CS 2566/53

รายงานความก้าวหน้าโครงหน้า ครั้งที่ 1

การประเมินราคาที่ดินจากข้อมูลกรมบังคับคดี โดยใช้อัลกอริทึมการเรียนรู้ของเครื่องกรณีศึกษาเขตอำเภอเมืองจังหวัดขอนแก่น

Appraisal of land prices from information from the Legal Execution Department By using machine learning algorithms, a case study in Muang District, Khon Kaen Province

โดย

633020566-7 นางสาวดุสิตา สังข์กลิ่นหอม

633021002-8 นางสาวโยษิตา ศรีวุฒิทรัพย์

อาจารย์ที่ปรึกษา : อ.ธนพล ตั้งชูพงศ์

รายงานนี้เป็นส่วนหนึ่งของการศึกษาวิชา SC314774 โครงงานวิทยาการคอมพิวเตอร์ 1

ภาคเรียนที่ 1 ปีการศึกษา 2566

วิทยาลัยการคอมพิวเตอร์

มหาวิทยาลัยขอนแก่น

(เดือน สิงหาคม พ.ศ. 2566)

CS 2566/53

รายงานความก้าวหน้าโครงหน้า ครั้งที่ 1

การประเมินราคาที่ดินจากข้อมูลกรมบังคับคดี โดยใช้อัลกอริทึมการเรียนรู้ของเครื่องกรณีศึกษาเขตอำเภอเมืองจังหวัดขอนแก่น

Appraisal of land prices from information from the Legal Execution Department By using machine learning algorithms, a case study in Muang District, Khon Kaen Province

โดย

633020566-7 นางสาวดุสิตา สังข์กลิ่นหอม

633021002-8 นางสาวโยษิตา ศรีวุฒิทรัพย์

อาจารย์ที่ปรึกษา : อ.ธนพล ตั้งชูพงศ์

รายงานนี้เป็นส่วนหนึ่งของการศึกษาวิชา SC314774 โครงงานวิทยาการคอมพิวเตอร์ 1

ภาคเรียนที่ 1 ปีการศึกษา 2566

วิทยาลัยการคอมพิวเตอร์

มหาวิทยาลัยขอนแก่น

(เดือน สิงหาคม พ.ศ. 2566)

โยษิตา ศรีวุฒิทรัพย์ และ ดุสิตา สังข์กลิ่นหอม. 2565. การประเมินราคาที่ดินจากข้อมูลกรมบังคับคดีโดยใช้อัลกอริทึมการเรียนรู้ของเครื่องกรณีศึกษาเขตอำเภอเมืองจังหวัดขอนแก่น. โครงงานคอมพิวเตอร์ปริญญาวิทยาศาสตร์บัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ วิทยาลัยการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น.

อาจารย์ที่ปรึกษา: อาจารย์ธนพล ตั้งชูพงศ์

บทคัดย่อ

โครงงานการประเมินราคาที่ดินจากข้อมูลกรมบังคับคดี โดยใช้อัลกอริทึมการเรียนรู้ของเครื่องกรณีศึกษาเขตอำเภอเมือง จังหวัดขอนแก่น กรณีศึกษาที่ดิน 412 แห่ง ในอำเภอเมือง จังหวัดขอนแก่น มีวัตถุประสงค์ เพื่อพัฒนาอัลกอริทึมโดยใช้ การเรียนรู้ของเครื่องคอมพิวเตอร์ สำหรับประเมินราคาที่ดิน กรณีศึกษาเขตอำเภอเมือง จังหวัดขอนแก่น เพื่อศึกษาปัจจัยที่เกี่ยวข้องกับการประเมินราคาที่ดิน เพื่อศึกษาและพัฒนาการสกัดคุณลักษณะสำคัญจากภาพถ่ายโดยอาศัยการแบ่งส่วนความหมาย เนื่องจากปัจจุบันราคาที่ดินนั่นมีมูลค่าที่สูงขึ้นเนื่องจากความเจริญของเมืองมีการขยายตัว มีห้างสรรพสินค้า รถไฟฟ้า โรงพยาบาล และโรงเรียน ยิ่งใกล้สถานที่สำคัญมากเท่าไร คนที่อาศัยอยู่ใกล้สถานที่สำคัญเหล่านั้นก็จะสะดวกสบายในการใช้ชีวิตมากขึ้น ดังนั้นราคาที่ดินในละแวกนั้นย่อมสูงขึ้นตามไปด้วย ทางคณะผู้จัดทำได้เห็นถึงความสำคัญในข้อนี้จึงได้จัดทำโครงงานนี้ขึ้นเพื่อประเมินราคาที่ดินใน อำเภอเมือง จังหวัดขอนแก่น ประกอบการตัดสินใจสำหรับผู้ที่ต้องการซื้อที่ดิน

โดยการประเมินราคาที่ดินใน อำเภอเมือง จังหวัดขอนแก่น โดยใช้อัลกอริทึมในการเรียนรู้ของเครื่องคอมพิวเตอร์ สามารถประเมินราคาที่ดินใน อำเภอเมือง จังหวัดขอนแก่น จากละติจูด ลองจิจูด ระยะทางจากสถานที่สำคัญ และรูปภาพที่ดินจากมุมสูงที่เห็นที่ดินบริเวณที่ต้องการประเมินราคา

โครงงานการประเมินราคาที่ดินใน อำเภอเมือง จังหวัดขอนแก่น โดยใช้อัลกอริทึมในการเรียนรู้ของเครื่องคอมพิวเตอร์ กรณีศึกษาที่ดิน 412 แห่ง ในอำเภอเมือง จังหวัดขอนแก่นสามารถแก้ปัญหาข้างต้นได้ โดยประเมินราคาต้องการซื้อที่ดินใน อำเภอเมือง จังหวัดขอนแก่น และเป็นแนวทางในการนำไปพัฒนาเพื่อนำไปใช้กับอุปกรณ์ มือถือ (Mobile) เว็บไซต์ (Website) ที่เป็น ฐาน (Platform) อื่นได้โดยใช้ฐานข้อมูลเดิม

คำสำคัญ: ประเมินราคาที่ดินในอำเภอเมืองจังหวัดขอนแก่นโดยใช้อัลกอริทึมในการเรียนรู้ของเครื่องคอมพิวเตอร์, การแบ่งส่วนความหมาย

Yosita Sriwuttisab and Dusita Sungklinhom. 2022. Appraisal of land prices from information from the Legal Execution Department By using machine learning algorithms, a case study in Muang District, Khon Kaen Province. computer science degree project Bachelor's Degree in Computer Science College of Computing Khon Kaen University

Advisor: Mr. Thanapon Tangchoopong

abstract

Project Appraisal of land prices from information from the Legal Execution Department By using machine learning algorithms, a case study in Muang District, Khon Kaen Province, a case study of 412 land plots in Mueang District, Khon Kaen Province. have a purpose to develop an algorithm using computer learning for land appraisal A case study in Muang District Khon Kaen to study factors related to land appraisal to study and develop the extraction of key features from photographs by means of semantic segmentation. Due to the current land price that has a higher value due to the growth of the city has expanded. There are shopping malls, train stations, hospitals and schools. The closer to the important places, the more. People who live near those important places will be more comfortable living. Therefore, the price of land in that area will increase as well. The organizing team has seen the importance of this issue and therefore has created this project to assess the price of land in Muang District, Khon Kaen Province. Make decisions for those who want to buy land.

By Appraisal of land in Mueang District, Khon Kaen Province using Machine Learning Algorithms can evaluate land prices in Mueang District, Khon Kaen Province from latitude, longitude, distance from important places. And a picture of the land from a high angle that sees the land in the area that we want to appraise.

Appraisal of land in Mueang District, Khon Kaen Province project using Machine Learning Algorithms, a case study of 412 land plots in Mueang District, Khon Kaen Province can solve the above problems. by appraising land prices in Muang District, Khon Kaen Province As specified in the required land information for decision making for those who want to buy land in Muang District, Khon Kaen Province And is a guideline for developing to be used with mobile devices (Mobile), websites (Websites) that are base (Platform) on other platforms by using the original database.

Keywords: Appraisal of land in Mueang District, Khon Kaen Province using Machine Learning Algorithms, semantic segmentation

กิตติกรรมประกาศ

ในการดำเนินโครงงานครั้งนี้ ผู้จัดทำโครงงานขอขอบคุณอาจารย์ธนพล ตั้งชูพงศ์ ที่เป็นที่ปรึกษา ให้คำชี้แนะและให้การสนับสนุนสำหรับการเดินทางเพื่อดำเนินงานต่าง ๆ ตลอดการทำโครงงานนี้ ขอขอบคุณบริษัท I-Net ที่ให้การสนับสนุน E-lab เพื่อเป็นพื้นที่ในการทำงาน ขอขอบคุณอาจารย์ ดร.ศักดิ์พจน์ ทองเลี่ยมนาค ที่ให้การสนับสนุนชุดข้อมูลภาพถ่ายดาวเทียมขอขอบคุณอาจารย์และรุ่นพี่ทุกท่านที่คอยอบรมสั่งสอน ขอขอบคุณ คุณพ่อ คุณแม่สำหรับกำลังใจรวมถึงความรักที่เป็นแรงผลักดันในการเรียนและการทำโครงงาน ขอขอบคุณเพื่อน ๆ ที่ให้คำปรึกษา และเป็นกำลังใจให้ซึ่งกันและกัน

ผู้จัดทำ
โยษิตา ศรีวุฒิทรัพย์
ดุสิตา สังข์กลิ่นหอม

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญภาพ

หน้า

สารบัญตาราง

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันราคาที่ดินนั่นมีมูลค่าที่สูงขึ้นเนื่องจากความเจริญของเมืองมีการขยายตัว มีห้างสรรพสินค้า รถไฟฟ้า โรงพยาบาล และโรงเรียน ยิ่งใกล้สถานที่สำคัญมากเท่าไร คนที่อาศัยอยู่ใกล้สถานที่สำคัญเหล่านั้นก็จะสะดวกสบายในการใช้ชีวิตมากขึ้น ดังนั้นราคาที่ดินในละแวกนั้นย่อมสูงขึ้นตามไปด้วย ทางคณะผู้จัดทำได้เห็นถึงความสำคัญในข้อนี้จึงได้จัดทำโครงงานนี้ขึ้นเพื่อประเมินราคาที่ดินใน อำเภอเมือง จังหวัดขอนแก่น ประกอบการตัดสินใจสำหรับผู้ที่ต้องการซื้อที่ดิน

โดยชุดข้อมูลของคณะผู้จัดทำเป็นปัญหาการถดถอย (Regression Problem) ทางคณะผู้จัดจึงเลือกใช้ 4 โมเดลที่เป็นประเภทการถดถอย (Regression) ได้แก่ Decision Tree จะสร้างเงื่อนไข If-else ขึ้นมาจากข้อมูลในตัวแปรเพื่อที่จะแบ่งข้อมูลออกเป็นกลุ่มใหม่ที่สามารถอธิบาย เป้าหมาย (Target) ได้ดีที่สุด โดยการสร้างเงื่อนไข If-else ในแต่ละตัวแปรจะถูกกำหนดด้วย Objective Function เป็น Residual sum of squares ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split Point) จากการลด (Minimize) ให้ RSS มีค่าน้อยที่สุด เพื่อแยกชุดข้อมูลเป็นสองส่วนเพื่อให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยให้อยู่ในจุดนั้นน้อยที่สุดอัลกอริทึมทำแบบนี้ซ้ำ และสร้างโครงสร้างเหมือนต้นไม้ ต่อมาคือ Random Forest เป็นเทคนิคการเรียนรู้แบบมีการสอน (Supervised Learning) ที่ใช้ในการแก้ไขปัญหาการจำแนกและการทำนาย (Classification and Regression) โดยอาศัยการรวมกันของหลาย ๆ Decision Trees ที่สร้างขึ้นแบบสุ่มเพื่อสร้างโมเดลทำนายที่แม่นยำและมีประสิทธิภาพ Gradient Boosted Trees เป็นเทคนิคการเรียนรู้แบบมีการสอน (Supervised Learning) ที่นำเสนอความสามารถในการทำนายและการจำแนกที่มีประสิทธิภาพ โดยใช้หลักการรวมกันของหลาย ๆ ต้นไม้การตัดสินใจ ที่สร้างขึ้นโดยต่อเนื่อง โดยให้ความสำคัญกับการแก้ไขความผิดพลาดของโมเดลก่อนหน้า Linear Regression เป็นเทคนิคทางสถิติที่ใช้ในการวิเคราะห์และทำนายความสัมพันธ์ระหว่างตัวแปรต้น (Explanatory Variables)หรือที่เรียกว่า "ตัวแปรอิสระ" (Independent Variables) กับตัวแปรตาม (Response Variable) หรือที่เรียกว่า "ตัวแปรตาม" (Dependent Variable) ที่เป็นค่าตัวเลข การถดถอยเชิงเส้นพยายามสร้างโมเดลเชิงเส้นที่อธิบายความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในลักษณะที่เป็นเส้นตรงที่เข้ากันได้ดีที่สุด แล้วใช้ เมทริกซ์ ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ (Mean Absolute Error) ค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง (Root Mean Square Error) และค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน (R-Squared) เพื่อที่จะวัดความถูกต้องของโมเดล อีกทั้งนำสถาปัตยกรรม U-net มาใช้ในการสกัดคุณลักษณะสำคัญจากภาพถ่ายโดยอาศัยการแบ่งส่วนความหมาย (Semantic Segmentation)

