

การคาดการณ์ราคาเที่ยวบิน

Prediction of Flight Price

อธิศ ฟองไข่มุก¹ และ ภาณุวัฒน์ เมฆะ^{1,*}

Athit Fongkhaimuk^{1*} and Panuwat Mekha^{1,*}

¹สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้ เชียงใหม่ 50290 ประเทศไทย

¹Department of Computer Science, Faculty of Science, Maejo University, Chiang Mai, 50290, Thailand

* ผู้นิพนธ์ประสานงาน: ผู้ช่วยศาสตราจารย์ภาณุวัฒน์ เมฆะ อีเมล: Panuwat_m@mju.ac.th

บทคัดย่อ:

การคาดการณ์ราคาเที่ยวบินเป็นกระบวนการที่สำคัญในการช่วยให้ผู้โดยสารสามารถวางแผนการเดินทางได้อย่างมีประสิทธิภาพ และประหยัดค่าใช้จ่ายมากขึ้น ในการวิจัยนี้ ใช้ชุดข้อมูลเกี่ยวกับราคาเที่ยวบินซึ่งประกอบไปด้วยปัจจัยต่างๆ เช่น สายการบิน, ต้นทาง และปลายทาง, จำนวนการหยุดพักเที่ยวบิน, วันเวลา, และระยะเวลาในการเดินทาง เพื่อพัฒนาระบบคาดการณ์ราคาตั๋วเครื่องบินผ่านโปรแกรม ออเรนจ์ ดาต้าไมนิ่ง ในการประมวลผลข้อมูลและทดลองกับอัลกอริธึมที่ต่างๆ ได้แก่ Decision Tree, k-Nearest Neighbor (kNN), Random Forest, และ Neural Network รวมถึงดำเนินการเปรียบเทียบความสามารถในการคาดการณ์ราคาตั๋วเครื่องบิน จากการวิจัยพบว่าแต่ละอัลกอริธึมมีความสามารถในการคาดการณ์ที่แตกต่างกัน โดย Random Forest และ Ada Boost มีความแม่นยำสูงสุดในการทำนายราคาตั๋วเครื่องบิน ข้อมูลเหล่านี้สามารถช่วยผู้โดยสารในการตัดสินใจเลือกจองตั๋วในช่วงเวลาที่เหมาะสมที่สุดและประหยัดค่าใช้จ่ายได้อย่างมีประสิทธิภาพ

คำสำคัญ: การคาดการณ์ ; ราคาเที่ยวบิน ; การเรียนรู้ด้วยเครื่องจักร

Abstract

Flight price prediction is an important process that helps passengers plan their trips more efficiently and save on costs. In this research, a dataset on flight prices, which includes factors such as airlines, departure and destination points, number of stops, date and time, and travel duration, is used to develop a flight ticket price prediction system. The data processing and algorithm testing were conducted through the Orange Data Mining

program, using various algorithms including Decision Tree, k-Nearest Neighbor (kNN), Random Forest, and Neural Network. The predictive performance of these algorithms was compared, with the research showing that each algorithm had different levels of accuracy. Random Forest and Ada Boost were found to have the highest accuracy in predicting flight prices. This information can assist passengers in deciding the best time to book tickets, thus allowing them to save costs more effectively.

Keywords: prediction; flight prices; machine learning

บทนำ

ในยุคปัจจุบันที่เทคโนโลยีสารสนเทศก้าวหน้าไปอย่างรวดเร็ว การเดินทางด้วยเครื่องบินกลายเป็นส่วนหนึ่งของชีวิตประจำวันของผู้คนมากขึ้น ความสะดวกสบายในการเดินทางและการเข้าถึงข้อมูลเกี่ยวกับการบินทำให้เราสามารถเลือกเที่ยวบินที่ตรงตามความต้องการได้อย่างง่ายดาย การค้นหาตั๋วเครื่องบินที่มีราคาที่ดีที่สุดยังคงเป็นความท้าทายสำหรับผู้โดยสารเนื่องจากราคาเที่ยวบินมีความผันผวนและอาจจะเปลี่ยนแปลงได้ตลอดเวลา

การคาดการณ์ราคาเที่ยวบินจึงเป็นกระบวนการที่สำคัญในการช่วยให้ผู้โดยสารสามารถวางแผนการเดินทางได้อย่างมีประสิทธิภาพ และประหยัดค่าใช้จ่ายได้มากขึ้น โดยการคาดการณ์ราคาเที่ยวบินคือการใช้ข้อมูลจากอดีตและปัจจัยต่างๆ เช่น สายการบิน, วันเวลา, สถานที่ต้นทางและปลายทาง, จำนวนการหยุดพักเที่ยวบิน, เวลาในการเดินทาง และปัจจัยอื่นๆ เพื่อคาดการณ์ราคาที่คาดว่าจะเกิดขึ้นในอนาคต

