

Technical Report

CASE 1

DATASET OF DIGITAL LITERACY OF UNIVERSITY
STUDENTS IN INDONESIA

DISUSUN :

RIDHO PRATAMA
WIDIANTORO

105222011



Latar Belakang

DI ERA DIGITAL SAAT INI, PENGGUNAAN INTERNET OLEH MAHASISWA SEMAKIN TINGGI, BAIK UNTUK KEPERLUAN BELAJAR MAUPUN KEHIDUPAN SEHARI-HARI. SAYANGNYA, TINGGINYA FREKUENSI PENGGUNAAN BELUM SELALU DIBARENGI DENGAN KEMAMPUAN UNTUK MEMILAH INFORMASI SECARA KRITIS, MENJAGA KEAMANAN DATA PRIBADI, MAUPUN BERINTERAKSI SECARA BIJAK DI RUANG DIGITAL. HAL INI MEMBUAT MAHASISWA RENTAN TERHADAP PENYEBARAN HOAKS, PELANGGARAN PRIVASI, DAN PERILAKU DARING YANG TIDAK ETIS. UNTUK ITU, KETERAMPILAN LITERASI DIGITAL MENJADI SANGAT PENTING DIMILIKI OLEH MAHASISWA. LITERASI DIGITAL MENCAKUP KEMAMPUAN MEMAHAMI TEKNOLOGI, MENJAGA KEAMANAN INFORMASI PRIBADI DAN PERANGKAT, BERPIKIR KRITIS TERHADAP INFORMASI, SERTA BERKOMUNIKASI SECARA EFEKTIF DI DUNIA MAYA.



Tujuan Prediksi

- Menganalisis sejauh mana keterampilan literasi digital mahasiswa Indonesia berdasarkan enam aspek utama: keterampilan teknologi, keamanan pribadi, berpikir kritis, keamanan perangkat, keterampilan informasi, dan komunikasi.
- Memprediksi jenis kelamin mahasiswa (laki-laki atau perempuan) berdasarkan profil keterampilan literasi digital mereka, dengan menggunakan pendekatan klasifikasi dalam pembelajaran mesin.



Exploratory Data Analysis

Insight 1 :

