Australia Car Market Data

Kelompok 3

Reyhan Arya Luki Saputra	22.11.4950
Joshua Juliandika Putra	22.11.4941
Claudio Gilang Wicaksana	22.11.4994
Barzy Ariel Nadzif	22.11.4993
Mahsa Athallah Zufar	22.11.4969

1. Berdasarkan apa yang sudah Anda pelajari, silahkan gunakan kemampuan anda untuk menyelesaikan sebuah menggunakan *classification* yang melibatkan penggunaan Machine Learning. (SCPMK 1534113, 50 Poin)

    a Pilih satu bidang yang Anda beserta rekan tim minati, jabarkan alasan pemilihan bidang tersebut dan jelaskan apa yang ingin dicapai dengan memilih topik ini.

➔ Memilih topik **pasar mobil di Australia** dapat memberikan manfaat dan tujuan tertentu, seperti analisis data pasar, tren industri otomotif, atau peluang bisnis. Berikut adalah beberapa alasan dan tujuan yang dapat dicapai dengan membahas topik ini:

1. Memahami Tren Konsumen dan Teknologi

- **Mengidentifikasi preferensi konsumen:** Misalnya, apakah konsumen lebih memilih kendaraan SUV, kendaraan listrik (EV), atau kendaraan hibrida.
- **Analisis teknologi baru:** EV dan hybrid menjadi fokus utama, sehingga analisis ini penting untuk memahami adopsi teknologi ramah lingkungan.

2. Menganalisis Faktor Ekonomi yang Mempengaruhi Pasar

- **Pengaruh nilai tukar mata uang:** Seperti dampaknya pada harga mobil impor.
- **Harga bahan bakar:** Bagaimana kenaikan harga bahan bakar memengaruhi pilihan konsumen.
- **Daya beli masyarakat:** Apakah pertumbuhan penjualan menunjukkan pemulihan ekonomi pasca-pandemi.

3. Mendukung Pengambilan Keputusan Bisnis

- **Strategi pemasaran:** Membantu produsen dan distributor mobil untuk menargetkan segmen pasar tertentu, seperti penjualan SUV yang tinggi.
- **Perencanaan ekspansi:** Memahami apakah pasar Australia adalah tempat yang menguntungkan untuk investasi dalam EV atau teknologi terkait lainnya.

4. Kontribusi pada Kebijakan Pemerintah dan Lingkungan

- **Kebijakan energi bersih:** Mendukung pemerintah dalam memahami adopsi EV dan mendesain strategi yang lebih efektif untuk mencapai target penurunan emisi.
- **Infrastruktur transportasi:** Memberikan wawasan tentang kebutuhan infrastruktur seperti stasiun pengisian daya EV.

5. Kesempatan Penelitian dan Edukasi

- **Studi pasar:** Meningkatkan pengetahuan tentang dinamika pasar otomotif global dengan fokus khusus pada Australia.
- **Inovasi teknologi:** Membuka peluang penelitian dalam pengembangan teknologi mobil ramah lingkungan.

b Ceritakan proses mendapatkan data dan informasi lengkap mengenai data tersebut (seperti waktu, penjelasan setiap kolom, sumber dll). Data yang digunakan harus data terbaru dengan range 1-4 tahun kebelakang.

➔ Proses pengambilan dataset tersebut dilakukan pada tanggal 5 januari 2025.

Dataset ini memiliki 16 kolom data yang terdiri dari kolom-kolom berikut :

