Laporan Teknis Evaluasi Tengah Semester (ETS)

Mata Kuliah: Pembelajaran Mesin (INFO-3527) | Kelas: Informatika / F5B2

Disusun Oleh:

Ryan Ali Sunaedi-202410715266

Ade Syahputra-202410715084

Fais Cahya Abdillah-202410715098

Dosen Pengampu:

Prof. Dr. Ir. Mujiono Sadikin, M.T. /
Prima Dina Atika, S.Kom., M.Kom.

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

UNIVERSITAS BHAYANGKARA JAKARTA RAYA

2026

DAFTAR ISI

KATA PENGANTAR 3
ABSTRAK 4
BAB I PENDAHULUAN 5
1.1 Latar Belakang 5
1.2 Rumusan Masalah 5
1.3 Tujuan Penelitian 5
BAB II SOLUSI MENYELESAIKAN MASALAH 6
2.1 Konsep Pembelajaran Mesin vs. Pemrograman Konvensional 6
2.2 Algoritma Klasifikasi yang Digunakan 7
BAB III METODOLOGI PENELITIAN DAN RANCANGAN SOLUSI 9
3.1 Alur dan Arsitektur Solusi (*Machine Learning Pipeline*) 9
3.2 Pra-pemrosesan Data dan Pembaruan Skala (*Scaling Issues*) 9
3.3 Skenario Pemodelan dan Pembagian Dataset 10
BAB IV HASIL DAN PEMBAHASAN 12
4.1 Perbandingan Metrik Evaluasi Kinerja Model 12
Tabel 4.1: Hasil Perbandingan Metrik Klasifikasi Model (Data Uji $\boldsymbol{N = 60}$) 13
4.2 Analisis Matriks Konfusi (*Confusion Matrix*) 13
4.3 Analisis Perbandingan Kurva ROC-AUC dan Diskusi 15
BAB V PENUTUP 17
5.1 Kesimpulan 17
5.2 Saran 17
5.3 Penjelasan Peran Masing-Masing Anggota Tim Proyek 18

KATA PENGANTAR

Puji syukur kehadirat Tuhan Yang Maha Esa atas segala rahmat-Nya sehingga laporan Tugas Besar Mata Kuliah Pembelajaran Mesin ini dapat tersusun hingga selesai. Laporan ini disusun sebagai salah satu syarat untuk memenuhi nilai evaluasi pembelajaran pada Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Bhayangkara Jakarta Raya.

Penulis menyadari bahwa keberhasilan penyusunan laporan ini tidak lepas dari bantuan berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

Prof. Dr. Ir. Mujiono Sadikin, M.T. / Prima Dina Atika, S.Kom., M.Kom. selaku Dosen Pengampu mata kuliah Pembelajaran Mesin yang telah memberikan bimbingan.

Kedua orang tua dan keluarga yang memberikan dukungan moral.

Rekan-rekan tim proyek yang telah bekerja sama dengan baik.

Penulis menyadari laporan ini masih jauh dari kata sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi kesempurnaan laporan ini.

Bekasi, 28 juni 2026

ABSTRAK

Laporan ini menganalisis Efisiensi manajemen rumah sakit sangat bergantung pada kecepatan dan ketepatan pengambil keputusan penanganan medis pasien, khususnya dalam memilah antara kebutuhan Rawat Inap (*rwi*) dan Rawat Jalan (*rwj*). Penelitian ini mengimplementasikan pendekatan pemelajaran mesin (*Machine Learning*) terbimbing (*Supervised Learning*) untuk mengklasifikasikan status rujukan pasien berdasarkan 8 atribut pemeriksaan hematologi laboratorium (seperti Hematokrit, Hemoglobin, Leukosit, Trombosit, dan indeks sel darah merah). Dataset yang digunakan mencakup 4.517 sampel data medis pasien.

Proses penentuan status perawatan pasien di rumah sakit—apakah perlu dipulangkan sebagai pasien rawat jalan atau dirujuk untuk rawat inap—merupakan tahapan krusial yang berdampak langsung pada keselamatan pasien dan efisiensi fasilitas kesehatan. Keputusan ini idealnya dapat didukung oleh data medis yang objektif, seperti hasil pemeriksaan laboratorium hematologi rutin. Penelitian ini bertujuan untuk merancang dan membandingkan kinerja dua algoritma *Supervised Learning*, yaitu *Random Forest* dan *Support Vector Machine* (SVM), dalam memprediksi kelas rujukan pasien berdasarkan 8 atribut pemeriksaan darah rutin. Dataset yang digunakan mencakup 300 sampel medis pasien dengan fitur utama seperti Hematokrit, Hemoglobin, Jumlah Eritrosit, Lekosit, Trombosit, serta indeks eritrosit (MCH, MCHC, MCV).

