Scheme of Evaluation

Internal Assessment Test 3– October 2024

Sub:	Data Mining & Business Intelligence							Sub Code:	22MCA252
Date:	14-10-24	Duration:	90 mins	Max Marks:	50	Sem:	2	Branch:	MCA

Q.NO	Description	Marks Distribution	Max Marks
1	**Explain Non- Linear regression prediction methods with example.** • Explanation of Non-linear regression methods	10	10
2	**Explain CART Decision tree induction classification method with example.** • Explanation of CART decision method • Example	6 4	10
3	**Explain Logistic regression prediction methods with example.** • Explanation of Logistic regression methods	10	10
4	**Write short notes on data mining tools.** • Description of data mining tools	10	10
5	**Write a short notes on Back propagation algorithm** • Explanation of Back propagation algorithm	10	10
6	**Discuss about Multi layer Neural network in detail** • Explanation of Multi layer neural network	10	10
7	**Write short notes on classification in detail.** • Definition of classification • Types of classification	2 8	10
8	**Explain the State of the practice in analytics role of data scientists.** • Explain the type of role in data analytics project	10	10
9	**Explain the main phases of Data Analytics Life Cycle in detail.** • List out the phases of life cycle in detail notes	10	10

10	**Explain the Data mining for business Applications for the following** **a) CRM** **b) Click stream Mining** • Description of CRM and Click stream mining	**10**	**10**

Internal Assessment Test 3 – October 2024

Data Mining and Business Intelligence							Sub Code:	22MCA252
14/10/2024	Duration:	90 min's	Max Marks:	50	Sem:	I	Branch:	MCA

PART I

1. Explain Non- Linear regression prediction methods with example.

Non-linear regression is a predictive modeling technique used when data shows a non-linear relationship between the independent and dependent variables. Unlike linear regression, which fits a straight line, non-linear regression fits curves or complex shapes to the data. These methods are commonly applied when patterns in the data are curvilinear or when relationships between variables are complex and not simply additive.

Key Characteristics:

- **Non-linear relationships**: These methods model non-linear associations between predictors and outcomes, allowing for greater flexibility in capturing complex data patterns.
- **Non-linear functions**: Non-linear regression can involve polynomial, exponential, logarithmic, or other complex functions to best represent the data.

Common Non-Linear Regression Methods

1. **Polynomial Regression**
   - Polynomial regression fits a polynomial equation to the data, where the relationship is modeled as a polynomial of a given degree.
   - The model can be represented as: Y=b0+b1X+b2X2+...+bnXnY = b_0 + b_1X + b_2X^2 + ... + b_nX^nY=b0+b1X+b2X2+...+bnXn, where nnn is the degree of the polynomial.
   - **Example**: Modeling the trajectory of a thrown object, where the height follows a parabolic path. A second-degree polynomial (quadratic) can model this pattern effectively.
2. **Exponential Regression**
   - Exponential regression models data that grows or decays at an increasing rate.
   - The model is of the form Y=a·ebXY = a \cdot e^{bX}Y=a·ebX, where eee is the base of natural logarithms, aaa is a scaling factor, and bbb determines the rate of growth or decay.
   - **Example**: Population growth over time, where the population grows faster as it increases, can be modeled using an exponential function.
3. **Logarithmic Regression**
   - Used when the relationship between the variables grows quickly initially and then levels off.
   - The model form is Y=a+b·ln⁡(X)Y = a + b \cdot \ln(X)Y=a+b·ln(X).

    - **Example**: Diminishing returns in productivity, where initial increases in effort lead to high productivity gains but eventually level off.
4. **Power Regression**
    - Power regression models data where the dependent variable changes at a rate proportional to a power of the independent variable.
    - The model form is Y=a·XbY = a \cdot X^bY=a·Xb.
    - **Example**: The relationship between the size of an organism and its metabolic rate, where larger organisms tend to have higher metabolic rates, often follows a power law.
5. **Sigmoidal (Logistic) Regression**
    - Useful for situations where the dependent variable asymptotically approaches an upper or lower limit.
    - A common sigmoidal model is the logistic function: Y=L1+e−k(X−X0)Y = \frac{L}{1 + e^{-k(X - X_0)}}Y=1+e−k(X−X0)L, where LLL is the curve's maximum value, kkk controls the steepness, and X0X_0X0 is the midpoint.
    - **Example**: Modeling population growth that stabilizes over time due to limited resources. It starts slow, grows rapidly, and then levels off, fitting a sigmoid curve.

