



**ข้อเสนอโครงการวิศวกรรมคอมพิวเตอร์**  
**วิชา 01076014 การเตรียมโครงการวิศวกรรมคอมพิวเตอร์**  
**ภาคเรียนที่ 2 ปีการศึกษา 2564**

1. ชื่อหัวข้อโครงการ (ไทย) .....ระบบแนะนำสำหรับการบริหารหลักสูตรวิศวกรรมคอมพิวเตอร์.....
2. ชื่อหัวข้อโครงการ (อังกฤษ) ...Recommendation System for CE Curriculum administrators..
3. Keyword 3 คำ .....
4. ประเภทโครงการ (ü)
 

☐ 1. HW+SW

☒ 2. SW\_Dev

☐ 3. Research
5. รายชื่อผู้ทำโครงการ
 

5.1. นางสาว .....ณิกานต์ สุขุมจิตพิทย.....รหัส .....62010299.....

5.2. นาย.....นนทพันธุ์ รุจิรกาล.....รหัส .....62010453.....

5.3. นาย.....นรวิชญ์ อยู่บัว.....รหัส .....62010465.....
6. อาจารย์ที่ปรึกษา
 

6.1. อาจารย์ที่ปรึกษาหลัก .....ผศ. ดร. ชาญชัย ตรีภาค.....

อาจารย์ที่ปรึกษาร่วม .....

## 1. ที่มาและความสำคัญของปัญหา (Motivation)

Data Analytics เป็นการวิเคราะห์ข้อมูลที่มีอยู่ตั้งแต่อดีตจนถึงปัจจุบัน เพื่อทำนายอนาคตที่เป็นประโยชน์ เพื่อพยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้น โดยใช้ข้อมูลที่ได้เกิดขึ้นแล้วกับแบบจำลองทางสถิติ รวมถึงการให้คำแนะนำทางเลือกต่างๆ ที่มีอยู่ และผลของแต่ละทางเลือก

โดยปัญหาที่ทางผู้จัดทำเล็งเห็นความสำคัญคือการที่นักศึกษาบางคนไม่รู้อนาคตหรือไม่สามารถตัดสินใจในการเลือกได้ ด้วยเกรดเฉลี่ยและวิชาเรียนที่มีอยู่ การวางแผนการเรียนและคำนวณเกรดเพื่อวางแผนนั้นเป็นสิ่งสำคัญสำหรับการไปต่อ หรือต่อยอดจากสิ่งที่มี

ซึ่งปัญหาการวางแผนในการเรียนของนักศึกษาไม่ได้เป็นปัญหาที่พบแค่ในตัวของนักศึกษา แต่อาจารย์และบุคลากรที่เกี่ยวข้องกับการศึกษานั้นก็มีความต้องการผลลัพธ์และการวางแผนนี้เช่นกัน เพื่อพัฒนานักศึกษาองค์กร และตลาดธุรกิจ

ดังนั้นผู้จัดทำจึงได้เสนอการใช้ระบบแนะนำการวางแผน การคาดการณ์ โดยใช้ความรู้ทางด้าน Predict หรือ Data Analytics และระบบ Data Visualization โดยใช้ข้อมูลจากในอดีตหรือข้อมูลที่เก็บไว้ในระบบ เข้ามาช่วยในการวิเคราะห์สิ่งต่างๆ เพื่อพัฒนาช่วยเหลือและตอบโจทย์ให้แก่นักศึกษาและบุคลากรทางการศึกษาหรือบุคคลที่เกี่ยวข้อง

## 2. วัตถุประสงค์ (Objectives)

2.1 เพื่อนำข้อมูลของนักศึกษาที่มีอยู่มาปรับใช้ให้เป็นระบบที่สามารถแก้ปัญหาการวางแผนการเรียนหรืออาชีพในอนาคตของตนเอง

2.2 เพื่อนำผลการพยากรณ์จากข้อมูลของนักศึกษาเหล่านั้นมาแสดงเป็นแผนภาพกราฟิกในการอำนวยความสะดวกให้อาจารย์ผู้สอนหรือกรรมการหลักสูตรในการทำงาน

2.3 เพื่อนำข้อมูลของนักศึกษาที่มีอยู่มาปรับใช้ให้เป็นระบบที่สามารถพยากรณ์โอกาสการพัฒนสภาพการเป็นนักศึกษาเพื่ออำนวยความสะดวกให้อาจารย์ผู้สอนหรือกรรมการหลักสูตรในการวางแผนรับมือเพื่อลดเปอร์เซ็นต์ของการเกิดการพัฒนสภาพการเป็นนักศึกษาของนักศึกษาในสถาบัน

2.4 เพื่อนำข้อมูลการพยากรณ์อาชีพในอนาคตของนักศึกษาในสถาบันมาแสดงเป็นแผนภาพกราฟิกในการอำนวยความสะดวกให้หน่วยงานภายนอกได้รับทราบว่าหลักสูตรปัจจุบันของสถาบันสามารถผลิตบุคลากรที่มีความชำนาญด้านใดบ้าง

2.5 เพื่อนำข้อมูลของนักศึกษาที่มีอยู่มาปรับใช้เป็นระบบแนะนำการเลือกเรียนวิชาของนักศึกษาให้นักศึกษาสามารถวางแผนวิชาเรียนตัวต่อไปได้อย่างมีประสิทธิภาพสูงสุด

### 3. ทฤษฎีที่เกี่ยวข้อง (Theoretical Background)

#### 3.1 Recommendation System

ระบบที่จะแนะนำสิ่ง (item) ที่ “เหมาะสม” ให้กับผู้ใช้ โดย item เป็นได้ตั้งแต่ ข้าว, เนื้อหา, เพลง, course เรียน, ไปจนถึงสินค้าที่ขายในร้าน online โดยระบบสามารถให้คำแนะนำที่ต่างกันตามความชอบของแต่ละผู้ใช้ได้

ข้อควรคำนึงในการทำ recommendation system

- Relevance: นำเสนอของที่คิดว่าผู้ใช้น่าจะสนใจ
- Diversity: สิ่งที่ต้องการจะนำเสนอ นั้นจะต้องคำนึงถึงความหลากหลาย
- Novelty: นำเสนอของที่ผู้ใช้ยังไม่เคยเห็นหรือมีความแตกต่างจากสิ่งที่ผู้ใช้ได้เห็นมาในอดีต
- Serendipity: มีทั้งความ relevant + novelty และสร้างความเซอร์ไพรส์ให้ผู้ใช้

ประเภทของ Recommendation System

1. Content-based Filtering เป็นรูปแบบโมเดลที่แนะนำลักษณะของตัวบริการหรือสินค้าเป็นตัวตั้ง แล้วแนะนำสินค้าและบริการอื่นที่มีคุณลักษณะคล้าย ๆ กัน

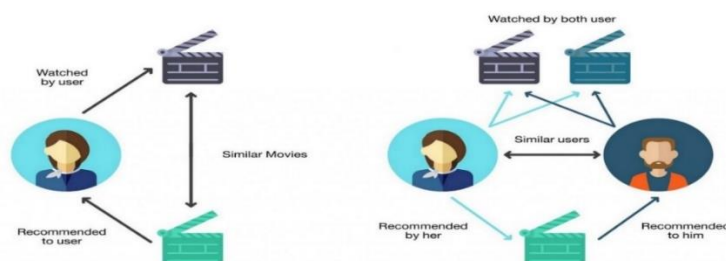
2. Collaborative Filtering เป็นรูปแบบโมเดลที่เรียนรู้จากพฤติกรรมของผู้ใช้กับผู้ใช้คนอื่น ๆ ที่คล้ายคลึงกัน

