การจำแนกประเภทวิดีโอ TikTok และการปรับค่าพารามิเตอร์ด้วย GridSearchCV

# วัตถุประสงค์

เข้าใจและนำการประเมินผลโมเดลและการปรับค่าพารามิเตอร์ (Hyperparameter Tuning) สำรวจกรณีศึกษาข้อมูลจาก TikTok เป้าหมายของโปรเจกต์นี้คือการจำแนกประเภทวิดีโอ TikTok โดยใช้โมเดลการเรียนรู้ของเครื่อง (Random Forest) และปรับปรุงประสิทธิภาพของโมเดลผ่านการปรับค่าพารามิเตอร์ด้วย GridSearchCV.

# 1. ปัญหา, สมมติฐานการฟิตโมเดล, และสรุปข้อมูล

## ปัญหา

การจำแนกประเภทวิดีโอ TikTok ตามคุณลักษณะต่าง ๆ เช่น จำนวนการดาวน์โหลด, จำนวนไลค์, จำนวนความคิดเห็น, และจำนวนการแชร์ โดยมีเป้าหมายในการทำนายสถานะการอ้างสิทธิ์ (claim status) ของวิดีโอว่าจะเป็น 'claim' หรือ 'opinion'

## สมมติฐานการฟิตโมเดล

สมมติฐานที่ทำขึ้นเมื่อฟิตโมเดลรวมถึง:  
1.ฟีเจอร์ที่เลือกสำหรับการฝึกโมเดลมีความเกี่ยวข้องและมีความสัมพันธ์กับตัวแปรเป้าหมาย (claim status).  
2. ข้อมูลที่ใช้เป็นตัวแทนของกลุ่มประชากรของวิดีโอ TikTok โดยรวม.  
3. โมเดล Random Forest เหมาะสมสำหรับการจัดการงานการจำแนกประเภทนี้ และการปรับค่าพารามิเตอร์จะช่วยในการเพิ่มประสิทธิภาพของโมเดล.

## สรุปข้อมูล

ชุดข้อมูลประกอบด้วยคุณลักษณะต่าง ๆ ที่เกี่ยวข้องกับวิดีโอ TikTok เช่น:  
1. video\_download\_count  
2. video\_like\_count  
3. video\_comment\_count  
4. video\_share\_count  
5. video\_transcription\_text\_length  
6. claim\_status (ตัวแปรเป้าหมาย[Y])

# 2. วิธีการเตรียมข้อมูล (Preprocessing Approach)

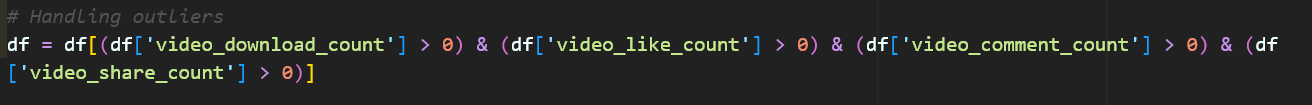
ขั้นตอนการเตรียมข้อมูลรวมถึง:  
1. การลบคอลัมน์ที่ไม่จำเป็น: คอลัมน์ '#', และ 'video\_id' ถูกลบออกเนื่องจากไม่มีส่วนช่วยในการทำนายโมเดล.

A black background with white text

Description automatically generated  
2. การจัดการค่าที่ขาดหายไป: แถวที่มีค่าที่ขาดหายไปถูกลบออกเพื่อรักษาความสมบูรณ์ของข้อมูล.

A black background with white text

Description automatically generated  
3. การลบข้อมูลซ้ำ: แถวข้อมูลที่ซ้ำกันถูกลบออกเพื่อหลีกเลี่ยงการเกิด bias ในโมเดล.

A black background with white text

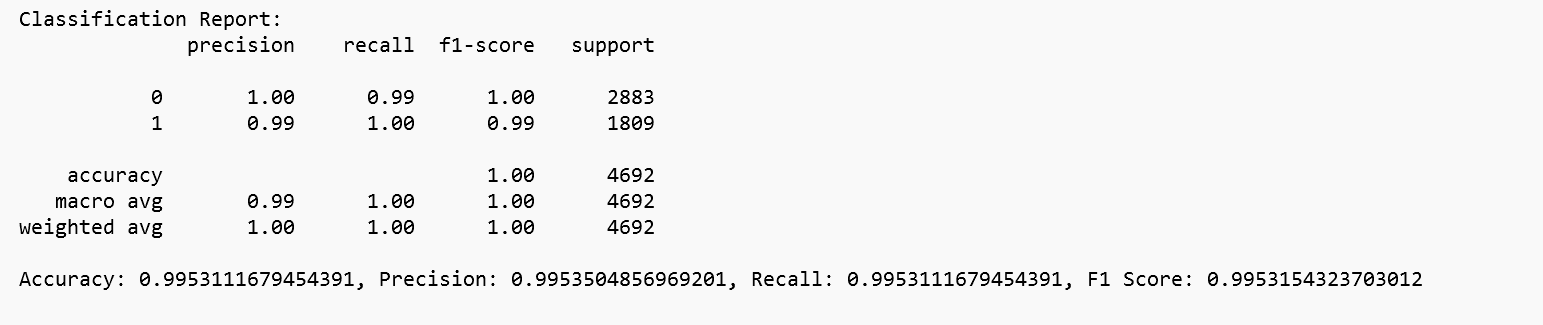
Description automatically generated  
4. การจัดการกับ outliers: แถวที่มีค่า outlier ในคอลัมน์ 'video\_download\_count', 'video\_like\_count', 'video\_comment\_count', และ 'video\_share\_count' ถูกลบออก โดยเฉพาะข้อมูลที่มีค่าน้อยกว่าหรือเท่ากับ 0.

