Nama : Erdianti Wiga Putri Andini

NIM : 13522053

Tugas 3 - EB3204 Pembelajaran Mesin dalam Teknik Biomedis

Analisis No 1

Lakukan pendekatan preprocessing data yang anda anggap perlu sesuai Chapter 4, terutama utk kolom yang non atomic. Sebutkan strategi anda dalam preprocessing data ini.

Jawaban:

Dalam preprocessing data ini, dilakukan beberapa tahapan:

1. **Data Cleaning**

   Langkah pertama adalah menangani nilai yang hilang (missing values). Untuk kolom numerik, digunakan strategi imputasi median agar tidak terpengaruh outlier. Sedangkan untuk kolom kategorikal, digunakan imputasi dengan nilai yang paling sering muncul (most frequent). Setelah proses ini, seluruh nilai kosong berhasil diatasi (missing values = 0).

2. **Feature Engineering**

   Kolom Komponen Nilai, yang merupakan kolom non-atomic (berisi lebih dari satu informasi dalam satu sel), diproses menjadi beberapa fitur baru:

   - Ada_PR, Ada_Kuis, Ada_Tubes, Ada_UTS, dan Ada_UAS menunjukkan keberadaan komponen tersebut dalam nilai.
   - Jumlah_Komponen_Nilai menghitung total jumlah komponen dalam satu mata kuliah.

   Hal ini membuat informasi dalam kolom tersebut menjadi lebih terstruktur dan bisa diproses oleh model.

3. **Feature Transformation**

   - One-hot encoding diterapkan untuk semua kolom kategorikal, termasuk Pelaksanaan kuliah dan kolom object lainnya, untuk mengubah data kategorikal menjadi representasi numerik.
   - Standarisasi (StandardScaler) diterapkan untuk seluruh fitur numerik agar berada dalam skala yang sama, yang penting untuk algoritma seperti SVM dan KNN.

4. **Feature Selection**

   Beberapa kolom yang tidak relevan dengan proses prediksi, seperti Nama Mata Kuliah dan Nilai (yang redundant terhadap Nilai_Numeric), dihapus dari dataset. Ini bertujuan untuk menghindari kebocoran data dan menjaga relevansi fitur.

5. **Split Data**

   Data akhir memiliki bentuk (815, 47) untuk fitur dan (815,) untuk target. Data kemudian dibagi menjadi data latih dan data uji dengan rasio 70:30. Ini memastikan bahwa model dilatih dan dievaluasi secara adil.

Analisis No 2

Terapkan berbagai pendekatan machine learning pada chapter 3 terhadap seluruh kolom setelah preprocessing (tidak perlu plot data region).

Jawaban:

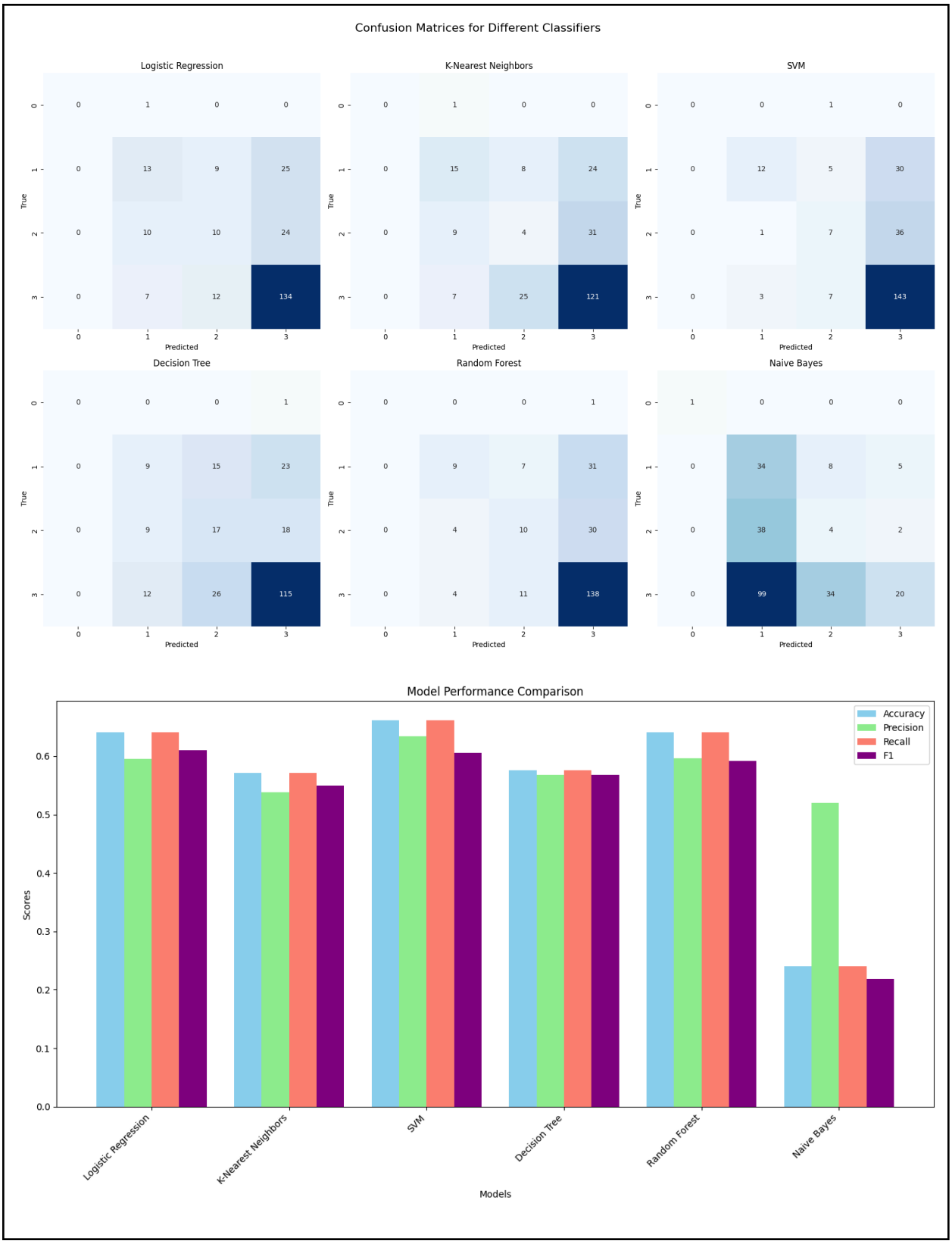
Pada eksperimen ini, telah diterapkan enam algoritma machine learning, yaitu **Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest,** dan **Naive Bayes**. Seluruh model diuji menggunakan data yang telah diproses sebelumnya.Hasil terbaik diperoleh oleh model **SVM** dengan nilai akurasi sebesar **66.12%**, serta skor precision, recall, dan F1 yang cukup baik dibandingkan model lainnya. Hal ini menunjukkan bahwa SVM mampu melakukan klasifikasi dengan cukup baik dan konsisten.

