โมเดลเพื่อการพยากรณ์สถานภาพทางการศึกษาของนักศึกษา

A model for forecasting educational status of students

พิชัย ระเวงวัน (Pichai Rawengwan)* ดร.พุธษดี ศิริแสงตระกูล (Dr.Pusadee Seresangtakul)**

บทคัดย่อ

งานวิจัยนี้มุ่งเน้นศึกษาปัจจัยที่มีผลต่อสถานภาพการศึกษาของนักศึกษา เพื่อทำการสร้างโมเดลเพื่อพยากรณ์สถานภาพทางการศึกษา ของนักศึกษาในสถาบันอุดมศึกษาเอกชนที่ตั้งในภาคตะวันออกเฉียงเหนือ โดยการประยุกต์ใช้วิธีการเหมืองข้อมูล ในการศึกษาผู้วิจัยได้ทำการ เก็บรวบรวมข้อมูลปัจจัยที่คาดว่าจะมีผลต่อสถานภาพทางการศึกษาจากงานวิจัยที่เกี่ยวข้องและได้รวมรวมข้อมูลของนักศึกษาระดับปริญญาตรีที่เข้าศึกษาระหว่างปีการศึกษา 2551 – 2553 จำนวนทั้งสิ้น 2,272 คน และวิเคราะห์ความสัมพันธ์ของปัจจัยที่มีผลต่อสถานภาพของนักศึกษาด้วยวิธีการวิเคราะห์สหสัมพันธ์ ผลจากศึกษาพบว่าปัจจัยที่มีผลต่อสถานภาพของนักศึกษามี 9 ปัจจัย ได้แก่ การกู้ยืมในระหว่างเรียน รายได้ผู้ปกครอง สถานภาพสมรสของผู้ปกครอง เพศ ประเภทโรงเรียนเดิมที่จบมา วุฒิการศึกษาเดิม กลุ่มโรงเรียนเดิมที่จบ อาชีพของมารดา และที่อยู่นักศึกษาระหว่างเรียนปัจจัยดังกล่าวจะถูกนำมาใช้ในการสร้างโมเดลเพื่อทำนายสถานภาพของนักศึกษา ด้วยเทคนิคต้นไม้ตัดสินใจ เบย์เซี่ยนบีลิฟเน็ตเวิร์ค และโลจิสติกรีเกรสชั่น ผลการศึกษาพบว่าค่าความถูกต้องมีค่า 82.85%, 78.98%และ 78.50% ด้วยเทคนิคต้นไม้ตัดสินใจ จิสติกรีเกรสชั่น และการเรียนรู้แบบเบย์เซี่ยนเน็ตเวิร์ค ตามลำดับ ซึ่งพบว่าโมเดลที่สร้างด้วยวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องดีที่สุด

ABSTRACT

This study emphasized on the study of factors affect student's status in order to develop a model to predict studying status in a Private University in the Northeast by applying data mining. In this study, we collected the data which might affect to studying status of 2,272 bachelor students enrolled during 2008-2010.We analyzed variables correlation by correlation analysis. The study results were nine factors consist of loan, income, marital, sex, types of school, educational background, school level, mother occupation and student address ,which would use to develop the model by decision tree, bayesian belief network, logistic regression. The result revealed that the accuracy were 82.85 %, 78.98 % and 78.50% for decision tree, logistic regression, bayesian belief network, respectively.

คำสำคัญ: ต้นไม้ตัดสินใจ เบย์เซี่ยนบีลิฟเน็ตเวิร์ค โลจิสติกรีเกรสชั่น

Keywords: Decision tree, Bayesian belief network, Logistic Regression

** นักศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

*** ผู้ช่วยศาสตราจารย์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

บทนำ

การศึกษาในปัจจุบันมีการเปลี่ยนแปลงพัฒนาอย่างต่อเนื่องภาครัฐเองก็ให้ความสำคัญทางด้านการศึกษามากขึ้น จากการดำเนินการด้านการศึกษาของไทย (สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานกระทรวงศึกษาธิการ, 2552) พบว่าในด้านการศึกษามีการสนับสนุนนักศึกษาทุกระดับทั้งการสนับสนุนการเรียนภาคบังคับ รวมถึงการจัดตั้งกองทุนให้กู้ยืมเพื่อการศึกษา (ราชกิจจานุเบกษา,2541) ทั้งสถาบันการศึกษาของรัฐและเอกชน แต่ยังมีนักเรียนนักศึกษาจำนวนมากที่มีการออกจากระบบการศึกษากลางคัน เป็นเหตุให้ไม่สำเร็จการศึกษาหรือสำเร็จการศึกษาช้า โดยมีปัจจัยสาเหตุก็แตกต่างกันไป เป็นเหตุให้เกิดการสูญเสียทั้งเงินและเวลา นำไปสู่การสูญเปล่าในการลงทุนด้านการศึกษา และยังเป็นสาเหตุที่นำไปสู่การขาดกำลังคนต่อภาคอุตสาหกรรมในอนาคต

จากการศึกษาพบว่ามีงานวิจัยหลายงานทั้งในและต่างประเทศ (ไก้รุ่งเฮงพระพรหม, 2557), (Tripti, 2014) ที่ได้ทำการศึกษาปัจจัยที่ผลต่อผลการเรียนของนักศึกษา หลายงานเป็นการศึกษาในสถาบันอุดมศึกษาของรัฐ (อ้อมใจ, 2557) สำหรับงานวิจัยนี้มุ่งเน้นที่ปัจจัยที่มีผลต่อสถานภาพของนักศึกษาของมหาวิทยาลัยเอกชนแห่งหนึ่งซึ่งเปิดสอนในภาคตะวันออกเฉียงเหนือ จากข้อมูลของนักศึกษาที่ศึกษา ระหว่างปี 2551-2557 พบว่ามีอัตราการตกออกอยู่ที่ 19.16 % ซึ่งสาขาคอมพิวเตอร์ธุรกิจมีอัตราการออกค่อนข้างสูงคิดเป็น 55.21 % (สำนักทะเบียนและประมวล, 2557) ของจำนวนนักศึกษาทั้งหมดและมีอัตราการออกเพิ่มมากขึ้นทุกๆ ปี ซึ่งปัญหาที่ทุกคนให้ความสำคัญอย่างมากภายในมหาวิทยาลัย หากทราบปัจจัยที่มีผลต่อสถานภาพของนักศึกษาแล้ว ผู้เกี่ยวข้องน่าจะใช้เป็นข้อมูลในการดูแล ติดตามช่วยเหลือเพื่อให้นักศึกษาสามารถสำเร็จการศึกษาได้ต่อไป

