การวิเคราะห์เชิงเปรียบเทียบแบบจำลองการถดถอย: กรณีศึกษา KNN Regression เทียบกับ Multiple Linear Regression

Comparative Analysis of Regression Models: A Case Study of KNN Regression vs. Multiple Linear Regression

กรรณิการ์ จะกอ [1] และ ปิยพันธ์ สุวรรณเวช [2*]

Kunnika Jakor [1] and Piyapan Suwannawach [2*]

[1]สาขาวิชาการเงิน คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลพระนคร

[1]Department of Finance, Faculty of Business Administration, Rajamangala University of Technology Phra Nakhon, Thailand

[2]สาขาวิชาระบบสารสนเทศ คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลพระนคร

[2]Department of Information Systems, Faculty of Business Administration, Rajamangala University of Technology Phra Nakhon, Thailand

***Corresponding Author E-mail**: piyapan.s@rmutp.ac.th

Received: 26/09/25, **Revised**: 11/11/25, **Accepted**: 14/11/25

บทคัดย่อ

บทความวิจัยนี้นำเสนอการประยุกต์ใช้ K-Nearest Neighbors (KNN) Algorithm ในฐานะแบบจำลองการถดถอย (Regression Model) สำหรับการทำนายค่าตัวเลข โดยนำเสนอ KNN Regression เป็นแบบจำลองทางเลือกสำหรับปัญหาที่ความสัมพันธ์ของข้อมูลไม่เป็นเชิงเส้นหรือเป็นเชิงเส้นอย่างไม่ชัดเจน ซึ่งแบบจำลองการถดถอยเชิงเส้นทั่วไปมักมีข้อจำกัด ผลการทดลองมุ่งเน้นการเปรียบเทียบประสิทธิภาพของ KNN Regression กับแบบจำลองการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression) กับชุดข้อมูลจริงจำนวน 5 ชุด เพื่อแสดงให้เห็นถึงศักยภาพของ KNN Regression ในการทำนายค่าตัวเลขได้อย่างแม่นยำเพิ่มขึ้น นอกจากนั้นยังได้ศึกษาถึงการปรับแต่งพารามิเตอร์ของ KNN Regression อย่างละเอียด เช่น ผลกระทบของการเลือกค่า K (จำนวนเพื่อนบ้าน) วิธีการวัดระยะห่าง (Distance Metrics) รวมทั้งเทคนิค Inverse Distance Weighting (IDW) ในการถ่วงน้ำหนักเพื่อนบ้าน ผลการวิจัยนี้ชี้ให้เห็นว่า KNN Regression เป็นทางเลือกของแบบจำลองการถดถอยที่มีประสิทธิภาพ และน่าสนใจสำหรับการทำนายค่าตัวเลข โดยเฉพาะอย่างยิ่งในสถานการณ์ที่ข้อมูลมีความซับซ้อนและไม่มีความสัมพันธ์เชิงเส้นตรงที่ชัดเจน โดยไม่จำเป็นต้องใช้แบบจำลองที่ซับซ้อนอย่างโครงข่ายประสาทเทียม

คำสำคัญ: การถดถอยแบบเพื่อนบ้านใกล้ที่สุด, การถดถอยเชิงเส้น, การถดถอยเชิงเส้นพหุคูณ, การสร้างแบบจำลองเชิงทำนาย, การศึกษาเชิงเปรียบเทียบ

Abstract

This research paper presents the application of the K-Nearest Neighbors (KNN) algorithm as a regression model for numerical prediction. We propose for KNN regression as a viable alternative for scenarios characterized by non-linear or ambiguously linear data relationships, where conventional linear regression models frequently underperform. Our experimental findings concentrate on evaluating the efficiency of KNN regression in comparison to established models such as multiple linear regression across five datasets. This illustrates the capability of KNN regression to achieve more accurate numerical predictions. In addition, we explore the effects of distance metrics, the inverse distance weighting (IDW) method for neighbor weighting, and K-value selection (number of neighbors) in our in-depth parameter tuning for KNN regression. The results suggest that KNN regression is an efficient and compelling alternative regression model for numerical prediction, particularly when dealing with complicated data and ambiguous linear correlations. Thus relieves the need for more complex models like artificial neural networks.

Keywords: KNN Regression, Linear Regression, Multiple Linear Regression, Predictive Modeling, Comparative Study

1. บทนำ

การวิเคราะห์การถดถอย (Regression Analysis) เป็นหนึ่งในเครื่องมือสำคัญในงานวิทยาการข้อมูลและการเรียนรู้ของเครื่อง โดยมีบทบาทอย่างยิ่งในการทำนายค่าตัวเลขต่อเนื่องในหลากหลายการประยุกต์ใช้งาน อาทิเช่น การคาดการณ์ราคาอสังหาริมทรัพย์ การประเมินประสิทธิภาพพลังงานในอาคาร หรือการทำนายกำลังอัดของวัสดุ เป็นต้น การทำนายค่าที่แม่นยำเหล่านี้เป็นหัวใจสำคัญในการสนับสนุนการตัดสินใจเชิงกลยุทธ์และการพัฒนาองค์ความรู้เพื่อทราบแนวโน้วความสัมพันธ์ของข้อมูล [1-2] อย่างไรก็ตามการสร้างแบบจำลองการถดถอยที่มีประสิทธิภาพสูงนั้นเป็นความท้าทายโดยเฉพาะเมื่อลักษณะความสัมพันธ์ของข้อมูลมีความซับซ้อน

แบบจำลองการถดถอยเชิงเส้น (Linear Regression: LR) และแบบจำลองการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression: MLR) เป็นวิธีการที่ได้รับความนิยมอย่างแพร่หลาย เนื่องจากมีความเรียบง่ายและง่ายต่อการตีความ แต่แบบจำลองเหล่านี้มีข้อจำกัดพื้นฐานที่สำคัญ คือ ต้องอาศัยสมมุติฐานว่าความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเป็นแบบเชิงเส้น ในทางปฏิบัติข้อมูลจำนวนมากมักสะท้อนความสัมพันธ์ที่มีลักษณะซับซ้อนหรือไม่เชิงเส้น (nonlinear) ซึ่งทำให้ประสิทธิภาพการทำนายของแบบจำลองเชิงเส้นลดลงและอาจนำไปสู่คำตอบหรือผลลัพธ์ที่คลาดเคลื่อน [3-4]

