**AI and DS QB1**

1. **Explain data analysis also different types of data analysis**
2. Data analysis is the process of systematically examining and interpreting raw data to extract meaningful insights and support informed decision-making, typically involving techniques like statistical modeling, visualization, and pattern recognition.
3. The main types of data analysis include descriptive, diagnostic, predictive, and prescriptive analysis.
4. Descriptive analysis: Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.
5. Diagnostic analysis: In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.
6. Predictive Analysis: Predictive analytics turn the data into valuable, actionable information. Predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring. Predictive analytics holds a variety of statistical techniques from modeling, machine learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event.
7. Prescriptive Analysis: Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.
8. **Difference between supervised & unsupervised learning**

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Supervised Learning** | **Unsupervised Learning** |
| Input Data | Uses labeled data (input features + corresponding outputs) | Uses unlabeled data (only input features, no outputs) |
| Goal | Predicts outcomes or classifies data based on known labels | Discovers hidden patterns, structures, or grouping in data |
| Computational Complexity | Less complex, as the model learns from labeled data with clear guidance. | More complex, as the model must find patterns without any guidance. |
| Types | Two types: Classification (for discrete outputs) or regression (for continuous outputs). | Clustering and association |
| Testing the model | Model can be tested and evaluated using labeled test data. | Cannot be tested in the traditional sense, as there are no labels |
| Type of dataset used | Uses training dataset | Uses just input dataset |
| Uses | Used for prediction | Used for Analysis |
| No. of classes | Known number of classes | Unknown number of classes |
| Data Classification | Data is classified based on training dataset | Uses properties of given data to classify it |
| Data analysis | Use off-line analysis of data | Use Real-time analysis of data |
| Graph |  |  |

1. **Explain in detail the following algorithm**
2. **Linear Regression:**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It provides valuable insights for prediction and data analysis.

The types of linear regression are:

1. Simple Linear Regression: Simple Linear regression is the simples form of linear regression and it involves only one independent variable and one dependent variable
2. Multiple Linear Regression: Multiple Linear Regression involves more than one independent variable and one dependent variable

Equation of linear regression is: Y = mX + b

Here,

Y = dependent variable

X = independent variable

m = slope of the line (how much there is a change in Y with the change in X)

b = intercept (value of Y when X =0)

The steps to calculate Linear Regression is:

* Step1: Data Collection
* Step2: Calculations
* Step3: Prediction
* Step4: Visualization

Example of Linear Regression: Consider the example of a pizza

1. **Logistics Regression:**

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors.

The different types of Logistic Regression are:

1. Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
3. Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

Example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It’s referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

1. **SVM**

* Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks.
* The algorithm maximizes the margin between the closest points of different classes.
* The different types of SVM’s are:
* Linear SVM: Linear SVMs use a linear decision boundary to separate the data points of different classes. When the data can be precisely linearly separated, linear SVMs are very suitable. This means that a single straight line (in 2D) or a hyperplane (in higher dimensions) can entirely divide the data points into their respective classes.
* Non-Linear SVM: Non-Linear SVM can be used to classify data when it cannot be separated into two classes by a straight line (in the case of 2D). By using kernel functions, nonlinear SVMs can handle nonlinearly separable data. The original input data is transformed by these kernel functions into a higher-dimensional feature space, where the data points can be linearly separated. A linear SVM is used to locate a nonlinear decision boundary in this modified space
* Advantages of SVM:

1. High-Dimensional Performance: SVM excels in high-dimensional spaces, making it suitable for image classification and gene expression analysis
2. Nonlinear Capability: Utilizing kernel functions like RBF and polynomial, SVM effectively handles nonlinear relationships.
3. Outliear Resilience: : The soft margin feature allows SVM to ignore outliers, enhancing robustness in spam detection
4. Binary and multiclass support: SVM is effective for both binary classification and multiclass classification,
5. Memory Efficiency: SVM focuses on support vectors, making it memory efficient compared to other algorithms.

* Disadvantages of SVM:

1. Slow training: SVM can be slow for large datasets, affecting performance
2. Noise Sensitivity: SVM struggles with noisy datasets and overlapping classes, limiting effectiveness in real-world scenarios.
3. **ID3 (Iterative Dichotomiser 3)**

* The ID3 algorithm is a popular decision tree method in machine learning.
* It builds a tree by selecting the best feature at each step based on information gain, aiming to create the most uniform groups. ID3 continues until a stopping condition, like a maximum tree depth, is met. More advanced versions, like C4.5 and CART, improve upon it.
* Its primary objective is to construct a tree that best explains the relationship between attributes in the data and their corresponding class labels.
* How does ID3 Work?

