PORTFOLIO

OPTIMASI NAIVE BAYES MENGGUNAKAN ALGORITMA GENETIKA PADA KLASIFIKASI KOMENTAR CYBERBULLYING PADA MEDIA SOSIAL X

USING PHYTON

BY SYIFA TAHIR

Project ini adalah Tugas Akhir saya yang berfokus pada Data Analyst dan mendapatkan nilai A ketika sidang Tugas Akhir

CRAWL DATA

Twitter Auth Token

```
#@title Twitter Auth Token

twitter_auth_token = '06fd705198fcba5eec5ed20ef39ebcdc61bbe7ad'
```

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo tee /etc/apt/sources.list.d/nodesource.list

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

Show hidden output

```
[ ] # Crawl Data

    filename = 'bodoh.csv'
    search_keyword = 'bodoh lang:id'
    limit = 100

    !npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

```
Tweet Harvest [v2.6.1]
```

```
[ ] # Cek jumlah data yang didapatkan

    num_tweets = len(df)
    print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")
```

```
Jumlah tweet dalam dataframe adalah 108.
```

Before
Pre - Processing

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count
0	1.800000e+18	Mon Jun 10 10:26:47 +0000 2024	2361	@kegblgnunfaedh KFC Malaysia Tutup 100 Gerai I...	1.800000e+18	NaN	kegblgnunfaedh	in	Jakarta	23	33	281
1	1.800000e+18	Mon Jun 10 08:58:26 +0000 2024	8936	@kegblgnunfaedh Si cireng ini licik dia videoi...	1.800000e+18	NaN	kegblgnunfaedh	in	Tangerang, Indonesia	15	42	276
2	1.800000e+18	Mon Jun 10 08:12:58 +0000 2024	14763	@kegblgnunfaedh matanya mirip bisa liat ke 2 a...	1.800000e+18	https://pbs.twimg.com/media/GPsqBpsbQAEWdlQ.jpg	kegblgnunfaedh	in	NaN	152	272	1117
3	1.800000e+18	Mon Jun 10 12:25:36 +0000 2024	1879	@kegblgnunfaedh Sudah terwakilin belum semua k...	1.800000e+18	https://pbs.twimg.com/amplify_video_thumb/1800...	kegblgnunfaedh	in	Verification X	2	6	160
4	1.800000e+18	Mon Jun 10 12:48:18 +0000 2024	6263	@kegblgnunfaedh @Andreakacamata mari kita bela...	1.800000e+18	https://pbs.twimg.com/ext_tw_video_thumb/18001...	kegblgnunfaedh	in	NaN	37	87	807
...	...	...	...	...	...	...	...	...	...	...	...	...
1171	1.810000e+18	Tue Jul 02 04:06:50 +0000 2024	0	ngomong want want want doang tapi ga gerak ga ...	1.810000e+18	https://pbs.twimg.com/media/GRdEq8daYAA7JAw.jpg	NaN	in	sunoo's heart.	0	0	0
1172	1.810000e+18	Tue Jul 02 04:02:48 +0000 2024	0	Sumpah demi allah illahi robbi lo tuh agensi p...	1.810000e+18	NaN	NaN	in	Under Donghae's Spell	0	1	0
1173	1.810000e+18	Tue Jul 02 03:38:33 +0000 2024	0	@ooheroin apaan si goblok minimal ngetik yg be...	1.810000e+18	NaN	ooheroin	in	mt after dm	0	0	0
1174	1.810000e+18	Tue Jul 02 03:35:17 +0000 2024	1	HEH GOBLOK LO GA USAH ASBUN YAH	1.810000e+18	NaN	NaN	in	ship nr & nomin x gg dni !!!	0	2	0
1175	1.810000e+18	Tue Jul 02 03:35:02 +0000 2024	0	@utdfocusid @bliblidotcom Gimana yah bingung s...	1.810000e+18	NaN	utdfocusid	in	Bandung, Jawa Barat	0	0	0

