Tugas 2: Praktikum Mandiri 2 Machine Learning

**Hisyam Wildan Alfath - 0110222206**

Teknik Informatika, STT Terpadu Nurul Fikri, Depok

E-mail: [hisy22206ti@student.nurulfikri.ac.id](mailto:hisy22206ti@student.nurulfikri.ac.id)

**Abstract.** Praktikum ini melatih kemampuan membaca, mengolah, dan membagi dataset menggunakan pandas dan scikit-learn di lingkungan Python. Mulai dari membaca file CSV yang berisi data kompleks terkait atribut cuaca dan pengguna, kemudian membagi data menjadi subset training, validation, dan testing untuk persiapan pemodelan machine learning. Visualisasi dan eksplorasi data menggunakan statistik deskriptif serta grafik juga dilakukan untuk memahami karakteristik data. Praktikum mandiri memperdalam pengalaman dengan dataset nyata yang lebih besar dan beragam, serta mengaplikasikan teknik pembagian data secara mandiri agar diperoleh model yang siap diuji dan dioptimalkan. Keseluruhan proses ini menyiapkan dasar kuat dalam alur kerja analisis data dan machine learning yang efektif dan terstruktur.

1. Praktikum 02

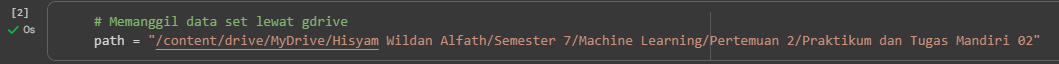
* Menghubungkan Colab dengan Google Drive

A screen shot of a computer

AI-generated content may be incorrect.

Kode ini memanggil modul drive dari google.colab untuk menghubungkan (mount) Google Drive ke lingkungan Google Colab, dan outputnya adalah muncul pesan “Mounted at /content/drive” yang menandakan folder Google Drive sekarang sudah bisa diakses pada path tersebut di Colab.

* Mendefinisikan Path Data



Baris ini mendefinisikan variabel bernama path yang berisi lokasi folder spesifik di Google Drive untuk menyimpan dataset atau dokumen tugas, dan tidak menghasilkan output karena hanya deklarasi variabel; variabel ini nantinya bisa digunakan untuk mengambil atau memproses file di lokasi tersebut.

* Membaca file csv menggunakan pandas

A black and orange background

AI-generated content may be incorrect.

Kode ini mengimpor library pandas sebagai pd, lalu menggunakan fungsi pd.read\_csv untuk membaca file CSV dari Google Drive berdasarkan path yang diberikan. Outputnya adalah DataFrame yang menampilkan data dari file CSV tersebut, berupa tabel dengan kolom Gender, Height, Weight, dan Index serta berisi 500 baris data.

* Mendapatkan info data pada file CSV menggunakan pandas

A screenshot of a computer program

AI-generated content may be incorrect.

Kode ini menjalankan fungsi df.info() untuk menampilkan ringkasan informasi DataFrame, termasuk tipe data, jumlah nilai non-null di setiap kolom, nama kolom, dan penggunaan memori. Outputnya menunjukkan DataFrame memiliki 4 kolom (Gender, Height, Weight, Index), masing-masing dengan 500 data non-null dan tipe data yang sesuai.

* Menghitung mean kolom numerik

A screenshot of a computer code

AI-generated content may be incorrect.

Kode ini menggunakan df['Height'].mean() untuk menghitung rata-rata (mean) dari seluruh nilai pada kolom Height di DataFrame. Outputnya adalah nilai mean tinggi, yaitu 169.944, dengan tipe data np.float64.

* Menghitung median kolom numerik

**A screen shot of a computer

AI-generated content may be incorrect.**

Kode ini menggunakan df['Height'].median() untuk mencari nilai tengah (median) dari semua data pada kolom Height di DataFrame. Outputnya adalah median tinggi, yaitu 170.5.

* Mencari modus kolom numerik

A screenshot of a computer program

AI-generated content may be incorrect.

