การหาประสิทธิภาพของเทคนิคการเรียนรู้สำหรับการจำแนกปริมาณฝุ่น โดยใช้เทคนิคเหมืองข้อมูล

Determining the Effectiveness of Learning Techniques for dust Classification using data mining Techniques

วศิน หาสอดส่อง[1*], มงคล ทะกอง[2], กริช สมกันธา[3] และ กฤษณพงศ์ สมสุข[4]

Wasin Hasodsong[1*], Mongkhon Thakong[2], Krit Somkantha[3] and Kritsanapong Somsuk[4]

*[*1,2,3,4]ภาควิชาวิทยาการข้อมูลและเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี จังหวัดอุดรธานี 41000*

*[*1,2,3,4]Department of Data Science and Information Technology, Faculty of Science, Udon Thani Rajabhat University, Udon Thani, 41000*

Received : 2024-09-16 Revised : 2024-11-12 Accepted : 2024-11-19

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองสำหรับการหาประสิทธิภาพฝุ่นละออง ในอากาศเพื่อเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับการการจำแนกฝุ่นละอองโดยใช้เทคนิคเหมืองข้อมูล 6 วิธี ซึ่งประกอบด้วย วิธีนาอีฟเบย์ (Naïve Bayes) วิธีแรนดอมฟอร์เรส (Random Forest) วิธีความใกล้เคียงกันที่สุด (K-Nearest Neighbor) และวิธีต้นไม้ตัดสินใจ (Decision Tree) โดยใช้ข้อมูลของสำนักงานสิ่งแวดล้อมและควบคุมมลพิษที่ 9 จังหวัดอุดรธานี สร้างชุดตัวแบบและชุดทดสอบตัวแบบ เป็นข้อมูลที่เกิดจากการทบทวนชุดข้อมูลคุณภาพอากาศย้อนหลัง จำนวน 724 รายการ 7 คุณลักษณะ จากนั้นทำการหาค่าความถูกต้อง โดยผลการวิจัยสามารถสรุปได้ดังนี้ วิธีแรนดอมฟอร์เรสมีประสิทธิภาพในการจำแนกมากที่สุดโดยมีค่าความถูกต้อง 98.03% วิธีต้นไม้ตัดสินใจมีค่าความถูกต้อง 91.83% วิธีนาอีฟเบย์มีค่าความถูกต้อง 88.35% วิธี K-NN เพื่อนบ้านที่ใกล้ที่สุดมีค่าความถูกต้อง 85.03% วิธี Neural Net มีค่าความถูกต้อง 91.06% และ วิธี SVM มีค่าความถูกต้อง 85.49% ตามลำดับ จะพบว่า วิธีแรนดอมฟอร์เรสมีประสิทธิภาพในการสร้างแบบจำลองมากที่สุดเมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกัน เนื่องจากเป็นวิธีที่ไม่มีการแจกแจงหรือไม่ใช้พารามิเตอร์ซึ่งไม่ได้ขึ้นอยู่กับสมมุติฐานการแจกแจง ความน่าจะเป็นเหมาะสมที่นำแบบจำลองไปใช้ในการการจำแนก เพื่อเป็นแนวในการสนับสนุนการตัดสินใจทางสิ่งแวดล้อมและควบคุมมลพิษ ในการจำแนกคุณภาพอากาศต่อไป

คำสำคัญ : การทำเหมืองข้อมูล, การจำแนกปริมาณฝุ่น, การเรียนรู้ของเครื่อง

Abstract

This research aims to create a model for dust efficiency in the air to compare the efficiency of models for dust forecasting using 4 data mining techniques, consisting of Naïve Bayes, Random Forest, K-Nearest Neighbor, and Decision Tree. The data from the Office of Natural Resources and Environmental Control 9, Udon Thani Province, were used to create a model set and a model test set, which were data from a review of a retrospective air quality data set of 724 items with 7 characteristics. The accuracy was then calculated by The research results can be summarized as follows: Random Forest Method The most efficient in forecasting with an accuracy of 98.03%, the decision tree method has an accuracy of 91.83%, the Naif Bayes method has an accuracy of 88.35%, the nearest neighbor K-NN method has

*วศิน หาสอดส่อง

E-mail address : 65120604107@udru.ac.th

an accuracy of 85.03%, the Neural Net method has an accuracy of 91.06%, and the SVM method has an accuracy of85.49%,respectively. It was found that the random forest method has the most efficient in modeling compared to the combined comparison methods because it is a non-distribution or non-parametric method that does not depend on the probability distribution assumption. It is suitable to use the model for forecasting as a guideline to support environmental decision-making and pollution control in air quality forecasting.

Keywords : Data, mining, Dust Classification

1.บทนำ

ในปัจจุบันปัญหามลพิษทางอากาศ ในประเทศไทยมีปริมาณเกินเกณฑ์มาตรฐานในหลายพื้นที่ ส่งผลให้คุณภาพอากาศอยู่ในระดับปานกลางถึงมีผลกระทบต่อสุขภาพ ประเทศไทยถูกจัดอยู่ในลำดับต้น ๆ ของเมืองที่มีคุณภาพอากาศแย่ที่สุดในโลก โดยเฉพาะใน 9 จังหวัดภาคเหนือ เช่น เชียงรายและเชียงใหม่ ซึ่งสาเหตุหลักมาจากการเผาในพื้นที่ป่าและการเผาวัสดุทางการเกษตร ส่วนจังหวัดสระบุรีมีปัญหาจากกิจการโรงโม่บดย่อยหินและโรงงานต่าง ๆ ขณะที่กรุงเทพมหานครและปริมณฑลมีปัญหาจากรถยนต์ดีเซลและการจราจร รัฐบาลได้ประกาศให้สถานการณ์นี้เป็นสภาวะฉุกเฉินในปี 2563 และได้มีมาตรการควบคุมต่าง ๆ เช่น การออกกฎหมายเกี่ยวกับการจราจรและลดการเผา เพื่อแก้ไขปัญหานี้ เนื่องจากมีผลกระทบต่อเศรษฐกิจ สุขภาพอนามัย และสิ่งแวดล้อม [1] จึงจำเป็นต้องพัฒนาควบคู่กับการควบคุมปริมาณฝุ่น เพื่อไม่ให้กระทบต่อสุขภาพประชาชนในอนาคต

