



Neha Jain

Data Scientist @ Paypal, Ex-Microsoft

Follow me on LinkedIn:

<https://www.linkedin.com/in/neha-jain-279b80118/>

Data Science Interview Questions and Answers

Statistics and Probability

1. What is the Central Limit Theorem and why is it important?

- The Central Limit Theorem states that the sampling distribution of the mean of a large number of independent, identically distributed variables approaches a normal distribution, regardless of the original distribution. It is important because it allows for statistical inference using the normal distribution.

2. Explain the difference between Type I and Type II errors.

- Type I error: Rejecting a true null hypothesis (false positive). A Type I error happens when you get false positive results: you conclude that the drug intervention improved symptoms when it actually didn't. These improvements could have arisen from other random factors or measurement errors.
- Type II error: Failing to reject a false null hypothesis (false negative). A Type II error happens when you get false negative results: you conclude that the drug intervention didn't improve symptoms when it actually did. Your study may have missed key indicators of improvements or attributed any improvements to other factors instead.

Type I and Type II Error		
Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

3. What is the p-value in hypothesis testing?

- The p-value measures the probability of obtaining test results as extreme as the observed results under the null hypothesis. A lower p-value indicates stronger evidence against the null hypothesis. P-value shows how likely it is that your set

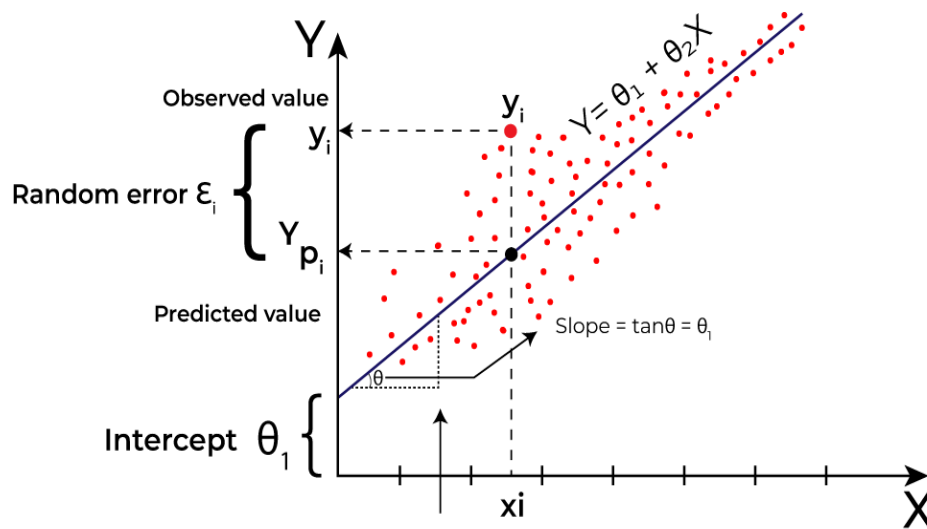
of observations could have occurred under the null hypothesis. P-Values are used in statistical hypothesis testing to determine whether to reject the null hypothesis.

4. How should a deployed model be maintained?

- To maintain a deployed model, the following steps should be taken:
 - **Monitor:** Continuously track the model's performance to ensure it is performing as expected. Regular monitoring helps assess the impact of any changes made to the system.
 - **Evaluate:** Assess the current model's performance using relevant evaluation metrics to determine if it requires improvement or if a new algorithm is needed.
 - **Compare:** Compare the performance of new models against the current one to identify the best-performing option.
 - **Rebuild:** Once the optimal model is identified, it should be rebuilt using the latest data to ensure it reflects current trends and information.

5. What is Linear Regression and what do the terms p-value, coefficient, and r-squared mean? What is the significance of each of these components?

- Linear regression is a statistical method used for predicting the relationship between a dependent variable and one or more independent variables. For instance, in predicting house prices, factors like size or location could be independent variables, and the price is the dependent variable. The model seeks to find the "line of best fit" between these variables, helping to determine if there is a positive or negative correlation.
- **P-value** measures the statistical significance of each predictor in the model, indicating whether the variable has a meaningful contribution.
- **Coefficient** represents the strength and direction of the relationship between an independent variable and the dependent variable.
- **R-squared** is a measure of how well the model explains the variance in the dependent variable, with a higher value indicating a better fit.



