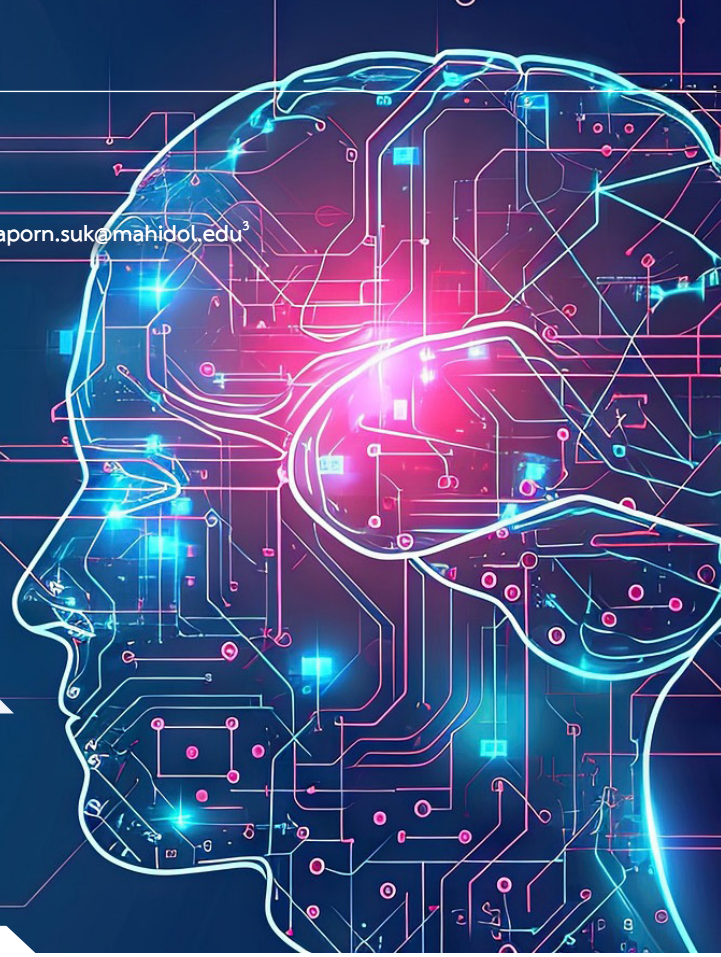
ศราวุฒิ สิทธารถ[1], ว่าที่ร้อยตรี วรพงศ์ แสงปิยะเพชร[2] และปรียาภรณ์ สุขงาม[3]
ฝ่ายสารสนเทศ คณะแพทยศาสตร์ศิริราชพยาบาล
อีเมล: sarawut.shi@mahidol.edu[1], vorapong.san@mahidol.ac.th[2], preeyaporn.suk@mahidol.edu[3]

การทำนายการเติบโตของฐานข้อมูลโดยการใช้ Machine Learning

Prediction of Database Growth Using Machine Learning



บทคัดย่อ

งานวิจัยเรื่อง การทำนายการเติบโตของฐานข้อมูลโดยการใช้ Machine Learning นำเสนอเทคนิคการเรียนรู้ของเครื่อง ด้วยอัลกอริทึมการเรียนรู้แบบมีการสอนมาพัฒนาแบบจำลองทำนายการเติบโตขนาดของไฟล์ข้อมูล โดยจะเป็นการศึกษาด้วยค่าตัวแปร 2 ตัว ซึ่งจะมีความสัมพันธ์ของตัวแปรทั้งสองในรูปแบบเส้นตรง (Simple Linear Regression) หรือเป็นเส้นโค้ง และสร้างส่วนที่โมเดล พร้อมทั้งประเมินความแม่นยำของโมเดล โดยมีวัตถุประสงค์ 1) เพื่อหาแบบจำลองการเรียนรู้ของเครื่องเพื่อทำนายการเติบโตของฐานข้อมูล และ 2) หาประสิทธิภาพด้านความถูกต้องของแบบจำลองการเรียนรู้ โดยได้ดำเนินการเก็บรวบรวมข้อมูล transaction log ในแต่ละวันของฐานข้อมูล SQL Server โดยการเขียนโปรแกรมภาษาไพธอนแล้วบันทึกในฐานข้อมูล เพื่อนำไปป้อนข้อมูลเข้าไปสอนให้คอมพิวเตอร์เรียนรู้ และสร้างส่วนที่เปรียบเหมือนสมองเพื่อกำหนดวิธีการในการคำนวณและการประมวลผล หลังจากนั้นประเมินความแม่นยำของโมเดลที่ได้จากการเรียนรู้ของเครื่อง แล้วดำเนินการทำนายการเติบโตของฐานข้อมูลและสร้างเป็นโมเดลออกมา โดยโมเดลที่ได้ออกมามีค่าสัมประสิทธิ์การตัดสินใจ เท่ากับ 0.852 โดยผลการวิจัยพบว่า ถ้าอัตราการเปลี่ยนแปลงข้อมูลในแต่ละวันไม่ต่างกันมาก จะทำให้โมเดลมีค่าใกล้เคียงกับ 1 ทำให้โมเดลมีความแม่นยำสูงขึ้น แต่ถ้าอัตราการเปลี่ยนแปลงของข้อมูลในแต่ละวันสูง จะทำให้โมเดลมีค่าห่างจาก 1 มาก และจะทำให้การทำนายมีความคลาดเคลื่อนสูง ทำให้โมเดลมีความไม่น่าเชื่อถือและผลในการทำนายอาจไม่ใกล้เคียงกับความเป็นจริง

