ประสิทธิภาพการจำแนกข้อมูลการเลือกอาชีพโดยอัตโนมัติด้วยเทคนิคเหมืองข้อมูล

Comparative Efficiency of Classification Choosing Career Automatic with Data Mining Techniques.

ชัชชฎา วันดี,[1] จิรัฏฐา ภูบุญอบ,[2*] ฉัตรเกล้า เจริญผล[3]
Chatchada Wandee,[1] Jiratta Phuboon-ob,[2*] Chatklaw Jareanpon[3]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา โดยในงานวิจัยนี้ได้ใช้ชุดข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2550-2554 จำนวน 12 คุณลักษณะ และ 2,515 ระเบียน ซึ่งได้นำเทคนิคแบบจำลองต้นไม้ตัดสินใจ เทคนิคโครงข่ายประสาทเทียม และเทคนิคการเรียนรู้แบบเบย์ มาทำการเปรียบเทียบประสิทธิภาพ ผลจากการศึกษาพบว่าประสิทธิภาพในการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ มีประสิทธิภาพในการจำแนกสูงสุดด้วยค่าเฉลี่ย 80.62% และปัจจัยสำคัญที่ทำให้การเลือกอาชีพตรงหรือไม่ตรงกับสาขา มี 4 ปัจจัย คือ สาขาวิชาที่เรียน เกรดเฉลี่ยเฉพาะวิชาสาขา เพศ และเกรดเฉลี่ยรวม ซึ่งผลการทดลองนี้สามารถนำไปประยุกต์ใช้กับคณะหรือหน่วยงานที่เกี่ยวข้อง เพื่อวางแผนพัฒนาโครงสร้างหลักสูตรหรือวางแผนการศึกษาให้กับนิสิตได้

คำสำคัญ: ข้อมูลภาวะการมีงานทำของบัณฑิต ข้อมูลระเบียนประวัติของนิสิต ประสิทธิภาพในการจำแนกข้อมูลการเลือกอาชีพ เทคนิคเหมืองข้อมูล

Abstract

This research aims to compare the performance of the classification for choosing the career for graduated students. This research used the data set of the job status and the personal data of graduated students from Faculty of Informatics Mahasarakham University from 2007-2011 years. The data set has 12 attributes and 2,515 records. This research used and compared the result of Decision tree techniques, Artificial Neural Network and Naive Bayes. The highest accuracy result is from the Decision tree with 80.62%. The important factors are Major, GPA specially in Major only, gender and overall GPA. The results in this experiment can be applied to the relevant authorities, to development plans course structure, or education plan.

Keyword: data set of the job status, the personal data of graduated students**,** performance of the classification for choosing the career, data Mining Techniques

---

[1]นิสิตปริญญาโท, [2,3]อาจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150
[1] Graduate student, [2,3]Lecturer, Faculty of Informatics, Mahasarakham University, KantharawichaiDistrict, Maha Sarakham. Thailand
*Corresponding author : Jiratta Phuboon-ob of Faculty of Informatics, Mahasarakham University, KantharawichaiDistrict, Maha Sarakham 44150. Thailand

บทนำ

การที่จะพัฒนาประเทศให้เจริญก้าวหน้าจะต้องมีการผลิตประชากรที่มีความรู้ความสามารถให้ครบทุกด้านตามความต้องการของประเทศที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา ทำให้การเลือกอาชีพนับว่าเป็นเรื่องสำคัญอย่างยิ่งในการดำรงชีวิต ซึ่งการเลือกอาชีพจะต้องเริ่มต้นด้วยการวางแผน ตั้งแต่วัยเรียน โดยเป็นการวางแผนระยะยาวที่ต้องใช้เวลานาน ซึ่งคนเรามีความถนัดความสามารถ และความสนใจในงานอาชีพแตกต่างกัน ดังนั้นหลายคนอาจต้องตัดสินใจเลือกอาชีพที่อาจตรงหรือไม่ตรงกับสาขาที่เรียน ด้วยเหตุผลที่แตกต่างกันออกไป ดังนั้นทุกมหาวิทยาลัยจึงได้จัดทำแบบสำรวจภาวะการมีงานทำของบัณฑิต เพื่อใช้ในการสำรวจปัญหาอุปสรรคในการหางานทำ และเพื่อให้ทราบว่านิสิต/ที่สำเร็จการศึกษาออกไปนั้นเลือกที่จะประกอบอาชีพอะไร โดยบัณฑิตจะต้องบันทึกข้อมูลให้กับมหาวิทยาลัยหลังจากสำเร็จการศึกษา ผู้วิจัยจึงได้นำข้อมูลภาวะการมีงานทำของบัณฑิต[1] และข้อมูลระเบียนประวัติของนิสิต[2] ที่สำเร็จการศึกษาตั้งแต่ปี พ.ศ. 2550-2554 ของคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม จำนวน 2,515 ระเบียน คุณลักษณะมาใช้ในงานวิจัยนี้ 12

