ระบบวิเคราะห์ และ พยากรณ์ สำหรับการบริหารหลักสูตรวิศวกรรมคอมพิวเตอร์

Analytics and Prediction System for CE Curriculum administrators

ณิชกานต์ สุขุมจิตพิทโยทัย
นรวิชญ์ อยู่บัว

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตร์บัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2565

ปริญญานิพนธ์ปี การศึกษา 2564

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง ระบบวิเคราะห์ และ พยากรณ์ สำหรับการบริหารหลักสูตรวิศวกรรมคอมพิวเตอร์

Analytics and Prediction System for CE Curriculum administrators

ผู้จัดทำ

1. นางสาวณิชกานต์ สุขุมจิตพิทโยทัย รหัสนักศึกษา 62010299

2. นายนรวิชญ์ อยู่บัว รหัสนักศึกษา 62010465

_____อาจารย์ที่ปรึกษา

(ผศ. ดร. ธนัญชัย ตรีภาค)

ระบบวิเคราะห์ และ พยากรณ์ สำหรับการบริหารหลักสูตรวิศวกรรมคอมพิวเตอร์

นางสาวณิชกานต์	สุขุมจิตพิทโยทัย	62010299
นายนรวิชญ์	อยู่บัว	62010465
ผศ. ดร. ธนัญชัย	ตรีภาค	อาจารย์ที่ปรึกษา

ปีการศึกษา 2565

บทคัดย่อ

โครงงานนี้จัดทำขึ้นเพื่อ พัฒนาระบบประมวลผลข้อมูลผลการเรียนของนักศึกษาในอดีต ข้อมูลของรายวิชาต่างๆ และข้อมูลจากแบบสำรวจการมีงานทำของบัณฑิต เพื่อนำเสนอข้อมูลสถิติต่างๆ วิเคราะห์ข้อมูลผลการผลิตบัณฑิตเพื่อให้ได้ผลลัพธ์ว่าที่ผ่านมาหลักสูตรสามารถผลิตบัณฑิตกลุ่มใดได้บ้าง มีจำนวนมากน้อยเพียงใด สามารถพยากรณ์ว่าในอนาคตหลักสูตรสามารถผลิตบัณฑิตกลุ่มใดได้เป็นจำนวนเท่าใด เพื่อเป็นประโยชน์และอำนวยความสะดวกให้กรรมการหลักสูตรในการวางแผนการบริหารหลักสูตรในอนาคต และแสดงเป็นแผนภาพกราฟิกในการอำนวยความสะดวกให้หน่วยงานภายนอกได้รับทราบว่าหลักสูตรปัจจุบันของสถาบันสามารถผลิตบุคลากรที่มีความชำนาญด้านใดได้บ้าง

Analytics and Prediction System for CE Curriculum administrators

Ms. Nichakan Sukhumjitpitayotai 62010299

Mr. Narawich Youbua 62010465

Mr. Thanunchai Threepak Advisor

Academic Year 2022

Abstract

กิตติกรรมประกาศ

โครงงานในภาคการศึกษานี้สำเร็จลุล่วงได้ด้วยดีจากความช่วยเหลือจากหลากหลายบุคคล โครงงานในภาคการศึกษานี้จะผ่านไปไม่ได้หากปราศจากความช่วยเหลือจากบุคคลเหล่านี้ขอขอบคุณ อาจารย์ที่ปรึกษา ผศ. ดร. ธนัญชัย ตรีภาค ที่ให้ความช่วยเหลือในเรื่องต่าง ๆ ไม่ว่าจะเป็นการให้คำแนะนำถึงแนวทางการทำงานที่ดี การให้คำปรึกษาเพื่อหาทางออกเมื่อพบเจอกับปัญหา รวมถึงให้ความรู้เกี่ยวกับตัวงานทำให้งานต่าง ๆ เมื่อเจอปัญหาก็สามารถผ่านไปได้ด้วยดี

ขอขอบคุณคณาจารย์ในภาควิชาวิศวกรรมคอมพิวเตอร์ ที่ประสาทวิชาการความรู้มาตลอด 4 ปี ซึ่งความรู้หลาย ๆ แขนงก็ถูกใช้เป็นพื้นฐาน และเป็นส่วนหนึ่งของโครงงานนี้

ขอขอบคุณเพื่อน ๆ วิศวกรรมคอมพิวเตอร์ที่ให้คำปรึกษา และแลกเปลี่ยนความรู้ซึ่งกันและกัน รวมถึงการรับฟังปัญหา

สุดท้ายนี้ขอขอบคุณบิดา มารดาและครอบครัว ที่เลี้ยงดูอบรมสั่งสอนและให้ความรู้คุณธรรม จริยธรรม และให้การสนับสนุนด้านการศึกษาจนได้มีโอกาสมาทำโครงงานนี้

ณิชกานต์ สุขุมจิตพิทโยทัย

นรวิชญ์ อยู่บัว

สารบัญ

หน้า

บทคัดย่อภาษาไทย

บทคัดย่อภาษาอังกฤษ

กิตติกรรมประกาศ

สารบัญ

สารบัญตาราง

สารบัญภาพ

บทที่ 1 บทนำ

- 1.1 ความเป็นมาของปัญหา
- 1.2 วัตถุประสงค์
- 1.3 ประโยชน์ของโครงงาน
- 1.4 ข้อจำกัดของโครงงาน
- 1.5 แผนการดำเนินงาน

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

- 2.1 ทฤษฎีที่เกี่ยวข้อง
- 2.2 เครื่องมือที่เกี่ยวข้อง
- 2.3 งานวิจัยที่เกี่ยวข้อง

บทที่ 3 การออกแบบ

บทที่ 4 ผลการดำเนินงาน

สารบัญ(ต่อ)

บทที่ 5 สรุป

5.1 บทสรุป

5.2 ปัญหาและอุปสรรคที่พบ

5.3 แนวทางในการพัฒนาต่อ

เอกสารอ้างอิง

ภาคผนวก

สารบัญตาราง

ตาราง **หน้า**

สารบัญรูป

รูป **หน้า**

บทที่ 1

บทนำ

1.1 ความเป็นมาของปัญหา

Data Analytics เป็นการวิเคราะห์ข้อมูลที่มีอยู่ตั้งแต่อดีตจนถึงปัจจุบัน ในกรณีที่ข้อมูลเพียงพอและเหมาะสมจะสามารถนำมาคาดการณ์แนวโน้ม ทำนายอนาคตที่เป็นประโยชน์ พยากรณ์สิ่งที่กำลังจะเกิดขึ้นหรือน่าจะเกิดขึ้นโดยใช้ข้อมูลในอดีตกับแบบจำลองทางสถิติรวมถึงการให้คำแนะนำทางเลือกต่าง ๆ และผลของแต่ละทางเลือก