คำสำคัญ : ข้อมูล ขับเคลื่อนด้วยข้อมูล ข้อมูลขนาดใหญ่ ปัญญาประดิษฐ์ การเติบโตของฐานข้อมูล

ABSTRACT

This research aims to prediction of database growth using machine learning presents techniques involving supervised learning algorithms to develop a model for predicting the growth in the size of data files. This study investigates the relationship between two variables, which may be linear (Simple Linear Regression) or non-linear, and constructs a model to evaluate its accuracy. The objective is 1) to develop a machine learning model to predict

database growth and 2) assess the model's accuracy. Data was collected from the daily Transaction Logs of a SQL Server database using a Python program, which logged the data for training purposes. The aim was to create a model that simulates brain function, allowing it to compute and process predictions. The accuracy of the machine learning model was then evaluated, and predictions of database growth were made. The resulting model achieved a coefficient of determination of 0.852. The findings indicate that if the daily data change rate is relatively consistent, the model's coefficient approaches 1, leading to higher accuracy. Conversely, if the daily data change rate varies significantly, the coefficient deviates from 1, resulting in higher prediction errors. This reduces the model's reliability and makes predictions less accurate.

Keywords: *Information, Data-Driven, Big Data, Artificial intelligence, and Growth Rate*

บทนำ

ปัจจุบันเทคโนโลยีการเรียนรู้ของเครื่องเป็นศาสตร์ที่ประยุกต์ใช้ปัญญาประดิษฐ์เพื่อให้ประมวลผล แก้ไขปัญหาแทนมนุษย์ ด้วยการป้อนข้อมูลเพื่อให้เครื่องเกิดการเรียนรู้และจดจำข้อผิดพลาด แล้วนำไปปรับปรุงกระบวนการเรียนรู้ให้ได้ผลลัพธ์ที่เกิดข้อผิดพลาดน้อยที่สุด (Nasteski 2017) ซึ่งปัจจุบันองค์กรและธุรกิจ เห็นถึงความสำคัญของการวิเคราะห์ค้นหาสิ่งที่ทำให้เกิดประโยชน์ต่อองค์กรและธุรกิจ จากข้อมูลขนาดใหญ่ (big data) ส่งผลให้การนำข้อมูลมาเก็บในฐานข้อมูล (database) มีมากขึ้น ดังนั้นการดูแลและวิเคราะห์ฐานข้อมูลขนาดใหญ่ ควรมีเครื่องมือที่ช่วยเหลือเพื่อให้เกิดประสิทธิภาพในการดูแล ยกตัวอย่าง การหาการเติบโตของฐานข้อมูล (growth rate) เราควรนำปัญญาประดิษฐ์ (Artificial Intelligence; AI) ซึ่งเป็นศาสตร์แขนงหนึ่งที่ทำการศึกษาและพัฒนาคอมพิวเตอร์ให้มีความสามารถอันชาญฉลาด เช่น การสร้างระบบประสาทรับรู้เลียนแบบมนุษย์ ฯลฯ ในปัญญาประดิษฐ์นั้น มีศาสตร์ย่อยแขนงหนึ่งคือ การเรียนรู้ของเครื่อง เป็นการนำข้อมูลมาสอนคอมพิวเตอร์ ให้เรียนรู้และสร้างส่วนที่เปรียบเหมือนสมอง เพื่อช่วยในการวิเคราะห์ฐานข้อมูล

งานวิจัยนี้ได้นำเทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning; ML) ประเภทการเรียนรู้แบบมีการสอน (Supervised Learning) ซึ่งจะเป็นการศึกษาความสัมพันธ์ของตัวแปร 2 ตัวขึ้นไป โดยจัดอยู่ในกลุ่ม Regression ซึ่งถ้าตัวแปรหนึ่งเปลี่ยนแปลงจะส่งผลกับอีกตัวแปรหนึ่งด้วย ไม่ว่าจะเพิ่มขึ้นหรือลดลง (รัสรินทร์ เมธาเฉลิมพัฒน์ 2565) นำมาใช้ทำนายการเติบโตของฐานข้อมูล SQL Server ที่ได้ดำเนินการเก็บรวบรวมข้อมูล Transaction Log ในแต่ละวัน โดยการเขียนโปรแกรมภาษาไพธอนแล้วบันทึกในฐานข้อมูลเพื่อนำไปป้อนข้อมูลเข้าไปสอนให้คอมพิวเตอร์เรียนรู้ และสร้างส่วนที่เปรียบเหมือนสมองเพื่อกำหนดวิธีการในการคำนวณและการประมวลผล หลังจากนั้นประเมินความแม่นยำของโมเดล

ดังนั้นคณะผู้วิจัยจึงได้พัฒนาแบบจำลองการทำนายการเติบโตของฐานข้อมูลโดยการใช้ Machine Learning โดยใช้ชุดข้อมูลสำหรับการสอนแบบจำลองร้อยละ 70 และชุดข้อมูลสำหรับการทดสอบแบบจำลองร้อยละ 30 และหาความถูกต้องของแบบจำลองด้วยวิธี R Square (จักรกฤษณ์ แสงแก้ว 2562) และมีความมุ่งหวังผู้สนใจจะนำ Model นี้ไปต่อยอดการพัฒนาเป็น Software และสามารถกรอกจำนวนวันที่ต้องการทำนายในอนาคต เพื่อที่จะใช้ในทำนายอัตราการเติบโตของฐานข้อมูลหรือประเภทของกลุ่มข้อมูลที่มีปริมาณมากให้ออกมาได้

