1. **What is Machine Learning?**
   * Machine Learning (ML) is a subset of artificial intelligence (AI) that enables systems to learn and improve from experience without being explicitly programmed. It focuses on developing algorithms that can process data, identify patterns, and make decisions with minimal human intervention. Example: A spam filter in email services.
2. **Mention the difference between Data Mining and Machine Learning.**
   * Data Mining is the process of discovering patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the web, and other data repositories.
   * Machine Learning focuses on using algorithms to learn from data and make predictions or decisions.
   * Difference: Data Mining is often about extracting patterns and knowledge from large datasets, whereas Machine Learning is about using those patterns to predict future data.
3. **What is ‘Overfitting’ in Machine Learning?**
   * Overfitting occurs when a machine learning model learns the training data too well, including its noise and outliers. This results in poor generalization to new, unseen data.
   * Example: A decision tree that is too complex and perfectly classifies the training data but fails on the test data.
4. **Why overfitting happens?**
   * Overfitting happens due to:
     + Too complex models with many parameters.
     + Insufficient training data.
     + Excessive training iterations.
     + Lack of regularization techniques.
5. **How can you avoid overfitting?**
   * Techniques to avoid overfitting include:
     + Simplifying the model.
     + Using more training data.
     + Applying regularization methods (e.g., L1, L2 regularization).
     + Using cross-validation techniques.
     + Pruning decision trees.
6. **What is inductive machine learning?**
   * Inductive machine learning involves learning a model from a set of observed instances (training data) and then making generalizations about new instances. It contrasts with deductive learning, where rules are manually specified.
7. **What are the five popular algorithms of Machine Learning?**
   * Popular algorithms include:
     + Linear Regression
     + Logistic Regression
     + Decision Trees
     + Support Vector Machines (SVM)
     + k-Nearest Neighbors (k-NN)
8. **What are the different Algorithm techniques in Machine Learning?**
   * Techniques include:
     + Supervised Learning (e.g., regression, classification)
     + Unsupervised Learning (e.g., clustering, dimensionality reduction)
     + Semi-supervised Learning
     + Reinforcement Learning
9. **What are the three stages to build the hypotheses or model in machine learning?**
   * The stages are:
     + Model Selection: Choosing the appropriate model or hypothesis for the problem.
     + Training: Training the model using a training dataset.
     + Evaluation: Evaluating the model's performance using test data.
10. **What is the standard approach to supervised learning?**
    * The standard approach involves:
      + Splitting the dataset into training and test sets.
      + Training the model on the training set.
      + Evaluating the model on the test set.
11. **What is ‘Training set’ and ‘Test set’?**
    * A training set is a subset of the dataset used to train the model.
    * A test set is a subset of the dataset used to evaluate the model's performance and generalization to new data.
12. **List down various approaches for machine learning?**
    * Approaches include:
      + Supervised Learning
      + Unsupervised Learning
      + Semi-supervised Learning
      + Reinforcement Learning
      + Transfer Learning
      + Active Learning
13. **What is not Machine Learning?**
    * Not every automated system is machine learning. Rule-based systems, traditional programming where rules are explicitly coded, and simple data analysis are not considered machine learning.
14. **Explain what is the function of ‘Unsupervised Learning’?**
    * Unsupervised Learning finds hidden patterns or intrinsic structures in input data without labeled responses. It includes clustering and association algorithms.
15. **Explain what is the function of ‘Supervised Learning’?**
    * Supervised Learning involves training a model on labeled data, where the algorithm learns from input-output pairs to make predictions on new data. It includes regression and classification algorithms.
16. **What is algorithm independent machine learning?**
    * Algorithm independent machine learning refers to techniques or methods that can be applied regardless of the specific learning algorithm used. These often involve data preprocessing, feature selection, and model evaluation strategies.
17. **What is the difference between artificial learning and machine learning?**
    * Artificial Intelligence (AI) is a broader concept that includes the simulation of human intelligence in machines.
    * Machine Learning (ML) is a subset of AI focused on the ability of machines to learn from data and improve performance over time.
18. **What is a classifier in machine learning?**
    * A classifier is an algorithm that maps input data to a specific category. It's used in classification tasks to assign labels to new instances based on learned patterns.
19. **What are the advantages of Naive Bayes?**
    * Advantages include:
      + Simplicity and efficiency.
      + Handles large datasets well.
      + Performs well with high-dimensional data.
      + Suitable for real-time predictions.
20. **What is Inductive Logic Programming in Machine Learning?**
    * Inductive Logic Programming (ILP) is a subfield of machine learning that uses logic programming (e.g., Prolog) to represent input data and background knowledge, aiming to induce hypotheses in the form of logical rules.
21. **What is Model Selection in Machine Learning?**
    * Model selection involves choosing the best model from a set of candidate models based on their performance on a validation set. It aims to find a model that generalizes well to new data.
22. **What are the two methods used for the calibration in Supervised Learning?**
    * Methods include:
      + Platt Scaling
      + Isotonic Regression
23. **Which method is frequently used to prevent overfitting?**
    * Cross-validation is frequently used to prevent overfitting by evaluating the model's performance on multiple subsets of the data and ensuring it generalizes well.
24. **Why instance-based learning algorithm sometimes referred as Lazy learning algorithm?**
    * Instance-based learning algorithms are referred to as lazy learning algorithms because they delay the generalization process until a query is made. They store the training data and only compute the decision boundaries at prediction time. Example: k-NN algorithm.
25. **What are the two classification methods that SVM (Support Vector Machine) can handle?**
    * SVM can handle:
      + Binary Classification: Classifying data into two distinct classes.
      + Multiclass Classification: Classifying data into more than two classes (using techniques like one-vs-one or one-vs-rest).
26. **What is ensemble learning?**
    * Ensemble learning is a technique where multiple models (often referred to as "weak learners") are trained and combined to solve the same problem, usually resulting in improved performance and robustness.
27. **When to use ensemble learning?**
    * Ensemble learning is used when:
      + Single models have high variance or bias.
      + Improved accuracy and robustness are needed.
      + Dealing with complex datasets where no single model performs optimally.
28. **What are the two paradigms of ensemble methods?**
    * The two main paradigms are:
      + Bagging (Bootstrap Aggregating)
      + Boosting
29. **What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?**
    * The general principle of ensemble methods is to combine multiple models to improve overall performance.
      + **Bagging**: Involves training multiple models independently on different subsets of the data (bootstrapped samples) and averaging their predictions. Example: Random Forest.
      + **Boosting**: Involves training models sequentially, where each model tries to correct the errors of the previous ones. Example: AdaBoost, Gradient Boosting.
30. **What is bias-variance decomposition of classification error in ensemble method?**
    * Bias-variance decomposition is a way of understanding and analyzing errors in machine learning models:
      + **Bias**: Error due to overly simplistic models that do not capture the underlying patterns of the data.
      + **Variance**: Error due to overly complex models that capture noise in the training data.
    * Ensemble methods like bagging and boosting help reduce variance and bias, respectively.
31. **What is an Incremental Learning algorithm in ensemble?**
    * Incremental Learning algorithms update the model as new data arrives without retraining from scratch. They are useful for streaming data and large datasets that cannot be processed in one go. Example: Online Gradient Descent.
32. **What is PCA, KPCA and ICA used for?**
    * **Principal Component Analysis (PCA)**: Used for dimensionality reduction by transforming data into a new coordinate system with orthogonal axes (principal components).
    * **Kernel PCA (KPCA)**: Extends PCA to non-linear data using kernel methods.
    * **Independent Component Analysis (ICA)**: Used to separate a multivariate signal into additive, independent components, often used in signal processing.
33. **What is dimension reduction in Machine Learning?**
    * Dimension reduction involves reducing the number of features in a dataset while preserving as much information as possible. Techniques include PCA, t-SNE, and LDA.
34. **What are support vector machines?**
    * Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks. They find

