การพัฒนาแบบจำลองเพื่ออนุมานรูปแบบกิจกรรมของการเดินทางโดยการใช้ระบบปัญญาประดิษฐ์
วิชา โครงงานทางวิศวกรรมโยธา
Development of Activity Inference Model by Machine Learning Method

กษมวัต จงพิพัฒน์วณิชย์[1] ขุนเดช พันธุโพธิ์[2] สุทิวัส จิตโสภักตร์[3] และ ผศ.ดร. พงษ์สันธ์ บัณฑิตสกุลชัย[3]

[1,2,3] ภาควิชาวิศวกรรมโยธา คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย จ.กรุงเทพฯ

บทคัดย่อ

ผู้คนเดินทางเพื่อทำกิจกรรมที่จุดหมาย โดยเฉพาะในกรุงเทพมหานครและปริมณฑล ซึ่งช่วยกระตุ้นเศรษฐกิจและปรับปรุงคุณภาพชีวิต การเข้าใจรูปแบบการเดินทางช่วยในการบริหารระบบขนส่งและวางแผนอนาคต ปัจจุบัน สำนักงานนโยบายและแผนการขนส่งและจราจรทำการสำรวจการเดินทางในกรุงเทพฯ ซึ่งมีค่าใช้จ่ายสูง งานวิจัยนี้จึงใช้ปัญญาประดิษฐ์เรียนรู้จากข้อมูลสำรวจเพื่ออนุมานวัตถุประสงค์การเดินทางจากข้อมูลสัญญาณโทรศัพท์ โดยใช้อัลกอริทึม Bayesian Network และ Random Forest พบว่าทั้งสองแบบจำลองมีความแม่นยำประมาณ 90% กับข้อมูลสำรวจ โดยแบบจำลอง Random Forest มีความแม่นยำมากกว่า Bayesian Network จากนั้นจึงนำแบบจำลองไปอนุมานวัตถุประสงค์จากชุดข้อมูลสัญญาณโทรศัพท์ ถึงแม้จะได้แบบจำลองที่มีความแม่นยำสูง แต่การนำไปใช้ต่อเนื่องนั้นต้องคำนึงถึงปัจจัยหลายประการเช่น ลักษณะและรูปแบบของข้อมูล รวมไปถึงวิธีการได้มาซึ่งข้อมูลนั้นด้วยว่ามีความแตกต่างกันมากน้อยเพียงใด และต้องมีการศึกษาเพิ่มเติมเพื่อชุดข้อมูลที่มีความแม่นยำกว่ามายืนยันผลของแบบจำลองนี้ต่อไป

คำสำคัญ: ข้อมูลการเดินทาง, อนุมานกิจกรรมการเดินทาง, ระบบปัญญาประดิษฐ์, แบบจำลองต้นไม้แบบสุ่ม, ทฤษฎีความน่าจะเป็นแบบเบย์

Abstract

People travel to engage in activities at their destinations, especially in densely populated areas like Bangkok, stimulating economic growth and improving quality of life. Understanding travel patterns can enhance transportation management and planning. regularly conducts travel surveys in Bangkok, but these are costly and time-consuming. This research investigates using machine learning to infer travel activities from mobile phone signal data, employing Bayesian Network and Random Forest algorithms. While both models achieve about 90% accuracy, the continued use must take factors such as data format and method of data processing. And additional studies are needed to obtain a more accurate model to evaluate this model's performance.

Key words: Travel data, Activity inference, Artificial intelligence, Random Forest model, Bayesian Network theory

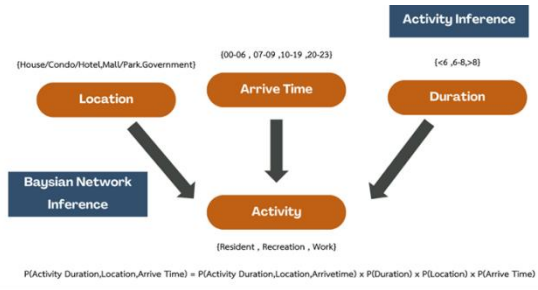
1. บทนำ

กรุงเทพมหานครและปริมณฑลเป็นพื้นที่ที่มีกิจกรรมทางเศรษฐกิจสูงที่สุดแห่งหนึ่งในประเทศไทย เนื่องจากเป็นแหล่งรวมศูนย์กลางของหน่วยงานราชการ, สถาบันการเงิน, สถาบันการศึกษา และ แหล่งความบันเทิงชั้นนำ ด้วยเหตุนี้เองจึงทำให้ผู้คนจากทั่วประเทศเข้ามารวมตัวกันในพื้นที่กรุงเทพมหานครและปริมณฑลเพื่อประกอบอาชีพและทำธุรกิจ พร้อมทั้งบริโภคสินค้าและบริการ ซึ่งก่อให้เกิดกิจกรรมและความต้องการเดินการเดินทางขึ้นอย่างมาก ดังนั้น สำนักงานนโยบายและแผนการขนส่งและจราจร (สนข.) จึงได้จัดทำ ชุดข้อมูลความต้องการการเดินทางของแบบจำลองด้านการขนส่งและจราจรระดับกรุงเทพมหานครและปริมณฑล (eBUM) เพื่อใช้ในการวางแผนและตัดสินใจในการแก้ไขปัญหาด้านการขนส่งและจราจรโดยมีข้อมูลประกอบด้วย เพื่อปรับเทียบแบบจำลองด้านการขนส่งและจราจรระดับกรุงเทพมหานครและปริมณฑล (extended

