

Nama : Erdianti Wiga Putri Andini

NIM : 13522053

## **Tugas 2 - EB3204 Pembelajaran Mesin dalam Teknik Biomedis**

### **Analisis No 1**

Pilih kira-kira 2 buah input yang menurut anda paling penting dan lakukan semua metode machine learning di ch03 terhadap data di atas. Buat analisa anda, mana yang metode yang menurut anda paling tepat?

#### **Jawaban:**

Input yang dipilih adalah:

- Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sangat tidak paham, 5 sangat paham)
- Minat/ketertarikan pada mata kuliah ini, skala 1 sd 5 (1 sangat tidak tertarik, 5 sangat tertarik)

Alasan memilih input ini adalah karena pada correlation matrix di atas, korelasi minat pemahaman materi terhadap nilai\_numeric paling tinggi yaitu 0.40. Diikuti dengan korelasi minat/ketertarikan terhadap nilai\_numeric sebesar 0.28.

Model yang dicoba pada kasus ini adalah Perceptron, Logistic Regression, Linear SVM, RBF SVM with different gamma values, Decision Tree, Random Forest, dan K-Nearest Neighbors. Berdasarkan model yang menggunakan 2 input tersebut, didapat result setiap model sebagai berikut.

#### **Perceptron Results:**

Accuracy: 0.5951

F1 Score: 0.5686

5-Fold CV Accuracy: 0.5276

Misclassified samples: 66

#### **Logistic Regression Results:**

Accuracy: 0.6442

F1 Score: 0.5425

5-Fold CV Accuracy: 0.6442

Misclassified samples: 58

**Linear SVM Results:**

Accuracy: 0.6442

F1 Score: 0.5048

5-Fold CV Accuracy: 0.6380

Misclassified samples: 58

**RBF SVM (gamma=0.1) Results:**

Accuracy: 0.6442

F1 Score: 0.5048

5-Fold CV Accuracy: 0.6368

Misclassified samples: 58

**RBF SVM (gamma=1) Results:**

Accuracy: 0.6380

F1 Score: 0.5492

5-Fold CV Accuracy: 0.6380

Misclassified samples: 59

**Decision Tree Results:**

Accuracy: 0.6319

F1 Score: 0.5421

5-Fold CV Accuracy: 0.6344

Misclassified samples: 60

**Random Forest Results:**

Accuracy: 0.6380

F1 Score: 0.5484

5-Fold CV Accuracy: 0.6442

Misclassified samples: 59

### K-Nearest Neighbors (KNN) Results:

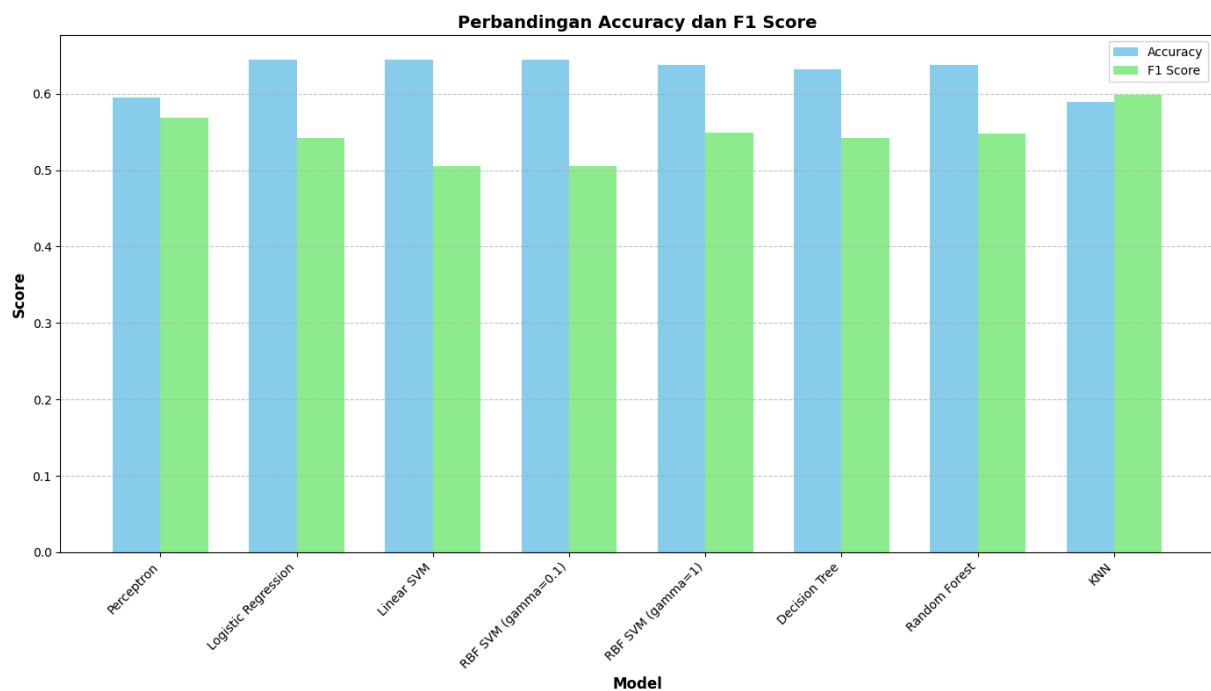
Accuracy: 0.5890

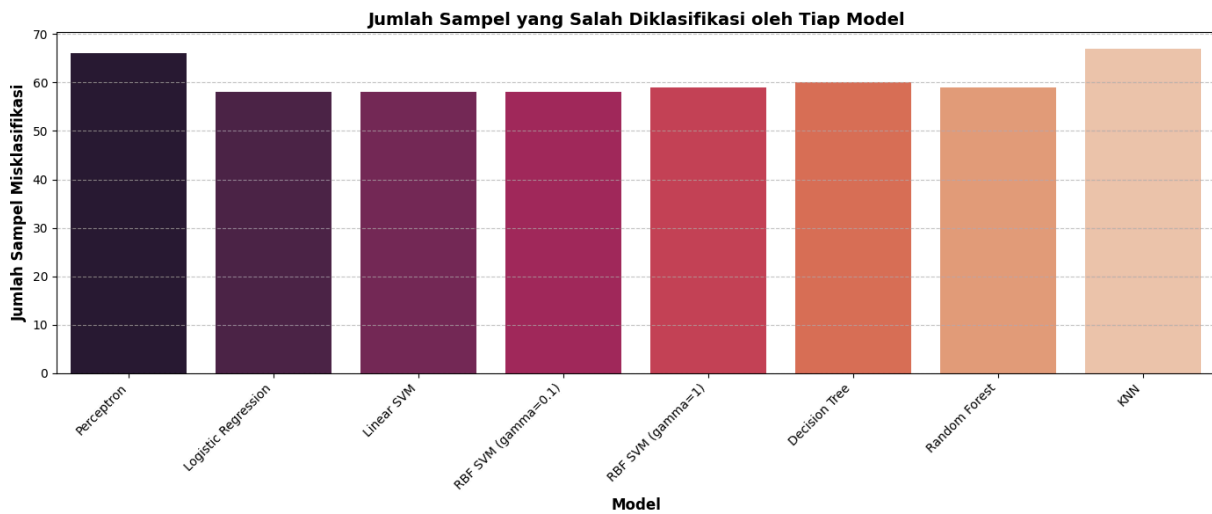
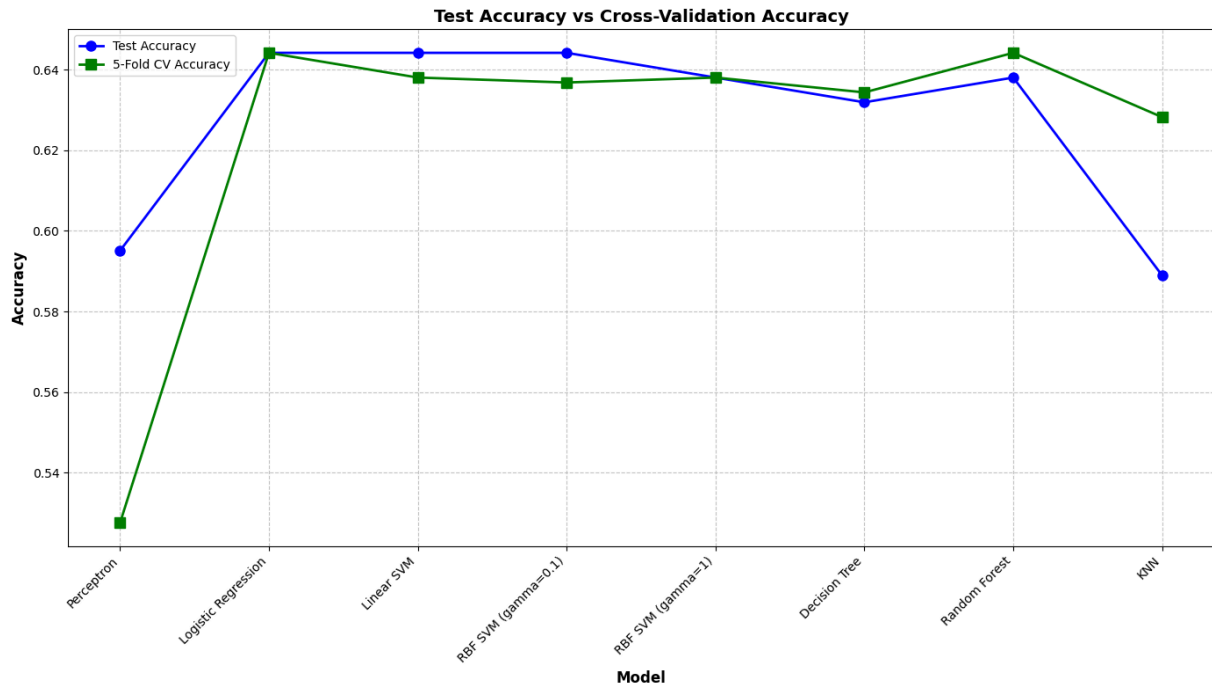
F1 Score: 0.5990