จังหวัดอุดรธานีก็เป็นหนึ่งจังหวัดที่ประสบปัญหามลพิษทางอากาศในระดับปานกลางและมีแนวโน้มที่จะเกิดความเสี่ยง [2] ซึ่งเป็นจังหวัดที่ตั้งอยู่ตอนบนของภาคตะวันออก เฉียงเหนือของประเทศไทย จังหวัดอุดรธานีเป็นศูนย์ปฏิบัติการกลุ่มจังหวัดภาคตะวันออกเฉียงเหนือตอนบน 1 และศูนย์กลางพื้นที่ลุ่มแม่น้ำโขง เดิมพื้นที่เมืองอุดรธานีในปัจจุบันคือบ้านเดื่อหมากแข้ง ซึ่งมีมานานร่วมสมัยกับเมืองเวียงจันทน์ อาณาจักรล้านช้าง เมืองอุดรธานี ปัญหาดังกล่าวส่งผลกระทบต่อคุณภาพชีวิตของประชาชนในพื้นที่ และการพัฒนาเศรษฐกิจของประเทศ การเฝ้าระวังและการจำแนกค่าฝุ่นละออง จึงเป็นเรื่องสำคัญอย่างยิ่ง เพื่อที่จะสามารถจัดการ และป้องกันปัญหาดังกล่าวได้อย่างมีประสิทธิภาพ การเรียนรู้ของเครื่อง (Machine Learning) [7] ได้รับการพิสูจน์แล้วว่าเป็นเครื่องมือที่มีประสิทธิภาพในการวิเคราะห์และการจำแนกข้อมูลต่าง ๆ รวมถึงข้อมูลมลพิษทางอากาศ โปรแกรม RapidMiner เป็นหนึ่งในเครื่องมือที่ช่วยให้การทำการเรียนรู้ของเครื่องเป็นไปได้อย่างง่ายดายและมีประสิทธิภาพ ด้วยการประมวลผลข้อมูลขนาดใหญ่และการสร้างแบบจำลองที่ซับซ้อน

ดังนั้น ผู้วิจัยได้สนใจศึกษาการสร้างตัวแบบและเปรียบเทียบประสิทธิภาพของตัวแบบเพื่อการจำแนกฝุ่นละอองในอากาศ โดยใช้เทคนิคการประมาณค่า 6 เทคนิค มาทำการสร้างตัวแบบการจำแนกที่ได้มาจากการศึกษางานวิจัยที่เกี่ยวข้องกับงานวิจัยในครั้งนี้ เพื่อหาตัวแบบที่เหมาะสมในการนำไปใช้สร้างตัวแบบการจำแนกฝุ่นและให้ได้ผลที่แม่นยำและสามารถนำไปใช้ประโยชน์ในการจัดการมลพิษทางอากาศในอนาคต

2.ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

2.1งานวิจัยที่เกี่ยวข้อง

หัวข้อนี้จะกล่าวถึงการทบทวนวรรณกรรมที่เกี่ยวข้องกับการวิจัยที่เกี่ยวข้องกับประสิทธิภาพของฝุ่นละออง งานวิจัยเหล่านี้ช่วยให้เข้าใจถึงข้อดีและข้อเสียของอัลกอริทึมต่าง ๆ ในการวิเคราะห์ข้อมูล

กฤติกา ทิพย์คำมี และคณะ นำเสนอเรื่อง “ประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่องเพื่อการจำแนกอนุภาคฝุ่นละเอียดในอากาศ” (Neural Network) โดยมีค่าความผิดพลาดกำลังสองเฉลี่ยน้อยที่สุด (MSE) 8.673, ค่าความผิดพลาดรูทสแควร์ (RMSE) 2.945, ค่าความผิดพลาดสัมบูรณ์ (AE) 2.126, กำลังสอง ค่าความผิดพลาด (SE) เท่ากับ 8.709 และค่าสัมประสิทธิ์การตัดสินใจพหุคูณ (R2) เท่ากับ 0.948 ประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่องจักรในการการจำแนกอนุภาคฝุ่นในอากาศ งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่องจักรในการทำนายอนุภาคในอากาศ (PM2.5) พร้อมตัวแปรที่ใช้ในการสร้างแบบจำลองฝุ่น นอกจากนี้ยังสามารถพัฒนาต่อไป

สู่ระบบสารสนเทศเพื่อทำนายฝุ่นละอองขนาดเล็กที่จะเกิดขึ้นในอนาคตได้อีกด้วย [1]

คมกฤต เมฆสกุล และคณะ ได้เสนอเรื่อง “แบบจำลองการการจำแนกปริมาณฝุ่นละออง PM2.5 ของจังหวัดพะเยา” กรณีศึกษาคือ ฝุ่นละอองในบรรยากาศขนาดเล็กไม่เกิน 2.5 ไมครอน (PM2.5) แบบจำลองการจำแนกปริมาณฝุ่น PM2.5 ในจังหวัดพะเยา การสร้างแบบจำลองการการจำแนกโดยใช้การวิเคราะห์การถดถอยโลจิสติกโดยใช้ชุดข้อมูลจากกรมควบคุมมลพิษทางอากาศ การจำแนกระดับความเข้มข้นของฝุ่นละอองในบรรยากาศขนาดเล็กไม่เกิน 2.5 ไมครอน โดยให้ค่าความไวและค่าความจำเพาะเท่ากับ 0.96 และ 0.72 ตามลำดับ โดยมีค่าพื้นที่ใต้เส้นโค้งเท่ากับ 0.98 [3]

