Nama: Muhammad Javier
NIM: 24/546674/PA/23210

TEORI - PENUGASAN OPEN RECRUITMENT

DIVISI DATA SCIENCE DAN ARTIFICIAL INTELLIGENCE

2024/2025

1. Dengan menggunakan kalimat Anda sendiri, jelaskan yang dimaksud dengan EDA (Exploratory Data Analysis)! Mengapa hal tersebut penting dalam proses analisis data?

   Jawaban: **EDA (Exploratory Data Analysis) adalah proses dimana kita memahami isi dari keseluruhan data dan mencari insight dari sebuah data, proses ini sangat penting karena menentukan langkah selanjutnya yang akan kita terapkan terhadap data kita, misal handling missing value, categorical encoding, memilih model yang sesuai, dll.**

2. Jelaskan perbedaan antara supervised learning, unsupervised learning, dan reinforcement learning! Termasuk kategori yang manakah problemset pada penugasan open recruitment ini?

   Jawaban: **Supervised Learning, data yang akan kita 'latih' memiliki label atau target column sehingga komputer akan dituntut untuk belajar berdasarkan training set yang diberi, contoh: regression dan classification. Lain halnya dengan Unsupervised Learning, training set yang akan kita 'latih' tidak memiliki output label sehingga komputer akan dibiarkan untuk belajar sendiri atau menemukan pola dari training set. Untuk Reinforcement Learning, simpelnya, komputer akan belajar sendiri dari lingkungan sekitar, algoritma komputer akan melakukan trial and error untuk menemukan pola.**

3. Apa yang dimaksud dengan overfitting dan underfitting dalam konteks machine learning? Apakah dalam pengerjaan penugasan praktek Anda mengalami salah satu atau kedua masalah tersebut? Bagaimana Anda menanganinya?

   Jawaban: **Overfitting adalah kondisi dimana model terlalu 'fit' dengan training data, kondisi ini akan memiliki hasil yang optimal di training data, namun akan buruk di test data. Kebalikan dari overfitting, underfitting adalah kondisi dimana model terlalu 'lemah' dalam melatih data, gagal menangkap pola dari data dan kondisi ini memiliki nilai evaluasi yang buruk di training data maupun test data. Untuk pengerjaan penugasan open recruitment ini, saya menemukan adanya overfitting, saya mencoba menggunakan teknik cross-validation. Simpelnya, cross-validation membagi training data menjadi beberapa subset dan mengujinya dengan model, lalu evaluation score pada masing masing subset akan ditinjau ulang.**

4. Seandainya dalam proses prediksi penugasan *problemset* diperbolehkan menambahkan data eksternal, apakah Anda akan menggunakan data eksternal? Jika iya, data apa yang akan Anda gunakan dan jelaskan alasannya! (NB: selain data primer harga laptop dengan spesifikasi yang sama, contoh: data harga laptop di marketplace)

   Jawaban: **Jika saya awam dengan data yang diberikan, mungkin menggunakan data eksternal akan sangat membantu pada tahap EDA. Case Understanding sangat penting dalam memahami situasi apa yang sebenarnya terjadi pada data. Dengan data eksternal, saya dapat memahami lebih luas terhadap apa yang akan saya lakukan pada dataset yang diberikan**.

5. Bagaimana tanggapan dan evaluasi Anda terhadap problem set pada penugasan praktek dan soal teori pada proses open recruitment ini?

Jawaban: **Problem yang mungkin sering ditemukan di dunia kerja, yaitu seputar gaji. Pada awal penugasan langsung mendapat gambaran seperti apa isi dataset, mungkin karena problem seperti ini tidak terlalu susah data understanding-nya. Struggle di bagian data preprocessing karena ada feature yang punya *high cardinality*, dan harus menentukan grouping nya seperti apa dan *encoding method*-nya. Mungkin next bakal cari tau penyelesaian yang baik dari dataset ini gimana.:)**

~ Selamat mengerjakan! ~