Propvivo – AI/ML Engineer Intern: Mock Online Assessment

Format: 2 Coding Questions + 1 AI/ML Practical Question + 5 Theory/MCQ Questions

---

Section A – Any 1 AI/ML Practical Question among below

Q1. Predict House Prices (ML Regression Task)

Problem Statement:
You are given a dataset with columns:
`['area', 'bedrooms', 'bathrooms', 'price']`

You need to build a simple **Linear Regression model** to predict house price based on area, bedrooms, and bathrooms.

Tasks:

1. Import and preprocess the dataset (handle missing values if any).
2. Split data into training and testing sets (80/20).
3. Train a linear regression model using **scikit-learn**.
4. Print R^2 score on test data.

Expected Output:

R2 Score on Test Data: 0.87

Tests: Data preprocessing, ML pipeline understanding, scikit-learn usage.

Q2. Image Classification (Conceptual Implementation)

Problem Statement:
You are training a Convolutional Neural Network (CNN) on the **MNIST digits dataset**.

Tasks:

1. Load the dataset from `tensorflow.keras.datasets.mnist`.
2. Normalize pixel values between 0 and 1.
3. Build a CNN model with at least:
    - One Conv2D layer
    - One MaxPooling2D layer
    - One Dense output layer with softmax
4. Print model summary.

Expected Output:

```
Model: "sequential"
```

```
 Layer (type)                 Output Shape
Param #

 conv2d (Conv2D)              (None, 26, 26, 32)
320

 ...

 dense (Dense)                (None, 10)
650

=================================================================

Total params: 8,450
```

Tests: Familiarity with TensorFlow/Keras, CNN basics, model design.

Q3. Data Analysis & Feature Engineering

Problem Statement:
You are given a DataFrame `df` with columns:
`['age', 'salary', 'city', 'purchased']`
You need to:

1. Handle missing values (fill numerical columns with mean, categorical with mode).
2. Encode categorical columns using LabelEncoder.
3. Scale the numeric features using StandardScaler.
4. Print the transformed data.

Expected Output Example:

```
  age  salary  city  purchased
0  0.23  -0.48   0       1
1 -0.77   0.62   1       0
```

Tests: Data preprocessing, feature engineering, pandas & sklearn familiarity.

2 Sample Questions of Coding like this

🧮 Coding Question 1 – Arrays

Problem:
You are given an integer array `nums` and an integer `k`.
Return the length of the longest subarray whose sum equals `k`.

Example:
Input: `nums = [1, 2, 3, -2, 5, 1], k = 6`
Output: `4`
Explanation: The subarray `[1, 2, 3]` and `[3, -2, 5]` both have sum `6`, the longest has length `4`.

Concept Tested: Prefix Sum + HashMap

🔁 Coding Question 2 – Strings

Problem:
Given a string `s`, return the length of the longest substring without repeating characters.

Example:
Input: `s = "abcabcbb"`
Output: `3`
Explanation: The answer is `"abc"`, with the length `3`.

Concept Tested: Sliding Window / Two Pointers

Section B – Conceptual / MCQs

Q4. (Machine Learning Concept)

Which of the following metrics is most suitable for **classification** problems?

Options:
A) Mean Squared Error
B) R^2 Score
C) Accuracy
D) Adjusted R^2

Q5. (Deep Learning)

What is the purpose of the **ReLU** activation function?

Options:
A) To introduce non-linearity
B) To prevent overfitting
C) To reduce dimensionality
D) To normalize input data

Q6. (Statistics & Data Analysis)

The **p-value** in hypothesis testing indicates:
Options:
A) Probability of Type II error
B) Strength of evidence against the null hypothesis
C) Confidence interval width
D) None of the above

Q7. (Cloud & MLOps)

Which of the following AWS services is primarily used for model deployment?