- Gender Mahasiswa
- rata rata Subskala

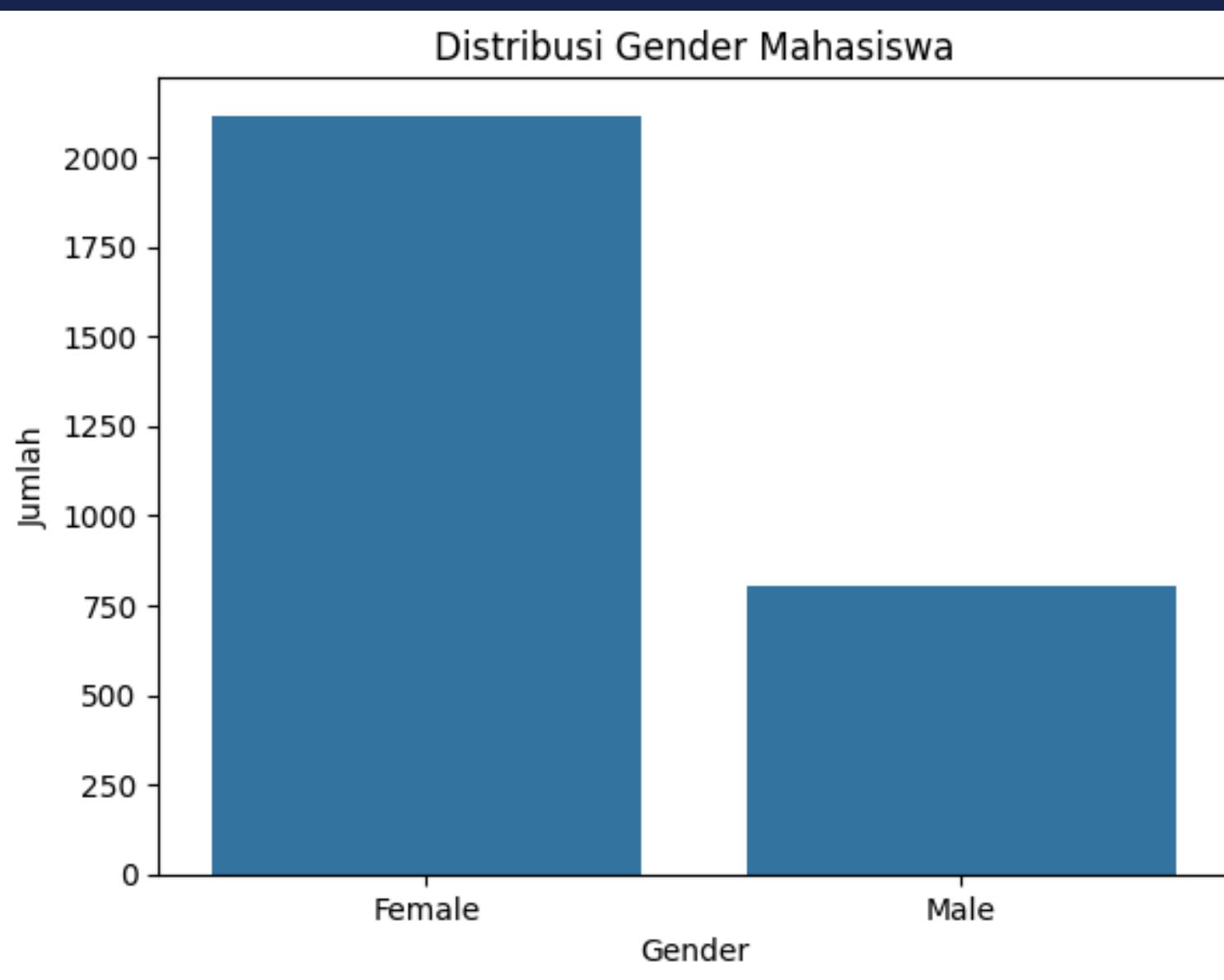


CODE Gender Mahasiswa

```
sns.countplot(data=df, x='Gender')
plt.title('Distribusi Gender Mahasiswa')
plt.xlabel('Gender')
plt.ylabel('Jumlah')
plt.show()
```



Visualisasi Gender Mahasiswa



EDA INSIGHT 1

BERDASARKAN HASIL ANALISIS DESKRIPTIF TERHADAP DATA LITERASI DIGITAL, DITEMUKAN BAHWA MAHASISWA PEREMPUAN MEMILIKI RATA-RATA SKOR YANG LEBIH TINGGI DIBANDINGKAN MAHASISWA LAKI-LAKI DI HAMPIR SEMUA ASPEK YANG DIUKUR. HAL INI MENUNJUKKAN BAHWA PEREMPUAN CENDERUNG MEMILIKI TINGKAT LITERASI DIGITAL YANG LEBIH BAIK DALAM KONTEKS DATAINI. SELAIN ITU, DARI KEENAM SUBSKALA YANG DIGUNAKAN UNTUK MENGUKUR LITERASI DIGITAL, ASPEK KEMAMPUAN KOMUNIKASI DIGITAL MENEMPATI POSISI TERTINGGI DALAM SKOR RATA-RATA. TEMUAN INI MENUNJUKKAN BAHWA SEBAGIAN BESAR MAHASISWA MERASA CUKUP PERCAYA DIRI DALAM BERKOMUNIKASI MELALUI MEDIA DIGITAL SEPERTI PESAN INSTAN, EMAIL, ATAU MEDIA SOSIAL.