1. Selecting the best attribute: ID3 uses entropy and information gain to find the best attribute for splitting data. Entropy measures the randomness in the dataset, and the algorithm selects the attribute that reduces uncertainty the most, ensuring better data separation.
2. Creating Tree Nodes: ID3 selects an attribute and splits the dataset into subsets based on its values. It then recursively finds the best attribute for each subset, forming branches and nodes.
3. Stopping Criteria: The process stops when all instances in a branch belong to the same class or when no more attributes are available for splitting.
4. Handling Missing Values: ID3 can handle missing attribute values by using various strategies like attribute mean/mode etc.
5. Tree Pruning: Pruning is a technique to prevent overfitting. While not directly included in ID3, post-processing techniques or variations like C4.5 incorporate pruning to improve the tree's generalization.

* Advantages of ID3:

1. Interpretability
2. Handles Categorical Data
3. Computationally Inexpensive

* Limitations of ID3:

1. Overfitting: D3 tends to create complex trees that may overfit the training data
2. Sensitive to noise: Noise or outliers in the data can lead to the creation of non-optimal or incorrect splits.
3. Binary trees only: ID3 constructs binary trees, limiting its ability to represent more complex relationships present in the data directly.
4. **Decision Tree:**

* A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It works by recursively splitting the dataset based on the best attribute, forming a tree-like structure where each internal node represents a decision, and each leaf node represents a final outcome or class label.
* A decision tree consist of different types of nodes:
* Root Node: the first node in the tree, representing the entire dataset. It is split on the child node based on the best attribute
* Internal Node: Nodes that represent the decision made on the attribute values
* Leaf Nodes: Terminal nodes that provide the final output
* Working of Decision tree:
* STEP1: SELECTING THE BEST ATTRIBUTES:

~To build a decision tree, we need to determine the best attribute at each step. This is done using:

~ Entropy: Measures the randomness or impurity of the data. Lower the entropy means purer the data

~ Information Gain: Measures how much an attribute reduces entropy when used for splitting

* STEP2: CREATING TREE NODES AND BRANCHES

~ The dataset is divided into subsets based on the selected attribute’s values

~ Each subset forms a child node, and the process repeats recursively

* STEP3: STOPPING CONDITIONS

~ All instances in a subset belong to the same class (pure node).

~ No more attributes remain for splitting (majority class is assigned to the leaf node).

~ A predefined tree depth or minimum subset size is reached (to prevent overfitting).

* The different types of decision tree are:
* Information Gain- Used in ID3: This measures the reduction in entropy after splitting and also the attributes with highest IG is choosen
* Gini Index- Used in CART: Measures impurity by calculating the probability of misclassification and the attribute with the lowest Gini Index is chosen
* Gain Ratio – Used in C4.5: Improves Information Gain by considering attribute value distribution

1. **K-Means Clustering**

* K-means clustering is a technique used to organize data into groups based on their similarity.
* The algorithm works by first randomly picking some central points called centroids and each data point is then assigned to the closest centroid forming a cluster.
* After all the points are assigned to a cluster the centroids are updated by finding the average position of the points in each cluster.
* This process repeats until the centroids stop changing forming clusters.
* The goal of clustering is to divide the data points into clusters so that similar data points belong to same group.
* The algorithm for K-mean works as follows:

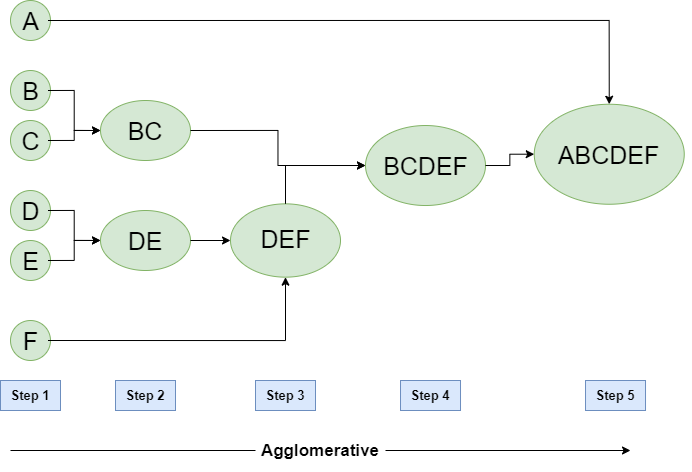
1. First, we randomly initialize k points, called means or cluster centroids
2. We categorize each item to its closest mean, and we update the mean’s coordinates, which are the averages of the items categorized in that cluster so far
3. We repeat the process for a given number of iterations and at the end, we have our clusters

* In conclusion, K-means clustering is a powerful unsupervised machine learning algorithm for grouping unlabeled datasets. Its objective is to divide data into clusters, making similar data points part of the same group. The algorithm initializes cluster centroids and iteratively assigns data points to the nearest centroid, updating centroids based on the mean of points in each cluster.