Options:
A) AWS Lambda
B) Amazon SageMaker
C) Amazon S3
D) Amazon EC2

Q8. (Programming / Python)

What will the following code output?

```
import numpy as np

a = np.array([1, 2, 3])
```

```
b = np.array([4, 5, 6])

print(a * b)
```

Options:
A) [4, 10, 18]
B) [5, 7, 9]
C) [1, 2, 3, 4, 5, 6]
D) Error

MCQ they might ask based on job profile-

Q1. What does a kernel do in an SVM?
A) Reduces dimensionality
✅ B) Transforms input data into higher-dimensional space
C) Normalizes data
D) Reduces bias

Q2. What is the difference between classification and regression?
A) Classification predicts continuous values, regression predicts categories
B) Both predict numeric outputs
✅ C) Classification predicts categories, regression predicts continuous values
D) Classification is unsupervised

Q3. What is bias in ML?
✅ A) Error due to simplifying assumptions in the model
B) Random noise in data
C) Difference between actual and predicted
D) Overfitting indicator

Q4. What does cross-validation help with?
✅ A) Checking model generalization
B) Measuring training speed
C) Hyperparameter scaling
D) Feature scaling

Q5. What are support vectors?
A) Hyperplanes in SVM
✅ B) Data points closest to the decision boundary
C) Outliers
D) Regularization parameters

Q6. What is PCA used for?
✅ A) Dimensionality reduction
B) Model selection
C) Clustering
D) Feature scaling

Q7. The term “Naive” in Naive Bayes refers to—
✅ A) The assumption of independence among features
B) Use of Gaussian distribution
C) Simplicity of implementation
D) Overfitting tendency

Q8. F1 Score is the harmonic mean of—
A) Accuracy and Recall
B) Accuracy and Precision
✅ C) Precision and Recall
D) Recall and Specificity

Q9. Random Forest is—
✅ A) An ensemble of decision trees using bagging
B) A clustering algorithm
C) A neural network
D) A regression-only model

Q10. What is the Bias-Variance tradeoff?
A) Tradeoff between accuracy and precision
✅ B) Balancing model simplicity and generalization
C) Tradeoff between data and model
D) Tradeoff between time and accuracy

🧩 Deep Learning (PyTorch & TensorFlow) – 10 Questions

Q11. What is an activation function used for?
✅ A) Introducing non-linearity
B) Optimizing weights
C) Increasing learning rate
D) Controlling epochs

Q12. Which function is most commonly used in hidden layers?
A) Sigmoid
✅ B) ReLU
C) Softmax
D) Linear

Q13. What is vanishing gradient problem?
✅ A) Gradients become too small to update weights effectively
B) Gradients explode
C) Loss doesn't change
D) Model overfits

Q14. Dropout helps to–
✅ A) Prevent overfitting
B) Increase batch size
C) Reduce learning rate
D) Improve gradient flow

Q15. CNNs are mainly used for–
✅ A) Image processing
B) Text summarization
C) Regression tasks
D) Clustering

Q16. LSTM networks solve–
✅ A) Vanishing gradient problem in RNNs
B) Classification problem
C) Reinforcement issues
D) Low learning rate

Q17. What is the difference between LSTM and GRU?
✅ A) GRU has fewer gates and is computationally efficient
B) LSTM has no gates
C) GRU cannot learn long-term dependencies
D) LSTM is unsupervised

Q18. What is the role of an attention mechanism?
✅ A) Focus on important parts of the input sequence
B) Reduces model size
C) Normalizes embeddings
D) Filters stopwords

Q19. What is BERT?
✅ A) Transformer-based language model using bidirectional context
B) RNN variant
C) GAN
D) Naive Bayes model

Q20. Transfer learning allows–
✅ A) Using a pre-trained model on a new related task
B) Compressing neural networks
C) Changing optimizer
D) Fine-tuning embeddings manually

💬 NLP (Natural Language Processing) – 10 Questions

Q21. What is Lemmatization?
✅ A) Reducing a word to its base dictionary form
B) Removing stopwords
C) Tokenizing text
D) Lowercasing

Q22. What is TF-IDF used for?
✅ A) Measuring word importance relative to a corpus
B) Stemming
C) Word embeddings
D) Text summarization

Q23. What are N-grams?
✅ A) Contiguous sequence of N items from text
B) Named entities
C) Stopwords
D) Lemmas

Q24. What is the key difference between stemming and lemmatization?
✅ A) Lemmatization uses linguistic rules, stemming is crude truncation
B) Both are same
C) Lemmatization is unsupervised
D) Stemming uses neural nets