After
Pre - Processing

	full_text	username	case_folded	cleaning	tokenizing	Filtering/stopword removal
0	@kegblgnunfaedh KFC Malaysia Tutup 100 Gerai I...	okezonenews	@kegblgnunfaedh kfc malaysia tutup 100 gerai i...	kfc malaysia tutup gerai imbas aksi boikot pr...	[kfc, malaysia, tutup, gerai, imbas, aksi, boi...	[kfc, malaysia, tutup, gerai, imbas, aksi, boi...
1	@kegblgnunfaedh Si cireng ini licik dia videoi...	wemutt_	@kegblgnunfaedh si cireng ini licik dia videoi...	si cireng ini licik dia videoin dia upload tp ...	[si, cireng, ini, licik, dia, videoin, dia, up...	[si, cireng, licik, videoin, upload, tp, ga, n...
2	@kegblgnunfaedh matanya mirip bisa liat ke 2 a...	odesaa_	@kegblgnunfaedh matanya mirip bisa liat ke 2 a...	matanya mirip bisa liat ke arah sekaligus	[matanya, mirip, bisa, liat, ke, arah, sekaligus]	[matanya, liat, arah]
3	@kegblgnunfaedh Sudah terwakilin belum semua k...	BaekBoy__	@kegblgnunfaedh sudah terwakilin belum semua k...	sudah terwakilin belum semua kemarahan kalian	[sudah, terwakilin, belum, semua, kemarahan, k...	[terwakilin, kemarahan]
4	@kegblgnunfaedh @Andreakacamata mari kita bela...	masgah_	@kegblgnunfaedh @andreakacamata mari kita bela...	mari kita belajar bersama	[mari, kita, belajar, bersama]	[mari, belajar]
...	...	...	...	...	...	...
1171	ngomong want want want doang tapi ga gerak ga ...	sunoobites	ngomong want want want doang tapi ga gerak ga ...	ngomong want want want doang tapi ga gerak ga ...	[ngomong, want, want, want, doang, tapi, ga, g...	[ngomong, want, want, want, doang, ga, gerak, ...
1172	Sumpah demi allah illahi robbi lo tuh agensi p...	MataHaeRi	sumpah demi allah illahi robbi lo tuh agensi p...	sumpah demi allah illahi robbi lo tuh agensi p...	[sumpah, demi, allah, illahi, robbi, lo, tuh, ...	[sumpah, allah, illahi, robbi, lo, tuh, agensi...
1173	@ooheroin apaan si goblok minimal ngetik yg be...	skrifsi	@ooheroin apaan si goblok minimal ngetik yg be...	apaan si goblok minimal ngetik yg bener caper ...	[apaan, si, goblok, minimal, ngetik, yg, bener...	[si, goblok, minimal, ngetik, yg, bener, caper...
1174	HEH GOBLOK LO GA USAH ASBUN YAH	JJLuv_Ra	heh goblok lo ga usah asbun yah	heh goblok lo ga usah asbun yah	[heh, goblok, lo, ga, usah, asbun, yah]	[heh, goblok, lo, ga, asbun, yah]
1175	@utdfocusid @bliblidotcom Gimana yah bingung s...	gakduluu_	@utdfocusid @bliblidotcom gimana yah bingung s...	gimana yah bingung sama fakta ini kadang aneh ...	[gimana, yah, bingung, sama, fakta, ini, kadan...	[gimana, yah, bingung, fakta, kadang, aneh, gw...

Stemming Data

Pre-Processing 6 - STEAMING DATA

```
[ ] !pip install Sastrawi

    from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
    from nltk.stem import PorterStemmer
    from nltk.stem.snowball import SnowballStemmer
    nltk.download('punkt')
```