ในการวิจัยจะต้องเลือกใช้เทคนิควิธีในการจำแนกข้อมูลที่เหมาะสมกับชุดข้อมูล เพื่อเป็นการเพิ่มประสิทธิภาพในการจำแนกให้มีความถูกต้องมากที่สุด ซึ่งเทคนิคการจำแนกข้อมูล เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เพื่อแสดงให้เห็นความแตกต่างระหว่างกลุ่มของข้อมูล และเพื่อทำนายข้อมูลว่าควรจัดอยู่ในคลาสใดตามโมเดลที่ใช้จำแนกข้อมูล จากนั้นจะปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลได้

งานวิจัยนี้จึงได้เสนอเทคนิคในการจำแนกข้อมูล 3 เทคนิควิธี ได้แก่แบบจำลองต้นไม้ตัดสินใจ แบบจำลองโครงข่ายประสาทเทียม แบบจำลองการเรียนรู้แบบเบย์ เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล และนำแบบจำลองที่มีประสิทธิภาพดีที่สุดมาหาปัจจัยที่ส่งผลต่อการเลือกอาชีพนิสิต อันจะทำให้สามารถนำมาพัฒนาปรับปรุงโครงสร้างหลักสูตรหรือวางแผนการศึกษาให้มีประสิทธิภาพมากยิ่งขึ้น

ทฤษฎีที่เกี่ยวข้อง

เทคนิคเหมืองข้อมูล (Data Mining)

เทคนิคเหมืองข้อมูล[3] (Data Mining) คือ การวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่เพื่อกลั่นกรองค้นหา ด้วยวิธีทางสถิติและคณิตศาสตร์ เพื่อให้ได้สารสนเทศที่มีรูปแบบ (Pattern) มีความสัมพันธ์ โดยสารสนเทศนั้นได้ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ และนำสารสนเทศที่ได้มาใช้เพื่อช่วยในการวางแผน การตัดสินใจในการบริหารหรือแก้ปัญหาด้านต่าง ๆ ซึ่งถือได้ว่าเป็นเครื่องมือที่ช่วยเพิ่มคุณค่าให้กับข้อมูลที่มีอยู่ ซึ่งประโยชน์หลักของเหมืองข้อมูล คือ การค้นหาความรู้ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่เพื่อให้ได้ซึ่งความรู้มาช่วยในการตัดสินใจ

การทำเหมืองข้อมูลมีความซับซ้อนเพราะเป็นการวิเคราะห์ข้อมูลจากฐานข้อมูลขนาดใหญ่จึงได้กำหนดขั้นตอนการทำเหมืองข้อมูล[4] ไว้ 6 ขั้นตอน ดัง Figure 1 ได้แก่

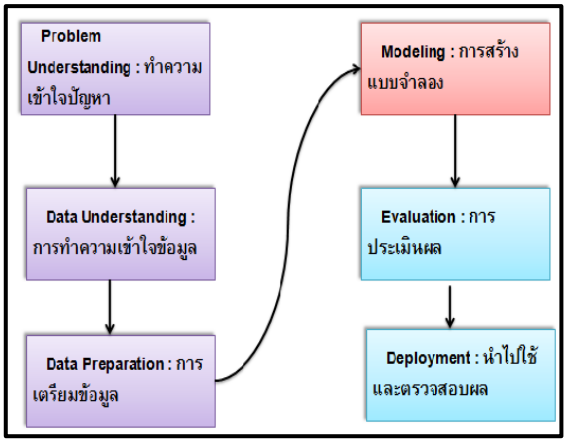

Figure 1 Data Mining.

1. Problem Understanding: ทำความเข้าใจปัญหา คือ การทำความเข้าใจปัญหาของหน่วยงาน และวัตถุประสงค์ที่จะนำกระบวนการทำเหมืองข้อมูลเข้ามาใช้

2. Data Understanding: การทำความเข้าใจข้อมูล คือการทำความเข้าใจกับข้อมูลที่มีหรือข้อมูลของ

หน่วยงาน ที่จะใช้สำหรับการวิเคราะห์ผ่านกระบวนการทำเหมืองข้อมูล

3. Data Preparation: การเตรียมข้อมูลคือ การจัดรูปแบบของข้อมูลใหม่ให้อยู่ในรูปแบบที่สามารถวิเคราะห์ข้อมูลได้ ถ้าข้อมูลนั้นไม่อยู่ในรูปแบบที่ถูกต้องหรือเหมาะสม จะต้องมีการปรับข้อมูลให้อยู่ในรูปแบบที่โปรแกรมที่ใช้วิเคราะห์ สามารถเรียกใช้งานได้

