การทำนายยอดการดูวิดีโอโดยใช้การแบ่งกลุ่มยอดการดูวิดีโอและแบบจำลองเชิงเส้นหลายตัวแปร

นายเอกพล วงศ์ศุภรัตน์กุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Prediction of View Count of Online Videos Using Clustering View Pattern with Multivariate Linear Model

Mr. Ekapol Wongsuparatkul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

หัวข้อวิทยานิพนธ์	การทำนายยอดการดูวิดีโอโดยใช้การแบ่งกลุ่มยอดการดูวิดีโอและแบบจำลองเชิงเส้นหลายตัวแปร
โดย	นายเอกพล วงศ์ศุภรัตน์กุล
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(รองศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ)

เอกพล วงศ์ศุภรัตน์กุล : การทำนายยอดการดูวิดีโอโดยใช้การแบ่งกลุ่มยอดการดูวิดีโอและแบบจำลองเชิงเส้นหลายตัวแปร. (Prediction of View Count of Online Videos Using Clustering View Pattern with Multivariate Linear Model) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุกรี สินธุภิญโญ

ในงานวิจัยนี้ เราตั้งเป้าหมายในการออกแบบแบบจำลองที่ทำนายยอดการดูระยะสั้นของวิดีโอบนยูทูบ เราเสนอแบบจำลองเอฟ7เอ็นเอ็มแอลซึ่งเป็นแบบจำลองที่สามารถจัดกลุ่มรูปแบบยอดการดูวิดีโอและกำจัดรูปแบบที่ผิดปกติ แบบจำลองนี้ประกอบด้วย 4 อย่าง อย่างแรกคือการจัดกลุ่มรูปแบบโดยใช้แบบจำลองการจัดกลุ่ม จากนั้นกลุ่มที่มีจำนวนน้อยซึ่งถูกกำหนดเป็นรูปแบบที่ไม่ได้เกิดขึ้นบ่อยจะถูกกำจัดออกไป ต่อมาจัดกลุ่มรูปแบบวิดีโอจากชุดข้อมูลทดสอบโดยใช้แบบจำลองเพื่อนบ้านใกล้เคียงที่สุด 1 อันดับ อย่างสุดท้ายคือรูปแบบแต่ละกลุ่มจะกลายเป็นชุดข้อมูลสำหรับแบบจำลองเชิงเส้นหลายตัวแปรซึ่งนำไปใช้ฝึกฝนเฉพาะกลุ่ม ผลการทดลองพบว่าแบบจำลองเอฟ7เอ็นเอ็มแอลที่ใช้แบบจำลองการจัดกลุ่มที่เหมาะสมทำให้ค่าความผิดพลาดจากการทำนายยอดการดูในวันที่ 30 ลดลง 27% จากแบบจำลองที่ดีที่สุดที่นำมาเปรียบเทียบจากงานวิจัยอื่น

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....
ปีการศึกษา 2562 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6170983121 : MAJOR COMPUTER SCIENCE
KEYWORD: clustering, regression, video view prediction, YouTube

Ekapol Wongsuparatkul : Prediction of View Count of Online Videos Using Clustering View Pattern with Multivariate Linear Model. Advisor: Asst. Prof. SUKREE SINTHUPINYO

In this research, we aim to design a model, which accurately predicts the short-term view count of videos on YouTube. We present F7NML, the First 7-day Normalization for clustering with Multi-variate Linear model, a predictive model that can group the patterns and remove outliers. First, it groups the patterns into many groups using the clustering model, which is presented in the paper. Then, it removes the groups of rare patterns, which are called outliers. Next, the video view count in the test dataset is matched into the groups using 1-nearest neighbor. Finally, Multivariate Linear model is trained for each group specifically. The experimental results show that F7NML with an appropriate clustering model reduces error when it was compared to the best baseline model from the literature by about 27% on the 30th-day view count prediction.

Field of Study: Computer Science Student's Signature .....
Academic Year: 2019 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์นี้ไม่มีทางสำเร็จได้ถ้าไม่มีอาจารย์ที่ปรึกษาอย่างผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ ซึ่งเป็นผู้ให้คำแนะนำเกี่ยวกับการแนะนำแนวทางการวิจัย การเก็บข้อมูลบนยูทูบ การตรวจโครงร่างวิทยานิพนธ์ งานวิจัย และวิทยานิพนธ์ เพื่อให้วิทยานิพนธ์นี้มีคุณภาพ การแนะนำเรื่องการทดสอบแบบจำลอง และการให้กำลังใจในการทำวิจัย ขอขอบพระคุณครับ

นอกจากนี้ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.นัททที นิภานันท์ สำหรับให้เกียรติเป็นประธานกรรมการ รองศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ และผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ สำหรับให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ ทั้ง 3 ท่านเป็นผู้ให้คำแนะนำการปรับปรุงวิทยานิพนธ์นี้ให้ดีและมีคุณภาพยิ่งขึ้น

สุดท้ายนี้ขอขอบคุณทุกคนที่ให้กำลังใจและให้คำแนะนำในการทำวิทยานิพนธ์ ทำให้ผ่านพ้นอุปสรรคในการทำวิทยานิพนธ์ไปได้ด้วยดี

เอกพล วงศ์ศุภรัตน์กุล

สารบัญ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันโซเชียลเน็ตเวิร์ก (Social Network) เป็นที่นิยมของคนทั่วโลก เป็นสิ่งที่สามารถใช้เผยแพร่ข้อมูล ใช้แสดงความคิดเห็น ให้ผู้ที่อยู่ในโซเชียลเน็ตเวิร์กได้รับรู้ เป็นสิ่งที่ใช้เชื่อมต่อระหว่างตัวเราเองกับผู้อื่นผ่านโซเชียลมีเดียและเป็นสิ่งที่ใช้เผยแพร่ข่าวสารได้ ประมาณ 64.5% ของผู้ที่ใช้อินเทอร์เน็ตรับข่าวล่าสุดจากโซเชียลมีเดีย (Social Media) แทนการอ่านหนังสือพิมพ์หรือดูโทรทัศน์ และ 50% ของผู้ใช้อินเทอร์เน็ตได้รับข่าวล่าสุดผ่านโซเชียลมีเดียก่อนที่จะถูกรายงานผ่านโทรทัศน์

โซเชียลมีเดียต่าง ๆ ที่เป็นที่นิยมในตอนนี้มีลักษณะการใช้งานแตกต่างกัน เช่น เฟซบุ๊ก (Facebook) สามารถใช้ได้ในเชิงส่วนบุคคลและเชิงธุรกิจ ความนิยมของเนื้อหาวัดจาก เช่น จำนวนไลก์ จำนวนความคิดเห็น เป็นต้น ยิ่งสิ่งเหล่านี้มีมากยิ่งปรากฏอยู่บนเฟซบุ๊กนานขึ้น ทวิตเตอร์ (Twitter) เป็นโซเชียลมีเดียที่เป็นที่นิยมโดยใช้แฮชแท็ก (#) เป็นตัวที่เข้าถึงข่าวสารของผู้ใช้อื่น ๆ สามารถเผยแพร่เนื้อหาได้มากขึ้น โดยการที่ผู้อื่นแชร์หรือรีทวิตเนื้อหา

ยูทูบ (YouTube) ก็เป็นโซเชียลมีเดียทางด้านการแชร์วิดีโอซึ่งมีหลากหลายหมวดหมู่ เช่น เพลง เกม ภาพยนตร์ ข่าว การศึกษา เป็นต้น สามารถแชร์วิดีโอได้ผ่านโซเชียลมีเดียอื่น ๆ จากสถิติของยูทูบพบว่ามีผู้ใช้ถึง 1 พันล้านคนซึ่งเป็น 1 ใน 3 ของผู้ที่ใช้อินเทอร์เน็ตทั่วโลก ซึ่งในแต่ละเดือนมีผู้ใช้ที่ลงชื่อเข้าสู่ระบบกว่า 1.9 พันล้านครั้งเข้าชมยูทูบและทุก ๆ วันมีผู้คนเข้าดูยูทูบมากกว่าพันล้านชั่วโมงและมีการดูนับพันล้านครั้ง แสดงให้เห็นว่ายูทูบเป็นโซเชียลมีเดียที่เป็นที่นิยมที่สุดในโลกทางด้านการแชร์วิดีโอ ในการดูวิดีโอบนยูทูบแต่ละครั้งจะมีค่าต่าง ๆ ที่น่าสนใจ เช่น จำนวนการดู จำนวนการกดถูกใจหรือไม่ถูกใจ การแสดงความคิดเห็นของผู้ที่ดูวิดีโอ จำนวนผู้ติดตามของช่องที่บรรจุขึ้นวิดีโอ เป็นต้น สิ่งเหล่านี้สามารถบ่งบอกถึงความนิยมของวิดีโอที่กำลังดูอยู่ ถ้าสามารถสร้างยอดการดูได้เร็ว ยิ่งมีโอกาสติดแท็บมาแรงบนยูทูบได้มากขึ้น และมีโอกาสที่วิดีโอจะไปอยู่หน้าแรกของยูทูบได้มากขึ้น การที่ผู้อื่นมาดูวิดีโอมากสามารถสร้างรายได้ให้กับเจ้าของช่องที่บรรจุขึ้นวิดีโออีกด้วย จึงเกิดการทำวิดีโออย่างจริงจังเพื่อดึงดูดให้ผู้อื่นมาดู นอกจากนี้ยังพบผู้ที่ประกอบอาชีพบนยูทูบ เช่น ผู้ให้ความบันเทิง เน้นการสร้างเนื้อหาของวิดีโอเพื่อดึงดูดให้ผู้อื่นมาดู มีทีมงานสำหรับการตัดต่อวิดีโอเพื่อให้วิดีโอมีคุณภาพยิ่งขึ้น นักสตรีมเกมก็เป็นอีกอาชีพที่เน้นอัดวิดีโอตอนเล่นเกมให้ผู้อื่นมาดูเพื่อความบันเทิง บางครั้งจะมีการสตรีมเกมสดเพื่อให้ผู้อื่นมาดูได้แบบถ่ายทอดสด และให้คนดูสามารถบริจาคเงินให้กับนักสตรีมเกมได้

การคาดการณ์รายได้จากการดูวิดีโอบนยูทูบเป็นเรื่องยาก เพราะยอดการดูวิดีโอเป็นปัจจัยที่ทำให้เกิดรายได้ ซึ่งการคาดการณ์ยอดการดูวิดีโอนั้นทำได้ยาก ถ้าสามารถคาดการณ์ยอดการดูวิดีโอได้ ก็จะสามารถประมาณรายได้จากวิดีโอนั้นได้ ก่อให้เกิดรายได้ที่เพิ่มขึ้น และทำให้สามารถวางแผนพัฒนาเนื้อหาของวิดีโอต่อ ๆ ไปเพื่อให้มียอดการดูวิดีโอเพิ่มขึ้นอีกด้วย

ในงานวิจัยนี้จะสร้างแบบจำลองสำหรับการทำนายยอดการดูวิดีโอโดยใช้การเรียนรู้ของเครื่อง (Machine Learning) ที่รับยอดการดูวิดีโอใน 7 วันแรกและวันที่ใช้ทำนายเป็นข้อมูลสำหรับการแบ่งกลุ่มลักษณะของยอดการดูวิดีโอสะสมและรับยอดการดูวิดีโอใน 7 วันแรกสำหรับสร้างแบบจำลองเชิงเส้นหลายตัวแปรของแต่ละกลุ่ม จากนั้นนำแบบจำลองมาเปรียบเทียบกับแบบจำลองในงานวิจัยอื่น ๆ

1.2 วัตถุประสงค์

1.2.1 เพื่อสร้างแบบจำลองที่มีความแม่นยำยิ่งขึ้นในการทำนายยอดการดูสะสมของวิดีโอเมื่อเปรียบเทียบกับแบบจำลองอื่น

1.2.2 เพื่อเป็นตัวช่วยสำหรับระบบแนะนำวิดีโอ ที่จะแนะนำวิดีโอที่มีแนวโน้มว่าจะมียอดการดูมาก

1.2.3 เพื่อเป็นตัวช่วยสำหรับคาดการณ์รายได้จากวิดีโอล่วงหน้าของผู้ที่หารายได้จากวิดีโอบนยูทูบ

1.3 ขอบเขตการดำเนินงาน

1.3.1 วิดีโอเป็นวิดีโอบนยูทูบที่เก็บรหัสระบุของวิดีโอผ่านยูทูบเอพีไอ ตั้งแต่วันที่ 14 กรกฎาคม 2562 เวลา 14.28น. ถึงวันที่ 15 กรกฎาคม 2562 เวลา 13.28น.

1.3.2 รหัสระบุของวิดีโอในข้อ 1 เป็นของวิดีโอที่บรรจุขึ้นก่อนหน้าที่จะเก็บรหัสระบุในช่วง 1 ชั่วโมง

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 สร้างแบบจำลองที่สามารถทำนายยอดการดูสะสมของวิดีโอได้แม่นยำยิ่งขึ้น

1.4.2 ช่วยให้ระบบแนะนำวิดีโอ สามารถแนะนำวิดีโอที่มีแนวโน้มว่าจะมียอดการดูมากได้

1.4.3 ช่วยคาดการณ์รายได้ล่วงหน้าของผู้ที่หารายได้จากการอัปโหลดวิดีโอบนยูทูบ

1.5 ขั้นตอนการดำเนินงาน

1.5.1 ศึกษาการเก็บข้อมูลของวิดีโอบนยูทูบผ่านยูทูบเอพีไอและเซเลเนียม

1.5.2 ศึกษาการใช้งานการประมวลผลกลุ่มก้อนเมฆ

1.5.3 เก็บข้อมูลต่าง ๆ ของวิดีโอ

1.5.4 ศึกษางานวิจัยที่เกี่ยวกับแบบจำลองการทำนายความนิยมของวิดีโอ

1.5.5 สร้างแบบจำลองสำหรับการแบ่งกลุ่มยอดการดูวิดีโอสะสม

1.5.6 สร้างแบบจำลองสำหรับการจัดกลุ่มยอดการดูวิดีโอสะสม

1.5.7 สร้างแบบจำลองเชิงเส้นหลายตัวแปรสำหรับการทำนายยอดการดูสะสมของวิดีโอ

1.5.8 ทดสอบและวัดค่าความผิดพลาดของแบบจำลองพื้นฐานและแบบจำลองในงานวิจัย

1.5.9 วิเคราะห์ผลการทดลอง

1.5.10 สรุปผลและเรียบเรียงวิทยานิพนธ์

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การแบ่งกลุ่ม (Clustering)

การแบ่งกลุ่มเป็นวิธีการของการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ในการเรียนรู้ของเครื่อง เป็นเทคนิคการแบ่งข้อมูลในชุดข้อมูลเป็นกลุ่มที่ขึ้นอยู่กับวิธีการของเทคนิคที่เลือกใช้

2.1.1 การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means Clustering)

การแบ่งกลุ่มข้อมูลแบบเคมีน [20] เป็นวิธีการแบ่งกลุ่มที่ต้องกำหนดจำนวนกลุ่มที่ต้องการเป็น k กลุ่ม แต่ละกลุ่มจะมีจุดเซนทรอยด์ (Centroid) เป็นจุดศูนย์กลางของกลุ่ม ระยะห่างระหว่างข้อมูลในกลุ่มและจุดเซนทรอยด์ของกลุ่มนั้นจะต้องเป็นระยะห่างที่น้อยที่สุด เมื่อเทียบกับจุดเซนทรอยด์ของกลุ่มอื่น ๆ สำหรับการแบ่งกลุ่มนี้ใช้การวัดระยะห่างแบบยูคลิด (Euclidean Distance) กระบวนการแบ่งกลุ่มข้อมูลแบบเคมีนมีขั้นตอนดังต่อไปนี้