โดยการคาดการณ์ราคาเที่ยวบินในครั้งนี้ใช้ โปรแกรม ออเรนจ์ ดาต้าไมนิ่ง ที่สามารถประมวลผลข้อมูลและค้นหาความสัมพันธ์ที่ซับซ้อนด้วยการพัฒนาของเทคโนโลยีและข้อมูลที่มีอยู่มากมาย การคาดการณ์ราคาเที่ยวบินได้กลายเป็นเครื่องมือที่มีประโยชน์ในการช่วยผู้โดยสารในการตัดสินใจเลือกจองตั๋วในเวลาที่เหมาะสมที่สุด

ดังนั้นคณะผู้วิจัยจึงต้องการสำรวจถึงวิธีการและเครื่องมือที่ใช้ในการคาดการณ์ราคาเที่ยวบิน รวมถึงข้อดีและข้อจำกัดของแต่ละวิธีการ เพื่อให้ผู้อ่านเข้าใจถึงความสำคัญของการคาดการณ์ราคาเที่ยวบินและวิธีการที่สามารถช่วยให้การเดินทาง เป็นไปได้อย่างราบรื่นและคุ้มค่ามากยิ่งขึ้น

2. วิธีการดำเนินงานวิจัย (Materials & Methods)

2.1 ชุดข้อมูล (Dataset)

ชุดข้อมูลประกอบไปด้วย สายการบิน(Airlines), ต้นทางและปลายทาง(Source and Destination), จำนวนการหยุดพัก (Total Stops), ราคา(Price), วัน เดือน และปี(Date, Month, and Year), เวลาออกเดินทาง และ เวลาถึงปลายทาง(Departure and Arrival Times), ระยะเวลาในการบิน(Duration) โดยชุดข้อมูลที่ใช้สำหรับการวิจัยนี้สามารถดาวน์โหลดได้จากลิงก์ <https://www.kaggle.com/datasets/viveksharmar/flight-price-data>

สายการบิน	จำนวนชุดข้อมูล
Air Asia	319
IndiGo	2053
Air India	1752
GoAir	194
Jet Airways	3849
Jet Airways Business	6
Multiple carriers	1196
รวม	10684

2.2 อัลกอริทึม ที่ใช้ทำนายราคา

2.2.1 **Decision Tree** เป็นอัลกอริทึมในการเรียนรู้ข้อมูลที่ใช้โครงสร้างของต้นไม้เพื่อตัดสินใจ จำแนกข้อมูล โดยจะสร้างโหนดสำหรับแต่ละขั้นตอนของการแบ่งข้อมูล ซึ่งโหนดนี้จะมีเงื่อนไขที่เกี่ยวข้องกับคุณสมบัติและค่าที่ต้องการจำแนก การแบ่งข้อมูลจะทำในทุกลโหนดจนกว่าจะได้คลาสจำแนกที่ถูกต้องที่สุด

อ้างอิง <https://www.bornodev.com/2022/09/15/รู้จักกับ-decision-tree>

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad D = - \sum_{k=1}^K \hat{p}_{mk} \log \hat{p}_{mk}$$

Gini Impurity(ค่าความไม่บริสุทธิ์)

Entropy(ค่าความไม่แน่นอน)

(1)

โดยที่ $\hat{p}$ คือ ผลรวมของค่าความน่าจะเป็นของเหตุการณ์ที่เราสนใจ

2.2.2 **k-Nearest Neighbor (kNN)** เป็นอัลกอริทึมในการจำแนกข้อมูล โดยจะค้นหา k ตัวอย่าง ที่ใกล้ที่สุดจากชุดข้อมูล เพื่อหาความคล้ายระหว่างตัวอย่างใหม่และตัวอย่างที่ใกล้ที่สุด จากนั้นจะใช้การ โหวต (voting) โดยนับจำนวนตัวอย่างในแต่ละคลาสที่อยู่ใน k ตัวอย่างที่ใกล้ที่สุด และคลาสที่มีจำนวน ตัวอย่างมากที่สุดจะกำหนดเป็นคลาสของตัวอย่างใหม่

อ้างอิง <https://kongruksiam.medium.com/สรุป-machine-learning-ep-4-เพื่อนบ้านใกล้ที่สุด-k-nearest-neighbors-787665f7c09d>

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Distance Function

(2)

โดยที่ p และ q แทน ชุดของข้อมูล ที่ต้องการวัดระยะห่างระหว่างกัน

2.2.3 Random Forest เป็นอัลกอริทึมที่ใช้การเรียนรู้จากคำตอบหลายๆ ต้นไม้ตัดสินใจ(Decision Trees) โดยจะสร้างต้นไม้ตัดสินใจแบบสุ่มคุณสมบัติ (features) จากชุดข้อมูลตัวอย่าง เพื่อใช้ในการจำแนกหรือทำนายข้อมูล

โดยที่แต่ละต้นไม้ตัดสินใจใช้สมการแบบเดียวกับ 2.2.1 Decision Tree แต่ Random Forest จะรวมผลลัพธ์จากหลายต้นเพื่อให้ได้การทำนายที่เสถียรและแม่นยำมากขึ้น

