Degree: B. Tech Semester: Vth — Course work

MID-SEMESTER EXAMINATION, September, 2024

Course Title: Data Handling and Visualization Tools

Course Code: CDCSE01

Duration: 1:30 Hours Max. Marks: 25

Note: - Attempt all questions in the given order only. Missing data/information (if any), maybe suitably assumed & mentioned in the answer.

Q. No.	Question	Marks	CO
1a	Explain the stages involved in a typical data science. Illustrate each stage with an example from a real-world application.	2.5	1
1b	A data scientist needs to collect and analyze data from a social media platform to study user engagement trends. Identify two potential data security issues that could arise in this context and propose solutions to mitigate these issues.	2.5	1
2a 2	You have a dataset with 1,000 records across 5 columns: A, B, C, D, and E. The dataset is structured as follows: 1. **Column A**: Contains 1,000 values. 2. **Column B**: Contains 1,000 values. 3. **Column C**: Contains 1,000 values. 4. **Column D**: Contains 1,000 values. 5. **Column E**: Contains 1,000 values. Upon analysis, you find the following: • 200 tuples (rows) have missing values in Column A. • 150 tuples have missing values in Column B. • 100 tuples have missing values in Column C. • 50 tuples have missing values in Column D. • 80 tuples have missing values in Column E. However, some tuples have missing values in more than one column. i. Determine the number of tuples with missing values in at least one column. Assume that no tuple is counted more than once, even if it has missing values in multiple columns. —max = 200 ii. Calculate the percentage of tuples with missing values in at least one column with respect to the total number of tuples. → $\frac{200}{1000} \times 100 = 20\%$ Discuss possible strategies for handling missing data without removing the rows with missing values. Explain how each strategy can be applied and the potential impact on the dataset. imputation, ignora	2.5	2
2b 4	Explain the importance of data cleaning in the data pre-processing stage. Provide an example where improper data	2.5	1

	cleaning led to misleading results.		
3a	Given a dataset with the following data points for the variable "X": [12, 15, 13, 20, 18, 14, 17, 16, 15, 19], calculate the mean, mode, standard deviation, skewness, and kurtosis. Interpret the skewness and kurtosis values in terms of the data distribution shape.	2.5	2
3b	How would you use a residual plot to evaluate the accuracy of a polynomial regression model? What specific patterns in the residual plot would indicate that the model is well-fitted or poorly fitted?	2.5	3
4a	Describe the purpose of a pivot table in data analysis. Provide an example of how a pivot table can be used to summarize data for a retail store.	2.5	2
4b	Discuss the interpretation of p-values in ANOVA. What does a low p-value indicate, and how should you proceed based on this result? Provide an example scenario where ANOVA might be used in a real-life situation.	2.5	3
5a	What is the difference between simple regression and multiple regression? Provide a scenario where multiple regression is preferred over simple regression.	2.5	3
5b	Explain a heat map and describe its typical use cases in data analysis. How does a heat map differ from other data visualization techniques like scatter plots or line graphs?	2.5	3

B.Tech. Semester: V- Course work

END-SEMESTER EXAMINATION, Nov-Dec 2024

Course Title: Data Handling and Visualization Tools

Course Code: CDCSE01

me: 03 Hours Max. Marks: 50

ote: - Attempt all the five questions. Missing data/ information (if any), maybe suitably assumed & entioned in the answer.