2.1 Memory-based ซึ่งจะเน้นไปในทางที่ดูข้อมูลแล้วหาความสัมพันธ์ ระหว่างผู้ใช้หรือสินค้า จากข้อมูลตรงๆ

2.2 Model-based ใช้เทคนิคของ machine learning เพื่อหา user embedding และ item embedding มาทำนาย rating ที่ผู้ใช้จะให้กับสินค้า หรือ relevance score

2.3 Hybrid ใช้หลายๆวิธีการมารวมกัน

3. Hybrid system เป็นการมัตรวมทั้งสองอัลกอริทึมด้านบนเอาไว้เพื่อให้ระบบการแนะนำ สมบูรณ์ขึ้น

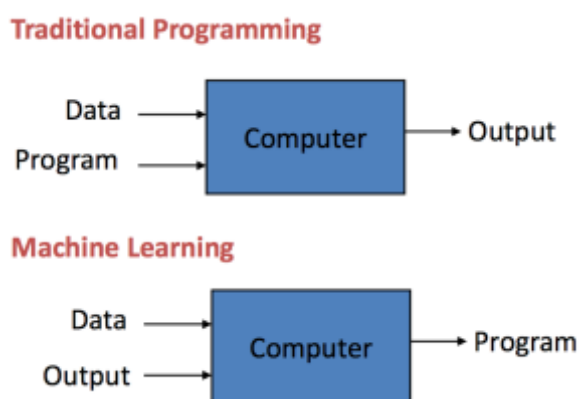


รูปที่ 3.1 ตัวอย่าง Recommendation System<sup>1</sup>

<sup>1</sup> <https://www.ldaworld.com/recommendation-system-lda/>

### 3.2 Machine Learning

การทำให้ระบบคอมพิวเตอร์เรียนรู้ได้ด้วยตนเอง โดยใช้ข้อมูล ซึ่งแตกต่างกับการเขียนโปรแกรมทั่วไป เพราะ Programming จะใส่ Data และ Program เข้าไปเพื่อให้ได้ Output แต่ Machine Learning เราไม่ได้ Program คำตอบ เราใส่ Data และ Output เข้าไปเพื่อให้หา Program ที่จะนำไปตอบในอนาคตได้ว่า Input แบบนี้ Output จะเป็นอะไร



รูปที่ 3.2 Machine Learning<sup>2</sup>

ประเภทของ Machine Learning

#### 1. Supervised Learning

เป็นการเรียนรู้ข้อมูลต่าง ๆ โดยมีผู้สอน อาศัยข้อมูลในการฝึกฝน เพื่อช่วยให้ตัวเทคโนโลยีสามารถเรียนรู้ผล และคาดคะเนผลลัพธ์ต่าง ๆ ได้อย่างแม่นยำมากยิ่งขึ้น โดยการเรียนรู้ในรูปแบบนี้ มักถูกนำมาใช้งานในเชิงธุรกิจทั้งการคำนวณราคาบ้าน การคาดคะเนค่าเงิน หรือแม้แต่การวิเคราะห์ผลการแข่งขันต่าง ๆ เป็นต้น

#### 2. Unsupervised Learning

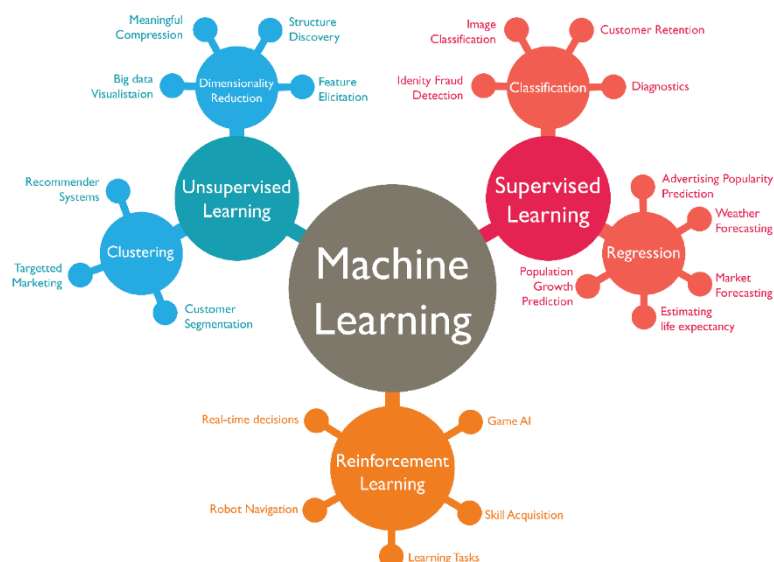
เป็นการเรียนรู้ที่ไม่มีผู้สอน โดยที่เครื่องจักรนั้น ๆ จะทำการเรียนรู้สิ่งต่าง ๆ ด้วยตัวเอง โดยไม่ต้องตั้งค่าเป้าหมายของแต่ละข้อมูล ระบบสามารถนำไปวิเคราะห์และสร้างแบบแผนจากข้อมูลที่ได้รับเข้าไป มักนำไปใช้ในการแนะนำผลิตภัณฑ์ต่างๆ หรือคัดเลือกข้อมูลหรือผลิตภัณฑ์ เป็นต้น

#### 3. Reinforcement Learning

การเรียนรู้จากการลองผิดลองถูกจากสถานการณ์ในอดีตหรือระบบจำลอง เพื่อพัฒนาระบบการตัดสินใจให้ดียิ่งขึ้น เช่น การพัฒนาระบบผู้เล่นอัตโนมัติให้ชนะผู้เล่นระดับโลก ระบบการจัดการข้อมูลเพื่อนำเสนอให้ตัดสินใจเลือกอัตราส่วนของสินทรัพย์ต่าง ๆ เป็นต้น

<sup>2</sup> [https://medium.com/investic/machine-learning-](https://medium.com/investic/machine-learning-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-fa8bf6663c07)

%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-fa8bf6663c07



รูปที่ 3.3 ประเภทของ Machine Learning<sup>3</sup>

### 3.3 Classification and Prediction

เทคนิคการจำแนกประเภทข้อมูล (Data Classification) เป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) หรือ ดาต้าไมน์นิ่ง (Data Mining) เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอนระบบ (training data) ที่แต่ละแถวของข้อมูลประกอบด้วยฟิลด์หรือแอททริบิวต์จำนวนมาก

ประเภทของปัญหาด้าน Classification

#### 1. Binary classification (การจำแนกแบบไบนารี)

เปรียบเทียบให้เข้าใจง่ายที่สุดก็คือ ตัวแปรที่แบ่งเป็นเพียงสองหมวดหมู่ เช่น ผลลัพธ์แบบ “ใช่” หรือ “ไม่ใช่” ช้อ หรือ ไม่ช้อ หากเปรียบเทียบเป็นตัวเลขก็คือ 0 กับ 1 อัลกอริทึมที่ใช้คู่กับการจำแนกแบบไบนารีก็มี k-Nearest Neighbors Decision Trees (ต้นไม้ตัดสินใจ) หรือ Naive Bayes