Example: Polynomial Regression in Predicting Sales Growth

Let's say a company wants to predict its sales based on advertising expenditure. If sales growth initially accelerates with more advertising but eventually slows down due to market saturation, a linear regression would be inadequate. A polynomial regression (e.g., quadratic or cubic) can capture this non-linear relationship by allowing the growth rate to vary with expenditure. For instance:

- **Quadratic Polynomial Model**:
  Sales=b0+b1·(Ad Spending)+b2·(Ad Spending)2\text{Sales} = b_0 + b_1 \cdot (\text{Ad Spending}) + b_2 \cdot (\text{Ad Spending})^2Sales=b0+b1·(Ad Spending)+b2 ·(Ad Spending)2.

This approach helps the company make better spending decisions based on a more realistic projection of sales growth dynamics.

Advantages and Disadvantages

- **Advantages**: More accurate for non-linear data, flexible, and provides a closer fit for complex relationships.
- **Disadvantages**: Harder to interpret, risk of overfitting with higher degrees, and requires more computational resources for complex functions.

2. Explain CART Decision tree induction classification method with example.

- A decision tree is a tree-like model that is used for making decisions. It consists of nodes that represent decision points, and branches that represent the outcomes of those decisions. The decision points are based on the values of the input variables, and the outcomes are the possible classifications or predictions.
  - A decision tree is constructed by recursively partitioning the input data into subsets based on the values of the input variables. Each partition corresponds to a node in the tree, and the partitions are chosen so as to minimize the impurity of the resulting subsets.

The algorithm works by recursively partitioning the training data into smaller subsets using binary splits. The tree starts at the root node, which contains all the training data, and recursively splits the data into smaller subsets until a stopping criterion is met.

- At each node of the tree, the algorithm selects a feature and a threshold that best separates the

training data into two groups, based on the values of that feature. This is done by choosing the feature and threshold that maximizes the information gain or the Gini impurity, which are measures of how well a split separates the data.

- The process continues recursively, with each node in the tree splitting the data into two smaller

subsets, until a stopping criterion is met. The stopping criterion could be a maximum depth for the tree, a minimum number of instances in each leaf node, or other criteria.

- Once the tree is built, it can be used to make predictions by traversing the tree from the root node to a leaf node that corresponds to the input data. For regression problems, the prediction is the average of the target values in the leaf node. For classification problems, the prediction is the majority class in the leaf node.

Calculate the Gini impurity for the entire dataset. This is the impurity of the root node.

- For each input variable, calculate the Gini impurity for all possible split points. The split point that results in the minimum Gini impurity is chosen.
- The data is split into two subsets based on the chosen split point, and a new node is created for each subset.
- Steps 2 and 3 are repeated for each new node, until a stopping criterion is met. This stopping criterion could be a maximum tree depth, a minimum number of data points in a leaf node, or a minimum reduction in impurity.
- The resulting tree is the decision tree.

To illustrate how the Gini impurity is calculated, consider the following example. Suppose we

have a binary classification problem, where we want to predict whether a person will buy a

particular product based on their age and income.

Age Income Buy Product?

30 20000 Yes
40 50000 Yes
20 30000 No
50 60000 No
60 80000 Yes

▸ We want to build a decision tree to predict whether a person will buy the product based on their age and income. Suppose we want to split the data based on age, with a threshold of 35.

The left child node will contain the data where age is less than or equal to 35, and the right child node will contain the data where age is greater than 35.

The Gini impurity for the left child node can be calculated as follows:

$G_{left} = 1 - (12)2 - (12)2 = 0.5$

There are two data points in the left child node, one of which buys the product, and one of which does not.

▸ Therefore, the probability of a randomly chosen element being labeled as "Yes" is 0.5, and the

probability of it being labeled as "No" is also 0.5.

▸ The Gini impurity for the right child node can be calculated as follows:

$G_{right} = 1 - (23)2 - (13)2 \approx 0.444$

There are three data points in the right child node, two of which buy the product, and one of which does not.

▸ Therefore, the probability of a randomly chosen element being labeled as "Yes" is 2/3, and the probability of it being labeled as "No" is 1/3.