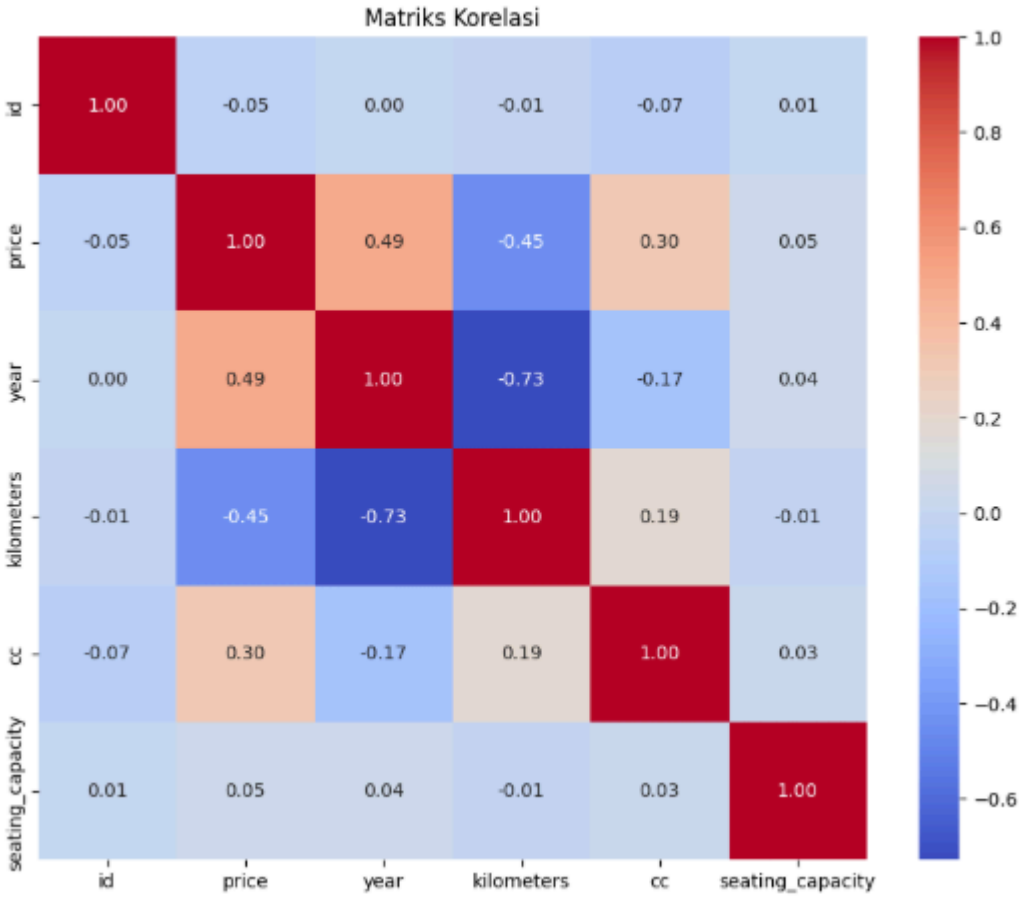
1. ID
2. Name
3. Price
4. Brand
5. Model
6. Variant
7. Series
8. Year
9. Kilometers
10. Type
11. Gearbox
12. Fuel
13. Status
14. CC
15. Color
16. Seating Capacity

Dataset tersebut diambil dari kaggle.com yang berjudul (Australia Car Market Data) dengan data yang berjumlah 17.048 data.

c Lakukan pre-processing data dengan memeriksa tipe data, mengganti nama kolom, memeriksa nilai null, mengubah tipe data (agar bisa di proses), menampilkan summary, dan menampilkan matriks korelasinya menggunakan metode metode yang pernah dipelajari.

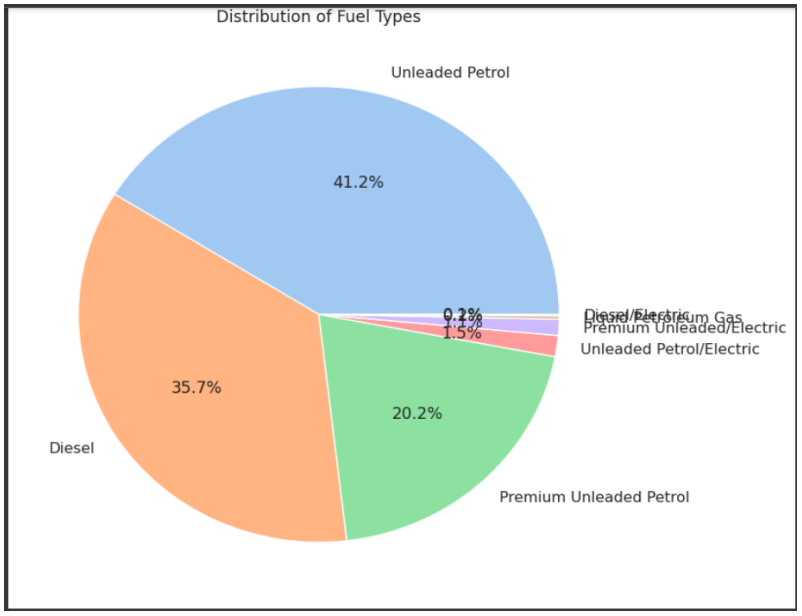
```
Tipe data awal:
ID                 int32
Name              object
Price              int32
Brand             object
Model             object
Variant           object
Series            object
Year               int32
Kilometers         int32
Type              object
Gearbox           object
Fuel              object
Status            object
CC                 int32
Color             object
Seating Capacity   int32
dtype: object

Jumlah nilai null per kolom:
id                0
name              0
price             0
brand             0
model             0
variant           0
series            0
year              0
kilometers        0
type              0
gearbox           0
fuel              0
status            0
cc                0
color             0
seating_capacity  0
dtype: int64
Hasil preprocessing telah disimpan ke /content/preprocessed_cars_info.csv.
<ipython-input-2-f886859b500b>:28: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.


  df[col].fillna(df[col].mean(), inplace=True)
```