# 3. การฝึกและปรับแต่งโมเดล, การประเมินผล, และผลลัพธ์

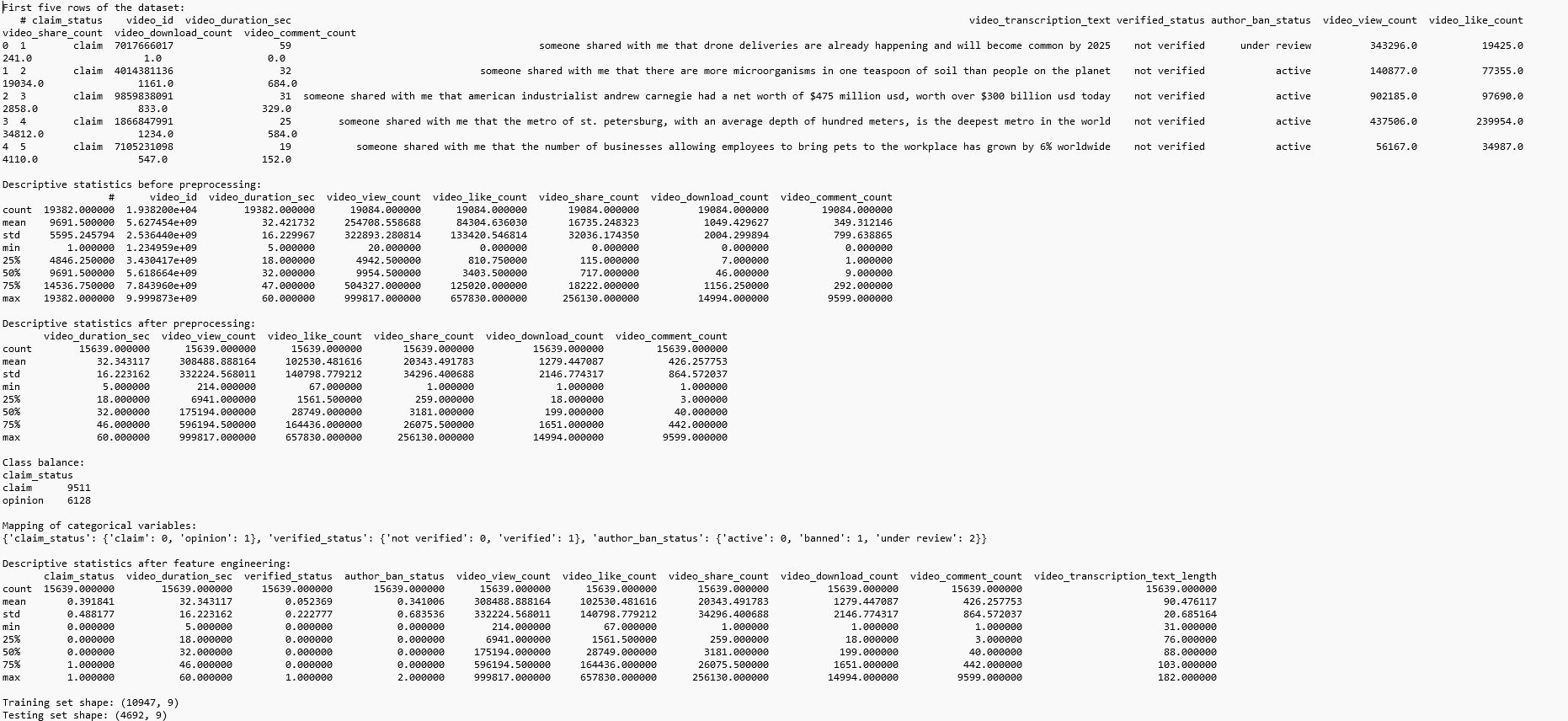
ใช้โมเดล Random Forest โดยมีการปรับค่าพารามิเตอร์ของโมเดลโดยใช้ GridSearchCVเพื่อค้นหาค่าพารามิเตอร์ที่ดีที่สุดที่ช่วยเพิ่มประสิทธิภาพของโมเดลขั้นตอนการปรับค่าพารามิเตอร์รวมถึงการค้นหาค่าที่ดีที่สุดในกริดของพารามิเตอร์ เช่น 'n\_estimators', 'max\_depth', 'max\_samples', 'min\_samples\_split', และ 'min\_samples\_leaf'.