4. Modeling: การสร้างแบบจำลองคือ การนำชุดข้อมูล ที่ได้ทำการเตรียมไว้มาทำการวิเคราะห์ ผ่านกระบวนการอัลกอริทึมที่เลือกใช้ เพื่อให้ได้มาซึ่งตัวแบบหรือ Model

5. Evaluation: การประเมินผล การประเมินอาจจะประเมินตัวแบบที่สร้างขึ้น ด้วยการลองนำไปใช้กับสถานการณ์จริง หรือกับสถานการณ์ที่จำลองขึ้น เพื่อดูว่าตัวแบบนี้ได้ผลหรือไม่เพียงใด และผิดพลาดตรงไหน ก็ทำการแก้ไขในกระบวนการก่อนหน้า ก่อนที่จะนำไปใช้งานจริง

6. Deployment: การนำไปใช้ เมื่อตัวแบบ (Model) มีความถูกต้อง สามารถนำตัวแบบ (Model) ที่ได้ไปใช้และตรวจสอบว่าบรรลุเป้าหมายที่ตั้งไว้หรือไม่

อัลกอริทึมในการทำเหมืองข้อมูล[5] สามารถแบ่งประเภทเป็น 2 ประเภท ได้แก่

1. การสร้างตัวแบบในการทำนาย (Predictive Modeling) หรือเรียกว่า Supervised Learning เช่น การที่มีข้อมูลในอดีต และนำข้อมูลมาสร้างโมเดลเพื่อทำการทำนายอนาคตโดยมีการใช้ข้อมูลในการสอน (Train) เช่น การจำแนกประเภทข้อมูล (Classification)

2. การสร้างตัวแบบในการบรรยาย (Descriptive Modeling) หรือเรียกว่า Unsupervised Learning คือ การนำข้อมูลที่มีอยู่มาศึกษา เช่น พฤติกรรมของลูกค้า เป็นการเรียนรู้จากข้อมูลที่มีอยู่และอธิบายให้เห็นภาพชัดเจน เทคนิคนี้เป็นลักษณะ เช่น Cluster Analysis, Association Rules

เทคนิคต้นไม้ตัดสินใจ (Decision tree: DT)

ต้นไม้ตัดสินใจ[6,7] คือ แบบจำลองที่มีลักษณะคล้ายกับต้นไม้ จะมีการสร้างกฎต่างๆ ขึ้นเพื่อใช้ในการตัดสินใจ ซึ่งแต่ละโหนด (Node) จะแสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล รูปแบบของต้นไม้จะประกอบด้วยโหนดแรกสุดที่เรียกว่า Root node จาก Root node จะแยกออกเป็นโหนดลูก และที่โหนดลูกก็จะมีลูกของตัวเอง ซึ่งโหนดในระดับสุดท้ายจะเรียกว่า Leaf node ซึ่งแสดงกลุ่มหรือคลาส (Class) ที่กำหนดไว้ สำหรับในงานวิจัยนี้ใช้อัลกอริทึม C4.5[8] ในการสร้างต้นไม้ตัดสินใจซึ่งมีวิธีการดังต่อไปนี้

ถ้าให้ชุดข้อมูล C ประกอบด้วยค่าที่เป็นไปได้ คือ $\{c_1,c_2 \ldots c_n\}$ และให้ค่าความน่าจะเป็นที่เกิดขึ้นเป็นค่า c_i ซึ่งมีค่าเท่ากับ $P(c_i)$ จะได้ค่า Information Gain ของ C เขียนแทนด้วย $I(C)$ ดังสมการที่ (1)

$$I(C) = \sum_{i=1}^{n} -P(c_i)\ log_2\ P(c_i) \qquad (1)$$

ถ้าให้ข้อมูลสอน คือ T โดยคุณลักษณะที่เป็นโหนด เช่นค่า x และมีค่าทั้งหมดที่เป็นไปได้เท่ากับ n ค่า ซึ่งโหนดปัจจุบันจะแบ่งตัวอย่างค่า T ออกตามกิ่งเป็น $\{t_1,\ t_2 \ldots\ t_n\}$ ตามค่าที่เป็นไปได้ของค่า x ดังนั้น จึงสามารถคำนวณค่า Information Gain หลังจากแบ่งตามคุณลักษณะ ดังสมการที่ (2)

$$I_x(T) = \sum_{i=1}^{n} \frac{t_i}{T} I(t_1) \qquad (2)$$

ค่า Gain ของคุณลักษณะ x ดังสมการที่ (3)

$$Gain(x) = I(T) - I_x(T) \qquad (3)$$

จากนั้นคำนวณค่า Information Gain ของ Split Information ตามคุณลักษณะแต่ละตัว ถ้าให้ T คือชุดของตัวอย่างเมื่อแบ่งตัวอย่างนี้ตามคุณลักษณะ x จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่ง คือ $\{t_1, t_2, \ldots, t_n\}$ จำนวน n ชุดตามค่าที่เป็นไปได้ในคุณสมบัติ x เมื่อคำนวณค่า Split Information ได้ ดังสมการที่ (4)

$$Split\ Information = \sum_{i=1}^{n} \frac{t_i}{T} \log_2 \frac{t_i}{T} \qquad (4)$$