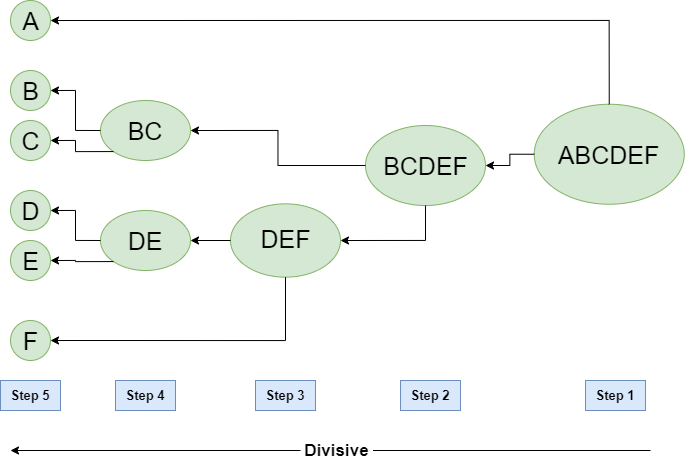
1. **Hierarchical Clustering**

* Hierarchical clustering is a technique used to group similar data points together based on their similarity creating a hierarchy or tree-like structure. The key idea is to begin with each data point as its own separate cluster and then progressively merge or split them based on their similarity.
* There are types of Hierarchical Clustering and they are:

1. Agglomerative Clustering: It is also known as the bottom-up approach or hierarchical agglomerative clustering (HAC).This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerate pairs of clusters until all clusters have been merged into a single cluster that contains all data.



1. Divisive Clustering: It is also known as a top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.



* Example for Hierarchical Clustering
* Imagine you have four fruits with different weights: an apple (100g), a banana (120g), a cherry (50g), and a grape (30g). Hierarchical clustering starts by treating each fruit as its own group.

1. It then merges the closest groups based on their weights.
2. First, the cherry and grape are grouped together because they are the lightest.
3. Next, the apple and banana are grouped together.

* Finally, all the fruits are merged into one large group, showing how hierarchical clustering progressively combines the most similar data points.

1. **What are the different univariate plots on EDA? Explain in detail**

* In Exploratory Data Analysis (EDA), common univariate plot used to visualize the distribution of a single variable include histogram, density plots, box plots, bar charts (for categorical data), stem-and-leaf plots, and frequency distribution tables, which help identify patterns, central tendency and outliers within the data
* Below are the most commonly used univariate plots, explained in detail:

1. Histogram: Shows the distribution of a continuous numerical variable by dividing it into bins (intervals) and counting the number of observations in each bin. Used to analyzing the age distribution of customers in a store

Key Features:

* X- axis represents the variable’s values
* Y-axis represents the frequency of occurrence
* Helps identifying skewness, modality (uni-modal, bi-modal, multi-modal) and outliners

1. Box Plot: Summarizes the distribution of a dataset using five-number summary: Minimum, First Quartile (Q1), Median (Q2), Third Quartile (Q3), and Maximum. Used in Understanding the spread of student exam scores in a class.

Key Features:

* The box represents the interquartile range (IQR = Q3 - Q1), where the middle 50% of data lies.
* The whiskers extend to the minimum and maximum values within 1.5 × IQR.
* Outliers are displayed as individual points beyond the whiskers.

1. Density Plot: Similar to a histogram, but provides a smooth estimate of the data’s distribution using a probability density function. Used in examining the salary distribution of employees in a company

Key Features:

* More refined than a histogram.
* Helps in identifying peaks and valleys in the distribution.
* Can be used to compare multiple distributions by overlaying different density plots.

1. Bar Plots: Represents categorical data using bars where the height indicates frequency. Used in displaying the number of products sold in different categories (electronics, furniture, clothing).

Key Features:

* Used for categorical variables.
* Can be vertical (column chart) or horizontal.
* Helps in understanding the most and least common categories.

1. Stem-and-Leaf Plots: Displays numerical data while preserving actual values, useful for small datasets. Used in analyzing student’s test scores in a class.

Key Features:

* The stem represents the leading digits.
* The leaves represent the trailing digits.
* Useful for quick visualization of distributions.

1. **What are the different issues in ML Algorithm?**

**-** Machine Learning (ML) algorithms can face several challenges that affect their performance, accuracy, and reliability. These issues can arise due to data quality, model selection, computational complexity, or ethical concerns. Below are some of the most common problems in ML:

1. Data-Related Issues

a) Insufficient Data

- ML models require a large amount of quality data to generalize well.