1.2 วัตถุประสงค์ของโครงงาน

1.2.1 เพื่อพัฒนาอัลกอริทึมโดยใช้ การเรียนรู้ของเครื่องคอมพิวเตอร์ สำหรับประเมินราคาที่ดินในเขตอำเภอเมือง จังหวัดขอนแก่น

1.2.2 เพื่อศึกษาปัจจัยที่เกี่ยวข้องกับการประเมินราคาที่ดิน

1.2.3 เพื่อศึกษาและพัฒนาการสกัดคุณลักษณะสำคัญจากภาพถ่ายโดยอาศัยการแบ่งส่วนความหมาย

1.3 ขอบเขตและข้อจำกัดของการวิจัย

โครงงานนี้เป็นการวิเคราะห์ราคาที่ดินโดยใช้การแบ่งส่วนภาพถ่ายดาวเทียม เพื่อประเมินราคาที่ดินของ อำเภอเมือง จังหวัดขอนแก่น

1.3.1 ใช้โมเดล Decision Tree, Random Forest, Gradient Boosted Trees, Linear Regression สำหรับประเมินราคาที่ดินกรณีศึกษาเขตอำเภอเมือง จังหวัดขอนแก่น

1.3.2 ฐานข้อมูล (Dataset) ชุดข้อมูลราคาประเมินที่ดินที่สนใจ 412 แห่งโดยข้อมูลได้จากกรมบังคับคดีและกรมที่ดินใน 1 ระเบียน (Record) ประกอบด้วย

1.3.2.1 ลำดับ

1.3.2.2 ล็อตที่/วันที่

1.3.2.3 หมายเลขคดี

1.3.2.4 ประเภททรัพย์

1.3.2.5 ไร่

1.3.2.6 งาน

1.3.2.7 ตารางวา

1.3.2.8 ราคาประเมิน

1.3.2.9 ตำบล

1.3.2.10 อำเภอ

1.3.2.11 จังหวัด

1.3.2.12 ที่ดินโฉนด

1.3.2.13 ราคาขายได้/ราคาเสนอสูงสุด

1.3.2.14 ระวาง

1.3.2.15 พิกัดแปลง

1.3.2.16 ชุดข้อมูลภาพถ่ายดาวเทียมอำเภอเมืองจังหวัดขอนแก่น

1.3.3 คุณลักษณะที่จะนำมาใช้ในการ regression

1.3.3.1 สกัดคุณลักษณะสำคัญจากภาพถ่ายโดยอาศัยการแบ่งส่วนความหมายโดยใช้สถาปัตยกรรมU-net

(1) สัดพื้นที่สีเขียว ที่อยู่อาศัย

(2) ขนาดพื้นที่มี 2 ขนาด 1 ไร่ และ 10 ไร่

1.3.3.2 สกัดคุณลักษณะสำคัญจาก ระยะทางจากสถานที่สำคัญในอำเภอเมือง จังหวัดขอนแก่น โดยวัดระยะทางจากละติจูดและลองติจูดจากที่ดินที่สนใจไปหาสถานที่สำคัญโดยใช้หน่วยเป็นเมตร

1.3.4 เป้าหมาย คือ ราคาขายได้/ราคาเสนอสูงสุด

1.3.5 วัดผลโมเดลโดย ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ ค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง และ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

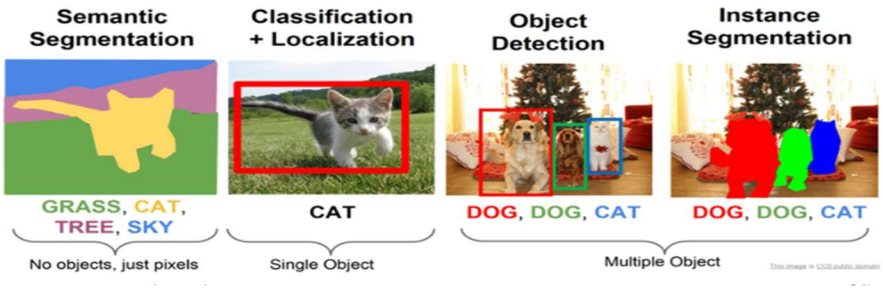
1.4.1 ประเมินราคาที่ดินใน อำเภอเมือง จังหวัดขอนแก่น ได้ตามที่ระบุข้อมูลที่ดินที่ต้องการลงไปเพื่อประกอบการตัดสินใจสำหรับผู้ที่ต้องการซื้อที่ดินใน อำเภอเมือง จังหวัดขอนแก่น

1.4.2 เป็นแนวทางในการนำไปพัฒนาเพื่อนำไปใช้กับอุปกรณ์ มือถือ (Mobile) เว็บไซต์ (Website) ที่เป็น ฐาน (Platform) อื่นได้โดยใช้ฐานข้อมูลเดิม

บทที่ 2
งานวิจัยและทฤษฎีที่เกี่ยวข้อง

2.1 การแบ่งส่วนความหมาย

การแบ่งส่วนความหมาย คือเทคนิคหนึ่งของวิสัยทัศน์คอมพิวเตอร์ที่มีหน้าที่ในการแยกส่วนและจำแนกวัตถุแต่ละวัตถุในภาพแยกจากกันโดยกำหนดคลาสในแต่ละจุดพิคเซลว่าเป็นคลาสอะไร [6] จะได้ผลออกมาเป็นแบ่งเป็นพื้นที่สีแบบต่าง ๆ ซึ่งแต่ละสีหมายความถึงลักษณะที่แตกต่างกัน เช่น บ้าน ถนน ต้นไม้ [7]



ภาพที่ 1 ภาพแสดงความแตกต่างระหว่าง การแบ่งส่วนความหมาย [7]

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องเป็นสาขาหนึ่งของปัญญาประดิษฐ์โดยเรียนรู้จากข้อมูลและทำนายข้อมูลได้โดยใช้อัลกอริทึมเรียนรู้ด้วยตนเองโดยการทำงานจะใช้แบบจำลอง (Model) ที่สร้างขึ้นจากชุดข้อมูลตัวอย่างในการทำนาย หรือ ตัดสินใจในภายหลังโดยอัลกอริทึมในการสอนชุดข้อมูลแบ่งได้ดังนี้ [5]

2.2.1 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

ถูกพัฒนาให้ใกล้เคียงกับการทำงานของสมองมนุษย์ยิ่งขึ้น ขั้นแรกคอมพิวเตอร์จะนำเอาข้อมูลเข้าไปจะไม่มีผลลัพธ์ออกมา ขั้นตอนต่อมาจะใช้กระบวนการเรียนรู้โดยใช้หลักทางสถิติหาค่าทางสถิติของชุดข้อมูลที่ฝึกสอนและทำการจัดกลุ่มข้อมูลออกเป็นระดับต่าง ๆ [5]

2.2.2 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

จะมีการเตรียมข้อมูลตัวอย่างและผลลัพธ์ที่ผู้สอนต้องการโดยขั้นแรกจะมีการนำข้อมูลมาสอนให้คอมพิวเตอร์เรียนรู้ด้วยผลลัพธ์ที่นำเข้าไปต่อมาคอมพิวเตอร์จะเชื่อมโยงข้อมูลและสร้างเป็นโมเดลไว้ทำนายผลลัพธ์ [5]

2.2.3 การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning)

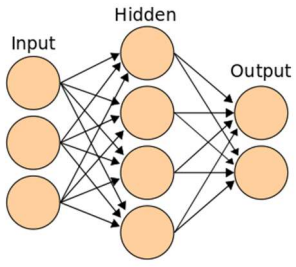
เป็นการเรียนรู้แบบลองผิดลองถูกและเป็นการเรียนรู้ว่าเป็นเส้นทางการทำงานแบบไหนที่จะทำให้ได้ผลลัพธ์ที่ดีที่สุด เช่น การเรียนรู้เพื่อเล่นเกม ตัวอย่างการใช้งานการเรียนรู้ของเครื่อง ได้แก่ การประเมินความต้องการของสินค้า โดยมีการคาดการณ์พฤติกรรมของลูกค้า โดยแนะนำสินค้าที่เหมาะสมให้ลูกค้า ซึ่งการเรียนรู้ของเครื่อง จะทำงานได้ดีกับข้อมูลที่เป็นระบบไม่ซับซ้อน เช่น แบบสอบถาม สถิติย้อนหลังผู้ป่วยโรคต่าง ๆ การทำงานที่มีความซับซ้อนสูงมากและไม่เป็นระบบจำเป็นต้องใช้การเรียนรู้เชิงลึก (Deep Learning) เข้ามาช่วยเพราะสามารถเรียนรู้ข้อมูลไปพร้อม ๆ กับพัฒนาตัวเองได้โดยไม่จำเป็นต้องให้การช่วยเหลือจากมนุษย์ [5]

2.3 การเรียนรู้เชิงลึก

เป็นอีกหนึ่งสาขาย่อยของการเรียนรู้ของเครื่องซึ่งเป็นการสร้างสถาปัตยกรรมโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) เพื่อเรียนรู้และจดจำความหมายของข้อมูลโดยสถาปัตยกรรมนั้นประกอบด้วยชั้นของประสาทเทียมจำนวนหลายชั้นและแต่ละชั้นนั้นมักจะมีการเปลี่ยนแปลงไม่เชิงเส้นผ่าน Activation Function เช่น การสร้างตัวแบบรู้จำรูปร่างต้นไม้ ถนน อีกทั้งการเรียนรู้เชิงลึกยังมีประสิทธิภาพสูงในการจัดการกับฟีเจอร์สำหรับการเรียนรู้แบบไม่มีหรือ การเรียนรู้แบบกึ่งมีผู้สอนโดยในใบงานนี้ได้เลือกใช้สถาปัตยกรรม U-Net [3]

2.4 โครงข่ายประสาทเทียม

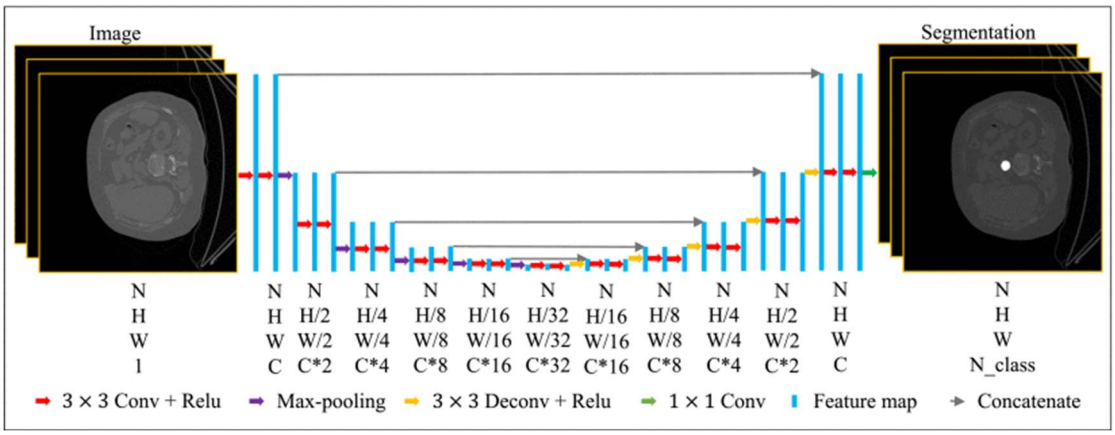
คือระบบคอมพิวเตอร์จากโมเดลทางคณิตศาสตร์ เพื่อจำลองการทำงานโครงข่ายประสาทชีวภาพที่อยู่ในสมองของสัตว์ โดยโครงข่ายประสาทเทียมยังสามารถเรียนรู้ที่จะทำงานที่มอบหมายได้โดยจากการที่ เรียนรู้ผ่านตัวอย่าง โดยไม่ถูกโปรแกรมด้วยกฎเกณฑ์ตายตัวแบบระบบอัตโนมัติ เช่น ในการประมวลผลภาพ คอมพิวเตอร์ที่ทำงานด้วยระบบโครงข่ายประสาทเทียมนั้นจะเรียนรู้การจำแนกรูปภาพบ้านได้จากการนำตัวอย่างของรูปภาพที่ถูกกำกับโดยผู้เขียนโปรแกรมว่าเป็นบ้านหรือไม่ใช่บ้านต่อไปก็นำผลลัพธ์ที่ได้ไปใช้ระบุภาพบ้านในตัวอย่างรูปภาพอื่นโดยโปรแกรมของโครงข่ายประสาทเทียมจะสามารถแยกแยะรูปภาพบ้านได้โดยไม่ต้องรู้ก่อนว่า ”บ้าน” คืออะไร อาทิ บ้านมีหลังคา มีรูปสี่เหลี่ยม มีหน้าต่าง มีประตู โครงข่ายประสาทเทียมจะทำการระบุตัวบ้านได้โดยอัตโนมัติเนื่องด้วยการระบุลักษณะเฉพาะ จากชุดข้อมูลตัวอย่างที่เคยได้ประมวลผล [8]