Kode ini menggunakan df['Height'].mode() untuk mencari nilai yang paling sering muncul (modus) pada kolom Height. Outputnya menunjukkan bahwa nilai modus pada kolom Height adalah 188, dan tipe datanya int64.

* Menghitung variansi kolom numerik

A screenshot of a computer

AI-generated content may be incorrect.

Kode ini menggunakan df.var(numeric\_only=True) untuk menghitung nilai variansi dari setiap kolom numerik pada DataFrame. Outputnya menunjukkan variansi untuk kolom Height adalah 268.149162, Weight adalah 1048.633267, dan Index adalah 1.836168, semuanya bertipe float64.

* Menghitung standar deviasi kolom numerik

A screenshot of a computer

AI-generated content may be incorrect.

Kode ini menggunakan df.std(numeric\_only=True) untuk menghitung nilai standar deviasi dari setiap kolom numerik pada DataFrame. Outputnya menunjukkan standar deviasi untuk kolom Height adalah 16.375261, Weight adalah 32.382607, dan Index adalah 1.355053, semuanya bertipe float64.

* Menghitung kuartil dan IQR kolom numerik

A screenshot of a computer program

AI-generated content may be incorrect.

Kode ini menggunakan df['Height'].quantile(0.25) untuk mencari kuartil pertama (Q1) dan df['Height'].quantile(0.75) untuk kuartil ketiga (Q3) dari kolom Height, lalu menghitung IQR (Interquartile Range) sebagai selisih Q3 - Q1. Outputnya menunjukkan nilai Q1 sebesar 156.0, Q3 sebesar 184.0, dan IQR sebesar 28.0.

* Statistik deskriptif kolom numerik

A screenshot of a computer

AI-generated content may be incorrect.

Kode ini menggunakan df.describe() untuk menghasilkan ringkasan statistik deskriptif pada semua kolom numerik DataFrame, seperti count, mean, std (standar deviasi), nilai minimum, persentil (25%, 50%, 75%), dan nilai maksimum untuk masing-masing kolom Height, Weight, dan Index. Outputnya berupa tabel ringkasan statistik dari ketiga kolom numerik tersebut.

* Matriks korelasi kolom numerik

A screenshot of a computer program

AI-generated content may be incorrect.

Kode ini menggunakan df.corr(numeric\_only=True) untuk menghitung matriks korelasi antar semua kolom numerik, dan hasilnya disimpan pada variabel correlation\_matrix. Output yang ditampilkan berupa tabel matriks korelasi yang menunjukkan hubungan antar kolom Height, Weight, dan Index, di mana masing-masing nilai menunjukkan derajat hubungan antar kolom tersebut.

* Visualisasi boxplot kolom Height dan Weight

A screenshot of a graph

AI-generated content may be incorrect.

Kode ini menggunakan df.boxplot(column=['Height', 'Weight']) untuk membuat visualisasi boxplot dua kolom, yaitu Height dan Weight, sehingga memudahkan analisis distribusi, letak median, serta deteksi outlier pada kedua data. Outputnya berupa grafik boxplot berdampingan untuk kolom Height dan Weight.

* Visualisasi histogram kolom Height

A screenshot of a computer

AI-generated content may be incorrect.

Kode ini membuat histogram untuk kolom Height menggunakan matplotlib dengan pembagian dalam 5 interval (bins), warna pink, dan label pada sumbu x serta y. Outputnya adalah grafik histogram yang memperlihatkan distribusi frekuensi nilai Height pada lima rentang interval berbeda.

* Scatter plot untuk korelasi positif

A screen shot of a computer

AI-generated content may be incorrect.

Kode ini membuat DataFrame sederhana dua kolom (Nilai1 dan Nilai2) lalu memvisualisasikan datanya dalam bentuk scatter plot menggunakan matplotlib, di mana setiap titik berwarna biru dan diberi marker bulat. Outputnya adalah grafik scatter plot yang memperlihatkan hubungan positif antara Nilai1 dan Nilai2, dilengkapi judul, label sumbu, dan grid pada plot.