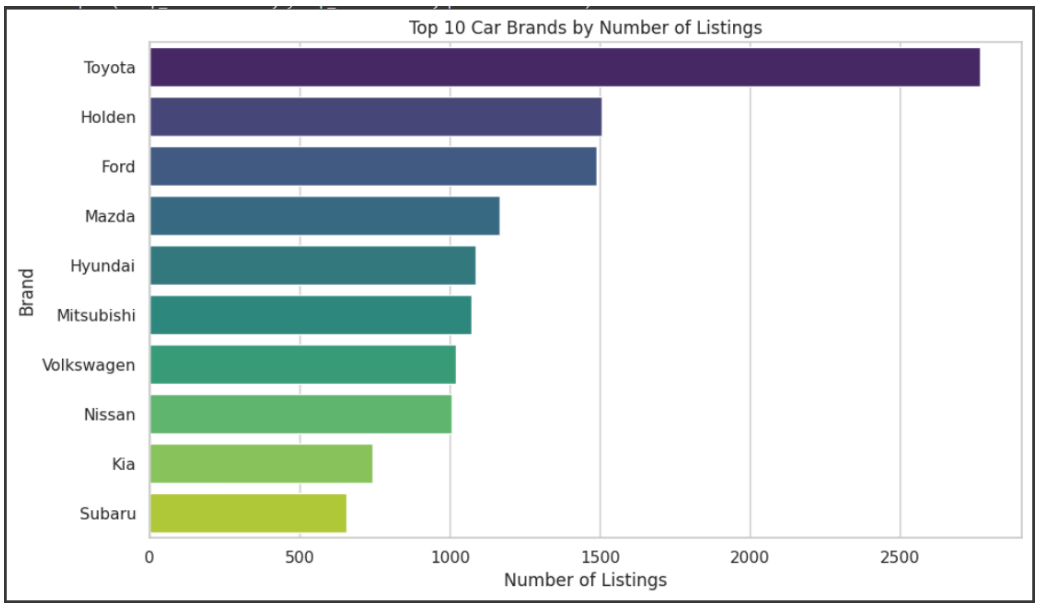
```
Ringkasan data:
                id          price          year     kilometers            cc  \
count  1.704800e+04   17048.000000  17048.000000   1.704800e+04  17048.000000
mean   1.279027e+07   36777.778038   2015.481288   1.032314e+05   2491.830303
std    5.051111e+04   30305.015328      4.721591   8.041313e+04    881.985562
min    1.153013e+07    1000.000000   1989.000000   1.000000e+00    875.000000
25%    1.275715e+07   18800.000000   2013.000000   4.450225e+04   1987.000000
50%    1.280207e+07   29990.000000   2016.000000   8.845400e+04   2354.000000
75%    1.283131e+07   45990.000000   2019.000000   1.488735e+05   2981.000000
max    1.285246e+07  999000.000000   2022.000000   2.700000e+06   7300.000000

       seating_capacity
count      17048.000000
mean           5.115849
std            1.121791
min            2.000000
25%            5.000000
50%            5.000000
75%            5.000000
max           14.000000

Matriks korelasi:
                          id     price      year  kilometers        cc  \
id                  1.000000 -0.047978  0.004708   -0.011543 -0.070773
price              -0.047978  1.000000  0.488033   -0.447490  0.298205
year                0.004708  0.488033  1.000000   -0.728515 -0.174578
kilometers         -0.011543 -0.447490 -0.728515    1.000000  0.185219
cc                 -0.070773  0.298205 -0.174578    0.185219  1.000000
seating_capacity    0.005320  0.046629  0.044151   -0.013088  0.029590

                   seating_capacity
id                         0.005320
price                      0.046629
year                       0.044151
kilometers                -0.013088
cc                         0.029590
seating_capacity           1.000000
```