1) สุ่มจุดเซนทรอยด์เริ่มต้นมา k จุด
2) ทำซ้ำกระบวนการต่อไปนี้ จนกระทั่งจุดเซนทรอยด์ทุกจุดไม่เปลี่ยนตำแหน่ง
    a. วัดระยะห่างระหว่างข้อมูลกับจุดเซนทรอยด์ โดยใช้การวัดระยะห่างแบบยูคลิด กลุ่มข้อมูลที่มีระยะห่างระหว่างข้อมูลในกลุ่มนั้นกับจุดเซนทรอยด์เดียวกันใกล้กว่าจุดเซนทรอยด์อื่น ๆ จะเป็นกลุ่มข้อมูลเดียวกัน
    b. เปลี่ยนจุดเซนทรอยด์ในแต่ละกลุ่ม โดยคำนวณจากค่าเฉลี่ยของข้อมูลในกลุ่ม ซึ่งคำนวณได้ดังนี้

$$\overline{x}_C = \frac{1}{n_C} \sum_{i=1}^{n_C} x_{iC}$$

โดยที่ $\overline{x}_C$ คือค่าเฉลี่ยของข้อมูลในกลุ่ม c

n_C คือจำนวนข้อมูลในกลุ่ม c

x_{iC} คือข้อมูลตัวที่ i ในกลุ่ม c

อัลกอริทึมการแบ่งกลุ่มข้อมูลแบบเคมีนทำได้ง่ายและแบ่งกลุ่มได้อย่างรวดเร็ว แต่อัลกอริทึมนี้จะมีปัญหากับการจัดการข้อมูลผิดปกติ (Noise) เพราะใช้การเปลี่ยนจุดเซนทรอยด์จากค่าเฉลี่ยของข้อมูลในกลุ่ม นอกจากนี้ต้องกำหนดค่า k และการสุ่มจุดเซนทรอยด์เริ่มต้น ซึ่งมีผลต่อประสิทธิภาพในการแบ่งกลุ่มในเชิงผลลัพธ์และเวลาในการคำนวณ

2.1.2 การแบ่งกลุ่มข้อมูลแบบเคมีเดียน (K-medians Clustering)

การแบ่งกลุ่มข้อมูลแบบเคมีเดียน [21] เป็นวิธีการแบ่งกลุ่มที่คล้ายกับการแบ่งกลุ่มข้อมูลแบบเคมีน แตกต่างที่จะใช้การวัดระยะห่างแบบแมนฮัตตัน (Manhattan Distance) และเปลี่ยนจุดเซนทรอยด์ใหม่โดยคำนวณจากค่ามัธยฐานของข้อมูลในกลุ่ม กระบวนการแบ่งกลุ่มข้อมูลแบบเคมีเดียนมีขั้นตอนดังต่อไปนี้

1) สุ่มจุดเซนทรอยด์เริ่มต้นมา k จุด
2) ทำซ้ำกระบวนการต่อไปนี้ จนกระทั่งจุดเซนทรอยด์ทุกจุดไม่เปลี่ยนตำแหน่ง
    a. วัดระยะห่างระหว่างข้อมูลกับจุดเซนทรอยด์ โดยใช้การวัดระยะห่างแบบแมนฮัตตัน กลุ่มข้อมูลที่มีระยะห่างระหว่างข้อมูลในกลุ่มนั้นกับจุดเซนทรอยด์เดียวกันใกล้กว่าจุดเซนทรอยด์อื่น ๆ จะเป็นกลุ่มข้อมูลเดียวกัน
    b. เปลี่ยนจุดเซนทรอยด์ในแต่ละกลุ่ม โดยคำนวณจากค่ามัธยฐานของข้อมูลในกลุ่ม

อัลกอริทึมการแบ่งกลุ่มข้อมูลแบบเคมีเดียนสามารถจัดการข้อมูลผิดปกติได้ดีกว่าแบบเคมีน เพราะใช้การเปลี่ยนจุดเซนทรอยด์จากค่ามัธยฐานของข้อมูลในกลุ่ม แต่ก็ต้องกำหนดค่า k และอัลกอริทึมนี้มีการคำนวณความซับซ้อนมากกว่าแบบเคมีน ซึ่งส่งผลให้อัลกอริทึมนี้ทำงานช้ากว่าการแบ่งกลุ่มแบบเคมีน

2.1.3 การแบ่งกลุ่มข้อมูลแบบดีบีสแกน (Density-Based Spatial Clustering of Applications with Noise / DBSCAN)

การแบ่งกลุ่มข้อมูลแบบดีบีสแกน [22] เป็นวิธีการแบ่งกลุ่มที่ขึ้นอยู่กับความหนาแน่นของข้อมูลในบริเวณต่าง ๆ บริเวณที่มีข้อมูลอยู่หนาแน่นและใกล้กันจะถูกจัดให้เป็นกลุ่มเดียวกัน ส่วนข้อมูลที่อยู่ไกลจากบริเวณที่มีข้อมูลหนาแน่นจะถูกกำหนดให้เป็นข้อมูลผิดปกติ เป็นการแบ่งกลุ่มที่ต้องกำหนดรัศมีความหนาแน่น (Epsilon / Eps) และจำนวนข้อมูลในช่วงรัศมีขั้นต่ำ (Minimum Points / MinPts) การวัดระยะห่างใช้การวัดระยะห่างแบบยูคลิด กระบวนการแบ่งกลุ่มข้อมูลแบบดีบีสแกนมีขั้นตอนดังต่อไปนี้

1) สุ่มข้อมูลเริ่มต้นจากชุดข้อมูล
2) กำหนดข้อมูลตามเงื่อนไขดังนี้
    a. ถ้าในช่วงรัศมีความหนาแน่นของข้อมูลมีจำนวนข้อมูลไม่น้อยกว่าจำนวนข้อมูลในช่วงรัศมีขั้นต่ำ ให้กำหนดข้อมูลดังกล่าวเป็นจุดแก่น (Core Point)
    b. ถ้าในช่วงรัศมีความหนาแน่นของข้อมูลมีจำนวนข้อมูลน้อยกว่าจำนวนข้อมูลในช่วงรัศมีขั้นต่ำ แต่มีจุดแก่นอยู่ในช่วงรัศมีความหนาแน่น ให้กำหนดข้อมูลดังกล่าวเป็นจุดชายแดน (Border Point)
    c. ถ้าในช่วงรัศมีความหนาแน่นของข้อมูลมีจำนวนข้อมูลน้อยกว่าจำนวนข้อมูลในช่วงรัศมีขั้นต่ำ แต่มีจุดชายแดนอยู่ในช่วงรัศมีความหนาแน่น ให้กำหนดข้อมูลดังกล่าวเป็นจุดชายแดนเช่นกัน
    d. มิฉะนั้นให้กำหนดข้อมูลดังกล่าวเป็นข้อมูลผิดปกติ
3) ข้อมูลที่ยังไม่ได้กำหนดตามเงื่อนไขข้อ 2) ซึ่งอยู่ใกล้เคียงกับข้อมูลที่กำหนดไปแล้ว มาพิจารณาตามเงื่อนไขข้อ 2) ต่อไป
4) ทำซ้ำข้อ 3) จนกระทั่งไม่มีข้อมูลที่ยังไม่ถูกกำหนดตามข้อ 2)
5) แบ่งกลุ่มข้อมูลจากจุดแก่นและจุดชายแดนที่อยู่ใกล้กัน จุดเหล่านี้จะถูกกำหนดให้เป็นกลุ่มข้อมูลเดียวกัน ส่วนข้อมูลผิดปกติจะเกิดจากข้อมูลซึ่งไม่ได้อยู่ใกล้จุดชายแดน จะไม่ถูกกำหนดให้อยู่ในข้อมูลกลุ่มใด ๆ

อัลกอริทึมการแบ่งกลุ่มข้อมูลแบบดีบีสแกนสามารถจัดการข้อมูลผิดปกติได้ดีกว่าแบบเคมีนและเคมีเดียน เพราะการแบ่งกลุ่มที่ขึ้นอยู่กับความหนาแน่นของข้อมูล ทำให้อัลกอริทึมสามารถจำแนกได้ว่าข้อมูลที่มีระยะห่างจากบริเวณที่มีความหนาแน่นเป็นข้อมูลผิดปกติ นอกจากนี้ไม่ต้องกำหนดค่า k เพื่อให้อัลกอริทึมแบ่งกลุ่มตามค่า k ทำให้การแบ่งกลุ่มมีความยืดหยุ่นตามความหนาแน่น แต่อัลกอริทึมนี้มีการคำนวณความซับซ้อนมากขึ้น ซึ่งส่งผลให้อัลกอริทึมนี้ทำงานช้ากว่าการแบ่งกลุ่มแบบเคมีนและเคมีเดียน

2.2 การจำแนก (Classification)

การทำนายค่าไม่ต่อเนื่องเป็นวิธีการของการเรียนรู้แบบมีผู้สอน (Supervised Learning) ในการเรียนรู้ของเครื่อง เป็นเทคนิคที่นำข้อมูลคุณลักษณะมาทำนายค่าไม่ต่อเนื่อง เรียกค่าที่ทำนายว่าป้ายกำกับ (Label)

2.2.1 การหาเพื่อนบ้านใกล้สุดเคอันดับ (k-Nearest Neighbor)

การหาเพื่อนบ้านใกล้สุดเคอันดับ [23] เป็นวิธีการทำนายค่าไม่ต่อเนื่อง โดยพิจารณาข้อมูลจากชุดข้อมูลเรียนรู้ที่มีระยะห่างกับข้อมูลที่จะจำแนกใกล้ที่สุด k ข้อมูล จากนั้นจำแนกป้ายกำกับด้วยการตรวจสอบข้อมูลทั้ง k ข้อมูลว่ามีป้ายกำกับใดมากที่สุด ให้จำแนกด้วยป้ายกำกับนั้น ถ้ามีป้ายกำกับที่มากที่สุดมากกว่า 1 ป้ายกำกับ ให้จำแนกป้ายกำกับด้วยการสุ่มเลือกป้ายกำกับ 1 ป้ายกำกับจากป้ายกำกับที่มากที่สุด การวัดระยะห่างใช้การวัดระยะห่างแบบยูคลิด

2.3 การทำนายค่าต่อเนื่อง (Regression)

การทำนายค่าต่อเนื่องเป็นวิธีการของการเรียนรู้แบบมีผู้สอนในการเรียนรู้ของเครื่อง เป็นเทคนิคที่นำข้อมูลคุณลักษณะมาทำนายค่าต่อเนื่อง

2.3.1 การถดถอยเชิงเส้น (Linear Regression)

การถดถอยเชิงเส้น [24] เป็นวิธีการทำนายค่าต่อเนื่อง ซึ่งสมมติว่าตัวแปรต้นและตัวแปรตามมีความสัมพันธ์เชิงเส้น นั่นคือแบบจำลองนี้สามารถทำนายค่าตัวแปรตามโดยใช้ตัวแปรต้น แบบจำลองนี้สามารถเขียนให้อยู่ในรูปสมการถดถอยที่แสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม ดังนี้

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots + w_k x_{ik} \quad ;i = 1,2,\ldots,n$$

โดยที่ y_i คือค่าทำนายตัวแปรตามข้อมูลที่ i

$w_0, w_1, w_2, \ldots, w_k$ คือพารามิเตอร์ของแบบจำลอง ซึ่งปรับค่าตามชุดข้อมูลฝึกสอน

$x_{i1}, x_{i2}, \ldots, x_{ik}$ คือตัวแปรต้นข้อมูลที่ i ซึ่งมีทั้งหมด k ตัวแปร

n คือจำนวนข้อมูลทั้งหมดที่ใช้ทำนาย

2.3.2 ต้นไม้ตัดสินใจแบบถดถอย (Regression Tree)

ต้นไม้ตัดสินใจแบบถดถอย [25] เป็นแบบจำลองที่แบ่งข้อมูลที่ใช้ทำนายเป็นส่วน ๆ ซึ่งการแบ่งข้อมูลสามารถสรุปได้เป็นโครงสร้างต้นไม้ทวิภาค (Binary Tree) จากนั้นทำนายค่าต่อเนื่องแต่ละส่วนโดยใช้ค่าเฉลี่ยของค่าที่ใช้ทำนายจากข้อมูลในส่วนนั้น กระบวนการสร้างต้นไม้ตัดสินใจแบบถดถอยมีขั้นตอนดังต่อไปนี้

1) แบ่งส่วนข้อมูลเป็นลักษณะโครงสร้างต้นไม้ทวิภาค โดยการแบ่งข้อมูลเป็น 2 ส่วน และแบ่งข้อมูลแต่ละส่วนเป็น 2 ส่วนย่อย มีหลักการแบ่งคือเลือกคุณลักษณะ X_j และค่า x

ที่เกิดเงื่อนไข $X_j < x$ สำหรับแบ่งข้อมูลเป็นฝั่งที่เงื่อนไขเป็นจริง (ซ้าย) และเป็นเท็จ (ขวา) การแบ่งข้อมูลที่ดีควรแบ่งให้ข้อมูลแต่ละส่วนมีความแปรปรวนของค่าที่ต้องการทำนายในข้อมูลมีค่าน้อยที่สุด เพื่อให้การทำนายโดยใช้ค่าเฉลี่ยมีความแม่นยำยิ่งขึ้น ดังนั้นการแบ่งข้อมูลที่ดีควรสอดคล้องเงื่อนไขด้านล่าง

$$\max \quad |Y|Var(Y) - \left(|Y_L|Var(Y_L) + |Y_R|Var(Y_R)\right)$$

โดยที่ Y, Y_L, Y_R คือกลุ่มค่าที่ต้องการทำนายของข้อมูลก่อนถูกแบ่ง ข้อมูลฝั่งซ้าย และข้อมูลฝั่งขวา ตามลำดับ

$|Y|$ คือจำนวนข้อมูลก่อนถูกแบ่ง

$Var(Y) = \frac{1}{|Y|}\sum_{i=1}^{|Y|}\left(y_i - \bar{y}\right)$ คือความแปรปรวนของค่าที่ต้องการทำนายในข้อมูลก่อนถูกแบ่ง

2) วนแบ่งข้อมูลซ้ำไปเรื่อย ๆ จนกระทั่งจำนวนข้อมูลในส่วนนั้นมีน้อยกว่าจำนวนที่กำหนด

2.4 งานวิจัยที่เกี่ยวข้อง

2.4.1 แบบจำลองเชิงเส้นตัวแปรเดียว (Univariate Linear Model / UL)

Szabo และ Huberman [3] ได้นำข้อมูลยอดการดูวิดีโอสะสม 30 วัน พบว่ามีค่าสัมประสิทธิ์สหสัมพันธ์ในทิศทางเดียวกันสูงระหว่างลอการิทึมฐาน 10 ของยอดการดูสะสมในวันที่ 7 และลอการิทึมฐาน 10 ของยอดการดูสะสมในวันที่ 30 จึงเสนอแบบจำลองเชิงเส้นตัวแปรเดียวที่ทำนายยอดการดูวิดีโอสะสมในวันที่ 30 โดยใช้ยอดการดูวิดีโอสะสมในวันที่ 7

แบบจำลองแบบจำลองเชิงเส้นตัวแปรเดียวดังนี้

$$\hat{N}_V(t) = w_V(7)x_V(7)$$

โดยที่ $\hat{N}_V(t)$ เป็นค่าทำนายยอดการดูวิดีโอสะสมของวันที่ต้องการทำนาย

$w_V(7)$ เป็นพารามิเตอร์

$x_V(7)$ เป็นยอดการดูวิดีโอสะสมในวันที่ 7

2.4.2 แบบจำลองเชิงเส้นหลายตัวแปร (Multivariate Linear Model / ML)

Pinto และคณะ [5] เสนอแบบจำลองเชิงเส้นหลายตัวแปรเป็นแบบจำลองที่ใช้ยอดการดูรายวันใน 7 วันแรกสำหรับทำนายยอดการดูวิดีโอสะสมในวันถัด ๆ ไป