เพื่อแก้ไขข้อจำกัดดังกล่าววิธีเพื่อนบ้านใกล้ที่สุด K ตัว (K-Nearest Neighbors: KNN) ซึ่งเป็นอัลกอริทึมที่รู้จักกันดีในงานจัดหมวดหมู่ (classification) [5-6] ได้รับการพิจารณาว่าเป็นทางเลือกที่มีศักยภาพสามารถนำมาประยุกต์ใช้เป็นแบบจำลองการถดถอยด้วยเพื่อนบ้านใกล้เคียงที่สุด K ตัว (K-Nearest Neighbors Regression: KNN Regression) ด้วยคุณสมบัติของการเป็นแบบจำลองที่ไม่ขึ้นกับพารามิเตอร์ (non-parametric) และการเรียนรู้เชิงตัวอย่าง (Instance-based Learning) ทำให้ KNN Regression ไม่จำเป็นต้องอาศัยสมมุติฐานเชิงเส้นตรงของข้อมูล และยังสามารถจับความสัมพันธ์ที่ซับซ้อนหรือไม่เป็นเชิงเส้นได้โดยธรรมชาติ ซึ่งสามารถให้ผลลัพธ์การทำนายที่ดีกว่าแบบจำลองเชิงเส้น โดยไม่ต้องขยับไปใช้แบบจำลองที่มีความซับซ้อนอย่างเช่น โครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) เป็นต้น ซึ่งมีหลายงานวิจัยที่ยืนยันว่า KNN Regression สามารถให้ผลลัพธ์ที่ดีกว่าแบบจำลองเชิงเส้นพหุคูณในหลายกรณี [7-10] อย่างไรก็ตามประสิทธิภาพของ KNN Regressor ยังคงขึ้นอยู่กับการเลือกพารามิเตอร์ (parameter) ที่เหมาะสม เช่น จำนวนเพื่อนบ้านที่ใกล้ที่สุด (K) และวิธีวัดระยะห่าง (Distance Metrics) เป็นต้น [11-13]

ดังนั้นงานวิจัยนี้จึงมีวัตถุประสงค์หลักเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง KNN Regression กับแบบจำลอง Multiple Linear Regression ในการทำนายค่าตัวเลขบนชุดข้อมูลจริงที่มีลักษณะแตกต่างกันทั้งหมด 5 ชุดข้อมูล ได้แก่ Auto MPG, Boston Housing, California Housing Prices, Concrete Compressive Strength และ Energy Efficient นอกจากนี้ยังมุ่งเน้นการศึกษาผลกระทบของการปรับแต่งพารามิเตอร์ของแบบจำลอง KNN Regression อย่างละเอียด ทั้งการเลือกจำนวนเพื่อนบ้านที่ใกล้ที่สุด (K) ที่เหมาะสมที่สุด การใช้วิธีการวัดระยะห่างที่หลากหลาย และการพิจารณาการถ่วงน้ำหนักโดยใช้เทคนิคการถ่วงน้ำหนักแบบผกผันระยะทาง (Inverse Distance Weighting: IDW) เพื่อเพิ่มความแม่นยำในการทำนาย

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การถดถอยเชิงเส้นพหุคูณ

แบบจำลอง Multiple Linear Regression เป็นการพัฒนาต่อยอดจากแบบจำลอง Linear Regression เพื่อให้สามารถใช้ตัวแปรอิสระได้หลายตัว $\left(x_1, x_2, ..., x_n\right)$ ในการทำนายตัวแปรตามเพียงหนึ่งตัว แบบจำลองนี้ยังคงพิจารณาความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม แต่แทนที่จะเป็นเส้นตรงในสองมิติจะเป็นการค้นหาระนาบ (plane) ของคำตอบหรือไฮเปอร์เพลน (hyperplane) ที่เหมาะสมที่สุดในพื้นที่ที่มีหลายมิติตามจำนวนของตัวแปรสมการพื้นฐานสำหรับแบบจำลองการถดถอยเชิงเส้นพหุคูณแสดงได้ดังสมการที่ (1)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \quad (1)$$

เมื่อ

Y คือ ตัวแปรตาม

$X_1, X_2, ..., X_n$ คือ ตัวแปรอิสระ n ตัว

β_0 คือ จุดตัดแกน Y

$\beta_1, \beta_2, ..., \beta_n$ คือ สัมประสิทธิ์การถดถอยสำหรับตัวแปรอิสระแต่ละตัว

ε คือ ค่าความคลาดเคลื่อน

2.2 การถดถอยแบบวิธีเพื่อนบ้านใกล้ที่สุด K ตัว

KNN เป็นหนึ่งในอัลกอริทึมการเรียนรู้ของเครื่องที่ไม่ขึ้นกับพารามิเตอร์และเป็นการเรียนรู้เชิงตัวอย่างที่เป็นที่เข้าใจง่ายและถูกนำไปใช้อย่างแพร่หลาย โดยทั่วไปแล้ว KNN มักจะเป็นที่รู้จักและใช้ในงานจัดหมวดหมู่ (Classification) ซึ่งมีหลักการคือการจำแนกข้อมูลใหม่ให้อยู่ในกลุ่มของเพื่อนบ้านที่ใกล้ที่สุดส่วนใหญ่ เช่น หากมีจุดข้อมูลใหม่ที่ต้องการจำแนกประเภท อัลกอริทึมจะค้นหา K จุดข้อมูลที่อยู่ใกล้ที่สุดในชุดข้อมูลการฝึกฝน และกำหนดประเภทให้กับจุดข้อมูลใหม่นั้นตามคะแนนเสียงส่วนใหญ่ (majority vote) ของเพื่อนบ้านทั้ง K ตัว อย่างไรก็ตาม KNN Algorithm ยังสามารถนำมาประยุกต์ใช้ในงานการถดถอย (regression) ได้เช่นกัน โดยมีหลักการพื้นฐานที่คล้ายคลึงกัน แต่แทนที่จะใช้การโหวตเพื่อกำหนดประเภท จะเป็นการหาค่าเฉลี่ยของค่าตัวแปรตาม โดยหลักการทำงานของ KNN Regression มีขั้นตอนดังต่อไปนี้

1) การคำนวณระยะห่าง เมื่อมีจุดข้อมูลใหม่ x_q ที่ต้องการทำนายค่าตัวแปรตาม ระบบจะคำนวณระยะห่าง (distance) ระหว่างจุดข้อมูลใหม่นี้กับทุกจุดข้อมูล x_i ในชุดข้อมูลการฝึกฝน

2) การระบุเพื่อนบ้าน ระบุและเลือกจุดข้อมูล K จุดที่อยู่ใกล้ที่สุด $\left(x_1, x_2, ..., x_K\right)$ กับจุดข้อมูลใหม่ โดยใช้เกณฑ์การวัดระยะห่างที่กำหนดไว้

