ML Questions

Q1. Why might a machine learning model fail even if you have a very large dataset? Give two possible causes.

Ans –

1. Poor Data Quality –  
   If the dataset is full of noise, mislabeled examples, missing values, or irrelevant features, then the model will learn misleading patterns.
2. Model Bias or Underfitting –  
   If the chosen model is too simple for the complexity of the problem, it won’t capture important patterns no matter how much data you feed it.

Q2. A model gives 95% accuracy on the training set but only 60% accuracy on test set. Which error (bias or variance) is more likely the problem, and why?

Ans - This usually means the model has memorized specifics of the training set—including noise—rather than learning general rules.

low bias (good fit to training data) but high variance (poor generalization). Condition of Overfitting.

Q3. In a medical dataset, age is missing for 20% of patients. Why might median imputation be better than mean imputation in this case?

Ans - Median imputation would likely be better here because age distributions in medical datasets are often skewed

Mean imputation is sensitive to extreme outliers, which could distort the imputed values.

Median imputation is more robust to outliers and better represents the “typical” patient’s age, preserving the distribution’s central tendency without being overly influenced by unusual ages.

In short: median = safer when the data is skewed or has outliers, which is common in age-related medical data.

Q4. You have a categorical feature with 1000 unique categories. Why might one-hot encoding be a poor choice?

Ans –

 High Dimensionality –  
1000 categories → 1000 new binary columns.  
This greatly increases the feature space, which can:

* Slow down training
* Increase memory usage
* Make the model prone to overfitting (curse of dimensionality)

 Sparsity –  
Each sample will have only 1 of those 1000 columns as “1” and the rest “0,” producing an extremely sparse (empty) matrix that doesn’t efficiently capture relationships between categories.

Q5. Can a model have both high bias and high variance at a same time? Give a example situation.

Ans - Yes — a model can have both high bias *and* high variance at the same time, though it’s less common than having one dominate.

Example situation:  
Imagine you’re using a very small, noisy dataset and training a simple linear regression model on a relationship that’s actually highly nonlinear.

* High bias: The linear model is too simple to capture the true complex relationship → even on the training set, predictions are systematically wrong.
* High variance: Because the dataset is tiny and noisy, the model’s parameters swing wildly if you collect a slightly different sample, leading to unstable predictions on the test set.

Q6. You trained a model with scaled features, but during prediction time you forget to scale the new input data. What happens, and why?

Ans - If you forget to scale the new input data at prediction time, the model’s outputs will likely be wrong or wildly off because the model’s learned parameters assume inputs are on the same scale as during training.

Why:

* During training, the features were transformed (e.g., standardized to mean = 0, std = 1).
* The model learned weights expecting those scaled values.
* If you feed raw, unscaled inputs, the magnitude of features will be completely different.
* This shifts them far outside the range the model was trained on, so even a well-trained model can make nonsense predictions.

Q7. Suppose your model's performance drops when you increase training data size. What could be causing this, and how would you diagnose it?

Ans –

Possible causes:

1. Data quality issues
2. Data leakage in smaller sets - Data leakage happens when your model accidentally gets information during training that it wouldn’t have in real-world predictions
3. Model capacity too small – If the model is underpowered (high bias), adding more diverse data exposes its inability to capture complexity, making accuracy drop.
4. Training procedure issues – Batch size, learning rate, or regularization might not be tuned for the larger dataset, leading to underfitting or poor convergence.

How to diagnose:

* Check data quality of the new portion
* Plot learning curves (train & validation accuracy vs. training set size) to see if the drop is due to underfitting, overfitting, or noise.
* Compare distributions (old vs. new data) using statistical tests or visualizations.
* Evaluate in chunks — train separately on old data and new data to see if the new portion is causing the drop.

Q8. You are working on an image classification task and reduce each image to an average pixel value as the only feature. Why might this make the model fail, even if the dataset is large?

Ans – Reducing each image to just its average pixel value destroys almost all the important information needed for classification.

* Loss of spatial structure: The arrangement of pixels (edges, shapes, textures) is crucial for distinguishing objects. Averaging flattens this into a single number, so a picture of a cat and a picture of a dog might have nearly the same average brightness.
* Loss of fine-grained features: Color distributions, gradients, and patterns are gone — the model can’t detect the actual visual differences between classes.
* Large dataset won’t help: Even with millions of samples, if the feature contains almost no class-specific information, the model can’t learn meaningful patterns.