**Experiment No.5**

**Title:** Applying similarity measures on the numeric datasets

**Batch: A4 Roll No.: 1914078 Experiment No.: 5**

**Aim:** Applying similarity measures on the numeric datasets

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed:** Any programming language, any data source (RDBMS/Excel/CSV)

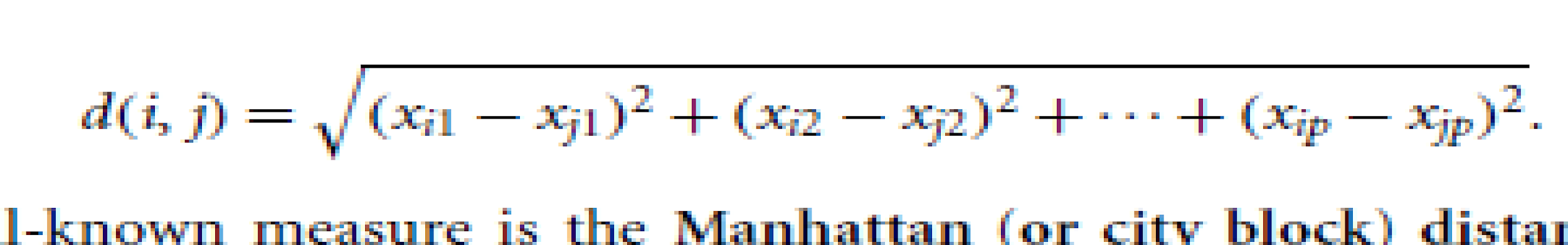
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

**Similarity measures:**

Similarity measures for numeric attributes include the *Euclidean, Manhattan*, and *Minkowski distances*.

The most popular distance measure is Euclidean distance (i.e., straight line or “as the crow flies”).  be two objects described by *p* numeric attributes. The Euclidean distance between objects *i* and *j* is defined as,

……………….…..……..(1)

Another well-known measure is the Manhattan (or city block) distance, named so because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks). It is defined as,

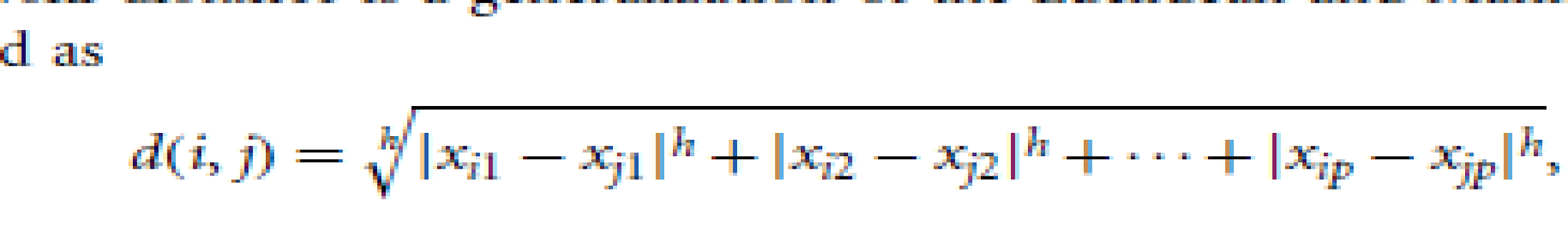
……………..........................(2)

Both the Euclidean and the Manhattan distance satisfy the following mathematical properties:

Non-negativity: Distance is a non-negative number.

Identity of indiscernible: The distance of an object to itself is 0.

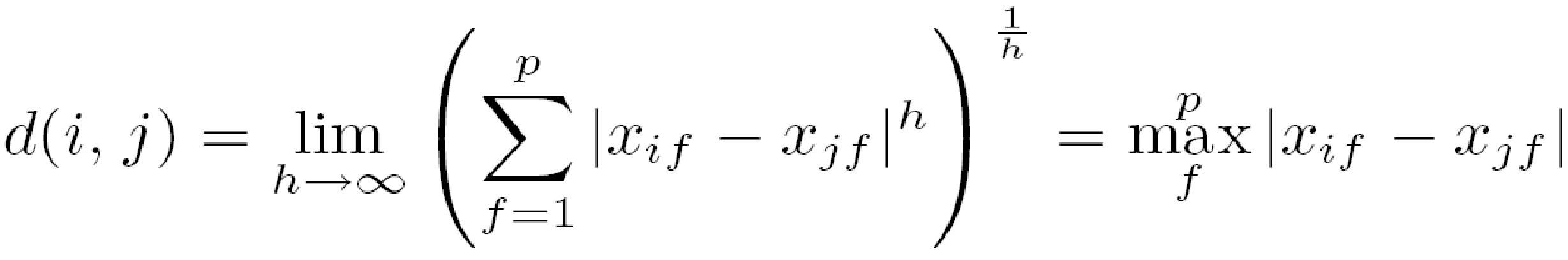
Minkowski distance is a generalization of the Euclidean and Manhattan distances. It is defined as,

………………………………….(3)

Where *h* is a real number such that *h* >= 1. It represents the Manhattan distance when *h* = 1 and Euclidean distance when *h* = 2.