6. **What are the assumptions required for Linear Regression?**

- Linear regression relies on four key assumptions to provide valid results:
 - **Linearity:** There must be a linear relationship between the dependent variable and the independent variables, meaning the model should fit the data accurately.
 - **Normality of errors:** The residuals (errors between predicted and observed values) should be normally distributed and independent of each other.
 - **No multicollinearity:** The independent variables should not be highly correlated with each other to avoid redundancy in the model.
 - **Homoscedasticity:** The variance of errors should remain constant across all values of the predictor variable, ensuring that the spread of residuals is even throughout the range of data.

7. **Define and differentiate correlation and covariance.**

- Correlation measures the strength and direction of a linear relationship between two variables, ranging from -1 to 1.
- Covariance measures how two variables vary together, but it does not indicate the strength of the relationship.

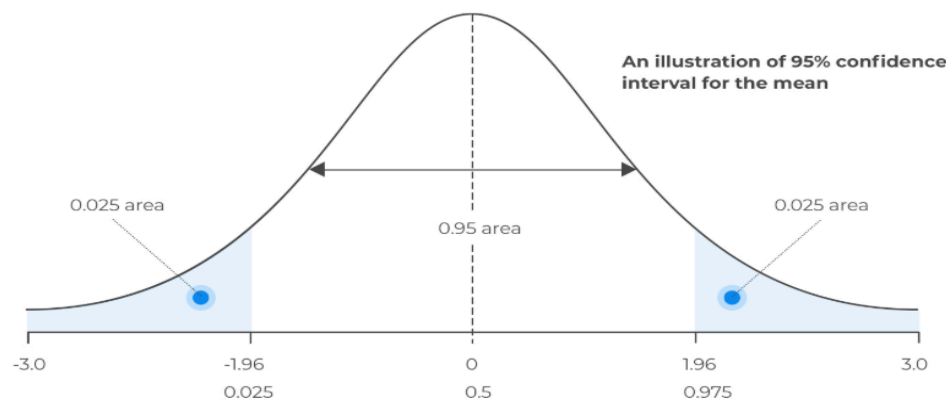
Covariance	Correlation
Indicates the direction of the linear relationship between variables	Indicates both the strength and direction of the linear relationship between two variables
Covariance values are not standard	Correlation values are standardized
Positive number being positive relationship and negative number being negative relationship	1 being strong positive correlation, -1 being strong negative correlation
Value between positive infinity to negative infinity	Value is strictly between -1 to 1

8. What is a confidence interval?

- A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by $1 - \alpha$, where α is the level of significance.

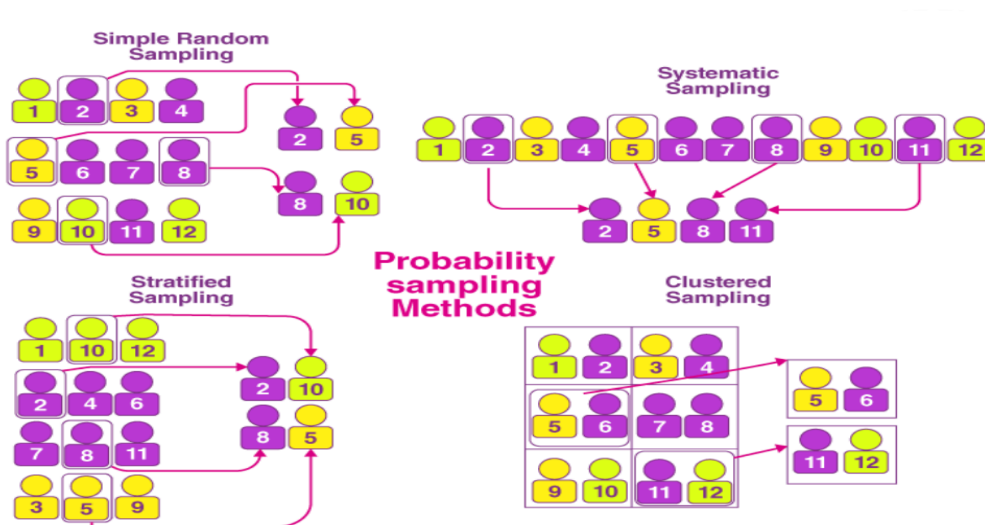


95% Interval



9. What are the different types of sampling methods?

- Random sampling, stratified sampling, systematic sampling, and cluster sampling.