EDA Insight 2 : Scatter plot Heatmap subskala literasi digital



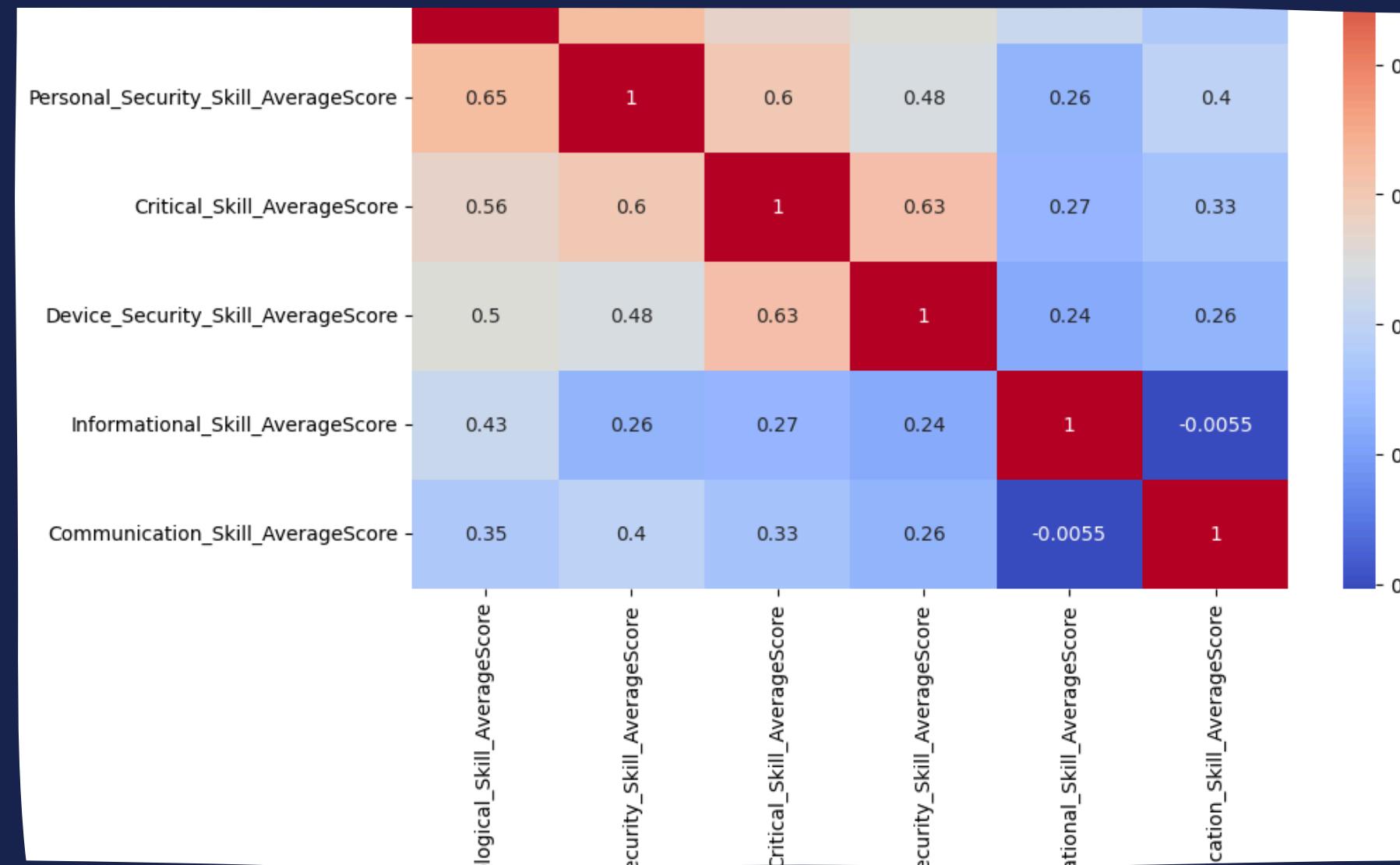
CODE

```
plt.figure(figsize=(10, 6))
sns.heatmap(X.corr(), annot=True, cmap='coolwarm')
plt.title('Korelasi antar Subskala Literasi Digital')
plt.show()
```

✓ 0.4s



Visualisasi Heatmap



EDA INSIGHT 2

INSIGHT KEDUA, HASIL ANALISIS MENUNJUKKAN BAHWA KETERKAITAN ANTARA SUBSKALA DAPAT DILIHAT DENGAN JELAS MELALUI SCATTER PLOT. BEBERAPA FITUR YANG PALING BERPENGARUH DALAM MODEL RANDOM FOREST ANTARA LAIN COMMUNICATION SKILL, PERSONAL SECURITY SKILL, DAN CRITICAL SKILL. ARTINYA, KEMAMPUAN KOMUNIKASI, KETERAMPILAN DALAM MENJAGA KEAMANAN PRIBADI, DAN KETERAMPILAN BERPIKIR KRITIS MEMILIKI PERAN YANG SIGNIFIKAN DALAM MEMPENGARUHI HASIL YANG DIANALISIS.



DATA PREPROCESSING



CODE DATA PREPROCESSING

```
kolom_skor = [
    'Technological_Skill_AverageScore',
    'Personal_Security_Skill_AverageScore',
    'Critical_Skill_AverageScore',
    'Device_Security_Skill_AverageScore',
    'Informational_Skill_AverageScore',
    'Communication_Skill_AverageScore'
]

for kolom in kolom_skor:
    df[kolom] = df[kolom].astype(str).str.replace(',', '.').astype(float)
] 0.0s

le = LabelEncoder()
df['Label_Gender'] = le.fit_transform(df['Gender'])
] 0.0s

X = df[kolom_skor]
y = df['Label_Gender']
] 0.0s
```



DATA PREPROCESSING

PADA TAHAP PREPROCESSING, BEBERAPA LANGKAH DILAKUKAN UNTUK MEMERSIAPKAN DATA AGAR SIAP DIGUNAKAN DALAM ANALISIS. PERTAMA, KOMA DIGANTI DENGAN TITIK PADA NILAI-NILAI NUMERIK UNTUK MEMASTIKAN KONSISTENSI FORMAT. SELANJUTNYA, NILAI-NILAI TERSEBUT DIKONVERSI MENJADI TIPE DATA FLOAT AGAR DAPAT DIPROSES LEBIH LANJUT. UNTUK VARIABEL GENDER, DILAKUKAN ENCODING AGAR DATA TERSEBUT DAPAT DITERIMA OLEH MODEL ANALISIS. TERAKHIR, FITUR (X) DAN LABEL (Y) DITETAPKAN, DENGAN FITUR BERISI VARIABEL-VARIABEL YANG DIGUNAKAN UNTUK PREDIKSI DAN LABEL BERISI HASIL ATAU KATEGORI YANG INGIN DIPREDIKSI, SESUAI DENGAN DATASET LITERASI DIGITAL MAHASISWA DI INDONESIA.