กำหนดให้ $x_V(i)$ เป็นยอดการดูวิดีโอวันที่ i นิยามเวกเตอร์ยอดการดูรายวัน r วันแรกดังนี้

$$X_V(r) = (x_V(1), x_V(2), \ldots, x_V(r))^T$$

จะได้แบบจำลองแบบจำลองเชิงเส้นหลายตัวแปรดังนี้

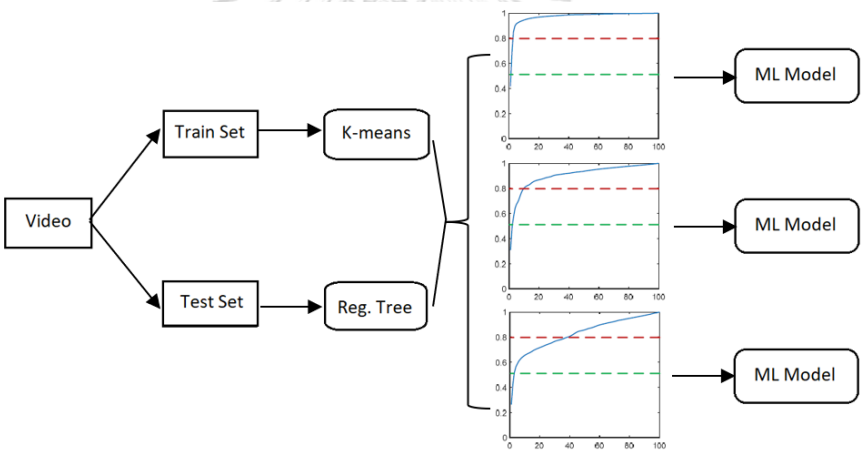
$$\hat{N}_V(t) = W_V(r) \cdot X_V(r)$$

โดยที่ $\hat{N}_V(t)$ เป็นค่าทำนายยอดการดูวิดีโอสะสมของวันที่ต้องการทำนาย

$W_V(r) = (w_V(1), w_V(2), \ldots, w_V(r))$ เป็นเวกเตอร์ของพารามิเตอร์

2.4.3 แบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต (Lifetime Aware Regression Model / LARM)

Ma และคณะ [12] เสนอแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตเป็นแบบจำลองหนึ่งที่ใช้การจัดกลุ่มของลักษณะของยอดการดูวิดีโอ และสร้างแบบจำลองเฉพาะกลุ่ม เพื่อไม่ให้แบบจำลองเชิงเส้นหลายตัวแปรต้องเรียนรู้ลักษณะยอดการดูวิดีโอที่หลากหลายเกินไป



รูปที่ 2.1 ภาพรวมของแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต

รูปที่ 2.1 แสดงภาพรวมของแบบจำลองนี้ ซึ่งเริ่มจากการนำยอดการดูสะสมของทุก ๆ วิดีโอหารด้วยยอดการดูสะสมวันที่ 30 ของแต่ละวิดีโอ เพื่อให้ลักษณะของยอดการดูวิดีโอมีค่าตั้งแต่ 0 ถึง 1 เท่านั้น นำข้อมูลเหล่านี้มาจัดกลุ่มโดยใช้การแบ่งกลุ่มข้อมูลแบบเคมีน หาค่าเฉลี่ยของอายุขัยในแต่ละกลุ่ม จากนั้นจัดกลุ่มให้กับชุดข้อมูลทดสอบโดยการนำคุณลักษณะต่าง ๆ ของวิดีโอ ได้แก่ ประเภทของวิดีโอ ความยาวของวิดีโอ จำนวนยอดการดูวิดีโอรวมของช่อง จำนวนความคิดเห็นรวมของช่อง จำนวนผู้ติดตามของช่อง จำนวนวิดีโอของช่อง จำนวนยอดการดูเริ่มต้น จำนวนความคิดเห็นเริ่มต้น จำนวนการกดถูกใจเริ่มต้น และจำนวนการกดไม่ถูกใจเริ่มต้น มาเรียนรู้กับแบบจำลองต้นไม้ตัดสินใจ

แบบถดถอย เพื่อทำนายอายุขัย แล้วนำผลการทำนายไปจัดกลุ่มให้กับชุดข้อมูลทดสอบ สุดท้ายนำแต่ละกลุ่มมาทำนายยอดการดูวิดีโอสะสมโดยใช้แบบจำลองเชิงเส้นหลายตัวแปร

2.4.4 งานวิจัยอื่น ๆ

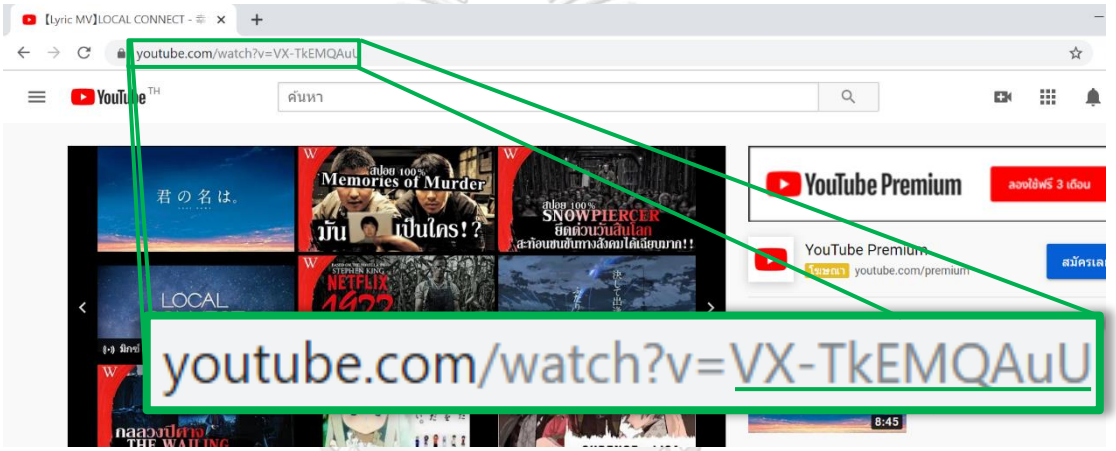
นอกจากนี้มีการนำเสนอแบบจำลองสำหรับการทำนายความนิยมของวิดีโอมากมาย ตัวอย่างเช่น Ding และคณะ [1] เสนอแบบจำลองกระบวนการฮอว์คแบบอารมณ์อ่อนไหวคู่ (Dual Sentimental Hawkes Process / DSHP) ซึ่งใช้หลักการถ่ายทอดอารมณ์ผ่านเครือข่ายที่แน่นอน และใช้ทำนายวิดีโอ 4 ประเภท ได้แก่ ภาพยนตร์ รายการโทรทัศน์ มิวสิกวิดีโอ และข่าวออนไลน์ Chen และคณะ [2] เสนอวิธีการทีมอลล์ (Transductive Multi-Modal Learning / TMALL) ซึ่งสกัดคุณลักษณะของวิดีโอ ได้แก่ คำบรรยายวิดีโอ เสียง ภาพ และสังคมออนไลน์ เพื่อทำนายวิดีโอสั้น จากงานวิจัยของ Crane และ Sornette [4] พบว่ายอดการดูวิดีโอรายวันมีรูปแบบที่หลากหลาย ทำให้ Pinto และคณะ [5] เสนอแบบจำลอง 2 อย่างคือแบบจำลองเชิงเส้นหลายตัวแปร และแบบจำลองเชิงเส้นหลายตัวแปรกับฟังก์ชันเรเดียลเบซิส (MRBF Model) ซึ่งเป็นแบบจำลองเชิงเส้นหลายตัวแปร ที่เพิ่มคุณลักษณะของการวัดความเหมือนของรูปแบบของยอดการดูวิดีโอโดยใช้ฟังก์ชันเกาส์เซียนเรเดียลเบสิส (Gaussian Radial Basis Function) สำหรับวัดความเหมือนระหว่างวิดีโอที่ต้องการทำนายกับวิดีโอที่ถูกสุ่มมาเป็นต้นแบบ Richier และคณะ [6] พบว่ารูปแบบยอดการดูวิดีโอสะสมจากยูทูบส่วนใหญ่มีลักษณะเป็นฟังก์ชันคณิตศาสตร์เชิงชีววิทยา ได้แก่ ฟังก์ชันเลขชี้กำลัง (Exponential) ฟังก์ชันเลขชี้กำลังแบบปรับปรุง (Modified Exponential) ฟังก์ชันซิกมอยด์ (Sigmoid) ฟังก์ชันซิกมอยด์แบบปรับปรุง (Modified Sigmoid) ฟังก์ชันกอมเพิร์ต (Gompertz) และฟังก์ชันกอมเพิร์ตแบบปรับปรุง (Modified Gompertz) ต่อมา Richier และคณะ [7] จึงสร้างระบบทำนายยอดการดูวิดีโอโดยการจำแนกตามฟังก์ชันจาก [6] กับฟังก์ชันเชิงเส้นและระดับความนิยมก่อนการทำนาย Trzci´nski และ Rokita [8] ได้แนวคิดจาก MRBF Model ในการสร้างแบบจำลอง Popularity-SVR ซึ่งก็คือแบบจำลองซัพพอร์ตเวกเตอร์รีเกรสชันใช้ฟังก์ชันเกาส์เซียนเรเดียลเบซิสสำหรับการแปลงปริภูมิ แบบจำลองนี้นอกจากใช้ยอดการดูวิดีโอแล้วยังใช้คุณลักษณะของวิดีโอช่วยในการทำนาย เช่น ความยาวของวิดีโอ โทนสีของวิดีโอ จานวนหน้าต่อฉาก เป็นต้น Li และคณะ [9] เสนอแบบจำลองเชิงเส้นหลายตัวแปรที่ขึ้นอยู่กับวิวัฒนาการรูปแบบและการทำนายการเพิ่มขึ้นอย่างมากของยอดการดูวิดีโอ (Evolution Pattern and Burst Prediction based Multivariate Linear Model / EPBP_ML) เป็นแบบจำลองที่แบ่งวิดีโอตามลำดับของสถานะของการเพิ่มขึ้นของยอดการดูวิดีโอรายวัน (Evolution Pattern) ช่วง 7 วันแรก ซึ่งสถานะแบ่งเป็นเพิ่มขึ้นอย่างมาก (Burst) และเพิ่มขึ้นน้อย (Slow) นำวิดีโอในแต่ละกลุ่มมาทำนายสถานะในวันต่อ ๆ

ไปโดยใช้คุณลักษณะของวิดีโอสำหรับการทำนายและสร้างแบบจำลองเชิงเส้นหลายตัวแปรของแต่ละกลุ่มเพื่อทำนายยอดการดูของวิดีโอ Ouyang และคณะ [10] แบ่งยอดการดูวิดีโอเป็น 4 ระดับ ใช้ในแบบจำลองโดยพิจารณาระดับของยอดการดูวิดีโอสะสมในวันที่ 7 และใช้ทำนายระดับของยอดการดูวิดีโอสะสมในวันที่ 30 โดยใช้คุณลักษณะของวิดีโอในการทำนายด้วย จากนั้นพิจารณาจากระดับ ถ้าอยู่ระดับเดียวกันให้ใช้แบบจำลองเชิงเส้นตัวแปรเดียว แต่ถ้าอยู่คนละระดับให้ใช้แบบจำลองเชิงเส้นหลายตัวแปร ในการทำนายยอดการดูวิดีโอสะสมในวันที่ 30 Mao และคณะ [11] เสนอแบบจำลองสำหรับการทำนายยอดการดูซีรีย์บนโทรทัศน์ที่เกิดจากโครงข่ายประสาทเทียม 3 ส่วนดังนี้ ส่วนแรกเป็นโครงข่ายประสาทเทียมแบบลึก ใช้สกัดคุณลักษณะที่เป็นการรวมกันที่ซับซ้อนของแต่ละคุณลักษณะ ส่วนที่สองเป็นโครงข่ายความสัมพันธ์ (Relation Network) เอาไว้ใช้หาความสัมพันธ์จากผลลัพธ์ของส่วนแรก ส่วนที่สามเป็นการเรียนรู้แบบหลายงาน (Multi-task Learning) ใช้ทำนายยอดการดู Chen และ Chang [13] เสนอแบบจำลองที่ใช้วิธีการร่วมกัน (Ensemble Method) ของการเรียนรู้ของเครื่องสำหรับทำนายระดับของความนิยมของยอดการดูวิดีโอ Gürsun และคณะ [14] ใช้แบบจำลองอัตถิภาวแบบเคลื่อนที่เฉลี่ย (ARMA) เพื่อทำนายยอดการดูวิดีโอระยะยาว ของวิดีโอ 2 ประเภท คือ วิดีโอที่เข้าถึงไม่บ่อย และวิดีโอที่เข้าถึงบ่อย Wu และคณะ [15] เสนอแบบจำลองวิวัฒนาการ (Evolution Model) เกิดจากการพิจารณาปัจจัยเหล่านี้ ได้แก่ การแนะนำวิดีโอโดยตรง การแนะนำจากปากต่อปาก และความนิยมที่แท้จริง นำมาใช้พิจารณาวิวัฒนาการของวิดีโอ ทำให้นำมาใช้ทำนายยอดการดูวิดีโอได้ Roy และคณะ [16] เสนออัลกอริทึมโซเชียลทรานสเฟอร์ (SocialTransfer) ซึ่งนำคุณลักษณะของโซเชียลมีเดียอื่นสำหรับทำนายการเพิ่มขึ้นของยอดการดูวิดีโอ Li และคณะ [17] เสนอแบบจำลองการทำนายวิดีโอที่ถูกช่วยจากโซเชียลเน็ตเวิร์ก (Social Network Assisted Video Prediction / SoVP) เพื่อการทำนายยอดการดูวิดีโอที่ถูกแชร์บนโซเชียลเน็ตเวิร์ก โดยการทำนายขึ้นอยู่กับการเผยแพร่ของวิดีโอ Ahmed และคณะ [18] เสนอแบบจำลองที่นำความนิยมในอดีตของวิดีโอมาเป็นสถานะต่อเนื่องกัน ซึ่งสถานะนั้นประกอบด้วยเวลาและอัตราการเปลี่ยนแปลงความนิยม แบบจำลองนี้ใช้การเปลี่ยนแปลงของสถานะเพื่อทำนายความนิยมในอนาคต Tan และคณะ [19] เสนอแบบจำลองพลวัตของยอดการดูวิดีโอ (View Count Dynamic Model / VCDM) ที่นำการเพิ่มขึ้นของยอดการดูวิดีโอมาเกี่ยวข้องกับยอดการดูวิดีโอในอดีต เพื่อสร้างเป็นสูตรสำหรับทำนายและใช้ยอดการดูวิดีโอแบบอนุกรมเวลาสำหรับกำหนดพารามิเตอร์ของสูตรนั้น

บทที่ 3
ระเบียบวิธีวิจัย

3.1 การเก็บข้อมูลของวิดีโอ

วิดีโอบนยูทูบจะมีรหัสระบุของวิดีโอสำหรับระบุตัวตนวิดีโอ ซึ่งแสดงในลิงก์ที่เข้าหน้าวิดีโอบนยูทูบผ่านเว็บเบราว์เซอร์ เช่น https://www.youtube.com/watch?v=xxxxxxxxxxx รหัสระบุของวิดีโอจากลิงก์ดังกล่าวจะอยู่ต่อท้าย ?v= เสมอ ดังรูปที่ 3.1 ทำให้สามารถเก็บข้อมูลต่าง ๆ ของวิดีโอได้จากรหัสระบุของวิดีโอ



รูปที่ 3.1 รหัสระบุของวิดีโอในลิงก์ซึ่งอยู่ต่อท้าย ?v= ของวิดีโอของยูทูบบนเว็บเบราว์เซอร์

การเก็บข้อมูลวิดีโอในงานวิจัยนี้มีอยู่ 3 อย่างคือ การเก็บรหัสระบุของวิดีโอ การเก็บข้อมูลค่าความนิยมของวิดีโอ และการเก็บข้อมูลคุณลักษณะของวิดีโอ

3.1.1 การเก็บรหัสระบุของวิดีโอ

การเก็บรหัสระบุของวิดีโอมีทั้งหมด 24 รอบ เก็บรอบละ 600 วิดีโอ โดยเว้นระยะห่าง 1 ชั่วโมงสำหรับการเก็บข้อมูลแต่ละรอบ แต่ละรอบต้องเป็นวิดีโอที่บรรจุขึ้นในช่วง 1 ชั่วโมงก่อนหน้าที่จะเก็บรหัสระบุของวิดีโอ เช่น ต้องการเก็บรหัสระบุของวิดีโอเวลา 12.00น. หมายความว่ารหัสระบุของวิดีโอที่ได้ต้องเป็นวิดีโอที่บรรจุขึ้นตั้งแต่เวลา 11.00น. ถึง 11.59น. ของวันนั้น เริ่มการเก็บรหัสระบุของวิดีโอตั้งแต่วันที่ 14 กรกฎาคม 2562 เวลา 14.28น. ถึงวันที่ 15 กรกฎาคม 2562 เวลา

