1. What is the difference between cost and loss function in machine learning?
    a. the loss function assesses the error for a single data point, while the cost function measures the overall performance of the model across the entire training dataset. During model training, the objective is to minimize the cost function by adjusting the model's parameters, which in turn minimizes the individual loss values for each data point.
2. What is BERT?
    a. Bidirectional Encoder Representations from Transformers (BERT) is a highly popular and influential natural language processing model because it's pre-trained on vast amounts of text data, understands context bidirectionally, achieves state-of-the-art NLP performance, and has open-source models that can be fine-tuned for various language tasks. It has set a new standard in NLP and inspired further innovations in the field.
3. What are some of the hyperparameters of BERT?
    a. some key hyperparameters for BERT (Bidirectional Encoder Representations from Transformers) include the model architecture, learning rate, batch size, number of training epochs, dropout rate, sequence length, and others. These hyperparameters can be adjusted to fine-tune BERT for specific natural language processing tasks.
4. What is the loss function for BERT?
    a. the loss function used during BERT pre-training is primarily the MLM loss, which involves predicting masked words in text. For fine-tuning on specific NLP tasks, task-specific loss functions are employed, which can vary depending on the nature of the task, such as text classification, sequence labeling, or question-answering. The choice of loss function is critical for optimizing the model's performance on these downstream tasks.
5. What is a transformer in machine learning?
    a. Transformers are a machine learning model architecture known for their effectiveness in processing sequential data. They employ self-attention mechanisms, which allow them to capture relationships and dependencies between elements in a sequence. This is particularly valuable in natural language processing, where understanding the context and relationships between words is crucial. Unlike older sequential models like RNNs, transformers can process data in parallel, making them more efficient for both training and inference. They consist of multiple layers, or transformer blocks, each containing self-attention and feedforward sub-layers. Stacking these layers enables the model to learn increasingly complex patterns and dependencies in the data. To account for the order of elements in a sequence, positional encoding is added to the input data. Transformers also often use multi-head attention, which allows them to focus on different parts of the input data simultaneously. One of the significant advantages of transformers is their scalability. They can handle sequences of varying lengths, making them versatile for a wide range of tasks.
6. What is the difference between a transformer and BERT?

   a. A "transformer" is a general-purpose deep learning model architecture designed for sequential data processing. "BERT" is a specific model based on the transformer architecture, pre-trained for natural language understanding tasks, and fine-tuned for NLP applications.
7. What are hyperparameters in machine learning?
   a. hyperparameters in machine learning are settings that determine a model's behavior. They are set before training and include values like learning rates, batch sizes, and the number of hidden layers. Examples:
      i. Learning Rate: Determines how quickly the model learns.
      ii. Number of Epochs: Controls the training duration.
      iii. Batch Size: Sets the number of training examples in each step.
      iv. Number of Layers: Defines the depth of neural networks.
      v. Regularization Strength: Balances model complexity and overfitting (e.g., L1 or L2 regularization).
      vi. Dropout Rate: Manages overfitting in neural networks.
8. What is autoencoder?
   a. an autoencoder is a type of neural network used for unsupervised learning. It compresses input data into a lower-dimensional representation and then tries to reconstruct the original data from that representation. It's used for tasks like data compression, feature learning, anomaly detection, and more
9. What are the log parsing techniques? Explain with example.
   a. Regular expression, template matching and count, Drain/Spell
10. What is the difference between structured and unstructured files? How to handle unstructured data?
   a. structured data follows a clear format and schema, making it easy to work with, while unstructured data lacks a specific structure and often requires specialized techniques to extract meaning and insights from it.
11. What is an anomaly in machine learning?
   a. an anomaly in machine learning refers to an unusual or unexpected data point that deviates significantly from the majority of the data. Anomaly detection is the process of identifying such anomalies, which can have applications in fraud detection, quality control, and various fields. It can be addressed using statistical, machine learning, or deep learning methods, depending on the data and application.
12. What models are appropriate for anomaly detection?
   a. Statistical Methods: Z-Score, Modified Z-Score, Grubbs' Test, and Hampel Identifier.
   b. Machine Learning Models: Isolation Forest, One-Class SVM, Local Outlier Factor, k-Nearest Neighbors.
   c. Deep Learning Models: Autoencoders, Variational Autoencoders, Recurrent Neural Networks (RNN).
   d. Clustering Techniques: DBSCAN, K-Means Clustering.
   e. Ensemble Methods: Combining multiple models for improved accuracy.

f. Time-Series Specific Techniques: Seasonality decomposition, moving averages, and exponential smoothing may be used for time-series data.

13. What is the F1 score?
    a. the F1 score is defined as the harmonic mean of precision (P) and recall (R). It is calculated as:
    b. $$F1 = \frac{2 \cdot P \cdot R}{P + R}$$
14. What is precision and recall?
    a. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.
    b. Precision is how good the model is at predicting a specific category. Recall tells you how many times the model was able to detect a specific category.
    c. $$P = \frac{TP}{TP + FP}$$
    d. $$R = \frac{TP}{TP + FN}$$
15. How to handle imbalanced data?
    a. Resampling: Oversample the minority class or undersample the majority class.
    b. Synthetic Data: Generate synthetic data for the minority class (e.g., SMOTE).
    c. Cost-Sensitive Learning: Assign different misclassification costs to classes.
    d. Ensemble Methods: Use ensemble techniques that handle class imbalance.
    e. Anomaly Detection: Treat the minority class as anomalies.
    f. Threshold Adjustment: Adjust classification thresholds.
    g. Data Augmentation: Enhance minority class data with techniques like rotation or noise.
    h. Collect More Data: If possible, gather more data for the minority class.
    i. Feature Selection: Focus on informative attributes and reduce noise.
    j. Algorithm Selection: Choose algorithms less sensitive to imbalance.
    k. Evaluation Metrics: Use metrics like F1-score or AUC-ROC.
    l. Cross-Validation: Apply stratified k-fold cross-validation.
    m. Expert Knowledge: Incorporate domain expertise.
    n. Class Weighting: Assign higher weights to the minority class during training.
16. What are some problems with pandas?
    a. Memory usage: Pandas can use a lot of memory when working with large datasets, which can lead to performance issues.

b. Slow performance: Pandas operations can be slow, especially when working with large datasets or when using certain operations like merging or groupby.
c. Data alignment: Misaligned data can lead to unexpected results when using Pandas, so it's important to ensure that data is properly aligned before performing operations.
d. Handling missing data: Pandas uses NaN to represent missing data, but this can lead to issues when working with certain types of data or when performing certain operations.
e. String operations: Pandas string operations can be slow, especially when working with large datasets or when using certain operations like string concatenation.
f. Limited support for certain data types: Pandas may not support certain data types, like datetime with timezone, so you may need to use additional libraries to handle these types of data.
g. Multi-threading: Pandas is not optimized for multi-threading and there are not many options to work around this.

17. What are some problems with numpy?
    a. **Using "nan" in Numpy:** "Nan" stands for "not a number". It was designed to address the problem of missing values. NumPy itself supports "nan" but lack of cross-platform support within Python makes it difficult for the user. That's why we may face problems when comparing values within the Python interpreter.
    b. **Require a contiguous allocation of memory:** Insertion and deletion operations become costly as data is stored in contiguous memory locations as shifting it requires shifting.
18. What is the difference between normalization and standardization in ML?
    a. Normalization scales data to a predefined range, often [0, 1], making it suitable when you want to maintain the relative relationships between data points and ensure all features have the same scale. Standardization, on the other hand, centers data around zero and gives it a standard deviation of 1, making it less sensitive to outliers and particularly useful when algorithms assume a Gaussian distribution of the data or when you aim to make variables more comparable. Standardization doesn't change the interpretation of the data, which can be crucial in some contexts, while normalization may alter the original scale and make interpretation more challenging.
19. Why do we need to normalize or standardize our features?
    a. Normalizing or standardizing features is essential in machine learning to ensure that all features have a similar scale. This improves model performance, aids optimization algorithms, enables feature comparability and interpretability, and enhances the robustness of the model, especially when dealing with outliers.
20. What are the assumptions of Naive Bayes model?
    a. Naive Bayes model assumes that features are conditionally independent given the class, which is a simplifying but often unrealistic assumption. It also assumes that the features used are relevant, the class distribution in the training data is representative of the population, and data is either continuous or discrete. It works best with a sufficient amount of data and when there are no missing

values. Understanding these assumptions is crucial when applying Naive Bayes to real-world problems.

