1. Core Machine Learning (ML)

A. Data Preprocessing & Feature Engineering

- Why scale features for SVM/K-means? When is it unnecessary?
- Why does PCA work better with standardized features?
- How to handle missing data (deletion vs. imputation)?
- Why might mean imputation or deleting rows be suboptimal?
- What happens if categorical variables are not encoded properly?
- How does feature scaling impact distance-based models (KNN, SVM)?
- Why is handling multicollinearity important? Detection methods.
- Why might removing an irrelevant feature decrease performance?
- How to detect and handle outliers?
- Why is feature selection important for interpretability in linear models?
- What are interaction features, and how do they improve models?
- Why should we avoid data leakage during preprocessing?

B. Model Evaluation & Metrics

- Why is accuracy misleading for imbalanced datasets? Alternatives (F1, AUC-PR).
- What's the difference between precision and recall? When to prioritize each?
- Why might a model with 95% accuracy still fail in production?
- How does a confusion matrix provide more insights than accuracy?
- Why might AUC-ROC be high but production performance poor?
- Why is R-squared misleading in regression?
- When to use MSE vs. MAE in regression?
- Why might a model with high accuracy lack business value?

C. Overfitting & Regularization

- How do L1/L2 regularization prevent overfitting? Differences?
- Why might decision trees overfit? How does pruning help?
- Why does increasing training data sometimes degrade performance?
- What is early stopping, and why is it helpful?
- Why might models perform worse after tuning/regularization?

- Can a model with high bias outperform a low-bias model?

D. Ensemble Methods

- Bagging vs. boosting: How to choose?
- Why does Random Forest's accuracy plateau with more trees?
- Why might XGBoost outperform DL on tabular data?
- Can ensemble methods overfit? Prevention strategies.
- Why might Random Forest perform worse than a single tree?
- What is stacking, and how does it work?

E. Imbalanced Data

- How to handle a heavily underrepresented class? (SMOTE, weighted loss).
- Why might SMOTE fail? Alternatives (ADASYN, cost-sensitive learning).
- Precision vs. recall for rare disease prediction.

F. Bias-Variance & Tradeoffs

- What is the bias-variance tradeoff? How to manage it?
- Why prefer simpler models (logistic regression) over complex ones?
- How do learning curves diagnose underfitting/overfitting?

G. Clustering & Dimensionality

- Why might K-means fail to find meaningful clusters?
- How to choose a distance metric in clustering?
- Why does the curse of dimensionality affect KNN? Solutions (PCA, t-SNE).

3. Time-Series Forecasting

- Why is stationarity important? Tests (ADF, KPSS).
- How does differencing help achieve stationarity?
- What if ADF says stationary but KPSS disagrees?
- How to handle missing timestamps or irregular intervals?

- Why might ARIMA struggle with structural breaks/outliers?

4. Practical Scenarios & Debugging

- **Fraud Detection**: High false positives? Adjust thresholds, optimize precision.

- **Chatbot Misunderstanding Dialects**: Add diverse training data, fine-tune embeddings.

- **Model Simplification**: Trade accuracy for interpretability (LIME, SHAP).

- **Cold-Start Recommendation**: Use content-based filtering or hybrid approaches.

- **Slow Training**: Check feature dimensions, redundant data, or algorithm choice.

- **Deployment Issues**: Why cloud-trained models fail on edge devices (quantization).