จากปัญหาที่ทางผู้จัดทำเล็งเห็นความสำคัญคือการนำข้อมูลผลการเรียนของนักศึกษาในอดีตมาใช้ประโยชน์ในการบริหารหลักสูตร และ นำมาวิเคราะห์ผลเพื่อช่วยในการวางแผนการเรียนของนักศึกษา ซึ่งการวางแผนในการเรียนของหลักสูตรจะสามารถช่วยอาจารย์และบุคลากรที่เกี่ยวข้องกับการศึกษาในด้านของการบริหารหลักสูตร เพื่อวางแผนการเพิ่มหรือลดจำนวนผู้เรียนในรายวิชาต่าง ๆ ซึ่งส่งผลต่อการผลิตบัณฑิตด้านต่าง ๆ ได้

ดังนั้นผู้จัดทำจึงได้เห็นถึงความสำคัญการประเมินสถานะขอหลักสูตร ของระบบแนะนำการวางแผนการคาดการณ์จากการใช้ความรู้ทางด้าน Data Analytics, Prediction และ Recommendation โดยใช้ข้อมูลผลการเรียนของนักศึกษาในอดีต เพื่อพัฒนาระบบช่วยเหลือ และตอบโจทย์ให้แก่นักศึกษาและบุคคลากรทางการศึกษาหรือบุคคลที่เกี่ยวข้องได้

1.2 วัตถุประสงค์

1) เพื่อนำข้อมูลของผลการเรียนของนักศึกษาในอดีตและข้อมูลจากแบบสำรวจการมีงานทำของบัณฑิตมาใช้ ในการวางแผนการเรียนหรือประเมินอาชีพในอนาคตของนักศึกษาได้
2) ประมวลผลข้อมูลผลการเรียนของนักศึกษาในอดีต และข้อมูลจากแบบสำรวจการมีงานทำของบัณฑิต และทำแผนภาพกราฟิกเพื่อนำเสนอข้อมูล อำนวยความสะดวกให้กรรมการหลักสูตรในการวางแผนการ ทำงาน
3) เพื่อนำข้อมูลผลการเรียนของนักศึกษาในอดีต มาพัฒนาเป็นระบบแนะนำและวางแผนการเรียนตัวของ นักศึกษาได้

4) เพื่อนำข้อมูลการพยากรณ์อาชีพในอนาคตของนักศึกษาในสถาบันมาแสดงเป็นแผนภาพกราฟิกในการ อำนวยความสะดวกให้หน่วยงานภายนอกได้รับทราบว่าหลักสูตรปัจจุบันของสถาบันสามารถผลิต บุคลากรที่มีความชำนาญด้านใดได้บ้าง

1.3 ประโยชน์ของโครงงาน

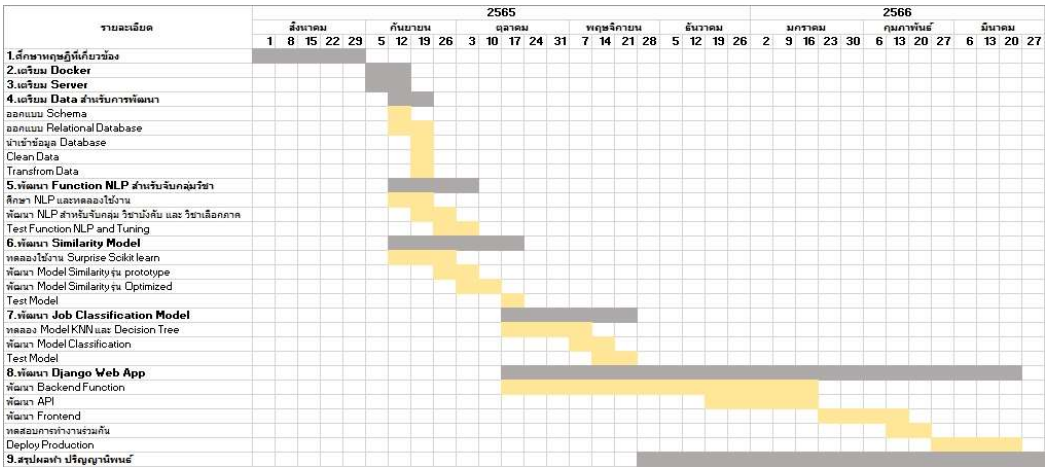
1) ได้ระบบรวบรวมข้อมูลผลการเรียนของนักศึกษาและข้อมูลแบบสำรวจการทำงานของบัณฑิต แล้วนำมาวิเคราะห์และนำเสนอข้อมูลที่เป็นประโยชน์ในการบริหารหลักสูตรของกรรมการหลักสูตร
2) มีระบบที่สามารถแนะนำ วางแผน และประเมินอาชีพในอนาคตจากผลการเรียนของนักศึกษา

1.4 ข้อจำกัดของโครงงาน

1) ข้อมูลผลการเรียนในอดีตย้อนหลังมีเพียง 2 ปี
2) ข้อมูลผลการเรียนในอดีตจะได้จากสำนักทะเบียนและประมวลผล โดยกรรมการหลักสูตรจะเป็นผู้ร้องขอข้อมูลดังกล่าวและนำเข้าระบบ
3) การทำนายต่าง ๆ จะใช้ข้อมูลเพียง 2 แหล่งคือข้อมูลผลการเรียนของนักศึกษาจากสำนักทะเบียนและประมวลผล และแบบสอบถามการมีงานทำของบัณฑิตเท่านั้น

1.5 แผนการดำเนินงาน

แผนการดำเนินงานในการพัฒนาโครงงานตลอดระยะเวลา 2 ภาคการศึกษา ตั้งแต่เดือนสิงหาคม พ.ศ. 2565 - มีนาคม พ.ศ. 2566 แสดงดังรูป 1.1



รูป 0.1 แผนการดำเนินงาน

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 Classification and Prediction