21. What are the assumptions of the linear regression model?
    a. Linearity: The relationship between X and the mean of Y is linear.
    b. Homoscedasticity: The variance of residual is the same for any value of X.
    c. Independence: Observations are independent of each other.
    d. Normality: For any fixed value of X, Y is normally distributed.
    e. Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. Autocorrelation measures the relationship between a variable's current value and its past values.
22. How to handles outliers in ML model? Why are they problematic?
    a. Outliers in machine learning can distort models, lead to biased results, and negatively impact predictive performance. To address outliers, start by identifying them through data inspection and statistical methods. You can transform data, truncate extreme values, use robust models, or remove outliers cautiously. The approach depends on the nature of the data and modeling technique, and addressing outliers during data preprocessing is vital for building accurate and robust machine learning models.
23. What is p value?
    a. The p-value is a statistical measure that tells you how likely your observed data is if a certain hypothesis is true. A small p-value indicates strong evidence against that hypothesis, while a large p-value suggests weak evidence. It's a critical tool in hypothesis testing and research significance.
24. What is t test?
    a. A t-test is a statistical test used to determine if there is a significant difference between the means of two groups. It's commonly employed when sample sizes are small or when the population standard deviation is unknown. The test yields a t-statistic and a p-value, with a small p-value indicating a significant difference between the groups. It's a widely used tool for comparing group means in various fields.
25. What is Z distribution?
    a. Z-distribution, or standard normal distribution, is a bell-shaped curve with a mean of 0 and a standard deviation of 1. It's used to standardize data and make statistical comparisons easier. Z-scores are common in statistics for hypothesis testing and confidence intervals.
26. What is cross validation?
    a. Cross-validation is a resampling technique used in machine learning to assess the performance of a model and reduce the risk of overfitting. It involves dividing a dataset into multiple subsets (often called "folds") and training and testing the model multiple times, each time using a different fold as the test set and the remaining folds as the training set. The results are then averaged to provide a more robust evaluation of the model's performance. Common types of cross-validation include k-fold cross-validation, leave-one-out cross-validation, and stratified cross-validation, each with its own advantages and trade-offs.

Cross-validation helps in estimating how well a model will generalize to new, unseen data and is a crucial step in model evaluation and selection.

27. How do we evaluate the performance of regression model?
    a. To evaluate the performance of a regression model, you can use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), and others. Additionally, analyze residuals and their distribution, check for autocorrelation and heteroscedasticity, and consider feature importance. Choose the evaluation metrics that best suit your specific problem and goals, and use a combination of metrics for a comprehensive assessment of the model's performance.
28. What is mean square error?
    a. Mean squared error (MSE) is a measure of how close a regression model's predictions are to the actual values. It is calculated by taking the average of the squared differences between the predicted and actual values.
29. How we evaluate the performance of classification model?
    a. To evaluate the performance of a classification model, use metrics like accuracy, precision, recall, F1 score, ROC AUC, and others. Analyze the confusion matrix, precision-recall curve, and ROC curve for detailed insights.
30. When to use least square method?
    a. the least squares method is a versatile and widely used technique in statistics, mathematics, and data analysis. It's particularly valuable when you want to find the best-fitting parameters to describe a relationship or model between variables while minimizing the sum of squared differences between observed and predicted values.
31. What is the confusion matrix?
    a. A confusion matrix is a table used to assess the performance of a classification model. It breaks down the model's predictions into true positives (correctly predicted positive), true negatives (correctly predicted negative), false positives (incorrectly predicted positive), and false negatives (incorrectly predicted negative). From these, various metrics are calculated to evaluate the model's performance. It's a useful tool for understanding the strengths and weaknesses of a classification model
32. What is the ROC plot?
    a. The Receiver Operating Characteristic (ROC) plot is a graph used to evaluate the performance of binary classification models. It shows the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) at different classification thresholds. The curve ranges from the bottom left (no true positives and no false positives) to the top right (all true positives and all false positives). The area under the ROC curve (AUC-ROC) summarizes the model's performance, with 0.5 indicating a random classifier and 1.0 indicating a perfect one. It's useful for comparing and selecting models and finding the right balance between sensitivity and specificity.
33. What are the difference between machine learning and deep learning?

   a. Machine learning focuses on manual feature engineering and can work with smaller datasets. Deep learning, a subset of machine learning, automatically learns features from data, requires large datasets, involves deep neural networks, and often lacks interpretability. Deep learning excels in applications like image and speech recognition. The choice depends on the problem, data, and computational resources

34. What are some of the regularization techniques? Can they be used for statistical methods and also for deep learning?
    a. Regularization techniques like L1, L2, Elastic Net, dropout, and others are used to prevent overfitting in both statistical methods (e.g., linear regression) and deep learning. They help models generalize better and are a valuable tool in improving model performance. The specific technique chosen depends on the problem and the model used.
35. What is Elastic Net?
    a. Combines both L1 and L2 regularization, offering a balance between feature selection and coefficient shrinkage. It's applicable to linear regression and some deep learning models.
36. What is dropout?
    a. A deep learning-specific regularization technique, dropout randomly "drops out" (deactivates) a proportion of neurons during training, which prevents the model from relying too heavily on any single neuron. It's commonly used in neural networks, including convolutional and recurrent networks.
37. What is weight decay?
    a. A general regularization technique that encourages smaller weights by adding a penalty term to the loss function. It can be applied to linear regression, logistic regression, and neural networks.
38. What is early stopping?
    a. In deep learning, training is stopped when the model's performance on a validation set starts to degrade, preventing overfitting. This can be used in some deep learning frameworks.
39. What is batch normalization?
    a. In deep learning, it normalizes the input of each layer to mitigate internal covariate shift, which helps with faster convergence and regularization.
40. What is data augmentation?
    a. data augmentation involves creating additional training data by applying transformations like rotation, scaling, and cropping. It can help regularize deep learning models by providing more diverse training examples.
41. What is the difference between Lasso and Ridge Regression?
    a. Lasso Regression (L1) and Ridge Regression (L2) are both used in linear regression to prevent overfitting, but they work differently. Lasso encourages sparsity by making some coefficients exactly zero, which is useful for feature selection. Ridge shrinks coefficients toward zero without forcing them to be exactly zero and is helpful when reducing the impact of all features is important, especially in the presence of collinearity.

42. What is SVM?
    a. Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. It's named after its key concept of identifying "support vectors," which are the data points closest to the decision boundary (hyperplane) that separates classes, and it maximizes the margin between classes. The term "Machine" signifies its automated learning nature.
43. WHat are the loff function of SVM?
    a. Hinge Loss: Measures misclassifications and encourages correct classification by penalizing errors.
    b. Regularization Term: Prevents overfitting by controlling model complexity.
44. What are the hyper parameters of SVM?
    a. C, Regularization parameter: This parameter controls the trade-off between maximizing the margin and minimizing the misclassification error. A larger value of C results in a smaller margin but a lower misclassification error.
    b. Kernel: This parameter specifies the type of kernel used in the SVM. Common kernel types include linear, polynomial, and radial basis function (RBF).
    c. Gamma: This parameter controls the width of the RBF kernel. A smaller value of gamma results in a wider kernel and a larger value results in a narrower kernel.
    d. Degree: This parameter controls the degree of the polynomial kernel.
    e. Tolerance: This parameter controls the tolerance for the stopping criterion.
    f. Coef0: This parameter controls the influence of higher-degree polynomials in the polynomial kernel.
    g. Shrinking: This parameter controls whether to use the shrinking heuristics.
    h. Probability: This parameter controls whether to enable probability estimates, which can be useful for certain types of classification tasks.
45. What is the loss function of Random Forest?
    a. For Classification: The "loss" is primarily measured by the out-of-bag error (OOB error), which is the error rate on the data points that are not used to build a particular decision tree. The lower the OOB error, the better the Random Forest model.
    b. For Regression: Random Forest typically uses mean squared error (MSE) to measure the variance between the predicted values and the actual values. The objective is to minimize MSE across the ensemble of trees.
46. What are the hyperparameters of Random Forest?
    a. Key hyperparameters for Random Forest include n_estimators (number of trees), max_depth (tree depth), min_samples_split and min_samples_leaf (split conditions), max_features (feature selection), and more. These parameters affect the model's performance and should be tuned for specific tasks.
47. What method to use to test the performance of the Random Forest model?
    a. To test the performance of a Random Forest model, use train-test splitting, cross-validation, OOB error (if enabled), confusion matrices, ROC and precision-recall curves for classification, mean squared error for regression, feature importance analysis, and error analysis. The choice of method depends on the problem and goals.

48. What is stemming?
    a. Stemming is a text processing technique that shortens words to their base form, such as converting "jumping" to "jump." It's used to group similar words together for text analysis but can have limitations in accuracy.
49. What is lemmatization?
    a. Lemmatization is a text processing technique that reduces words to their base or dictionary form, preserving their meaning and linguistic structure. It's more accurate than stemming but can be computationally intensive.
50. What is the difference between lemmatization and stemming?
    a. Stemming is a faster and less precise text processing technique that reduces words to a common root form. Lemmatization is a more accurate but slower technique that reduces words to their dictionary form, considering the word's part of speech and context. The choice depends on the specific NLP task and desired accuracy.
51. What is tf-idf? Why it is important?
    a. TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical measure used in natural language processing to assess the importance of words in documents. It helps identify key terms, rank document relevance in search engines, categorize documents, and more. It combines term frequency (how often a term appears in a document) with inverse document frequency (how rare a term is across all documents) to assign weights to terms. Higher weights are given to terms that are frequent in a document but rare in the corpus, indicating their significance. TF-IDF is crucial for tasks like information retrieval, document classification, and content recommendation.
52. What are the assumptions of logistic regression model?
    a. Linearity of the logit.
    b. Independence of errors.
    c. No multicollinearity.
    d. Large sample size.
    e. Linearity of independent variables and log-odds.
    f. Absence of outliers.
    g. Binary dependent variable.
    h. No endogeneity.
    i. No perfect separation or prediction.
53. How to assess the performance of logistic regression model?
    a. Use a confusion matrix to measure true positives, true negatives, false positives, and false negatives.
    b. Calculate accuracy, precision, recall, and F1 score for classification performance.
    c. Plot ROC and Precision-Recall curves and compute AUC-ROC and AUC-PR.
    d. Consider log-loss, Brier score, and calibration for probability estimates.
    e. Evaluate residuals and conduct cross-validation to check for overfitting.
    f. Use metrics like AIC and BIC for model selection.
    g. Choose the appropriate metrics based on the problem and dataset characteristics.

