Roll No.----------

प्रश्नपुस्तिका क्रमांक
Question Booklet No.

615643

O.M.R. Serial No. | | | | | | | |

BCA (Sixth Semester) Examination, 2024-25

(NEP)

(BCA6004)

DATA SCIENCE AND MACHINE LEARNING

K-704

Paper Code							
A	6	0	0	0	1	4	T

(To be filled in the OMR Sheet)

प्रश्नपुस्तिका सीरीज
Question Booklet Series
C

SEA

Time : 1:30 Hours] [Maximum Marks-75

Instructions to the Examinee :

1. Do not open the booklet unless you are asked to do so.
2. The booklet contains 100 questions. Examinee is required to answer 75 questions in the OMR Answer-Sheet provided and not in the question booklet. All questions carry equal marks.
3. Examine the Booklet and the OMR Answer-Sheet very carefully before you proceed. Faulty question booklet due to missing or duplicate pages/questions or having any other discrepancy should be got immediately replaced.

(Remaining instructions on the last page)

परीक्षार्थियों के लिए निर्देश :

1. प्रश्न–पुस्तिका को तब तक न खोलें जब तक आपसे कहा न जाए।
2. प्रश्न–पुस्तिका में 100 प्रश्न हैं। परीक्षार्थी को 75 प्रश्नों को केवल दी गई OMR आन्सर–शीट पर ही हल करना है, प्रश्न–पुस्तिका पर नहीं। सभी प्रश्नों के अंक समान हैं।
3. प्रश्नों के उत्तर अंकित करने से पूर्व प्रश्न–पुस्तिका तथा OMR आन्सर–शीट को सावधानीपूर्वक देख लें। दोषपूर्ण प्रश्न–पुस्तिका जिसमें कुछ भाग छपने से छूट गए हों या प्रश्न एक से अधिक बार छप गए हो या उसमें किसी अन्य प्रकार की कमी हो, तो उसे तुरन्त बदल लें।

(शेष निर्देश अन्तिम पृष्ठ पर)

1. Why is stationarity important in time series analysis?

(A) It ensures data completeness

(B) It stabilizes variance

(C) It allows for accurate forecasting

(D) It reduces data size

2. Which of the following is commonly used to detect seasonality in time series data?

(A) Histogram

(B) Autocorrelation

(C) Scatter plot

(D) PCA

3. What is a key characteristic of time series data?

(A) Random observations

(B) Data without timestamps

(C) Sequential observations over time

(D) Categorical data

4. A dataset has missing values for important features. What is the best approach to address this?

(A) Remove the rows

(B) Impute values

(C) Drop the feature

(D) Ignore the missing data

5. A numerical feature has a skewed distribution. What transformation can address this?

(A) Log transformation

(B) Drop the feature

(C) One-hot encoding

(D) Normalize values

6. A dataset has highly correlated features. How should you handle this issue?

(A) Normalize features

(B) Drop one of the correlated features

(C) Encode features

(D) Use PCA

7. Which method in scikit-learn is used for dimensionality reduction?

(A) PCA()

(B) StandardScaler()

(C) KMeans()

(D) OneHotEncoder()

8. What is the main purpose of sampling in data science?

(A) To reduce data redundancy

(B) To avoid collecting data

(C) To make inferences about a population using a subset

(D) To create complex models

9. Which scikit-learn function is used to normalize data?

(A) normalize()

(B) standardize()

(C) scale()

(D) transform()

10. When should feature extraction be used instead of feature selection?

(A) When raw features are sufficient

(B) When features need transformation

(C) When data is balanced

(D) When model accuracy is high

11. What is feature selection?

(A) Adding new features

(B) Choosing the best features

(C) Removing outliers

(D) Scaling data

12. Why is feature scaling important in machine learning?

(A) Reduces model size

(B) Improves convergence during training

(C) Handles missing values

(D) Reduces overfitting

13. Which technique is commonly used to handle categorical data in feature engineering?

(A) Normalization

(B) One-hot encoding

(C) PCAD

(D) Standardization

14. What is the primary goal of feature engineering in machine learning?

(A) Improve model interpretability

(B) Reduce dataset size

(C) Enhance model performance

(D) Avoid overfitting

15. After applying KMeans, one cluster has very few data points. What should you consider next?

(A) Increase cluster count

(B) Decrease cluster count

(C) Visualize clusters

(D) Change the algorithm

16. A clustering model produces inconsistent results. What could be the likely cause?

(A) Wrong feature scaling

(B) Labeled data

(C) High accuracy

(D) Balanced dataset

17. A supervised model performs poorly on unseen data. What is the likely issue?

(A) Data leakage

(B) Underfitting

(C) Incorrect loss function

(D) Missing labels

18. How do you specify the number of clusters in the KMeans algorithm using scikit-learn?

(A) KMeans(n_clusters=n)