yes

ChatGPT

Sure, I'll continue answering the questions and then compile everything into a document for you.

1. **Differentiate between inductive learning and deductive learning.**
   * **Inductive Learning**: Involves learning from examples and generalizing from observed data to form rules or models. Example: A machine learning model trained on a dataset to predict future data.
   * **Deductive Learning**: Involves applying existing knowledge or rules to deduce new facts or outcomes. Example: Using known mathematical theorems to solve a problem.
2. **What is the difference between Data Mining and Machine Learning?**
   * **Data Mining**: Focuses on discovering patterns, correlations, and anomalies within large datasets using statistical and computational methods. It often involves a human interpreting the results.
   * **Machine Learning**: Focuses on developing algorithms that can learn from data and make predictions or decisions. It involves the model continuously improving itself based on new data.
3. **Differentiate supervised and unsupervised machine learning.**
   * **Supervised Learning**: Models are trained on labeled data, where the input comes with corresponding output labels. Example: Classification tasks.
   * **Unsupervised Learning**: Models are trained on unlabeled data and need to identify patterns or structures. Example: Clustering tasks.
4. **How does Machine Learning differ from Deep Learning?**
   * **Machine Learning**: Encompasses a broad range of algorithms and techniques to make predictions or decisions based on data. Example: Linear Regression, Decision Trees.
   * **Deep Learning**: A subset of machine learning focused on neural networks with many layers (deep networks) capable of learning from vast amounts of data. Example: Convolutional Neural Networks (CNNs) for image recognition.
5. **How is KNN different from k-means?**
   * **k-Nearest Neighbors (k-NN)**: A supervised learning algorithm used for classification and regression by identifying the k-nearest data points to the input.
   * **k-means**: An unsupervised learning algorithm used for clustering data into k groups based on feature similarity.
6. **What are the different types of Algorithm methods in Machine Learning?**
   * **Supervised Learning**: Algorithms include Linear Regression, Logistic Regression, Decision Trees, SVM.
   * **Unsupervised Learning**: Algorithms include k-means, Hierarchical Clustering, PCA.
   * **Semi-supervised Learning**: Combines labeled and unlabeled data.
   * **Reinforcement Learning**: Algorithms include Q-Learning, Deep Q-Networks.
7. **What do you understand by Reinforcement Learning technique?**
   * Reinforcement Learning (RL) involves training agents to make sequences of decisions by rewarding desirable behaviors and punishing undesirable ones. Example: Training a robot to navigate a maze.
8. **What is the trade-off between bias and variance?**
   * **Bias**: Error due to overly simplistic models that miss the underlying patterns. High bias can cause underfitting.
   * **Variance**: Error due to overly complex models that capture noise in the training data. High variance can cause overfitting.
   * The trade-off involves finding a balance where the model is complex enough to capture the data patterns but simple enough to generalize well.
9. **How do classification and regression differ?**
   * **Classification**: Predicts discrete labels or categories. Example: Email spam detection.
   * **Regression**: Predicts continuous values. Example: House price prediction.
10. **What are the three stages of building the hypotheses or model in machine learning?**
    * **Model Selection**: Choosing the right algorithm and model for the problem.
    * **Training**: Fitting the model to the training data.
    * **Evaluation**: Assessing the model's performance using a validation or test set.
11. **Describe 'Training set' and 'Test set'.**
    * **Training Set**: The dataset used to train the model.
    * **Test Set**: The dataset used to evaluate the model's performance and generalization to new data.
12. **What are the common ways to handle missing data in a dataset?**
    * **Imputation**: Filling in missing values with mean, median, mode, or using advanced techniques like k-NN imputation.
    * **Deletion**: Removing rows or columns with missing values.
    * **Model-based methods**: Using algorithms that handle missing values naturally.
13. **What are the necessary steps involved in Machine Learning Project?**
    * Defining the problem.
    * Collecting and preparing the data.
    * Selecting and training a model.
    * Evaluating and tuning the model.
    * Deploying the model.
    * Monitoring and maintaining the model.
14. **Describe Precision and Recall.**
    * **Precision**: The ratio of true positive predictions to the total predicted positives. Measures accuracy of positive predictions.