2.2.4 Neural Network เป็นระบบที่จำลองการทำงานของเซลล์ประสาทในสมองมนุษย์ โดยใช้อัลกอริทึมทางคณิตศาสตร์ประกอบด้วยหลายชั้น (Layers) โดยแต่ละชั้นประกอบด้วยโหนด (Nodes) หรือเซลล์ประสาทเทียม ที่มีความสัมพันธ์ทางคณิตศาสตร์ระหว่างกัน

อ้างอิง <https://medium.com/@dopplerz/bayesian-neural-network-ตอนที่-3-อะไรคือ-deep-learning-และอะไรคือ-neural-network ฉบับมือใหม่-cce9227ab599>

$$\text{predicted_Y} = w * X + \text{bias}$$

(3)

โดยที่ bias คือตัวเลขที่บวกเข้าไปเพื่อปรับให้ค่าที่คำนวณออกมาถูกต้องมากขึ้น โดยค่านี้จะไม่ขึ้นอยู่กับ X ใดๆ

2.2.5 Gradient Boosting เป็นเทคนิคการเรียนรู้แบบยกระดับ (boosting) ซึ่งจะสร้างชุดของโมเดลที่อ่อนแอ (weak models) เช่น ต้นไม้ตัดสินใจแบบตื้น (shallow decision trees) โดยการฝึกแต่ละโมเดลตามข้อผิดพลาดที่เกิดจากโมเดลก่อนหน้า โดยรวมโมเดลเหล่านี้เพื่อการทำนายที่แม่นยำยิ่งขึ้น เทคนิคนี้จะพยายามลดค่าความผิดพลาดของโมเดลแต่ละตัวในแต่ละรอบการฝึก

อ้างอิง <https://guopai.github.io/ml-blog11.html>

$$h(x) = h_1(x) + h_2(x) + \dots + h_J(x)$$

(4)

โดย J คือจำนวน Instance ทั้งหมด

2.2.6 AdaBoost (Adaptive Boosting) เป็นเทคนิคการเรียนรู้แบบยกระดับที่สร้างชุดของโมเดลอ่อนแอ โดยจะปรับน้ำหนักของตัวอย่างข้อมูลที่ถูกจำแนกผิดในแต่ละรอบการฝึก โมเดลที่สร้างในแต่ละรอบจะมีความสำคัญมากขึ้นหากสามารถปรับปรุง

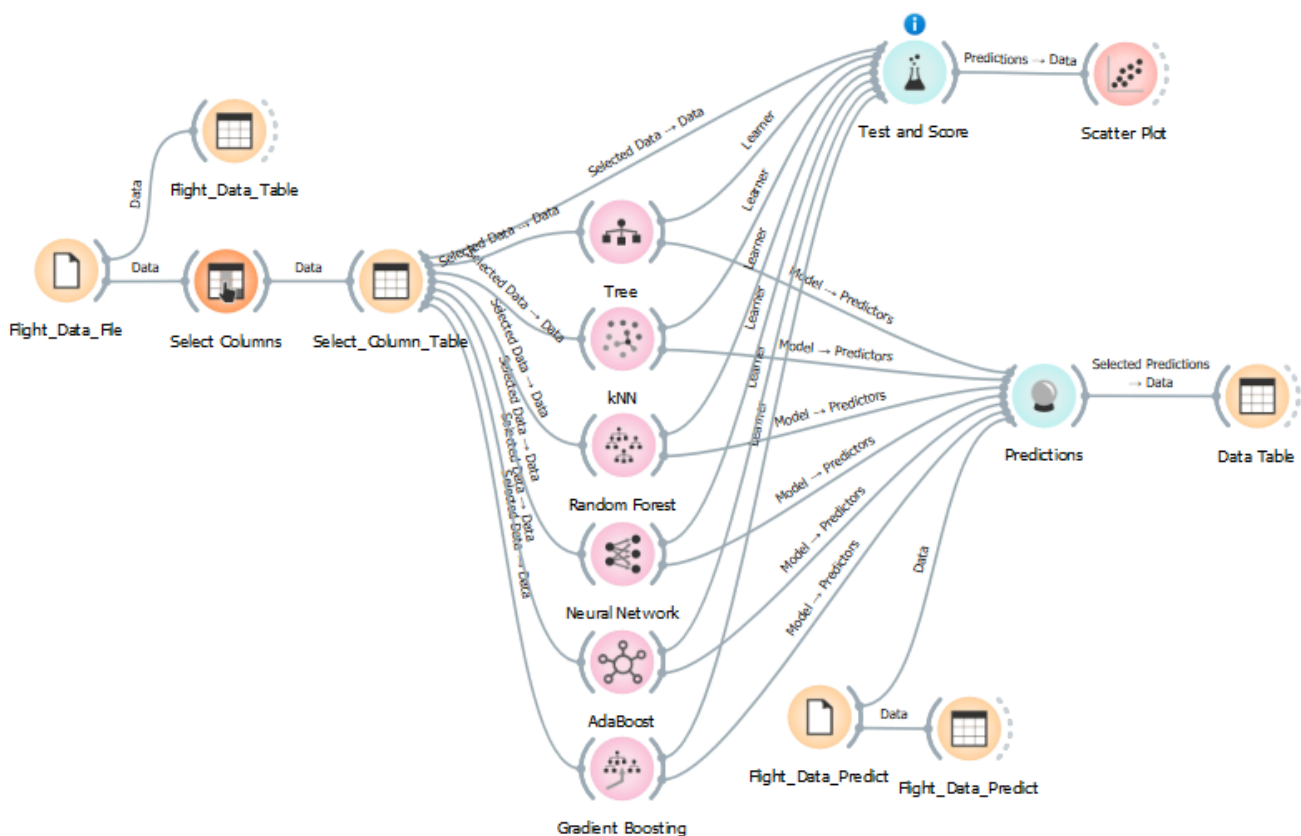
ข้อผิดพลาดของโมเดลก่อนหน้าได้อย่างมีประสิทธิภาพ โมเดลสุดท้ายจะเป็นการรวมผลลัพธ์ของโมเดลแต่ละตัวในชุด

