การวิเคราะห์เชิงเปรียบเทียบความสัมพันธ์ระหว่างข้อมูล เพื่อคัดเลือกตัวแปรที่มีศักยภาพในการพยากรณ์ด้วยการถดถอยเชิงเส้นอย่างง่าย

Comparative Analysis of Data Correlations for Selecting Predictive Variables Using Simple Linear Regression

กุลิสรา ทรัพย์ประกอบ[1] , นาวิน ดวงสุวรรณ์[2] และ ปยุต มีมานะ[3]

[1]คณะบริหารธุรกิจและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ,

E-Mail 267345221002-st@rmutsb.ac.th

[2]คณะบริหารธุรกิจและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ,

E-Mail 267345221013-st@rmutsb.ac.th

[3]คณะบริหารธุรกิจและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ,

E-Mail 267345221015-st@rmutsb.ac.th

วันที่รับบทความ 18 มิถุนายน 2568

วันแก้ไขบทความ 30 มิถุนายน 2568

วันตอบรับบทความ 30 มิถุนายน 2568

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) วิเคราะห์ความสัมพันธ์เชิงเส้นระหว่างตัวแปรจากชุดข้อมูลหลายชุด โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน 2) สร้างแบบจำลองการถดถอยเชิงเส้นอย่างง่าย 3) ประเมินประสิทธิภาพของแบบจำลองผ่านค่าตัวชี้วัดทางสถิติ ด้วยค่า R2 ค่า RMSE และ MAE โดยที่ผู้วิจัยใช้ข้อมูลทุติยภูมิจากแหล่งเปิดเผยจำนวน 5 ชุด ได้แก่ *Heart.csv, StudentMarks.csv, TestEnergyData.csv, Salary_dataset.csv* และ *Advertising.csv* โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน และค่าตัวชี้วัดประสิทธิภาพของแบบจำลอง ได้แก่ ค่า R^2, RMSE และ MAE ในการประเมิน ผลการวิเคราะห์พบว่า ชุดข้อมูล *Salary_dataset.csv* แสดงค่าความสัมพันธ์สูงที่สุด (R^2 = 0.9570) ซึ่งชี้ให้เห็นว่าตัวแปร “ประสบการณ์ทำงาน” มีศักยภาพสูงในการทำนาย “เงินเดือน” ขณะที่ชุดข้อมูล *StudentMarks.csv* และ *Advertising.csv* ก็มีค่าความสัมพันธ์ในระดับสูงเช่นกัน (R^2 > 0.80) ส่วนชุดข้อมูล *Heart.csv* และ *TestEnergyData.csv* แสดงความสัมพันธ์ในระดับต่ำและปานกลางตามลำดับ การศึกษาครั้งนี้แสดงให้เห็นถึงความสำคัญของการวิเคราะห์สหสัมพันธ์เชิงเปรียบเทียบในการคัดเลือกตัวแปรเชิงปริมาณที่เหมาะสมก่อนการสร้างแบบจำลองทางสถิติ ซึ่งสามารถนำไปประยุกต์ใช้ในงานวิเคราะห์ข้อมูลเชิงพยากรณ์ในหลายบริบท เช่น เศรษฐกิจ การศึกษา และพฤติกรรมผู้บริโภค

คำสำคัญ: การถดถอยเชิงเส้นอย่างง่าย ความสัมพันธ์เชิงบวกสูง สัมประสิทธิ์สหสัมพันธ์

ABSTRACT

This study aimed to 1) analyze linear relationships among variables across multiple datasets using Pearson’s correlation coefficient, 2) construct simple linear regression models, and 3) evaluate model

performance using statistical metrics including R^2, RMSE, and MAE. Secondary data from five publicly available datasets—Heart.csv, StudentMarks.csv, TestEnergyData.csv, Salary_dataset.csv, and Advertising.csv—were analyzed. The study employed Pearson's correlation coefficient and model evaluation metrics, specifically R^2, RMSE, and MAE, to assess performance. The analysis revealed that Salary_dataset.csv demonstrated the highest level of linear association (R^2=0.9570), indicating that "Years of Experience" is a highly effective predictor of "Salary." StudentMarks.csv and Advertising.csv also exhibited strong correlations (R^2>0.80), while Heart.csv and TestEnergyData.csv showed lower to moderate correlations. This study highlights the importance of comparative correlation analysis in selecting appropriate quantitative variables before building statistical models. Such an approach can be applied to predictive data analysis in various domains, including economics, education, and consumer behavior.

KEYWORDS: Simple Linear Regression, High Positive Correlation, Correlation Analysis

1. บทนำ

ในยุคของการตัดสินใจบนพื้นฐานข้อมูล (Data-Driven Decision Making) ความสามารถในการค้นหาความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ถือเป็นสิ่งสำคัญอย่างยิ่งสำหรับการสร้างแบบจำลองที่มีประสิทธิภาพ โดยเฉพาะในการวิเคราะห์ข้อมูลเชิงปริมาณ การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) โดยเฉพาะค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation coefficient) ได้รับการยอมรับอย่างแพร่หลายในการประเมินความแข็งแรงและทิศทางของความสัมพันธ์เชิงเส้นระหว่างตัวแปร (Mukaka, 2012) แม้ว่าการถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) จะเป็นเทคนิคพื้นฐานที่มีความเรียบง่ายในเชิงคณิตศาสตร์ แต่ก็ยังคงมีบทบาทสำคัญในงานวิจัยและการวิเคราะห์ในหลากหลายสาขา เนื่องจากสามารถอธิบายและตีความได้อย่างชัดเจน (Montgomery, Peck, & Vining, 2012) ทั้งนี้ ความแม่นยำของแบบจำลองดังกล่าวขึ้นอยู่กับการเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญ

จากการวิเคราะห์ความสัมพันธ์เชิงเปรียบเทียบก่อนการสร้างแบบจำลอง (Pre-model Correlation Analysis) ช่วยให้สามารถคัดกรองตัวแปรที่มีศักยภาพได้อย่างมีประสิทธิภาพ อีกทั้งยังลดความเสี่ยงของปัญหา Overfitting และเพิ่มความสามารถในการทำนายของแบบจำลอง (James, Witten, Hastie, & Tibshirani, 2021)

การศึกษานี้จึงมุ่งเน้นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรจากหลายชุดข้อมูล เพื่อคัดเลือกคู่ตัวแปรที่มีความสัมพันธ์เชิงบวกสูง (High Positive Correlation) ซึ่งเหมาะสมสำหรับการสร้างแบบจำลองถดถอยเชิงเส้นอย่างง่าย การวิเคราะห์เชิงเปรียบเทียบลักษณะนี้ไม่เพียงส่งเสริมการเลือกตัวแปรอย่างมีหลักการ แต่ยังเป็นแนวทางที่สอดคล้องกับกระบวนการวิเคราะห์ข้อมูลอย่างเป็นระบบและมีประสิทธิภาพ

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อวิเคราะห์ความสัมพันธ์เชิงเส้นระหว่างตัวแปรจากชุดข้อมูลหลายชุด โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's Correlation Coefficient)

2.2 เพื่อสร้างแบบจำลองการถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) สำหรับคัดเลือกคู่ความสัมพันธ์เชิงบวกสูง

2.3 เพื่อประเมินประสิทธิภาพของแบบจำลองผ่านค่าตัวชี้วัดทางสถิติ ด้วยค่า R^2 ค่า RMSE และ MAE

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 สหสัมพันธ์ (Correlation Theory)

การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) เป็นเทคนิคทางสถิติที่ใช้ตรวจสอบความสัมพันธ์เชิงเส้นระหว่างตัวแปรอย่างน้อยสองตัว โดยเฉพาะ ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation coefficient) ซึ่งเป็นเครื่องมือที่นิยมใช้ในงานวิจัยเชิงปริมาณ เพื่อวัดความสัมพันธ์ระหว่างตัวแปร (Mukaka, 2012) ซึ่งเหมาะกับตัวแปรที่มีข้อมูลต่อเนื่อง (Continuous Variables) โดยค่า r อยู่ในช่วง -1 ถึง +1 ซึ่งค่าที่เข้าใกล้ +1 หมายถึงความสัมพันธ์เชิงบวกในระดับสูง

3.2 การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) เป็นแบบจำลองเชิงสถิติที่ใช้ทำนายค่าของตัวแปรตาม (Dependent Variable) จากค่าของตัวแปรอิสระ (Independent Variable) หนึ่งตัว ผ่านสมการเชิงเส้นในรูปแบบ

โดยที่ $\beta 0$ คือค่าคงที่ $\beta 1$ คือค่าสัมประสิทธิ์ของตัวแปร X และ ϵ คือค่าคลาดเคลื่อน (Montgomery, Peck, & Vining, 2012)

3.3 ความสัมพันธ์เชิงบวกสูง (High Positive Correlation) ความสัมพันธ์เชิงบวกสูง หมายถึง การที่ตัวแปรทั้งสองตัวมีแนวโน้มที่จะเปลี่ยนแปลงไปในทิศทางเดียวกันอย่างชัดเจน กล่าวคือ เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้น อีกตัวแปรก็จะมีแนวโน้มเพิ่มขึ้นตาม โดยมีค่าสัมประสิทธิ์สหสัมพันธ์ (Pearson's correlation coefficient) ใกล้เคียงกับ +1 ซึ่งสะท้อนถึงความสัมพันธ์เชิงเส้นที่มีความแข็งแรงในทิศทางบวก การตีความค่านี้นิยมใช้ในงานวิจัยที่ต้องการประเมินระดับความสัมพันธ์ระหว่างตัวแปรอย่างเป็นรูปธรรม เช่น ในสาขาวิทยาศาสตร์สุขภาพ สังคมศาสตร์ หรือพฤติกรรมศาสตร์

Mukaka (2012) แนะนำว่า ค่าความสัมพันธ์ที่อยู่ระหว่าง 0.70 ถึง 0.90 ถือว่าเป็นความสัมพันธ์เชิงบวกที่สูง (High positive correlation) และควรระวังการตีความให้ถูกต้องว่าเป็นความสัมพันธ์ ไม่ใช่เหตุและผล

การเลือกตัวแปรที่นำมาใช้ในแบบจำลองมีความสำคัญอย่างยิ่งต่อความแม่นยำของการพยากรณ์ การคัดเลือกโดยอิงจากความสัมพันธ์เชิงสถิติที่มีนัยสำคัญระหว่างตัวแปรสามารถช่วยลดปัญหา Overfitting และเพิ่มประสิทธิภาพของแบบจำลอง (James, Witten, Hastie, & Tibshirani, 2021)