ดังนั้นหากมีระบบที่ใช้ปัจจัยที่เกี่ยวข้องมาช่วยในการคาดการณ์สถานภาพทางการศึกษา ก็จะมีประโยชน์ในการดูแล ติดตามช่วยเหลือนักศึกษาได้ การวิเคราะห์และข้อมูลด้านการศึกษา ส่วนใหญ่นิยมใช้ ต้นไม้ตัดสินใจ (Tripti, 2014), (Bunkar,2012) เนื่องจากมีความสามารถในการจำแนกข้อมูลได้ดี (Al-Radaideh,2006) โดยวิเคราะห์ข้อมูลปัจจัยที่มีผลต่อการศึกษา พบว่าต้นไม้ตัดสินใจมีความถูกต้องกว่าวิธีอื่น แต่ก็มีงานวิจัย (อ้อมใจ,2557) โมเดลเพื่อการพยากรณ์นักศึกษาที่มีผลการเรียนต่ำกว่าเกณฑ์ โดยใช้ เบย์เซี่ยนบีลิฟเน็ตเวิร์ค โครงข่ายประสาทเทียม และ โลจิสติกรีเกรสชั่น ผลการศึกษาพบว่า เบเซี่ยนบีลิฟเน็ตเวิร์ค ให้ผลการทำนายดีกว่าวิธีอื่นและมีงานวิจัย (ปฐมา, 2553) มีการศึกษาผลสัมฤทธิ์ทางการเรียนระดับมัธยมศึกษาตอนปลาย ผลการสอบคัดเลือกเข้ามหาวิทยาลัยโดยใช้การทำนาย โลจิสติกรีเกรสชั่น ปัจจัยมีความสัมพันธ์กับผลสัมฤทธิ์ทางการเรียนอย่างมีนัยสำคัญทางสถิติ งานวิจัยนี้จึงได้นำเอาวิธี ต้นไม้ตัดสินใจ เบเซี่ยนบีลิฟเน็ตเวิร์ค มาประยุกต์ใช้ในการสร้างโมเดลการพยากรณ์สถานภาพทางการศึกษา พร้อมทั้งได้นำวิธีโลจิสติกรีเกรสชั่นเข้ามาเปรียบเทียบประสิทธิภาพในการทำนาย

วัตถุประสงค์การวิจัย

1. ศึกษาคุณลักษณะและปัจจัยที่มีผลต่อการเรียนของนักศึกษา
2. เพื่อสร้างโมเดลที่ใช้ในการพยากรณ์สถานภาพทางการศึกษาของนักศึกษา

วิธีการวิจัย

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ต้นไม้ตัดสินใจ (Decision tree)

ต้นไม้ตัดสินใจ (Han,2006) เป็นวิธีที่ใช้ในการจำแนกกฎ โดยลักษณะการทำงานเหมือนโครงสร้างต้นไม้ ที่แต่ละโหนด (Node) แสดงคุณลักษณะ (Attribute) โดยโหนดแรกสุดจะเป็นรากของของต้นไม้ (Root node)

ประกอบด้วยกิ่งของต้นไม้ (branch) และใบ (leaf)การทำงานต้นไม้ตัดสินใจจะเริ่มจากการนำข้อมูลมาแบ่งเป็นโครงสร้างต้นไม้นั้นจะเป็นโครงสร้างที่เป็นกฎต่างๆ ส่วนใหญ่ไม่รองรับแบบต่อเนื่อง (Continuous Data) จะต้องมีการแบ่งให้เป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) (Qui,1998) ในการสร้างอัลกอริทึมกฎการตัดสินใจแบบต้นไม้สำหรับจัดแบ่งกลุ่มข้อมูลและมีการใช้หลักการหาคลาสที่มีความสัมพันธ์ของคลาสนั้นๆ ตามสมการที่ 1

$$\mathrm{Gain}(S,A) = E(S) - \sum_{v=\mathrm{value}(A)} \frac{|S_v|}{|S|} E(S_v) \qquad (1)$$

เมื่อ S คือ ตัวอย่างที่ประกอบด้วยชุดข้อมูลตัวแปรต้นและตัวแปรตาม

A คือตัวแปรต้นที่พิจารณา

E คือ เอนโทปีของตัวอย่าง

value = *(A)* คือเซตของค่า A

S_v คือตัวอย่างที่ A มีค่า V ทั้งหมด

เบย์เชี่ยนบีลิฟเน็ตเวิร์ค (Bayesian belief network)

