Introduction to Data Science – Mid-Term Special Exam (May 7th, 2024)

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

Exercise No.	Points
1	
2	
3	
4	
Total:	20/20

Ex 1: General definitions and concepts

1. What are the key distinctions between supervised and unsupervised learning in ML? (**1 pt**)
2. What are the similarities and differences between Anscombe's Quartet and the Unstructured Quartet in statistical analysis? (**1 pt**)
3. What are the fundamental steps in an exploratory data analysis (EDA) workflow, and how does each step contribute to gaining insights into the data's characteristics? (**1 pt**)

Ex 2: Similarity between objects

Let's consider two vectors: $A = (3, -2, 5)$ and $B = (1, 4, -1)$

1. Provide the formal definition of L_1 norm and calculate the L_1 norm between A and B (**2 pts**)
2. Under which conditions a distance measure is considered as a metric? Verify that the L_1 norm is a metric (**provide proofs to support each condition**). (**3 pts**)

Ex 3: Unsupervised learning - Clustering

1. What are the hyper-parameters of DBSCAN and how do we set them up? (**2 pts**)
2. How do the complete and single linkage methods differ in agglomerative clustering, and how does this difference affect the formation and structure of clusters in the resulting dendrogram? (**2 pts**)
3. How does the selection of initial centroids affect the convergence and final clustering results in K-means, and what techniques can be used to minimize this influence? (**2 pts**)

Ex 4: Supervised learning - Regression

1. Describe the objective function of a simple linear regression and explain how the parameters are chosen to optimize its objective function. (**2 pts**)
2. Explain the concept of least squares estimation and its significance in linear regression modeling. (**2 pts**)
3. Describe polynomial regression and discuss its advantages and limitations compared to linear regression models. (**2 pts**)

Good Luck!