Machine Learning and Data Analytics
(Unit 1 & Unit 2)

1. Differentiate between simple linear regression and multiple linear regression with suitable examples.
2. What is the role of the cost function in linear regression? Why is Mean Squared Error (MSE) commonly used?
3. Explain the problem of overfitting and underfitting in regression models. How can regularization techniques (L1 & L2) help to overcome overfitting? Illustrate with examples.
4. From the given dataset, compute the regression equation using the least squares method and predict the value of y when x=100:

Student	xi	yi
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

(*Hint: Use regression line y=b0+b1x.*)

5. A retail company wants to forecast its monthly sales revenue using regression analysis. The independent variables include advertising spend, price discount percentage, and number of promotional campaigns.
   - Formulate how you would set up a multiple linear regression model for this case.
   - Which factors could cause bias and variance in your model?
   - How would you evaluate the model's accuracy?

6. A researcher is studying the spread of a disease epidemic and finds that the relationship between time (in days) and number of cases is non-linear.
   - Why would polynomial regression be more suitable than linear regression in this scenario?
   - Discuss one advantage and one disadvantage of using polynomial regression for this case.

7. What is the role of entropy and information gain in the ID3 algorithm for decision tree construction?
8. Differentiate between pre-pruning and post-pruning in decision trees with an example.
9. Discuss the strengths and weaknesses of decision trees. Why do decision trees often face the problem of overfitting, and how can pruning or ensemble methods (like Random Forests) help?
10. Using the Gini Index method, calculate the best attribute to split from the following dataset (simplified example from the notes):

Gender	Play Cricket
Male	Yes
Male	No
Male	Yes

Gender	Play Cricket
Female	Yes
Female	No

(*Hint: Compute Gini for split on Gender and decide if it is a good split.*)

11. A bank wants to predict whether a loan applicant will default or repay based on attributes such as income, credit score, and employment status.
    - Explain how you would build a decision tree classifier for this case.
    - Which splitting method (ID3 or Gini Index) would you choose and why?
    - How would you handle overfitting in this model?
12. Suppose a dataset contains many outliers and noise in the training samples.
    - Explain how this might affect the structure and accuracy of a decision tree.
    - Which pruning technique (pre-pruning vs. post-pruning) would be more effective here, and why?

13. Define support vectors in the context of SVM. Why are they critical in determining the optimal hyperplane?
14. Differentiate between linear SVM and non-linear SVM with suitable examples.
15. Explain the role of the kernel trick in SVM. How do different kernels (Linear, Polynomial, and RBF) transform the input space, and what are the implications of choosing the wrong kernel?
16. Consider a 2D dataset with two features (x1, x2) where two classes are linearly separable. The equations of the separating hyperplanes are:
    - Positive class: $w \cdot x + b = +1$
    - Negative class: $w \cdot x + b = -1$

If $w=(2,1)$ and $b=-3$, calculate:

- The equation of the optimal hyperplane.
- The margin width between the two classes.

17. A company wants to build an email spam classifier using SVM.
    - Which kernel would you prefer for text classification and why?
    - How would you tune the C and gamma parameters to avoid overfitting or underfitting?
    - Discuss one advantage and one disadvantage of using SVM for this task.
18. In Support Vector Regression (SVR), what is the role of the epsilon-insensitive zone? Explain how SVR differs from traditional linear regression in handling outliers and noisy data.

19. Why is KNN considered a lazy learning algorithm? How does it differ from model-based learning techniques?
20. List any two advantages and two disadvantages of using KNN for classification tasks.