- **Simple Random Sampling:** In simple random sampling technique, every item in the population has an equal and likely chance of being selected in the sample. Since the item selection entirely depends on the chance, this method is known as “Method of chance Selection”. As the sample size is large, and the item is chosen randomly, it is known as “Representative Sampling”.
- **Systematic Sampling:** In the systematic sampling method, the items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size.
- **Stratified Sampling:** In a stratified sampling method, the total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.
- **Clustered Sampling:** In the clustered sampling method, the cluster or group of people are formed from the population set. The group has similar significatory characteristics. Also, they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population.

10. What is a null hypothesis?

- A null hypothesis is a default assumption that there is no significant effect or relationship between variables. The null hypothesis, also known as “the conjecture,” is used in quantitative analysis to test theories about markets, investing strategies, and economies to decide if an idea is true or false. In other words, the null hypothesis is a hypothesis in which the sample observations results from the chance. It is said to be a statement in which the surveyors wants to examine the data. It is denoted by H_0 .

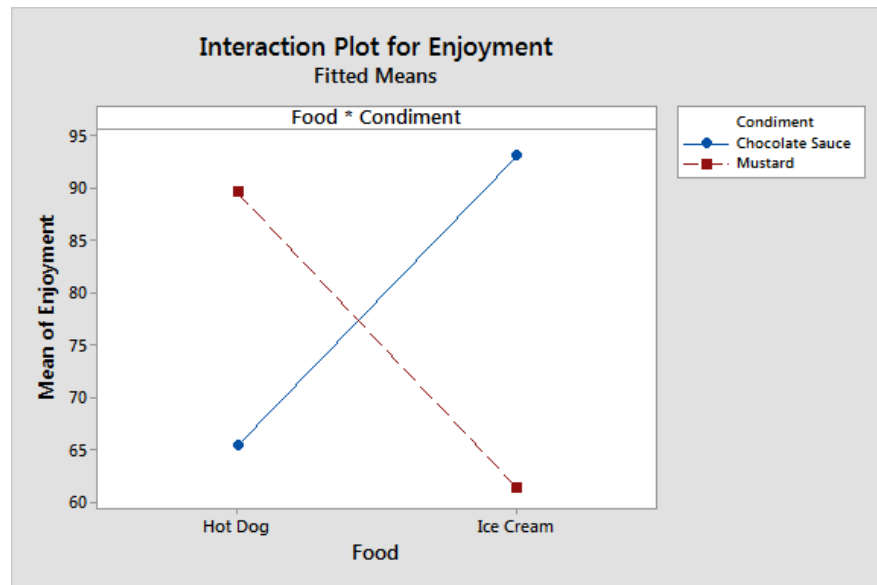
11. How do you Find the Null Hypothesis?

- The null hypothesis says there is no correlation between the measured event (the dependent variable) and the independent variable. We don't have to believe that

the null hypothesis is true to test it. On the contrast, you will possibly assume that there is a connection between a set of variables (dependent and independent).

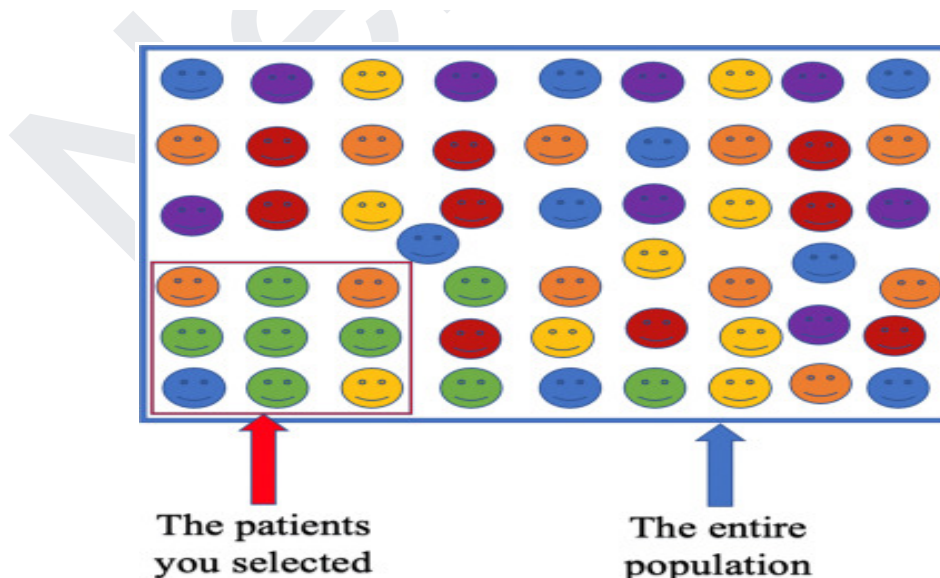
12. What is a statistical interaction?

