 11 หลังจากนั้นจึงทำการแก้สมการต่อด้วย Lagrangian dual problem ในสมการที่ 12

$$\begin{aligned}
 W^T x^+ + b &= 1 \\
 W^T x^- + b &= -1 \\
 W^T (x^+ - x^-) &= 2 \\
 M &= \left(\frac{w}{\|w\|} \right)^T (x^+ - x^-) \\
 M &= \frac{2}{\|w\|}
 \end{aligned} \tag{11}$$

เมื่อ M คือ ความกว้างของเส้นขอบ

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i ((w_i)^2 + b) - 1] \tag{12}$$

เมื่อ $\alpha_i \geq 0$; $i = 1, 2, 3, \dots, N$



จากสมการที่ 12 เมื่อนำมาอนุพันธ์ เพื่อทำการหาค่าน้ำหนักที่ดีที่สุด (w) ดังแสดงในสมการที่ 13 และ สมการที่ 14

$$\frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^N y_i \alpha_i x_i = 0 \quad (13)$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^N y_i \alpha_i = 0 \quad (14)$$

เมื่อได้สมการที่ผ่านการแก้ปัญหามาจากวิธีการ Dual Problem จะได้ค่า w ที่มีการลดรูปลงแล้ว และสามารถหาได้จากสมการที่ 15

$$w = \sum_{i=1}^N \alpha_i y_i x_i \quad (15)$$

เมื่อ α คือ สัมประสิทธิ์คงที่

ค่า b จะไม่ปรากฏในการแก้ไขปัญหามาแบบ Dual Problem ซึ่งหาได้จากสมการที่ 16

$$b = -\frac{\max_{y_i=-1} (w^T x_i) + \min_{y_i=1} (w^T x_i)}{2} \quad (16)$$

จากนั้นนำค่า w ที่แก้ปัญหามาแล้วแทนในสมการที่ 15 ซึ่งเป็นสมการที่ใช้ในการหาค่าสัมประสิทธิ์ α เพื่อนำมาใช้บอกการวางตัวของเส้นแบ่งข้อมูล ซึ่งจะได้

$$y_i (\sum_{i=1}^N \alpha_i y_i (x_i^T x_i) + b) - 1 \geq 0 \quad (17)$$

ในการหา SVM จะใช้ในกรณีที่ชุดข้อมูลไม่เกิดข้อผิดพลาดใดๆ จากค่าของข้อมูลที่วางตัวในแกนของ SVM ซึ่งเมื่อเกิดปัญหาจะยอมให้เกิดข้อผิดพลาดนี้ได้ในระยะหนึ่งที่สามารถยอมรับได้ ซึ่งเรียกค่านั้นว่า เวกเตอร์อนุโลม (Slack Vector)

เมื่อนำค่าเวกเตอร์อนุโลมเข้าในสมการที่ 18 ซึ่งเป็นกระบวนการแก้ปัญหามาแบบ Dual Problem จึงทำให้เกิดค่าคงที่ค่าหนึ่งซึ่งแทนด้วยตัวแปร C ดังแสดงในสมการที่ 18

$$L(w, b, \zeta, \alpha) = \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^N \zeta_i^2 + \sum_{i=1}^N \alpha_i [y_i ((w_i \cdot w_i) + b) - 1 + \zeta_i] \quad (18)$$

จากนั้นเมื่อนำเอาสมการดังกล่าวมาหาค่าอนุพันธ์ เช่นเดียวกับการแก้สมการในสมการที่ 12 ซึ่งหลังจากที่ได้สมการที่ผ่านการแก้ปัญหามาด้วย Dual Problem แบบมีเวกเตอร์อนุโลมแล้วจะได้ค่า w คงเดิมแต่จะได้ขอบเขตของ α ขึ้นมา ดังแสดงในสมการที่ 19



$$w = \sum_{i=1}^N \alpha_i y_i x_i \quad (19)$$

เมื่อ $C \geq \alpha \geq 0$

ค่า C จึงถือได้ว่าเป็นค่าพารามิเตอร์อีกค่าหนึ่งที่จำเป็นต้องกำหนดขึ้นมาเพื่อใช้ในการหาค่าสัมประสิทธิ์ α ซึ่งใช้สำหรับเพิ่มประสิทธิภาพของการแบ่งกลุ่มข้อมูล ซึ่งมีผลทำให้ระยะทางระหว่างเวกเตอร์มีความกว้างยิ่งขึ้นและยังทำหน้าที่ช่วยลดข้อผิดพลาดที่เกิดขึ้นด้วย

2.1.6 การวัดประสิทธิภาพแบบจำลอง

2.1.6.1 K-fold cross validation [9] คือ การเลือกสุ่มข้อมูลแบบความเที่ยงตรงโดยเป็นวิธีที่นิยมในการทำงานวิจัยเพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลอง โดย K-fold cross validation จะสุ่มเลือกข้อมูลออกเป็น K ชุดเท่ากัน ในการทดลองครั้งแรกข้อมูลชุดที่ 1 เป็นชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ในการทดลองครั้งที่สอง ข้อมูลชุดที่ 2 เป็นชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบซึ่งจะต้องมีการทดลองทั้งหมด K ครั้ง ตัวอย่างการเลือกสุ่มข้อมูล ความเที่ยงตรง K กลุ่ม เมื่อ K=5 แสดงดังตารางที่ 2.2

ตารางที่ 2.2 แสดงตัวอย่างการทดสอบประสิทธิภาพ เมื่อ K=5

การทดลอง	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 4	ครั้งที่ 5
ข้อมูลชุดสอน	ชุดที่ 5	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4
	ชุดที่ 4	ชุดที่ 5	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3
	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5	ชุดที่ 1	ชุดที่ 2
	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5	ชุดที่ 1
ข้อมูลชุดทดสอบ	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5

จากนั้นนำค่าความถูกต้อง (Accuracy) ที่ได้จากการทดสอบประสิทธิภาพในแต่ละรอบมาหาค่าเฉลี่ยเพื่อเป็นค่าความถูกต้องโดยรวมของการจำแนกข้อมูลของแบบจำลอง ซึ่งสามารถคำนวณค่าความถูกต้องได้ดังสมการที่ 20

$$\text{Overall accuracy} = \text{Average}(\sum_{i=1}^K \text{Accuracy}_i) \quad (20)$$

เมื่อ K คือจำนวนชุดข้อมูลที่ทดสอบทั้งหมด



2.1.6.2 การวิเคราะห์ประสิทธิภาพ [9] คือการวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล โดยค่าของผลลัพธ์ที่ได้จากการจำแนกคือ ค่า True Positive (TP) ค่า True Negative (TN) ค่า False Positive (FP) ค่า False Negative (FN) และสามารถหาค่าความถูกต้อง (Accuracy) จากสมการที่ 21 ค่าความแม่นยำ (Precision) จากสมการที่ 22 และค่าความระลึก (Recall) ได้จากสมการที่ 23

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (21)$$

$$Precision = \frac{TP}{TP + FP} \quad (22)$$

$$Recall = \frac{TP}{TP + FN} \quad (23)$$

2.2 งานวิจัยที่เกี่ยวข้อง

ธีรพงษ์ สังข์ศรี [31] ได้เสนอการวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียน และการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งประกอบด้วยเทคนิค C4.5 และเทคนิค Artificial Neural Network โดยทำการวิเคราะห์ข้อมูลของผู้สมัครเข้าศึกษาต่อมหาวิทยาลัยราชภัฏนครราชสีมา จำนวน 70,000 ข้อมูล ผลการทดลองพบว่าเทคนิค C4.5 ให้ค่าความถูกต้อง 93.76% ซึ่งมากกว่าเทคนิค Artificial Neural Network ให้ค่าความถูกต้อง 93.60%

ลาภ พุ่มหิรัญ และมาลีรัตน์ โสตานิล [32] ได้นำเทคนิค Decision Tree มาใช้ในการสร้างตัวแบบการพยากรณ์สำหรับทำนายจำนวนนักศึกษาใหม่ โดยเทคนิคสำหรับสร้างแบบจำลอง Decision Tree ที่เลือกใช้คือ C4.5 แล้วนำไปทดสอบด้วยชุดข้อมูลทดสอบด้วยวิธีการสุ่มข้อมูล 3 วิธี คือ 1. K-fold cross-validation 2. Percentage split 3. Training set and test set ผลการทดลองพบว่าเมื่อใช้ C4.5 กับ Training set and test set ได้ค่าความถูกต้อง 94% ซึ่งมากกว่าเมื่อใช้ C4.5 กับ fold cross-validation ที่ได้ค่าความถูกต้อง 93.97% และเมื่อใช้ C4.5 กับ Percentage split ที่ได้ค่าความถูกต้อง 93.73%

พรณิภา บุตรเอก และสุรเดช บุญลือ [33] ได้เสนอการพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาโดยใช้ SVM ซึ่งการพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษา จำนวน 138 ระเบียบ 18 คุณลักษณะ ด้วยเทคนิค SVM เทคนิค C4.5 และเทคนิค BP-ANN ผลการทดสอบพบว่าเทคนิค SVM ให้ความแม่นยำในการพยากรณ์ร้อยละ 89.13 ซึ่งสูงกว่าเทคนิค BP-ANN ที่ให้ความแม่นยำในการพยากรณ์ร้อยละ 84.36 และเทคนิค C4.5 ที่ให้ความแม่นยำในการพยากรณ์ร้อยละ 83.33



เยาวภา การสำเร็จ และคณะ [34] ได้เสนอ การเปรียบเทียบอัลกอริทึมเหมือนข้อมูล เพื่อวิเคราะห์ปัจจัยที่ส่งผลกระทบต่อระดับผลการเรียนของนักศึกษา โดยการค้นหาปัจจัยที่ส่งผลกระทบต่อระดับผลการเรียนของนักศึกษา จำนวน 4,591 ชุดข้อมูล 17 แอตทริบิวต์ ด้วยเทคนิค 3 เทคนิคคือ C4.5, Naive Bayes และ k-NN ผลการเปรียบเทียบพบว่าเทคนิค C4.5 ให้ค่าประสิทธิภาพสูงที่สุดร้อยละ 73.55 มากกว่าเทคนิค k-NN ที่ให้ค่าประสิทธิภาพร้อยละ 66.63 และเทคนิค Naive Bayes ที่ให้ค่าประสิทธิภาพร้อยละ 49

วันเพ็ญ วิโรจน์เจริญวงศ์ และบุศรินทร์ เอี่ยมมณกุล [35] ได้เสนอการเปรียบเทียบการ จำแนกข้อมูลของแบบจำลองเทคนิค Decision Tree, Rule Base และ k-NN กับชุดข้อมูลจำนวน 3 ชุดจาก UCI Machine Learning Repository คือ Congressional Voting Records, Ecoli และ Car Evaluation ผลการทดลองพบว่า 1) ชุดข้อมูล Congressional Voting Records เทคนิค k-NN ได้ค่า ความถูกต้องร้อยละ 100 มากกว่าเทคนิค Rule Base ได้ค่าความถูกต้องร้อยละ 97.69 และเทคนิค Decision Tree ได้ค่าความถูกต้องร้อยละ 96.92 2.) ชุดข้อมูล Ecoli เทคนิค k-NN ได้ค่าความถูกต้อง ร้อยละ 100 มากกว่าเทคนิค Decision Tree ได้ค่าความถูกต้องร้อยละ 97 และเทคนิค Rule Base ได้ค่าความถูกต้องร้อยละ 92 3.) ชุดข้อมูล Car Evaluation เทคนิค k-NN ได้ค่าความถูกต้องร้อยละ 100 มากกว่าเทคนิค Rule Base ได้ค่าความถูกต้องร้อยละ 98.07 และเทคนิค Decision Tree ได้ค่า ความถูกต้องร้อยละ 97.79