TEKNIK PEMBAGIAN DATA



CODE Teknik Pembagian Data

```
x_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)  
✓ 0.0s
```



Teknik Pembagian Data

DALAM PROSES PEMBAGIAN DATA, DIGUNAKAN TEKNIK TRAIN_TEST_SPLIT DENGAN PROPORSI 80% DATA UNTUK PELATIHAN (TRAIN) DAN 20% UNTUK PENGUJIAN (TEST). PADA TAHAPINI, KOLOM GENDER DIJADIKAN SEBAGAI TARGET ATAU LABEL YANG AKAN DIPREDIKSI, BERDASARKAN DATASET LITERASI DIGITAL MAHASISWA DI INDONESIA. PEMBAGIANINI BERTUJUAN AGAR MODEL DAPAT BELAJAR DARI SEBAGIAN BESAR DATA, LALU DIUJI KEMAMPUANNYA MENGGUNAKAN DATA YANG BELUM PERNAH DILIHAT SEBELUMNYA.



MODEL MACHINE LEARNING



Decision Tree

Model Decision Tree digunakan dengan pengaturan kedalaman default, artinya tidak ada batasan khusus pada seberapa dalam pohon keputusan dapat berkembang. Selain itu, random_state diatur sebesar 42 untuk menjaga konsistensi hasil setiap kali model dijalankan. Penggunaan Decision Tree ini bertujuan untuk membangun model yang mampu memahami pola dari berbagai fitur yang ada di dataset literasi digital mahasiswa di Indonesia, sehingga dapat mengklasifikasikan gender berdasarkan karakteristik literasi digital yang dimiliki mahasiswa. Dengan pengaturan ini, model diharapkan mampu memberikan gambaran awal tentang bagaimana data dapat dipisahkan secara sederhana namun tetap efektif.



CODE Decision Tree

```
model_dt = DecisionTreeClassifier(random_state=42)
model_dt.fit(X_train, y_train)
y_pred_dt = model_dt.predict(X_test)
```



Random Forest

Model Random Forest digunakan dengan membangun 100 pohon keputusan (`n_estimators=100`) dan pengaturan `random_state` sebesar 42 untuk memastikan hasil yang konsisten setiap kali model dijalankan. Random Forest dipilih karena kemampuannya dalam menggabungkan banyak pohon keputusan untuk meningkatkan akurasi prediksi dan mengurangi risiko overfitting. Dengan pendekatan ini, model dapat membuat keputusan yang lebih stabil dan andal berdasarkan pola yang ditemukan dalam dataset literasi digital mahasiswa di Indonesia. Setiap pohon dalam Random Forest memberikan kontribusi terhadap keputusan akhir, sehingga hasil klasifikasi gender menjadi lebih kuat dan tidak bergantung pada satu pohon saja.



CODE Random Forest

```
model_rf = RandomForestClassifier(n_estimators=100, random_state=42)
model_rf.fit(X_train, y_train)
y_pred_rf = model_rf.predict(X_test)
```



SVM (Support Vector Machine)

Model Support Vector Machine (SVM) digunakan dengan menggunakan kernel linear, yang berarti model ini mencoba memisahkan kategori data menggunakan sebuah garis lurus (atau bidang datar untuk data berdimensi lebih tinggi). Pemilihan kernel linear dilakukan karena dianggap sesuai untuk data yang dapat dipisahkan dengan batas sederhana. Dalam konteks dataset literasi digital mahasiswa di Indonesia, SVM bertugas untuk menemukan garis pemisah terbaik yang dapat membedakan mahasiswa berdasarkan gender, dengan memaksimalkan jarak antara data dari dua kategori tersebut. Pendekatan ini diharapkan mampu menghasilkan model yang sederhana, namun tetap efektif dalam melakukan klasifikasi.

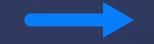


CODE SVM

```
model_svm = SVC(kernel='linear')
model_svm.fit(X_train, y_train)
y_pred_svm = model_svm.predict(X_test)
```



Hasil Evaluasi



Hasil Evaluasi

Pada tahap evaluasi model, digunakan beberapa metrik untuk mengukur kinerjanya, yaitu Accuracy, Classification Report, dan Confusion Matrix. Metrik-metrik ini dipilih karena sesuai dengan karakteristik target yang bersifat biner, yaitu membedakan gender mahasiswa. Dengan menggunakan dataset literasi digital mahasiswa di Indonesia, evaluasi ini membantu untuk memahami seberapa baik model dalam mengklasifikasikan data, melihat distribusi prediksi benar dan salah, serta menganalisis performa model secara lebih rinci melalui nilai precision, recall, dan f1-score.