54. What are the hyperparameters of logistic regression model?
    a. Key hyperparameters for logistic regression include "Penalty" (L1 or L2 regularization), "C" (inverse of regularization strength), "Solver" (optimization algorithm), "Multi-Class Handling," "Class Weight," "Max Iterations," and "Tolerance." The choice of hyperparameters depends on the problem and dataset.
55. What is multilabel binarization? Why is it important?
    a. Multilabel binarization is a technique to convert multilabel classification problems into a binary format. It transforms the data into a binary matrix where each column represents a label, allowing the use of traditional binary classifiers for each label. This is important for tasks where instances can have multiple associated labels, such as text classification and image tagging.
56. What are some of the parallel processing techniques? How to use them?
    a. Parallel processing techniques include multithreading, multiprocessing, distributed computing, SIMD, GPU parallelism, and pipeline processing. To use them, consider your task's nature, available hardware, and select the appropriate technique based on programming languages, libraries, and platforms to improve performance and efficiency.
57. What is data leakage? How we ensure there is no data leakage?
    a. Data leakage is when external information is mistakenly used to create or evaluate a model, leading to unreliable results. Prevent it by:
        i. Cleaning and transforming data only within the training dataset.
        ii. Avoiding future information in feature engineering or selection.
        iii. Proper cross-validation.
        iv. Being cautious with external data.
        v. Continuously monitoring model performance and data sources.
        vi. Documenting data processes to track potential leakage sources.
58. How to handle if there is data leakage?
    a. Identify the source of leakage.
    b. Remove the leakage source from the data.
    c. Reevaluate models using corrected data.
    d. Retrain and revalidate models if needed.
    e. Document the issue and solutions.
    f. Prevent recurrence with best practices.
    g. Monitor model performance regularly.
59. What is multinomial distribution?
    a. The multinomial distribution is a probability distribution used when there are multiple categories in an experiment. It describes the probabilities of observing specific combinations of outcomes in a fixed number of trials, where each outcome has an associated probability.
60. What is multicollinearity?
    a. Multicollinearity is **a statistical concept where several independent variables in a model are correlated**. Two variables are considered to be perfectly collinear

if their correlation coefficient is +/- 1.0. Multicollinearity among independent variables will result in less reliable statistical inferences.

61. How to check whether multicollinearity occurs?
    a. The first simple method is to plot the correlation matrix of all the independent variables.
    b. The second method to check multi-collinearity is to use the Variance Inflation Factor(VIF) for each independent variable.
62. How to deal multicollinearity?
    a. Remove some of the highly correlated independent variables.
    b. Linearly combine the independent variables, such as adding them together.
    c. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.
63. Why is collinearity a problem for ML models?
    a. Collinearity is a problem for ML models because it introduces instability, reduces model interpretability, and leads to unreliable predictions. It can cause coefficient estimates to be unstable, increase model variance, and make it harder to identify the true importance of predictor variables. Addressing collinearity is crucial to building reliable and interpretable models.
64. What is power in statistics?
    a. Power is **the probability of rejecting the null hypothesis when in fact it is false**.
65. What is type 1 and type 2 error?
    a. Type 1 Error (False Positive) is when a true null hypothesis is mistakenly rejected, and Type 2 Error (False Negative) is when a false null pothesis is not rejected. The significance level (α) controls Type 1 errors, and the power (1 - β) relates to Type 2 errors. Balancing these errors depends on the specific goals of a hypothesis test.
66. What is chi-square test?
    a. A chi-square test is a statistical test used **to compare observed results with expected results**. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.
67. What is the loss function for logistic regression, LDA, Naive Bayes and linear regression?
    a. The loss function used in logistic regression is the log loss, also known as cross-entropy loss. It measures the performance of a classification model whose output is a probability value between 0 and 1.
    b. The loss function used in LDA is the Bayesian criterion, also known as the Bayes decision rule. The goal of LDA is to find a linear combination of features that maximizes the ratio of between-class variance to within-class variance.
    c. The loss function used in Naive Bayes is the negative log-likelihood. The negative log-likelihood measures the difference between the true label and the predicted probability distribution.

   d. The loss function used in linear regression is the mean squared error (MSE). The mean squared error measures the average squared difference between the predicted and actual values. Alternatively, another common loss function for linear regression is the mean absolute error (MAE). Both MSE and MAE measure the difference between the true value and the predicted value and the difference is that MSE is sensitive to the outliers while MAE is not.

68. What is the difference between correlation and covariance?
   a. Correlation and covariance are both measures of the relationship between two variables, but they are calculated differently.
   b. Correlation is a standardized measure of the linear relationship between two variables, expressed as a value between -1 and 1. A correlation of 1 indicates a perfect positive linear relationship, a correlation of -1 indicates a perfect negative linear relationship, and a correlation of 0 indicates no linear relationship.
   c. Covariance, on the other hand, is a measure of the degree to which two variables change together. It is not standardized, and its value can be any real number. A positive covariance indicates that the variables increase or decrease together, and a negative covariance indicates that one variable increases as the other decreases. Covariance of A and B 0 does not imply independence. If data are multivariate normal, then only the Covariance of A and B 0 imply independence
69. What is a multivariate normal distribution?
   a. A multivariate normal distribution is a probability distribution that describes the joint behavior of multiple random variables. It's characterized by a mean vector and a covariance matrix, showing the average values and relationships between these variables. This distribution is used in various fields to model correlated, multidimensional data.
70. What are some advantages of ML models compared to DL models?
   a. Advantages of Machine Learning (ML) models over Deep Learning (DL) models include better interpretability, data efficiency, faster training, lower resource requirements, less feature engineering, reduced risk of overfitting, and alignment with human expertise. The choice depends on the problem and dataset characteristics.
71. How KL divergence loss is different from cross entropy loss?
   a. KL Divergence (KLD Loss) measures the difference between two probability distributions and is used in generative modeling, while Cross-Entropy Loss (CE Loss) measures the discrepancy between predicted and true class labels and is used in classification tasks. KLD is asymmetric, and CE is symmetric. They serve different purposes
72. What is cross entropy?
   a. Cross-entropy is a loss function used in classification tasks that measures the dissimilarity between predicted probabilities and true class labels. It encourages models to assign high probabilities to the correct classes.
73. How to handle if decision tree models are over fitted?
   a. Prune the tree by setting depth limits.
   b. Apply a minimum samples per leaf/split threshold.

c. Limit the number of features considered at each split.
d. Use cross-validation to assess and fine-tune the model.
e. Consider ensemble methods like Random Forest or Gradient Boosting.
f. Carefully select and preprocess features.
g. Experiment with regularization parameters.
h. Augment the training dataset with more data if possible.
i. Implement early stopping for gradient-boosted trees.
j. Simplify the tree by reducing depth or increasing minimum samples per leaf.
k. Collect more data if feasible.
l. Engage in feature engineering to create more informative features.

74. What is the difference between bootstrap and random forest?
    a. Bootstrap is a resampling technique used to create subsets from data. Random Forest, on the other hand, is an ensemble learning method that combines multiple decision trees by using bootstrap sampling and random feature selection to enhance predictive accuracy. Random Forest incorporates bootstrap as part of its modeling process.
75. Explain naive in Naive Bayes?
    a. This algorithm is based on the Bayes theorem. It describes the probability of an event based on the prior knowledge of conditions related to that event.
76. What is pruning?
    a. Pruning is a technique in Machine Learning and search algorithms that reduce the size of decision trees. It removes sections of the tree that provide little power to classify instances. Thus, the removal of sub-nodes of a decision node is called pruning.
77. Explain cost function?
    a. The Cost Function is also referred to as "loss" or "error." It is a measure to check how good the model's performance is. It's used to compute the error of the output layer during backpropagation.
78. How do you interpret the logistic regression?
    a. Interpreting logistic regression involves examining coefficients (beta values), which represent the impact of predictor variables on the log-odds of a binary outcome. Positive coefficients suggest increased odds, while negative coefficients imply decreased odds. Assess significance, direction, magnitude, and confidence intervals. Also, evaluate model fit, goodness of fit, ROC curves, and real-world implications for a comprehensive interpretation.
79. How does dropout work?
    a. Dropout randomly deactivates some neurons during training to prevent overfitting. It acts like training and testing multiple models and promotes robustness and generalization. During testing, all neurons are used with adjusted weights.
80. What is the difference between bagging and boosting?
    a. Bagging creates multiple base models in parallel from bootstrapped data and combines their predictions by averaging or voting. Boosting builds base models sequentially, focusing on misclassified examples, and combines their predictions

with weighted voting. Bagging reduces variance, while boosting aims to reduce bias and for improved performance.