ค่า Gain Ratio สามารถคำนวณได้จาก *Gain Ratio* = *Gain - Split Information* ค่า Gain Ratio สูงสุดจะถูกเลือกเป็นคุณลักษณะเริ่มต้น และเลือกคุณลักษณะถัดไปตามค่า Gain Ratio น้อยลงตามลำดับ

เทคนิคโครงข่ายประสาทเทียม (Artificial neural network: ANN)

โครงข่ายประสาทเทียม[9,10] มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ด้วยโปรแกรมคอมพิวเตอร์ซึ่ง Back Propagation Algorithm เป็นอัลกอริทึมที่ใช้ในการเรียนรู้ของโครงข่ายประสามเทียมวิธีหนึ่งที่นิยมใช้ใน Multilayer Perceptron เพื่อปรับค่าน้ำหนักสำหรับข่ายงานไปข้างหน้าหลายชั้น โดยเคลื่อนจากข้อมูลชั้นนำเข้า ชั้นซ่อน ไปจนถึงชั้นแสดงผล ดัง Figure 2 ซึ่งจะคำนวณค่าผิดพลาดระหว่างเอาต์พุตของข่ายงานและค่าจริง เพื่อใช้ในการปรับค่าเวกเตอร์น้ำหนัก และจะทำซ้ำไปซ้ำมาจนได้ค่าเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดน้อยที่สุด ดังสมการที่ (5)

$$n = \sum_{i=1}^{z} x_i w_i + b \qquad (5)$$

โดยที่ n คือ ผลรวมที่ได้จากฟังก์ชันผลรวม

x_i คือ ค่าข้อมูลเข้าตัวที่ i

w_i คือ ค่าน้ำหนักของนิวรอนตัวที่ i

z คือ จำนวนนิวรอนชั้นข้อมูลเข้า

b คือ ค่าความโน้มเอียง

i คือ มีค่าตั้งแต่ 1 ถึง z

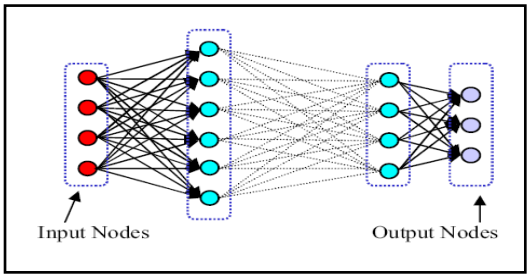

Figure 2 Artificial neural network.

เทคนิคการเรียนรู้แบบเบย์ (Naive bayes: NB)

การเรียนรู้แบบเบย์[11,12] เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพอีกวิธีหนึ่งซึ่งใช้งานได้ดีและเหมาะสมกับกรณีของเซตตัวอย่างที่มีจำนวนมากและมี Attribute ของตัวอย่างไม่ขึ้นต่อกัน เป็นการเรียนรู้ที่ใช้หลักการของความน่าจะเป็นซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes theorem) เข้ามาช่วยในการเรียนรู้จุดมุ่งหมายก็เพื่อต้องการสร้างอัลกอริทึมที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกตจากนั้นนำอัลกอริทึมมาหาว่าสมมติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย ข้อดีของวิธีการเรียนรู้แบบนี้ คือ เราสามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior knowledge) เข้ามาช่วยในการเรียนรู้ได้ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดีไม่ด้อยกว่าวิธีการเรียนรู้ประเภทอื่น ดังสมการที่ (6)

$$P(H|E) = \frac{[P(E|H) \times P(H)]}{P(E)} \qquad (6)$$

โดยที่ $P(H|E)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E

งานวิจัยที่เกี่ยวข้อง

ชัชชฎา วันดี และคณะ[13] ทำการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล โดยใช้ชุดข้อมูลภาวะการมีงานทำของบัณฑิต และชุดข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2550-2554 ด้วยเทคนิคแบบจำลองต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และการเรียนรู้แบบเบย์ ผลจากการศึกษาการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ ซึ่งงานวิจัยนี้ยังไม่มีการหาปัจจัยที่ส่งผลต่อการเลือกอาชีพของนิสิต