เบย์เชี่ยนบีลิฟเน็ตเวิร์คหรือ Baye net เป็นวิธีการวิเคราะห์สาเหตุของปัญหา โดยโหนดในเครือข่ายเบย์เชี่ยนบีลิฟเน็ตเวิร์ค แต่ละตัวจะถูกสุ่มและมีความสัมพันธ์กันกับการคำนวณความน่าจะเป็น นอกจากนี้ยังสามารถใช้ความน่าจะเป็นตรวจสอบเหตุการณ์ที่จะเกิดขึ้นก่อนหรือหลังเหตุการณ์ได้ และอธิบายความสัมพันธ์ระหว่างตัวแปรและแสดงออกมาในรูปแบบโมเดล สามารถอธิบายความไม่ขึ้นต่อกันอย่างมีเงื่อนไข (Condition Independent)ระหว่างตัวแปร หรือปัจจัยเป็นบริบทของข่ายงานความเชื่อเบย์เพื่อกระบวนการเรียนรู้มีประสิทธิภาพ โดยสามารถที่จะใส่ความรู้ก่อนในข่ายงานความเชื่อเบย์ให้อยู่ในรูปของโครงสร้างข่ายงานและตารางความน่าจะเป็นมีเงื่อนไขโหนดทั้งหมดในเบย์เชี่ยนบีลิฟเน็ตเวิร์ค แต่ละโหนดจะแทนด้วยตัวแปร ต่างๆ ที่เกี่ยวข้องกับเหตุการณ์หรือข้อมูลที่สนใจ (Mehran,1996) ดังสมการที่ 2

$$P(Y_i|X) = \frac{P(Y_i \cap X)}{\sum_{i=1}^{n} P(X|Y_i)P(Y_i)} \qquad (2)$$

โดย $P(Y_i|X)$ คือ ความน่าจะเป็นที่เหตุการณ์ Y_i จะเกิดขึ้นถ้าเหตุการณ์ X เกิดขึ้นแล้ว

$P(Y_i \cap X)$ คือ ความน่าจะเป็นของY_iเมื่อเกิดเหตุการณ์ X

$P(X|Y_i)$ คือ ความน่าจะเป็นที่เหตุการณ์ X จะเกิด ถ้าเหตุการณ์ Y_iเกิดขึ้นแล้ว

$P(Y_i)$ คือ ความน่าจะเป็นที่เหตุการณ์ Y_iจะเกิด

n คือจำนวนเหตุการณ์ทั้งหมด

การถดถอยโลจิสติค (Logistic Regression)

เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุ (David,2009) เป็นการทำนายค่าความน่าจะเป็นของการเกิดเหตุการณ์หรือไม่เกิดเหตุการณ์ ความสัมพันธ์จะอยู่ในรูปแบบของสมการเส้นถดถอย (Regression Equation) มีตัวแปรตามที่มีค่าเพียง 2 ค่าคือ เป็น 0 และ1ตัวแปรอิสระอาจจะมีตัวเดียวหรือหลายตัวก็ได้ค่าพยากรณ์หรือค่าตัวแปรจะมีค่าอยู่ที่ 0 กับ 1 โดยที่ถ้าค่าความน่าจะเป็น >0.5 ถือว่าค่าพยากรณ์คือ 1 ถ้าค่าความน่าจะเป็น <0.5 ถือว่าค่าพยากรณ์คือ 0 ดังสมการที่ 3

$$L(Y) = b_0 + b_1X + \cdots + b_kx_k \quad (3)$$

โดย L(Y) คือ ค่าพยากรณ์ของตัวแปรตาม

Y คือ ตัวแปรตาม

b_0 คือ ค่าคงที่

b_1 คือ ค่าสัมประสิทธิ์การถดถอยสำหรับตัวแปร

b_k คือ ค่าสัมประสิทธิ์การถดถอยสำหรับตัวแปร

X คือ ค่าของตัวแปรอิสระ

การวิเคราะห์สหสัมพันธ์ (Correlation)

เป็นการวิเคราะห์ความสัมพันธ์ที่ไม่มีการกำหนดตัวแปรใดเป็นตัวแปรอิสระหรือตัวแปรตาม ความสัมพันธ์ที่พบจึงเป็นเพียงสิ่งที่บ่งบอกว่าตัวแปรทั้งสอง (Rodgers, 1988) มีการแปรเปลี่ยนสัมพันธ์ดังสมการที่ 4

$$\square_{\square\square} = \frac{\square\sum\square\square-(\sum\square)(\sum\square)}{\sqrt{\square\sum_{\square}2-(\sum\square)^2}\sqrt{\square\sum\square^2-(\sum\square)^2}} \quad (4)$$

เมื่อ X คือ ตัวแปรอิสระ Y คือ ตัวแปรตามและ N คือ ขนาดของกลุ่มตัวอย่าง

ระดับหรือขนาดของความสัมพันธ์จะใช้ตัวเลขของค่าสัมประสิทธิ์สหสัมพันธ์หากค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเข้าใกล้ -1 หมายถึง การมีความสัมพันธ์กันไปในทิศทางตรงกันข้าม หากมีค่าเข้าใกล้ 1 หมายถึง การมีความสัมพันธ์กันไปในทิศทางเดียวกัน มีความสัมพันธ์กันในระดับสูง แต่หากมีค่าเข้าใกล้ 0 แสดงถึงการมีความสัมพันธ์กันในระดับน้อย หรือไม่มีเลย สำหรับการพิจารณาค่าสหสัมพันธ์ใช้เกณฑ์ ดังตารางที่ 1

ตารางที่ 1 ระดับของความสัมพันธ์

ค่า r	ระดับความสัมพันธ์
.90 - 1.00	มีความสัมพันธ์กันสูงมาก
.70 - .90	มีความสัมพันธ์กันในระดับสูง
.50 - .70	มีความสัมพันธ์กันในระดับปานกลาง
.30 - .50	มีความสัมพันธ์กันในระดับต่ำ
.00 - .30	มีความสัมพันธ์กันในระดับต่ำมาก

งานวิจัยที่เกี่ยวข้อง

Tripti Mishra และคณะ (Tripti, 2014)ได้ใช้เหมืองข้อมูลในการทำนายประสิทธิภาพของนักศึกษาที่จะไม่สำเร็จการศึกษา ของศึกษา ในประเทศอินเดียโดยเก็บข้อมูล เพศ การศึกษาของพ่อแม่ อาชีพพ่อแม่ รายได้ครอบครัว การกู้ยืมเงิน เกรดที่จบการศึกษา ประเภทการศึกษาที่จบมา เวลาในการเรียน ในการศึกษาใช้เทคนิคต้นไม้ตัดสินใจ J48 และการ Random Tree ซึ่งในการสร้างโมเดลเพื่อเปรียบเทียบรูปแบบ พร้อมกับการหาปัจจัยที่มีผลต่อการเรียนผลการทดลองพบว่าRandom tree มีความถูกต้องของการทำนายมากกว่า J 48 โดยมีค่าความถูกต้อง 94.41% ส่วน J48 มีค่าความถูกต้องอยู่ที่ 88.37%