Classification and Prediction คือการจำแนกประเภทของข้อมูล โดยจะนำมาใช้ในการวิเคราะห์ข้อมูล ซึ่งกระบวนการดั่งกล่าวสามารถแบ่งได้เป็น 2 ส่วน

1) Training Data คือการนำข้อมูลมาที่ได้มาทำการเรียนรู้ให้กับคอมพิวเตอร์เพื่อนำไปสร้างเป็นโมเดลแบบจำลองและวัดประสิทธิภาพของโมเดลแบบจำลองนั้น โดยจะทำการสร้างโมเดลซึ่งจะมีด้วยกันหลายวิธี เช่น Decision Tree, Naive Bayes, K Nearest Neighbors และ Neural Network เป็นต้น
2) Predict คือการนำข้อมูลใหม่ที่รับมานำเข้าโมเดลแบบจำลองที่เป็นผลลัพท์จากการผ่านกระบวนการ Training Data ไปทำการคำนวณหรือพยากรณ์

ประเภทของปัญหาในด้าน Classification

1) Binary classification (การจำแนกแบบไบนารี)
   เปรียบให้ดีที่สุดคือ ตัวแปรที่อยู่ในรูปแบบสองหมวดหมู่ เช่น ผลลัพธ์แบบใช่ หรือ ไม่ใช่ ตก หรือ ผ่าน หากเปรียบในรูปแบบของตัวเลขก็คือ 0 กับ 1 อัลกอริทึมที่ใช้คู่กับการจำแนกแบบไบนารี จะมีดังนี้ k-Nearest Neighbors Decision Trees หรือ Naive Bayes
2) Multi-Class Classification (การจำแนกประเภทหลายคลาส)
   ในการจำแนกรูปแบบนี้จะต่างกับการจำแนกแบบไบนารี โดยจะมีหมวดหมู่มากกว่าสอง ตัวอย่างของการจำแนกประเภทนี้ เช่น รูปภาพที่มีองค์ประกอบคล้ายคลึงกับรูปภาพที่อยู่ในฐานข้อมูลเพื่อค้นหาคำศัพท์ที่คาดว่าจะพิมพ์ใน predictive keyboard โดยผลลัพธ์ที่อาจเกิดนั้นจะมีได้มากกว่า 2 หมวดหมู่

อัลกอริทึมที่ใช้คู่ไปกับการจำแนกประเภทนี้สามารถใช้อัลกอริทึมคล้ายกับกับการจำแนกแบบไบนารีได้

3) Multi-Label Classification (การจำแนกประเภทหลายเลเบล)
เปรียบให้เข้าใจง่ายโดยการยกตัวอย่างเช่น รูปภาพรูปหนึ่งสามารถมีรูปดอกไม้ ท้องฟ้าก้อนเมฆได้ แต่รูปภาพรูปนั้นจะจัดว่าเป็นหมวดหมู่รูปวาด รูปถ่าย หรือรูปเสีย Multi-Label Classification ก็คือการทำเลเบลให้กับชุดข้อมูล หรือการติดฉลากให้รูปนั้น ๆ ว่ามีดอกไม้หรือเปล่ามีก้อนเมฆหรือไม่ ส่วน Multi-Class Classification จะจำแนกว่ารูปนั้นเป็นรูปที่เกิดจากการวาดหรือรูปที่เกิดจากการถ่ายหรือรูปเสีย

4) Imbalanced Classification (การจำแนกแบบข้อมูลไม่เท่าเทียม)
คือปัญหาที่เกิดจากข้อมูลที่มีไม่เท่าเทียมกัน (Imbalanced dataset) ตัวอย่างเช่นข้อมูลของการทุจริตโดยข้อมูลส่วนใหญ่ย่อมเป็นข้อมูลที่จัดว่า “ไม่ทุจริต” และจะมีเปอร์เซ็นต์น้อยที่จัดว่าเป็น “ทุจริต” เป็นต้น โดยจะเปรียบโดยง่ายคือกรณีที่ชุดข้อมูลมีการแยกประเภทกันแต่จำนวนของประเภทนั้นมีอัตราส่วนของข้อมูลที่ห่างกันค่อนข้างมาก

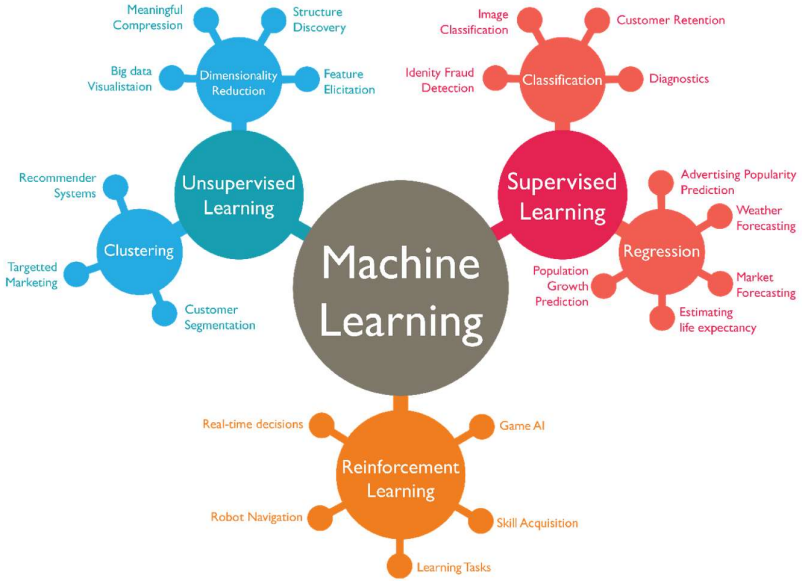
2.1.2 Machine Learning

Machine Learning คือ การทำให้ระบบของคอมพิวเตอร์นั้นสามารถเรียนรู้ได้ด้วยตนเอง โดยจะใช้ข้อมูล ด้วยวิธีการใส่ข้อมูลและผลลัพธ์เข้าไป เพื่อทำให้โปรแกรมนำผลลัพท์นั้นไปประมวลผลและพยากรณ์ Output และ Input ของข้อมูลใหม่ โดยแบ่ง Machine Learning ออกได้เป็น 3 ประเภท คือ