#### 2. Multi-Class Classification (การจำแนกประเภทหลายคลาส)

ต่างกับการจำแนกแบบไบนารีตรงที่มีหมวดหมู่มากกว่าสอง ยกตัวอย่างเช่น รูปภาพที่คล้ายคลึงกับรูปภาพที่อัปโหลดเพื่อค้นหาคำศัพท์ที่คาดว่าจะพิมพ์ใน predictive keyboard คือผลลัพธ์ที่มีได้มากกว่า 2 หมวดหมู่ อัลกอริทึมที่ใช้คู่กับการจำแนกประเภทหลายคลาสก็สามารถใช้อัลกอริทึมคล้ายกับที่ใช้กับการจำแนกแบบไบนารีได้ โดยอาศัยกลยุทธ์การแปลงคลาสแบบ One-vs-Rest (OvR) หรือ One-vs-One (OvO)

<sup>3</sup> <https://medium.com/invetic/machine-learning-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-fa8bf6663c07>

### 3. Multi-Label Classification (การจำแนกประเภทหลายเลเบล)

เพื่อเปรียบเทียบให้เข้าใจง่ายขึ้น ขอยกตัวอย่างเช่น รูปภาพรูปหนึ่งสามารถมีรูปดอกไม้ ท้องฟ้า ก้อนเมฆ ได้ แต่รูปภาพนั้นจะจัดว่าเป็นหมวดหมู่รูปวาด รูปถ่าย หรือรูปเสีย Multi-Label Classification ก็คือการเลเบล หรือตัดสินกว่าในรูปนั้นๆ มีดอกไม้หรือเปลามีก้อนเมฆหรือไม่ ส่วน Multi-Class Classification จะจำแนกว่ารูปนั้นเป็นรูปวาดรูปถ่ายรูปเสีย

### 4. Imbalanced Classification (การจำแนกแบบข้อมูลไม่เท่าเทียม)

เป็นปัญหาที่เกิดจากข้อมูลที่ไม่เท่าเทียม (Imbalanced dataset) ตัวอย่างเช่นข้อมูลการทุจริต โดยข้อมูลส่วนใหญ่จะเป็นข้อมูลที่จัดว่า “ไม่ทุจริต” และจะมีเปอร์เซ็นต์น้อยที่จัดว่าเป็น “ทุจริต” เป็นต้น การจัดเตรียมข้อมูลสำหรับการจำแนกและการทำนายข้อมูล

1. การทำความสะอาดข้อมูล (Data cleansing) จะเกี่ยวข้องกับการประมวลผลข้อมูลเบื้องต้นที่จะลบหรือลดข้อมูลที่มีสิ่งรบกวน (noise)

2. ความเกี่ยวเนื่องของข้อมูล (Relevance analysis) จะทำการตรวจสอบข้อมูลแอทริบิวต์ต่างๆว่ามีความเกี่ยวเนื่องหรือซ้ำซ้อนกันมากน้อยเพียงใด ซึ่งโดยปกติของชุดข้อมูลจะมีแอทริบิวต์ที่ซ้ำซ้อนกันดังนั้น เพื่อที่จะหลีกเลี่ยงความซ้ำซ้อน

3. การเปลี่ยนแปลง/เปลี่ยนรูปข้อมูลและการลดจำนวนข้อมูล (Data transformation and reduction) ข้อมูลที่เป็นอินพุตอาจมีช่วงของข้อมูลหรือค่าของข้อมูลที่มีระยะห่างค่อนข้างมาก ดังนั้น เราอาจทำการเปลี่ยนแปลง/เปลี่ยนรูปด้วยวิธีการ normalization ที่ทำการปรับเปลี่ยนค่าในแอทริบิวต์หนึ่งๆให้อยู่ในช่วงที่กำหนด

## Decision Tree

เป็นหนึ่งในวิธีการเรียนรู้ซึ่งใช้ในสถิติ, การเรียนรู้ของเครื่อง และการทำเหมืองข้อมูล โดยพิจารณาการสังเกตการแบ่งแยกข้อมูลโดยพิจารณาข้อมูล

ในการเรียนรู้ของเครื่อง (machine learning) ต้นไม้ตัดสินใจ เป็นโมเดลทางคณิตศาสตร์ที่ใช้ทำนายประเภทของวัตถุโดยพิจารณาจากลักษณะของวัตถุ บัพภายใน (inner node) ของต้นไม้จะแสดงตัวแปร ส่วนกิ่งจะแสดงค่าที่เป็นไปได้ของตัวแปร ส่วนบัพใบ (leaf node) จะแสดงประเภทของวัตถุ

ต้นไม้ตัดสินใจที่บัพใบแสดงถึงข้อมูลที่เป็นข้อมูลไม่ต่อเนื่อง (discrete values) จะเรียกว่าต้นไม้ตัดสินใจแบบจำแนก (classification trees) และต้นไม้ตัดสินใจที่บัพใบเป็นข้อมูลต่อเนื่อง (continuous values) จะเรียกว่าต้นไม้ตัดสินใจแบบถดถอย (regression trees)

### ลักษณะของต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจจะทำการจัดกลุ่ม (classify) ชุดข้อมูลนำเข้าในแต่ละกรณี (Instance) แต่ละบัพ (node) ของต้นไม้การตัดสินใจคือตัวแปร (attribute) ต่างๆของชุดข้อมูล เช่นหากต้องการตัดสินใจว่าจะไปเล่นกีฬหรือไม่ก็จะมีตัวแปรต้นที่จะต้องพิจารณาคือ ทัศนียภาพ ลม ความชื้น อุณหภูมิ เป็นต้น และมีตัวแปรตามซึ่งเป็นผลลัพธ์จากต้นไม้คือการตัดสินใจว่าจะไปเล่นกีฬารึเปล่า ซึ่งแต่ละตัวแปรนั้นก็จะมีค่าของตัวเอง (value) เกิดเป็นชุดของตัวแปรค่าของตัวแปร (attribute-value pair) เช่น ทัศนียภาพเป็นตัวแปร ก็อาจมีค่าได้เป็น ฝนตก แดดออก หรือการตัดสินใจว่าจะไปเล่นกีฬารึเปล่านั้นก็อาจมีค่าได้เป็นใช่ กับ ไม่ใช่ เป็นต้น การทำนายประเภทด้วยต้นไม้ตัดสินใจ จะเริ่มจากบัพราก โดยทดสอบค่าตัวแปรของบัพ แล้วจึงตามกิ่งของต้นไม้ที่กำหนดค่าเพื่อไปยังบัพลูกถัดไป การทดสอบนี้จะกระทำไปจนกระทั่งเจอบัพใบซึ่งจะแสดงผลการทำนาย

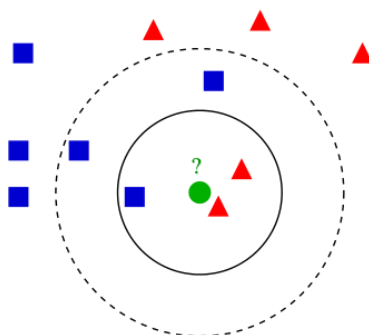
### K-Nearest Neighbors (KNN)

ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (อังกฤษ: K-Nearest Neighbors Algorithm) เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน (“K” ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด) ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด

กำหนดให้จุดที่พิจารณาคือ วงกลมสีเขียว ควรจัดกลุ่มให้จุดที่สนใจไปอยู่ใน คลาสแรกของสี่เหลี่ยมสีน้ำเงิน หรือ คลาสสองของสามเหลี่ยมสีแดง

ถ้า  $k=3$  แล้ว วงกลมสีเขียวจะอยู่ในคลาสสอง เพราะมี สี่เหลี่ยม 1 รูป และ สามเหลี่ยม 2 รูป อยู่ในวงกลมวงใน

ถ้า  $k=5$  แล้ว วงกลมสีเขียวจะอยู่ในคลาสแรก เพราะมี สี่เหลี่ยม 3 รูป และ สามเหลี่ยม 2 รูป อยู่ในวงกลมวงนอก



รูปที่ 3.4 ตัวอย่างการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด<sup>4</sup>

<sup>4</sup> [https://th.wikipedia.org/wiki/ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด\\_k\\_ตัว](https://th.wikipedia.org/wiki/ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด_k_ตัว)

## ขั้นตอนวิธี

การนำเทคนิคของขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดไปใช้นั้น เป็นการหาระยะห่างระหว่างแต่ละตัวแปร (Attribute) ในข้อมูล จากนั้นก็คำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น อย่างเช่น ถ้าเป็นเรื่องของสี เราจะใช้อะไรวัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว ต่อจากนั้นเราต้องมีวิธีในการรวมค่าระยะห่างของ Attribute ทุกค่าที่วัดมาได้ เมื่อสามารถคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่างๆ ได้ จากนั้นก็เลือกชุดของเงื่อนไขที่ใช้จัดคลาส มาเป็นฐานสำหรับการจัดคลาสในเงื่อนไขใหม่ๆ ได้แล้วเราจะตัดสินใจได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้น ควรมีขนาดใหญ่มากแค่ไหน และอาจมีการตัดสินใจได้ด้วยว่าจะนับจำนวนจุดข้างเคียงตัวมันได้หรือไม่ โดยขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดมีขั้นตอนโดยสรุป ดังนี้

1. กำหนดขนาดของ K (ควรกำหนดให้เป็นเลขคี่)
2. คำนวณระยะห่าง (Distance) ของข้อมูลที่ต้องการพิจารณากับกลุ่มข้อมูลตัวอย่าง
3. จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการพิจารณาตามจำนวน K ที่กำหนดไว้
4. พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่ม (class) ไหนที่ใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด
5. กำหนด class ให้กับจุดที่พิจารณา (class) ที่ใกล้จุดพิจารณามากที่สุด

### การดำเนินการหลัก

ฟังก์ชันระยะทาง(Distance Function) เป็นการคำนวณค่าระยะห่างระหว่างสองเรคคอร์ด เพื่อที่จะมาวัดความคล้ายคลึงกันของข้อมูล

ฟังก์ชันการแจกแจง(Combination Function) เป็นการรวมกันของผลลัพธ์ที่ได้จากการคำนวณค่าระยะห่าง(Distance) โดยทำการเรียงลำดับค่าระยะห่าง(Distance) จากน้อยไปมาก หลังจากนั้นดูจากค่า “K” ว่ากำหนดเป็นเท่าไร แล้วนำลำดับที่เรียงได้มาเทียบกับคลาสข้อมูลที่เรียงแล้วนำมาตอบ

### คุณสมบัติของฟังก์ชันระยะทาง(Distance Function)

- ค่าระยะทาง(ความห่าง) ที่คำนวณได้ต้องไม่ติดลบ
- ถ้าตำแหน่งเดียวกัน ฟังก์ชันต้องเป็นศูนย์(ค่าเหมือนกัน)
- การคำนวณวัดระยะทางไปกลับต้องเท่ากัน



การคำนวณค่าฟังก์ชันระยะทาง

- ใส่ค่าสัมบูรณ์(Absolute)ให้กับค่าระยะห่าง :  $|A-B|$
- ยกกำลังสองให้กับค่าระยะห่าง :  $(A-B)^2$
- ทำการปรับให้เป็นค่ามาตรฐาน:  $|(A-\text{mean})/(SD)-(B-\text{mean})/(SD)|$

การรวมค่าระยะทาง(Distance) ในเรคคอร์ด(Record)

การวัดระยะแบบแมนฮัตตัน(Manhattan distance)เป็นการนำค่าที่คำนวณได้ในหนึ่งเรคคอร์ด(Record)มารวมกัน

ระยะทางแบบยูคลิด(Euclidean distance)เป็นการหารากที่สอง(Square Root)ในแต่ละตัวแปร(attribute) แล้วนำมาบวกกัน แล้วนำค่าที่คำนวณได้ในหนึ่งเรคคอร์ด(Record) มารวมกัน

Naive Bayesian Classifier

ตัวจำแนกแบบเบย์อย่างง่าย (Naive Bayesian Classifier) คือโมเดลการจำแนกประเภทข้อมูลที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem และสมมติฐานที่ให้การเกิดของเหตุการณ์ต่างๆเป็นอิสระต่อกัน(Independence) กำหนดให้  $P(h)$  ความน่าจะเป็นที่จะเกิดเหตุการณ์  $h$  และ  $P(h|D)$  คือความน่าจะเป็นที่จะเกิด

เหตุการณ์  $h$  เมื่อเกิดเหตุการณ์  $D$  จากตัวแปรที่กำหนดและแนวคิดของ Bayes' Theorem นั้นเราสามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่างๆ ได้ดังสมการ

$$P(h|D) = [P(D|h) * P(h)]/P(D)$$

### 3.4 Visualization

ในทศวรรษที่ผ่านมา ข้อมูลมหาศาลถูกสร้างขึ้นโดยอัตโนมัติ และเพื่อให้ได้มาซึ่ง Insight จากข้อมูลเหล่านั้น รูปแบบและวิธีการวิเคราะห์ข้อมูลต่างๆได้เกิดขึ้นจำนวนมากและหลากหลาย อย่างไรก็ตามความสามารถทางสัมผัสและสติปัญญาของมนุษย์ที่จะเข้าใจธรรมชาติของข้อมูลที่ซับซ้อนเหล่านั้นมีอยู่จำกัด จะมีหนทางหรือวิธีการใดที่ทำให้มนุษย์สามารถเข้าใจข้อมูลได้อย่างง่ายและรวดเร็วได้บ้าง

#### Visual Analytics

การวิเคราะห์ด้วยภาพ คือการใช้เหตุผลเชิงวิเคราะห์ (Analytical Reasoning) ซึ่งว่าด้วยการมองดูข้อมูลในเชิงปริมาณหรือปริมาตรแล้วสามารถสกัดเอาความรู้หรือข้อสรุปออกมาได้ ประกอบด้วยหลายศาสตร์มากมาย เช่น Cognitive Science, Information Systems, Graphic & Interactive Design, Human-Computer Interaction, Mathematics & Statistical Methods, Data Analysis ซึ่งมี Interactive Visual Interfaces เป็นเครื่องมือและสื่อกลาง



รูปที่ 3.5 ตัวอย่าง Visual Analytics สรุปการขาย<sup>5</sup>

การวิเคราะห์ด้วยภาพเป็นการวิเคราะห์ข้อมูลโดยใช้รูปภาพเพื่อสื่อความ ไม่มีส่วนเกี่ยวข้องและเป็นคนละเรื่องกับศาสตร์ของการประมวลผลภาพดิจิทัล หรือ Digital Image Processing Visual Analytics จะช่วยเพิ่มกำลังความสามารถในการรับรู้และเข้าใจของมนุษย์ได้ต่อไปนี้