Kamal Bunkar และคณะ (Bunkar,2012) นำเสนอโมเดลทำนายเกรดของนักศึกษาโดยใช้การจัดหมวดหมู่ในการทำเหมืองข้อมูล ในการศึกษาผู้วิจัยใช้คุณลักษณะประกอบด้วย คะแนนแต่ละรายวิชา ประวัติการศึกษา พฤติกรรม

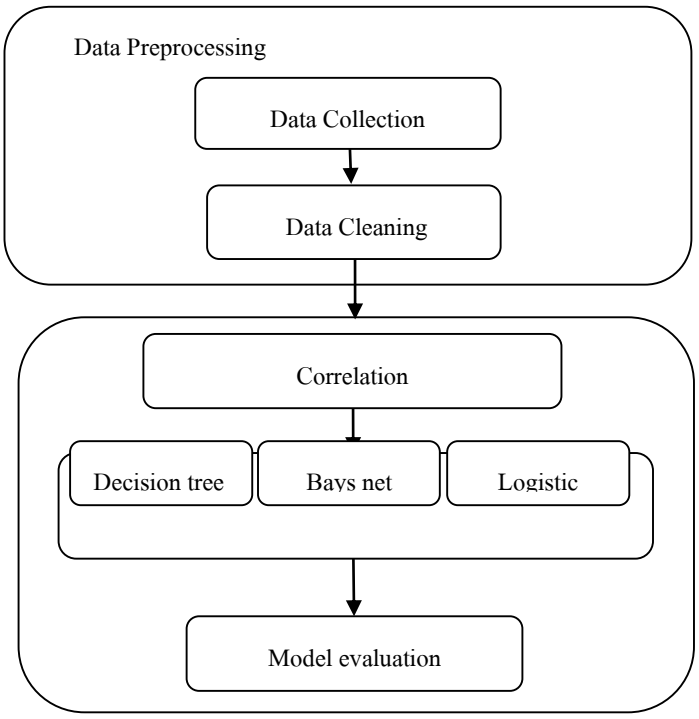
การเรียน อาจส่งผลกระทบการเรียนของนักศึกษาของมหาวิทยาลัย Ujjain โดยใช้เทคนิคต้นไม้การตัดสินใจประกอบด้วย C4.5, ID3, CART ผลการทดลองพบว่านักศึกษาที่มีเกณฑ์การเรียนที่ต้องปรับปรุงค่าความถูกต้องเท่ากับ 0.22% ส่วนเกณฑ์ที่ผ่าน 0.856% ผลการศึกษาถูกนำไปใช้เพื่อปรับปรุงการเรียนการสอนในระดับมหาวิทยาลัย

ไก้รุ่งเฮงพระพรหมและคณะ (ไก้รุ่งเฮงพระพรหม, 2557) ได้นำเสนอการค้นหาการค้นหาความรู้คุณลักษณะของนักศึกษาที่มีต่อผลสัมฤทธิ์ทางการเรียนในกลุ่มโปรแกรมคอมพิวเตอร์มหาวิทยาลัยราชภัฏนครปฐมด้วยเทคนิคเหมืองข้อมูลโดยใช้ข้อมูล เพศ วุฒิการศึกษาก่อนเข้า ข้อมูลเกี่ยวกับครอบครัว ข้อมูลสุขภาพ ข้อมูลเจตคติต่อรายวิชาโดยใช้เทคนิคต้นไม้ตัดสินใจและข่ายงานความเชื่อแบบเบย์เพื่อเปรียบเทียบการทำงานในการวัดประสิทธิภาพของคุณลักษณะในการจำแนกประเภทข้อมูลจาการทดลองต้นไม้ตัดสินใจ ให้ค่าความถูกต้องในการจำแนกข้อมูลได้ดีกว่าเมื่อเปรียบเทียบกับข่ายความเชื่อเบย์

อ้อมใจ เขาทอง และ พุธษดี ศิริแสงตระกูล (อ้อมใจ,2557) นำเสนอโมเดลการพยากรณ์นักศึกษาที่มีผลการเรียนต่ำกว่าเกณฑ์ โดยเก็บรวบรวมข้อมูลเป็นแบบสอบถาม 72 ปัจจัยและใช้การวิเคราะห์ถดถอยแบบพหุคูณแบบขั้นบันได้พบว่ามี 10 ปัจจัยที่มีผลต่อผลการเรียนของนักศึกษา คือ เพศ สถานะของผู้ปกครอง อาชีพบิดา การทำ กิจกรรมในห้องเรียน การอ่านหนังสือ การคบเพื่อนต่างเพศ การดื่มแอลกอฮอร์ ครอบครัวไม่สนับสนุน หลักสูตรการศึกษาและวิธีสอนของอาจารย์ที่เร็วเกินไป และสร้างโมเดลด้วยเทคนิควิธี เบย์เซี่ยนบีลิฟเน็ตเวิร์ค โครงข่ายประสาทเทียมและโลจิสติกรีเกรสชั่น ผลการทดลองพบว่า การพยากรณ์ด้วยโมเดลเบย์เซี่ยนบีลิฟเน็ตเวิร์ค มีความถูกต้องมากกว่าโครงข่ายประสาทเทียม และโลจิสติกรีเกรสชั่น

