Roll Number: _____

	Course:	Data Science	**CourseCode:**	CS 4048
	Program:	BS(CS)	**Semester:**	Fall 2023
	Duration:	3 Hour	**Total Marks:**	50
	Paper Date:		**Page(s):**	6
	Section:		**Section:**	
	Exam:	Final	**Roll No:**	

Instructions: Answer in the space provided. You can ask for rough sheets, but they will not be graded or marked. In case of confusion or ambiguity make a reasonable assumption.

Problem 1: **10 Marks**

a. Compare and contrast the use cases of Pandas GroupBy and Crosstab.

Crosstab outputs a contingency table where rows and columns represent categories and values represent count while groupby outputs a grouped object that can be used with different aggregation functions.

Default agg function of crosstab is count while you can specify other agg function using aggfunc parameter. Groupby supports various agg function directly.

b. Explain the role of lemmatization in NLP and how it contributes to the creation of feature vectors.

Lemmatization is a natural language processing technique that involves reducing words to their base or root form, known as the lemma. It plays a crucial role in feature vector creation by ensuring that different inflected forms of a word are represented by a common lemma, consolidating related words, thus contributes to control the sparsity.

c. Explain the concept of histogram equalization in image processing. How does it enhance the contrast of an image?

Histogram equalization is a technique in image processing that enhances the contrast of an image by redistributing the intensity values across the entire available range. It works by transforming the pixel intensities in such a way that the cumulative distribution function (CDF) of the image becomes more uniform.

d. Differentiate image classification, object detection and segmentation with examples.

Image classification involves assigning a label to an entire image, such as categorizing it as a cat or a dog. Object detection goes a step further, identifying and localizing multiple objects within an image, providing bounding boxes and labels for each. Segmentation, on the other hand, precisely delineates object boundaries by assigning a label to each pixel, separating foreground from background.

e. Explain the concept of a radar chart in data visualization. Discuss the potential use cases of using a radar chart to display data.

A radar chart, also known as a spider or star chart, is a data visualization tool that displays multivariate data in a two-dimensional plane with three or more quantitative variables represented on axes starting from the same point. It is particularly useful for showcasing patterns and relationships between variables, making it suitable for comparing performance across multiple dimensions simultaneously.

Problem 2: **10 Marks**

Calculate the output dimensions and number of trainable parameters for each layer and fill in the table.

```
=================================================================
 conv2d_5 (Conv2D)           (None, 55, 55, 96)        11712

 max_pooling2d_3 (MaxPoolin  (None, 27, 27, 96)        0
 g2D)

 conv2d_6 (Conv2D)           (None, 27, 27, 256)       614656

 max_pooling2d_4 (MaxPoolin  (None, 13, 13, 256)       0
 g2D)

 conv2d_7 (Conv2D)           (None, 13, 13, 384)       885120

 conv2d_8 (Conv2D)           (None, 13, 13, 384)       1327488

 conv2d_9 (Conv2D)           (None, 13, 13, 256)       884992

 max_pooling2d_5 (MaxPoolin  (None, 6, 6, 256)         0
 g2D)

 dropout_2 (Dropout)         (None, 6, 6, 256)         0

 flatten_1 (Flatten)         (None, 9216)              0

 dense_3 (Dense)             (None, 4096)              37752832

 dropout_3 (Dropout)         (None, 4096)              0

 dense_4 (Dense)             (None, 4096)              16781312

 dense_5 (Dense)             (None, 1000)              4097000

=================================================================
Total params: 62355112 (237.87 MB)
Trainable params: 62355112 (237.87 MB)
Non-trainable params: 0 (0.00 Byte)
```