5-Fold CV Accuracy: 0.6282

Misclassified samples: 67

Berdasarkan hasil accuracy, F1 score, CV accuracy, dan misclassified samples, berikut adalah hasil analisis dan perbandingan model.





Peringkat Keseluruhan Model (berdasarkan gabungan semua metrik):

	Model	Accuracy	F1 Score	5-Fold CV Accuracy	Misclassified Samples	Accuracy Rank	F1 Rank	CV Rank	Misclassified Rank	Overall Rank
1	Logistic Regression	0.644172	0.542460	0.644172	58	2.0	5.0	1.5	2.0	2.625
6	Random Forest	0.638037	0.548357	0.644172	59	4.5	4.0	1.5	4.5	3.625
4	RBF SVM (gamma=1)	0.638037	0.549225	0.638037	59	4.5	3.0	3.0	4.5	3.750
2	Linear SVM	0.644172	0.504761	0.638037	58	2.0	7.5	4.0	2.0	3.875
3	RBF SVM (gamma=0.1)	0.644172	0.504761	0.636810	58	2.0	7.5	5.0	2.0	4.125
0	Perceptron	0.595092	0.568645	0.527607	66	7.0	2.0	8.0	7.0	6.000
5	Decision Tree	0.631902	0.542143	0.634356	60	6.0	6.0	6.0	6.0	6.000
7	KNN	0.588957	0.598970	0.628221	67	8.0	1.0	7.0	8.0	6.000

**Logistic Regression:** Model ini menunjukkan performa terbaik secara keseluruhan dengan overall rank 2.625. Memiliki akurasi tertinggi (0.644172) yang konsisten dengan CV accuracy-nya, serta jumlah sampel misklasifikasi terendah (58). Meskipun F1 Score-nya (0.542460) tidak setinggi beberapa model lain, Logistic Regression menawarkan keseimbangan yang baik antar semua metrik dan memiliki keunggulan dalam hal interpretabilitas model.

**Random Forest:** Model ini menduduki posisi kedua terbaik dengan overall rank 3.625 serta memiliki akurasi yang cukup baik (0.638037), F1 Score yang relatif tinggi (0.548357), dan CV accuracy tertinggi (0.644172). Dengan 59 sampel misklasifikasi, Random Forest memberikan alternatif yang solid jika ingin sedikit mengorbankan akurasi demi F1 Score yang lebih baik, terutama untuk data dengan kelas tidak seimbang.