งานวิจัยที่มีความสอดคล้องกับงานวิจัย "การวิเคราะห์เชิงเปรียบเทียบความสัมพันธ์ระหว่างข้อมูล เพื่อคัดเลือกตัวแปรที่มีศักยภาพในการพยากรณ์ด้วยการถดถอยเชิงเส้นอย่างง่าย" ประกอบด้วยงานวิจัยของ *Zou, Hastie และ Tibshirani (2005)* ชี้ให้เห็นว่า การใช้เทคนิควิเคราะห์สหสัมพันธ์ก่อนการสร้างแบบจำลองมีความสำคัญอย่างยิ่งต่อการเลือกตัวแปรที่เหมาะสม และสามารถเพิ่มความสามารถในการทำนายของแบบจำลองทางสถิติได้อย่างมีนัยสำคัญ งานวิจัยของ *Mukaka (2012)* ได้อธิบายแนวทางการเลือกใช้ค่าสัมประสิทธิ์สหสัมพันธ์ในบริบททางการแพทย์ โดยชี้ว่า การใช้ Pearson's r ในการตรวจสอบความสัมพันธ์ระหว่างตัวแปรเป็นจุดเริ่มต้นที่ดีในการพัฒนาแบบจำลองถดถอยที่มีความแม่นยำ และงานวิจัยของ *Rahman และ Zhang (2016)* ได้เสนอการวิเคราะห์เปรียบเทียบตัวแปรหลายคู่ในชุดข้อมูลขนาดใหญ่เพื่อค้นหาความสัมพันธ์ที่แข็งแรง และพบว่า การคัดเลือกคู่ที่มีค่าความสัมพันธ์สูงมีผลต่อความแม่นยำในการพยากรณ์ของแบบจำลองอย่างมีนัยสำคัญ รวมถึงงานวิจัยของ *James et al. (2021)* ได้กล่าวถึงความสำคัญของการเลือกตัวแปร (Feature Selection) โดยระบุว่า ขั้นตอนการวิเคราะห์ความสัมพันธ์ก่อนสร้างโมเดลจะช่วยจำกัดจำนวนตัวแปรให้แคบลงและรักษาความเรียบง่ายของแบบจำลอง ในขณะที่ยังรักษาความสามารถในการพยากรณ์ไว้อย่างมีประสิทธิภาพ งานวิจัยเรื่อง *ปัจจัยที่มีความสัมพันธ์กับความสำเร็จในการเป็นหมู่บ้านเศรษฐกิจพอเพียง* โดยนิสรา ใจซื่อ (2557) มีวัตถุประสงค์เพื่อศึกษาระดับความสำเร็จ ปัจจัยที่มีความสัมพันธ์ ปัญหาอุปสรรค และแนวทางพัฒนาหมู่บ้านเศรษฐกิจพอเพียง โดยเก็บข้อมูลจากสมาชิกสองหมู่บ้านในจังหวัดสระบุรี ใช้สถิติร้อยละ ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน และ ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient) วิเคราะห์ข้อมูล ผลการวิจัยพบว่า ปัจจัย 8 ด้าน

ได้แก่ เทคโนโลยี เศรษฐกิจ ทรัพยากรธรรมชาติ จิตใจ สังคม–วัฒนธรรม การมีส่วนร่วม ผู้นำหมู่บ้าน และการสนับสนุนจากรัฐ มีความสัมพันธ์อย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.01 และ 0.05 กับความสำเร็จของหมู่บ้านเศรษฐกิจพอเพียง โดยสรุป การพัฒนาอย่างยั่งยืนต้องอาศัยความร่วมมือระหว่างรัฐ ผู้นำ และสมาชิกหมู่บ้านที่ร่วมกันวางแผน วิเคราะห์ แก้ไขปัญหา และประเมินผลการพัฒนาอย่างต่อเนื่อง (นิสรา ใจซื่อ, 2557)

4. วิธีดำเนินการวิจัย/ระเบียบวิธีวิจัย/Research Methodology

การวิจัยนี้เป็นการวิจัยเชิงปริมาณ (Quantitative Research) โดยใช้ระเบียบวิธีแบบไม่ทดลอง (Non-experimental Method) ในลักษณะของการวิเคราะห์ข้อมูลทุติยภูมิ (Secondary Data Analysis) โดยมุ่งเน้นการวิเคราะห์เชิงเปรียบเทียบความสัมพันธ์ระหว่างตัวแปรจากหลายชุดข้อมูล ด้วยวิธีการวิเคราะห์สหสัมพันธ์ (Correlation Analysis) และการถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)

4.1 ประชากรของการวิจัยคือชุดข้อมูลทุติยภูมิที่ประกอบด้วยตัวแปรเชิงปริมาณจากหลายแหล่ง เช่น ข้อมูลพฤติกรรมผู้ใช้ ข้อมูลสุขภาพ ข้อมูลเศรษฐกิจ หรือข้อมูลภาครัฐที่เผยแพร่สู่สาธารณะ (Open Data)
กลุ่มตัวอย่าง ได้มาจากการเลือกชุดข้อมูลจำนวน 5 ชุดที่มีคุณสมบัติเหมาะสมในการวิเคราะห์ ดังนี้

