LAPORAN AKHIR SEMESTER MACHINE LEARNING

DETEKSI EMOSI PENGGUNA TWEETER

Diajukan Untuk Memenuhi Tugas

Mata Kuliah Machine Learning

Yang diampu oleh:

Ibu Adevian Fairuz Pratama, S.S.T, M.Eng.

Semester Genap Tahun Akademik 2022/2023

Disusun Oleh:

Iqri Mannisa' Buchori **(2041720066 / 12)**

Wazir Qorni Abud **(2041720124 / 21)**

PROGRAM STUDI D-IV TEKNIK INFORMATIKA

JURUSAN TEKNOLOGI INFORMASI

POLITEKNIK NEGERI MALANG

2022

1. Prepocessing Data

Data yang akan diolah untuk project UAS kali ini adalah data *tweet_emotions.csv*. Tahap pertama yang dapat dilakukan adalah load dataset kedalam Dataframe menggunakan Pandas.

Project UAS

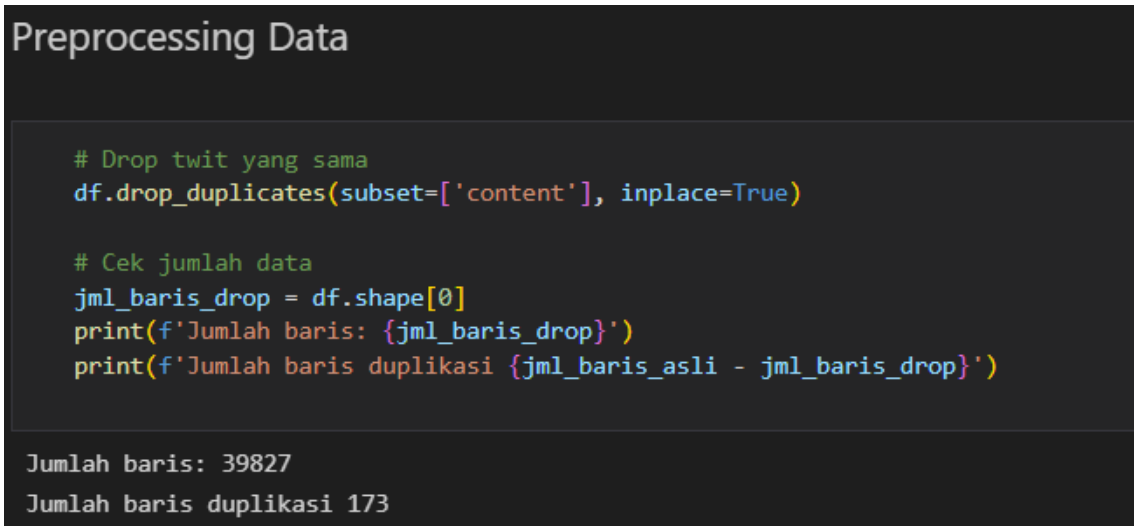
```
[2]
import numpy as np
import pandas as pd
```

```
[3]
df = pd.read_csv('tweet_emotions.csv')
display(df.head())

jml_baris_asli = df.shape[0]
print(f'Jumlah baris: {jml_baris_asli}')
```

	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...

```
Jumlah baris: 40000
```

Setelah data terbaca oleh fungsi python, didapati jumlah baris yang terdapat pada data *tweet_emotions.csv* berjumlah 40.000. Kemudian masuk pada tahap Prepocessing Data dimana akan dilakukan inisialisasi terlebih dahulu apakah didapati data yang terduplikasi dari 40.000 data yang ada. Data yang terduplikasi akan dihapus dengan menggunakan method drop_duplicates.



Preprocessing Data

```
# Drop twit yang sama
df.drop_duplicates(subset=['content'], inplace=True)

# Cek jumlah data
jml_baris_drop = df.shape[0]
print(f'Jumlah baris: {jml_baris_drop}')
print(f'Jumlah baris duplikasi {jml_baris_asli - jml_baris_drop}')
```

```
Jumlah baris: 39827
Jumlah baris duplikasi 173
```

Setelah menghapus data yang terduplikasi masuk pada tahap operasi dasar yang digunakan pada tahap pra pengolahan data adalah Case Folding, Tokenizing, Filtering, dan Stemming. Namun sebelum itu hal yang perlu diperhatikan lagi adalah menghapus mention @.