Model **Logistic Regression** dan **Random Forest** juga menunjukkan performa yang cukup baik, masing-masing dengan akurasi **64.08%**. Kedua model ini memberikan hasil evaluasi yang seimbang pada metrik lainnya.

Model **KNN** dan **Decision Tree** memiliki performa yang lebih rendah, dengan akurasi sekitar **57%**. Nilai precision, recall, dan F1-nya juga masih bisa ditingkatkan.

Sementara itu, model **Naive Bayes** menunjukkan performa paling rendah, dengan akurasi hanya **24.08%**. Meskipun nilai precision-nya terlihat tinggi, hal ini tidak mencerminkan kinerja keseluruhan karena distribusi prediksi yang tidak seimbang.

Secara umum, hasil ini menunjukkan bahwa pemilihan algoritma sangat mempengaruhi performa model. Selain itu, pendekatan lanjutan seperti **penyeimbangan kelas**, **pemilihan fitur yang lebih relevan**, dan **tuning hyperparameter** berpotensi meningkatkan hasil model secara keseluruhan.


Confusion Matrices for Different Classifiers
Logistic Regression
K-Nearest Neighbors
SVM
Decision Tree
Random Forest
Naive Bayes
True
Predicted
Model Performance Comparison
Accuracy
Precision
Recall
F1
Scores
Models
Logistic Regression
K-Nearest Neighbors
SVM
Decision Tree
Random Forest
Naive Bayes

Analisis No 3

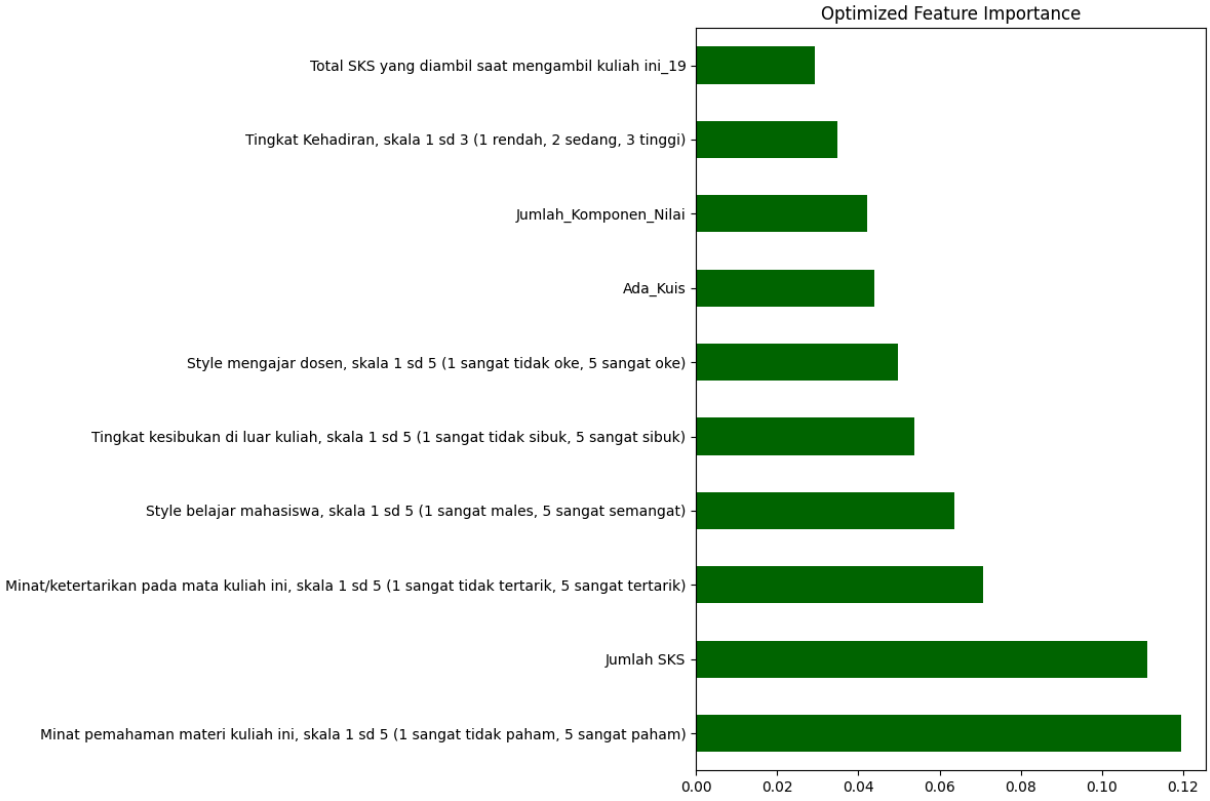
Tentukan fitur input terpenting dengan SBS dan Random Forest. Dari kedua hasil tersebut, buat Tree nya dan lihat decision tree nya. Bagaimana menurut anda?