```
Collecting Sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 kB 4.7 MB/s eta 0:00:00
Installing collected packages: Sastrawi
Successfully installed Sastrawi-1.0.1
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
[ ] factory = StemmerFactory()
    stemmer = factory.create_stemmer()

    def stem_text(text):
      tokens = nltk.word_tokenize(str(text))
      stemmed_words = [stemmer.stem(word) for word in tokens]
      stemmed_text = ' '.join(stemmed_words)
      return stemmed_text

    df['stemming_data'] = df['Filtering/stopword removal'].apply(stem_text)

    print(df.head(1176))
```

```
                                          full_text     username  \
0     @kegblgnunfaedh KFC Malaysia Tutup 100 Gerai I...  okezonenews
1     @kegblgnunfaedh Si cireng ini licik dia videoi...      wemutt_
2     @kegblgnunfaedh matanya mirip bisa liat ke 2 a...      odesaa_
3     @kegblgnunfaedh Sudah terwakilin belum semua k...    BaekBoy__
4     @kegblgnunfaedh @Andreakacamata mari kita bela...      masgah_
...                                                 ...          ...
1171  ngomong want want want doang tapi ga gerak ga ...    sunoobites
1172  Sumpah demi allah illahi robbi lo tuh agensi p...    MataHaeRi
1173  @ooheroin apaan si goblok minimal ngetik yg be...       skrifsi
1174                  HEH GOBLOK LO GA USAH ASBUN YAH      JJLuv_Ra
1175  @utdfocusid @bliblidotcom Gimana yah bingung s...     gakduluu_

                                       case_folded  \
0     @kegblgnunfaedh kfc malaysia tutup 100 gerai i...
1     @kegblgnunfaedh si cireng ini licik dia videoi...
2     @kegblgnunfaedh matanya mirip bisa liat ke 2 a...
```

Labelling

```
from sklearn.preprocessing import LabelEncoder

# Fungsi untuk mendeteksi kata-kata bullying
def deteksi_bullying(teks):
    bullying_keywords = ['bodoh', 'jelek', 'goblok', 'tolol', 'sialan', 'anjing', 'bangsat', 'mati', 'k
    teks = teks.lower()
    if any(word in teks for word in bullying_keywords):
        return 1
    return 0

# Terapkan fungsi untuk menambahkan kolom label setelah stemming
df['label'] = df['stemming_data'].apply(deteksi_bullying)

# Ubah nilai 1 menjadi 'bullying' dan 0 menjadi 'nonbullying'
df['label'] = df['label'].map({1: 'bullying', 0: 'nonbullying'})

# Hitung jumlah 'nonbullying' dan 'bullying'
jumlah_komentar = df['label'].value_counts()
print(jumlah_komentar)
```

```
label
bullying       709
nonbullying    467
Name: count, dtype: int64
```

```
[ ] # Inisialisasi label encoder
    label_encoder = LabelEncoder()

    # Fit dan transform label
    df['encoded_label'] = df['label'].map({'bullying': 1, 'nonbullying': 0})

    # Tampilkan beberapa baris pertama setelah labeling
    print(df.head(1176))
```

```python
# Fungsi untuk dioptimalkan (fitness function)
def fitness_function(params):
    alpha = params[0]
    model = MultinomialNB(alpha=alpha)
    scores = cross_val_score(model, X_train_tfidf, y_train, cv=StratifiedKFold(n_splits=10), scoring='accuracy')
    return -scores.mean()  # Dikonversi menjadi masalah minimisasi

# Batasan untuk parameter alpha
varbound = np.array([[0.1, 1.0]])

# Membuat objek geneticalgorithm
model = ga(function=fitness_function, dimension=1, variable_type='real', variable_boundaries=varbound)

# Menjalankan algoritma genetika
model.run()
```

```python
print(f"Parameter Optimal: alpha = {optimal_alpha}")
print(f"Akurasi Optimal: {accuracy_optimal}")
print(f"Precision Optimal: {precision_optimal}")
print(f"Recall Optimal: {recall_optimal}")
print(f"F1 Score Optimal: {f1_optimal}")