81. Explain in detail how 1D CNN works?
    a. A 1D CNN processes one-dimensional data, like time series or text sequences. It uses convolutional filters to capture local patterns in the data, followed by pooling to reduce dimensionality. After flattening, fully connected layers learn complex relationships. Activation functions introduce non-linearity, and the output layer produces the network's result. Training involves adjusting filter weights to minimize a loss function. 1D CNNs excel in tasks involving ordered data and have applications in speech recognition, natural language processing, and time series analysis.
82. Having a categorical variable with thousands of distinct values, how would you encode it?
    a. Consider using frequency or count encoding.
    b. Explore target encoding (mean encoding) if you're working on classification tasks.
    c. Group rare categories into an "other" category to reduce dimensionality.
    d. Use feature hashing to map categories into a fixed number of hash buckets.
    e. If using deep learning models, employ embeddings to represent categories.
    f. Explore domain-specific techniques, such as word embeddings for text data.
    g. Apply dimensionality reduction techniques or feature selection to manage dimensionality. The choice depends on the specific problem and the nature of the categorical variable.
83. How to handle imbalanced dataset?
    a. Resample data: Oversample the minority class or undersample the majority class.
    b. Adjust misclassification costs for different classes.
    c. Augment data with synthetic samples.
    d. Collect more data for the minority class.
    e. Use ensemble methods, like bagging or boosting.
    f. Consider anomaly detection techniques.
    g. Change the decision threshold for classification.
    h. Evaluate with appropriate metrics (precision, recall, F1-score).
    i. Generate synthetic data with GANs.
    j. Apply class weighting in some algorithms.
    k. Experiment with hybrid approaches combining methods.
84. What is LSTM? Why use LSTM?
    a. LSTM (Long Short-Term Memory) is a type of neural network used for tasks involving sequences, like text, speech, and time series data. It's favored because it can handle long sequences, capture complex patterns, and mitigate training problems seen in traditional RNNs. LSTMs are widely used in applications where sequential information is essential and have achieved state-of-the-art performance in many such tasks.
85. What is VIF

   a. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.
   b. A rule of thumb for interpreting the variance inflation factor:
      i. 1. 1 = not correlated.
      ii. 2. Between 1 and 5 = moderately correlated.
      iii. 3. Greater than 5 = highly correlated.

86. Explain different time series analysis models. What are some time series models other than Arima?
   a. ARIMA for basic forecasting.
   b. Exponential Smoothing for weighted averages.
   c. State Space Models for hidden state estimation.
   d. Prophet for data with strong seasonality.
   e. GARCH for financial volatility.
   f. STL for decomposition.
   g. SARIMA for seasonal data.
   h. VAR for interrelated series.
   i. Neural Networks (LSTM, GRU) for complex dependencies.
   j. ARIMAX with exogenous variables.
   k. TBATS for complex data patterns.
   l. Neural Prophet for deep learning-based forecasting.
87. What is Arima?
   a. ARIMA (AutoRegressive Integrated Moving Average) is a time series forecasting model. It combines past values (AR), differencing (I), and past error terms (MA) to make predictions. It's effective for data with trends and seasonality, like stock prices and demand forecasting. ARIMA is defined by three parameters: p (AR order), d (differencing order), and q (MA order).
88. What is Anova?
   a. ANOVA (Analysis of Variance) is a statistical method used to compare means in two or more groups to determine if there are significant differences between them. It's commonly used in experimental research to assess the impact of different factors on a dependent variable.
89. How does a neural network with one layer and one input and output compare to a logistic regression?
   a. A neural network with one layer, one input, and one output neuron has the advantage of capturing non-linear relationships, thanks to its activation function. In contrast, logistic regression is a linear model and is suitable for simpler problems. The choice depends on the complexity of the problem.
90. How would you explain hypothesis testing for a newbie?
   a. Hypothesis testing is a method to make decisions based on data. It involves forming two hypotheses (null and alternative), collecting data, performing statistical tests, and comparing the results to a significance level. If the test

shows strong evidence against the null hypothesis, you accept the alternative. It's a way to evaluate beliefs or questions using evidence.

91. What is OLS regression?
    a. OLS (Ordinary Least Squares) regression is a statistical method for modeling the relationship between a dependent variable (what you want to predict) and one or more independent variables (predictors). It finds the line or hyperplane that best fits the data by minimizing the sum of squared differences between observed and predicted values. It's used for prediction, hypothesis testing, and understanding variable relationships.
92. How do you interpret OLS regression results?
    a. Examine coefficients: Positive means predictor variable increases with the dependent variable; negative means decrease.
    b. Consider the intercept: Represents the expected value when predictors are zero.
    c. Check R-squared: A higher value indicates a better fit.
    d. Look at p-values: Small p-values (<0.05) suggest significance.
    e. Analyze confidence intervals: Narrow intervals are more precise.
    f. Assess residuals: Ensure they are normally distributed and have constant variance.
    g. Evaluate F-statistic: Tests overall model significance.
    h. Check for multicollinearity.
    i. Verify regression assumptions.
    j. Consider practical significance in addition to statistical significance.
93. What is a confidence interval?
    a. A confidence interval is a range of values that estimates a population parameter from a sample, with a specified level of confidence. It provides a measure of uncertainty and is used to make inferences about the population based on sample data. The confidence level determines the range's width, and a higher level means a wider range.
94. What is the difference between linear regression and a t-test?
    a. Purpose: Linear regression models relationships between variables and is used for prediction, while a t-test compares means of two groups to test for differences.
    b. Number of Variables: Linear regression involves multiple predictors, while a t-test typically involves only one dependent variable and two groups.
    c. Output: Linear regression provides coefficients and model fit measures, while a t-test gives a t-statistic and p-value for mean comparisons.
    d. Use Cases: Linear regression is for prediction and understanding relationships; t-tests are for comparing group means.
    e. Assumptions: Both have different assumptions; linear regression assumes linearity and normality, while t-tests assume normality and equal variances in groups.
95. Why is cross-validation so popular?
    a. Assesses model performance robustly.
    b. Helps avoid overfitting.

c. Optimizes hyperparameters effectively.
d. Makes better use of data.
e. Provides reliable generalization estimates.
f. Reduces sampling bias.
g. Works in diverse scenarios.
h. Detects issues and enhances transparency and reproducibility in model assessment.

96. What is variance? What does it imply? Give example.'
    a. Variance measures the spread or variability in a dataset. High variance means data points are more spread out, while low variance indicates they are close to the mean. For example, if test scores vary widely, the variance is high; if they are similar, the variance is low.
97. What is standard deviation? What does it mean?
    a. Standard deviation measures the spread or variability in a dataset. High standard deviation means data points are more spread out, while low standard deviation indicates they are close to the mean. It's expressed in the same units as the data and is more interpretable than variance.
    b. It is the square root of variance.
    c. For normal distribution
        i. 68% of the data lie within 1 SD
        ii. 95% of the data lie within 2 SD
        iii. 98% of the data lie within 3 SD
98. How to measure the dispersion of data?
    a. Quartiles: Q1, Q2, Q3
    b. IQR: Q3 - Q1, i.e., the height of the box
    c. Five number summary: min, Q1, median, Q3, max
    d. Boxplot
    e. Outliers: value lower/higher than 1.5 * IQR
99. What is a quantile plot? What do we get from the quantile plot? Why is it important?
    a. A quantile plot (Q-Q plot) is a graphical tool to check if a dataset follows a specific theoretical distribution (e.g., normal distribution). It compares data quantiles to expected quantiles. A straight line on the plot suggests a good fit, deviations indicate differences from the theoretical distribution. It's important for validating statistical assumptions and assessing data quality.
100. Some notes
    a. Histogram gives univariate plot meaning plot against a single variable
    b. Scatter plot gives bivariate plot
    c. Redundant attributes may be detected by correlation analysis
    d. Preprocessing changes the data and introduces bias

Problem Type	Last-layer Output Nodes	Hidden-layer activation	Last-layer activation	Loss function
Binary classification	1	RELU (first choice), Tanh (for RNNs)	Sigmoid	Binary Crossentropy
Multi-class, single-label classification	Number of classes		Softmax	Categorical Crossentropy
Multi-class, multi-label classification	Number of classes		Sigmoid (one for each class)	Binary Crossentropy
Regression to arbitrary values	1		None	MSE
Regression to values between 0 and 1	1		Sigmoid	MSE/Binary Crossentropy

e.

101. What are some data cleaning techniques?
    a. Filling missing values
        i. Fill by global constant, mean value of all sample belonging to same class class or infered based on decision tree
    b. Smooth noisy data
        i. Example of noisy data maybe some values incomplete or -ve value for salary or age
        ii. Smooth by bin means or fitting into regression function or clustering
    c. Identify or remove outliers
    d. Resolve inconsistencies caused by data integration
102. What is bayesian statistics? What are the assumptions? Why is it important?
    a. Bayesian statistics is a statistical approach that combines prior knowledge and observed data to update beliefs and make predictions. It assumes prior information, allows for subjectivity, and ensures coherent probability reasoning. It's important for incorporating prior knowledge, quantifying uncertainty, informed decision-making, modeling flexibility, and predictive inference.
103. What are some data transformation techniques?
    a. Smoothing: remove noise from data
    b. Aggregation
    c. Generalization: concept hierarchy climbing
    d. Normalization
    e. Attribute or feature construction: new feature from given ones
104. Formula for min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

105. The formula for Z-score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

106. What are some benefits of feature selection?
    a. Simply data/models
    b. Improve interpretability
    c. Avoid overfitting
    d. Reduce running time (train and test)
107. What are some benefits of feature generation?
    a. Increase discriminative power
    b. Most beneficial for simple models
108. What is A/B testing?
    a. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.
    b. Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.
109. how dropout works?
    a. Dropout works by randomly blocking off a fraction of neurons in a layer during training. Then, during prediction (after training), Dropout does not block any neurons.
110. how would you improve a classification model that suffers from low precision?
    a. Raising the classification threshold typically increases precision; however, precision is not guaranteed to increase monotonically as we raise the threshold. Probably increase. In general, raising the classification threshold reduces false positives, thus raising precision.
111. What is the difference between using 1D CNN and LSTM?
    a. An LSTM is designed to work differently than a CNN because an LSTM is usually used to process and make predictions given sequences of data (in contrast, a CNN is designed to exploit "spatial correlation" in data and works well on images and speech).
112. What is misclassification error?

a. Misclassification is the error in categorical data in which the observed category is different from the underlying one.
b. Here is how to calculate the misclassification rate for the model: Misclassification Rate = # incorrect predictions / # total predictions.

