P.S.R. ENGINEERING COLLEGE
(An Autonomous Institution, Affiliated to Anna University, Chennai)
Sevalpatti (P.O), Sivakasi– 626140

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
QUESTION BANK

Programme:	B.E	Department:	Computer Science and Engineering
Academic Year:	2024-2025	Year/Sem/Sec:	II/ III/ I & II
Course Code:	231CS33	Course Name:	PROGRAMMING FOR DATA SCIENCE
Course Tutor:	Dr.S.Singaravelan Prof/CSE, Dr.R.Arun ASP/CSE.		

UNIT-I INTRODUCTION

PART-A (1 Mark)

S.No	Questions	COs	BT
1.	Select the correct statement. a) Raw data is original source of data b) Preprocessed data is original source of data c) Raw data is the data obtained after processing steps d) None of the mentioned	CO1	K1
2.	Which of the following does Data Scientist perform? a) Define the question b) Create reproducible code c) Challenge results d) All of the mentioned	CO1	K1
3.	List the components of data science? a) Statistics b) Data expertise c) Data engineering d) Visualization	CO1	K1
4.	Which of the following approach should be used to ask Data Analysis question? a) Find only one solution for particular problem b) Find out the question which is to be answered c) Find out answer from dataset without asking question d) None of the mentioned	CO1	K1
5.	In the following which is NOT a part of the Data Science Life Cycle? a) Problem Identification b) Data Visualization c) Model Deployment d) System Architecture Design	CO1	K1
6.	What is it called when the data source is gathered and compiled with others? a) Primary Data b)Quantitative data c)Secondary data d)None of the above	CO1	K1
7.	What are the four steps of data preparation? a) Data cleaning>Data reduction>Data transformation>Data integration b)Data cleaning>Data reduction> Data integration>Data transformation c)Data reduction> Data cleaning>Data transformation>Data integration	CO1	K1

	d)Data cleaning> Data transformation> Data reduction>Data integration		
8.	Which of the following are the applications of data science? a)Risk detection b)Image recognition c)Speech recognition d)All of the above	CO1	K1
9.	Which type of data analysis gives a summary of the raw data set? a)Descriptive data analysis b)Diagnostic data analysis c)Predictive data analysis d)Prescriptive data analysis	CO1	K1
10.	What do you mean by the model planning phase in the life cycle of data analytics? a)This phase involves creating data sets for training for testing, production, and training purposes b)This phase involves the processing of big raw data c)This Phase involves the team which is responsible for evaluating the tools	CO1	K1
11.	Which of the following is one of the key data science skills? a) Data Visualization b) Machine Learning c) Statistics d) All of the mentioned	CO1	K1
12.	Which of the following is a good way of performing experiments in data science? a) Generalize to the problem b) Have Replication c) Measure variability d) All of the mentioned	CO1	K1
13.	Which of the following is the top most important thing in data science? a) data b) question c) answer d) none of the mentioned	CO1	K1
PART-B (2 Marks)			
1.	Define Data Science.	CO1	K1
2.	Infer Data Source Nomenclature.	CO1	K2
3.	List the different key components in digital universe.	CO1	K1
4.	What is a Data Scientist? What Does a Data Scientist Do?	CO1	K1
5.	Define Data repositories.	CO1	K1
6.	What is data?	CO1	K1
7.	What is machine - generated data?	CO1	K1
8.	List the stages of data science process.	CO1	K1
9.	What are the advantages of data repositories?	CO1	K1
10.	List the stages of data science process.	CO1	K1
PART C (14 Marks)			
11.	Recall the concept of Information commons.	CO1	K1
12.	Explain Life cycle of data science	CO1	K2

13.	Organize the various applications of data science.	CO1	K3
14.	Explain the Digital universe and sources of data.	CO1	K2
15.	Demonstrate Data Science Project Life Cycle.	CO1	K2
16.	Explain briefly about various data sources.	CO1	K2

UNIT-II DATA PREPROCESSING

PART-A (1 Mark)