    * **Recall**: The ratio of true positive predictions to the total actual positives. Measures the ability to capture all positive instances.
15. **What do you understand by Decision Tree in Machine Learning?**
    * A Decision Tree is a model used for classification and regression tasks. It splits the data into subsets based on feature values, creating a tree structure where each node represents a feature and each leaf represents an outcome.
16. **What do you understand by algorithm independent machine learning?**
    * Algorithm independent machine learning involves techniques that can be applied across different algorithms, focusing on data preprocessing, feature selection, and model evaluation, rather than the specifics of any single algorithm.
17. **Describe the classifier in machine learning.**
    * A classifier is an algorithm that assigns a label to input data based on learned patterns from labeled training data. Examples include Logistic Regression, Decision Trees, and SVM.
18. **What is SVM in machine learning? What are the classification methods that SVM can handle?**
    * Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It finds the hyperplane that best separates different classes in the feature space. SVM handles:
      + Binary Classification
      + Multiclass Classification (using one-vs-one or one-vs-rest methods)
19. **What do you understand by the Confusion Matrix?**
    * A Confusion Matrix is a table used to evaluate the performance of a classification model by comparing predicted and actual values. It includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).
20. **Explain True Positive, True Negative, False Positive, and False Negative in Confusion Matrix with an example.**
    * **True Positive (TP)**: Correctly predicted positive cases. Example: Predicting spam emails correctly.
    * **True Negative (TN)**: Correctly predicted negative cases. Example: Predicting non-spam emails correctly.
    * **False Positive (FP)**: Incorrectly predicted positive cases. Example: Predicting non-spam emails as spam.
    * **False Negative (FN)**: Incorrectly predicted negative cases. Example: Predicting spam emails as non-spam.
21. **What according to you, is more important between model accuracy and model performance?**
    * Model performance is more important as it encompasses various metrics (accuracy, precision, recall, F1-score) that provide a comprehensive evaluation of the model's effectiveness, especially in imbalanced datasets where accuracy alone can be misleading.
22. **What is Bagging and Boosting?**
    * **Bagging**: An ensemble method that builds multiple models independently on bootstrapped samples of the data and combines their predictions. Example: Random Forest.
    * **Boosting**: An ensemble method that builds models sequentially, where each model focuses on correcting errors made by the previous ones. Example: AdaBoost, Gradient Boosting.
23. **What are the similarities and differences between bagging and boosting in Machine Learning?**
    * **Similarities**: Both are ensemble methods aiming to improve model performance by combining multiple models.
    * **Differences**:
      + **Bagging**: Models are trained independently on random subsets of data. Reduces variance.
      + **Boosting**: Models are trained sequentially, focusing on errors of previous models. Reduces bias.
24. **What do you understand by Cluster Sampling?**
    * Cluster Sampling is a sampling technique where the population is divided into clusters (groups), and a random sample of these clusters is selected. All members of the chosen clusters are included in the sample. It's useful when dealing with large, geographically spread populations.
25. **What do you understand by the F1 score?**
    * The F1 score is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when dealing with imbalanced datasets.
26. **How is a decision tree pruned?**
    * Pruning a decision tree involves removing parts of the tree that do not provide additional power to classify instances. Techniques include pre-pruning (setting a limit on tree depth) and post-pruning (removing branches after the tree is created).
27. **What are the Recommended Systems?**
    * Recommender Systems are algorithms designed to suggest products, services, or information to users based on their preferences and behavior. Types include collaborative filtering, content-based filtering, and hybrid methods. Example: Amazon's product recommendations.
28. **When does regularization become necessary in Machine Learning?**
    * Regularization is necessary when models are overfitting the training data, indicated by high variance and poor generalization to new data. Regularization techniques add a penalty to complex models to prevent overfitting.
29. **What is Regularization? What kind of problems does regularization solve?**
    * Regularization adds a penalty to the loss function to discourage complex models and prevent overfitting. Common