Q. No.	Question	Marks	CO
Q 1	**Attempt any 2 parts of the following.**		
1a	Outline the typical stages in a data science project lifecycle. i. Explain the importance of each stage, from problem definition to deployment. ii. What challenges might arise at each stage, and how can they be addressed?	5	1
1b	i. Compare and contrast different data collection strategies. What are the advantages and disadvantages of each method in terms of data quality and relevance? ii. Discuss how data integration and transformation can enhance the usability of data from multiple sources.	5	1
1c	Analyse the ethical implications of data science applications. i. How can data privacy concerns be addressed while leveraging data for decision-making? ii. What frameworks or guidelines should organizations follow to ensure ethical use of data?	5	1
Q 2	**Attempt any 2 parts of the following.**		
2a	Discuss the concept of in-sample evaluation and its importance in model development. i. What measures (e.g., R-squared, adjusted R-squared) can be used to evaluate the performance of a regression model? ii. How do you interpret these metrics in the context of your model?	5	2
2b	Explain the ANOVA (Analysis of Variance) and describe a real-time scenario where it would be appropriate to use ANOVA. i. How does ANOVA help in comparing means across multiple groups? ii. What assumptions must be met for ANOVA to be valid?	5	2
2c	Given a dataset with the following values: [12, 14, 15, 15, 16, 18, 19, 20, 22, 30], calculate the skewness and kurtosis. i. Interpret the results: What do the values of skewness and kurtosis suggest about the distribution? ii. How would you visually confirm your findings using a histogram or a box plot?	5	2
Q 3	**Attempt any 2 parts of the following.**		
3a	Define what is meant by "integrity in visualization." i. Why is integrity important in the context of data visualization? ii. Provide an example of a visualization that lacks integrity and discuss the potential implications of misleading visualizations.	5	3

<table>
<tr><td>3b ✓</td><td>List and describe the seven common quantitative relationships in graphs.
i. Explain the significance of each relationship in data analysis.
ii. Provide an example of a dataset that would be appropriate for illustrating each relationship.</td><td>5</td><td>3</td></tr>
<tr><td>3c</td><td>Explain how improper visualization techniques can mislead viewers.
i. What are some common pitfalls in data visualization that can distort information?
ii. Describe how intentional or unintentional biases in visualization can impact decision-making.</td><td>5</td><td>3</td></tr>
<tr><td>Q 4</td><td><b>Attempt any 2 parts of the following.</b></td><td></td><td></td></tr>
<tr><td>4a ✓</td><td>Given two datasets, one containing sales data and another containing customer information, explain the steps to merge them in Python using Pandas. Then, demonstrate reshaping techniques to pivot this data based on customer regions.</td><td>5</td><td>4</td></tr>
<tr><td>4b</td><td>Explain what regular expressions are and provide an example of how they can be used for string matching in Python.
i. Describe how to use the <b><i>remodule</i></b> in Python to search for a pattern in a string.
ii. What is the significance of special characters in regular expressions? Provide examples of at least three special characters and their meanings.</td><td>5</td><td>4</td></tr>
<tr><td>4c ✓</td><td>Explain the concept of data aggregation and its importance in data analysis.
i. How does the <b><i>GroupBy</i></b> operation work in pandas, and what is its purpose?
ii. Provide a scenario where grouping data would be beneficial, including an example of a <b><i>DataFrame</i></b> you might use.</td><td>5</td><td>4</td></tr>
<tr><td>Q 5</td><td><b>Attempt any 2 parts of the following.</b></td><td></td><td></td></tr>
<tr><td>5a ✓</td><td>Identify and explain at least three factors that can lead to overfitting in a machine learning model.
i. How does the tradeoff between bias and variance relate to overfitting and underfitting?
ii. How can the size of the training dataset contribute to overfitting?
iii. Describe the role of noise in the training data and its effect on overfitting.</td><td>5</td><td>5</td></tr>
<tr><td>5b ✓</td><td>Suppose you have trained a binary classifier to predict whether emails are spam (positive class) or not spam (negative class). After evaluating your classifier on a test set of 100 emails, you obtained the following confusion matrix:
<table>
<tr><td></td><td>Predicted not Spam</td><td>Predicted Spam</td></tr>
<tr><td>Actually not Spam</td><td>65</td><td>5</td></tr>
<tr><td>Actually Spam</td><td>10</td><td>20</td></tr>
</table>
</td><td>5</td><td>5</td></tr>
</table>

	Using this confusion matrix, calculate the following performance metrics for the classifier: i. Accuracy ii. Precision for the spam class iii. Recall for the spam class iv. F1-score for the spam class		
5c	Define cross-validation and explain how it can help to prevent overfitting. Use the concept of Ridge Regression to show how regularization can improve model generalization.	5	5