วัตถุประสงค์การวิจัย

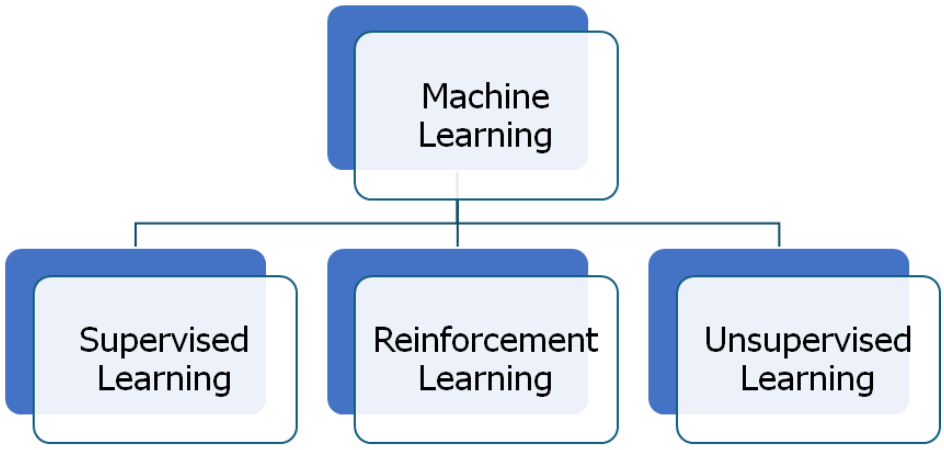
1. เพื่อหาแบบจำลองการเรียนรู้ของเครื่องเพื่อทำนายการเติบโตของฐานข้อมูล
2. หาประสิทธิภาพด้านความด้านความถูกต้องของแบบจำลองการเรียนรู้

แนวคิดของทฤษฎีการเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องเป็นศาสตร์หนึ่งของปัญญาประดิษฐ์ ที่เกี่ยวข้องกับการใช้อัลกอริทึมทางสถิติ (statistical algorithms) และโมเดลคอมพิวเตอร์เพื่อสร้างโปรแกรมคอมพิวเตอร์ที่สามารถเรียนรู้จากข้อมูลเพื่อแก้โจทย์ปัญหาแทนการเขียนโปรแกรมที่กำหนดขั้นตอนการแก้ปัญหา (ธนัชพร ศรีอาจ 2565)

โมเดลการเรียนรู้ของเครื่องโดยทั่วไปจะมีประสิทธิภาพที่ดีขึ้นเมื่อใช้ข้อมูลจำนวนมากในการฝึกฝน และข้อมูลที่ใช้ในการฝึกฝนสามารถมีรูปแบบได้หลากหลาย ตัวอย่างเช่น ตัวอักษร รูปภาพ วิดีโอ เสียง และข้อมูลเซ็นเซอร์ประเภทต่างๆ การเรียนรู้ของเครื่องถูกประยุกต์ใช้ในแอปพลิเคชันต่างๆ มากมาย ตัวอย่างเช่น การประมวลภาษาธรรมชาติ (natural language processing) การรู้จำด้วยคำพูด (speech recognition) การเข้าใจรูปภาพ (image recognition) และการตรวจจับทุจริต (fraud detection) เป็นต้น (ธนัชพร ศรีอาจ 2565)

ในการพัฒนาโปรแกรมหรือเขียนโปรแกรม คณะผู้วิจัยจะเป็นผู้กำหนดวิธีการในการคำนวณหรือการประมวลผล และกำหนดการทำงานทั้งหมด แต่ในส่วนการเรียนรู้ของเครื่อง (Machine Learning) จะมีกระบวนการที่แตกต่างจากการพัฒนาโปรแกรมโดยปกติ คือจะเป็นการนำข้อมูลที่มีอยู่มาทำการสอนให้คอมพิวเตอร์ (machine) ทำการเรียนรู้และสร้างส่วนที่เป็นสมอง (model) เพื่อให้มันกำหนดวิธีการในการคำนวณ การประมวลผล และกำหนดการทำงาน (วินน์ วรวุฒิคุณชัย 2564) การเรียนรู้ของเครื่อง (Machine Learning) แบ่งออกเป็น 3 ประเภท ดังนี้



รูปที่ 1. การเรียนรู้ของเครื่อง (Machine learning)

1. Supervised Learning: การเรียนรู้แบบมีการสอน

การเรียนรู้แบบมีการสอนนั้น เป็นการให้คอมพิวเตอร์เรียนรู้ จากการป้อนข้อมูลเข้าไปสอน การที่เรานำข้อมูลไปสอนเราจะเรียกว่า Training set หรือ Training data การเรียนรู้แบบมีการสอน คล้ายๆ กับการสอนเด็กๆ คุณครูก็จะนำรูปภาพเช่น นำรูปสุนัขให้ดูและบอกว่ารูปนี้เป็นสุนัข นำรูปแมวให้ดูและบอกว่านี่คือแมว พอนำรูปมาให้ดูมากๆ เด็กก็จะสามารถเรียนรู้และสามารถแยกแยะออกว่าเป็นสุนัขหรือแมว การเรียนรู้แบบมีการสอนจะสามารถคัดแยกหรือทำนายได้ดีก็จะต้องมีจำนวน Training data ที่เพียงพอ และมีคุณภาพข้อมูลที่ดี เราสามารถแยกย่อย Supervised Learning ได้เป็น 2 กลุ่ม คือ