# Confusion Matrix
cm_optimal = confusion_matrix(y_test, y_pred_optimal)
disp = ConfusionMatrixDisplay(confusion_matrix=cm_optimal, display_labels=['nonbullying', 'bullying'])
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix (Genetic + MultinomialNB)')
plt.show()

# Menampilkan hasil dalam bentuk diagram bar
metrics = ['Accuracy', 'Precision', 'Recall', 'F1 Score']
scores = [accuracy_optimal, precision_optimal, recall_optimal, f1_optimal]

plt.figure(figsize=(8, 4))
bars = plt.bar(metrics, scores)

# Menambahkan nilai pada tiap bar
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2 - 0.1, yval + 0.01, round(yval, 2))

plt.ylim(0, 100)
plt.xlabel('Metrics')
plt.ylabel('Nilai Rata-rata (%)')
plt.title('')
plt.show()
```

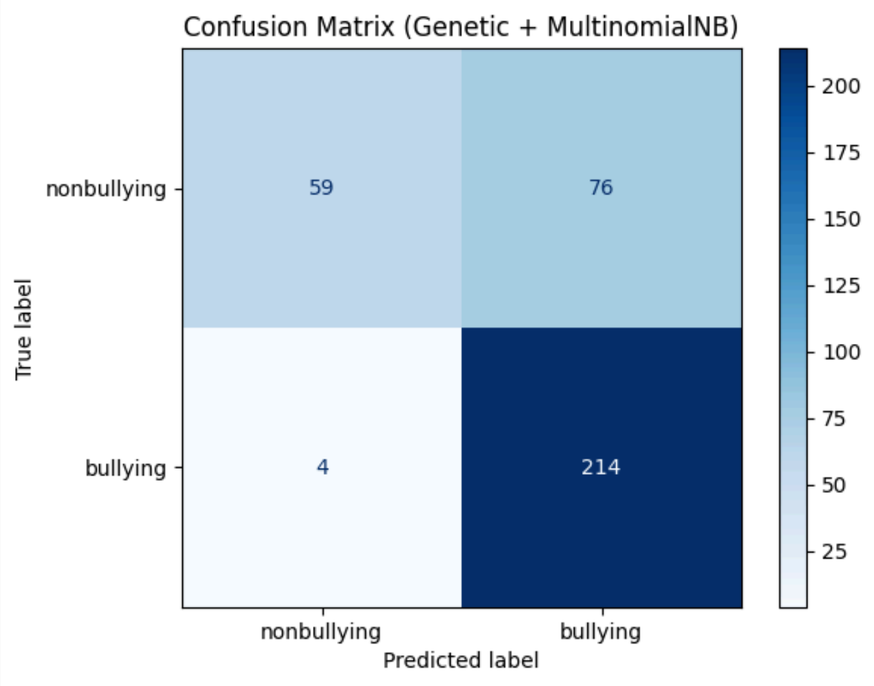
Pengujian Algoritma Genetika + Naive Bayes

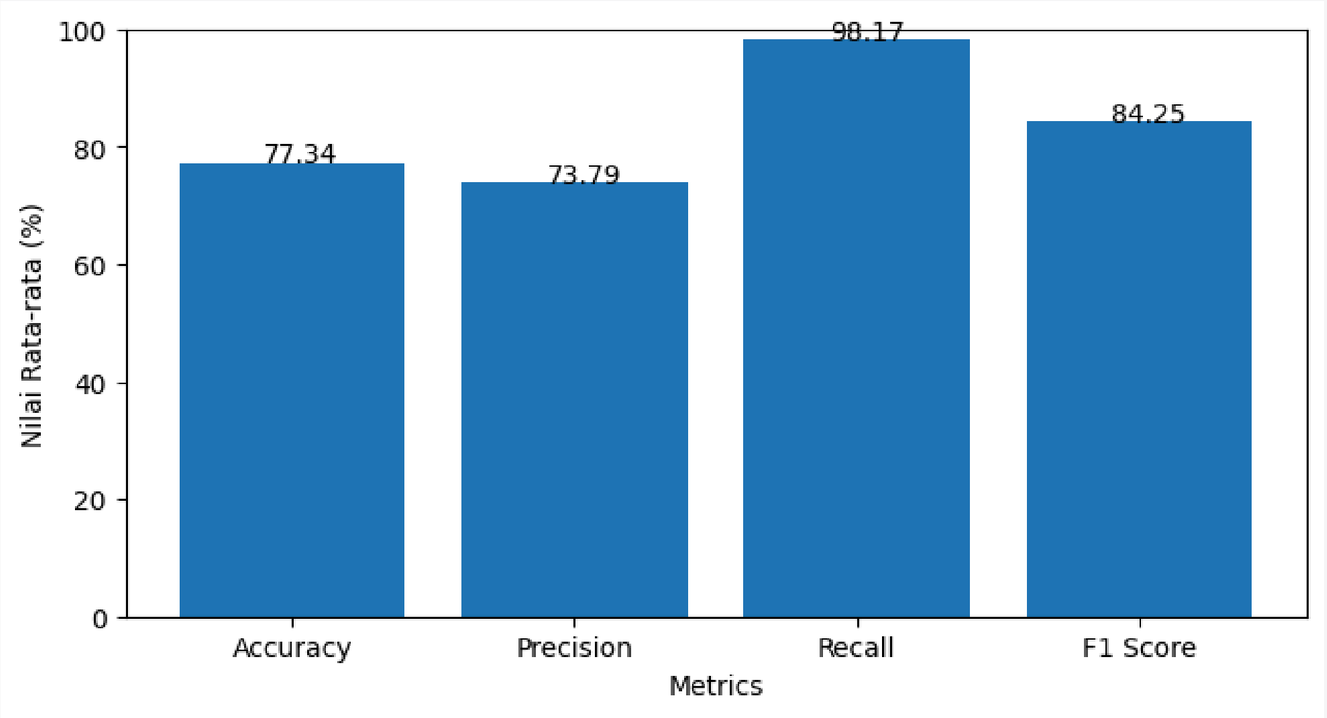
```
[ ]  # Latih model Naive Bayes tanpa optimasi parameter
     model_nb_multinomial = MultinomialNB()
     model_nb_multinomial.fit(X_train_tfidf, y_train)

