## 1. What is the purpose of using Python for machine learning?

**Answer**: Python is widely used for machine learning due to its simplicity, readability, and vast ecosystem of libraries (like NumPy, Pandas, Scikit-learn, and Matplotlib) that make data manipulation, model building, and evaluation easier.

## 2. What are the key packages required for machine learning in Python?

**Answer**: The key packages for machine learning in Python are:

- **NumPy**: Used for numerical computing and array handling.
- **Pandas**: Used for data manipulation and analysis, working with DataFrames.
- **Scikit-learn**: Provides simple and efficient tools for machine learning tasks.
- **Matplotlib**: Used for plotting graphs and visualizations.

## 3. What is linear regression, and how is it used in machine learning?

**Answer**: Linear regression is a statistical model that assumes a linear relationship between input features and the target variable. It is used in machine learning for predicting continuous values, such as sales prediction.

## 4. What is the difference between Linear and Logistic Regression?

**Answer**: Linear regression is used for predicting continuous values (regression problems), while logistic regression is used for binary classification (predicting categories like 0 or 1).

## 5. Explain how you would implement a Linear Regression model for sales prediction.

**Answer**: You would:

- Collect historical data on factors influencing sales (e.g., marketing budget, store traffic).
- Preprocess the data (e.g., handle missing values, normalize features).
- Split the data into training and testing sets.
- Train a linear regression model using the training set.
- Evaluate the model using metrics like Mean Squared Error (MSE) on the test set.

## 6. What is the role of the train-test split in machine learning?

**Answer**: The train-test split is used to evaluate a model's performance. The data is divided into two sets: a training set (for model training) and a test set (for evaluating the model's performance on unseen data).

## 7. What is logistic regression, and how does it work?

**Answer**: Logistic regression is a machine learning algorithm used for binary classification. It predicts the probability of a binary outcome (0 or 1) by applying the logistic function (sigmoid) to a linear combination of input features.

## 8. What is the k-Nearest Neighbor (k-NN) algorithm?

**Answer**: k-NN is a supervised learning algorithm that classifies a data point based on the majority label of its k-nearest neighbors in the feature space. It is used for both classification and regression tasks.

## 9. How would you evaluate the performance of a classification model?

**Answer**: The performance of a classification model is evaluated using metrics like:

- **Accuracy**: Proportion of correctly predicted instances.
- **Precision**: Proportion of true positive predictions among all positive predictions.
- **Recall**: Proportion of true positives among all actual positives.
- **F1-Score**: Harmonic mean of precision and recall.

## 10. What is the purpose of using K-means clustering?

**Answer**: K-means clustering is an unsupervised learning algorithm used for partitioning data into k clusters. It is used for customer segmentation, anomaly detection, and other grouping tasks.

## 11. How do you choose the value of k in K-means clustering?

**Answer**: The value of k can be chosen using methods like the **Elbow Method**, where the sum of squared distances to the centroids (inertia) is plotted against k. The "elbow point" indicates the optimal k.

## 12. What is hierarchical clustering?

**Answer**: Hierarchical clustering is an unsupervised learning algorithm that builds a hierarchy of clusters. It can be agglomerative (bottom-up) or divisive (top-down), and produces a dendrogram to visualize the clustering process.

## 13. Explain the concept of decision trees.

**Answer**: Decision trees are supervised learning algorithms used for both classification and regression. They split the data based on feature values, creating a tree structure where each node represents a decision rule, and leaves represent the output.

## 14. What are Random Forests, and how do they improve upon decision trees?

**Answer**: Random Forests are an ensemble learning method that builds multiple decision trees on random subsets of the data. They reduce overfitting and improve accuracy by averaging the predictions from multiple trees.

## 15. What is the importance of cross-validation in model evaluation?

**Answer**: Cross-validation helps in assessing the generalization ability of a model by splitting the dataset into multiple folds, training and testing the model on different folds, and averaging the results to reduce bias and variance.

## 16. What is K-fold cross-validation?

**Answer**: K-fold cross-validation divides the dataset into k equal parts (folds). For each fold, the model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, and the average performance is reported.

## 17. What is the confusion matrix, and how is it used in classification?

**Answer**: A confusion matrix is a table that shows the actual versus predicted classifications. It is used to evaluate the performance of classification models by displaying true positives, false positives, true negatives, and false negatives.

## 18. What is the difference between precision and recall?

**Answer**:

- **Precision**: Proportion of true positive predictions out of all positive predictions (minimizing false positives).
- **Recall**: Proportion of true positives out of all actual positives (minimizing false negatives).

## 19. What is the F1-Score, and why is it important?

**Answer**: The F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is especially useful when you need to balance the importance of false positives and false negatives.