- Classification เป็นการคัดแยก จำแนก เช่น เป็นสุนัขหรือแมว ดีหรือเสีย
- Regression เป็นการทำนาย คำนวณเป็นค่าตัวเลข เช่น ทำนายผู้ติดเชื้อผู้ป่วย COVID-19

2. Reinforcement Learning: การเรียนรู้อาศัยการป้อนกลับของผลลัพธ์

เป็นการให้คอมพิวเตอร์เรียนรู้ จากผลลัพธ์ที่เกิดขึ้นและปรับปรุงกระบวนการในการเรียนรู้ เช่น หุ่นยนต์เรียนรู้ในการทรงตัว พอหุ่นยนต์ล้มก็จะนำผลลัพธ์ในการล้มมาเรียนรู้และพัฒนาการเรียนรู้ปรับปรุงตัวเอง จนทำให้หุ่นยนต์สามารถเรียนรู้จนสามารถทรงตัวได้

3. Unsupervised Learning: การเรียนรู้แบบไม่มีการสอน

การเรียนรู้แบบไม่มีการสอน จะตรงข้ามกับการเรียนรู้แบบมีการสอน (Supervised Learning) ก็คือ เครื่องคอมพิวเตอร์เรียนรู้ได้โดยไม่มีการสอน เราเพียงมีชุดข้อมูลที่ป้อนเข้ามาเท่านั้น โดยปราศจากการให้ผลลัพธ์ที่เกิดขึ้น การเรียนรู้แบบไม่มีการสอนนี้ เครื่องคอมพิวเตอร์จะหาแบบแผน (pattern) ของข้อมูลนั้น และแบ่งประเภทข้อมูล

โดยในบทวิจัยนี้จะเป็นการเรียนรู้ของเครื่อง ประเภทการเรียนรู้แบบมีการสอน (Supervised Learning) ซึ่งจะเป็นการศึกษาความสัมพันธ์ของตัวแปร 2 ตัวขึ้นไป โดยจัดอยู่ในกลุ่ม Regression (จักรกฤษณ์ แสงแก้ว 2562) ซึ่งถ้าตัวแปรหนึ่งเปลี่ยนแปลงจะส่งผลกับอีกตัวแปรหนึ่งด้วย ไม่ว่าจะเพิ่มขึ้นหรือลดลง ซึ่ง Regression แบ่งออกเป็น 2 ประเภท คือ

1) Simple Regression ประกอบด้วยตัวแปร 2 ตัว โดยจะมีความสัมพันธ์ของตัวแปรทั้งสอง เป็นแบบเส้นตรง (Simple Linear Regression) หรือเป็นเส้นโค้ง ก็ได้

2) Multiple Regression จะคล้ายกับ Simple Regression แต่จะประกอบด้วยตัวแปรมากกว่า 2 ตัว จะกล่าวถึง Simple Linear Regression ซึ่งเราจะศึกษาความสัมพันธ์ของตัวแปร 2 ตัว ซึ่งมีความสัมพันธ์กันเป็นลักษณะเป็นเชิงเส้นตรง เช่น เมื่อจำนวนวันในการเก็บข้อมูลเพิ่มขึ้น (x) อัตราขนาดของพื้นที่ไฟล์ข้อมูล Data file (y) ก็จะเพิ่มขึ้น โดยการเพิ่มขึ้นจะมีลักษณะเป็นเส้นตรง ตัวแปรหรือข้อมูล (feature) จะประกอบด้วยดังนี้

- ตัวแปรอิสระ (predictor) จะเป็นค่าที่กำหนดข้างต้น เรามักจะแทนด้วยสัญลักษณ์ (x) ส่วนค่าประมาณการหรือค่าเป้าหมายมักจะแทนด้วยสัญลักษณ์ (y)
- ตัวแปรตาม (response) เราจะใช้สัญลักษณ์ (y) จากที่เราเรียกว่าตัวแปรตาม หมายความว่า ตัวแปรตามจะแปรผันตามตัวแปรอิสระ เช่น ตัวแปรอิสระมีการเปลี่ยนแปลงโดยที่มีค่าเพิ่มสูงขึ้น ก็จะมีผลทำให้ตัวแปรตามเปลี่ยนแปลงตามไปด้วย จากที่กล่าวมาเราสามารถเขียนสูตรสมการของความสัมพันธ์ของตัวแปรทั้งสองเป็นเชิงเส้นตรง ดังนี้

โดยที่ $y = m * x + b$

y = ค่าประมาณการหรือค่าเป้าหมาย โดยค่า y แปรผันตามค่าของ x
x = ค่าที่กำหนดตั้งต้น เพื่อนำไปหาค่า
m = ค่าความชันของเส้นตรง (slope) ซึ่งจะเป็นตัวบ่งบอกความสัมพันธ์ของสองตัวแปร ว่าเป็นแบบใด แปรผันตามกัน หรือแปรผันกลับ
b = คือจุดตัดบนแกน y เมื่อค่า x เป็นศูนย์ (intercept) คือ ถ้า x = 0, y จะมีค่าเท่ากับ b