- Small datasets can lead to overfitting, where the model performs well on training data but poorly on unseen data.

Solution: Collect more data, use data augmentation, or apply transfer learning.

b) Imbalanced Data

- When one class has significantly more samples than another, the model becomes biased toward the majority class.

- Example: Fraud detection, where fraudulent transactions are rare compared to normal transactions.

Solution: Use oversampling, undersampling, SMOTE (Synthetic Minority Over-sampling Technique), or class-weighted loss functions.

c) Noisy and Incomplete Data

- Noisy data contains errors or outliers, which can mislead the model.

- Missing data can reduce the effectiveness of the model.

Solution: Use data cleaning techniques, imputation methods (mean, median, mode), or more robust ML algorithms.

2. Model-Related Issues

a) Overfitting

- The model learns patterns too well, including noise, making it perform well on training data but poorly on test data.

- Happens when the model is too complex.

Solution: Use regularization techniques (L1, L2), cross-validation, dropout in neural networks, and simpler models.

b) Underfitting

- The model is too simple and fails to capture underlying patterns in data.

- Leads to high bias and poor performance on both training and test data.

Solution: Use a more complex model, increase training time, or add more relevant features.

c) Hyperparameter Tuning Issues

- Choosing the right hyperparameters (learning rate, number of layers, regularization strength) is difficult.

- Poor tuning can lead to poor performance.

Solution: Use Grid Search, Random Search, or Bayesian Optimization.

3. Algorithm-Related Issues

a) Computational Complexity

- Some ML algorithms require high computational power, making them infeasible for large datasets.

- Example: Deep learning models require GPUs/TPUs for efficient training.

Solution: Use dimensionality reduction (PCA, t-SNE), optimized algorithms, and cloud computing.

b) Feature Selection and Engineering

- Too many irrelevant features can slow down training and reduce accuracy.

- Choosing the right features is critical for model performance.

Solution: Use feature selection techniques (Recursive Feature Elimination, Mutual Information) or create new meaningful features.

4. Ethical and Interpretability Issues

a) Bias and Fairness

- If training data is biased, the model will make biased predictions.

- Example: AI hiring tools may favor certain demographics based on historical hiring data.

Solution: Use fair ML techniques, diverse datasets, and debiasing methods.

b) Lack of Interpretability

- Complex models (like deep learning) act as black boxes, making it hard to understand why they make certain predictions.

- This is a problem in high-stakes applications like healthcare and finance.

Solution: Use explainability tools (SHAP, LIME) and prefer interpretable models where possible.

5. Deployment and Real-World Challenges

a) Concept Drift

- Data distribution changes over time, causing model performance to degrade.

- Example: Spam detection models need updates as new spam patterns emerge.

Solution: Regularly retrain and update models with new data.

b) Scalability

- A model trained on small datasets might not scale well when deployed in a real-world system.

Solution: Use distributed computing (Hadoop, Spark), cloud-based ML platforms (AWS, GCP, Azure).

c) Security and Adversarial Attacks

- ML models can be fooled by adversarial attacks, where attackers modify input slightly to trick the model.

- Example: Self-driving cars misidentifying stop signs due to adversarial perturbations.

Solution: Implement robust training techniques, adversarial training, and model security practices.

1. **Describe the architecture of ML Application with an example (using an application)**

* A Machine Learning (ML) application follows a structured architecture that includes data collection, preprocessing, model training, deployment, and monitoring.
* ML Application Architecture: Key Components
* Data Collection Layer: The first step is gathering data from various sources such as databases, APIs, sensors, or user inputs.

Example sources: CSV files, SQL databases, cloud storage, web scraping.

* Data Processing & Feature Engineering Layer: Cleansing and transforming raw data into a suitable format for training.

Its Common tasks are: Handling missing values, Encoding categorical data, Scaling numerical features, Removing outliers

* Model Training & Evaluation Layer: A machine learning algorithm is selected based on the problem type (e.g., classification, regression, clustering).

• The model is trained using the preprocessed data and optimized using techniques like hyperparameter tuning.

• Performance is evaluated using metrics like:

Accuracy, Precision, Recall (for classification)

RMSE, MAE (for regression)

* Model Deployment Layer: After successful training, the model is deployed for real-world use.

•Deployment options: Web applications (Flask, FastAPI, Django), Mobile applications, Cloud-based APIs (AWS, GCP, Azure)

* Prediction & Inference Layer: The model takes real-time input data and generates predictions. Predictions are sent back to users or applications.
* Monitoring & Continuous Learning: Ensures the model remains accurate over time.

Monitors data drift and concept drift (when patterns in data change).