13.28น. โดยใช้ครอนแท็บ (Crontab) สำหรับตั้งเวลาให้ดำเนินการไฟล์ไพทอน (Python) ที่ใช้ยูทูบเอพีไอ (YouTube API) เก็บรหัสระบุของวิดีโอตามช่วงเวลาที่ตั้ง ผ่านการประมวลผลบนกลุ่มก้อนเมฆของกูเกิลคลาวด์ (Google Cloud Platform) หลังจากเก็บรหัสระบุของวิดีโอครบ 24 รอบ พบว่าในบางรอบไม่สามารถเก็บรหัสระบุได้ถึง 600 วิดีโอ

3.1.2 การเก็บข้อมูลค่าความนิยม

ข้อมูลค่าความนิยมของวิดีโอประกอบด้วยจำนวนการกดถูกใจ จำนวนการกดไม่ถูกใจ จำนวนความคิดเห็น และยอดการดูวิดีโอ การเก็บข้อมูลค่าความนิยมของวิดีโอจะเริ่มเก็บหลังจากที่เก็บรหัสระบุของวิดีโอนั้นมาแล้ว 24 ชั่วโมง โดยเก็บทุก 24 ชั่วโมง ทั้งหมด 30 รอบ สมมติว่ารหัสระบุของวิดีโอถูกเก็บวันที่ 14 กรกฎาคม 2562 เวลา 14.28น. การเก็บข้อมูลค่าความนิยมของวิดีโอนี้จะเริ่มตั้งแต่วันที่ 15 กรกฎาคม 2562 เวลา 14.28น. ถึงวันที่ 13 สิงหาคม 2562 เวลา 14.28น. โดยใช้ครอนแท็บตั้งเวลาให้ดำเนินการไฟล์ไพทอนที่ใช้เซเลเนียม (Selenium) เก็บข้อมูลค่าความนิยมของวิดีโอ ผ่านการประมวลผลบนกลุ่มก้อนเมฆของกูเกิลคลาวด์

3.1.3 การเก็บข้อมูลคุณลักษณะของวิดีโอ

การเก็บข้อมูลคุณลักษณะของวิดีโอจะเริ่มหลังจากที่เก็บรหัสระบุของวิดีโอเสร็จทุกรอบ เนื่องจากข้อมูลคุณลักษณะของวิดีโอไม่ได้เป็นข้อมูลแบบอนุกรมเวลา โดยใช้จูไพเตอร์ โน้ตบุ๊ก (Jupyter Notebook) ในเครื่องคอมพิวเตอร์ส่วนตัวดำเนินการไฟล์ไพทอนที่ใช้ยูทูบเอพีไอ เก็บข้อมูลคุณลักษณะของวิดีโอจากรหัสระบุของวิดีโอ ตัวอย่างของข้อมูลคุณลักษณะของวิดีโอ เช่น ความยาว (หน่วยชั่วโมง นาที และวินาที) หมวดหมู่ ชื่อ คำบรรยาย เป็นต้น

3.2 การจัดการข้อมูลที่มีปัญหา

หลังจากเก็บข้อมูลเสร็จได้ทั้งหมด 14,395 วิดีโอ แต่พบปัญหาของชุดข้อมูลนี้ 3 อย่าง คือ การเก็บรหัสระบุของวิดีโอซ้ำกัน ค่าความนิยมของวิดีโอบางวันเป็นค่าว่าง และยอดการดูวิดีโอสะสมบางวิดีโอบางวันมีค่ามากกว่าวันถัดไป

3.2.1 การเก็บรหัสระบุของวิดีโอซ้ำกัน

การเก็บรหัสระบุของวิดีโอซ้ำกันถือว่าเป็นข้อผิดพลาดของยูทูบเอพีไอ ที่เก็บรหัสระบุของวิดีโอมาซ้ำกันใน 1 รอบ ทำให้การเก็บข้อมูลอื่นของวิดีโอซ้ำกันด้วย ตัดสินใจเก็บข้อมูลที่มาจากรหัสระบุของวิดีโอที่ถูกดึงมาครั้งแรก และลบข้อมูลของวิดีโอที่ซ้ำกันนี้ เว้นแต่ว่าถ้าข้อมูลค่าความนิยมที่จะเก็บไว้มีค่าว่าง จึงพิจารณาข้อมูลวิดีโอที่ซ้ำกันลำดับถัดไป

3.2.2 ค่าความนิยมของวิดีโอบางวันเป็นค่าว่าง

สาเหตุของค่าความนิยมของบางวันมีค่าว่าง ได้แก่ วิดีโอถูกลบเนื่องจากผิดกฎของยูทูบหรือผู้ใช้ที่เป็นเจ้าของวิดีโอลบวิดีโอด้วยตนเอง เจ้าของวิดีโอปิดการแสดงค่าความนิยมบนหน้าแสดงวิดีโอ และความหน่วงของการแสดงผลบนเว็บเบราว์เซอร์ทำให้โหลดค่าความนิยมบางค่าในเวลานั้นไม่ได้ ทำให้ไม่สามารถดึงค่าความนิยมโดยใช้เซเลเนียมได้ จึงพิจารณาทุกค่าความนิยมดังนี้

1) ถ้าค่าความนิยมบางค่ามีค่าว่างตั้งแต่ 4 ค่าเป็นต้นไป ต้องลบข้อมูลวิดีโอนั้น

2) ถ้าค่าความนิยมบางค่ามีค่าว่างติดกันตั้งแต่ 3 วันขึ้นไป ต้องลบข้อมูลวิดีโอนั้น

3) ถ้าค่าความนิยมบางค่ามีค่าว่างในวันที่ 1 หรือวันที่ 30 ต้องลบข้อมูลวิดีโอนั้น

4) ถ้าค่าความนิยมบางค่ามีค่าว่างโดยที่ค่าความนิยมของวันก่อนหน้าและวันถัดไปไม่เป็นค่าว่าง จะแทนที่ค่าว่างด้วยค่าเฉลี่ยระหว่างค่าความนิยมของวันก่อนหน้าและวันถัดไป ให้ x เป็นค่าความนิยมและ i เป็นวันที่ค่าความนิยมเป็นค่าว่าง จะได้ว่า

$$x_i = \frac{x_{i-1} + x_{i+1}}{2} \tag{2}$$

5) ถ้าค่าความนิยมบางค่ามีค่าว่างติดกันเพียง 2 วัน ให้ x เป็นค่าความนิยมและ i, $i+1$ เป็นวันที่ค่าความนิยมเป็นค่าว่าง จะแทนที่ค่าว่าง ดังนี้

$$x_i = x_{i-1} + d \tag{3}$$

$$x_{i+1} = x_i + d \tag{4}$$

โดยที่ $d = \frac{x_{i+2} - x_{i-1}}{3}$

ถ้าค่าความนิยมที่ได้จากการแทนที่ค่าว่างเป็นทศนิยม ให้ปัดเศษค่าเหล่านั้นให้เป็นจำนวนเต็ม

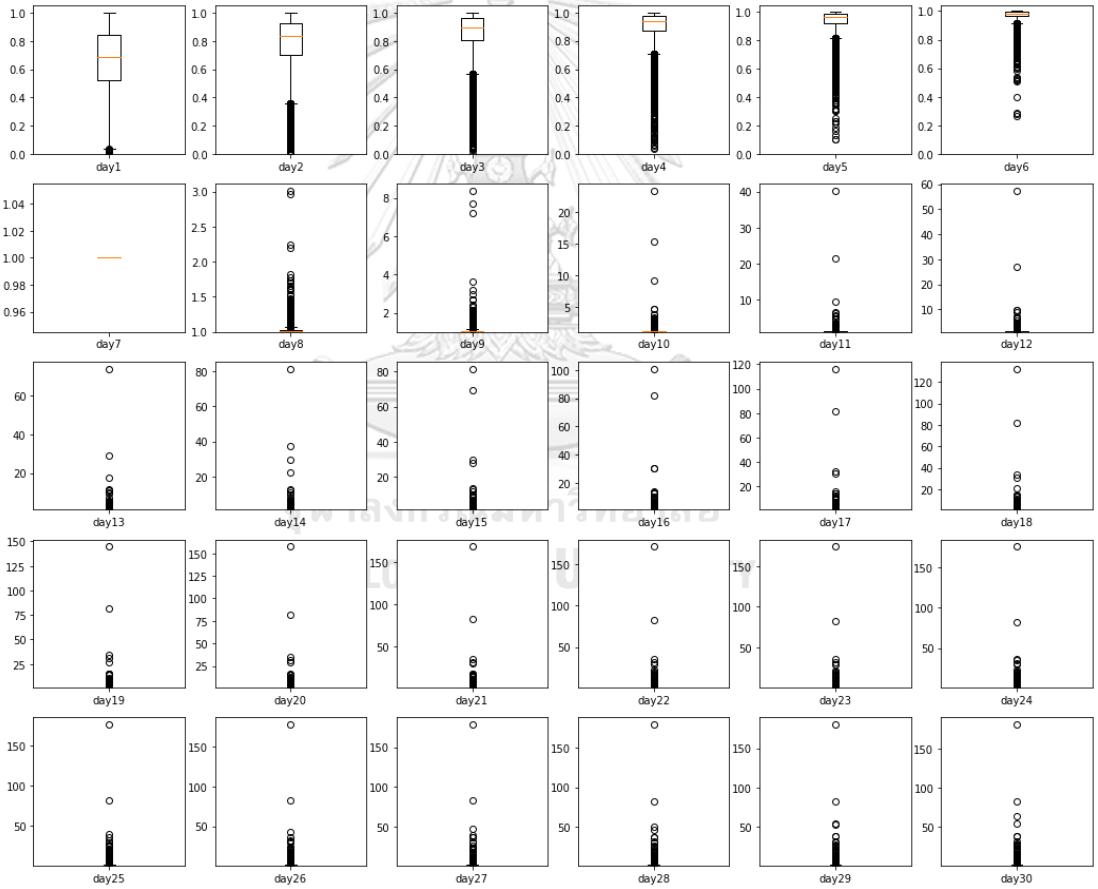
3.2.3 ยอดการดูวิดีโอสะสมบางวิดีโอบางวันมีค่ามากกว่าวันถัดไป

ปัญหาสุดท้ายคือยอดการดูวิดีโอสะสมมีค่าน้อยลงในวันถัดไป พบว่าเป็นปัญหาที่สามารถเกิดขึ้นได้ของวิดีโอบนยูทูบ จึงตัดสินใจลบข้อมูลวิดีโอเหล่านั้นทั้งหมด

3.3 ชุดข้อมูล

หลังจากจัดการข้อมูลที่มีปัญหาทั้งหมด ทำให้เหลือวิดีโออยู่ 10,234 วิดีโอ นำวิดีโอเหล่านั้นมาใช้วิเคราะห์และสร้างแบบจำลอง

พิจารณาการกระจายตัวของยอดการดูวิดีโอสะสมที่ถูกหารด้วยยอดการดูวิดีโอสะสมในวันที่ 7 โดยการสร้างแผนภาพกล่อง (Box Plot) ตั้งแต่วันที่ 1 ถึงวันที่ 30 ได้ออกมาเป็นดังรูปที่ 3.2

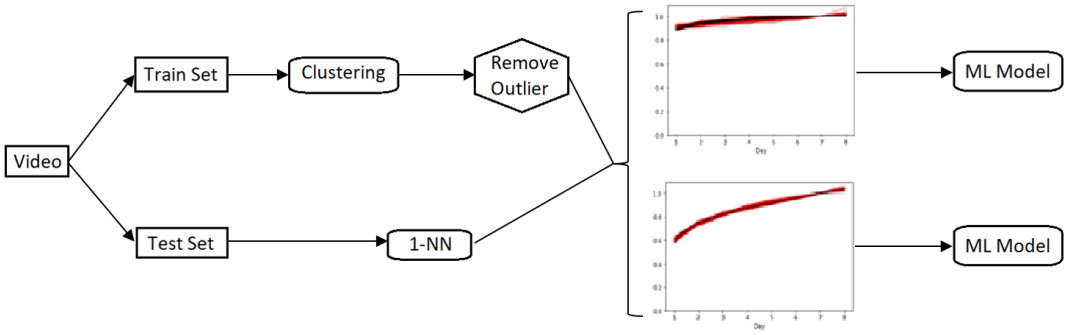


รูปที่ 3.2 แผนภาพกล่องแสดงการกระจายตัวของ
ยอดการดูวิดีโอสะสมที่ถูกหารด้วยยอดการดูวิดีโอสะสมในวันที่ 7

จากรูปที่ 3.2 พบว่ายิ่งจำนวนวันมากขึ้น ก็จะพบค่าผิดปกติ (Outlier) มากขึ้น แบบจำลองที่เหมาะสมกับชุดข้อมูลนี้ควรจัดการกับค่าเหล่านี้ได้ดี เพื่อไม่ให้แบบจำลองเรียนรู้ค่าเหล่านี้ ซึ่งนำไปสู่ผลการทำนายที่แย่

3.4 แบบจำลองทำนายยอดการดูวิดีโอ

งานวิจัยนี้ขอเสนอแบบจำลองเอฟ7เอ็นเอ็มแอล (the First 7-day Normalization for clustering with Multivariate Linear model / F7NML) ซึ่งแบบจำลองนี้ประกอบด้วย 4 ส่วน ดังนี้ ส่วนแรกเป็นการแบ่งลักษณะของยอดการดูวิดีโอสะสม ส่วนที่สองเป็นการกำจัดกลุ่มของลักษณะของยอดการดูวิดีโอที่มีลักษณะผิดปกติ ส่วนที่สามเป็นการจัดกลุ่มของลักษณะของยอดการดูวิดีโอให้กับชุดข้อมูลทดสอบ และส่วนสุดท้ายเป็นการทำนายยอดการดูวิดีโอสะสมเฉพาะกลุ่ม ภาพรวมของแบบจำลองนี้ถูกแสดงรูปที่ 3.3



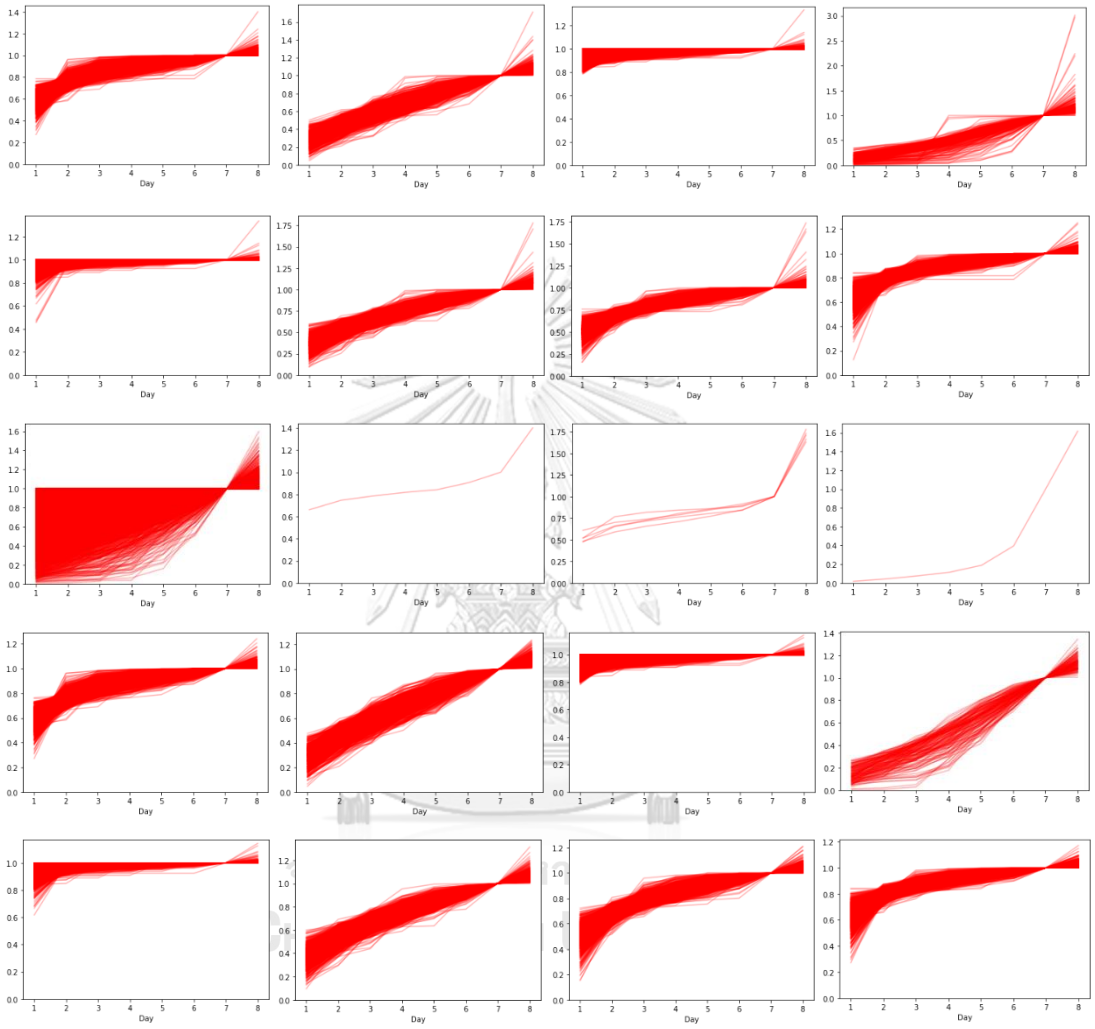
รูปที่ 3.3 ภาพรวมของแบบจำลองเอฟ7เอ็นเอ็มแอล