1) Heart.csv เป็นชุดข้อมูลทางการแพทย์ที่บันทึกข้อมูลผู้ป่วยโรคหัวใจ โดยมีทั้งตัวแปรเชิงปริมาณและเชิงหมวดหมู่ เช่นอายุ (Age), อัตราการเต้นหัวใจสูงสุด (MaxHR), เพศ และผลการวินิจฉัยโรค ขนาดข้อมูลประมาณ 300 แถว และมีตัวแปรประมาณ 13–14 ตัวแปร ข้อมูลนี้มักใช้เป็นตัวอย่างในงานวิจัยเพื่อพัฒนาโมเดลการพยากรณ์ความเสี่ยงโรคหัวใจ

2) StudentMarks.csv เป็นข้อมูลเชิงปริมาณแบบต่อเนื่องที่ประกอบด้วยตัวแปรหลักสองตัวคือเวลาในการศึกษา (time_study) และคะแนนสอบ (Marks) มีประมาณ 20–25 แถวและ 2 คอลัมน์ ใช้เพื่อสอนและแสดงตัวอย่างการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย โดยแสดงถึงความสัมพันธ์เชิงเส้นระหว่างเวลาเรียนและผลสัมฤทธิ์ทางการเรียน

3) TestEnergyData.csv เป็นชุดข้อมูลเกี่ยวกับคุณลักษณะทางกายภาพของอาคารและการใช้พลังงาน เช่น ขนาดพื้นที่ใช้สอย (Square Footage), ความสูง, จำนวนชั้น และค่าการใช้พลังงาน ข้อมูลเป็นเชิงปริมาณต่อเนื่องทั้งหมด ขนาดประมาณ 768 แถวและ 8 ตัวแปร เหมาะสำหรับงานวิจัยด้านวิศวกรรมพลังงานเพื่อสร้างแบบจำลองทำนายการใช้พลังงาน

4) Salary_dataset.csv เป็นชุดข้อมูลเชิงปริมาณต่อเนื่องที่ประกอบด้วยตัวแปรประสบการณ์ทำงาน (YearsExperience) และเงินเดือน (Salary) มีขนาดประมาณ 30 แถวและ 2 คอลัมน์ ใช้เป็นตัวอย่างคลาสสิกในการสอนการถดถอยเชิงเส้นอย่างง่าย เนื่องจากแสดงความสัมพันธ์เชิงเส้นในระดับสูงระหว่างปีประสบการณ์และเงินเดือน

5) Advertising.csv เป็นชุดข้อมูลเชิงปริมาณที่รวบรวมข้อมูลค่าใช้จ่ายโฆษณาผ่านสื่อ 3 ช่องทาง (TV, Radio, Newspaper) และยอดขาย (Sales) มีขนาดประมาณ 200 แถวและ 4 คอลัมน์ มักใช้สอนการถดถอยเชิงเส้นทั้งแบบง่ายและพหุคูณ โดยเน้นการวิเคราะห์ผลกระทบของงบโฆษณาต่อยอดขายสินค้า

4.2 เครื่องมือที่ใช้ในการวิจัยประกอบด้วย

1) ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson Correlation Coefficient)
2) แบบจำลองการถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)
3) ค่า R^2 ค่า p-value ค่า MAE/MSE สำหรับประเมินความแม่นยำของโมเดล

4.3 ขั้นตอนการดำเนินงานวิจัย

1) การคัดเลือกข้อมูล ด้วยการรวบรวมชุดข้อมูลจากแหล่ง Open Data ที่น่าเชื่อถือ

2) การจัดเตรียมข้อมูล ด้วยหลักการของการทำความสะอาดข้อมูล (Data Cleaning) และการจัดรูปแบบข้อมูลให้อยู่ในรูปแบบเชิงปริมาณที่เหมาะสม

3) การวิเคราะห์สหสัมพันธ์ระหว่างคู่ตัวแปรทั้งหมดในแต่ละชุดข้อมูลด้วย Pearson's r

4) การคัดเลือกคู่ตัวแปรที่มีความสัมพันธ์เชิงบวกสูง เช่น r > 0.80 ซึ่งเป็นเกณฑ์ถูกใช้เพื่อระบุความสัมพันธ์เชิงบวกในระดับสูงมาก (Evans, 1996; Dancey & Reidy, 2004) เพื่อเข้าสู่การทดสอบด้วยการถดถอย

5) การถดถอยเชิงเส้นอย่างง่าย ด้วยการทดสอบความสามารถของตัวแปรในการพยากรณ์ โดยใช้โมเดล Simple Linear Regression และวิเคราะห์ผลลัพธ์

6) การเปรียบเทียบประสิทธิภาพ ด้วยค่า R^2 ค่า RMSE และค่า MAE ของคู่ตัวแปรที่ผ่านการคัดเลือกเพื่อหาโมเดลที่เหมาะสมที่สุด

7) สรุปผลการวิเคราะห์ ด้วยการนำเสนอคู่ตัวแปรที่เหมาะสมในการพยากรณ์ พร้อมข้อเสนอแนะ

5. ผลการวิจัยและอภิปรายผลการวิจัย

จากการนำชุดข้อมูลจากแหล่งข้อมูลตัวอย่างต่าง ๆ เมื่อนำมาหาค่าความสัมพันธ์ด้วยเทคนิคการวิเคราะห์สหสัมพันธ์เพียร์สัน (Pearson Correlation) คณะผู้วิจัยได้ดำเนินการวิเคราะห์เชิงเปรียบเทียบความสัมพันธ์ระหว่างตัวแปรในชุดข้อมูลจำนวน 5 ชุด (UCI Machine Learning Repository, 1988; Kaggle, 2021; James et al., 2021) เพื่อคัดเลือกคู่ตัวแปรที่มีความสัมพันธ์เชิงบวกสูง เหมาะสมในการนำไปสร้างแบบจำลองการพยากรณ์ โดยมีผลการวิจัยตามวัตถุประสงค์ดังนี้