S.No	Questions	COs	BT
1.	Processing data includes sub setting, formatting and merging only. a) True b) False	CO2	K1
2.	Interpret the technique used to handle missing data by filling in the missing values with the estimated data? a. Data Normalization b. Data Imputation c. Data Transformation d. Data Encoding	CO2	K1
3.	How to convert categorical data with two categories into numerical form? a) One-hot encoding b) Standardization c) Label encoding d) Smoothing	CO2	K1
4.	Interpret when .groupby() function return used with .size() a) The total sum of each group b) The average size of each group c) The number of elements in each group d) The unique values in each group	CO2	K1
5.	Data often contain _____ a)Target Class b)Uncertainty c)Methods d)Keywords	CO2	K1
6.	To remove noise and inconsistent data ____ is needed. **a**)Data Cleaning b)Data Transformation c)Data Reduction d)Data Integration	CO2	K1
7.	Which of the following is not a form of data transformation? a) Normalization b) Discretization c) Concept hierarchy d) Compression	CO2	K1
8.	What is the process of filtering and cleaning data called? a) Data Mining b) Data Wrangling c) Machine Learning d)All of the above	CO2	K1

9.	What is the purpose of the merge() function in Pandas? a) To combine two or more data frames based on a common key b) To sort a data frame based on one or more columns c)To calculate the mean of a column in a data frame d)To rename the columns of a data frame	CO2	K1
10.	In data preprocessing, what does the term “outlier” refer to? a) A feature with the highest correlation b) A duplicate record in the dataset c) An observation significantly different from others d) A column with missing values	CO2	K1
PART-B (2 Marks)			
1.	Compare Numerical filtering and Text filtering	CO2	K2
2.	Outline how to Rank the data.	CO2	K2
3.	Identify various methods of data ingestion.	CO2	K3
4.	Recall the concept of Text filtering.	CO2	K1
5.	What is data preprocessing, and why is it important in data science?	CO2	K1
6.	Name two techniques used for handling missing data in a dataset.	CO2	K1
7.	What is the purpose of feature scaling in data preprocessing?	CO2	K1
8.	What is the difference between normalization and standardization?	CO2	K1
9.	What are the common methods for handling imbalanced datasets?	CO2	K1
10.	How data can read from the various data sources.	CO2	K1
11.	How will a data can be ranked?	CO2	K1
PART C (14 Marks)			
12.	What are all the Key steps involved in data pre-processing.	CO2	K1
13.	Explain Manipulation of data briefly.	CO2	K2
14.	Illustrate the concept behind the Sorting of data.	CO2	K2
15.	Show the Life cycle of Data Preprocessing	CO2	K1
16.	Recall the ideas behind the following Grouping and Rearranging of data.	CO2	K1
17.	Recall the ideas behind the following Ranking Data.	CO2	K1
18.	Explain various methods involved in handling the missing data’s.	CO2	K2

UNIT III ESSENTIALS OF R

PART-A (1 Mark)			
S.No	**Questions**	**COs**	**BT**
1.	What is the result of the following R expression? x <- c(2, 4, 6), y <- c(1, 2, 3) , x + y a) c(3, 6, 9) b) c(1, 2, 3) c) c(2, 4, 6) d) Error	CO3	K1
2.	In R, which function is commonly used for One-Hot Encoding? a) model.matrix()	CO3	K1

	b) dummyVars() c) oneHotEncoder() d) get_dummies()		
3.	Which of the following can be stored in a data frame in R? a) Only numeric data b) Only character data c) Numeric, character, and logical data d) Only matrix data	CO3	K1
4.	To remove duplicate rows in a DataFrame, which method is used? a) .drop_duplicates() b) .remove_duplicates() c) .delete_duplicates() d) .unique()	CO3	K1
5.	In One-Hot Encoding, each unique category in a categorical feature is represented by: a) A single binary column b) A unique integer value c) Multiple binary columns d) A scaled numerical value	CO3	K1
6.	What is the purpose of the term "one-hot encoding" in machine learning? a)Handling missing values b)Scaling numerical features c)Encoding categorical variables into binary vectors d)Reducing dimensionality	CO3	K1
7.	What is the purpose of the term "feature engineering" in machine learning? a)Extracting valuable information from the target variable b)Creating new features or modifying existing ones to improve model performance c)Selecting the most important features for model training d)Normalizing feature values to have zero mean and unit variance	CO3	K1
8.	In R, what does the str() function do? a. It converts an object to a string b. It provides compact information about an object c. It concatenates multiple strings together d. It checks if an object is a string	CO3	K1
9.	How do you install a package in R? a) install.package("package_name") b) library("package_name") c) install.packages("package_name") d) require("package_name")	CO3	K!
10.	How can you add a new column to an existing data frame df in R? a) df$new_col <- values b) addColumn(df, values) c) df <- add.new.col(df, values) d) df.new_col <- values	CO3	K1
	PART-B (2 Marks)		
1.	List the keywords in R.	CO3	K1
2.	Define Data Frames.	CO3	K1
3.	List the features of R.	CO3	K1
4.	Outline R data frame structure.	CO3	K1
5.	What is R, and why is it popular in data science?	CO3	K1
6.	What is the purpose of the summary () function in R?	CO3	K1
7.	What is the difference between a vector and a list in R?	CO3	K2