Periodic model retraining is done using new data.

* Example:

Problem: Predict house prices based on location, size, and number of rooms

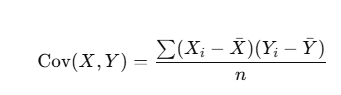
Steps:

1. Data Collection: Gather real estate data (price, location, area, rooms)
2. Preprocessing: Handle missing values, scale, numerical features
3. Model Training: use Liner regression or random forest to learn patterns
4. Deployment: Create an API to take inputs and return price prediction
5. Prediction: User inputs location=Mumbai, sqft=1500, bedrooms=3, model predicts ₹75,00,000.
6. Monitoring: Retrain the model periodically with new data.
7. **Distinguish Between Business Intelligence and Data Science**

|  |  |  |
| --- | --- | --- |
| Aspect | Business Intelligence | Data Science |
| Def | Analyzes past and present data for decision-making. | Uses ML and statistical techniques to predict future trends. |
| Purpose | Reporting, monitoring, and performance tracking. | Prediction, automation, and pattern discovery. |
| Data Type | Structured Data (SQL databases, spreadsheets) | Structured, semi-structured and unstructured data (text, images, videos) |
| Time Orientation | Historical and current data | Past, present and future data |
| Techniques Used | Data visualization, dashboards, SQL, OLAP | Machine Learning, AI, deep learning, statistical modelling |
| Tools & Technologies | Power BI, Tableau, Excel, Looker, SAP BI | Python, R, TensorFlow, PyTorch, Scikit-learn |
| Complexity | Easier, focuses on data summarization & visualization. | More complex, involves coding, modeling, and algorithm development. |
| User | Business analysts, executives, managers. | Data scientists, AI engineers, researchers |
| Example Use Cases | Analyzing sales trends and creating dashboards. | Predicting customer churn using machine learning. |
| End Goal | Better decision-making based on past data. | Building AI-driven solutions for automation and efficiency. |

1. **Explain the role played by correlation and covariance in EDA**

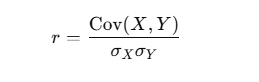
* Covariance:
* Definition:
  + Measures the direction of the relationship between two variables
  + A positive covariance means both variables increase together
  + A negative covariance means one increase while the other decreases
* Formula:



* Limitations:
* Covariance does not indicate the strength of the relationship
* Its value are not standardized, making comparisons difficult
* Example in EDA:

If the height and weight of people have a positive covariance, it suggests taller people tend to weigh more.

* Correlation:
* Definition:
* Measures both direction and strength of the relationship between two variables
* It is the scaled version of covariance and ranges from -1 to +1
* Types of Correlation:
* Positive Correlation (r > 0): One variable increases, the other also increases
* Negative Correlation (r<0): One variable increases, the other decreases
* No correlation ( r ≈ 0): No relationship between variables
* Formula:



* Advantages:
* Standardized between -1 to +1, making it easier to interpret
* Helps in feature selection by identifying redundant variables in ML models
* Example In EDA:

If hours studied and exam scores have r = 0.85, it means they have a strong positive correlation.

1. **Explain various stages in data analystic lifecycle**

* The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project.
* Phase 1: Discovery
* The data science team learns and investigates the problem
* Develop context and understanding
* Come to know about data sources needed and available for the project.
* The team formulates the initial hypothesis that can be later tested with data.
* Phase 2: Data Preparation:
* Steps to explore, preprocess, and condition data before modeling and analysis.
* It requires the presence of an analytic sandbox, the team executes, loads, and transforms, to get data into the sandbox.
* Data preparation tasks are likely to be performed multiple times and not in predefined order.
* Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.
* Phase 3: Model Planning:
* The team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
* In this phase, the data science team develops data sets for training, testing, and production purposes.
* Team builds and executes models based on the work done in the model planning phase.
* Several tools commonly used for this phase are – Matlab and STASTICA.
* Phase 4: Model Building:
* Team develops datasets for testing, training, and production purposes.
* Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
* Free or open-source tools – Rand PL/R, Octave, WEKA.
* Commercial tools – Matlab and STASTICA.
* Phase 5: Communication Results:
* After executing model team need to compare outcomes of modeling to criteria established for success and failure.
* Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
* Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.
* Phase 6: Operationalize:
* The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
* This approach enables team to learn about performance and related constraints of the model in production environment on small scale which make adjustments before full deployment.
* The team delivers final reports, briefings, codes.
* Free or open source tools – Octave, WEKA, SQL, MADlib

1. **Explain histogram? Can we perform univarient graphical analysis using histogram**

* A histogram is a graphical representation of numerical data that shows the frequency distribution of a single variable.
* It consists of bars, where the height of each bar represents the number of data points that fall within a specific range (bin). Unlike a bar chart, the bars in a histogram are continuous since they represent numerical intervals.
* Univariate analysis focuses on analyzing a single variable at a time.
* A histogram visualizes the distribution of that variable, helping to identify:
* Central tendency (mean, median, mode).
* Spread and variability (range, standard deviation).
* Skewness and symmetry (left-skewed, right-skewed, or normal).
* Outliers or gaps in data.
* Example of Univariate Analysis Using a Histogram:

Suppose we analyze the exam scores of 100 students.