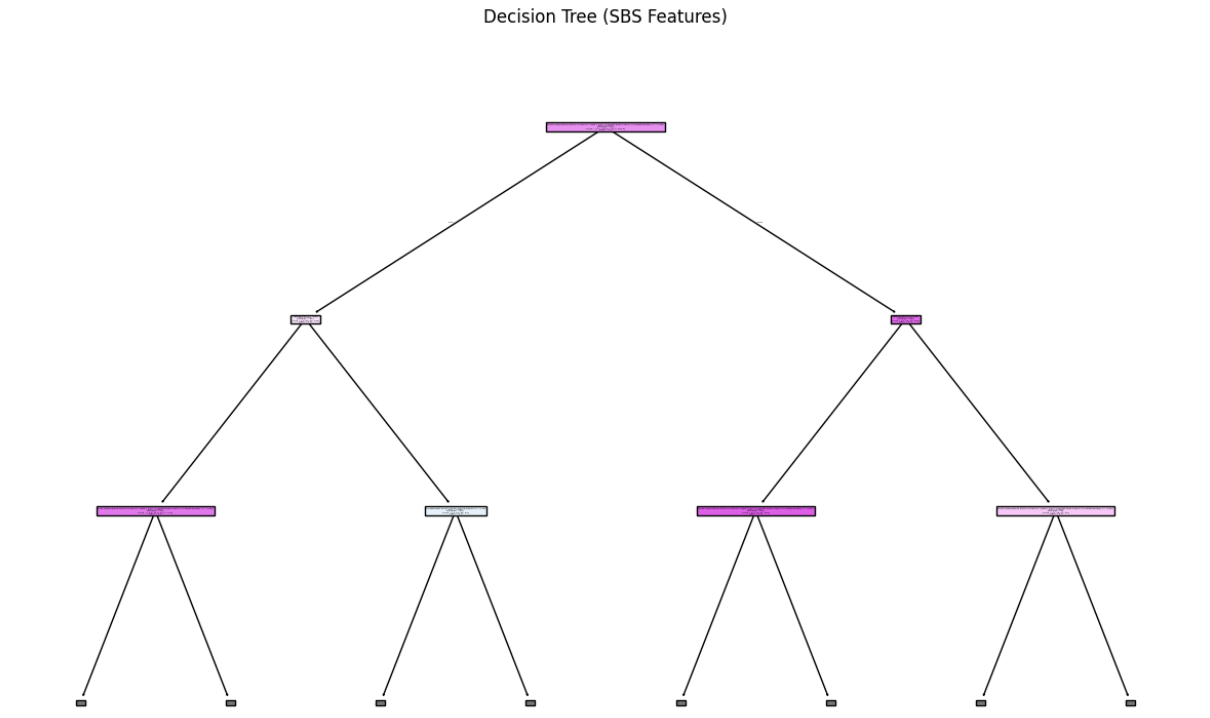
Jawaban:

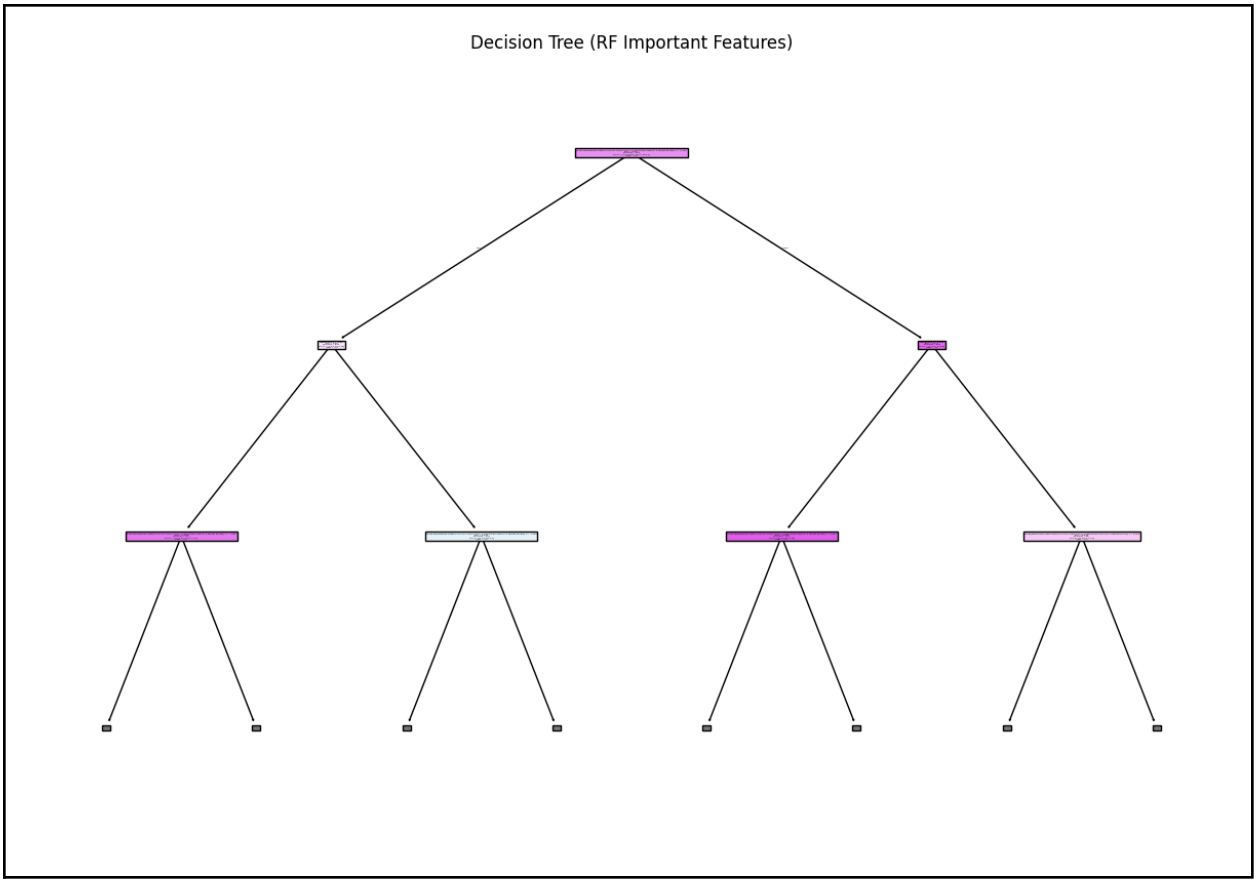
Dua pendekatan telah digunakan untuk menentukan fitur paling berpengaruh dalam model klasifikasi, yaitu **Sequential Backward Selection (SBS)** dan **Random Forest**. Dari metode SBS, lima fitur utama yang teridentifikasi adalah: **jumlah SKS**, **minat pemahaman materi kuliah**, **gaya mengajar dosen**, **pelaksanaan kuliah (luring/daring sinkron)**, serta **total SKS yang diambil**. Sementara itu, Random Forest menunjukkan lima fitur dengan nilai *importance* tertinggi, yaitu: **jumlah SKS**, **minat pemahaman materi kuliah**, **minat atau ketertarikan pada mata kuliah**, **gaya belajar mahasiswa**, dan **tingkat kesibukan di luar kuliah**. Menariknya, dua fitur yaitu **jumlah SKS** dan **minat pemahaman materi kuliah**, muncul **konsisten di kedua metode**, menunjukkan bahwa keduanya **sangat berpengaruh** terhadap label yang diprediksi.