8.	How do you create a data frame in R?	CO3	K1
9.	How do you read a CSV file in R?	CO3	K2
10.	How do you merge two data frames in R?	CO3	K1
11.	List the different control structure in R.	CO3	K1
PART C (14 Marks)			
12.	Recall the concept of R Programming Basics.	CO3	K1
13.	Explain Feature Engineering in detail.	CO3	K2
14.	Summarize different Control Structures in R.	CO3	K2
15.	Organize various data types and objects in R.	CO3	K3
16.	Relate Label Encoding and One Hot Encoding.	CO3	K1
17.	Explain data frames in R.	CO3	K2

CANVAS

UNIT-IV MODEL FIT USING R

PART-A (1 Mark)			
S.No	**Questions**	**COs**	**BT**
1.	What does the k in k-Nearest Neighbors (k-NN) represent in classification? a) The number of input features b) The number of nearest neighbors considered for classification c) The number of clusters d) The number of trees in the forest	CO5	K1
2.	What is the main purpose of clustering in machine learning? a) Predicting future values b) Grouping similar data points c) Reducing data dimensionality d) Classifying data into predefined categories	CO5	K1
3.	The relationship between the independent and dependent variables, in Linear Regression, is assumed to be: a) Linear b) Non-linear c) Exponential d) Random	CO5	K1
4.	How each tree trained in Random Forest,? a) Using the entire dataset b) Using a subset of features and samples c) Using only categorical features d) Using only the most important features	CO5	K1
5.	What is the key principle behind the k-nearest neighbors (KNN) algorithm? a)Minimizing the sum of squared errors b)Maximizing the margin between classes c)Assigning a data point the most common label among its k-nearest neighbors d)Minimizing the cross-entropy loss	CO5	K1
6.	In linear regression, what does the coefficient of a variable represent?	CO5	K1

	a) The correlation between variables b) The intercept of the regression line c) The change in the dependent variable for a unit change in the independent variable d) The error term		
7.	What does a "leaf node" represent in a decision tree? a) A splitting point b) A final decision or outcome c) An intermediate node d) The root of the tree	CO5	K1
8.	How does a Random Forest algorithm create multiple trees? a) By using the same subset of data for all trees b) By using bootstrap sampling and random feature selection c) By splitting the data equally for each tree d) By clustering the data before creating the trees	CO4	K1
9.	What is the difference between Agglomerative and Divisive Hierarchical Clustering? a) Agglomerative starts with individual points and merges clusters, while Divisive starts with one cluster and splits it. b) Agglomerative uses mean values, and Divisive uses median values. c) Agglomerative is used for classification, and Divisive is used for regression. d) They are the same but use different distance metrics.	CO4	K1
10.	Which of the following is a popular clustering algorithm? a) K-Means b) Support Vector Machine c) Linear Regression d) Random Forest	CO5	K1
PART-B (2 Marks)			
1.	What is Regression Model?	CO4	K1
2.	Interpret Model fitting	CO4	K2
3.	Compare Hard Clustering and Soft Clustering.	CO4	K2
4.	Infer Regression model.	CO4	K2
5.	Interpret Random forest algorithm.	CO4	K2
6.	How do you fit a linear regression model in R?	CO4	K1
7.	How do you check the goodness-of-fit of a regression model in R?	CO4	K1
8.	What function in R is used to fit a logistic regression model?	CO4	K2
9.	What is cross-validation, and which R package is commonly used for it?	CO4	K1
10.	How do you visualize the fit of a linear regression model in R?	CO4	K2
11.	How can you evaluate the performance of a model in R?	CO4	K1
12.	What is overfitting in classification models?	CO4	K2
13.	What is the difference between precision and recall?	CO4	K2
14.	Name two algorithms used for classification problems.	CO4	K1
15.	How is K-Nearest Neighbors (KNN) used in classification?	CO4	K2
PART C - (14 Marks)			
16.	Demonstrate Classification Model with examples.	CO4	K2
17.	Outline the concept of Hierarchical and K Means Clustering.	CO4	K1