Bangkok Urban Area Model: eBUM) ให้มีความทันสมัยและแม่นยำ สอดคล้องกับพฤติกรรมและลักษณะการเดินทางที่เปลี่ยนแปลงไป สำนักงานนโยบายและแผนการขนส่งและจราจร (สนข.) จึงได้ดำเนินการสำรวจความต้องการการเดินทางและจัดทำฐานข้อมูลเพื่อใช้ในการวิเคราะห์และปรับปรุงแบบจำลองด้านการขนส่งและจราจรระดับกรุงเทพมหานครและปริมณฑล (Bangkok Travel Demand Survey) เป็นประจำทุกประมาณ 5 ปี เพื่อให้ข้อมูลมีความทันสมัยต่อการเปลี่ยนแปลงในปัจจุบัน โดยปีล่าสุดที่จัดทำคือประจำปี 2565 และอีกโครงการหนึ่งของ สนข. ก็คือ โครงการศึกษาพัฒนานวัตกรรมระบบวิเคราะห์ฐานข้อมูลขนาดใหญ่ (Big Data Analytics) เพื่อการขนส่งสินค้าด้วยรถบรรทุกและการเดินทางของคนในเขตกรุงเทพมหานครและปริมณฑล ทั้งสองโครงการมีวัตถุประสงค์ร่วมกันก็คือเพื่อนำข้อมูลมาวางแผนด้านคมนาคมให้มีประสิทธิภาพสูงสุดเพื่อเพิ่มขีดความสามารถในการแข่งขันของประเทศ ถึงแม้ว่าทั้ง 2 โครงการจะมีวัตถุประสงค์แบบเดียวกัน แต่วิธีที่นำมาซึ่งผลลัพธ์นั้นมีความแตกต่างกันอย่างมาก โครงการ Bangkok Travel Demand Survey ใช้วิธีการเก็บข้อมูลจริง (Ground truth) จากการลงพื้นที่สำรวจแบบสอบถามทั้งรูปแบบ Onsite และ Online ทำให้ได้ข้อมูลที่มีความถูกต้อง แม่นยำ และน่าเชื่อถือสูง แต่ก็มีต้นทุนในการดำเนินการที่สูงเช่นกัน ทั้งเวลา บุคลากรและค่าใช้จ่าย ทำให้ไม่สามารถที่จะจัดทำได้เป็นประจำทุกปี ส่งผลให้ข้อมูลที่ได้อาจจะล้าสมัยเกินไปเมื่อต้องนำไปใช้งานต่อเนื่อง ขณะที่โครงการ Big Data Analytics ที่ต้องการนำฐานข้อมูลขนาดใหญ่ซึ่งถูกสร้างขึ้นแบบ Real time เช่น สัญญาณโทรศัพท์ มาสร้างแบบจำลอง ข้อดีก็คือสามารถทำได้ในเวลาที่รวดเร็ว ต้นทุนต่ำกว่าการจัดเก็บแบบดั้งเดิม และได้ข้อมูลที่ทันสมัยมากกว่า แต่นั่นก็นำมาสู่ข้อจำกัดที่ว่า ข้อมูลที่ได้นั้นไม่ได้อ้างอิงอยู่บนความเป็นจริง (Ground truth) เช่น ข้อมูลประเภทของสถานที่ รูปแบบการเดินทาง วัตถุประสงค์การเดินทางที่ถูกทำนายด้วยข้อมูลประเภทอาคาร ช่วงเวลาที่การเดินทางนั้นเกิดขึ้น เป็นต้น ข้อมูลบางส่วนนั้นได้มาจากการใช้การอนุมานอย่างง่ายด้วย Bayesian Network ทั้งสิ้น แม้ว่าจะมีหลักการแต่ก็มีข้อบกพร่องอยู่ ซึ่งทำให้มีความน่าเชื่อถือต่ำกว่าข้อมูลที่ได้จากโครงการ BTDS

ดังนั้นงานวิจัยชิ้นนี้จึงมุ่นเน้นที่จะใช้ประโยชน์จากข้อดีของทั้ง 2 โครงการ โดยการสร้างแบบจำลองปัญญาประดิษฐ์ด้วยอัลกอริทึม Random Forest และ Bayesian network จากนั้นนำมาฝึกฝนและทดสอบด้วยข้อมูลการเดินทางในพื้นที่กรุงเทพมหานครและปริมณฑลจากข้อมูลจริง (Ground truth) ที่ได้จากโครงการ BTDS แล้วนำแบบจำลองที่ได้มาอนุมานวัตถุประสงค์การเดินทาง จากข้อมูลที่ได้จากโครงการ Big Data Analytics

รูปที่ 1 โครงการศึกษาสำรวจการเดินทางในเขตกรุงเทพมหานครและปริมณฑล 2565 Bangkok Travel Demand Survey - BTDS 2565



รูปที่ 2 การอนุมานความน่าจะเป็นเพื่อให้ได้วัตถุประสงค์การเดินทาง ของชุดข้อมูล Big Data Analytics ด้วย Bayesian Network

1.1 วัตถุประสงค์ของการศึกษา

- สร้างแบบจำลองเพื่ออนุมานวัตถุประสงค์การเดินทางจากชุดข้อมูล BTDS และอนุมานวัตถุประสงค์การเดินทางของชุดข้อมูล Big Data Analytics
- เปรียบเทียบประสิทธิภาพของแบบจำลองระหว่าง Random Forest และ Bayesian Network

1.2 ขอบเขตการศึกษา

- โครงการสำรวจการเดินทางในกรุงเทพมหานครและปริมณฑล (BTDS) ปี 2565
- ข้อมูลจากสัญญาณโทรศัพท์มือถือในโครงการ Big Data Analytics

2. ทบทวนวรรณกรรม

2.1 ศึกษาข้อมูลในปัจจุบัน

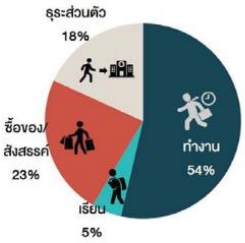
2.1.1 โครงการสำรวจการเดินทางในกรุงเทพมหานครและปริมณฑล (BTDS) ปี 2565

โครงการศึกษาสำรวจการเดินทางในเขตกรุงเทพมหานครและปริมณฑล (Bangkok Travel Demand Survey) มีวัตถุประสงค์เพื่อศึกษาและสำรวจความต้องการการเดินทางและพฤติกรรมการเดินทางของคนและการขนส่งสินค้า โดยพัฒนาแบบจำลองด้านการขนส่งและจราจรให้สอดคล้องกับพฤติกรรมการเดินทางที่เปลี่ยนแปลงไป นอกจากนี้ยังวิเคราะห์สภาพการจราจรบนโครงข่ายการเดินทางทั้งในปัจจุบันและ