Q25. What is Bag of Words (BoW)?
✅ A) Representing text as a frequency count of words
B) Sequence encoding
C) One-hot encoding
D) TF-IDF model

Q26. Perplexity measures–
✅ A) How well a language model predicts a sample
B) Word frequency
C) Data variance
D) Text entropy

Q27. What is Word2Vec?
✅ A) Neural embedding model for words
B) Stemming technique
C) Clustering method
D) Tokenization tool

Q28. Difference between NLP and NLU?
✅ A) NLP = processing, NLU = understanding
B) NLP = understanding, NLU = generation
C) Both are identical
D) NLU comes before NLP

Q29. What does Masked Language Modeling do?
✅ A) Predicts missing words in a sentence
B) Classifies text
C) Detects sentiment
D) Removes stopwords

Q30. POS tagging identifies–
✅ A) Grammatical role of each word
B) Sentiment
C) Named entities
D) Text summary

🤖 Generative AI – 10 Questions

Q31. Generative AI differs from traditional AI because–
✅ A) It generates new content rather than only predicting outcomes
B) It's rule-based
C) It cannot learn
D) It uses linear regression

Q32. GAN consists of–
✅ A) Generator and Discriminator
B) Encoder and Decoder
C) Attention layers
D) Tokenizers

Q33. What is mode collapse in GANs?
✅ A) Generator produces limited variety of outputs
B) Discriminator fails
C) Generator stops learning
D) Training diverges

Q34. What are diffusion models?
✅ A) Models that iteratively denoise data to generate samples
B) Variants of CNNs
C) Clustering models
D) RNNs with attention

Q35. What is latent space?
✅ A) Compressed representation of data features
B) Hidden neurons
C) Overfitting zone
D) Training buffer

Q36. What are hallucinations in LLMs?
✅ A) Generating incorrect but plausible outputs
B) Memory leaks
C) Misaligned attention
D) Syntax errors

Q37. Encoder-only transformers are used for—
✅ A) Understanding tasks (e.g., BERT)
B) Text generation
C) Translation
D) Diffusion

Q38. What is self-attention used for?
✅ A) Computing relationships between all words in a sequence
B) Regularization
C) Batch normalization
D) Memory reduction

Q39. What is RAG (Retrieval-Augmented Generation)?
✅ A) Combines document retrieval with generation
B) Only retrieves text
C) Only generates text
D) Fine-tunes embeddings

Q40. KL divergence in VAEs measures–
✅ A) Difference between learned and prior distribution
B) Loss gradient
C) Distance between centroids
D) Encoder accuracy

☁️ Data, Cloud & Coding – 10 Questions

Q41. One-hot encoding is used for–
✅ A) Categorical variable representation
B) Numerical scaling
C) Feature reduction
D) Feature selection

Q42. Outliers can be detected using–
✅ A) Boxplot or Z-score
B) Accuracy score
C) Gradient boosting
D) Confusion matrix

Q43. AWS S3 is used for–
✅ A) Object storage
B) Compute services
C) Serverless hosting
D) Real-time streaming

Q44. EC2 provides–
✅ A) Virtual machine instances
B) Container orchestration

C) Data storage
D) File sharing

Q45. What is Docker used for?
✅ A) Containerizing applications
B) File compression
C) Version control
D) Cloud storage

Q46. In SQL, finding duplicates can be done using—
✅ A) `GROUP BY` with `HAVING COUNT(*) > 1`
B) `SELECT DISTINCT`
C) `JOIN`
D) `WHERE`

Q47. Feature scaling ensures—
✅ A) All features contribute equally to model training
B) Model accuracy increases automatically
C) Data normalization
D) PCA success

Q48. Kubernetes is mainly used for—
✅ A) Container orchestration
B) Cloud storage
C) Model evaluation
D) Version control

Q49. ETL stands for—
✅ A) Extract, Transform, Load

B) Encode, Train, Learn
C) Evaluate, Test, Learn
D) Extract, Train, Load

Q50. Which of the following algorithms uses dynamic programming?
A) Binary Search
B) DFS
✅ C) Longest Common Subsequence
D) Greedy Knapsack