CODE Hasil evaluasi Decision Tree

```
print("\n[Laporan Decision Tree]")
print(classification_report(y_test, y_pred_dt))
ConfusionMatrixDisplay.from_estimator(model_dt, X_test, y_test)
plt.title('Confusion Matrix - Decision Tree')
plt.show()
```



Hasil evaluasi Decision Tree

[Laporan Decision Tree]				
	precision	recall	f1-score	support
0	0.72	0.70	0.71	414
1	0.32	0.34	0.33	171
accuracy			0.59	585
macro avg	0.52	0.52	0.52	585
weighted avg	0.60	0.59	0.60	585



CODE Hasil evaluasi Random Forest

```
print("\n[Laporan Random Forest]")
print(classification_report(y_test, y_pred_rf))
ConfusionMatrixDisplay.from_estimator(model_rf, X_test, y_test)
plt.title('Confusion Matrix - Random Forest')
plt.show()
```



Hasil evaluasi Random Forest

[Laporan Random Forest]				
	precision	recall	f1-score	support
0	0.72	0.89	0.80	414
1	0.39	0.18	0.24	171
accuracy			0.68	585
macro avg	0.56	0.53	0.52	585
weighted avg	0.63	0.68	0.64	585



CODE Hasil evaluasi SVM

```
print("\n[Laporan SVM]")
print(classification_report(y_test, y_pred_svm)) (variable) X_test:
ConfusionMatrixDisplay.from_estimator(model_svm, X_test, y_test)
plt.title('Confusion Matrix - SVM')
plt.show()
```



Hasil evaluasi SVM

[Laporan SVM]				
	precision	recall	f1-score	support
0	0.71	1.00	0.83	414
1	0.00	0.00	0.00	171
accuracy			0.71	585
macro avg	0.35	0.50	0.41	585
weighted avg	0.50	0.71	0.59	585



ANALISIS EVALUASI

Berdasarkan hasil evaluasi, model Support Vector Machine (SVM) menunjukkan akurasi tertinggi, yaitu sekitar 70% secara desimal 0.7077, dibandingkan dengan Random Forest yang mencapai sekitar 68% secara desimal yaitu 0.6803 dan *Decision Tree* sekitar 59% maupun dalam desimal diangka 0.5932. Hal ini menunjukkan bahwa SVM merupakan model yang paling optimal untuk dataset literasi digital mahasiswa di Indonesia, karena mampu menangani kompleksitas data dengan lebih baik dan menghasilkan pemisahan kategori yang lebih akurat dibandingkan model lainnya.