วิธีการดำเนินงาน

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาคุณลักษณะและปัจจัยที่ส่งผลกระทบต่อผลการเรียนและสร้างโมเดลของเทคนิคต้นไม้ในการตัดสินใจ เทคนิควิธีเบย์เซี่ยนเน็ตเวิร์ค และเทคนิควิธีการถดถอยโลจิสติคในการทำนายสถานะภาพทางการศึกษาของนักศึกษา โดยมีขั้นตอนในการทำงานดังรูปที่ 1



ภาพที่ 1 ขั้นตอนการดำเนินการ

การรวบรวมข้อมูล (Data collection)

การดำเนินงานวิจัย ผู้วิจัยได้ทำการรวบรวมข้อมูลจากงานทะเบียนและการประมวลผลร่วมกับสำนักงานกองทุนเพื่อการศึกษาโดยใช้ข้อมูลของนักศึกษาที่สำเร็จการศึกษาและตกออกของมหาวิทยาลัยเอกชนแห่งหนึ่งในภาคตะวันออกเฉียงเหนือจังหวัดขอนแก่น ระหว่างปีการศึกษา 2551 ถึงปี 2553 เป็นจำนวน 2,575ตัวอย่าง มีผู้สำเร็จการศึกษา 1,020 คน และ ไม่สำเร็จการศึกษา 1,555 คน และทำการรวมปัจจัยที่เกี่ยวข้องและจากฐานข้อมูลของมหาวิทยาลัยมีคุณลักษณะทั้งหมด 13 คุณลักษณะและผลสัมฤทธิ์ทางการศึกษา พร้อมทั้งทำการเข้ารหัสข้อมูลดังแสดงในตารางที่ 2 เพื่อใช้การทำนายสถานภาพทางการศึกษาของนักศึกษาโดยใช้เหมืองข้อมูล ประกอบด้วยชุดข้อมูลของนักศึกษาที่สำเร็จการศึกษาแล้วและไม่สำเร็จการศึกษา

การทำความสะอาดข้อมูล (Data cleaning)

การทำความสะอาดข้อมูลเป็นการทำให้ข้อมูลสมบูรณ์มีคุณภาพในการนำไปวิเคราะห์ มีขั้นตอนดังนี้

1. กรณีข้อมูลไม่สมบูรณ์ (incomplete data) สำหรับรายการใดที่ปรากฏเฉพาะรหัสประจำตัวและชื่อนักศึกษาส่วนคุณสมบัติอื่นไม่ปรากฏในงานวิจัยจะตัดออกไป

2. ใช้เฉลี่ยของคุณลักษณะของตัวอย่างที่อยู่ในประเภทเดียวกันเพื่อเติมค่าข้อมูลที่ขาดหาย เช่น กรณีรายได้ของครอบครัวขาดหายไป จะใช้ค่าข้อมูลรายได้เฉลี่ยของกลุ่มข้อมูลที่ผู้ปกครองที่มีอาชีพเดียวกัน

จากข้อมูลจำนวนทั้งสิ้น 2,575 รายการ เมื่อทำความสะอาดข้อมูลแล้วพบว่าคงเหลือข้อมูลจำนวน 2,272 รายการแยกเป็น สำเร็จการศึกษา 960 คนไม่สำเร็จการศึกษา 1,312 คน

การวิเคราะห์และหาความสัมพันธ์ของตัวแปร

จากปัจจัยที่รวบรวมมาจำนวน 13 ปัจจัยรวมผลสัมฤทธิ์ทางการศึกษา (จำนวน 2,272 รายการ) เมื่อทำการวิเคราะห์หาความสัมพันธ์ของปัจจัยทั้งหมดด้วยวิธีการวิเคราะห์สหสัมพันธ์ (Correlation) ผลการศึกษาพบว่าปัจจัยที่มีความสัมพันธ์และส่งผลต่อสถานะภาพทางการศึกษาของนักศึกษา มีจำนวน 9 ปัจจัย ได้แก่ รายได้ครอบครัว สถานะการกู้ยืมระหว่างเรียนสถานะภาพของบิดามารดา เพศ โรงเรียนเดิมที่จบ อาชีพมารดาวุฒิการศึกษาเดิม ประเภทของโรงเรียนเดิม ที่พักอาศัยระหว่างศึกษาโดยค่า R จะมีความสัมพันธ์กันอย่างมีนัยสำคัญที่ระดับ 0.05 และ 0.01 เพราะเป็นระดับนัยสำคัญทางสถิติ คือค่าที่แสดงความมั่นใจต่อการสรุปผลได้อย่างถูกต้อง และมีโอกาสที่จะเกิดความคลาดเคลื่อนในการสรุปผลมีเพียง .05 และ .01 ซึ่งโดยทั่วไปแล้วในการวิจัยจะมีการกำหนดค่าของระดับความเชื่อมั่นไว้เท่ากับ 95 % หรือ 99%