3.4.1 การแบ่งลักษณะของยอดการดูวิดีโอสะสม

การแบ่งลักษณะของยอดการดูวิดีโอสะสม แบ่งโดยการนำข้อมูลยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันที่ 7 และวันที่ใช้ทำนายยอดการดูวิดีโอสะสม หารด้วยยอดการดูวิดีโอสะสมวันที่ 7 จากนั้นนำไปเรียนรู้กับแบบจำลองการแบ่งกลุ่มของการเรียนรู้ของเครื่อง

ในงานวิจัยนี้ ใช้แบบจำลองการแบ่งกลุ่มพื้นฐาน ซึ่งประกอบด้วยการแบ่งกลุ่มข้อมูลแบบเคมีน การแบ่งกลุ่มข้อมูลแบบเคมีเดียน และการแบ่งกลุ่มข้อมูลแบบดีบีสแกน นอกจากนี้ขอเสนอแบบจำลองการแบ่งกลุ่มแบบผสมซึ่งเป็นแบบจำลองการแบ่งกลุ่มข้อมูลที่ใช้การแบ่งกลุ่มข้อมูลแบบ

เคมีนหรือเคมีเดียนในการแบ่งกลุ่มข้อมูลก่อน จากนั้นนำแต่ละกลุ่มที่แบ่งจากการแบ่งกลุ่มข้อมูลก่อนหน้านี้มาใช้การแบ่งกลุ่มข้อมูลแบบดีบีสแกน



รูปที่ 3.4 ผลลัพธ์ของบางกลุ่มจากการแบ่งลักษณะของยอดการดูวิดีโอสะสม
โดยใช้การแบ่งกลุ่มข้อมูลแบบเคมีน (แถวที่ 1) เคมีเดียน (แถวที่ 2) ดีบีสแกน (แถวที่ 3)
เคมีนกับดีบีสแกน (แถวที่ 4) และเคมีเดียนกับดีบีสแกน (แถวสุดท้าย)

รูปที่ 3.4 แสดงผลลัพธ์บางกลุ่มของแต่ละแบบจำลอง พบว่าการแบ่งกลุ่มข้อมูลแบบเคมีนและเคมีเดียนให้ผลลัพธ์แต่ละกลุ่มเป็นรูปแบบที่ชัดเจน แต่แบบจำลองทั้งสองอย่างจัดการกับค่าผิดปกติได้ไม่ดี กล่าวคือแบบจำลองทั้งสองอย่างมีวิดีโอจำนวนหนึ่งในกลุ่มที่มีรูปแบบแตกต่างจาก

วิดีโออื่น ๆ ในทางตรงกันข้าม การแบ่งกลุ่มข้อมูลแบบดีบีสแกนสามารถจัดการกับค่าผิดปกติได้ดี แต่ให้ผลลัพธ์การแบ่งกลุ่มเป็นรูปแบบได้ไม่ดี กล่าวคือมีบางกลุ่มที่รวมหลาย ๆ รูปแบบไว้ในกลุ่มเดียว ซึ่งเป็นรูปแบบที่ไม่ชัดเจนเหมือนกับการแบ่งกลุ่มข้อมูลทั้งสองอย่างก่อนหน้านี้ ส่วนแบบจำลองการแบ่งกลุ่มแบบผสมสามารถแบ่งกลุ่มเป็นรูปแบบได้ชัดเจนกว่าและจัดการกับค่าผิดปกติได้ดีกว่ากลุ่มแบบจำลองพื้นฐาน

3.4.2 การกำจัดกลุ่มของลักษณะของยอดการดูวิดีโอที่มีลักษณะผิดปกติ

เนื่องจากความหลากหลายของลักษณะของยอดการดูวิดีโอ ดังนั้นผลลัพธ์จากการแบ่งลักษณะของยอดการดูวิดีโอเป็นกลุ่ม ๆ อาจมีบางกลุ่มที่มีจำนวนวิดีโอน้อยซึ่งกำหนดให้เป็นกลุ่มของลักษณะของยอดการดูวิดีโอที่มีลักษณะผิดปกติ การนำกลุ่มเหล่านี้มาใช้สร้างแบบจำลองการเรียนรู้ของเครื่อง อาจทำให้แบบจำลองเกิดการเรียนรู้ที่เฉพาะเจาะจงกับข้อมูลเรียนรู้มากเกินไป (Overfitting) ในงานวิจัยนี้จึงกำหนดให้กำจัดกลุ่มข้อมูลเมื่อมีวิดีโอน้อยกว่า 6 รายการ

3.4.3 การจัดกลุ่มของลักษณะของยอดการดูวิดีโอให้กับชุดข้อมูลทดสอบ

ส่วนต่อมาเป็นการจัดกลุ่มของลักษณะของยอดการดูวิดีโอให้กับชุดข้อมูลทดสอบ ในงานวิจัยจัดกลุ่มโดยการนำข้อมูลยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันที่ 7 หารด้วยยอดการดูวิดีโอสะสมวันที่ 7 ทั้งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ และกำหนดป้ายกำกับของชุดข้อมูลเรียนรู้เป็นกลุ่มของลักษณะของยอดการดูวิดีโอ จากนั้นนำชุดข้อมูลเรียนรู้ไปเรียนรู้กับแบบจำลองการหาเพื่อนบ้านใกล้สุดเคอันดับ โดยกำหนดให้ k=1 ซึ่งทำนายว่าลักษณะของยอดการดูวิดีโอของชุดข้อมูลทดสอบมีลักษณะที่ใกล้เคียงกับลักษณะไหน เมื่อทราบว่าใกล้เคียงกับลักษณะไหนแล้ว จะกำหนดว่าลักษณะทั้ง 2 เป็นลักษณะเดียวกัน จึงควรอยู่ในกลุ่มเดียวกัน

3.4.4 การทำนายยอดการดูวิดีโอสะสมเฉพาะกลุ่ม

การใช้ยอดการดูวิดีโอวันที่ 1 ถึงวันที่ 7 ทำนายยอดการดูวิดีโอสะสมวันถัด ๆ ไป เป็นส่วนสุดท้ายหลังจากจัดกลุ่มให้กับชุดข้อมูลทดสอบ ทำนายโดยการใช้แบบจำลองเชิงเส้นหลายตัวแปรจาก [5] สำหรับแต่ละกลุ่ม เพื่อไม่ให้แบบจำลองเชิงเส้นหลายตัวแปรต้องเรียนรู้ลักษณะยอดการดูวิดีโอที่หลากหลายเกินไป

3.5 ความแตกต่างระหว่างแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตและแบบจำลองเอฟ7เอ็นเอ็มแอล

เมื่อพิจารณาภาพรวมของแบบจำลองทั้งสองอย่างตามรูปที่ 2.1 และรูปที่ 3.3 พบว่ามีลักษณะที่ใกล้เคียงกัน แต่ทั้งสองอย่างแตกต่างกันในแง่ของการแบ่งลักษณะของยอดการดูวิดีโอสะสมของชุดข้อมูลเรียนรู้ การกำจัดกลุ่มที่เป็นลักษณะผิดปกติและการจัดกลุ่มของลักษณะของยอดการดูวิดีโอให้กับชุดข้อมูลทดสอบ ซึ่งสรุปไว้ในตารางที่ 3.1

ตารางที่ 3.1 สรุปความแตกต่างระหว่างแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตและแบบจำลองเอฟ7เอ็นเอ็มแอล

ความแตกต่าง	แบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต	แบบจำลองเอฟ7เอ็นเอ็มแอล
1.) การแบ่งลักษณะของยอดการดูวิดีโอสะสมของชุดข้อมูลเรียนรู้	1.1) ใช้แบบจำลองการแบ่งกลุ่มข้อมูลแบบเคมีน 1.2) ใช้ยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันสุดท้ายที่ต้องการทำนาย หารด้วยยอดการดูวิดีโอสะสมวันสุดท้าย สำหรับแบบจำลองในข้อ 1.1	1.1) ใช้แบบจำลองการแบ่งกลุ่มข้อมูลหลาย ๆ แบบ เพื่อเพิ่มการรองรับการกำจัดกลุ่มที่เป็นลักษณะผิดปกติ 1.2) ใช้ยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันที่ 7 และวันที่ต้องการทำนายหารด้วยยอดการดูวิดีโอสะสมวันที่ 7 สำหรับแบบจำลองในข้อ 1.1
2.) การกำจัดกลุ่มที่เป็นลักษณะผิดปกติ	2.1) ไม่มีการกำจัดกลุ่มที่เป็นลักษณะผิดปกติ	2.1) มีการกำจัดกลุ่มที่เป็นลักษณะผิดปกติ โดยกำหนดให้กลุ่มที่มีวิดีโอน้อยกว่า 6 วิดีโอเป็นกลุ่มที่เป็นลักษณะผิดปกติ
3.) การจัดกลุ่มของลักษณะของยอดการดูวิดีโอให้กับชุดข้อมูลทดสอบ	3.1) ใช้แบบจำลองต้นไม้ตัดสินใจแบบถดถอย เพื่อทำนายอายุขัยซึ่งเป็นตัวแทนของกลุ่ม	3.1) ใช้แบบจำลองการหาเพื่อนบ้านใกล้สุด 1 อันดับ เพื่อทำนายว่าลักษณะของยอดการดูวิดีโอของชุดข้อมูลทดสอบมีลักษณะที่

	3.2) ใช้คุณลักษณะของวิดีโอ ได้แก่ ประเภทของวิดีโอ ความยาวของวิดีโอ จำนวนยอดการดูวิดีโอรวมของช่อง จำนวนความคิดเห็นรวมของช่อง จำนวนผู้ติดตามของช่อง จำนวนวิดีโอของช่อง จำนวนยอดการดูเริ่มต้น จำนวนความคิดเห็นเริ่มต้น จำนวนการกดถูกใจเริ่มต้น และจำนวนการกดไม่ถูกใจเริ่มต้น สำหรับแบบจำลองในข้อ 3.1	ใกล้เคียงกับลักษณะไหน 3.2) ใช้ยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันสุดท้ายซึ่งถูกหารด้วยยอดการดูวิดีโอสะสมวันสุดท้าย สำหรับแบบจำลองในข้อ 3.1

บทที่ 4

การทดลองและผลการทดลอง

4.1 แบบจำลองที่ใช้เปรียบเทียบ

แบบจำลองที่นำมาใช้เปรียบเทียบกับแบบจำลองของงานวิจัยนี้มีทั้งหมด 5 อย่าง ได้แก่ แบบจำลองการถดถอยเชิงเส้นที่มีตัวแปรต้นคือยอดการดูวิดีโอตั้งแต่วันที่ 1 ถึง 7 และตัวแปรตามคือยอดการดูวิดีโอสะสมของวันที่ต้องการทำนาย แบบจำลองต้นไม้ตัดสินใจแบบถดถอยที่ทำนายโดยใช้ยอดการดูวิดีโอตั้งแต่วันที่ 1 ถึงวันที่ 7 แบบจำลองเชิงเส้นตัวแปรเดียว [3] แบบจำลองเชิงเส้นหลายตัวแปร [5] และแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต [12] รายละเอียดของแบบจำลองอย่างอธิบายในหัวข้อที่ 2.3.1 2.3.2 2.4.1 2.4.2 และ 2.4.3 ตามลำดับ

4.2 การวัดประสิทธิภาพ

งานวิจัยนี้วัดประสิทธิภาพว่าแบบจำลองใดดีที่สุด โดยใช้ 10 โฟลด์ครอสวาลิเดชัน (10-Fold Cross Validation) สำหรับแบ่งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ การทดสอบแต่ละครั้งใช้ค่าความผิดพลาดกำลังสองที่สัมพันธ์กันแบบเฉลี่ย (Mean Relative Squared Error / MRSE) สำหรับวัดประสิทธิภาพ ซึ่งมีสูตรดังนี้

$$MRSE = \frac{1}{|V|}\sum_{V \in V}\left(\frac{N_V(t) - \hat{N}_V(t)}{N_V(t)}\right)^2 \quad (1)$$

โดยที่ $N_V(t)$ เป็นยอดการดูวิดีโอสะสมจริงของวันที่จะทำนาย

$\hat{N}_V(t)$ เป็นค่าทำนายยอดการดูวิดีโอสะสมของวันที่จะทำนาย

V เป็นเซตของวิดีโอที่ใช้ในการวัดประสิทธิภาพ

เนื่องจากแบบจำลองทั้งหมดใช้แบบจำลองเชิงเส้นหลายตัวแปรสำหรับการทำนายยอดการดูวิดีโอสะสม ดังนั้นวิธีการหาพารามิเตอร์ที่ดีที่สุดของแต่ละแบบจำลองจึงเหมือนกัน

กำหนดให้ C เป็นเซตของวิดีโอสำหรับเรียนรู้ เนื่องจากเป็นการวัดประสิทธิภาพของแบบจำลองโดยใช้ค่าความผิดพลาดกำลังสองที่สัมพันธ์กันแบบเฉลี่ย ดังนั้นสามารถหาพารามิเตอร์ $W_V(r)$ ที่ดีที่สุดได้ดังนี้

$$\underset{W_V(r)}{\operatorname{argmin}} \frac{1}{|C|} \sum_{V \in C} \left(\frac{W_V(r) \cdot X_V(r)}{N_V(t)} - 1 \right)^2 \quad (2)$$

กำหนดให้ $X_V^* = \frac{X_V(r)}{N_V(t)}$ จาก (2) สามารถแสดงปัญหาการหาค่าที่ดีที่สุดได้ดังนี้

$$\underset{W_V(r)}{\operatorname{argmin}} \frac{1}{|C|} \sum_{V \in C} \left(W_V(r) \cdot X_V^* - 1 \right)^2 \quad (3)$$

จาก (3) พบว่าเป็นปัญหาการหาค่าที่ดีที่สุด สามารถคำนวณหาพารามิเตอร์ได้โดยใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares / OLS)

4.3 การทดลองใช้แบบจำลองการแบ่งกลุ่มต่าง ๆ กับแบบจำลองเอฟ7เอ็นเอ็มแอล

การทดลองนี้ทดลองใช้แบบจำลองการแบ่งกลุ่มข้อมูล 5 อย่าง ได้แก่ การแบ่งกลุ่มข้อมูลแบบเคมีน การแบ่งกลุ่มข้อมูลแบบเคมีเดียน การแบ่งกลุ่มข้อมูลแบบดีบีสแกน การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกน และการแบ่งกลุ่มข้อมูลแบบเคมีเดียนกับดีบีสแกน ในแบบจำลองเอฟ7เอ็นเอ็มแอล ซึ่งกำหนดไฮเปอร์พารามิเตอร์ตามตารางที่ 4.1.1 การทดลองนี้ใช้การค้นหาแบบตะแกรง (Grid Search) สำหรับหาไฮเปอร์พารามิเตอร์ที่ดีที่สุดในแต่ละรอบ จากนั้นนำไฮเปอร์พารามิเตอร์นั้นมาใช้ทำนายและวัดประสิทธิภาพ