5.1 ผลการวิเคราะห์ความสัมพันธ์เชิงเส้นระหว่างตัวแปรจากชุดข้อมูลหลายชุด โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ดังต่อไปนี้

ตารางที่ 1 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร MaxHR และ Age

Feature	Age	MaxHR
Age	1.000000	-0.382045
MaxHR	-0.382045	1.000000

จากตารางที่ 1 ผลการวิเคราะห์แสดงว่าตัวแปร Age และ MaxHR มีความสัมพันธ์เชิงลบในระดับต่ำถึงปานกลาง คือ r = -0.382 แสดงว่าเมื่ออายุเพิ่มขึ้น อัตราการเต้นหัวใจสูงสุดมีแนวโน้มลดลง แม้จะสอดคล้องตามทฤษฎีทางสรีรวิทยา แต่ความสัมพันธ์ในระดับนี้อาจไม่เพียงพอสำหรับการทำนายด้วยโมเดลเชิงเส้นอย่างง่ายเพียงตัวแปรเดียว

ตารางที่ 2 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร Marks และ time_study

Feature	Marks	time_study
Marks	0.942254	1.000000
time_study	1.000000	0.942254

จากตารางที่ 2 ผลการวิเคราะห์แสดงว่าตัวแปร Marks และ time_study มีความสัมพันธ์เชิงบวกในระดับสูงมาก คือ r = 0.942 แสดงว่าการเพิ่มเวลาในการศึกษาเป็นปัจจัยที่มีอิทธิพลสำคัญต่อการเพิ่มคะแนนสอบ ทำให้ time_study เป็นตัวแปรต้นที่เหมาะสมอย่างยิ่งสำหรับการสร้างแบบจำลองการพยากรณ์ด้วยการถดถอยเชิงเส้นอย่างง่าย

ตารางที่ 3 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร Square Footage และ Energy Consumption

Feature	Square Footage	Energy Consumption
Square Footage	1.000000	0.724039
Energy Consumption	0.724039	1.000000

จากตารางที่ 3 ผลการวิเคราะห์แสดงว่าตัวแปร Square Footage และ Energy Consumption มีความสัมพันธ์เชิงบวกในระดับปานกลางถึงสูง คือ r = 0.724 แสดงว่าการเพิ่มพื้นที่ใช้สอยสัมพันธ์กับการใช้พลังงานที่มากขึ้น เหมาะสำหรับใช้เป็นตัวแปรต้นในแบบจำลองถดถอยเชิงเส้นอย่างง่าย แต่ควรพิจารณาปัจจัยอื่นเพิ่มเติมเพื่อเพิ่มความแม่นยำในการพยากรณ์

ตารางที่ 4 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร Salary และ YearsExperience

Feature	Salary	YearsExperience
Salary	0.978242	1.000000
YearsExperience	1.000000	0.978242

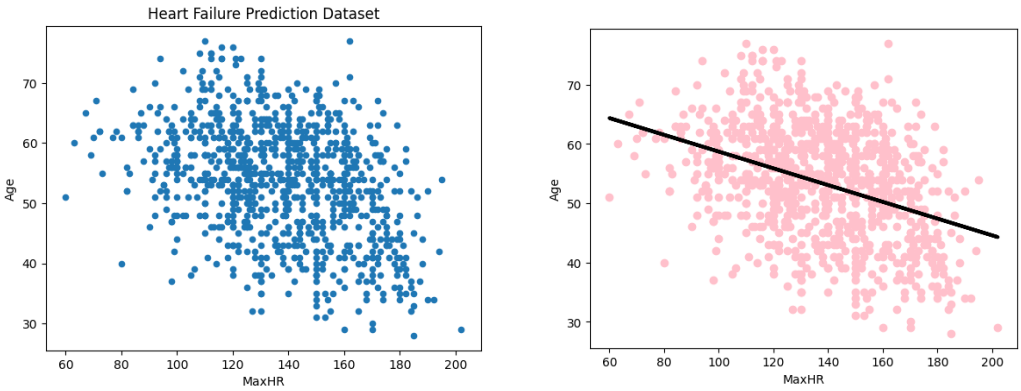
จากตารางที่ 4 ผลการวิเคราะห์แสดงว่าตัวแปร Salary และ YearsExperience มีความสัมพันธ์เชิงบวกในระดับที่สูงมาก คือ r = 0.978 แสดงว่าประสบการณ์ทำงานเป็นปัจจัยสำคัญที่ทำนายเงินเดือนได้อย่างแม่นยำ เหมาะสำหรับใช้เป็นตัวแปรต้นในแบบจำลองถดถอยเชิงเส้นอย่างง่ายเพื่อการพยากรณ์เงินเดือน