(B) KMeans(clusters=n)

(C) KMeans(n=n)

(D) KMeans(n_cluster=n)

19. In systematic sampling, if the sampling interval is 10, what does it mean?

(A) Every 10th element is included in the sample

(B) Only 10 elements are sampled

(C) 10 elements are skipped after each selection

(D) The first 10 elements are taken

20. Which of the following techniques ensures each subgroup in the population is properly represented?

(A) Cluster Sampling

(B) Systematic Sampling

(C) Stratified Sampling

(D) Snowball Sampling

21. Which Python library provides the K-Means function for clustering?

(A) NumPy

(B) Pandas

(C) scikit-learn

(D) Matplotlib

22. Why is clustering considered an unsupervised learning technique?

(A) It requires labeled data

(B) It uses supervised models

(C) It finds patterns in unlabeled data

(D) It predicts outcomes

23. What metric is commonly used to evaluate a regression model in supervised learning?

(A) Accuracy

(B) Mean Squared Error (MSE)

(C) Precision

(D) Silhouette score

24. Which task is best suited for unsupervised learning?

(A) Predicting house prices

(B) Identifying customer segments

(C) Spam classification

(D) Stock price prediction

25. Which of the following is an example of a supervised learning algorithm?

(A) K-Means

(B) Linear Regression

(C) Hierarchical Clustering

(D) PCA

26. What is the key difference between supervised and unsupervised learning?

(A) Supervised uses labeled data, unsupervised does not

(B) Both use labeled data

(C) Both use unlabeled data

(D) Unsupervised requires labels

27. After training a regression model, the residuals show a clear pattern. What does this imply?

(A) Model is accurate

(B) Model assumptions are violated

(C) Feature scaling is wrong

(D) Data is balanced

28. A classification model achieves 99% accuracy on the training set but only 60% on the test set. What is the issue?.

(A) Overfitting

(B) Underfitting

(C) Data imbalance

(D) Feature scaling

29. A model's predictions have high bias. What could be the likely issue?

(A) Overfitting

(B) Underfitting

(C) Feature scaling

(D) Incorrect testing data

30. Which scikit-learn function is used to calculate the accuracy of a classification model?

(A) classification_report

(B) accuracy_score

(C) score

(D) confusion_matrix

31. Which of the following is a probability sampling method?

(A) Convenience Sampling

(B) Judgmental Sampling

(C) Simple Random Sampling

(D) Quota Sampling

32. Which Python library provides the train_test_split function?

(A) NumPy

(B) Pandas

(C) Scikit-learn

(D) Matplotlib

33. Why is it important to split data into training and testing datasets?

(A) To increase dataset size

(B) To evaluate model performance on unseen data

(C) To clean data

(D) To preprocess features

34. What is the purpose of a loss function in machine learning?

(A) To evaluate model predictions

(B) To split datasets

(C) To improve visualization

(D) To standardize data

35. What is overfitting in machine learning?

(A) Model performs poorly on training data

(B) Model performs well on training data but poorly on new data

(C) Model is too simple

(D) Model has no bias

36. Which of the following is a supervised learning algorithm?

(A) K-Means

(B) Decision Trees

(C) DBSCAN

(D) Principal Component Analysis

37. What is the primary objective of machine learning?

(A) To clean data

(B) To make predictions based on data

(C) To create databases

(D) To improve hardware

38. A line chart is difficult to interpret due to too many data points. What is the best approach to simplify it?

(A) Aggregate data

(B) Remove the chart

(C) Use larger axes

(D) Switch to bar chart

39. A scatter plot shows overlapping points, making it hard to interpret. What can improve its readability?

(A) Increase marker size

(B) Add jitter

(C) Use smaller axes

(D) Change chart type

40. A pie chart in Matplotlib displays incorrect proportions. What could be the issue?

(A) Wrong data labels

(B) Missing data

(C) Incorrect sum of values

(D) Invalid chart type

41. Which Python library allows for creating highly interactive visualizations with minimal coding?

(A) Seaborn

(B) Matplotlib

(C) Plotly

(D) Pandas

42. How do you create a bar chart in Matplotlib?

(A) plt.bar(x, y)

(B) plt.plot(x, y)

(C) plt.hist(x)

(D) plt.scatter(x, y)

43. Which Matplotlib function is used to create a simple line chart?

(A) plt.scatter()

