การวิเคราะห์และทำนายพฤติกรรมการดาวน์โหลดเสียงเพลงรอสายจากเพศ ช่วงอายุ และภูมิภาค

Analysis and Forecasting of RingBack-Tone Downloading Behavior Based on Genders, Age Ranges and Region

ณัฐวุฒิ แก้วอัสดร[1] และ เทพฤทธิ์ บัณฑิตวัฒนาวงศ์[2]

[1,2]หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีปทุม 61 ถ.พหลโยธิน จตุจักร กรุงเทพ 10900
Email: [1]hunternut@hotmail.com

บทคัดย่อ

งานวิจัยนี้นำเสนอกลุ่มแบบจำลองการวิเคราะห์และทำนายพฤติกรรมการใช้บริการเสียงเพลงรอสายด้วยเทคนิคการทำเหมืองข้อมูล โดยแบบจำลองแรกสร้างด้วยวิธีการจัดหมวดหมู่ประเภทของข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจ เพื่อวิเคราะห์หาปัจจัยที่มีผลต่อการเลือกใช้ประเภทเสียงเพลงรอสาย เช่น เพศ, อายุ และภูมิภาคของผู้ใช้บริการเสียงเพลงรอสาย ซึ่งปัจจัยเหล่านี้ได้ถูกนำเข้าสู่กระบวนการวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอนด้วยหลักการแบ่งกลุ่มแบบเคมีนเพื่อค้นหากลุ่มพฤติกรรมของผู้ใช้บริการเสียงเพลงรอสายสำหรับใช้เป็นแบบจำลองที่สอง ผลการวิจัยโดยใช้ข้อมูลจริงแสดงให้เห็นว่าแบบจำลองแรกสามารถจัดหมวดหมู่ผู้ใช้บริการเสียงเพลงรอสายได้ค่าความถูกต้องสูงสุด 75.09% ส่วนแบบจำลองที่สองสามารถแบ่งกลุ่มผู้ใช้บริการเสียงเพลงรอสายที่ดีที่สุด 2 กลุ่มได้แก่ กลุ่มผู้ใช้งานเสียงเพลงรอสายประเภทพ็อพที่เป็นเพศหญิง อายุอยู่ในช่วง 21-40 ปี และอยู่ที่ภาคกลาง โดยกลุ่มนี้มีขนาดเป็น 67.42% ส่วนแบบจำลองที่สองสามารถแบ่งกลุ่มผู้ใช้เสียงเพลงรอสาย เป็นประเภทคันทรี่ที่เป็นเพศชาย อายุอยู่ในช่วง 21-40 ปี และอยู่ที่ภาคกลาง โดยกลุ่มนี้มีขนาดเป็น 32.58%

คำสำคัญ— เหมืองข้อมูล, ต้นไม้ตัดสินใจ, การแบ่งกลุ่มแบบเคมีน

ABSTRACT

This research presents a set of models to analyze and predict the behavior of ringback-tone downloading with data mining techniques. The first model has been created by means of decision tree-based classification to analyze factors affecting the selection of ringback tone types by customers such as gender, age and region. The factors were brought into a nonhierarchical cluster analysis process based on K-means clustering technique to find out the behavior groups of the customers serving as the second model. The results based on real data sets have showed that the first model can classify the type of the customer with the highest accuracy of 75.09% while the second model can cluster the best two groups : the first group is 67.42% of size contains females between 21 to 40 years-old living in central region used the ringback tones of pop type and the second group is 32.58% of size contains males between 21 to 40 years-old living in central region used ringback tone of country type.

Keywords – Data Mining, Decision Tree, K-Mean Clustering

1. บทนำ

ในปัจจุบันบริการเสียงเพลงรอสายเป็นอีกกลุ่มหนึ่งของบริการเสริมของธุรกิจโทรศัพท์เคลื่อนที่ โดยมีบทบาทสำคัญส่วนหนึ่งในการสร้างรายได้ให้แก่ดีแทคมากกว่า 100 ล้านบาทต่อเดือน ทำให้ผู้วิจัยมีความสนใจที่จะศึกษาเกี่ยวกับพฤติกรรมการดาวน์โหลดเสียงเพลงรอสายของผู้ใช้บริการโทรศัพท์เคลื่อนที่ของบริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) [1] เพื่อวิเคราะห์และทำนายพฤติกรรมการดาวน์โหลดเสียงเพลงรอสาย โดยการนำเทคโนโลยีสารสนเทศมาช่วยในการจัดการวิเคราะห์ข้อมูล และหาปัจจัยที่สำคัญที่ส่งผลต่อพฤติกรรมการดาวน์โหลดเสียงเพลงรอสายของลูกค้า เพื่อนำข้อมูลที่ผ่านการวิเคราะห์จากงานวิจัยมาใช้เป็นประโยชน์สำหรับบริษัทผู้ประกอบการในการพัฒนาบริการ และทราบแนวโน้มพฤติกรรมการดาวน์โหลดเสียงเพลงรอสาย เพื่อนำไปใช้ในการกำหนดกลยุทธ์ให้ตรงกับความต้องการของผู้ใช้บริการให้มากที่สุด

2. ทฤษฎี เทคโนโลยี และงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการประยุกต์ใช้เทคนิคเหมือง เพื่อพัฒนาสร้างตัวแบบการวิเคราะห์ความสัมพันธ์ของปัจจัยต่างๆกับพฤติกรรมการดาวน์โหลดเสียงเพลงรอสายของผู้บริโภค โดยมีการเลือกใช้เทคนิคการทำเหมืองข้อมูล , เทคนิคต้นไม้ตัดสินใจและเลือกใช้การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอนโดยหลักกงานของ K-mean Clustering มาช่วยการจัดกลุ่มข้อมูลการใช้งานของลูกค้าที่มีลักษณะเดียวกันให้อยู่ในกลุ่มเดียวกันและเลือกแบบที่มีประสิทธิภาพสูงสุด

การทำเหมืองข้อมูล คือการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล ซึ่งสัมพันธ์และรูปแบบเหล่านั้นได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากที่มีอยู่เหมืองข้อมูล จะทำการสำรวจและวิเคราะห์ข้อมูลให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์หน่วยนี้แสดงให้เห็นถึงความรู้ต่าง ๆ (Knowledge) ที่มีประโยชน์ในฐานข้อมูลในปัจจุบันองค์กรธุรกิจ ขั้นตอนการทำเหมืองข้อมูลมีดังนี้คือ