**RBF SVM (gamma=1):** Model ini berada di posisi ketiga dengan overall rank 3.750 serta memiliki akurasi menengah (0.638037), F1 Score cukup baik (0.549225), CV accuracy menengah (0.638037), dan 59 sampel misklasifikasi. Keunggulannya adalah F1 Score yang relatif tinggi, namun CV accuracy-nya tidak setinggi Logistic Regression.

**Linear SVM:** Menduduki posisi keempat dengan overall rank 3.875. Memiliki akurasi tinggi (0.644172) dan sampel misklasifikasi rendah (58), tetapi F1 Score-nya relatif rendah (0.504761) dan CV accuracy-nya menengah (0.638037). Model ini baik untuk memaksimalkan akurasi, namun perlu dipertimbangkan jika keseimbangan antara presisi dan recall (F1 Score) penting.

**RBF SVM (gamma=0.1):** Berada di posisi kelima dengan overall rank 4.125. Model ini memiliki akurasi tinggi (0.644172) dan sampel misklasifikasi rendah (58), tetapi F1 Score-nya rendah (0.504761) dan CV accuracy-nya menengah (0.636810). Performa model ini mirip dengan Linear SVM, tetapi dengan generalisasi yang sedikit lebih rendah.

**Perceptron:** Menduduki posisi terendah dengan overall rank 6.000. Meskipun memiliki F1 Score cukup tinggi (0.568645), model ini memiliki akurasi rendah (0.595092), CV accuracy terendah (0.527607), dan sampel misklasifikasi tinggi (66). Perceptron tidak direkomendasikan untuk kasus ini karena stabilitas dan generalisasinya rendah.

**Decision Tree:** Sama-sama di peringkat bawah dengan overall rank 6.000. Model ini menunjukkan performa seimbang di semua metrik dengan akurasi menengah (0.631902), F1 Score menengah (0.542143), CV accuracy menengah (0.634356), dan 60 sampel misklasifikasi. Meskipun mudah diinterpretasi, Decision Tree tidak unggul di metrik manapun.

**KNN:** Juga menduduki posisi terendah dengan overall rank 6.000. Meskipun memiliki F1 Score tertinggi (0.598970), model ini memiliki akurasi rendah (0.588957), CV accuracy menengah (0.628221), dan sampel misklasifikasi tertinggi (67). KNN bisa dipertimbangkan jika F1 Score menjadi prioritas utama, tetapi perlu dicatat bahwa akurasi dan jumlah kesalahan klasifikasinya kurang optimal.

### Kesimpulan

Berdasarkan analisis di atas, **Logistic Regression** adalah **model yang paling tepat** untuk kasus ini. Model ini menawarkan keseimbangan terbaik antara akurasi prediksi, F1 score, stabilitas cross-validation, dan minimalisasi kesalahan klasifikasi. Selain itu, Logistic Regression juga lebih mudah diinterpretasi, memberikan keuntungan tambahan untuk memahami faktor-faktor yang mempengaruhi prediksi.

### Analisis No 2

Buat machine learning dengan tree tapi kali ini pakai seluruh input. Berdasarkan hasil tersebut mana kira-kira input yang paling penting?

#### Jawaban:

Input yang paling penting berdasarkan decision tree adalah:

##### 1. Jumlah SKS

- **Importance** : 0.123027 (tertinggi)
- **Interpretasi** : Fitur ini paling berpengaruh dalam model Decision Tree. Jumlah SKS mencerminkan beban studi suatu mata kuliah. Nilai importance tertinggi menunjukkan bahwa semakin banyak SKS, semakin signifikan dampaknya terhadap hasil pembelajaran. Jumlah SKS yang diambil kemungkinan besar berhubungan erat dengan target, seperti nilai akhir atau kepuasan belajar. Mata kuliah dengan SKS tinggi biasanya membutuhkan lebih banyak waktu, tugas, atau kedalaman materi, sehingga mempengaruhi pemahaman dan performa mahasiswa. Beban SKS juga bisa berkaitan dengan tingkat kesulitan atau prioritas mahasiswa dalam mengalokasikan waktu belajar.