3) การทำนายค่า ค่าทำนาย $\hat{y}_q$ ของตัวแปรตามสำหรับจุดข้อมูลใหม่นั้น จะได้มาจากการหาค่าเฉลี่ยของค่าตัวแปรตาม (y_i) ของ K จุดข้อมูลเพื่อนบ้านที่ถูกเลือกมา ดังสมการที่ (2)

$$\hat{y}_q = \frac{1}{K}\sum_{i=1}^{K} y_i \tag{2}$$

เมื่อ

$\hat{y}_q$ คือ ค่าทำนายของตัวแปรตามสำหรับจุดข้อมูลใหม่ x_q

K คือ จำนวนเพื่อนบ้านที่ใกล้ที่สุด

y_i คือ ค่าตัวแปรตามของเพื่อนบ้านแต่ละจุด

นอกจากนี้ความแม่นยำในการทำนายอาจเพิ่มขึ้นได้โดยการใช้การถ่วงน้ำหนักด้วยระยะห่างผกผัน (Inverse Distance Weighting - IDW) ซึ่งจะให้ความสำคัญกับเพื่อนบ้านที่อยู่ใกล้กว่า โดยให้น้ำหนักที่มากกว่ากับเพื่อนบ้านที่อยู่ใกล้ และให้น้ำหนักที่น้อยกว่ากับเพื่อนบ้านที่อยู่ห่างออกไป ซึ่งการทำนายค่าด้วยการหาค่าเฉลี่ยถ่วงน้ำหนักด้วยระยะห่างผกผันสามารถแสดงได้ดังสมการที่ 3

$$\hat{y}_q = \frac{\sum_{i=1}^{K} w_i y_i}{\sum_{i=1}^{K} w_i} \tag{3}$$

เมื่อ

w_i คือ น้ำหนักของเพื่อนบ้านแต่ละจุด โดยทั่วไปกำหนดให้

$$w_i = \frac{1}{d\left(x_q, x_i\right)}$$

$d\left(x_q, x_i\right)$ คือ ระยะห่างระหว่างจุดข้อมูลใหม่ x_q กับเพื่อนบ้าน x_i

KNN Regression มีข้อดีที่สำคัญคือไม่ต้องอาศัยสมมติฐานเชิงเส้นตรงของข้อมูล ทำให้สามารถจัดการกับความสัมพันธ์ที่ซับซ้อนหรือไม่เป็นเชิงเส้นได้อย่างเป็นธรรมชาติ อย่างไรก็ตามประสิทธิภาพของ KNN Regression ขึ้นอยู่กับการเลือกค่า K และวิธีการวัดระยะห่างที่เหมาะสม

2.3 การวัดระยะห่าง (Distance Metrics)

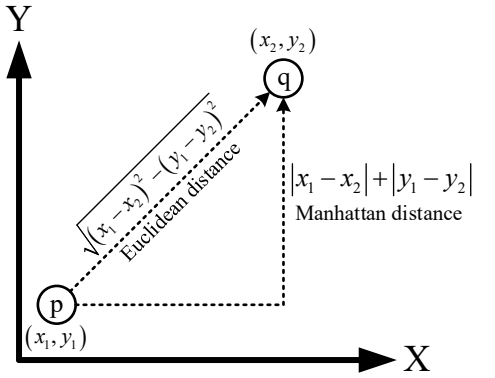
การเลือกวิธีการวัดระยะห่างที่เหมาะสมมีผลต่อประสิทธิภาพของ KNN เนื่องจากเป็นตัวกำหนดว่า “เพื่อนบ้าน” ถูกนิยามอย่างไร ระยะห่างที่นิยมใช้ได้แก่ ระยะทางแบบยูคลิด (Euclidean Distance) เป็นระยะทางที่ใช้กันแพร่หลายที่สุด ซึ่งสามารถคำนวณได้โดยกฎของปีทากอรัสสามารถเขียนเป็นสมการได้ดังนี้ $c^2 = a^2 + b^2$ เมื่อ c เป็นความยาวของด้านตรงข้ามมุมฉาก และ a กับ b เป็นด้านประกอบมุมฉากทั้งสอง ซึ่งถ้าให้ a และ b เป็นมิติสองมิติ จะได้ค่า c เป็นระยะทางยูคลิดใน 2 มิติ และเมื่อทำการขยายผลไปโดยการเพิ่มจำนวนมิติที่สูงขึ้นเป็น 3 มิติ โดยให้ g เป็นมิติ และ h เป็นระยะทางยูคลิดในปริภูมิยูคลิดที่มี 3 มิติ โดยอาศัยกฎของปีทากอรัสจะสามารถเขียนเป็นสมการดังนี้ $h^2 = g^2 + c^2 = g^2 + a^2 + b^2$ ซึ่งสามารถอธิบายเป็นคำพูดได้ว่าระยะทางยูคลิดระหว่างจุดพิกัดใด ๆ ยกกำลังสองจะมีค่าเท่ากับผลรวมของระยะทางยูคลิดในแต่ละมิติยกกำลังสอง หรือกล่าวอีกนัยนึงว่าเป็นการวัดระยะทางเส้นตรง (straight-line distance) ระหว่างสองจุดในปริภูมิ โดยมีนิยามดังสมการที่ (4)

$$d\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(p_i - q_i\right)^2} \tag{4}$$

โดยที่ $d\left(p,q\right)$ คือ ระยะทางระหว่างจุด p และ q n คือ จำนวนมิติ p_i และ q_i คือค่าพิกัดของ p และ q ในมิติที่ i

แต่ยังมีระยะทางอีกแบบที่เป็นที่นิยมเช่นกันคือ ระยะทางแบบแมนฮัตตัน (Manhattan Distance) หรือที่เรียกอีกอย่างหนึ่งว่าระยะทางแบบบล็อกของเมือง (city-block distance) หรือหน่วยวัดระยะของรถแท็กซี่ (taxicab metric) เป็นระยะทางที่คำนวณได้จากการรวมผลต่างค่าสัมบูรณ์ของแต่ละแกนของจุดสองจุด โดยคำนวณเฉพาะในแนวแกนตั้งและแนวนอนเท่านั้น โดยมีนิยามดังสมการที่ (5)

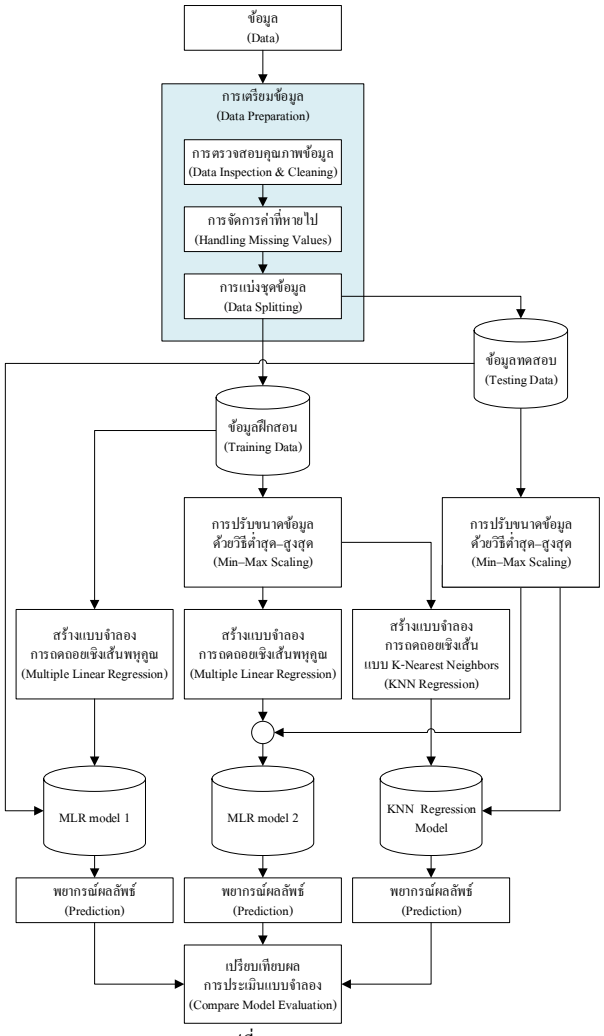
$$d\left(p,q\right) = \sum_{i=1}^{n}\left|q_i - p_i\right| \tag{5}$$



