LAPORAN MACHINE LEARNING

UAS / Tugas Besar

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:

1. **Firdaus Bia Firmansyah** **(2041720255 / 11)**
2. **Nesa Itfirul Lail** **(2041720004 / 17)**

PROGRAM STUDI D-IV TEKNIK INFORMATIKA

JURUSAN TEKNOLOGI INFORMASI

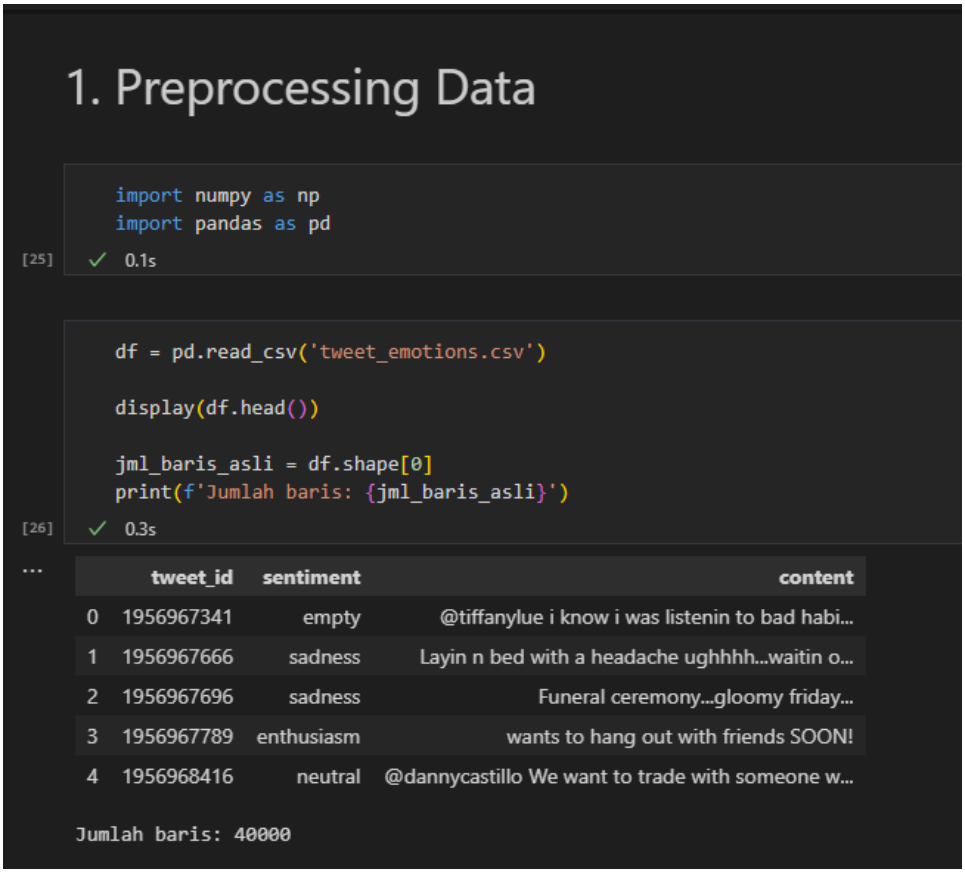
POLITEKNIK NEGERI MALANG

2022

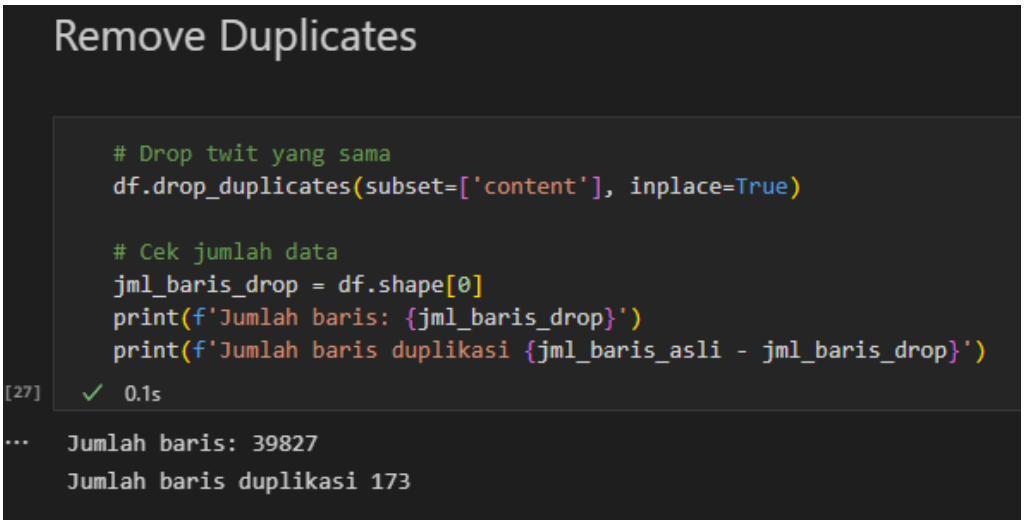
UAS akan dinilai berdasarkan 6 proses yang akan Anda lakukan, yaitu preprocessing data, clustering, labeling, classification (pembuatan model machine learning), prediction, dan evaluasi.

