United International University

B.Sc. in Data Science

DS 3885: Data Wrangling

Final Examination: Spring 2025 Time: 2 Hours Marks: 30

Any examinee found adopting unfair means will be expelled from the trimester/program as per UIU disciplinary rules.

1. (a) What is the difference between *normalization* and *standardization* in data pre-processing? Briefly explain when each method is preferred. [2]

   (b) Given the following data points for a feature X:

   $$X = [10,\ 20,\ 30,\ 40,\ 50]$$

   Standardize the data using z-score normalization. Show your calculation steps. [3]

2. (a) Briefly explain the difference between the Bag of Words (BoW) model and the TF-IDF method in text feature extraction. Which one helps reduce the influence of common but less informative words? [2]

   (b) You are given the following two documents:

   - Doc1: "Data wrangling is an important step in data science."
   - Doc2: "Cleaning and preparing data are key tasks in data wrangling."

   (a) Construct a combined vocabulary from both documents (ignore case and punctuation). [2]

   (b) Using the frequency-based Bag-of-Words model, represent each document as a vector based on the vocabulary. [2]

   (c) Compute the TF-IDF score of the word "science" in Doc1. Show the formula, calculation, and provide a brief interpretation. [2]

   (d) Suppose you are tasked with detecting near-duplicate text entries in a dataset. Using the frequency-based Bag-of-Words vectors constructed in part (b) for Doc1 and Doc2:

   (i) Compute the Cosine Similarity between Doc1 and Doc2. Show all necessary steps. [2]

   (ii) Compute the Jaccard Similarity using the sets of unique words only, ignoring word frequency. [2]

Based on your results, which similarity measure do you think better captures the semantic overlap in this case? [1]

3. (a) Briefly explain the role of the following components in extracting features from an image:

   i. Convolution layer – how does it help detect patterns? [1]

   ii. Padding – how does it affect the output feature map and edge information? [1]

   (b) Given the following 4×4 feature map from a convolution layer:

$$\begin{bmatrix} -2 & 4 & 1 & -1 \\ 6 & -3 & 2 & 0 \\ -1 & 5 & -2 & 3 \\ 2 & -4 & 6 & 1 \end{bmatrix}$$

   i. Apply the ReLU activation function to this feature map. [2]

   ii. Apply 2×2 max pooling with a stride of 2 on the activated feature map and write the resulting 2×2 output. [2]

4. A diagnostic AI model is used to classify patients into three health categories: Class 0 - No Risk, Class 1 - At Risk, and Class 2 - Critical Risk.

   The confusion matrix below summarizes the model's performance on a dataset of 360 patients:

	Predicted: 0	**Predicted: 1**	**Predicted: 2**
Actual: 0	80	10	10
Actual: 1	20	70	10
Actual: 2	0	30	130

   As a hospital data analyst, you are asked to evaluate how well the model identifies Critical Risk (Class 2) patients.

   - Calculate the values of True Positives, False Positives, False Negatives, and True Negatives for Class 2 when considered as the positive class. [2]
   - Compute the Precision and Recall for Class 2. Show the formula and values used. [2]
   - Considering that missing a Critical Risk patient is a serious concern, explain which metric (Precision or Recall) is more important and why. [2]