1) Supervised Learning คือการเรียนรู้ที่เครื่องจักรหรือคอมพิวเตอร์นั้นจำเป็นต้องอาศัยข้อมูลใน การฝึกฝน เปรียบเสมือนกับการเรียนการสอนของเด็ก ซึ่งจำเป็นที่จะต้องอาศัยชุดของข้อมูล ซึ่งประกอบไปด้วยชุดของข้อมูลและชุดของผลลัพธ์ของข้อมูลที่ต้องการจะนำมาให้ เครื่องจักรหรือคอมพิวเตอร์ในการเรียนรู้
2) Unsupervised Learning เป็นการเรียนรู้ที่ให้เครื่องจักรหรือคอมพิวเตอร์นั้นสามารถเรียนรู้ได้ด้วยตนเอง โดยไม่จำเป็นต้องมีค่าเป้าหมายของแต่ละชุดข้อมูล ซึ่งวิธีการนี้คือการที่มนุษย์นั้นจะเป็นผู้ใส่ชุดข้อมูล และกำหนดสิ่งที่

ต้องการจากชุดข้อมูลเหล่านั้น โดยให้เครื่องจักรหรือคอมพิวเตอร์วิเคราะห์จากการจำแนกและทำการสร้างแบบแผนจากข้อมูลที่ได้รับมา

3) Reinforcement Learning เป็นการเรียนรู้สิ่งต่าง ๆ ผ่านจากการลองผิดลองถูกภายใต้แนวคิดที่ว่าจะเลือกกระทำสิ่งใดที่ทำให้ได้ผลลัพธ์มากที่สุด โดยจะทำการเรียนรู้จากการลองผิดลองถูกในสถานการณ์ในอดีตหรือระบบจำลองและพยายามที่จะพัฒนาระบบการตัดสินใจของตัวเองให้ดีขึ้นอย่างต่อเนื่องโดยที่อาจจะสามารถพัฒนาด้วยการพยายามสร้างแบบจำลองสถานการณ์ต่าง ๆ ขึ้นมา



รูป 0.1 ประเภทของ Machine Learning

(ที่มา : medium.com, 2018)

2.1.3 Extract-Transform-Load (ETL)

Extract-Transform-Load คือ กระบวนการ กระกวนการหนึ่งซึ่งอยู่ในระบบของ Data Warehouse ซึ่งเป็นระบบที่ออกแบบมาเพื่อที่จะสามารถดึงข้อมูลออกมาจากหลายแหล่ง โดยจะนำกระบวนการตรวจสอบคุณภาพของชุดข้อมูลมาประยุกต์ร่วมใช้ ซึ่งมีการเชื่อมโยงและปรับชุดของข้อมูลให้เป็นไปในรูปแบบเดียวกันทั้งหมดเพื่อให้ ชุดของข้อมูลจากหลากหลายแหล่งสามารถใช้งานร่วมกันได้ และทำการส่งมอบ

1) Extract เป็นกระบวนการเริ่มต้นของระบบที่ดึงข้อมูลจากแหล่งของข้อมูล จะประกอบด้วยข้อมูลจากหลากหลายแหล่งที่มา ข้อมูลที่อยู่ต่างที่กันนั้นอาจจะอยู่ในรูปแบบที่แตกต่างกัน ยกตัวอย่างเช่น อาจจะอยู่ในรูปแบบของฐานข้อมูลคนละชนิด หรือ ไม่ใช่ฐานข้อมูลแท้จริงซึ่งอาจจะเป็นระบบไฟล์ข้อมูลธรรมดา
2) Transforming ขั้นตอนการแปลงรูปแบบของข้อมูลนี้จะมีการใช้กฎหรือฟังชั่น (Function) มากมายเพื่อที่จะแปลงข้อมูลให้ได้อยู่ในรูปแบบตามที่ต้องการก่อนที่จะนำข้อมูลเหล่านั้นเข้าไปยังปลายทาง ข้อมูลจากต้นทางนั้นบางแหล่งข้อมูลมีความจำเป็นน้อยมากหรือแทบจะไม่ต้องการ การแปลงข้อมูลเลย แต่ในบางแหล่งอาจจะต้องการกระบวนการที่ซับซ้อนในการแปลงข้อมูล ซึ่งจะกินทรัพยากรของระบบที่ใช้และเวลาในการประมวลผลของระบบ ซึ่งความซับซ้อนของข้อมูลจะขึ้นอยู่กับความต้องการของเชิงธุรกิจ หรือ เป้าหมายของการนำข้อมูลไปใช้งาน โดยจะมีกระบวนการตัวอย่างต่อไปนี้
    1) Selection คือ การเลือก Column ที่ต้องการที่จะนำไปใช้งานหรือเก็บลงฐานข้อมูล ยกตัวอย่าง เช่น ถ้าต้นทางของข้อมูลมีอยู่ด้วยกัน 3 Column หรือ 3 attributes เช่น enroll_num, age และ salary จะมีการแปลงข้อมูลเกิดขึ้นและ เลือกที่จะไม่มีการแปลงข้อมูลหากพบว่า record นั้นมีค่าของข้อมูล column salary เป็นค่าว่าง
    2) Translation คือ การแปลข้อมูล ตัวอย่างเช่น หากข้อมูลต้นทางนั้นมีการเก็บข้อมูลของเพศโดยให้ 1 เป็นเพศชาย และ 2 เป็นเพศหญิง จะต้องมีการแปลจากชุดตัวเลขที่กำหนดก่อนหน้านี้ให้ 1 = Male และ 2 = Female กระบวนการนี้เรียกว่า data cleaning หรือ กระบวนการทำความสะอาดข้อมูล
    3) Encoding free form ยกตัวอย่างเช่นการ mapping จาก “Male” ไปเป็น “1” และ “Mr” ไปเป็น “M”
    4) Filtering คือ กระบวนการกรองเฉพาะข้อมูลที่กำหนด
    5) Sorting คือ กระบวนการเรียงข้อมูลที่ต้องการ
    6) Joining คือ กระบวนการเชื่อมโยงข้อมูลระหว่างตารางข้อมูล

7) Aggregation คือ กระบวนการรวบรวม และ สรุปชุดข้อมูล ยกตัวอย่าง เช่น การรวมยอด (summarize) ข้อมูลจากหลาย ๆ ระเบียบจนได้มาเป็น ยอดขายรวม เป็นต้น