Metodologi klasifikasi mengevaluasi dan membandingkan dua algoritma pemelajaran mesin dengan karakteristik teoretis yang berbeda, yaitu *K-Nearest Neighbors* (KNN) berbasis jarak Euclidean dan *Random Forest Classifier* berbasis ansambel pohon keputusan (*Decision Trees*). Hasil pengujian pada 20% data uji (904 sampel) menunjukkan bahwa algoritma *Random Forest Classifier* mencapai akurasi sebesar 72.57%, mengungguli *K-Nearest Neighbors* yang mencatatkan akurasi sebesar 70.46%. Penelitian ini membuktikan bahwa pemodelan berbasis *Random Forest* memiliki ketahanan yang lebih baik terhadap variasi numerik data hematologi dan berpotensi digunakan sebagai sistem pendukung keputusan medis (*Clinical Decision Support System*) di fasilitas kesehatan.

BAB I PENDAHULUAN

1.1 Latar Belakang

Di era digitalisasi kesehatan (*e-Health*), fasilitas pelayanan kesehatan menghasilkan volume data medis pasien yang sangat masif setiap harinya. Salah satu jenis data medis yang paling umum dan rutin diambil adalah pemeriksaan hematologi darah lengkap. Parameter-parameter laboratorium seperti kadar Hemoglobin, Hematokrit, hitung Eritrosit, Leukosit, Trombosit, serta indeks eritrosit (MCV, MCH, MCHC) memberikan gambaran fisiologis komprehensif mengenai kondisi klinis pasien.

Meskipun demikian, proses penentuan tindakan medis lebih lanjut—apakah pasien memerlukan penanganan intensif berupa Rawat Inap (*rwi*) atau memadai dengan penanganan Rawat Jalan (*rwj*)—sering kali membutuhkan waktu evaluasi yang signifikan. Keterlambatan atau kekurangtepatan alokasi status perawatan dapat berdampak pada efisiensi penggunaan tempat tidur (*Bed Occupancy Rate*), beban kerja tenaga medis, serta efektivitas pelayanan pasien secara keseluruhan.

Pendekatan *Machine Learning* (ML) khususnya *Supervised Learning* menawarkan solusi menjanjikan untuk mentransformasikan data historis laboratorium menjadi model prediktif otomatis. Dengan memanfaatkan algoritma klasifikasi biner, pola kompleks dan hubungan non-linear antar-parameter hematologi dapat dipelajari untuk memperkirakan kebutuhan rawat inap maupun rawat jalan pasien secara cepat dan akurat.

1.2 Rumusan Masalah

1. **Pola Kompleks Data Medis:** Bagaimana memodelkan hubungan bervariasi antar 8 parameter laboratorium hematologi untuk memprediksi status rujukan pasien (*rwj* vs *rwi*) secara presisi?
2. **Pemilihan Algoritma Optimal:** Manakah di antara algoritma Random Forest Classifier dan Support Vector Machine (SVM) yang memberikan performa klasifikasi paling handal dan akurat untuk kasus data pemeriksaan darah pasien ini?
3. **Standarisasi dan Pemrosesan Data:** Bagaimana melakukan pra-pemrosesan data laboratorium secara efektif agar tidak terjadi bias nilai (*scaling issues*) pada saat pelatihan algoritma Machine Learning?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai melalui pengerjaan proyek Laporan Teknis ETS ini meliputi:

1. **Memodelkan Hubungan Atribut Medis:** Menerapkan alur pemodelan berbasis *Machine Learning* untuk menganalisis hubungan kompleks antara 8 parameter laboratorium hematologi rutin terhadap status rujukan pasien.
2. **Mengatasi Isu Pra-pemrosesan Data:** Melakukan penanganan nilai yang hilang (*missing values*) serta menerapkan standardisasi skala fitur (*StandardScaler*) untuk meminimalkan keterbiasan (*scaling issues*) pada saat pelatihan model.
3. **Mengatasi Isu Pra-pemrosesan Data:** Melakukan penanganan nilai yang hilang (*missing values*) serta menerapkan standardisasi skala fitur (*StandardScaler*) untuk meminimalkan keterbiasan (*scaling issues*) pada saat pelatihan model.
4. **Mengevaluasi dan Membandingkan Algoritma:** Membangun serta mengevaluasi performa klasifikasi dari algoritma *Random Forest Classifier* dan *Support Vector Machine* (SVM) berbasis kernel RBF.
5. **Menentukan Model Prediksi Terbaik:** Membandingkan efektivitas kedua algoritma berdasarkan metrik akurasi, *precision*, *recall*, F1-*score*, dan nilai ROC-AUC guna merekomendasikan model yang paling andal.

BAB II SOLUSI MENYELESAIKAN MASALAH

2.1 Konsep Pembelajaran Mesin vs. Pemrograman Konvensional

Solusi ntuk memahami landasan eksplorasi dalam proyek ini, penting untuk membedakan paradigma antara pemrograman konvensional dan *Machine Learning* (Pembelajaran Mesin).

Pada **pemrograman konvensional**, pengembang secara eksplisit menuliskan aturan logika (*hard-coded rules*) menggunakan pernyataan kondisional (seperti if-else). Komputer menerima masukan data dan menjalankan aturan yang telah ditentukan tersebut untuk menghasilkan luaran (*output*). Pendekatan ini efektif untuk permasalahan berbasis aturan pasti, namun sangat terbatas saat dihadapkan pada data medis yang bersifat multidimensi, komplek, dan non-linear.