รูปที่ 1 การคำนวณระยะทางแบบ Euclidean และ Manhattan

3. ระเบียบวิธีวิจัย

การศึกษานี้มุ่งเปรียบเทียบประสิทธิภาพของ Multiple Linear Regression เทียบกับ KNN Regression โดยทำการปรับแต่งพารามิเตอร์ของ KNN Regressor ได้แก่ การเลือกค่า K ที่เหมาะสมที่สุด การใช้วิธีการวัดระยะทางที่หลากหลาย รวมถึงการประยุกต์ใช้เทคนิค Inverse Distance Weighting (IDW) เพื่อถ่วงน้ำหนักของเพื่อนบ้านในการเพิ่มความแม่นยำในการทำนาย ขั้นตอนการดำเนินการสามารถสรุปได้ตามรูปที่ 2



รูปที่ 2 ภาพรวมระบบ

3.1 ข้อมูล

ในการศึกษาครั้งนี้ได้เลือกใช้ชุดข้อมูลจำนวน 5 ชุดที่มีความหลากหลาย ทั้งในด้านลักษณะข้อมูลและบริบทการใช้งาน เพื่อทดสอบแบบจำลองการถดถอย โดยชุดข้อมูลทั้ง 5 ชุดมีรายละเอียดดังนี้

1) Auto MPG ชุดข้อมูลนี้รวบรวมจากรถยนต์หลากหลายรุ่น โดยมุ่งเน้นการวิเคราะห์อัตราสิ้นเปลืองเชื้อเพลิง (Miles per Gallon หรือ MPG) ตัวแปรต้นประกอบด้วยน้ำหนักรถยนต์ ความจุเครื่องยนต์ จำนวนลูกสูบ ปีผลิต และต้นกำเนิด ชุดข้อมูลนี้มักใช้ในงานเปรียบเทียบโมเดลทำนายทางกลศาสตร์ยานยนต์

2) Boston Housing ชุดข้อมูลนี้ใช้สำหรับวิเคราะห์ราคาบ้านในเขตเมืองบอสตัน โดยอ้างอิงจากคุณลักษณะต่าง ๆ เช่น อัตราการเกิดอาชญากรรม สัดส่วนพื้นที่ที่อยู่อาศัย และระยะห่างจากศูนย์กลางธุรกิจ ตัวแปรเป้าหมายคือราคาบ้านเฉลี่ย ซึ่งเป็นข้อมูลเชิงต่อเนื่อง เหมาะสมสำหรับการทดลองด้านการพยากรณ์แบบ regression

3) California Housing Prices เป็นชุดข้อมูลที่รวบรวมรายละเอียดของที่พักอาศัยในรัฐแคลิฟอร์เนีย เช่น จำนวนห้องนอน รายได้เฉลี่ยของผู้อยู่อาศัย และจำนวนผู้อยู่อาศัยต่อยูนิต จุดเด่นของชุดข้อมูลนี้คือมีขนาดใหญ่ และสะท้อนลักษณะตลาดอสังหาริมทรัพย์จริง เหมาะสำหรับทดสอบความสามารถของโมเดลในงานที่มีความซับซ้อนเชิงพื้นที่และสังคม

4) Concrete Compressive Strength ชุดข้อมูลนี้รวบรวมจากการทดลองในห้องปฏิบัติการ เพื่อศึกษาความแข็งแรงของคอนกรีต โดยพิจารณาจากองค์ประกอบ เช่น ปริมาณซีเมนต์ น้ำ ทราย สารผสมเพิ่ม และอายุของคอนกรีต (วัน) ตัวแปรเป้าหมายคือค่ากำลังอัดของคอนกรีต ซึ่งสามารถใช้เปรียบเทียบความแม่นยำของโมเดลในงานวัสดุก่อสร้างได้

5) Energy Efficiency ชุดข้อมูลนี้เกี่ยวข้องกับการออกแบบอาคารประหยัดพลังงาน โดยมีตัวแปรต้นอย่างเช่น ความหนาของผนัง ทิศทางอาคาร อัตราส่วนพื้นที่กระจก และประเภทหลังคา ตัวแปรเป้าหมายคือประสิทธิภาพด้านความร้อนและการใช้พลังงานรวม เหมาะกับงานวิเคราะห์ในสาขาวิศวกรรมสิ่งแวดล้อมและพลังงาน

ในการทดลองนี้เลือกใช้เฉพาะตัวแปรเชิงตัวเลข เนื่องจากตัวแปรประเภทข้อความ (string) จำเป็นต้องผ่านกระบวนการแปลงเป็นค่าตัวเลขก่อนจึงจะสามารถนำไปใช้กับแบบจำลองการถดถอยได้ ซึ่งเป็นการเพิ่มความซับซ้อนโดยไม่จำเป็น จากชุดข้อมูลทั้ง 5 ชุดพบว่ามีตัวแปรประเภทข้อความจำนวน 2 ตัวแปร คือ ตัวแปร car model name ที่แสดงชื่อรุ่นรถยนต์ในชุดข้อมูล Auto MPG และตัวแปร ocean_proximity ที่แสดงถึงทำเลที่ตั้งของที่อยู่อาศัยตามระยะห่างจากชายฝั่งทะเลในชุดข้อมูล California Housing Prices ดังนั้นจึงไม่นำตัวแปรทั้งสองมาใช้ในงานวิจัยนี้

3.2 การเตรียมข้อมูล

การเตรียมข้อมูล (Data Preparation) เป็นกระบวนการพื้นฐานสำหรับการวิเคราะห์ข้อมูล ซึ่งช่วยให้ข้อมูลมีความถูกต้องและพร้อมใช้งาน ในงานวิจัยนี้แบ่งกระบวนการเตรียมข้อมูลออกเป็น 3 ส่วน คือ

1) การตรวจสอบคุณภาพข้อมูล (Data Inspection & Cleaning) เป็นการตรวจสอบข้อมูลเบื้องต้นก่อนการนำไปใช้งาน โดยจะมีการตรวจสอบค่าที่หายไป (Missing values) ตรวจหาค่าผิดปกติ (Outliers) แก้ไขค่าซ้ำซ้อน (Duplicates) และตรวจสอบชนิดของตัวแปร (Data types) ให้ถูกต้อง เช่น จำนวนเต็ม ทศนิยม ข้อความ (string) เป็นต้น