Step1: Collect exam scores

Step2: Plot a histogram with:

X-axis: score ranges (bins)

Y-axis: Frequency (number of students)

Step3: Analyze the shape of the histogram:

If bell-shaped, scores are normally distributed.

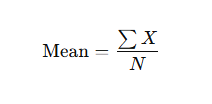
If right-skewed, most students scored lower

Id left-skewed, most students scored higher

1. **Explain various measures of central tendencies of statistical analysis using histogram**

* Central tendency refers to the statistical measures that summarize the center of a dataset. The three main measures of central tendency are: Mean, Median and Mode
* These measures help in understanding the distribution of data, which can be effectively visualized using a histogram
* Mean:

The mean is the sum of all values divided by the total number of values. It represents the center of the data.



Where:

X = Data Values

N = Number of values

*Interpretation in a Histogram:*

* The mean is affected by outliers and skewed distributions.
* In a normal distribution, the mean is located at the center.
* In a right-skewed histogram, the mean is greater than the median.
* In a left-skewed histogram, the mean is less than the median.

Example: In an exam score histogram, if the mean score is 75, most students scored around this value

* Median:

The median is the middle value when the data is arranged in ascending order. It divides the dataset into two equal halves.

*Interpretation in a Histogram:*

* The median is not affected by outliers and is useful for skewed data.
* In a normal distribution, the median is at the center, equal to the mean.
* In a right-skewed histogram, the median is less than the mean.
* In a left-skewed histogram, the median is greater than the mean.

Example: In an income distribution histogram, if the median salary is $50,000, half of the people earn below and half earn above this value.

* Mode:

The mode is the most frequently occurring value in a dataset

*Interpretation in a Histogram:*

* The mode is represented by the tallest bar in a histogram.
* A dataset can have:

Unimodal Distribution → One peak (single mode).

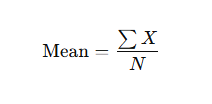
Bimodal Distribution → Two peaks (two modes).

Multimodal Distribution → Multiple peaks (more than two modes)

1. **Write with respect to quantitative data analysis:**
2. **Measures of central tendency:**

* Central tendency refers to statistical measures that identify the central or typical value in a dataset. The three key measures are Mean, Median, and Mode, and they help summarize large datasets effectively.
* Mean:

The mean is the sum of all values divided by the total number of values. It represents the center of the data.



Where:

X = Data Values

N = Number of values

*Interpretation in a Histogram:*

* The mean is affected by outliers and skewed distributions.
* In a normal distribution, the mean is located at the center.
* In a right-skewed histogram, the mean is greater than the median.
* In a left-skewed histogram, the mean is less than the median.

Example: In an exam score histogram, if the mean score is 75, most students scored around this value

* Median:

The median is the middle value when the data is arranged in ascending order. It divides the dataset into two equal halves.

*Interpretation in a Histogram:*

* The median is not affected by outliers and is useful for skewed data.
* In a normal distribution, the median is at the center, equal to the mean.
* In a right-skewed histogram, the median is less than the mean.
* In a left-skewed histogram, the median is greater than the mean.

Example: In an income distribution histogram, if the median salary is $50,000, half of the people earn below and half earn above this value.

* Mode:

The mode is the most frequently occurring value in a dataset

*Interpretation in a Histogram:*

* The mode is represented by the tallest bar in a histogram.
* A dataset can have:

Unimodal Distribution → One peak (single mode).

Bimodal Distribution → Two peaks (two modes).

Multimodal Distribution → Multiple peaks (more than two modes)

1. **Measures of spread**

* Measures of spread (dispersion) describe how much the data values vary from the central tendency (mean, median, or mode). These measures help understand the distribution, variability, and consistency of data.
* The main measures of spread are:

1. Range: The range is the difference between the maximum and minimum values in a dataset.

Formula: Range = Max Value – Min Value

Example: If the scores in a test are 45,50,60,75,90 then:

Range = 90-45 = 45

1. Interquartile Range (IQR): The IQR s the range of the middle 50% of data, found by subtracting the 1st quartile (Q1) from the 3rd quartile (Q3).