(B) plt.line()

(C) plt.plot()

(D) plt.bar()

44. Which of the following is a common mistake in data visualization?

(A) Using appropriate scales

(B) Choosing the right chart type

(C) Overloading charts with data

(D) Labeling axes

45. What does a boxplot help identify in a dataset?

(A) Outliers

(B) Correlations

(C) Clusters

(D) Trends

46. Which chart is most effective for comparing parts of a whole?

(A) Scatter plot

(B) Pie chart

(C) Boxplot

(D) Line chart

47. Which visualization is best suited for showing data distribution?

(A) Line chart

(B) Scatter plot

(C) Histogram

(D) Pie chart

48. What is the primary purpose of data visualization?

(A) To analyze data

(B) To predict outcomes

(C) To represent data visually

(D) To encode data

49. Which of the following is NOT an advantage of sampling?

(A) Cost-effective

(B) Faster data collection

(C) Full population coverage

(D) Manageable data size

50. A dataset's p-value is 0.01 after running a statistical test. What does this imply?

(A) Strong evidence against the null hypothesis

(B) No evidence against the null hypothesis

(C) Data is normally distributed

(D) Data has no variance

51. A dataset has a column with skewed numerical data. What is the best approach to normalize it?

(A) Use log transformation

(B) Drop outliers

(C) Encode values

(D) Use boxplot

52. What is the main drawback of non-probability sampling methods?

(A) They are too expensive

(B) They require complex algorithms

(C) They do not allow generalization to the population

(D) They are always biased

53. In cluster sampling, the population is divided into:

(A) Equal-sized samples

(B) Strata based on characteristics

(C) Clusters, then some clusters are randomly selected

(D) Individual observations only

54. What assumption is made about data in a parametric statistical test?

(A) Data is categorical

(B) Data follows a normal distribution

(C) Data has no missing values

(D) Data is continuous

55. What type of statistical analysis helps identify relationships between variables?

(A) Correlation analysis

(B) Variance analysis

(C) Skewness analysis

(D) Descriptive statistic

56. What does the standard deviation indicate in a dataset?

(A) The central tendency

(B) The variability

(C) The skewness

(D) The correlation

57. When is the p-value considered statistically significant in hypothesis testing?

(A) When $p > 0.05$

(B) When $p < 0.05$

(C) When $p = 0.1$

(D) When $p > 1$

58. Which statistical measure represents the spread of data values around the mean?

(A) Variance

(B) Mean

(C) Median

(D) Skewness

59. What is the primary purpose of hypothesis testing in statistics?

(A) To clean data

(B) To test assumptions

(C) To visualize trends

(D) To encode features

60. During EDA, an outlier is identified in a boxplot. What is the best course of action?

(A) Remove the outlier

(B) Keep the outlier

(C) Investigate the outlier

(D) Ignore the outlier

61. A dataset shows a perfect correlation of +1 between two variables. What is the likely issue?

(A) Multicollinearity

(B) Outliers

(C) No issue

(D) Wrong visualization

62. If a dataset contains missing values in a column, what is the simplest way to visualize its impact?

(A) Use a scatter plot

(B) Use a heatmap

(C) Drop the column

(D) Fill missing values

63. Which visualization technique is useful for identifying clusters in data during EDA?

(A) Scatter plot

(B) Heatmap

(C) Boxplot

(D) Pairplot

64. How do you compute the correlation matrix for a DataFrame in Python?

(A) df.corr()

(B) df.describe()

(C) df.cov()

(D) df.plot()

65. Which Python library is used for creating basic visualizations such as line and bar charts?

(A) NumPy

(B) Pandas

(C) Matplotlib

(D) Seaborn

66. Why is it critical to detect multicollinearity during EDA?

(A) It improves model accuracy

(B) It ensures independence among predictors

(C) It removes missing values

(D) It selects important features

67. Which visualization is best suited for analyzing the relationship between two numerical variables?

(A) Histogram

(B) Boxplot

(C) Scatter plot

(D) Bar chart

68. What is the significance of identifying skewness in data during EDA?

(A) It helps in feature scaling

(B) It determines model type

(C) It affects data distribution assumptions

(D) It improves visualization

69. Which of the following is a common technique used during EDA?

(A) Clustering

(B) PCA

(C) Descriptive statistics

(D) Feature selection

70. What is the primary goal of Exploratory Data Analysis?

(A) Predict outcomes

(B) Summarize data characteristics

(C) Visualize predictions

(D) Build models

71. After standardizing a dataset, a model performs poorly. What could be a possible issue?

(A) Data leakage

(B) Over fitting

(C) Outliers

(D) Incorrect scaling

72. A column contains both numerical and non-numerical values. How should you preprocess it for numerical analysis?

(A) Drop the column

(B) Impute missing values

(C) Use encoding techniques .