2) การจัดการค่าที่หายไป (Handling Missing Values) เมื่อมีการตรวจพบข้อมูลบางแถวมีค่าในบางคอลัมน์ขาดหายไป จะดำเนินการเติมข้อมูลด้วยค่าเฉลี่ยของคอลัมน์ดังกล่าวจากข้อมูลที่มีอยู่ทั้งหมด

3) การแบ่งชุดข้อมูล (Data Splitting) ถูกใช้สำหรับการตรวจสอบความถูกต้องของแบบจำลอง ซึ่งในงานวิจัยนี้ใช้มีการแบ่งข้อมูลออกเป็นชุดฝึกสอน (training set) และชุดทดสอบ (test set) โดยใช้เทคนิคการตรวจสอบไขว้แบบ K ตอน (K-fold Cross Validation) กล่าวคือ ชุดข้อมูลทั้งหมดจะถูกแบ่งออกเป็น 5 ส่วน (folds) เท่า ๆ กัน จากนั้นจะทำการวนซ้ำ 5 รอบ โดยในแต่ละรอบจะเลือกข้อมูล 1 ส่วนเป็นชุดทดสอบ และใช้ส่วนที่เหลืออีก 4 ส่วนเป็นชุดฝึกสอน เมื่อครบทั้ง 5 รอบจะได้ผลการประเมิน 5 ค่า แล้วนำมาหาค่าเฉลี่ยเพื่อวัดประสิทธิภาพของแบบจำลอง

3.3 การปรับขนาดข้อมูลการปรับขนาดข้อมูลด้วยวิธีต่ำสุด–สูงสุด

การปรับขนาดข้อมูลด้วยวิธีต่ำสุด–สูงสุด (Min–Max Scaling) คือเทคนิคการปรับค่าของตัวแปรเชิงตัวเลขให้อยู่ในช่วง [0, 1] โดยสามารถคำนวณได้จากสมการที่ (6)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{6}$$

เมื่อ x คือค่าเดิม และ x' คือค่าที่ถูกปรับขนาดข้อมูลแล้ว วิธีนี้ช่วยลดความแตกต่างของสเกลระหว่างตัวแปร ทำให้ตัวแปรทุกตัวมีน้ำหนักเท่าเทียมกันในกระบวนการวิเคราะห์ เหมาะสำหรับอัลกอริทึมที่อ่อนไหวต่อระยะทางหรือขนาดของข้อมูล เช่น KNN, Neural Networks เป็นต้น เนื่องจากช่วยให้การคำนวณระยะทางหรือการหาค่าความสัมพันธ์มีความถูกต้องและมีประสิทธิภาพมากขึ้น

3.4 การประเมินผล

งานวิจัยนี้ใช้ตัวชี้วัดในการประเมินประสิทธิภาพของแบบจำลองการถดถอยเชิงเส้นพหุคูณและการถดถอยแบบวิธีเพื่อนบ้านใกล้ที่สุด K ตัว จำนวน 4 ค่า ดังนี้

1) ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ (Mean Absolute Error: MAE) เป็นตัวชี้วัดที่ใช้ประเมินความแม่นยำของแบบจำลองโดยคำนวณจากค่าเฉลี่ยของผลต่างเชิงสัมบูรณ์ระหว่างค่าที่แบบจำลองทำนายได้กับค่าจริง ดังสมการที่ (7)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

เมื่อ y_i คือค่าจริง $\hat{y}_i$ คือค่าที่ทำนายได้ และ n คือจำนวนข้อมูลทั้งหมด ค่านี้แสดงถึงความคลาดเคลื่อนเฉลี่ยในหน่วยเดียวกับข้อมูลต้นฉบับ โดยค่าต่ำแสดงว่าแบบจำลองสามารถทำนายได้ใกล้เคียงกับค่าจริง

2) ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE) เป็นตัวชี้วัดที่ใช้ประเมินประสิทธิภาพของแบบจำลอง โดยคำนวณจากค่าเฉลี่ยของผลต่างระหว่างค่าจริงกับค่าที่แบบจำลองทำนายได้ยกกำลังสองก่อนนำมาเฉลี่ย ดังสมการที่ (8)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

เมื่อ y_i คือค่าจริง $\hat{y}_i$ คือค่าที่ทำนายได้ และ n คือจำนวนข้อมูลทั้งหมด ซึ่งค่าที่น้อยแสดงถึงความแม่นยำของแบบจำลองที่มาก ตัวชี้วัดนี้เหมาะในการตรวจจับโมเดลที่มีการทำนายผิดพลาดรุนแรง เพราะให้ความสำคัญกับความผิดพลาดที่มีขนาดใหญ่

3) รากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE) ตัวชี้วัดนี้อ่านค่าได้ง่ายกว่า MSE เนื่องจากค่าความผิดพลาดถูกคำนวณกลับมาในหน่วยเดิมของข้อมูล ซึ่งสามารถคำนวณได้ดังสมการที่ (9)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

เมื่อ y_i คือค่าจริง $\hat{y}_i$ คือค่าที่ทำนายได้ และ n คือจำนวนข้อมูลทั้งหมด ซึ่งค่าที่น้อยแสดงถึงความแม่นยำของแบบจำลองที่มาก เหมาะในการตรวจจับโมเดลที่มีการทำนายผิดพลาดรุนแรงเช่นเดียวกับ MSE

