Capstone Project Briefing Document

Title: 'Instilit' - Scalable Global Salary Intelligence System (Enhanced Dataset)

1. Problem Statement

A multinational HR analytics company Instilit is building an AI platform, a predictive engine to help recruiters and employers with benchmark compensation globally.

The system should predict total salary packages (base + bonus + stock) for different roles, geographies, and experience levels.

The Team has challenges in dealing with inconsistencies due to the variations in job roles and geographical standards.

2. Business Case

- Offer data-backed salary benchmarks to employers across industries
- Improve equity in compensation decisions
- Integrate with HRMS platforms (PostgreSQL) and job boards for real-time suggestions
- Analyze patterns in compensation by location, experience, education, and company size

3. Dataset Description

File: Software_Salaries.csv

Column Name	Description
job_title	Title or role of the employee (e.g., Software Engineer, Data Scientist). May contain typos/noise.
experience_level	Seniority level of the employee (e.g., Entry, Mid, Senior). 20% values may be missing.
employment_type	Type of employment contract (e.g., Full-time, Contract, Internship). May contain missing values.
company_size	Size of the employing organization (e.g., Small, Medium, Large).
company_location	Country where the company is headquartered.
remote_ratio	Percentage of work done remotely (0 = on-site, 100 = fully remote).
salary_currency	Original currency in which salary was reported.
years_experience	Total years of professional experience (numeric).
base_salary	Fixed annual salary before bonuses or stock. May include outliers.
bonus	Annual bonus awarded to the employee (in original currency).
stock_options	Monetary value of stock options or RSUs (in original currency).
total_salary	Sum of base_salary + bonus + stock_options in original currency.
salary_in_usd	Normalized total salary converted to USD using external rates (baseline).
currency	Randomly assigned currency to simulate inconsistencies (e.g., USD, INR, EUR).
conversion_rate	Exchange rate used to convert salary figures to USD.
adjusted_total_usd	Final total compensation (base + bonus + stock) after applying conversion_rate.

4. Tools & Technologies

- pandas, scikit-learn, xgboost, numpy, SHAP
- MLflow for model tracking and versioning
- Apache Airflow for orchestration
- Creating pipelines for capturing Data Drift, Model Drift, and Concept Drift.
- Flask or FastAPI for model deployment