##### 2. Minat Pemahaman Materi Kuliah (Skala 1-5)

- **Importance** : 0.103128 (tertinggi kedua)
- **Interpretasi** : Skala minat atau ketertarikan mahasiswa terhadap materi kuliah (1-5) menjadi faktor kedua terpenting. Pemahaman materi kuliah sangat berpengaruh pada prediksi. Mahasiswa yang lebih paham cenderung memiliki hasil lebih baik. Minat yang tinggi cenderung meningkatkan motivasi belajar, partisipasi, dan retensi pemahaman.

### Decision Tree (All Features) Results:

Accuracy: 0.5890

F1 Score: 0.5833

Mean Squared Error: 0.7914

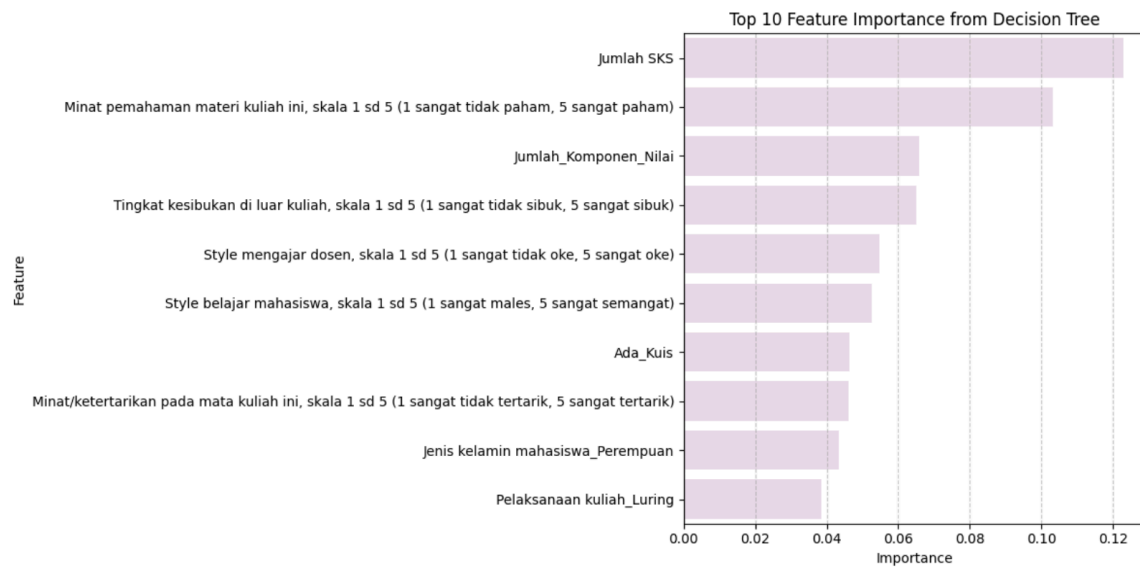
R<sup>2</sup> Score: -0.2356

5-Fold CV Accuracy: 0.5497

Misclassified samples: 67

### Feature Importance DT:

	Feature	Importance
0	Jumlah SKS	0.123027
3	Minat pemahaman materi kuliah ini, skala 1 sd ...	0.103128
12	Jumlah_Komponen_Nilai	0.065851
1	Tingkat kesibukan di luar kuliah, skala 1 sd 5...	0.064952
4	Style mengajar dosen, skala 1 sd 5 (1 sangat t...	0.054613
5	Style belajar mahasiswa, skala 1 sd 5 (1 sanga...	0.052554
8	Ada_Kuis	0.046269
2	Minat/ketertarikan pada mata kuliah ini, skala...	0.046127
20	Jenis kelamin mahasiswa_Perempuan	0.043465
16	Pelaksanaan kuliah_Luring	0.038438



### Analisis No 3

Apabila 2 input paling penting dari soal no 2 berbeda dengan yang anda pilih di no 1, lakukan kembali no 1 dengan 2 input terpenting berdasarkan hasil soal nomor 2. Bagaimana hasil analisa anda?

#### **Jawaban:**

Dari hasil training feature importance pada no 2, didapatkan dua input terpenting adalah 'Jumlah SKS' dan 'Minat pemahaman materi kuliah ini, skala 1 sd 5 (1 sangat tidak paham, 5 sangat paham)'. Setelah menggunakan input dari no 2 untuk metode no 1, didapat result sebagai berikut.