Aitkenhead M.J.[14] ได้นำเสนอการพัฒนาการจำแนกร่วมกับต้นไม้ตัดสินใจจากปัญหาการจำแนกและจัดหมวดหมู่นั้นมีลักษณะที่แตกต่างกันหรือคุณสมบัติของระบบต่างกันและข้อมูลยังมีการสูญหายหรือเกิดการรบกวนทำให้ข้อมูลไม่มีคุณภาพจึงวิเคราะห์ข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจโดยเลือกใช้อัลกอริทึม C4.5 ซึ่งผลการวิจัยพบว่าอัลกอริทึมต้นไม้ตัดสินใจแสดงให้เห็นถึงวิวัฒนาการหรือโครงสร้างของข้อมูลได้ง่ายและสามารถจัดการกับช่วงของค่าและชนิดของมูลได้และอัลกอริทึมนี้ยังมีความเข้าใจกว่าวิธีอื่นๆ

Lawrence O. Hall และคณะ[15] ได้นำเสนองานวิจัยที่ทดสอบเทคนิค Pruning สำหรับอัลกอริทึม C4.5 คือ Error Based Pruning (EBP) ผลที่ได้คือต้นไม้มีขนาดใหญ่แต่ความถูกต้องไม่เพิ่มขึ้นทั้งนี้เพราะงานวิจัยส่วนใหญ่ใช้ค่าพารามิเตอร์ Certainty Factorเท่ากับ 0.25

โดยงานวิจัยนี้ได้ทำการทดลองกับข้อมูลหลายชุด เพื่อทดลองว่ามีผลต่อขนาดของต้นไม้และความถูกต้องหรือไม่ ผลที่ได้พบว่ามีค่าความถูกต้องใกล้เคียงกัน

S.chen และคณะ[16] ได้นำเสนอการจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนที่มีการกำหนดค่าคอร์เนลฟังก์ชันรวมทั้งค่าพารามิเตอร์ที่หลากหลาย ทำให้ตัวแทนที่สร้างขึ้นสามารถนำไปแก้ปัญหาที่หลากหลายเช่นเดียวกัน รวมทั้งยังเป็นการเพิ่มประสิทธิภาพโดยรวมของโมเดลอีกด้วย

วิธีการดำเนินการวิจัย

ข้อมูลที่ใช้ในการวิจัย

งานวิจัยนี้นำข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม โดยใช้ชุดข้อมูลระหว่าง ปี พ.ศ.2550 – 2554 มีจำนวน 12 คุณลักษณะ และ 2,515 ระเบียน ดัง Table 1

Table 1 Attributes used in research.

No.	Attributes	No.	Attributes
1.	major	7.	father's career
2.	gender	8.	Income Father / year
3.	Old school	9.	mothers's career
4.	GPA in old school	10.	Income mothers / year
5.	Overall GPA	11.	Jobs.
6	GPA specially in major only	12.	Yes or No

ขั้นตอนในการดำเนินงานวิจัย

การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลได้ดำเนินการตามเทคนิคการทำเหมืองข้อมูล โดยเริ่มจากการเตรียมข้อมูล เนื่องจากข้อมูลที่ได้มาอาจมีความไม่สมบูรณ์ เช่น ข้อมูลแปลกปลอม หรือ ข้อมูลขาดหาย จึงต้องทำการกลั่นกรองข้อมูล (Data Cleaning) และการแปลงข้อมูล (Data Trasformation)[17] จากนั้นแบ่งข้อมูลออกเป็นชุดข้อมูลสำหรับสอนและชุดข้อมูลสำหรับทดสอบ ดัง Figure 3

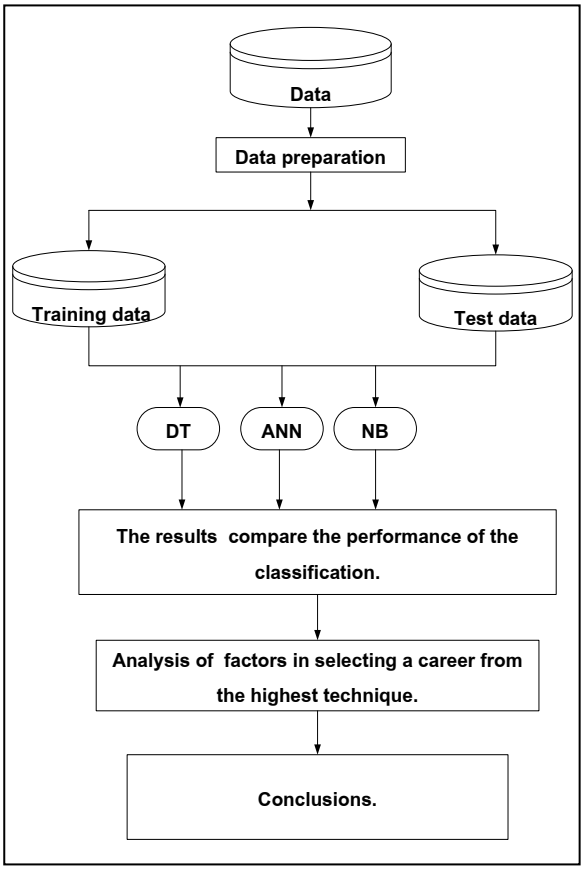

Figure 3 Step operation.