นิเวศ จิระวิจิตชัย [12] ได้เสนอการค้นหาเทคนิคเหมือนข้อมูลเพื่อสร้างโมเดลการวิเคราะห์ โรคอัตโนมัติ โดยทดสอบประสิทธิภาพในการจำแนกสำหรับข้อมูลทางการแพทย์ ซึ่งทดลองกับ 7 เทคนิค ซึ่งประกอบด้วย Naive Bayes, Multilayer Perceptron, Radial Basis Function Network, SVM, k-NN, C4.5 และ Ripper ผลการทดลองพบว่า กลุ่มข้อมูล Hypothyroid เทคนิค C4.5 ให้ ประสิทธิภาพดีที่สุดร้อยละ 99.57 กลุ่มข้อมูล Leukemia เทคนิค Naive Bayes กับ SVM ให้ ประสิทธิภาพดีที่สุดร้อยละ 98.61 กลุ่มข้อมูล Breast-w เทคนิค SVM ให้ประสิทธิภาพดีที่สุดร้อยละ 96.99 กลุ่มข้อมูล Lymphography เทคนิค SVM ให้ประสิทธิภาพดีที่สุดร้อยละ 86.48 กลุ่มข้อมูล Hepatitis เทคนิค Radial Basis Function กับ k-NN ให้ประสิทธิภาพดีที่สุดร้อยละ 85.80 กลุ่มข้อมูล Heart-c เทคนิค SVM ให้ประสิทธิภาพดีที่สุดร้อยละ 84.15 กลุ่มข้อมูล Heart-statlog เทคนิค Radial Basis Function กับ SVM ให้ประสิทธิภาพดีที่สุดร้อยละ 84.07

Abdelghani และ Guven [36] ได้เสนอการทำนายการรอดชีวิตจากมะเร็งเต้านมโดยใช้ เทคนิคการทำเหมือนข้อมูล โดยใช้ 3 เทคนิค คือ Naive Bayes, BP-ANN และ C4.5 ผลการทดลอง พบว่าการทำนายการรอดชีวิตจากมะเร็งเต้านมโดยใช้เทคนิค C4.5 ให้ความถูกต้องร้อยละ 86 มากกว่า การทำนายด้วยเทคนิค BP-ANN ที่ให้ค่าความถูกต้องร้อยละ 86.5 และเทคนิค Naive Bayes ที่ให้ค่า ความถูกต้องร้อยละ 84.5

นิเวศ จิระวิจิตชัย [37] ได้เสนอการพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทย แบบอัตโนมัติ โดยลดคุณลักษณะของเอกสารก่อนประมวลผล เพื่อประโยชน์ในการลดมิติของข้อมูล ลดระยะเวลาประมวลผล ประหยัดทรัพยากรของระบบ และเพิ่มประสิทธิภาพในการจัดหมวดหมู่ เอกสารภาษาไทย จากการทดลองพบว่าเทคนิค SVM ให้ประสิทธิภาพสูงที่สุดร้อยละ 94.3 รองลงมาเป็น เทคนิค Naive Baye ร้อยละ 86.2 เทคนิค RBF ร้อยละ 86.1 เทคนิค C4.5 ร้อยละ 79.7 เทคนิค Ripper ร้อยละ 78.9 และเทคนิค k-NN ร้อยละ 69.5



บทที่ 3

วิธีดำเนินการศึกษา

วิธีดำเนินการในการศึกษาค้นคว้าอิสระได้นำเอาขั้นตอนการทำเหมืองข้อมูลมาใช้ เพื่อสร้างแบบจำลองพยากรณ์การสำเร็จการศึกษาของผู้สนใจที่จะเข้าศึกษาต่อในสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ซึ่งมีขั้นตอนการดำเนินการดังนี้ การเตรียมข้อมูล การทำกระบวนการก่อนการสร้างแบบจำลอง การสร้างแบบจำลอง และการวัดประสิทธิภาพแบบจำลอง

3.1 การเตรียมข้อมูล

การเตรียมข้อมูลของการศึกษาค้นคว้าอิสระครั้งนี้ ได้ข้อมูลมาจากการสำรวจข้อมูลจำนวนนักเรียนจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ประจำปีการศึกษา 2557-2558 จำนวน 11 แห่ง ดังตารางที่ 3.1

ตารางที่ 3.1 แสดงจำนวนนักเรียนประจำปีการศึกษา 2557-2558

รายชื่อสถานศึกษา	จำนวนนักเรียน ปีการศึกษา 2557 (คน)	จำนวนนักเรียน ปีการศึกษา 2558 (คน)
1. วิทยาลัยเทคโนโลยีพณิชยการกาฬสินธุ์	176	197
2. วิทยาลัยเทคโนโลยีกมลาลัย	38	35
3. วิทยาลัยเทคโนโลยีเพิ่มพูนบริหารธุรกิจ	21	19
4. วิทยาลัยอาชีวศึกษาลำปาว	13	14
5. วิทยาลัยเทคโนโลยีสหพันธ์	15	16
6. วิทยาลัยเทคโนโลยีสามชัย	9	11
7. วิทยาลัยอาชีวศึกษาไทย-เทศกาฬสินธุ์	12	15
8. วิทยาลัยอาชีวศึกษาไทยเทคนิส์	23	12
9. วิทยาลัยเทคโนโลยีสมเด็จพณิชยการ	132	268
10. วิทยาลัยเทคโนโลยีพัฒนบัณฑิต	18	10
11. วิทยาลัยอาชีวศึกษาไทยเทคนิส์	130	82
รวม	587	679
รวมทั้งสิ้น	1,266	



ซึ่งข้อมูลที่ใช้การสำรวจ ได้แก่

- | | |
|---|-----------------------------------|
| 1. เพศ | 2. อายุ |
| 3. จำนวนพี่น้อง | 4. สถานภาพครอบครัว |
| 5. อาชีพบิดา | 6. อาชีพมารดา |
| 7. อำเภอ | 8. เกรดเฉลี่ยกลุ่มภาษาไทย |
| 9. เกรดเฉลี่ยกลุ่มคณิตศาสตร์ | 10. เกรดเฉลี่ยกลุ่มวิทยาศาสตร์ |
| 11. เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม | |
| 12. เกรดเฉลี่ยกลุ่มสุขศึกษา พลศึกษา | 13. เกรดเฉลี่ยกลุ่มศิลปะ |
| 14. เกรดเฉลี่ยกลุ่มการทำงานอาชีพและเทคโนโลยี | 15. เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ |
| 16. ประเภทวิชาที่จบ | 17. จบการศึกษา/ไม่จบการศึกษา |

3.2 การทำกระบวนการก่อนการสร้างแบบจำลอง

3.2.1 การแปลงข้อมูล

การแปลงข้อมูล (Data transformation) จากข้อมูลที่ได้จากการสำรวจข้อมูลจำนวนนักเรียนจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ประจำปีการศึกษา 2557-2558 จำนวน 11 แห่ง มีทั้งหมด 17 ตัวแปร ดังตารางที่ 3.2

ตารางที่ 3.2 แสดงการจำแนกชนิดของตัวแปร

ชื่อตัวแปร	คำอธิบายตัวแปร	ชนิดของตัวแปร
Sex	เพศ	Nominal
Age	อายุ	Numeric
Number of siblings	จำนวนพี่น้อง	Numeric
Family Status	สถานภาพครอบครัว	Nominal
Father Career	อาชีพบิดา	Nominal
Mother Career	อาชีพมารดา	Nominal
District	อำเภอ	Nominal
Thai	เกรดเฉลี่ยกลุ่มภาษาไทย	Nominal
Math	เกรดเฉลี่ยกลุ่มคณิตศาสตร์	Nominal
Science	เกรดเฉลี่ยกลุ่มวิทยาศาสตร์	Nominal
Social	เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม	Nominal
Health	เกรดเฉลี่ยกลุ่มสุขศึกษา พลศึกษา	Nominal



ตารางที่ 3.2 แสดงการจำแนกชนิดของตัวแปร (ต่อ)

ชื่อตัวแปร	คำอธิบายตัวแปร	ชนิดของตัวแปร
Art	เกรดเฉลี่ยกลุ่มศิลปะ	Nominal
Career and Technology	เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี	Nominal
Foreign language	เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ	Nominal
Type of course	ประเภทวิชาที่จบ	Nominal
Graduate (Class)	จบการศึกษา	Nominal

ตารางที่ 3.3 แสดงการแปลงค่าตัวแปร

ชื่อตัวแปร	ข้อมูล	ค่าตัวแปร	ข้อมูลในตาราง
Sex:เพศ	ชาย	Male	1
	หญิง	Female	2
Age:อายุ	อายุ	ใช้ตัวเลขจริง	ใช้ตัวเลขจริง
Number of siblings: จำนวนพี่น้อง	จำนวนพี่น้อง	ใช้ตัวเลขจริง	ใช้ตัวเลขจริง
Parents Status: สถานภาพครอบครัว	อยู่ด้วยกัน	Family	1
	แยกกันอยู่	Separated	2
	ถึงแก่กรรม	Deceased	3
Father Career:อาชีพบิดา	รับราชการ	Government	1
	พนักงานรัฐวิสาหกิจ	State Enterprises	2
	นักธุรกิจ-ค้าขาย	Trade	3
	เกษตรกรรม	Agriculture	4
	รับจ้าง	Contractors	5
	พนักงาน/เจ้าหน้าที่ของรัฐ	Government Employee	6
	ข้าราชการเกษียณอายุราชการ	Retired Government	7
	พระ/นักบวช	Priest	8
	ไม่ได้ประกอบอาชีพ	Jobless	9
	อื่นๆ	Others	10
	เสียชีวิต	Deceased	11



ตารางที่ 3.3 แสดงการแปลงค่าตัวแปร (ต่อ)

ชื่อตัวแปร	ข้อมูล	ค่าตัวแปร	ข้อมูลในตาราง
Mother Career:อาชีพ มารดา	รับราชการ	Government	1
	พนักงานรัฐวิสาหกิจ	State Enterprises	2
	นักธุรกิจ-ค้าขาย	Trade	3
	เกษตรกรรม	Agriculture	4
	รับจ้าง	Contractors	5
	พนักงาน/เจ้าหน้าที่ของรัฐ	Government Employee	6
	ข้าราชการเกษียณอายุ ราชการ	Retired Government	7
	พระ/นักบวช	Priest	8
	ไม่ได้ประกอบอาชีพ	Jobless	9
	อื่นๆ	Others	10
	เสียชีวิต	Deceased	11
Amphoe:อำเภอ	อำเภอ	อำเภอเมืองกาฬสินธุ์	1
		อำเภอกมลาไสย	2
		อำเภอยางตลาด	3
		อำเภอห้วยเม็ก	4
		อำเภอสหัสขันธ์	5
		อำเภอสามชัย	6
		อำเภอท่าคันโท	7
		อำเภอสมเด็จ	8
		อำเภอคำม่วง	9
		อำเภอห้วยผึ้ง	10
		อำเภอกุฉินารายณ์	11
		อำเภอเขาวง	12
Thai:เกรดเฉลี่ย กลุ่มภาษาไทย	เกรดเฉลี่ยกลุ่มภาษาไทย	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4



ตารางที่ 3.3 แสดงการแปลงค่าตัวแปร (ต่อ)

ชื่อตัวแปร	ข้อมูล	ค่าตัวแปร	ข้อมูลในตาราง
Math:เกรดเฉลี่ย กลุ่มคณิตศาสตร์	เกรดเฉลี่ยกลุ่มคณิตศาสตร์	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4
Science:เกรดเฉลี่ย กลุ่มวิทยาศาสตร์	เกรดเฉลี่ยกลุ่มวิทยาศาสตร์	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4
Social:เกรดเฉลี่ย กลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม	เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4
Health:เกรดเฉลี่ย กลุ่มสุขศึกษา พลศึกษา	เกรดเฉลี่ยกลุ่มสุขศึกษา พลศึกษา	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4
Art:เกรดเฉลี่ยกลุ่มศิลปะ	เกรดเฉลี่ยกลุ่มศิลปะ	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3