Hasil Analisis Evaluasi model

Decision Tree: 0.5932

Random Forest: 0.6803

SVM: 0.7077

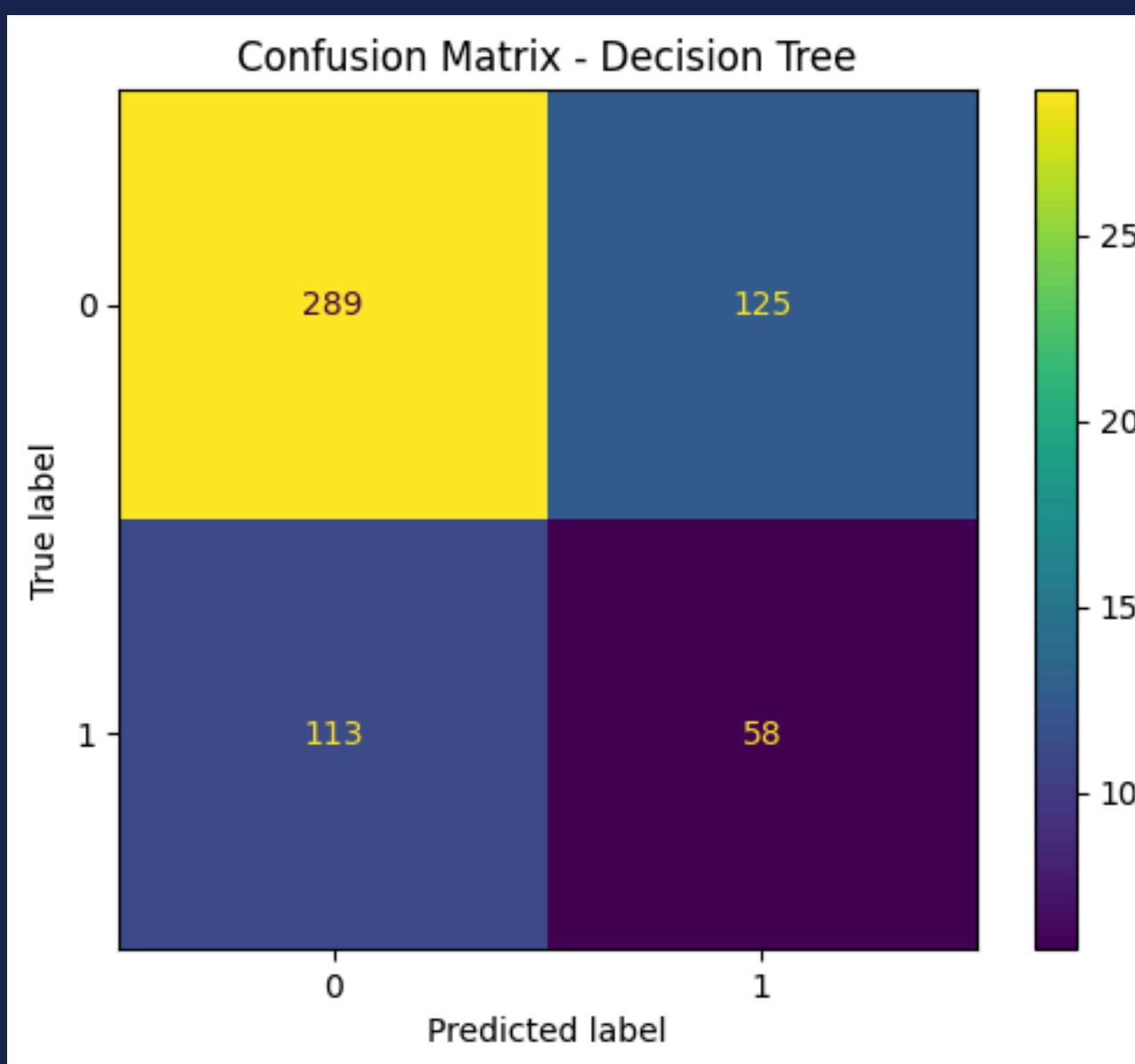


Tambahan

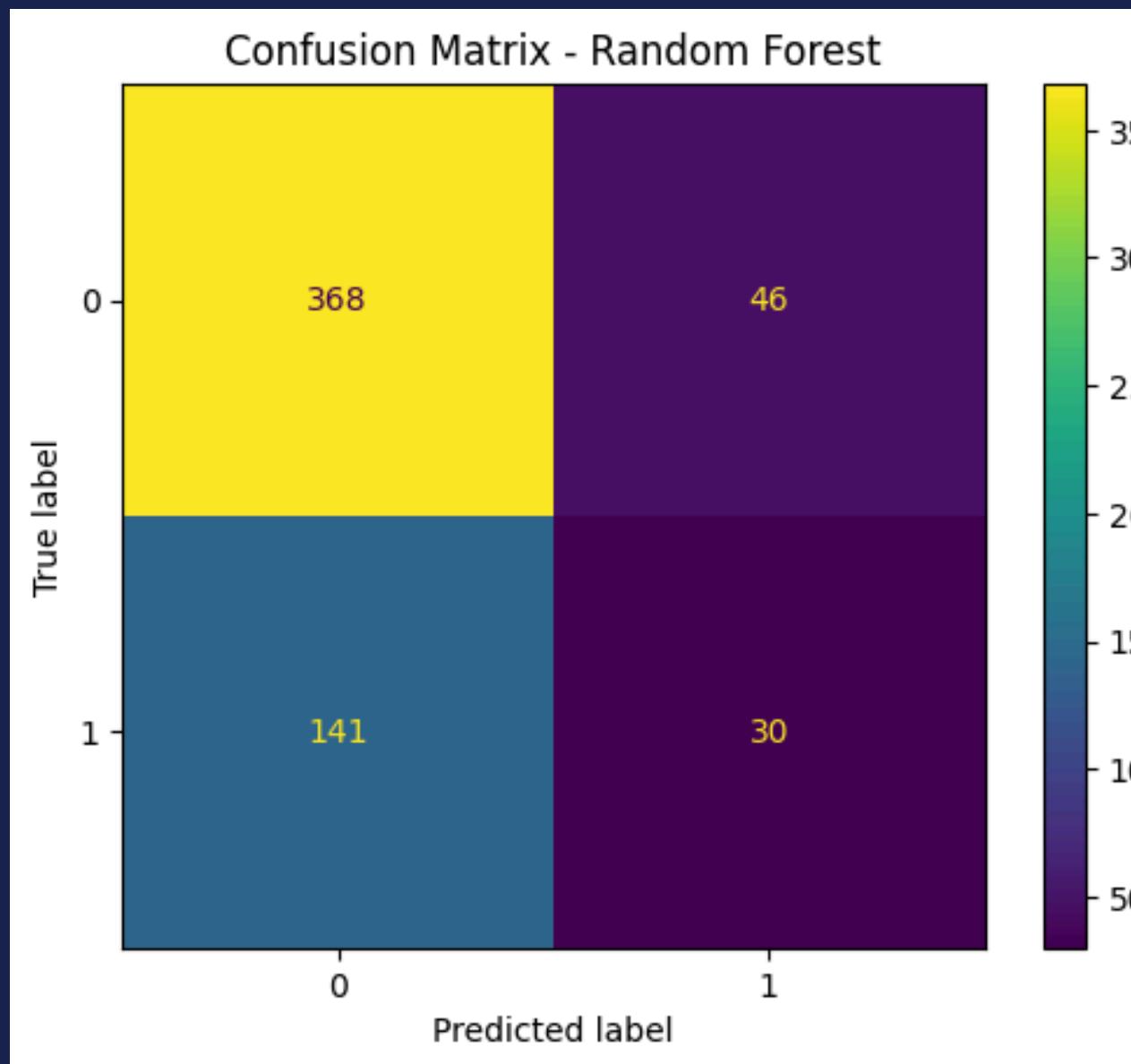
Untuk mengevaluasi performa masing-masing model, dilakukan visualisasi Confusion Matrix untuk setiap model yang digunakan. Visualisasi ini menunjukkan jumlah prediksi benar dan salah, sehingga memudahkan dalam mengidentifikasi pola kesalahan yang terjadi. Selain itu, dibuat grafik perbandingan akurasi antar model untuk memberikan gambaran yang lebih jelas mengenai kinerja masing-masing model. Khusus untuk model Support Vector Machine (SVM), dilakukan analisis feature importance dengan mengamati bobot (koefisien) dari setiap fitur, sehingga dapat diketahui fitur-fitur yang paling berkontribusi dalam membedakan kategori gender berdasarkan dataset literasi digital mahasiswa di Indonesia.