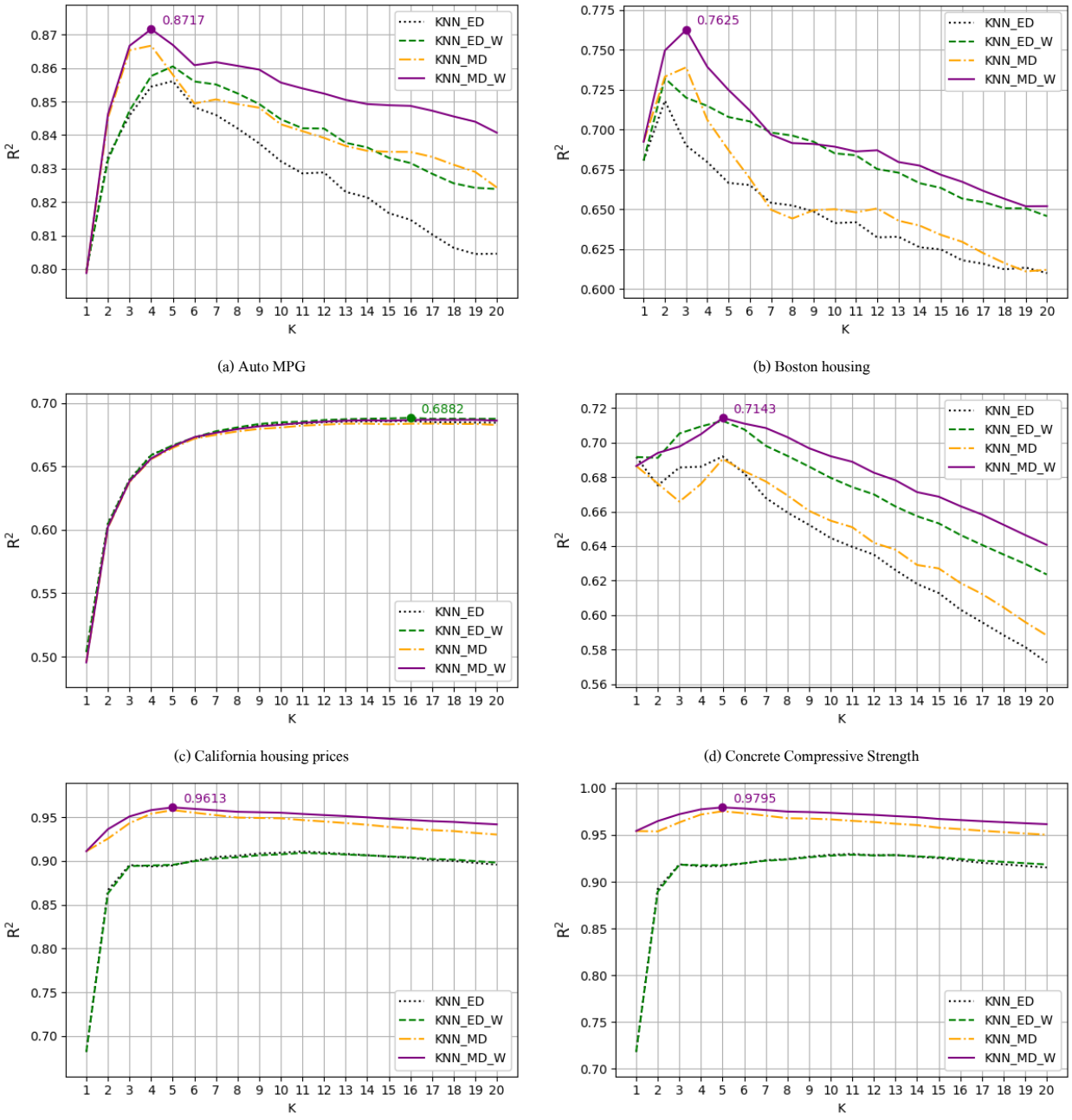
4) ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2) เป็นตัวชี้วัดที่ใช้บ่งบอกว่าแบบจำลองสามารถอธิบายความแปรปรวนของตัวแปรตามได้มากน้อยเพียงใด โดยมีค่าตั้งแต่ $-\infty$ ถึง 1 ซึ่งค่าที่ใกล้ 1 หมายถึงแบบจำลองสามารถอธิบายข้อมูลได้ดี ในขณะที่ค่าที่ใกล้ศูนย์หรือเป็นลบสะท้อนว่าแบบจำลองมีประสิทธิภาพต่ำหรืออาจแย่กว่าการใช้ค่าเฉลี่ยของข้อมูลเป็นตัวทำนาย ซึ่งสามารถคำนวณได้จากสมการที่ (10)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

เมื่อ y_i คือค่าจริง $\hat{y}_i$ คือค่าที่ทำนายได้ และ n คือจำนวนข้อมูลทั้งหมด ซึ่งค่าที่น้อยแสดงถึงความแม่นยำของแบบจำลองที่มาก ตัวชี้วัดนี้เหมาะในการตรวจจับโมเดลที่มีการทำนายผิดพลาดรุนแรง

4. ผลการทดลองและอภิปรายผล

ในส่วนนี้ ผลลัพธ์ที่ได้จากแบบจำลองการถดถอยระหว่าง Multiple Linear Regression เทียบกับ KNN Regression โดยที่ KNN Regression มีการปรับค่าพารามิเตอร์ 3 แบบ ดังนี้ การเลือกค่า K (1 ถึง 20) การคำนวณระยะทาง (Euclidean และ Manhattan) และการถ่วงน้ำหนักเพื่อนบ้าน (Inverse Distance Weighting) จะถูกนำมาอภิปรายโดยพิจารณาจากประสิทธิภาพการทำนาย ประสิทธิภาพของแบบจำลองได้รับการประเมินโดยพิจารณาจากคะแนน MAE, MSE, RMSE และ R^2 ของแต่ละแบบจำลองและแต่ละชุดข้อมูล



(a) Auto MPG

(b) Boston housing

(c) California housing prices

(d) Concrete Compressive Strength

(e) Energy Efficiency (Cooling Load)

(f) Energy Efficiency (Heating Load)

รูปที่ 3 ค่า R^2 ของการปรับค่าพารามิเตอร์ KNN Regression ในการทำนายของแต่ละชุดข้อมูล

ตารางที่ 1 ตัวชี้วัดประสิทธิภาพของแบบจำลองการถดถอยของแต่ละชุดข้อมูล

Dataset	MLR				MLR (min-max)				KNN				KNN parameter		
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2	K	distance	weight
Auto MPG	2.569	11.550	3.391	0.8030	2.6503	11.806	3.428	0.7990	1.9721	7.372	2.701	0.8717	4	Manhattan	IDW
Boston	3.178	20.985	4.505	0.7461	3.6177	27.271	5.162	0.6664	2.9033	19.584	4.340	0.7625	3	Manhattan	IDW
California	50,832.41	4,851M	69,621.288	0.6358	69,415.33	8,897M	93,515.503	0.3324	43,579.14	4,155M	64,425.66	0.6882	16	Euclidean	IDW
Concrete	8.326	109.759	10.469	0.5998	8.367	112.219	10.583	0.5897	6.7858	78.703	8.859	0.7143	5	Manhattan	IDW
Energy (Cooling)	2.2702	10.3827	3.2044	0.8859	2.2702	10.3827	3.2044	0.8859	1.2743	3.5618	1.8618	0.9613	5	Manhattan	IDW
Energy (Heating)	2.0931	8.7028	2.9452	0.9140	2.0931	8.7028	2.9452	0.9140	0.9636	2.0712	1.4359	0.9795	5	Manhattan	IDW

* หมายเหตุ : IDW คือการถ่วงน้ำหนักแบบ Inverse Distance Weighting

การเปรียบเทียบแบบจำลองการถดถอยระหว่าง Multiple Linear Regression เทียบกับ KNN Regression ที่มีการปรับค่าพารามิเตอร์แบบต่าง ๆ มีทั้งหมด 6 แบบดังนี้

1) MLR คือการทำนายด้วย Multiple Linear Regression โดยใช้ข้อมูลที่ไม่มีการปรับช่วงค่าข้อมูล

2) MLR_MM คือการทำนายด้วย Multiple Linear Regression โดยใช้ข้อมูลที่มีการปรับช่วงค่าข้อมูลด้วยวิธีต่ำสุด–สูงสุด

3) KNN_ED คือการทำนายด้วย KNN Regression โดยใช้วิธีการคำนวณระยะห่างแบบ Euclidean แต่ไม่มีการถ่วงน้ำหนัก