ธรณินทร์ สัจวิริยทรัพย์ ได้เสนอเรื่อง ตัวแบบผสม Arima และโมเดล Support Vector Machine เน้นองค์ประกอบเชิงบวกสำหรับการการจำแนกปริมาณฝุ่นละอองขนาดไม่เกิน 10 ไมครอน”ด้วยความสัมพันธ์เชิงเส้นของแบบจำลอง Arima เพื่ออธิบายองค์ประกอบเชิงเส้นของข้อมูลอนุกรมเวลา ความสัมพันธ์เชิงเส้นหรือไม่เชิงเส้นอย่างชัดเจน การเพิ่มอนุกรมเวลา โมเดลไฮบริดที่พัฒนาแล้วได้รับการเปรียบเทียบประสิทธิภาพการการจำแนกกับโมเดลการการจำแนกเดี่ยว Arima และโมเดล Support Vector Machine โดยใช้เกณฑ์การประเมินประสิทธิภาพการการจำแนกทั้งหมด 6 รายการ ดังนั้นตัวแบบผสมระหว่างตัวแบบ Arima และโมเดล Support Vector Machine จึงเป็นตัวแบบที่มีความเหมาะสมที่จะใช้สำหรับการการจำแนกปริมาณฝุ่นละอองขนาดเล็กในเขตพื้นที่ตำบลหน้าพระลาน เพื่อใช้เป็นสารสนเทศที่เป็นประโยชน์ต่อการเฝ้าระวัง และวางแผนเพื่อลดผลกระทบจากมลพิษทางฝุ่นละอองลงอย่างมีประสิทธิภาพซึ่งไม่เพียงแต่เป็นงานที่ท้าทายแต่ยังเป็นปัญหาสาธารณะที่สำคัญสำหรับสุขภาพของประชาชนอีกด้วย รวมทั้งยังเป็นตัวแบบที่อาจมีประโยชน์ในการประยุกต์ใช้สำหรับการการจำแนกปริมาณฝุ่นละอองในเขตพื้นที่อื่น ๆ ได้ [5]

2.2การทำ (K-Fold Cross Validation)

เป็นขั้นตอนสำหรับการประเมินแบบจำลอง การกำหนด Baseline ในการประเมินแบบจำลอง และเทคนิคการสุ่มตัวอย่างสำหรับการแบ่งข้อมูลออกเป็น Training และ Test Sets ด้วยการใช้วิธี K-Fold Cross Validationที่ใช้ทดสอบกับ Final Model โดยทดสอบ 10 รอบ เพื่อนำมาหาประสิทธิภาพของเทคนิคการเรียนรู้ ที่มีค่าความถูกต้องสูงสุด [6]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Precision = \frac{TP}{TP + TN} \quad (1)$$
$$Recall = \frac{TP}{TP + FN}$$
$$F - Measure = \frac{Precision \times Recall}{Precision + Recall}$$

FP = อัตราความถูกต้องเชิงบวก

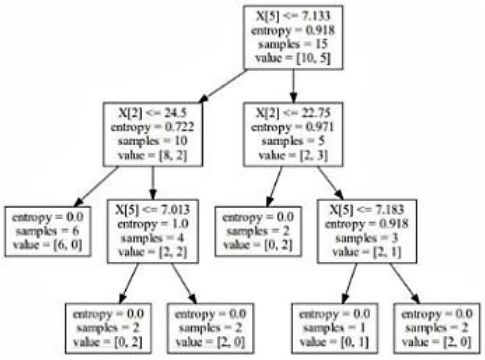
FN = อัตราความถูกต้องเชิงลบ

TP = อัตราความผิดพลาดเชิงบวก

TN = อัตราความผิดพลาดเชิงลบ

2.3การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือ การวิเคราะห์ข้อมูลเพื่อหารูปแบบ (Patterns) หรือความสัมพันธ์ของข้อมูลสารสนเทศซึ่งปัจจุบันมีการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลมาใช้ในด้านการหาคุณภาพ เพื่อหาสาเหตุหรือปัจจัยที่มีความสัมพันธ์ต่อความเสี่ยงต่าง ๆ เช่น การจำแนกฝุ่นละอองในสภาพอากาศ เพื่อนำข้อมูลมาใช้ในการบริหารจัดการและควบคุมดูแล และการวางแผน เพื่อบริหารจัดการดูแล



รูปที่ 1 ขั้นตอนการทำงานของวิธีต้นไม้ตัดสินใจ

*2.3.1*วิธีต้นไม้ตัดสินใจ (Decision Tree)

เป็นการจำแนกค่าคุณลักษณะของข้อมูล โดยจะประกอบด้วยบัพ (Node) และกิ่ง (Link) ที่ต่อกับบัพ บัพที่ปลายสุดจะเรียกว่าบัพใบ (Leaf) ต้นไม้ตัดสินใจจะทำโดยสร้างบัพทีละบัพ เพื่อตรวจสอบคุณสมบัติของตัวอย่าง แล้วแยกตัวอย่างลงตาม

ค่าของกิ่ง ทำจนกระทั่งตัวอย่างในใบแต่ละใบอยู่ในประเภทเดียวกันทั้งหมด [6] แสดงดังรูปที่ 2