Roll Number: _____

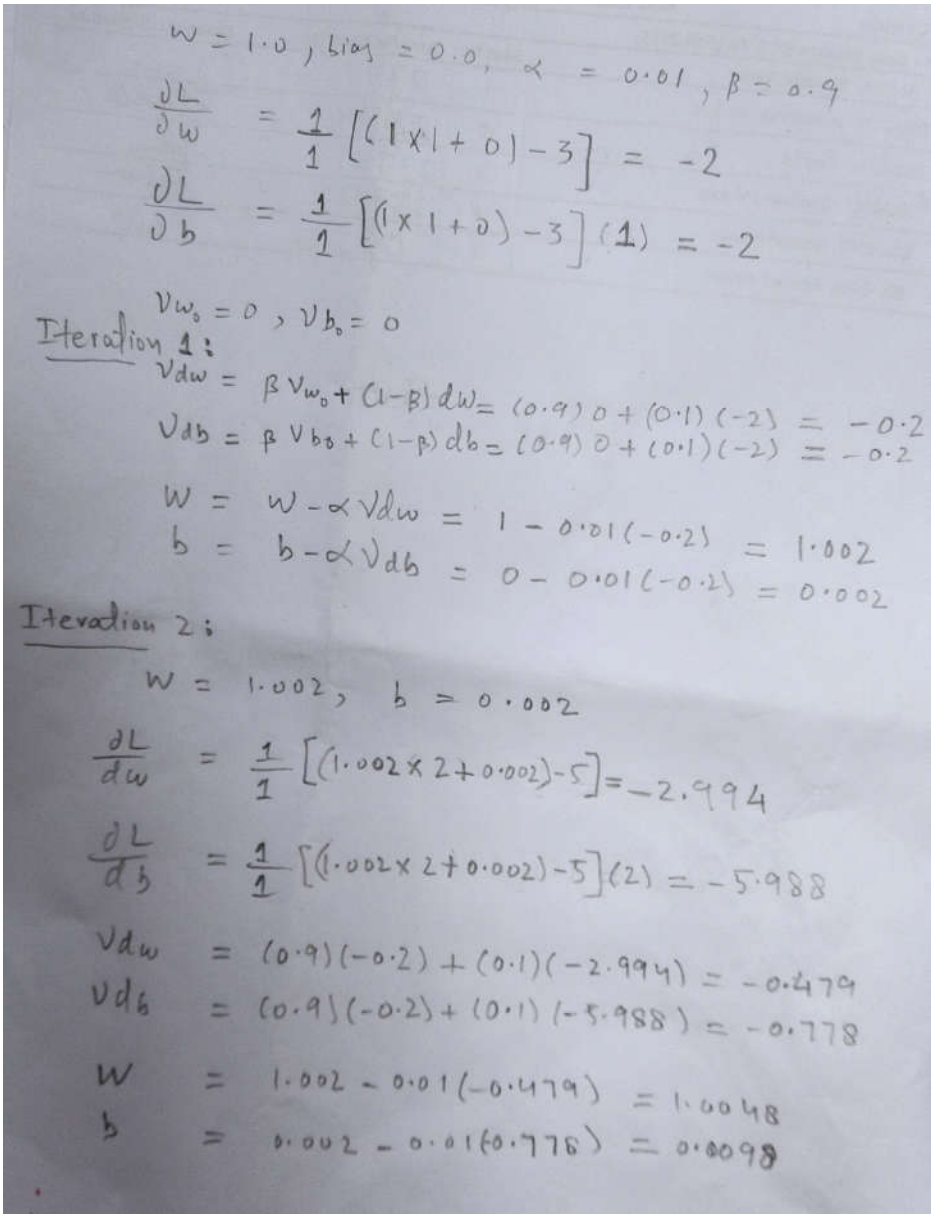
Problem 3: **10 Marks**

Apply Stochastic Gradient Descent (SGD) with momentum on the given dataset and calculate the updated parameters for each iteration . In this example, we'll use a linear regression problem with one independent feature (X). Our goal is to minimize the MSE loss. Perform 1 epoch and show updated parameters after each iteration.

	1	**2**	**3**	**4**	**5**
X	1	2	3	5	7
Y	3	5	7	10	13
Parameters	Weight (W) = 1.0, Bias (b) = 0.0, $\alpha = 0.01$, $\beta = 0.9$				

$$\frac{\partial L}{\partial W} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \qquad \frac{\partial L}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)}$$

Roll Number: _____



$w = 1.0$, bias $= 0.0$, $\alpha = 0.01$, $\beta = 0.9$

$$\frac{\partial L}{\partial w} = \frac{1}{1}[(1 \times 1 + 0) - 3] = -2$$

$$\frac{\partial L}{\partial b} = \frac{1}{1}[(1 \times 1 + 0) - 3](1) = -2$$

$v_{w_0} = 0$, $v_{b_0} = 0$

Iteration 1:

$v_{dw} = \beta v_{w_0} + (1-\beta) dw = (0.9)0 + (0.1)(-2) = -0.2$

$v_{db} = \beta v_{b_0} + (1-\beta) db = (0.9)0 + (0.1)(-2) = -0.2$

$w = w - \alpha v_{dw} = 1 - 0.01(-0.2) = 1.002$

$b = b - \alpha v_{db} = 0 - 0.01(-0.2) = 0.002$

Iteration 2:

$w = 1.002$, $b = 0.002$

$$\frac{dL}{dw} = \frac{1}{1}[(1.002 \times 2 + 0.002) - 5] = -2.994$$

$$\frac{dL}{db} = \frac{1}{1}[(1.002 \times 2 + 0.002) - 5](2) = -5.988$$

$v_{dw} = (0.9)(-0.2) + (0.1)(-2.994) = -0.479$

$v_{db} = (0.9)(-0.2) + (0.1)(-5.988) = -0.778$

$w = 1.002 - 0.01(-0.479) = 1.0048$

$b = 0.002 - 0.01(0.778) = 0.0098$

Roll Number: _____

Problem 4: **10 Marks**

a. Using the given input image of 6x6 and a 3x3 filter, compute output feature map. Apply 0 padding on borders so that output feature map has the same dimensions. Apply RELU activation and show the resultant matrix. (4+1)

0	0	0	0	0	0
0	1	1	1	1	0
0	1	0	0	1	0
0	1	0	0	1	0
0	1	1	1	1	0
0	0	0	0	0	0

-1		-1	-1		-1
-2	1	-2	-2	1	-2
-3	2	-3	-3	2	-3
-3	2	-3	-3	2	-3
-2	1	-2	-2	1	-2
-1		-1	-1		-1

	1			1	
	2			2	
	2			2	
	1			1	

-1	1	-1
-1	1	-1
-1	1	-1

b. Apply a 2x2 max pooling filter on the resultant feature map with stride=2 and show the output matrix. (2)

1 0 1
2 0 2
1 0 1

c. Flatten the resultant output from the MaxPool and pass it to a output layer with one neuron. Weights for output layer are -0.1 for all, bias=0 and activation is sigmoid. Calculate the output value. (3)

Calculate the weighted sum in the output layer:
Weighted Sum=$(-0.1\times1)+(-0.1\times0)+(-0.1\times1)+(-0.1\times2)+(-0.1\times0)+(-0.1\times2)+(-0.1\times1)+(-0.1\times0)+(-0.1\times1)+0$

Weighted Sum= **-0.8**

Apply the sigmoid activation function $=1/1+e-(-0.8)$

Output≈0.31

-

Roll Number: _____

Problem 5: **10 Marks**

Astronomical researchers aim to predict whether celestial bodies are hazardous or not based on observed features. Two independent features, Feature_1 and Feature_2, are considered for the prediction. The researchers have devised a predictive model represented by the equation:

Prediction = −1.5 + 0.8 × Feature1 + 1.2 × Feature2

The prediction is binary: 1 if the celestial body is predicted to be hazardous, and 0 otherwise. The decision boundary is set at a threshold of 5.

a) Calculate the predicted value for the remaining records and then fill the predicted label column.

Feature_1	Feature_2	Hazardous (Actual Label)	Prediction	Prediction (Threshold = 8)
4.2	7.1	0	10.38	1
9.5	3.8	1	10.66	1
2.8	8.6	0	10.06	1
6.1	4.5	0	8.86	1
7.3	2.1	0	4.66	0
3.7	9.2	1	**12.5**	**1**
1.9	5.1	0	**6.14**	**0**
5.4	6.3	1	**10.38**	**1**
8.6	1.7	0	**7.42**	**0**
3.0	7.8	0	**10.26**	**1**

b) Create a confusion matrix using the predicted labels computed above. Calculate Accuracy. Sensitivity and Specificity. Interpret the results.

True Positives (TP) = 3

True Negatives (TN) = 3

False Positives (FP) = 4

False Negatives (FN) = 0

	Actual	
	P	N
P	3	4
N	0	3

Accuracy= 60, Sensitivity= 100, Specificity= 43

The model demonstrates a moderate overall accuracy of 60% however, it effectively identifies all positive cases but struggling badly in predicting negative cases.