การประเมินความแม่นยำของโมเดล ที่ได้จากการเรียนรู้ของเครื่อง นับเป็นสิ่งที่มีความสำคัญมาก เพราะจะเป็นตัวบอกว่าเหมาะกับการใช้งานหรือไม่มีความแม่นยำมากน้อยเพียงไร โดยการประเมินหรือวัดประสิทธิภาพความแม่นยำ คือการหาความคลาดเคลื่อนจากข้อมูลจริง ถ้าการทำนาย (predict) ใกล้กับจำนวนจริงมากๆ แสดงว่าโมเดล มีความแม่นยำสูง ซึ่งในบทความวิจัยนี้จะกล่าวถึงการประเมินประสิทธิภาพ สำหรับ Regression โดยใช้สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination) หรือ R Square ซึ่งได้ใช้ไลบารีของ Scikit-learn มาช่วยคำนวณค่าให้ โดยค่าที่ได้จากการคำนวณจะมีค่าระหว่าง 0 – 1 ถ้า R Square มีค่าเท่ากับหรือใกล้เคียง 1 แสดงว่าโมเดลมีค่าแม่นยำสูง แต่ถ้าค่าของ R Square มีค่าต่ำใกล้เคียงหรือเท่ากับ 0 แสดงว่าโมเดลมีค่าแม่นยำต่ำหรือมีความผิดพลาดสูงนั่นเอง

วิธีดำเนินการวิจัย

ในบทความวิจัยวิจัยนี้จะเป็นการอ่านค่าขนาดไฟล์ข้อมูล (data file) ซึ่งเป็นไฟล์ที่ใช้ในการจัดเก็บข้อมูลต่างๆ ในฐานข้อมูล โดยจะมีนามสกุลหลักเป็น .mdf (Primary data file) และมีนามสกุลรอง .ndf (Secondary data file) และไฟล์ Transaction Log เป็นไฟล์ที่ใช้สำหรับเก็บข้อมูลการทำงานในฐานข้อมูลทั้งหมด โดยการเขียนข้อมูลในฐานข้อมูลจะต้องทำการเขียนบน Transaction Log ก่อนและรอ Check Point จาก SQL Server เพื่อทำการเขียนในไฟล์ข้อมูล (data file) โดยนามสกุลของ Transaction Log จะเป็นนามสกุล .ldf

โดยคณะผู้วิจัยจะเขียนโปรแกรมภาษาไพธอน เพื่อเก็บข้อมูลขนาดของไฟล์ข้อมูลในแต่ละวัน บันทึกในฐานข้อมูล เพื่อนำไปป้อนข้อมูลเข้าไปสอนให้คอมพิวเตอร์เรียนรู้ และสร้างส่วนที่เปรียบเหมือนสมอง (model) เพื่อให้มันกำหนดวิธีการในการคำนวณ การประมวลผล และกำหนดการทำงาน โดยขั้นตอนนี้ได้ใช้โปรแกรม Jupyter Lab เวอร์ชัน 2.2.6 ในการสร้างส่วนที่เปรียบเสมือนสมอง (model) หลังจากนั้นประเมินความแม่นยำของโมเดล ที่ได้การเรียนรู้ของเครื่อง และดำเนินการทำนายการเติบโตของฐานข้อมูล โดยดำเนินการตามลำดับกระบวนการวิทยาการข้อมูล (Data science process) ดังต่อไปนี้

1. การตั้งคำถาม (ask an interesting question) ซึ่งในบทความวิจัยวิจัยนี้ เราต้องการทำนายว่า ในเวลาที่เปลี่ยนแปลงไป จำนวนการเติบโตของฐานข้อมูลจะเป็นอย่างไร เช่น ในอีก 120 วัน ข้อมูลขนาดไฟล์ข้อมูล (data file) จะเป็นเท่าไร

2. การเก็บรวบรวมข้อมูล (get the data) จะต้องมีการเก็บรวบรวม ขนาดข้อมูลของไฟล์ข้อมูล (data file) ทุกๆ วัน ซึ่งพัฒนาโปรแกรมโดยใช้ภาษา Python ในการอ่านข้อมูลจาก SQL Server เวอร์ชัน 2016 และเก็บข้อมูลในฐานข้อมูล โดยมีโครงสร้างตารางเก็บข้อมูล ดังแสดงในตารางที่ 1 ดังนี้

ตารางที่ 1. โครงสร้างตารางเก็บข้อมูล

Column Name	Data Type
Date Insert	datetime
Segment Name	nvarchar(50)
Group ID	nchar(10)
Filename	nvarchar(50)
AllocatedSizeinGB	decimal(19,2)
SpaceUsedinGB	decimal(19,2)
MaxinMB	decimal(19,2)
AvailableSpaceinMB	decimal(19,2)
PrecentUsed	decimal(19,2)
RemainderInGB	decimal(19,2)

3. การเตรียมข้อมูล (data preparation) จะต้องมีการทำความสะอาดข้อมูล (data cleansing) ให้มีความเหมาะสมให้เครื่องคอมพิวเตอร์ได้เรียนรู้จัดการข้อมูลที่มีความผิดปกติ (outlier)

