

Data Analysis and Visualization Lab (D 3001)

Date: December 31st 2024


Instructor(s):

Eesha-tur-Razia, Amna Zulifqar, Usman Anwer
Mateen Fatima, Bassam Ahmad

22L-7803
Roll No

BDS-5A
Section

Total Time (Hrs): 3
Total Marks: 85
Total Questions: 3


Student Signature

Do not write below this line

- The starter notebook, datasets are given in the xeon folder.
- Submission path: `\cactus1\xeon\Fall 2024\Usman Anwer\DAV Lab Final\Submissions**YOUR SECTION**`
- You can only drag/drop in the folder.
- Submit in respective section only. Submit the file with your student id e.g: 22L-1234.ipynb. Submissions without student ids will be credited against zero.
- Double side handwritten A4 cheat sheet is allowed.
- Understanding the question is also part of it. Make assumptions wherever required.

CLO #1:

Question 1: Quantitative data and image data preprocessing and visualization [30 marks]

Q1a:

- a) Load the provided "exam_data" data into a pandas DataFrame and Display the first few rows of the dataset [0.5 Mark] (Remember its an excel file)
- b) Display statistics summary of the dataset [0.5 Mark]
- c) Detect and remove outliers in Height (cm) column [2 Mark]
- d) Create a correlation heatmap for all numerical features. [2 Marks]
- e) fill missing values in numerical column with the mean of the column [2 Marks]
- f) Count the values of each Department in the Department column [1 Mark]
- g) Create a bar plot of Salary By Department [1 Mark]
- h) Create a histogram for Employee Age Distribution [1 Mark]

Q1b:

- 1) Load the given "parrot_picture" on opencv, Load the image in grayscale, define Sobel filter, apply convolution on image using Sobel filter, show original and convoluted image. [5 marks]
- 2) Load the given "house.tiff". In this task you have to convert the image into Gray Scaling and separate the 3 color from image i.e red green and blue. Implement color channel separation and visualize each channel [5 marks]
- 3) Load house.tiff image. Create a translation matrix to shift the image by 100 pixels along the x-axis and 50 pixels along the y-axis. Extract the number of rows and columns from the image. Apply the Translation. Convert the translated image from BGR to RGB format. Use matplotlib to display the translated image without axis labels. [5 marks]

- 4) Make four copies of house.tiff image. Rotate each image by 90, 180, 270, and 360 degrees. Display rotated images. Apply Canny edge detection on each rotated image and display all the resultant images [5 marks]

CLO # 2

Question 2: Email Classification Using Cosine Similarity, Naive Bayes, and Logistic Regression (20 marks)

Load the given "emails" data set.

Notes:

Ensure that the dataset is preprocessed before applying the models.
Feel free to add visualizations (e.g., confusion matrix) to support the analysis.
Use built-in functions for Cosine Similarity, Naïve Bayes, Logistic Regression
Write the code for preprocessing, vectorization, and similarity computation.

Question 2a: Cosine Similarity (5 marks)

Given the dataset, preprocess the text (subject and body) to compute the Term Frequency-Inverse Document Frequency (TF-IDF) vectors.

- (a) Compute the cosine similarity between two specific emails in the dataset.
- (b) Identify the most similar email to Email ID 1 based on cosine similarity.

Write the code for preprocessing, vectorization, and similarity computation.

Question 2b: Naive Bayes Classifier (5 marks)

Train a Naive Bayes classifier on the dataset to predict whether an email is spam or not.

- (a) Split the data into training and testing sets.
- (b) Train the Naive Bayes model using the preprocessed data.
- (c) Evaluate the accuracy of the model on the testing set.

Write the code for data splitting, training, and evaluation.

Question 2c: Logistic Regression (5 marks)

Use Logistic Regression to classify emails as spam or non-spam.

- (a) Train a Logistic Regression model on the dataset.
- (b) Use feature importance to analyze which words contribute most to classifying spam emails.
- (c) Compare the accuracy of Logistic Regression with Naive Bayes.

Write the code for model training, feature analysis, and evaluation.

Question 2d: Combined Analysis (5 marks)

Compare the results of Cosine Similarity, Naive Bayes, and Logistic Regression.

- (a) Which method performs better in terms of accuracy and why?
- (b) Provide a brief analysis of when each technique might be more appropriate to use.

Write the code or explanation for combined analysis.

CLO # 3

Question 3: R language, plotly, RNN [35 marks]

Question 3a: R language (5 marks)

You are given a matrix of numeric values representing the scores of students in different subjects. Each row corresponds to a student, and each column corresponds to a subject. Your task is to write an R function

`compute_sums(matrix, by)` that computes either the row-wise or column-wise sums of the matrix based on a user-defined parameter. If `by=1` then it computes the row-wise sum and for `by=0` it computes the column-wise sum. (only R language is allowed).

Question 3b: Plotly (10 marks)

You have to analyze sales performance of a company across its various product categories. The data contains individual sales transactions, including the Product Category, Region, and the Sales Amount in dollars. Your goal is to create a pie chart using Plotly to visualize the total sales contribution of each product category to the company & overall revenue.

Initial data is given as:

* `categories = [Electronics, Clothing, Home Appliances, Books, Sports Equipment]`

* `regions = [North, South, East, West]`

Pie plot will contain %age share of each category. Total sales of all categories is summed up to 100. Label of each category in pieplot will contains the \$ values of sales for said category. (Only use plotly. Use of matplotlib is not allowed)

Question 3c: RNN (20 marks)

The MNIST handwritten dataset contains 60,000 grayscale images for training and 10,000 for testing, each representing digits from 0 to 9. Each image is 28x28 pixels, making it a benchmark dataset for machine learning and deep learning tasks. You have to apply RNN (Recurrent Neural Networks) on the above dataset by dividing the dataset into training-testing subsets. Also find the testing accuracy by creating confusion matrix.