The overall Gini impurity for the split can be calculated as a weighted average of the Gini impurities for the child nodes, where the weights are proportional to the number of data points

in each node.

$G_{split} = 25G_{left} + 35G_{right} \approx 0.48$

▸ The decision tree algorithm will try different thresholds and different features to find the split that minimizes the Gini impurity.

Let's consider an example of using the CART algorithm for a binary classification problem.

▸ Suppose we have a dataset of patients with information about their age, gender, blood pressure, and cholesterol level, and whether or not they have heart disease. We want to build a decision tree to predict whether a new patient will have heart disease based on their age,gender, blood pressure, and cholesterol level.

▸ To start, we calculate the Gini impurity of the entire dataset. Suppose there are 500 patients in the dataset, and 200 of them have heart disease, and 300 of them do not. The Gini impurity is:

$G = 1 - (200/500)2 - (300/500)2 = 0.48$

▸ Next, we consider each input variable and each possible split point. Suppose we choose age as the first split variable. We consider all possible split points and calculate the Gini impurity for each split. Suppose the split point that results in the minimum Gini impurity is 50 years.

We split the data into two subsets: patients who are 50 years old or younger, and patients who are older than 50. We create two new nodes for these subsets and calculate the Gini impurity for each node.

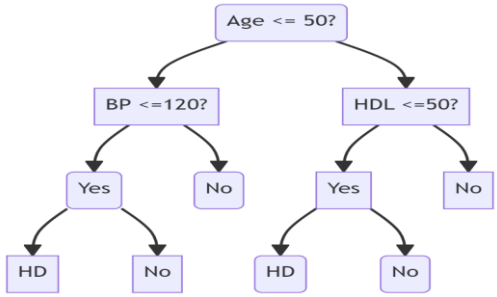
▸ Suppose the first node contains 300 patients, of which 100 have heart disease and 200 do not.

The Gini impurity of this node is:

$G1 = 1 - (100/300)2 - (200/300)2 = 0.44$

- Suppose the second node contains 200 patients, of which 100 have heart disease and 100 do not.
  The Gini impurity of this node is:
  $G2 = 1 - (100/200)2 - (100200)2 = 0.5$
- We choose the split that results in the minimum Gini impurity, which is the split on age at 50.
- We create two new nodes for the subsets and continue the process until we meet a stopping criterion.

- Suppose we set the stopping criterion to be a maximum tree depth of 2. The resulting decision tree would look like this:



- This decision tree can be used to predict whether a new patient will have heart disease based on their age, blood pressure, and cholesterol level.

Advantages
- It is a simple and intuitive algorithm that is easy to understand and interpret.
- It can handle both numerical and categorical data.
- It can handle missing values by imputing them with surrogate splits.
- It can handle multi-class classification problems by using an extension called the multi-class CART.

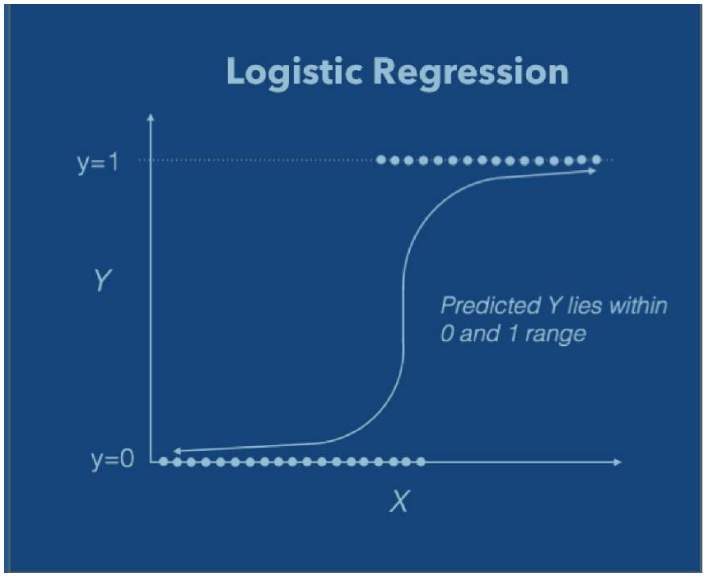
Disadvantages
- It tends to overfit the data, especially if the tree is allowed to grow too deep.
- It is a greedy algorithm that may not find the optimal tree.
- It may be biased towards predictors with many categories or high cardinality.
- It may produce unstable results if the data is sensitive to small changes or noise.