Algorithm: Naïve-Bayes

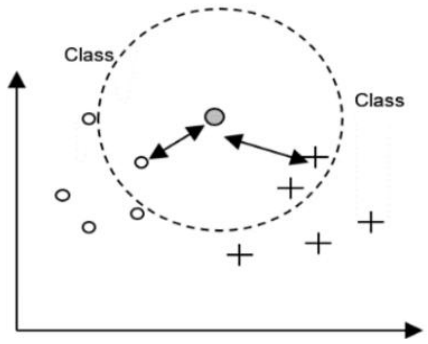
- Naive_Bayes_Learn(examples)
  FOR EACH target value v DO
  $\overline{P}(v_j) \leftarrow \text{estimate } P(v_j)$
  FOR EACH attribute value a of each attribute DO
  $\overline{P}(a_i \mid v_j) \leftarrow \text{estimate } P(a_i \mid v_j)$
- Classify_New_Example(x)

$$v_{NB} = \arg\max_{v_j \in V} \overline{P}(v_j) \times \prod_{i=1}^{n} \overline{P}(a_i \mid v_j)$$

รูปที่ 2 ขั้นตอนการทำงานของวิธีนาอีฟเบย์

*2.3.2*วิธี (Naïve Bayes)

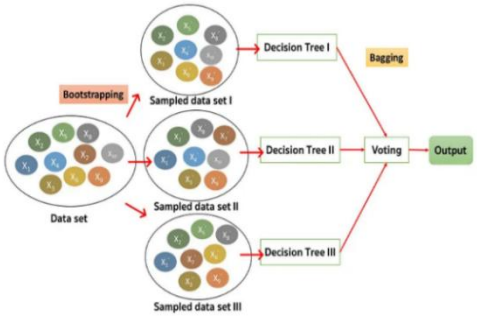
เป็นตัวจำแนกที่เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมาก และคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน มีการนำตัวจำแนกประเภทเบย์ไปประยุกต์ใช้งานด้านการจำแนกประเภท ข้อความ (Text Classification) [6] และพบว่า ใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีอื่น ๆ มีขั้นตอนการทำงาน แสดงดังรูปที่ 3



รูปที่ 3 ขั้นตอนการทำงานของวิธีที่ใกล้เคียงกันที่ส (K-Nearest Neighbor)

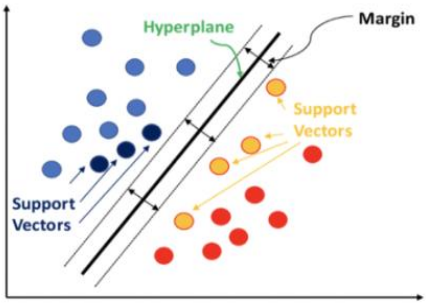
*2.3.3*วิธีความใกล้เคียงกันที่สุด

(K-Nearest Neighbor)เป็นวิธีการหนึ่งสำหรับแก้ปัญหาการประมาณค่าที่ไม่ใช้พารามิเตอร์สำหรับการจำแนกกลุ่ม ซึ่งเหมาะสมกับข้อมูลที่เป็นรูปร่างที่ดี หลักการคือวัดค่าของข้อมูลที่ใกล้ที่สุดเพื่อหาจุดได้เท่ากับ K ตัวที่ต้องการ เช่น k = 10 ก็จะขยายเคอร์เนล ไปเรื่อย ๆ จนกว่าจะเจอข้อมูลที่เป็น k = 10 แสดงดังรูปที่ 4



รูปที่ 4 ขั้นตอนการทำงานของแรนดอมฟอร์เรส

*2.3.4*หลักการทำงานของ (Random Forest) คือ สร้าง model จาก Decision Tree หลาย ๆ model ย่อย ๆ (ตั้งแต่ 10 model ถึง มากกว่า 1000 model) โดยแต่ละ model จะได้รับ data set ไม่เหมือนกัน ซึ่งเป็น subset ของ data set ทั้งหมด ตอนทำ prediction ก็ให้แต่ละ Decision Tree ทำ prediction sample ข้อมูล (bootstrapping) จาก data set ทั้งหมด ให้ได้ข้อมูลออกมา n ชุด ที่ไม่เหมือนกัน ตามจำนวน Decision Tree แสดงดังรูปที่ 5

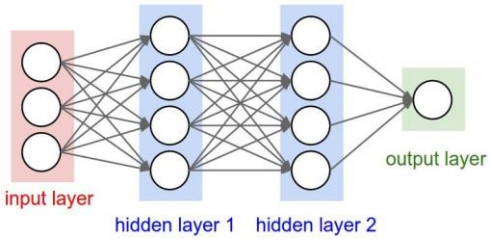


รูปที่ 5 ขั้นตอนการทำงานของ SVM

*2.3.5*ลักษณะขยายจาก Support Vector Machine (SVM) โดยใช้ Kernel trick เพื่อจัดการข้อมูลที่ไม่เชิงเส้น เหมาะสำหรับข้อมูลที่มีความซับซ้อนและต้องการโมเดลที่มีความยืดหยุ่นสูงในการปรับเส้นคาดเดา ฟังก์ชันการทำนาย [1] (Prediction Function) สมการหลักของ SVR มีรูปแบบดังนี้

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b \qquad (2)$$

$\mathbf{w}$ คือเวกเตอร์น้ำหนัก (Weight vector)
$\mathbf{x}$ คือตัวแปรต้น (Feature vector)
b คือค่าไบแอส (Bias)