1. Preprocessing Data: Data tweeter yang ada dapatkan merupakan sebuah data mentah, maka beberapa hal dapat Anda lakukan (namun tidak terbatas pada) yaitu

   Membuka file dataset “tweet_emotion.csv” dan menampilkan jumlah baris keseluruhan pada dataset tersebut



   Menghapus baris-baris yang terdapat duplikasi atau kesamaan pada dataset



Menghapus (@) dan URL pada dataset

Remove Mention (@) and URL

```
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)
```

28] ✓ 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...
...	...	...	...	...
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

- Case Folding
  Mengubah isi dataset menjadi lowercase semuanya

Case Folding

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

df.tail(10)
```

✓ 0.1s

	tweet_id	sentiment	content	content_clean
39990	1753918829	neutral	@shonali I think the lesson of the day is not ...	i think the lesson of the day is not to have ...
39991	1753918846	neutral	@lovelylisaj can you give me the link for the ...	can you give me the link for the kimba diarie...
39992	1753918881	neutral	@jasimmo Ooo showing of your French skills!! I...	ooo showing of your french skills lol thing...
39993	1753918892	neutral	@sendsome2me haha, yeah. Twitter has many uses...	haha yeah twitter has many uses for me it ...
39994	1753918900	happiness	Succesfully following Tayla!!	succesfully following tayla
39995	1753918954	neutral	@JohnLloydTaylor	johnlloydtaylor
39996	1753919001	love	Happy Mothers Day All my love	happy mothers day all my love
39997	1753919005	love	Happy Mother's Day to all the mommies out ther...	happy mother s day to all the mommies out ther...
39998	1753919043	happiness	@niariley WASSUP BEAUTIFUL!!! FOLLOW ME!! PEE...	wassup beautiful follow me peep out my ...
39999	1753919049	love	@mopedronin bullet train from tokyo the gf ...	bullet train from tokyo the gf and i have ...

- Tokenizing
  Memisahkan kalimat menjadi kata

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()
```

✓ 5.7s

	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,...

- Filtering
  Mengambil kata-kata inti atau penting menggunakan fungsi stopwords

Filtering

```
df_stem = df.copy()
from nltk.corpus import stopwords

# ---------------------- get stopword from NLTK stopword -----------------------------
# get stopword english
list_stopwords = stopwords.words('english')
# read txt stopword using pandas
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['content_filtering'] = df_stem['content_clean'].apply(stopwords_removal)

df_stem.head()
```

[33] ✓ 0.6s

	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...

- Stemming
  Memperkecil jumlah indeks yang berbeda dari suatu data sehingga kata yang memiliki suffix atau prefix akan Kembali ke bentuk dasarnya

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_clean'].apply(lambda x: stemming(x))

df_stem.head()
```

✓ 2.9s

	tweet_id	sentiment	content	content_clean	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, , k, n, o, w, , , i, , w, a, s, , ...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[l, a, y, i, n, , n, , b, e, d, , w, i, t, ...
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[f, u, n, e, r, a, l, , c, e, r, e, m, o, n, ...
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[w, a, n, t, s, , t, o, , h, a, n, g, , o, ...
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[, w, e, , w, a, n, t, , t, o, , t, r, a, ...

2. Clustering: Pengelompokan data ke dalam beberapa kategori atau cluster, yaitu komentar positif, netral, dan negatif.

   Menampilkan kelompok data sesuai kolom sentiment

2. Clustering