Model **decision tree** kemudian dibangun berdasarkan masing-masing kumpulan fitur. Dengan fitur hasil seleksi **SBS**, **akurasi model mencapai 68,59%**. Model ini cenderung lebih sederhana dan fokus pada aspek-aspek yang **berkaitan langsung dengan mata kuliah**, seperti metode pengajaran, pelaksanaan perkuliahan, serta persepsi mahasiswa terhadap pemahaman dan kualitas dosen. Di sisi lain, model berbasis fitur dari **Random Forest** menghasilkan **akurasi sebesar 67,98%**, dengan cakupan fitur yang lebih luas mencakup **faktor pribadi mahasiswa** seperti minat umum dan tingkat kesibukan.

Secara keseluruhan, **metode SBS memberikan sedikit keunggulan dalam akurasi** dibandingkan Random Forest. Kombinasi dari fitur-fitur penting pada kedua metode **dapat dipertimbangkan untuk membangun model yang lebih kompleks dan akurat**.


Optimized Feature Importance
Total SKS yang diambil saat mengambil kuliah ini_19
Tingkat Kehadiran, skala 1 sd 3 (1 rendah, 2 sedang, 3 tinggi)
Jumlah_Komponen_Nilai
Ada_Kuis
Style mengajar dosen, skala 1 sd 5 (1 sangat tidak oke, 5 sangat oke)
Tingkat kesibukan di luar kuliah, skala 1 sd 5 (1 sangat tidak sibuk, 5 sangat sibuk)
Style belajar mahasiswa, skala 1 sd 5 (1 sangat males, 5 sangat semangat)
Minat/ketertarikan pada mata kuliah ini, skala 1 sd 5 (1 sangat tidak tertarik, 5 sangat tertarik)
Jumlah SKS
Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sangat tidak paham, 5 sangat paham)
0.00
0.02
0.04
0.06
0.08
0.10
0.12


Decision Tree (SBS Features)



Analisis No 4

Lakukan reduksi dimensi dengan PCA atau LDA dan terapkan berbagai pendekatan machine learning pada chapter 3 terhadap dua fitur hasil transformasi, dan plot decision regionnya. Apakah ada peningkatan akurasi ? Bagaimana analisis anda?

Jawaban:

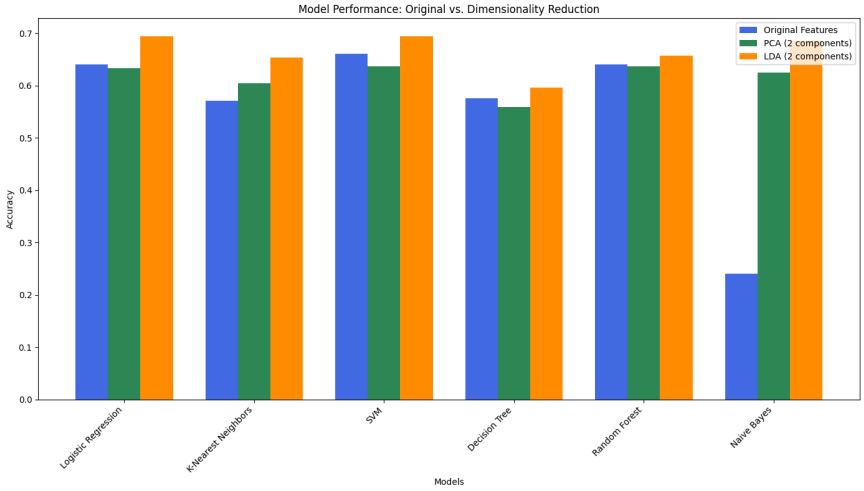
Reduksi dimensi menggunakan **PCA** hanya mempertahankan **13.14% informasi dari data asli** (PC1: 6.64%, PC2: 6.50%), sehingga berdampak negatif pada kinerja model. **Akurasi tertinggi** yang dicapai adalah **63.67% oleh SVM** dan **Random Forest**, tetapi precision yang rendah (**45.81% untuk SVM**) menunjukkan **banyak prediksi salah. F1 Score yang rendah (52.97% untuk Logistic Regression)** juga mengindikasikan ketidakseimbangan antara precision dan recall.