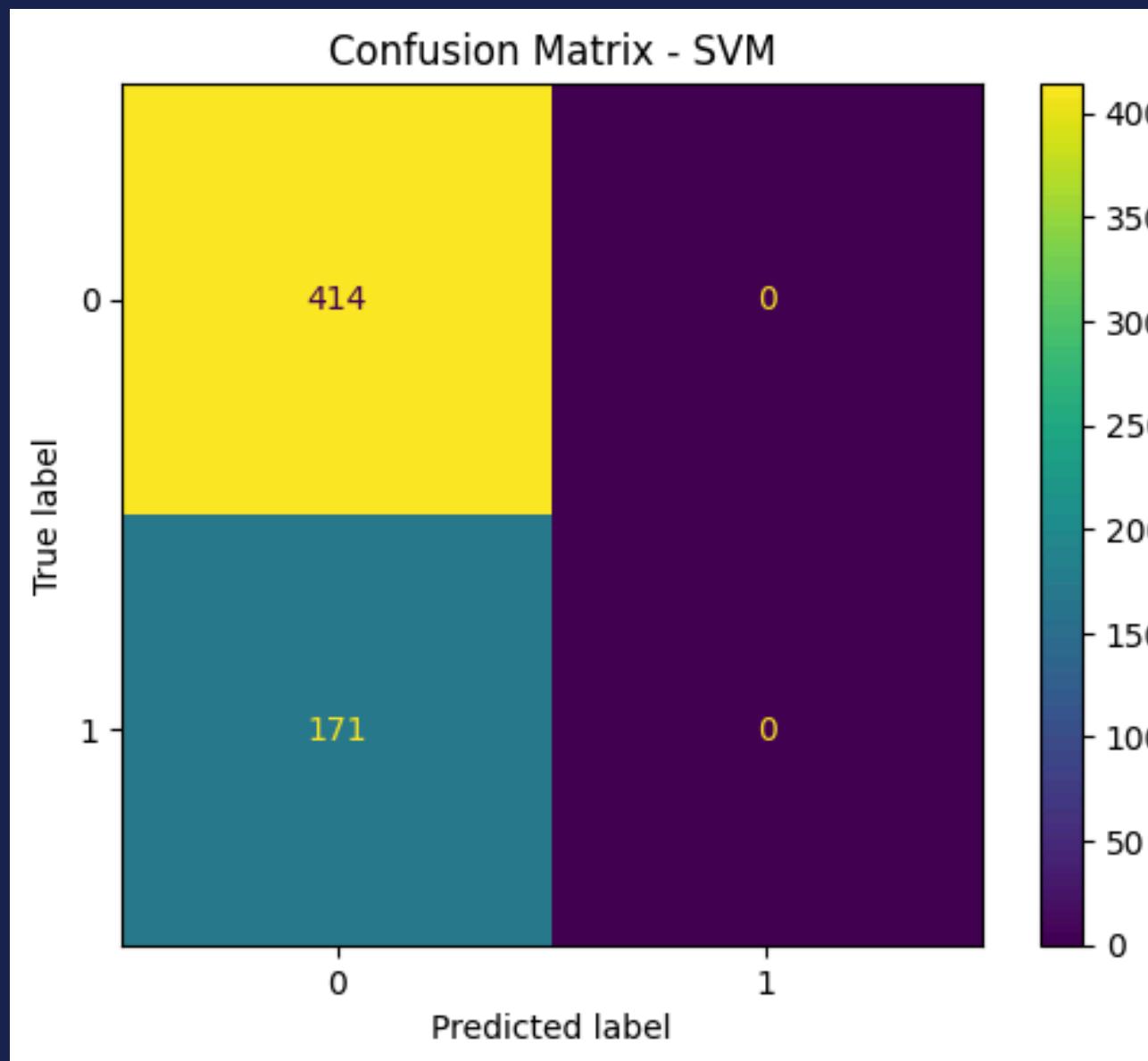
Visualisasi confusion matrix decision tree



