Name: _____ Roll#_____

National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	**Data Analysis and Visualization**	**Course Code:**	**DS3001**
	Degree Program:	**BS (Data Science)**	**Semester:**	**Fall 2023**
	Exam Duration:	**180 Minutes**	**Total Marks:**	**100**
	Paper Date:	**30-December-2023**	**Weight**	**40%**
	Section:	**ALL**	**Page(s):**	**14**
	Exam Type:	**Final Exam**		

Name:_____ Roll No._____ Section:_____

Instruction/Notes:
- Attempt all questions on the provided question paper.
- Space for rough work is provided at the end of the paper.
- Even if you do use extra rough sheets, they should **NOT** be attached with final paper.
- **Show all calculations and steps, direct answers are not acceptable**.

Question #	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Total Marks	**20**	**15**	**20**	**15**	**15**	**15**	**100**
Obtained Marks							

Question#1

a) Perform the histogram equalization on the provided input image.

4	1	3	2
3	1	1	1
0	1	5	2
1	1	2	2

input image

Input Image Histogram:

Number of levels L=_____

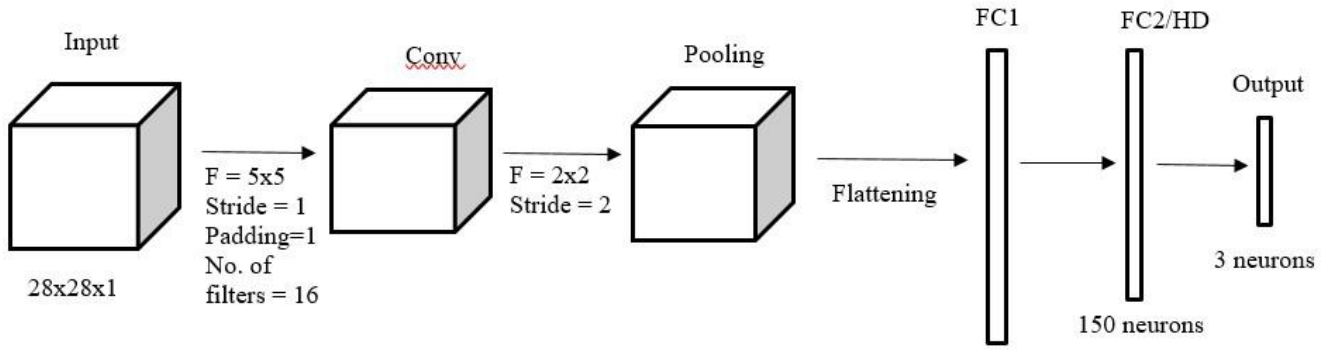
Transformation= s_k=_____

Name: _____ Roll#_____

Gray Level r_k	Frequency n_k	Normalized Histogram $p(r_k)$	Transformation s_k	Transformation s_k (rounded values)	Equalized Level $p(s_k)$

Equalized Histogram	**Equalized Image**

b) You come up with a CNN classifier shown in following figure. For each layer, calculate the number of weights, number of biases and the size of the associated feature maps.



The notation follows the convention:

- F: Filter size
- Conv denotes a convolutional layer
- Pool indicates pooling layer with stride mentioned in figure
- FC-N stands for a fully connected layer with N neurons
- output-N indicates output layer with N neurons

Layer	Dimensions	Number of weights	Number of biases
Input	**28x28x1**	**0**	**0**
Conv			
Pool			
FC1			
FC2-150			
Output-3			

Name: _____ Roll#_____

Question#2

Consider a dataset containing text documents labeled with categories. You are tasked with implementing a Naive Bayes classifier for text classification.

Training Data:

Sentence ID	Text	Label
1	the sun is shining brightly	Positive
2	It is raining heavily today	Negative
3	learning is an exciting journey	Positive
4	the traffic is unbearable	negative
5	the puppy is cute and playful	Positive

Test Data:

Sentence ID	Text	Label
6	the sun is shining beautifully	**?**
7	the computer crashed frustratingly	**?**

Vocabulary: ['the', 'sun', 'shining', 'brightly', 'is', 'raining', 'heavily', 'learning', 'an', 'exciting', 'journey', 'unbearable', 'UNK']

This vocabulary includes 12 words, and any other word encountered will be replaced with the "UNK" token.

a) Create Bag of Words representation for training dataset

ID	the	sun	shining	brightly	is	raining	heavily	learning	traffic	exciting	journey	unbearable	UNK
1													
2													
3													
4													
5													

b) Compute Jaccard similarity between sentence ID1 and ID2.

c) Compute the most likely class label for test data using Naive Bayes classifier with Laplace smoothing.

