Tugas Modul 8 (20 Oktober 2025)

Tugas Versi A/1

Kelompok 14 A STARTERPACK :

1. Aditya Winarto

Link Google Colab : Link Google Colab

Penjelasan dataset

Dataset Auto MPG yang Anda kirim (dari UCI Machine Learning Repository) adalah kumpulan data multivariat yang digunakan untuk tugas regresi, yang mencatat konsumsi bahan bakar mobil dalam Miles Per Gallon (MPG) dari tahun 1970 hingga 1982. Dataset ini terdiri dari 398 observasi (mobil) dan 9 atribut, termasuk fitur kontinu seperti displacement, horsepower, dan weight (yang merupakan prediktor utama), serta fitur diskret/kategorikal seperti cylinders, model year, dan origin (asal negara: Amerika, Eropa, atau Jepang).

Exploratory Data Analysis

- **Menampilkan sample data**
  Pada bagian ini digunakan untuk melakukan verifikasi beberapa hal penting seperti nama dan urutan kolom.

	displacement	mpg	cylinders	horsepower	weight	acceleration	\
0	307.0	18.0	8	130.0	3504	12.0	
1	350.0	15.0	8	165.0	3693	11.5	
2	318.0	18.0	8	150.0	3436	11.0	
3	304.0	16.0	8	150.0	3433	12.0	
4	302.0	17.0	8	140.0	3449	10.5	

	model_year	origin	car_name
0	70	1	chevrolet chevelle malibu
1	70	1	buick skylark 320
2	70	1	plymouth satellite
3	70	1	amc rebel sst
4	70	1	ford torino

Berdasarkan hasil tersebut dataset mpg memiliki beberapa kolom seperti displacement, mpg, dan cylinder. Kolom ini nantinya digunakan sebagai variable dependen dan juga variable independen untuk membangun sebuah model prediksi.

- **Menggunakan statistik deskriptif**
  Digunakan untuk mengetahui gambaran sebuah data secara umum seperti jumlah data rata rata nilai min dan max serta standar deviasi.

	displacement	mpg	cylinders	horsepower	weight	acceleration	model_year	origin
count	398.000000	398.000000	398.000000	398.000000	398.000000	398.000000	398	398.000000
mean	193.425879	23.513693	5.454774	103.790201	2970.424623	15.552010	1976-01-05 01:30:27.135678400	1.572864
min	68.000000	9.000000	3.000000	46.000000	1613.000000	8.800000	1970-01-01 00:00:00	1.000000
25%	104.250000	17.500000	4.000000	76.000000	2223.750000	13.825000	1973-01-01 00:00:00	1.000000
50%	148.500000	23.000000	4.000000	93.500000	2803.500000	15.500000	1976-01-01 00:00:00	1.000000
75%	262.000000	29.000000	8.000000	125.000000	3608.000000	17.175000	1979-01-01 00:00:00	2.000000
max	455.000000	46.250000	8.000000	198.500000	5140.000000	22.200000	1982-01-01 00:00:00	3.000000
std	104.269838	7.813400	1.701004	36.770468	846.841774	2.693089	NaN	0.802055

Berdasarkan tabel statistik deskriptif dari dataset Auto MPG, dapat disimpulkan bahwa data terdiri dari 398, dengan rentang nilai yang luas di antara fitur-fitur teknis utama: rata-rata konsumsi bahan bakar (MPG) adalah sekitar 23.45 dengan deviasi standar 7.80, menunjukkan variasi yang signifikan; mobil-mobil tersebut memiliki rata-rata berat (Weight) sekitar 2977 pon, dan rata-rata daya kuda (Horsepower) sekitar 104.47, dengan nilai maksimum mencapai 230 (Horsepower) dan 5140 (Weight).

car_name	count
ford pinto	6
ford maverick	5
amc matador	5
toyota corolla	5
amc hornet	4
...	...
amc concord dl	1
volkswagen rabbit l	1
mazda glc custom l	1
mazda glc custom	1
chevy s-10	1

Selain numerik juga terdapat data kategorikal yaitu nama nama mobil yang ada dalam dataset tersebut seperti ford pinto, dalam dataset tersebut terdapat 6 mobil jenis tersebut.