ตารางที่ 4.1.1 ไฮเปอร์พารามิเตอร์สำหรับแบบจำลองการแบ่งกลุ่ม

แบบจำลอง	**ไฮเปอร์พารามิเตอร์**
การแบ่งกลุ่มข้อมูลแบบเคมีน	$k \in \{2,3,4,5,6\}$
การแบ่งกลุ่มข้อมูลแบบเคมีเดียน	$k \in \{2,3,4,5,6\}$
การแบ่งกลุ่มข้อมูลแบบดีบีสแกน	$epsilon \in \{0.1,0.2,0.3,0.4,0.5\}, minPts = 1$
การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกน	$k \in \{2,3,4,5,6\}$, $epsilon \in \{0.1,0.2,0.3,0.4,0.5\}, minPts = 1$
การแบ่งกลุ่มข้อมูลแบบเคมีเดียนกับดีบีสแกน	$k \in \{2,3,4,5,6\}$, $epsilon \in \{0.1,0.2,0.3,0.4,0.5\}, minPts = 1$

จากตารางที่ 4.1.2 พบว่าการแบ่งกลุ่มข้อมูลแบบเคมีนให้ค่าความผิดพลาดสูงเมื่อทำนายวันที่มากขึ้น การจัดการกับลักษณะที่ผิดปกติจึงสำคัญมาก การแบ่งกลุ่มข้อมูลแบบดีบีสแกนให้ผลการทำนายที่ดีกว่าการแบ่งกลุ่มข้อมูลแบบเคมีนและเคมีเดียน เพราะอัลกอริทึมนี้สามารถจัดการกับลักษณะที่ผิดปกติได้ดีกว่า อย่างไรก็ตามแบบจำลองการแบ่งกลุ่มแบบผสมทำนายได้ดีที่สุด เพราะสามารถแบ่งกลุ่มของลักษณะได้ชัดเจนและจัดการกับลักษณะที่ผิดปกติได้เท่ากับการแบ่งกลุ่มข้อมูลแบบดีบีสแกน นอกจากนี้พบว่าการแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนให้ค่าความผิดพลาดน้อยกว่าการแบ่งกลุ่มข้อมูลแบบเคมีเดียนกับดีบีสแกนประมาณ 4.6% เมื่อเปรียบเทียบในวันที่ 30 ดังนั้นจึงนำการแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนไปใช้ในแบบจำลองเอฟ7เอ็นเอ็มแอล ไปใช้ในหัวข้อที่ 4.4 เพื่อเปรียบเทียบกับแบบจำลองอื่นที่ใช้เปรียบเทียบ

ตารางที่ 4.1.2 ผลลัพธ์จากการใช้แบบจำลองการแบ่งกลุ่มสำหรับแบบจำลองเอฟ7เอ็นเอ็มแอล

	เคมีน	**เคมีเดียน**	**ดีบีสแกน**	**เคมีนกับ ดีบีสแกน**	**เคมีเดียนกับ ดีบีสแกน**
วันที่ 8	0.000495	0.000507	0.000506	0.000452	0.000495
วันที่ 9	0.001500	0.001894	0.001884	0.001393	0.001452
วันที่ 10	0.004816	0.004010	0.002805	0.002511	0.002586
วันที่ 11	0.010481	0.006468	0.004273	0.003626	0.003813
วันที่ 12	0.020387	0.009143	0.006040	0.005072	0.005201
วันที่ 13	0.046376	0.011547	0.008120	0.006350	0.007030
วันที่ 14	0.023058	0.014673	0.009470	0.008181	0.008323
วันที่ 15	0.048254	0.020207	0.011507	0.009525	0.009926
วันที่ 16	0.082839	0.027982	0.013683	0.010797	0.011486
วันที่ 17	0.040887	0.033881	0.015864	0.012764	0.014720
วันที่ 18	0.032991	0.040349	0.018073	0.014632	0.015220
วันที่ 19	0.121942	0.046293	0.020521	0.017655	0.017538
วันที่ 20	0.078229	0.053347	0.022887	0.017477	0.018520
วันที่ 21	0.057860	0.060496	0.025432	0.017866	0.020895
วันที่ 22	0.400621	0.067197	0.028203	0.019023	0.020713
วันที่ 23	0.414948	0.071983	0.030437	0.021520	0.022372

วันที่ 24	0.441538	0.076721	0.033336	0.021556	0.023200
วันที่ 25	0.506341	0.079753	0.035707	0.023475	0.024471
วันที่ 26	0.545410	0.081882	0.038893	0.025544	0.025618
วันที่ 27	0.521360	0.084549	0.039402	0.025619	0.027465
วันที่ 28	0.704461	0.088216	0.040827	0.025737	0.029156
วันที่ 29	0.751762	0.092229	0.041232	0.028023	0.031827
วันที่ 30	0.812408	0.096430	0.042328	0.030391	0.031884

4.4 การทดลองเปรียบเทียบกับแบบจำลองที่ใช้เปรียบเทียบ

เป้าหมายของงานวิจัยนี้ คือ เพื่อสร้างแบบจำลองที่ใช้ยอดการดูวิดีโอในวันที่ 1 ถึงวันที่ 7 ทำนายยอดการดูวิดีโอสะสมตั้งแต่วันที่ 8 ถึงวันที่ 30 โดยได้ค่าความผิดพลาดน้อยกว่าแบบจำลองเหล่านี้ ได้แก่ แบบจำลองการถดถอยเชิงเส้น แบบจำลองต้นไม้ตัดสินใจแบบถดถอย แบบจำลองเชิงเส้นตัวแปรเดียว แบบจำลองเชิงเส้นหลายตัวแปร และแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต สำหรับแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตและแบบจำลองเอฟ7เอ็นเอ็มแอลจะกำหนดไฮเปอร์พารามิเตอร์ตามตารางที่ 4.2.1 การทดลองนี้ใช้การค้นหาแบบตะแกรง สำหรับหาไฮเปอร์พารามิเตอร์ที่ดีที่สุดในแต่ละรอบ จากนั้นนำไฮเปอร์พารามิเตอร์นั้นมาใช้ทำนายและวัดประสิทธิภาพ

ตารางที่ 4.2.1 ไฮเปอร์พารามิเตอร์สำหรับแบบจำลองการทำนายยอดการดูวิดีโอ

แบบจำลอง	**ไฮเปอร์พารามิเตอร์**
ต้นไม้ตัดสินใจแบบถดถอย	$\text{max\_depth} \in \{5,6,7,\ldots,20\}$
ทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต	$\alpha \in \{0.5,0.8\}, k \in \{2,3,4,5,6\}$
เอฟ7เอ็นเอ็มแอล โดยใช้การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกน	$k \in \{2,3,4,5,6\}$, $\text{epsilon} \in \{0.1,0.2,0.3,0.4,0.5\}, \text{minPts} = 1$

จากตารางที่ 4.2.2 พบว่าแบบจำลองการถดถอยเชิงเส้นและต้นไม้ตัดสินใจแบบถดถอยให้ผลลัพธ์ที่แย่ที่สุด ไม่เหมาะสำหรับการใช้ทำนายยอดการดูวิดีโอสะสม แบบจำลองเชิงเส้นตัวแปรเดียวทำนายได้ดีกว่าแบบจำลองก่อนหน้านี้ แต่ยังไม่ดีพอที่จะทำนายได้ดีเท่าแบบจำลองอื่น ส่วนแบบจำลองเอฟ7เอ็นเอ็มแอลให้ค่าความผิดพลาดน้อยที่สุดสำหรับการทำนายทุกวัน เมื่อวัด

ประสิทธิภาพในวันที่ 30 แบบจำลองนี้มีค่าความผิดพลาดน้อยกว่าแบบจำลองเชิงเส้นหลายตัวแปร และแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตประมาณ 34% และ 27% ตามลำดับ ข้อเสียของแบบจำลองเชิงเส้นหลายตัวแปรคือแบบจำลองเรียนรู้ทุก ๆ ลักษณะโดยที่ไม่มีการแบ่งกลุ่มและจัดการกับลักษณะที่ผิดปกติ ซึ่งสิ่งเหล่านี้ทำให้ค่าความผิดพลาดสูงยิ่งขึ้น แบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตทำนายได้ดีกว่าแบบจำลองเชิงเส้นหลายตัวแปรประมาณ 9% เมื่อเปรียบเทียบค่าความผิดพลาดในวันที่ 30 เพราะว่าแบบจำลองดังกล่าวสามารถแบ่งกลุ่มของลักษณะยอดการดูวิดีโอโดยใช้แบบจำลองการแบ่งกลุ่มข้อมูลแบบเคมีน แต่จะจัดการกับลักษณะที่ผิดปกติได้ไม่ดี ซึ่งจากรูปที่ 3.2 ชุดข้อมูลมีลักษณะที่ผิดปกติอยู่มาก แบบจำลองที่ดีสำหรับชุดข้อมูลนี้ควรจะจัดการกับลักษณะที่ผิดปกติได้ดีก่อนการเรียนรู้ แบบจำลองเอฟ7ดีเอ็นซีเอ็มแอลที่ใช้การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนเป็นแบบจำลองที่เหมาะสม เพราะว่าแบบจำลองนี้สามารถแบ่งกลุ่มลักษณะยอดการดูวิดีโอสะสมได้โดยใช้การแบ่งกลุ่มข้อมูลแบบเคมีนซึ่งแบ่งแต่ละลักษณะได้อย่างชัดเจน กับการแบ่งกลุ่มข้อมูลแบบดีบีสแกนซึ่งจัดการกับลักษณะที่ผิดปกติได้อย่างดี แบบจำลองเหล่านี้ทำให้แบบจำลองเอฟ7เอ็นเอ็มแอลทำนายได้ดีกว่าแบบจำลองอื่น

ตารางที่ 4.2.2 ผลลัพธ์จากการใช้แบบจำลองการทำนายยอดการดูวิดีโอสะสมตั้งแต่วันที่ 8 ถึง 30

	การถดถอยเชิงเส้น	**ต้นไม้ตัดสินใจแบบถดถอย**	**เชิงเส้นตัวแปรเดียว**	**เชิงเส้นหลายตัวแปร**	**ทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต**	**เอฟ7เอ็นเอ็มแอล**
วันที่ 8	7.308197	0.001879	0.001190	0.000592	0.000582	0.000452
วันที่ 9	25.886014	0.008895	0.004231	0.002512	0.002441	0.001393
วันที่ 10	62.801536	0.008429	0.015433	0.007374	0.006978	0.002511
วันที่ 11	118.782370	0.086764	0.017862	0.012979	0.012513	0.003626
วันที่ 12	176.959093	0.065566	0.029331	0.019111	0.017322	0.005072
วันที่ 13	222.880112	0.084217	0.042645	0.022304	0.020217	0.006350
วันที่ 14	278.819891	0.116659	0.044942	0.023844	0.022266	0.008181
วันที่ 15	368.420342	0.277852	0.047213	0.025644	0.022822	0.009525
วันที่ 16	124.582381	0.385533	0.049429	0.027388	0.024960	0.010797

วันที่ 17	150.267012	0.746322	0.061491	0.029282	0.026422	0.012764
วันที่ 18	150.240941	0.691737	0.063499	0.031358	0.028488	0.014632
วันที่ 19	179.606989	1.049833	0.065482	0.033143	0.030528	0.017655
วันที่ 20	196.949207	0.512755	0.067368	0.034360	0.030590	0.017477
วันที่ 21	222.873726	3.539733	0.069168	0.035601	0.032039	0.017866
วันที่ 22	267.817387	0.971608	0.070879	0.036833	0.032981	0.019023
วันที่ 23	318.361864	0.371905	0.072547	0.038120	0.035000	0.021520
วันที่ 24	375.879667	0.438172	0.074130	0.039320	0.036682	0.021556
วันที่ 25	431.945614	1.183004	0.075635	0.040473	0.038412	0.023475
วันที่ 26	446.656963	1.014633	0.077141	0.041745	0.039356	0.025544
วันที่ 27	464.505307	4.777031	0.078734	0.042902	0.040316	0.025619
วันที่ 28	482.759306	1.088572	0.080327	0.043970	0.041101	0.025737
วันที่ 29	484.254071	1.399194	0.082113	0.044950	0.041698	0.028023
วันที่ 30	507.150139	5.011909	0.083725	0.045894	0.041752	0.030391

4.5 การทดลองใช้แบบจำลองทำนายยอดการดูวิดีโอสะสมปัจจุบัน

นอกจากนี้มีการเก็บยอดการดูวิดีโอสะสมทุกวิดีโอในชุดข้อมูลในวันที่ 27 พฤษภาคม 2563 เวลา 8.30น. จากยูทูบ โดยใช้ไลบรารีบิวตี้ฟูลซุป (Beautiful Soup) ของไพธอน เพื่อทดสอบประสิทธิภาพของแบบจำลองกับการทำนายยอดการดูในปัจจุบัน กำหนดไฮเปอร์พารามิเตอร์ตามตารางที่ 4.1.1 และตารางที่ 4.2.1 การทดลองนี้ใช้การค้นหาแบบตะแกรง สำหรับหาไฮเปอร์พารามิเตอร์ที่ดีที่สุดในแต่ละรอบ จากนั้นนำไฮเปอร์พารามิเตอร์นั้นมาใช้ทำนายและวัดประสิทธิภาพ ผลลัพธ์ถูกแสดงในตารางที่ 4.3

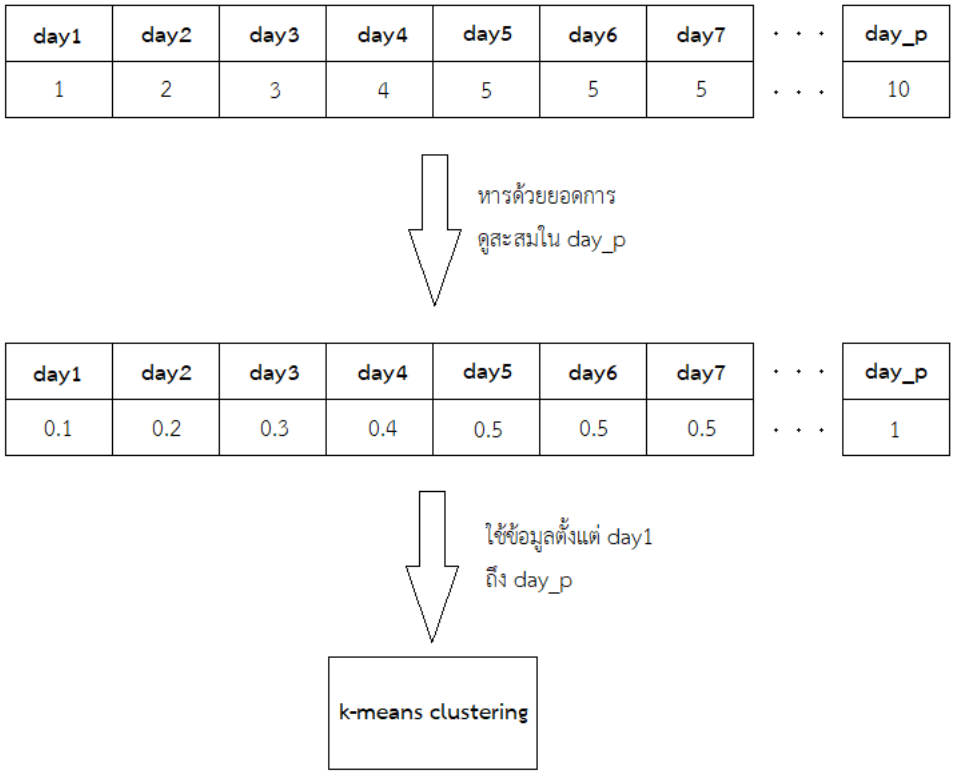
ตารางที่ 4.3 ผลลัพธ์จากการใช้แบบจำลองการทำนายยอดการดูวิดีโอสะสมปัจจุบัน