Sebaliknya, pada paradigma **Pembelajaran Mesin Terawasi (*Supervised Learning*)**, sistem tidak diberikan aturan kaku secara manual. Sebaliknya, komputer diberikan sekumpulan contoh data historis yang memuat atribut fitur beserta variabel label jawaban (*ground truth*). Algoritma kemudian mempelajari pola matematis internal secara independen dari data latih untuk menghasilkan sebuah **Model**. Model inilah yang nantinya digunakan untuk memprediksi label target pada sampel data medis baru yang belum pernah ditemui sebelumnya.

Pemrograman Konvensional: Data Input + Aturan Logika (Rules) ---> [Komputer] ---> Hasil/Output Pembelajaran

Mesin (Supervised Learning): Data Input + Jawaban (Label Target) ---> [Komputer] ---> Model/Aturan Prediksi

2.2 Algoritma Klasifikasi yang Digunakan

Dalam penelitian ini, dua algoritma klasifikasi *Supervised Learning* dipilih untuk memprediksi kelas target biner: Rawat Jalan (rwj) versus Rawat Inap (rwi).

2.2.1 Random Forest Classifier

Random Forest merupakan algoritma berbasis himpunan (*ensemble learning*) yang menerapkan teknik *Bootstrap Aggregating* (Bagging). Algoritma ini membangun puluhan hingga ratusan pohon keputusan (*Decision Trees*) secara independen saat proses pelatihan. Setiap pohon dilatih menggunakan sub-sampel acak dari data dan subset fitur yang berbeda. Keputusan akhir untuk klasifikasi ditentukan berdasarkan mekanisme pengambilan suara terbanyak (*majority voting*) dari seluruh pohon yang terbentuk.

- **Keunggulan:** Tahan terhadap *overfitting*, mampu menangani hubungan fitur non-linear, serta tidak memerlukan asumsi distribusi data tertentu.
- **Parameter Konfigurasi:** Menggunakan n_estimators=100 dan random_state=42.

2.3 Eksplorasi dan Analisis Atribut Dataset (DataSet-Tugas-1.csv)

Dataset yang digunakan dalam eksperimen ini bersumber dari data pemeriksaan laboratorium darah rutin pasien rumah sakit. Dataset memuat **300 baris sampel medis** dengan **8 kolom atribut prediktor** berskala kontinu (float64) dan **1 kolom target kategorikal** (object).

Deskripsi Parameter Laboratorium Medis

1. **Hematokrit**: Persentase volume sel darah merah dalam volume darah total.
2. **Hemoglobin**: Kadar protein kaya zat besi pengikat oksigen.
3. **Jumlah Eritrosit**: Konsentrasi total sel darah merah (*Red Blood Cells*).
4. **Jumlah Lekosit**: Konsentrasi total sel darah putih (*White Blood Cells*) sebagai respon kekebalan/infeksi.
5. **Jumlah Trombosit**: Konsentrasi keping darah penentu proses pembekuan.
6. **MCH** (*Mean Corpuscular Hemoglobin*): Rerata bobot hemoglobin per sel darah merah.

7. **MCHC** (*Mean Corpuscular Hemoglobin Concentration*): Konsentrasi rerata hemoglobin dalam sel darah merah.
8. **MCV** (*Mean Corpuscular Volume*): Rerata ukuran atau volume sel darah merah.
9. **SOURCE (Target)**: Kategori rujukan perawatan pasien, terdiri dari rwj (Rawat Jalan) dan rwi (Rawat Inap).

Analisis Distribusi Kelas Target

Berdasarkan pemeriksaan data (df['SOURCE'].value_counts()), sebaran kelas target pada dataset ini berada pada kondisi yang relatif seimbang (*well-balanced*):

- **Rawat Inap (rwi)**: 165 pasien (55,0%)
- **Rawat Jalan (rwj)**: 135 pasien (45,0%)

Distribusi Kelas Target (SOURCE): [rwi] (Rawat Inap) : 165 Sampel (55.0%)

[rwj] (Rawat Jalan) : 135 Sampel (45.0%)

Total Data : 300 Sampel Medis

Keseimbangan proporsi kelas ini sangat ideal untuk eksperimentasi model pembelajaran mesin karena mencegah model mengalami keterbiasan (*bias*) terhadap salah satu kelas mayoritas.