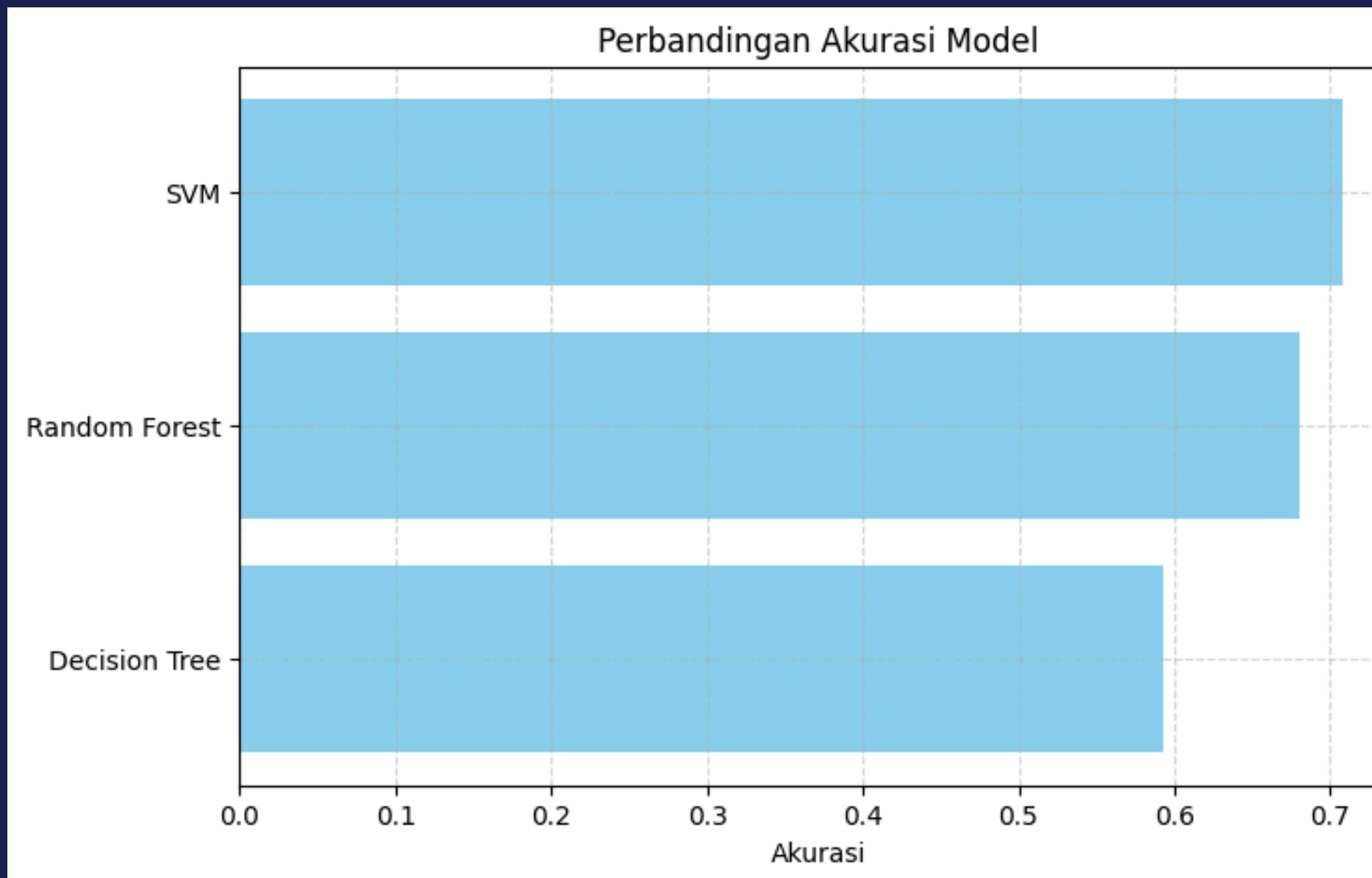
Visualisasi confusion matrix Random Forest



Visualisasi confusion matrix SVM



Visualisasi Perbandingan Akurasi Model



THANK YOU