ภาพที่ 2 ข่ายงานประสาทเทียมมีการเชื่อมต่อกันผ่านกลุ่มโหนด [8]

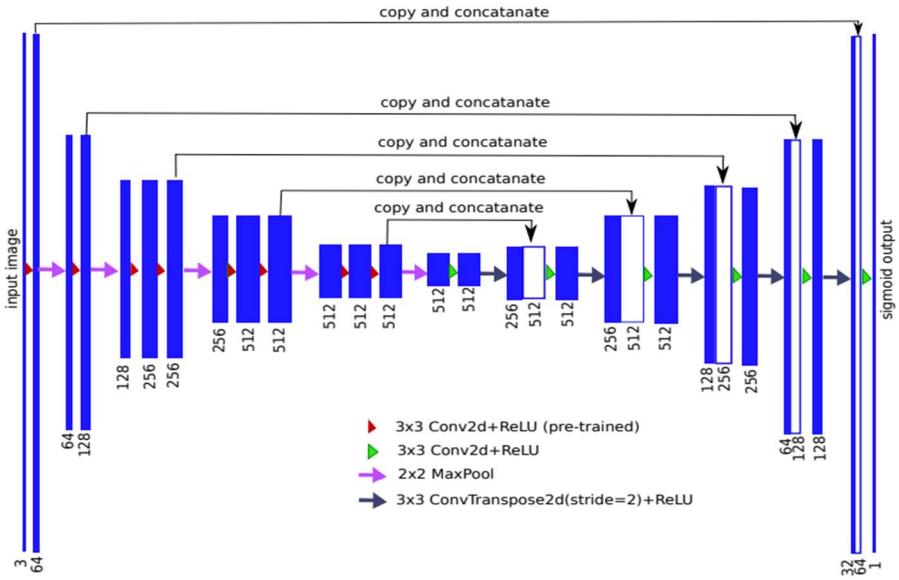
2.5 สถาปัตยกรรม U-net

U-Net ซึ่งเป็นโครงสร้าง DCNN โดยที่การทำงานจะค่อย ๆ เพิ่มฟิลด์การรับพิกเซล (Pixel) ที่เห็นด้วยเลเยอร์ (Layer) รวมสูงสุดส่งผลให้มิติเชิงพื้นที่ลดลงจากนั้น U-Net จะกู้คืนและเพิ่มมิติเชิงพื้นที่ด้วยชั้น Deconvolutional โครงสร้างของ U-Net ตัวอย่างใน ภาพที่ 3 ที่ใช้คือ Conv - Convolution, Deconv –Deconvolution [4]



ภาพที่ 3 โครงสร้างของ U-Net ที่ใช้ Conv - Convolution, Deconv – Deconvolution [4]

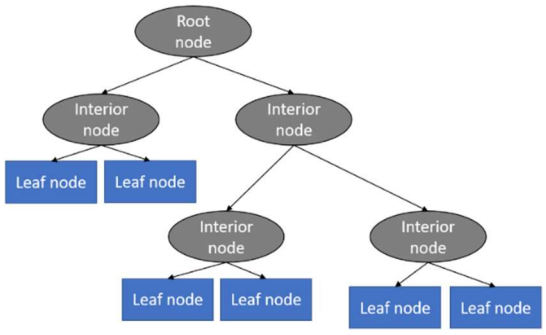
U-Net สามารถเรียนรู้จากชุดเทรนที่ค่อนข้างเล็ก [2] มักใช้ในงานด้านการมองเห็นและการประมวลผลภาพโดยเฉพาะในการแบ่งส่วนของภาพที่มีความละเอียดสูงมีบทบาทสำคัญอย่างมาก โมเดลประกอบด้วยเลเยอร์การพลิกกลับทั้งหมด 23 ชั้นซึ่งประกอบด้วยเส้นทางการหดตัวตัวเข้ารหัส (Encode) และเส้นทางการขยายตัวถอดรหัส (Decode) ส่วนเข้ารหัสประกอบด้วยบล็อกการบิดซ้ำ ๆ แต่ละบล็อกประกอบด้วยเลเยอร์การบิดสองชั้นพร้อมตัวกรองขนาด (3× 3) แต่ละอันจะตามด้วยการเปิดใช้งาน ReLU (Rectified Linear Unit) และการทำงานของพูลสูงสุด (2× 2) [1] ผลลัพธ์ของโมเดลคือระบุจุดแบบพิกเซลต่อพิกเซลที่แสดงคลาสของแต่ละพิกเซลโมเดลนี้ถูกพิสูจน์แล้วว่ามีประโยชน์มากสำหรับปัญหาการแบ่งส่วนที่มีข้อมูลจำนวนจำกัดในเครือข่าย U-Net ใช้ CNN ที่ค่อนข้างง่ายของตระกูล VGG ที่ประกอบด้วย 11 เลเยอร์ตามลำดับและรู้จักกันในชื่อ VGG11 ดูภาพที่ 4 VGG11vจะประกอบด้วยเลเยอร์ที่บิดเบี้ยวเจ็ดชั้น แต่ละชั้นนั้นตามด้วยฟังก์ชันการเปิดใช้งาน ReLU และการทำโพลสูงสุด 5 ครั้ง แต่ละฟีเจอร์ลดแปลง 2 ชั้น ชั้นคอนโวลูชัน (Convolutional Layers) ทั้งหมดมีเคอร์เนล 3×3 และจำนวนช่องสัญญาณจะแสดงภาพที่ 4 ชั้นคอนโวลูชันแรกสร้าง 64 ช่องสัญญาณจากนั้นเมื่อเครือข่ายลึกขึ้น จำนวนช่องสัญญาณจะเพิ่มเป็นสองเท่าหลังจากการดำเนินการรวมยอดแต่ละครั้งจนไปถึง 512 เลเยอร์ [2]



ภาพที่ 4 สถาปัตยกรรมโครงข่ายประสาทเทียมตัวถอดรหัสหรือที่เรียกว่า U-Net [2]

2.6 Decision Tree

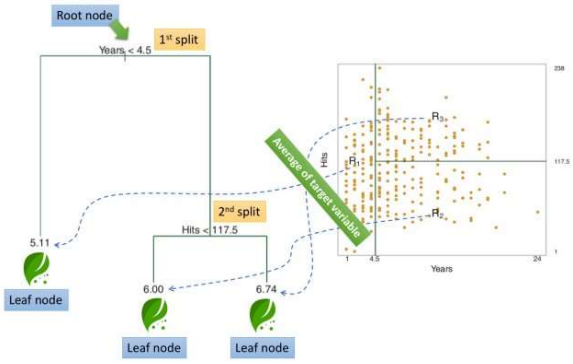
จะสร้างเงื่อนไข If-else ขึ้นมาจากข้อมูลในตัวแปร เพื่อที่จะแบ่งข้อมูลออกเป็นกลุ่มใหม่ที่สามารถอธิบายเป้าหมายได้ดีที่สุด โดยการสร้างเงื่อนไข If-else ในแต่ละตัวแปรจะถูกกำหนดด้วย Objective Function [4] โครงสร้างแบบต้นไม้มีโหนดสามประเภท โหนดราก (Root Node) เป็นโหนดเริ่มต้นซึ่งแสดงถึงตัวอย่างทั้งหมดและอาจแยกออกเป็นโหนดแบบอื่น ๆ เช่น โหนดภายใน (Interior Nodes) แสดงถึงคุณสมบัติของชุดข้อมูลและสาขา (Branches) เป็นตัวแทนของกฎการตัดสินใจสุดท้ายตัวแทนของผลลัพธ์ (Leaf Nodes) [9]



ภาพที่ 5 Source [9]

2.7 Regression Tree

ต้นไม้การตัดสินใจที่ใช้ในการแก้โจทย์ประเภทถดถอย (Regression) โดย ฟังก์ชันวัตถุประสงค์ (Objective function) ของต้นไม้ถดถอยเป็น RSS (Residual Sum Of Squares) ในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split Point) จากการลด (Minimize) ให้ RSS มีค่าน้อยที่สุด [11]



ภาพที่ 6 ตัวอย่างการทำ prediction [11]

Residual (Error term, Ei) คือ ค่าความคลาดเคลื่อนหรือค่าผิดพลาด (Error) ระหว่าง y ทุก ๆ จุดในข้อมูล กับ y_hat ที่ได้มาจากการประมาณค่า (Prediction) ขึ้นมาการคำนวณค่าความคลาดเคลื่อนของข้อมูลตัวที่ i [11]

$$e_i = y_i - \hat{y}_i \tag{1}$$

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 \tag{2}$$

หลักการในการแบ่งข้อมูล

1. เลือก 1 feature จาก k feature มาทำ sorting ข้อมูล ด้วยค่าของ feature ที่เลือกมา
2. หาจุดแบ่งข้อมูลที่เป็นไปได้ทั้งหมด จากข้อมูล n ตัวอย่าง (Observation)
   สามารถหาจุดแบ่งข้อมูลที่เป็นไปได้ n-1 จุด
3. สำหรับการแบ่งข้อมูลแต่ละแบบที่เป็นไปได้ คำนวณค่า RSS
4. เลือกจุดแบ่งข้อมูลที่ให้ค่า RSS น้อยที่สุดเมื่อสิ้นสุดการแบ่งข้อมูล (Split) แล้ว จะประมาณค่าผลลัพธ์ของการทำนาย (Target Variable) จากค่าค่าเฉลี่ย (Mean) ของผลลัพธ์ของการทำนายภายใน Node ของตัวเอง [10]

ข้อดี คือ รองรับข้อมูลแบบ Non-Linear เนื่องจากข้อมูลจริงในธรรมชาติไม่จำเป็นต้องเป็นเชิงเส้น (Linear) เสมอไปและให้ผล Trained Model ที่สามารถตีความได้ง่าย รวดเร็ว สามารถฝึกข้อมูลที่มีข้อผิดพลาด หรือค่าที่ขาดหายไปปรับขนาดได้ทั้งจำนวนตัวแปรและขนาดของชุดฝึก

ข้อเสีย คือ การเปลี่ยนแปลงข้อมูลการฝึกเพียงเล็กน้อยอาจส่งผลให้เกิดการเปลี่ยนแปลงครั้งใหญ่ในแผนผังและส่งผลให้มีการคาดคะเนขั้นสุดท้าย ต้องการข้อมูลจำนวนมากเพื่อผลลัพธ์ที่แม่นยำ สามารถนำไปสู่โครงสร้างต้นไม้ขนาดใหญ่และซับซ้อนมากเกินไป เมื่อชุดข้อมูลมีขนาดใหญ่ [12]

2.8 Multiple Regression

เป็นการวิเคราะห์การถดถอยเชิงเส้นแบบพหุคูณ มีวัตถุประสงค์ในการพยากรณ์หรือทำนายตัวแปรที่ต้องการศึกษาด้วยการสร้างสมการพยากรณ์ โดยมีตัวแปรเกณฑ์ (Y) ที่เป็นตัวแปรต่อเนื่องเพียง 1 ตัวและตัวแปรพยากรณ์ (X) หลายตัวแปร โดยมีรูปแบบสมการในการวิเคราะห์ ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \qquad (3)$$