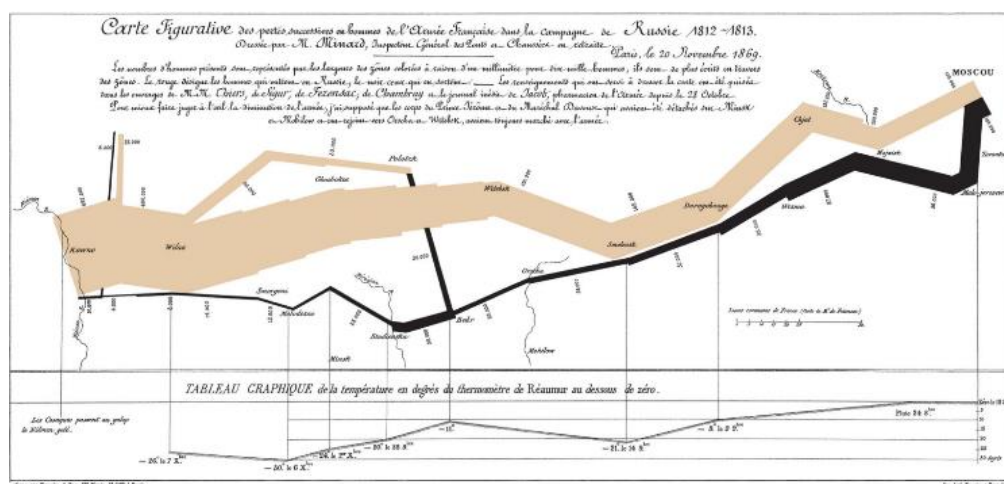
- ลดการทำงานในส่วนความจำ โดยใช้ภาพที่มองเห็นด้วยตาเปล่าเป็นสื่อรับรู้
- ลดการค้นหาค้นหาที่หนักและซับซ้อน โดยการนำเสนอข้อมูลที่มีขนาดใหญ่ในพื้นที่แสดงขนาดเล็ก
- ใช้รูปแบบเฉพาะเพื่อให้เราสามารถเข้าใจความหมายได้อย่างรวดเร็ว เช่น แผนภูมิวงกลม (Pie chart) หมายถึงสัดส่วนของข้อมูล

<sup>5</sup> <https://bigdata.go.th/big-data-101/visual-analytics/>

- ช่วยในการอนุมานความสัมพันธ์ให้ง่ายขึ้น
- ติดตามเหตุการณ์หรือปรากฏการณ์ในข้อมูลหลายเรื่อง ๆ ให้สามารถมองเห็นได้ด้วยตาเปล่า
- มีเครื่องมือเพื่อให้จัดการเปลี่ยนแปลงค่าพารามิเตอร์ต่าง ๆ ทำให้ทำการสำรวจได้มากยิ่งขึ้น ต่างจากรูปภาพนิ่งที่ไม่สามารถเปลี่ยนแปลงค่าใด ๆ ได้

## Data Visualization

การแสดงผลข้อมูลในรูปแบบรูปภาพนิ่ง เช่น รูปกราฟแท่ง รูปพายกราฟ เป็นต้น เพื่อให้มนุษย์สามารถเข้าใจข้อมูลได้ง่ายขึ้น มีจุดประสงค์หลัก ๆ ได้แก่ ถ่ายทอดสิ่งที่ข้อมูลต้องการสื่อสารอย่างชัดเจน ช่วยอธิบายแนวโน้มทางสถิติ แสดงรูปแบบเฉพาะในข้อมูลที่ยากจะเห็นด้วยตาเปล่า



รูปที่ 3.6 แผนภูมิรูปภาพการยกทัพไปรัสเซียของจักรพรรดินโปเลียน

โดย Charles Joseph Minard ในปี 1869<sup>6</sup>

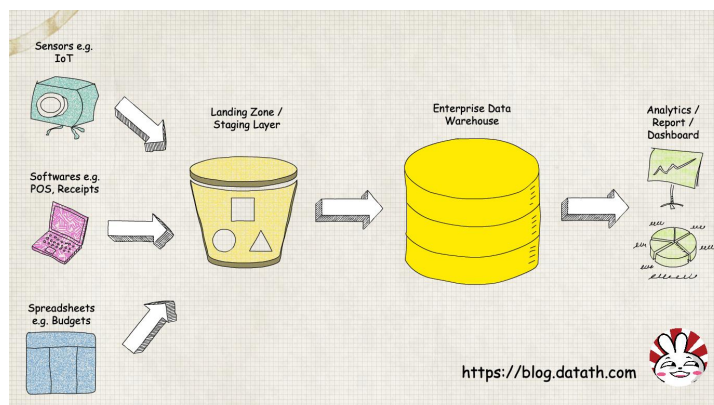
ปัจจุบันมีเครื่องมือสำหรับ Visual Analytics หลายหลายยี่ห้อ เช่น Tableau, Microsoft PowerBI, SAS เป็นต้น ซึ่งสามารถสร้างสรรค์ Interactive Dashboards หรือรายงานให้ตอบคำถามของผู้ใช้และสามารถเข้าใจได้ง่าย นอกจากนี้ยังมีความสามารถในการปรับแต่งค่าหรือกรองผลลัพธ์ต่าง ๆ ได้อย่างทันใจ ซึ่งจะเป็นผลดีเมื่อเจอสถานการณ์ที่ผู้ใช้ต้องการวิเคราะห์ข้อมูลเพื่อทำการตัดสินใจอย่างรวดเร็วทันที

ด้วยการที่ Visual Analytics สามารถเข้าถึงหลากหลายกลุ่มผู้ใช้งานไม่ได้จำกัดเฉพาะกลุ่มนักวิเคราะห์ข้อมูลเท่านั้น ทำให้เครื่องมือและเทคโนโลยีในสาขานี้มีความก้าวหน้าและถูกนำไปใช้ในงานหลายแขนงมากขึ้นด้วย เช่น งานด้านวิจัยงานด้านการศึกษา งานด้านการเกษตร เป็นต้น

<sup>6</sup> <https://bigdata.go.th/big-data-101/visual-analytics/>

### 3.5 Data pipeline

Data pipeline คือกระบวนการ หรือขั้นตอนในการ “ย้ายข้อมูลจากต้นทาง ไปยังปลายทาง” โดยมี 4 Phase หลักๆ คือ นำเข้าข้อมูล (Ingest), เก็บข้อมูล (Store), จัดการข้อมูล (Process) และ นำข้อมูลไปใช้ (Consume) ซึ่งในแต่ละขั้นตอนจะมีวิธีการย่อยๆ และใช้เครื่องมือแตกต่างกันออกไป