4) KNN_ED_W คือการทำนายด้วย KNN Regression โดยใช้วิธีการคำนวณระยะห่างแบบ Euclidean ที่มีการถ่วงน้ำหนักระยะห่างแบบผกผันระยะทาง (IDW)

5) KNN_MD คือการทำนายด้วย KNN Regression โดยใช้วิธีการคำนวณระยะห่างแบบ Manhattan แต่ไม่มีการถ่วงน้ำหนัก

6) KNN_MD_W คือการทำนายด้วย KNN Regression โดยใช้วิธีการคำนวณระยะห่างแบบ Manhattan ที่มีการถ่วงน้ำหนักแบบผกผันระยะทาง (IDW)

โดยการทำนายด้วย KNN Regression ทุกวิธีจะมีการปรับค่าพารามิเตอร์จำนวนเพื่อนบ้าน (K) ตั้งแต่ 1 ถึง 20 และข้อมูลที่ใช้ใน KNN Regression จะถูกปรับช่วงค่าข้อมูลด้วยวิธีต่ำสุด-สูงสุดก่อนที่นำไปใช้งาน เนื่องจาก KNN Regression มีการคำนวณระยะทางในการค้นหาเพื่อนบ้านที่ใกล้ที่สุดเพื่อนำมาทำนายผลลัพธ์ ดังนั้นความแตกต่างของช่วงค่า (scale) ของแต่ละตัวแปรย่อมส่งผลโดยตรงต่อการคำนวณหาเพื่อนบ้านและผลการทำนายด้วย

รูปที่ 2 แสดงผลการเปรียบเทียบค่าสัมประสิทธิ์การตัดสินใจ (R^2) จากการปรับค่าพารามิเตอร์ KNN regression จำนวน 3 ค่า คือ การเลือกจำนวนเพื่อนบ้านที่ใกล้ที่สุด (K) วิธีการวัดระยะห่างแบบ Euclidean และ Manhattan และการถ่วงน้ำหนักแบบผกผันระยะทางการคำนวณระยะทาง (IDW) โดยที่ชุดข้อมูล Auto-MPG, Boston Housing, Concrete Compressive Strength และ Energy Efficiency จะมีค่าสัมประสิทธิ์การตัดสินใจดีที่สุด เมื่อใช้วิธีการคำนวณระยะห่างแบบ Manhattan ที่มีการถ่วงน้ำหนักระยะห่างแบบผกผันระยะทาง และ K มีค่าอยู่ระหว่าง 3 ถึง 5 แต่ถ้ามีการเพิ่มค่า K ขึ้นไปมากกว่านี้ประสิทธิภาพจะมีแนวโน้มลดลง ส่วนชุดข้อมูล California Housing Prices จะมีค่าสัมประสิทธิ์การตัดสินใจดีที่สุด เมื่อใช้วิธีการคำนวณระยะห่างแบบ Euclidean ที่มีการถ่วงน้ำหนักระยะห่างแบบผกผันระยะทาง และใช้ K = 16 แต่เมื่อพิจารณาโดยละเอียดจะพบว่าค่าสัมประสิทธิ์การตัดสินใจจะเริ่งคงที่เมื่อ $K \geq 10$

ดังนั้นเมื่อได้ค่าของพารามิเตอร์ของ KNN Regression จากผลลัพธ์ในรูปที่ 2 แล้ว จึงสามารถนำผลการทำนายที่ได้มาเปรียบเทียบกับ Multiple Linear Regression ทั้งแบบที่ไม่มีการปรับช่วงค่าข้อมูลและแบบมีการปรับช่วงค่าข้อมูลด้วยด้วยวิธีต่ำสุด–สูงสุด ได้ดังตารางที่ 1 ซึ่งผลลัพธ์พบว่า KNN Regression ให้ผลการทำนายดีกว่า Multiple Linear Regression ทั้งแบบที่ไม่มีการปรับช่วงค่าข้อมูล (MLR) และแบบที่มีการปรับช่วงค่าข้อมูลด้วยด้วยวิธีต่ำสุด-สูงสุด (MLR min-max) ในทุกชุดข้อมูล โดยพิจารณาจากค่า MAE, MSE และ RMSE ที่มีค่าต่ำกว่า และค่า R^2 ที่สูงกว่า เมื่อพิจารณาในชุดข้อมูล Auto MPG และ Boston Housing จะพบว่าค่า R^2 ของ KNN Regression มีค่าอยู่ที่ 0.8717 และ 0.7625 ตามลำดับ ซึ่งสูงกว่า Multiple Linear Regression ทั้งสองแบบ ส่วนชุดข้อมูล California Housing Prices จะพบว่าแบบจำลองการถดถอยทุกแบบมีความคลาดเคลื่อนสูงเนื่องมาจากตัวแปรมีช่วงค่าข้อมูลที่กว้าง แต่ KNN Regression ยังคงให้ค่า R^2 ที่ดีกว่า (R^2=0.6882) เมื่อเทียบกับ Multiple Linear Regression แบบที่มีการปรับช่วงค่าข้อมูลด้วยด้วยวิธีต่ำสุด-สูงสุด (MLR min-max) ที่ให้ค่า R^2 ลดลงเหลือเพียง 0.3324 ซึ่งผลดังกล่าวนี้สะท้อนให้เห็นว่าการปรับช่วงค่าข้อมูลไม่ได้ส่งผลให้การทำนายมีความแม่นยำเพิ่มมากขึ้น แต่อาจจะทำให้ผลลัพธ์แย่ลงมากกว่าเดิม สำหรับผลลัพธ์ของชุดข้อมูล Concrete Compressive Strength ยังคงสอดคล้องกับชุดข้อมูล Auto MPG และ Boston Housing โดย KNN Regression ให้ค่า R^2 เท่ากับ 0.7143 ซึ่งสูงกว่า Multiple Linear Regression ทุกแบบ ส่วนชุดข้อมูล Energy Efficiency ทั้งแบบที่ทำนาย Cooling load และ Heating load นั้น KNN Regression ให้ค่า R^2 มากกว่าในทุกชุดข้อมูล โดยให้ค่า R^2 เท่ากับ 0.9613 และ 0.9795 ตามลำดับ

เมื่อพิจารณาพารามิเตอร์ของ KNN Regression ในตารางที่ 1 จะพบว่าการใช้ค่าถ่วงน้ำหนักแบบผกผันระยะทาง (IDW) ให้ประสิทธิภาพดีกว่าในทุกชุดข้อมูล โดยที่การคำนวณระยะทางส่วนใหญ่จะคำนวณด้วยวิธี Manhattan ยกเว้นชุดข้อมูล California housing prices ที่ใช้วิธีการคำนวณระยะทางแบบ Euclidean เมื่อพิจารณาที่จำนวนเพื่อนบ้าน (K) ส่วนใหญ่จะใช้ค่าอยู่ระหว่าง 3 ถึง 5 ยกเว้น California housing prices ที่ใช้ค่า K เท่ากับ 16