Remove Mention (@)

```
import re # python regex lib

df = df.copy()

# Membuat kolom baru untuk kebutuhan berbandingan
df['content_clean'] = df['content']

# Membuat fungsi lambda untuk membuat mention, url
rm_rt_url = lambda x: re.sub('(@[A-Za-z0-9\w]+) | (@\w+:) | (\w+:\/\/\S+) | (www.\S+)',' ',x)
rm_punct = lambda x: re.sub('\W', ' ', x)

# Membuat fungsi untuk membuang protocol internet

# Map filter
df['content_clean'] = df.content_clean.map(rm_rt_url).map(rm_punct)
df.head(100)
```

[5]

	tweet_id	sentiment	content	content_clean
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	Layin n bed with a headache ughhhh waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...	Funeral ceremony gloomy friday
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends SOON
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	We want to trade with someone who has Houston...
...	...	...	...	...
95	1956989514	sadness	@sweeetnspicy hiii im on my ipod...i cant fall...	hiii im on my ipod i cant fall asleep
96	1956989526	sadness	dont wanna work 11-830 tomorrow but i get paid	dont wanna work 11 830 tomorrow but i get paid
97	1956989560	sadness	feels sad coz i wasnt able to play with the gu...	feels sad coz i wasnt able to play with the gu...
98	1956989561	neutral	PrinceCharming	PrinceCharming
99	1956989601	hate	@ cayogial i wanted to come to BZ this summer ...	cayogial i wanted to come to BZ this summer ...

100 rows × 4 columns

a) Case Folding

Case Folding digunakan untuk mengubah semua bentuk huruf dalam sebuah teks atau mengubah isi dokumen menjadi huruf kecil semua. Sementara itu, karakter lain yang bukan termasuk huruf dan angka, seperti tanda baca dan spasi dianggap sebagai delimiter. Delimiter ini bisa juga dihapus atau diabaikan dengan menggunakan perintah yang ada di Python.

Case Folding

```
df['content_clean'] = df.content_clean.str.lower()
df.head(10)
```

[33] ✓ 0.7s

	tweet_id	sentiment	content	content_clean
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...
5	1956968477	worry	Re-pinging @ghostridah14: why didn't you go to...	re pinging why didn t you go to prom bc my bf...
6	1956968487	sadness	I should be sleep, but im not! thinking about ...	i should be sleep but im not thinking about ...
7	1956968636	worry	Hmmm. http://www.djhero.com/ is down	hmmm is down
8	1956969035	sadness	@charviray Charlene my love. I miss you	charlene my love i miss you
9	1956969172	sadness	@kelcouch I'm sorry at least it's Friday?	i m sorry at least it s friday

b) Tokenizing

Ditahap ini akan dilakukan proses number removal, whitecase removal, puntuation removal dan word_tokenize() untuk memecah string kedalam tokens. Pandas Dataframe atau Series mampu menjalankan function external untuk di terapkan pada kolom atau baris dengan menggunakan fungsi *.apply*().

Tokenizing

```
from nltk.tokenize import TweetTokenizer
df_stem = df.copy()

tweet_token = TweetTokenizer()
df_stem['content_token'] = df_stem['content_clean'].apply(tweet_token.tokenize)

df_stem.head()
```

	tweet_id	sentiment	content	content_clean	content_token
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...	[i, know, i, was, listenin, to, bad, habit, ea...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[layin, n, bed, with, a, headache, ughhhh, wai...
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[funeral, ceremony, gloomy, friday]
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[wants, to, hang, out, with, friends, soon]
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[we, want, to, trade, with, someone, who, has,...

c) Filtering

Pada tahap ini kita akan menggunakan stopword bahasa English yang didapatkan dari library NLTK untuk filtering terhadap Dataframe. Tahapan filtering yang digunakan untuk mengambil kata-kata yang penting dari hasil token tadi. Kata umum yang biasanya muncul dan tidak memiliki makna disebut dengan stopword.