แบบจำลอง	**ค่าความผิดพลาดกำลังสองที่สัมพันธ์กันแบบเฉลี่ย**
การถดถอยเชิงเส้น	50357.181230
ต้นไม้ตัดสินใจแบบถดถอย	66.184436
เชิงเส้นตัวแปรเดียว	0.318225
เชิงเส้นหลายตัวแปร	0.225577

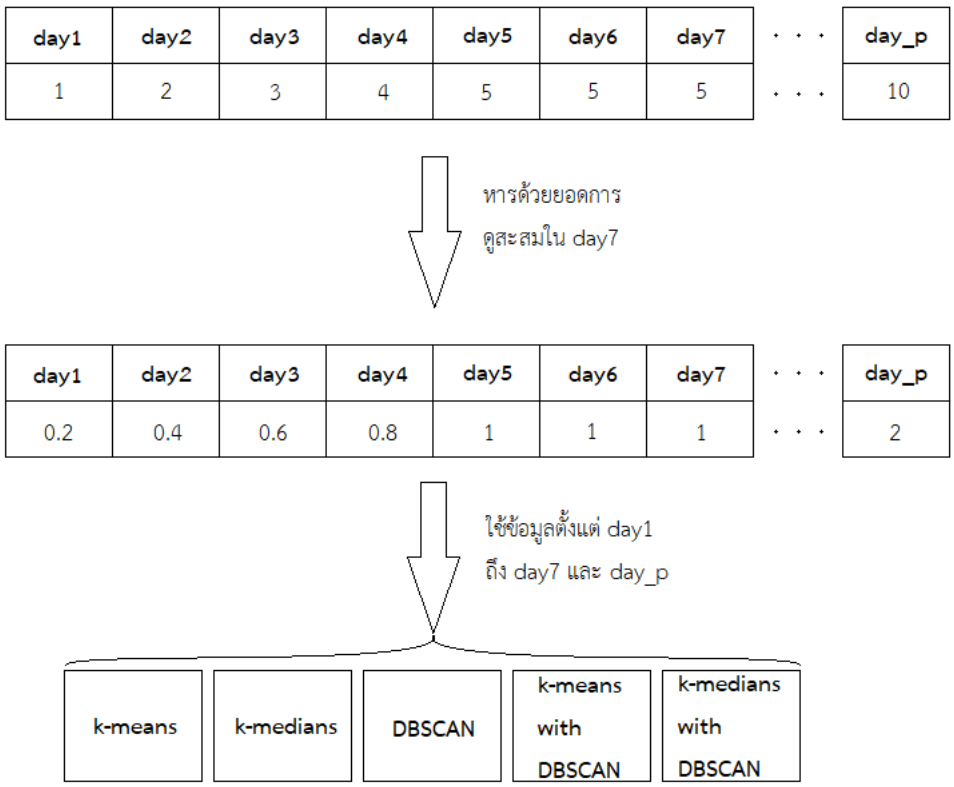
ทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต	0.736071
เอฟ7เอ็นเอ็มแอลที่ใช้เคมีน	9.774807
เอฟ7เอ็นเอ็มแอลที่ใช้เคมีเดียน	15.530089
เอฟ7เอ็นเอ็มแอลที่ใช้ดีบีสแกน	0.521993
เอฟ7เอ็นเอ็มแอลที่ใช้เคมีนกับดีบีสแกน	0.183841
เอฟ7เอ็นเอ็มแอลที่ใช้เคมีเดียนกับดีบีสแกน	0.242841

4.6 การทดลองใช้วิธีการแบ่งกลุ่มลักษณะยอดการดูวิดีโอสะสมจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตกับแบบจำลองเอฟ7เอ็นเอ็มแอล

ความแตกต่างที่ชัดเจนระหว่างแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตกับแบบจำลองเอฟ7เอ็นเอ็มแอล คือ การแบ่งลักษณะของยอดการดูวิดีโอสะสมของชุดข้อมูลเรียนรู้ ซึ่งมีการจัดการลักษณะข้อมูลชุดข้อมูลเรียนรู้ที่แตกต่างกันก่อนนำข้อมูลเหล่านั้นไปใช้แบ่งกลุ่มลักษณะของยอดการดูวิดีโอสะสม



รูปที่ 4.1.1 วิธีการแบ่งกลุ่มจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิต



รูปที่ 4.1.2 วิธีการแบ่งกลุ่มจากแบบจำลองเอฟ7เอ็นเอ็มแอล

แบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตใช้ยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันสุดท้ายที่ต้องการทำนาย หารด้วยยอดการดูวิดีโอสะสมวันสุดท้าย ส่วนแบบจำลองเอฟ7เอ็นเอ็มแอลใช้ยอดการดูวิดีโอสะสมตั้งแต่วันที่ 1 ถึงวันที่ 7 และวันที่ต้องการทำนายหารด้วยยอดการดูวิดีโอสะสมวันที่ 7 ก่อนนำชุดข้อมูลไปใช้แบ่งกลุ่มลักษณะยอดการดูวิดีโอสะสมโดยใช้แบบจำลองการแบ่งกลุ่ม แสดงให้เห็นชัดเจนตามรูปที่ 4.1.1 และ 4.1.2

การทดลองนี้ทดลองใช้วิธีการแบ่งกลุ่มลักษณะยอดการดูวิดีโอสะสมจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตประยุกต์ใช้กับแบบจำลองเอฟ7เอ็นเอ็มแอล จากการแบ่งกลุ่มข้อมูล 5 อย่าง ได้แก่ การแบ่งกลุ่มข้อมูลแบบเคมีน การแบ่งกลุ่มข้อมูลแบบเคมีเดียน การแบ่งกลุ่มข้อมูลแบบดีบีสแกน การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกน และการแบ่งกลุ่มข้อมูลแบบเคมีเดียนกับดีบีสแกน กำหนดไฮเปอร์พารามิเตอร์ตามตารางที่ 4.1.1 ใช้การค้นหาแบบตะแกรงสำหรับหาไฮเปอร์พารามิเตอร์ที่ดีที่สุดในแต่ละรอบ จากนั้นนำไฮเปอร์พารามิเตอร์นั้นมาใช้ทำนายและวัดประสิทธิภาพผลลัพธ์การทำนายเป็นตารางที่ 4.4 แสดงเป็นค่าความผิดพลาดกำลังสองที่สัมพันธ์กันแบบเฉลี่ย

ตารางที่ 4.4 ผลลัพธ์จากการใช้วิธีการแบ่งกลุ่มจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตกับแบบจำลองเอฟ7เอ็นเอ็มแอล

	เคมีน	**เคมีเดียน**	**ดีบีสแกน**	**เคมีนกับดีบีสแกน**	**เคมีเดียนกับดีบีสแกน**
วันที่ 8	0.000511	0.000505	0.000506	0.000464	0.000579
วันที่ 9	0.001912	0.001891	0.001884	0.001502	0.001998
วันที่ 10	0.004028	0.004047	0.003961	0.002580	0.003019
วันที่ 11	0.006754	0.006601	0.006417	0.003787	0.004782
วันที่ 12	0.009939	0.009637	0.009238	0.005294	0.007263
วันที่ 13	0.012650	0.012237	0.011697	0.007105	0.009112
วันที่ 14	0.015960	0.015514	0.014813	0.009588	0.012422
วันที่ 15	0.022007	0.021042	0.020196	0.010424	0.014087
วันที่ 16	0.030598	0.029002	0.027562	0.012996	0.017732
วันที่ 17	0.037217	0.035415	0.033088	0.014053	0.021016
วันที่ 18	0.044829	0.042109	0.039257	0.015473	0.022161
วันที่ 19	0.052187	0.048777	0.045306	0.016023	0.021561
วันที่ 20	0.059222	0.055692	0.051602	0.018402	0.022315
วันที่ 21	0.066501	0.062848	0.057825	0.020463	0.024957
วันที่ 22	0.072151	0.068003	0.062495	0.020245	0.025486
วันที่ 23	0.078467	0.073003	0.066861	0.022403	0.027891
วันที่ 24	0.084334	0.078781	0.071493	0.023430	0.028111
วันที่ 25	0.089469	0.083333	0.075387	0.025603	0.030560
วันที่ 26	0.093579	0.088126	0.079109	0.025284	0.030798
วันที่ 27	0.098820	0.092711	0.082925	0.027571	0.034217
วันที่ 28	0.105266	0.097537	0.086654	0.030384	0.036244
วันที่ 29	0.110099	0.102171	0.090409	0.031285	0.035819
วันที่ 30	0.115133	0.108903	0.094247	0.030649	0.036705

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

วิทยานิพนธ์นี้ได้ออกแบบและพัฒนาแบบจำลองสำหรับทำนายยอดการดูวิดีโอสะสม ในบทที่ 4 มีการทดสอบประสิทธิภาพของการแบ่งกลุ่มลักษณะยอดการดูวิดีโอ และเปรียบเทียบกับแบบจำลองอื่น ๆ นอกจากนี้มีการนำมาใช้ทำนายยอดการดูวิดีโอสะสมปัจจุบัน จากผลการทดลองทั้งหมดสามารถสรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1 สรุปผลการวิจัย

หัวข้อที่ 4.3 เป็นการทดลองเลือกแบบจำลองการแบ่งกลุ่มข้อมูลของเอฟ7เอ็นเอ็มแอลเพื่อใช้เปรียบเทียบกับแบบจำลองอื่น ๆ พบว่าการแบ่งกลุ่มข้อมูลแบบเคมีนให้ค่าความผิดพลาดน้อยในช่วงวันแรก ๆ แต่ให้ค่าความผิดพลาดสูงมากเมื่อทำนายยอดการดูวิดีโอสะสมในวันหลัง ๆ นอกจากนี้ยังมีความไม่แน่นอนในค่าความผิดพลาด จึงไม่ควรเลือกใช้การแบ่งกลุ่มข้อมูลแบบเคมีนสำหรับปัญหานี้ ในขณะที่การแบ่งกลุ่มข้อมูลแบบเคมีเดียนให้ค่าความผิดพลาดที่ค่อนข้างแน่นอน แต่ยังมีค่าความผิดพลาดในการทำนายค่อนข้างสูง การแบ่งกลุ่มข้อมูลแบบดีบีสแกนให้ค่าความผิดพลาดที่แน่นอนมากกว่าและให้ค่าความผิดพลาดน้อยกว่าแบบจำลองการแบ่งกลุ่มที่ผ่านมา ส่วนแบบจำลองการแบ่งกลุ่มแบบผสมให้ค่าความผิดพลาดที่ค่อนข้างแน่นอนและน้อยกว่าแบบจำลองพื้นฐานที่ผ่านมา พบว่าการแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนให้ค่าความผิดพลาดน้อยที่สุด รองลงมาคือการแบ่งกลุ่มข้อมูลแบบเคมีเดียนกับดีบีสแกน ซึ่งให้ค่าความผิดพลาดน้อยกว่าประมาณ 4.6% เมื่อเปรียบเทียบการทำนายยอดการดูวิดีโอในวันที่ 30 จึงเลือกการแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนสำหรับหัวข้อที่ 4.4

หัวข้อที่ 4.4 เป็นการทดลองเปรียบเทียบกับแบบจำลองที่ใช้เปรียบเทียบ พบว่าแบบจำลองการถดถอยเชิงเส้นให้ค่าความผิดพลาดที่มากที่สุด และรองลงมาคือแบบจำลองต้นไม้ตัดสินใจแบบถดถอย จึงไม่ควรนำมาใช้แบบจำลองเหล่านี้ทำนายยอดการดูวิดีโอ แบบจำลองเชิงเส้นตัวแปรเดียวทำนายได้ดีกว่าแบบจำลองพื้นฐานก่อนหน้านี้ มีค่าความผิดพลาดที่แน่นอน แต่ยังทำนายได้ไม่ดีพอเมื่อเปรียบเทียบกับแบบจำลองในงานวิจัยอื่น แบบจำลองเชิงเส้นหลายตัวแปรทำนายได้ดีกว่าแบบจำลองเชิงเส้นตัวแปรเดียว แบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตทำนายได้ดีว่าแบบจำลองเชิงเส้นหลายตัวแปร ส่วนแบบจำลองเอฟ7เอ็นเอ็มแอลทำนายได้ดีที่สุด

หัวข้อที่ 4.5 เป็นการทดลองใช้แบบจำลองทำนายยอดการดูวิดีโอสะสมปัจจุบัน ซึ่งเก็บยอดการดูวิดีโอสะสมทุกวิดีโอในชุดข้อมูลในวันที่ 27 พฤษภาคม 2563 เวลา 8.30น. จากยูทูบ พบว่า

แบบจำลองเอฟ7เอ็นเอ็มแอลที่ใช้เคมีนกับดีบีสแกนให้ค่าความผิดพลาดน้อยที่สุด รองลงมาคือแบบจำลองเชิงเส้นหลายตัวแปรและเชิงเส้นตัวแปรเดียว ส่วนแบบจำลองการถดถอยเชิงเส้นและต้นไม้ตัดสินใจแบบถดถอยยังคงให้ค่าความผิดพลาดที่มากที่สุด

หัวข้อที่ 4.6 เป็นการทดลองใช้วิธีการแบ่งกลุ่มลักษณะยอดการดูวิดีโอสะสมจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตกับแบบจำลองเอฟ7เอ็นเอ็มแอล พบว่าการใช้การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนให้ค่าความผิดพลาดการทำนายน้อยที่สุดในทุกวัน แต่เมื่อนำไปเปรียบเทียบกับผลลัพธ์ของแบบจำลองเอฟ7เอ็นเอ็มแอลจากตารางที่ 4.2.2 พบว่าแบบจำลองเอฟ7เอ็นเอ็มแอลจากตารางที่ 4.2.2 ให้ค่าความผิดพลาดการทำนายน้อยกว่าเพียงเล็กน้อย

5.2 อภิปรายผลการทดลอง

การแบ่งกลุ่มลักษณะของยอดการดูวิดีโอเป็นขั้นตอนที่จำแนกลักษณะเป็นกลุ่มเพื่อสร้างแบบจำลองเฉพาะกลุ่มสำหรับทำนายยอดการดูวิดีโอสะสม ขั้นตอนนี้ทำให้ประสิทธิภาพของการทำนายยอดการดูวิดีโอสะสมดีขึ้น แต่การแบ่งกลุ่มลักษณะเหล่านั้นควรมีการจัดการลักษณะที่ผิดปกติที่ดี เพื่อไม่ให้แบบจำลองเฉพาะกลุ่มต้องเจอกับลักษณะที่ผิดปกตินั้นซึ่งนำไปสู่ผลการทำนายที่แย่ลง ซึ่งเห็นผลลัพธ์ได้อย่างชัดเจนในหัวข้อที่ 4.3 แบบจำลองการแบ่งกลุ่มข้อมูลแบบเคมีนมีปัญหากับการเลือกจำนวนกลุ่มที่เหมาะสมและการจัดการกับลักษณะที่ผิดปกติ จึงได้นำแบบจำลองการแบ่งกลุ่มแบบเคมีเดียนและดีบีสแกนมาใช้ในการทดลองการแบ่งกลุ่มลักษณะยอดการดูวิดีโอ ซึ่งแบบจำลองที่เหล่านั้นสนับสนุนการจัดการกับลักษณะที่ผิดปกติได้ดีกว่าการแบ่งกลุ่มข้อมูลแบบเคมีน ซึ่งปรากฏว่าการนำการแบ่งกลุ่มข้อมูลแบบดีบีสแกนมาใช้กับแบบจำลองเอฟ7เอ็นเอ็มแอลสามารถทำนายได้ค่าความผิดพลาดน้อยกว่า สิ่งที่สังเกตได้จากการแบ่งกลุ่มลักษณะยอดการดูวิดีโอคือการแบ่งกลุ่มข้อมูลแบบเคมีนและเคมีเดียนจะได้ลักษณะที่ชัดเจนกว่าแบบดีบีสแกนที่ลักษณะส่วนใหญ่มารวมในกลุ่มเดียว ส่วนกลุ่มที่เหลือเป็นลักษณะที่ผิดปกติ ดังนั้นจึงมีการนำเสนอแบบจำลองการแบ่งกลุ่มแบบผสมซึ่งเป็นแบบจำลองการแบ่งกลุ่มข้อมูลที่ใช้การแบ่งกลุ่มข้อมูลแบบเคมีนหรือเคมีเดียนในการแบ่งกลุ่มข้อมูลก่อน จากนั้นนำแต่ละกลุ่มที่แบ่งจากการแบ่งกลุ่มข้อมูลก่อนหน้านี้มาใช้การแบ่งกลุ่มข้อมูลแบบดีบีสแกน ทำให้แบบจำลองนี้รวมคุณสมบัติด้านการแบ่งกลุ่มที่ชัดเจนและการจัดการลักษณะที่ผิดปกติได้ดี ดังผลลัพธ์ที่เห็นในรูปที่ 3.4 เมื่อนำแบบจำลองการแบ่งกลุ่มแบบผสมมาใช้ในแบบจำลองเอฟ7เอ็นเอ็มแอล พบว่าได้ผลการทำนายที่ดีกว่าการแบ่งกลุ่มแบบดีบีสแกนและการแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนให้ค่าความผิดพลาดในการทำนายน้อยที่สุด