ตารางที่ 3.3 แสดงการแปลงค่าตัวแปร (ต่อ)

ชื่อตัวแปร	ข้อมูล	ค่าตัวแปร	ข้อมูลในตาราง
		3.50-3.99	3.5
		4.00	4
Career and Technology:เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี	เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4
Foreign language: เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ	เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ	1.00-1.49	1
		1.50-1.99	1.5
		2.00-2.49	2
		2.50-2.99	2.5
		3.00-3.49	3
		3.50-3.99	3.5
		4.00	4
Type of course: ประเภทวิชา	พาณิชยกรรม	1	1
	อุตสาหกรรม	2	2
Graduate:จบการศึกษา	จบการศึกษา	1	1
	ไม่จบการศึกษา	2	2

3.2.2 การวิเคราะห์ปัจจัย

การวิเคราะห์ปัจจัยเป็นการคัดเลือกแอตทริบิวต์ที่สามารถเป็นตัวแทนของกลุ่มแอตทริบิวต์เพื่อลดจำนวนแอตทริบิวต์ในการพยากรณ์ซึ่งงานวิจัยนี้ได้นำเทคนิคการวิเคราะห์ปัจจัยมาใช้ จำนวน 3 เทคนิค คือ

1. เทคนิค Cfs Subset Evaluation
2. เทคนิค Gain Ratio Attribute Evaluation
3. เทคนิค Wrapper Subset Evaluation



3.3 การสร้างแบบจำลอง

การสร้างแบบจำลองพยากรณ์การสำเร็จการศึกษาของผู้ที่สนใจเข้าศึกษาต่อสถานศึกษา อาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ เพื่อช่วยในการตัดสินใจว่าเมื่อเข้าศึกษาต่อแล้วจะสามารถสำเร็จ การศึกษาได้หรือไม่ด้วยตัวแปรทั้ง 17 ตัวแปรที่ได้จากการสำรวจข้อมูลจำนวนนักเรียนในปีการศึกษา 2557-2558 นั้น ด้วยการทำเหมืองข้อมูลและเทคนิคที่นำมาใช้ในการสร้างแบบจำลองมีจำนวน 6 เทคนิคคือ

1. เทคนิค C4.5
2. เทคนิค Random Forest
3. เทคนิค Random Tree
4. เทคนิค REP Tree
5. เทคนิค K-Nearest Neighbour (k-NN)
6. เทคนิค Support Vector Machine (SVM)

3.4 การวัดประสิทธิภาพของแบบจำลอง

ในการวัดประสิทธิภาพของแบบจำลองนั้น ได้มีการใช้เทคนิคแบบ K-fold cross validation โดยทำการวัดประสิทธิภาพของแบบจำลองจำนวน 3 ครั้ง ครั้งที่ 1 วัดประสิทธิภาพด้วย 5-fold cross validation ครั้งที่ 2 วัดประสิทธิภาพด้วย 7-fold cross validation ครั้งที่ 3 วัดประสิทธิภาพด้วย 10-fold cross validation แล้วก็นำการวัดประสิทธิภาพทั้ง 3 ครั้งมาเปรียบเทียบ ว่าครั้งใดวัด ประสิทธิภาพได้แม่นยำที่สุด

3.5 การพัฒนาระบบ

การพัฒนาระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษา จังหวัดกาฬสินธุ์ หลังจากที่ได้มีการเตรียมข้อมูล ทำความสะอาดข้อมูล วิเคราะห์ปัจจัยที่ส่งผลต่อ การสำเร็จการศึกษาด้วย 3 เทคนิค คือ 1.เทคนิค Cfs Subset Evaluation 2.เทคนิค Gain Ratio Attribute Evaluation 3.เทคนิค Wrapper Subset Evaluation แล้วก็ทำการทดลองเพื่อหา แบบจำลองที่มีประสิทธิภาพมากที่สุดด้วย 6 เทคนิค คือ 1. เทคนิค C4.5 2. เทคนิค Random Forest 3.เทคนิค Random Tree 4. เทคนิค REP Tree 5. เทคนิค K-Nearest Neighbour (k-NN) 6. เทคนิค Support Vector Machine (SVM) และวัดประสิทธิภาพแบบจำลองด้วย k-fold cross validation วิเคราะห์ประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความ ระลึก (Recall) เมื่อได้แบบจำลองที่มีประสิทธิภาพมากที่สุดแล้ว ผู้ศึกษาได้นำแบบจำลองนั้นมาพัฒนา ส่วนติดต่อกับผู้ใช้งานแบบ GUI ใช้โปรแกรมภาษา PHP ในการสร้างเว็บเพจและสร้างโมดูลเกี่ยวกับ งานทั้งหมดซึ่งการพัฒนาระบบจะใช้ระบบปฏิบัติการ Microsoft Windows 7 ขึ้นไป



บทที่ 4

ผลการศึกษา

ในการดำเนินการศึกษาค้นคว้าครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาปัจจัยที่มีผลต่อการสำเร็จการศึกษาของนักเรียนในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ และพัฒนาโปรแกรมพยากรณ์การสำเร็จการศึกษาของนักเรียนที่ต้องการศึกษาต่อในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ ผู้ศึกษาจะกล่าวถึงผลการศึกษาตามลำดับดังนี้

1. ผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา
2. ผลการวัดประสิทธิภาพแบบจำลองทั้ง 6 เทคนิค
3. ผลการพัฒนาระบบพยากรณ์การสำเร็จการศึกษา

4.1 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา

การวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา จำนวน 17 ปัจจัย ที่ผู้ศึกษาได้ทำการรวบรวมไว้ ด้วย 3 เทคนิคในโปรแกรม WEKA คือ

1. เทคนิค Cfs Subset Evaluation
2. เทคนิค Gain Ratio Attribute Evaluation
3. เทคนิค Wrapper Subset Evaluation

ซึ่งผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษาเป็นไปตามตารางที่ 4.1

ตารางที่ 4.1 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา

เทคนิค Cfs Subset Evaluation		เทคนิค Gain Ratio Attribute Evaluation		เทคนิค Wrapper Subset Evaluation	
เลือก	ไม่เลือก	เลือก	ไม่เลือก	เลือก	ไม่เลือก
District	Sex	District	Number of siblings		Sex
Graduate (Class)	Age	Father Career	Age		Age
	Number of siblings	Mother Career			Number of siblings
	Family Status	Career and Technology			Family Status
	Father Career	Math			Father Career



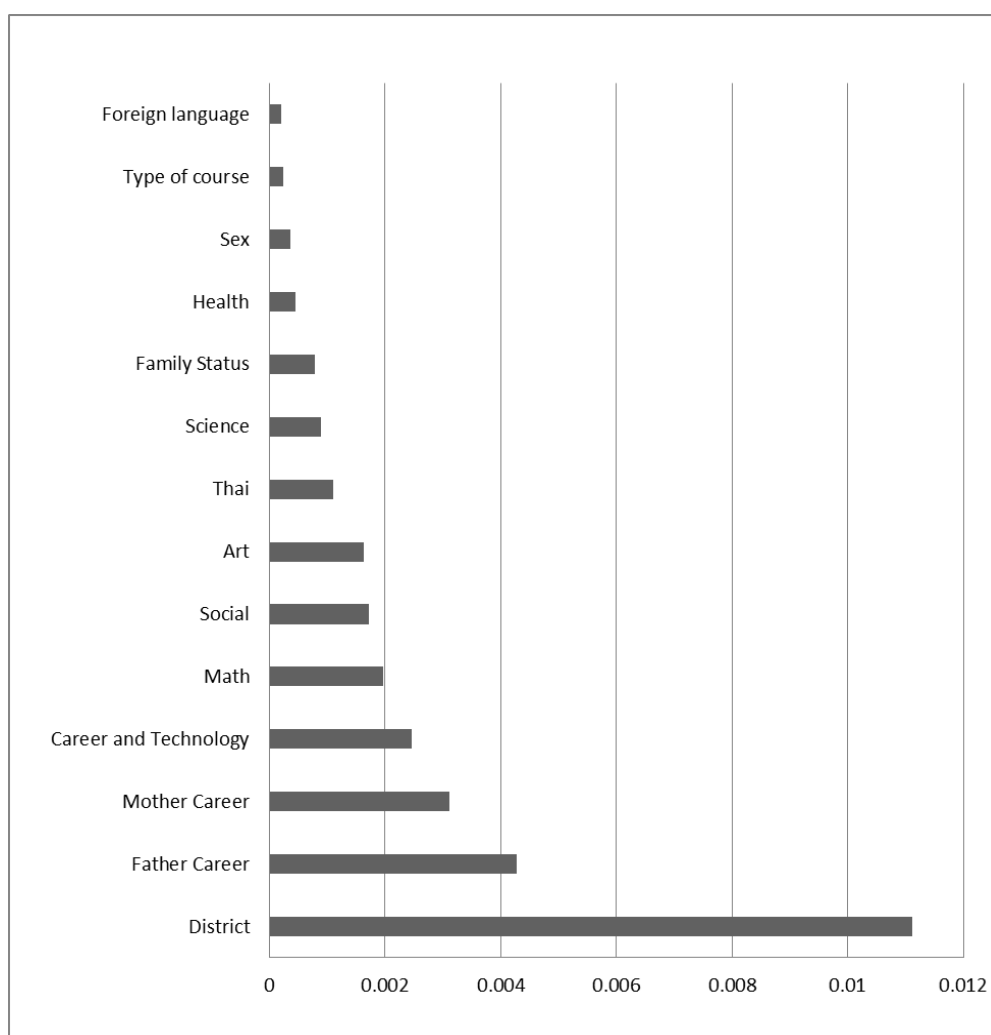
ตารางที่ 4.1 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา (ต่อ)

เทคนิค Cfs Subset Evaluation		เทคนิค Gain Ratio Attribute Evaluation		เทคนิค Wrapper Subset Evaluation	
	Mother Career	Social			Mother Career
	Thai	Art			District
	Math	Thai			Thai
	Science	Science			Math
	Social	Family Status			Science
	Health	Health			Social
	Art	Sex			Health
	Career and Technology	Type of course			Art
	Foreign language	Foreign language			Career and Technology
	Type of course	Graduate			Foreign language
					Type of course
					Graduate (Class)

จากตารางที่ 4.1 พบว่าการวิเคราะห์ปัจจัยด้วยเทคนิค Cfs Subset Evaluation ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษามา 2 ปัจจัยคือ 1.District 2.Graduate แต่ไม่สามารถทำการทดลองได้เนื่องจากปัจจัยน้อย

เทคนิค Gain Ratio Attribute Evaluation ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษามา 15 ปัจจัยดังภาพที่ 4.1





ภาพที่ 4.1 การวิเคราะห์ปัจจัยด้วย Gain Ratio

จากภาพที่ 4.1 แสดงให้เห็นค่าความสำคัญของแต่ละตัวแปรตามลำดับดังนี้ 1. District ด้วยค่า 0.011115 2. Father Career ด้วยค่า 0.004282 3. Mother Career ด้วยค่า 0.003105 4. Career and Technology ด้วยค่า 0.002459 5. Math ด้วยค่า 0.001971 6. Social ด้วยค่า 0.001721 7. Art ด้วยค่า 0.001634 8. Thai ด้วยค่า 0.001109 9. Science ด้วยค่า 0.000891 10. Family Status ด้วยค่า 0.000776 11. Health ด้วยค่า 0.000445 12. Sex ด้วยค่า 0.000353 13. Type of course ด้วยค่า 0.000239 14. Foreign language ด้วยค่า 0.000208 15. Graduate (Class) และเทคนิค Wrapper Subset Evaluation ไม่ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษา

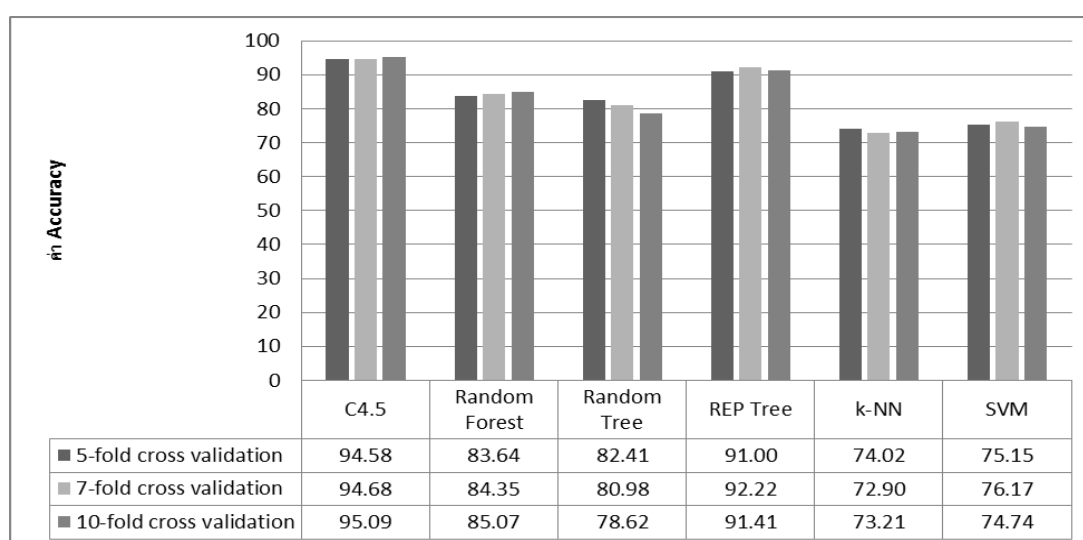


4.2 ผลการวัดประสิทธิภาพแบบจำลองทั้ง 6 เทคนิค

จากการทดลองครั้งแรกผู้ศึกษาได้ทดลองโดยการให้ตัวแปรเกรดเฉลี่ยกลุ่มภาษาไทย เกรดเฉลี่ยกลุ่มคณิตศาสตร์ เกรดเฉลี่ยกลุ่มวิทยาศาสตร์ เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม เกรดเฉลี่ยกลุ่มสุขศึกษา พลศึกษา เกรดเฉลี่ยกลุ่มศิลปะ เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี และเกรดเฉลี่ยกลุ่มภาษาต่างประเทศ เป็นตัวแปรชนิด Numeric ใช้เกรดจริงในการทดลอง เมื่อทำการวัดประสิทธิภาพแบบจำลองแล้วผลการทดลองปรากฏว่า เทคนิค C4.5 มีประสิทธิภาพมากที่สุด ด้วยค่าความถูกต้อง (Accuracy) 87.73% ค่าความแม่นยำ (Precision) 87.50% และค่าความระลึก (Recall) 87.70% แล้วได้ทำการทดลองต่อโดยการเปลี่ยนตัวแปรข้างต้น เป็นตัวแปรชนิด Nominal โดยทำการจัดกลุ่มเกรดเฉลี่ยดังนี้ 1.00-1.49=1, 1.50-1.99=1.5, 2.00-2.49=2, 2.50-2.99=2.5, 3.00-3.49=3, 3.50-3.99=3.5, 4.00=4 ผลการทดลองปรากฏว่า เทคนิค C4.5 มีประสิทธิภาพมากที่สุดและสูงกว่าการที่ให้ตัวแปรกลุ่มนี้เป็นตัวแปรชนิด Numeric ด้วยค่าความถูกต้อง (Accuracy) 95.09% ค่าความแม่นยำ (Precision) 95.10% และค่าความระลึก (Recall) 95.10%

จากข้อ 4.1 ผลการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษา แล้วจึงมาทำการวัดประสิทธิภาพแบบจำลองทั้ง 6 เทคนิคโดยทำการเปรียบเทียบโดย 5-fold cross validation, 7-fold cross validation และ 10-fold cross validation และทำการวิเคราะห์ประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ด้วยการวิเคราะห์ปัจจัยทั้ง 3 เทคนิคดังภาพที่ 4.2, 4.3 และ 4.4

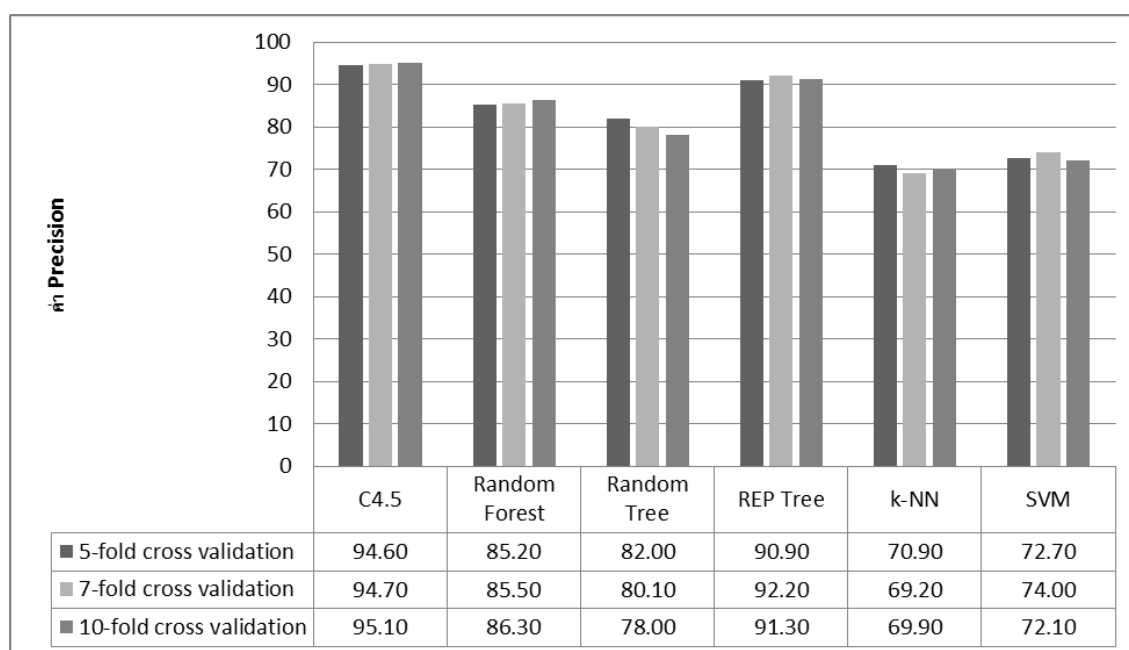
การวิเคราะห์ปัจจัยด้วยเทคนิค Cfs Subset Evaluation ซึ่งเทคนิคนี้ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษา มาเพียง 2 ปัจจัย ซึ่งมีจำนวนน้อยจึงไม่สามารถทำการทดลองหาผลการวัดประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ได้



ภาพที่ 4.2 ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ทั้ง 6 เทคนิค
จากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation



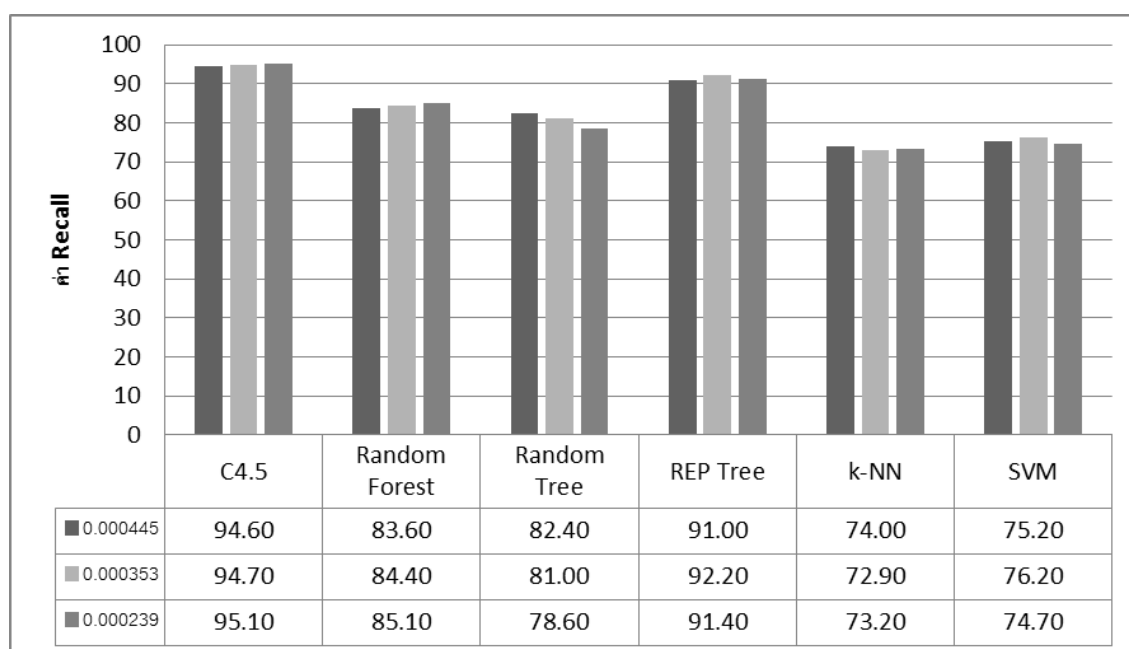
จากภาพที่ 4.2 แสดงผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation ผลปรากฏว่าเทคนิค C4.5 ให้ค่าความถูกต้องมากที่สุด 95.09% ด้วย 10-fold cross validation เทคนิค REP Tree ให้ค่าความถูกต้อง 92.22% ด้วย 7-fold cross validation เทคนิค Random Forest ให้ค่าความถูกต้อง 85.07% ด้วย 10-fold cross validation เทคนิค Random Tree ให้ค่าความถูกต้อง 82.41% ด้วย 5-fold cross validation เทคนิค SVM ให้ค่าความถูกต้อง 76.17% ด้วย 7-fold cross validation และเทคนิค k-NN ให้ค่าความถูกต้อง 74.02% ด้วย 5-fold cross validation ตามลำดับ



ภาพที่ 4.3 ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความแม่นยำ (Precision) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation

จากภาพที่ 4.3 แสดงผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความแม่นยำ (Precision) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation ผลปรากฏว่าเทคนิค C4.5 ให้ค่าความแม่นยำมากที่สุด 95.10% ด้วย 10-fold cross validation เทคนิค REP Tree ให้ค่าความแม่นยำ 92.20% ด้วย 7-fold cross validation เทคนิค Random Forest ให้ค่าความแม่นยำ 86.30% ด้วย 10-fold cross validation เทคนิค Random Tree ให้ค่าความแม่นยำ 82.00% ด้วย 5-fold cross validation เทคนิค SVM ให้ค่าความแม่นยำ 74.00% ด้วย 7-fold cross validation และเทคนิค k-NN ให้ค่าความแม่นยำ 70.90% ด้วย 5-fold cross validation ตามลำดับ





ภาพที่ 4.4 ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความระลึก (Recall) ทั้ง 6 เทคนิค
จากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation

จากภาพที่ 4.4 แสดงผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความระลึก (Recall) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation ผลปรากฏว่า เทคนิค C4.5 ให้ค่าความระลึกมากที่สุด 95.10% ด้วย 10-fold cross validation เทคนิค REP Tree ให้ค่าความระลึก 92.20% ด้วย 7-fold cross validation เทคนิค Random Forest ให้ค่าความระลึก 85.10% ด้วย 10-fold cross validation เทคนิค Random Tree ให้ค่าความระลึก 82.40% ด้วย 5-fold cross validation เทคนิค SVM ให้ค่าความระลึก 76.20% ด้วย 7-fold cross validation และเทคนิค k-NN ให้ค่าความระลึก 74.00% ด้วย 5-fold cross validation ตามลำดับ

ส่วนการวิเคราะห์ปัจจัยด้วยเทคนิค Wrapper Subset Evaluation ซึ่งเทคนิคนี้ไม่ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษา จึงไม่สามารถทำการทดลองหาผลการวัดประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ได้

4.3 ผลการพัฒนาระบบพยากรณ์การสำเร็จการศึกษา

เมื่อได้แบบจำลอง C4.5 ที่ให้ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) สูงกว่าแบบจำลองอื่น แล้วก็นำมาสร้างระบบพยากรณ์การสำเร็จการศึกษา แสดงดังภาพที่ 4.5



ระบบพยากรณ์การสำเร็จการศึกษาระดับ ปวช.