Filtering

```
from nltk.corpus import stopwords
list_stopwords = stopwords.words('english')

txt_stopword = pd.read_csv("tweet_emotions.csv", names= ["stopwords"], header = None)

# convert stopword string to list & append additional stopword
list_stopwords.extend(txt_stopword["stopwords"][0].split(' '))

# convert list to dictionary
list_stopwords = set(list_stopwords)

#remove stopword pada list token
def stopwords_removal(words):
    return [word for word in words if word not in list_stopwords]

df_stem['content_filtering'] = df_stem['content_token'].apply(stopwords_removal)

df_stem.head()
```

✓ 0.5s

	tweet_id	sentiment	content	content_clean	content_token	content_filtering
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[layin, n, bed, with, a, headache, ughhhh, wai...	[layin, n, bed, headache, ughhhh, waitin, call]
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[funeral, ceremony, gloomy, friday]	[funeral, ceremony, gloomy, friday]
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[wants, to, hang, out, with, friends, soon]	[wants, hang, friends, soon]
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[we, want, to, trade, with, someone, who, has,...	[want, trade, someone, houston, tickets, one]

d) Stemming

Stemming adalah proses mengurangi infleksi kata-kata ke bentuk akarnya, seperti memetakan sekelompok kata ke batang yang sama, bahkan jika batang itu sendiri bukan kata yang valid dalam Bahasa.

Stemming

```
# Stemming
from nltk.stem import SnowballStemmer

stemmer = SnowballStemmer("english")

def stemming(text):
    stem_text = [stemmer.stem(word) for word in text]
    return stem_text

df_stem['content_stem'] = df_stem['content_filtering'].apply(lambda x: stemming(x))

df_stem.head()
```

✓ 8.3s

	tweet_id	sentiment	content	content_clean	content_token	content_filtering	content_stem
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...	[know, listenin, bad, habit, earlier, start, f...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[layin, n, bed, with, a, headache, ughhhh, wai...	[layin, n, bed, headache, ughhhh, waitin, call]	[layin, n, bed, headach, ughhhh, waitin, call]
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[funeral, ceremony, gloomy, friday]	[funeral, ceremony, gloomy, friday]	[funer, ceremoni, gloomi, friday]
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[wants, to, hang, out, with, friends, soon]	[wants, hang, friends, soon]	[want, hang, friend, soon]
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[we, want, to, trade, with, someone, who, has,...	[want, trade, someone, houston, tickets, one]	[want, trade, someon, houston, ticket, one]

2. Clustering

Pengelompokan data ke dalam beberapa kategori atau cluster, yaitu komentar positif, netral, dan negatif. Clustering adalah sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum.

Clustering

```
# Import TextBlog Package

from textblob import TextBlob

# Membuat fungsi untuk menghitung polarity
def get_polarity(text):
    return TextBlob(text).sentiment.polarity

df_stem['polarity'] = df_stem['content_clean'].apply(get_polarity)
```

13.0s Python

```
def condition(c):
    if c>0:
        return "Positif"
    elif c==0:
        return "Neutral"
    else:
        return 'Negatif'

df_stem['sentiment_cluster'] = df_stem['polarity'].apply(condition)

df_stem.head()
```

0.1s Python

	tweet_id	sentiment	content	content_clean	content_token	content_filtering	content_stem	polarity	sentiment_cluster
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...	[know, listenin, bad, habit, earlier, start, f...	-0.35	Negatif
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[layin, n, bed, with, a, headache, ughhhh, wai...	[layin, n, bed, headache, ughhhh, waitin, call]	[layin, n, bed, headach, ughhhh, waitin, call]	0.00	Neutral
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[funeral, ceremony, gloomy, friday]	[funeral, ceremony, gloomy, friday]	[funer, ceremoni, gloomi, friday]	0.00	Neutral
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[wants, to, hang, out, with, friends, soon]	[wants, hang, friends, soon]	[want, hang, friend, soon]	0.20	Positif
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[we, want, to, trade, with, someone, who, has,...	[want, trade, someone, houston, tickets, one]	[want, trade, someon, houston, ticket, one]	0.00	Neutral

3. Labeling

Tahap selanjutnya adalah Labeling dimana hasil dari pengelompokan atau cluster diatas diberikan label hasil dari pengelompokan example melalui clustering. Seperti hasil kode program berikut yang memberikan label dengan keterangan Positif =1, Neutral = 0, Negatif = -1.