Data preparation

- **Mengubah format tanggal dan tipe jenis datanya**

Digunakan untuk memastikan bahwa *temporal information* (informasi waktu) diperlakukan secara benar, memungkinkan model untuk menangkap tren linear seiring berjalannya waktu—seperti korelasi positif antara tahun model yang lebih baru dan efisiensi MPG yang lebih baik—atau untuk memplot data secara akurat pada sumbu waktu.

```
--- Verifikasi Tipe Data Setelah Transformasi ---
0   1970-01-01
1   1970-01-01
2   1970-01-01
3   1970-01-01
4   1970-01-01
Name: model_year, dtype: datetime64[ns]

Tipe data kolom 'model_year' baru: datetime64[ns]
```

Hasil transformasi menunjukkan bahwa kolom model_year yang awalnya berisi integer dua digit (misalnya, 70 dan 71) telah berhasil dikonversi menjadi objek tanggal lengkap (datetime64[ns]), yang secara default diinterpretasikan sebagai tanggal 1 Januari pada tahun 1900-an (yaitu, 1970-01-01 dan 1971-01-01).

- **Menangani missing value pada kolom horsepower**
  Digunakan untuk mengurangi ketidakakuratan karena nilainya kosong sehingga jika terdapat missing value maka akan menyebabkan model tidak akurat.

  ```
  --- Proses Imputasi Selesai ---
  Nilai Median Horsepower yang digunakan: 93.50
  Jumlah baris NaN yang diimputasi: 6
  ```

  Berdasarkan hasil tersebut terdapat bahwa ada 6 missing value. Karena datanya tergolong sedikit hanya 398 sehingga dipertahankan baris yang kosong tersebut dengan cara mengisinya menggunakan median. Median dipilih untuk menghindari outlier yang semakin banyak sehingga dengan adanya median outlier yang dihasilkan tidak terlalu banyak dan mempermudah penanganan outlier jika ada masalah tentang outlier

- **Mendeteksi outlier menggunakan iqr**
  Digunakan untuk mendeteksi *outlier* dengan menentukan batas toleransi normal pada distribusi dataMetode ini dianggap *robust* karena tidak terlalu sensitif terhadap nilai ekstrem, berbeda dengan metode yang menggunakan rata-rata atau standar deviasi.

```
--- Kolom: mpg ---
Q1: 17.50, Q3: 29.00, IQR: 11.50
Batas Bawah: 0.25
Batas Atas: 46.25
Jumlah Outlier Ditemukan: 1 dari 398 total data.
Persentase Outlier: 0.25 %
----------------------------------------
--- Kolom: cylinders ---
Q1: 4.00, Q3: 8.00, IQR: 4.00
Batas Bawah: -2.00
Batas Atas: 14.00
Jumlah Outlier Ditemukan: 0 dari 398 total data.
Persentase Outlier: 0.00 %
----------------------------------------
--- Kolom: displacement ---
Q1: 104.25, Q3: 262.00, IQR: 157.75
Batas Bawah: -132.38
Batas Atas: 498.62
Jumlah Outlier Ditemukan: 0 dari 398 total data.
Persentase Outlier: 0.00 %
----------------------------------------
----------------------------------------
--- Kolom: horsepower ---
Q1: 76.00, Q3: 125.00, IQR: 49.00
Batas Bawah: 2.50
Batas Atas: 198.50
Jumlah Outlier Ditemukan: 11 dari 398 total data.
Persentase Outlier: 2.76 %
----------------------------------------
--- Kolom: weight ---
Q1: 2223.75, Q3: 3608.00, IQR: 1384.25
Batas Bawah: 147.38
Batas Atas: 5684.38
Jumlah Outlier Ditemukan: 0 dari 398 total data.
Persentase Outlier: 0.00 %
----------------------------------------
--- Kolom: acceleration ---
Q1: 13.83, Q3: 17.18, IQR: 3.35
Batas Bawah: 8.80
Batas Atas: 22.20
Jumlah Outlier Ditemukan: 7 dari 398 total data.
Persentase Outlier: 1.76 %
----------------------------------------
```