4. การวิเคราะห์ข้อมูล (analyze the data) ทำการวิเคราะห์ข้อมูล จากการเก็บรวบรวมข้อมูล ว่าจะมีข้อมูลอะไรบ้างที่จะให้เครื่องคอมพิวเตอร์เรียนรู้ เพื่อนำข้อมูลไปวิเคราะห์จากโมเดล ในการทำนายการเติบโตของฐานข้อมูลต่อไป

5. การสร้างแบบจำลอง (modelling) เกิดจากการเรียนรู้จากการป้อนข้อมูลให้กับ Machine แล้วให้ Machine กำหนดเป็นหลักเกณฑ์ในการทำนาย เมื่อมีอินพุตเข้ามา หลังจากนั้นประเมินความแม่นยำของโมเดล ที่ได้การเรียนรู้ของเครื่อง (Machine Learning)

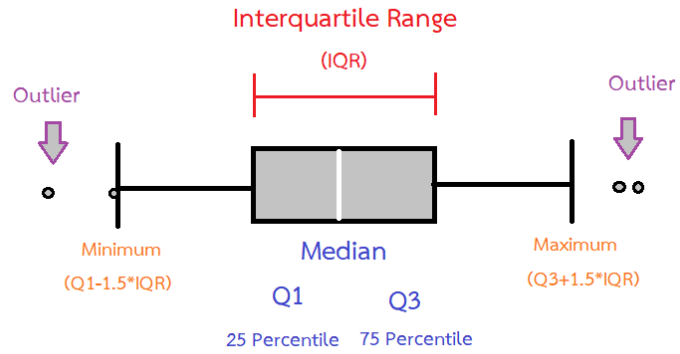
6. การสื่อสารและการทำผลลัพธ์ให้เป็นภาพ (communicate and visualize the results) เป็นการนำสิ่งที่ได้ไปใช้งาน เวลาใช้งานก็ใส่ข้อมูลจำนวนวันที่ต้องการทำนายเข้าไปเป็นอินพุต เข้าไปในโมเดล แล้วโมเดลก็จะทำนายคาดการณ์ผลลัพธ์ของขนาดไฟล์ข้อมูล (data file) ออกมา

ผลการทดลองและวิเคราะห์

ในบทความวิจัยวิจัยนี้ เราต้องการทำนายว่า ในเวลาที่เปลี่ยนแปลงไป จำนวนการเติบโตของฐานข้อมูลจะเป็นอย่างไร เช่น ใน 30, 60, 120, 180 และ 360 วัน จะมีขนาดของไฟล์ข้อมูลจะเป็นเท่าไร โดยการเก็บรวบรวมข้อมูล พัฒนาโปรแกรมในการดึงข้อมูลขนาดไฟล์ข้อมูล และขนาดของไฟล์ Transaction log บันทึกในฐานข้อมูล

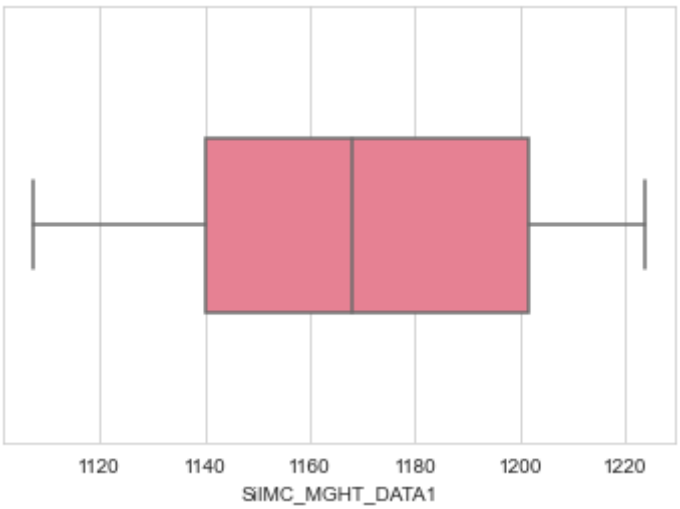
เมื่อได้ข้อมูลมา ต้องมีการทำความสะอาดข้อมูล (data cleansing) ให้มีความเหมาะสมเพื่อให้เครื่องคอมพิวเตอร์ได้เรียนรู้ โดยการจัดการข้อมูลที่มีความผิดปกติ (outlier) ซึ่งได้ใช้วิธีการตรวจจับ outliers ในข้อมูลด้วย Boxplot และ Interquartile Range (IQR) ดังแสดงในรูปที่ 2

IQR	=	Q3 - Q1
ค่าบน	=	Q3 + 1.5 * IQR
ค่าล่าง	=	Q1 – 1.5 * IQR
Q1	=	ควอไทล์ 3 หรือ Q3 ร้อยละ 25 คือค่าของข้อมูลตำแหน่ง 1/4 ของข้อมูลทั้งหมด
Q3	=	ควอไทล์ 3 หรือ Q3 ร้อยละ 75 คือค่าของข้อมูลตำแหน่ง 3/4 ของข้อมูลทั้งหมด



รูปที่ 2. Interquartile Range (IQR)