A screen shot of a computer code

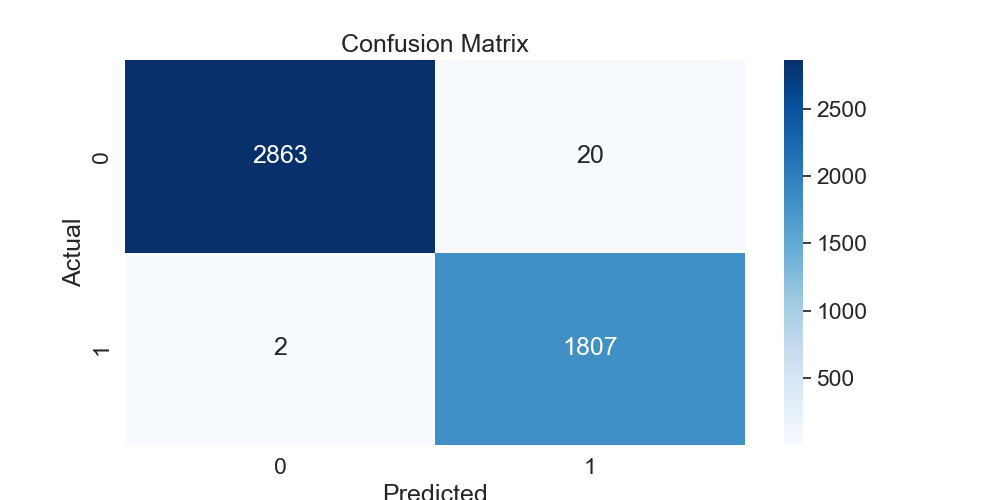
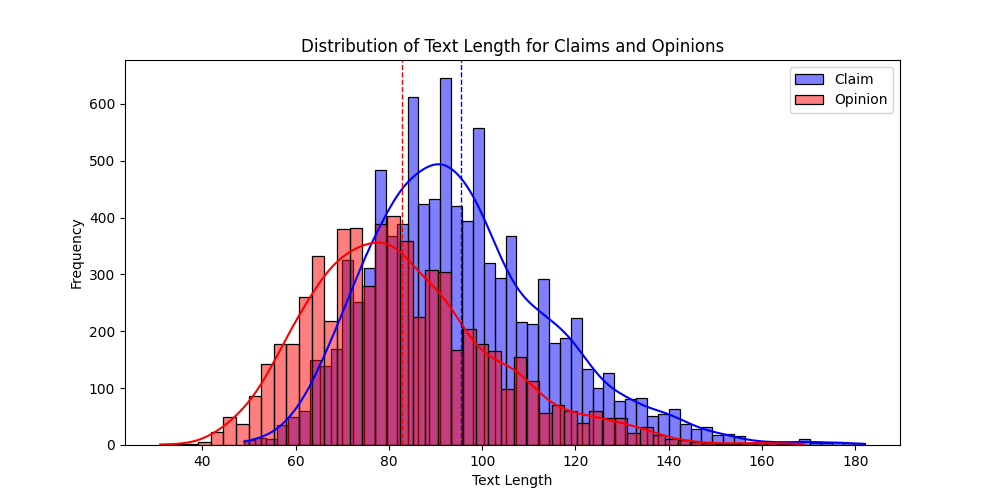
Description automatically generated  
โมเดลที่ดีที่สุดถูกเลือกโดยดูจากค่า accuracy และประสิทธิภาพของโมเดลถูกประเมินโดยใช้ metrics ต่าง ๆ เช่น precision, recall, F1 score และ accuracy.

การประเมินผลของโมเดลทำได้บนชุดข้อมูลทดสอบ และผลลัพธ์แสดงให้เห็นว่าโมเดลสามารถจำแนกประเภทวิดีโอได้อย่างแม่นยำ นอกจากนี้ยังได้สร้าง Confusion Matrix เพื่อแสดงผลการทำนายของโมเดลในการแยกแยะระหว่างวิดีโอ 'claim' และ 'opinion'.

**การบันทึกผลลัพธ์ลงไฟล์ output.txt**:

 ผลลัพธ์ที่ได้จากการพิมพ์จะถูกบันทึกลงในไฟล์ output.txt เพื่อให้สามารถเปิดดูได้ภายหลัง ซึ่งช่วยป้องกันการสูญหายของข้อมูลที่พิมพ์ในกรณีที่ Terminal ไม่สามารถแสดงผลทั้งหมดได้(มีการแก้ไข code ให้บันทึกผลลัพธ์ลง output.txt เนื่องจาก buffer ของ terminal ไม่เพียงพอในการแสดงผลลัพธ์ทั้งหมด)

# 4. ภาพกราฟที่ได้



# 5. แหล่งอ้างอิง

https://github.com/mlkklkhl/Machine-Learning-Course/tree/main