Problem

Consider the following entry set:

1	7	-2	4.5
2	3	-5	2.3
1	6	-2	1.2
2	5	-2	4.5
1	6	-5	2.3

- Each line represents a pattern
- Each column represents a feature

Requirements:

a) Read the entry set from a file saved on your local drive (in.txt) The values on each line will be separated by space. Handle exceptions that may occur

b) Calculate the Euclidian distance (generalized form - using all the feature columns) between all the patterns and write in a new file the distance matrix.(the size of the distance matrix will be: n x n). Truncate the distance matrix values to have only two decimals in the output file

For calculating the distance matrix values, you can use the following formula:

$$d[x, y] = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

where:
d[x][y] - is the distance between pattern x and pattern y;
d[x][x] = 0.00 – the distance from a patter to itself is 0
d[x][y] = d[y][x]
0 <= x,y < n (n is the total number of patterns);
p – the total number of features

Example:
Consider the following entry set

4	5	1	2
2	-1	1.6	3
1	3	4	1
0	6	-1.5	3
3	-2.1	3	4

The Euclidian generalized distance between the first two patterns is :

$d[0][1] = \sqrt{(4-2)^2 + (5-(-1))^2 + (1-1.6)^2 + (2-3)^2}$ = 6.4311

please note that d[0][1] = d[1][0]

using similar calculations for d[0][2], d[0][3], d[0][4], d[1][2], d[1][3], d[1][4], d[2][3], d[2,4], d[2,3] the distance matrix will be:

distance	pattern 0	pattern 1	pattern 2	pattern 3	pattern 4
pattern 0	0.00	**6.43**	4.79	4.92	7.70
pattern 1	**6.43**	0.00	5.17	7.91	2.27
pattern 2	4.79	5.17	0.00	6.65	6.32
pattern 3	4.92	7.91	6.65	0.00	9.79
pattern 4	7.70	2.27	6.32	9.79	0.00

c) Add the following column to the initial entry set:

1	7	-2	4.5	1
2	3	-5	2.3	2
1	6	-2	1.2	2
2	5	-2	4.5	1
1	6	-5	2.3	

The new column represents the class of each pattern

Important note: Please note that there is no value for the last pattern. This is because we will have to calculate it using the 1-NN rule

d) Calculate and display the last pattern class using the 1-NN rule and the distance matrix you obtained in section b)

What is k-NN rule ?

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification. The input consists of the k closest training examples in the feature space. (Wikipedia)

1-NN is a particular case where we take into consideration only the first nearest neighbor. In this case, the pattern will take the class of its nearest neighbor.

Example:
Consider the following entry set with the following classes

Pattern 0:	4	5	1	2	1
Pattern 1:	2	-1	1.6	3	2
Pattern 2:	1	3	4	1	2
Pattern 3:	0	6	-1.5	3	1
Pattern 4:	3	-2.1	3	4	?

The class of pattern 5 is the class of its nearest neighbor. Considering that we have previously calculated the distance matrix, we can easily find the closest neighbor for pattern4 by identifying the minimum distance to pattern4 in the distanceMatrix :

$$minimumDistance_4 = \min_{0 \le i < 4}(distanceMatrix[4][i] \ )$$

In our case, the minimum distance is 2.27 and the closest neighbor is pattern1. Considering this, pattern4 takes the class of pattern1 which is **2**