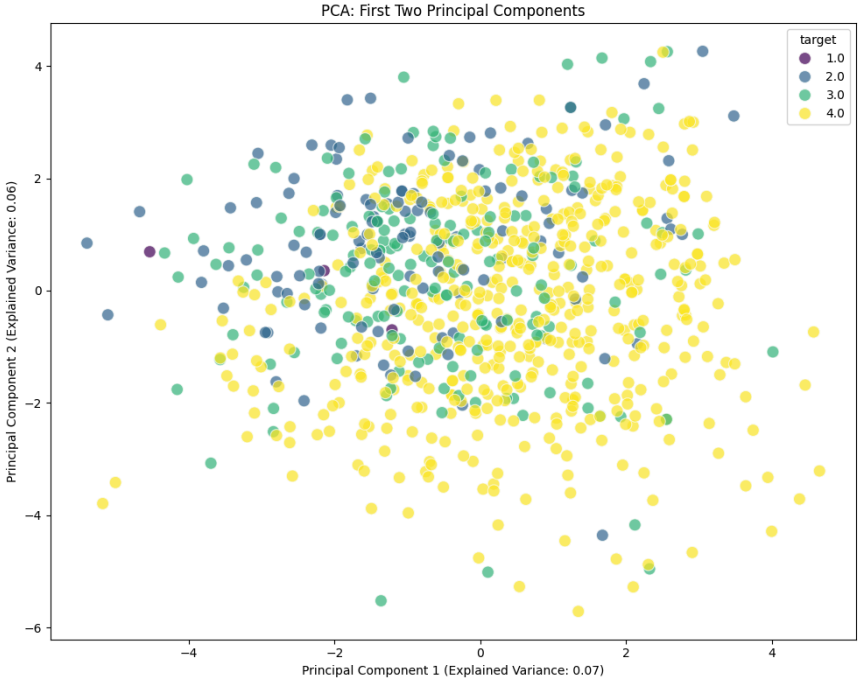
Sebaliknya, **LDA terbukti lebih efektif** karena fokus pada pemisahan kelas. Model seperti **Logistic Regression dan SVM mencapai akurasi 69.39%**, lebih tinggi dibandingkan PCA, dengan **F1 Score lebih baik (66.39%)**, menandakan klasifikasi yang lebih stabil. Visualisasi decision region pada LDA juga menunjukkan pemisahan kelas yang lebih jelas dibandingkan PCA, yang cenderung noise akibat hilangnya informasi.

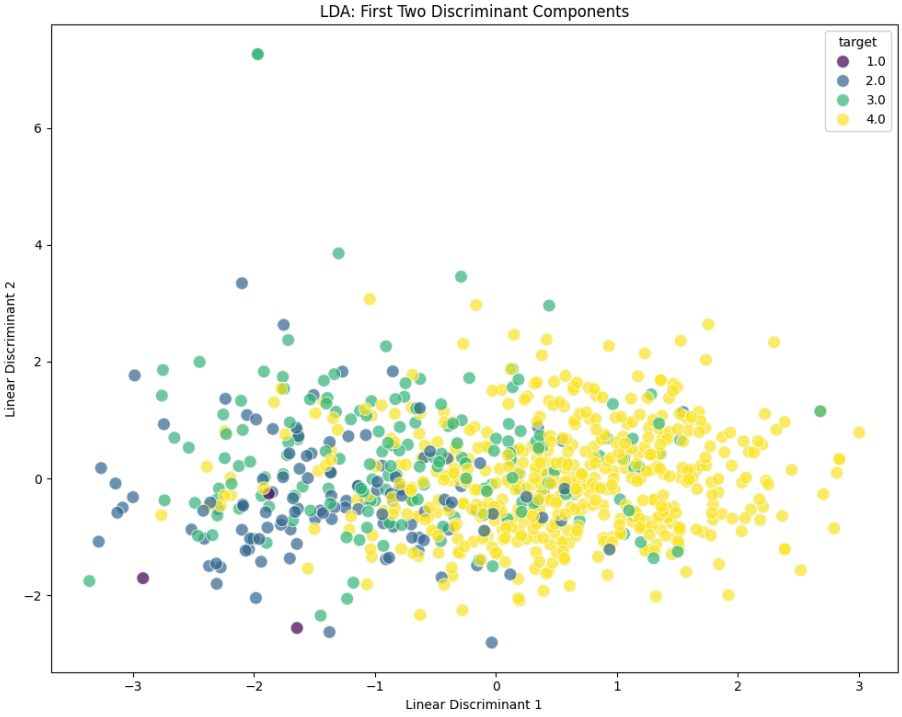
Kesimpulan:

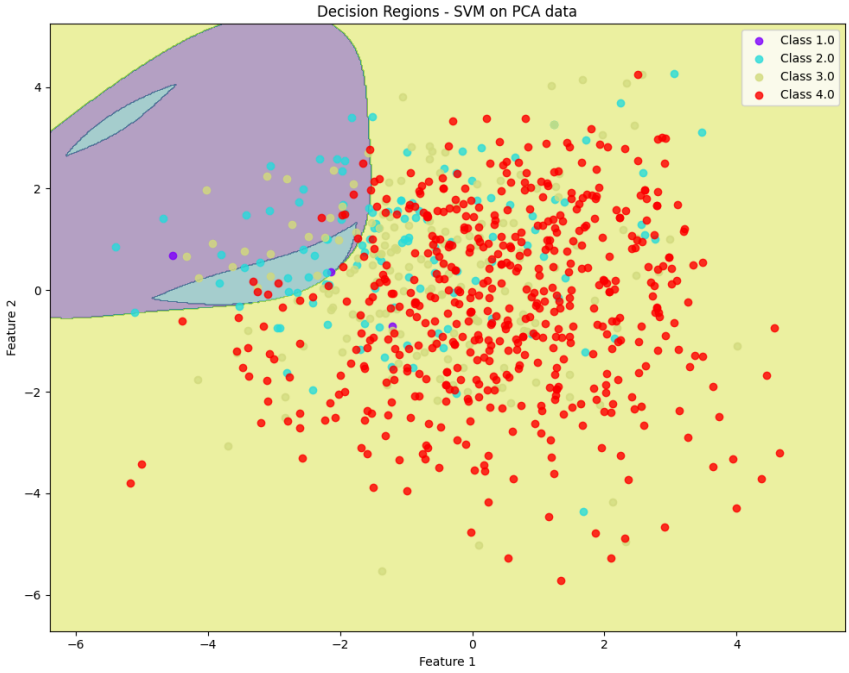
- LDA meningkatkan akurasi dan kualitas klasifikasi untuk data ini.
- PCA dengan 2 komponen tidak direkomendasikan karena explained variance terlalu rendah, kecuali jumlah komponen ditambah.
- Decision region LDA lebih terdefinisi, mendukung penggunaan reduksi dimensi berbasis diskriminasi untuk tugas klasifikasi.

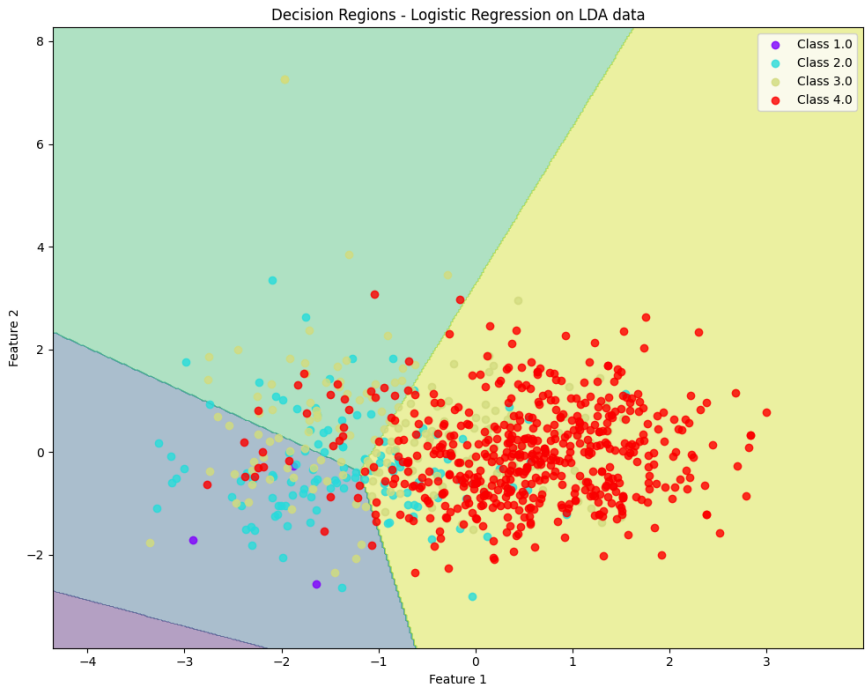
LDA diprioritaskan jika tujuan utama adalah klasifikasi, dan hindari PCA dengan explained variance di bawah 80% tanpa optimasi lebih lanjut.






LDA: First Two Discriminant Components
target
1.0
2.0
3.0
4.0
Linear Discriminant 2
Linear Discriminant 1
6
4
2
0
−2
−3
−2
−1
0
1
2
3


Decision Regions - SVM on PCA data
Class 1.0
Class 2.0
Class 3.0
Class 4.0
Feature 2
Feature 1
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4



Analisis No 5

Lakukan pengecekan korelasi seperti bab 2 buku Aurelien, apakah sama dengan poin no 3? kalua, berbeda buat Tree nya, apakah ada peningkatan akurasi?

Jawaban:

Berdasarkan analisis korelasi, ditemukan bahwa fitur seperti **Minat Pemahaman Materi** (korelasi: 0.40) dan **Minat/Ketertarikan terhadap Mata Kuliah** (korelasi: 0.28) memiliki hubungan yang cukup kuat secara linear terhadap variabel target. Namun, ketika model **Decision Tree** dibangun hanya menggunakan fitur-fitur yang memiliki korelasi tinggi ini, akurasi yang dicapai hanya sebesar **65.03%**.