yes

ChatGPT

Let's continue with the explanations:

1. **What is Regularization? What kind of problems does regularization solve?**
   * Regularization adds a penalty to the loss function to discourage complex models, which helps prevent overfitting. It essentially reduces the model complexity by shrinking the coefficients. Common regularization techniques include L1 (Lasso) and L2 (Ridge) regularization.
2. **Why do we need to convert categorical variables into factors? Which functions are used to perform the conversion?**
   * Converting categorical variables into factors is necessary because most machine learning algorithms require numerical input. Converting them into factors allows algorithms to interpret categorical data correctly. Common functions for conversion in Python are **pd.get\_dummies()** in pandas and **LabelEncoder** in scikit-learn.
3. **Do you think that treating a categorical variable as a continuous variable would result in a better predictive model?**
   * Treating categorical variables as continuous variables can lead to incorrect model predictions because the ordinal relationships may not reflect the true nature of the data. For example, the categories "Red", "Blue", and "Green" do not have a natural order or numerical relationship.
4. **How is machine learning used in day-to-day life?**
   * Machine learning is used in various day-to-day applications, including:
     + Email spam filtering
     + Personalized recommendations on streaming services
     + Fraud detection in banking
     + Customer support chatbots
     + Autonomous vehicles
     + Health diagnostics
5. **How Do You Handle Missing or Corrupted Data in a Dataset?**
   * Handling missing or corrupted data can involve:
     + **Imputation**: Replacing missing values with mean, median, mode, or using advanced techniques like k-NN imputation.
     + **Deletion**: Removing rows or columns with missing values if the amount is small.
     + **Using models that handle missing data**: Some algorithms can handle missing values natively.
6. **How Can You Choose a Classifier Based on a Training Set Data Size?**
   * Choosing a classifier can depend on the size of the training data:
     + **Small Datasets**: Naive Bayes, k-NN, and Decision Trees can perform well.
     + **Large Datasets**: SVM, Neural Networks, and ensemble methods like Random Forest and Gradient Boosting are suitable.
7. **What Are the Applications of Supervised Machine Learning in Modern Businesses?**
   * Applications include:
     + Predictive maintenance in manufacturing
     + Customer segmentation and targeting in marketing
     + Risk assessment in finance
     + Medical diagnosis in healthcare
     + Fraud detection in banking
     + Sentiment analysis in social media monitoring
8. **What is Semi-supervised Machine Learning?**
   * Semi-supervised learning involves using both labeled and unlabeled data for training. It is useful when obtaining labeled data is expensive or time-consuming, and there is an abundance of unlabeled data.
9. **Compare K-means and KNN Algorithms.**
   * **k-means**: An unsupervised clustering algorithm that partitions the data into k clusters based on feature similarity.
   * **k-NN**: A supervised classification algorithm that predicts the class of a new data point based on the majority class among its k-nearest neighbors.
10. **What Is ‘naive’ in the Naive Bayes Classifier?**
    * The term 'naive' refers to the assumption that features are independent of each other given the class label. This simplification makes the computation easier, though it is often not true in real-world data.
11. **How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?**
    * The choice of algorithm can depend on:
      + **Data size and dimensionality**
      + **Nature of the problem (linearity, complexity)**
      + **Performance on validation data**
      + **Computational resources**
      + **Algorithm interpretability**
12. **How is Amazon Able to Recommend Other Things to Buy? How Does the Recommendation Engine Work?**
    * Amazon uses a recommendation engine that combines collaborative filtering, content-based filtering, and hybrid methods to suggest products based on user behavior, purchase history, and similar users' activities.
13. **When Will You Use Classification over Regression?**
    * Use classification when the output variable is categorical (e.g., spam or not spam).
    * Use regression when the output variable is continuous (e.g., predicting house prices).
14. **How Do You Design an Email Spam Filter?**
    * Designing an email spam filter involves:
      + Collecting and labeling a dataset of spam and non-spam emails.
      + Extracting features (e.g., word frequency, presence of certain keywords).
      + Training a classification model (e.g., Naive Bayes, SVM).