สำหรับ อาชีวศึกษาเอกชนกาฬสินธุ์

แบบฟอร์มกรอกข้อมูล

ชื่อ	นายพัฒนพงษ์	นามสกุล	ดลรัตน์
เพศ	ชาย	สถานภาพบิดา มารดา	อยู่ด้วยกัน
อาชีพบิดา	พนักงานรัฐวิสาหกิจ	อาชีพมารดา	นักธุรกิจ-ค้าขาย
อำเภอ	อำเภอยางตลาด		

เกรดเฉลี่ยกลุ่มภาษาไทย	2.50-2.99=2.5
เกรดเฉลี่ยกลุ่มคณิตศาสตร์	2.50-2.99=2.5
เกรดเฉลี่ยกลุ่มวิทยาศาสตร์	2.00-2.49=2.0
เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม	2.00-2.49=2.0
เกรดเฉลี่ยกลุ่มสุขศึกษา พลศึกษา	2.50-2.99=2.5
เกรดเฉลี่ยกลุ่มศิลปะ	2.50-2.99=2.5
เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี	3.00-3.49=3.0
เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ	1.50-1.99=1.5

ประเภทวิชาที่ต้องการเรียน	อุตสาหกรรม
---------------------------	---

ภาพที่ 4.5 แสดงหน้าระบบพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชน จังหวัดกาฬสินธุ์



ข้อมูลผู้พยากรณ์

ชื่อ - นามสกุล นายพัฒนพงษ์ ตลรัตน์

เพศ ชาย

สถานภาพบิดา มารดา อยู่ด้วยกัน

อาชีพบิดา พนักงานรัฐวิสาหกิจ

อาชีพมารดา นักธุรกิจ-ค้าขาย

อำเภอ อำเภอขามเฒ่า

เกรดเฉลี่ยรายวิชาต่างๆ							
ภาษาไทย	คณิตศาสตร์	วิทยาศาสตร์	สังคมศึกษา	สุขศึกษา	ศิลปะ	การงาน อาชีพและเทคโนโลยี	ภาษาต่างประเทศ
2.5	2.5	2	2	2.5	2.5	3	1.5
ประเภทวิชาที่ต้องการเรียน		อุตสาหกรรม					
ผลการพยากรณ์ปรากฏว่า		คุณมีสิทธิ์จบ		ด้วยค่าความถูกต้องร้อยละ		100	

ภาพที่ 4.6 แสดงหน้าผลการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชน จังหวัดกาฬสินธุ์



บทที่ 5

วิจารณ์ผลการทดลอง สรุป และข้อเสนอแนะ

ในการศึกษาและพัฒนาแบบจำลองที่เหมาะสมสำหรับการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์ โดยใช้เทคนิคในการเปรียบเทียบได้แก่ C4.5, Random Forest, Random Tree, REP Tree, k-NN และ SVM ผลการทดลองพบว่าเทคนิค C4.5 มีความเหมาะสมมากที่สุดในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์ โดยการวิเคราะห์ประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

5.1 วิจารณ์ผลการทดลอง

การวิเคราะห์ปัจจัยด้วยเทคนิค Cfs Subset Evaluation ซึ่งเทคนิคนี้ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษา มาเพียง 2 ปัจจัย ซึ่งมีจำนวนน้อยจึงไม่สามารถทำการทดลองหาผลการวัดประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ได้

การวัดประสิทธิภาพแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation ผลปรากฏว่า เทคนิค C4.5 ให้ค่าความถูกต้องมากที่สุด 95.09% ด้วย 10-fold cross validation เทคนิค REP Tree ให้ค่าความถูกต้อง 92.22% ด้วย 7-fold cross validation เทคนิค Random Forest ให้ค่าความถูกต้อง 85.07% ด้วย 10-fold cross validation เทคนิค Random Tree ให้ค่าความถูกต้อง 82.41% ด้วย 5-fold cross validation เทคนิค SVM ให้ค่าความถูกต้อง 76.17% ด้วย 7-fold cross validation และเทคนิค k-NN ให้ค่าความถูกต้อง 74.02% ด้วย 5-fold cross validation ตามลำดับ

การวัดประสิทธิภาพแบบจำลองด้วยค่าความแม่นยำ (Precision) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation ผลปรากฏว่า เทคนิค C4.5 ให้ค่าความแม่นยำมากที่สุด 95.10% ด้วย 10-fold cross validation เทคนิค REP Tree ให้ค่าความแม่นยำ 92.20% ด้วย 7-fold cross validation เทคนิค Random Forest ให้ค่าความแม่นยำ 86.30% ด้วย 10-fold cross validation เทคนิค Random Tree ให้ค่าความแม่นยำ 82.00% ด้วย 5-fold cross validation เทคนิค SVM ให้ค่าความแม่นยำ 74.00% ด้วย 7-fold cross validation และเทคนิค k-NN ให้ค่าความแม่นยำ 70.90% ด้วย 5-fold cross validation ตามลำดับ

การวัดประสิทธิภาพแบบจำลองด้วยค่าความระลึก (Recall) ทั้ง 6 เทคนิคจากการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation ผลปรากฏว่า เทคนิค C4.5 ให้ค่าความระลึกมากที่สุด 95.10% ด้วย 10-fold cross validation เทคนิค REP Tree ให้ค่าความระลึก 92.20% ด้วย 7-fold cross validation เทคนิค Random Forest ให้ค่าความระลึก 85.10% ด้วย 10-fold cross validation เทคนิค Random Tree ให้ค่าความระลึก 82.40% ด้วย 5-fold cross validation



เทคนิค SVM ให้ค่าความระลึก 76.20% ด้วย 7-fold cross validation และเทคนิค k-NN ให้ค่าความระลึก 74.00% ด้วย 5-fold cross validation ตามลำดับ

ส่วนการวิเคราะห์ปัจจัยด้วยเทคนิค Wrapper Subset Evaluation ซึ่งเทคนิคนี้ไม่ได้เลือกปัจจัยที่ส่งผลต่อการสำเร็จการศึกษา จึงไม่สามารถทำการทดลองหาผลการวัดประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ได้

5.2 สรุป

สรุปได้ว่า เทคนิค C4.5 มีความเหมาะสมในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนกรณีศึกษาจังหวัดกาฬสินธุ์มากที่สุดเมื่อใช้ร่วมกับการวิเคราะห์ปัจจัยด้วยเทคนิค Gain Ratio Attribute Evaluation

5.3 ข้อเสนอแนะ

หากผู้ที่สนใจศึกษาหรือสนใจที่จะพัฒนางานวิจัยนี้ ควรเพิ่มระบบการแนะนำความเหมาะสมกับสาขาที่ควรเข้าศึกษาต่อ เช่น สาขาคอมพิวเตอร์ธุรกิจ สาขาการบัญชี สาขาการขาย สาขานานยนต์ สาขาอิเล็กทรอนิกส์ สาขาไฟฟ้า เป็นต้น เพื่อให้ผู้สนใจเข้าศึกษาต่อมีข้อมูลที่จะทำการตัดสินใจเลือกสาขาเรียนต่อไป



เอกสารอ้างอิง



เอกสารอ้างอิง

- [1] ข้าราชการศึกษา. “การเพิ่มสัดส่วนสายอาชีพเป็น 60% สายสามัญเหลือ 40%”. [ออนไลน์]. [สืบค้นเมื่อ 28 สิงหาคม 2559]. <http://www.unigang.com/Article/5170>.
- [2] วันฉัตร ทิพย์มาศ. “ปัจจัยที่มีผลต่อการตัดสินใจเข้าศึกษาต่อวิทยาลัยอาชีวศึกษาเอกชนของนักเรียนในเขตภาคใต้ตอนบน”. วารสารศึกษาศาสตร์ 2556; 24[2]: 150-160.
- [3] Kavitha R.K. and Rangasamy D. Dorai. “Predicting Breast Cancer Survivability Using Naive baysein Classifier And C4.5 Algorithm”. Elysium Journal Engineering Research & Management 2014; 1[1]: 61-63.
- [4] เพียงฤทัย หนูสวัสดิ์. “การสร้างโมเดลทำนายอัตราการใช้พลังงานของแบตเตอรี่มือถือโดยใช้เทคนิคเหมืองข้อมูล” [วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต]. กรุงเทพฯ: มหาวิทยาลัยศิลปากร; 2556.
- [5] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. Top 10 Algorithms in Data Mining”. Knowledge and Information Systems 2008; 14: 1–37.
- [6] จงจิตต์ ฤทธิรงค์ และรีนา ต๊ะดี. “ข้อท้าทายในการผลิตแรงงานฝีมือไทยเพื่อเข้าสู่ตลาดแรงงานประชาคมเศรษฐกิจอาเซียน”. วารสารสถาบันวิจัยประชากรและสังคม มหาวิทยาลัยมหิดล 2558; 2: 129-147.
- [7] พิเชษฐ์ สุขเสกสรรค์. “บทบาทใหม่ของวิศวกรในภาวะการณ์การขาดแคลนแรงงานช่างฝีมือหรือช่างเทคนิค”. รายงานการวิจัย. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2549.
- [8] ส่วนเศรษฐกิจภาค ธนาคารแห่งประเทศไทย สำนักงานภาคตะวันออกเฉียงเหนือ. “ความไม่สมดุลของตลาดแรงงานไทยนัยของการขาดแคลนแรงงาน”. ขอนแก่น: คลังนาโนวิทยา; 2555.
- [9] Han J. & Kamber M. “Data Mining Concepts and Techniques”. CA: Morgan Kaufmann, San Francisco; 2006.
- [10] Buathong W. “Effect of data dimension reduction Analysis and Data Classification Performance of Decision Tree, Support Vector Machine and Naïve Bayes”. [Online]. 2012 [cited 22 August 2015] ; Available from: http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20140511180257.pdf.
- [11] Hall M. A. “Correlation-based Feature Subset Selection for Machine Learning”. [Ph.D. thesis]. Waikato: The University of Waikato; 1998.
- [12] นิเวศ จิระวิชิตชัย. “การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ”. รายงานการวิจัย. กรุงเทพฯ: มหาวิทยาลัยราชภัฏสวนสุนันทา; 2553.