อนาคต เพื่อเสนอแนวทางหรือมาตรการด้านการขนส่งและจราจรที่ควรพัฒนา และจัดทำข้อมูลสารสนเทศเพื่อการวางแผนพัฒนาระบบการขนส่งและจราจร

การสำรวจข้อมูลครัวเรือน (Household Travel Survey, HTS) ใช้วิธีการสัมภาษณ์ส่วนบุคคล ซึ่งเป็นวิธีที่นิยมเพราะสามารถอธิบายข้อมูลรายละเอียดได้ดี ผู้ตอบแบบสอบถามมักให้ความร่วมมือและข้อมูลที่ถูกต้อง อย่างไรก็ตาม การสำรวจแบบนี้ใช้เวลานานและมีค่าใช้จ่ายสูง ส่วนการสำรวจทางโทรศัพท์จะใช้การโทรสอบถามข้อมูล ซึ่งเจ้าหน้าที่จะเป็นผู้บันทึกข้อมูลในแบบสอบถาม

ส่วนที่สนใจ คือ วัตถุประสงค์การเดินทางของกลุ่มตัวอย่างส่วนใหญ่เป็นการเดินทางเพื่อไปทำงาน คิดเป็นร้อยละ 54 ของกลุ่มตัวอย่างทั้งหมด



รูปที่ 3 วัตถุประสงค์การเดินทาง

2.1.2 โครงการศึกษาการพัฒนานวัตกรรมระบบวิเคราะห์ฐานข้อมูลขนาดใหญ่ (Big Data Analytics) เพื่อการขนส่งสินค้าด้วยรถบรรทุก และการเดินทางของคนในเขตกรุงเทพมหานครและปริมณฑล ของสำนักงานนโยบายและแผนการขนส่งและจราจร (สนข.)

เป็นโครงการที่เกิดขึ้นจากนโยบายรัฐบาล ที่ต้องการให้หน่วยงานภาครัฐขับเคลื่อนการใช้ประโยชน์จาก ข้อมูลขนาดใหญ่ (Big Data) และประยุกต์สร้างเป็นบริการที่เหมาะสม โดยวิเคราะห์ฐานข้อมูลขนาดใหญ่เพื่อหา ข้อมูลพฤติกรรมการขนส่งสินค้าและการเดินทางของคนเพื่อใช้ประโยชน์ในการวางแผน การปรับปรุงและก่อสร้างโครงสร้างพื้นฐานด้านคมนาคมขนส่ง และเพิ่มขีดความสามารถในการแข่งขันของประเทศ โดยข้อมูลการเดินทางของคนนั้นเก็บจากวิธีที่หลากหลาย ได้แก่ ข้อมูลพิกัดโทรศัพท์เคลื่อนที่ แอปพลิเคชันสำรวจข้อมูล เหมาะสมกับการนำมาวิเคราะห์การเดินทาง โดยทำการประมวลผลข้อมูลสัญญาณโทรศัพท์จากผู้ให้บริการโทรศัพท์เคลื่อนที่ เป็นข้อมูลเกี่ยวกับวันที่ ชั่วโมงที่ พิกัดเสาสัญญาณที่เชื่อมต่อ จังหวัด อำเภอ/เขต ตำบล/แขวง โดยจะเรียงลำดับพิกัดต่าง ๆ ในแต่ละชั่วโมงโดยใช้ traveling sales man problem มาดำเนินการปัญหาในลักษณะนี้ แล้วจึงนำ probability มาแปลงเป็น ค่าเวลา เพื่อระบุการเคลื่อนที่หรือหยุดนิ่งของโทรศัพท์ในกรณีเคลื่อนที่จะนำมาระบุรูปแบบการเดินทาง แต่หากหยุดนิ่งจะนำมาระบุกิจกรรมที่ทำ

รูปที่ 4 แสดงตัวอย่างหน้าต่างการตัวอย่างหน้าต่างการบันทึกข้อมูล

2.2 ทฤษฎีที่เกี่ยวข้องภายในงานวิจัย

2.2.1 Machine Learning

Machine Learning (ML) คือ การประยุกต์คณิตศาสตร์และวิทยาศาสตร์คอมพิวเตอร์เพื่อทำให้คอมพิวเตอร์ สามารถเรียนรู้สิ่งต่างๆ ตัดสินใจได้โดยไม่ต้องกำหนดกฎและเงื่อนไขตายตัว และประมวลผลข้อมูลนั้นเพื่อให้สามารถตัดสินใจตามที่ต้องการ โดยการตัดสินใจที่ ML ทำได้ มีหลายรูปแบบดังนี้

1. Supervised Learning (การเรียนรู้แบบมีการควบคุม): แบบจำลองเรียนรู้จากข้อมูลที่มีการควบคุมหรือป้ายกำกับ เช่น การทำนายยอดขายในอนาคตจากข้อมูลในอดีต จำแนกประเภทของอีเมลว่าเป็นสแปมหรือไม่
2. Unsupervised Learning (การเรียนรู้แบบไม่มีการควบคุม): แบบจำลองเรียนรู้จากข้อมูลที่ไม่มีการควบคุมหรือป้ายกำกับ เช่น การจัดกลุ่มลูกค้าที่คล้ายคลึงกันตามพฤติกรรมการซื้อของพวกเขา การค้นหากฎที่หาได้จากข้อมูลส่วนใหญ่
3. Reinforcement Learning (การเรียนรู้แบบเสริมสร้าง): แบบจำลองเรียนรู้ด้วยการทดลองและรับรู้จาก สิ่งแวดล้อม โดยได้รับรางวัลหรือลบล้างค่าโดยรวมของการกระทำ เช่น การสอนหุ่นยนต์ทำการเดินโดยให้ได้รับรางวัลเมื่อเดินถึงจุดหมาย