18.	Examine briefly about Decision Tree, Naïve Baye algorithm.	CO4	K4
19.	Inspect K Means and Hierarchical clustering.	CO4	K4
20.	Explain regression model	CO4	K1
21.	Differentiate Linear and Logistic Regression model	CO4	K4
22.	Summarizes the concept of SVM and Random forest	CO4	K2

UNIT-V VISUALIZATION

PART-A (1 Mark)			
S.No	**Questions**	**Cos**	**BT**
1.	Infer the type of chart is best suited to show the distribution of a single variable? a) Line chart b) Histogram c) Pie chart d) Scatter plot	CO5	K2
2.	Identify the following methods commonly used for data balancing? a) Dimensionality reduction b) Label encoding c) Oversampling the minority class d) Scaling the data	CO5	K3
3.	What is the primary purpose of a scatter plot? a)Showing the distribution of a single variable b)Representing hierarchical relationships c)Visualizing the relationship between two variables d)Displaying the summary statistics of a dataset	CO5	K1
4.	What is the purpose of a box plot in data analysis? a)It shows the correlation between two variables b) It presents the frequency of data points c)It provides a five-number data summary d) It indicates the variance of a data set	CO5	K1
5.	In Matplotlib, what does the plt.show() function do? a). It generates a plot b). It displays a plot c). It saves a plot d). It clears the current plot	CO5	K1
6.	What is a heatmap typically used for in data visualization? a). To represent data in a table format b). To visualize correlation between variables c). To show geographical data d). To plot time series data	CO5	K1
7.	What is the primary goal of data visualization? a) To clean and preprocess data b) To represent data visually for better understanding c) To train machine learning models d) To perform complex statistical analysis	CO5	K1
8.	Which of the following is a common tool used for interactive data visualization? a) Power BI	CO5	K1

	b) Excel c) Tableau d) All of the above		
9.	Which chart type is most appropriate for comparing parts of a whole? a) Scatter plot b) Stacked bar chart c) Line chart d) Histogram	CO6	K1
10.	What is a common drawback of using a pie chart? a) It cannot display categorical data. b) It becomes difficult to interpret when there are too many slices. c) It does not show percentages. d) It requires data normalization.	CO6	K1
PART-B (2 Marks)			
1.	What is data visualization and its types?	CO5	K1
2.	Compare balanced and imbalance data.	CO5	K2
3.	What is the primary goal of data visualization in data science?	CO5	K1
4.	What is the difference between a histogram and a bar chart?	CO5	K1
5.	How do box plots help in data visualization?	CO5	K1
6.	What is the purpose of a scatter plot?	CO5	K1
7.	What is a heatmap, and when would you use it in data visualization?	CO5	K1
8.	What is a line chart used for in data visualization?	CO5	K1
9.	What is the purpose of using a pie chart in data visualization?	CO5	K1
10.	How do stacked bar charts differ from grouped bar charts?	CO6	K1
11.	Why is color important in data visualization?	CO6	K1
12.	What is a density plot, and how is it different from a histogram?	CO6	K1
13.	Term "data aggregation" refer to in data visualization?	CO6	K1
PART C (14 Marks)			
14.	Importance of Box plot	CO5	K5
15.	Discuss about Data Balancing	CO5	K6
16.	Explain the following Outlier detection	CO5	K2
17.	Explain the following scatter plot	CO5	K2
18.	Elaborate Data Visualization.	CO5	K6
19.	Write short notes on Histogram	CO6	K1
20.	Summaries the concept of Tableau	CO6	K2

Note: K1-Remembering, K2-Understanding, K3-Applying, K4-Analyzing, K5-Evaluating, K6-Creating

CO1:	Recall the basic knowledge in data science.
CO2:	Organize real time data into suitable form for analysis.
CO3:	Apply encoding techniques based on project objectives
CO4:	Construct algorithmic data models using machine learning techniques.
CO5:	Identify necessary requirement for data visualization to plot graphs using R.
CO6:	Solve real-world problems through case studies.

Prepared by
(Dr.S.SingaravelanP/CSE)
(Dr.R.Arun, ASP/CSE)

Approved
HOD/CSE