(D) Normalize data

73. A dataset has duplicate rows causing issues in analysis. Which Pandas method will you use to fix this?

(A) drop_duplicates()

(B) dropna()

(C) fillna()

(D) groupby()

74. Which Python library is best suited for outlier detection using clustering techniques?

(A) Scikit-learn

(B) NumPy

(C) Pandas

(D) Matplotlib

75. What is the first step in stratified sampling?

(A) Select clusters

(B) Define strata

(C) Choose a sample size

(D) Collect data

76. In Python, which Pandas method removes rows with missing values?

(A) drop_duplicates()

(B) dropna()

(C) fillna()

(D) replace()

77. When dealing with a dataset containing multiple irrelevant features, which method is most effective?

(A) Data cleaning

(B) Feature selection

(C) One-hot encoding

(D) Standardization

78. Which preprocessing step ensures categorical variables are suitable for numerical models?

(A) Scaling

(B) One-hot encoding

(C) Outlier detection

(D) Normalization

79. What is the effect of standardization in data preprocessing?

(A) It removes duplicates

(B) It ensures data values are centered around zero

(C) It improves data cleaning

(D) It removes missing values

80. Which technique can be used to handle outliers in numerical data?

(A) Removing them

(B) Normalizing data

(C) Imputation

(D) All of the above

81. Why is handling missing values important during data preprocessing?

(A) It ensures model interpretability

(B) It improves model accuracy

(C) It increases data storage

(D) It simplifies code

82. What is the primary goal of data cleaning in Data Science?

(A) To remove duplicates

(B) To visualize data

(C) To identify and fix data quality issues

(D) To split data

83. A Data Scientist's model performs poorly in production compared to testing. What could be the most likely cause?

(A) Over fitting

(B) Clean data

(C) Balanced dataset

(D) Simple model

84. In Python, which library is commonly used for splitting datasets during the Data Preparation phase?

(A) Scikit-learn

(B) NumPy

(C) Pandas

(D) Matplotlib

85. What is a major challenge during the evaluation phase of the Data Science Life Cycle?

(A) Selecting the right metric

(B) Collecting data

(C) Training models

(D) Understanding business goals

86. Why is the deployment phase critical in the Data Science Life Cycle?

(A) It ensures the model is trained

(B) It makes the model accessible for users

(C) It removes irrelevant data

(D) It generates reports

87. Which step in the Data Science Life Cycle involves feature engineering and transformation?

(A) Problem Definition

(B) Data Cleaning

(C) Data Preparation

(D) Evaluation

88. What happens during the Data Collection phase of the Data Science Life Cycle?

(A) Data is stored in a database

(B) Data is gathered from multiple sources

(C) Data is split into training and test sets

(D) Data is discarded

89. Which step in the Data Science Life Cycle involves determining if the model meets project objectives?

(A) Data Collection

(B) Model Deployment

(C) Evaluation

(D) Visualization

90. Which phase in the Data Science Life Cycle involves cleaning and preparing data for analysis?

(A) Model Evaluation

(B) Data Cleaning

(C) Data Analysis

(D) Visualization

91. What is the first step in the Data Science Life Cycle?

(A) Model Building

(B) Data Cleaning

(C) Problem Definition

(D) Evaluation

92. Which of the following statements is TRUE about simple random sampling?

(A) It gives preference to certain groups

(B) It involves subjective judgment

(C) It requires a complete list of the population

(D) It does not represent the population well

93. In Python, which library is commonly used for numerical computations in Data Science?

(A) NumPy

(B) Matplotlib

(C) Flask

(D) Pandas

94. Which of the following is a critical component of a Data Science pipeline?

(A) Web hosting

(B) Feature selection

(C) Presentation design

(D) Software installation

95. What role does domain expertise play in Data Science?

(A) It is optional

(B) It provides data storage solutions

(C) It helps understand data context

(D) It prevents coding errors

96. What is a key challenge faced in Data Science projects?

(A) Lack of storage

(B) Model over fitting

(C) Manual calculations

(D) System downtime

97. Which of these domains does Data Science NOT directly involve?

(A) Machine learning

(B) Database optimization

(C) Statistics

(D) Data visualization

98. What type of data does Data Science primarily handle?

(A) Only structured

(B) Only unstructured

(C) Both structured and unstructured

(D) None

99. Which of the following is a key aspect of data science?

(A) Building dashboards

(B) Cleaning and analyzing data

(C) Developing web pages

(D) Writing blogs

100. What is Data Science primarily focused on?

(A) Data storage

(B) Data visualization

(C) Insight extraction

(D) App development

******

4. Four alternative answers are mentioned for each question as – A, B, C & D in the question booklet. The candidate has to choose the correct answer and mark the same in the OMR Answer-Sheet as per the direction :