      + Evaluating the model using metrics like precision, recall, and F1-score.
      + Deploying the model and continuously updating it with new data.
15. **What is a Random Forest?**
    * Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.
16. **What is Pruning in Decision Trees, and How Is It Done?**
    * Pruning in decision trees involves removing branches that have little importance to reduce complexity and improve generalization. Pruning can be done:
      + **Pre-pruning**: Setting a maximum depth or minimum number of samples required at a node.
      + **Post-pruning**: Removing branches after the tree is fully grown based on performance on validation data.
17. **Briefly Explain Logistic Regression.**
    * Logistic Regression is a supervised learning algorithm used for binary classification. It models the probability of the default class using a logistic function, which outputs a value between 0 and 1.
18. **Explain the K Nearest Neighbor Algorithm.**
    * k-NN is a non-parametric, instance-based learning algorithm used for classification and regression. It classifies a new data point based on the majority class among its k-nearest neighbors in the feature space.
19. **What is Kernel SVM?**
    * Kernel SVM is an extension of Support Vector Machines that uses kernel functions to transform the input data into a higher-dimensional space where it is easier to find a separating hyperplane. Common kernels include the polynomial, radial basis function (RBF), and sigmoid kernels.
20. **What Are Some Methods of Reducing Dimensionality?**
    * Methods include:
      + **Principal Component Analysis (PCA)**
      + **t-Distributed Stochastic Neighbor Embedding (t-SNE)**
      + **Linear Discriminant Analysis (LDA)**
      + **Autoencoders**
21. **What is Principal Component Analysis?**
    * Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the data into a new coordinate system where the axes (principal components) are orthogonal and ordered by the amount of variance they explain.
22. **What do you understand by Type I vs Type II error?**
    * **Type I Error (False Positive)**: Incorrectly rejecting the null hypothesis when it is true.
    * **Type II Error (False Negative)**: Failing to reject the null hypothesis when it is false.
23. **Explain Correlation and Covariance?**
    * **Correlation**: Measures the strength and direction of the linear relationship between two variables, ranging from -1 to 1.
    * **Covariance**: Measures the extent to which two variables change together, but it is not normalized, so its magnitude depends on the data scale.
24. **What are Support Vectors in SVM?**
    * Support vectors are the data points closest to the decision boundary (hyperplane) in an SVM model. They are crucial in defining the position and orientation of the hyperplane.
25. **What is Cross-Validation?**
    * Cross-validation is a technique used to evaluate the performance of a model by splitting the data into multiple subsets and training/testing the model on different combinations of these subsets. Common methods include k-fold cross-validation and leave-one-out cross-validation.
26. **What are the different methods to split a tree in a decision tree algorithm?**
    * Common methods include:
      + **Gini Impurity**: Measures the impurity of a node.
      + **Entropy**: Measures the information gain.
      + **Variance Reduction**: Used in regression trees to measure the variance reduction after a split.
27. **How does the Support Vector Machine algorithm handle self-learning?**
    * SVMs do not inherently have self-learning capabilities, but they can be combined with other algorithms and techniques for semi-supervised learning or reinforcement learning scenarios.
28. **What are the assumptions you need to take before starting with linear regression?**
    * Assumptions include:
      + **Linearity**: The relationship between the input and output is linear.
      + **Independence**: Observations are independent of each other.
      + **Homoscedasticity**: Constant variance of errors.
      + **Normality**: Errors are normally distributed.
29. **What is the difference between Lasso and Ridge regression?**
    * **Lasso Regression (L1 Regularization)**: Adds a penalty equal to the absolute value of the coefficients, which can shrink some coefficients to zero, effectively performing feature selection.
    * **Ridge Regression (L2 Regularization)**: Adds a penalty equal to the square of the coefficients, which shrinks coefficients but does not eliminate any.
30. **What is Entropy in Machine Learning?**
    * Entropy is a measure of the randomness or impurity in a dataset. In decision trees, it is used to determine the best split by measuring the information gain.
31. \*\*What is Epoch in Machine Learning?