## 20. How do you evaluate the performance of a regression model?

**Answer**: The performance of a regression model is evaluated using metrics like:

- **Mean Absolute Error (MAE)**: Average of the absolute differences between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: Square root of the average of squared differences between predicted and actual values.
- **R-squared**: Proportion of the variance in the dependent variable that is predictable from the independent variables.

## 21. What is the purpose of clustering in machine learning?

**Answer**: Clustering is an unsupervised learning technique used to group similar data points together based on their features. It is useful for customer segmentation, anomaly detection, and exploratory data analysis.

## 22. What is the elbow method in K-means clustering?

**Answer**: The elbow method is used to determine the optimal number of clusters (k) by plotting the inertia (within-cluster sum of squares) against k and looking for the "elbow point" where the inertia starts decreasing at a slower rate.

## 23. How do decision trees handle missing data?

**Answer**: Decision trees can handle missing data by either:

- Ignoring missing values and proceeding with available data.
- Imputing missing values using techniques like mean or median imputation.
- Splitting the data based on surrogate splits for missing values.

## 24. What is feature scaling, and why is it important?

**Answer**: Feature scaling refers to the process of standardizing or normalizing features so they have the same scale. It is important for algorithms that are sensitive to the magnitude of features, such as KNN, SVM, and gradient descent-based methods.

## 25. Explain the difference between supervised and unsupervised learning.

**Answer**:

- **Supervised Learning**: The model is trained on labeled data, where the target variable is known (e.g., classification or regression).
- **Unsupervised Learning**: The model is trained on unlabeled data, and the goal is to find patterns or groupings (e.g., clustering or dimensionality reduction).

## 26. What are the challenges with imbalanced datasets, and how can they be addressed?

**Answer**: Imbalanced datasets can lead to biased models. Techniques to address this include:

- Resampling methods (oversampling the minority class or undersampling the majority class).
- Using algorithms like Random Forest or XGBoost that handle class imbalance well.
- Applying cost-sensitive learning methods or using different evaluation metrics like precision and recall.

## 27. What is the purpose of feature selection in machine learning?

**Answer**: Feature selection aims to reduce the number of input features by selecting the most relevant ones, improving model performance, reducing overfitting, and making the model easier to interpret.

## 28. How do you handle categorical data in machine learning?

**Answer**: Categorical data can be handled by techniques like:

- **One-Hot Encoding**: Converts categorical variables into a binary vector.

- **Label Encoding**: Assigns an integer value to each category.
- **Ordinal Encoding**: For ordinal data, assigns numerical values based on the order.

## 29. What is the difference between bagging and boosting?

**Answer**:

- **Bagging (Bootstrap Aggregating)**: Reduces variance by training multiple models independently on random subsets of data and combining their predictions (e.g., Random Forest).
- **Boosting**: Reduces bias by training models sequentially, where each model corrects the errors of the previous one (e.g., AdaBoost, Gradient Boosting).

## 30. What is the significance of using Random Forest over a single decision tree?

**Answer**: Random Forest improves upon a single decision tree by reducing overfitting through ensemble learning. It aggregates predictions from multiple decision trees, leading to better generalization and accuracy.

These questions cover a wide range of machine learning concepts, algorithms, and evaluation techniques, providing a solid foundation for your viva preparation.

**Key:** **tp** = True Positive, **tn** = True Negative, **fp** = False Positive, **fn** = False Negative

| Metric Name | Metric Forumla | Code | When to use |
|---|---|---|---|
| Accuracy | **Accuracy** $= \frac{tp + tn}{tp + tn + fp + fn}$ | `tf.keras.metrics.Accuracy()` or `sklearn.metrics.accuracy_score()` | Default metric for classification problems. Not the best for imbalanced classes. |
| Precision | **Precision** $= \frac{tp}{tp + fp}$ | `tf.keras.metrics.Precision()` or `sklearn.metrics.precision_score()` | Higher precision leads to less false positives. |
| Recall | **Recall** $= \frac{tp}{tp + fn}$ | `tf.keras.metrics.Recall()` or `sklearn.metrics.recall_score()` | Higher recall leads to less false negatives. |
| F1-score | **F1-score** $= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ | `sklearn.metrics.f1_score()` | Combination of precision and recall, usually a good overall metric for a classification model. |
| Confusion matrix | NA | `Custom function` or `sklearn.metrics.confusion_matrix()` | When comparing predictions to truth labels to see where model gets confused. Can be hard to use with large numbers of classes. |

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

$\hat{y}$ $-$ *predicted value of* $y$

$\bar{y}$ $-$ *mean value of* $y$

## Regression model evaluation metrics

The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.
- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.