8) Transposing or pivoting คือการสลับทิศทางตำแหน่งของการแสดง ข้อมูล เช่นการย้ายระเบียบไปเป็น Column หรือ ย้าย Column มาเป็น ระเบียบ เพื่อให้ง่ายต่อการนำข้อมูลไปใช้

3) Loading กระบวนการโหลดข้อมูลเข้า โดยทั่วไปจะนำข้อมูลเข้าไปในระบบ Data Warehouse ทั้งนี้ขึ้นอยู่กับความต้องการขององค์กร หรือ ธุรกิจว่าจะให้ข้อมูลไหลไปในทิศทางใด บางองค์กร หรือ บางงานจะมีการสะสมของข้อมูล ความถี่ของการนำข้อมูลเข้าสู่ระบบ อาจจะมีการล้างข้อมูลแล้วทับข้อมูลใหม่ โดยทั่วไปแล้วข้อมูลของ Data Warehouse จะมีการใช้กันปีต่อปี เมื่อขึ้นปีใหม่แล้วจะมีการล้างข้อมูลของปีเก่า และ เก็บไว้ในระบบข้อมูลสำรอง เนื่องจากว่ากระบวนการนำข้อมูลเข้าจะต้องปฏิสัมพันธ์กับฐานข้อมูล (Database) ดังนั้นจะต้องมีประเด็นเรื่องของ Database Constraints, Referential Integrity, Database Trigger เข้ามาเกี่ยวข้องด้วยในกระบวนการนำข้อมูลเข้า ซึ่งสิ่งเหล่านี้รวม ๆ แล้วเรียกว่า กระบวนการควบคุมคุณภาพของข้อมูล (Data Quality performance of E-T-L process)

2.1.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) เป็นเครื่องมือที่ให้คอมพิวเตอร์เข้าใจภาษาของมนุษย์ที่มีความซับซ้อน เป็นศาสตร์หนึ่งที่สำคัญทางด้าน Machine Learning โดยเป็นสาขาวิชาหนึ่งที่ประกอบด้วยองค์ความรู้จากหลากหลายแขนง อาทิ ภาษาศาสตร์ (Linguistics) วิทยาการคอมพิวเตอร์ (Computer Science) ปัญญาประดิษฐ์ (Artificial Intelligence: AI) รวมไปถึงสถิติ (Statistics) โดย NLP มีมาตั้งแต่ช่วงกลางศตวรรษที่ 19 และมีการพัฒนามาเรื่อย ๆ จนถึงปัจจุบัน โดยแบ่งออกเป็น 3 ยุค ดังนี้

1) ยุค Rule-based Method (ช่วง ค.ศ.1950-1990)

ในยุคแรกของ NLP มีการใช้งานตามกฎ (Rule-based Method) โดยนักภาษาศาสตร์ผู้มีความเชี่ยวชาญโครงสร้างของภาษาที่สนใจ จะเป็นผู้เขียน

กฎต่าง ๆ ขึ้นมาเพื่อให้คอมพิวเตอร์สามารถคำนวณข้อความของโจทย์ต่างๆ ได้

2) ยุค Machine Learning (ช่วง ค.ศ.1990-2010)
ในยุคนี้ พบว่ามีการเขียนกฎด้วยมือไม่สามารถตอบโจทย์ที่มีความซับซ้อนได้ จึงมีสิ่งที่ได้มาทดแทนในยุคนี้คือ ความสามารถของเครื่องคอมพิวเตอร์ รวมถึงความรู้ทางด้านสถิติ และ Machine Learning ซึ่งได้ถูกนำมาพัฒนาเพื่อใช้ในการทำงานด้าน NLP โดยมีการนำเข้าข้อมูลเพื่อให้คอมพิวเตอร์สามารถเรียนรู้ด้วยตนเองแทนการใช้ผู้เชี่ยวชาญทางด้านภาษา

3) ยุค Deep Learning (ช่วง ค.ศ.2010-ปัจจุบัน)
ในยุคปัจจุบัน ด้วยพลังการคำนวนของคอมพิวเตอร์ที่มีการพัฒนาสูงขึ้นอย่างต่อเนื่อง ทำให้เทคโนโลยีที่มีความซับซ้อนสูงอย่าง การเรียนรู้เชิงลึก (Deep Learning) ถูกนำมาใช้งานแทนที่ Machine Learning ซึ่งใช้ความรู้ทางด้านสถิติแบบดั้งเดิมอย่างแพร่หลายมากขึ้น รวมถึงในงานด้าน NLP ด้วยเช่นกัน อาทิ การสร้างแบบจำลองทางภาษา (Language Model) และการวิเคราะห์โครงสร้างของข้อความ (Parsing)

ตัวอย่างการประยุกต์ใช้ NLP ในด้านต่าง ๆ

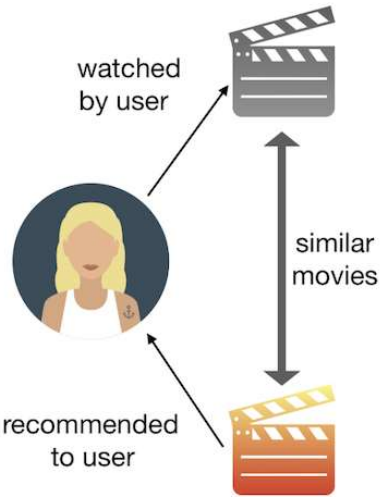
1) ด้านการทำงานวิจัย การวิจัยมีแหล่งของข้อมูลทางภาษาขนาดใหญ่ ซึ่งทำให้ NLP สามารถเข้ามามีบทบาทได้อย่างหลากหลาย ตัวอย่างเช่น การใช้ Topic Model ในการจัดหมวดหมู่บทความ
2) ด้านพาณิชย์อิเล็กทรอนิกส์ การซื้องของผ่านช่องทางออนไลน์ เข้ามามีบทบาทสำคัญเป็นอย่างมากในระบบเศรษฐกิจ ซึ่งทำให้เกิดปริมาณธุรกรรมขนาดใหญ่ ไม่ว่าจะเป็น คำอธิบายสินค้าและบริการ การแสดงความคิดเห็นของผู้บริโภค รวมถึงการสนทนากันระหว่างผู้ซื้อและผู้ขายผ่านทางช่องแชท
3) ด้านการแพทย์ ข้อมูลทางการแพทย์มีการบันทึกข้อมูลด้วยข้อความ ตัวอย่างเช่น บทสนทนาระหว่างแพทย์และผู้ป่วย การวินิจฉัยโรคโดยแพทย์ และประวัติการรักษาของผู้ป่วย
4) ด้านกฎหมาย สำหรับงานด้าน มีข้อมูลทางด้านภาษาที่แตกต่างและหลากหลาย เช่นเดียวกัน เช่น ประมวลกฎหมายต่าง ๆ คำร้องต่อศาล คำให้การของคู่ความ และคำพิพากษาของศาล ซึ่งสามารถประยุกต์ใช้

