Data Pre-processing/Visualisation/Modelling/Analysis, Feature Engineering Quiz Bank

Quiz Exercise:

➔ Statistical Modelling is a modelling is a mathematical framework used to understand, explain and predict outcomes from complex data.

Importance in Data Analysis:
i) **Understand Relationship**: Identify and understand the relationships between different variables in a dataset.
ii) **Make predictions**: Use historical data to predict future outcomes or trends.
iii) **Drive Decisions**: Inform decision-making processes with empirical evidence derived from data analysis.

Applications of critical roles in Healthcare:

i) Predictive Healthcare
ii) Treatment Effectiveness
iii) Resource Optimization
iv) Policy Making

Mode: Easy
1) What is the primary purpose of data modelling?
   a) To design user interfaces
   b) To create data backups
   c) To represent the relationship between data elements.
   d) To perform data analysis.

   Answer: C)

   Explanation: Data Modelling illustrates the types of data stored within the system, the relationship among these data types, the ways the data can be grouped and organised with it format and attributes.

2) In data modelling, what is the term for the process of selecting relevant features or attributes for a specific task?

a) Data Aggregation
b) Data Cleaning
c) Feature Selection
d) Data Transformation

Answer C

Explanation:

Feature Engineering: Transforming raw data into features that are suitable for machine learning models. Therefore process of selecting, extracting and transforming the relevant features.

3) Data Model is a representation of data.

- False:

    Explanation:
    A data model is not representation of data but rather a conceptual or logical structure that defines how data is organised, stored and accessed.

    o It provides a framework for designing databases and assists in understanding the relationships between different data elements.

4) In data modelling, what is the term for the process of selecting a subset of data for model training and testing.

   a) Data Sampling
   b) Data Cleaning
   c) Feature Engineering
   d) Data Transformation

Answer: A

Explanation:
Data Sampling involves the procession of selection of subset of data from the dataset usually utilised in perseverance of important information incorporated in model training and testing.

5) What is the primary goal of regression analysis in data modelling?
   a) To group similar data points
   b) To predict numerical values.
   c) To improve model accuracy
   d) To calculate summary statistics

Answer: B

Explanation:
Regression analysis used to analyse the relationship between a dependent variable (target variable) and one or more independent variable. This resonates as in order to predict the

dependent variable or in tangent effect of independent variable in relation with the dependent variables.

6) What is the primary goal of time series model in data modelling?

   a) To predict numerical values
   b) To group similar data points
   c) To improve model accuracy
   d) To calculate summary statistics

Answer: A

Explanation:

Time series model underscores management of underlying causes/triggers of trends or systemic patterns over time {; identification nature of phenomenon represented by observations} and forecasting (predicting future values of time series variables).

7) In data modelling, what is the term for the process of transforming data into numerical format for machine learning models?

   a) Data Encoding
   b) Data Normalisation
   c) Data Imputation
   d) Data Transformation

Answer: A

Explanation:
Data Encoding Transforming categorical or textual data into numerical representation to enable computational analysis and modelling.

8) What is the primary purpose of data mart in data modelling?

   a) To visualise the data distribution
   b) To store and manage data for a specific task.
   c) To define data attributes and properties.
   d) To calculate summary statistics.

Answer : B

Explanation:

Data Marts is a data storage system built to provide quick and easy access to the data that is most relevant to the specific needs of department or business unit.

9) In data modelling, what is the term for the process of filling in missing data values by estimating them from available data?

a) Data Imputation
b) Data Cleaning
c) Data Transformation
d) Data Encoding

Answer: A

Explanation:

Imputation involves estimating or filling in missing values based on the available information. Various imputation methods can be utilised, depending on the nature of the data and underlying assumptions.

10) What is the primary goal of ensemble modelling technique in data modelling?

a) To group similar data points
b) To predict future data.
c) To improve model accuracy
d) To calculate summary statistics.

Answer: C

Explanation:

Ensemble methods are techniques that generated multiple models and then combine them to generate improved results. Ensemble methods in machine learning usually generate accurate solutions than a single models.

11) In data modelling, what is the term for the process of reducing the dimensionality of data while preserving its important features?

a) Data Imputation
b) Data Reduction Answer: B
c) Feature Engineering
d) Data Normalization

Answer: B : Data Reduction

Explanation:

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible.

12) What is the primary purpose of data pipeline in data modelling?
a) To visualise data relationships
b) To prepare and transform data for analysis.
c) To define data attributes and properties.
d) To calculate summary statistics.

Answer: B: To prepare and transform data for analysis.

Explanation:

A data pipeline is a series of processes that extract data from multiple sources, transform the data into a standardised format, and load it into a target system for analysis.

The primary purpose of a data pipeline in data modelling is to prepare and transform data into analysis.

13) In data modelling, what is the term for process of selecting and combining data from multiple sources for analysis?

a) Data Aggregation
b) Data Integration
c) Data Sampling
d) Data Transformation

Answer: B: Data Integration

Explanation:
Data integration refers to the process of aggregating data from multiple sources across an organisation to provide a complete, accurate and up to date dataset.

14) Which of the following is Data Modelling incorporated with?
a) Application
b) Designing
c) Data Requirement Analysis Answer: C
d) Rewriting

Explanation: Data modelling is a technique used in the field of data management to analyse and define the data requirements of a system or organization. It involves identifying the

entities, attributes, relationships and constraints within the data and creating a conceptual representation of these elements.

15) The process of creating a data model for an information system is termed?
a) Programming
b) Data Analysis
c) Database Management
d) Data Modelling

Answer D: Data Modelling

Explanation:
Data Modelling; Refers to the process of creating a representation of the data requirements for a specific information system. It involves identifying the entities, attributes, relationships and constraints of the system and organising them into structures format.

16) In machine learning what is 'feature engineering'?

a) The process of creating new features from existing data.
b) Engineering a new algorithm for feature selection
c) Designing engineering-based features for industrial data.
d) The process of selecting the best engineering tools for data analysis.

Answer A: The process of creating new features from existing data.

Explanation:

Feature Engineering is the process of generating new features or transforming existing features to improve the performance of a machine learning model. Perform feature engineering to i) improve model's predictive performance.
ii) reduce computational or data needs
iii) improve interpretability of results.

17) In data visualisation, what does a "Box-plot" typically show?

a) Geographic Data
b) Time series data
c) The distribution, median and range of data. Answer C
d) Correlation between two variables

Answer : C

Explanation:

A Box-plot showcases the distribution range of data highlighting the distinctive quartile , median denotation {Q,, Q2 ,Q3} respectively.

18) What is 'regularisation' in machine learning?

a) A method to speed up the training of a model
b) A technique to reduce overfitting by adding a penalty to the loss function Answer: B
c) A process to regularise the intervals at which data is fed into the model.
d) A technique to normalise the input data.

Answer: B

Explanation:

Regularisation is a set of methods for reducing overfitting in machine learning models. Regularisation encompasses a range of techniques to correct for overfitting in machine learning models.