Sebaliknya, metode seleksi fitur berbasis model seperti **Sequential Backward Selection (SBS)** dan **Random Forest (RF)**, yang digunakan pada tahap berikutnya, berhasil mengidentifikasi fitur berbeda—di antaranya **Jumlah SKS** dan **Pelaksanaan Kuliah (Luring/Daring Sinkron)**. Fitur-fitur ini sebelumnya tidak terdeteksi sebagai signifikan oleh analisis korelasi sederhana. Dengan kombinasi fitur hasil SBS dan RF, akurasi model meningkat menjadi:

- **68.59%** (menggunakan fitur hasil SBS)
- **67.98%** (menggunakan fitur hasil Random Forest)

Kenaikan akurasi sekitar **3%** ini menunjukkan bahwa pemilihan fitur melalui pendekatan berbasis model lebih efektif untuk meningkatkan performa prediktif.

Perbedaan hasil antara analisis korelasi dan seleksi fitur berbasis model disebabkan oleh:

1. **Keterbatasan Korelasi Linear**

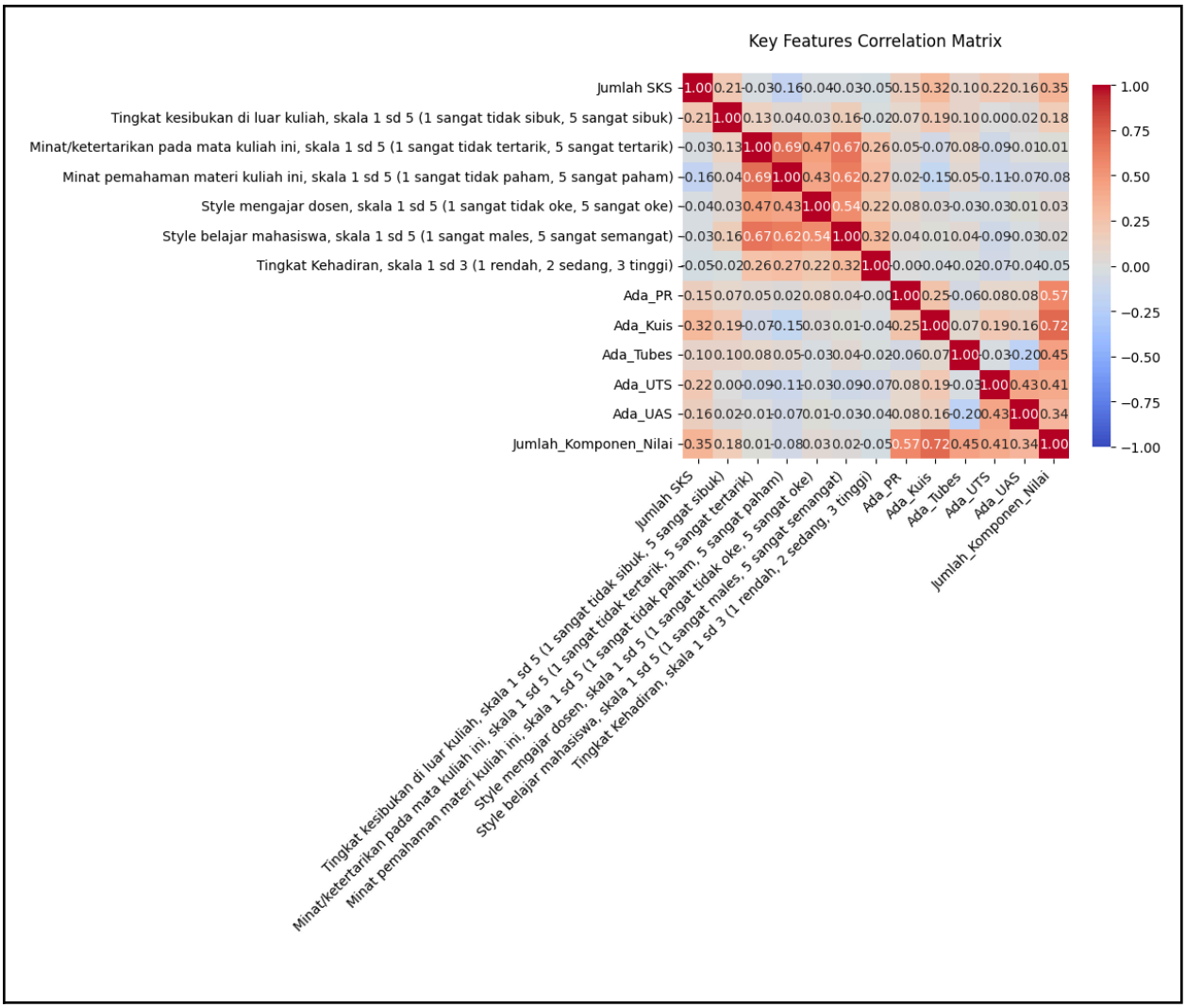
   Analisis korelasi hanya mengukur hubungan linier antara dua variabel. Fitur yang memiliki pengaruh non-linear atau interaksi kompleks antar fitur tidak akan terdeteksi sebagai signifikan oleh metode ini.