โดยผลการทดลองทั้งหมดสามารถสรุปได้ว่า KNN Regression ที่ใช้เทคนิคการถ่วงน้ำหนักแบบผกผันระยะทาง (IDW) และเลือกค่าพารามิเตอร์ที่เหมาะสม ให้ผลการทำนายได้ดีกว่า Multiple Linear Regression ทั้งในกรณีที่ไม่ปรับช่วงค่าข้อมูลและแบบที่มีการปรับช่วงค่าข้อมูลด้วยวิธีต่ำสุด-สูงสุด

5. สรุป

จากการทดลองเปรียบเทียบประสิทธิภาพของแบบจำลอง Multiple Linear Regression ทั้งแบบที่มีและไม่มีการปรับช่วงค่าข้อมูลกับแบบจำลอง KNN Regression ในหลายชุดข้อมูล พบว่าแบบจำลอง KNN Regression ที่ใช้เทคนิคการถ่วงน้ำหนักแบบผกผันระยะทาง (IDW) และการเลือกจำนวนเพื่อนบ้าน (K) ให้ผลลัพธ์ที่ดีกว่าแบบจำลอง Multiple Linear Regression ทั้งในแง่ของค่าความ

คลาดเคลื่อน (MAE, MSE, RMSE) ที่ต่ำกว่า และค่า R^2 ที่สูงกว่า โดยเฉพาะในชุดข้อมูล Auto-MPG, Boston Housing, Concrete Compressive Strength และ Energy Efficiency แบบจำลอง KNN Regression แสดงประสิทธิภาพการทำนายที่ดีกว่าอย่างชัดเจน แสดงให้เห็นถึงความยืดหยุ่นและศักยภาพของแบบจำลอง KNN Regression ในการจัดการข้อมูลที่มีความซับซ้อน อย่างไรก็ตาม ผลการทดลองยังสะท้อนให้เห็นข้อจำกัดของแบบจำลอง Multiple Linear Regression ที่ไม่สามารถจัดการข้อมูลที่มีความไม่เชิงเส้นได้ดีนัก แม้จะมีการปรับช่วงค่าข้อมูลด้วยวิธีต่ำสุด-สูงสุด ซึ่งสามารถสรุปได้ว่าแบบจำลอง KNN Regression เป็นแบบจำลองที่มีความเหมาะสมและมีประสิทธิภาพสูงในการพยากรณ์ข้อมูลที่มีความซับซ้อนหลากหลาย โดยเฉพาะเมื่อใช้ร่วมกับการน้ำหนักแบบผกผันระยะทาง (IDW) และการเลือกจำนวนเพื่อนบ้าน (K) ที่เหมาะสม

แนวทางในการวิจัยต่อไปคือการเปรียบเทียบ KNN regression กับโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) บนปัญหาที่มีความไม่เป็นเชิงเส้นสูง (Highly nonlinear) เพื่อประเมินแนวโน้มในการจัดการปัญหา โดยใช้กระบวนการเตรียมข้อมูลแบบเดียวกัน แบ่งชุดข้อมูลแบบ k-fold cross-validation และปรับค่าพารามิเตอร์ต่าง ๆ อย่างเป็นระบบ จากนั้นประเมินประสิทธิภาพจากค่าตัวชี้วัด คือ MAE, MSE, RMSE และ R^2 รวมทั้งการพิจารณาในด้านเวลาการเรียนรู้ เวลาในการทำนาย และการใช้หน่วยความจำ ซึ่งโดยทั่วไปแล้วโครงข่ายปราสาทเทียมอาจจะให้ผลลัพธ์ที่ดีกว่า KNN Regression แต่เนื่องจาก KNN Regression มีรูปแบบการทำงานที่ซับซ้อนน้อยกว่า กล่าวคือ KNN Regression ไม่จำเป็นที่จะต้องฝึกฝนโมเดล (train model) ใหม่เมื่อมีข้อมูลใหม่เข้ามา ซึ่งแตกต่างกับโครงข่ายปราสาทเทียม ดังนั้นจึงทำให้ KNN Regression ยังคงมีความน่าสนใจที่ถูกนำไปใช้งานได้เหมาะสมมากกว่าในบางสถานการณ์

เอกสารอ้างอิง

[1] U. Grömping, "Variable importance in regression models," *Wiley Interdisciplinary Reviews: Computational Statistics.*, vol. 7, no. 2, pp. 137-152, 2015.

[2] A. Zapf, C. Wiessner, and I. R. König, "Regression analyses and their particularities in observational studies: Part 32 of a series on evaluation of scientific publications," *Deutsches Ärzteblatt International.*, vol. 121, no. 4, pp. 128-134, 2024.

[3] A. Khazaee Poul, M. Shourian, and H. Ebrahimi, "A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly stream flow prediction," *Water Resources Management*., vol. 33, no. 8, pp. 2907-2923, 2019.

[4] D. N. Cosenza, L. Korhonen, M. Maltamo, P. Packalen, J. L. Strunk, E. Næsset, T. Gobakken, P. Soares, and M. Tomé, "Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock," *Forestry: An International Journal of Forest Research.*, vol. 94, no. 2, pp. 311-323, 2021.

[5] H. A. Abu Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. E. Salman, and V. S. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review," *Big Data.*, vol. 7, no. 4, pp. 221-248, 2019.

[6] A. İ. Çetin and A. H. Büyüklü, "A new approach to K-nearest neighbors distance metrics on sovereign country credit rating," *Kuwait Journal of Science.*, vol. 52, no. 1, Article 100324, 2025.

[7] M. M. Kumbure and P. Luukka, "A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance," *Granular Computing.*, vol. 7, no. 3, pp. 657-671, 2022.

[8] G. Zhao, Z. Li, and M. Yang, "Comparison of twelve machine learning regression methods for spatial decomposition of demographic data using multisource geospatial data: An experiment in Guangzhou City, China," *Applied Sciences.,* vol. 11, no. 20, Article 9424, 2021.

[9] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports.*, vol. 12, no. 1, Article 6256, 2022.

[10] M. J. Hosen and I. Ahmed, "Comparison of Regression, K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) Models for the Prediction of Weight, Gender and Body Mass Index Status," *International Journal of Computer Applications.*, vol. 185, no. 29, Article 8887, 2023.

[11] Z. N. Liu, X. Y. Yu, L. F. Jia, Y. S. Wang, Y. C. Song, and H. D. Meng, "The influence of distance weight on the inverse distance weighted method for ore-grade estimation," *Scientific Reports.*, vol. 11, no. 1, Article 2689, 2021.

[12] P. Srisuradetchai and K. Suksrikran, "Random kernel k-nearest neighbors regression," *Frontiers in Big Data.*, vol. 7, Article 1402384, 2024.

[13] A. M. Raheem, I. J. Naser, M. O. Ibrahim, and N. Q. Omar, "Inverse distance weighted (IDW) and kriging approaches integrated with linear single and multi-regression models to assess particular physico-consolidation soil properties for Kirkuk city," *Modeling Earth Systems and Environment.*, vol. 9, no. 4, pp. 3999-4021, 2023.