ตารางที่ 5 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร Sales และ TV

Feature	Sales	TV
Sales	0.901208	1.000000
TV	1.000000	0.901208

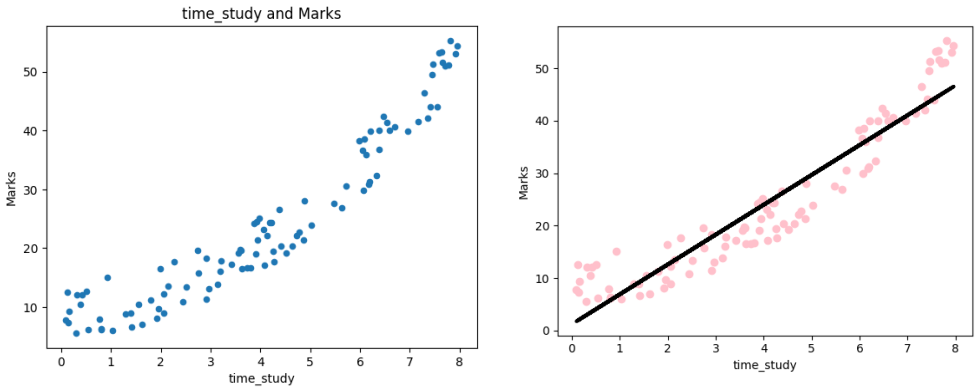
จากตารางที่ 5 ผลการวิเคราะห์แสดงว่าตัวแปร Sales และ TV มีความสัมพันธ์เชิงบวกในระดับที่สูงมาก คือ r = 0.901 แสดงว่าการใช้งบโฆษณาทางโทรทัศน์เป็นตัวแปรต้นที่เหมาะสมสำหรับทำนายยอดขายผ่านแบบจำลองถดถอยเชิงเส้นอย่างง่ายได้อย่างมีประสิทธิภาพ

5.2 ผลการสร้างแบบจำลองการถดถอยเชิงเส้นอย่างง่าย ได้ผลการวิเคราะห์ดังรูปที่ 1-5



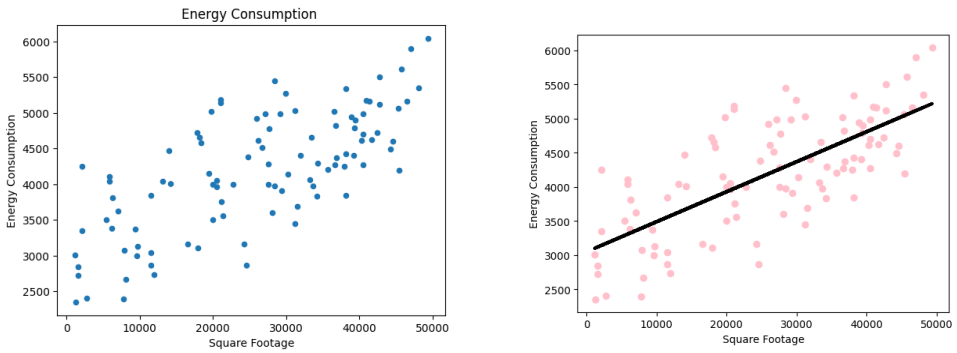
(a) (b)

รูปที่ 1 แผนภาพการกระจายตัวของตัวแปร MaxHR และ Age



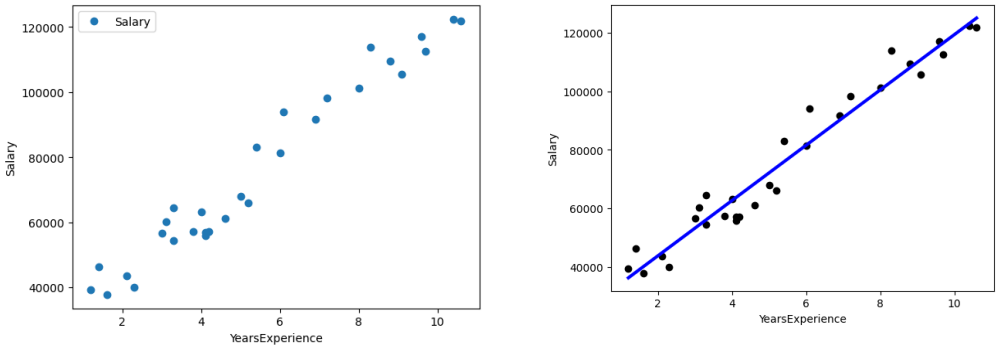
(a) (b)

รูปที่ 2 แผนภาพการกระจายตัวของตัวแปร time_study และ Marks



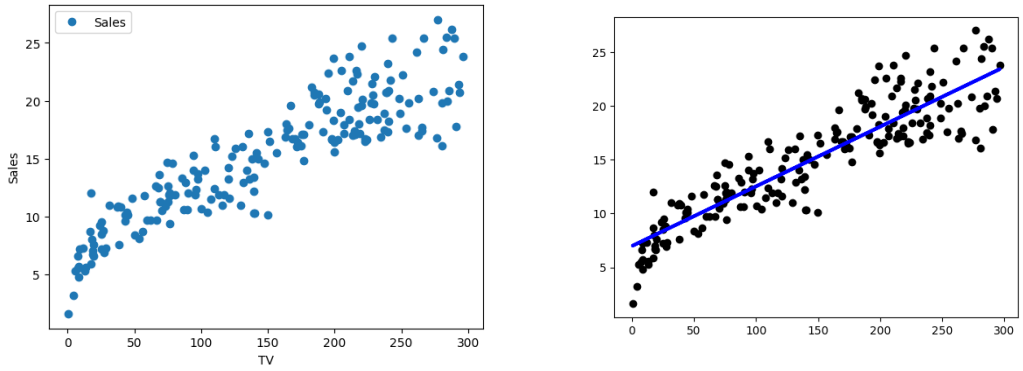
(a) (b)

