Nama : Widad Muhammad Rafi
NIM : 24/545635/PA/23190

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?

Jawab :

EDA (Exploratory Data Analysis) adalah proses eksplorasi data untuk memahami isi dataset secara keseluruhan, mendeteksi pola, menganalisis hubungan antar fitur, serta mengidentifikasi masalah seperti missing values, outliers, atau distribusi data yang tidak normal. Tahapan ini penting untuk menentukan langkah selanjutnya dalam mengolah data, seperti pemilihan metode preprocessing (handling missing values, categorical encoding), algoritma yang sesuai, serta evaluasi model. EDA juga membantu dalam menentukan relevansi fitur, sehingga model yang dikembangkan dapat bekerja lebih optimal.

2. Jelaskan perbedaan antara *supervised learning*, *unsupervised learning*, dan

reinforcement learning! Termasuk kategori yang manakah *problemset* pada penugasan *open recruitment* ini?

Jawab :

- **Supervised Learning**: Proses *machine learning* dengan data yang memiliki label atau target. Model dilatih untuk mempelajari hubungan antara fitur (input) dan label (output). Contohnya adalah regresi dan klasifikasi.

  Analogi sederhana : Bayangkan anda sedang belajar mengenali jenis-jenis buah. Teman anda memberikan foto-foto buah seperti apel, jeruk, dan pisang, lalu memberi tahu nama masing-masing buahnya. Setelah melihat banyak contoh, anda mulai bisa mengenali dan menebak jenis buah di foto baru berdasarkan apa yang telah Anda pelajari. Perumpamaan ini mirip dengan *supervised learning*, di mana model belajar dari data yang sudah diberi jawaban (label).

- **Unsupervised Learning**: Proses pembelajaran mesin tanpa label. Model diminta untuk menemukan pola, struktur, atau pengelompokan data secara mandiri. Contohnya adalah clustering

  Analogi sederhana : Bayangkan kita masuk ke perpustakaan besar yang belum pernah kita kunjungi sebelumnya. Buku-buku di sana tidak memiliki kategori

atau label, seperti novel, sejarah, atau sains. Kita mulai memperhatikan kesamaan di antara buku-buku tersebut, seperti ukuran, warna sampul, atau gaya tulisan, lalu kita mengelompokkan buku-buku tersebut berdasarkan kemiripannya. Perumpamaan ini mirip dengan *unsupervised learning*, di mana model mencoba mengelompokkan atau mengorganisasi data tanpa bantuan label.

- **Reinforcement Learning**: Proses pembelajaran mesin dengan sistem reward (hadiah) dan punishment (hukuman). Model belajar melalui trial and error, sehingga akhirnya dapat mengambil tindakan yang menghasilkan reward dalam jangka panjang.

  Analogi sederhana : Bayangkan Anda sedang mengajarkan anak kecil cara bermain game. Anak tersebut mencoba berbagai tombol untuk menjalankan karakter dalam game. Ketika tombol yang ditekan membuat karakter bisa mengalahkan musuh, Anda memujinya (reward). Tapi jika tombol yang ditekan membuat karakter kalah, Anda memberi tahu bahwa itu salah (punishment). Setelah mencoba berkali-kali, anak tersebut akan tahu tombol mana yang memungkinkan dia untuk memenangkan level-level selanjutnya dalam video game tersebut.

Problemset pada penugasan ini termasuk *supervised learning*, karena data latih memiliki label berupa target (salary). Model bertujuan untuk mempelajari hubungan antara fitur-fitur dengan target tersebut.

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?

Jawab :

- **Overfitting**: Terjadi ketika model terlalu 'fit' atau terlalu menghafal dengan data latih sehingga hasilnya sangat baik di data latih, tetapi performanya buruk di data uji. Hal ini disebabkan model terlalu fokus pada detail dan noise di data latih, sehingga kehilangan generalisasi untuk data baru.

  Analogi sederhana : Bayangkan anda akan menjalani ujian. Sebelum ujian dimulai, Anda telah diberi latihan soal oleh dosen untuk belajar mandiri, tetapi daripada Anda mempelajari dan mencoba mengerti bagaimana cara pengerjaan soal dari latihan soal itu, Anda malah menghafal jawaban pada latihan soal itu sehingga anda mengalami *struggle* saat anda mengerjakan ujian karena bertemu dengan soal yang tipe yang berbeda atau versi modifikasi dari latihan

soal anda sebelumnya.

- **Underfitting**: Terjadi ketika model gagal menangkap pola dari data training karena kurangnya sampel atau model kurang belajar pada data training, bisa juga karena faktor-faktor lainnya sehingga model memiliki performa buruk baik di data latih maupun data uji.

  Analogi sederhana : Bayangkan anda akan menjalani ujian kalkulus, tetapi sejauh ini hal yang baru anda pelajari tentang matematika hanya dasar-dasarnya saja, seperti tambah, kali, kurang, bagi. Tentu saja dalam hal ini Anda pasti akan mendapatkan hasil yang buruk pada ujian kalkulus ataupun latihan-latihan soal atau kuis sebelum ujian datang.

Pada penugasan ini, saya menemukan adanya *overfitting*. Untuk mengatasinya, saya menghindari menggunakan teknik one hot-encoding atau teknik encoding lainnya yang bisa menyebabkan ledakan kolom, saya juga menggunakan model-model yang lumayan *robust* terhadap *overfitting*, seperti XGBoost, LightGBM, Catboost, GradientBoosting dan Randomforest. Saya juga melakukan hyperparameter tuning untuk masing-masing model ini, lalu saya juga menggabungkan model-model tersebut menggunakan **Ensemble Learning Stacking**, yaitu menggabungkan beberapa *base model* menggunakan *meta model.*

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*)

Jawab :

Ya, saya akan menggunakan data eksternal. Salah satu data yang relevan adalah kurs mata uang (salary_currency), terutama untuk pekerja yang tidak menggunakan USD sebagai mata uang pembayaran. Data eksternal ini akan membantu model mengenali variasi gaji berdasarkan mata uang yang berbeda, sehingga prediksi menjadi lebih akurat. Selain itu, data eksternal juga dapat memberikan wawasan tambahan dalam tahap EDA, terutama karena saya awam dengan konteks dataset yang diberikan.

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

Jawab :

Datasetnya cukup menarik, sekaligus menantang karena kolom-kolom yang tersedia merupakan kolom-kolom bertipe kategorikal dan ada beberapa kolom data yang merupakan *high cardinality* yang menjadi salah satu penyebab data menjadi

overfitting. Saya banyak mendapatkan pelajaran juga terutama pada bagian encoding dan juga modelling, seperti ensemble learning dll.

~ Selamat mengerjakan! ~