  Outlier yang terdeteksi pada variabel target mpg (1 outlier atau 0.25\%), horsepower (11 outlier atau 2.76%), dan acceleration (7 outlier atau 1.76%). Hasil ini mengindikasikan bahwa terdapat beberapa kendaraan yang memiliki nilai efisiensi bahan bakar yang sangat tinggi atau sangat rendah, daya kuda yang ekstrem, atau waktu akselerasi yang tidak biasa, yang perlu ditangani
- **Menangani outlier menggunakan winsorization**
  Metode Winsorization (sering disebut *capping* atau *clamping*) adalah teknik statistik untuk menangani *outlier* pada data numerik dengan mengganti (bukan menghapus) nilai-nilai ekstrem tersebut dengan nilai ambang batas terdekat. Prosesnya bekerja dengan menentukan batas Bawah dan Batas Atas—yang biasanya dihitung berdasarkan persentil tertentu.

```
--- Menerapkan Winsorization (Capping) ---
✅ Kolom 'mpg' telah di-capping.
✅ Kolom 'horsepower' telah di-capping.
✅ Kolom 'acceleration' telah di-capping.
```

  Berdasarkan hasil tersebut kolom kolom yang memiliki nilai outlier sudah berhasil ditangani dengan metode tersebut. metode ini digunakan agar data tersebut tetap dapat digunakan dan tidak perlu dihapus.

Bangun model

- **Menentukan variable independent menggunakan r**

  Digunakan dalam pemilihan fitur pada Regresi Linier karena ia secara kuantitatif mengukur kekuatan dan arah hubungan linier antara variabel independen dan variabel target . Nilai yang mendekati |1| (baik positif maupun negatif) mengindikasikan bahwa variabel tersebut sangat relevan dan memiliki potensi prediktif yang tinggi

```
--- 📈 Korelasi Variabel dengan MPG (Diurutkan) ---
mpg             1.000000
weight         -0.831901
displacement   -0.804353
horsepower     -0.785846
cylinders      -0.775556
acceleration    0.418333
```

  Berdasarkan hasil tersebut selain mpg sendiri nilai r yang memiliki hubungan kuat adalah kolom weight dengan -0.8 menunjukkan hubungan negatif sebagai contoh semakin kecil weightnya maka semakin besar mpgnya kemudiann disusul displacement, horsepower, cylinders dan acceleration dengan hubungan yang negatif juga.