#### **Perceptron Results:**

Accuracy: 0.6503

F1 Score: 0.6270

5-Fold CV Accuracy: 0.5926

Misclassified samples: 57

#### **Logistic Regression Results:**

Accuracy: 0.6319

F1 Score: 0.5407

5-Fold CV Accuracy: 0.6528

Misclassified samples: 60

#### **Linear SVM Results:**

Accuracy: 0.6380

F1 Score: 0.5475

5-Fold CV Accuracy: 0.6577

Misclassified samples: 59

#### **RBF SVM (gamma=0.1) Results:**

Accuracy: 0.6319

F1 Score: 0.5407



5-Fold CV Accuracy: 0.6601

Misclassified samples: 60

**RBF SVM (gamma=1) Results:**

Accuracy: 0.6503

F1 Score: 0.6072

5-Fold CV Accuracy: 0.6564

Misclassified samples: 57

**Decision Tree Results:**

Accuracy: 0.6503

F1 Score: 0.6072

5-Fold CV Accuracy: 0.6528

Misclassified samples: 57

**Random Forest Results:**

Accuracy: 0.6503

F1 Score: 0.6072

5-Fold CV Accuracy: 0.6564

Misclassified samples: 57

**K-Nearest Neighbors (KNN) Results:**

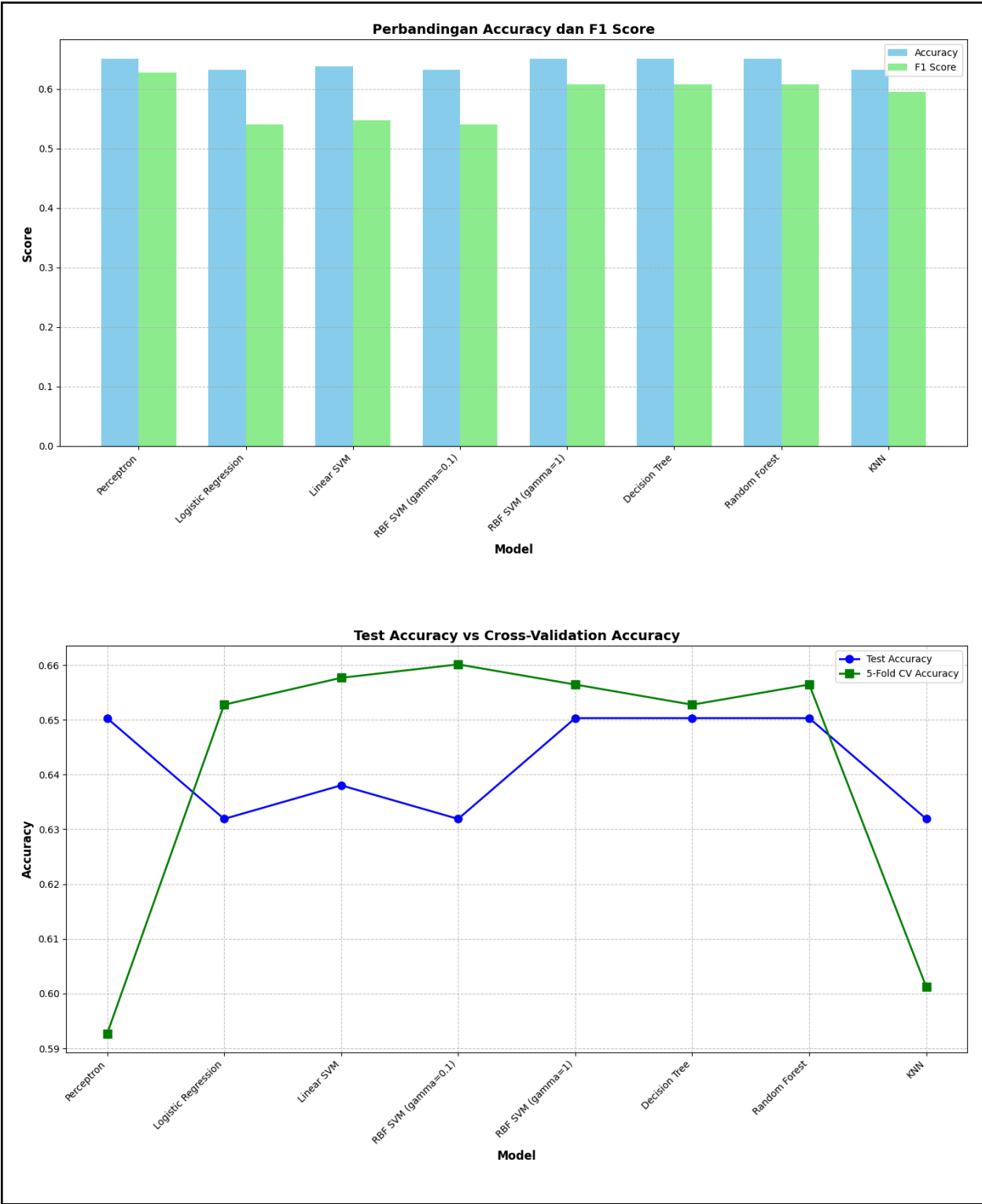
Accuracy: 0.6319

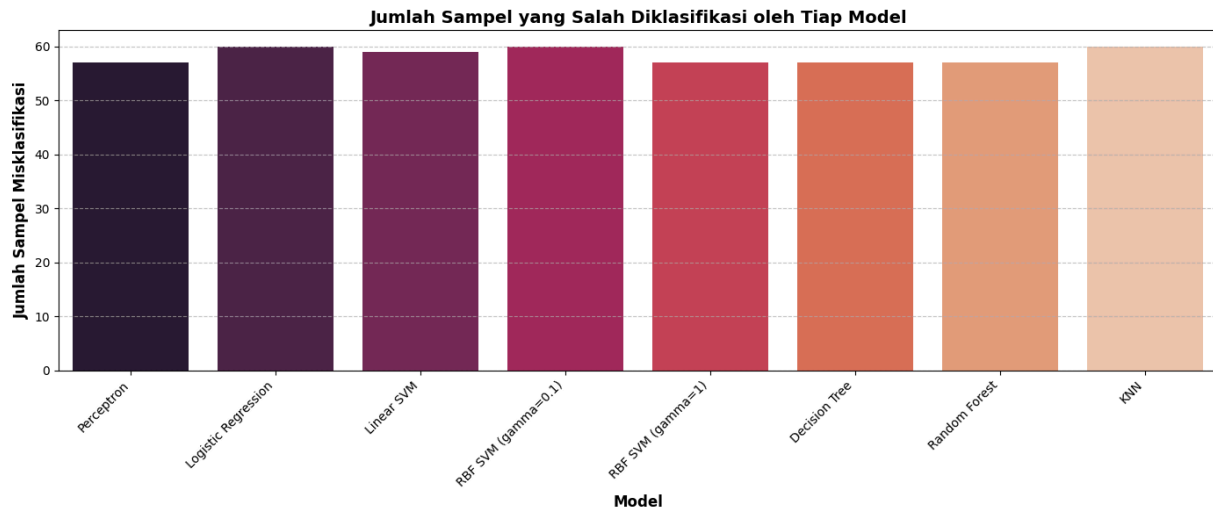
F1 Score: 0.5954

5-Fold CV Accuracy: 0.6012

Misclassified samples: 60