1.การทำข้อมูลให้สมบูรณ์ (Data Cleaning) เป็นขั้นตอนสำหรับการทำความสะอาดข้อมูล เป็นกระบวนการตรวจสอบและการแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล ตารางหรือฐานข้อมูล ซึ่งเป็นหลักสำคัญของฐานข้อมูล เพราะหมายถึงความไม่สมบูรณ์ ความไม่ถูกต้อง ความไม่สัมพันธ์กับข้อมูลอื่นๆ เป็นต้นจึงต้องมีการแทนที่ การปรับปรุง หรือการลบข้อมูลที่ไม่ถูกต้องเหล่านี้ออกไป เพื่อให้ข้อมูลมีคุณภาพ [2]

1.1 Ignore the tuple หมายถึง การตัดทิ้งรายการที่มีข้อมูลสูญหาย นิยมใช้กับการทำเหมืองข้อมูลแบบจำแนก ประเภท (classification) ในกรณีที่ค่าคุณลักษณะขาดหายไปเป็นจำนวนมาก

1.2 Fill in the missing value manually หมายถึง วิธีการเติมค่าที่ขาดหายด้วยมือ วิธีนี้ไม่เหมาะสมกรณีที่ชุดข้อมูลขนาดใหญ่และมีข้อมูลขาดหายจำนวนมาก เพราะอาจจะทำให้เกิดความผิดพลาดในข้อมูลที่ใส่ลงไป และเสียเวลาในการทำค่อนข้างมาก

1.3 Use a global constant to fill in the missing value หมายถึง การเติมคุณลักษณะของข้อมูลที่ขาดหายทุกค่า ด้วยค่าคงที่ค่าหนึ่ง เช่น unknown

1.4 Use the attribute mean to fill in the missing value หมายถึง การใช้ค่าเฉลี่ยของคุณลักษณะ เติมค่าข้อมูลที่ขาดหาย เช่น ถ้าทราบว่าลูกค้าที่รายได้เฉลี่ยเดือนละ 15000 บาท จะใช้ค่านี้แทนค่ารายได้ของลูกค้าที่ขาดหาย

1.5 Use the attribute mean for all samples belonging to the same class as the given tuple หมายถึง การใช้ค่าเฉลี่ยคุณลักษณะของตัวอย่างที่จัดอยู่ในประเภทเดียวกัน เพื่อเติมค่าข้อมูลที่ขาดหาย เช่น เติมค่ารายได้ของลูกค้าที่ขาดหายด้วยค่าเฉลี่ยของลูกค้าที่อยู่ในกลุ่มอาชีพเดียวกัน

1.6 Use the most probable value to fill in the missing value หมายถึง การใช้ค่าที่เป็นไปได้มากที่สุด เติมแทนค่าข้อมูลที่ขาดหายไป เช่น ค่าที่ได้จากสมการความถดถอย (Regression) ค่าที่ได้จากการอนุมาน โดยใช้สูตรของ เบย์ (Bayesian formula) หรือต้นไม้ตัดสินใจ (Decision tree) เช่นใช้ข้อมูลลูกค้า มาสร้างต้นไม้ตัดสินใจเพื่อทำนายรายได้ของลูกค้า แล้วนำไปแทนค่าที่ขาดหาย วิธีนี้นิยมกันอย่างแพร่หลาย เนื่องจากทำนายค่าข้อมูลที่ขาดหาย โดยพิจารณาจากค่าของข้อมูลชุดปัจจุบัน และความสัมพันธ์ระหว่างคุณลักษณะ ในชุดข้อมูล

2. การผสานข้อมูล (Data Integration) เป็นการรวบรวมข้อมูลจากแหล่งที่เก็บต่างๆมาไว้ที่เดียวกัน

การผสานข้อมูลจากแหล่งต่างๆ เพื่อช่วย ลดหรือหลีกเลี่ยงความซ้ำซ้อนของข้อมูล (Data Redundancies) ซึ่งจะนำไปสู่ปัญหาความไม่สอดคล้องกันของข้อมูล (Data inconsistencies) และเพิ่มความเร็วและคุณภาพในการทำเหมืองข้อมูล

2.1 การผสานโครงสร้างการเก็บข้อมูล (Schema Integration) โดยใช้ metadata ช่วยในการบ่งชี้คุณลักษณะในแหล่งเก็บข้อมูลต่างๆ เช่น คุณลักษณะ Cusid ในแหล่งข้อมูล A เป็นคุณลักษณะเดียวกับ CustNumber ในแหล่งข้อมูล B หรือไม่

2.2 ตรวจหาและแก้ไข ค่าข้อมูลที่ขัดแย้ง เช่น ค่าคุณลักษณะเดียวกัน แต่ใช้หน่วยวัดต่างๆกัน ในแต่ละแหล่งข้อมูล

2.3 การกำจัดค่าข้อมูลซ้ำซ้อน เช่น ค่าคุณลักษณะในแหล่งข้อมูลหนึ่ง อาจเป็นค่าที่ได้จากการคำนวณในอีกแหล่งข้อมูลหนึ่ง อาทิ ค่ารายได้ รายปี

3. การคัดเลือกข้อมูล (Data Selection) เป็นขั้นตอนการดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกข้อมูล โดยการกำหนดขนาดกลุ่มตัวอย่างหมายถึง จำนวนกลุ่มตัวอย่าง ซึ่งผู้วิจัยจะต้องกำหนดกลุ่มตัวอย่างว่าจะใช้จำนวนเท่าใด เนื่องจากข้อมูลตัวอย่างที่นำมา มีขนาดของกลุ่มตัวอย่างจำนวนมาก

หลักการทฤษฎีของ ทาโร ยามาเน่ (TARO YAMANE) (พวงรัตน์ ทวีรัตน์ 2543 : 285) [3] จะทำให้ผลการวิเคราะห์ข้อมูลมีความเชื่อมั่นสูงโอกาสที่จะเกิดความคลาดเคลื่อนมีน้อยเนื่องจากทฤษฎีนี้ได้ใช้กันมากในงานวิจัยต่างๆ ซึ่งหลักการทฤษฎีของทาโร ยามาเน่ ได้เสนอสูตรการคำนวนขนาดตัวอย่างสัดส่วน1 กลุ่มโดยสมมุติค่าสัดส่วนเท่ากับ 0.5 และที่ระดับความเชื่อมั่น 95 % และทำการสุ่มตัวอย่างด้วยวิธีสุ่มอย่างง่าย