- [13] พฤฒิพงศ์ เพ็งศิริ, สุริยะ พินิจการ, ณัฐชดา มงคลชาติ, นวพร วิสิฐพงศ์พันธ์ และพยุง มีสัจ. “การลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์ทโฟน”. [Online]. 2014. [cited 10 August 2015]; Available from: http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20140511180257.pdf.
- [14] พรเทพ คงไชย และรัชฎา คงคะจันทร์. “การศึกษาเชิงเปรียบเทียบในการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับการทำเหมืองข้อมูลเพื่อพยากรณ์โอกาสการสำเร็จการศึกษาของนักศึกษา”. presented at the การประชุมวิชาการข่ายงานวิศวกรรมอุตสาหกรรม ประจำปี 2554; 20-21 ตุลาคม 2554; กรุงเทพฯ, 2554. หน้า 2012-2016.
- [15] Quinlan R. “C4.5: Programs for Machine Learning: Morgan Kaufmann Publishers”. Morgan kaufmann Publishers In.: San Francisco, CA; 1993.
- [16] วิระ จิรภิจนุสรณ์ และชูลีรัตน์ จรัสกุลชัย. “การคาดการณ์ภาษีมูลค่าเพิ่มด้วยเทคนิคของเหมืองข้อมูล”. The 6 th National Conference On Computing And Information Technology 2553; 1: 193-198.
- [17] สุภโชค เรืองศรี. “การพัฒนาระบบประมาณการใช้กระดาษภายในแผนกผู้โดยสารด้วยวิธีโครงข่ายประสาทเทียม”. รายงานการวิจัย. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2555.
- [18] Kohavi R. & John G. H. “Wrappers for Feature Subset Selection”. Artificial Intelligence 2557; 97: 273-324.
- [19] สุนทรทิพย์ วงศ์พันธ์. “การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา” [วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต]. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์; 2551.
- [20] Rodriguez, D., Ruiz, R., Cuadrado-Gallego, J., Aguilar-Ruiz, J. & Garre, M. “Attribute selection in software engineering datasets for detecting fault modules”. 33rd EUROMICRO Conference on Software Engineering and Advanced Applications 2007; 418-423.
- [21] Breiman L. “Random Forest”. Machine Learning 2001; 45: 5-32.
- [22] Krishnaveni. S & Hemalatha. M. “A Perspective Analysis of Traffic Accident using Data Mining Techniques”. International Journal of Computer Applications 2011;; 23[7]: 40-48.
- [23] ภรณ์ยา ปาลวิสุทธิ. “การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูล การเป็นโรคติดเชื้อในท่อน้ำดี”. วารสารเทคโนโลยีสารสนเทศ 2559; 12[1]: 54-63.
- [24] Weka. [Online]. 2016 [cited 3 Decemver 2016]; Available from: <http://www.cs.waikato.ac.nz/ml/index.html>.



- [25] จิตติมา ช่วงชัย. “การวิเคราะห์หารูปแบบการเรียนรู้โดยใช้เหมืองข้อมูลของนักศึกษาต่อการจัดทำปฏิญานพนธ์”. วารสารบัณฑิตศึกษา มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์ 2559; 7[1]: 53-62.
- [26] Kalmegh S. “Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News”. International Journal of Innovative Science, Engineering & Technology 2015; 2[2]: 438-446.
- [27] Kapoor P. & Rani R. “Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning”. International Journal of Engineering Research and General Science 2015; 3[3]: 1613-1621.
- [28] Wisaeng K. “A Comparison of Decision Tree Algorithms For UCI Repository Classification”. International Journal of Engineering Trends and Technology 2013; 4[8]: 3393-3397.
- [29] ศลิษา หนูเสมียน. “ระบบแนะนำการเลือกสาขาเพื่อศึกษาต่อระดับอาชีวศึกษาโดยเทคนิคการคัดกรองข้อมูลแบบผสมระหว่างการคัดกรองข้อมูลแบบอิงเนื้อหากับการคัดกรองแบบพึ่งพาผู้ใช้ร่วมกรณีศึกษา วิทยาลัยสารพัดช่างระยอง”. [วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต]. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2554.
- [30] Zaki M. J., Meira Jr. W. & Meira, M. “Data Mining And Analysis Fundamental Concepts and Algorithms”. Cambridge: Cambridge University Press; 2014.
- [31] ชีรพงษ์ สังข์ศรี. “การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล”. [Online]. 2014 [cited 3 Decemver 2016]; Available http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20141111160105.pdf.
- [32] ลาก พุ่มหิรัญ และมาลีรัตน์ โสตานิล. “การพัฒนาตัวแบบการพยากรณ์จำนวนนักศึกษาใหม่โดยใช้กฎการจำแนกเทคนิคต้นไม้ตัดสินใจ”. วารสารวิชาการและวิจัย มหาวิทยาลัยเทคโนโลยีราชมงคลพระนคร 2554; ฉบับพิเศษ:
- [33] พรณิภา บุตรเอก และสุรเดช บุญลือ. “การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาโดยใช้ซอฟต์แวร์เวกเตอร์แมชชีน”. วารสารวิทยาศาสตร์และเทคโนโลยีมหาวิทยาลัยศิลปากร 2557; 1[6]: 40-49.
- [34] ยาวภา ภารสำเร็จ, จิรัฐา ภูบุญอบ และวิรัตน์ พงษ์ศิริ. “การเปรียบเทียบอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษา”. วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม 2556; ฉบับพิเศษ การประชุมวิชาการ “มหาวิทยาลัยมหาสารคามครั้งที่ 9”. มหาสารคาม: มหาวิทยาลัยมหาสารคาม; 2556. 281-289.



- [35] วันเพ็ญ วิโรจน์เจริญวงศ์ และบุศรินทร์ เอี่ยมธนากุล. “การเปรียบเทียบการจำแนกข้อมูลของแบบจำลองเทคนิคต้นไม้ตัดสินใจ กฎพื้นฐาน และเคเนียร์สเนอเบอร์”. [Online]. 2012 [cited 3 Decemver 2016]; Available http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20142810143631.pdf
- [36] Abdelghani, B. & Guven E. “Predicting Breast Cancer Survivability Using Data Mining Techniques”. [Online]. 2006 [cited 3 Decemver 2016]; Available from: <http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf>
- [37] นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุ่ง มีสัจ. “การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ”. วารสารพัฒนบริหารศาสตร์ 2554; 51[3]: 187-205.



ภาคผนวก



ผลการตีพิมพ์เผยแพร่





NCCIT2017: 005

Author(s): Pattanaphong Donrat, Jaree Thongkam

Title: The Comparison of Data Mining Techniques for Predicting Student Success in Vocation Education

Your paper has been ACCEPTED for ORAL presentation at the NCCIT2017.

Information of NCCIT2017 is available on the conference web site <http://www.nccit.net> and more will come very soon. The conference information will be updated shortly to include the full technical program.

CRITICAL INFORMATION: At least one author MUST REGISTER for the conference by May 31st, 2017 with the registration fee of 5,500 Baht.

Failure to have at least one author registered to attend the conference will result in removal of the paper from the Technical Program. Also, the paper MUST be presented at the conference by one of the authors.

For your oral presentation, you must prepare PowerPoint slides to present your work to the audiences about 20 minutes (15-17 min talk and 3-5 min for questions).

We look forward to welcoming you on 6th - 7th July 2017 at NCCIT2017, Amoma Grand Hotel Bangkok, Thailand

Sincerely,

NCCIT2017 Committee



**การเปรียบเทียบเทคนิคเหมืองข้อมูลในการพยากรณ์ความสำเร็จการศึกษาของนักเรียน
ระดับประกาศนียบัตรวิชาชีพ**
**The Comparison of Data Mining Techniques for Predicting Student Success in
Vocation Education**

พัฒนพงษ์ คลรรัตน์ (Pattanaphong Donrat)¹ และจारी ทองคำ (Jaree Thongkam)²

สาขาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

¹s2s9890@gmail.com, ²jaree.thongkam@gmail.com

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษา ซึ่งแบบจำลองสามารถนำไปใช้ในการแนะนำการเข้าศึกษาต่อของนักเรียนระดับมัธยมศึกษาตอนต้นเพื่อศึกษาต่อในระดับอาชีวศึกษาเอกชน ข้อมูลถูกรวบรวมจากสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ในปีการศึกษา 2557-2558 ข้อมูลประกอบด้วย 1,266 ข้อมูล 4 เทคนิคที่เป็นที่นิยมและมีประสิทธิภาพในการสร้างแบบจำลอง คือ C4.5, Random Forest, Random Tree และ Reduced Error Pruning Tree (REP Tree) โดยใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพการพยากรณ์ของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) จากการศึกษาพบว่าแบบจำลอง C4.5 มีประสิทธิภาพในการพยากรณ์ด้วยค่าความถูกต้อง (Accuracy) สูงถึง 87.73% ค่าความแม่นยำ (Precision) 87.5% และค่าความระลึก (Recall) 87.7%

คำสำคัญ: แบบจำลองการพยากรณ์ความสำเร็จการศึกษา
ประกาศนียบัตรวิชาชีพ ต้นไม้การตัดสินใจ
เหมืองข้อมูล

Abstract

This research is to compare the efficiency of data mining techniques in modeling for predicting the Graduate in Private Vocational Education Institutes. The models could guide the study of junior high school

students to be educated at the Private vocational. Data were collected from Private Vocational Education Institutes in Kalasin Province from 2014-2015. The data consist of 1,266 records. Four efficient techniques including C4.5, Random Forest, Random Tree and Reduced Error Pruning Tree (REP Tree) are used for building the prediction models. 10-fold cross validation is applied to segment the data into training and testing sets. Moreover, accuracy, precision and recall are used as the criteria for selecting the most effective model. According to the research, the experimental results showed that C4.5 is the most efficient prediction model which has accuracy up to 87.73%, precision 87.50%, and recall 87.70%, respectively.

Keyword: Predicting Graduation Model, Vocational Education Certificate, Decision Tree, Data Mining.

1. บทนำ

ปัจจุบันประเทศไทยนั้นอยู่ใต้อาณาเขตแดนแรงงานฝีมือ กระทรวงศึกษาธิการจึงได้มีนโยบายในปีพุทธศักราช 2554 การเปลี่ยนแปลงสัดส่วนของผู้เรียนอาชีวศึกษาต่อผู้เรียนสายสามัญ จากเดิม 40:60 เป็น 60:40 ภายในปีพุทธศักราช 2561 ซึ่งเป็นปีสิ้นสุดของการปฏิรูปการศึกษาทศวรรษที่ 2 [1]

การเพิ่มจำนวนผู้สำเร็จการศึกษาระยะเวลาของหลักสูตรของอาชีวศึกษา ซึ่งหลักสูตรอาชีวศึกษาแบ่งออกได้เป็น 2 หลักสูตรคือ 1) หลักสูตรประกาศนียบัตรวิชาชีพ (ปวช.)



มีระยะเวลาการศึกษาตามหลักสูตร 3 ปี 2) หลักสูตรประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) มีระยะเวลาการศึกษาตามหลักสูตร 2 ปี ซึ่งการจัดให้ผู้เรียนสำเร็จการศึกษาตามระยะเวลาของหลักสูตรนั้น เป็นตัวบ่งชี้ตัวหนึ่งแสดงว่าสถานศึกษาได้จัดการศึกษาอย่างมีคุณภาพและประสิทธิภาพ ตามตัวบ่งชี้ของการประกันคุณภาพการศึกษาทั้งการประกันคุณภาพภายในและการประกันคุณภาพภายนอก [1]

สถานศึกษาอาชีวศึกษาเอกชน [2] ก็เป็นหน่วยงานหนึ่งในการจัดการศึกษาอาชีวศึกษาที่มีบทบาทสำคัญในการให้บริการการศึกษาทางด้านวิชาชีพเช่นเดียวกับสถานศึกษาอาชีวศึกษารัฐบาลซึ่งล้วนแต่มีบทบาทและความสำคัญอย่างยิ่งต่อการพัฒนาประเทศเพราะเป็นการศึกษาที่จัดเตรียมบุคคลให้มีอาชีพเป็นหลัก ในอนาคต และช่วยให้ผู้ที่มีอาชีพอยู่แล้วมีความก้าวหน้าในอาชีพนั้นเพิ่มขึ้น

ในการนี้ได้นำนักวิจัยที่นำเอาเหมืองข้อมูล ซึ่งเป็นกระบวนการในการค้นหาความรู้จากข้อมูล มาใช้ในการวิเคราะห์ เช่น เยาวภา การสำเร็จและคณะ [3] ได้ทำการเปรียบเทียบ 3 เทคนิค ในเหมืองข้อมูล ประกอบด้วย C4.5, Naïve Bayes และ k-Nearest Neighbour เพื่อใช้ในการจำแนกปัจจัยที่ส่งผลกระทบต่อผลการเรียนของนักศึกษา ผลการเปรียบเทียบพบว่าเทคนิค C4.5 ให้ค่าประสิทธิภาพสูงที่สุด 73.55% ธีรญา ปาลวิสุทธิ [4] พบว่าเทคนิค Random Forest มีค่าความแม่นยำร้อยละ 87.15 สามารถพยากรณ์ได้ดีกว่า J48 ID3 LMT และ CART ส่วน Kalmegh [5] พบว่า Random Tree มีค่าความถูกต้อง 100% ซึ่งมากกว่า REP Tree และ Simple Cart ในการจำแนกข่าวอินเดียบ Kittipol Wisaeng [6] พบว่า REP Tree ให้ค่าความถูกต้องที่ 92.87% ในการจำแนก UCI Repository ได้แก่ Nursery, Ins, Anneal, Shuttle_tm, Voting, Waveform และ Sick