PART II

3. Explain Logistic regression prediction methods with example.

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.
For example,

➢ To predict whether an email is spam (1) or (0)
➢ whether the tumor is malignant (1) or not (0)



4. Write short notes on data mining tools.

Data mining tools help extract patterns, trends, and useful information from large datasets. They support various data mining tasks, such as classification, clustering, association, and regression. Below are some widely used data mining tools and their applications.

1. RapidMiner

- An open-source platform that integrates data preparation, machine learning, and predictive analytics.
- Offers a drag-and-drop interface with support for over 1500 functions and pre-built templates.
- Commonly used for data preprocessing, model building, and evaluation.

2. Weka (Waikato Environment for Knowledge Analysis)

- Open-source and ideal for educational and research purposes.
- Supports various data mining tasks like classification, clustering, regression, and association.
- Provides a graphical interface, making it easy for beginners, and integrates well with Java for custom solutions.

3. KNIME (Konstanz Information Miner)

- Open-source, focusing on data integration, processing, and analytics.
- Known for its modular workflow and drag-and-drop interface.
- Used extensively in pharmaceutical research, data science, and business intelligence.

4. Orange

- Open-source tool with an intuitive, visual programming interface.
- Best for beginners and suitable for exploratory data analysis and visualization.
- Used in various applications like bioinformatics and education, focusing on visual data mining.

5. Tableau

- Not strictly a data mining tool but widely used for data visualization and reporting.
- Helps in visually exploring data patterns and trends, often used after data mining for presenting findings.
- Popular in business intelligence and analytics for creating interactive dashboards.

6. SAS Enterprise Miner

- A commercial tool used in large organizations for advanced analytics and predictive modeling.
- Provides a comprehensive suite for data preparation, exploration, and modeling.
- Often used in sectors like finance, healthcare, and retail for robust data analysis.

7. IBM SPSS Modeler

- A commercial tool by IBM used for statistical analysis, data mining, and machine learning.
- Offers powerful predictive analytics capabilities with an easy-to-use interface.
- Commonly used in market research, healthcare, and social sciences for in-depth data analysis.

8. Python and R (with Libraries)

- Python libraries (e.g., Scikit-Learn, Pandas, TensorFlow) and R packages (e.g., Caret, Dplyr) are versatile and widely used for custom data mining solutions.
- Both are open-source and ideal for data manipulation, machine learning, and statistical analysis.
- Extensively used in research, data science, and industry applications for flexible, programmable data mining.
-

PART III

5. Write a short notes on Back propagation algorithm

- Backpropagation is a neural network learning algorithm.
- A neural network is a set of connected input/output units inwhich each connection has a weightassociated with it.
- During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples.
- Neural network learning is also referred to as connectionist learning due to the connections between units.
- Neural networks involve long training times and are therefore more suitable for applicationswhere this is feasible.
- Backpropagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value.
- The target value may be the known class label of the training tuple (for classification problems) or a continuous value (for prediction).
- For each training tuple, the weights are modified so as to minimize the mean squared errorbetween the network's prediction and the actual target value. These modifications are made in the ―backwards‖ direction, that is, from the output layer, through each hidden layer down to the first hidden layer hence the name is backpropagation.
- Although it is not guaranteed, in general the weights will eventually converge, and the learning process stops.

Advantages:

- It include their high tolerance of noisy data as well as their ability to classify patterns on which they have not been trained.
- They can be used when you may have little knowledge of the relationships between attributesand classes.
- They are well-suited for continuous-valued inputs and outputs, unlike most decision tree algorithms.
- They have been successful on a wide array of real-world data, including handwritten character recognition, pathology and laboratory medicine, and training a computer to pronounce English text.
- Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process.