สูตรทาโรยามาเน่

$$\frac{N}{1 + Ne^2}$$

โดยที่

n คือ ขนาดกลุ่มตัวอย่าง

N คือ ขนาดประชากร

e คือ คลาดคลาดเคลื่อนของกลุ่มตัวอย่าง

4. การแปลงข้อมูล (Data Transformation) เป็นการแปลงข้อมูลที่พบบ่อยในการทำเหมืองข้อมูลคือ การทำนอร์มอลไลซ์ (Normalization) โดยแปลงค่าข้อมูลให้อยู่ในช่วงสั้นๆ ที่อัลกอริทึมการทำเหมืองข้อมูลสามารถนำไปใช้ประมวลผลได้

5. เหมืองข้อมูล (Data Mining) เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่ ซึ่งใช้เทคนิคของการทำเหมืองข้อมูลดังต่อไปนี้

การจัดหมวดหมู่ (Classification)

เป็นขั้นการสร้างตัวแบบโดยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาส (Class)ไว้เรียบร้อยแล้วซึ่ง ตัวแบบที่ได้อาจแสดงในรูปดังต่อไปนี้

5.1 เทคนิคต้นไม้ตัดสินใจ (Decision Tree Technique)

การเรียนรู้ของต้นไม้ตัดสินใจ (Decision tree) เป็นการเรียนรู้โดยการจำแนกประเภท (Classification) ข้อมูลออกเป็นกลุ่มคลาส (class) ต่าง ๆ โดยใช้คุณลักษณะ (attribute) ข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่า คุณลักษณะใดเป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญมากน้อยต่างกันอย่างไรโดยต้นไม้ตัดสินใจ เป็นโครงสร้างข้อมูลชนิดเป็นลำดับชั้น (hierarchy) ใช้สนับสนุนการตัดสินใจ โดยจะมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดรากอยู่ด้านบนสุดและโหนดใบอยู่ล่างสุดของต้นไม้ [4]

5.2 การจัดกลุ่มฐานข้อมูล (Database Clustering)

เป็นการลดขนาดข้อมูล (Data reduction) ด้วยเทคนิคที่แม่นยำและมีประสิทธิภาพเป็นอีกแนวทางหนึ่งเพื่อจัด การกับปัญหาดังกล่าว หลักการที่สำคัญของการลดขนาดข้อมูลคือ การทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด เนื่องจากข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีจะ ทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ ข้อมูลที่มีการรวมกลุ่มกันอย่างหนาแน่นจะเป็นข้อมูลที่มีความสำคัญต่อการจัดกลุ่มข้อมูลในอนาคต

6. รูปแบบการประเมินค่า (Pattern Evaluation) เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล ว่าขั้นตอนการประเมินผลของแบบจำลองที่ได้เกิดความผิดพลาดมากน้อยเพียงใด ผลลัพธ์จากการทำนายที่ได้นั้นเป็นไปตามความต้องการหรือไม่ โดยใช้เทคนิควิธีการประเมินความแม่นยำด้วยวิธีการไขว้ข้าม (k-fold cross-validation) เป็นวิธีการในการตรวจสอบค่าความผิดพลาดในการคาดการณ์ของแบบจำลองข้อมูล

7. การแทนความรู้ (Knowledge Representation) เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ เมื่อทำตามขั้นตอนอย่างถูกต้องก็จะส่งผลให้ข้อมูลที่สืบค้นจากเหมืองข้อมูลมีความถูกต้องแม่นยำและสามารถใช้ช่วยสนับสนุนการตัดสินใจจริงได้ตามสิ่งที่ธุรกิจนั้นต้องการ

เทคโนโลยีที่ใช้ในการทำสารนิพนธ์ คือ ซอฟต์แวร์เวก้า ซึ่งเขียนมาเพื่อเน้นกับงานทางด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และการทำเหมืองข้อมูล ซอฟต์แวร์จะประกอบด้วยโมดูลย่อยๆ สำหรับใช้ในการจัดการข้อมูล และเป็นซอฟต์แวร์ที่สามารถใช้ Graphic User Interface (GUI) และใช้คำสั่งในการให้ซอฟต์แวร์ประมวลผล และสามารถประมวลผลได้หลายระบบปฏิบัติการ และสามารถพัฒนาต่อยอดได้ เป็นเครื่องมือที่ใช้ทำงานด้านการทำดาต้าไมน์นิ่งที่รวบรวมแนวคิดขั้นตอนวิธีมากมาย ซึ่งขั้นตอนวิธีสามารถเลือกใช้งานโดยตรงได้จาก 2 ทางคือจาก ชุดเครื่องมือที่มีขั้นตอนวิธีมาให้ หรือเลือกใช้จากขั้นตอนวิธีที่ได้เขียนเป็นซอฟต์แวร์ลงไปเป็นชุดเครื่องมือเพิ่มเติม และชุดเครื่องมือมีฟังก์ชันสำหรับการทำงานร่วมกับข้อมูลได้แก่ การเตรียมข้อมูล (Pre-Processing), การทำเหมืองข้อมูลด้วยเทคนิคการจำแนกข้อมูล(Classification) , การทำเหมืองข้อมูลด้วยเทคนิคการจัดกลุ่ม(Clustering), การทำเหมืองข้อมูลด้วยเทคนิคการวิเคราะห์ความสัมพันธ์(Association rules), เทคนิคการคัดเลือกข้อมูล(Selection Attributes) และ เทคนิคการนำเสนอข้อมูลด้วยรูปภาพ(Visualization)

ซึ่งผู้วิจัยจะนำซอฟต์แวร์เวก้ามาช่วยในด้านเทคนิคการทำเหมืองข้อมูล แบบการจำแนก เทคนิคการจำแนกที่ใช้คือ J48 เป็นขั้นตอนวิธีในการจำแนกโดยใช้ต้นไม้ตัดสินใจ C4.5 ซึ่งจะใช้สำหรับการสร้างต้นไม่ตัดสินใจ โดยแต่ละโหนด หมายถึงแอททริบิวต์ แต่ละกิ่งของต้นไม้เป็นผลในการทดสอบและโหนดใบแสดงคลาส เพื่อการวิเคราะห์และทำนายพฤติกรรมการดาวน์โหลดเสียงเพลงรอสายจากเพศ ช่วงอายุ และภูมิภาค

งานวิจัยที่เกี่ยวข้อง

วาลุกา วงษ์กราน (2552) ศึกษาเรื่อง ความพึงพอใจของลูกค้า AIS ต่อการใช้บริการเพลงรอสาย (Calling Melody) กลุ่มตัวอย่างที่ใช้ในการวิจัยคือ พนักงานบริษัทและลูกค้า อาคารชินวัตร 2 เขต พญาไท จำนวน 200 คน โดยใช้แบบสอบถามเป็นเครื่องมือในการเก็บข้อมูล ตัวแปรอิสระ คือ สถานภาพส่วนบุคคลของประชากรที่ใช้ในการวิจัย ได้แก่ เพศ อายุ ระดับการศึกษา อาชีพ รายได้เฉลี่ยต่อเดือน และสถานภาพสมรส ตัวแปรตาม คือ พฤติกรรมการใช้บริการเพลงรอสายของลูกค้า AIS และความพึงพอใจต่อการใช้บริการเพลงรอสาย ของลูกค้า AIS สถิติที่ใช้ คือ ใช้ค่าเฉลี่ย X ส่วนเบี่ยงเบนมาตรฐาน (S.D) ไคแสควส์ ค่า t-test, F-test, One-Way ANOVA โดยใช้โปรแกรมสำเร็จรูป