113. What is the difference between stemming and lemmatization?
a. Stemming and lemmatization are methods used by search engines and chatbots to analyze the meaning behind a word. Stemming uses the stem of the word, while lemmatization uses the context in which the word is being used. For instance, stemming the word 'Caring' would return 'Car'. For instance, lemmatizing the word 'Caring' would return 'Care'.

114. What does a high TF-IDF score mean?
a. The higher the TF-IDF score the more important or relevant the term is; as a term gets less relevant, its TF-IDF score will approach 0.

115. What loss function used in logistic regression?
a. Any deep learning or machine learning model tries to optimize a loss function. The loss function for logistic regression is Binary cross entropy. Cross-entropy can be used as a loss function when optimizing classification models like logistic regression and artificial neural networks.

116. what is binary cross entropy?
a. Binary cross entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value. That means how close or far from the actual value.

117. What is loss function for linear regression?
a. The most commonly used loss function for Linear Regression is Least Squared Error, and its cost function is also known as Mean Squared Error(MSE).

118. What is the loss function of random forest?
a. For the random forest classifier, this is the Gini impurity.

119. What is Gini impurity?
a. Gini impurity is a measure of impurity or disorder in a dataset, often used in decision trees and machine learning. It ranges from 0 (pure set) to 0.5 (max impurity) and helps find the best splits for classifying data. It's particularly useful for creating decision rules in classification tasks.

120. What is Gini Index?
a. The Gini index is a measure of economic inequality. It ranges from 0 (perfect equality) to 1 (perfect inequality). It's used to assess and quantify disparities in income, wealth, or any ranked attribute within a population. A lower Gini index indicates more equal distribution, while a higher index suggests greater inequality.

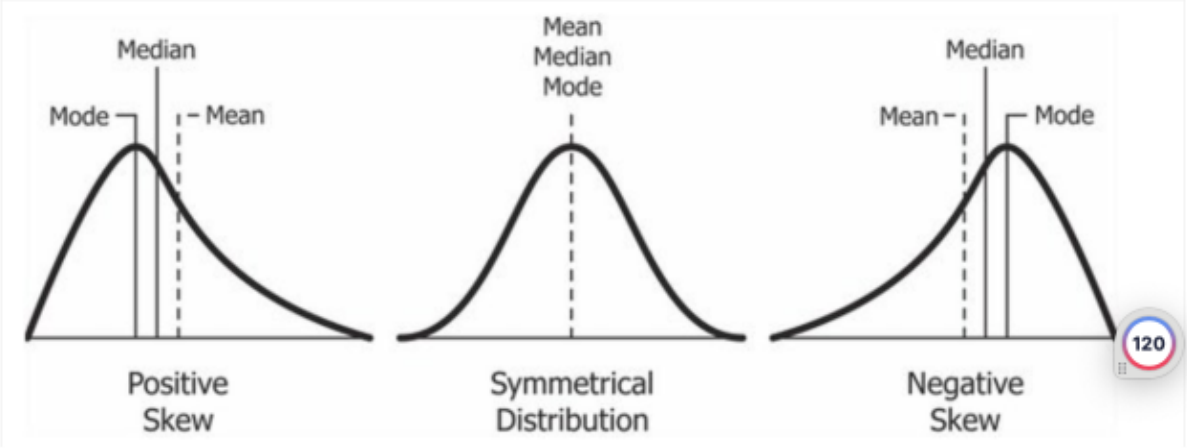
121. What is the difference between gini index and gini impurity?
a. The Gini index measures income or wealth inequality in economics, while Gini impurity assesses the impurity of data for classification in machine learning. They have different purposes and usage contexts.
b. The lower the Gini impurity, the better the feature is for splitting the dataset.

Select the attribute that gives the highest information gain. Information gain requires measure of impurity. Impurity measure that it uses is the entropy of the class distribution, which is a measure from the information theory.

When node is pure, measure should be zero for impurity. When impurity maximal, measure should be maximal. Measure should ideally obey multistage property. Entropy is the only function that satisfies all three properties.

Information gain is biased towards choosing attributes with a large number of values. Gain ratio is a modification of the information gain that reduces its bias towards attributes with many values.

122. what is the role of optimizer in neural network?
    a. An optimizer is a function or an algorithm that modifies the attributes of the neural network, such as weights and learning rate. Thus, it helps in reducing the overall loss and improve the accuracy.
123. What is skewness?
    a. In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.



124. What is the difference between privacy and security?
    a. privacy is primarily concerned with an individual's right to personal data and its confidentiality, while security has a broader scope, focusing on safeguarding entire systems and organizations from a wide range of threats. Both privacy and security are crucial, and they often overlap in efforts to protect data and individuals' rights.
125. How do we decide on k-means clustering?

a. Elbow Method: Plot within-cluster sum of squares (WCSS) against k and look for the "elbow" point.
b. Silhouette Score: Calculate the silhouette score for different k values and choose the one with the highest score.
c. Gap Statistics: Compute the gap statistic and select the k with the largest gap.
d. Domain Knowledge: Use your domain expertise to make an informed guess.
e. Hierarchical Clustering: Build a dendrogram and determine k from its structure.
f. Cross-Validation: Employ cross-validation to assess k-means performance.
g. Visual Inspection: Visualize clusters with different k values to make a subjective judgment.

126. What are other clustering methods?
    a. Hierarchical Clustering:
        i. Agglomerative (bottom-up) or divisive (top-down) approaches.
        ii. Produces a tree-like structure (dendrogram) that can be cut at different levels to obtain clusters.
    b. Agglomerative Clustering:
        i. A hierarchical clustering approach that starts with each data point as a single cluster and merges them iteratively.
        ii. Agglomerative clustering can be used with different linkage criteria, such as Ward's, single, complete, or average linkage.
    c. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):
        i. Groups together data points that are close to each other and separates outliers.
        ii. Works well with unevenly distributed clusters and can find clusters of arbitrary shapes.
127. Why choose k-means clustering?
    a. Your data is numerical.
    b. Clusters are roughly spherical and similar in size.
    c. You want a simple, efficient method.
    d. Hard data point assignment to clusters is needed.
    e. Dealing with large datasets.
    f. You have knowledge or can estimate initial cluster centers.
    g. Initial data exploration or hypothesis generation is required.
    h. You need quantitative evaluation metrics.
    i. Clusters should be easy to interpret.
    j. Scalability is important.
128. What are the limitations of k-means clustering?
    a. k-means has limitations, such as sensitivity to the initial cluster centers and a tendency to converge to local optima. If your data does not meet the assumptions of k-means (e.g., clusters are not spherical), or you need more flexibility in cluster shapes, you might consider other clustering methods like hierarchical clustering, DBSCAN, or Gaussian Mixture Models (GMM).
129. What are some classification algorithms and when to use them?
    a. Logistic Regression:

   i. Use when you have a binary or multinomial outcome variable and want to model the probability of an individual belonging to a specific population.
   ii. It's suitable for cases where the relationship between predictors and population membership is assumed to be linear.

b. Decision Trees:
   i. Use when you want an interpretable model for classifying individuals into populations.
   ii. Decision trees are helpful when the relationship between predictors and population membership is non-linear and can handle both binary and multi-class classification.

c. Random Forest:
   i. Appropriate when you need a more accurate and robust classification model than a single decision tree.
   ii. It's especially useful when dealing with a large number of predictors and you want to reduce overfitting.

d. Gradient Boosting:
   i. Useful for improving the accuracy of population classification by combining the predictions of multiple weak learners.
   ii. Gradient boosting algorithms like XGBoost or LightGBM are effective when high accuracy is crucial.

e. Naive Bayes:
   i. Appropriate for text classification problems, such as sentiment analysis or document categorization.
   ii. Works well when you have a limited amount of training data and predictors are conditionally independent.

f. k-Nearest Neighbors (k-NN):
   i. Use when you want to classify individuals based on the similarity of their attributes to those of their neighbors.
   ii. K-NN is suitable when the underlying data distribution is not well-defined and you want to take into account local patterns.

g. Support Vector Machines (SVM):
   i. Suitable for both binary and multi-class classification when you want to find the optimal hyperplane that maximizes the margin between different populations.
   ii. Works well when you have a relatively small dataset and want a powerful classifier.

h. Neural Networks:
   i. Employ deep learning techniques when you have large datasets, complex data, and the potential for discovering intricate patterns.
   ii. Neural networks can handle various types of classification problems but require substantial data and computational resources.

i. Ensemble Methods:
   i. Consider ensemble methods like AdaBoost or Bagging when you want to combine multiple classifiers to improve classification performance.

   ii. They are useful in scenarios where a single algorithm may not be sufficient.
j. Cluster Analysis:
   i. When you don't have predefined categories but want to group individuals into populations based on similarities in their attributes, use clustering algorithms like k-means, hierarchical clustering, or DBSCAN.