จะเห็นว่าเทคนิค Decision Tree เป็นเทคนิคที่มีประสิทธิภาพในการจำแนก ดังนั้นงานวิจัยนี้ได้ใช้ 4 เทคนิคในกลุ่มของ Decision Tree ที่มีประสิทธิภาพในการสร้างแบบจำลองเพื่อการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ คือ C4.5, Random Forest, Random Tree และ Reduced Error Pruning Tree (REP Tree) คณะผู้วิจัยได้ใช้หลักการ 10-fold cross validation ในการ

แบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพการพยากรณ์ของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ในการแนะนำการเข้าศึกษาต่อของนักเรียนระดับมัธยมศึกษาตอนต้นเพื่อศึกษาต่อในระดับอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์

2. งานวิจัยที่เกี่ยวข้อง

2.1 สถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์

สถานศึกษาอาชีวศึกษาเอกชนสังกัดสำนักงานคณะกรรมการส่งเสริมการศึกษาเอกชน (สช.) และมีหน่วยงานต้นสังกัดคือสำนักงานเขตพื้นที่การศึกษาประถมศึกษากาฬสินธุ์ เขต 1 เขต 2 และเขต 3 ตามพื้นที่ที่สถานศึกษาดังอยู่ แต่ใช้หลักสูตรของสำนักงานคณะกรรมการการอาชีวศึกษา (สอศ.) จนมาถึงวันที่ 12 กุมภาพันธ์ 2559 เว็บไซต์ราชกิจจานุเบกษาได้เผยแพร่คำสั่งหัวหน้าคณะรักษาความสงบแห่งชาติที่ 8/2559 เรื่องการบริหารจัดการรวมสถานศึกษาอาชีวศึกษาภาครัฐและภาคเอกชน โดยให้รวมสถานศึกษาทั้งภาครัฐและเอกชนไว้ด้วยกันภายใต้สังกัดอาชีวศึกษาจังหวัดกาฬสินธุ์ (อศจ.กาฬสินธุ์)

2.2 เทคนิคในเหมืองข้อมูล

เทคนิคในเหมืองข้อมูลได้ถูกนำมาใช้ในการสร้างแบบจำลองกันอย่างแพร่หลายและมีประสิทธิภาพ เช่น C4.5, Random Forest, Random Tree และ Reduced Error Pruning Tree (REP Tree) ซึ่งสามารถอธิบายได้ดังต่อไปนี้

1. เทคนิค C4.5 [7] เป็นเทคนิคในการสร้างต้นไม้การตัดสินใจพัฒนาโดย J. Ross Quinlan ในปี 1993 โดยนำเอา ID3 มาปรับปรุงให้มีความสามารถมากขึ้นใช้วิธีการ Information Gain เพิ่มเดิมการจัดการกับข้อมูล, ตัวเลข, ข้อมูลที่ขาดไปและไม่สมบูรณ์ และการ Prune ด้วยการแทนกิ่ง (Branch) ที่ไม่ช่วยในการตัดสินใจด้วย Leaf Node ที่ตัดสินใจได้ดีกว่า การแบ่งของ Tree ในการทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 คือต้องหา Info และ Gain ออกมาก่อนซึ่งมีนักวิจัยหลายท่านได้นำเอาเทคนิคนี้มาใช้ในการพยากรณ์



2. เทคนิค Random Forest [8] เป็นเทคนิคการสุ่มเลือกใช้ข้อมูลและคุณลักษณะ Decision Tree ซึ่งถูกสร้างจากการนำข้อมูลไปสุ่มเลือกตัวอย่างแบบเลือกแล้วใส่กลับ (Sampling with Replacement) แล้วนำมาสร้างเป็น Tree ซึ่งจะมีตัวอย่างส่วนหนึ่งที่未被เลือก ซึ่งข้อมูลส่วนนี้เรียกว่า Out-of-Bag (OOB) จะถูกนำมาใช้ในการทดสอบ Decision Tree วิธีการดังกล่าวนี้เรียกว่า Bagging ผลลัพธ์ที่ได้จะอิสระจาก Decision Tree ในแต่ละต้นถูกนำมาคิดเป็นผลการโหวต ผลโหวตที่มากที่สุดจะใช้ระบุสถานะของคลาส เทคนิค Random Forest ไม่จำเป็นต้องมีข้อมูลทดสอบ เพื่อประมาณความผิดพลาดเพราะข้อมูล OOB นั้นถูกนำมาใช้ทดสอบ Decision Tree นั้นแล้ว

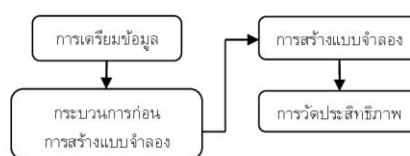
3. เทคนิค Random Tree [9] คือ เทคนิคที่ใช้ในการจำแนกหมวดหมู่เช่นเดียวกับ C4.5 โดยมีหลักการสร้าง Tree จากการสุ่ม Tree หลายๆ แบบในแต่ละโหนดแล้วเลือกมาประมวลผล โดยไม่ใช้การ Prune และเนื่องจากจำนวนของ Tree เพิ่มขึ้นอย่างรวดเร็ว ซึ่งยากแก่การแก้ปัญหา การสร้าง Tree ที่เป็นไปได้ทั้งหมด ส่วนประกอบสามารถสร้างชุดสุ่มของ Tree ออกมาจากการกระจายชุดต่างๆ ของ Tree Random Tree เป็นการสุ่มวาดที่สุ่มจากชุดของ Tree ที่เป็นไปได้ ในบริบทนี้ "สุ่ม" หมายความว่า Tree ในชุดของ Tree แต่ละมีโอกาสเท่าเทียมกันของการเป็นตัวอย่าง วิธีที่บ่งชี้ก็คือว่าการกระจายของ Tree คือ "ชุด" Tree สุ่มแบบต่อเนื่อง (CRT) เป็นแบบสุ่ม Tree จริง T_i และมีระยะในการเดินทางสั้นที่สุด Tree ชุดย่อยประกอบด้วย Tree ที่มี n ชุด แต่ละองค์ประกอบของโครงสร้างที่แท้จริงคือราก ซึ่งมีนักวิจัยจำนวนมากได้นำเทคนิค Random Tree มาใช้ในการจำแนก

4. เทคนิค Reduced Error Pruning Tree (REP Tree) [9] คือ เทคนิคที่ใช้ regression tree logic และสร้าง Tree หลายๆ ต้นที่แตกต่างกัน หลังจากนั้นก็เลือกที่ดีที่สุดจาก Tree ที่สร้างทั้งหมดมาเป็นตัวแทนของ Tree ทั้งหมด ในการตัดกิ่งใช้ค่า mean square error ในการพยากรณ์เป็นพื้นฐานการวัด Reduced Error Pruning Tree (REP Tree) เป็น Decision Tree ที่มี การเรียนรู้และสร้างแบบจำลองอย่างรวดเร็วบนพื้นฐานของ Information gain หรือ reducing the variance และตัดกิ่งโดยใช้ การลดข้อผิดพลาดในการตัดแต่ใช้เฉพาะตัวแปรที่เป็น

ตัวเลขเท่านั้น มีนักวิจัยจำนวนมากได้นำเทคนิค REP Tree มาใช้ในการจำแนก

3. ขั้นตอนวิธีการดำเนินการวิจัย

ในการทำวิจัยคณะผู้วิจัยได้มีแบ่งวิธีการดำเนินการวิจัยออกเป็น 4 ขั้นตอนหลัก 1) การเตรียมข้อมูล 2) กระบวนการก่อนการสร้างแบบจำลอง 3) การสร้างแบบจำลอง และ 4) การวัดประสิทธิภาพแบบจำลอง ดังภาพที่ 1



ภาพที่ 1 The process of data mining

3.1 การเตรียมข้อมูล

การเตรียมข้อมูลของการศึกษาค้นคว้าอิสระครั้งนี้ ได้ข้อมูลมาจากการสำรวจข้อมูลจำนวนนักเรียนจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ประจำปีการศึกษา 2557-2558 จำนวน 1,266 ข้อมูล จากสถานศึกษา 12 แห่ง ดังตารางที่ 1

ตารางที่ 1 จำนวนนักเรียนประจำปีการศึกษา 2557-2558

รายชื่อสถานศึกษา	2557	2558
1. วิทยาลัยเทคโนโลยีพนิชการกาฬสินธุ์	176	197
2. วิทยาลัยเทคโนโลยีกมลไธย	38	35
3. วิทยาลัยเทคโนโลยีเพิ่มพูนบริหารธุรกิจ	21	19
4. วิทยาลัยอาชีวศึกษาลำปาว	13	14
5. วิทยาลัยเทคโนโลยีสหพันธ์	15	16
6. วิทยาลัยเทคโนโลยีสามชัย	9	11
7. วิทยาลัยอาชีวศึกษาไทย-เทคนิกกาฬสินธุ์	12	15
8. วิทยาลัยอาชีวศึกษาไทยเทคโนธานี	23	12
9. วิทยาลัยเทคโนโลยีสมเด็จพนิชการ	132	268
10. วิทยาลัยเทคโนโลยีพัฒนบัณฑิต	18	10
11. วิทยาลัยอาชีวศึกษาไทยเทคโนเอเชีย	130	82
12. วิทยาลัยเทคโนโลยีนิพนธ์	-	-
รวม	587	679
รวมทั้งสิ้น	1,266	



3.2 การทำกระบวนการก่อนการสร้างแบบจำลอง

ในขั้นตอนนี้ผู้วิจัยได้ทำการแปลงข้อมูลและทำการวิเคราะห์ปัจจัย

1. การแปลงข้อมูล (Data transformation) จากข้อมูลที่ได้จากการสำรวจข้อมูลจำนวนนักเรียนจากสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ ประจำปีการศึกษา 2557-2558 จำนวน 12 แห่ง มีทั้งหมด 20 ตัวแปร ดังตารางที่ 2

ตารางที่ 2 ตัวแปรในงานวิจัย

ชื่อตัวแปร	คำอธิบายตัวแปร	ชนิดของตัวแปร
Sex	เพศ	Nominal
Age	อายุ	Numeric
Number of siblings	จำนวนพี่น้อง	Numeric
Parent Status	สถานภาพครอบครัว	Nominal
Father Career	อาชีพบิดา	Nominal
Mother Career	อาชีพมารดา	Nominal
Village	หมู่บ้าน	Nominal
Tumbol	ตำบล	Nominal
Amphoe	อำเภอ	Nominal
Province	จังหวัด	Nominal
Thai	เกรดเฉลี่ยกลุ่มภาษาไทย	Numeric
Math	เกรดเฉลี่ยกลุ่มคณิตศาสตร์	Numeric
Science	เกรดเฉลี่ยกลุ่มวิทยาศาสตร์	Numeric
Social	เกรดเฉลี่ยกลุ่มสังคมศึกษา ศาสนา และวัฒนธรรม	Numeric
Health	เกรดเฉลี่ยกลุ่มสุขภาพ พลศึกษา	Numeric
Art	เกรดเฉลี่ยกลุ่มศิลปะ	Numeric
Career and Technology	เกรดเฉลี่ยกลุ่มการงานอาชีพและเทคโนโลยี	Numeric
Foreign language	เกรดเฉลี่ยกลุ่มภาษาต่างประเทศ	Numeric
Type of course	ประเภทวิชาที่จบ	Nominal
Graduate	จบการศึกษา	Nominal