ผลการวิจัยพบว่า ผู้ใช้บริการส่วนใหญ่เป็นเพศหญิง มีอายุ 26-30 ปี มีระดับการศึกษาปริญญาตรี หรือเทียบเท่า มีอาชีพเป็นพนักงานบริษัทเอกชน มีรายได้เฉลี่ยต่อเดือน 20,001-30,000 บาท และสภานภาพเป็นโสด ระบบ SIM ทั้ง GSM Advance และ 1-2 Call ปัจจุบันนิยมใช้พอๆ กัน การใช้บริการดาวน์โหลดนิยมอยู่ในรูปแบบเสียงรอสาย ความถี่ในการดาวน์โหลดเสียงรอสายโดยเฉลี่ยต่อเดือนอยู่ในความถี่นานๆ ครั้ง ขึ้นอยู่กับเพลงที่ชอบ ช่องทางการใช้บริการดาวน์โหลดเสียงรอสายที่นิยมมากที่สุดคือผ่านทางเว็บไซต์ www.mobilelife.co.th วัตถุประสงค์ของการใช้บริการดาวน์โหลดเสียยงรอสายเพื่อความเพลิดเพลิน/ผ่อนคลาย ผู้ใช้บริการให้ระดับความสำคัญเกี่ยวกับปัจจัยด้านส่วนประสมทางการตลาดบริการ โดยรวมอยู่ในระดับพึงพอใจมาก เมื่อพิจารณาจากหลายด้าน พบว่า ด้านผลิตภัณฑ์ด้านช่องทางการจัดจำหน่าย ด้านบุคลากร ด้านลักษณะทางกายภาพ และด้านกระบวนการ อยู่ในระดับมาก และทางด้านราคา ด้านการส่งเสริมการตลาดอยู่ในระดับพอใจปานกลาง

ธัญพร ชัยสุขโกศล (2553) ศึกษาเรื่อง การวิเคราะห์ข้อมูลผู้บริโภคเพื่อทำนายพฤติกรรมการซื้อสินค้า กรณีศึกษา “ร้านค้าสวัสดิการ” งานวิจัยนี้นำเสนอการวิเคราะห์ความสัมพันธ์ของโรคไขมันและหัวใจกับพฤติกรรมผู้บริโภคในการเลือกซื้อสินค้าโดยประเมินค่าจากการวัดประเภทระดับไขมันในเลือด 4 ประเภทคือ HDL, LDL, คลอเรสเตอรอล และไตรกลีเซอไรด์ โดยใช้สองเทคนิคในการทำเหมืองข้อมูล ได้แก่ เทคนิคกฎความสัมพันธ์ เพื่อหาความสัมพันธ์ของผู้บริโภคที่เกิด

ความเสี่ยงต่อโรคกับพฤติกรรมการซื้อสินค้า และการแบ่งกลุ่มเพื่อจัดกลุ่มผู้บริโภคจากลักษณะประจำหรือพฤติกรรมผู้บริโภค จากการทดลองพบว่าการใช้เทคนิคกฎความสัมพันธ์สามารถทำนายพฤติกรรมและลักษณะผู้บริโภคที่เกิดความเสี่ยงต่อโรคได้ถูกต้อง 74% สำหรับการแบ่งกลุ่มผู้บริโภคนั้นใช้วิธีการแบ่งกลุ่มแบบ 2 ขั้นตอน คือใช้ Hierarchical Clustering เพื่อหาค่า k ที่เหมาะสมที่สุด หลังจากนั้นจึงนำค่า k มากำหนดใน K-means Clustering ผลการทดลองสามารถแบ่งกลุ่มผู้บริโภคได้ดีที่สุด 2 กลุ่ม กลุ่มแรกเป็นกลุ่มที่มีการศึกษาระดับมัธยมศึกษาตอนปลายถึงประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) เพศชายอายุ 31-35 ปี เป็นพนักงานระดับหัวหน้าไลน์ผลิต, เสมียน, ช่างเทคนิค มักซื้อสินค้าประเภทน้ำผลไม้ มีความเสี่ยงต่อการเกิดโรค กลุ่มสุดท้ายเป็นกลุ่มที่มีการศึกษาระดับมัธยมศึกษาตอนปลายถึงประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) เพศหญิงอายุ 26-30 ปีเป็นพนักงานฝ่ายผลิต, แม่บ้าน, คนสวน มักซื้อสินค้าประเภทน้ำผลไม้ มีความเสี่ยงต่อการเกิดโรค

3. วิธีการพัฒนา

การศึกษาการวิเคราะห์พัฒนาสร้างตัวแบบความสัมพันธ์ของปัจจัยต่างๆ เพื่อวิเคราะห์และทำนายพฤติกรรมการดาวน์โหลดเสียงเพลงรอสายจากเพศ ช่วงอายุ และภูมิภาค ผู้วิจัยได้กำหนดขั้นตอนการดำเนินงานประกอบด้วย 3 ขั้นตอนดังนี้

3.1 การรวบรวมข้อมูลที่เกี่ยวข้องกับการวิจัย

3.2 การเตรียมข้อมูลประชากรและกลุ่มตัวอย่าง

3.3 กำหนดขั้นตอนวิธีที่ใช้ในดำเนินการวิจัย

3.1 การรวบรวมข้อมูลที่เกี่ยวข้องกับการวิจัย

ผู้วิจัยได้รวมรวมข้อมูลที่เกี่ยวข้องสำหรับการวิจัย ประกอบด้วยฐานข้อมูล ที่ใช้ในการวิเคราะห์ ผ่านกระบวนการคลังข้อมูล เพื่อให้ได้ข้อมูลที่เหมาะสมต่อการวิจัย ดังต่อไปนี้

3.1.1 ฐานข้อมูลรายละเอียดเพลง เป็นการเก็บข้อมูลชื่อเพลงและประเภทเพลง ทั้งหมดที่มีอยู่ในระบบ