ตารางที่ 2 ปัจจัยที่ทำนายสถานภาพทางการศึกษาของนักศึกษา

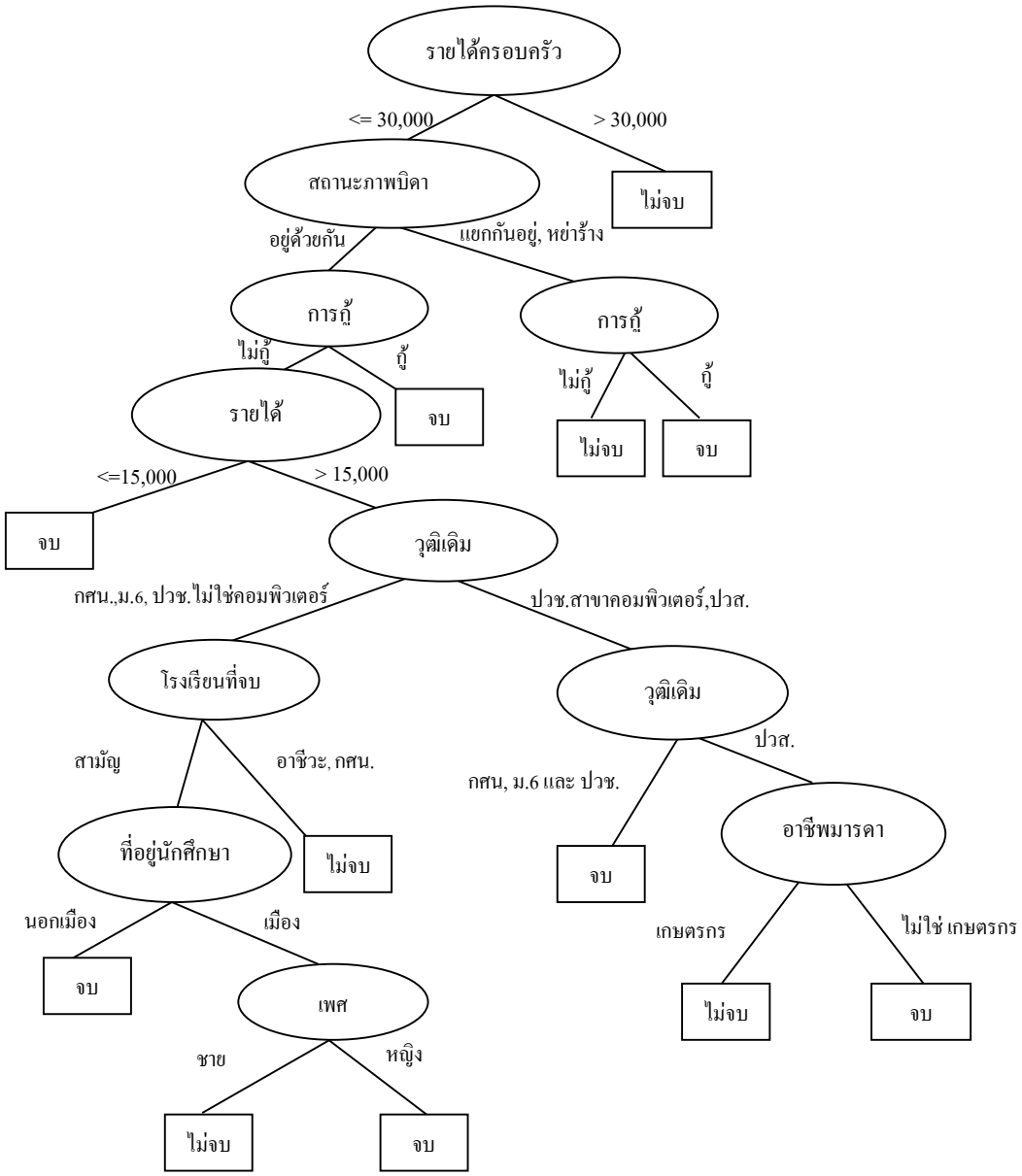
ปัจจัย	ตัวแปร	การเข้ารหัสข้อมูล	Correlation	
			R	Sig.
เพศ	Sex	0 = หญิง,1 = ชาย	-.149**	.000
เชื้อชาติ	Race	0 = ไทย ,1 = อื่นๆ	.014	.260
ศาสนา	Religion	0 = พุทธ,1 = อื่นๆ	-.019	.494
สถานการณ์กู้ยืมระหว่างเรียน	Loan	0 = ไม่กู้,1 = กู้ยืม	.580**	.000
อายุ	Age	1 = ต่ำกว่า 18 ปี, 2 =18-25 ปี, 3 = 26 ปีขึ้นไป	0.15	.474
ที่พักอาศัยในระหว่างที่ศึกษา	StudAddress	0 = ในเขตเมือง,1 = นอกเมือง	-.135*	.000

ตารางที่ 2 ปัจจัยที่ทำนายสถานภาพทางการศึกษาของนักศึกษา (ต่อ)

ปัจจัย	ตัวแปร	การเข้ารหัสข้อมูล	Correlation	
			R	Sig.
อาชีพบิดา	DadOccu	0 = เกษตรกรรม,1 = ข้าราชการ, 2 = ธุรกิจส่วนตัว, 3 = ค้าขาย,4 = อื่นๆ	-.024	.251
อาชีพมารดา	MomOccu	1 = ข้าราชการ, 2 = ธุรกิจส่วนตัว, 3 = ค้าขาย,4 = อื่นๆ	-.051*	.015
รายได้ครอบครัว	FamilyIncome	1 =ต่ำกว่า 5000, 2 = 5,001 – 10,000, 3 = 10,001 – 15000, 4=15,001 – 30,000, ขึ้นไป 30,001 =5	-.755**	.000
สถานะภาพของบิดามารดา	ParentMarital	0 = อยู่ด้วยกัน, 1 = แยกกันอยู่, 2 = อย่าร้าง, 3= บิดามารดาเสียชีวิต/	-.532**	.000
วุฒิการศึกษาเดิม	PrevDegree	0 = ม.6, 1 = การศึกษานอกโรงเรียน 2 = ปวช.ไม่ใช่สาขาคอม, 3 = ปวช.สาขาคอม 4 = ปวส. ไม่ใช่สาขาคอม, 5= ปวส.สาขาคอม	.105**	.000
ประเภทของโรงเรียนเดิม	SchoolType	0= รัฐบาล, 1 =เอกชน	.094**	.000
โรงเรียนที่จบมา	GradSchool	0 = มัธยม , 1=อาชีวะ, 2 = การศึกษานอกโรงเรียน	.120**	.000
ผลสัมฤทธิ์ทางการศึกษา	Gradstatus	0 =ไม่สำเร็จการศึกษา, 1=สำเร็จการศึกษา		

