Lab Exercise – 7: Data Similarity - Dissimilarity & Handling Missing Values

Note:

* Prepare a PDF document and name the file as "Lab7_RegisterNo.pdf".
* PDF file should consist Question No, Code, and Result for each Question.
* File Should be headed with your Register number, Slot number, Lab Exercise number.

* * *

1. Develop User Defined functions to calculate dissimilarity matrices for Nominal attributes, Binary attributes, Numeric attributes, Ordinal attributes, Mixed attributes. Apply these functions on the following data:

Obj Id	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1	4	D3	0	24	T	Y	P	poor	A	14
2	5	D1	1	36	F	Y	N	average	B	16
3	6	D4	1	43	T	N	P	good	C	18
4	7	D6	0	13	T	N	N	average	E	12
5	8	D1	1	22	F	N	P	poor	B	21

Hint:

- All Binary attributes are symmetric.
- All Numeric attributes are in Euclidean space
- Use the following ordinal relationships in the following attributes:
  A8: poor < average < good **A9:** A < B < C < D < E **A2:** D6 < D5 < D4 < D3 < D2 < D1

2. Calculate Term Frequency Vector (Document Vectors) for the following documents and calculate Cosine Similarity between every pair of documents to identify which documents are more similar.

Data 1:

doc1 = "I want to start learning to charge something in life"
doc2 = "reading something about life no one else knows"
doc3 = "Never stop learning"
doc4 = "life learning"

Data 2:

doc1 = "Mr. Imran Khan win the president seat after winning the National election 2020-2021. Though he lost the support of some republican friends, Imran Khan is friends with President Nawaz Sharif"

doc2 = "President Imran Khan says Nawaz Sharif had no political interference is the election outcome. He claimed President Nawaz Sharif is a friend who had nothing to do with the election"

doc3 = "Post elections, Vladimir Nawaz Sharif win the president seat of Russia. President Nawaz Sharif had served as the Prime Minister earlier in his political career"

3.

a. Load ***crx.data*** into a data frame and do the following operations: (The data has no headers)

- Change the column names to A1 to A16
- Replace all '?' marks with np.nan
- Convert A2 and A14 attributes to float data type
- Convert '+' to 1 and "–" to 0 of A16 attribute
- Replace values of "A3, A8, A9, A10" attributes to np.nan in 50 random objects
- Save the file as `Transformed_crx.csv`

b. Ignoring missing values:

- Load the Credit Approval Data Set `Transformed_crx.csv`
- Calculate the percentage of missing values for each variable and sort them in ascending order
- Remove the observations with missing data in any of the variables
- Print and compare the size of the original and complete case datasets

c. Performing mean and median imputation:

- Load the Credit Approval Data Set `Transformed_crx.csv`
- Replace the missing values with the median in five numerical variables 'A2', 'A3', 'A8', 'A11', 'A15' using pandas
- Replace the missing values with the mean in five numerical variables 'A2', 'A3', 'A8', 'A11', 'A15' using pandas
- Use SimpleImputer() of scikit-learn to fill the missing values with median and mean, separately.

d. Performing mode or frequent category imputation:

- Load the Credit Approval Data Set `Transformed_crx.csv`
- Replace the missing values with the mode in the attributes 'A4', 'A5', 'A6', 'A7' using pandas
- Use SimpleImputer() of scikit-learn to fill the missing values with mode.

e. Performing most probable value imputation: **[Optional]**

- Load the Credit Approval Data Set `Transformed_crx.csv`
- Replace the missing values with probable values using linear regression in the attributes 'A2', 'A3', 'A8', 'A11', 'A15' using pandas