3.1.2 ฐานข้อมูลลูกค้า เป็นการเก็บประวัติส่วนตัวของลูกค้า เช่น เบอร์โทรศัพท์ลูกค้า ,อายุ ,เพศ, ภูมิภาค

3.1.3 ฐานข้อมูลกล่องเพลงของลูกค้า เป็นการเก็บเลขอ้างอิงเพลง ที่ลูกค้ามีในระบบ

ภาพประกอบที่ 1.แหล่งที่มาของข้อมูลพฤติกรรมของผู้ใช้บริการเสียงเพลงรอสาย

3.2 การเตรียมข้อมูลประชากรกลุ่มตัวอย่าง

การเตรียมข้อมูลประชากรและกลุ่มตัวอย่างเพื่อนำมาใช้สร้างตัวแบบการวิเคราะห์ความสัมพันธ์ของปัจจัยต่างๆ กับพฤติกรรมการดาวน์โหลดเสียงเพลงรอสาย ผู้วิจัยได้นำข้อมูลการดาวน์โหลดเสียงเพลงรอสายของบริษัทผู้ให้บริการโทรศัพท์เคลื่อนรายหนึ่งมาใช้ในการวิเคราะห์ โดยนำข้อมูล ณ. วันที่ 18 กรกฎาคม 2557 โดยประกอบด้วย 4 ข้อมูลปัจจัย คือ ข้อมูล อายุ, เพศ, ภูมิภาค, ประเภทเพลง การวิเคราะห์เพื่อให้ได้ข้อมูลที่เหมาะสมกับเทคนิคที่ใช้ในการนำมาวิเคราะห์ แบ่งเป็น 5 ขั้นอตนย่อยๆ ดังนี้

3.2.1 การเตรียมข้อมูล

ในงานวิจัยนี้ผู้วิจัยได้คัดเลือกเฉพาะข้อมูลในส่วนที่จำเป็นสำหรับงานวิจัย เพื่อให้สะดวกต่อการนำมาใช้งาน เช่น อายุ, เพศ, ภูมิภาค, ประเภทเสียงเพลงรอสาย ซึ่งเป็นข้อมูลตั้งแต่ ปี 2548 – 2557 จำนวน 192,924 ระเบียน

3.2.2 การทำข้อมูลให้สมบูรณ์

การทำข้อมูลลูกค้าเสียงเพลงรอสายให้สมบูรณ์ โดยการเติมค่าของข้อมูลให้สมบูรณ์ เช่น ชื่อจังหวัดที่เขียนไม่ครบ ให้เติมข้อมูลให้สมบูรณ์, อายุลูกค้าที่เป็นคริสต์ศักราชนำไปบวก 543 ปีและลบด้วยพุทธศักราชปัจจุบัน จะได้อายุปัจจุบันของลูกค้า, และอายุลูกค้าที่ปีเกิดเป็นพุทธศักราชอยู่แล้วให้มา มาทำการลบด้วยพุทธศักราชปัจจุบัน จึงจะได้อายุของลูกค้า

3.2.3 การผสานข้อมูล

การรวบรวมข้อมูลผู้ดาวน์โหลดเสียงเพลงรอสายจากหลายๆ แหล่งข้อมูล ให้มาเป็นข้อมูลชุดเดียว เพื่อให้สะดวกในการนำมาใช้งาน และช่วยลดความซ้ำซ้อนของข้อมูล และยังเพิ่มความเร็วในการทำเหมืองข้อมูล

3.2.4 การคัดเลือกข้อมูล

เนื่องจากข้อมูลที่นำมาใช้สำหรับงานวิจัยนี้ มาจากแหล่งที่มีข้อมูลเป็นจำนวนมาก ซึ่งผู้วิจัยจึงดึงจำนวนของกลุ่มตัวอย่างมาเพื่อใช้

ในการวิเคราะห์ข้อมูล โดยใช้หลักการของทฤษฎี ทาโร ยามาเน่ จะทำให้ผลการวิเคราะห์ข้อมูลมีความเชื่อมั่นสูงโอกาสที่จะเกิดความคลาดเคลื่อนมีน้อย เนื่องจากทฤษฎีนี้ได้ใช้กันมากในงานวิจัยต่างๆ การคำนวน ขนาดตัวอย่างสัดส่วน 1 กลุ่ม โดยสมมุติค่าสัดส่วนเท่ากับ 0.5 และที่ระดับความเชื่อมั่น 95 %

ในการวิจัยครั้งนี้มีข้อมูลลูกค้าเสียงเพลงรอสาย จำนวน 192,924 ระเบียน โดยยอมให้เกิดความคลาดเคลื่อนร้อยละ 5 จากสูตรการคำนวนของทาโร ยามาเน่ สามารถ คำนวนได้ดังด้านล่างนี้

N = 192,924

E = 0.05

$$N = \frac{192,924}{1+192,924\ (0.05)^2}$$

$$= \frac{192,924}{1+482.31}$$

$$= 399.17$$

กลุ่มตัวอย่างที่ใช้ในการวิจัย = 399 คน

3.2.5 การแปลงข้อมูล มีขั้นตอนดังนี้

การปรับเปลี่ยนรูปแบบข้อมูล ผู้วิจัยปรับเปลี่ยนรูปแบบข้อมูลโดยการนำข้อมูลมาเชื่อมต่อกัน ด้วยเทคนิคการทำคลังข้อมูลและเปลี่ยนรูปแบบข้อมูลให้ครอบคลุมตัวชี้วัดที่กำหนด โดยงานวิจัยนี้เลือกใช้ Oracle 11g มาใชในการออกแบบคลังข้อมูล

3.2.6 การลดขนาดของข้อมูล ผู้วิจัยได้ทำการลดขนาดของข้อมูลโดย การปรับเปลี่ยนรูปแบบข้อมูลให้ง่ายต่อการวิเคราะห์ โดยการเลือกข้อมูลเฉพาะที่สนใจ และทำการจัดกลุ่ม เพื่อให้เหมาะสมต่องานวิจัย จนได้ข้อมูลที่จะนำไปวิเคราะห์โดยซอฟต์แวร์

3.3 ขั้นตอนการดำเนินการวิจัย

งานวิจัยนี้ใช้เทคนิคการทำเหมืองข้อมูลแบบการจัดหมวดหมู่ โดยใช้เทคนิคต้นไม้ตัดสินใจโดยมีขั้นตอนและรายละเอียดดังนี้

3.3.1 เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจผู้วิจัยจะนำข้อมูลลูกค้าจำนวน 192,924 ระเบียน และใช้สูตรคำนวนจำนวนตัวอย่างของ ทาโร ยามาเน่ และได้ผลของกลุ่มตัวอย่างจำนวนคือ 399 คน เป็นกลุ่มทดสอบประสิทธิภาพ โดยมีขั้นตอนการทำงานดังนี้