รูปที่ 6 ขั้นตอนการทำงานของ Neural Net

*2.3.6*Neural Network Regression

ใช้ Neural Networks เพื่อจับความสัมพันธ์ที่ซับซ้อนในข้อมูลหลายมิติหรือที่มีรูปแบบซ่อนเร้น โดยการทำนายค่าจะส่งข้อมูลผ่านโครงข่ายนิวรอนหลายชั้น เพื่อสร้างฟังก์ชันที่เชื่อมโยงตัวแปรต้นและตัวแปรตาม สมการประกอบด้วยน้ำหนัก, การรวมเชิงเส้น, ฟังก์ชันกระตุ้น และกระบวนการทำนาย ฟังก์ชันพื้นฐานของนิวรอนเดี่ยว (Neuron) นิวรอนหนึ่งตัวใน Neural Network คำนวณผลลัพธ์จากการรวมเชิงเส้นของค่าอินพุตตามสมการ [3]

$$z = \mathbf{w} \cdot \mathbf{x} + b$$

$\mathbf{w}$ คือ เวกเตอร์น้ำหนัก (Weight vector) ที่เชื่อมโยงระหว่างอินพุตแต่ละตัวกับนิวรอนนั้น

$\mathbf{x}$ คือ ตัวแปรอินพุต (Input vector)

b คือ ค่าคงที่ที่เรียกว่า ไบแอส (Bias)

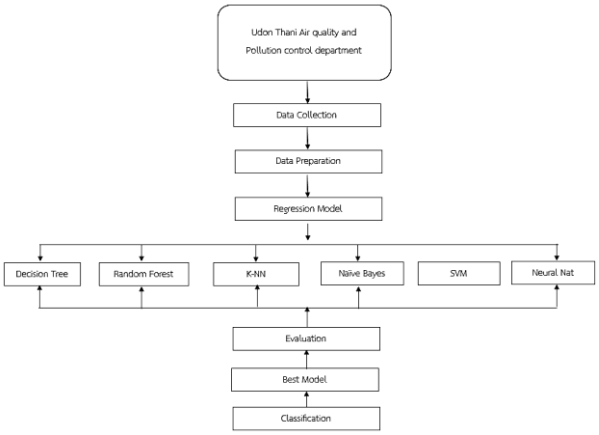
z คือ ค่าผลลัพธ์ที่ได้จากการรวมเชิงเส้นของ อินพุต

2.4การสร้างแบบจำลอง

ในการสร้างแบบจำลองงานวิจัยนี้ผู้วิจัยได้ใช้เครื่องมือวิเคราะห์ข้อมูล คือ โปรแกรม RapidMiner Studio และเทคนิควิธีการหาประสิทธิภาพของข้อมูลที่ใช้ประกอบด้วยวิธีนาอีฟเบย์ (Naïve Bayes) วิธีแรนดอมฟอร์เรส (Random Forest) วิธีความใกล้เคียงกันที่สุด (K-Nearest Neighbor) และวิธีต้นไม้ตัดสินใจ (Decision Tree) และการสร้างแบบจำลอง แสดงดังรูปที่ 8

3.วิธีดำเนินการวิจัย

ผู้วิจัยได้ศึกษาเกี่ยวกับเทคนิคการเรียนรู้ของเครื่องสำหรับนำมาจำแนกฝุ่นละอองขนาดเล็กในอากาศโดยผ่านกระบวนการการวิเคราะห์ข้อมูล (Data Analysis Process) โดยการแบ่งข้อมูลออกเป็น 2 ส่วนเพื่อใช้สำหรับการสร้างตัวแบบ (Training Set) และการทดสอบ ประสิทธิภาพของตัวแบบ (Testing Set) ด้วยเทคนิค Cross Validation โดยผู้วิจัยได้ทำการแบ่ง ข้อมูลออกเป็น 10 ส่วนเท่า ๆ กัน (10-fold cross validation) [6] ซึ่งเทคนิคที่ใช้สำหรับการสร้าง ตัวแบบการจำแนกข้อมูล ผู้วิจัยได้ใช้เทคนิคในการประมาณค่าข้อมูลทั้งหมด 6 เทคนิค ได้แก่ วิธีนาอีฟเบย์ (Naïve Bayes) วิธีแรนดอมฟอร์เรส (Random Forest) วิธีความใกล้เคียงกันที่สุด (K-Nearest Neighbor) และวิธีต้นไม้ ตัดสินใจ (Decision Tree) วิธี Neural Net และวิธี SVM จากนั้นทำการทดสอบ ประสิทธิภาพเพื่อหาตัวแบบที่ มีความเหมาะสมสำหรับการนำไปใช้ในการจำแนกฝุ่นละอองขนาดเล็กดังภาพที่ 7



รูปที่ 7 ขั้นตอนการดำเนินงาน [1]

3.1การเตรียมข้อมูล

ผู้วิจัยได้ดำเนินการเก็บและรวบรวมข้อมูล [9] ที่ได้จากสถานีตรวจวัดคุณภาพอากาศ พ.ศ. 2565–2566 จำนวน 724 เรคคอร์ด โดยดำเนินการวัดคุณภาพอากาศ รวม 1 จุดตรวจวัด ได้แก่ สวนสาธารณะหนองประจักษ์ศิลปาคม จังหวัดอุดรธานี ในพื้นที่รับผิดชอบของสำนักงานสิ่งแวดล้อมและควบคุมมลพิษที่ 9 จังหวัดอุดรธานี

1)ข้อมูลฝุ่นละอองขนาดไม่เกิน 2.5 ไมครอน

2)ข้อมูลฝุ่นละอองขนาดไม่เกิน 10 ไมครอน

3)ข้อมูลก๊าซโอโซน (O_3)

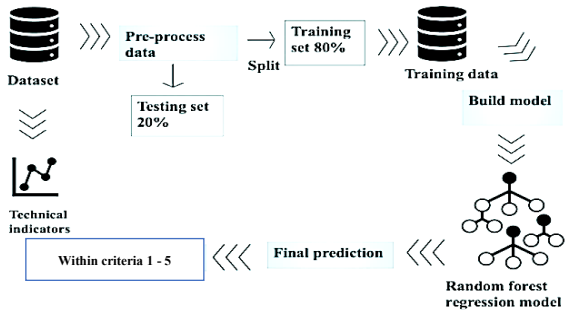
4)ข้อมูลก๊าซคาร์บอนมอนอกไซด์ (CO)