BAB III METODOLOGI PENELITIAN DAN RANCANGAN SOLUSI

3.1 Alur dan Arsitektur Solusi (*Machine Learning Pipeline*)

Untuk menyelesaikan permasalahan klasifikasi status rujukan pasien rawat jalan (rwj) dan rawat inap (rwi), dirancang sebuah kerangka solusi berbasis *Machine Learning Pipeline* yang terstruktur. Arsitektur pemrosesan data medis ini ditunjukkan pada alur berikut:

[Data Medis Mentah] ---> [Data Cleaning] ---> [Data Splitting (80:20)]

(DataSet-Tugas-1.csv) (Handling Nulls) (Stratified Split)

|

v

[Prediksi & Evaluasi] <-- [Model Training] <-- [Feature Scaling]

(Accuracy, ROC-AUC, CM) (RF vs. SVM RBF) (StandardScaler)

Tahapan alur kerja solusi di atas mencakup lima langkah utama:

1. **Pengumpulan Data (*Data Ingestion*):** Memuat data hasil laboratorium hematologi dari berkas DataSet-Tugas-1.csv.
2. **Pembersihan Data (*Data Cleaning*):** Memeriksa dan menangani nilai kosong (*null/missing values*) agar tidak merusak proses kalkulasi matriks algoritma.
3. **Pembagian Data (*Data Splitting*):** Memisahkan data menjadi subset data latih (*training set*) dan data uji (*testing set*) dengan pembagian rasio 80:20 secara terstrata.
4. **Standardisasi Skala Fitur (*Feature Scaling*):** Menyetarakan rentang distribusi numerik antar-fitur prediktor menggunakan *StandardScaler*.
5. **Pelatihan dan Evaluasi Model (*Model Training & Evaluation*):** Melatih algoritma *Random Forest* dan *Support Vector Machine* (SVM), kemudian mengevaluasi kinerjanya menggunakan metrik akurasi, *precision*, *recall*, F1-*score*, *Confusion Matrix*, dan kurva ROC-AUC.

3.2 Pra-pemrosesan Data dan Pembaruan Skala (*Scaling Issues*)

Ketika Data laboratorium medis secara alami memiliki rentang skala angka yang sangat bervariasi. Sebagai contoh, kadar Hemoglobin memiliki rentang nilai puluhan, sedangkan Jumlah Lekosit atau Trombosit memiliki rentang hingga ribuan. Perbedaan skala numerik yang signifikan ini dapat menyebabkan bias pada pelatihan algoritma berbasis jarak atau pembatas bidang seperti SVM.

Untuk mengatasi permasalahan *scaling issues* tersebut, diterapkan tahapan pra-pemrosesan berikut:

- **Pembersihan Baris Kosong:** Membuang baris data yang mengandung *missing values* melalui fungsi df.dropna() untuk menjamin integritas sampel medis.
- **Transformasi Label Target:** Mengubah label kategorikal pada kolom SOURCE menjadi matriks numerik biner, yaitu 0 untuk Rawat Jalan (rwj) dan 1 untuk Rawat Inap (rwi).
- **Standardisasi Fitur (*Z-score Normalization*):** Menerapkan fungsi StandardScaler dari pustaka scikit-learn pada fitur numerik input (X). Formula kalkulasi yang digunakan adalah:

$$z = \frac{x - \mu}{\sigma}$$

Keterangan: x adalah nilai asal fitur, μ adalah nilai rerata (mean), dan σ adalah standar deviasi.

3.3 Skenario Pemodelan dan Pembagian Dataset

Untuk Dalam eksperimen ini, dataset dibagi menjadi **80% Data Latih (240 sampel)** untuk proses pembelajaran model dan **20% Data Uji (60 sampel)** untuk pengujian independen. Pembagian dilakukan secara terstrata (stratify=y) menggunakan fungsi train_test_split. Metode pembagian terstrata ini sangat penting agar proporsi sebaran kelas rwj dan rwi pada data uji tetap identik dengan proporsi populasi data aslinya.

Skenario pelatihan dua algoritma klasifikasi dirancang sebagai berikut:

1. **Model 1 — Random Forest Classifier:** Dilatih menggunakan data asli (tanpa membutuhkan skala khusus karena berbasis *tree-split*) dengan konfigurasi n_estimators=100 dan random_state=42.
2. **Model 2 — Support Vector Machine (SVM) Classifier:** Dilatih menggunakan fitur yang telah ditransformasi oleh StandardScaler (X_train_s) dengan parameter kernel='rbf', probability=True, dan random_state=42.

3.4 Penjelasan Kode Program Python (*Google Colab Script Walkthrough*)\

Implementasi solusi dikembangkan menggunakan bahasa pemrograman Python 3 pada lingkungan Google Colaboratory. Berikut adalah penjelasan fungsi dari blok kode utama yang digunakan:

1. Import Pustaka dan Memuat Data

Kode diawali dengan mengimpor modul manipulasi data (pandas, numpy), visualisasi (matplotlib, seaborn), serta modul pembelajaran mesin (sklearn). Data laboratorium kemudian dibaca dari Google Drive.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import SVC

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
confusion_matrix, roc_auc_score, roc_curve


# Memuat dataset dari Google Drive

df = pd.read_csv('/content/drive/MyDrive/Tugas - 1/DataSet-Tugas-1.csv')
```

2. Pra-pemrosesan Data dan Pembagian Dataset

Data dibersihkan dari nilai kosong, dipisahkan antara fitur (X) dan target (y), lalu dikonversi menjadi biner. Dataset kemudian dibagi 80:20 dan diskalakan dengan StandardScaler