*When**h* → ∞., its a “supremum” (Lmax norm, L∞ norm) distance.

* + This is the maximum difference between any component (attribute) of the vectors

****

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

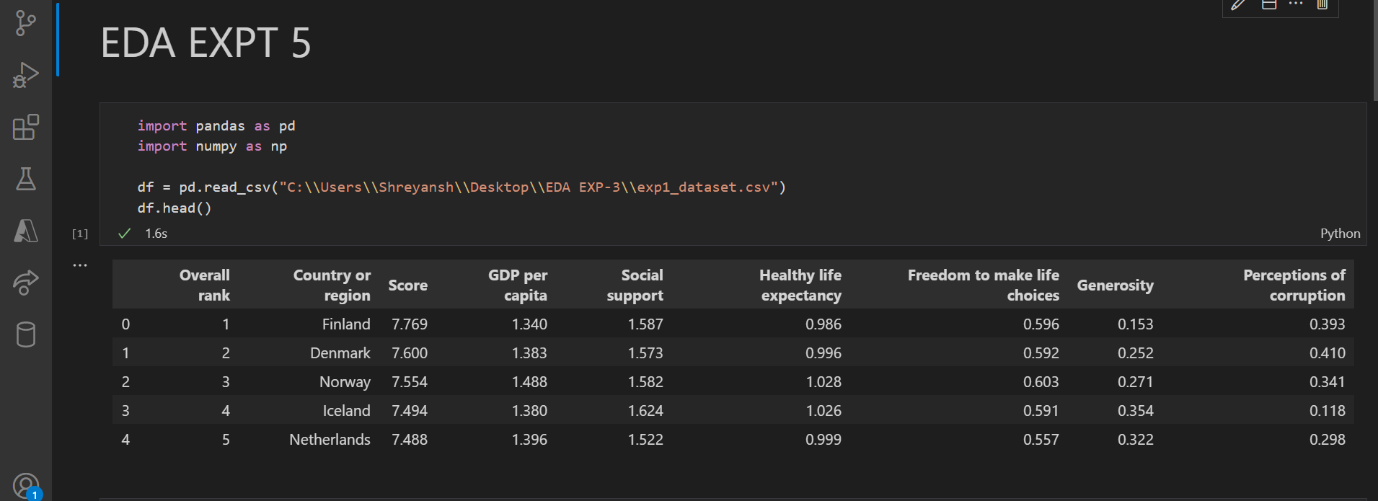
**Procedure / Approach /Algorithm / Activity Diagram:**

1. Identify the suitable attributes to apply the numeric similarity measures and write python code to calculate Euclidean, Manhattan similarity measures on it.
2. Identify the suitable attributes to apply the textual similarity measures and write python code to calculate Longest common subsequence, edit distance similarity measures on it.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

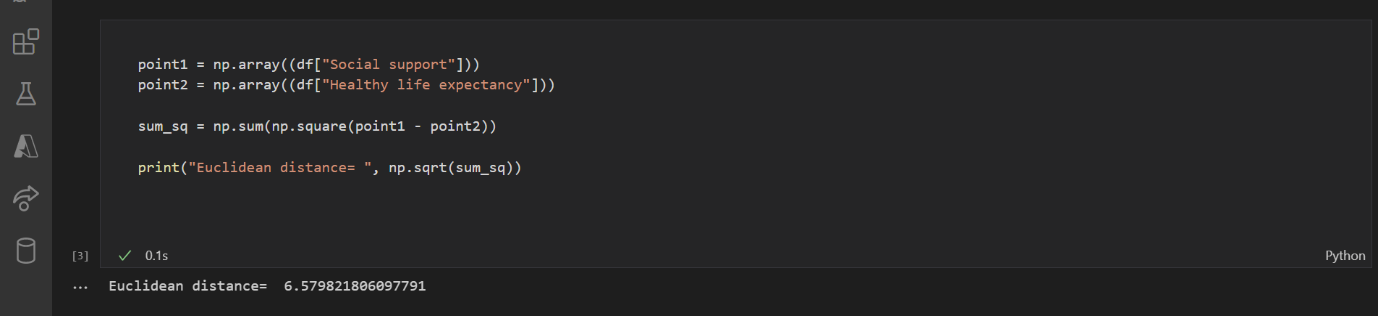
**Results: (Program printout with output / Document printout as per the format)**