Labeling

```
# Labeling sentiment_cluster and make new column with name labeling from sentiment_cluster
df_stem['labeling'] = df_stem['sentiment_cluster'].map({'Positif': 1, 'Neutral': 0, 'Negatif': -1})
df_stem.head()
```

0.2s Python

	tweet_id	sentiment	content	content_clean	content_token	content_filtering	content_stem	polarity	sentiment_cluster	labeling
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...	i know i was listenin to bad habit earlier a...	[i, know, i, was, listenin, to, bad, habit, ea...	[know, listenin, bad, habit, earlier, started,...	[know, listenin, bad, habit, earlier, start, f...	-0.35	Negatif	-1
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[layin, n, bed, with, a, headache, ughhhh, wai...	[layin, n, bed, headache, ughhhh, waitin, call]	[layin, n, bed, headach, ughhhh, waitin, call]	0.00	Neutral	0
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[funeral, ceremony, gloomy, friday]	[funeral, ceremony, gloomy, friday]	[funer, ceremoni, gloomi, friday]	0.00	Neutral	0
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[wants, to, hang, out, with, friends, soon]	[wants, hang, friends, soon]	[want, hang, friend, soon]	0.20	Positif	1
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[we, want, to, trade, with, someone, who, has,...	[want, trade, someone, houston, tickets, one]	[want, trade, someon, houston, ticket, one]	0.00	Neutral	0

Setelah melakukan labeling perlu dicek kembali jumlah data yang telah dibuatkan labeling sesuai dengan pengelompokan sebelumnya dengan memanfaatkan *method .value_counts().*

```
# Cek jumlah data Pastikan Sesuai
print(df_stem['sentiment_cluster'].value_counts())
print(df_stem['labeling'].value_counts())
```

0.2s

```
Positif    18027
Neutral    13619
Negatif     8181
Name: sentiment_cluster, dtype: int64
 1    18027
 0    13619
-1     8181
Name: labeling, dtype: int64
```

4. Classification

Pada tahapn classification ini menggunakan ekstraksi fitur TfIdfVEctorizer dan metode klasifikasi Naïve Bayes. Hal ini digunakan untuk klasifikasi teks yang melibatkan set data pelatihan dimensi tinggi.

Classification

```
# Buat Clasification with naive bayes
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer

# Split data
X_train, X_test, y_train, y_test = train_test_split(df_stem['content_clean'], df_stem['labeling'], test_size=0.2, random_state=42)

# Vectorize
tfidf = TfidfVectorizer()

X_train = tfidf.fit_transform(X_train)
X_test = tfidf.transform(X_test)

# Import Naive Bayes
from sklearn.naive_bayes import MultinomialNB

# Train model
model = MultinomialNB()
model.fit(X_train, y_train)

# Predict
y_pred = model.predict(X_test)

# Evaluation
label = {1: 'Positif', 0: 'Neutral', -1: 'Negatif'}
y_test = y_test.map(label)
y_pred = pd.Series(y_pred).map(label)
```

✓ 2.2s

Proses yang dilakukan mulai dari import library yang akan digunakan, kemudian melakukan split data, modeling dengan menggunakan TfidfVectorizer(), train model, melakukan prediksi atas model yang dibuat dan melakukan evaluasi.

5. Predict

Tahapan ini adalah melakukan uji coba apakah machine learning yang dibuat bekerja dengan baik. Caranya dengan melihat hasil atau prediksi yang dihasilkan. Apakah sesuai dengan *input data*. Maka tahapan ini diputuskan untuk membuat sebuah prediksi yang didapatkan dari inputan kalimat baru seperti pada kode program berikut.

Prediction

```
# Make Prediction with new data
new_data = ['I love you so much', 'I hate you so much', 'I am so happy', 'I am so sad', 'I am so angry', 'I am so bored']

# new_data = input('Masukkan teks: ')
# new_data = [new_data]

# Vectorize
new_data = tfidf.transform(new_data)

# Predict
new_pred = model.predict(new_data)

# Evaluation
new_pred = pd.Series(new_pred).map(label)
print(new_pred)
```

✓ 0.9s

```
0    Positif
1    Negatif
2    Positif
3    Negatif
4    Negatif
5    Negatif
dtype: object
```

6. Evaluasi

Pada proses evaluasi, menggunakan metric akurasi dan juga menambahkan metric lain seperti Recall, Precision, F1-Score, detail Confussion Metric, ataupun Area Under Curve (AUC).