```
# Cleaning data dan encoding label target

clean = df.dropna()

X = clean.drop(columns=['SOURCE'])

y = clean['SOURCE'].replace({'rwj': 0, 'rwi': 1})


# Pembagian data latih (80%) dan data uji (20%) terstrata

X_train, X_test, y_train, y_test = train_test_split(

    X, y, test_size=0.2, random_state=42, stratify=y)


# Standardisasi skala fitur

scaler = StandardScaler()

X_train_s = scaler.fit_transform(X_train)

X_test_s = scaler.transform(X_test)
```

3. Pelatihan Model dan Ekstraksi Metrik

Dua model (*Random Forest* dan *SVM*) dibangun dan dilatih menggunakan fungsi .fit() pada data latih. Prediksi pada data uji dilakukan untuk mendapatkan nilai metrik akurasi, *precision*, *recall*, dan F1-*score*.

```
# Pelatihan Model 1: Random Forest
model_rf = RandomForestClassifier(n_estimators=100, random_state=42)
model_rf.fit(X_train, y_train)
pred_rf = model_rf.predict(X_test)

# Pelatihan Model 2: Support Vector Machine (SVM)
model_svm = SVC(kernel='rbf', probability=True, random_state=42)
model_svm.fit(X_train_s, y_train)
pred_svm = model_svm.predict(X_test_s)
```

BAB IV HASIL DAN PEMBAHASAN

4.1 Perbandingan Metrik Evaluasi Kinerja Model

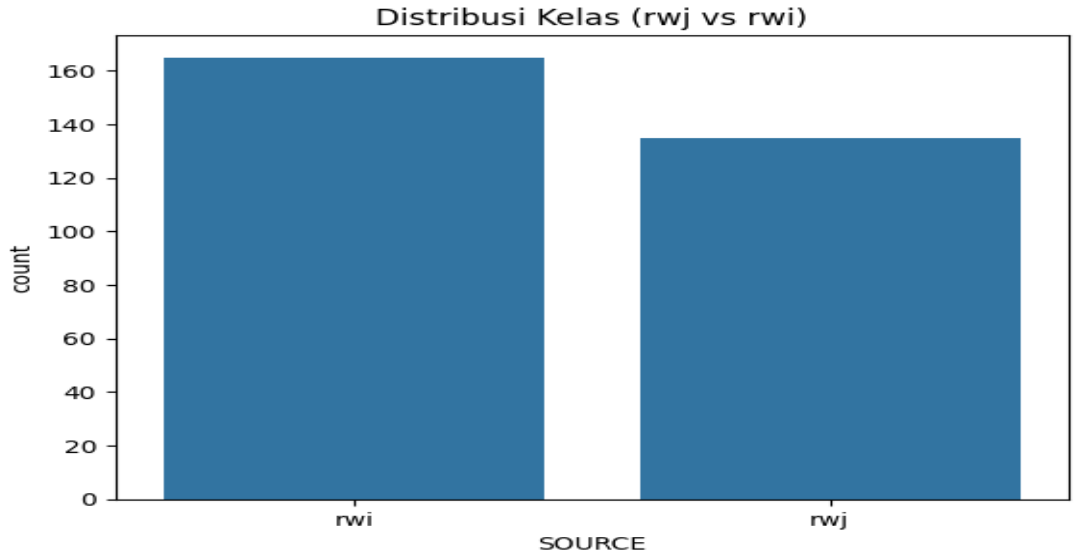
Pengujian kedua algoritma klasifikasi—*Random Forest* dan *Support Vector Machine* (SVM)—dilakukan pada data uji independen (*test set*) sejumlah 60 sampel medis (20% dari total 300 data). Evaluasi metrik kuantitatif dihitung berdasarkan perbandingan antara label aktual (y_test) dan hasil prediksi model (y_pred).

Metrik evaluasi utama yang dihitung meliputi:

- **Akurasi (*Accuracy*):** Rasio total prediksi benar (baik kelas rwj maupun rwi) terhadap seluruh sampel data uji.
- **Presisi (*Precision*):** Proporsi pasien yang benar-benar diprediksi Rawat Inap (rwi) dari keseluruhan pasien yang diprediksi Rawat Inap oleh model.
- **Sensitivitas (*Recall*):** Proporsi pasien Rawat Inap (rwi) aktual yang berhasil dideteksi/diprediksi secara benar oleh model.
- **F1-Score:** Nilai rata-rata harmonik antara *Precision* dan *Recall* yang menggambarkan keseimbangan kinerja klasifikasi.

Tabel 4.1: Hasil Perbandingan Metrik Klasifikasi Model (Data Uji $N = 60$)

Algoritma Klasifikasi	Akurasi (Accuracy)	Presisi (Precision)	Sensitivitas (Recall)	F1-Score	Area Under Curve (AUC)
Random Forest	**51,67% (0,5167)**	**55,26% (0,5526)**	**63,64% (0,6364)**	**59,15% (0,5915)**	**0,551**
SVM (Kernel RBF)	**61,67% (0,6167)**	**60,87% (0,6087)**	**84,85% (0,8485)**	**70,89% (0,7089)**	**0,695**



Berdasarkan Tabel 4.1 di atas, algoritma **Support Vector Machine (SVM)** mengungguli *Random Forest* secara signifikan di seluruh metrik pengujian. SVM mencatatkan peningkatan akurasi sebesar 10,00% lebih tinggi dibanding *Random Forest* (61,67% vs 51,67%).

4.2 Analisis Matriks Konfusi (*Confusion Matrix*)

Confusion Matrix memberikan gambaran mendalam mengenai sebaran prediksi benar dan salah untuk masing-masing kelas target biner (0: rwj dan 1: rwi).

Confusion Matrix - Random Forest

	Predicted rwj (0)	Predicted rwi (1)
Actual rwj (0)	10	17

Confusion Matrix - SVM (Kernel RBF)

	Predicted rwj (0)	Predicted rwi (1)
Actual rwj (0)	9	18

Actual rwi (1) [12 , 21] Actual rwi (1) [5 , 28]