Berdasarkan hasil accuracy, F1 score, CV accuracy, dan misclassified samples, berikut adalah hasil analisis dan perbandingan model.





Peringkat Keseluruhan Model (berdasarkan gabungan semua metrik):

	Model	Accuracy	F1 Score	5-Fold CV Accuracy	Misclassified Samples	Accuracy Rank	F1 Rank	CV Rank	Misclassified Rank	Overall Rank
6	Random Forest	0.650307	0.607211	0.656442	57	2.5	3.0	3.5	2.5	2.875
4	RBF SVM (gamma=1)	0.650307	0.607211	0.656442	57	2.5	3.0	3.5	2.5	2.875
5	Decision Tree	0.650307	0.607211	0.652761	57	2.5	3.0	5.5	2.5	3.375
0	Perceptron	0.650307	0.627036	0.592638	57	2.5	1.0	8.0	2.5	3.500
2	Linear SVM	0.638037	0.547527	0.657669	59	5.0	6.0	2.0	5.0	4.500
3	RBF SVM (gamma=0.1)	0.631902	0.540723	0.660123	60	7.0	7.5	1.0	7.0	5.625
7	KNN	0.631902	0.595365	0.601227	60	7.0	5.0	7.0	7.0	6.500
1	Logistic Regression	0.631902	0.540723	0.652761	60	7.0	7.5	5.5	7.0	6.750

**Random Forest:** Salah satu model terbaik dengan peringkat 2.875. Akurasinya tinggi (0.650), F1 Score bagus (0.607), cross-validation sangat baik (0.656), dan kesalahan klasifikasi paling sedikit (57 sampel). Performanya sangat seimbang di semua metrik.