Question#3

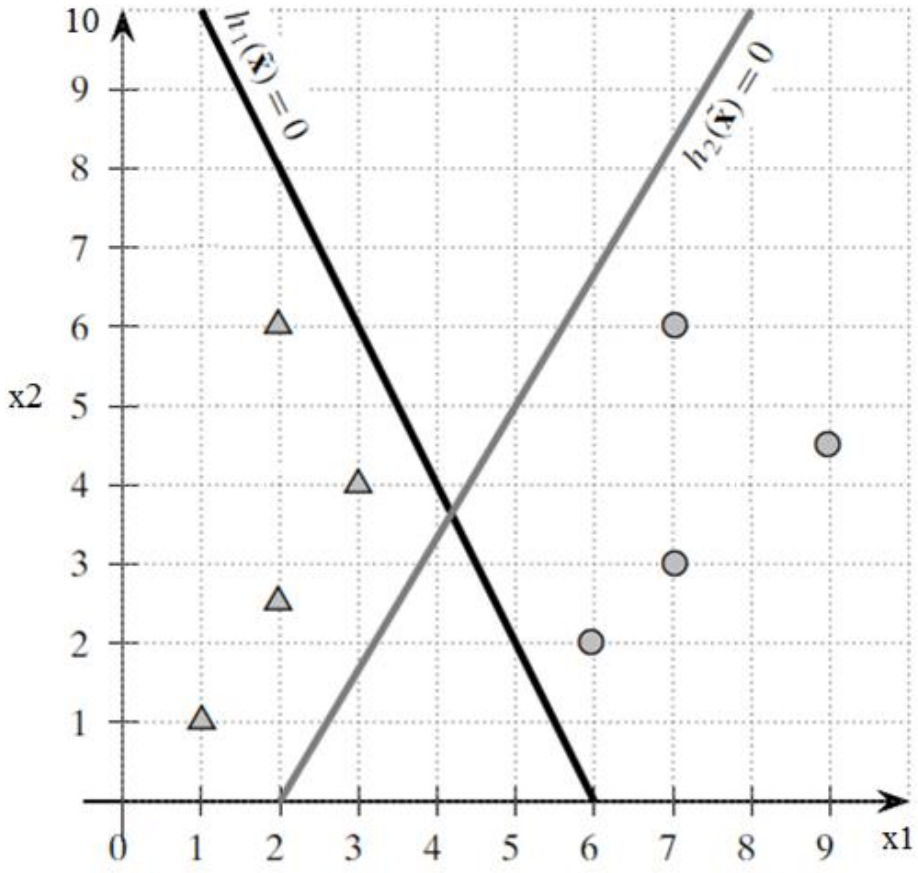
Imagine you are developing a Sentiment Analysis model using a Recurrent Neural Network (RNN). The goal is to classify the sentiment of short text sequences into two categories: "Positive" (1) or "Negative" (0). The RNN architecture consists of an input layer, a hidden layer, and an output layer. The vocabulary includes four words: **["happy", "sad", "good", <unk>]**, <unk> is used for unknown words. Use One Hot encoding for embedding input. Hidden layer size = 3, output layer size = 1. Activation function for hidden layer is **tanh** and for output layer is **sigmoid**. Weight matrices are given:

W^0				W^{11}			W^1		
0.2	0.4	0.1	0.3	0.1	0.6	0.5	0.9	0.8	0.3
0.6	0.7	0.8	0.0	0.7	0.1	0.2	b^1		
0.1	0.8	0.3	0.4	0.3	0.2	0.1			
b^0							0.2		
0.1	0.3	0.2							

You are required to compute the forward pass of the given RNN model and predict the sentiment of the following sentence. "**He good good sad**".

Question#4

Consider the dataset in figure below, which has points from two classes $\mathbf{c_1}$ (triangles) and $\mathbf{c_2}$ (circles).



a) Find the equations for the two hyperplanes h_1 and h_2.