     # Menguji model pada data test
     y_pred_nb_multinomial = model_nb_multinomial.predict(X_test_tfidf)

     # Menghitung metrik evaluasi
     accuracy_nb_multinomial = accuracy_score(y_test, y_pred_nb_multinomial) * 100
     precision_nb_multinomial = precision_score(y_test, y_pred_nb_multinomial, average='weighted') * 100
     recall_nb_multinomial = recall_score(y_test, y_pred_nb_multinomial, average='weighted') * 100
     f1_nb_multinomial = f1_score(y_test, y_pred_nb_multinomial, average='weighted') * 100

     print(f"Akurasi (Naive Bayes Multinomial saja): {accuracy_nb_multinomial}")
     print(f"Precision (Naive Bayes Multinomial saja): {precision_nb_multinomial}")
     print(f"Recall (Naive Bayes Multinomial saja): {recall_nb_multinomial}")
     print(f"F1 Score (Naive Bayes Multinomial saja): {f1_nb_multinomial}")

     # Confusion Matrix
     cm_nb_multinomial = confusion_matrix(y_test, y_pred_nb_multinomial)
     disp = ConfusionMatrixDisplay(confusion_matrix=cm_nb_multinomial, display_labels=['nonbullying', 'bullying'])
     disp.plot(cmap=plt.cm.Blues)
     plt.title('Confusion Matrix (Naive Bayes Multinomial saja)')
     plt.show()
```

```
# Menampilkan hasil dalam bentuk diagram bar
metrics = ['Accuracy', 'Precision', 'Recall', 'F1 Score']
scores = [accuracy_nb_multinomial, precision_nb_multinomial, recall_nb_multinomial, f1_nb_multinomial]

plt.figure(figsize=(8, 4))
bars = plt.bar(metrics, scores)