```
print(df['sentiment'].value_counts())
print('\n')
```

[36] ✓ 0.2s

```
neutral       8598
worry         8437
happiness     5184
sadness       5154
love          3785
surprise      2181
fun           1775
relief        1522
hate          1322
empty          822
enthusiasm     758
boredom        179
anger          110
Name: sentiment, dtype: int64
```

Mengimport textblob package untuk menghitung polaritas pada kolom content_clean

```
# 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)
```

38] ✓ 8.9s

Mengelompokkan hasil polaritas menjadi 3 kategori yaitu positif, negatif dan netral

```
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()
```

[39] ✓ 0.1s Python

	tweet_id	sentiment	content	content_clean	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, , k, n, o, w, , , i, , w, a, s, , ...	-0.35	Negatif
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...	layin n bed with a headache ughhhh waitin o...	[l, a, y, i, n, , n, , b, e, d, , w, i, t, ...	0.00	Neutral
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[f, u, n, e, r, a, l, , c, e, r, e, m, o, n, ...	0.00	Neutral
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[w, a, n, t, s, , t, o, , h, a, n, g, , o, ...	0.20	Positif
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[, w, e, , w, a, n, t, , t, o, , t, r, a, ...	0.00	Neutral

3. Labeling: Label merupakan hasil dari pengelompokan example melalui clustering. Sebagai contoh, machine learning yang berfungsi menyaring email spam, melabeli setiap example dengan ‘spam’ atau ‘not spam’.

Memberi label berdasarkan kategori hasil hitungan polaritas

3. 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()
```

✓ 0.1s Python

	tweet_id	sentiment	content	content_clean	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, , k, n, o, w, , , i, , w, a, s, , ...	-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...	[l, a, y, i, n, , n, , b, e, d, , w, i, t, ...	0.00	Neutral	0
2	1956967696	sadness	Funeral ceremony...gloomy friday...	funeral ceremony gloomy friday	[f, u, n, e, r, a, l, , c, e, r, e, m, o, n, ...	0.00	Neutral	0
3	1956967789	enthusiasm	wants to hang out with friends SOON!	wants to hang out with friends soon	[w, a, n, t, s, , t, o, , h, a, n, g, , o, ...	0.20	Positif	1
4	1956968416	neutral	@dannycastillo We want to trade with someone w...	we want to trade with someone who has houston...	[, w, e, , w, a, n, t, , t, o, , t, r, a, ...	0.00	Neutral	0

Menampilkan jumlah data masing-masing kategori dan label

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

[41] ✓ 0.1s

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

4. Classification: Anda dibebaskan dalam memilih algoritma klasifikasi. Anda dapat menggunakan algoritma yang telah diajarkan didalam kelas atau yang lain, namun dengan catatan. Berdasarkan asas akuntabilitas pada pengembangan model machine learning, Anda harus dapat menjelaskan bagaimana model Anda dapat menghasilkan nilai tertentu.

Menggunakan algoritma MultinomialNB dari Naïve Bayes untuk mengklasifikasikan data

4. Classification

Naive Bayes

```
# 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)
```

✓ 1.3s

5. Predict: Uji coba apakah machine learning yang Anda buat bekerja dengan baik. Caranya dengan melihat hasil atau prediksi yang dihasilkan. Apakah sesuai dengan input data.

Melakukan prediksi dengan mencoba memasukkan data baru ke dalam dataset dan mengkategorikan sesuai dengan ketentuan sebelumnya

6. Evaluasi: Pada proses evaluasi, minimal Anda harus menggunakan metric akurasi. Akan tetapi Anda juga dapat menambahkan metric lain seperti Recall, Precision, F1- Score, detail Confussion Metric, ataupun Area Under Curve (AUC)

Menggunakan fungsi Accuracy, Precision, dan Recall untuk menentukan hasil prediksi sebelumnya