อ้างอิง <https://kongruksiam.medium.com/สรุป-machine-learning-ep-4-เพื่อนบ้านใกล้ที่สุด-k-nearest-neighbors-787665f7c09d>

$$f(x) = \sum_{t=1}^T \alpha_t h_t(x)$$

โดยที่ $h_t(x)$ คือ ผลลัพธ์ของ weak classifier ส่วน Alpha คือ weight ที่กำหนดให้กับตัว Classifier แต่ละรอบที่มีการคำนวณ

2.3 ขั้นตอนดำเนินการวิจัย



รูปที่ 1 ขั้นตอนการการเปรียบเทียบประสิทธิภาพของอัลกอริทึมทำนายราคา

จากรูปที่ 1 ขั้นตอนการเปรียบเทียบประสิทธิภาพของอัลกอริทึมทำนายราคาตัวเครื่องบิน ทดลองนี้ผ่านโปรแกรมออร์เจเนดา ต้าไมนิ่ง โดยเริ่มจากนำ Dataset มา Select Column และ กำหนด ชื่อ Column (target) ที่เราต้องการในโดยการทดลองนี้ได้

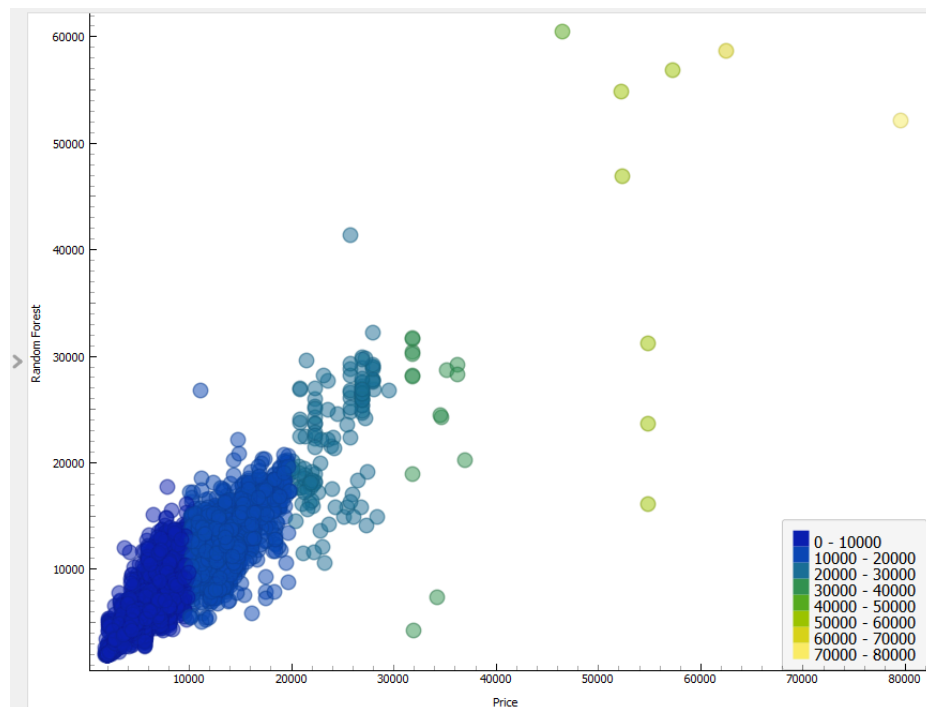
กำหนด Target ให้กับ Column ที่ชื่อว่า price และ เลือก ignored-> Arrive_min, Arrive_hours, Duration_min, Duration_hours เนื่องจาก เวลาที่ถึงที่หมายและ ระยะเวลาในการเดินทางไม่มีผลต่อการทำนายราคาตัวจากนั้นจึงนำ Dataset อีกชุดหนึ่งที่ต้องการนำมาทำนายราคาตัวต่อเข้ากับ Predictions (Widget) และดำเนินการใช้อัลกอริทึมเรียนรู้ด้วยเครื่องจักรแบบต่างๆ เพื่อทำนายราคาตัว และนำผลลัพธ์ที่ได้มาแสดงผล