**RBF SVM (gamma=1):** Berbagi peringkat terbaik dengan Random Forest (2.875). Memiliki nilai yang sama persis untuk semua metrik dengan Random Forest: akurasi 0.650, F1 Score 0.607, CV accuracy 0.656, dan 57 sampel salah klasifikasi.

**Decision Tree:** Peringkat ketiga (3.375). Akurasi sama dengan dua model teratas (0.650), F1 Score juga sama (0.607), tetapi CV accuracy sedikit lebih rendah (0.652). Jumlah kesalahan klasifikasi tetap rendah (57 sampel).

**Perceptron:** Peringkat keempat (3.500). Akurasinya tinggi (0.650) dan F1 Score tertinggi (0.627), tetapi CV accuracy-nya rendah (0.592). Kesalahan klasifikasi tetap rendah (57 sampel). Performa cross-validation yang buruk menunjukkan model ini mungkin overfitting.

**Linear SVM:** Peringkat kelima (4.500). Akurasi menengah (0.638), F1 Score rendah (0.547), tetapi CV accuracy cukup baik (0.657). Kesalahan klasifikasi agak tinggi (59 sampel).

**RBF SVM ( $\gamma=0.1$ ):** Peringkat keenam (5.625). Akurasi lebih rendah (0.631), F1 Score rendah (0.540), tetapi memiliki CV accuracy tertinggi (0.660). Kesalahan klasifikasi tinggi (60 sampel).

**KNN:** Peringkat ketujuh (6.500). Akurasi rendah (0.631), F1 Score cukup baik (0.595), CV accuracy rendah (0.601), dan kesalahan klasifikasi tinggi (60 sampel).

**Logistic Regression:** Peringkat terendah (6.750). Akurasi rendah (0.631), F1 Score rendah (0.540), CV accuracy menengah (0.652), dan kesalahan klasifikasi tinggi (60 sampel).

### Kesimpulan

Dengan input baru "Jumlah SKS" dan "Minat pemahaman materi kuliah", Random Forest dan RBF SVM ( $\gamma=1$ ) menunjukkan performa terbaik dengan nilai yang sama persis. Kedua model ini menawarkan keseimbangan optimal antara semua metrik: akurasi tinggi, F1 Score bagus, stabilitas cross-validation yang baik, dan kesalahan klasifikasi paling sedikit.

Perubahan input sangat berpengaruh pada hasil performa model:

- **Perubahan Peringkat Model:** Pada input sebelumnya, Logistic Regression adalah model terbaik. Dengan input baru, Logistic Regression justru menjadi model terburuk, sementara Random Forest dan RBF SVM ( $\gamma=1$ ) naik menjadi yang terbaik.
- **Perubahan Performa Perceptron:** Pada input baru, Perceptron menunjukkan peningkatan signifikan dengan akurasi tinggi dan F1 Score tertinggi, meski cross-validation-nya masih lemah.
- **Peningkatan Akurasi Keseluruhan:** Semua model menunjukkan peningkatan akurasi dengan input baru, menunjukkan bahwa kombinasi "Jumlah SKS" dan "Minat pemahaman" lebih prediktif.
- **Pengurangan Kesalahan Klasifikasi:** Input baru menghasilkan jumlah kesalahan klasifikasi yang lebih rendah pada model-model teratas (57 vs 58 pada input sebelumnya).

Kesimpulannya, kombinasi input baru menghasilkan model yang lebih akurat dan seimbang, dengan Random Forest dan RBF SVM ( $\gamma=1$ ) sebagai pilihan terbaik. Ini menunjukkan bahwa pemilihan fitur yang tepat sangat penting untuk performa model machine learning.