Pada pengevaluasian menggunakan fungsi accuracy_score, classification_report, precision_score dari library sklearn.metrics untuk mendapatkan nilai-nilai matriks dari data yang digunakan.

Evaluation

```
# Import library for evaluation
from sklearn.metrics import classification_report,precision_score,recall_score,accuracy_score

print(classification_report(y_test, y_pred))

print(f'Accuracy\t: {accuracy_score(y_test, y_pred)}')
print(f'Precision\t: {precision_score(y_test, y_pred, average="macro")}')
print(f'Recall\t\t: {recall_score(y_test, y_pred, average="macro")}')
```

[125] ✓ 0.7s

```
              precision    recall  f1-score   support

     Negatif       0.98      0.17      0.29      1623
     Neutral       0.92      0.32      0.48      2752
     Positif       0.53      0.99      0.69      3591

    accuracy                           0.59      7966
   macro avg       0.81      0.49      0.49      7966
weighted avg       0.76      0.59      0.54      7966

Accuracy        : 0.5937735375345217
Precision       : 0.8113139597528978
Recall          : 0.49474918340400365
```

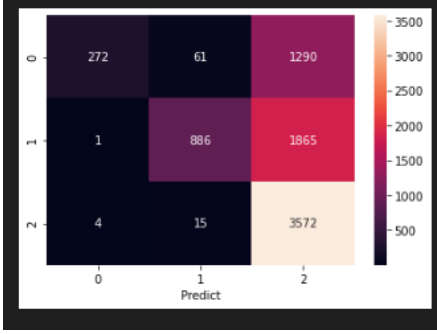
Dari perhitungan matrik diatas memunculkan nilai akurasi yang kurang baik yakni 0.59.

Kemudian membuat confusion Matrix Confusion Matrix merupakan metode evaluasi yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran dari proses klasifikasi seperti berikut

```
#  Confussion Metric
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

print('Confusion Matrix' , confusion_matrix(y_test, y_pred), sep='\n')

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d')
plt.xlabel('Predict')
plt.show()
```

✓ 0.7s

```
Confusion Matrix
[[ 272   61 1290]
 [   1  886 1865]
 [   4   15 3572]]
```

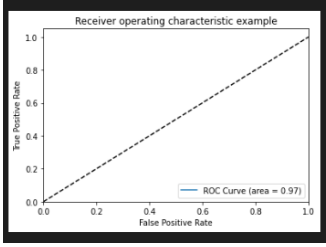


ROC Curve dibuat berdasarkan nilai telah didapatkan pada perhitungan dengan confusion matrix, yaitu antara False Positive Rate dengan True Positive Rate. Sehingga dihasilkan seperti pada gambar beriku ini.

```
# Create AUC Evaluation
from sklearn.metrics import roc_auc_score, roc_curve

# Predict Probability
y_pred_proba = model.predict_proba(X_test)

# Get AUC Score
auc = roc_auc_score(y_test, y_pred_proba, multi_class='ovr')

# Get ROC Curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba[:,1], pos_label=1)

# Plot ROC Curve
plt.plot(fpr, tpr, label='ROC Curve (area = %0.2f)' % auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic example')
plt.legend(loc="lower right")
plt.show()
```

✓ 0.6s

```
c:\Users\IQRIMA\anaconda3\lib\site-packages\sklearn\metrics\_ranking.py:1018: UndefinedMetricWarning: No positive samples in y_true, true positive value should be meaningless
  warnings.warn(
```



Kesimpulan :

Ukuran besaran precision, recall, dan accuracy biasanya diberi nilai dalam bentuk presentase antara 1 sampai 100%. Sebuah sistem akan dianggap baik jika tingkat precision, recall, dan accuracy-nya tinggi.

Sedangkan dari hasil keluaran data tersebut menunjukkan bahwa keakuratan data training dan data testing rendah. hasil presisi juga menunjukkan rendah karena kurang dari satu. Bobot Akurasi atau tingkat kedekatan antara nilai yang didapat terhadap nilai sebenarnya dari data testing dan data training juga rendah. Presisi atau kecocokan antara bagian data yang diambil dengan informasi yang dibutuhkan rendah. Kemudian Recall atau tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi juga rendah. Intinya sistem ini dapat dianggap tidak baik karena memiliki nilai keakuratan yang rendah.

Link github : https://github.com/iqrima4422/UAS_ProyekML