Formula: IQR = Q3 – Q1

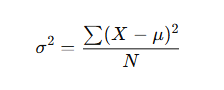
Example: If Q1 = 50 and Q3 = 80 then:

IQR = 80 – 50 = 30

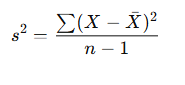
1. Variance: The variance measures the average squared deviation of each data point from the mean. It shows how data points spread around the mean.

Formula:

For a population:



For a sample:



Where:

X = Each data point

μ = Mean (for population), X = Sample mean

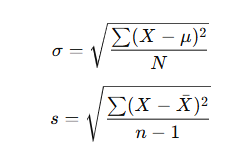
N, n = number of observations

Example:

If exam scores 50, 55, 60, 70, 80, variance is calculated as the average of squared differences from the mean

1. Standard Deviation: The standard deviation is the square root of variance. It measures the typical amount by which data points deviate from the mean.

Formula:



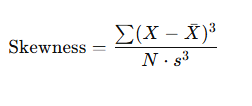
Example: If the exam scores have a standard deviation of 10, most student’s scores deviate by ±10 points from the mean

1. **Measures of skewness and kurtosis**

* Skewness: Skewness measures how asymmetric a dataset is compared to a normal distribution. It shows whether the data is skewed (shifted) left or right.
* Types of Skewness:

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Description | Skewness Value | Graph Shape |
| Symmetrical | Data is evenly distributed around the mean | 0 | Bell shaped |
| Positive Skew | More values are concentrated on the left, with a long right tail | >0 | Tail on right |
| Negative Skew | More values are concentrated on the right, with a long left tail | < 0 | Tail on left |

* Formula for Skewness:



Where,

X = Data Values

X = Mean

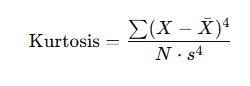
N = Number of values

s = Standard deviation

* Example: If exam scores have a right-skewed distribution, most students scored low, but a few high scores pull the mean higher than the median.
* Kurtosis: Kurtosis measures how heavy or light the tails of a distribution are compared to a normal distribution. It helps detect extreme values (outliers).
* Types of Kurtosis:

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Description | Kurtosis Value | Graph Shape |
| Mesokurtic | Moderate peak, normal tails. | ≈ 3 | Bell shaped |
| Leptokurtic | High peak, thick tails (many outliers). | >3 | Tall, thin peak |
| Platykurtic (Light-Tailed) | Low peak, thin tails (few outliers). | < 3 | Broad, flat peak |

* Formula for Kurtosis:



* Example: A leptokurtic distribution appears in financial markets, where extreme price fluctuations (outliers) are frequent.

1. **Explain the rules of convergence from predicate to CNF (All rules explain with example)**

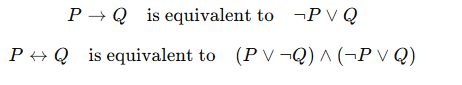
Conjunctive Normal Form (CNF) is a standard format for logical expressions used in automated theorem proving and logic programming. It is a conjunction (AND) of disjunctions (OR) of literals.

Steps to Convert a Predicate Logic Statement to CNF

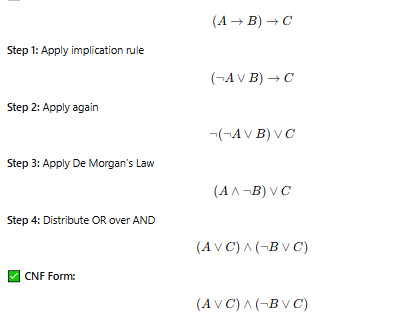
Rule1: Eliminate Implications (--->) and Bi-conditional (<--->)

Implication and bi-conditional operators are not allowed in CNF.

* Use the following transformations:

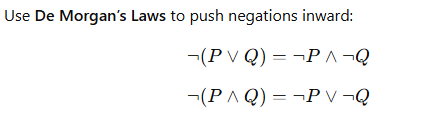


* Example:

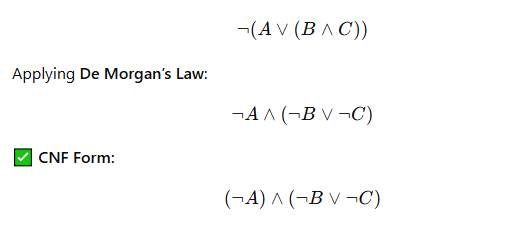


Rule2: Move NOT (¬) inward

Negation should only apply to individual predicates, not to entire expressions.