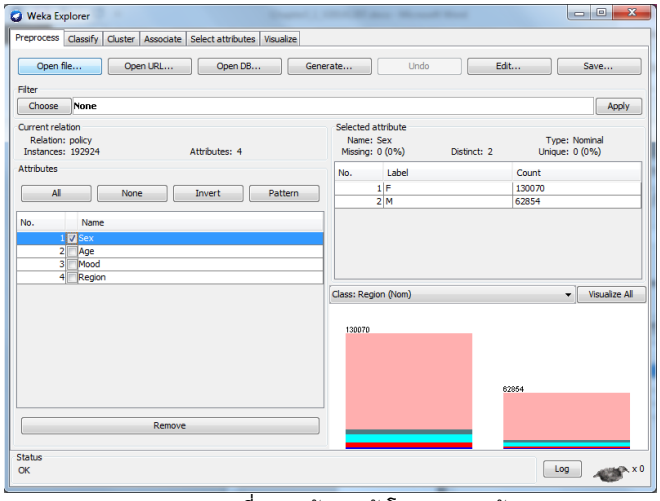
ภาพประกอบที่ 2 ลำดับขั้นตอนการดำเนินงานด้วยเทคนิคต้นไม้ตัดสินใจ

ในต้นไม้ตัดสินใจนี้ทางผู้วิจัยได้ทำการเตรียมข้อมูลในรูปแบบไฟล์นามสกุล ARFF และจึงนำไปทดสอบผ่านซอฟต์แวร์เวก้าเวอร์ชั่น 3.6.11 [5] เป็นเครื่องมือช่วยในการทดสอบดังภาพประกอบที่ 3

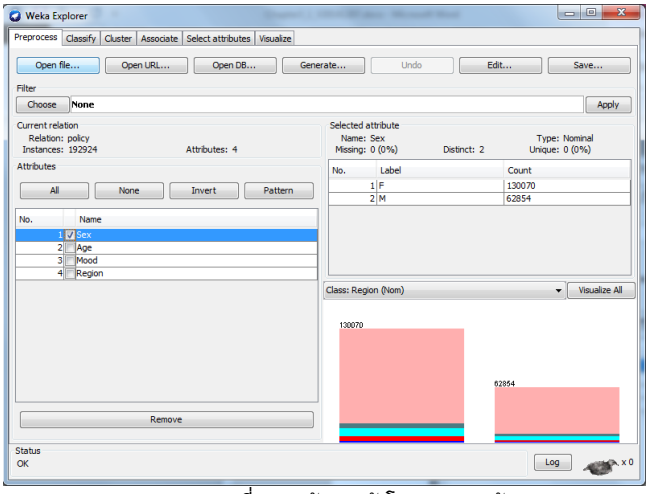
```
@relation policy
@attribute 'Sex' {F,M}
@attribute 'Age' {1,2,3,4,5}
@attribute 'Mood' {1,2,3,4,5,6,7,8,9}
@attribute 'Region' {Northeast,East,South,North,Central,West}

@data

F,1,5,Central
M,1,5,Central
F,1,6,Central
F,1,6,Central
F,1,6,Central
F,1,6,Central
F,1,6,Central
F,1,5,Central
F,1,5,Central
F,1,5,Central
M,1,6,Central
M,1,6,Central
F,1,6,Central
M,1,6,Central
F,1,6,Central
M,1,5,Central
F,1,5,Central
```

รูปภาพประกอบที่ 3 การเตรียมข้อมูลในรูปแบบไฟล์ ARFF

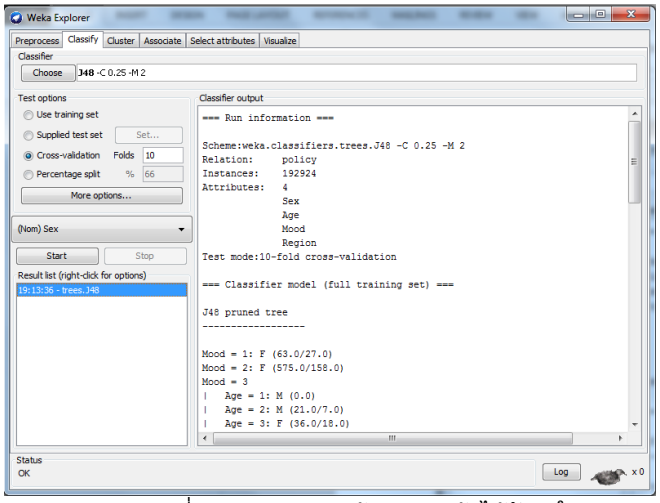
หลังจากได้เตรียมข้อมูลเสร็จแล้ว ผู้วิจัยจะนำมาเข้าสู่โปรแกรมเวก้า เพื่อทำการประมวลผลด้วยเทคนิค ต้นไม้ตัดสินใจ ผลลัพธ์ที่ได้ออกมาดังรูปภาพ ที่ 4 และ 5