- **Menentukan variable independent menggunakan AIC dan R2**

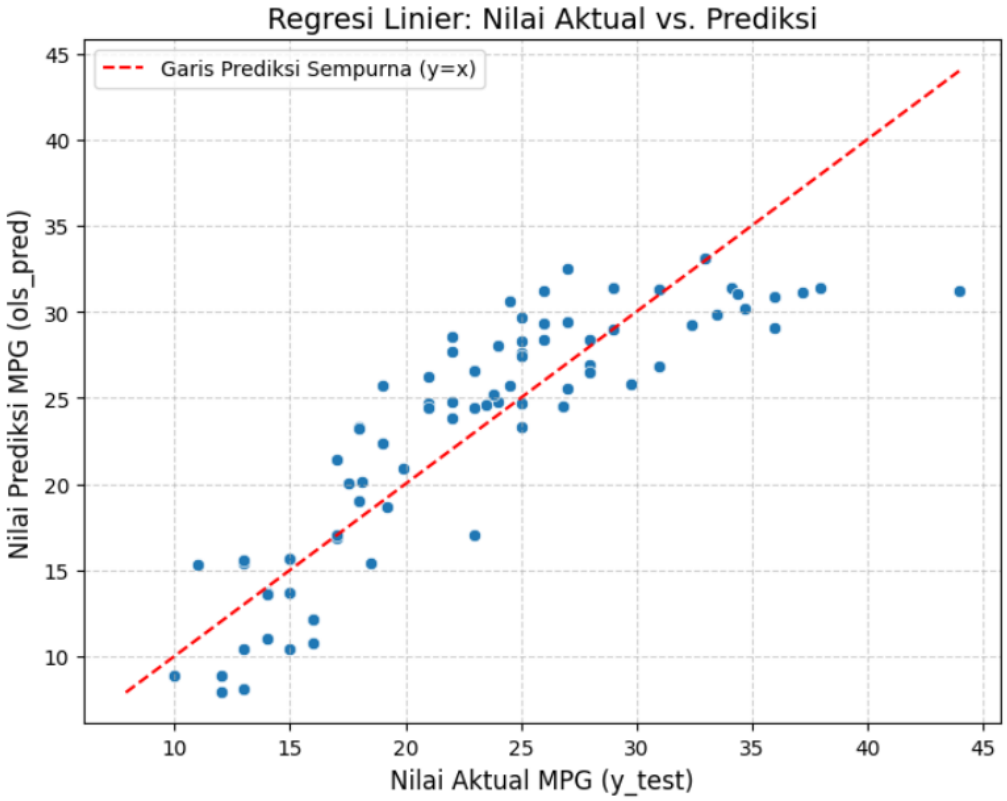
  Pemilihan variabel independen menggunakan AIC (Akaike Information Criterion) dan R2 (Koefisien Determinasi) dilakukan untuk menemukan model regresi yang paling optimal dan sederhana. Metrik R2 mengukur seberapa baik variasi dalam variabel target dijelaskan oleh prediktor—tujuannya adalah memaksimalkan R2. AIC adalah metrik yang menyeimbangkan kecocokan model R2 dengan kompleksitasnya (jumlah prediktor), di mana model yang lebih baik dicirikan oleh nilai AIC terendah.

```
--- Riwayat Evaluasi Kombinasi ---
Semua model yang diuji di setiap langkah:
            AIC  R-squared                          Variables
7   2282.793667   0.706669                 weight + horsepower
9   2283.559607   0.707577     weight + horsepower + cylinders
10  2284.065666   0.707205  weight + horsepower + displacement
11  2284.124333   0.707162  weight + horsepower + acceleration
6   2294.064868   0.698244               weight + displacement
8   2295.042048   0.697502               weight + acceleration
5   2296.130462   0.696673                  weight + cylinders
3   2300.138166   0.692060                              weight
1   2354.509302   0.646983                        displacement
2   2386.377605   0.617554                          horsepower
0   2402.755995   0.601488                           cylinders
4   2692.357416   0.175002                        acceleration
```

  Berdasarkan hasil tersebut maka variable independent yang dipilih adalah weight dan horsepower karena memiliki nilai AIC yang terendah serta memiliki nilai R2 yang cukup baik meski tidak tertinggi dari kombinasi variable yang lain.