* Scatter plot untuk korelasi negatif

A screenshot of a computer screen

AI-generated content may be incorrect.

Kode ini membuat DataFrame sederhana dua kolom (Nilai1 dan Nilai2), lalu divisualisasikan dalam scatter plot menggunakan matplotlib dengan warna merah dan marker bulat. Outputnya adalah grafik scatter plot yang menunjukkan hubungan negatif antara Nilai1 dan Nilai2, lengkap dengan judul, label pada sumbu, dan grid pada plot.

Kesimpulan praktikum ini adalah bahwa library Python seperti pandas, numpy, dan matplotlib sangat powerful dan efisien untuk membangun alur kerja analisis data mulai dari membaca data, eksplorasi data deskriptif, hingga visualisasi statistik dan hubungan antar variabel. Dengan alat-alat ini, data dapat diolah secara cepat dan visualisasi seperti boxplot, histogram, scatter plot sangat membantu dalam memahami distribusi data dan korelasi antar variabel. Praktikum ini memperlihatkan pentingnya langkah awal eksplorasi data sebagai fondasi analisis lebih lanjut di bidang machine learning atau ilmu data secara umum. Visualisasi yang tepat akan mempercepat insight dan pengambilan keputusan berdasarkan data.

1. **Praktikum Mandiri 02**

* Membaca file CSV dengan pandas

A screenshot of a computer

AI-generated content may be incorrect.

Kode ini menggunakan pandas untuk membaca file CSV dari Google Drive dan menampilkannya sebagai DataFrame dengan 731 baris dan 16 kolom, berisi data harian yang komprehensif termasuk tanggal, musim, temperatur, kelembaban, dan jumlah pengguna casual serta terdaftar. Outputnya adalah tabel DataFrame yang merepresentasikan data lengkap tersebut.

* Membagi dataset menjadi data training, validation, dan testing

A screen shot of a computer program

AI-generated content may be incorrect.

Kode ini mengimpor library pandas dan fungsi train\_test\_split dari sklearn untuk membagi dataset menjadi subset train, validation, dan test. Pengaturan display di pandas dibuat agar semua kolom terlihat secara penuh. File CSV dibaca dari Google Drive dan disimpan ke df2. Dataset diformat menjadi tiga bagian dengan ukuran 80% untuk training dan 20% testing, lalu dari data training dibuat lagi subset validation sebesar 10%. Selanjutnya, dibuat gaya tabel agar hasil tampilan DataFrame lebih rapi. Terakhir, dicetak jumlah data dan ditampilkan lima baris awal dari masing-masing subset data tersebut dengan format yang sudah ditata agar mudah dibaca secara terpisah.

* Menampilkan data training, validation, dan testing hasil pembagian dataset

A screenshot of a computer screen

AI-generated content may be incorrect.

Kode ini mencetak jumlah data dan menampilkan lima baris awal dari masing-masing subset dataset training, validation, dan testing dengan format tabel yang sudah diberi gaya agar mudah dibaca dan dipahami. Outputnya menampilkan data dengan kolom lengkap seperti tanggal, musim, suhu, kelembaban, dan jumlah pengguna sehingga pengguna dapat memantau hasil pembagian data secara jelas dan terpisah.

Kesimpulan setelah mengerjakan praktikum mandiri adalah bahwa proses membaca dan membagi dataset menggunakan pandas dan scikit-learn penting dilakukan untuk mempersiapkan data dalam machine learning. Pembagian menjadi data training, validation, dan testing memungkinkan pelatihan model yang lebih handal dan evaluasi yang valid. Melalui praktikum ini, pemahaman mengenai alur persiapan data yang benar serta kemampuan menggunakan teknik pembagian data secara tepat dapat diperkuat sebagai dasar penting dalam proyek machine learning selanjutnya.

Link Praktikum dan Tugas Mandiri 02:

<https://github.com/HisyamWildan/TI03_HisyamW.A_0110222206/blob/main/Pertemuan%202/Praktikum%20dan%20Tugas%20Mandiri%2002/notebooks/praktikum02.ipynb>