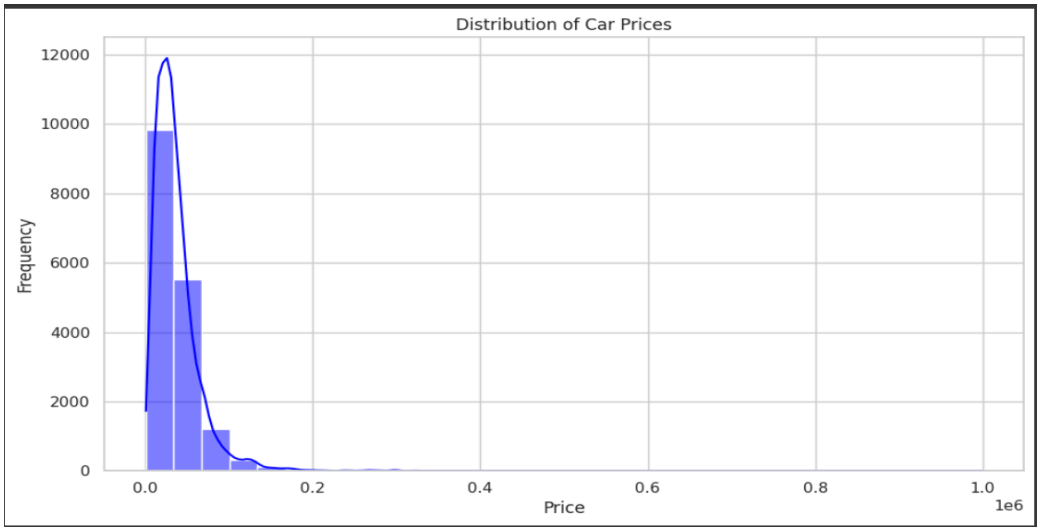


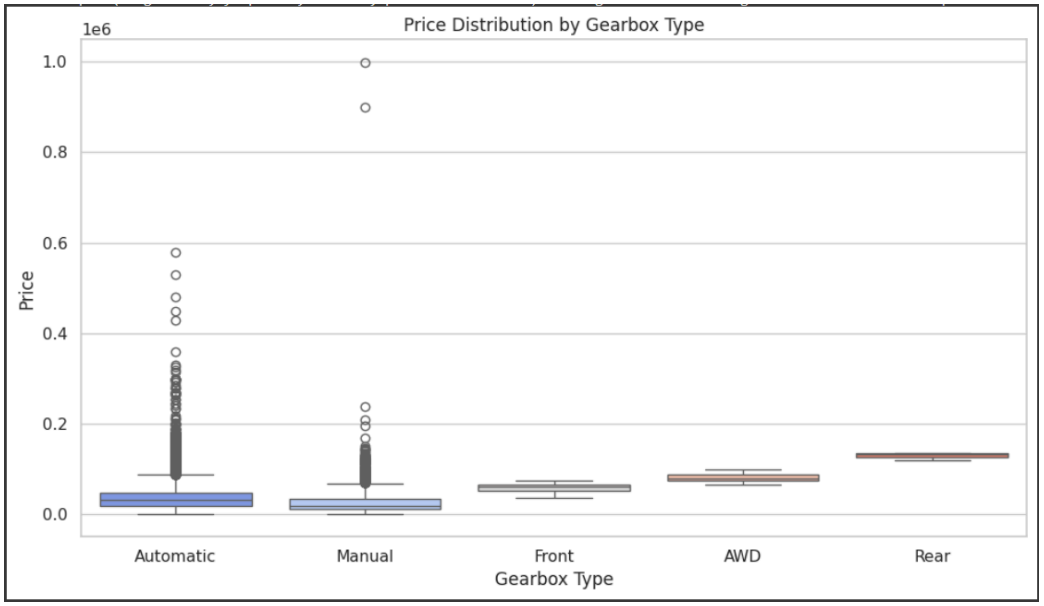
d Gunakan exploratory data analysis (EDA) untuk melihat sudut padang yang ada mengenai data (minimal 4) dua diantaranya bar dan pie chart, 2 diantaranya bebas. Berikan penjelasan.