5)ก๊าซไนโตรเจนไดออกไซด์ (NO_2)

6)ก๊าซซัลเฟอร์ไดออกไซด์ (SO_2)

7)ค่าดัชนีคุณภาพอากาศ (AQI)

โดยข้อมูลได้จากสำนักงานสิ่งแวดล้อมและควบคุมมลพิษที่ 9 จังหวัดอุดรธานี ช่วงปี พ.ศ. 2565 ถึง ปี พ.ศ. 2566 มีจำนวน 724 เรคคอร์ด รายการ 7 คุณลักษณะตาม พรบ. มาตราฐานฝุ่น 2566 [2] โดยดำเนินการวัดคุณภาพอากาศรวม 1 จุด ตรวจวัด ได้แก่ สวนสาธารณะหนองประจักษ์ศิลปาคม



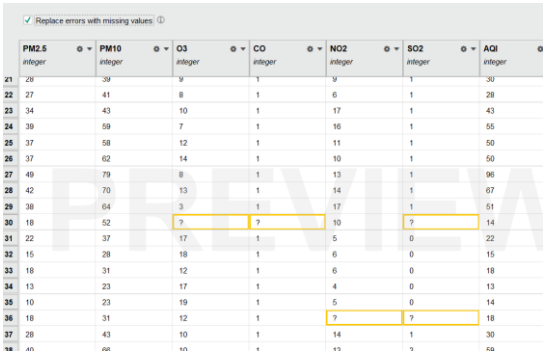
รูปที่ 8 ขั้นตอนการสร้างแบบจำลอง

ตารางที่ 1 คุณลักษณะสำคัญของข้อมูลฝุ่นละออง

ลำดับที่	คุณลักษณะ	คำอธิบาย
1	PM2.5	ค่าปริมาณฝุ่น
2	PM10	ค่าปริมาณฝุ่น
3	O3	เป็นก๊าซที่ไม่มีสีหรือมีสีฟ้าอ่อน
4	CO	เป็นก๊าซที่ไม่มีสี กลิ่น และรส
5	NO2	เป็นก๊าซที่ไม่มีสีและกลิ่น ละลายน้ำได้เล็กน้อย
6	SO2	เป็นก๊าซที่ไม่มีสี หรืออาจมีสีเหลืองอ่อน ๆ
7	AQI	ค่าดัชนีคุณภาพอากาศ

3.2ทำความสะอาดข้อมูล (Data Cleansing)

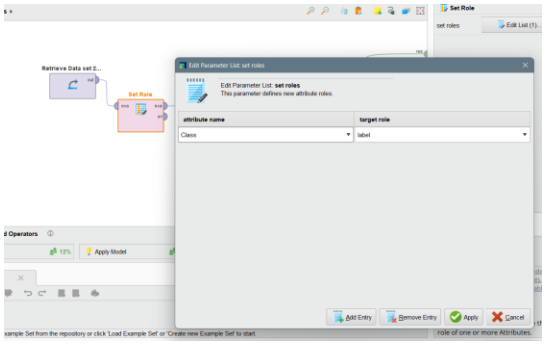
ใช้เครื่องมือโปรแกรม RapidMiner Studio ในการ Cleansing โดยใช้ Tools Replace errors with missing values แทนที่ข้อมูลที่มีข้อผิดพลาดด้วยค่าที่ขาดหายไปใน Data set



รูปที่ 9 ขั้นตอน Replace errors with missing values

3.3เลือกข้อมูล (Select data)

จากข้อมูลที่ได้นำมาจัดเก็บให้อยู่ในรูปแบบของไฟล์ .CSV จากนั้นนำมา Set Role กำหนด Class เพื่อนำมาใช้ในการทดสอบและวิเคราะห์ด้วยเครื่องมือการเรียนรู้ได้ให้ได้ข้อมูลที่มีความถูกต้อง

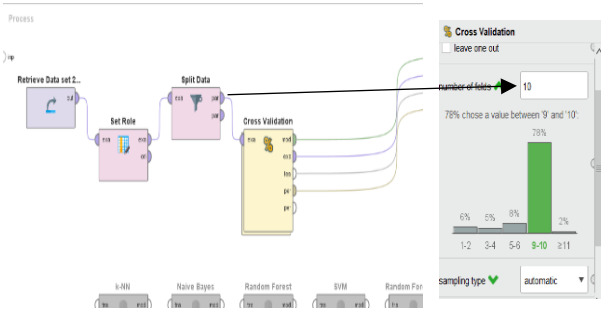


รูปที่ 10 ขั้นตอน Edit Parameter List: Set Roles

3.4วิเคราะห์โดย K-Fold Cross Validation

เป็นขั้นตอนสำหรับการประเมินแบบจำลองการกำหนด Baseline ในการประเมินแบบจำลอง และเทคนิคการสุ่มตัวอย่างสำหรับการแบ่งข้อมูลออกเป็น Training และ Test Sets ด้วยการใช้วิธี K-Fold Cross Validationที่ใช้ทดสอบกับ Final Model โดยทดสอบ 10 รอบ และแต่ละรอบ จะเก็บค่าเฉลี่ยไว้ ทำแบบนี้ไปเรื่อย ๆ จนกว่าจะครบ จากนั้นนำ

ค่าเฉลี่ยในแต่ละรอบมาหาค่าเฉลี่ยทั้งหมด ก็จะได้ค่าความถูกต้องในการทำนายของแบบจำลองแต่ละเทคนิค แสดงดังตารางที่ 2 และการเปรียบเทียบค่าความถูกต้องแบบจำลองของสภาพอากาศ แสดงดังตารางที่ 3