เครื่องมือ NLP ได้ในหลายมิติไม่ว่าจะเป็นการใช้ PoS Tagging และ NER เพื่อช่วยในการตีความประมวลกฎหมาย

2.1.5 Recommendation System

Recommendation System เป็นระบบที่จะทำการแนะนำสิ่ง (item) ที่ “เหมาะสม” ให้แก่ผู้ใช้ โดย item เป็นได้ตั้งแต่ ข่าว เนื้อหา เพลง course เรียน ไปจนถึงสินค้าที่ขายในร้าน online โดยสามารถแนะนำสิ่งที่ผู้ใช้สนใจได้ผ่านโมเดลที่ส่วนใหญ่มักจะถูกใช้กันมีอยู่ด้วยกันสามประเภท ได้แก่

1) Content-based Filtering เป็นรูปแบบของโมเดลที่จะแนะนำลักษณะของตัวบริการหรือสินค้าเป็นตัวตั้งและทำการแนะนำสิ่งค้าที่มีลักษณะที่คล้ายกัน

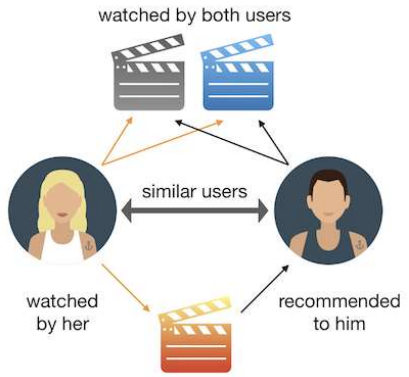


รูป 2.6 รูปแบบของ Content-based Filtering

(ที่มา : towardsdatascience.com, 2018)

2) Collaborative Filtering เป็นรูปแบบโมเดลที่เรียนรู้จากพฤติกรรมของผู้ใช้กับผู้ใช้คนอื่น ๆ ที่คล้ายคลึงกัน

1) Memory-based เป็นการดูข้อมูลแล้วหาความสัมพันธ์ ระหว่างผู้ใช้หรือสินค้าจากข้อมูลโดยตรง



รูป 2.6 รูปแบบของ Memory-based

(ที่มา : towardsdatascience.com, 2018)

2) Model-based ใช้เทคนิคของ machine learning เพื่อหา user embedding และ item embedding มาทำการทำนาย rating ที่ผู้ใช้จะให้กับสินค้า หรือ relevance score

3) Hybrid ใช้หลาย ๆ วิธีการมารวมกัน Hybrid system เป็นการมัดรวมทั้งสองอัลกอริทึมของ Model-based และ Memory-based เอาไว้เพื่อทำให้ระบบการแนะนำสมบูรณ์ขึ้น ซึ่งระบบนี้ถูกนำไปใช้ในปัจจุบันมากที่สุด แทบจะทุกแพลทฟอร์มใหญ่ที่มีการแนะนำสินค้าและบริการ

3) Hybrid system เป็นการมัดรวมทั้งสองระหว่าง Content-based Filtering และ Collaborative Filtering เพื่อทำให้ระบบการแนะนำสมบูรณ์ขึ้น

2.2 เครื่องมือที่เกี่ยวข้อง

2.2.1 Docker

Docker เป็นเครื่องมือแบบ open-source ที่ช่วยจำลองสภาพแวดล้อม ในการรัน service หรือ server โดยการสร้าง container เพื่อจัดการกับ library ต่างๆ และยังช่วยจัดการในเรื่องของ version control เพื่อให้ง่ายต่อการจัดการกับปัญหาต่างๆ ที่เกิดขึ้นองค์ประกอบต่างๆ ของ Docker

1) Docker image

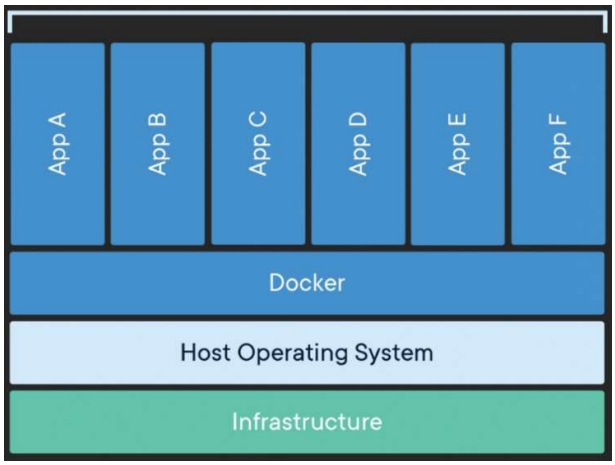
   คือต้นแบบของ Container โดยข้างในจะเป็นระบบปฏิบัติการ Linux ที่มีการติดตั้ง Application และ มีการ Configuration เอาไว้ ซึ่งเกิดจากการ build ไฟล์ Docker file ขึ้นมาเป็น image

2) Docker container

   Docker container จะถูกสร้างมาจาก Docker Image ที่เป็นต้นแบบหรือแม่พิมพ์ เกิดเป็น container และจะได้ Service หรือ Application ที่สามารถเรียกใช้งานได้ทันที

3) Docker registry

   การสร้าง Docker Image แล้วนำไปเก็บรวบรวมไว้บน server (ลักษณะเดียวกับการเก็บ Source Code ไว้บน (Github) โดย Docker registry ณ ปัจจุบันก็มีให้เลือกใช้งานได้อย่างหลากหลายโดยมี Docker Hub เป็น Docker registry หลักในการเรียกใช้(pull) Docker Image และนอกจากนี้ยังมีผู้ให้บริการ docker registry อื่นๆด้วย เช่น Gitlab, Quay.io, Google Cloud เป็นต้น