**Importing required libraries and reading the dataset**

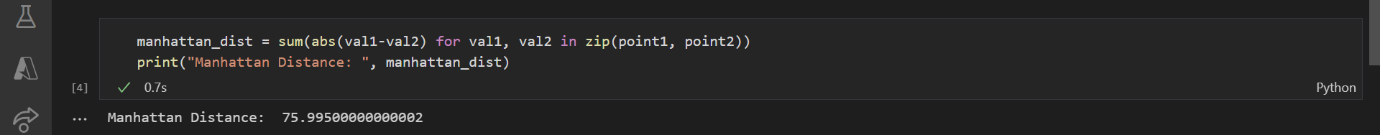


**Numeric Similarity:**

**Euclidean distance:**

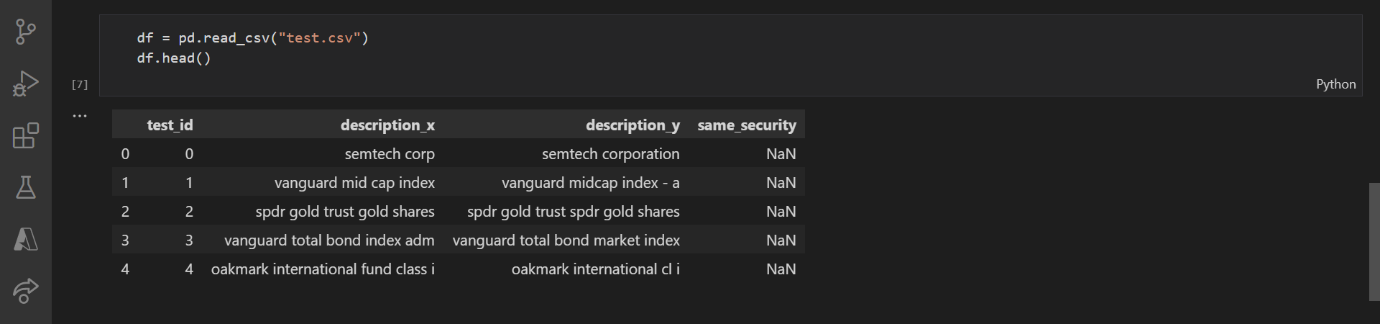


**Manhattan Distance:**



**Text similarity:**

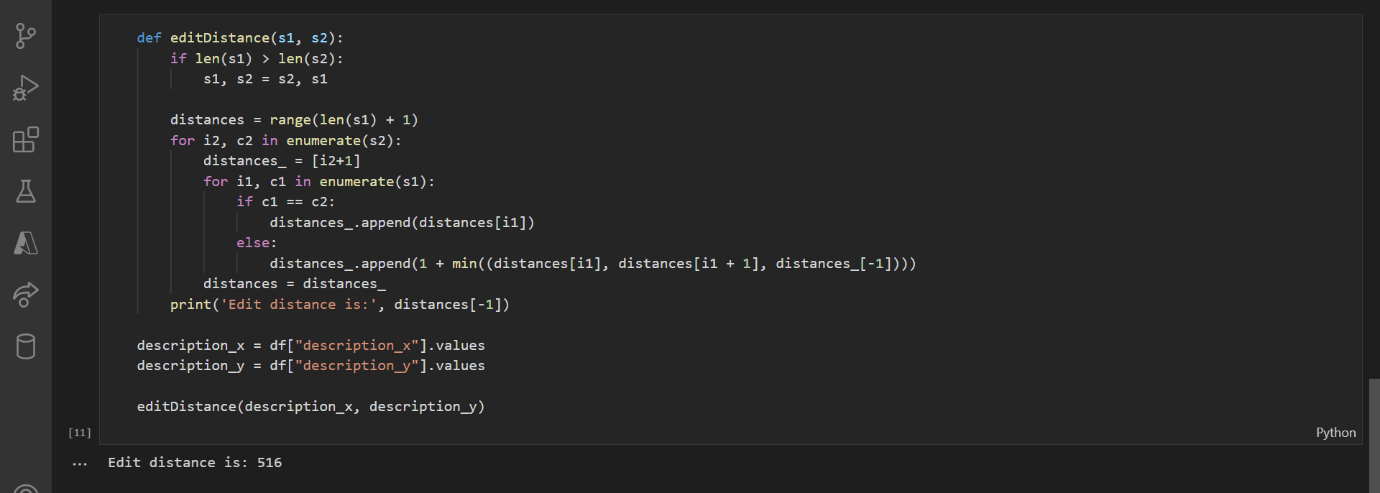
**Reading dataset for text similarity:**



**Average Length of Longest Common Subsequence:**



**Edit distance:**



**Questions:**

1. **What are the different applications of Numeric similarity measure?**

Similarity measures are used to identify:

* Duplicate data that may have differences due to typos
* Equivalent instances from different data sets
* Groups of data that are very close (clusters)

1. **What are the different applications of finding similarity between textual attributes?**

Many real-world applications make use of similarity measures to see how two objects are related together. We can use these measures in the applications involving Computer vision and Natural Language Processing, for example, to find and map similar documents. One important use case here for the business would be to match resumes with the Job Description saving a considerable amount of time for the recruiter. Another important use case would be to segment different customers for marketing campaigns using the K Means Clustering algorithm which also uses similarity measures. It is used for checking similarity between 2 documents, so it can be used to detect plagiarism. It can be used in auto correction tools and recommendation systems.

**Outcomes:**

CO2: Comprehend descriptive and proximity measures of data

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

Thus, calculated the numeric similarity measures like manhattan distance and Euclidean distance for Student Marks dataset and calculated the text similarity measures for the appropriate dataset without using predefined functions/ libraries in python.

**Grade: AA / AB / BB / BC / CC / CD /DD**

Signature of faculty in-charge with date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**References:**

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
2. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.