2. การวิเคราะห์ปัจจัยเป็นการคัดเลือกแอตทริบิวต์ที่สามารถเป็นตัวแทนของกลุ่มแอตทริบิวต์เพื่อลดจำนวนแอตทริบิวต์ในการพยากรณ์ซึ่งงานวิจัยนี้ได้นำเทคนิค Gain Ratio Attribute Eval มาใช้ในการวิเคราะห์ปัจจัยผลการวิเคราะห์ทำการตัดแอตทริบิวต์ที่มีผลต่อการพยากรณ์น้อยออกจำนวน 7 แอตทริบิวต์คือ Tumbol, Amphoe, Age, Career and Technology, Thai, Number of siblings และ Province คงเหลือ 13 แอตทริบิวต์ ดังนี้ Father Career, Mother Career, Village, Parent Status, Sex, Type of course, Science, Math, Social, Art, Foreign language, Health และ Graduate(Class)

3.3 การสร้างแบบจำลอง

การสร้างแบบจำลองพยากรณ์การสำเร็จการศึกษาของผู้ที่สนใจเข้าศึกษาต่อสถานศึกษาอาชีวศึกษาเอกชนในจังหวัดกาฬสินธุ์ เพื่อช่วยในการตัดสินใจว่าเมื่อเข้าศึกษาต่อแล้วจะสามารถสำเร็จการศึกษาได้หรือไม่ ด้วยตัวแปรทั้ง 20 ตัวแปรที่ได้จากการสำรวจข้อมูลจำนวนนักเรียนในปีการศึกษา 2557-2558 จำนวน 1,266 ข้อมูล แบ่งเป็น 2 คลาส คือ สำเร็จการศึกษาจำนวน 812 ข้อมูลและไม่สำเร็จการศึกษาจำนวน 454 ข้อมูล ด้วยการทำให้เหมือนข้อมูลและเทคนิคที่นำมาใช้ในการสร้างแบบจำลองมีจำนวน 4 เทคนิคคือ

1. เทคนิค C4.5
2. เทคนิค Random Forest
3. เทคนิค Random Tree
4. เทคนิค REP Tree

3.4 การวัดประสิทธิภาพของแบบจำลอง

ในการวัดประสิทธิภาพของแบบจำลองนั้น ได้มีการใช้เทคนิคแบบ 10-fold cross validation โดยจะทำการแบ่งข้อมูลออกเป็น 10 ชุดเท่าๆกัน จากนั้นจะทำการทดสอบทั้งหมด 10 รอบ โดยในแต่ละรอบจะใช้ข้อมูล 1 ชุดเป็นชุดทดสอบและอีก 9 ชุดที่เหลือเป็นชุดฝึกสอน ในรอบต่อไปก็ใช้ชุดข้อมูลถัดไปเป็นชุดทดสอบจนครบทั้ง 10 ชุดข้อมูล ซึ่งข้อมูลทั้งหมดจำนวน 1266 ข้อมูล แบ่งออกเป็น 10 ชุดข้อมูล ซึ่งคิดเป็นอัตราข้อมูลทดสอบต่อข้อมูลฝึก เป็นอัตราส่วน 10:90



การวิเคราะห์ประสิทธิภาพ คือ การวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล โดยค่าของผลลัพธ์ที่ได้จากการจำแนกคือ ค่า True Positive (TP) ค่า True Negative (TN) ค่า False Positive (FP) ค่า False Negative (FN) และค่าความถูกต้อง (Accuracy) คือ ค่าที่แบบจำลองสามารถพยากรณ์ความสำเร็จการศึกษาหรือไม่สำเร็จ ได้ถูกต้อง คือนักเรียนทั้งหมด ดังสมการที่ 1

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

ค่าความแม่นยำ (Precision) คือ ค่าที่แบบจำลองสามารถพยากรณ์นักเรียนที่สำเร็จการศึกษาได้ถูกต้อง ต่อจำนวนนักเรียนที่พยากรณ์สำเร็จการศึกษาทั้งหมด ดังสมการที่ 2

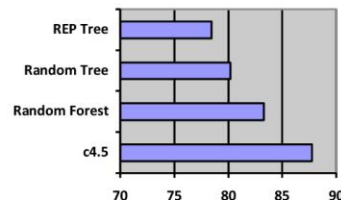
$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

และค่าความระลึก (Recall) คือ ค่าที่แบบจำลองสามารถพยากรณ์นักเรียนที่สำเร็จการศึกษาได้ถูกต้อง ต่อจำนวนนักเรียนที่สำเร็จการศึกษจริงทั้งหมด ดังสมการที่ 3

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

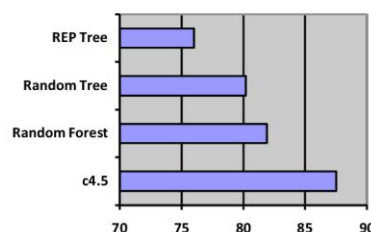
4. ผลการดำเนินการวิจัย

ในการศึกษาประสิทธิภาพของแบบจำลองนี้คณะผู้วิจัยได้นำเอาโปรแกรม WEKA เวอร์ชัน 3.9.1 มาเป็นเครื่องมือมาใช้ในการสร้างแบบจำลองด้วยเทคนิค C4.5, Random Forest, Random Tree และ REP Tree โดยใช้ค่าพารามิเตอร์เริ่มต้นของโปรแกรม WEKA ในงานวิจัยนี้คณะผู้วิจัยได้นำเอา 10 fold cross-validation มาใช้ในการแบ่งข้อมูลออกเป็นชุดเรียนรู้ และชุดทดสอบ และได้วัดประสิทธิภาพของแบบจำลองด้วย Accuracy, Precision และ Recall ผลการทดลองสามารถแสดงได้ดัง ภาพที่ 2, ภาพที่ 3 และภาพที่ 4



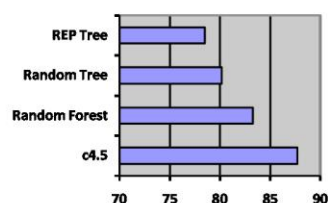
ภาพที่ 2 ผลการวัดประสิทธิภาพของแบบจำลองด้วย Accuracy

ภาพที่ 2 แสดงให้เห็นค่า Accuracy ของแบบจำลองได้แก่ เทคนิค C4.5 ให้ค่า Accuracy มากที่สุด 87.73% ต่อมา เทคนิค Random Forest ให้ค่า Accuracy 83.29% เทคนิค Random Tree ให้ค่า Accuracy 80.20% เทคนิค REP Tree ให้ค่า Accuracy 78.45% ตามลำดับ



ภาพที่ 3 ผลการวัดประสิทธิภาพของแบบจำลองด้วย Precision

ภาพที่ 3 แสดงให้เห็นค่า Precision ของแบบจำลองได้แก่ เทคนิค C4.5 ให้ค่า Precision มากที่สุด 87.5% ต่อมา เทคนิค Random Forest ให้ค่า Precision 81.9% เทคนิค Random Tree ให้ค่า Precision 80.2% เทคนิค REP Tree ให้ค่า Precision 76.0% ตามลำดับ



ภาพที่ 4 ผลการวัดประสิทธิภาพของแบบจำลองด้วย Recall



ภาพที่ 4 แสดงให้เห็นค่า Recall ของแบบจำลองที่ได้แก่เทคนิค C4.5 ให้ค่า Recall มากที่สุด 87.7% ต่อมาเทคนิค Random Forest ให้ค่า Recall 83.3% เทคนิค Random Tree ให้ค่า Recall 80.2% เทคนิค REP Tree ให้ค่า Recall 78.5% ตามลำดับ

5. สรุป

ในการศึกษาและเปรียบเทียบแบบจำลองที่เหมาะสมสำหรับการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ โดยแบบจำลองที่ใช้ในการเปรียบเทียบ 4 เทคนิค ได้แก่ C4.5, Random Forest, Random Tree และ REP Tree ผลการทดลองพบว่าเทคนิค C4.5 มีความเหมาะสมมากที่สุดในการพยากรณ์การสำเร็จการศึกษาในสถานศึกษาอาชีวศึกษาเอกชนจังหวัดกาฬสินธุ์ โดยการวัดประสิทธิภาพการทำงานของแบบจำลองด้วยค่า Accuracy ได้ 87.73% ค่า Precision ได้ 87.5% และค่า Recall ได้ 87.7% เพราะ C4.5 สามารถจำแนกตัวแปรชนิด nominal ได้อย่างเหมาะสม

เอกสารอ้างอิง

- [1] ข่าวการศึกษา. การเพิ่มสัดส่วนสายอาชีพเป็น 60% สายสามัญเหลือ 40%. Available: [http://www.unigang.com/Article/5170,\(2559,28 สิงหาคม\)](http://www.unigang.com/Article/5170,(2559,28 สิงหาคม)).
- [2] วันจักร ทัพย์มาศ, "ปัจจัยที่มีผลต่อการตัดสินใจเข้าศึกษาต่อวิทยาลัยอาชีวศึกษาเอกชนของนักเรียนในเขตภาคใต้ตอนบน," *วารสารศึกษาศาสตร์*, vol. 24, กุมภาพันธ์-สิงหาคม 2556.
- [3] เขียวภา การสำเร็จ, จิรฎฐา ญูญอมและวิรัตน์ พงษ์ศิริ, "การเปรียบเทียบอัลกอริทึมเหมืองข้อมูลเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อระดับผลการเรียนของนักศึกษา," *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม*, 2556.
- [4] ภรณ์ยา ปาลวิสุทธิ์ "การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่มีสมดุล โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูล การเป็นโรคติดเชื้อเอดส์ " *วารสารเทคโนโลยีสารสนเทศ*, vol. 12, pp. 54-63, 2559.
- [5] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, pp. 438-446, 2015.
- [6] K. Wisang, "A Comparison of Decision Tree Algorithms For UCI Repository Classification," *International Journal of Engineering Trends and Technology*, vol. 4, pp. 3393-3397, 2013.
- [7] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second Edition ed.: Morgan Kaufmann, 2006.
- [8] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [9] Weka. Available: [http://www.cs.waikato.ac.nz/ml/index.html,\(2016,3 December\)](http://www.cs.waikato.ac.nz/ml/index.html,(2016,3 December)).





ประวัติย่อผู้ศึกษา



ประวัติย่อผู้ศึกษา

ชื่อ นามสกุล	นายพัฒนพงษ์ ดลรัตน์
วัน เดือน ปีเกิด	วันที่ 23 ธันวาคม พ.ศ. 2526
จังหวัด และประเทศที่เกิด	อำเภอยางตลาด จังหวัดกาฬสินธุ์ ประเทศไทย
ประวัติการศึกษา	พ.ศ. 2542 มัธยมศึกษาตอนต้น โรงเรียนยางตลาดวิทยาคาร อำเภอยางตลาด จังหวัดกาฬสินธุ์ พ.ศ. 2545 ประกาศนียบัตรวิชาชีพ วิทยาลัยเทคนิคกาฬสินธุ์ อำเภอเมือง จังหวัดกาฬสินธุ์ พ.ศ. 2547 ประกาศนียบัตรวิชาชีพชั้นสูง วิทยาลัยเทคนิคกาฬสินธุ์ อำเภอเมือง จังหวัดกาฬสินธุ์ พ.ศ. 2550 ปริญญาวิศวกรรมศาสตรบัณฑิต (วศ.บ.) สาขาวิชาวิศวกรรมอิเล็กทรอนิกส์ มหาวิทยาลัยภาคตะวันออกเฉียงเหนือ พ.ศ. 2560 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม
ตำแหน่ง สถานที่งาน	ช่างเครื่องกล องค์การส่งเสริมกิจการโคนมแห่งประเทศไทย ตำบลท่าพระ อำเภอเมือง จังหวัดขอนแก่น 40260
ที่อยู่ที่สามารถติดต่อได้	บ้านเลขที่ 476 หมู่ 4 ตำบลยางตลาด อำเภอยางตลาด จังหวัดกาฬสินธุ์ 46120