- An interaction occurs when the effect of one factor (input variable) on the dependent variable (output variable) varies depending on the level of another factor. Essentially, the impact of one factor is not consistent across the different levels of another factor, indicating a combined effect that cannot be understood independently from each factor.



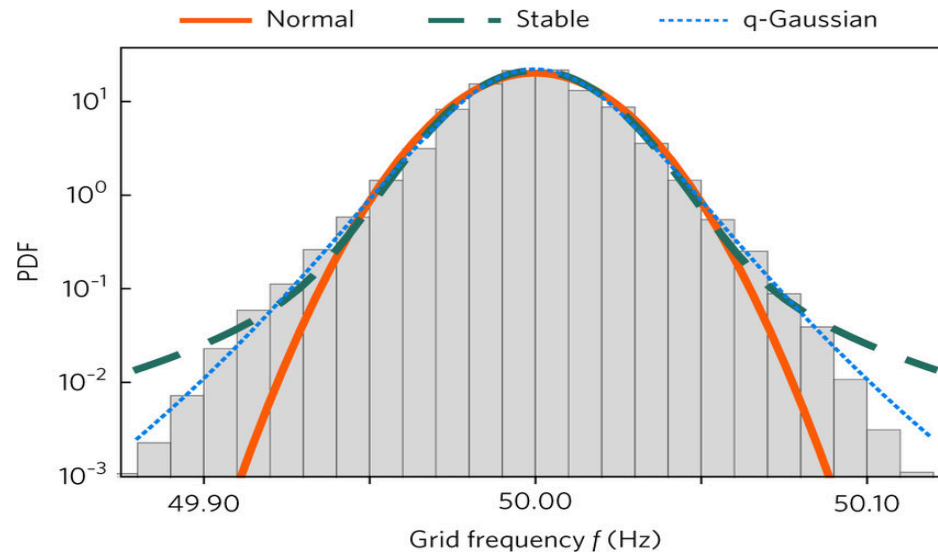
13. What is selection bias?

- Selection bias, or sampling bias, happens when the data selected for analysis does not accurately represent the larger population it is meant to reflect. This form of bias arises when certain subsets of the data are systematically excluded, meaning the sample is not randomly chosen, which can lead to skewed or inaccurate conclusions.



14. What is an example of a dataset with a non-Gaussian distribution?

- While the Gaussian distribution is part of the Exponential family, there are numerous other distributions that are also commonly used in statistical modeling. A dataset with a non-Gaussian distribution could follow a different pattern, such as a Poisson or exponential distribution, which can be more appropriate in certain scenarios, depending on the underlying data characteristics.



15. What is the Binomial Probability Formula?

- The binomial probability formula calculates the probability of a certain number of successes in a fixed number of independent trials, where each trial has the same probability of success. The binomial distribution is particularly useful in scenarios where the outcomes are binary (e.g., success or failure) and the trials are independent.

16. What is a z-score and how is it used?

- A z-score measures how many standard deviations a data point is from the mean. It is used in standardizing data and conducting hypothesis tests. It is a way to compare the results from a test to a “normal” population.

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score

x = observed value

μ = mean of the sample

σ = standard deviation of the sample

17. Explain the concept of statistical power.

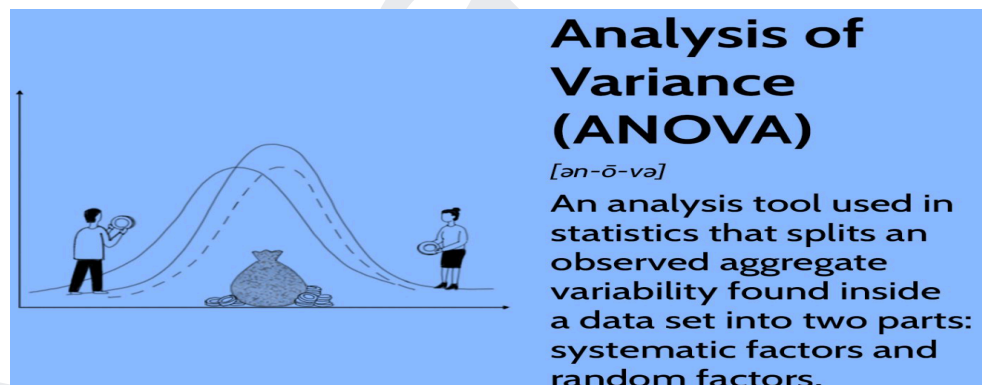
- Statistical power is the probability of correctly rejecting a false null hypothesis. Higher power reduces the likelihood of Type II errors.

