**Nama : Satya Wira Pramudita**

**NIM : 24/543649/PA/23102**

**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?
   1. EDA merupakan visualisasi dari bentuk data dan analisis hubungan antar data. Visualisasi membantu kita melihat bentuk distribusi data dan menunjukkan data *outlier*. Untuk menentukan hubungan antara fitur dalam dataset, biasanya digunakan heatmapping. Dengan mengetahui hal-hal tersebut kita dapat menentukan cara untuk memproses data. Semisal ada data yang kosong, dengan melihat distribusi kita dapat menentukan metode apa yang digunakan untuk mengisi data kosong tersebut, seperti mean, median, atau modus.
2. Jelaskan perbedaan antara *supervised learning*, *unsupervised learning*, dan *reinforcement learning*! Termasuk kategori yang manakah *problemset* pada penugasan *open recruitment* ini?
   1. Supervised learning adalah metode machine learning dimana dataset yang digunakan untuk melatih model memiliki input dan output yang telah diketahui. Informasi yang telah didapatkan dari data akan dipetakan ke dataset yang serupa untuk memprediksi output dari dataset serupa tersebut.
   2. Unsupervised learning adalah metode machine learning dimana model dilatih menggunakan dataset yang tidak memiliki known output. Unsupervised learning digunakan untuk mengelompokkan data atau mencari pola dalam data.
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?
   1. Overfitting terjadi Ketika suatu model menjadi terlalu baik dalam mempelajari satu dataset. Model akan sangat baik dalam memprediksi data dari dataset tersebut, namun ketika diberi dataset yang tidak familier, prediksi model akan menjadi tidak akurat.
   2. Underfitting terjadi Ketika model tidak mendapatkan informasi yang cukup saat mempelajari data. Hal ini dapat terjadi apabila model yang digunakan terlalu sederhana untuk dataset atau iterasi training yang terlalu sedikit.
   3. Dalam pengerjaan tugas, saya menggunakan model yang berbasis Decision Tree untuk menghindari *underfitting*. Namun penggunaan model Decision Tree rentan Terhadap overfitting sehingga saya menggunakan *cross-validation* method yaitu K-Fold untuk memitigasi overfitting.
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*)
   1. Menurut saya data eksternal dapat menambah akurasi prediksi model. Data eksternal yang mungkin ingin saya tambahkan yaitu work\_hour. Dengan menambahkan work\_hour, kita dapat memeberikan model data kuantitatif yang dapat mebantu model memahami trend yang kontinu.
5. Bagaimana tanggapan dan evaluasi Anda terhadap *problem set* pada penugasan.
   1. Menurut saya problem set yang diberikan cukup baik. Dengan arti dataset yang diberikan cukup bersih dan tidak memiliki banyak data null. Yang menurut saya cukup *challenging* dalam penugasan ini yaitu memilih metode encoding. Sebab ada fitur yang memiliki banyak sekali unique values. Menurut saya metode encoding yang saya gunakan sudah cukup, saya menggabungkan one-hot encoding, label encoding, dan target encoding.