130. What is the difference between Bagging and Boosting?
   a. Bagging (Bootstrap Aggregating):
      i. Trains multiple independent models in parallel.
      ii. Averages (regression) or majority vote (classification) for the final prediction.
      iii. Reduces variance and overfitting.
   b. Boosting:
      i. Trains a sequence of models sequentially, each correcting errors of the previous.
      ii. Assigns weights to models based on performance.
      iii. Reduces bias, often leading to higher accuracy.
      iv. Common algorithms: AdaBoost, Gradient Boosting (e.g., XGBoost, LightGBM).

131. For which models we need to standardize or normalize the data?
   a. typically need to standardize or normalize data for models that rely on distances or gradients, such as K-Means, PCA, SVM, K-NN, and neural networks. Models like decision trees and random forests are less sensitive to feature scales and may not require this preprocessing.

132. Is searchable encryption a homomorphic encryption?

   Searchable encryption and homomorphic encryption are distinct cryptographic techniques that serve different purposes within the realm of encrypted data processing.

   Searchable Encryption (SE) is designed to allow encrypted data to be searchable without revealing the data's plaintext. It focuses on the specific task of performing search queries, such as keyword searches, on data that remains encrypted during the search process. This capability is particularly useful for ensuring data privacy in environments like cloud storage, where users may need to retrieve information without exposing the actual contents to the service provider.

   Homomorphic Encryption (HE), in contrast, is a more versatile encryption scheme that enables a wide range of computations on encrypted data. With homomorphic encryption, it's possible to perform arithmetic operations and other computations on ciphertexts in a way that, when decrypted, the result matches what would have been obtained if the operations had been performed on the plaintext. This enables complex data analysis and processing while maintaining the confidentiality of the underlying data.

While both technologies aim to enhance data security and privacy by allowing certain types of operations on encrypted data, searchable encryption is specialized for search functionality, whereas homomorphic encryption supports a broader scope of computational operations. They are not interchangeable but rather complementary, depending on the specific privacy and security requirements of a given application.

133. What is the formula for sample variance

Sample Variance (s^2)

The formula for sample variance is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Where:

- x_i represents each value in the sample,
- $\bar{x}$ is the sample mean (average),
- n is the number of observations in the sample,
- $\sum$ denotes the summation over all n observations.

134. Difference between API and endpoint

Imagine you're at a big, bustling food market that sells all kinds of foods and ingredients. The market itself is like an API (Application Programming Interface), a place where you can request various services or information. When you go to a specific stall to buy fruits, another stall to get vegetables, and yet another to pick up spices, each stall represents an "endpoint." Each

endpoint is a specific location within the market (API) where you can ask for something specific, like getting the price of apples, the variety of carrots available, or the types of spices on offer.

So, in simple words:

API is the entire system or service that provides access to a range of functionalities or data. Endpoint is a specific address (like a URL for a web service) within that system where you can access a particular function or piece of data.

135. how can skewness affect the performance of the predictive models

Skewness in the distribution of data can significantly impact the performance of predictive models, particularly those that assume or perform better with normally distributed data. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The effects of skewness on predictive models vary depending on the type of model and the nature of the data, but several common impacts include:

Model Accuracy: Many algorithms, especially linear models like linear regression, assume that the predictors are normally distributed. Skewness can lead to biases in the estimation process, affecting accuracy. Models may not adequately capture the relationship between features and the target variable if the data is heavily skewed.

Model Interpretability: Skewness can make it harder to interpret model coefficients because the relationship between predictors and the response variable might not be linear or might vary differently across the range of the predictor values.

Convergence Issues: Algorithms that rely on gradient descent for optimization may have difficulty converging to a solution if the data is highly skewed. This is because skewness can lead to outliers that disproportionately affect the loss function, making it harder for the model to find the global minimum.

Impact on Variance and Bias: Skewness can increase the variance of the model predictions for certain areas of the input space while increasing bias in others. This imbalance can make it harder to achieve a good trade-off between variance and bias, essential for building models that generalize well to new data.

Outlier Sensitivity: Skewed data is often accompanied by outliers. Many models, particularly those that assume data normality, are sensitive to outliers. This sensitivity can lead to overfitting, where the model learns the noise in the training data instead of the actual signal, reducing its generalization capability.

Effect on Model Validation: When using skewed data, traditional model validation techniques like cross-validation might give misleading results because the model's

performance can vary significantly across different parts of the data distribution. This variation makes it challenging to assess the model's true predictive power.

To mitigate these effects, data scientists often apply transformations to the data to reduce skewness, such as logarithmic, square root, or Box-Cox transformations. These transformations can help make the data more symmetric, potentially improving model performance. Additionally, using algorithms that are robust to outliers or that do not assume normality can also help manage the effects of skewness.

136. What do you know about the concept of confounding variables?

“Confounding variables are external factors that can distort the true relationship between the independent and dependent variables, leading to misleading conclusions. For example, in a study examining the relationship between exercise and heart health, age could be a confounding variable as it affects both the level of exercise and the risk of heart disease.

Various methods, such as stratification, matching, and multivariate regression analysis, can be used to identify and control for confounding variables. Carefully evaluating confounding variables is essential to ensuring the validity and reliability of data analysis and predictive modeling results.”

137. What are some of the implications of false positive and false negative for a company?

False Positives Implications:

Wasted Resources: Time and money spent addressing non-issues.
Operational Inefficiency: Unnecessary actions reduce efficiency.
Customer Irritation: Inappropriate responses can annoy customers.
Distrust in Systems: Over time, users may ignore important alerts.

False Negatives Implications:

Security Risks: Undetected threats can lead to breaches and data loss.
Quality Issues: Defective products could reach customers, harming reputation.
Missed Opportunities: Failure to identify leads or issues affects sales and satisfaction.
Compliance Risks: Overlooking non-compliance can result in fines and legal trouble.
Balancing the minimization of false positives and negatives is key to operational success and maintaining trust.

138. boosting uses original data or bootsrap data?

Boosting uses the original dataset, adjusting weights of instances based on previous errors, while bootstrapping, used in bagging, creates multiple samples from the original data for training different models.

139. How to interpret coefficients of the logistic regression model

Intercept: The value of the log-odds of the outcome when all the predictors are held at zero. It's the baseline log-odds of the outcome.
Coefficients : Each coefficient represents the change in the log-odds of the outcome for a one-unit change in the predictor, holding all other predictors constant.
Positive Coefficient: A positive coefficient means that as the predictor increases, the log-odds of the outcome occurring increases, and thus the probability of the outcome occurring also increases.
Negative Coefficient: A negative coefficient means that as the predictor increases, the log-odds of the outcome occurring decreases, and thus the probability of the outcome occurring decreases.

140. What is the difference between standard error and standard deviation.

The standard deviation describes variability within a single sample. The standard error estimates the variability across multiple samples of a population.

141. When to use t distribution?

when working problems when the population standard deviation (σ) is not known and the sample size is small (n<30). General Correct Rule: If σ is not known, then using t-distribution is correct. If σ is known, then using the normal distribution is correct.

142. What is the difference between t distribution and normal distribution?

The normal distribution assumes that the population standard deviation is known. The t-distribution does not make this assumption. The t-distribution is defined by the degrees of freedom. These are related to the sample size.

The normal distribution is the most commonly used distribution in all of statistics and is known for being symmetrical and bell-shaped. A closely related distribution is the t-distribution, which is also symmetrical and bell-shaped but it has heavier "tails" than the normal distribution.

143. When to use t distribution

when the population standard deviation (σ) is not known and the sample size is small (n<30). General Correct Rule: If σ is not known, then using t-distribution is correct. If σ is known, then using the normal distribution is correct.

144. What is residual standard error?

The residual standard error is used to measure how well a regression model fits a dataset. In simple terms, it measures the standard deviation of the residuals in a regression model.

145. What is R^2?

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

146. how to interpret the r squared value in regression?

The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model.

147. how to interpret the residual standard error ?

The residual standard error is the standard deviation of the residuals – Smaller residual standard error means predictions are better

148. What is f statistic in linear regression?

This f statistic indicates whether the regression model provides a better fit to the data than a model that contains no independent variables. In essence, it tests if the regression model as a whole is useful.

F is a test for statistical significance of the regression equation as a whole. It is obtained by dividing the explained variance by the unexplained variance.

149. how to interpret f statistic in linear regression?

A large F-statistic value proves that the regression model is effective in its explanation of the variation in the dependent variable and vice versa. On the contrary, an F-statistic of 0 indicates that the independent variable does not explain the variation in the dependent variable.

150. difference between sd and variance?

Standard deviation measures how far apart numbers are in a data set. Variance, on the other hand, gives an actual value to how much the numbers in a data set vary from the mean. Standard deviation is the square root of the variance and is expressed in the same units as the data set.

151. Formula for variance

Formula

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

S^2 = sample variance

x_i = the value of the one observation

$\bar{x}$ = the mean value of all observations

n = the number of observations

152. What does the p value tell you about

The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.

153. what is the standard error of a coefficient?

The standard deviation of an estimate is called the standard error. The standard error of the coefficient measures how precisely the model estimates the coefficient's unknown value.