6. Evaluasi

```
# 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")}')
```

[56] ✓ 0.4s

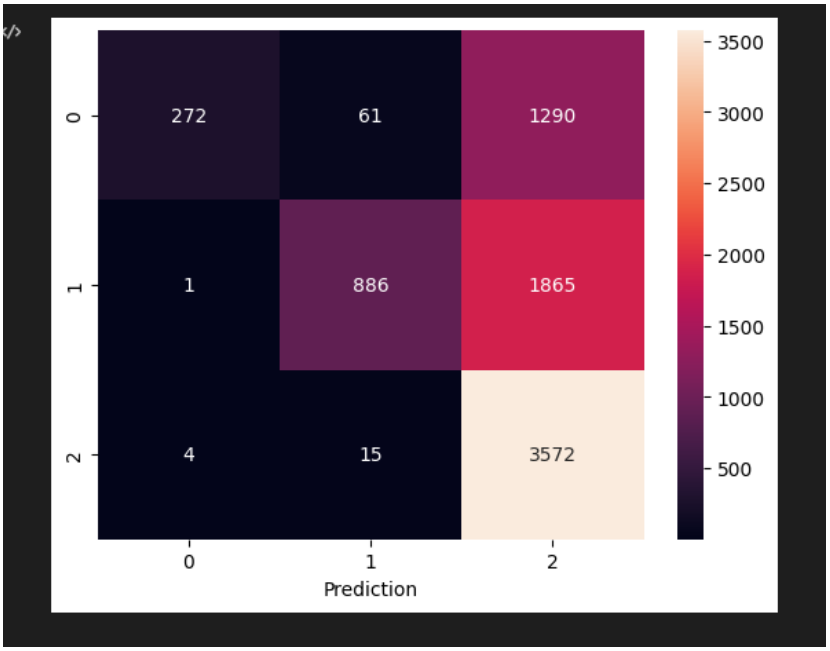
```
              precision    recall  f1-score   support

      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
```

Menggunakan confusion matrix untuk menentukan hasil prediksi dan mengimplementasikan kedalam grafik

```
#  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('Prediction')
plt.show()
```

[59] ✓ 0.4s

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

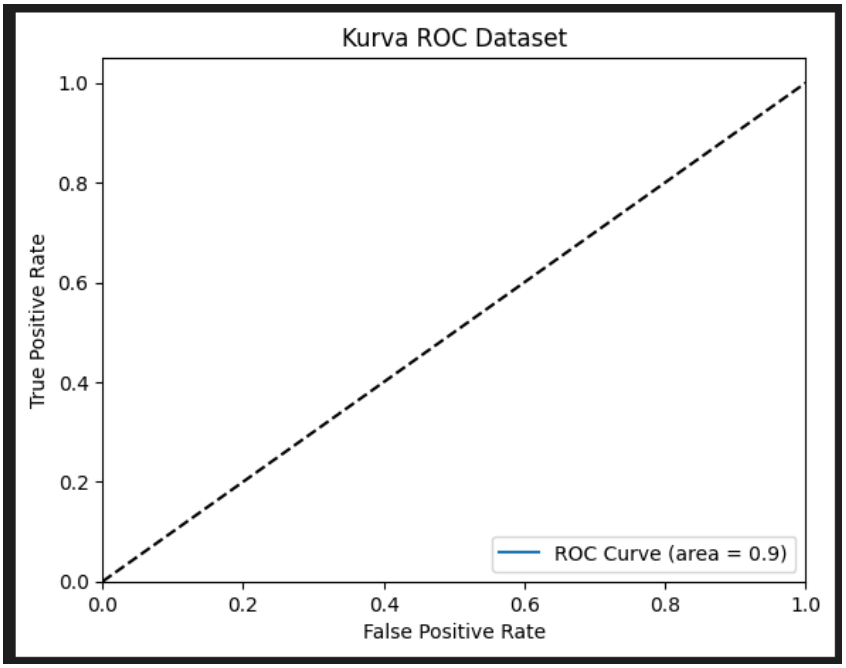


Menggunakan kurva ROC untuk menentukan hasil prediksi dan mengimplementasikan kedalam grafik

```
# 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 = %0.1f)' % 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('Kurva ROC Dataset')
plt.legend(loc="lower right")
plt.show()
```

[62] ✓ 0.3s

```
c:\Users\HP\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\metrics\_ranking.py:1018: UndefinedMetricWarning: No positive samples in y_true, true positive value should be meaningless
  warnings.warn(
```



Link repositori :

https://github.com/nesaitfirullail12/UAS_ML_TI3F