# Menambahkan nilai pada tiap bar
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2 - 0.1, yval + 0.01, round(yval, 2))

plt.ylim(0, 100)
plt.xlabel('Metrics')
plt.ylabel('Nilai Rata-rata (%)')
plt.title('')
plt.show()
```

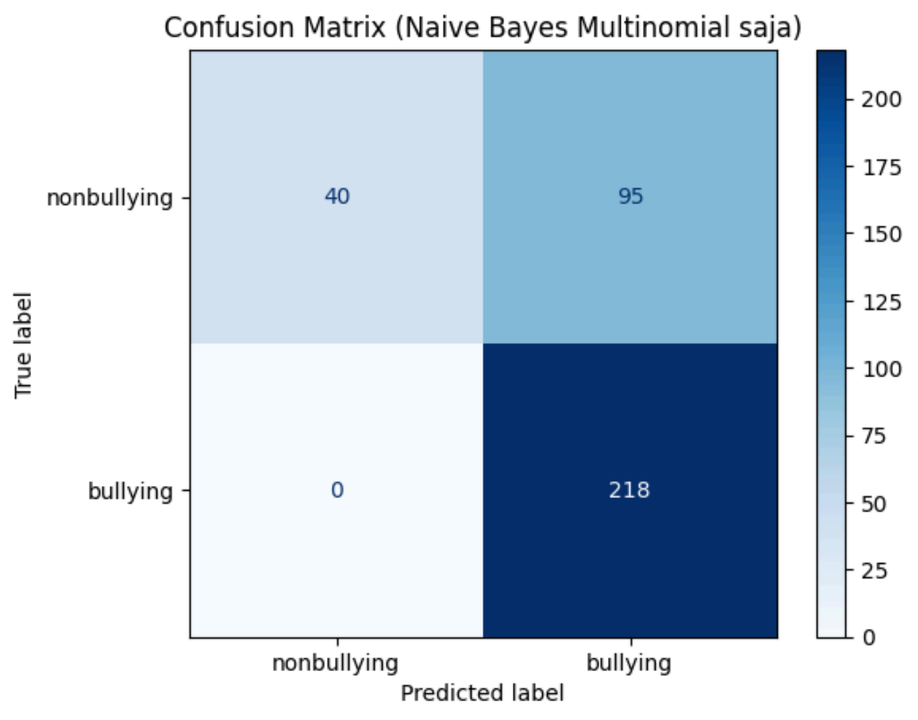
Pengujian Naive Bayes Tanpa Optimasi (MultinomialNB)

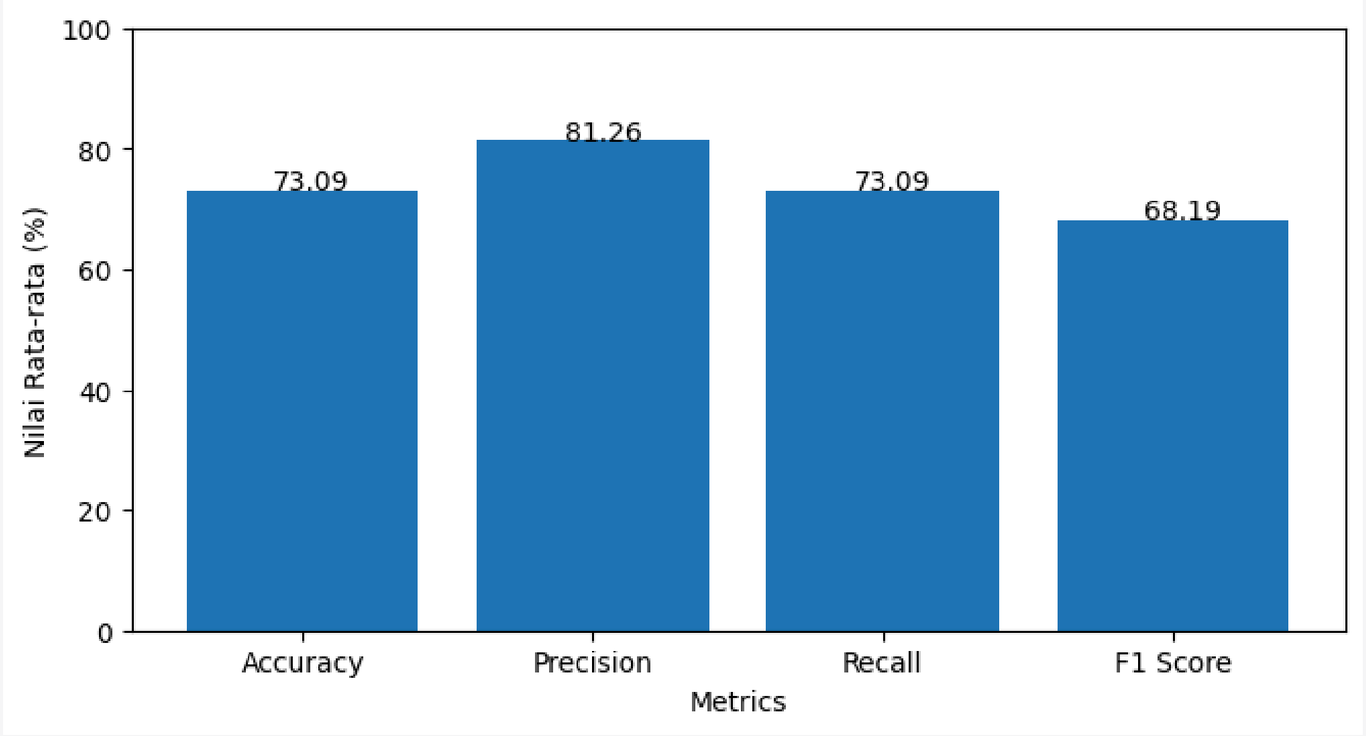
Hasil Pengujian Algoritma Genetika + Naive Bayes





Hasil Pengujian Naive Bayes Tanpa Optimasi (MultinomialNB)





Perbandingan hasil antara yang di Optimasi dan tidak