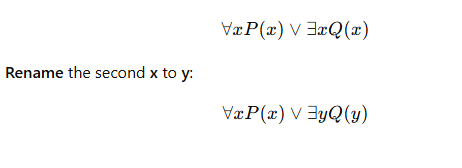
Example:



Rule 3: Rename Variables

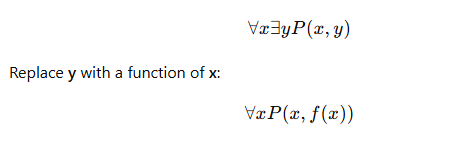
Each quantifier should have unique variable names to avoid confusion

Example:



Rule4: Remove Existential Quantifiers

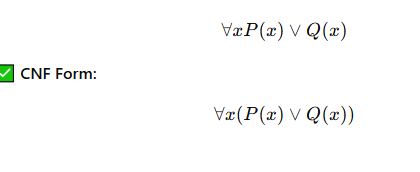
* Existential quantifiers (∃) must be eliminated.
* Replace ∃x with a Skolem function or constant.
* Example:



Rule5: Move Universal Quantifiers Left

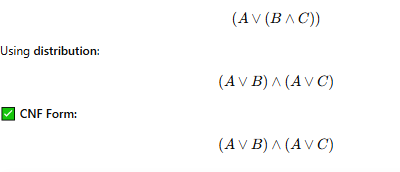
All universal quantifiers (∀) should be moved to the left of the expression.

Example:



Rule6: Convert to Conjunctive Form

Apply distributive law to get a conjunction of disjunctions



1. **Differentiate between uni-variate, bi-variate, multi-variate**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Univariate | Bivariate | Multivariate |
| No. of variables | 1 | 2 | 3 or more |
| Purpose | Study one variable at a time | Find relationship between two variables | Analyze how multiple variables interact |
| Example | Examining students test scores | Checking how temp affects ice cream sales | Studying how weather, traffic, and fuel price affect car sales |
| Statistical data | Z-test , t-test | Correlation test, regression test | ANOVA, factor analysis |
| ML use | Checking feature importance | Simple predictions (e.g., one input, one output) | Complex models like deep learning, AI-based forecasting |
| Dependency Consideration | No relationships considered | Examines how two variables affect each other | Studies interactions between many variables |
| Difficulty Level | Easy | Medium | Hard |
| Real-World Example | Checking the average salary of employees | Finding the link between exercise and weight loss | Predicting house prices based on location, size, and facilities |

1. **Explain confusion matrix w.r.t. ML. Also explain false positive and false negative and how are they significant?**

* A Confusion Matrix is a table used to evaluate the performance of a classification model by comparing actual values with predicted values. It helps in understanding the accuracy and types of errors the model makes.
* Structure of a Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Actual/ Predicted | Predicted Positive (1) | Predicted Negative (0) |
| Actual Positive | True Positive (Correct Prediction) | False Negative (FN) (Missed Positive) |
| Actual Negative (0) | False Positive (FP) (Wrongly Predicted as Positive) | True Negative (TN) (Correct Prediction) |

* A False Positive (Type I Error) occurs when the model incorrectly predicts a positive result for a negative case.
* The Significance of False Positive is:
* Can lead to unnecessary actions (e.g., unnecessary medical treatment).
* In fraud detection, wrongly blocking a genuine transaction can inconvenience users.
* Example:
* A spam filter marks an important email as spam.
* A COVID-19 test wrongly detects a healthy person as infected.
* A False Negative (Type II Error) occurs when the model incorrectly predicts a negative result for a positive case.
* The Significance of False Negative is:
* Can be more dangerous than false positives in critical applications like healthcare.
* In fraud detection, a fraudulent transaction being marked as genuine can lead to financial loss.
* Example:
* A cancer detection model fails to detect cancer in a patient.
* A security system does not flag a cyberattack.

1. **Explain data visualization and its importance in Data Analysis**

* Data Visualization is the graphical representation of data using charts, graphs, and maps. It helps in understanding complex datasets by presenting them in an easy-to-interpret visual format. Tools like matplotlib, seaborn, Tableau, and Power BI are commonly used for data visualization.
* Better Data Understanding – Helps identify patterns, trends, and correlations quickly.
* Simplifies Complex Data – Converts large datasets into easy-to-interpret visuals.
* Faster Decision-Making – Enables quick insights for business and research.
* Detecting Outliers & Anomalies – Identifies unusual patterns or errors.
* Enhances Communication – Makes reports and presentations more effective.
* Supports Predictive Analytics – Helps forecast future trends.
* Aids in Comparative Analysis – Allows comparison of different datasets.
* Interactive Exploration – Enables dynamic filtering and deeper insights.
* Industry-Wide Applications – Used in finance, healthcare, marketing, etc.
* Improves ML Interpretability – Helps in feature selection and model evaluation.