รูป 2.6 การทำงานของแอพพลิเคชันต่าง ๆ บน Docker Engine

(ที่มา : docker.com)

2.2.2 Django

Django Framework เป็นชุดของเครื่องมือ Framework สำหรับ การนำไปพัฒนาเว็บไซต์ด้วยภาษาของ Python โดยทุกวันนี้ Framework สำหรับการเขียนเว็บไซต์ด้วยภาษา Python มีค่อนข้างที่จะเยอะ ซึ่ง Django Framework ก็เป็นหนึ่งใน Framework สำหรับการพัฒนาเว็บไซต์ และทำเว็บไซต์ด้วยภาษา Python ด้วยเช่นกัน

คุณสมบัติของ Django Framework

1) Object-relational mapper คือ การกำหนด Data Model ในภาษา Python เพื่อใช้ในการทำงานด้านข้อมูล และช่วยสนับสนุน dynamic database-access API
2) Automatic admin interface คือ ส่วนในการสร้าง Interface อัตโนมัติสำหรับการ add, edit, delete และ search ด้วย Django Framework
3) Elegant URL design คือ การทำให้ URL มีความสั้น กระชับ สวยงาม และสื่อความหมายของหน้านั้น ๆ ได้อย่างชัดเจน
4) Template system คือ Django นั้นมีการออกแบบ Template Language เพื่อการเขียนแยกส่วนระหว่าง Design และ Business Logic

5) Cache system คือ ส่วนของการบันทึก หรือจัดการข้อมูลที่มีการดาวน์โหลดไปแล้ว เพื่อเพิ่มประสิทธิภาพการทำงานของเว็บไซต์ด้านความเร็ว และด้านอื่น ๆ

6) Internationalization คือ Django สนับสนุน Application ที่มีความหลากหลายด้านภาษาในการแสดงผล

2.2.3 Scikit-learn

Scikit-learn เป็นโมดูลหนึ่งของภาษา Python เป็นแพ็กเกจที่รวบรวม Library ด้าน การเรียนรู้ของเครื่อง (Machine Learning) เอาไว้ และถูกออกแบบมาให้ทำงานร่วมกับ Library ของภาษา Python อย่าง NumPy และ SciPy ได้ดี

Scikit-learn ยังเป็น Open Source ที่เปิดให้สามารถเข้าไปพัฒนาต่อยอดได้และเป็นแหล่งรวม Library และอัลกอริทึมที่เน้นไปในด้านของ การเรียนรู้ของเครื่อง (Machine Learning) ซึ่งมีส่วนในการทำ แบบจำลองข้อมูล (Data Modeling) อีกหนึ่งปัจจัยที่ทำให้มีผู้ใช้เยอะ เพราะเป็น Interface ระดับสูง ทำให้มือใหม่สามารถเข้าใจภาพรวมและ ขั้นตอนการทำงาน ของการเรียนรู้ของเครื่อง (Machine Learning) ได้เครื่องมือที่ผู้ใช้งานสามารถนำไปใช้ในได้

2.2.4 MariaDB

MariaDB คือ เป็น Open Source สำหรับจัดการกับฐานข้อมูล MariaDB เป็นหนึ่งในฐานข้อมูลที่ได้รับความนิยมมากที่สุดในโลก MariaDB ถูกพัฒนาขึ้นโดยนักพัฒนาเดิมของ MySQL เนื่องจากความกังวลที่เกิดขึ้นเมื่อ MySQL ถูกซื้อโดย Oracle Corporation ในปี 2009 ตอนนี้นักพัฒนาและผู้ดูแลของ MariaDB ได้รวมรายเดือนกับฐานรหัส MySQL เพื่อให้แน่ใจว่า MariaDB มีการแก้ไขข้อบกพร่องที่เกี่ยวข้องเพิ่มลงใน MySQL

MariaDB ได้รับการพัฒนาเป็นซอฟต์แวร์โอเพ่นซอร์ส และเป็นฐานข้อมูลเชิงสัมพันธ์แบบ SQL สำหรับการเข้าถึงข้อมูล เวอร์ชันล่าสุดของ MariaDB มีคุณลักษณะ GIS และ JSON ด้วย

MariaDB เปลี่ยนข้อมูลเป็นฐานข้อมูลที่มีโครงสร้างในหลากหลายแอพพลิเคชั่น ตั้งแต่ธนาคารไปจนถึงเว็บไซต์ต่างๆ เป็นการปรับปรุงและแทนที่ด้วยการแทนที่ของ MySQL เนื่องจากมีความรวดเร็วและสามารถปรับขนาดได้และมีระบบแวดล้อมที่อุดมไปด้วยปลั๊กอินเอนจินและเครื่องมืออื่น ๆ ทำให้สามารถใช้งานได้หลากหลาย

2.2.5 React

React เป็น JavaScript library ที่ใช้สำหรับสร้าง user interface ที่ให้เราสามารถเขียนโค้ดในการสร้าง UI ที่มีความซับซ้อนแบ่งเป็นส่วนเล็กๆออกจากกันได้ ซึ่งแต่ละส่วนสามารถแยกการทำงานออกจากกันได้อย่างอิสระ และทำให้สามารถนำชิ้นส่วน UI เหล่านั้นไปใช้ซ้ำได้