Y	คือ ตัวแปรเกณฑ์
$X_1 X_2 \ldots X_n$	คือ ค่าของตัวแปรอิสระแต่ละตัว
β_0	คือ ส่วนตัดแกน Y เมื่อกำหนดให้ X_1=X_2=...X_n=0
$\beta_1\ \beta_2 \ldots \beta_n$	คือ ค่าสัมประสิทธิ์ความถดถอย
ε	คือ ค่าความคลาดเคลื่อน (Error or Residual)

โดยที่ β_i เป็นค่าที่แสดงถึงการเปลี่ยนแปลงของตัวแปรตาม Y เมื่อตัวแปรอิสระ X_i เปลี่ยนไป 1 หน่วยโดยที่ตัวแปรอิสระ X ตัวอื่น ๆ มีค่าคงที่ เช่น ถ้า X_1 เปลี่ยนไป 1 หน่วยค่า Y จะเปลี่ยนไป β_i หน่วย โดยที่ $X_1 X_2 \ldots X_n$ มีค่าคงที่

ข้อดี คือ ผลการวิเคราะห์ข้อมูลจะสามารถบอกขนาด ทิศทางอิทธิพลของตัวแปรต้นแต่ละตัวที่มีต่อตัวแปรตาม และสร้างสมการพยากรณ์ตัวแปรตามได้เมื่อรู้ค่าตัวแปรต้น

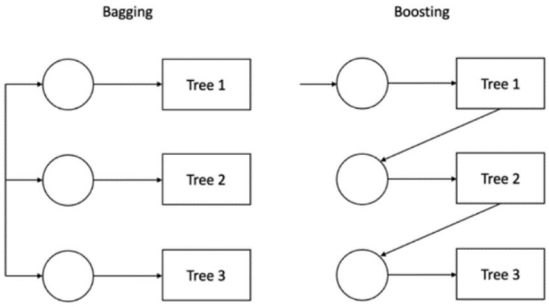
ข้อเสีย คือ อาจเกิดความสัมพันธ์ระหว่างตัวแปรพยากรณ์ (Multicollinearity) ด้วยกันที่ถือเป็นข้อตกลงเบื้องต้นข้อหนึ่งของการวิเคราะห์จากหลักการของการวิเคราะห์ถดถอยพหุคูณที่กล่าวมาแล้วว่าใช้ตัวแปรพยากรณ์หลายตัวในการทำนายตัวแปรเกณฑ์ตัวเดียว ทำให้ตัวแปรพยากรณ์บางตัวที่ไม่มีส่วนในการอธิบายการผันแปรต่อตัวแปรเกณฑ์ไม่มีความสำคัญต่อสมการพยากรณ์ ดังนั้นวิธีการคัดเลือกตัวแปรพยากรณ์จึงมีความจำเป็น เพื่อให้ได้สมการพยากรณ์ที่ดีที่สุด [13]

2.9 Random Forest

Random Forest เป็นเทคนิคการเรียนรู้ของเครื่องที่พัฒนามาจากแนวคิดของต้นไม้การตัดสินใจ ซึ่งเป็นส่วนหนึ่งของแนวทางการเรียนรู้แบบ Ensemble ซึ่งเป็นการรวมหลาย ๆ โมเดลเข้าด้วยกันเพื่อแก้ไขปัญหาที่มีความซับซ้อน Random Forest สร้างโมเดล โดยใช้หลาย ๆ ต้นไม้การตัดสินใจที่สร้างขึ้นโดยการสุ่มข้อมูลตัวอย่างและตัวแปรเข้ามาในแต่ละต้นไม้ แล้วรวมผลลัพธ์จากทุกต้นไม้เพื่อให้คำตอบสุดท้าย Random Forest สามารถแบ่งออกเป็นสองประเภทหลักคือ Bagging และ Boosting

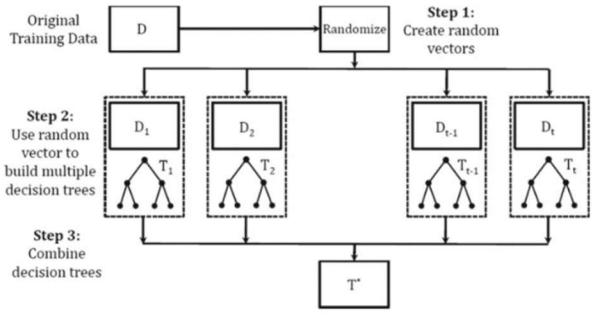
2.9.1. Bagging (Bootstrap Aggregating) : เป็นแนวทางการเรียนรู้แบบแยกจากกันโดยใช้การสุ่มข้อมูลตัวอย่างในการสร้างแต่ละต้นไม้ แล้วรวมผลลัพธ์จากทุก ๆ ต้นไม้ เพื่อให้การทำนายมีความน่าเชื่อถือสูงขึ้น

2.9.2. Boosting: เป็นแนวทางการเรียนรู้ที่จะทำการเรียนรู้และปรับปรุงแบบจำลองทีละขั้นตอน โดยให้ความสำคัญมากกับตัวอย่างที่ทำนายผิดพลาดมากในขั้นตอนก่อนหน้า เพื่อให้โมเดลพยายามแก้ไขความผิดพลาดเหล่านั้น [15]



ภาพที่ 7 ตัวอย่าง Bagging และ Boosting [15]

ต้นไม้แต่ละต้นใน Random Forest จะถูกฝึกสอนด้วยชุดข้อมูลย่อยและเซ็ตของคุณลักษณะ (Feature) ที่สุ่มเลือกมาทำให้แต่ละต้นไม้มีความหลากหลายและไม่เกิดการเรียนรู้จากข้อมูลเดิมทุกครั้งในขั้นตอนการทำนาย Random Forest จะให้ทุกต้นไม้ทำการทำนายจากนั้นส่งผลลัพธ์ที่ได้มาให้กับโมเดลหลักและโมเดลจะเลือกผลลัพธ์ที่ได้รับการโหวตมากที่สุดเป็นผลลัพธ์สุดท้าย[16]



ภาพที่ 8 อัลกอริทึมของ Random Forest [16]

ข้อดีของ Random Forest

1. ความแม่นยำและเสถียรภาพสูง Random Forest สามารถให้การทำนายที่แม่นยำและเสถียรได้ เนื่องจากมีการรวมผลลัพธ์จากหลาย ๆ ต้นไม้
2. การจัดการข้อมูลที่หายไป (Missing Data) Random Forest สามารถจัดการกับข้อมูลที่หายไปได้โดยที่ไม่ต้องลบข้อมูลตัวอย่าง ซึ่งช่วยให้เกิดข้อมูลที่ครอบคลุมมากขึ้น
3. ความยืดหยุ่นและการปรับแต่ง Random Forest มีพารามิเตอร์หลายอย่างที่สามารถปรับแต่งได้เพื่อให้สอดคล้องกับข้อมูลและเป้าหมายของงาน

ข้อเสียของ Random Forest

1. การทำงานของ Random Forest มีความซับซ้อนมากกว่าโมเดลเชิงเส้น เนื่องจากต้องสร้างและรวมผลลัพธ์จากหลายต้นไม้ซึ่งอาจทำให้การปรับปรุงและการทดสอบจะใช้เวลานานขึ้นอีกทั้งค่าความแม่นยำขึ้นอยู่กับการปรับแต่งพารามิเตอร์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด
2. ต้นไม้แต่ละต้นภายใน Random Forest อาจไม่มีความสามารถในการอธิบายของโมเดลเชิงเส้นแบบเดียว ทำให้คำอธิบายที่มาจากการใช้ Random Forest ยากต่อการเข้าใจ
3. การใช้งานทรัพยากรที่มากกว่าการสร้างโมเดลเชิงเส้น ทำให้ต้องพิจารณาความสมควรในการใช้งาน

2.10 Gradient Boosted Trees

เกิดจากพื้นฐานของ Decision Tree และใช้กระบวนการปรับปรุงเพื่อเพิ่มประสิทธิภาพของโมเดลในการทำนาย Gradient Boosted Trees เป็นเทคนิคการเรียนรู้ที่ใช้ในการแก้ปัญหาการจำแนกและการสร้างโมเดลทำนายที่มีการสอนโดยการสร้างโมเดลที่ใช้ส่วนประกอบที่อ่อนแอ (Weak Learning) โมเดลจะถูกสร้างขึ้นเป็นลำดับ[16] ค่าความผิดพลาดของโมเดลต้นที่ผ่านมาจะถูกใช้ในการกำหนดค่าในโมเดลต้นที่ถัดไป มักถูกนำมาใช้เนื่องจากมีความแม่นยำที่มีประสิทธิภาพ สามารถจัดการกับข้อมูลที่ขาดหายได้ และมีความยืดหยุ่นสูง อีกทั้งยังมีพารามิเตอร์ต่าง ๆ ที่สามารถปรับแต่งได้ เช่น จำนวนต้นไม้หรือจำนวนรอบการทำซ้ำ ความลึกสูงสุดของต้นไม้ ฟังก์ชันสูญเสีย (Loss Function) และอัตราการเรียนรู้ (Learning Rate) [17]

ข้อดีของ Gradient Boosted Trees

1. ความแม่นยำสูง Gradient Boosted Trees สามารถให้ผลลัพธ์ที่แม่นยำและเหมาะสมกับข้อมูลได้มาก เนื่องจากมีการปรับปรุงโมเดลจากต้นไม้เข้าไปเรื่อย ๆ โดยอิงค่าความผิดพลาดของโมเดลก่อนหน้า
2. การจัดการ Overfitting ด้วยกระบวนการปรับปรุงที่ทำให้โมเดลเรียนรู้จากข้อผิดพลาดของโมเดลก่อนหน้า
3. การจัดการคุณลักษณะ (Feature) สามารถเรียนรู้ความสำคัญของคุณลักษณะในการทำนายและสามารถให้คะแนนความสำคัญให้กับแต่ละคุณลักษณะ

4. ความยืดหยุ่นในการปรับแต่ง Gradient Boosted Trees มีพารามิเตอร์ต่าง ๆ ที่สามารถปรับแต่งได้ เช่น จำนวนต้นไม้ ความลึกของต้นไม้ ฟังก์ชันสูญเสีย และอัตราการเรียนรู้

ข้อเสียของ Gradient Boosted Trees

1. การปรับแต่งและการสร้างจำเป็นต้องใช้พารามิเตอร์หลายอย่าง ทำให้กระบวนการติดตั้งและปรับปรุงโมเดลอาจมีความซับซ้อน

2. เนื่องจากต้องสร้างต้นไม้หลาย ๆ ต้นและปรับปรุงในแต่ละขั้นตอน Gradient Boosted Trees อาจใช้เวลานานกว่าโมเดลอื่น ๆ ในกรณีข้อมูลมากหรือต้นไม้มาก

3. ในบางกรณี Gradient Boosted Trees อาจมีความเสี่ยงในเรื่องของการเจาะจงข้อมูล เนื่องจากการใช้งานต้นไม้หลาย ๆ ต้นที่ได้รับการปรับปรุงในข้อมูลที่เหลืออาจทำให้โมเดลเรียนรู้ข้อมูลส่วนนั้นได้

4. การเลือกค่าพารามิเตอร์ที่เหมาะสมสำหรับ Gradient Boosted Trees เป็นกระบวนการที่ซับซ้อนและอาจต้องใช้การทดลองหลายครั้ง

2.11 แฮเวอร์ซีน (Haversine)

ระยะทาง Haversine คือระยะห่างเชิงมุมระหว่างจุดสองจุดบนพื้นผิวของทรงกลม พิกัดแรกของแต่ละจุดจะถือว่าเป็นละติจูด พิกัดที่สองคือลองจิจูด กำหนดเป็นเรเดียน โดยมีสูตรการคำนวณคือ [14]

$$D_{(x,y)} = 2aresin[\sqrt{sin^2\ \left(\frac{x1-y1}{2}\right) = \ +\cos(x1)\cos(y1)\,sin^2(\frac{x2-y2}{2})]} \qquad (4)$$