รูปที่ 11 ขั้นตอน K-Fold Cross Validation

4.ผลการทดลอง

ผลการทดลองงานวิจัยฉบับนี้ได้วิเคราะห์ถึงค่าความถูกต้องของแบบจำลองการจำแนกคุณภาพอากาศด้วยเทคนิคเหมืองข้อมูลสามารถแสดงค่าความถูกต้องของการการจำแนกได้ดังตารางที่ 2

ตารางที่ 2 ผลการวัดประสิทธิภาพของโมเดลด้วย 10-Fold Cross Validation ในแต่ละรอบ

Number	Decision Tree	K-NN	Naïve Bayes	Random Forest	SVM	Neural Net
1	0.9179	0.7411	0.7911	0.9607	0.8439	0.8909
2	0.9641	0.8439	0.8798	0.9820	0.8508	0.90.74
3	0.9765	0.8481	0.8895	0.9793	0.8494	0.9171
4	0.9848	0.8688	0.8964	0.9820	0.8619	0.9088
5	0.9820	0.8618	0.9864	0.9807	0.8537	0.9171
6	0.9793	0.8633	0.8867	0.9848	0.8578	0.9144
7	0.9848	0.8702	0.9005	0.9848	0.8576	0.9227
8	0.9862	0.8757	0.8964	0.9820	0.8577	0.8965
9	0.9807	0.8604	0.8950	0.9834	0.8619	0.9254
10	0.9834	0.8700	0.9032	0.9834	0.8549	0.9060
Accuracy	**0.9739**	**0.8503**	**0.8835**	**0.9803**	**0.8549**	**0.9106**

จากตารางที่ 2 จะเห็นได้ว่าผลการทดสอบประสิทธิภาพของโมเดล แต่ละส่วนจะมีค่าความถูกต้องในแต่ละรอบ เมื่อ ทดสอบครบจำนวน 10 รอบ (Fold) จึงจะเฉลี่ยผลค่าความถูกต้องของแบบจำลองแต่ละวิธี

ตารางที่ 3 การเปรียบเทียบประสิทธิภาพการจำแนกของแบบจำลอง

	Precision	Recall	Accuracy	F1_Score
Decision Tree	92.31%	91.35%	97.39%	91.83%
K-NN	82.76%	79.50%	85.03%	81.10%
Naïve Bayes	80.45%	80.34%	88.35%	80.39%
Random Forest	94.17%	92.87%	98.03%	93.52%
SVM	69.27%	82.57%	85.49%	75.34%
Neural Net	57.35%	73.59%	91.06%	64.46%

จากตารางที่ 3 เมื่อพิจารณาจากค่าความถูกต้อง พบว่าวิธี แรนดอมฟอร์เรส มีประสิทธิภาพในการการจำแนกมากที่สุด โดยมีค่า Accuracy 98.03%, Precision 94.17% , Recall 92.87% และ F1_Score 93.52% วิธีต้นไม้ตัดสินใจ มีค่า Accuracy 91.83%, Precision, 92.31%, Recall 91.35% และ F1_Score 91.83% ตามลำดับ จากผลการวิจัย ดังตารางที่ 2 พบว่า วิธีแรนดอมฟอร์เรส มีประสิทธิภาพในการสร้างแบบจำลองมากที่สุดเมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกัน

5.อภิปรายการวิจัย

งานวิจัยนี้มุ่งเน้นการสร้างแบบจำลองสำหรับการจำแนกข้อมูล โดยใช้เทคนิคเหมืองข้อมูลหลายวิธีเพื่อเปรียบเทียบความแม่นยำในการจำแนกและหาวิธีที่เหมาะสมที่สุดในการนำไปใช้ควบคุมมลพิษทางอากาศ โดยจากการทดลองพบว่าวิธีแรนดอมฟอร์เรส (Random Forest) มีค่าความถูกต้องสูงสุด ซึ่งบ่งชี้ว่ามีประสิทธิภาพในการจำแนกข้อมูลได้ดีที่สุดเมื่อเทียบกับเทคนิคอื่น ๆ อย่างชัดเจน เช่น วิธีต้นไม้ตัดสินใจ (Decision Tree) และวิธีอื่น ๆ ที่มีความถูกต้องลดหลั่นกันไป ได้แก่ Naive Bayes, K-Nearest Neighbor (K-NN), Neural Network และ Support Vector Machine (SVM) การใช้วิธี 10-Fold Cross Validation นับว่าเหมาะสมในการประเมินประสิทธิภาพของแบบจำลอง เนื่องจากช่วยลดการเกิด overfitting และทำให้ค่าความแม่นยำที่ได้จากแต่ละวิธีมีความน่าเชื่อถือเพิ่มขึ้น การที่แรนดอมฟอเรสมีประสิทธิภาพสูงสุด