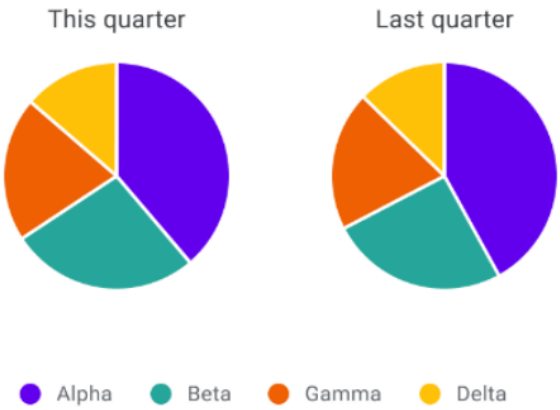
b) Which of the two hyperplanes is better at separating the two classes based on the margin computation?

c) The optimal hyperplane equation for this dataset is $h(x) = \binom{1}{0}^T \boldsymbol{x} - 4.5 = 0$. Classify the data point $x = \binom{3}{1}$ using this optimal hyperplane into class c1 or c2.

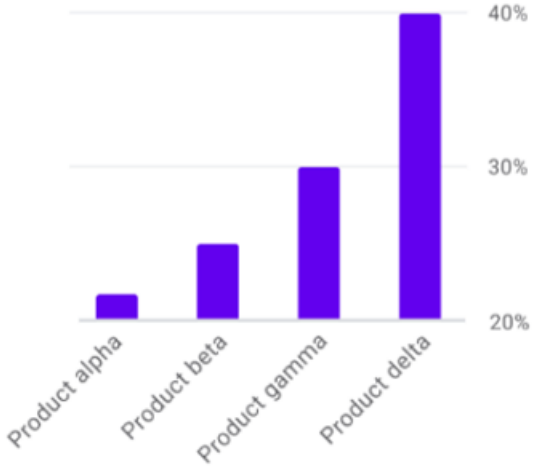
Question#5

Answer the following questions about the provided visualizations.

a) The analyst has dataset of one year monthly sales of different products (i.e. alpha, beta, gamma and delta products) of a company. The manager wants to see the comparison of sales of current quarter and previous quarter of the year for each product. The analyst provided the following graph. Point out the mistakes and suggest and alternate.



b) Point out the mistake/s in the following graph



c) You are provided with the following dataset. In order to see the trends of CO_2 over the time from the data, what type of graph would you suggest and why?

Month	Carbon Emission
Jan	646
Feb	820
Mar	950
Apr	730
May	889
Jun	467
Jul	250
Aug	359
Sep	402
Oct	690

d) Draw the graph that you have suggested in part **c**.

Question#6

In table 1, smoothed bigram probabilities were calculated on a mini corpus for language modelling. Using the given probabilities, answer the following:

a) Utilizing the given bigram probabilities for language generation, complete the sentence where '**<s>**' symbol is used for beginning of the sentence and **'<\s>'** symbol marks the ending of the sentence.

<s> sunsets _

b) Find the probability of next word given all the previous words.

P(distance| dog barked loudly in the)=

c) Compute the probability of these two sentences and determine their appropriateness using the perplexity measure.

i) P(the cat sat on the mat)
ii) P(on the mat the cat sat)

	<s>	the	cat	sat	on	mat	in	are	by	dog	barked	loudly	distance	sunsets	always	beautiful	beach	children	played	happily	park	<\s>
<s>	0.02	0.07	0.02	0.02	0.04	0.02	0.04	0.02	0.05	0.02	0.02	0.04	0.02	0.04	0.02	0.02	0.02	0.05	0.02	0.02	0.02	0.02
the	0.02	0.02	0.05	0.02	0.02	0.07	0.02	0.02	0.02	0.07	0.02	0.02	0.03	0.03	0.02	0.02	0.08	0.03	0.02	0.02	0.05	0.02
cat	0.02	0.02	0.02	0.07	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
sat	0.02	0.02	0.02	0.02	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.06
on	0.02	0.10	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
mat	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.07
in	0.02	0.09	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
are	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.07	0.02	0.02	0.02	0.02	0.02	0.02	0.02
by	0.02	0.09	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
dog	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
barked	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04
loudly	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
distance	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04
sunsets	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.07	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
always	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.07	0.02	0.02	0.02	0.02	0.02	0.02
beautiful	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.05	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04
beach	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.06
children	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.08	0.02	0.02	0.02
played	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.06	0.02	0.02
happily	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04
park	0.02	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04
<\s>	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Table 1: Bigram probabilities

ROUGH PAGE