บทที่ 3
วิธีการดำเนินงาน

3.1 วิธีการดำเนินงาน

3.1.1 กำหนดหัวข้อโครงงานที่สนใจ

3.1.2 ศึกษาค้นคว้าข้อมูลทฤษฎี งานวิจัยที่เกี่ยวข้อง และเครื่องมือที่ใช้ดำเนินงานวิจัย

3.1.2.1 ทฤษฎีที่เกี่ยวข้อง ประกอบไปด้วย

(1) การแบ่งส่วนความหมาย

(2) การเรียนรู้เชิงลึก

(3) โครงข่ายประสาทเทียม

(4) สถาปัตยกรรม U-net

(5) Decision Tree

(6) Regression Tree

(7) Random Forest

(8) Gradient Boosted Trees

(9) แฮเวอร์ซีน

3.1.2.2 งานวิจัยที่เกี่ยวข้อง ประกอบไปด้วย

(1) Road Segmentation using U-Net architecture

(2) U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation

(3) Accuracy Improvement for Segmentation and Classification of Wound Tissues through Region-Focus Training

(4) Normalization in Training U-Net for 2-D Biomedical Semantic Segmentation

(5) ระบบจำแนกถนนชำรุด

(6) Naive Bayes Classifier, Decision Tree and AdaBoost Ensemble Algorithm– Advantages and Disadvantages

(7) การคัดเลือกตัวแปรในตัวแบบการถดถอยเชิงเส้นพหุโดยใช้วิธีการค้นหาแบบต้องห้าม

(8) การคัดเลือกตัวแปรพยากรณ์เข้าในสมการถดถอยพหุคูณ

3.1.2.3 เครื่องมือที่ใช้ดำเนินงานวิจัย ประกอบไปด้วย

(1) เครื่องคอมพิวเตอร์ พร้อมเชื่อมต่ออินเทอร์เน็ต

(2) Colab

(3) E-lab

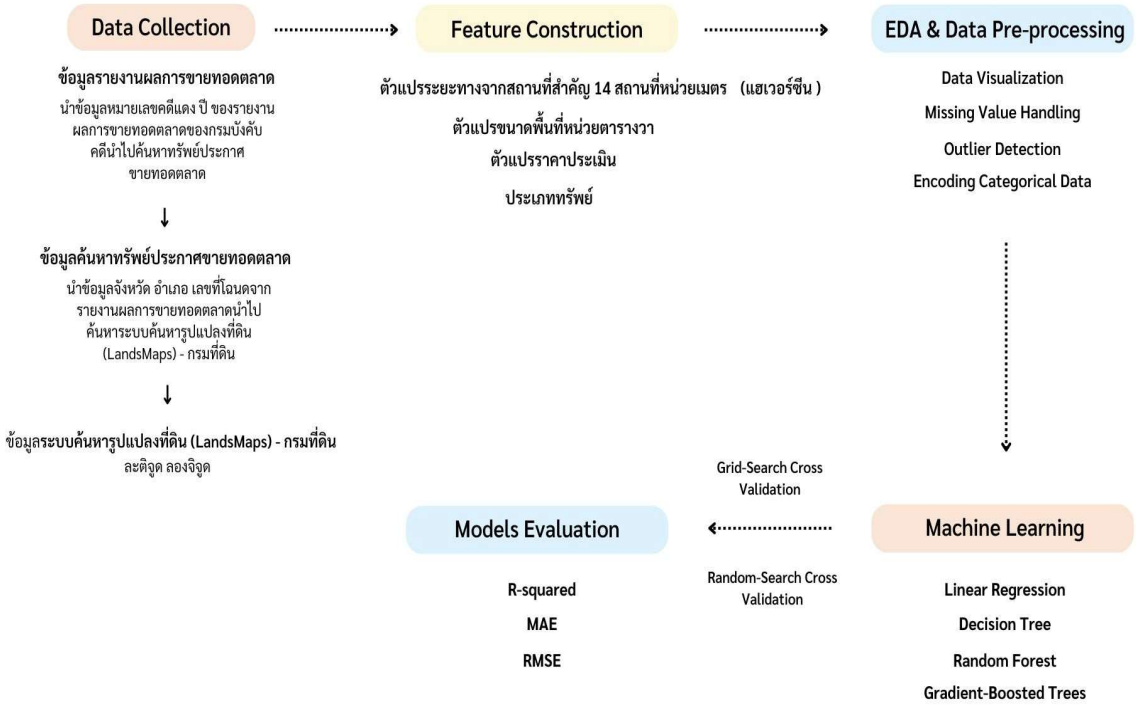
(4) ระบบค้นหารูปแปลงที่ดิน (LandsMaps) – กรมที่ดิน

(5) กรมบังคับคดี

(6) Selenium

3.1.3 กำหนดขอบเขตและเป้าหมายของโครงงานเพื่อวิเคราะห์ความเป็นไปได้

3.1.4 โครงสร้างในการประมวลผลข้อมูลในกระบวนการวิเคราะห์ข้อมูล



ภาพที่ 9 โครงสร้างในการประมวลผลข้อมูลในกระบวนการวิเคราะห์ข้อมูล

3.1.5 หาฐานข้อมูลที่ดิน

ผู้จัดทำใช้ชุดข้อมูลราคาประเมินที่ดินที่ดินที่สนใจ 412 แห่งโดยข้อมูลได้จากกรมบังคับคดีและกรมที่ดิน ใน 1 ระเบียนประกอบด้วย ลำดับ ล็อตที่/วันที่ หมายเลขคดี ประเภททรัพย์ ไร่ งาน ตารางวา ราคาประเมิน ตำบล อำเภอ จังหวัด ที่ดินโฉนด ราคาขายได้/ราคาเสนอสูงสุด ระวาง และ พิกัดแปลง

ตารางที่ 1 รายละเอียดชุดข้อมูลจากกรมบังคับคดีและกรมที่ดิน

ลำดับ	ล็อตที่/วันที่	หมายเลขคดี	ประเภททรัพย์	ไร่	งาน	ตารางวา
1 - 17	21 มิ.ย/2566 - ขอนแก่น	ผบ.1003	ที่ดินพร้อมสิ่งปลูกสร้าง	-	-	73.4
1 - 42	มหกรรม ก.ย/2566 - ขอนแก่น	ผบ.1096	ที่ดินพร้อมสิ่งปลูกสร้าง	-	-	82
1 - 39	มหกรรม ก.ย/2566 - ขอนแก่น	ผบE.4122	ที่ดินพร้อมสิ่งปลูกสร้าง	-	-	46
1 - 37	21 มิ.ย/2566 - ขอนแก่น	อ.948	ที่ดินพร้อมสิ่งปลูกสร้าง	-	-	60.2
1 - 44	มหกรรม / 2566 - ขอนแกน	ผบ.1109	ที่ดินว่างเปล่า	-	2	-

ตารางที่ 2 รายละเอียดชุดข้อมูลจากกรมบังคับคดีและกรมที่ดิน

ลำดับ	ราคาประเมิน	ตำบล	อำเภอ	จังหวัด	ที่ดินโฉนด
1 - 17	3,196,000.00	บ้านเป็ด	เมืองขอนแก่น	ขอนแก่น	254492
1 - 42	759,600.00	ดอนหัน	เมืองขอนแก่น	ขอนแก่น	284389
1 - 39	427,280.00	เมืองเก่า	เมืองขอนแก่น	ขอนแก่น	76003
1 - 37	525,600.00	บ้านทุ่ม	เมืองขอนแก่น	ขอนแก่น	184279
1 - 44	500,000.00	ท่าพระ	เมืองขอนแก่น	ขอนแก่น	186591

ตารางที่ 3 รายละเอียดชุดข้อมูลจากกรมบังคับคดีและกรมที่ดิน

ลำดับ	ราคาขายได้/ราคาเสนอสูงสุด	ระวาง	พิกัดแปลง
1 - 17	3,350,000	5541 I 6414-05 (1000)	16.41020062,102.78767544
1 - 42	540,000	5541 I 7006-04 (1000)	16.34306450,102.85952879
1 - 39	122,000	5541 I 7010-03 (1000)	16.37864085,102.85478639
1 - 37	390,000	5541 IV 5020-00 (4000)	16.45639701,102.66242311
1 - 44	440,000	5541 I 6604-00 (4000)	16.30971574,102.80944771

3.1.6 หาชุดข้อมูลภาพถ่ายดาวเทียม อำเภอเมือง จังหวัดขอนแก่น

ผู้จัดทำใช้ชุดข้อมูลภาพถ่ายดาวเทียมอำเภอเมือง จังหวัดขอนแก่นโดยกำหนดละติจูด ลองติจูดจากที่ดิโดยการสุ่มตัวอย่างแบบชั้นภูมิโดยกำหนดกลุ่มของราคาประเมินเป็น 6 กลุ่มตามตัวเลขเริ่มต้นและตัวเลขสิ้นสุดของราคาที่ได้จากกรมบังคับคดี

ภาพที่ 10 ชุดข้อมูลภาพถ่ายดาวเทียม อำเภอเมือง จังหวัดขอนแก่น

3.1.7 หาคุณลักษณะเด่นของข้อมูล (Features) เพื่อนำไปใช้ในการเรียนรู้ของเครื่อง

ผู้จัดทำสกัดคุณลักษณะสำคัญจากระยะทางจากสถานที่สำคัญในอำเภอเมืองจังหวัดขอนแก่นดังตารางที่ 7 โดยวัดระยะทางจากละติจูด ลองติจูดจากที่ดินที่สนใจไปหาสถานที่สำคัญโดยใช้หน่วยเป็นกิโลเมตร

ตารางที่ 4 ละติจูด ลองติจูดของจากสถานที่สำคัญ

ละติจูด	ลองติจูด	สถานที่
16.4652775	102.7868931	ท่าอากาศยานนานาชาติขอนแก่น
16.4423117	102.8200274	โรงพยาบาลกรุงเทพขอนแก่น
16.4680574	102.8300541	โรงพยาบาลศรีนครินทร์
16.4321938	102.8236214	มหาวิทยาลัยขอนแก่น
16.41681	102.818995	ตลาดต้นตาล
16.4081492	102.8342183	พระมหาธาตุแก่นนคร
16.4255311	102.8347574	ตลาดโต้รุ่งรื่นรมย์ ขอนแก่น
16.4329489	102.8257675	Central Khonkaen
16.3713146	102.832789	KK Falabella Horse

ตารางที่ 4 ละติจูด ลองติจูดของจากสถานที่สำคัญ (ต่อ)

ละติจูด	ลองติจูด	สถานที่
16.4170864	102.8351271	Bueng Kaen Nakhon Public Park (km)
16.4290746	102.8301506	Pullman Khon Kaen Raja Orchid (km)
16.4304997	102.8318962	Tukcom Khonkaen
16.4321938	102.8236214	Big C Supercenter Khon Kaen 2
16.4461011	102.8386202	Khon Kaen National Museum

สกัดคุณลักษณะสำคัญจากระยะทางแบบแฮเวอร์ซีนซึ่งเป็นระยะห่างเชิงมุมระหว่างจุดสองจุดบนพื้นผิวของทรงกลม จากละติจูด ลองติจูดจากที่ดินที่สนใจไปหาสถานที่สำคัญโดยใช้หน่วยเป็นเมตร

ตารางที่ 5 ระยะทางจากละติจูด ลองติจูดจากที่ดินที่สนใจไปหาสถานที่สำคัญ

ID	Latitude	Longitude	ท่าอากาศยานนานาชาติขอนแก่น	โรงพยาบาลกรุงเทพขอนแก่น	โรงพยาบาลศรีนครินทร์
1	16.41020062	102.78767544	6.1248	4.9654	7.8623
2	16.34306450	102.85952879	15.643	11.8129	14.2497
3	16.37864085	102.85478639	12.0517	7.9919	10.2866
4	16.45639701	102.66242311	13.3098	16.8803	17.9226

ตารางที่ 6 ระยะทางจากละติจูด ลองติจูดจากที่ดินที่สนใจไปหาสถานที่สำคัญ

ID	มหาวิทยาลัยขอนแก่น	ตลาดต้นตาล	พระมหาธาตุแก่นนคร	ตลาดโต้รุ่งรื่นรมย์ขอนแก่น	Central Khonkaen	KK Falabella Horse
1	4.5475	3.4205	4.9698	5.3033	4.7859	6.4697
2	10.6252	9.2704	7.7244	9.543	10.6238	4.2435
3	6.8199	5.7088	3.9472	5.6347	6.7859	2.4842
4	17.4006	17.269	19.0918	18.6969	17.6141	20.487