➔




Top 10 Car Brands by Number of Listings
Toyota
Holden
Ford
Mazda
Hyundai
Mitsubishi
Volkswagen
Nissan
Kia
Subaru
Brand
0
500
1000
1500
2000
2500
Number of Listings


Distribution of Car Prices
12000
10000
8000
6000
4000
2000
0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Price


Price Distribution by Gearbox Type
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Price
Automatic
Manual
Front
AWD
Rear
Gearbox Type

e Berdasarkan analisis data tersebut, jelaskan alasan pemilihan kolom/fitur yang relevan untuk menyelesaikan permasalahan yang ingin dicapai.

berikut adalah analisis dan alasan pemilihan fitur yang relevan untuk memprediksi harga mobil:

Analisis Statistik Deskriptif

1. **Kolom `price`:**
   - Kolom ini adalah target prediksi yang menjadi fokus analisis.
   - Harga berkisar antara 1.000 hingga 999.000, dengan nilai rata-rata 36.777,78 dan distribusi yang lebar (terlihat dari nilai standar deviasi yang tinggi).
2. **Kolom `year`:**
   - Terdapat korelasi positif dengan harga (`0.4883`), menunjukkan bahwa tahun produksi mobil memengaruhi harga secara signifikan.
   - Mobil yang lebih baru cenderung memiliki harga lebih tinggi.
3. **Kolom `kilometers`:**
   - Korelasi negatif dengan harga (`-0.4475`), yang mengindikasikan bahwa mobil dengan jarak tempuh yang lebih jauh cenderung memiliki harga lebih rendah.
   - Jarak tempuh adalah indikator keausan dan umur pemakaian mobil.
4. **Kolom `cc` (kapasitas mesin):**
   - Korelasi positif dengan harga (`0.2982`), menunjukkan bahwa mobil dengan kapasitas mesin lebih besar cenderung lebih mahal.
   - Kapasitas mesin mencerminkan performa kendaraan, yang relevan untuk penentuan harga.
5. **Kolom `seating_capacity`:**
   - Korelasi kecil tetapi positif (`0.0466`) dengan harga.
   - Walaupun pengaruhnya kecil, fitur ini tetap relevan dalam konteks mobil yang ditujukan untuk keluarga atau penumpang banyak.

2. Pengembangan model machine learning. (SCPMK 1534114, 50 Poin)

   a. Gunakan minimal 4 model Machine Learning dari library Spark untuk menyelesaikan masalah yang Anda pilih. 2 Model sesuai dengan instruksi (Random Forest, Gradient Boost Tree) dan dua model lain bebas (belum pernah dibahas). Lalu bandingkan hasilnya menggunakan metriks seperti AUC (ROC Curve), Akurasi, F1 Score, Presisi, dan Recall.

```
Random Forest Accuracy: 0.9302186878727634
Gradient Boosted Tree Accuracy: 0.9196819085487078
KNN Accuracy: 0.8543499511241447
MLP Accuracy: 0.8294234592445328

Model Accuracy Comparison:
Random Forest: 0.9302
Gradient Boosted Tree: 0.9197
KNN: 0.8543
MLP: 0.8294
```

Dari keempat model tersebut yang memiliki hasil akurasi tertinggi adalah Random Forest dan Gradient Boost Tree dengan jumlah akurasi 93% dan 91%.

b. Dari ke-4 model classification tersebut, pilih dua model dengan performa terbaik dan lakukan hyperparameter tuning untuk melihat perubahan performa yang dihasilkan. Lalu tentukan model terbaik yang bisa menjadi solusi pada masalah yang Anda tetapkan diawal.

```
Random Forest Accuracy: 0.9302186878727634
Random Forest F1 Score: 0.9299073126100834
```

```
Gradient Boosted Tree Accuracy: 0.9196819085487078
Gradient Boosted Tree F1 Score: 0.9202978324654103
```

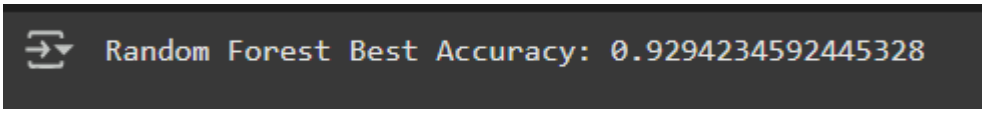
Performa Kedua Model:

1. **Random Forest**:
   - **Akurasi**: 0.9302
   - **F1-Score**: 0.9299
   - Random Forest memiliki performa yang sangat baik dalam akurasi dan F1-score, menunjukkan kemampuan yang kuat untuk menangani ketidakseimbangan kelas (jika ada).
2. **Gradient Boosted Tree**:
   - **Akurasi**: 0.9197
   - **F1-Score**: 0.9203
   - Gradient Boosted Tree sedikit tertinggal dibanding Random Forest, meskipun tetap memiliki kinerja tinggi.