18. Why does power matter in statistics?

- Having enough statistical power is necessary to draw accurate conclusions about a population using sample data.
- In hypothesis testing, you start with null and alternative hypotheses: a null hypothesis of no effect and an alternative hypothesis of a true effect (your actual research prediction).
- The goal is to collect enough data from a sample to statistically test whether you can reasonably reject the null hypothesis in favor of the alternative hypothesis.

19. What is ANOVA and when is it used?

- Analysis of Variance (ANOVA) tests whether there are significant differences between the means of three or more groups.
- You might use ANOVA when you want to test a particular hypothesis between groups, determining – in using one-way ANOVA – the relationship between an independent variable and one quantitative dependent variable.
- An example could be examining how the level of employee training impacts customer satisfaction ratings. Here the independent variable is the level of employee training; the quantitative dependent variable is customer satisfaction.



20. What is the role of Bayesian Optimization in hyperparameter tuning?

- Bayesian Optimization is a sequential search strategy used to optimize objective functions that are expensive to evaluate. It builds a probabilistic model of the function and selects hyperparameters by balancing exploration (trying new values) and exploitation (using known good values).

21. How does ANOVA work?

- ANOVA works by analysing the levels of variance within more than two groups through samples taken from each of them. In an ANOVA test you first examine

the variance within each group defined by the independent variable – this variance is calculated using the values of the dependent variable within each of these groups. Then, you compare the variance within each group to the overall variance of the group means.

- In general terms, a large difference in means combined with small variances within the groups signifies a greater difference between the groups. Here the independent variable significantly varies by dependent variable, and the null hypothesis is rejected.
- On the flip side, a small difference in means combined with large variances in the data suggests less variance between the groups. In this case, the independent variable does not significantly vary by the dependent variable, and the null hypothesis is accepted.

Data Analysis and Visualization

22. Explain the steps involved in data cleaning.

- Handling missing values
- Removing duplicates
- Correcting data types
- Handling outliers
- Normalizing data

23. What is data normalization? Why is it important?

- Data normalization rescales values into a range, usually 0 to 1, to ensure that features contribute equally to the analysis. Data normalization ensures that your data remains clean, consistent, and error-free by breaking it into smaller tables and linking them through relationships. This process reduces redundancy, improves data integrity, and optimizes database performance
- If there is no normalization in SQL, there will be many problems, such as:
 - Insert Anomaly: This happens when we cannot insert data into the table without another.
 - Update Anomaly: This is due to data inconsistency caused by data redundancy and data update.
 - Delete exception: Occurs when some attributes are lost due to the deletion of other attributes.

24. What is the difference between long and wide data formats?

- Long format: Each row is a single observation.
- Wide format: Each subject has a single row with multiple columns for measurements.

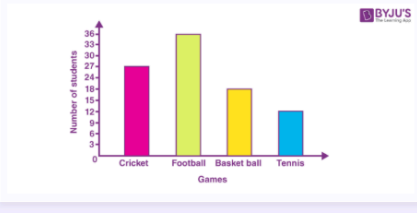
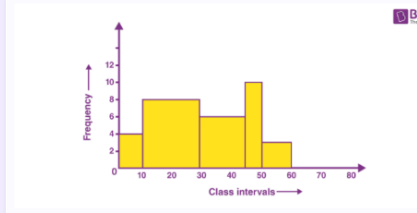
25. What are some common data visualization tools?

- Matplotlib, Seaborn, Tableau, Power BI, and Plotly.

