Data Scientist Interview Task: Multi-Output Flower Analysis System

Introduction & Scenario:

Imagine you're working with a botanical research institute or a premium floristry service. The goal is to develop a data-driven system that analyzes flower images to provide insights including flower type, dominant color, and biochemical compound estimates. As a Data Scientist, your role focuses on model development, data analysis, and deriving insights from the data.

🎯 Objective:

Design, train, and evaluate a multi-output machine learning model that can:

- Classify the flower type (e.g., daisy, rose, tulip).
- Predict the dominant color category (e.g., red, white, yellow).
- Estimate concentrations of key essential oils (Linalool, Geraniol, Citronellol).

💾 Dataset Guidance:

You may use a dataset from Kaggle or similar sources. If unavailable, simulate or prepare a sample dataset based on publicly available flower image datasets and biochemical annotations.

🛠️ Core Task Requirements:

Input: Flower image.

Output (JSON format preferred):

- predicted_flower_type: (string, e.g., 'rose')
- predicted_flower_color: (string, e.g., 'red')
- estimated_oil_concentrations:
  - Linalool: (float)
  - Geraniol: (float)
  - Citronellol: (float)

Modeling Approach:

- Use Python with scikit-learn, TensorFlow, or PyTorch or as preferred.
- Extract features using pre-trained CNN backbones (e.g., ResNet) as preferred.
- Employ multi-head architecture or ensemble models for multi-task learning.

Data Preprocessing:

- Preprocess images (resize, normalize).
- Augment data for robustness.
- Normalize oil concentration values (e.g., scale 0–1).
- Stratified split for training, validation, and testing.

Training & Evaluation:

- Losses:
  - Classification: Cross-Entropy
  - Regression: MSE or MAE
- Combine losses via weighted sum.
- Evaluation metrics:
  - Classification: Accuracy, F1-score, Confusion Matrix
  - Regression: MAE, MSE, R^2 Score

📊 Deliverables:

- Jupyter Notebook or Python scripts (.ipynb or .py)
- Trained model weights (optional)
- requirements.txt
- README.md with:
  - Project overview
  - Setup and installation
  - Dataset description
  - Preprocessing details
  - Modeling and evaluation steps
  - Inference steps
  - Key challenges and decisions

✨ Optional Enhancements:

- Model explainability (e.g., SHAP, Grad-CAM)
- Hyperparameter tuning (Optuna, GridSearchCV)
- Statistical analysis of feature importance
- Visualization of results and predictions