ขั้นตอนของ Data pipeline

1. เขียนโค้ดเก็บข้อมูลจากแหล่งต่าง ๆ = ข้อมูลมาได้จากหลายทาง เช่น เซ็นเซอร์เก็บข้อมูล, โปรแกรมต่าง ๆ เช่น Point of Sale (เครื่องคิดเงินตามร้านค้า) หรือแม้แต่ Excel ที่คนกรอก งานนี้อาจจะต้องเขียน NodeJS ดึงจาก API บ้าง, เขียน JavaScript ไปดึงข้อมูลในเว็บบ้าง, หรือแม้แต่เขียน Bash Script ไปดึงไฟล์
2. เขียนโค้ด Ingestion = ข้อมูลที่เก็บมาส่วนใหญ่จะเป็นไฟล์ดิบ ๆ ใน Landing Zone (เช่น HDFS) เอาไปทำอะไรยังไม่ได้ เราก็ต้องทำการโหลดข้อมูลเรานี้เข้าระบบฐานข้อมูลที่เราเอามาจัดการต่อได้ง่ายก่อน งานในส่วนนี้ก็คือการโหลดข้อมูลเข้าฐานข้อมูลที่เป็น Staging Layer หรือ Data Lake (เช่น Hive)
3. เขียนโค้ด Integration = ข้อมูลใน Staging Layer จะเป็นข้อมูลที่ค่อนข้างดิบ แต่อยู่ในฐานข้อมูลทำให้เราดัดแปลงได้ง่ายแล้ว ส่วนนี้ก็จะเป็นการแปลงข้อมูลให้เหมาะสมกับ Data Warehouse ที่จะเก็บ เช่น แปลงหน่วยของวันที่ให้เหมาะสม หรือแปลงเลขให้มีทศนิยมตามที่ต้องการ พอแปลงเสร็จก็เก็บลง Data Warehouse ซึ่งองค์กรใหญ่ ๆ จะเรียกว่า EDW (Enterprise Data Warehouse) และบางครั้งก็มีการแบ่งย่อยไปเป็น Data Mart ตามหน่วยธุรกิจอีก ขั้นตอนนี้แหละที่เค้ามักจะเรียกกันว่า ETL
4. นำข้อมูลไปวิเคราะห์, ทำ Report, ทำ Dashboard = งานส่วนนี้ Data Scientist จะเป็นคนทำมากกว่า แต่ก็อาจจะมีบางส่วนที่ยังต้องใช้ความช่วยเหลือจาก Data Engineer เพื่อเตรียมข้อมูลให้พร้อมไปทำ Data Wrangling (ทำความสะอาดข้อมูล) ซึ่งส่วนนี้เป็นส่วนที่พูดถึงกันบ่อย ๆ ไม่ว่าจะเป็นการทำ Data Visualization, Predictive Model, Neural Network, Deep Learning ฯลฯ

<https://blog.datath.com/data-pipeline/>

### 3.6 Scikit-learn

Scikit-learn เป็นโมดูลหนึ่งของภาษาโปรแกรมมิ่ง Python เป็นแพ็คเกจที่รวบรวม Library ด้าน Machine Learning เอาไว้ และถูกออกแบบมาให้ทำงานร่วมกับ Library ของภาษาโปรแกรมมิ่ง Python อย่าง Numpy และ Scipy ได้ดี

นอกจากนี้ Scikit-learn ยังเป็น Open Source ที่เปิดให้ผู้ที่สนใจสามารถเข้าไปพัฒนาต่อยอดได้ และที่ทำให้ทุกคนต่างยอมรับคือเป็นแหล่งรวม Library และอัลกอริทึมที่เน้นไปในด้านของ Machine Learning ซึ่งมีส่วนในการทำ Data Modeling ทั้งนี้ อีกหนึ่งปัจจัยที่ทำให้มีผู้ใช้เยอะ และเป็นเครื่องมือที่แนะนำสำหรับมือใหม่ เพราะเป็น Interface ระดับสูง ให้มือใหม่สามารถเข้าใจภาพรวมและ Workflow ของ Machine Learning ได้

เพื่อให้เห็นภาพการใช้งานของ Scikit-learn วันนี้ Achieve plus ขอพูดถึง 5 เครื่องมือที่ User สามารถนำไปใช้ในโปรเจกต์ได้

#### Classification

หนึ่งในตัวอย่างของ Classification คือการแยกอีเมลว่าเป็นสแปมหรือไม่ classification ถือว่าเป็นหนึ่งในแขนงของ Supervised Learning การเรียนรู้ของอัลกอริทึมจากชุดข้อมูล(Datasets) ที่มีคำตอบที่ถูกต้อง

#### Regression

เป็นเครื่องมือเพื่อเข้าใจความสัมพันธ์ระหว่างข้อมูล Input กับ Output ซึ่งก็ถือว่าเป็นอีกหนึ่งแขนงของ Supervised Learning เช่นกัน ตัวอย่างของอัลกอริทึมนี้คือการเข้าใจพฤติกรรมของราคาหุ้น

#### Clustering

คือการจำแนกกลุ่มข้อมูลที่มีคุณสมบัติคล้ายกัน เครื่องมือนี้ถือว่าเป็นแขนงของ Unsupervised Learning และตัวอย่างของการใช้อัลกอริทึมนี้คือการจำแนกกลุ่มลูกค้าตามพื้นที่

#### Model selection

อัลกอริทึมเพื่อใช้เปรียบเทียบ ตรวจสอบ และเลือกโมเดลและ Parameter ที่เหมาะสมกับชุดข้อมูลที่สุ่มในโปรเจกต์ จะช่วยเพิ่มความแม่นยำของอัลกอริทึม Machine Learning ได้

#### Pre-processing

ในขั้นตอนของ Data Analysis หรือการเข้าใจและวิเคราะห์ข้อมูล อาจต้องมีการแก้ไขให้ข้อมูลอยู่ในรูปแบบที่เรานำไปใช้งานได้ เครื่องมือตัวนี้ของ Scikit-learn จะสามารถช่วยจัดการกับข้อมูลได้

<https://medium.com/achieve-space/รู้จักใช้-scikit-learn-เหมือนมีโปรในเกม-4cebd3195518>

#### 4. งานวิจัยที่เกี่ยวข้อง (Related Works)

- 4.1 วิเคราะห์และพัฒนาระบบแนะนำหนังสือคอมพิวเตอร์ แบบออนไลน์โดยใช้เทคนิคการกรองแบบอิงเนื้อหา Development and Analysis of an Online Computer Books Recommendation System Using Content-Based Filtering
- 4.2 ระบบทำนายการพ้นสภาพของนักศึกษาระดับปริญญาตรี คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ ด้วยเทคนิคการทำเหมืองข้อมูล A Prediction System for Undergraduate Student Dropout at Faculty of Science, Buriram Rajabhat University using Data Mining Techniques

#### 5. ขอบเขตของโครงการ (Scope)

1. สร้างระบบวางแผนการเรียนของนักศึกษาโดยรับข้อมูลผ่านแบบกรอกฟอร์มใน Excel และนำมาคำนวณวิเคราะห์ผลลัพธ์แก่นักศึกษา ได้แก่ ความถนัดของนักศึกษา, เกรดที่คาดว่าจะได้รับ, อาชีพที่เป็นไปได้ในอนาคต
2. สร้างแบบแผนภาพที่สามารถแสดงจำนวนงานของบัณฑิตในแต่ละปี
3. สร้างระบบที่สามารถคาดการณ์จำนวนสายงานของบัณฑิตในปีถัด ๆ ไป
4. สร้างระบบแสดงรายชื่อนักศึกษาในรายวิชาเรียนต่าง ๆ
5. สร้างระบบคาดการณ์เกรดของนักศึกษาในรายวิชาเรียนต่าง ๆ
6. สร้างระบบจำกัดสิทธิ์กลุ่มผู้ใช้งานในแต่ละประเภท ได้แก่ อาจารย์, คณะกรรมการหลักสูตร
7. สร้างระบบของการรับข้อมูลนักศึกษาด้วยไฟล์ประเภท Excel
8. สร้างระบบที่สามารถคาดการณ์นักศึกษาที่คาดว่าจะเรียนไม่จบ หรือลาออก
9. ระบบสามารถดูข้อมูลย้อนหลังบัณฑิตได้นานสุดไม่เกิน 5 ปี