26. Explain the difference between bar charts and histograms.

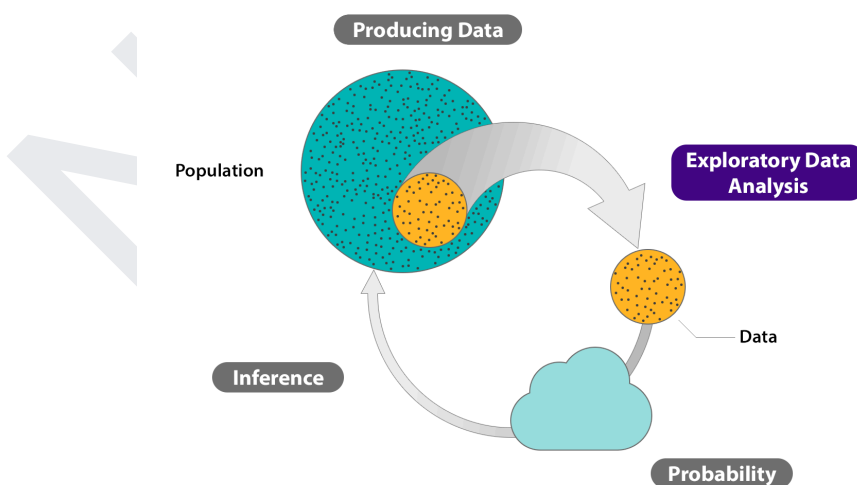
- Bar charts compare categorical data, while histograms show frequency distributions of continuous variables.

- The major difference between Bar Chart and Histogram is the bars of the bar chart are not just next to each other. In the histogram, the bars are adjacent to each other. In statistics, bar charts and histograms are important for expressing a huge or big number of data.

Bar graph	Histogram
The bar graph is the graphical representation of categorical data.	A histogram is the graphical representation of quantitative data.
There is equal space between each pair of consecutive bars.	There is no space between the consecutive bars.
The height of the bars shows the frequency, and the width of the bars are same.	The area of rectangular bars shows the frequency of the data and the width of the bars need not to be same.
	

27. What is exploratory data analysis (EDA)?

- EDA involves summarizing the main characteristics of data through visualizations and statistical measures to gain insights.
- Exploratory Data Analysis (EDA) is a statistical method for analyzing data sets to identify their main characteristics. It often uses data visualization techniques, such as charts and graphs, to help make the data easier to understand

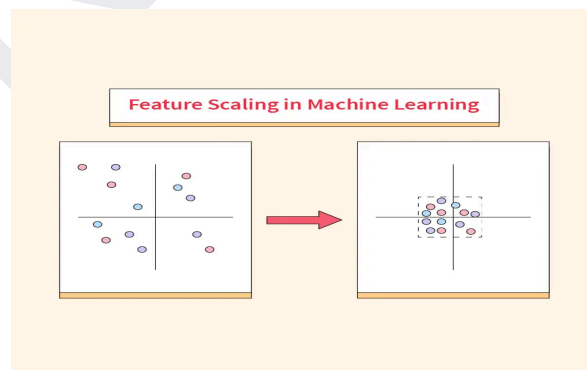


28. What are the key metrics used to evaluate model performance?

- Model performance is evaluated using various metrics depending on the type of task, such as classification, regression, or clustering. Some key metrics include:
- **Classification Metrics:**
 - **Accuracy:** The percentage of correctly predicted instances.
 - **Precision:** The ratio of true positives to all predicted positives, indicating model reliability.
 - **Recall (Sensitivity):** The ratio of true positives to actual positives, showing the model's ability to detect relevant instances.
 - **F1-Score:** The harmonic mean of precision and recall, balancing the two metrics.
 - **ROC-AUC:** Measures the model's ability to distinguish between classes.
- **Regression Metrics:**
 - **Mean Absolute Error (MAE):** The average absolute difference between predicted and actual values.
 - **Mean Squared Error (MSE):** The average squared difference, penalizing larger errors.
 - **Root Mean Squared Error (RMSE):** The square root of MSE, making the error metric interpretable.
 - **R-Squared (Coefficient of Determination):** Indicates how well the model explains the variance in the target variable.
- **Clustering Metrics:**
 - **Silhouette Score:** Measures how well instances are clustered based on cohesion and separation.
 - **Davies-Bouldin Index:** Evaluates cluster separation and compactness (lower is better).
 - **Adjusted Rand Index (ARI):** Compares the agreement between predicted and true labels in clustering tasks.

29. Explain the concept of feature scaling.

- Feature scaling standardizes the range of independent variables to ensure all features contribute equally.
- Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.



30. What is cross-validation and why is it important?

- Cross-validation splits data into training and validation sets to assess model performance and prevent overfitting.
- Cross-validation is a statistical technique used in machine learning to evaluate the performance of a model by repeatedly splitting the data into training and testing sets, training the model on different subsets of the data, and testing it on the remaining subsets, which helps assess how well the model generalizes to unseen data and prevents overfitting, making it a crucial tool for selecting the best model for a given problem