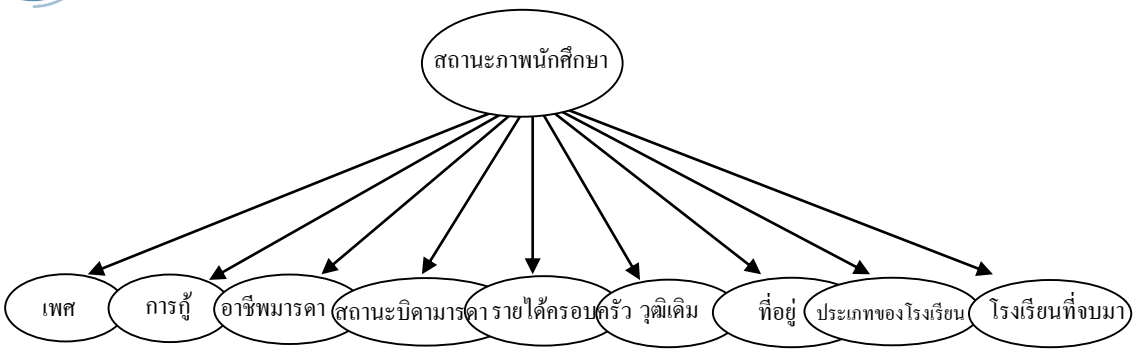
การสร้างโมเดล

ในการศึกษานี้ผู้วิจัยได้ทำการสร้างโมเดลด้วยวิธี ต้นไม้ตัดสินใจ เบย์เซี่ยนบีลิฟเน็ตเวิร์ค และ วิธีการถดถอยโลจิสติกโดยใช้ปัจจัยได้จากการหาความสัมพันธ์ทั้งหมด 9 ปัจจัยประกอบด้วย สถานะการกู้ยืมระหว่างเรียน รายได้ครอบครัว สถานะภาพของบิดามารดา เพศ ประเภทโรงเรียนเดิมที่จบมา วุฒิการศึกษาเดิม โรงเรียนที่จบมา อาชีพของมารดา ที่พักอาศัยในระหว่างเรียน โดยใช้ K-Fold Cross Validation ผลลัพธ์ของต้นไม้ตัดสินใจและเบย์เซี่ยนบีลิฟเน็ตเวิร์ค ดังภาพที่ 2 และภาพที่ 3



ภาพที่ 2 ผลลัพธ์โมเดลต้นไม้ตัดสินใจ

ภาพที่ 2 เป็นกฎการตัดสินใจของเทคนิควิธีต้นไม้ตัดสินใจ พบว่าปัจจัยที่เป็นโหนดหลักที่มีอิทธิพลต่อสถานภาพทาทางการศึกษาคือ รายได้ครอบครัวเป็นตัวกำหนดการทำงานของโมเดล



ภาพที่ 3 ผลลัพธ์เบย์เซี่ยนบีลิฟเน็ตเวิร์ค

จากภาพที่ 3 พบว่า รายได้ครอบครัว สถานะการกู้ยืมระหว่างเรียนสถานะภาพของบิดามารดา เพศ โรงเรียนเดิมที่จบ อาชีพมารดาวุฒิการศึกษาเดิม ประเภทของโรงเรียนเดิม ที่พักอาศัยระหว่างศึกษา มีความสัมพันธ์กับสถานะภาพนักศึกษา

ผลการวิจัย

ในการทดลองหาโมเดลเพื่อพยากรณ์สถานภาพสถานภาพทางการศึกษาของนักศึกษาระดับปริญญาตรีใช้ข้อมูลนักศึกษาตั้งแต่ปี 2551-2553 โดยใช้ข้อมูลทั้งสิ้น 2,575 ชุดข้อมูล ทำความสะอาดข้อมูลแล้วเหลือข้อมูลทั้งสิ้น 2,272 ชุดข้อมูล และทำการสร้างโมเดลเทคนิค ต้นไม้ตัดสินใจ เบย์เซี่ยนบีลิฟเน็ตเวิร์คและโลจิสติกรีเกรสชั่นโดยใช้ปัจจัยที่ได้จากการวิเคราะห์ความพันธ์ทั้งสิ้น 9 ปัจจัย ประมวลผลโดยใช้ K-Fold Cross Validation ผลจากการทดสอบประสิทธิภาพการทำนายของแบบจำลอง แสดงในตารางที่ 4 ผลการศึกษาพบว่า โมเดลต้นไม้ตัดสินใจมีค่าความแม่นยำสูงสุด หรือค่า Accuracy เท่ากับ 94.27 % โดยมีค่า True Positive: TP หรือ ค่าที่ทำนายว่าไม่สำเร็จการศึกษาตามข้อมูลจริง เท่ากับ 93.67 % และค่า True Negative: TN หรือ ค่าที่ทำนายว่าสำเร็จการศึกษาตามข้อมูลจริง เท่ากับ 95.19 % เบย์เซี่ยนบีลิฟเน็ตเวิร์คและโลจิสติกรีเกรสชั่น มีค่าความแม่นยำ 93.88 %, 93.70 % ตามลำดับ

ตารางที่ 3 ประสิทธิภาพการทำนายด้วยแบบจำลองต้นไม้ตัดสินใจ เบย์เซี่ยนบีลิฟเน็ตเวิร์คและโลจิสติกรีเกรสชั่น

Model	K-Flod	TP (%)	TN (%)	Accuracy (%)
Decision tree	10 k	93.67	95.19	94.27
Bays	4 k	95.42	91.77	93.88
Logistic	9 k	92.68	95.10	93.70

จากโมเดลที่ได้ ผู้วิจัยได้นำโมเดลที่ได้มาใช้ในการทำนายสถานะภาพของนักศึกษาที่เข้าศึกษาในปีการศึกษา 2554 จำนวน 414 คน ผลการทำนายแสดงในตารางที่ 5 ประสิทธิภาพในการทำนายแต่ละโมเดลพบว่า โมเดลต้นไม้ตัดสินใจค่าความถูกต้อง เท่ากับ 82.85 % โมเดลเบย์เซี่ยนบีลิฟเน็ตเวิร์คค่าความถูกต้อง เท่ากับ 78.50 % และโลจิสติกรีเกรสชั่นค่าความถูกต้องเท่ากับ 78.98 % ซึ่งมีค่าที่มีความแตกต่างนัยสำคัญ