154. Formula for standard error

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

SE = standard error of the sample

σ = sample standard deviation

n = number of samples

155. How does SVM works?

Support Vector Machines (SVM) are a type of supervised learning algorithm used for classification and regression. They work by identifying the best hyperplane that separates data points into classes with the maximum margin, which is the maximum distance between the data points of both classes. The closest points to the hyperplane, called support vectors, define the margin. SVMs can handle both linear and non-linear data by using different kernel functions, such as linear, polynomial, and radial basis function (RBF), to project non-linearly separable data into a higher-dimensional space where it becomes linearly separable. The optimization problem of finding the best hyperplane is solved using techniques like Sequential Minimal Optimization (SMO). For multi-class classification, strategies like one-vs-one or one-vs-all are used.

156. Difference between gini index, cross entropy, information gain and gain ratio

Gini Index and Entropy are measures of impurity used to determine how a single attribute can split the data into the most homogeneous subsets.
Information Gain uses Entropy to measure how effectively an attribute separates the classes.
Gain Ratio adjusts Information Gain by normalizing it with the intrinsic information of the split, aiming to reduce bias towards attributes with many levels.

157. Role of c regularizartion parameter in SVM

In Support Vector Machines (SVM), the regularization parameter C
C controls the trade-off between achieving a low error on the training data and maintaining a simple model to avoid overfitting. A high C value tries to minimize the training error, potentially leading to a complex model that fits noise and outliers (overfitting). Conversely, a low C value allows more misclassifications, favoring a simpler model that may generalize better to new data. The optimal C is usually chosen via cross-validation to balance model complexity and training data accuracy.

158. how to decide on a feature to split on in a decision tree for classification and regression?

Deciding on a feature to split on in a decision tree, whether for classification or regression tasks, involves assessing how effectively each feature divides the training data into subsets that are more homogeneous or have less variability with respect to the target variable. The methodology for this decision is similar across both tasks but uses different metrics to assess the quality of a split. Here's a detailed approach for both:

For Classification Trees
Metrics:

Gini Impurity: A measure of the frequency at which any element of the dataset will be mislabeled when it is randomly labeled according to the distribution of labels in the subset. The best split is the one that minimizes the impurity.
Entropy and Information Gain: Entropy is a measure of the randomness or uncertainty. Information Gain is the reduction in entropy caused by partitioning the data according to a particular feature. A higher information gain indicates a more effective feature at reducing uncertainty.
Calculation Steps:
Evaluate each feature:
For categorical features: Test each category or combinations of categories.
For continuous features: Try splitting at various points, typically at the midpoint between unique values sorted in order.
For each potential split, calculate the impurity or entropy for the two resulting child nodes.
Compute the Information Gain or the decrease in Gini impurity for the split.
Choose the feature and threshold that provide the maximum Information Gain or the greatest reduction in Gini impurity.

For Regression Trees
Metric:
Variance Reduction: This is similar to the concept of Information Gain but involves reduction in variance as the criterion. The idea is to choose a split that results in subsets having the lowest possible variance in their target values, which indicates more homogeneity.
Calculation Steps:
Evaluate each feature:
For continuous features: Consider splits at various points.
For categorical features: Consider each category or combinations of categories.
Calculate the total variance in the target variable for each subset that would result from a split at each possible point.
Determine the reduction in variance from the parent node to the weighted average variance of the two child nodes.
Select the feature and point that offer the greatest reduction in variance.
Common Considerations for Both
Overfitting Prevention: Decision trees can easily overfit the data:
Limit the tree depth.
Set a minimum number of samples per node.
Prune the tree after building it to remove splits that have little impact on performance.
Feature Types:
Continuous features are split by choosing a cutoff value.
Categorical features can be split by grouping categories.
Stopping Criteria:
Maximum depth of the tree.
Minimum gain in impurity reduction.
Minimum number of samples in a node or leaf.

In summary, for both classification and regression, the decision on which feature to split involves calculating a metric (like Gini impurity, entropy, or variance) that assesses how well a feature organizes the data into groups that are as pure as possible regarding the output variable. By repeatedly applying these criteria at each node, decision trees aim to construct a model that captures the underlying patterns in the data effectively.

159. weak learners and strong learners in decision tree

The relationship between weak and strong learners is most prominently utilized in ensemble methods, where multiple weak learners are combined to form a strong learner. Techniques like boosting and bagging use this strategy:

Boosting: This is an ensemble technique that aims to create a strong classifier from a number of weak classifiers. It works by sequentially applying weak classifiers, adjusting the weights of incorrectly classified instances so that subsequent classifiers focus more on difficult cases. AdaBoost is a popular boosting algorithm.

Bagging: Bagging (Bootstrap Aggregating) involves training multiple weak learners on different subsets of the original training set, then averaging their predictions to form a final verdict. Random forests, which consist of many decision trees (each tree being a weak learner), use bagging.

160. What is the Connection between p value, significance level and confidence interval.

Connection Between P-value, Significance Level, and Confidence Interval

1. P-value:
    - The probability of observing the data (or something more extreme) under the null hypothesis.
    - Small p-values suggest rejecting the null hypothesis.
2. Significance Level (α):
    - The threshold to reject the null hypothesis, often set at 0.05.
    - If $p<\alpha$, reject H0.
3. Confidence Interval (CI):
    - A range of plausible values for a parameter.
    - A 95% CI corresponds to $\alpha=0.05$

Example:

A drug's effect on blood pressure shows:

- Sample mean: 2, SE: 1, $\alpha=0.05$
- P-value: $p=0.029$ (reject H0).
- 95% CI: $(-0.064,4.064)$ (does not include 0, aligns with rejecting H0).

Both approaches show the drug likely affects blood pressure.

161. What is SHAP

SHAP (SHapley Additive exPlanations) is an explainability tool used to interpret machine learning model predictions. It assigns each feature a **Shapley value**, representing its contribution to a specific prediction. SHAP is based on cooperative game theory and provides consistent, fair, and additive explanations.

Key Features of SHAP:

1. **Feature Contribution**:
   - SHAP explains the impact of each feature on a model's prediction, indicating whether it increases or decreases the predicted value.
2. **Model-Agnostic**:
   - It can explain any machine learning model, including tree-based models (e.g., LightGBM, XGBoost) and deep learning models.
3. **Global and Local Explanations**:
   - **Local Explanation**: Explains individual predictions by showing feature contributions.
   - **Global Explanation**: Summarizes overall feature importance across the entire dataset.
4. **Fair and Consistent**:
   - Ensures that the sum of all feature contributions equals the model's predicted output.

Why SHAP is Useful in Fraud Models:

- **Transparency**: Provides insights into why a transaction is flagged as fraudulent.
- **Trust**: Builds stakeholder confidence in the model by explaining its behavior.
- **Feature Refinement**: Helps identify unimportant or redundant features to improve model performance.

Example:

If a fraud model predicts a transaction as fraudulent, SHAP can show which features (e.g., unusual transaction amount or location) contributed most to the decision and by how much.

162. What is maximum likelihood estimation in logistic regression

Maximum likelihood estimation in logistic regression finds the parameters (coefficients) that maximize the probability of observing the given data. logistic regression predicts probabilities using the sigmoid function. The likelihood function combines the probabilities for all observations. Log likelihood is the logarithm of the likelihood used for optimization. Maximum

likelihood estimation maximizes the log likelihood to estimate the parameters ensuring the model fits the observed data.

163. What is standard error in stats

The **standard error (SE)** is a measure of the variability or precision of a sample statistic, such as the mean or regression coefficient. It quantifies how much a sample statistic is expected to fluctuate from one sample to another if the sampling process is repeated.

Example:

Suppose you take a sample of 100 people to measure their average height. The sample has:

- Mean height = 170 cm.
- Standard deviation = 10 cm.

The SE of the mean is:

$$SE = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1\,\text{cm}.$$

This indicates the sample mean is expected to vary by about 1 cm across repeated samples of size 100.

In Practice:

- Smaller SE indicates more precise estimates of the population parameter.
- Larger sample sizes reduce SE, improving the reliability of statistical inferences.

164. What is MSE? Is it sum of bias and variance?

Not exactly! The **Mean Squared Error (MSE)** is not simply the sum of bias and variance, but it is related to both in a mathematical way.

MSE Breakdown:

Prepared by Redwan Walid (rwalid1@umbc.edu)

The **MSE** of a model can be expressed as:

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Here's what each term means:

1. **Bias²**: This measures the error due to incorrect assumptions in the model.
    - A high bias means the model is underfitting and failing to capture the true relationship.
2. **Variance**: This measures how much the model's predictions change when trained on different datasets.
    - A high variance means the model is overfitting and too sensitive to small fluctuations in the data.
3. **Irreducible Error**: This is the noise in the data that cannot be eliminated, no matter how good the model is.
    - It's the inherent randomness in the data, like measurement errors or unpredictable influences.

165. What is the difference between standard deviation and standard error

1. Standard Deviation (SD):

- **Definition:** Measures the amount of variation or dispersion of individual data points in a dataset from the mean.
- **Purpose:** Reflects the variability of the entire dataset.

- **Formula:**

$$SD = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

Where:

- x_i: Each individual data point
- $\bar{x}$: Mean of the dataset
- N: Number of data points

- **Use Case:** To understand the spread of data in a population or sample.
- **Interpretation:** A larger SD indicates data points are more spread out from the mean, while a smaller SD means they are closer to the mean.

---

2. Standard Error (SE):

- **Definition:** Measures the variability of the sample mean from the true population mean.
- **Purpose:** Indicates the precision of the sample mean as an estimate of the population mean.