Machine learning ใช้ทฤษฎีและเทคนิคจากคณิตศาสตร์และสถิติเพื่อให้แบบจำลองสามารถเรียนรู้และทำนายผลได้อย่างแม่นยำมากขึ้น โดยมีเครื่องมือที่ใช้ในการพัฒนาและประยุกต์ใช้ machine learning ได้แก่ Python (กับไลบรารี Scikit-learn, TensorFlow, Keras), R (กับไลบรารี Caret, MLR), Weka, RapidMiner ฯลฯ

2.2.2 Decision tree model (Bayesian Network)

Decision Tree Model หรือแผนภาพต้นไม้ตัดสินใจเป็น Classification Model ที่ใช้กระบวนการ Rule-Based คือการสร้างกฎถ้า - แล้ว (If – Else) ขึ้นมา โดยเป็นวิธีการแยกข้อมูลที่กรณีข้อมูลตรงตามกฎหรือเงื่อนไขที่ตั้งขึ้นก็จะถูกจำแนกไปอยู่กลุ่มหนึ่ง ถ้าไม่ตรงตามกฎหรือเงื่อนไขที่ตั้งขึ้นก็จะถูกจำแนกไปอยู่อีกกลุ่มหนึ่ง

แต่ก็ยังมีข้อเสียของการใช้งานอัลกอริทึมนี้อยู่คือ เมื่อประยุกต์ใช้กับข้อมูลที่ซับซ้อนมากขึ้นนั้นจะไม่มีประสิทธิภาพในการวิเคราะห์ มีข้อดี คือ ง่ายต่อการใช้งานและความเข้าใจที่เรียบง่าย ส่วนข้อเสีย คือ ไม่มีความสามารถในการวิเคราะห์ และทำนายข้อมูลที่มีความซับซ้อนมาก

2.2.3 Random Forest (RF) Model

อัลกอริทึมหนึ่งของ Machine Learning ที่ถูกพัฒนามาจาก Decision Tree โดยจะมีการเพิ่มจำนวน Tree มากกว่า ทำให้มีประสิทธิภาพและความแม่นยำสูงมากขึ้น

2.2.4 Naïve Bayes

Naive Bayes Classification (NB) เป็นหนึ่งใน model สำหรับใช้ในการทำนายว่า ตัวแปร หรือตัวอย่างนั้นอยู่ในกลุ่มไหน ซึ่งเป็น supervised model ที่ต้องมีการ label ก่อน (คล้ายกับ linear classification / logistic regression) โดยหลักการของ Naive Bayes Classification จะใช้หลักการเรื่อง ความน่าจะเป็น (probability) ในการทำนายว่าเป็น กลุ่มไหน ในขณะที่ linear classification จะใช้หลักการทางเลขาคณิต (Geometry) ในการทำนายว่าเป็น กลุ่มไหน

2.2.5 การประเมินประสิทธิภาพแบบจำลอง

งานวิจัยที่เกี่ยวข้องกับการประเมินแบบจำลอง (model evaluation) มักนิยามและใช้เกณฑ์ต่างๆ เพื่อวัด ประสิทธิภาพของแบบจำลอง สามเกณฑ์ที่ใช้แพร่หลายในการประเมินแบบจำลองได้แก่ Accuracy, Sensitivity และ Precision

Accuracy (ความถูกต้อง) คือสัดส่วนของการทำนายที่ถูกต้องทั้งหมดต่อจำนวนทั้งหมดของข้อมูล คำนวณโดย

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ predictions}$$

Sensitivity (ความไวต่อการตรวจพบ) คือสัดส่วนของผลลัพธ์ที่ถูกต้องบวกกับจำนวนทั้งหมดของผลลัพธ์ ที่เป็น Positive หรือที่ถูกต้อง คำนวณโดย

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Precision (ความแม่นยำในการพยากรณ์) คือสัดส่วนของผลลัพธ์ที่เป็น Positive ที่ถูกต้องต่อทั้งหมดของ ผลลัพธ์ที่เป็น Positive คำนวณโดย

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

ในงานวิจัยที่เกี่ยวข้องกับทฤษฎีนี้ เราใช้ Accuracy เพื่อวัดประสิทธิภาพของระบบการทำนายโดยรวม ในขณะที่ Sensitivity และ Precision นั้นใช้ในการวัดประสิทธิภาพของระบบในกรณีที่เกี่ยวข้องกับสิ่งที่เราสนใจมากขึ้น

2.2.6 Type I Error และ Type II Error

จาก confusion matrix สามารถแบ่งประเภทของการทำนายได้ดังนี้

- True Positives, TP: ถูกทำนายว่าเป็นคลาสที่ถูกต้อง
- False Positives, FP: ถูกทำนายว่าเป็นคลาสใดคลาสหนึ่ง แต่เป็นคลาสอื่น
- False Negatives, FN: เป็นคลาสใดคลาสหนึ่ง แต่ถูกทำนายว่าเป็นคลาสอื่น
- True Negatives, TN: ถูกทำนายว่าไม่เป็นคลาสใดคลาสหนึ่งอย่างถูกต้อง

Type I Error: False Positive ทำนายว่าเป็นสิ่งหนึ่ง แต่จริงๆแล้วไม่ได้เป็นสิ่งนั้น เช่น ความจริงไม่ได้เป็นมะเร็ง แต่ผลบอกว่าเป็นมะเร็ง

Type II Error: False Negative ทำนายว่าไม่ได้เป็นสิ่งหนึ่ง แต่จริงๆแล้วเป็นสิ่งนั้น เช่น ความจริงเป็นมะเร็ง แต่ผลบอกว่าไม่ได้เป็นมะเร็ง