3. ผลการทดลองและวิจารณ์ผลการทดลอง:

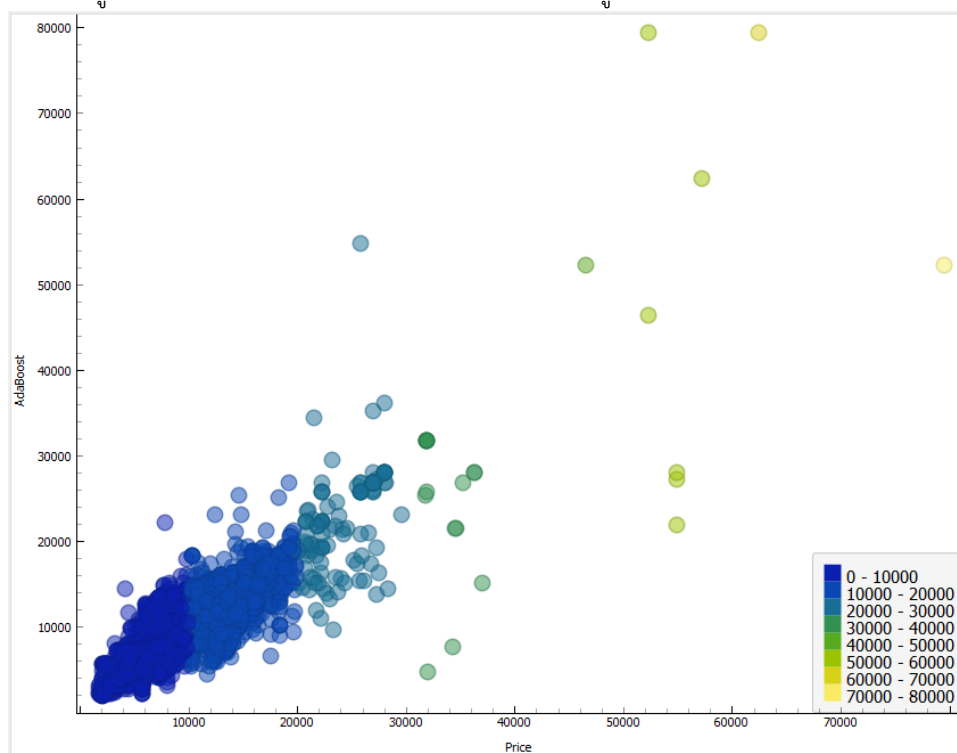
ตารางที่ 2 การเปรียบเทียบประสิทธิภาพการทำนายราคาของอัลกอริทึมแบบต่างๆ

Model	MSE (ล้าน)	RMSE	MAE	R ² Score (%)	Explained Variance Score (%)
Neural Network	8.35	2,890.40	2,001.57	24.14	60.71
Gradient Boosting	5.00	2,236.37	1,539.24	17.99	76.48
AdaBoost	4.13	2,032.11	1,175.07	12.99	80.58
Random Forest	3.67	1,916.35	1,144.13	12.61	82.73
kNN	9.99	3,159.96	1,994.13	22.58	53.04
Tree	4.56	2,135.45	1,246.29	13.53	78.55

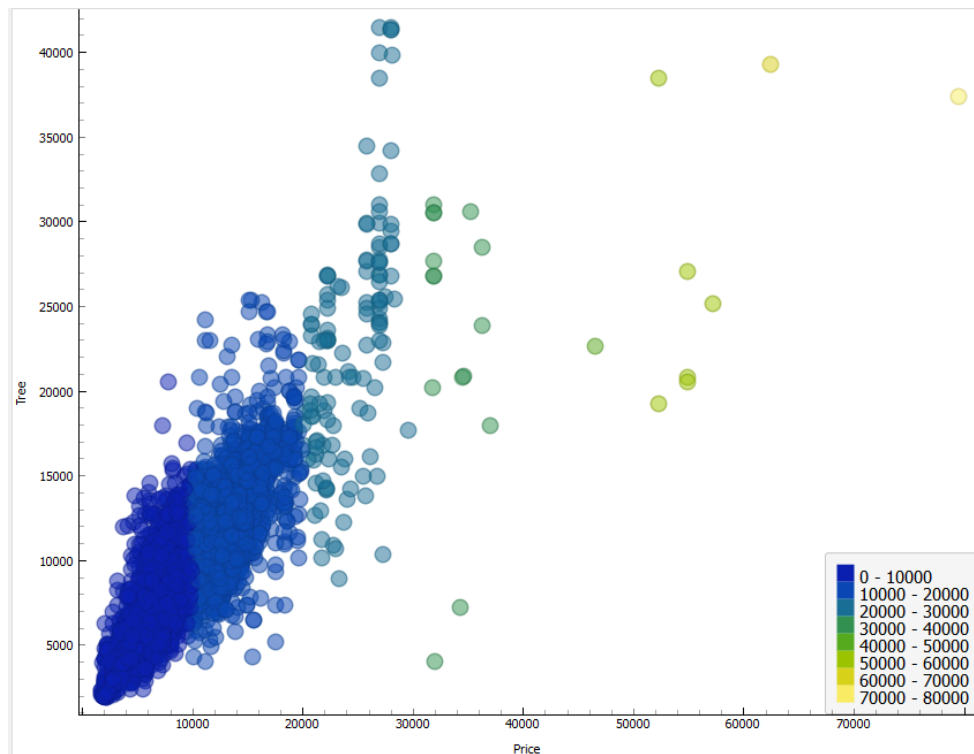
จากตารางที่ 2 AdaBoost และ Random Forest เป็นอัลกอริทึมที่มีประสิทธิภาพดีที่สุดในการทำนายราคาตามข้อมูลที่น่ามาใช้ โดยที่ Neural Network และ kNN มี ผลลัพธ์ที่ต่ำที่สุดในการทำนายราคา