2.2.6 Node.JS

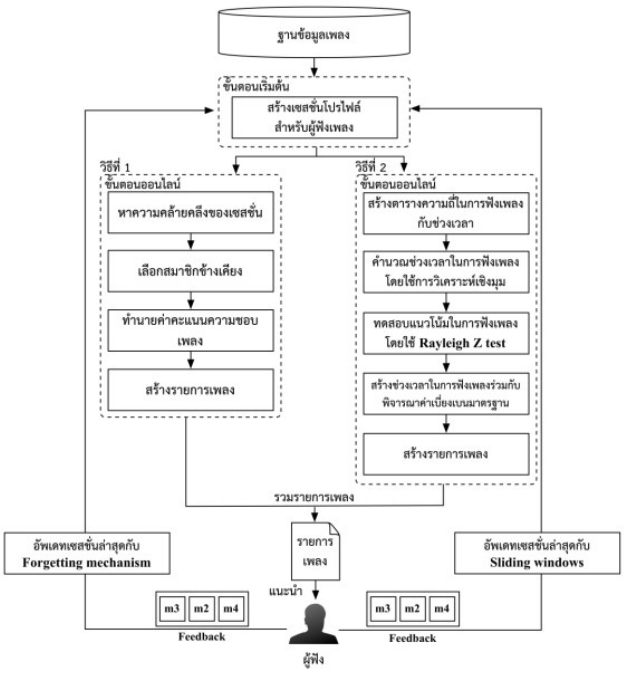
Node.js คือสภาพแวดล้อมการทำงานของภาษา JavaScript นอกจากเว็บเบราว์เซอร์ที่ทำงานด้วย V8 engine นั่นหมายความว่าเราสามารถใช้ Node.js ในการพัฒนาแอพพลิเคชันแบบ Command line แอพพลิเคชัน Desktop หรือแม้แต่เว็บเซิร์ฟเวอร์ได้ โดยที่ Node.js จะมี APIs ที่จะสามารถใช้สำหรับทำงานกับระบบปฏิบัติการ เช่น การรับค่าและการแสดงผล การอ่านเขียนไฟล์ และการทำงานกับเน็ตเวิร์ก และยังเป็นเป็นโปรแกรมที่สามารถใช้ได้ทั้งบน Windows, Linux และ Mac OS X โดยสามารถเขียนโปรแกรมในภาษา JavaScript และนำไปรันได้ทุกระบบปฏิบัติการที่สนับสนุนโดย Node.js

2.3 งานวิจัยที่เกี่ยวข้อง

2.3.1 การสร้างรายการเพลงโดยใช้การกรองร่วมแบบเซสชั่นที่เพิ่มขึ้นด้วยกลไกการลืมและการวิเคราะห์สถิติเชิงมุม

สุเมธ ดาราพิสุท นำเสนองานวิจัยเรื่อง การสร้างรายการเพลงโดยใช้การกรองร่วมแบบเซสชั่นที่เพิ่มขึ้นด้วยกลไกการลืมและการวิเคราะห์สถิติเชิงมุม โดยใช้ 2 วิธีร่วมกัน 1 การสร้างรายการเพลงจะพิจารณาการฟังเพลงในเซสชั่นปัจจุบันที่คล้ายกับเซสชั่นในอดีตของผู้ฟัง 2 สร้างรายการเพลงแนะนำโดยพิจารณาช่วงเวลา เฉพาะในการฟังเพลงซึ่งแตกต่างจากช่วงเวลาอื่นอย่างมีนัยสำคัญทางสถิติในรอบวันของผู้ฟังโดยใช้ การวิเคราะห์สถิติเชิงมุม และวัดประสิทธิภาพโดย ประสิทธิภาพ HitRatio และ Precision จากการทดลองพบว่าการใช้ 2 วิธีแยกกันนั้นได้ผลลัลพท์ที่น้อยกว่านำมาใช้ร่วมกัน 0.18-0.22 % โดยวัตถุประสงค์ในการทำเพื่อวิเคราะห์ปัญหาที่เกิดขึ้นในการสร้างรายการเพลงแนะนำแบบออฟไลน์ พัฒนาขั้นตอนวิธีการสร้างรายเพลงแนะนำให้มีประสิทธิภาพเพิ่มขึ้นทั้งทางด้านความเร็วและความถูกต้องในการสร้างรายการเพลง

โดยในโครงงานของผู้จัดทำนั้นได้นำสวนของการออกแบบ Diagram ในงานวิจัยนี้มาใช้งาน โดยใช้วิธีที่ 1ซึ่งของผู้จัดทำจะเป็น 1 หาความคล้ายคลึงของหมวดหมู่วิชา 2 เลือกสมาชิกข้างเคียง 3 ทำนายค่าผลลัพท์การเรียนหรือเกรด 4 นำไปสร้างรายการสำหรับขั้นตอนต่อไป



รูป 0.2 ประเภทของ Machine Learning

(ที่มา : ดาราพิสุทธิ์, 2016**)**

2.3.2 การสร้างรายการเพลงโดยใช้การกรองร่วมแบบเซสชั่นที่เพิ่มขึ้นด้วยกลไกการลืมและการวิเคราะห์สถิติเชิงมุม

นิภาภรณ์ พันธ์นาม นำเสนองานวิจัย ระบบแนะนำสินค้าอาหารโดยใช้ระบบแนะนำแบบผสมผสาน ใช้เทคนิค Content based filtering แบบหลักการ Cosine และสร้างแบบจำลองโดยใช้ lib Surprise ซึ่งมีอัลกอริทึม SVD, NMF, Baseline และ KNN และวัดประสิทธิภาพโดย RMSE, MAE จากการทดลองพบว่า 1 เทคนิคการกรองแบบอิงเนื้อหานำวิธีการ TF-IDF เข้ามาช่วยในการทำ Vectorization ส่วนใหญ่ค่าความเหมือนออกมาค่อนข้างที่จะตำเนื่องมาจากข้อมูลที่น้อยเกินไป 2 เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ผ่าน library Surprise ของ Scikitlearn ซึ่งโมเดลที่มีผลคะแนนโดยรวมดีที่สุดคืออัลกอริทึมของ SVD ซึ่งได้ค่ำ RMSE 1.2528 และ MAE 0.9376 และ 3 ระบบแนะนำแบบผสมผสาน โดยผลลัพท์นั้นจะไม่ชัดเจนเนื่องจากวิธีนี้ได้มีการทำนายค่า Rating ซึ่งวิธีการของระบบแนะนำแบบผสมผสานนั้น ได้มีนำเทคนิคการกรองแบบอิงเนื้อหา ที่ไม่ได้มีการทำนายค่าอะไรมารวมในการทำงานของแบบจำลองด้วย ซึ่งถ้าต้องการวัดผลลัพท์สามารถอ้างอิงจากค่า RMSE, MAE ได้

โดยในงานโครงงานของผู้จัดทำนั้นได้นำผลลัพท์การทดลองของงานวิจัยนี้ที่สรุปได้ว่าผลคะแนนโดยรวมดีที่สุดคืออัลกอริทึม SVD เป็นตัวตัดสินในการเลือกใช้อัลกอริทึมนี้และได้นำวิธีการการวัดประสิทธิภาพของแบบจำลองนี้จากงานวิจัยมาปรับใช้ในรูปแบบเดียวกันกับตัวโครงงาน