รูปที่ 3 แผนภาพการกระจายตัวของตัวแปร SquareFootage และ EnergyConsumption



(a) (b)

รูปที่ 4 แผนภาพการกระจายตัวของตัวแปร YearsExperience และ Salary



(a) (b)

รูปที่ 5 แผนภาพการกระจายตัวของตัวแปร TV และ Sales

5.3 ผลประเมินประสิทธิภาพของโมเดลด้วยตัวชี้วัด ได้แก่ ค่า R^2 (Coefficient of Determination) ค่า RMSE (Root Mean Squared Error) และค่า MAE (Mean Absolute Error) ดังต่อไปนี้

ตารางที่ 6 การเปรียบเทียบชุดข้อมูลด้วยตัวชี้วัดการประเมินโมเดล

No.	Datasets	Predictor (X)	Response (y)	R^2	RMSE	MAE
1	Heart.csv	MaxHR	Age	0.1460	8.71	7.06
2	StudentMarks.csv	time_study	Marks	0.8878	4.77	3.89
3	TestEnergyData.csv	SquareFootage	EnergyConsumption	0.5242	571.39	477.20
4	Salary_dataset.csv	YearsExperience	Salary	0.9570	5592.04	4644.20
5	Advertising.csv	TV	Sales	0.8122	2.28	1.83

จากการเปรียบเทียบค่าตัวชี้วัดทั้งหมด พบว่าแบบจำลองจาก ชุดข้อมูล Salary_dataset.csv มีประสิทธิภาพสูงสุด โดยมีค่า R^2 = 0.9570 ซึ่งหมายถึง ตัวแปร “ประสบการณ์ทำงาน (Years of Experience)” สามารถอธิบายความแปรปรวนของ “เงินเดือน (Salary)” ได้ถึง 95.7% และมีค่า RMSE และ MAE อยู่ในระดับที่ยอมรับได้ แม้จำนวนเต็มจะสูง แต่เป็นไปตามขนาดของค่าจริงในชุดข้อมูล รองลงมาคือแบบจำลองจาก ชุดข้อมูล StudentMarks.csv และ Advertising.csv ที่มีค่า R^2 สูงกว่า 0.80 แสดงว่าตัวแปรต้นสามารถทำนายตัวแปรตามได้อย่างแม่นยำในระดับสูง ในทางตรงกันข้าม ชุดข้อมูล Heart.csv มีความสัมพันธ์ต่ำมาก (R^2 = 0.1460) จึงไม่เหมาะแก่การนำมาใช้สร้างแบบจำลองเชิงพยากรณ์ โดยทั่วไปค่า RMSE มักมีค่ามากกว่า MAE โดยเฉพาะเมื่อชุดข้อมูลมีค่าคลาดเคลื่อนที่แตกต่างกันสูงหรือมี Outlier ในขณะที่หากข้อมูลมี Error ที่กระจายตัวสม่ำเสมอ ค่า RMSE และ MAE จะใกล้เคียงกัน การเปรียบเทียบค่า RMSE และ MAE จึงช่วยให้ผู้วิจัยสามารถประเมินรูปแบบการกระจายของ Error ได้ หาก RMSE สูงกว่า MAE อย่างมาก แสดงว่าโมเดลอาจมีปัญหาความคลาดเคลื่อนขนาดใหญ่บางจุดซึ่งควรได้รับการแก้ไขหรือตรวจสอบเพิ่มเติม

อภิปรายผลการวิจัย

ผลการวิจัยสอดคล้องกับหลักการทางสถิติ ในด้านของตัวแปรต้นและตัวแปรตามควรมีความสัมพันธ์เชิงเส้นในระดับสูง จึงจะสามารถนำไปสร้างแบบจำลองที่มีค่าความแม่นยำได้ (Montgomery et al., 2012; James et al., 2021) โดยเฉพาะในกรณีของ Salary_dataset.csv ซึ่งค่า R^2 สูงมากสะท้อนถึงความเป็นเชิงเส้นที่ชัดเจนระหว่างจำนวนปีประสบการณ์และเงินเดือน นอกจากนี้ การที่ StudentMarks.csv มีความสัมพันธ์สูงระหว่างเวลาในการศึกษาและคะแนนสอบ สอดคล้องกับทฤษฎีการเรียนรู้ที่ว่า เวลาในการฝึกฝนส่งผลต่อผลสัมฤทธิ์ทางการเรียน ซึ่งสนับสนุนแนวคิดในทางจิตวิทยาการศึกษา (Bloom, 1976) อย่างไรก็ตาม ชุดข้อมูลบางชุด เช่น Heart.csv มีค่า R^2 ต่ำ ซึ่งอาจเนื่องมาจากความสัมพันธ์ที่ไม่เป็นเชิงเส้น หรือมีตัวแปรแทรกซ้อนจำนวนมาก ส่งผลให้การถดถอยเชิงเส้นอย่างง่ายไม่สามารถอธิบายได้เพียงพอ ดังนั้น ในกรณีเหล่านี้ควรพิจารณาใช้ การถดถอยพหุคูณ หรือ แบบจำลองที่ไม่เป็นเชิงเส้นแทน