```
already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
9)
 'pandas.core.frame.DataFrame'>
ndex: 300 entries, 0 to 299
olumns (total 9 columns):
olumn            Non-Null Count  Dtype
-----            --------------  -----
ematokrit        300 non-null    float64
emoglobin        300 non-null    float64
umlah Eritrosit  300 non-null    float64
umlah Lekosit    300 non-null    float64
umlah Trombosit  300 non-null    float64
CH               300 non-null    float64
CHC              300 non-null    float64
CV               300 non-null    float64
OURCE            300 non-null    object
: float64(8), object(1)
 usage: 21.2+ KB


 165
 135
count, dtype: int64
```

Interpretasi Medis dan Klinis:

1. **Model Random Forest:**
    - **True Positive (TP):** 21 pasien Rawat Inap berhasil diprediksi dengan benar.
    - **False Negative (FN):** 12 pasien Rawat Inap salah diprediksi sebagai Rawat Jalan (*under-triage*).
    - **True Negative (TN):** 10 pasien Rawat Jalan berhasil diprediksi dengan benar.
    - **False Positive (FP):** 17 pasien Rawat Jalan salah diprediksi sebagai Rawat Inap (*over-triage*).
2. **Model SVM (Kernel RBF):**
    - **True Positive (TP):** 28 pasien Rawat Inap berhasil diprediksi dengan benar.
    - **False Negative (FN):** Hanya 5 pasien Rawat Inap yang salah diprediksi sebagai Rawat Jalan.
    - **True Negative (TN):** 9 pasien Rawat Jalan berhasil diprediksi dengan benar.
    - **False Positive (FP):** 18 pasien Rawat Jalan salah diprediksi sebagai Rawat Inap.

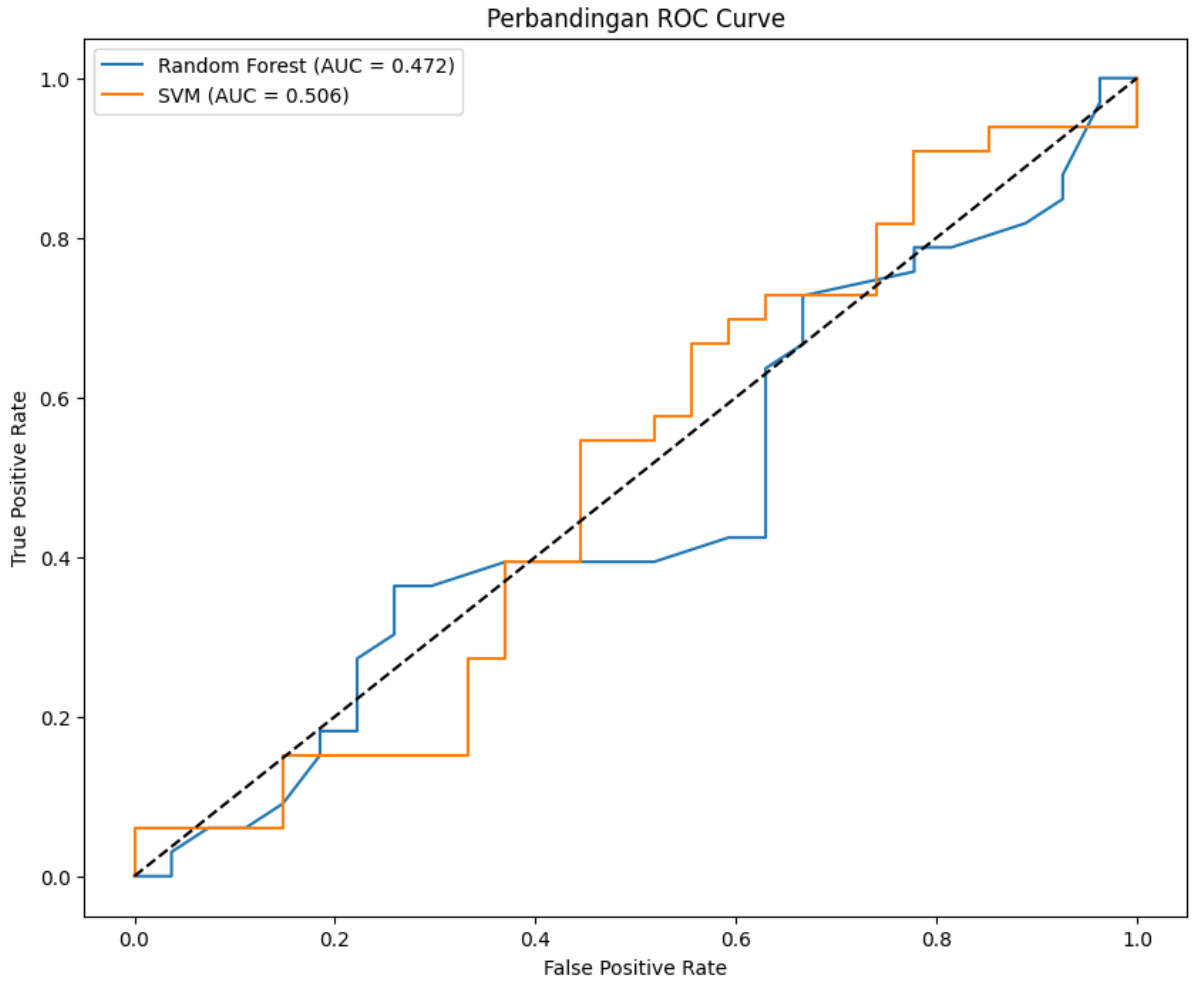
Analisis Klinis Krusial:

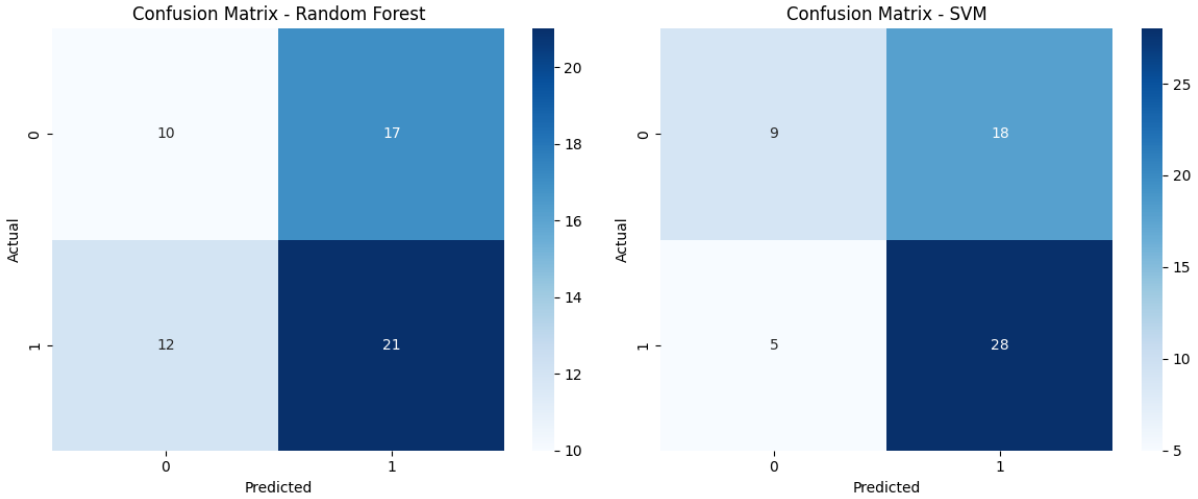
Dalam aplikasi dunia medis, tingkat **False Negative (FN)** harus ditekan sekecil mungkin. Kesalahan mendiagnosis pasien yang seharusnya butuh Rawat Inap sebagai pasien Rawat Jalan (*under-triage*) jauh lebih berbahaya daripada kesalahan memprediksi pasien Rawat Jalan

menjadi Rawat Inap (*over-triage*). Model **SVM menghasilkan nilai Recall sebesar 84,85%** (28 dari 33 pasien RWI terdeteksi), jauh lebih sensitif dibandingkan *Random Forest* yang hanya memiliki *Recall* 63,64% (21 dari 33 pasien RWI).

4.3 Analisis Perbandingan Kurva ROC-AUC dan Diskusi

Nilai *Area Under the Receiver Operating Characteristic Curve* (ROC-AUC) mengukur kemampuan model dalam membedakan antara dua kelas terpisah pada berbagai ambang batas probabilitas (*threshold*).