Pilihan Model Terbaik:

- **Random Forest** dipilih sebagai model terbaik berdasarkan:
  - Akurasi yang lebih tinggi (0.9302 vs 0.9197).
  - F1-Score yang sedikit lebih baik, menunjukkan keseimbangan antara presisi dan recall dalam menangani klasifikasi.

c. Jabarkan karakteristik model terbaik yang Anda dapatkan terhadap korelasinya dengan data. Apakah ada sifat tertentu dari data yang ternyata cocok dengan model dan sebaliknya?



Berdasarkan hasil akhir yang menunjukkan bahwa **Random Forest** memiliki akurasi terbaik setelah hyperparameter tuning (**0.9294**), berikut adalah analisis terkait karakteristik model ini dan hubungannya dengan data yang digunakan:

Karakteristik Random Forest

1. **Kemampuan Menangani Data dengan Fitur Beragam:**
    - Random Forest adalah model berbasis ensemble yang menggabungkan hasil dari beberapa decision tree.
    - Model ini sangat baik dalam menangani data dengan banyak fitur yang saling berinteraksi.
    - Cocok untuk data yang memiliki kombinasi fitur numerik dan kategorikal.
2. **Robust terhadap Outlier dan Noise:**
    - Random Forest relatif tahan terhadap outlier dan noise karena menggunakan rata-rata dari prediksi beberapa pohon.
3. **Kemampuan Mengatasi Ketidakseimbangan Data:**
    - F1-score menunjukkan bahwa model ini mampu menyeimbangkan presisi dan recall, yang mengindikasikan bahwa model dapat menangani distribusi kelas yang tidak seimbang.
4. **Mengurangi Risiko Overfitting:**
    - Dengan mekanisme pemilihan fitur acak dan rata-rata hasil dari beberapa pohon, Random Forest cenderung tidak overfit dibandingkan decision tree tunggal.

Korelasi Model dengan Data

1. **Struktur Data yang Mungkin Mendukung Kinerja Random Forest:**
    - **Fitur Independen:** Jika data memiliki fitur yang kurang berkorelasi, Random Forest cenderung bekerja lebih baik karena setiap pohon dalam ensemble mencoba belajar dari subset fitur.
    - **Non-Linearitas:** Jika data memiliki hubungan non-linear antar fitur, Random Forest dapat menangkap pola-pola kompleks ini melalui banyak pohon.
2. **Skala Fitur Tidak Penting:**

    - Tidak seperti model berbasis gradien seperti SVM atau Logistic Regression, Random Forest tidak sensitif terhadap skala fitur. Ini mempermudah pengolahan data tanpa memerlukan normalisasi atau standarisasi.
3. **Redundansi dan Fitur Tidak Penting:**
    - Random Forest secara inheren memilih subset fitur secara acak untuk setiap pohon, sehingga fitur yang kurang penting atau redundan tidak terlalu memengaruhi performa keseluruhan model.
4. **Distribusi Kelas:**
    - Jika data memiliki distribusi kelas yang relatif tidak terlalu seimbang, algoritma Random Forest dapat bekerja dengan baik dengan menyesuaikan parameter seperti class_weight.

Sifat Data yang Tidak Cocok dengan Random Forest

1. **Volume Data yang Sangat Besar:**
    - Jika data memiliki jumlah fitur atau sampel yang sangat besar, Random Forest bisa menjadi lambat karena proses pembentukan banyak pohon.
2. **Kebutuhan Interpretasi:**
    - Jika interpretasi mendalam untuk setiap prediksi sangat penting, Random Forest mungkin kurang cocok karena sifat ensemble membuat hasilnya lebih sulit dijelaskan dibandingkan model linear.

Kesimpulan

Random Forest menjadi model terbaik karena:

- Data mendukung karakteristik Random Forest, seperti menangani hubungan non-linear antar fitur dan ketidakseimbangan kelas.
- Model ini memberikan keseimbangan antara akurasi dan generalisasi, sehingga dapat menangkap pola kompleks tanpa overfitting.

Link colab : uas6.ipynb