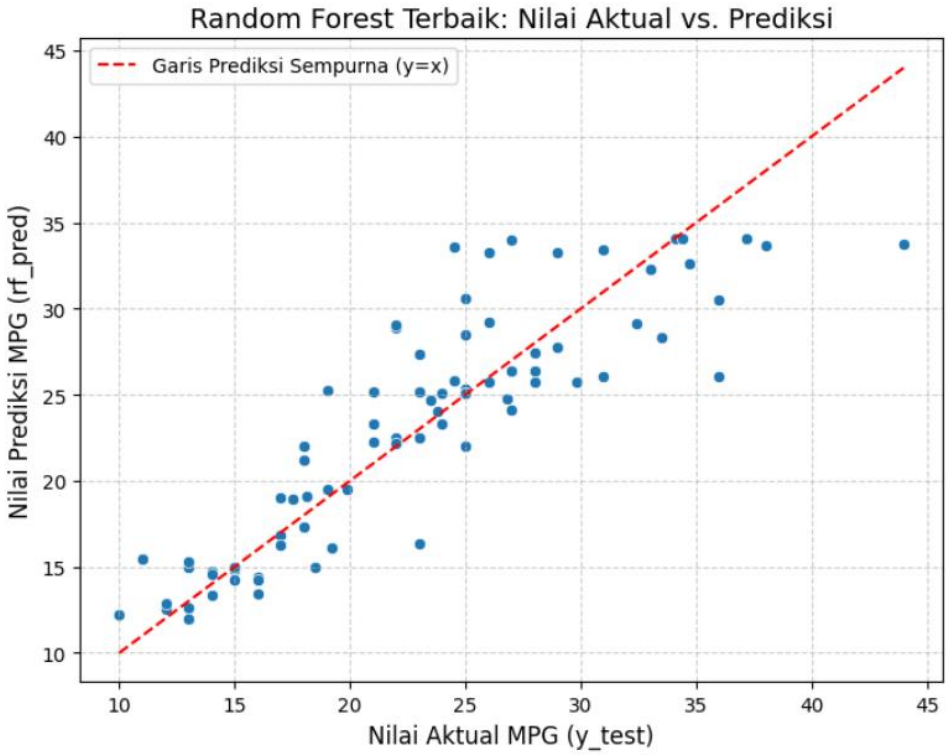
- **Menggunakan multilinier regresi**

  Model statistik yang digunakan untuk memprediksi nilai dari variabel dependen kontinu tunggal (Y) berdasarkan hubungan liniernya dengan dua atau lebih variabel independen. Tujuan model ini adalah menemukan koefisien (bobot) terbaik untuk setiap variabel independen yang meminimalkan jumlah kuadrat kesalahan (selisih antara nilai aktual dan nilai prediksi) di seluruh data. Secara esensial, model ini memodelkan data sebagai sebuah hyperplane multidimensi, di mana setiap koefisien menunjukkan seberapa besar perubahan variabel independen memengaruhi variabel dependen, dengan mengasumsikan variabel lain konstan.



Berdasarkan hasil tersebut secara visualisasi model regresi linier berganda cukup baik digunakan untuk memprediksi mpg dengan variable independent weight dan horsepower. Data training pada model regresi linier berganda memiliki porsi 80% dari keseluruhan data dan data testing sebanyak 20% sisanya.

- **Menggunakan random forest regressor**
  Random Forest Regressor bekerja dengan cara membangun sejumlah besar pohon keputusan (*decision trees*) secara independen dan acak, yang disebut sebagai "forest" (hutan). Proses ini memanfaatkan teknik yang dikenal sebagai bagging (Bootstrap Aggregating). Dalam *bagging*, setiap pohon dilatih menggunakan subset data pelatihan yang berbeda yang diambil secara acak dengan penggantian (bootstrap sample). Selain itu, ketika setiap pohon memecah (*split*) simpulnya, ia hanya mempertimbangkan subset acak dari fitur yang tersedia (misalnya, hanya menggunakan 2 dari 5 fitur yang ada) . Setelah semua pohon selesai dilatih, hasil prediksi akhir untuk nilai target kontinu diperoleh dengan mengambil rata-rata dari semua prediksi individu yang dihasilkan oleh setiap pohon di dalam *forest*.



Pada random forest regression cenderung memiliki hasil visualisasi yang lebih baik dibandingkan regresi linier berganda. Hal tersebut dikarenakan adanya tuning hyperparameter menggunakan grid search, dimana parameter yang dilakukan tunning yaitu ada n_estimator yaitu jumlah pohon pada model tersebut. kemudian max_depth yaitu kedalaman tree nya dan yang terakhir yaitu min_sample_split dimana sample tersebut dibagi menjadi beberapa bagian untuk menemukan hasil model yang terbaik.

Evaluasi model

- **Metrik Evaluasi**
    - R-squared (R2) atau Koefisien Determinasi: Metrik ini mengukur proporsi varians dalam variabel dependen (Y) yang dapat dijelaskan oleh variabel independen (X). Nilai R2 berkisar dari 0 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa model sangat baik dan mampu menjelaskan hampir semua variasi dalam data.
    - RMSE (Root Mean Squared Error): RMSE adalah akar kuadrat dari rata-rata kuadrat kesalahan (*Mean Squared Error*). Metrik ini mewakili rata-rata magnitudo kesalahan yang dibuat model dalam satuan variabel dependen (Y) (misalnya, MPG). Karena kesalahan dikuadratkan, metrik ini memberikan penalti yang lebih besar pada kesalahan prediksi yang besar (*outlier*).
    - MSE (Mean Squared Error): MSE adalah rata-rata kuadrat dari kesalahan. Ini adalah metrik yang digunakan sebagai fungsi kerugian utama untuk melatih model linier, dan nilainya selalu non-negatif, dengan nilai mendekati 0 menunjukkan kinerja yang sempurna.
    - MAE (Mean Absolute Error): MAE adalah rata-rata nilai absolut dari semua kesalahan prediksi. Metrik ini kurang sensitif terhadap *outlier* dibandingkan RMSE dan memberikan indikasi yang lebih langsung tentang rata-rata kesalahan aktual.
    - R (Koefisien Korelasi Ganda): Meskipun sering digunakan pada konteks model linier, R adalah akar kuadrat dari R2 (hanya diambil nilai positifnya), yang menunjukkan koefisien korelasi antara nilai aktual dan nilai prediksi. Nilai mendekati 1 menunjukkan korelasi yang sangat kuat antara nilai prediksi model dengan nilai aktual yang sebenarnya.

