Introduction to Data Science – Mid-Term Retake Exam (May 16th, 2024)

Name: **Practical Group:** (1) (2) (3) (4) **Neptun Code:**

Ex 1: You are given the following categorical dataset with features f_0, f_1, and f_2

f_0	f_1	f_2
c	b	x
a	a	z
c	c	y
a	a	y

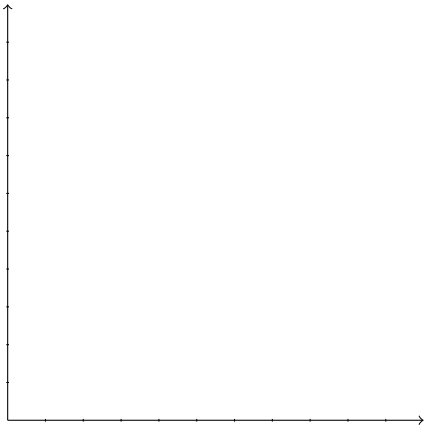
1. Provide the formal definition of Jaccard similarity.
2. What dataset do you obtain by applying 1-hot encoding to all of its features? Specify the column names in the form "attribute_value".
3. Compute J(x_1, x_2) and J(x_3, x_4) after one-hot encoding.
4. Define the Jaccard distance and prove that it can be used as a metric.

Ex 2:

1. What are the initialization methods commonly used in K-means? Discuss their pros and cons.
2. Explain how K-means determines convergence and identifies when the algorithm has finished.

Ex 3: You are provided with a dataset containing information about the number of hours spent studying (X) and the corresponding scores achieved (Y) by a group of students in a particular exam. Your task is to perform a simple linear regression analysis on this dataset.

Hours Studied (X)	**Exam Score (Y)**
2.5	85
3.0	88
3.5	90
4.0	92
4.5	94



1. Calculate the mean, variance, and standard deviation of both X and Y. Reflect on the significance of these statistical measures in understanding the distribution of the data.
2. Plot a scatter plot of the data to visualize the relationship between hours studied and exam scores.
3. Fit a simple linear regression model to the dataset to predict exam scores based on the number of hours studied. Use the least squares method to estimate the regression coefficients β_0 and β_1.
4. Reflect on the interpretation of the regression coefficients β_0 and β_1. What do they represent in the context of this problem?
5. Use the fitted regression model to predict the exam score for a student who studies for 5 hours.

Good Luck!