ML Questions

Q1. You have model with high training accuracy but low testing accuracy. What does it indicate?

Ans :- If a model has high training accuracy but low testing accuracy, it indicates overfitting.

Overfitting happens when a model learns the training data *too well*, including its noise and specific patterns, which don’t generalize to unseen (test) data.

* Training accuracy high → Model performs very well on the data it was trained on.
* Testing accuracy low → Model fails to generalize to new, unseen data.

Q2. How can you improve the generalization ability of model?

Ans :- You can improve a model’s generalization ability by using techniques that prevent overfitting and help it perform well on unseen data.

Explanation with Techniques:

1. Regularization
2. Cross-validation
3. Early Stopping
4. Data Augmentation
5. More Training Data
6. Simplify the Model
7. Dropout (for neural networks)
8. Feature Selection/Dimensionality Reduction

Q3. Explain with example when a programming based approach would fail and a machine learning based approach would success.

Ans :- A programming-based approach is likely to fail when the problem involves complex, vague, or unpredictable patterns that are hard to describe with fixed rules. A machine learning approach succeeds in such cases by learning patterns automatically from data.

Example :- In a rule-based (programming) system, you would write specific conditions like “if the email contains ‘win money’ or ‘free offer’, mark it as spam.” However, spammers can easily bypass these rules by changing words or using new tricks, and it becomes very hard to keep updating rules manually.

In contrast, a machine learning system is trained on thousands of labeled emails (spam or not spam). It learns patterns like word usage, frequency, sender behavior, etc., and can identify spam—even when the exact words change.

Why ML Succeeds Here:

* Can handle huge amounts of data.
* Learns complex, hidden patterns.
* Adapts better to new or changing inputs.

Q4. Is it possible for a model to have both low bias and low variance? Explain.

Ans :- In practice, it is very difficult (but not impossible) for a model to have both low bias and low variance.

Explanation:

* Bias means the model is too simple and makes errors even on training data (underfitting).
* Variance means the model is too complex and performs well on training data but poorly on test data (overfitting).

Usually, there is a trade-off between bias and variance:

* Reducing bias increases variance.
* Reducing variance increases bias.

However, with:

* Good model selection
* Enough quality data
* Proper regularization

…it’s possible to achieve a balance where both bias and variance are low, and the model generalizes well.

Q5. How would you evaluate whether your machine learning model is good or not?

Ans :- To evaluate whether a machine learning model is good, you use performance metrics, validation techniques, and check how well it generalizes to new data.

Key Evaluation Methods:

1. Use Appropriate Metrics (depends on the problem):
   * Classification: Accuracy, Precision, Recall, F1-score, ROC-AUC
   * Regression: Mean Squared Error (MSE), Mean Absolute Error (MAE), R² Score
2. Check for Overfitting/Underfitting:
   * Large gap between training and test performance → overfitting.
   * Poor performance on both → underfitting.
3. Confusion Matrix (for classification)

Q6. The ability of a model to perform well on unseen data is called :- Generalization

There are 4 cases in ML

1. High bias high variance

2. High bias low variance

3. Low bias high variance

4. Low bias low variance

| Case | Feasibility | Suitability | Example |
| --- | --- | --- | --- |
| 1. High Bias, High Variance | Worst Case | Practically undesirable | A linear model trained on very noisy non-linear data with few samples. |
| 2. High Bias, Low Variance | Possible | Sometimes okay (if simplicity is key) | Using linear regression to model housing prices without enough features. |
| 3. Low Bias, High Variance | Common in practice | May work with regularization | Deep neural nets trained on small datasets. |
| 4. Low Bias, Low Variance | Best Case | Ideal, but hard to achieve | Well-tuned model with sufficient data & complexity (e.g., Random Forest on rich dataset). |

* Why it's bad: Model is too simple (can't capture complexity) and also changes too much with different data.
* Analogy: A bad student who memorizes wrong concepts — performs poorly on both practice and real tests.

Case 2: High Bias, Low Variance

* Why it's not ideal: Model is too simple, so it underfits — but at least it’s consistent across datasets.
* Analogy: A student who only studies summaries — always gives the same wrong answers.

Case 3: Low Bias, High Variance

* Why it’s risky: Model fits training data perfectly but fails to generalize (overfits).
* Analogy: A student who memorizes all practice questions but fails when new questions are asked.

Case 4: Low Bias, Low Variance (Best)

* Why it’s ideal: Model captures true relationships and generalizes well.
* Analogy: A smart student who deeply understands the subject — does well on any kind of exam.