Process: Initialize the weights:

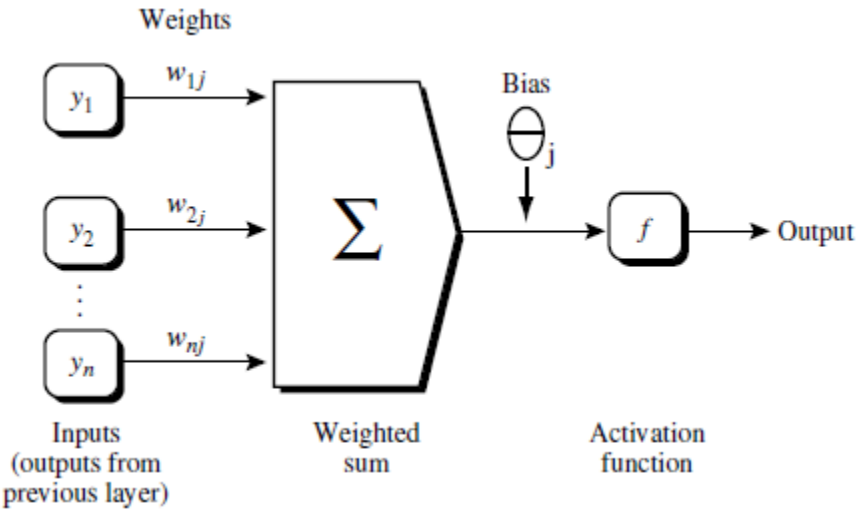
- The weights in the network are initialized to small random numbers ranging from-1.0 to 1.0, or -0.5 to 0.5. Each unit has a *bias* associated with it. The biases are similarly initialized to small random numbers.

- Each training tuple, ***X***, is processed by the following steps.

- **Propagate the inputs forward:** First, the training tuple is fed to the input layer of thenetwork. The inputs pass through the input units, unchanged. That is, for an input unit*j*, its output, *Oj*, is equal to its input value, *Ij*. Next, the net input and output of eachunit in the hidden and output layers are computed. The net input to a unit in the hiddenor output layers is computed as a linear combination of its inputs. Each such unit has anumber of inputs to it that are, in fact, the outputs of the units connected to it in theprevious layer. Each connection has a weight. To compute the net input to the unit, each input connected to the unit ismultiplied by its correspondingweight, and this is summed.

$$I_j = \sum_i w_{ij} O_i + \theta_j,$$

where$w_{i,j}$is the weight of the connection from unit iin the previous layer to unit j; $O_i$is the output of unit ifrom the previous layer $\theta_j$is the bias of the unit & it actsas a threshold in that it serves to vary the activity of the unit. Each unit in the hidden and output layers takes its net input and then applies an activation function to it.



Backpropagate the error:
The error is propagated backward by updating the weights and biases to reflect the error of the network's prediction. For a unit *j* in the output layer, the error $Err_j$is computed by

$$Err_j = O_j(1 - O_j)(T_j - O_j)$$

where$O_j$is the actual output of unit j, and $T_j$is the known target value of the giventraining tuple. The error of a hidden layerunit *j* is

$$Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk}$$

Where w_{jk} is the weight of the connection from unit j to a unit k in the next higher layer, and Err_k is the error of unit k. Weights are updatedby the following equations, where Dw_{ij} is the change in weight w_{ij}:

$$\Delta w_{ij} = (l) Err_j O_i$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

Biases are updated by the following equations below

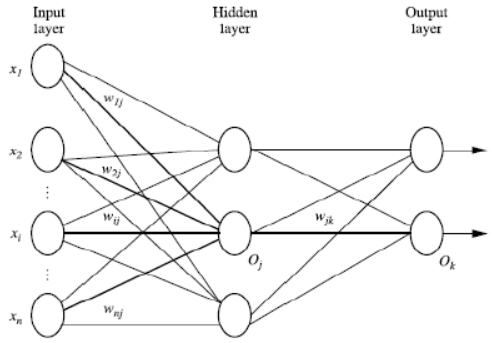
$$\Delta\theta_j = (l) Err_j$$

$$\theta_j = \theta_j + \Delta\theta_j$$

6. Discuss about Multi layer Neural network in detail

- The back propagation algorithm performs learning on a multilayer feed-forward neural network.
- It iteratively learns a set of weights for prediction of the class label of tuples.
- A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.

Example:



- The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer known as a hidden layer.

- The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary.
- The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples

PART IV

7. Write short notes on classification in detail.

Classification is a key data mining technique used for predictive modeling, where the goal is to assign items to predefined classes or categories. It is a type of supervised learning that relies on labeled data to train models for predicting outcomes on new, unlabeled data. Commonly used in spam detection, medical diagnosis, sentiment analysis, and fraud detection.

