Machine Learning Interview Questions and Answers

📌 Step 1: Fundamentals of Machine Learning

1. What are feature selection techniques, and why are they important?

Feature selection techniques help in selecting the most relevant features from a dataset to improve model performance and reduce overfitting. Techniques include:

- **Filter Methods** (e.g., correlation, mutual information)
- **Wrapper Methods** (e.g., Recursive Feature Elimination - RFE)
- **Embedded Methods** (e.g., Lasso Regularization)

2. What is the difference between L1 and L2 regularization?

- **L1 Regularization (Lasso Regression)**: Shrinks some feature coefficients to zero, leading to sparse models.
- **L2 Regularization (Ridge Regression)**: Distributes penalties among coefficients, preventing large values but keeping all features.

3. What are Type 1 and Type 2 errors in hypothesis testing?

- **Type 1 Error (False Positive)**: Rejecting a true null hypothesis.
- **Type 2 Error (False Negative)**: Accepting a false null hypothesis.

4. What is the bias-variance tradeoff, and how does it affect model performance?

- **High Bias**: Underfitting (model too simple, poor training & testing accuracy).
- **High Variance**: Overfitting (model too complex, good training but poor testing accuracy).
- The goal is to balance bias and variance for optimal performance.

📌 Step 2: Data Preprocessing & Handling

5. How do you handle outliers in a dataset?

- **Statistical methods**: Z-score, IQR method
- **Transformation**: Log transformation, Winsorization

- **Model-based approaches**: Isolation Forest, DBSCAN

6. What are different preprocessing techniques used in data science?

- Handling missing values (imputation, deletion)
- Encoding categorical variables (One-Hot, Label Encoding)
- Scaling numerical data (Standardization, Normalization)

7. What are imputation techniques, and when should you use them?

- **Mean/Median/Mode imputation**: For numerical data with small missing values.
- **KNN imputation**: Uses nearest neighbors to estimate missing values.
- **Predictive modeling**: Regression or ML models predict missing values.

8. What are encoding techniques in Machine Learning, and which ones should be used for categorical data?

- **One-Hot Encoding**: For nominal categorical data.
- **Label Encoding**: For ordinal categorical data.
- **Target Encoding**: Encoding based on target variable distribution.

9. When should you use standard scaling vs. min-max scaling?

- **Standard Scaling (Z-score normalization)**: When data follows a normal distribution.
- **Min-Max Scaling**: When data has a fixed range and is not normally distributed.

📌 Step 3: Model Training & Evaluation

10. How can you prevent overfitting in Machine Learning (ML) and Deep Learning (DL)?

- **Regularization**: L1/L2 penalties
- **Dropout**: Randomly dropping neurons in deep learning
- **Cross-validation**: Ensuring model generalizability
- **Ensemble methods**: Using multiple models to reduce variance

11. What is cross-validation, and what are its different types?

- **K-Fold Cross Validation**: Divides data into K subsets for training/testing.
- **Stratified K-Fold**: Ensures class distribution remains balanced.
- **Leave-One-Out CV**: Uses one sample for testing, rest for training.

12. What are hyperparameter tuning techniques in Machine Learning?

- **Grid Search**: Tries all parameter combinations.
- **Random Search**: Randomly samples parameter space.
- **Bayesian Optimization**: Uses probabilistic models to find the best hyperparameters.

13. What are different performance metrics used for evaluating models?

- **Regression**: RMSE, MAE, R^2
- **Classification**: Accuracy, Precision, Recall, F1-score

14. What is the AUC/ROC curve, and how is it used to evaluate classification models?

- AUC (Area Under Curve) measures how well a classifier distinguishes between classes.
- ROC (Receiver Operating Characteristic) plots True Positive Rate vs. False Positive Rate.

15. Explain the concepts of Precision and Recall with examples.

- **Precision**: TP / (TP + FP) – How many predicted positives are actually positive?
- **Recall**: TP / (TP + FN) – How many actual positives are correctly identified?

📌 Step 4: Supervised Learning Algorithms

16. Explain the following Machine Learning algorithms:

- **Linear Regression**: Predicts a continuous target variable.
- **Logistic Regression**: Binary classification using a sigmoid function.
- **Naïve Bayes**: Probabilistic classifier based on Bayes' Theorem.
- **Support Vector Machines (SVM)**: Uses kernel tricks for classification.
- **K-Nearest Neighbors (KNN)**: Classifies based on closest neighbors.
- **Decision Trees**: Tree-based classification.
- **Random Forest (RF)**: Multiple decision trees for robust classification.
- **Boosting techniques**: XGBoost, AdaBoost, Gradient Boosting.

17. What is the difference between Decision Trees and Random Forests?

- Decision Trees are prone to overfitting.
- Random Forests use multiple trees to improve accuracy and reduce overfitting.

18. What is Random Forest, and how does it work?

- An ensemble of decision trees trained on different data subsets.
- Uses bagging (bootstrap aggregating) to reduce variance.

19. What are the key assumptions of Linear Regression and Naïve Bayes?

- Linear Regression: Linearity, Independence, Homoscedasticity, No multicollinearity.
- Naïve Bayes: Independence assumption among features.

20. Why is logistic regression called "regression" if it is used for classification? What functions are used in logistic regression?

- It models the probability of class membership using the **sigmoid function**.
- Despite classification, it estimates probabilities using regression.

📌 Step 5: Unsupervised Learning & Dimensionality Reduction

21. Explain the following unsupervised learning techniques:

- **K-Means Clustering**: Groups data into K clusters based on distance.
- **Principal Component Analysis (PCA)**: Reduces dimensionality by transforming features into orthogonal components.

📌 Step 6: Advanced Topics & Optimization Techniques

22. What is R^2 (R-squared), and how is adjusted R^2 different from it?

- R^2 measures model fit, but adjusted R^2 penalizes adding irrelevant features.

23. What is SMOTE, and how is it used in handling imbalanced datasets?

- Synthetic Minority Over-sampling Technique generates synthetic samples for minority classes.

24. What is the kernel trick in Support Vector Machines (SVM)?

- A mathematical technique to transform data into higher dimensions for better separation.

25. Explain GINI and Entropy in the context of Decision Trees.

- GINI: Measures impurity (low GINI means better splits).
- Entropy: Measures information gain (higher entropy reduction leads to better splits).

End of Document.