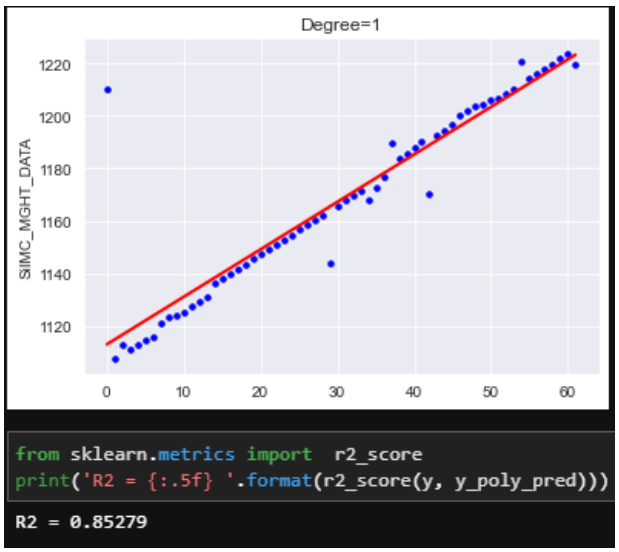
โดยการทำความสะอาดข้อมูล (data cleansing) ในโปรแกรม Jupyter Lab จะแสดง Visualization เป็น Boxplot ดังแสดงในรูปที่ 3 ดังนี้



รูปที่ 3. แสดง Visualization เป็น Boxplot

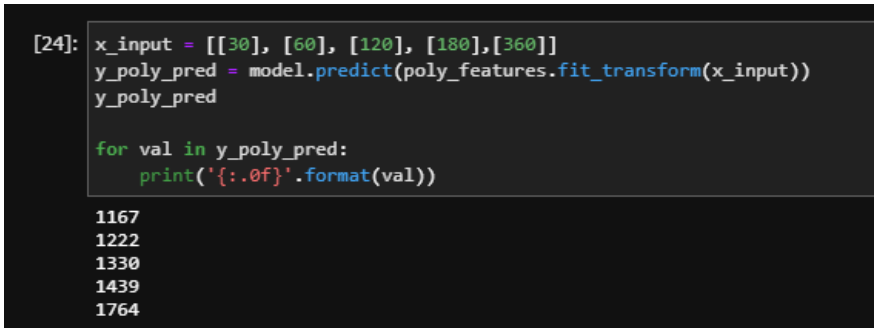
นำข้อมูลไปให้ Machine Learning สร้างแบบจำลอง (Modelling) ออกมา โดยเราจะนำค่าขนาดของไฟล์ข้อมูล (data file) ฟิลด์ SpaceUsedinGB ให้ Machine เรียนรู้ไปสร้างแบบจำลองออกมา โดยโมเดล ที่ได้ออกมาจะเป็นไฟล์นามสกุล .pkl หลังจากนั้นจะต้องประเมินความแม่นยำของโมเดล ที่ได้การเรียนรู้ของเครื่อง ซึ่งในบทความวิจัยนี้จะกล่าวถึงการประเมินประสิทธิภาพ สำหรับ Regression โดยใช้สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination) หรือ R Square ซึ่งได้ใช้ไลบารีของ Scikit-learn มาช่วยคำนวณค่าให้

โดยค่าที่ได้จากการคำนวณจะมีค่าระหว่าง 0 – 1 ถ้า R Square มีค่าเท่ากับหรือใกล้เคียง 1 แสดงว่าโมเดล มีค่าแม่นยำสูง แต่ถ้าค่าของ R Square มีค่าต่ำใกล้เคียงหรือเท่ากับ 0 แสดงว่าโมเดล มีค่าแม่นยำต่ำหรือมีความผิดพลาดสูงนั่นเอง โดยผลลัพธ์ได้ออกมา ดังแสดงในรูปที่ 4 ดังนี้



รูปที่ 4. แสดง Visualization เป็น Boxplot

หลังจากประเมินความแม่นยำของโมเดลการนำเสนอแล้วนำไปใช้งาน (Communication & Deployment) เป็นการนำสิ่งที่ได้ไปใช้งาน โดยเวลาใช้งานก็ใส่ข้อมูลจำนวนวันที่ต้องการทำนายเข้าไปเป็นอินพุตในโมเดล แล้วโมเดล ก็จะทำนายคาดการณ์ผลลัพธ์ของขนาดไฟล์ข้อมูล (data file) ออกมา โดยเราจะป้อนข้อมูลให้โมเดลทำนายโดยป้อนข้อมูลว่า ถ้าฐานข้อมูลมีการเก็บข้อมูลในวันที่ 30, 60, 120, 180 และ 360 ข้อมูลที่ทำนายจะเป็นอย่างไร โดยผลลัพธ์จะ แสดงดังในรูปที่ 5



รูปที่ 5. แสดง Visualization เป็น Boxplot

จากรูปที่ 5 เป็นการทำนายว่า ถ้าฐานข้อมูลมีการเก็บข้อมูลจำนวน 30 วัน จะมีขนาดข้อมูลเท่ากับ 1,167 GB และถ้ามีการเก็บข้อมูล 60 วัน จะมีการเก็บข้อมูลเท่ากับ 1,222 GB และถ้าเก็บข้อมูล 120 วัน จะมีการเก็บข้อมูลเท่ากับ 1,330 GB เป็นต้น ซึ่งจะทำให้เราสามารถทำนายอัตราการเติบโตของข้อมูลได้ และสามารถเตรียมพื้นที่ของฐานข้อมูล ให้สอดคล้องกับการเปลี่ยนแปลงได้