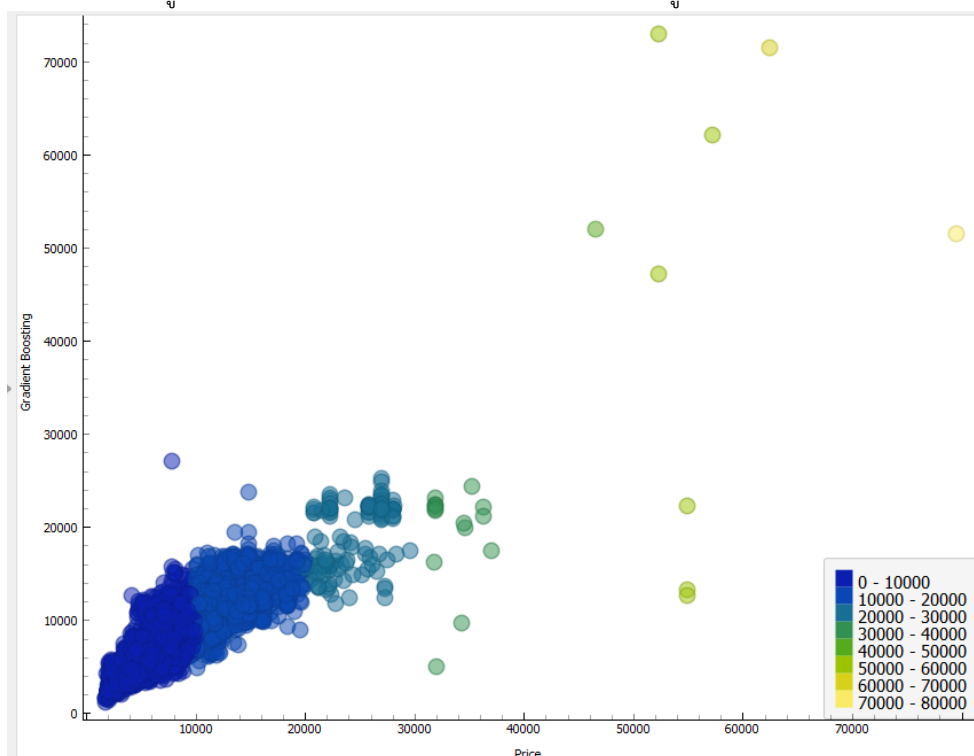
รูปที่ 2 กราฟ Scatter Plot ของการกระจายของข้อมูลจริง กับ Random Forest



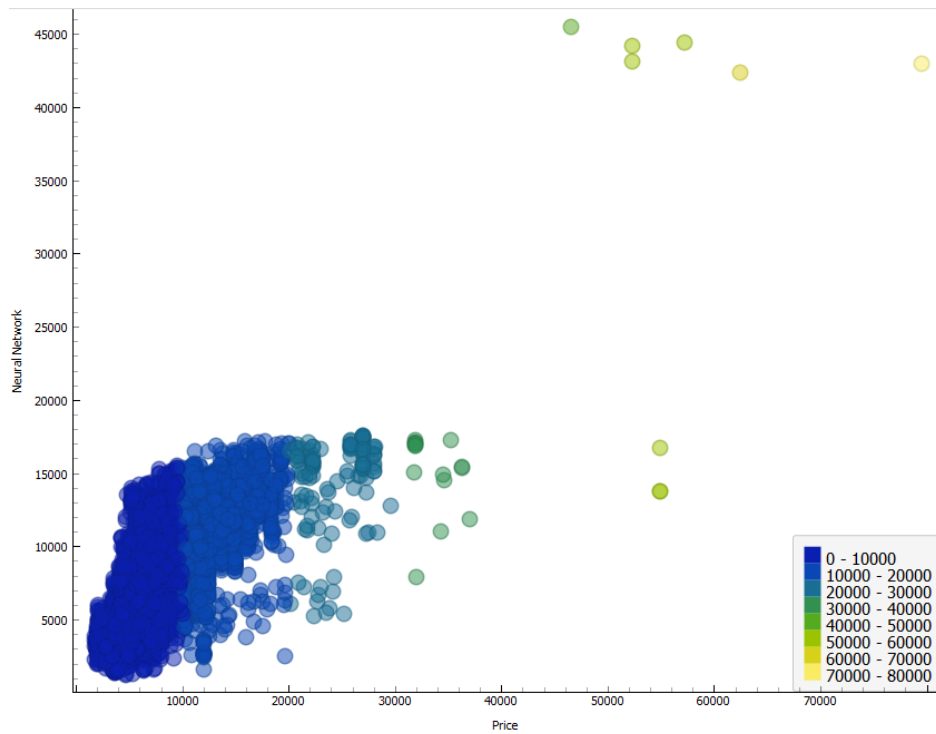
รูปที่ 3 กราฟ Scatter Plot ของการกระจายของข้อมูลจริง กับ AdaBoost