จากภาพที่ 3 เมื่อเตรียมข้อมูลเรียบร้อยแล้วจะทำการเลือกตัวอย่างด้วยการแบ่งข้อมูลตามสัดส่วนชุดข้อมูลการเรียนรู้และชุดข้อมูลการทดสอบเป็น 70:30 โดยใช้เทคนิคการสุ่มอย่างง่าย[18] (Simplerandom sampling) สมาชิกทั้งหมดของประชากรเป็นอิสระซึ่งกันและกัน แล้วสุ่มหน่วยของการสุ่ม(Sampling unit) จนกว่าจะได้จำนวนตามที่ต้องการ แต่ละครั้งที่สุ่มสมาชิกของแต่ละหน่วยของประชากรมีโอกาสถูกเลือกเท่าๆ กัน เพื่อนำชุดข้อมูลการเรียนรู้มาจำแนกข้อมูลด้วยเทคนิค DT, ANN และ NB จากนั้นทำการทดสอบประสิทธิภาพในการจำแนกข้อมูลด้วยชุดข้อมูลทดสอบ เพื่อวิเคราะห์หาค่าความถูกต้องของการจำแนกข้อมูล ซึ่งในการทดลองจะทำการคำนวณหาค่าความแม่นยำของแบบจำลอง (Accuracy) ของชุดข้อมูลการเรียนรู้ และคำนวณหาค่าสัมบูรณ์ของค่าคลาดเคลื่อนเฉลี่ย (Mean Absolute Error: MAE) ของชุดข้อมูลทดสอบ เพื่อวัดค่าประสิทธิภาพของทั้ง 3 เทคนิควิธี ดังสมการที่ 7 และ 8 ตามลำดับ

ค่าความแม่นยำของแบบจำลอง[19] (Accuracy) ของชุดข้อมูลเรียนรู้

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (7)$$

โดยที่

TP คือ ค่าที่ทำนายถูกต้อง (ข้อมูลบอกว่าจริง ทำนายว่าจริง)

TN คือ ค่าที่ทำนายถูกต้อง (ข้อมูลบอกว่าไม่จริง ทำนายว่าไม่จริง)

FP คือ ค่าที่ทำนายไม่ถูกต้อง (ข้อมูลบอกว่าจริง ทำนายว่าไม่จริง)

FN คือ ค่าที่ทำนายไม่ถูกต้อง (ข้อมูลบอกว่าไม่จริง ทำนายว่าจริง)

ค่าสัมบูรณ์ของค่าคลาดเคลื่อนเฉลี่ย[20] (Mean Absolute Error: MAE) ของชุดข้อมูลทดสอบ

$$MAE = \frac{\sum_{i=1}^{n} |e_i|}{n} \quad (8)$$

โดยที่

e_i คือผลต่างระหว่างค่าข้อมูลจริงและค่าพยากรณ์

n คือข้อมูลในการพยากรณ์

ผลการดำเนินงานวิจัย

ผลการวิเคราะห์ประสิทธิภาพวิธีการจำแนกข้อมูล การเลือกอาชีพ ระหว่างปี พ.ศ. 2550 – 2554 ของคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ผลการแสดงดัง Table 2

Table 2 The results compared of classification data.

Performance accuracy. (Average)		
Technique	Accuracy	MAE
DT	80.62	0.38
ANN	78.94	0.30
NB	70.47	0.26

จากตารางที่ 2 พบว่าประสิทธิภาพในการจำแนกประเภทข้อมูล ด้วยเทคนิค DT มีประสิทธิภาพการจำแนกข้อมูลเฉลี่ย 80.62% เทคนิค ANN 78.94% และเทคนิค NB 70.47% ตามลำดับ และเมื่อนำข้อมูลมาแสดงในรูปของกราฟ เพื่อเห็นความแตกต่างของผลความแม่นยำที่ชัดเจนยิ่งขึ้น ดัง Figure 4

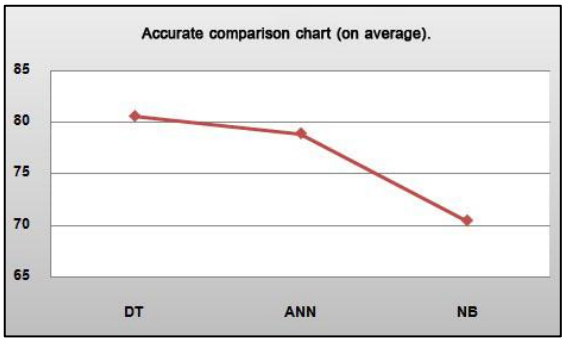

Figure 4 Accurate comparison chart (on average).