```
[ ]  # Tampilkan perbandingan hasil untuk MultinomialNB dengan optimasi
     print(f'\nPerbandingan Hasil MultinomialNB dengan Optimasi:')
     print(f'Akurasi (Genetic + MultinomialNB): {accuracy_optimal}')
     print(f'Precision (Genetic + MultinomialNB): {precision_optimal}')
     print(f'Recall (Genetic + MultinomialNB): {recall_optimal}')
     print(f'F1 Score (Genetic + MultinomialNB): {f1_optimal}')

     # Tampilkan perbandingan hasil untuk MultinomialNB tanpa optimasi
     print(f'\nPerbandingan Hasil MultinomialNB Tanpa Optimasi:')
     print(f'Akurasi (MultinomialNB saja): {accuracy_nb_multinomial}')
     print(f'Precision (MultinomialNB saja): {precision_nb_multinomial}')
     print(f'Recall (MultinomialNB saja): {recall_nb_multinomial}')
     print(f'F1 Score (MultinomialNB saja): {f1_nb_multinomial}')
```

```
Perbandingan Hasil MultinomialNB dengan Optimasi:
Akurasi (Genetic + MultinomialNB): 77.33711048158641
Precision (Genetic + MultinomialNB): 73.79310344827587
Recall (Genetic + MultinomialNB): 98.1651376146789
F1 Score (Genetic + MultinomialNB): 84.25196850393701

Perbandingan Hasil MultinomialNB Tanpa Optimasi:
Akurasi (MultinomialNB saja): 73.08781869688386
Precision (MultinomialNB saja): 81.25605263872423
Recall (MultinomialNB saja): 73.08781869688386
F1 Score (MultinomialNB saja): 68.19048228756782
```