📊 Project: Employee Income & Segmentation Analysis

🧩 MAJU JAYA DIGITAL TECHNOLOGY – Data Science Portfolio Project

1. Business Understanding

Latar Belakang:  
Perusahaan ingin memahami faktor-faktor yang mempengaruhi pendapatan karyawan, memprediksi pendapatan karyawan baru, dan/atau mengelompokkan karyawan berdasarkan karakteristik tertentu.  
Informasi ini dapat digunakan untuk strategi kompensasi, rekrutmen, pelatihan, dan retensi karyawan.

Tujuan Proyek:

* Membangun model prediktif untuk memperkirakan monthly\_income (Regresi) dan/atau income\_class (Klasifikasi).
* Melakukan clustering untuk menemukan kelompok karyawan dengan pola karakteristik yang mirip.
* Memberikan insight yang dapat digunakan untuk pengambilan keputusan manajemen.

Pilihan Modeling:  
Siswa boleh memilih salah satu atau kombinasi dari:

1. Regresi – memprediksi pendapatan bulanan.
2. Klasifikasi – memprediksi kelas pendapatan (Low/High).
3. Clustering – mengelompokkan karyawan tanpa label untuk segmentasi.

2. Data Understanding

Dataset Kolom:

* employee\_id : ID unik karyawan
* age : Usia karyawan
* gender : Jenis kelamin
* marital\_status : Status pernikahan
* city : Kota tempat bekerja
* education\_level : Tingkat pendidikan terakhir
* years\_experience : Lama pengalaman kerja
* weekly\_hours : Jam kerja per minggu
* department : Departemen tempat bekerja
* bonus\_percentage : Persentase bonus tahunan
* performance\_score : Skor kinerja
* overtime\_hours : Jam lembur per bulan
* monthly\_income : Pendapatan bulanan (target regresi)
* income\_class : Kelas pendapatan (target klasifikasi)

Tugas Siswa:

* Lakukan EDA untuk memahami distribusi, hubungan antar variabel, dan potensi masalah data.
* Identifikasi variabel numerik vs kategorikal.

3. Data Preparation

Langkah minimal yang perlu dilakukan:

* Cleaning: Tangani missing/null values & outliers.
* Encoding: Konversi variabel kategorikal ke numerik.
* Scaling: (Opsional) untuk data numerik, terutama untuk clustering & model sensitif skala.
* Feature Engineering: Buat fitur baru yang relevan (opsional).

4. Modeling

Pilih pendekatan sesuai tujuan:

A. Regresi

* Linear Regression
* Random Forest Regressor
* XGBoost Regressor

B. Klasifikasi

* Logistic Regression
* Random Forest Classifier
* XGBoost Classifier

C. Clustering

* K-Means
* Hierarchical Clustering
* DBSCAN

Catatan:

* Jika memilih clustering, pastikan menjelaskan metode menentukan jumlah cluster (misal elbow method, silhouette score) dan interpretasi hasilnya.
* Boleh menggabungkan dua pendekatan (misalnya clustering → klasifikasi).

5. Evaluation

Gunakan metrik yang tepat:

* Regresi: MAE, MSE, RMSE, R²
* Klasifikasi: Accuracy, Precision, Recall, F1-Score, ROC-AUC
* Clustering: Silhouette Score, Davies-Bouldin Index, interpretasi visual cluster

6. Deployment

**Opsional** untuk portofolio:

* Deploy ke Streamlit, Flask API, atau buat dashboard Tableau/Power BI.
* Untuk clustering: buat visualisasi interaktif untuk melihat segmen karyawan.

📌 Deliverables

1. Notebook / Script – proses lengkap (data → model → evaluasi).
2. Laporan / Slide – insight dan rekomendasi.
3. Deployment – opsional, untuk nilai tambah portofolio.