จะสังเกตเห็นว่าเทคนิค DT มีประสิทธิภาพในการจำแนกสูงที่สุด เหมาะสมที่จะนำไปทำการทดลองวิเคราะห์เพื่อหาความสัมพันธ์ของปัจจัยที่ส่งผลต่อการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา ซึ่งในการตรวจสอบหาความสำคัญของปัจจัย จะลดตัวแปรทีละตัวแปร และวัดประสิทธิภาพจากค่า MAE เพื่อให้ทราบว่าปัจจัยใดมีความสำคัญมากที่สุด โดยเรียงความสำคัญจากค่ามากไปหาค่าน้อย กล่าวคือ หากค่า MAE มีค่าความคาดเคลื่อนมากแสดงว่าตัวแปรนั้นมีความสำคัญมาก ซึ่งปัจจัยสำคัญที่ส่งผลต่อการเลือกอาชีพของนิสิตหลังสำเร็จการศึกษาระดับปริญญา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม มีปัจจัยที่สำคัญ 4 ปัจจัย ดัง Table 3

Table 3 Test results data reduction imported with decision tree technique.

No.	Factors	MAE
1	major	0.3982
2	GPA specially in major only	0.3911
3	gender	0.3021
4	Overall GPA	0.3013

จากตารางที่ 3 แสดงผลการวิเคราะห์ปัจจัยที่ส่งผลต่อการเลือกอาชีพของนิสิต ซึ่งวัดจากการคำควณค่า พบว่าปัจจัยที่สำคัญ 4 ปัจจัย คือ สาขาที่เรียน เกรดเฉพาะสาขาที่เรียน เพศ เกรดเฉลี่ยรวม และเมื่อนำข้อมูลมาแสดงในรูปของกราฟ เพื่อเห็นความแตกต่างของผลความแม่นยำที่ชัดเจนยิ่งขึ้น ดัง Figure 5

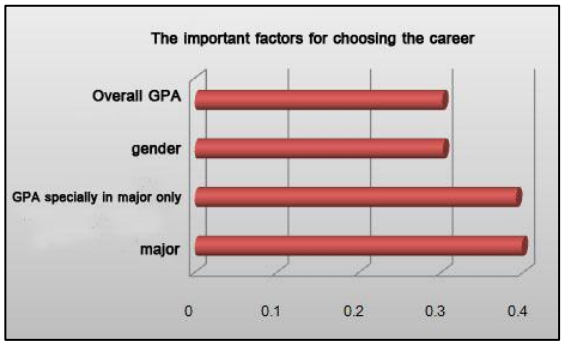

Figure 5 Graph factors that affect career choices.

สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้ทำการวัดประสิทธิภาพของการจำแนกประเภทข้อมูล ซึ่งผู้วิจัยได้เลือกเทคนิควิธีที่ได้รับความนิยมสำหรับการจำแนกข้อมูล 3 เทคนิควิธี ได้แก่เทคนิคต้นไม้ตัดสินใจ เทคนิคโครงข่ายประสาทเทียม และเทคนิคการเรียนรู้แบบเบย์ โดยทดลองกับข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2550 – 2554 จำนวน 12 คุณลักษณะ และ 2,515 ระเบียน จากผลการทดลอง พบว่าค่าเฉลี่ยความถูกต้องของเทคนิค DT ให้ค่าความถูกต้องสูงสุดโดยมีค่าเฉลี่ย 80.62% รองลงมาคือ ANN และNB มีค่าความถูกต้องสูงเฉลี่ย 78.94% และ 70.47% ตามลำดับ ดังนั้นจึงสรุปได้ว่าอัลกอริทึม DT ให้ประสิทธิภาพในการจำแนกข้อมูลสูงสุดเหมาะกับข้อมูลที่ใช้ในงานวิจัยนี้มากที่สุด และเมื่อวิเคราะห์ปัจจัยที่ส่งผลต่อการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามจากเทคนิค DT พบว่ามีปัจจัยที่สำคัญ 4 ปัจจัย คือ สาขาที่เรียน เกรดเฉพาะวิชาสาขา เพศ และเกรดเฉลี่ยรวม ซึ่งผลการทดลองงานวิจัยนี้สามารถนำไปประยุกต์ใช้กับคณะหรือหน่วยงานที่เกี่ยวข้อง เพื่อวางแผนพัฒนาโครงสร้างหลักสูตรหรือวางแผนการศึกษาให้กับนิสิตได้

เอกสารอ้างอิง

1 กองแผนงาน. มหาวิทยาลัยมหาสารคาม. . ระบบภาวะการมีงานทำของบัณฑิต. [ออนไลน์]. 2554 [สืบค้นเมื่อ 28 พฤศจิกายน 2554], http://www.survey.msu.ac.th/.

