ML Assignment-1

1. Define Artificial Intelligence (AI)
AI refers to the simulation of human intelligence in machines.

2. Explain the differences between AI, ML, DL, and DS
AI is the broad field, ML is a subset focused on data-driven models, DL is a subset of ML with deep neural networks, and DS is the practice of extracting insights from data.

3. How does AI differ from traditional software development?
AI adapts and improves based on data, while traditional software follows predefined rules.

4. Provide examples of AI, ML, DL, and DS applications
AI: Autonomous vehicles, ML: Spam filtering, DL: Image recognition, DS: Customer insights.

5. Discuss the importance of AI, ML, DL, and DS in today's world
They drive automation, personalization, and data-driven decisions across industries.

6. What is Supervised Learning?
Supervised learning uses labeled data to train models.

7. Provide examples of Supervised Learning algorithms
Linear Regression, Decision Trees, SVM.

8. Explain the process of Supervised Learning
It involves training a model on labeled data to make predictions.

9. What are the characteristics of Unsupervised Learning?
Unsupervised learning finds patterns in data without labels.

10. Give examples of Unsupervised Learning algorithms
K-means, DBSCAN, PCA.

11. Describe Semi-Supervised Learning and its significance
Semi-supervised learning uses both labeled and unlabeled data to improve learning efficiency.

12. Explain Reinforcement Learning and its applications
Reinforcement learning teaches models by rewarding actions that lead to desired outcomes (e.g., robotics, game AI).

13. How does Reinforcement Learning differ from Supervised and Unsupervised Learning?
Reinforcement learning involves action and feedback, while the others focus on prediction or pattern discovery.

14. What is the purpose of the Train-Test-Validation split in machine learning?
To ensure models are trained, validated, and tested on separate data to avoid overfitting.

15. Explain the significance of the training set
The training set teaches the model how to make predictions.

16. How do you determine the size of the training, testing, and validation sets?
Use a typical ratio like 70% train, 15% test, and 15% validation.

17. What are the consequences of improper Train-Test-Validation splits?
It can lead to overfitting, underfitting, or misleading performance metrics.

18. Discuss the trade-offs in selecting appropriate split ratios
Higher training data improves model learning, while more test/validation data ensures generalizability.

19. Define model performance in machine learning
Model performance refers to how accurately a model predicts or classifies data.

20. How do you measure the performance of a machine learning model?
Using metrics like accuracy, precision, recall, F1 score, and AUC.

21. What is overfitting and why is it problematic?
Overfitting occurs when a model learns noise instead of patterns, leading to poor generalization.

22. Provide techniques to address overfitting
Cross-validation, regularization, and pruning.

23. Explain underfitting and its implications
Underfitting happens when the model is too simple, leading to poor performance.

24. How can you prevent underfitting in machine learning models?
Use more complex models or increase training time.

25. Discuss the balance between bias and variance in model performance
Bias refers to errors from overly simplistic models, and variance refers to errors from overly complex models. A good model balances both.

26. What are the common techniques to handle missing data?
Imputation, deletion, or using algorithms that handle missing values.

27. Explain the implications of ignoring missing data
It can lead to biased models or inaccurate results.

28. Discuss the pros and cons of imputation methods.
Imputation fills missing data but may introduce bias or inaccuracies.

29. How does missing data affect model performance?
It can reduce model accuracy or cause bias.

30. Define imbalanced data in the context of machine learning
Imbalanced data refers to unequal class distributions in classification tasks.

31. Discuss the challenges posed by imbalanced data
It leads to biased models that favor the majority class.

32. What techniques can be used to address imbalanced data?
Up-sampling, down-sampling, and synthetic data generation (e.g., SMOTE).

33. Explain the process of up-sampling and down-sampling
Up-sampling increases the minority class, and down-sampling reduces the majority class.

34. When would you use up-sampling versus down-sampling?
Up-sampling is used when the minority class is too small, down-sampling when the majority class is too large.

35. What is SMOTE and how does it work?
SMOTE generates synthetic samples for the minority class.

36. Explain the role of SMOTE in handling imbalanced data?
SMOTE balances the class distribution by creating synthetic examples.

37. Discuss the advantages and limitations of SMOTE
Advantages: Balances classes. Limitations: Can introduce noise or overfitting.

38. Provide examples of scenarios where SMOTE is beneficial?
Imbalanced binary classification problems like fraud detection.

39. Define data interpolation and its purpose?
Data interpolation estimates missing values based on available data.

40. What are the common methods of data interpolation?
Linear, polynomial, and spline interpolation.

41. Discuss the implications of using data interpolation in machine learning?
It may lead to unrealistic assumptions or overfitting.