เพื่อเพิ่มประสิทธิภาพของการพยากรณ์ในอนาคต ควรพิจารณาการเปรียบเทียบผลลัพธ์จากการถดถอยเชิงเส้นอย่างง่ายกับ เทคนิคการถดถอยแบบไม่เป็นเชิงเส้น เช่น Polynomial Regression หรือ Spline Regression ซึ่งสามารถจับความสัมพันธ์ที่มีรูปแบบโค้งได้ดีกว่า นอกจากนี้ ควรใช้ Machine Learning เช่น Decision Trees หรือ Random Forest เพื่อรองรับความซับซ้อนของข้อมูล พร้อมทั้งประเมินค่า R^2, RMSE และ MAE อย่างเป็นระบบ รวมถึงตรวจสอบ Residual Plot เพื่อเลือกโมเดลที่เหมาะสมที่สุดต่อข้อมูลแต่ละชุด และนักวิเคราะห์ข้อมูลควรเริ่มต้นด้วยการตรวจสอบ ความสัมพันธ์เชิงสถิติ ระหว่างตัวแปรโดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson Correlation Coefficient) เพื่อคัดเลือกคู่ตัวแปรที่มี ความสัมพันธ์เชิงเส้นในระดับสูง ซึ่งสามารถช่วยลดจำนวนตัวแปรที่ไม่สำคัญและเพิ่มประสิทธิภาพของแบบจำลองพยากรณ์ ควรใช้ ค่าเกณฑ์ (threshold) ที่เหมาะสม เช่น $r > 0.80$ เพื่อบ่งบอกถึงความสัมพันธ์เชิงบวกในระดับสูง (Evans, 1996) และพิจารณาตัวแปรที่มีความหมายในเชิงสาเหตุหรือเชิงทฤษฎีร่วมด้วย ไม่ควรอาศัยค่า Correlation เพียงอย่างเดียว นอกจากนี้ ควรทำ Residual Analysis หลังสร้างโมเดลถดถอย เพื่อตรวจสอบสมมติฐานเชิงเส้น และหากพบความสัมพันธ์ไม่เป็นเชิงเส้น ควรพิจารณาเทคนิคอื่น เช่น การถดถอยแบบไม่เป็นเชิงเส้น หรือวิธี Machine Learning เพื่อเพิ่มความแม่นยำของการพยากรณ์ในบริบทข้อมูลจริง

6. สรุปผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรในชุดข้อมูลจำนวน 5 ชุด โดยใช้การวิเคราะห์สหสัมพันธ์ และการถดถอยเชิงเส้นอย่างง่าย เพื่อคัดเลือกคู่ตัวแปรที่มีความสัมพันธ์เชิงบวกสูง และมีศักยภาพในการนำไปพยากรณ์ค่าของตัวแปรตาม จากผลการวิเคราะห์ พบว่า ชุดข้อมูล Salary_dataset.csv มีค่าความสัมพันธ์สูงสุด โดยมีค่า

R^2=0.9570 แสดงว่าตัวแปร “Years of Experience” สามารถอธิบายการเปลี่ยนแปลงของ “Salary” ได้มากถึง 95.7% ส่วนชุดข้อมูล StudentMarks.csv และ Advertising.csv ก็แสดงความสัมพันธ์ ในระดับสูงเช่นกัน โดยมีค่า R^2=0.8878 และ R^2=0.8122 ตามลำดับ ขณะที่ชุดข้อมูล Heart.csv และ TestEnergyData.csv มีค่า R^2 ต่ำกว่า 0.60 แสดงว่าไม่เหมาะแก่การนำไปใช้สร้างแบบจำลองเชิงพยากรณ์

7. กิตติกรรมประกาศ

ขอขอบคุณผู้ช่วยศาสตราจารย์ณัฐกานต์ โตนวล อาจารย์ประจำวิชา Data Innovation for Digital Business ที่ได้ให้คำแนะนำในการเขียนบทความวิจัยในครั้งนี้เป็นอย่างสูง

8. เอกสารอ้างอิง

นิสรา ใจซื่อ. (2557). *ปัจจัยที่มีความสัมพันธ์กับความสำเร็จในการเป็นหมู่บ้านเศรษฐกิจพอเพียง*. มหาวิทยาลัยธุรกิจบัณฑิตย์.

Bloom, B. S. (1976). Human characteristics and school learning. McGraw-Hill.

Dancey, C. P., & Reidy, J. (2004). *Statistics without maths for psychology* (3rd ed.). Pearson Education.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Brooks/Cole.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Advertising Dataset* [Data set]. In *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://www.statlearning.com/resources-second-edition

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kaggle. (2018). *Simple Linear Regression: Salary Prediction Dataset*. Kaggle Datasets. https://www.kaggle.com/datasets/krantisurve/simple-linear-regression

Kaggle. (2021). *Student Study Hours vs Marks Dataset*. Kaggle Datasets. https://www.kaggle.com/datasets/sagnik1511/student-study-hours-vs-marks

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71. https://doi.org/10.4314/mmj.v24i3.84676

Rahman, M., & Zhang, Y. (2016). An empirical analysis of correlation-based feature selection techniques for high dimensional bioinformatics data. *BMC Bioinformatics, 17*(1), 1–17. https://doi.org/10.1186/s12859-016-1257-2

Zou, H., Hastie, T., & Tibshirani, R. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

UCI Machine Learning Repository. (1988). *Heart Disease Data Set*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/heart+Disease

UCI Machine Learning Repository. (2012). *Energy Efficiency Data Set*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Energy+efficiency