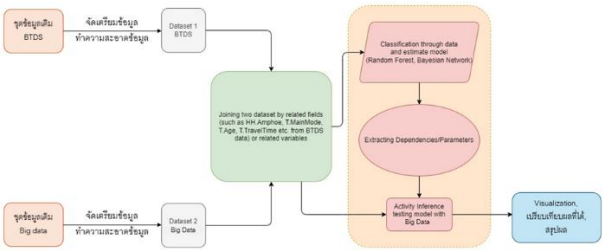
$$Type\ I\ Error = \frac{False\ Positives}{Total\ Predicted}$$

$$Type\ II\ Error = \frac{False\ Negatives}{Total\ Actual}$$

3. ระเบียบวิธีวิจัย

3.1 แผนรายละเอียดของข้อมูล

ขั้นตอนการทำงานขั้นแรก เมื่อได้ข้อมูลมาต้องมีการจัดเตรียมและทำความสะอาดข้อมูลทั้งสองชุดก่อน แล้วจึงเริ่มต้นเลือกตัวแปรที่จะนำมาใช้ในการสร้างแบบจำลอง โดยตัวแปรเหล่านั้นจะต้องเป็นตัวแปรที่มีอยู่ในทั้งสองชุดข้อมูล จากนั้นเลือกอัลกอริทึมที่ต้องการใช้และนำมาสร้างแบบจำลอง โดยจะให้แบบจำลองเรียนรู้วัตถุประสงค์ของการเดินทางจากชุดข้อมูล BTDS เพื่อนำไปทำนายวัตถุประสงค์ของการเดินทางในชุดข้อมูล Big Data โดยแบบจำลองจะใช้อัลกอริทึม Bayesian Network และ Random Forest ในโปรแกรม R จากนั้นนำผลลัพธ์ที่ได้มาเปรียบเทียบประสิทธิภาพ ดังแผนภูมิขั้นตอนการทำงานด้านล่าง รวมทั้งมีการทำรายละเอียดชุดข้อมูลและการวิเคราะห์ข้อมูลเชิงบรรยาย



รูปที่ 5 แผนภาพขั้นตอนการทำงานกับข้อมูลจากสองแหล่งที่มา

3.2 การจัดเตรียมและทำความสะอาดข้อมูล

โดยทำการคัดกรองและทำความสะอาดข้อมูลจากชุดข้อมูลดิบที่ได้รับมาให้มีความถูกต้อง

- คัดกรองข้อมูลในการระบุรูปแบบการเดินทาง จะกรองข้อมูล transit ออก ซึ่งหมายถึงการเดินทางที่อยู่ระหว่างจุดเริ่มต้นและจุดปลายทาง
- ทำการรวมข้อมูลแต่ละจุดเข้าด้วยกัน ระบุกิจกรรมหลัก โหมดการเดินทางหลัก เวลาออก/ถึง และสถานที่จุดหมาย รวมถึงเวลาที่อยู่ ณ สถานที่นั้นๆ
- ทำการคัดกรองข้อมูลที่ไม่ได้ระบุอย่างชัดเจน ข้อมูลที่ไม่มีการเดินทาง และข้อมูลที่อยู่นอกเหนือขอบเขตของงานวิจัยออกเหลือเป็นข้อมูล ดังนี้

ตารางที่ 1 ข้อมูลที่มีหลังการคัดกรองและทำความสะอาดข้อมูล

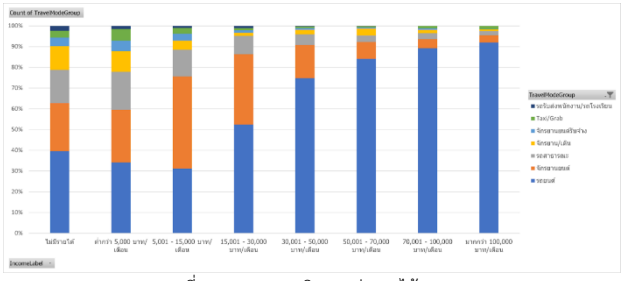
ชื่อตัวแปร	ความหมาย	ตัวอย่าง
M_Normalized_order	สัดส่วนการเดินทางของวัน	0.25
M_activity_type	วัตถุประสงค์การเดินทาง	work
M_Dep_Week_Ttype	ประเภทช่วงเวลา และวันในสัปดาห์ที่ออกเดินทาง	MP-WE
M_Arr_Week_Ttype	ประเภทช่วงเวลา และวันในสัปดาห์ที่ถึงปลายทาง	DOP-WE
M_zone_o	เขตในกรุงเทพมหานครของต้นทาง	OT

M_zone_d	เขตในกรุงเทพมหานครของปลายทาง	TS
M_Mode	รูปแบบยานพาหนะในการเดินทาง	TX
M_Building_type	ประเภทของอาคารจุดหมายปลายทาง	Off/Shop
M_stay_time_type	ประเภทของระยะเวลาที่อยู่ที่ปลายทาง	2
M_travel_time_type	ประเภทของระยะเวลาที่ใช้เดินทาง	5
M_distance_type	ประเภทของระยะทางที่เดินทาง	6

3.3 การวิเคราะห์ข้อมูลเชิงบรรยาย

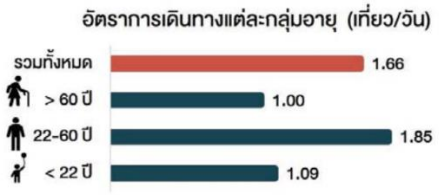
จากการสำรวจชุดข้อมูลโครงการศึกษาสำรวจความต้องการการเดินทางในเขตกรุงเทพมหานครและปริมณฑล (Bangkok Travel Demand Survey (BTDS), 2565) มีการสำรวจและเก็บข้อมูลในระยะเวลาช่วงวันที่ 1 พฤศจิกายน พ.ศ. 2564 ถึงวันที่ 6 เมษายน พ.ศ. 2565 สามารถสรุปรายละเอียดข้อมูลที่ได้จากการสำรวจและวิเคราะห์ผล ดังนี้

รูปแบบการเดินทางต่อรายได้ส่วนบุคคล พบว่า 3 ลำดับแรกที่มีการเลือกเดินทางมากสุดคือ รถยนต์ส่วนบุคคล รถจักรยานยนต์ และรถสาธารณะตามลำดับ