#### 6 การพัฒนาโครงการ (Project Development)

##### 6.1 ขั้นตอนการพัฒนา (Methodology)

- 6.1.1 กำหนดขอบเขต เป้าหมาย และวัตถุประสงค์ในการทำโครงการ
- 6.1.2 ศึกษา และทดลองตามขอบเขตของโครงการที่ได้กำหนดไว้
- 6.1.3 วางแผนการพัฒนาระบบ
- 6.1.4 ออกแบบระบบที่จะพัฒนา
- 6.1.5 ระบบที่จะใช้ใน ขั้นตอนการพัฒนา

- 6.1.5.1 Python เป็นภาษาระดับสูง ถูกออกแบบเพื่อให้มีโครงสร้างและ ไวยากรณ์ของภาษาที่ไม่ซับซ้อน มี Library ที่เหมาะแก่การทำ Classification Prediction และ Machine Learning
- 6.1.5.2 Django เป็นโปรแกรม Open Source สำหรับพัฒนา Web Application ซึ่งเขียนในภาษา Python
- 6.1.5.3 Scikit-learn เป็นไลบรารีในภาษาไพธอน จุดเด่นคือฟังก์ชันในการแบ่งประเภทข้อมูล การแบ่งกลุ่มข้อมูล การวิเคราะห์การถดถอยหลายอย่างไม่ว่าจะเป็น ซัพพอร์ตเวกเตอร์แมชชีน การเรียนรู้ต้นไม้ตัดสินใจ และการแบ่งกลุ่มข้อมูลแบบเคมีน
- 6.1.5.4 Firebase เป็นหนึ่งในผลิตภัณฑ์ของ Google โดย Firebase คือ Platform ที่รวบรวมเครื่องมือต่าง ๆ ใช้ในส่วนของ Database และ Authentication
- 6.1.5.5 React เป็น JavaScript library ที่ใช้สำหรับสร้าง User Interface

#### 6.1.6 แบ่งการทำงานออกเป็นดังนี้

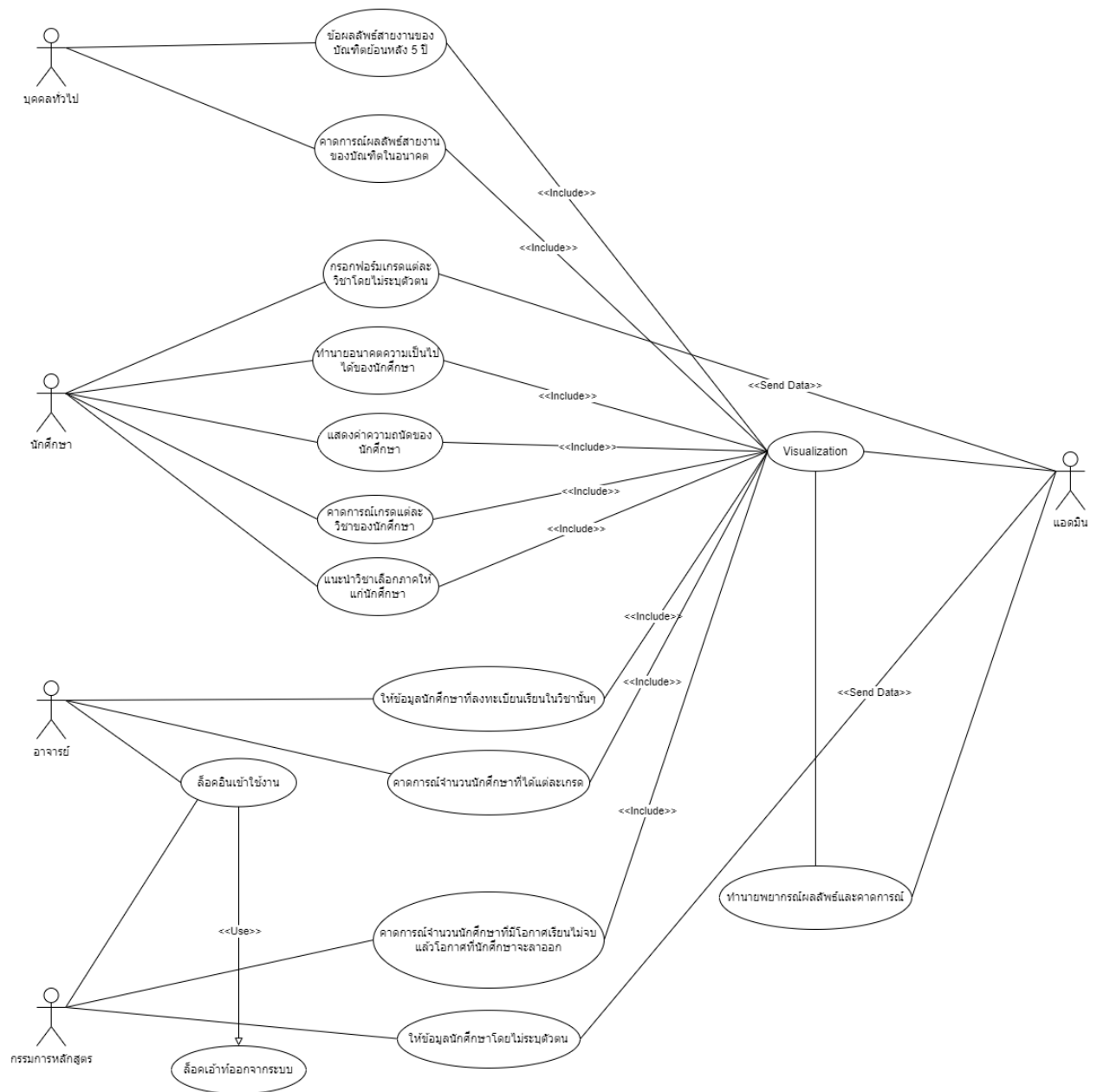
- 6.1.6.1 ส่วนของการรับข้อมูลจาก User
- 6.1.6.2 ส่วนของเซิร์ฟเวอร์สำหรับการประมวลผล Classification และ Prediction
- 6.1.6.3 ส่วนของเซิร์ฟเวอร์สำหรับดูแลรับส่งข้อมูลของ User
- 6.1.6.4 ส่วนของเซิร์ฟเวอร์สำหรับแสดงผลของ Web Application
- 6.1.6.5 ส่วนของฐานข้อมูลสำหรับจัดเก็บ model ที่ผ่านการ Trained ของ Classification และ Prediction
- 6.1.6.6 ส่วนของฐานข้อมูลสำหรับจัดเก็บข้อมูลที่ใช้ Train model