นั้นเป็นผลมาจากคุณสมบัติที่ไม่ขึ้นกับการแจกแจงของข้อมูลและความสามารถในการจัดการกับข้อมูลที่มีมิติสูง ซึ่งทำให้แบบจำลองนี้เหมาะสมต่อการใช้งานกับข้อมูลที่มีความซับซ้อนอย่างการจำแนกฝุ่นละอองที่มีคุณลักษณะหลากหลาย ทั้งนี้ปัจจัยที่เกี่ยวข้องกับคุณลักษณะความเสี่ยงที่นำไปใช้ในสร้างแบบจำลองมีการใช้ตัวแปรที่มีความสัมพันธ์ในการเกิดมลพิษทางอากาศ เช่น O_3, CO, NO_2 และ SO_2 [2] จากการทบทวนวรรณกรรม พบว่า ปัจจัยเหล่านี้ไม่ได้ถูกนำมาเพื่อประมวลผลในเครื่องจักรการเรียนรู้แต่ผู้เชี่ยวชาญยืนยันว่ามีผลกระทบทำให้เกิดมลพิษได้ ยกตัวอย่าง เช่น ตัวแปร O_3 หากมีค่ามากกว่า 50 หมายความว่าค่าดัชนีคุณภาพอากาศ AQI จะอยู่ในระดับเสี่ยงปานกลางและตัวแปรอื่น ๆ ข้างต้น ที่ได้กล่าวมาไม่ได้มีน้ำหนักมากเท่าตัวแปร PM2.5 หรือ PM10 ซึ่งส่วนใหญ่จะใช้ตัวแปรนี้ในการทำนาย ผู้วิจัยจึงได้นำผลลัพธ์ที่ได้จากแบบจำลองของต้นไม้ตัดสินใจที่สร้างจากแรนดอมฟอร์เรสไปใช้ในการการจำแนก [8] เพื่อเป็นแนวในการสนับสนุนการตัดสินใจทางสิ่งแวดล้อมและควบคุมมลพิษในการจำแนกคุณภาพอากาศและได้นำผลการวิจัยมาใช้ในด้านการป้องกันและพัฒนาคุณภาพสังคมจะช่วยให้สังคมมีความพร้อมในการรับมือกับมลพิษ [11] มีระบบแจ้งเตือนและมาตรการป้องกันที่เหมาะสม อีกทั้งยังช่วยส่งเสริมให้ประชาชนมีความรู้ความเข้าใจเกี่ยวกับปัญหามลพิษ เพื่อให้เกิดการดูแลสิ่งแวดล้อมร่วมกัน

6.สรุปผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองการการจำแนก โดยใช้เทคนิคเหมืองข้อมูลและเปรียบเทียบประสิทธิภาพของการหาค่าความถูกต้อง (Accuracy) ด้วยเทคนิคเหมืองข้อมูล 6 วิธี โดยใช้วิธี 10–Fold Cross Validation เพื่อหาแบบจำลองที่เหมาะสมที่สุดในการจำแนก

จากผลการวิจัยสามารถสรุปได้ดังนี้ วิธีแรนดอมฟอร์เรสมีประสิทธิภาพในการการจำแนกมากที่สุดโดยมีค่าความถูกต้อง 98.03% วิธีต้นไม้ตัดสินใจมีค่าความถูกต้อง 91.83% วิธีนาอีฟเบย์มีค่าความถูกต้อง 88.35% วิธี K-NN เพื่อนบ้านที่ใกล้ที่สุดมีค่าความถูกต้อง 85.03% วิธี Neural Net มีค่าความถูกต้อง 91.06% และ วิธี SVM มีค่าความถูกต้อง 85.49% ตามลำดับ จะพบว่า วิธีแรนดอมฟอร์เรสมีประสิทธิภาพในการสร้างแบบจำลองมากที่สุดเมื่อเทียบกับวิธีที่ใช้เปรียบเทียบร่วมกันเนื่องจากเป็นวิธีที่ไม่มีการแจกแจงหรือไม่ใช้พารามิเตอร์ซึ่งไม่ได้ขึ้นอยู่กับสมมุติฐานการแจกแจงความน่าจะเป็นอีกทั้งสามารถจัดการกับข้อมูลที่มีมิติสูงได้อย่างแม่นยำเหมาะสมที่จะนำแบบจำลองไปทำการการจำแนกฝุ่นละออง

เอกสารอ้างอิง

[1] k. Kritika Thipphakhammee, “Article on the efficiency of machine learning techniques for forecasting fine dust particles in air”, Kalasin University Journal of Science Technology and Innovation,58th ed., No. 1, 2023, pp. 6 - 16. (in thai)

[2] P. Pollution Control Department, Announcement of the National Environment Board on setting air quality index standards for Thailand, 2023, [Online]. https: //www.aqi.in/th/dashboard/thailand/udon- thani/udon-thani. (Accessed 28 January 2024). (in thai)

[3] K. Khomkrit Meksakul, “Article on PM2.5 dust forecasting model of Phayao Province”, 11th ed., Issue 2, *Journal of Engineering and Digital Technology (JEDT)*, 2023, pp. 3 - 11. (in thai)

[4] J. Jintana Prachumphan, “*PM2.5, fine dust particles in the air and the health crisis that Thai people will have to exchange”*, [Online].https:// thestandard.com. (Accessed : 20 January 2023). (in thai)

[5] T. Thoranin Satchaviriyatra, Article on the hybrid model Arima-Support Vector Machine emphasizing positive components for the forecasting of PM2.5 particle size less than 10 microns, 27th ed., No. 3, *Science and Technology*, 2019, pp. 5 - 13. (in thai)

[6] N. Nopparat Nonsiri, Article on Data Forecasting for Diagnosing Disease Risk 2023, 33rd ed., No. 2, pages 543 – 545.

[7] N. Naruchon Rojanaburanon, “Data analysis guide with Rapid Miner Studio”, [Online] https://www.gotoknow.org/posts/660127. (Accessed : 15 March 2024).

[8] S. Sanya Pantipitham, “An article on the application of time series analysis and geographic information system to assess particulate matter less than 10 microns in Bangkok 2016”, 44th ed., No. 4, pp. 8 – 20. (in thai)

[9] S. Sutharat Muenmee and Susira Butdee, Article on Health Risk Assessment from PM Dust Exposure, 2022, 14th ed., No. 3, pp. 95 – 108. . (in thai)

[10] S. Samitivej, “Understanding and dust”, [Online]. https://samitivejchinatown.com /th/article /health/what -is-pm-2.5. (Accessed : 28 January 2024).

[11] I. Inthachat Sukkasem, Article on the forecast of particulate matter not exceeding 2.5 microns (PM2.5) in Nakhon Ratchasima Province 2020, 9th printing, issue 21, pages 8-10. (in thai)