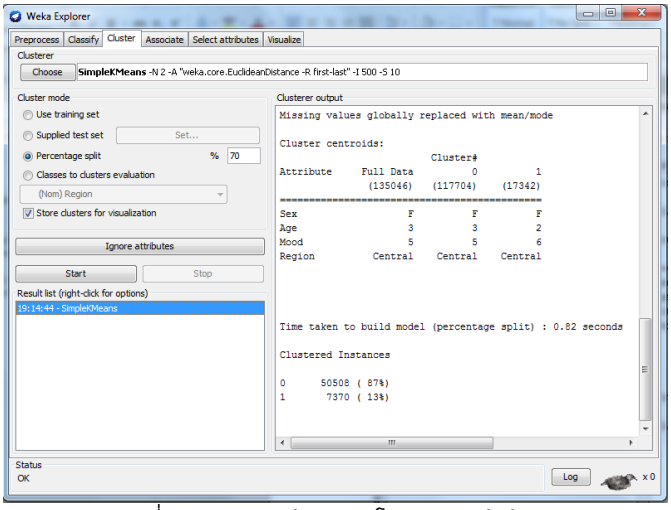
ภาพประกอบที่ 4 นำข้อมูลเข้าโปรแกรมเวก้า



ภาพประกอบที่ 7 นำข้อมูลเข้าโปรแกรมเวก้า



ภาพประกอบที่ 5 การประมวลผลด้วยเทคนิคต้นไม้ตัดสินใจ



ภาพประกอบที่ 8 ประมวลผลข้อมูลผ่านโปรแกรมเวก้าด้วยเทคนิค SimpleKmeans

3.3.2 การจัดกลุ่มฐานข้อมูล

ผู้วิจัยได้ใช้การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน โดยการใช้หลักการของ K-Means Clustering กำหนดกลุ่มที่เหมาะสม โดยผู้วิจัยเลือกใช้โปรแกรมเวก้า 3.6.11 เป็นเครื่องมือในการทดลอง ดังภาพประกอบที่ 6,7 และ 8

```
1  @relation policy
2  @attribute 'Sex' {F,M}
3  @attribute 'Age' {1,2,3,4,5}
4  @attribute 'Mood' {1,2,3,4,5,6,7,8,9}
5  @attribute 'Region' {Northeast,East,South,North,Central,West}
6
7  @data
8
9  F,1,5,Central
10 M,1,5,Central
11 F,1,6,Central
12 F,1,6,Central
13 F,1,6,Central
14 F,1,6,Central
15 F,1,6,Central
16 F,1,5,Central
17 F,1,5,Central
18 F,1,5,Central
19 M,1,6,Central
20 M,1,6,Central
21 F,1,6,Central
22 M,1,6,Central
23 F,1,6,Central
24 M,1,5,Central
25 F,1,5,Central
```

ภาพประกอบที่ 6 ขั้นตอนการเตรียมข้อมูลก่อนเข้าโปรแกรมเวก้า

4. ผลการพัฒนา

ผลการวิจัยจากการที่นำข้อมูลที่มีจำนวน 192,924 รายการ มาทำการทดสอบแบบจำลองต้นไม้ตัดสินใจ โดยการนำเข้าการประมวลผลในโปรแกรม Weka และผลการทดลองสรุปได้ว่ามีความถูกต้องที่ 75.09% โดยมีผลการทดลองตามด้านล่างนี้

ผลการทดลองชุดข้อมูลจากโปรแกรม Weka

Age = 1

| Region = Northeast: 5 (506.0/136.0)

| Region = East: 5 (531.0/144.0)

| Region = South: 5 (902.0/268.0)

| Region = North: 5 (140.0/42.0)

| Region = Central: 6 (9929.0/3807.0)

| Region = West: 5 (7.0/1.0)

Age = 2: 5 (142364.0/33627.0)

Age = 3: 5 (36609.0/9436.0)

Age = 4: 5 (1936.0/596.0)

อธิบายผลการทดลองของชุดข้อมูล

บรรทัดที่ 1

Age = 1 : ลูกค้าช่วงอายุ 1-20ปี

บรรทัดที่ 2

| Region = Northeast: 5 (506.0/136.0) : ลูกค้าช่วงอายุ 1-20 ปี ในภาคตะวันออกเฉียงเหนือ ใช้ประเภทเสียงเพลงรอสายเป็นแนว Pop มากที่สุดจำนวน 370 คนและแนวอื่นๆอีก 136 คน

บรรทัดที่ 3

| Region = East: 5 (531.0/144.0) : ลูกค้าช่วงอายุ 1-20 ปี ในภาคตะวันออก ใช้ประเภทเสียงเพลงรอสายเป็น แนว Pop มากที่สุดจำนวน 387 คนและแนวอื่นๆ อีก 144 คน

บรรทัดที่ 4

| Region = South: 5 (902.0/268.0) : ลูกค้าช่วงอายุ 1-20 ปี ในภาคใต้ ใช้ประเภทเสียงเพลงรอสายเป็น แนว Pop มากที่สุดจำนวน 634 คนและแนวอื่นๆ อีก 268 คน

บรรทัดที่ 5

| Region = North: 5 (140.0/42.0) : ลูกค้าช่วงอายุ 1-20 ปี ในภาคเหนือ ใช้ประเภทเสียงเพลงรอสายเป็น แนว Pop มากที่สุดจำนวน 98 คนและแนวอื่นๆ อีก 42 คน

บรรทัดที่ 6

| Region = Central: 6 (9929.0/3807.0) : ลูกค้าช่วงอายุ 1-20 ปี ในภาคกลาง ใช้ประเภทเสียงเพลงรอสายเป็น แนว Country มากที่สุดจำนวน 6,122 คนและแนวอื่นๆ อีก 3,807 คน

บรรทัดที่ 7

| Region = West: 5 (7.0/1.0) : ลูกค้าช่วงอายุ 1-20 ปี ในภาคตะวันตก ใช้ประเภทเสียงเพลงรอสายเป็น แนว Country มากที่สุดจำนวน 6 คนและแนวอื่นๆ อีก 1 คน

บรรทัดที่ 8

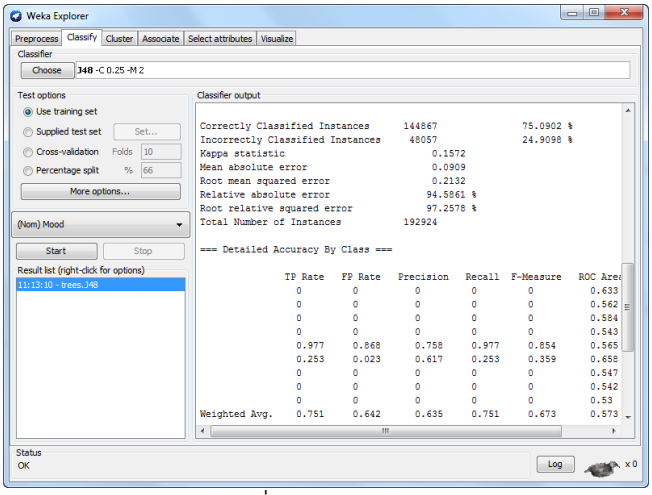
Age = 2: 5 (142364.0/33627.0) : ลูกค้าช่วงอายุ 21-40 ปีของทุกภาคใช้เสียงเพลงรอสายประเภท Pop จำนวน 108,737 คนและเป็นเสียงเพลงรอสายประเภทอื่นอีก 33,627 คน

บรรทัดที่ 9

Age = 3: 5 (36609.0/9436.0) : ลูกค้าช่วงอายุ 41-60 ปีของทุกภาคใช้เสียงเพลงรอสายประเภท Pop จำนวน 27,173 คนและเป็นเสียงเพลงรอสายประเภทอื่นอีก 9,436 คน