#### 6.1.7 ทดสอบการทำงานในแต่ละส่วน

#### 6.1.8 รวมแต่ละส่วนของระบบเข้าด้วยกัน

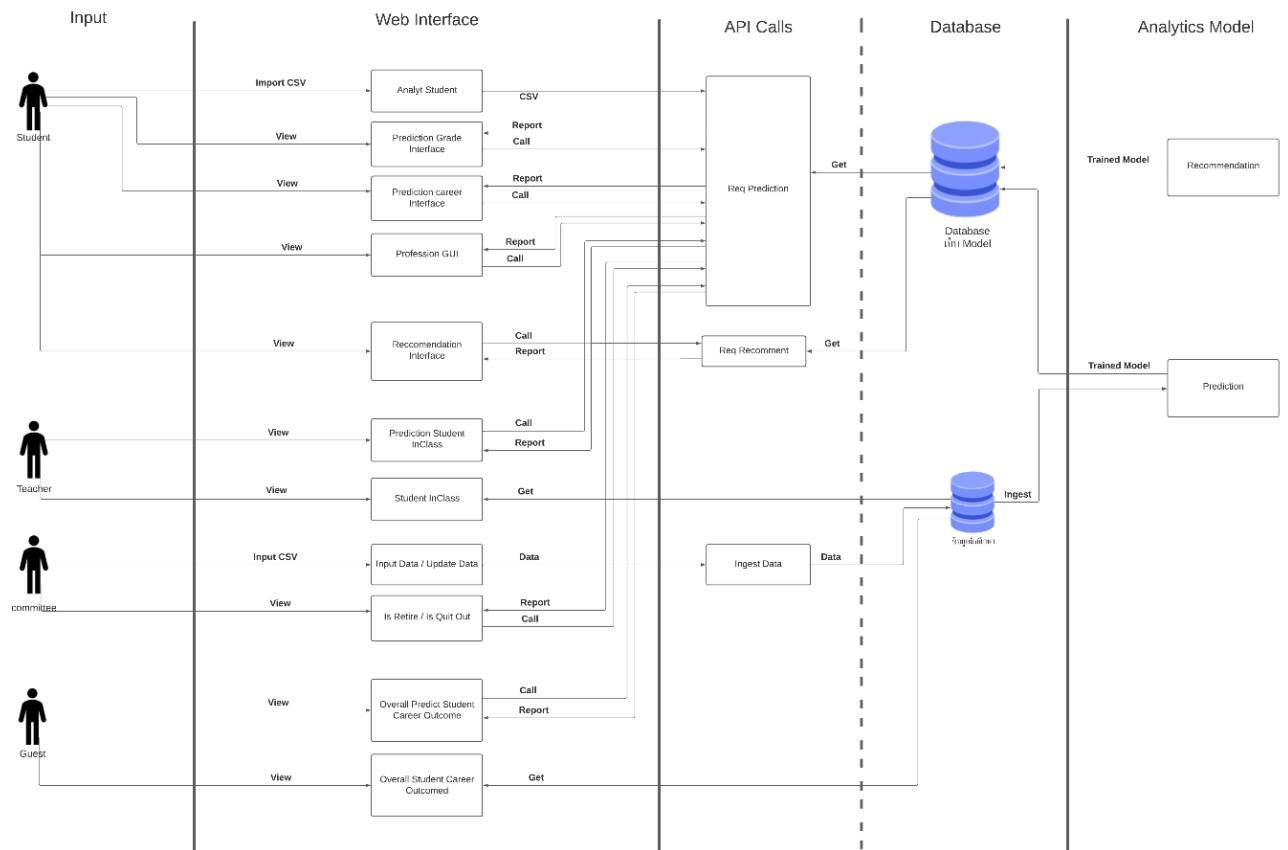
#### 6.1.9 ทดสอบการทำงานร่วมกันของระบบ

## 6.2 การออกแบบ (Design)



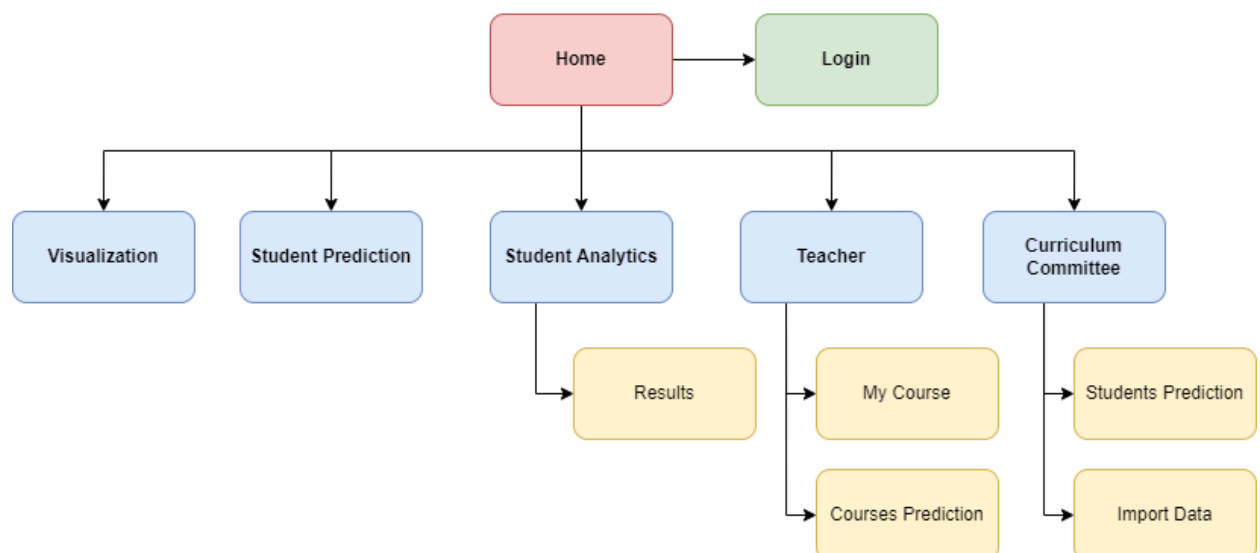
use case diagram





แผนผังงาน

### Site Map





### 6.3 แนวทางการทดสอบและการวัดประสิทธิภาพ (Test and Performance Evaluation Approaches )

## 7 แผนการดำเนินโครงการ (Gantt Chart)

## 8 ประโยชน์ที่คาดว่าจะได้รับ (Expected Benefits)

8.1 นักศึกษาสามารถวางแผนการเรียนรู้ การตั้งเป้าหมาย และเรียนรู้ตัวเองได้มากขึ้น

8.2 อาจารย์สามารถวางแผนการจัดการเรียนการสอนเพื่อให้เหมาะสมกับนักศึกษา

8.3 บุคคลภายนอกสามารถเข้ามาชมผลงานผลของบัณฑิตของหลักสูตรได้

8.4 คณะกรรมการหลักสูตรสามารถวางแผนปรับหลักสูตรและรับมือกับนักศึกษาที่มีโอกาสเรียนไม่จบหรือลาออกได้

## 9 ผลการศึกษาเทคโนโลยีที่ใช้พัฒนา

## 10 เอกสารอ้างอิง (Reference)

## 11 ใช้ Main Topic Style

### 11.1 ใช้ Sub Topic 1 Style

ใช้ Normal Style



รูป 11.1 ใช้ Caption Style<sup>7</sup>

ตาราง 11.1 ใช้ Table Caption Style


<sup>7</sup> ใช้ Footnote S