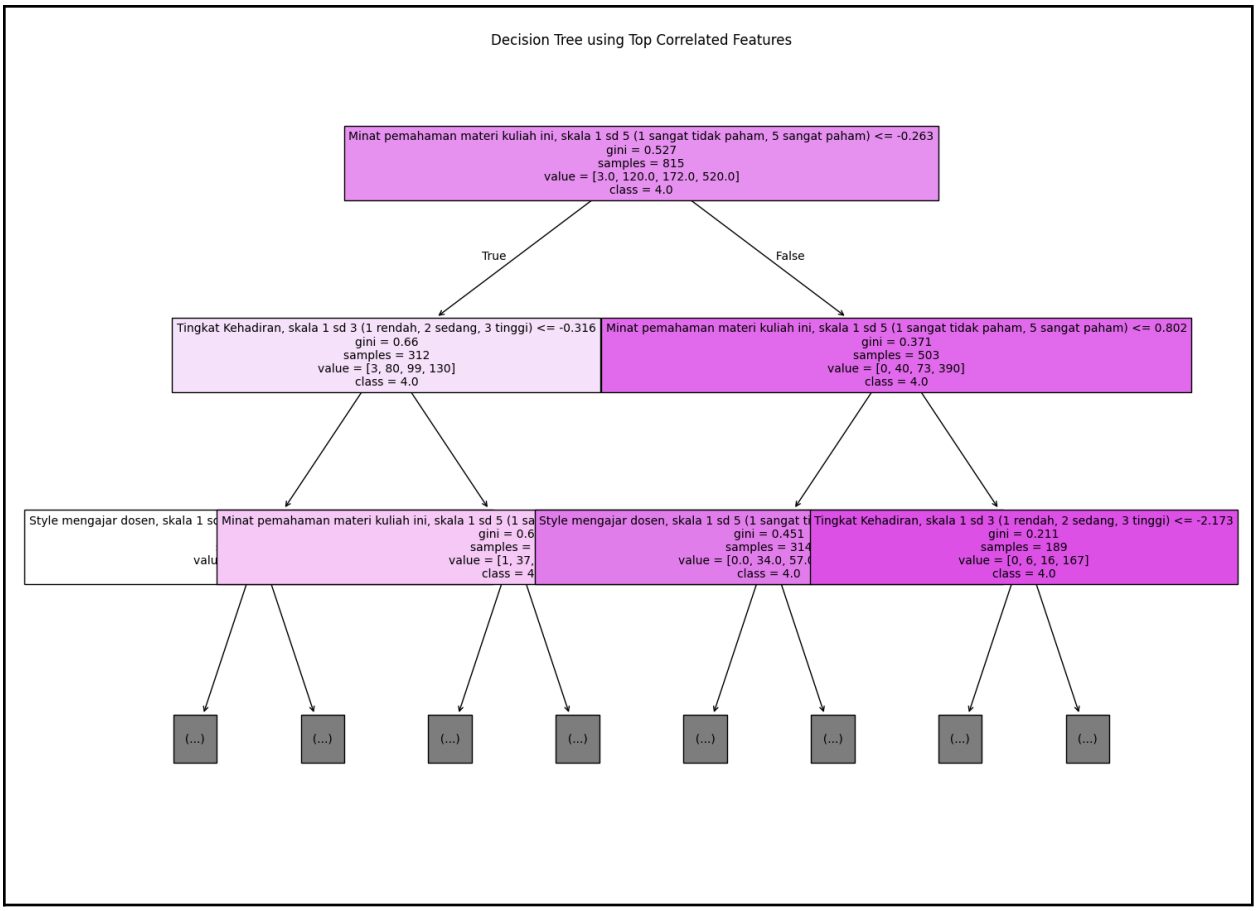
2. **Evaluasi Berbasis Kinerja Model**

   Metode SBS dan Random Forest mengevaluasi kontribusi masing-masing fitur secara langsung terhadap kinerja model prediktif. Hal ini memungkinkan identifikasi fitur-fitur yang mungkin tidak terlihat signifikan secara statistik, tetapi penting secara fungsional dalam konteks prediksi.

Kesimpulan:

- Analisis korelasi tetap berguna sebagai langkah awal eksplorasi data untuk memahami hubungan antar variabel.
- Namun, untuk keperluan seleksi fitur yang optimal dalam model prediktif, pendekatan berbasis model seperti SBS dan Random Forest **lebih disarankan**.
- Peningkatan akurasi sebesar 3% membuktikan bahwa kombinasi fitur yang tepat, termasuk fitur dengan pengaruh tidak langsung seperti **Jumlah SKS** dan **konteks pelaksanaan kuliah**, memiliki dampak signifikan terhadap performa model.


Key Features Correlation Matrix
Jumlah SKS
Tingkat kesibukan di luar kuliah, skala 1 sd 5 (1 sangat tidak sibuk, 5 sangat sibuk)
Minat/ketertarikan pada mata kuliah ini, skala 1 sd 5 (1 sangat tidak tertarik, 5 sangat tertarik)
Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sangat tidak paham, 5 sangat paham)
Style mengajar dosen, skala 1 sd 5 (1 sangat tidak oke, 5 sangat oke)
Style belajar mahasiswa, skala 1 sd 5 (1 sangat males, 5 sangat semangat)
Tingkat Kehadiran, skala 1 sd 3 (1 rendah, 2 sedang, 3 tinggi)
Ada_PR
Ada_Kuis
Ada_Tubes
Ada_UTS
Ada_UAS
Jumlah_Komponen_Nilai
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Decision Tree using Top Correlated Features
Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sangat tidak paham, 5 sangat paham) <= -0.263
gini = 0.527
samples = 815
value = [3.0, 120.0, 172.0, 520.0]
class = 4.0
True
False
Tingkat Kehadiran, skala 1 sd 3 (1 rendah, 2 sedang, 3 tinggi) <= -0.316
gini = 0.66
samples = 312
value = [3, 80, 99, 130]
class = 4.0
Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sangat tidak paham, 5 sangat paham) <= 0.802
gini = 0.371
samples = 503
value = [0, 40, 73, 390]
class = 4.0
Style mengajar dosen, skala 1 sd
valu
Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sa
gini = 0.6
samples =
value = [1, 37,
class = 4
Style mengajar dosen, skala 1 sd 5 (1 sangat t
gini = 0.451
samples = 314
value = [0.0, 34.0, 57.
class = 4.0
Tingkat Kehadiran, skala 1 sd 3 (1 rendah, 2 sedang, 3 tinggi) <= -2.173
gini = 0.211
samples = 189
value = [0, 6, 16, 167]
class = 4.0
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)