

SILABUS TRAINING DATA SCIENCE

Data Analysis using Python

Data tidak akan bermanfaat apabila tidak memberikan informasi atau pengetahuan yang dapat mempengaruhi pengambilan keputusan sebuah perusahaan atau organisasi. Oleh karena itulah seorang Data Analyst berperan penting disini dalam mengambil data dari beragam sumber dan beragam format, eksplorasi data, pembersihan dan transformasi data, menganalisa data secara deskriptif, memodelkan data, dan memvisualisasikannya agar mudah untuk dianalisa sehingga dapat memberikan wawasan yang dapat dijadikan dasar untuk menjawab pertanyaan bisnis yang spesifik.

Materi training ini membahas tentang implementasi bahasa Python untuk penggunaan package-package yang berhubungan dengan siklus data analysis. Kemampuan data analysis ini merupakan tahap prasyarat untuk melangkah lebih lanjut ke sub topik Data Science lainnya seperti Machine Learning dan Deep Learning.

Setelah mengikuti training ini peserta diharapkan dapat memahami konsep dan mengimplementasikan bahasa Python untuk kebutuhan analisa data yang dapat membantu dalam pengambilan keputusan bisnis organisasi atau perusahaan.

Materi training disusun dari berbagai sumber dan media pembelajaran. Training disampaikan dalam bentuk pengenalan konsep, teori, dan praktek dalam porsi yang seimbang, serta dipandu oleh trainer atau instruktur yang telah berpengalaman di bidangnya.

Durasi dan Waktu

- Durasi Pelaksanaan: 3 (Tiga) Hari
- Waktu Pelaksanaan: 09.00 - 15.00 WIB

Prasyarat

- Peserta harus sudah memiliki pengetahuan dasar pemrograman Python seperti: penggunaan library, variabel, tipe data, struktur data (list, tuple, dictionary, set, array), struktur keputusan, struktur perulangan, pembuatan fungsi, dan yang lainnya.

Outline Materi

■ Instalasi dan Konfigurasi Tools

Awal pertemuan training ini akan diawali dengan pembahasan tentang instalasi dan konfigurasi tools yang akan digunakan selama training. Tools yang akan digunakan yaitu JupyterLab yang dapat diinstal melalui [Pip](#) (packager installer for Python) atau distribusi [Anaconda](#). Ekstensi JupyterLab dan Google Colab juga dapat diinstal di [Visual Studio Code](#) sehingga dapat dijadikan sebagai pilihan utama tools lainnya yang dapat digunakan.

■ Definisi dan Alur Kerja Data Analisis

Sebelum masuk ke pembahasan materi inti dari analisa data kita harus mengetahui terlebih dahulu definisi dan alur kerja atau proses dari analisa data.

■ Import Dataset

Objek utama dari analisa data yaitu dataset yang diimport atau di-load ke dalam data frame. Akuisisi atau load data dapat dibaca dari beragam format salah satunya yaitu file csv. Data frame diimport dari library Pandas yang nantinya diolah untuk data cleaning yang dikombinasikan dengan library Numpy. Disini kita akan eksplorasi fitur-fitur yang dimiliki oleh data frame seperti menampilkan data dengan beragam opsi, konfigurasi header, pembersihan data null, menampilkan informasi struktur data frame, menyimpan dataset, dan menampilkan deskripsi statistik. Dari hasil eksplorasi ini akan didapatkan wawasan dasar (basic insight) dari dataset yang telah diimport.

■ Pembersihan dan Transformasi Data

Disini kita akan membahas tentang data wrangling (cleaning dan transformasi) untuk mengidentifikasi dan menangani nilai yang kosong atau null yang dapat diganti dengan nilai lain seperti rata-rata, frekuensi, atau nilai lainnya. Format tipe data kolom ke tipe data yang lain agar dapat diproses dengan baik dan memberikan hasil yang sesuai. Standarisasi data juga diperlukan apabila data set memiliki format yang tidak umum yang kita anggap akan memberikan kesulitan interpretasi data. Data yang sifatnya kategori akan lebih mudah dianalisa dengan fitur binning yang hasilnya dapat divisualisasikan dengan grafik yang terdapat di dalam module matplotlib. Teknik lain yang cukup penting juga yaitu normalisasi data yang akan mengubah dataset ke dalam interval yang konsisten untuk mengurangi bias yang disebabkan oleh adanya perbedaan skala antar variabel yang ada.

■ Analisa Data Eksploratif

Sebelum dataset dianalisa lebih lanjut maka harus dilakukan eksplorasi data terlebih dahulu untuk mendapatkan pemahaman umum. Pada tahap ini akan

didapatkan beberapa informasi seperti karakteristik fitur-fitur yang ada untuk memprediksikan nilai, menganalisa pola fitur dengan visualisasi, menampilkan korelasi antar fitur untuk melihat seberapa berpengaruh variabel-variabel yang ada dalam memprediksikan nilai. Korelasi dapat ditampilkan secara visual dengan plot regres (regplot) yang merupakan gabungan scatter plot dengan garis regresi linear. Selain regplot juga akan dipelajari penggunaan boxplot untuk menampilkan grafik dari kolom yang sifatnya kategori. Pada tahap ini akan disinggung kembali statistik deskriptif untuk analisa data. Terkadang untuk kasus tertentu diperlukan penggunaan visualisasi grafik yang sesuai dengan kebutuhan seperti grafik heatmap yang mengambil data dari hasil pengelompokan atau hasil pivot. Terakhir yang tidak kalah pentingnya yaitu metrik korelasi yang akan menggunakan Pearson Correlation dan p-values.

■ Pengembangan Model

Data yang sudah dibersihkan, ditransformasi, dan dieksplorasi telah siap untuk digunakan sebagai sumber data untuk pengembangan model. Model adalah sebuah persamaan matematika yang digunakan untuk keperluan khusus misalnya memprediksi harga mobil berdasarkan fitur-fitur tertentu. Pada sesi ini dibahas beberapa model seperti regresi linier, multi regresi linear, dan regresi polinomial. Dari beberapa model yang digunakan akan dipilih model yang paling akurat untuk kasus tertentu. Tentunya pemilihan model tersebut merupakan hasil evaluasi dari perhitungan metrik-metrik seperti R-squared dan Mean Squared Error (MSE). Visualisasi disini juga cukup penting untuk melihat seberapa cocok atau akurasi model terhadap data prediksi. Satu fitur yang cukup membantu dalam penulisan kode yang dibahas di sesi ini yaitu Pipeline yang berfungsi untuk menyederhanakan proses-proses seperti cleaning dan transformasi.

■ Evaluasi dan Perbaikan Model

Evaluasi dan perbaikan model lebih lanjut dibahas pada sesi terakhir ini. Beberapa teknik untuk meningkatkan performa dataset yang dibahas disini yaitu memisahkan data training dan data test, dan validasi silang. Selain itu juga dibahas tentang penggunaan Regresi Ridge dengan mengatur hyperparameter dan Grid Search untuk identifikasi hyperparameter mana yang memberikan hasil yang terbaik. Semua teknik tersebut ditujukan agar model tidak overfitting dan underfitting. Pada akhirnya semua alur kerja atau proses telah dilakukan sehingga kita akan memiliki wawasan untuk memberikan keputusan berdasarkan hasil yang didapatkan dari semua proses tersebut.

Note

- Reguler Training dan Online Training tidak ada minimum kuota peserta.
- In-House Training dan Off-Site Training minimum kuota peserta 8 orang.
- Online Training menggunakan software Zoom Meeting.
- Hubungi kami untuk biaya training serta pembuatan Surat Penawaran.