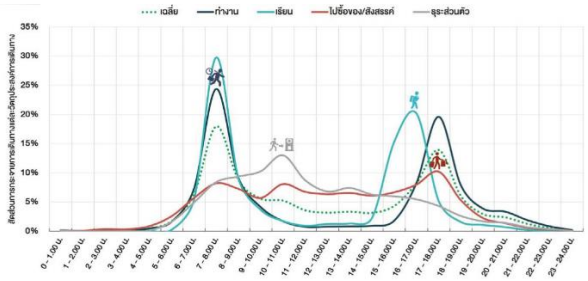
รูปที่ 6 รูปแบบการเดินทางต่อรายได้บุคคล

นอกจากนี้ยังพบว่าการแบ่งกลุ่มอายุจากนั้นนำมาคำนวณอัตราการเดินทางแต่ละกลุ่มอายุของกลุ่มตัวอย่าง พบว่าอัตราการเดินทางเฉลี่ย 1.66 เที่ยว/วัน ซึ่งรูปแบบกลุ่มอายุที่มีอัตราการเดินทาง สูงสุด ได้แก่ กลุ่มอายุมากกว่า 22 ปี ถึง 60 ปี โดยมีอัตราการเดินทางอยู่ที่ 1.85 เที่ยว/วัน



รูปที่ 7 อัตราการเดินทางแต่ละกลุ่มอายุ (เที่ยว/วัน)

สำหรับช่วงเวลาที่มีการเดินทางในแต่ละรูปแบบสูงสุด พบว่าช่วงเช้าตั้งแต่เวลา 07.00 – 08.00 น. และช่วงเย็นเวลา 16.00 – 17.00 น. โดยช่วงเช้ามีปริมาณจราจรสูงสุดคิดเป็นร้อยละ 18 และช่วงเย็นมีปริมาณจราจรสูงสุด คิดเป็นร้อยละ 14



รูปที่ 8 อัตราการเดินทางแต่ละกลุ่มอายุ (เที่ยว/วัน)

3.4 การนำอัลกอริทึมมาใช้ในการสร้างแบบจำลอง

การสร้างแบบจำลองเพื่ออนุมานวัตถุประสงค์การเดินทางจากชุดข้อมูลที่ได้จัดเตรียมไว้ อัลกอริทึมที่เลือกมาใช้ ได้แก่ Naive Bayes classifier จาก Bayesian Network และ Random Forest โดยเริ่มจากชุดข้อมูล BTDS ที่จะนำข้อมูล 80% ของชุดข้อมูลทั้งหมดมาฝึกแบบจำลอง และ 20% ที่เหลือมาใช้ทดสอบ เพื่อดูประสิทธิภาพของแบบจำลองก่อนหลังจากนั้นนำแบบจำลองนำไปใช้ไปทดสอบการอนุมานวัตถุประสงค์การเดินทางกับชุดข้อมูล Big Data

```
{r Naive Bayes with 5 activity types and unknown}
### Naive Bayes 5 Activity Type with Unknown
btds_dfU_5index <- sample(2, nrow(btds_dfU_5), replace = TRUE, prob = c(0.8, 0.2))
btds_dfU_5train <- btds_dfU_5[btds_dfU_5index==1,]
btds_dfU_5test <- btds_dfU_5[btds_dfU_5index==2,]

# Fitting Naive Bayes to the train dataset
set.seed(222)
btds_dfU_5_nb <- naiveBayes(x = btds_dfU_5train[, 1:10],
                            y = btds_dfU_5train$M_activity_type_2,
                            laplace = 1)
btds_dfU_5_nb

# Predicting the Test set results
btds_dfU_5_nb_test <- predict(btds_dfU_5_nb, # NB model
                              newdata = btds_dfU_5test[, 1:10], # predictors
                              type = "class")


# Test Confusion Matrix
btds_dfU_5_nbtest_conf <- confusionMatrix(data = btds_dfU_5_nb_test, # predictions
                                          reference = btds_dfU_5test$M_activity_type_2, # actual
                                          mode = "everything")
btds_dfU_5_nbtest_conf

## Try mobile data mb_dfU_5 predictions
mb_dfU_5_nb_prd <- predict(btds_dfU_5_nb, # NB model
                           newdata = mb_dfU_5[, 1:10], # predictors
                           type = "class")

cf_mtx_mb_dfU_5_nb <- confusionMatrix(data = mb_dfU_5_nb_prd, # predictions
                                      mb_dfU_5$M_activity_type_2, # actual
                                      mode = "everything")
cf_mtx_mb_dfU_5_nb
```

รูปที่ 9 การฝึกและอนุมานผลแบบจำลอง Naïve Bayes ในโปรแกรม R

```
# 5 Activity Type
set.seed(11)
ind <- sample(2,nrow(BTDS_Imp2), replace = TRUE, prob = c(0.8,0.2))
train.btds <- BTDS_Imp2[ind == 1,]
test.btds <- BTDS_Imp2[ind == 2,]

rf.w.unk.5 <- randomForest(
  x = train.btds[,c("M_Dep_Week_Ttype","M_Arr_Week_Ttype",
                    "M_zone_o","M_zone_d","M_distance_type",
                    "M_travel_time_type","M_Mode",
                    "M_stay_time_type2","M_trip_proportion","M_node_type")],
  y = train.btds$M_activity_type_2,
  ntree = 200)
pre.rf.w.unk.5 <- predict(rf.w.unk.5,test.btds)
```

รูปที่ 10 การฝึกและอนุมานผลแบบจำลอง Random Forest ในโปรแกรม R

4. ผลการดำเนินการวิจัย

4.1 การดำเนินการพัฒนาแบบจำลอง

4.1.1 Naïve Bayes Classifier

นำข้อมูลที่ได้ทำความสะอาดและจัดรูปแบบแล้วมาสร้างแบบจำลองด้วยอัลกอริทึม Naïve Bayes Classifier แล้วทดสอบได้ตัวอย่างออกมาดังรูปด้านล่างนี้

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = btds_df_43train[, 1:10], y = btds_df_43train$M_activity_type_3,
    laplace = 1)