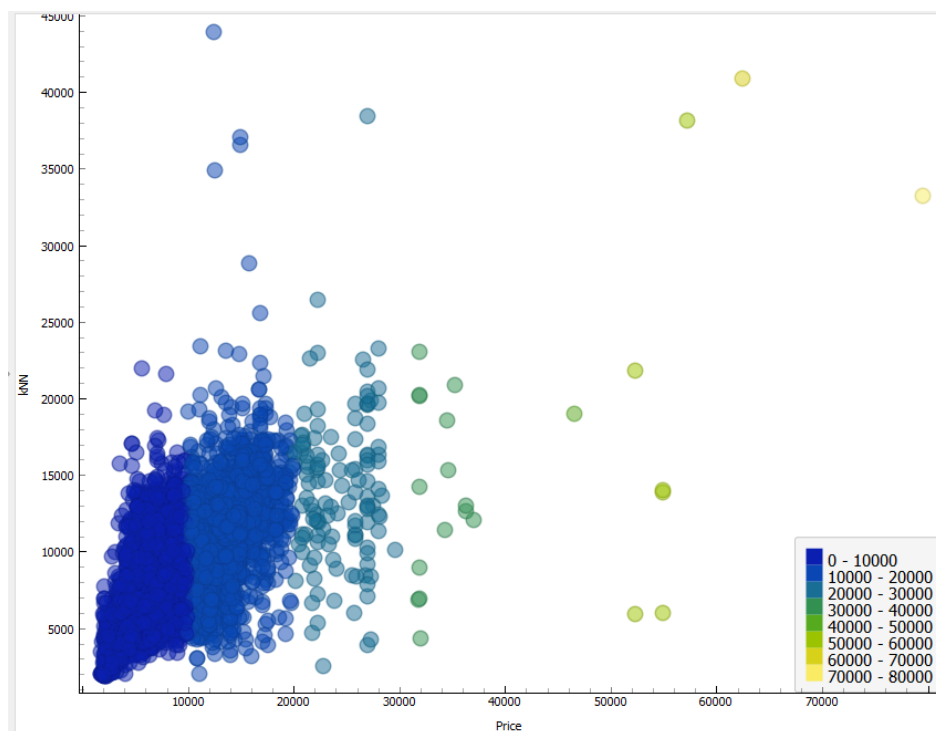
รูปที่ 4 กราฟ Scatter Plot ของการกระจายของข้อมูลจริง กับ Tree



รูปที่ 5 กราฟ Scatter Plot ของการกระจายของข้อมูลจริง กับ Gradient Boosting



รูปที่ 6 กราฟ Scatter Plot ของการกระจายของข้อมูลจริง กับ Neural Network



รูปที่ 7 กราฟ Scatter Plot ของการกระจายของข้อมูลจริง กับ kNN

จากรูปที่ 2-7 จะเห็นได้ว่า **Random Forest** มีการกระจายของข้อมูลจริงกับผลทำนายน้อยที่สุดเมื่อเทียบกับโมเดลอื่น ๆ ซึ่งสามารถสรุปได้จากค่า **R² Score** ที่สูงสุด (12.61%) และ **Explained Variance Score** ที่สูงสุด (82.73%) แสดงให้เห็นว่าโมเดลนี้สามารถอธิบายความแปรปรวนของข้อมูลได้ดีที่สุดและทำนายค่าได้ใกล้เคียงกับข้อมูลจริงมากที่สุด

ในขณะเดียวกัน **Neural Network** และ **k-Nearest Neighbor (kNN)** มีการกระจายของข้อมูลมากที่สุด โดยมี **R² Score** ต่ำที่สุด (24.14% และ 22.58% ตามลำดับ) แสดงถึงความแม่นยำที่ต่ำกว่าในการทำนายข้อมูลจริง

โมเดลอื่น ๆ เช่น **Gradient Boosting** และ **AdaBoost** มีผลลัพธ์อยู่ในระดับกลาง โดยทั้งสองโมเดลมีค่า **Explained Variance Score** ที่ใกล้เคียงกันที่ 76.48% และ 80.58% ซึ่งแสดงถึงความสามารถในการทำนายที่ดีกว่าโมเดล **Neural Network** และ **kNN** แต่ยังไม่ดีกว่า **Random Forest**