ตารางที่ 7 ระยะทางจากละติจูด ลองติจูดจากที่ดินที่สนใจไปหาสถานที่สำคัญ

ID	Bueng Kaen Nakhon Public Park	Pullman Khon Kaen Raja Orchid	Tukcom Khonkaen	Big C Supercenter Khon Kaen 2	Khon Kaen National Museum
1	5.1189	4.9929	5.2288	4.5475	6.7423
2	8.6327	10.0643	10.1594	10.6252	11.6722
3	4.7616	6.1931	6.2621	6.8199	7.6969
4	18.9306	18.1439	18.3018	17.4006	18.8250

3.1.8 นำข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์ (Training set) แล้วได้แบบจำลอง

3.1.8.1 รายละเอียดชุดข้อมูล

ผู้จัดทำใช้ชุดข้อมูลที่ดินที่สนใจ 412 แห่งโดยข้อมูลแสดงรายละเอียดเกี่ยวกับข้อมูล (Data) เพื่อดูจำนวนแถวและคอลัมน์ (Column) โดยละลอง 5 ปีคือข้อมูลตั้งแต่ปี 60 ถึง 65 แต่ละลอง 66 ปีคือปี 66

ตารางที่ 8 รายละเอียดเกี่ยวกับจำนวนแถวและคอลัมน์

ข้อมูล	จำนวนแถวและคอลัมน์
ละลอง 5 ปี	(86, 30)
ละลอง 66 ปี	(355, 30)

ข้อมูลแสดงรายละเอียดชื่อของคอลัมภ์ทั้งหมดพร้อมทั้งเพื่อแสดงประเภทของข้อมูลในแต่ละคอลัมน์ ประเภทข้อมูลรวมถึงตัวแปรที่ใช้ในแต่ละคอลัมน์ เช่น int, float, string, datetime, bool เป็นต้น การรู้ประเภทของข้อมูลในแต่ละคอลัมน์มีความสำคัญเพราะมันช่วยให้คุณทราบว่าคอลัมน์แต่ละคอลัมน์เป็นตัวแปรประเภทใด ซึ่งสามารถช่วยในการปรับแก้ปัญหาที่เกี่ยวข้องกับการแปลงและการจัดการข้อมูลได้

ตารางที่ 9 รายละเอียดเกี่ยวกับประเภทของข้อมูล

ข้อมูล	ประเภท
ลำดับที่	Object
ล็อตที่/วันที่	Object
หมายเลขคดี	Object
ประเภททรัพย์	Object
ไร่	Float64
งาน	Float64
ตารางวา	Float64
ราคาประเมิน	Float64
ตำบล	Object
อำเภอ	Object
จังหวัด	Object
ที่ดินโฉนด	Int64
ราคาขายได้/ราคาเสนอสูงสุด	Int64
ระวาง	Object
พิกัดแปลง	Object
ท่าอากาศยานนานาชาติขอนแก่น	Float64
โรงพยาบาลกรุงเทพขอนแก่น (km)	Float64
โรงพยาบาลศรีนครินทร์ (km)	Float64
มหาวิทยาลัยขอนแก่น (km)	Float64
ตลาดต้นตาล (km)	Float64
พระมหาธาตุแก่นนคร (km)	Float64
ตลาดโต้รุ่งรื่นรมย์ ขอนแก่น (km)	Float64
Central Khonkaen (km)	Float64
KK Falabella Horse (km)	Float64
Bueng Kaen Nakhon Public Park (km)	Float64
Pullman Khon Kaen Raja Orchid (km)	Float64
Tukcom Khonkaen	Float64
Big C Supercenter Khon Kaen 2	Float64
Khon Kaen National Museum	Float64

3.1.8.2 การจัดการกับชุดข้อมูล

ผู้จัดทำจัดเตรียมข้อมูลก่อนเข้าแบบจำลองโดย เปลี่ยนประเภทของคอลัมน์ ราคาประเมิน ราคาขายได้/ราคาเสนอสูงสุด ไร่ งาน ตารางวา

ตารางที่ 10 รายละเอียดเกี่ยวกับประเภทของข้อมูลหลังจัดการกับชุดข้อมูล

ข้อมูล	**ประเภท**
ลำดับที่	Object
ล็อตที่/วันที่	Object
หมายเลขคดี	Object
ประเภททรัพย์	Object
ไร่	Float64
งาน	Float64
ตารางวา	Float64
ราคาประเมิน	Float64
ตำบล	Object
อำเภอ	Object
จังหวัด	Object
ที่ดินโฉนด	Int64
ราคาขายได้/ราคาเสนอสูงสุด	Int64
ระวาง	Object
พิกัดแปลง	Object
ท่าอากาศยานนานาชาติขอนแก่น	Float64
โรงพยาบาลกรุงเทพขอนแก่น (km)	Float64
โรงพยาบาลศรีนครินทร์ (km)	Float64
มหาวิทยาลัยขอนแก่น (km)	Float64
ตลาดต้นตาล (km)	Float64
พระมหาธาตุแก่นนคร (km)	Float64
ตลาดโต้รุ่งรื่นรมย์ ขอนแก่น (km)	Float64
Central Khonkaen (km)	Float64
KK Falabella Horse (km)	Float64
Bueng Kaen Nakhon Public Park (km)	Float64

ตารางที่ 10 รายละเอียดเกี่ยวกับประเภทของข้อมูล (ต่อ)

Pullman Khon Kaen Raja Orchid (km)	Float64
Tukcom Khonkaen	Float64
Big C Supercenter Khon Kaen 2	Float64
Khon Kaen National Museum	Float64
Area size	Float64

ตรวจสอบและจัดการกับข้อมูลที่หายไป (Missing data) เป็นขั้นตอนสำคัญในการวิเคราะห์ข้อมูล เนื่องจากข้อมูลที่หายไปอาจส่งผลกระทบต่อการวิเคราะห์และการสรุปของผลได้ ดังนั้นผู้จัดทำตรวจสอบ

ตารางที่ 11 ตรวจสอบและจัดการกับข้อมูลที่หายไป

ข้อมูล	**ข้อมูลที่หาย**
ลำดับที่	0
ล็อตที่/วันที่	0
หมายเลขคดี	0
ประเภททรัพย์	0
ไร่	0
งาน	0
ตารางวา	0
ราคาประเมิน	0
ตำบล	0
อำเภอ	0
จังหวัด	0
ที่ดินโฉนด	0
ราคาขายได้/ราคาเสนอสูงสุด	0
ระวาง	0
พิกัดแปลง	0
ท่าอากาศยานนานาชาติขอนแก่น	0
โรงพยาบาลกรุงเทพขอนแก่น (km)	0

ตารางที่ 11 ตรวจสอบและจัดการกับข้อมูลที่หายไป (ต่อ)

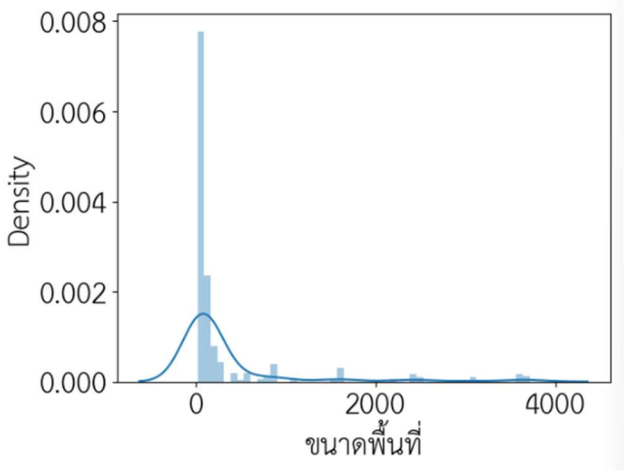
ข้อมูล	ข้อมูลที่หาย
โรงพยาบาลศรีนครินทร์ (km)	0
มหาวิทยาลัยขอนแก่น (km)	0
ตลาดต้นตาล (km)	0
พระมหาธาตุแก่นนคร (km)	0
ตลาดโต้รุ่งรื่นรมย์ ขอนแก่น (km)	0
Central Khonkaen (km)	0
KK Falabella Horse (km)	0
Bueng Kaen Nakhon Public Park (km)	0
Pullman Khon Kaen Raja Orchid (km)	0
Tukcom Khonkaen	0
Big C Supercenter Khon Kaen 2	0
Khon Kaen National Museum	0
Area size	0

ผู้จัดทำได้ตัดแถวที่ราคาขายได้/ราคาเสนอสูงสุดต้องมากกว่า 0 เมื่อเจ้าพนักงานเปิดประมูล ผู้ที่สนใจสามารถยกป้ายเพื่อขอเสนอราคาตามราคาเริ่มต้น หรือยกป้ายสู้ราคากับผู้ประมูลอื่น โดยเจ้าพนักงานจะเป็นผู้กำหนดว่าจะเพิ่มราคาครั้งละเท่าใด แต่ผู้ประมูลสามารถเพิ่มราคาได้เท่ากับหรือมากกว่าที่กำหนดไว้ก็ได้

ตารางที่ 12 รายละเอียดเกี่ยวกับประเภทของข้อมูล

ประเภทของข้อมูล	ราคาขายได้/ราคาเสนอสูงสุด
Count	4.120000e+02
Mean	1.580473e+06
STD	1.749492e+06
Min	3.000000e+04
25%	2.550000e+05
50%	9.500000e+05
75%	2.152500e+06
Max	8.210000e+06

ผู้จัดทำได้ทำการคำนวนไร่ งาน ตารางวาได้คำนวณ 1 ไร่ เท่ากับ 4 งาน และ 1 งาน เท่ากับ 100 ตารางวา โดยสร้างคอลัมภ์ขนาดพื้นที่ มาเก็บข้อมูลคำนวนให้เป็นหน่วยตารางวา



ภาพที่ 11 กราฟ Histogram ของขนาดพื้นที่

ผู้จัดทำได้การแปลงข้อมูลที่เป็นข้อมูลที่มีค่าในรูปแบบของกลุ่ม (Categorical data) เป็นรูปแบบที่เหมาะสำหรับการประมวลผลด้วยโมเดลทางสถิติหรือแบบจำลองข้อมูลโดยใช้ OneHotEncoder

ตารางที่ 13 รายละเอียดเกี่ยวกับการแปลงข้อมูลที่เป็นข้อมูลที่มีค่าในรูปแบบของกลุ่ม

ประเภททรัพย์ที่ดินพร้อมสิ่งปลูกสร้าง	ประเภททรัพย์ที่ดินว่างเปล่า
1	0
1	0
1	0
1	0
...	...
0	1
0	1
1	0
0	1
1	0

3.1.8.3 การฝึกสอนแบบจำลอง

ผู้จัดทำการแบ่งชุดข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลจะถูกแบ่งเป็น 80% สำหรับข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์ (X_train, y_train) และ 20% สำหรับชุดข้อมูลทดสอบ (test set) (X_test, y_test) โดยตัวแปรต้นจะมีประเภทของข้อมูลดังนี้

ตารางที่ 14 รายละเอียดเกี่ยวกับข้อมูล

ตัวแปรต้น	**ประเภทของข้อมูล**
ประเภททรัพย์	ข้อมูลเชิงคุณภาพ
ท่าอากาศยานนานาชาติขอนแก่น	ข้อมูลเชิงปริมาณ
โรงพยาบาลกรุงเทพ	ข้อมูลเชิงปริมาณ
โรงพยาบาลศรีนครินทร์	ข้อมูลเชิงปริมาณ
มหาวิทยาลัยขอนแก่น	ข้อมูลเชิงปริมาณ
ตลาดต้นตาล	ข้อมูลเชิงปริมาณ
พระมหาธาตุแก่นนคร	ข้อมูลเชิงปริมาณ
ตลาดโต้รุ่งรื่นรมย์ ขอนแก่น	ข้อมูลเชิงปริมาณ
Central Khonkaen	ข้อมูลเชิงปริมาณ
KK Falabella Horse	ข้อมูลเชิงปริมาณ
Bueng Kaen Nakhon Public Park	ข้อมูลเชิงปริมาณ
Pullman Khon Kaen Raja Orchid	ข้อมูลเชิงปริมาณ
Tukcom Khonkaen	ข้อมูลเชิงปริมาณ
Big C Supercenter Khon Kaen 2	ข้อมูลเชิงปริมาณ
Khon Kaen National Museum	ข้อมูลเชิงปริมาณ
ขนาดพื้นที่	ข้อมูลเชิงปริมาณ