ตารางที่ 4 ประสิทธิภาพการทำนายผล

Model	TP (%)	TN (%)	Accuracy (%)
Decision tree	80.51	85.90	82.85
Bays	72.00	87.07	78.50
Logistic	67.70	93.80	78.98

อภิปรายและสรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อสถานภาพทางการศึกษาของนักศึกษาเพื่อสร้างตัวแบบในการพยากรณ์สถานะภาพของนักศึกษา จากการศึกษาครั้งนี้พบว่าปัจจัยที่มีผลต่อสถานะภาพทางการศึกษาของนักศึกษามีทั้งหมด 9 ปัจจัยประกอบไปด้วย สถานะการกู้ยืมระหว่างเรียน รายได้ครอบครัว สถานะภาพของบิดามารดา เพศ ประเภทโรงเรียนเดิมที่จบมา วุฒิการศึกษาเดิม โรงเรียนที่จบมาและอาชีพของมารดา ที่พักอาศัยในระหว่างเรียนซึ่งมีปัจจัยดังกล่าวถูกใช้ในการสร้างโมเดลเพื่อพยากรณ์โดยวิธี ต้นไม้ตัดสินใจ วิธีการเรียนรู้แบบเบย์เชี่ยนเน็ตเวิร์ค และวิธีการถดถอยโลจิสติค ผลการศึกษาพบว่าเมื่อทดสอบการพยากรณ์โมเดลที่สร้างขึ้นโดยใช้ข้อมูลนักศึกษาที่เข้าศึกษาในปี 2554 ผลการพยากรณ์ของโมเดลต้นไม้ตัดสินใจ วิธีการเรียนรู้แบบเบย์เชี่ยนเน็ตเวิร์คและวิธีการถดถอยโลจิสติคมีค่าความถูกต้อง 82.85 %,78.50 % ,78.98 % ตามลำดับ

เอกสารอ้างอิง

ไก้รุ่งเฮงพระพรหมและคณะ. (2557). การค้นหาความรู้คุณลักษณะของนักศึกษาที่มีต่อผลสัมฤทธิ์ทางการเรียนในกลุ่มโปรแกรมคอมพิวเตอร์มหาวิทยาลัยราชภัฏนครปฐมด้วยเทคนิคเหมืองข้อมูล.วารสารวิชาการราชภัฏตะวันตก (ฉบับบัณฑิตศึกษา);9(1), 71-80

ปฐมา อาแวและคณะ. (2553). การศึกษาความสัมพันธ์ระหว่างผลสัมฤทธิ์ทางการเรียนระดับมัธยมศึกษาตอนปลายผลการสอบคัดเลือกเข้ามหาวิทยาลัย และผลการเรียนระดับมหาวิทยาลัย กับผลสัมฤทธิ์ทางการเรียนของบัณฑิตมหาวิทยาลัยสงขลานครินทร์วิทยาเขตปัตตานีปีการศึกษา 2546-2550.รายงานการวิจัยสถาบันของสำนักงานอธิการบดีมหาวิทยาลัยสงขลานครินทร์วิทยาเขตปัตตานี

ยุทธ ไกยวรรณ. (2555). หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย.วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย;4(1), 1-12

ราชกิจจานุเบกษา.(2541). พระราชบัญญัติกองทุนเงินยืมเพื่อการศึกษา พ.ศ.2541

สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐานกระทรวงศึกษาธิการ. (2552). สถานการณ์การออกกลางคันของนักเรียนปีการศึกษา 2552.

สำนักทะเบียนและประมวลผล.(2557). ข้อมูลจำนวนนักศึกษารับเข้าและลาออก ปีการศึกษา2551-2557. ขอนแก่น: มหาวิทยาลัยเอกชนภาคตะวันออกเฉียงเหนือ

อ้อมใจ เขาทอง และพุธษดี ศิริแสงตระกูล.(2557). โมเดลเพื่อการพยากรณ์นักศึกษาที่มีผลการเรียนต่ำกว่าเกณฑ์. การประชุมวิชาการ Knowledge and Smart Technology :(6),60-65

Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar. Mining student data using decision trees. International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan. 2006.; 3(2) 43-47

Al,R., Qasem A., Emad, M., Al, S., & Mustafa, N. (2006).Mining student data using decision trees. International Arab Conference on Information Technology, 15-17 December. (pp. 43-47). Jordan: Zarqa Private University

Bunkar, K., Bunka, R., Umesh, r., &Bhupendra, P. (2012). Data Mining: Prediction for Performance Improvement of Graduate Students using Classification. Ninth International Conference on Wireless and Optical Communications Networks (WOCN), 20-22 Sept. (pp. 1 - 5). India: Indore.

David, A. (2009). Statistical ModelsTheory and Practice. Cambridge University Press. p. 128.

Han, J., &Kamber, M. (2006). Data Mining Concept and Techniques.Seconded,Morgan Kaufmann Publishers,4 Jun (pp. 285-382). Ireland: Uk.

Mehran, S. Learning Limited Dependence Bayesian Classifiers.KDD-96 Proceedings, 2–4 August (pp. 81–106). Portland: Oregon;1996.

Qui, R.Diction of decision trees. Mach Learn,1(1), 81–106;1998

Rodgers, J. L.; Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician. 42 (1): 59–66.

Suwanarattaphoom P. An approach for improving associative classification in imbalanced dataset [M Eng Thesis].Bangkok: Kasetsart University; 2010.Thai

Tripti, M, K., Dharminder, Sangeeta G. (2014). Mining Student s' Data for Performance Prediction. Fourth International Conference on Advanced Computing & Communication Technologies,20-22 Sept. (pp. 30- 37). India: Indore.