Example :

Question :

Q. 1 (A) ● (C) (D)
Q. 2 (A) (B) ● (D)
Q. 3 (A) ● (C) (D)

Illegible answers with cutting and over-writing or half filled circle will be cancelled.

5. Each question carries equal marks. Marks will be awarded according to the number of correct answers you have.
6. All answers are to be given on OMR Answer Sheet only. Answers given anywhere other than the place specified in the answer sheet will not be considered valid.
7. Before writing anything on the OMR Answer Sheet, all the Instructions given in it should be read carefully.
8. After the completion of the examination candidates should leave the examination hall only after providing their OMR Answer Sheet to the invigilator. Candidate can carry their Question Booklet.
9. There will be no negative marking.
10. Rough work, if any, should be done on the blank pages provided for the purpose in the booklet.
11. To bring and use of log-book, calculator, pager and cellular phone in examination hall is prohibited.
12. In case of any difference found in English and Hindi version of the question, the English version of the question will be held authentic.

Impt. On opening the question booklet, first check that all the pages of the question booklet are printed properly. If there is any discrepancy in the question booklet, then after showing it to the invigilator, get another question booklet of the same series.

4. प्रश्न–पुस्तिका में प्रत्येक प्रश्न के चार सम्भावित उत्तर– A, B, C एवं D हैं। परीक्षार्थी को उन चारों विकल्पों में से एक सही उत्तर छाँटना है। उत्तर को OMR आन्सर–शीट में सम्बन्धित प्रश्न संख्या में निम्न प्रकार भरना है :

उदाहरण :

प्रश्न :

प्रश्न 1 (A) ● (C) (D)
प्रश्न 2 (A) (B) ● (D)
प्रश्न 3 (A) ● (C) (D)

अपठनीय उत्तर या ऐसे उत्तर जिन्हें काटा या बदला गया है, या गोले में आधा भरकर दिया गया, उत्तर निरस्त कर दिया जाएगा।

5. प्रत्येक प्रश्न के अंक समान हैं। आपके जितने उत्तर सही होंगे, उन्हीं के अनुसार अंक प्रदान किये जायेंगे।
6. सभी उत्तर केवल ओ. एम. आर. उत्तर–पत्रक (OMR Answer Sheet) पर ही दिये जाने हैं। उत्तर–पत्रक में निर्धारित स्थान के अलावा अन्यत्र कहीं पर दिया गया उत्तर मान्य नहीं होगा।
7. ओ. एम. आर. उत्तर–पत्रक (OMR Answer Sheet) पर कुछ भी लिखने से पूर्व उसमें दिये गये सभी अनुदेशों को सावधानीपूर्वक पढ़ लिया जाये।
8. परीक्षा समाप्ति के उपरान्त परीक्षार्थी कक्ष निरीक्षक को अपनी OMR Answer Sheet उपलब्ध कराने के बाद ही परीक्षा कक्ष से प्रस्थान करें। परीक्षार्थी अपने साथ प्रश्न–पुस्तिका ले जा सकते हैं।
9. निगेटिव मार्किंग नहीं है।
10. कोई भी रफ कार्य, प्रश्न–पुस्तिका के अन्त में, रफ–कार्य के लिए दिए खाली पेज पर ही किया जाना चाहिए।
11. परीक्षा–कक्ष में लॉग–बुक, कैलकुलेटर, पेजर तथा सेल्युलर फोन ले जाना तथा उसका उपयोग करना वर्जित है।
12. प्रश्न के हिन्दी एवं अंग्रेजी रूपान्तरण में भिन्नता होने की दशा में प्रश्न का अंग्रेजी रूपान्तरण ही मान्य होगा।

महत्वपूर्ण : प्रश्नपुस्तिका खोलने पर प्रथमतः जाँच कर देख लें कि प्रश्न–पुस्तिका के सभी पृष्ठ भलीभाँति छपे हुए हैं। यदि प्रश्नपुस्तिका में कोई कमी हो, तो कक्षनिरीक्षक को दिखाकर उसी सिरीज की दूसरी प्रश्न–पुस्तिका प्राप्त कर लें।