yes

ChatGPT

Let's continue:

1. **What is Epoch in Machine Learning?**
   * An epoch in machine learning refers to one complete pass through the entire training dataset. It is a measure used in iterative training processes, such as in neural networks, where the algorithm updates its parameters multiple times across several epochs to minimize the loss function.
2. **Differentiate between Classification and Regression in Machine Learning.**
   * **Classification**: Predicts categorical labels or classes. Examples include spam detection (spam or not spam) and image recognition (cat, dog, etc.).
   * **Regression**: Predicts continuous values. Examples include predicting house prices or stock market trends.
3. **How is the suitability of a Machine Learning Algorithm determined for a particular problem?**
   * The suitability of a machine learning algorithm is determined by:
     + **Nature of the Problem**: Whether it is classification, regression, clustering, etc.
     + **Data Characteristics**: Size, dimensionality, and quality of the dataset.
     + **Performance Metrics**: Accuracy, precision, recall, F1-score, etc.
     + **Computational Resources**: Available processing power and memory.
     + **Interpretability**: The need for understanding the model's decisions.
4. **What is ROC Curve and what does it represent?**
   * An ROC (Receiver Operating Characteristic) Curve is a graphical representation of a classification model's performance. It plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings. The area under the ROC Curve (AUC) indicates the model's ability to distinguish between classes.
5. **Both being Tree-based Algorithms, how is Random Forest different from Gradient Boosting Machine (GBM)?**
   * **Random Forest**: An ensemble method that builds multiple decision trees independently using bootstrapped samples and averages their predictions. It reduces variance and improves robustness.
   * **Gradient Boosting Machine (GBM)**: An ensemble method that builds trees sequentially, where each tree corrects errors made by the previous ones. It reduces bias and variance but is more prone to overfitting and typically requires careful tuning.
6. **What do you understand about the P-value?**
   * The P-value in statistical hypothesis testing measures the probability of obtaining test results at least as extreme as the observed results, assuming the null hypothesis is true. A low P-value (< 0.05) indicates strong evidence against the null hypothesis, leading to its rejection.
7. **Suppose you found that your model is suffering from high variance. Which algorithm do you think could handle this situation and why?**
   * Algorithms like Random Forest, Gradient Boosting, or Regularized models (Lasso, Ridge) can handle high variance. These methods either combine multiple models to reduce variance or use regularization to prevent overfitting by penalizing complex models.
8. **What is Rescaling of Data and how is it done?** - Rescaling (or normalization) involves adjusting the range of data features to a standard scale, typically 0 to 1 or -1 to 1. Methods include Min-Max scaling, Standardization (z-score normalization), and Robust Scaling. Rescaling ensures that features contribute equally to the model's performance and improves convergence speed in algorithms like gradient descent.