Key Steps in Classification:

1. **Data Preprocessing**: Prepare the data by cleaning, normalizing, and handling missing values to improve model performance.
2. **Model Training**: Use a labeled training dataset to train the model. The model learns patterns and relationships between input features and output classes.
3. **Model Evaluation**: Evaluate the model's accuracy and effectiveness on a test set using metrics like accuracy, precision, recall, and F1 score.
4. **Prediction**: Use the trained model to classify new data instances.

Common Classification Algorithms:

- **Decision Trees**: A tree-like structure where each node represents a decision on a feature, making it easy to interpret.
- **k-Nearest Neighbors (k-NN)**: Classifies an instance based on the majority class among its k closest neighbors.
- **Naive Bayes**: Based on Bayes' theorem, assuming predictor independence, often used in text classification.
- **Support Vector Machines (SVM)**: Finds an optimal hyperplane that best separates different classes.
- **Neural Networks**: Layers of neurons that capture complex, nonlinear relationships, suitable for large and complex datasets.

Performance Metrics:

- **Accuracy**: The percentage of correctly classified instances.
- **Precision**: Proportion of true positive results among predicted positives, important for identifying relevant results.
- **Recall**: Proportion of true positives correctly identified, useful in capturing all relevant cases.
- **F1 Score**: Harmonic mean of precision and recall, balancing both metrics.

Challenges:

- **Imbalanced Data**: When one class is more frequent, it can bias the model. Solutions include resampling or adjusting class weights.
- **Overfitting**: When the model performs well on training data but poorly on new data; mitigated by cross-validation.
- **Feature Selection**: Choosing relevant features improves model accuracy and reduces processing time.

Applications:

- **Healthcare**: Disease diagnosis and prediction.
- **Finance**: Detecting fraudulent transactions.
- **Email Filtering**: Spam vs. non-spam classification.
- **Sentiment Analysis**: Analyzing customer feedback for positive, neutral, or negative sentiment.

Classification is a foundational technique in data mining, providing predictive insights for a variety of real-world applications.

8. Explain the State of the practice in analytics role of data scientists.

Key Roles for a Data analytics project :

Business User :

- The business user is the one who understands the main area of the project and is also basically benefited from the results.
- This user gives advice and consult the team working on the project about the value of the results obtained and how the operations on the outputs are done.
- The business manager, line manager, or deep subject matter expert in the project mains fulfills this role.

Project Sponsor :

- The Project Sponsor is the one who is responsible to initiate the project. Project Sponsor provides the actual requirements for the project and presents the basic business issue.
- He generally provides the funds and measures the degree of value from the final output of the team working on the project.
- This person introduce the prime concern and brooms the desired output.

Project Manager :

- This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

Business Intelligence Analyst :

- Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs), key matrix, and business intelligence from a reporting point of view.
- This person generally creates fascia and reports and knows about the data feeds and sources.

Database Administrator (DBA) :

- DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.
- His responsibilities may include providing permission to key databases or tables and making sure that the appropriate security stages are in their correct places related to the data repositories or not.

Data Engineer :

- Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.
- The data engineer works jointly with the data scientist to help build data in correct ways for analysis.

Data Scientist :

- Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling, and applying correct analytical techniques for a given business issues.
- He ensures overall analytical objectives are met.
- Data scientists outline and apply analytical methods and proceed towards the data available for the concerned project.

PART V

9. Explain the main phases of Data Analytics Life Cycle in detail.

- **Data Analytics Lifecycle :**
  The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step–by–step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

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.