21. Explain the effect of different values of K in the KNN algorithm. What issues can occur when K is too small or too large, and how can we decide the optimal K for a dataset?
22. For the given dataset and query point X=(Maths=5,Science=7), with K=3, use Euclidean distance to classify the test sample:

Sr. No.	Maths	Science	Result
1	3	2	Fail
2	5	6	Pass
3	6	7	Pass

Sr. No.	Maths	Science	Result
4	4	4	Fail
5	7	7	Pass

(*Show all distance calculations and final classification.*)

23. A hospital wants to predict whether a patient has diabetes or not using KNN. The dataset includes features like glucose level, BMI, blood pressure, and age.
    - Explain how KNN can be applied to this problem.
    - Which distance metric (Euclidean, Manhattan, Minkowski, Hamming) would you choose and why?
    - What precautions should the hospital take regarding noise in data and choice of K?
24. Suppose you are working with a large dataset (millions of rows) for image recognition using KNN.
    - Why might KNN struggle in this case?
    - Suggest two possible strategies or alternatives to make KNN more efficient.

25. Why is Naïve Bayes called "naïve"? Explain with an example.
26. List any two pros and two cons of the Naïve Bayes classifier.
27. Explain how prior probability, likelihood, and posterior probability interact in the Naïve Bayes algorithm. Use the example of predicting whether a cloudy morning leads to rain to illustrate your answer.

28. Consider the following weather dataset:

- P(Sunny∣Yes)=3/9=0.33
- P(Yes)=9/14=0.64
- P(Sunny)=5/14=0.36

Use the Naïve Bayes formula to calculate P(Yes∣Sunny). Based on the result, decide whether players are likely to play on a sunny day.

29. A company wants to build an email spam filter using Naïve Bayes.
    - Which type of Naïve Bayes model (Gaussian, Multinomial, or Bernoulli) would be most appropriate, and why?
    - What problem might occur if a word appears in the test email but was not observed in the training data?
    - Suggest a solution to handle this issue.
30. Suppose you are using Naïve Bayes to predict whether a person has flu based on symptoms like fever, cough, and sore throat.
    - Why might the independence assumption of Naïve Bayes lead to incorrect predictions in this case?
    - Despite this limitation, why is Naïve Bayes still widely used in practice?
31. Define clustering. How does it differ from classification in machine learning?
32. List two advantages of using clustering in real-world applications such as marketing or recommendation systems.
33. Compare partitioning, hierarchical, and density-based clustering approaches. Explain with suitable examples when each is most effective.
34. Consider the following 2D points: (2,3),(3,3),(6,8),(7,9).
    Apply one iteration of K-means clustering with k=2.

- Initialize centroids as C1=(2,3) and C2=(6,8).
- Assign points to the nearest cluster using Euclidean distance.

- Compute the new centroids after this iteration.

35. A bank wants to segment its customers using clustering based on income and debt levels.
    - Explain how clustering can help in targeted marketing strategies.
    - Which clustering algorithm (K-means, Hierarchical, DBSCAN) would you recommend in this case and why?
    - How would you evaluate whether the clusters formed are meaningful?
36. The Dunn Index and Inertia are two evaluation metrics for clustering.
    - Explain how each metric works and what it tries to optimize.
    - If you have compact clusters that are still very close to each other, what will happen to these two metrics?

37. Define dimensionality reduction. Why is it important in machine learning?
38. Differentiate between feature selection and feature extraction with one example each.

39. Explain the step-by-step process of Principal Component Analysis (PCA) with mathematical justification. Why is standardization required before applying PCA?
40. Consider the following two variables X and Y:

- X=[2,4,6]
- Y=[1,3,5]

    (a) Compute the mean of X and Y.
    (b) Compute the covariance between X and Y.
    (c) Based on your result, state whether X and Y are positively or negatively correlated.

41. A company is working with a dataset containing 1,000 features for customer behavior prediction.
    - Why might applying dimensionality reduction improve both model performance and training speed?
    - Which method would you recommend between PCA and Random Forest Feature Importance, and why?
    - How would visualization in 2D/3D after PCA help business decision-making?
42. Compare PCA, ICA, and LDA in terms of:
    - Their objective (variance, independence, or class separation)
    - Whether they are supervised or unsupervised
    - One real-world application for each