เมื่อนำแบบจำลองเอฟ7เอ็นเอ็มแอลที่ใช้การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนไปเปรียบเทียบกับแบบจำลองอื่น ๆ ตามการทดลองในหัวข้อที่ 4.4 พบว่าทำนายได้ดีที่สุด แบบจำลอง

ส่วนใหญ่ซึ่งรวมไปถึงแบบจำลองเชิงเส้นหลายตัวแปรไม่มีการแบ่งกลุ่มลักษณะของยอดการดูวิดีโอและไม่มีการจัดการกับลักษณะที่ผิดปกติ ทำให้ทำนายได้ไม่ดีที่สุด แบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตมีการแบ่งกลุ่มลักษณะทำให้ทำนายได้ดีขึ้น แต่ไม่มีการจัดการกับลักษณะที่ผิดปกติทำให้ผลการทำนายไม่ดีที่สุด ส่วนแบบจำลองเอฟ7เอ็นเอ็มแอลให้ผลการทำนายที่ดีที่สุด เพราะนอกจากมีการแบ่งลักษณะของยอดการดูวิดีโอแล้วยังมีการจัดการกับลักษณะที่ผิดปกติได้ดีทำให้มีการทำนายออกมาได้ดีที่สุด

เมื่อนำแบบจำลองต่าง ๆ ไปใช้ทำนายยอดการดูวิดีโอสะสมในปัจจุบัน ถึงแม้ว่าแบบจำลองเอฟ7เอ็นเอ็มแอลที่ใช้เคมีนกับดีบีสแกนให้ผลลัพธ์ที่ดีที่สุด แต่แบบจำลองเอฟ7เอ็นเอ็มแอลที่ใช้การแบ่งกลุ่มแบบอื่น ๆ รวมไปถึงแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตก็ให้ผลการทำนายออกมาแย่กว่าแบบจำลองเชิงเส้นหลายตัวแปร หมายความว่าการแบ่งกลุ่มลักษณะของยอดการดูวิดีโอก่อนทำนายยังทำนายได้ไม่ดีเสมอไปกับการทำนายยอดการดูวิดีโอสะสมในระยะยาว นั่นคือต้องมีการเลือกแบบจำลองการแบ่งกลุ่มที่เหมาะสม เพื่อใช้ทำนายยอดการดูวิดีโอสะสมระยะยาว ส่วนแบบจำลองอื่น ๆ ยังถือว่าสอดคล้องกับการทำนายของตั้งแต่วันที่ 8 ถึงวันที่ 30

เมื่อนำวิธีการแบ่งกลุ่มจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตมาใช้กับแบบจำลองเอฟ7เอ็นเอ็มแอล พบว่าการแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกนให้ค่าความผิดพลาดการทำนายน้อยที่สุดในทุกวัน แต่ให้ค่าความผิดพลาดมากกว่าแบบจำลองเอฟ7เอ็นเอ็มแอลที่ใช้การแบ่งกลุ่มข้อมูลแบบเคมีนกับดีบีสแกน หมายความว่าวิธีการแบ่งกลุ่มจากแบบจำลองทำนายค่าต่อเนื่องแบบทราบช่วงชีวิตไม่ได้พัฒนาแบบจำลองเอฟ7เอ็นเอ็มแอลให้ทำนายได้ดีขึ้น และมีการใช้พื้นที่ความจำของเครื่องมากกว่า เพราะใช้จำนวนข้อมูลในการแบ่งลักษณะข้อมูลมากขึ้นเมื่อต้องการทำนายยอดการดูสะสมในวันที่ถัด ๆ ไป

5.3 ปัญหาและอุปสรรคในการดำเนินงาน

1) การเก็บรหัสระบุของวิดีโอผ่านยูทูบเอพีไอบางครั้งได้รหัสระบุของวิดีโอซ้ำกัน จึงต้องจัดการให้รหัสระบุไม่ซ้ำกัน

2) บางวิดีโอไม่สามารถเก็บข้อมูลได้หลังจากเก็บรหัสระบุของวิดีโอเนื่องจากสาเหตุบางอย่าง เช่น วิดีโอถูกลบ วิดีโอผิดกฎของยูทูบ เป็นต้น

3) บางวิดีโอมียอดการดูวิดีโอสะสมน้อยลงในวันถัดไป พบว่าเป็นปัญหาที่สามารถเกิดขึ้นได้ของวิดีโอบนยูทูบ จึงตัดสินใจลบข้อมูลวิดีโอเหล่านั้นทั้งหมด

4) ในการทดลองใช้แบบจำลองทำนายยอดการดูวิดีโอสะสมปัจจุบัน มีบางวิดีโอที่ไม่สามารถเก็บยอดการดูวิดีโอสะสมได้ เนื่องจากวิดีโอเหล่านั้นถูกลบ

5.4 ข้อเสนอแนะ

นอกจากการแบ่งกลุ่มลักษณะของยอดการดูวิดีโอจะทำให้พัฒนาประสิทธิภาพในการทำนายของแบบจำลองเพื่อการทำนายยอดการดูวิดีโอสะสมแล้ว การจัดการกับลักษณะของยอดการดูวิดีโอที่ผิดปกติยังทำให้แบบจำลองทำนายได้ดีขึ้นอีกด้วย เนื่องจากลักษณะที่ผิดปกติเหล่านี้ทำให้เมื่อมีการแบ่งกลุ่มลักษณะและสร้างแบบจำลองเฉพาะกลุ่มเพื่อทำนายลักษณะเฉพาะกลุ่ม แบบจำลองย่อยเหล่านั้นจะเรียนรู้ลักษณะที่ผิดปกติเหล่านั้น ก่อให้เกิดการทำนายที่แย่ลง ทางผู้วิจัยจึงนำเสนอแบบจำลองการแบ่งกลุ่มแบบเคมีนกับดีบีสแกนที่สามารถแบ่งกลุ่มลักษณะและจัดการกับลักษณะที่ผิดปกติได้ดี

แบบจำลองเอฟ7เอ็นเอ็มแอลเน้นใช้อัลกอริทึมสำหรับยอดการดูวิดีโอแต่ละวันในการทำนายยอดการดูวิดีโอสะสม เพราะฉะนั้นแนวทางในการวิจัยถัดไปควรมีการพิจารณาเรื่องคุณลักษณะของวิดีโออื่น ๆ เช่น ความยาวของวิดีโอ ประเภทของวิดีโอ ยอดการแชร์วิดีโอสู่โซเชียลมีเดียอื่น เป็นต้น นอกจากนี้อาจมีการพัฒนาการแบ่งกลุ่มลักษณะของยอดการดูวิดีโอหรือการจัดการลักษณะที่ผิดปกติให้ดียิ่งขึ้น เพื่อให้แบบจำลองที่ถูกพัฒนาทำนายยอดการดูวิดีโอสะสมได้ดียิ่งขึ้น รวมไปถึงควรหาแนวทางในการพัฒนาแบบจำลองที่ทำนายยอดการดูวิดีโอในระยะยาวที่ดีกว่าเดิม

บรรณานุกรม

1. Ding, W., Shang, Y., Guo, L., Hu, X., Yan, R., and He, T. 2015. Video Popularity Prediction by Sentiment Propagation Via Implicit Network. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (Melbourne, Australia, October 19-23, 2015). CIKM '15. ACM, New York, NY, 1621-1630. DOI=https://doi.org/10.1145/2806416.2806505.
2. Chn, J., Song, X., Nie, L., Wang, X., Zhang, H., and Chua, T. S. 2016. Micro Tell Macro: Predicting the Popularity of Micro-Videos Via a Transductive Model. In Proceedings of the 2016 ACM on Multimedia Conference (Amsterdam, Netherlands, October 15-19, 2016). MM '16. ACM, New York, NY, 898-907. DOI=http://dx.doi.org/10.1145/2964284.2964314.
3. Szabo, G. and Huberman, B. A. 2010. Predicting the Popularity of Online Content. Communications of the ACM, 53, 8 (Aug. 2010). ACM, New York, NY, 80-88. DOI=https://doi.org/10.1145/1787234.1787254.
4. Crane R. and Sornette, D. 2008. Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System. In Proceedings of the National Academy of Sciences of the United States of America, 105, 41 (Oct. 2008). National Academy of Sciences, Washington, DC, 15649-15653. DOI=https://doi.org/10.1073/pnas.0803685105.
5. Pinto, H., Almeida, J. M., and Gonçalves, M. A. 2013. Using Early View Patterns to Predict the Popularity of YouTube Videos. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining (Rome, Italy, February 4-8, 2013). WSDM '13. ACM, New York, NY, 365-374. DOI=https://doi.org/10.1145/2433396.2433443.
6. Richier, C., Altman, E., Elazouzi, R., Jimenez, T., Linarès, G., and Portilla, Y. 2014. Bio-Inspired Models for Characterizing YouTube Viewcout. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (Beijing, China, August 17-20, 2014). ASONAM'14. IEEE, Piscataway, New Jersey, 297-305. DOI= https://doi.org/10.1109/ASONAM.2014.6921600.

7. Richier, C., Elazouzi, R., Jimenez, T., Altman, E., and Linarès, G. 2015. Forecasting Online Contents' Popularity.

8. Trzciński, T. and Rokita, P. 2017. Predicting Popularity of Online Videos Using Support Vector Regression. IEEE Transactions on Multimedia, 19, 11 (Nov. 2017). IEEE, Piscataway, New Jersey, 2561-2570. DOI=https://doi.org/10.1109/TMM.2017.2695439.

9. Li, C., Liu, J., and Ouyang, S. 2016. Characterizing and Predicting the Popularity of Online Videos. IEEE Access, 4 (Apr. 2016). IEEE, Piscataway, New Jersey, 1630-1641. DOI=https://doi.org/10.1109/ACCESS.2016.2552218.

10. Ouyang, S., Li, C., and Li, X. 2016. A Peek into the Future: Predicting the Popularity of Online Videos. IEEE Access, 4 (Jun. 2016). IEEE, Piscataway, New Jersey, 3026-3033. DOI=https://doi.org/10.1109/ACCESS.2016.2580911.

11. Mao, Y., Shen, Y., Qin, G., and Cai, L. 2017. Predicting the Popularity of Online Videos via Deep Neural Networks.

12. Ma, C., Yan, Z., and Chen, C. W. 2017. LARM: A Lifetime Aware Regression Model for Predicting YouTube Video Popularity. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (Singapore, November 6-10, 2017). CIKM '17. ACM, New York, NY, 467-476. DOI=https://doi.org/10.1145/3132847.3132997.

13. Chen, Y. and Chang, C. 2019. Early Prediction of The Future Popularity of Uploaded Videos. Expert Systems with Applications, 133 (May. 2019). 59-74. DOI=https://doi.org/10.1016/j.eswa.2019.05.015.

14. Gürsun, G., Crovella, M., and Matta, I. 2011. Describing and Forecasting Video Access Patterns. In 2011 Proceedings IEEE INFOCOM (Shanghai, China, April 10-15, 2011). IEEE, Piscataway, New Jersey, 16-20. DOI=https://doi.org/10.1109/INFCOM.2011.5934965.

15. Wu, J., Zhou, Y., Chiu, D. M., and Zhu, Z. 2016. Modeling Dynamics of Online Video Popularity. IEEE Transactions on Multimedia, 18, 9 (Sep. 2016). IEEE, Piscataway, New Jersey, 1882-1895. DOI=https://doi.org/10.1109/TMM.2016.2579600.

16. Roy, S. D., Mei, T., Zeng, W., and Li, S. 2013. Towards Cross-Domain Learning for Social Video Popularity Prediction. IEEE Transactions on Multimedia, 15, 6 (May. 2013). IEEE, Piscataway, New Jersey, 1255-1267. DOI=https://doi.org/10.1109/TMM.2013.2265079.

17. Li, H., Ma, X., Wang, F., Liu, J., and Xu, K. 2013. On Popularity Prediction of Videos Shared in Online Social Networks. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (San Francisco, CA, USA, October 27-November 1, 2013). CIKM '13. ACM, New York, NY, 169-178. DOI=https://doi.org/10.1145/2505515.2505523.

18. Ahmed, M., Spagna, S., Huici, F., and Niccolini, S. 2013. A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining (Rome, Italy, February 4-8, 2013). WSDM '13. ACM, New York, NY, 607-616. DOI=https://doi.org/10.1145/2433396.2433473.

19. Tan, Z., Wang, Y., Zhang, Y., and Zhou, J. 2016. A Novel Time Series Approach for Predicting the Long-Term Popularity of Online Videos. IEEE Transactions on Broadcasting, 62, 2 (Apr. 2016). IEEE, Piscataway, New Jersey, 436-445. DOI=https://doi.org/10.1109/TBC.2016.2540522.

20. กัณฐิกา พรมมา. **การเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึมการจัดกลุ่มแบบ 2 ขั้นตอน**. วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. 2556.

21. ณัฐกานต์ เอี่ยมอ่อน. **การจัดกลุ่มบทคัดย่อทางวิชาการด้วยอัลกอริทึมเคมีเดียนส์**. วิทยานิพนธ์ปริญญามหาบัณฑิต, มหาวิทยาลัยเชียงใหม่. 2547.

22. วรนัญ โกวิทวณิชกานนท์. **การเปรียบเทียบวิธีการระบุค่าพารามิเตอร์ในวิธี DBSCAN**. วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. 2558.

23. พัชร์ประภา ตั้งเพียร. **ระบบการระบุลายพิมพ์ฝ่ามือโดยใช้การเทียบรูปร่างและขั้นตอนวิธีข้อมูลข้างเคียงที่ใกล้ที่สุด k ตัว**. วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. 2554.

24. Montgomery, C. D., Peck, A. E., and Vining, G. G. 2012. **Introduction to Linear Regression Analysis, 5th Edition**. John Wiley and Sons.

25. Steorts, C. R. **Tree Based Methods: Regression Trees.** Duke University.

ประวัติผู้เขียน

ชื่อ-สกุล	เอกพล วงศ์ศุภรัตน์กุล
วัน เดือน ปี เกิด	16 กรกฎาคม 2537
สถานที่เกิด	สมุทรปราการ
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ มหาวิทยาลัยธรรมศาสตร์ ในปีการศึกษา 2560
ที่อยู่ปัจจุบัน	327/206 หมู่ 9 ตำบลหนองปรือ อำเภอบางละมุง จังหวัดชลบุรี 20150
ผลงานตีพิมพ์	Wongsuparatkul, E. and Sinthupinyo, S. 2020. View Count of Online Videos Prediction Using Clustering View Count Patterns with Multivariate Linear Model. In Proceedings of the 8th International Conference on Computer and Communications Management (Singapore, July 17-19, 2020). ICCCM'20. ACM, New York, NY, 123-129. DOI=https://doi.org/10.1145/3411174.3411186.
รางวัลที่ได้รับ	-