A-priori probabilities:
btds_df_43train$M_activity_type_3
work_study   shopping       rest      other
0.16801458 0.13615588 0.61421432 0.08161522

Conditional probabilities:
                              M_Dep_Week_Ttype
btds_df_43train$M_activity_type_3        MP-WE        EP-WE       DOP-WE
                        work_study 0.6063865831 0.0230391326 0.0540826698
                        shopping   0.1977843967 0.1892146104 0.2467471391
                        rest       0.0486168252 0.3712146673 0.1119751527
                        other      0.5203903799 0.0385151621 0.2100034855
                              M_Dep_Week_Ttype
btds_df_43train$M_activity_type_3        NOP-WD        EP-WD        MP-WD
                        work_study 0.0013975944 0.0086396747 0.2348805692
                        shopping   0.0136385013 0.0710142656 0.0819877724
                        rest       0.0684807677 0.1497444575 0.0180676116
                        other      0.0027012896 0.0139421401 0.1287905193
                              M_Dep_Week_Ttype
btds_df_43train$M_activity_type_3        NOP-WE        EM-WE         N-WE
                        work_study 0.0048704049 0.0343892936 0.0006776215
                        shopping   0.0377802163 0.0450958875 0.0004702931
                        rest       0.1729345093 0.0061538818 0.0099667389
                        other      0.0081038689 0.0261415127 0.0008713838
```

รูปที่ 11 ตัวอย่าง Conditional Probabilities ของการสร้างแบบจำลอง Naïve Bayes ในโปรแกรม R ที่ฝึกฝนจากชุดข้อมูล BTDS

การแสดงผลการดำเนินการสร้างโมเดลออกมาในรูปแบบของ Conditional Probabilities ของ features แต่ละตัวเทียบกับค่า target คือ M_activity_type_2 ที่ประกอบด้วย other, rest, shopping, และ work_study โดยมีการระบุ probabilities ตามประเภทของกิจกรรมหรือวัตถุประสงค์การเดินทาง ซึ่งเมื่อนำไปทดสอบกับชุดข้อมูลเดียวกันแล้วได้ผลดังตารางด้านล่างนี้

ตารางที่ 2 ผลการวัดประสิทธิภาพของ Naïve Bayes Classifier Model ที่ฝึกฝนจากชุดข้อมูล BTDS มาทดสอบกับชุดข้อมูลเดิม

กิจกรรม	Precision	Sensitivity
other	75.93%	69.55%
rest	99.63%	99.69%
shopping	80.65%	86.91%
work_study	85.99%	83.86%
	Accuracy	**92.85%**

4.1.2 Random Forest

นำข้อมูลที่ได้ทำความสะอาดและจัดรูปแบบแล้วมาสร้างแบบจำลองประเภท Random Forest แล้วทดสอบได้ตัวอย่างออกมาดังรูปด้านล่างนี้

```
Confusion Matrix and Statistics

          Reference
Prediction other  rest shopping  work
  other     2064    27      219   284
  rest        16 21568       10    17
  shopping   330     6     4350   321
  work       347    16      249  5248

Overall Statistics

               Accuracy : 0.9475
                 95% CI : (0.9451, 0.9498)
    No Information Rate : 0.6164
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9073

 Mcnemar's Test P-Value : 2.135e-07

Statistics by Class:

                     Class: other Class: rest Class: shopping Class: work
Sensitivity              0.74864      0.9977          0.9010      0.8940
Specificity              0.98360      0.9968          0.9783      0.9790
Pos Pred Value           0.79568      0.9980          0.8688      0.8956
Neg Pred Value           0.97866      0.9964          0.9841      0.9787
Prevalence               0.07861      0.6164          0.1377      0.1674
Detection Rate           0.05885      0.6150          0.1240      0.1496
Detection Prevalence     0.07396      0.6162          0.1428      0.1671
Balanced Accuracy        0.86612      0.9973          0.9396      0.9365
```

รูปที่ 12 ผลการทดสอบแบบจำลอง Random Forest ในโปรแกรม R ด้วยชุดข้อมูลเดิม

ผลการทดสอบได้ค่า Precision และ Sensitivity ดังนี้

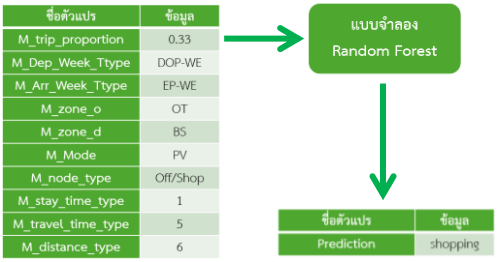
ตารางที่ 3 ผลการวัดประสิทธิภาพของ Random Forest Model ที่ฝึกฝนจากชุดข้อมูล BTDS มาทดสอบกับชุดข้อมูลเดิม

กิจกรรม	Precision	Sensitivity
other	79.57 %	74.86 %
rest	99.80 %	99.77 %
shopping	86.88 %	90.10 %
work_study	89.56 %	89.40 %
	Accuracy	**94.75 %**

5. บทสรุป

5.1 สรุปภาพรวมการวิจัย

จากบท 4 เมื่อให้แบบจำลองทั้งสองเรียนรู้และทดสอบบนชุดข้อมูล BTDS พบว่า Random Forest มีประสิทธิภาพโดยรวมสูงสุด มีความถูกต้องอยู่ที่ 94.75% เมื่อดูแต่ละกิจกรรม พบว่ากิจกรรม rest มีความไวต่อการตรวจพบและความแม่นยำในการพยากรณ์สูงที่สุด เนื่องจากเป็นกิจกรรมที่คาดเดาได้ง่ายจากประเภทอาคาร ช่วงเวลาของวัน และสัดส่วนการเดินทางของวัน เพราะมักเป็นกิจกรรมที่เกิดที่อาคารประเภท home เกิดเวลาช่วงเย็นถึงค่ำ และเป็นการเดินทางลำดับท้ายๆ ของวัน ส่วนกิจกรรม other มีค่าทั้งสองต่ำที่สุดในสองแบบจำลองเนื่องจากเป็นกิจกรรมที่ไม่มีแบบแผนตายตัว จึงคาดเดาได้ยาก เมื่อนำแบบจำลองที่สร้างจากชุดข้อมูล BTDS มาอนุมานวัตถุประสงค์การเดินทางในชุดข้อมูล Big Data ได้ตัวอย่างการอนุมาน ดังนี้



รูปที่ 13 ตัวอย่างการอนุมานวัตถุประสงค์การเดินทางของแบบจำลอง

5.2 ข้อเสนอแนะ

เนื่องจากเดิมทีชุดข้อมูลมีรูปแบบและตัวแปรที่แตกต่างกัน ในอนาคตถ้าหากมีความต้องการใช้ข้อมูลการเดินทางจากเสาสัญญาณโทรศัพท์มาใช้อย่างเต็มรูปแบบมากขึ้น ควรคำนึงและพยายามทำให้รูปแบบตัวแปรของข้อมูลมีความคล้ายคลึงกัน และมีการสื่อสารระหว่างสองฝ่าย จะทำให้แบบจำลองมีประสิทธิภาพในการอนุมานกิจกรรมการเดินทางมากขึ้น และในตอนนี้ถ้าหากต้องการให้กิจกรรมการเดินทางเดิมของชุดข้อมูล Big Data อยู่บนฐานความจริงมากขึ้น อาจนำแบบจำลองที่ได้จากชุดข้อมูล BTDS ไปอนุมานตัวแปรในชุดข้อมูล Big Data จะทำให้ชุดข้อมูลมีความน่าเชื่อถือมากขึ้น

เอกสารอ้างอิง

[1] พรหมพร ล่ามกิจจา, และพลอยสวรรค์ กิตติบุญญาทิวากร. (2566). การพัฒนาแบบจำลองเพื่ออนุมานรูปแบบกิจกรรมของการเดินทาง (วิทยานิพนธ์ปริญญาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย)

[2] สำนักงานนโยบายและแผนการขนส่งและจราจร. (2565). โครงการศึกษาสำรวจการเดินทางในเขตกรุงเทพมหานครและปริมณฑล 2565. สืบค้นจาก https://btds-otp.com/index.php/trainning/

[3] สำนักงานนโยบายและแผนการขนส่งและจราจร. (2565). แบบสำรวจการเดินทางในเขตกรุงเทพมหานครและปริมณฑล (Bangkok Travel Demand Survey) 2564 – 2565. สืบค้นจาก https://survey.btds-otp.com/

[4] สำนักงานนโยบายและแผนการขนส่งและจราจร. (ม.ป.ป.). โครงการศึกษาการพัฒนานวัตกรรมระบบวิเคราะห์ฐานข้อมูลขนาดใหญ่ (Big Data Analytics) เพื่อการขนส่งสินค้าด้วยรถบรรทุก และการเดินทางของคนในเขตกรุงเทพมหานครและปริมณฑล ระยะที่ 1 กรุงเทพมหานคร. สืบค้นจาก https://www.bigdata -otp.com/

[5] boom626. (3 ตุลาคม 2562). Confusion Matrix. สืบค้นจาก https://medium.com/@mirthful_suns et _cattle_231/confusion-matrix-48cc396b1b58

[6] KongRuksiam Studio. (27 มีนาคม 2563). Machine Learning(EP.6)-การจัดหมวดหมู่ด้วย Naive Bayes. สืบค้นจาก https://kongruksiam.medium.com/สรุป-machine-learning-ep-5-การจัดหมวดหมู่ด้วย-naive-bayes-eb9ce0e1b010

[7] KongRuksiam Studio. (27 มีนาคม 2563). ตัวอย่างการคำนวณ Posterior probability [รูป]. สืบค้นจาก https://miro.medium.com/v2/resize:fit:640/format:webp/0*CuQyw_lpE78dR9BS.png

[8] AnalystPrep. (2021). Supervised Machine Learning, Unsupervised Machine Learning, and Deep Learning. Retrieved from https://analystprep.com/study-notes/cfa-level-2/quantitative-method/supervised-machine-learning-unsupervised-machine-learning-deep-learning/

[9] Evidently AI Team. (n.d.). Accuracy vs. precision vs. recall in machine learning: what's the difference?. Retrieved from https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall#:~:text=Accuracy%20shows%20how%20often%20a,when%20choosing% 20the%20suitable%20metric.

[10] Graells-Garrido, E., Caro, D. & Parra, D. Inferring modes of transportation using mobile phone data. EPJ Data Sci. 7, 49 (2018). https://doi.org/10.1140/epjds/s13688-018-0177-1

[11] Mohammed Mohammed, Jimi Oke, Origin-destination inference in public transportation systems: A comprehensive review, International Journal of Transportation Science and Technology, Volume 12, Issue 1, 2023, Pages 315-328, ISSN 2046-0430, https://doi.org/10.1016/j.ijtst.2022.03.002.

[12] Nilufer, S. A., Mohamed, R. I., Tao, C., Huanfa, C., & Yang, Z. (2021). Neural networks to predict public transport trip purposes from individual smart card data and POIs. Retrieve from https://doi.org/10.1080/10095020.2021.1985943

[13] Ravish Raj. (n.d.). Supervised, Unsupervised and Semi-supervised Learning with Real-life Usecase. Retrieved from https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning

[14] scikit-learn developers. (n.d.). Cross-validation: evaluating estimator performance. Retrieved from https://scikit-learn.org/stable/modules/cross_validation.html

[15] Ximing, C., Jianjun, W., Hao, L., Xiaoyong, Y., Huijun, S., & Yunchao, Q. (2019). Travel mode choice: a data fusion model using machine learning methods and evidence from travel diary survey data. Retrieved from https://doi.org/10.1080/23249935.2019.1620380

[16] Zhu, Y. (2022). Inference of activity patterns from urban sensing data using conditional random fields. Environment and Planning B: Urban Analytics and City Science, 49(2), 549-565. https://doi.org/10.1177/23998083211016863