42. What are outliers in a dataset?
Outliers are data points significantly different from others.

43. Explain the impact of outliers on machine learning models?
They can distort predictions and bias the model.

44. Discuss techniques for identifying outliers?
Z-scores, IQR, and visualization methods like box plots.

45. How can outliers be handled in a dataset?
By removing, capping, or transforming them.

46. Compare and contrast Filter, Wrapper, and Embedded methods for feature selection?
Filter methods use statistical tests, wrapper methods use model performance, and embedded methods select features during model training.

47. Provide examples of algorithms associated with each method?
Filter: Chi-square, Wrapper: Recursive Feature Elimination, Embedded: Lasso Regression.

48. Discuss the advantages and disadvantages of each feature selection method?
Filter: Fast but independent of model; Wrapper: More accurate but slower; Embedded: Efficient but model-dependent.

49. Explain the concept of feature scaling?
Feature scaling normalizes features to a similar range.

50. Describe the process of standardization?
Standardization transforms data to have zero mean and unit variance.

51. How does mean normalization differ from standardization?
Mean normalization centers data around 0, while standardization adjusts for variance.

52. Discuss the advantages and disadvantages of Min-Max scaling?
Advantages: Normalizes within a fixed range. Disadvantages: Sensitive to outliers.

53. What is the purpose of unit vector scaling?
It scales data to have a magnitude of 1.

54. Define Principal Component Analysis (PCA)?
PCA is a dimensionality reduction technique that transforms data into principal components.

55. Explain the steps involved in PCA?
Center the data, compute the covariance matrix, find eigenvalues/vectors, and project data onto principal components.

56. Discuss the significance of eigenvalues and eigenvectors in PCA?
Eigenvalues represent variance, and eigenvectors define directions of maximum variance.

57. How does PCA help in dimensionality reduction?
It reduces the number of features by selecting the principal components that capture the most variance.

58. Define data encoding and its importance in machine learning?
Data encoding transforms categorical data into numerical form.

59. Explain Nominal Encoding and provide an example.
Nominal encoding assigns a unique number to each category (e.g., colors: red=1, blue=2).

60. Discuss the process of One Hot Encoding?
One Hot Encoding creates binary columns for each category.

61. How do you handle multiple categories in One Hot Encoding?
By creating a separate binary column for each category.

62. Explain Mean Encoding and its advantages?
Mean encoding replaces categories with the mean of the target variable.

63. Provide examples of Ordinal Encoding and Label Encoding?
Ordinal Encoding: Low, Medium, High. Label Encoding: Red = 0, Blue = 1.

64. What is Target Guided Ordinal Encoding and how is it used?
It orders categories based on the target variable's mean.

65. Define covariance and its significance in statistics?
Covariance measures the directional relationship between two variables.

66. Explain the process of correlation check?
It evaluates the linear relationship between two variables using correlation coefficients.

67. What is the Pearson Correlation Coefficient?
It measures the strength and direction of a linear relationship between two variables.

68. How does Spearman's Rank Correlation differ from Pearson's Correlation?
Spearman measures monotonic relationships, while Pearson measures linear relationships.

69. Discuss the importance of Variance Inflation Factor (VIF) in feature selection?
VIF assesses multicollinearity by measuring how much a feature is correlated with others.

70. Define feature selection and its purpose?
Feature selection involves choosing the most relevant features to improve model performance.

71. Explain the process of Recursive Feature Elimination?
It iteratively removes features and builds a model to find the best feature subset.

72. How does Backward Elimination work?
It starts with all features and removes the least significant ones.

73. Discuss the advantages and limitations of Forward Elimination?
Advantages: Simple and interpretable. Limitations: May not find the optimal solution.

74. What is feature engineering and why is it important?
Feature engineering involves creating new features from raw data to improve model performance.

75. Discuss the steps involved in feature engineering?
Data cleaning, transformation, extraction, and selection.

76. Provide examples of feature engineering techniques?
Log transformations, creating interaction features, encoding categorical variables.

77. How does feature selection differ from feature engineering?
Feature selection involves choosing relevant features, while feature engineering creates new ones.

78. Explain the importance of feature selection in machine learning pipelines?
It reduces complexity, improves model accuracy, and prevents overfitting.

79. Discuss the impact of feature selection on model performance?
It can improve performance by removing irrelevant features and reducing noise.

80. How do you determine which features to include in a machine-learning model?
Based on data analysis, domain knowledge, and feature importance techniques.