- **Formula:**

$$SE = \frac{SD}{\sqrt{n}}$$

Where:

- SD: Standard deviation of the sample
- n: Sample size

- **Use Case:** Used in inferential statistics to construct confidence intervals or conduct hypothesis tests.
- **Interpretation:** A smaller SE indicates that the sample mean is a more precise estimate of the population mean. SE decreases as the sample size increases.

166. What is epoch in deep learning?

Epoch:

- Refers to one complete pass through the entire training dataset.
- In one epoch, every training sample is seen by the model exactly once (assuming no data augmentation or duplication).

Iteration:

- Refers to one pass through a **single batch** of data during training.
- It is a single update of the model's weights using the samples in the batch.

Summary of Relationship

- Epochs: Measure how many times the entire dataset is processed.
- Iterations: Measure how many batches are processed.

167. What are activation functions and why are they needed?

Activation functions are mathematical functions used in neural networks to introduce non-linearity into the model. They are applied to the output of neurons in a layer to determine the final output passed to the next layer.

Why Are Activation Functions Needed?

- **Introduce Non-Linearity**:
  - Neural networks solve complex problems like image recognition, language processing, and more, which often involve non-linear relationships between inputs and outputs.
  - Without activation functions, the network would be essentially a linear mapping (matrix multiplication), no matter how many layers are added.
  - Non-linear activation functions allow the network to learn and approximate complex functions.
- **Enable Deep Learning**:
  - Non-linear activation functions allow stacking multiple layers of neurons. Each layer can learn different aspects or features of the data, enabling deep architectures to solve intricate problems.
- **Control Output Range**:
  - Activation functions can constrain neuron outputs to specific ranges (e.g., [0,1][0,1] or [−1,1][−1,1]), which can stabilize training and make the network's behavior more predictable.
- **Gradient-Based Optimization**:
  - Certain activation functions help mitigate vanishing or exploding gradients, ensuring gradients are meaningful during backpropagation for learning.

Types of Activation Functions

- **Linear Activation**:
  - Rarely used because it doesn't introduce non-linearity.
- **Sigmoid**:
  - Range: (0,1)
  - Suitable for binary classification, but prone to vanishing gradient problems.
- **Tanh (Hyperbolic Tangent)**:
  - Range: (−1,1)
  - Zero-centered output, but still suffers from vanishing gradients for large inputs.
- **ReLU (Rectified Linear Unit)**:

    - Non-linear, computationally efficient, and avoids vanishing gradients for positive inputs.
    - Can suffer from **dying neurons** if many outputs become zero.
- **Leaky ReLU**:
    - Addresses the "dying ReLU" problem by allowing small gradients for negative inputs.
- **Softmax**:
    - Outputs probabilities for multi-class classification problems.
- **Swish**:
    - Smooth and non-monotonic, showing better performance in some deep networks.
- **ELU (Exponential Linear Unit)**:
    - Similar to ReLU but smoothens transitions for negative inputs.

Choosing an Activation Function

- **ReLU**: Most common for hidden layers due to simplicity and efficiency.
- **Sigmoid/Softmax**: For output layers in binary/multi-class classification problems.
- **Tanh**: Sometimes preferred over Sigmoid for zero-centered output.
- **Advanced Functions (e.g., Swish, ELU)**: In specific scenarios for better performance.

168. what is object oriented programming. explain in a simple and easy way

Object-Oriented Programming (OOP) is a way of organizing and writing code that makes it easier to manage and reuse. It is based on the idea of **objects**, which represent real-world things like a car, a person, or a bank account.

✅ Makes code **organized and reusable**
✅ Helps in **reducing errors**
✅ Allows for **scalability** (can grow easily)

OOP is widely used in **game development, web applications, and software engineering**.

```
# Defining a class (blueprint)
class Car:
    def __init__(self, brand, color):
        self.brand = brand
        self.color = color
```

```
    def drive(self):
        return f"{self.brand} is driving."

# Creating objects (real-world instances)
car1 = Car("Toyota", "Red")
car2 = Car("BMW", "Blue")

print(car1.drive())  # Output: Toyota is driving.
print(car2.drive())  # Output: BMW is driving.
```

169. How does cross entropy work?

Cross-entropy is a loss function used in classification tasks, especially in logistic regression and neural networks. It measures how different the predicted probability distribution is from the actual (true) distribution.

Simple Explanation:

1. **Prediction vs. Reality:** It compares the predicted probabilities (output of a model) to the actual class labels (0 or 1 in binary classification).
2. **Penalty for Wrong Predictions:** If the predicted probability is far from the actual class, the loss is high. If the prediction is accurate, the loss is low.
3. **Formula (for binary classification:**

$$L = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$$

- y = actual class (0 or 1)
- $\hat{y}$ = predicted probability
- Logarithm ensures higher penalties for wrong confident predictions.

Example:

- If the actual class is **1** and the model predicts **0.9**, the loss is **small**.
- If the actual class is **1** but the model predicts **0.1**, the loss is **large**.

Cross-entropy encourages the model to give higher confidence to correct predictions.

170. What is the equation for softmax?

The **softmax** function is used in multi-class classification to convert raw model outputs (logits) into probabilities. The formula for softmax is:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$

Where:

- z_i is the i-th raw score (logit) from the model.
- N is the total number of classes.
- e^{z_i} exponentiates the logit to make it positive.
- The denominator ensures that the sum of all outputs equals **1**, making them valid probabilities.

How It Works:

- Larger values get higher probabilities.
- Values are normalized so that they sum to **1**.
- Used in the final layer of classification models to assign probabilities to different classes.

171. What is the difference between quantiles and percentiles

Percentiles are given as percent values, values such as 95%, 40%, or 27%. Quantiles are given as decimal values, values such as 0.95, 0.4, and 0.27. the 50th percentile is median.

172. difference between standard deviation and standard error

1. Standard Deviation (SD): It measures the spread (or variability) of data points in a sample or population. It tells how much individual data points deviate from the mean.

2. Standard Error (SE): It measures how accurately a sample mean estimates the true population mean. It is derived from the standard deviation and depends on the sample size.

Prepared by Redwan Walid (rwalid1@umbc.edu)

Simple Path to Understanding

Imagine you are measuring the weight of apples in a basket:

1. **Standard Deviation (SD) → Spread of Apples in One Basket**
    - You take all the apples in **one basket** and measure their weight.
    - Some apples are heavy, some are light. The standard deviation tells you how much the weights vary **within that basket**.
2. **Standard Error (SE) → Accuracy of the Average Weight from Multiple Baskets**
    - You now take multiple samples (baskets of apples) and calculate the average weight of apples in each basket.
    - The standard error tells you how much these sample averages **fluctuate** if you were to take many different samples.

Formula	$SD = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$	$SE = \frac{SD}{\sqrt{n}}$

Final Thought

- **Use SD** when you want to know the variation in **individual data points**.
- **Use SE** when you want to know how precise your **sample mean** is in estimating the true population mean.

Standard error is used to calculate the confidence interval

173. What are the different ways to increase the precision of a ML Model

- Adjust the classification threshold
    - Higher threshold = higher precision but lower recall
- Feature engineering
    - Add more relevant features or remove noisy ones
    - Better features = model has more context to avoid false positive
- Model complexity/regularization
    - Try complex models like random forest or XGBoost
- Resample the data
    - If the dataset is imbalanced, undersample the majority class

174. What are the differences between parametric and nonparametric model. Give examples.

Here's a breakdown of the key differences and examples:

1. Assumptions about Data Distribution:

- **Parametric:** Assumes data follows a specific distribution (e.g., normal, exponential) and uses a fixed number of parameters to describe it.
- **Nonparametric:** Makes minimal or no assumptions about the underlying data distribution.
-

2. Flexibility and Complexity:

Parametric:
Less flexible as they are bound by the assumed distribution, potentially limiting their ability to model complex relationships.

Nonparametric:
More flexible and can model complex and non-linear relationships because they don't assume a predetermined form.

3. Examples:

Parametric:

- **Linear Regression:** Assumes a linear relationship between variables and a normally distributed error term.
-
- **t-test:** Assumes data is normally distributed and used to compare means of two groups.
-
- **ANOVA:** Assumes data is normally distributed and used to compare means of multiple groups.
-
- **Normal Distribution:** A specific probability distribution with parameters like mean and standard deviation.
-
- **Exponential Distribution:** A specific probability distribution often used to model waiting times or durations.
-

Nonparametric:

- **Decision Trees:** A model that uses a tree-like structure to make predictions, without assuming a specific distribution.
- **K-Nearest Neighbors (k-NN):** A model that classifies data points based on the majority class of their nearest neighbors, without assuming a specific distribution.
- **Support Vector Machines (SVM):** A model that finds the optimal hyperplane to separate data points, without assuming a specific distribution.
- **Wilcoxon Test:** A non-parametric test used to compare two groups when data is not normally distributed.
- **Mann-Whitney U Test:** A non-parametric test used to compare two groups when data is not normally distributed.
- **Kaplan-Meier Survival Analysis:** A non-parametric method used to estimate survival probabilities.

4. Advantages and Disadvantages:

Parametric:

- **Advantages:** Can be computationally efficient and require less data.
- **Disadvantages:** May not be suitable for complex data or when assumptions are violated.

Nonparametric:

- **Advantages:** More flexible and can handle complex data and non-linear relationships.
- **Disadvantages:** Can be computationally intensive and require more data.

175.