ดังนั้น โมเดล **Random Forest** มีความสามารถสูงสุดในการทำนายข้อมูลที่ใกล้เคียงกับข้อมูลจริง ตามด้วย **AdaBoost** และ **Gradient Boosting** ในขณะที่ **Neural Network** และ **kNN** มีการกระจายของข้อมูลมากที่สุด

4. สรุปผลการทดลอง

จากผลการทดลองของการเปรียบเทียบประสิทธิภาพของอัลกอริทึมทำนายราคาตัวเครื่องบิน จะเห็นได้ว่า อัลกอริทึมที่มีประสิทธิภาพในการทำนายราคาตัวเครื่องบินได้ดีที่สุด คือ **Random Forest** โดยมีค่า **Explained Variance Score** เท่ากับร้อยละ 82.73% และอัลกอริทึมที่มี ประสิทธิภาพลำดับที่ 2 คือ **AdaBoost** โดยมีค่า **Explained Variance Score** เท่ากับร้อยละ 80.58% และอัลกอริทึมที่มีประสิทธิภาพลำดับที่ 3 คือ **Tree** โดยมีค่า **Explained Variance Score** เท่ากับร้อยละ 78.55% และอัลกอริทึมที่มี ประสิทธิภาพลำดับที่ 4 คือ **Gradient Boosting** โดยมีค่า **Explained Variance Score** เท่ากับร้อยละ 76.48% และอัลกอริทึมที่มี ประสิทธิภาพลำดับที่ 5 คือ **Neural Network** โดยมีค่า **Explained Variance Score** เท่ากับร้อยละ 60.71% มีค่า **R² Score** สูงที่สุด แต่ค่า **Explained Variance Score** ต่ำกว่าโมเดลอื่นๆ และอัลกอริทึมที่มีประสิทธิภาพลำดับที่ 6 คือ **kNN** โดยมีค่า **Explained Variance Score** เท่ากับร้อยละ 53.04%

ดังนั้นคณะผู้วิจัยจึงสรุปได้ว่า **Random Forest** มีประสิทธิภาพในการทำนายราคาตัวเครื่องบินได้ดีที่สุด ดังนี้

1. ค่า **MSE**, **RMSE**, และ **MAE** ของ **Random Forest** มีค่า **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, และ **Mean Absolute Error (MAE)** ที่ต่ำที่สุดในบรรดาโมเดลทั้งหมด ซึ่งบ่งชี้ว่าสามารถทำนายราคาตัวเครื่องบินได้อย่างแม่นยำและเชื่อถือได้มากที่สุด
2. ค่า **Explained Variance Score** ของ **Random Forest** มีค่า **Explained Variance Score** สูงสุดที่ 82.73% ซึ่งหมายความว่าสามารถอธิบายความแปรปรวนของข้อมูลได้ดีที่สุด และให้การอธิบายที่ชัดเจนเกี่ยวกับการเปลี่ยนแปลงของราคาตัวเครื่องบิน
3. **Random Forest** เป็นอัลกอริทึมที่มีความยืดหยุ่นสูงในการจัดการกับข้อมูลที่มีความซับซ้อนและหลายมิติ ซึ่งทำให้เหมาะสมในการทำนายราคาตัวเครื่องบินที่อาจได้รับผลกระทบจากหลายปัจจัย
4. ประสิทธิภาพในการเรียนรู้จากข้อมูลจำนวนมากของ **Random Forest** สามารถเรียนรู้จากข้อมูลจำนวนมากและหลากหลายลักษณะได้ดี ซึ่งเป็นประโยชน์อย่างยิ่งในกรณีที่มีข้อมูลขนาดใหญ่และความหลากหลาย

ด้วยเหตุผลเหล่านี้คณะผู้วิจัยจึงถือว่า RandomForestเป็นโมเดลที่มีประสิทธิภาพสูงที่สุดในการทำนายราคาตั๋วเครื่องบินในกรณีนี้

5. กิตติกรรมประกาศ:

ขอขอบคุณสาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้

6. เอกสารอ้างอิง

Biswas, P., Chakraborty, R., Mallik, T., Uddin, S. I., Saha, S., Das, P., & Mitra, S. (2022). Flight price prediction: a case study. *International Journal of Research in Applied Science and Engineering Technology (IJRASET)*, 10(6).

Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied Energy*, 262, 114566.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21.

Panigrahi, A., Sharma, R., Chakravarty, S., Paikaray, B., & Bhoyar, H. (2022). Flight Price Prediction Using Machine Learning. In *ACI@ISIC* (pp. 172-178).

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, 47(1), 31-39.

Suthaharan, S., & Suthaharan, S. (2016). Decision tree learning. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, 237-269.

Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. (2017, August). Airfare prices prediction using machine learning techniques. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1036-1039). IEEE.

Wellman, M. P., Reeves, D. M., & Lochner, K. M. (2003, June). Price prediction in a trading agent competition. In *Proceedings of the 4th ACM conference on Electronic commerce* (pp. 216-217).

Wu, Y. C., & Feng, J. W. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, 102, 1645-1656.