2 กองทะเบียนและประมวลผล. มหาวิทยาลัยมหาสารคาม. งานทะเบียนและประมวลผล. [ออนไลน์]. 2554 [สืบค้นเมื่อ 28 พฤศจิกายน 2554], http:// www.regpr.msu.ac.th.

3 Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2st ed. Oxford Morgan Kaufmann Publishers, 2005.

4 บุญเสริม กิจศิริกุล. รายงานวิจัยฉบับสมบูรณ์ : โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ 2545 โครงการย่อยที่ 7 อัลกอริทึมการทาเหมืองข้อมูล. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัย, 2546.

5 บุญมา เพ่งซ่วน. การใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบสนับสนุนการตัดสินใจ ด้านการผลิตบัณฑิตระดับปริญญาตรี.วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต.มหาวิทยาลัยศิลปากร,2548.

6 มลธิดา ฤทธิ์สมบูรณ์, สุชา สมานชาติ. การพัฒนาระบบสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อโดยใช้เทคนิคต้นไม้ตัดสินใจ.วารสารเทคโนโลยีสารสนเทศ 2551:4(7):9-14.

7 กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ, ธนาวินท์ รักธรรมานนท์. การใช้เทคนิคการทำเหมืองข้อมูล (Data Mining)เพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์.วารสารเทคโนโลยีสารสนเทศ 2544;3(11):134-142.

8 Quinlan John Ross. C4.5: programs for machine learning. London: England, 1988.

9 เรวดี ศกดิ์ดุลยธรรม. การใช้เทคนิคดาต้าไมน์นิ่งในการสร้างฐานความรู้ เพื่อการทำนายสัมฤทธิ์ ผล

ทางการเรียนของนักศึกษา.วิทยาลัยราชพฤกษ์, 2552.

10 นรินทร์ พนาวาส, นิเวศ จิระวิชิตชัย. การจำแนกมะเร็งเม็ดเลือดขาวโดยใช้โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น. เอกสารงานประชุมสัมมนาวิชาการมหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก ครั้งที่ 4 ; 27 พฤษภาคม 2554; ชลบุรี. p. 154-161

11 กฤตยา ทองผาสุข. การเปรียบเทียบเทคนิคต้นไม้ตัดสินใจ กฎนาอีฟเบย์ และเคเนียรเรสเนเบอร์ เพื่อการจำแนกข้อมูล. National Conference on Computer Informastion Technolgies 2011; 26-28 มกราคม 2554; นครปฐม. p. 30-35.

12 ธัญญาภรณ์ บุญยัง, เอกรัฐ หล่อพิเชียร. การเปรียบเทียบข้อมูลของแบบจำสองเคเนียรเรสเนเบอร์ นาอีฟเบย์ ต้นไม้ตัดสินใจ และกฎพื้นฐาน. National Conference on Computer Informastion Technolgies 2011; 26-28 มกราคม 2554; นครปฐม. p. 19-23.

13 ชัชชฎา วันดี, จิรัฏฐา ภูบุญอบ, ฉัตรเกล้า เจริญผล. การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเลือกอาชีพโดยใช้เทคนิคเหมืองข้อมูล. Proceedings of National Conference on Computing and Information Technology:: NCCIT; 9 - 10 พฤษภาคม 2556; กรุงเทพฯ. p. 161 - 166.

14 Aitkenhead MJ. A co-evolving decision tree classification method. Expert Systems With Applications - ESWA 2008; 34 p. 18-25.

15 Lawrence O. Hall, Richard Collins, Bowyer KW. Error-Based Pruning of Decision Trees Grown on Very Large Data Sets Can Work. IEEE Transactions on Applications and Industry 2002; p. 233 - 238.

16 S.Chen, W.Wang, H.V.Zuylen. Construct support vector machine ensemble to detect traffic incidnt. Expert Systems With Applications - ESWA 2009; 36 p. 10976-10986.

17 กิตติ ภักดีวัฒนะกุล. การออกแบบและพัฒนาคลังข้อมูล = Data warehouse. กรุงเทพฯ: วี.ซี.พี, 2552.

18 สิรินทร์ นิยมางกูร. เทคนิคการสุ่มตัวอย่าง. กรุงเทพฯ สานักพิมพ์มหาวิทยาลัยเกษตรศาสตร์, 2541.

19 ธานินทร์ ศิลป์จารุ. การวิเคราะห์ข้อมูลทางสถิติ SPSS 11st ed. กรุงเทพฯ ธรรมสาร 2553.

20 รวินรัฒน์ จารุกำเนิดกนก. การพยากรณ์มูลค่าการส่งออกอัญมณีและเครื่องประดับโดยวิธีอารีม่า. เชียงใหม่: มหาวิทยาลัยเชียงใหม่, 2552.