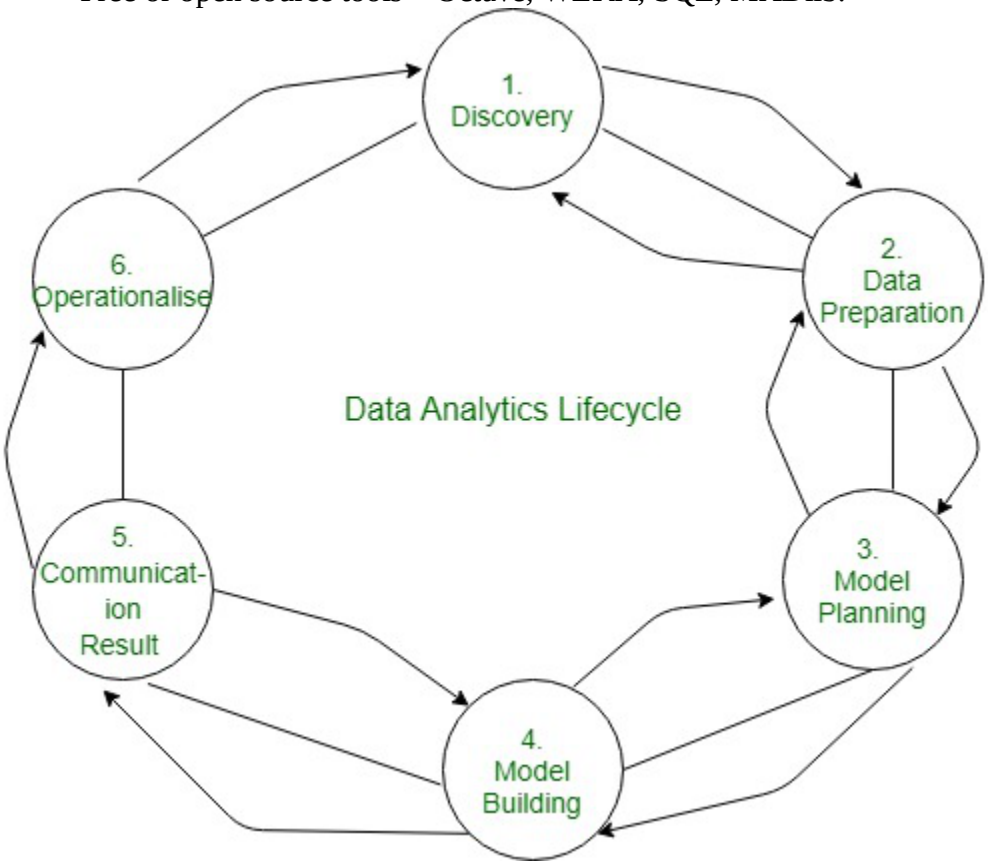
Phase 5: Communication Results –

- 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.



10. Explain the Data mining for business Applications for the following
a) CRM b) Click stream Mining

a) CRM

Customer Relationship Management(CRM) emerged in the last decade to reflect the central role of the customer for the strategic positioning of a company.

- It encompasses all measures for understanding the customers and for exploiting this knowledge to design and implement marketing activities, align production and coordinate the supply chain.
- CRM puts emphasis on the coordination of such measures, also implying the integration of customer-related data, meta-data and knowledge and the centralized planning and evaluation of measures to increase customer lifetime value.
- CRM gains in importance for companies that serve multiple groups of customers and exploit different interaction channels for them.
- CRM is a broadly used term, and covers a wide variety of functions.

These functions include:

- marketing automation (e.g. campaign management, cross and up-sell, customer segmentation, customer retention),
- sales force automation (e.g. contact management , lead generation , sales analytics, generation of quotes, product configuration) and
- contact centre management(e.g. call management , integration of multiple contact channels, problem escalation and resolution, metrics and monitoring , logging interactions and auditing), among others.

- Data mining helps marketing professionals improve their understanding of customer behavior.
- In turn, this better understanding allows them to target marketing campaigns more accurately and to align campaigns more closely with needs, wants and attitudes of customers and prospects.

b) Clickstream Mining

1. The approach which is used by most of the people for surfing information on Websites is difficult to analyze and understand.

2. Quantitative data can lack information about what a user actually intends to do, while qualitative data tends to be localized and is impractical to gather for large samples.

3. Once a website is made public, the user is in ultimate control of their own navigation, often employing a variety of different strategies for browsing.

4. These strategies also vary over time depending, not only on the user's goals, but also on factors such as expertise, familiarity with the site, time pressures and perceive cost of information.

5. Given this continually shifting nature of browsing strategies, the question arises how can these strategies be identified in the use made of an existing Website.

6. One solution is to use the clickstream logs, which contain the address of each page visited, the date and time of the visit and the referring page and are potentially rich source of data on Internet user activity.

7. Clickstream logs can be generate either by software hosted by the client application or directly from the server logs.