ผู้จัดทำการแบ่งชุดข้อมูลออกเป็น 2 ส่วนโดยจะทำให้การแบ่งชุดข้อมูลเป็นแบบสุ่มแบบเดียวกันทุกครั้งโดยแสดงการแบ่งข้อมูลตามตารางที่ 13 เพื่อนำมาฝึกสอนแบ่งออกเป็น 4 แบบจำลองได้แก่ Decision Tree, Random Forest, Gradient Boosted Trees, Linear Regression

ตารางที่ 15 ขนาดของชุดข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์และชุดข้อมูลทดสอบ

ชุดข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์	ขนาดของชุดข้อมูล
Training Features Shape	(329,18)
Training Labels Shape	(329,)
Testing Features Shape	(83, 18)
Testing Labels Shape	(83,)

3.1.9 วัดผลแบบจำลองการเรียนรู้ของเครื่อง

งานวิจัยนี้ประเมินประสิทธิภาพของแบบจำลองโดยใช้ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง ค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง และค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน ซึ่งค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ วัดความแตกต่างสัมบูรณ์เฉลี่ยระหว่างค่าจริง และค่าที่คาดการณ์ คำนวณโดยใช้ค่าเฉลี่ยของความแตกต่างสัมบูรณ์ระหว่างค่าที่คาดการณ์และค่าจริงค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง วัดค่าเฉลี่ยของความแตกต่างกำลังสองระหว่างค่าที่คาดการณ์และค่าจริงคำนวณโดยใช้ค่าเฉลี่ยของผลต่างกำลังสองระหว่างค่าที่คาดการณ์และค่าจริง ค่าความคลาดเคลื่อนเฉลี่ยรากที่สองคือค่ารากที่สองของ ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง และเป็นเมตริกยอดนิยมเนื่องจากอยู่ในหน่วยเดียวกันกับตัวแปรเป้าหมาย คำนวณโดยใช้รากที่สองของ ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนคือการวัดทางสถิติที่แสดงถึงสัดส่วนของความแปรปรวนในตัวแปรตามซึ่งอธิบายโดยตัวแปรอิสระในแบบจำลอง มีค่าตั้งแต่ 0 ถึง 1 โดย 1 ระบุความพอดีระหว่างโมเดลและข้อมูล รวมถึง 0 ระบุว่าไม่มีความสัมพันธ์ระหว่างโมเดลและข้อมูล

โดยสรุปแล้ว ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง วัดข้อผิดพลาดระหว่างค่าที่คาดการณ์และค่าจริง ในขณะที่ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน วัดสัดส่วนของความแปรปรวนในตัวแปรตามซึ่งอธิบายโดยตัวแปรอิสระในแบบจำลอง

3.2 แผนงานและระยะเวลาดำเนินงาน

ในการจัดทำโครงงาน การประเมินราคาที่ดินจากข้อมูลกรมบังคับคดีโดยใช้อัลกอริทึมการเรียนรู้ของเครื่อง กรณีศึกษาเขตอำเภอเมืองจังหวัดขอนแก่นมีแผนงานและระยะเวลาในการดำเนินงานดังนี้

ตารางที่ 16 แผนงานและระยะเวลาดำเนินงาน

การดำเนินการ	สัปดาห์ที่/เดือน																			
	ปี 2566																			
	มิถุนายน				กรกฎาคม				สิงหาคม				กันยายน				ตุลาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
กำหนดหัวข้อโครงงานที่สนใจ																				
กำหนดขอบเขตและเป้าหมายของโครงงานเพื่อวิเคราะห์ความเป็นไปได้																				
เตรียมฐานข้อมูล																				
หาคุณลักษณะเด่นของข้อมูล เพื่อนำไปใช้ในการเรียนรู้ของเครื่อง																				

ตารางที่ 16 แผนงานและระยะเวลาดำเนินงาน (ต่อ)

การดำเนินการ	สัปดาห์ที่/เดือน																			
	ปี 2566																			
	มิถุนายน				กรกฎาคม				สิงหาคม				กันยายน				ตุลาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
นำข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์แล้วได้แบบจำลอง																				
วัดผลแบบจำลองการเรียนรู้ของเครื่อง																				
เขียนรายงาน																				

3.3 งบประมาณ

หมวดวัสดุอุปกรณ์

- ค่าวัสดุสำนักงาน (กระดาษ หมึก ฯลฯ) 500 บาท

หมวดค่าใช้สอย

- ค่าถ่ายเอกสาร จัดรูปเล่ม 500 บาท

หมวดค่าใช้จ่ายอื่นๆ 1,000 บาท

รวม 2,000 บาท

บทที่ 4
ผลการดำเนินงาน

4.1 ผลการศึกษาแบบจำลอง

4.1.1 การศึกษาแบบจำลอง

ผู้จัดทำได้ศึกษาแบบจำลอง 4 แบบจำลอง ได้แก่ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees โดยได้สรุปไว้ตารางที่ 17

ตารางที่ 17 ลักษณะเด่นของแต่ละแบบจำลอง

แบบจำลอง	ลักษณะเด่น
Linear Regression	ผลการวิเคราะห์ข้อมูลจะสามารถบอกขนาด ทิศทางอิทธิพลของตัวแปรต้นแต่ละตัวที่มีต่อตัวแปรตาม และสร้างสมการพยากรณ์ตัวแปรตามได้เมื่อรู้ค่าตัวแปรต้น
Regression Tree	สามารถจำแนกโมเดลที่ไม่เชิงเส้นได้ ซึ่งสามารถใช้ในการพยากรณ์ที่มีความซับซ้อนมากขึ้น ความยืดหยุ่นในการเลือกโมเดลโดยการปรับขนาดและปรับพารามิเตอร์ของโมเดลเพื่อให้เหมาะสมกับข้อมูล รวมถึงมีความสามารถในการทำโมเดลที่มีความซับซ้อน
Random Forest	ความแม่นยำและเสถียรภาพสูง Random Forest สามารถให้การทำนายที่แม่นยำและเสถียรได้ เนื่องจากมีการรวมผลลัพธ์จากหลายๆ ต้นไม้ความยืดหยุ่นและการปรับแต่ง Random Forest มีพารามิเตอร์หลายอย่างที่สามารถปรับแต่งได้
Gradient-boosted Trees	ความแม่นยำสูง Gradient Boosted Trees สามารถให้ผลลัพธ์ที่แม่นยำและเหมาะสมกับข้อมูลได้มาก เนื่องจากมีการปรับปรุงโมเดลจากต้นไม้เข้าไปเรื่อย ๆ โดยอิงค่าความผิดพลาดของโมเดลก่อนหน้า

4.1.2 ผลการฝึกสอนและประเมินประสิทธิภาพ

ผลที่ได้จากLinear Regressionค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.862 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 494106.484 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 661839.738 ผลที่ได้จาก Regression Tree โดยใช้ Grid Search ในการหาค่าพารามิเตอร์ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.968 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 149634.343 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 319120.274

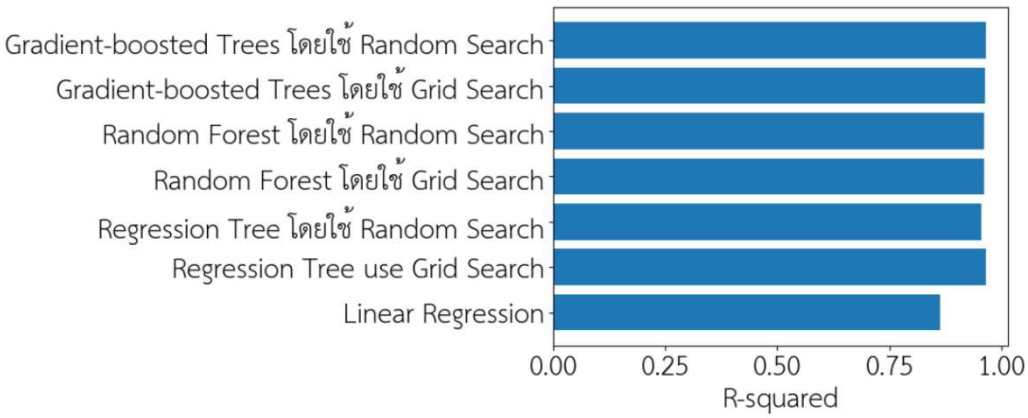
ผลที่ได้จาก Regression Tree โดยใช้ Random Search ในการหาค่าพารามิเตอร์ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.960 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 131785.943 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 354059.528 ผลที่ได้จาก Random Forest โดยใช้ Grid Search ในการหาค่าพารามิเตอร์ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.959 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 207564.036 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 360597.362 ผลที่ได้จาก Random Forest โดยใช้ Random Search ในการหาค่าพารามิเตอร์ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.960 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 189149.745 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 352946.525 ผลที่ได้จาก Gradient-boosted Trees โดยใช้ Grid Search ในการหาค่าพารามิเตอร์ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.965 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 151003.267 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 332627.454 ผลที่ได้จาก Gradient-boosted Trees โดยใช้ Random Search ในการหาค่าพารามิเตอร์ ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน 0.962 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ 139552.939 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 345263.109

ตารางที่ 18 ผลการฝึกสอนและประเมินประสิทธิภาพของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees

วัดประสิทธิภาพของโมเดลโดย	R-squared	MAE	RMSE
Multiple Regression	0.862	494106.484	661839.738
Regression Tree โดยใช้ Grid Search	0.968	149634.343	319120.274
Regression Tree โดยใช้ Random Search	0.960	131785.943	354059.528
Random Forest โดยใช้ Grid Search	0.959	207564.036	360597.362
Random Forest โดยใช้ Random Search	0.960	189149.745	352946.525
Gradient-boosted Trees โดยใช้ Grid Search	0.965	151003.267	332627.454
Gradient-boosted Trees โดยใช้ Random Search	0.962	139552.939	345263.109

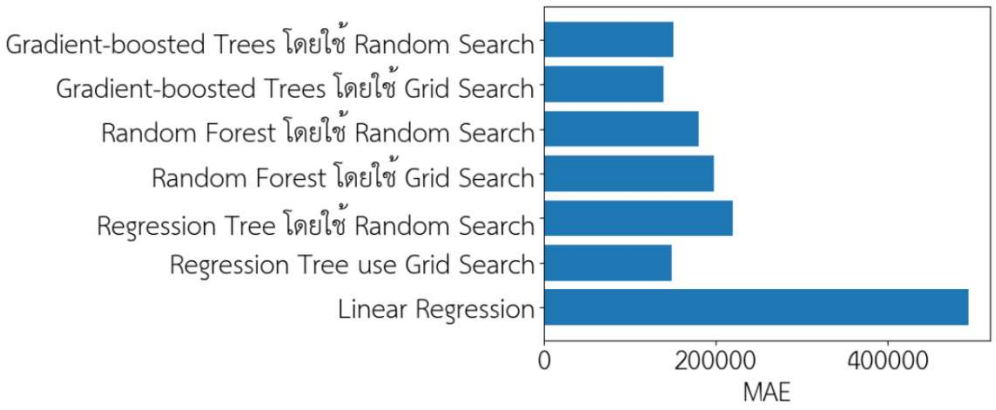
4.1.3 ผลการเปรียบเทียบผลลัพธ์จากตัวแบบจำลองทั้ง 3 แบบ

การประเมินประสิทธิภาพของแต่ละตัวแบบโดยที่แกน x จะเป็นโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees โดยใช้ Grid Search หรือ Random Searchใน การหาค่าพารามิเตอร์ที่ดีที่สุดสำหรับโมเดลเครื่องมือการเรียนรู้เชิงลึกหรือโมเดลอื่น ๆ เพื่อให้ได้ประสิทธิภาพที่มากที่สุด การประเมินประสิทธิภาพของแต่ละตัวแบบได้นำค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน มาทำกราฟแท่งเปรียบเทียบแต่ละโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees



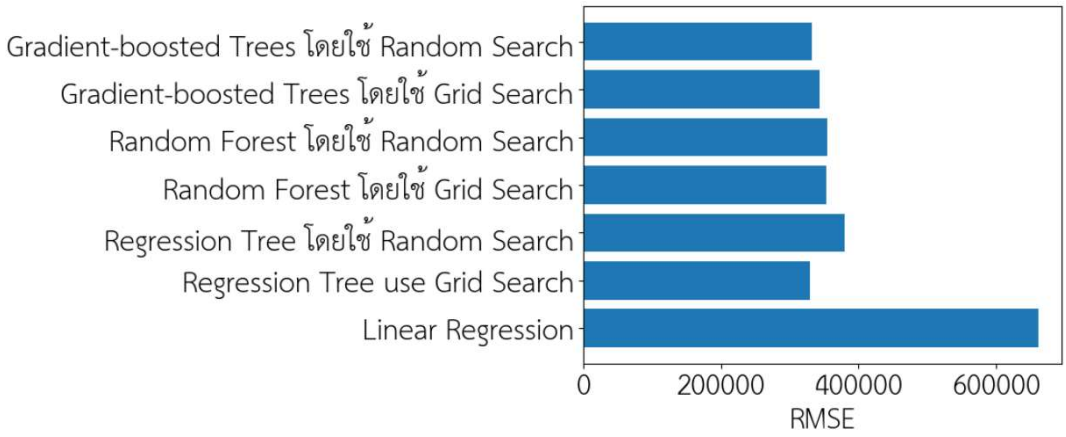
ภาพที่ 12 ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน ของโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees

การประเมินประสิทธิภาพของแต่ละตัวแบบได้นำค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์มาทำกราฟแท่งเปรียบเทียบแต่ละโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees



ภาพที่ 13 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ ของโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees

การประเมินประสิทธิภาพของแต่ละตัวแบบได้นำค่าความคลาดเคลื่อนเฉลี่ยรากที่สองมาทำกราฟแท่งเปรียบเทียบแต่ละโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees



ภาพที่ 14 ค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง ของโมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees

บทที่ 5
บทสรุป

5.1 สรุปการดำเนินโครงงาน

จากการศึกษาและดำเนินการทดลองเพื่อพัฒนาอัลกอริทึมโดยใช้ การเรียนรู้ของเครื่องคอมพิวเตอร์สำหรับหาราคาขายได้/ราคาเสนอสูงสุดกรณีศึกษาเขตอำเภอเมือง จังหวัดขอนแก่นโดยจากค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง ที่ได้จากการประมาณค่าโดยใช้ โมเดลของ Linear Regression, Regression Tree, Random Forest และ Gradient-boosted Trees โดย Regression Tree, Random Forest และ Gradient-boosted Trees ใช้ Grid Search หรือ Random Searchในการหาค่าพารามิเตอร์ที่ดีที่สุดสำหรับโมเดลเครื่องมือการเรียนรู้เชิงลึกหรือโมเดลอื่นๆ เพื่อให้ได้ประสิทธิภาพที่มากที่สุด สามารถสรุปการดำเนินโครงงานได้ดังนี้

Linear Regression แสดง ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนที่ 0.862 แสดงถึงความสามารถในการอธิบายความเปลี่ยนแปลงของตัวแปรตามตัวแปรต้นที่ประมาณ 86.2% และมีค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ และ ค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง ที่สูงกว่าที่ค่าอื่นๆ คือ 494106.484 และ 661839.738 ตามลำดับ จึงสังเกตได้ว่าโมเดลนี้มีประสิทธิภาพต่ำในการทำนายข้อมูล Regression Tree ที่ใช้ Grid Search มี ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนที่ 0.968 Regression Tree ใช้ Random Search แสดง ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวน ที่ 0.960 Regression Tree ที่ใช้ Grid Search แสดงถึงความสามารถในการอธิบายข้อมูลที่ดีกว่าการใช้ Random Search สำหรับ Regression Tree ที่ใช้ Grid Search มีค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ เท่ากับ 149634.343 Regression Tree ที่ใช้ Random Search มีค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ เท่ากับ 131785.943 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ ของ Regression Tree ที่ใช้ Random Search มีค่าน้อยกว่า แสดงว่าโมเดลนี้มีความคลาดเคลื่อนในการทำนายที่น้อยกว่าเมื่อเทียบกับ Regression Tree ที่ใช้ Grid Search สำหรับ Regression Tree ที่ใช้ Grid Search มีค่า RMSE เท่ากับ 319120.274 สำหรับ Regression Tree ที่ใช้ Random Search มีค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง เท่ากับ 354059.528 ค่าความคลาดเคลื่อนเฉลี่ยรากที่สองของ Regression Tree ที่ใช้ Grid Search มีค่าน้อยกว่า แสดงว่าจะเกิดจากการมี outlier หรือข้อมูลผิดปกติที่มีค่ามากๆ ที่มีผลกระทบต่อค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง RMSE เนื่องจากมีการนำค่า Error มายกกำลังสองแตกต่างกันทำให้ ค่าความคลาดเคลื่อนเฉลี่ยรากที่สองของ Regression Tree ที่ใช้ Random Search น้อยกว่าค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนของ Regression Tree ที่ใช้ Grid Search Random Forest ที่ใช้ Grid Search มีค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนเท่ากับ 0.959 Random Forest ที่ใช้ Random Search มีค่า R-squared เท่ากับ 0.960 ค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนของทั้งสองโมเดลใกล้เคียงกันมาก

แสดงว่าทั้งสองโมเดลมีความสามารถในการอธิบายข้อมูลอย่างใกล้เคียงกัน Gradient-boosted Trees ที่ใช้ Grid Search มีค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนเท่ากับ 0.965 Gradient-boosted Trees ที่ใช้ Random Search มีค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนเท่ากับ 0.962 มีค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนของทั้งสองโมเดลใกล้เคียงกัน แสดงว่าทั้งสองโมเดลมีความสามารถในการอธิบายข้อมูลอย่างใกล้เคียงกัน Gradient-boosted Trees ที่ใช้ Grid Search ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์เท่ากับ 151003.267 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สองเท่ากับ 332627.454 Gradient-boosted Trees ที่ใช้ Random Search มีค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์เท่ากับ 139552.939 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สองเท่ากับ 345263.109 ค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์ของ Gradient-boosted Trees ที่ใช้ Random Search น้อยกว่า แสดงว่าโมเดลนี้มีความคลาดเคลื่อนในการทำนายที่น้อยกว่าเมื่อเทียบกับ Gradient-boosted Trees ที่ใช้ Grid Search สรุปผลได้ว่าโมเดลที่มีประสิทธิภาพดีที่สุดคือ Regression Tree ที่ใช้ Grid Search โดยมีค่าสัมประสิทธิ์กำลังสองของการอธิบายความแปรปรวนสูงสุดที่ 0.968 และมีค่าความคลาดเคลื่อนเฉลี่ยสัมบูรณ์149634.343 และค่าความคลาดเคลื่อนเฉลี่ยรากที่สอง 319120.274

5.2 ปัญหาและอุปสรรค

การหาพื้นที่สำคัญที่ส่งผลต่อราคาขายได้/ราคาเสนอสูงสุด

5.3 ข้อเสนอแนะ

ใช้ semantic segmentation เขามาช่วยในการบอกพื้นที่ว่าพื้นที่แถบนี้เป็นอะไร เช่น พื้นที่สีเขียว ที่อยู่อาศัย

เอกสารอ้างอิง

[1] N. Y. Q. Abderrahim, S. Abderrahim และ A. Rida, “Road Segmentation using U-Net architecture,” ใน 2020 IEEE International conference of Moroccan Geomatics (Morgeo), 2020.

[2] V. Iglovikov และ A. Shvets, TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation, arXiv, 2018.

[3] C. PATHOMPATAI และ others, “Accuracy Improvement for Segmentation and Classification of Wound Tissues through Region-Focus Training,” 2021.

[4] X.-Y. Zhou และ G.-Z. Yang, “Normalization in Training U-Net for 2-D Biomedical Semantic Segmentation,” IEEE Robotics and Automation Letters, เล่มที่ 4, pp. 1792-1799, 2019.

[5] นายพิชยุทธ บุญตน และ นายยุทธนา สีขวา, (2564). ระบบจำแนกถนนชำรุด. ค้นเมื่อ 2564. 15 กันยายน 2565, จาก http://digital.csmsu.net:8080/library/handle/123456789/137

[6] Posted bySurapong Kanoktipsatharporn. (2019). Image Segmentation คืออะไร Image Segmentation แยกส่วนภาพภาพถ่ายบนท้องถนน CamVid ด้วย Deep Learning – Image Segmentation แยกส่วนภาพ ภาพถ่ายบนท้องถนน CamVid ด้วย Deep Learning – Image Segmentation ep.1. ค้นเมื่อ 15 กันยายน 2565, จาก https:// www.bualabs.com/archives/835/what-is-image-segmentation- semantic-segmentation-camvid-machine learning-unet-deep--image-segmentation-ep-1/

[7] Nick Untitled. (2562). วิธีเทรน Segmentation โดย mmSegmentation. ค้นเมื่อ 29 กันยายน 2565, จาก https://nickuntitled.com/2022/04/10/semantic-segmentation-by-openmmlab/

[8] โครงข่ายประสาทเทียม. (2564). ค้นเมื่อ 29 กันยายน 2565, จาก https://th.wikipedia.org/wiki/%E0%B9%82%E0B8%84%E0%B8%A3%E0%B8%87%E0%B8%82%E0%B9%88%E0%B8%B2%E0%B8%A2%E0%B8%9B %E0%B8%A3%E0%B8%B0%E0%B8%AA%E0%B8%B2%E0%B8%97%E0%B9%80%E0%B8%97%E0%B8%B5%E0%B8%A2%E0%B8%A1

[9] Gurucharan M K. (2020). Machine Learning Basics: Decision Tree Regression. ค้นเมื่อ 29 กันยายน 2565, จาก https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

[10] Witchapong Daroontham. (2018). รู้จัก Decision Tree, Random Forest, และ XGBoost!!! — PART 1. ค้นเมื่อ 29 กันยายน 2565, จากhttps://medium.com/@witchapongdaroontham/%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B8%88%E0%B8%B1%E0%B8%81-decision-tree-random-forrest-%E0%B9%81%E0%B8%A5%E0%B8%B0-xgboost-part-1-cb49c4ac1315

[11] Chaiyaphop Jamjumra. (2565). รู้จักกับ Decision Tree มันคือต้นไม้อะไร ทำงานอย่างไร ? – BorntoDev เริ่มต้น เรียน เขียนโปรแกรม ขั้นเทพ !. ค้นเมื่อ 29 กันยายน 2565, จาก https://www.borntodev.com /2022/09/15/%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B8%88%E0%B8%B1%E0%B8%81%E0%B8%81%E0%B8%B1%E0%B8%9A-decision-tree/

[12] N. Kalcheva, M. Todorova and G. Marinova, "Naive Bayes Classifier, Decision Tree and Ada BoostEnsemble Algorithm–Advantages and Disadvantages," KNOWLEDGE BASED SUSTAINABLE DEVELOPMENT (2020), vol. 153, 2020.

[13] กานต์ ณัฐณ บางช้าง. (2011). การคัดเลือกตัวแปรในตัวแบบการถดถอยเชิงเส้นพหุโดยใช้วิธีการค้นหาแบบต้องห้าม.

[14] นักพัฒนา scikit-learn. (2022). sklearn.metrics.pairwise.haversine_distances. ค้นเมื่อ 29 กันยาย 2565, จาก https://scikit-learn.org/stable/ modules/ Generatedmodules/generated/Sklearn.metrics. pairwise.haversine_distances.html

[16] วิรากานต์ กิตติบวรกุล, ศรายุทธ นนท์ศิริ, พิชิตชัย คาอินทร์. (2565). วารสารวิชาการสมาคมสถาบันอุดมศึกษาเอกชนแห่งประเทศไทย (สสอท.), 11(1). 6-7.

[17] สัพพัญญู ชูแก้ว, ประสงค์ ประณีตพลกรัง, พายัพ ศิรินาม. (2566). การพัฒนาโมเดลภัยคุกคามทางไซเบอร์ในกองทัพอากาศด้วยการเรียนรู้ของเครื่อง, 11(1). 36.

[18] Sinlapasorn, Saranchai, et al. "MODELING TO PREDICT THE PATIENTS'POSTOPERATIVE WOMAC SCORE BY FEATURES ENGINEERING AND GRADIENT BOOST TREE." Suranaree Journal of Science & Technology 30.3 (2023).

ลงชื่อผู้ทำโครงงาน .....
(นางสาวดุสิตา สังข์กลิ่นหอม)
ลงชื่อผู้ทำโครงงาน .....
(นางสาวโยษิตา ศรีวุฒิทรัพย์)
วันที่ 31 มีนาคม 2566

การตรวจสอบจากอาจารย์ที่ปรึกษาโครงงาน

.....

(ลงชื่อ) .....

(.....)

วันที่...../...../.....