สรุปผลการวิจัยและอภิปรายผล

บทความวิจัยนี้จะเป็นการนำ Historical Data ของอัตราการเปลี่ยนแปลงของ Data File ในฐานข้อมูล SQL Server เวอร์ชัน 2016 โดยการเก็บข้อมูล ได้มีการพัฒนาโปรแกรมภาษาไพธอน เก็บข้อมูลลงฐานข้อมูล และใช้โปรแกรม Jupyter Lab เวอร์ชัน 2.2.6 เขียนโปรแกรมเพื่อให้ Machine Learning เรียนรู้แบบมีการสอน (Supervised Learning) จาก Historical Data ซึ่งจะเป็นการศึกษาความสัมพันธ์ของตัวแปร 2 ตัวขึ้นไป โดยจัดอยู่ในกลุ่ม Regression และสร้างเป็นโมเดลออกมา โดยโมเดลที่ออกมามีค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination) หรือ R Square เท่ากับ 0.852 ความหมายคือ มีความแม่นยำถูกต้องประมาณร้อยละ 85 พูดง่ายๆ ก็คือ ถ้านำโมเดลไปทำนาย (predict) หรือพยากรณ์หรือคำนวณประมาณการ จะมีความถูกต้องร้อยละ 85 หลังจากนั้นจะนำโมเดลไปพัฒนาเป็น Software และนำข้อมูลจำนวนวันที่ต้องการทำนายในอนาคต มากรอกจะสามารถทำนายอัตราการเติบโตของฐานข้อมูลออกมาให้

ซึ่งจากการวิเคราะห์ผลลัพธ์ที่ออกมา ถ้าอัตราการเปลี่ยนแปลงข้อมูลในแต่ละวันไม่ต่างกันมาก จะทำให้โมเดลมีค่า R Square ใกล้เคียงกับ 1 มากขึ้น จะทำให้โมเดลมีความแม่นยำสูงขึ้น แต่ถ้าอัตราการเปลี่ยนแปลงของข้อมูลในแต่ละวันมีการเปลี่ยนแปลงข้อมูลที่มีความแตกต่างกันมากจะทำให้ค่า R Square ในโมเดลมีค่าห่างจาก 1 มาก จะทำให้การทำนายจะมีความคลาดเคลื่อนสูง ทำให้โมเดลมีความไม่น่าเชื่อถือ และผลในการทำนายอาจไม่ใกล้เคียงกับความเป็นจริง

ข้อเสนอแนะ

ในการทำวิจัยฉบับนี้ จะพบว่าถ้าข้อมูลที่มีความผิดปกติ (outlier) อยู่ในการทำแบบจำลองจำนวนมาก จะทำให้โมเดลที่ออกมามีความแม่นยำน้อยลง มีความคาดเคลื่อนสูง ดังนั้นกระบวนการที่สำคัญที่สุดในการทำวิจัยนี้ ต้องให้ความสำคัญกับการเตรียมข้อมูล (data preparation) การทำความสะอาดข้อมูล (data cleansing) เพื่อคุณภาพของข้อมูลที่จะไปสอนให้คอมพิวเตอร์เรียนรู้ และออกมาเป็นโมเดล ที่มีคุณภาพ และในการนำข้อมูลมาสอนคอมพิวเตอร์ให้เรียนรู้ จำเป็นจะต้องมีข้อมูลจำนวนมากเพียงพอ เพื่อให้เกิดโมเดลที่ดีด้วย

เอกสารอ้างอิง

จักรกฤษณ์ แสงแก้ว. 2562. ลิเนียร์รีเกรสชั่น (Linear Regression). [ออนไลน์]. เข้าถึงได้จาก: https://dsdi.msu.ac.th/?article=data-science&fn=linear-regression, [เข้าถึงเมื่อ 20 เมษายน 2567].

รัสรินทร์ เมธาเฉลิมพัฒน์. 2565. การประยุกต์ใช้ Machine Learning กับงานในภาคอุตสาหกรรม (ตอนที่ 1). [ออนไลน์]. เข้าถึงได้จาก: https://www.nectec.or.th/smc/machine-learning/, [เข้าถึงเมื่อ 1 มิถุนายน 2567].

วินน์ วรวุฒิคุณชัย. 2564. Machine Learning. [ออนไลน์]. เข้าถึงได้จาก: https://medium.com/botnoi-classroom/machine-learning-4d96d5ee886d, [เข้าถึงเมื่อ 1 มิถุนายน 2567].

ศรีอาจ ธนัชพร. 2565. กลยุทธ์การประมูลวิชาเรียนด้วยทรัพยากรที่จำกัดโดยใช้อัลกอริทึมเชิงพันธุกรรม. [ออนไลน์]. เข้าถึงได้จาก: https://digital.car.chula.ac.th/chulaetd/6487, [เข้าถึงเมื่อ 31 พฤษาภาคม 2567].

อนุวัฒน์ พานิพัด. 2565. กระบวนการวิทยาการข้อมูล. [ออนไลน์]. เข้าถึงได้จาก: https://www.teachernu.com/2022/10/15/21/13/10/7553/, [เข้าถึงเมื่อ 31 พฤษาภาคม 2567].

Nasteski, V., 2017. An overview of the supervised machine learning methods. *Horizons*, b, **4**, pp. 51-62.