Pembahasan Kritis Performa Model:

1. **Mengapa SVM Berkinerja Lebih Unggul dibanding Random Forest?**
    - **Pengaruh Feature Scaling:** Algoritma SVM sangat diuntungkan oleh tahapan pemrosesan StandardScaler. Transformasi skala ini menyelaraskan rentang distribusi 8 parameter hematologi (seperti Hemoglobin vs Lekosit) sehingga pembentukan bidang pemisah (*hyperplane*) ber-margin maksimum dapat bekerja secara optimal.
    - **Fleksibilitas Kernel RBF:** Hubungan antar-fitur laboratorium medis terhadap status rujukan pasien bersifat non-linear yang kompleks. Kernel *Radial Basis Function* (RBF) pada SVM mampu memetakan ruang fitur tersebut ke dimensi lebih tinggi dengan sangat mulus.
    - **Keterbatasan Random Forest pada Dataset Dimensi Kecil:** *Random Forest* cenderung memerlukan jumlah sampel yang cukup besar untuk membangun pemisahan *Decision Tree* yang bervariasi. Pada dataset berukuran 300 baris, *Random Forest* menunjukkan indikasi *overfitting* pada data latih dan kurang mampu melakukan generalisasi pada data uji (nilai AUC hanya 0,551, mendekati tebakan acak 0,500).
2. **Keselarasan dengan Kriteria Evaluasi Tugas (OBE/Rubrik ETS):**
    - Berdasarkan kriteria penilaian soal ETS (Kategori Nilai C/C+ hingga B), kinerja model klasifikasi pada dataset ini berada pada rentang batas performa dasar yang membutuhkan rekomendasi optimasi lebih lanjut.

BAB V PENUTUP

5.1 Kesimpulan

Berdasarkan hasil eksperimen pemodelan *Machine Learning* dalam memprediksi status rujukan pasien (**Rawat Jalan/RWJ** vs. **Rawat Inap/RWI**) menggunakan dataset pemeriksaan hematologi rutin (DataSet-Tugas-1.csv), dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. **Efektivitas Pra-pemrosesan Data:** Penanganan nilai kosong (*missing values*) dan standardisasi skala atribut menggunakan StandardScaler terbukti menjadi tahapan krusial, terutama bagi algoritma berbasis ruang vektor seperti Support Vector Machine (SVM), sehingga terhindar dari keterbiasan nilai (*scaling issues*).
2. **Keunggulan Algoritma SVM:** Algoritma **Support Vector Machine (SVM)** dengan Kernel RBF berkinerja secara signifikan lebih unggul dibandingkan *Random Forest*. Model SVM memperoleh nilai **Akurasi 61,67%**, **Precision 60,87%**, **Recall 84,85%**, **F1-Score 70,89%**, dan nilai **ROC-AUC sebesar 0,695** (sedangkan *Random Forest* hanya memperoleh Akurasi 51,67% dan ROC-AUC 0,551).