- **Multilinier regresi**

```
--- Metrik Kinerja Pada Data Test (Metrik Error) ---
2. R-squared (Koefisien Determinasi - Test): 0.7320
3. RMSE (Root Mean Squared Error): 3.80
4. MSE (Mean Squared Error): 14.41
5. MAE (Mean Absolute Error): 3.11
```

Metrik evaluasi untuk Regresi Linier Berganda pada data uji menunjukkan bahwa model memiliki R2 sebesar 0.7320, yang berarti sekitar 73% dari variasi MPG pada data yang belum pernah dilihat berhasil dijelaskan oleh fitur weight dan horsepower. Meskipun ini adalah kecocokan yang kuat, nilai R2 yang sangat baik ini datang dengan rata-rata kesalahan prediksi (error) sebesar RMSE 3.80 MSE 14.41 dan MAE 3.11, ini mengindikasikan bahwa terdapat beberapa kesalahan prediksi yang lebih besar (*outlier errors*) yang memengaruhi kinerja model secara signifikan.

- **Random forest regressor**

```
--- Metrik Kinerja Random Forest Terbaik Pada Data Test ---
1. R-squared (Koefisien Determinasi): 0.7719
2. RMSE (Root Mean Squared Error): 3.50
3. MSE (Mean Squared Error): 12.26
4. MAE (Mean Absolute Error): 2.52
```

Berdasarkan hasil metrik evaluasi dari model Random Forest Regressor pada data uji, model ini menunjukkan kinerja prediksi yang sangat kuat dan akurat. Nilai R2 sebesar 0.7719 berarti model ini berhasil menjelaskan 77% dari total variasi yang ada pada nilai MPG, menandakan kecocokan yang superior pada data. Meskipun terdapat R2 yang tinggi, rata-rata kesalahan prediksi model ditunjukkan oleh RMSE sebesar 3.50 MSE 12.26 dan MAE sebesar 2.52 , menunjukkan bahwa *outlier errors* (kesalahan prediksi besar) jarang terjadi, menegaskan Random Forest sebagai model yang andal dan akurat untuk memprediksi efisiensi bahan bakar.

Kesimpulan

Berdasarkan perbandingan metrik evaluasi kedua model, Random Forest Regressor menunjukkan kinerja yang superior dibandingkan dengan Regresi Linier Berganda dalam memprediksi efisiensi bahan bakar (MPG). Random Forest mencapai nilai R^2 yang lebih tinggi (0.7719 vs 0.7320), menjelaskan 4% lebih banyak variasi dalam data, sementara secara bersamaan menghasilkan kesalahan prediksi yang lebih rendah di semua metrik evaluasi (RMSE: 3.50 vs 3.80, MSE: 12.26 vs 14.41, dan MAE: 2.52 vs 3.11). Keunggulan Random Forest terletak pada kemampuannya menangani hubungan non-linier yang kompleks antara weight, horsepower, dan MPG melalui ensemble dari multiple decision trees, serta lebih robust terhadap outlier dibandingkan dengan pendekatan linier yang lebih sederhana. Meskipun kedua model menunjukkan performa yang baik dengan R^2 di atas 70%, Random Forest terbukti lebih andal dan akurat untuk aplikasi prediksi efisiensi bahan bakar, dengan konsistensi kesalahan yang lebih baik dan kemampuan generalisasi yang superior pada data yang belum pernah dilihat.