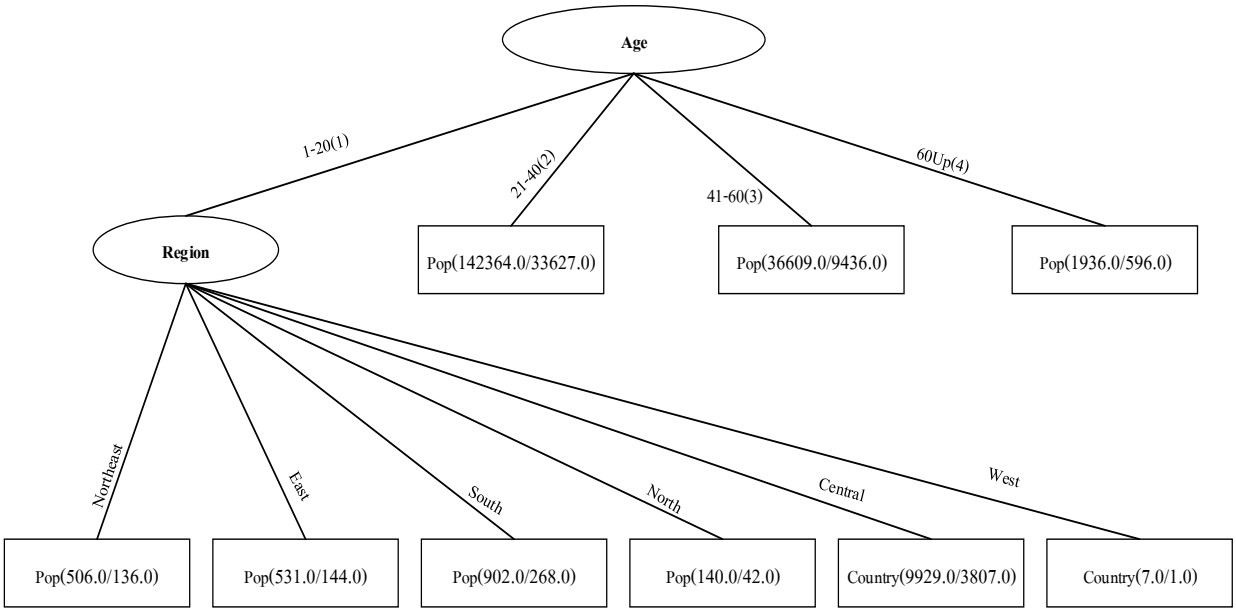
บรรทัดที่ 10

Age = 4: 5 (1936.0/596.0): ลูกค้าช่วงอายุ 60 ปีขึ้นไป ของทุกภาคใช้เสียงเพลงรอสายประเภท Pop จำนวน 1,340 คนและเป็นเสียงเพลงรอสายประเภทอื่นอีก 596 คน



ภาพประกอบที่ 9 ผลการวิจัยกับชุดข้อมูล

จากผลการวิจัยกับชุดข้อมูล สามารถวาดออกมาเป็นต้นไม้ตัดสินใจได้ดังนี้



ภาพประกอบที่ 10 แบบจำลองต้นไม้ตัดสินใจจากชุดข้อมูล ของพฤติกรรมการใช้เสียงเพลงรอสาย

5. สรุป

งานวิจัยนี้นำเสนอแบบจำลองสำหรับทำนายพฤติกรรมการเลือกใช้ประเภทเสียงเพลงรอสาย ผลการทำนายพฤติกรรมการเลือกใช้ประเภทเสียงเพลงรอสาย ได้ค่าความถูกต้องสูงสุดคือ 75.09 ซึ่งทำนายถูก 75.09% โดยสามารถอธิบายผลการทดลองได้ตามด้านล่างนี้คือ

ผู้ใช้เสียงเพลงรอสายช่วงอายุ 1-20 ปี

- ผู้ใช้งานเสียงเพลงรอสายที่อยู่ในภาคตะวันออกเฉียงเหนือจะใช้งานเสียงเพลงรอสายประเภท Pop จำนวน 370 คนและแนวอื่นๆอีก 136 คน
- ผู้ใช้งานเสียงเพลงรอสายที่อยู่ในภาคตะวันออกจะใช้งานเสียงเพลงรอสายประเภท Pop จำนวน 387 คนและแนวอื่นๆ อีก 144 คน
- ผู้ใช้งานเสียงเพลงรอสายที่อยู่ในภาคใต้จะใช้งานเสียงเพลงรอสายประเภท Pop จำนวน 634 คนและแนวอื่นๆ อีก 268 คน
- ผู้ใช้งานเสียงเพลงรอสายที่อยู่ในภาคเหนือจะใช้งานเสียงเพลงรอสายประเภท Pop จำนวน 98 คนและแนวอื่นๆ อีก 42 คน
- ผู้ใช้งานเสียงเพลงรอสายที่อยู่ในภาคกลางจะใช้เสียงเพลงรอสายประเภท Country จำนวน 6,122 คนและแนวอื่นๆ อีก 3,807 คน
- ผู้ใช้งานเสียงเพลงรอสายที่อยู่ในภาคตะวันตกจะใช้เสียงเพลงรอสายประเภท Country จำนวน 6 คนและแนวอื่นๆ อีก 1 คน

ผู้ใช้เสียงเพลงรอสายช่วงอายุ 21-40 ปี

- ผู้ใช้งานเสียงเพลงรอสายของทุกภาคใช้เสียงเพลงรอสายประเภท Pop จำนวน 108,737 คนและเป็นเสียงเพลงรอสายประเภทอื่นอีก 33,627 คน

ผู้ใช้เสียงเพลงรอสายช่วงอายุ 41-60 ปี

- ผู้ใช้งานเสียงเพลงรอสายของทุกภาคใช้เสียงเพลงรอสายประเภท Pop จำนวน 27,173 คนและเป็นเสียงเพลงรอสายประเภทอื่นอีก 9,436 คน

ผู้ใช้เสียงเพลงรอสายที่อายุมากกว่า 60 ปีขึ้นไป

- ผู้ใช้งานเสียงเพลงรอสายของทุกภาคใช้เสียงเพลงรอสายประเภท Pop จำนวน 1,340 คนและเป็นเสียงเพลงรอสายประเภทอื่นอีก 596 คน

เอกสารอ้างอิง

[1] ข้อมูลผู้ใช้บริการโทรศัพท์เคลื่อนที่ของบริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน), สืบค้นเมื่อ 18 กรกฎาคม 2557

[2] Jiawei Hun and Micheline Kamber. (2001). Data Mining Concept and Techniques. United Stated of America.

[3] Yamane, Taro (1967). Statistics, An Introductory Analysis, 2nd Ed., New York: Harper and Row.

[4] Quinlan, J.R. (1986). Induction of decision trees, pp. 81-106, Machine Learning 1.

[5] The University of Waikato. WEKA. [ออนไลน์]. สืบค้นเมื่อ 23 กรกฎาคม 2557 จาก http://www.cs.waikato.ac.nz/ml/weka/