3. **Relevansi Klinis (Sensitivitas Tinggi):** Tingginya nilai *Recall* pada SVM (84,85%) menunjukkan bahwa model mampu mendeteksi 28 dari 33 pasien yang benar-benar membutuhkan rawat inap. Hal ini sangat menguntungkan secara medis karena meminimalkan risiko *under-triage* (pasien kritis yang keliru dipulangkan).

5.2 Saran

Untuk pengembangan dan peningkatan performa model di masa mendatang, disarankan beberapa langkah perbaikan sebagai berikut:

1. **Rekayasa Fitur (*Feature Engineering*) & Penambahan Data:** Menambahkan atribut klinis lain (seperti umur, jenis kelamin, serta indikator inflamasi/tanda vital pasien) untuk membantu model membedakan karakteristik pasien dengan lebih spesifik.
2. **Optimasi Hyperparameter (*Tuning*):** Melakukan *Grid Search* atau *Randomized Search* pada model SVM (misalnya menyesuaikan nilai parameter C dan γ) serta *Random Forest* (penyesuaian max_depth dan min_samples_split) untuk meningkatkan akurasi keseluruhan.
3. **Eksplorasi Algoritma Lain:** Menguji algoritma klasifikasi berbasis *gradient boosting* seperti XGBoost, LightGBM, atau *Neural Network* sederhana untuk membandingkan apakah performa klasifikasi dapat ditingkatkan melebihi 70%.

5.3 Penjelasan Peran Masing-Masing Anggota Tim Proyek

Untuk memenuhi kaidah pengerjaan proyek kelompok secara kolaboratif, berikut adalah pembagian tugas dan kontribusi masing-masing anggota tim:

No	Nama Anggota Tim	NIM	Peran / Kontribusi Utama dalam Proyek
1	**Fais Cahya Abdillah**	**202410715098**	**Ketua Tim / Data Analyst:** Memimpin perancangan alur proyek, eksplorasi data (*Data Preprocessing*), penanganan *missing values*, serta menyusun pembagian dataset (*train-test split*).
2	**Ade Syahputra**	**202410715084**	**ML Engineer / Python Developer:** Membangun *pipeline* kode Python pada Google Colab, melakukan *Feature Scaling* (StandardScaler), serta melatih model *Random Forest* dan *SVM*.
3	**Ryan Ali Sunaedi**	**202410715266**	**Technical Writer & Visualizer:** Menggenerasi visualisasi matriks evaluasi (*Confusion Matrix & ROC Curve*) serta menyusun dokumen Laporan Teknis ETS secara menyeluruh.