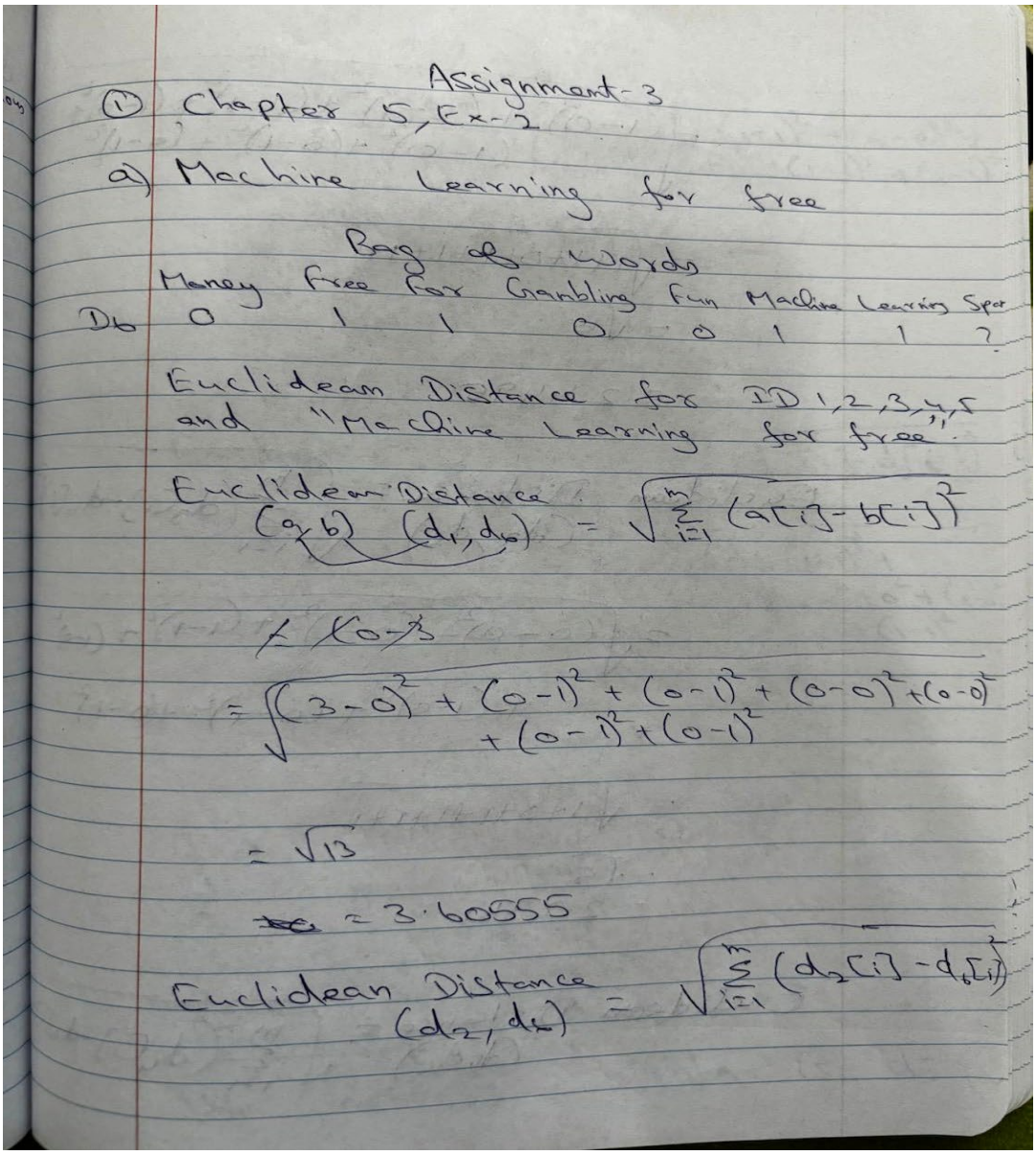
Assignment-3-Aparna Bharathi Suresh

1.(15%) Chapter 5, exercise 2

Question:a)

What target level would a nearest neighbor model using Euclidean distance return for the following email: “machine learning for free”?

Solution:



$$= \sqrt{(1-0)^2 + (2-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2}$$

$$= \sqrt{1+1+0+1+1+1+1}$$

$$= \sqrt{6}$$

$$= 2.4495$$

Euclidean Distance (d_3, d_6) $= \sqrt{\sum_{i=1}^{n} (d_3[i] - d_6[i])^2}$

$$= \sqrt{(0-0)^2 + (0-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2}$$

$$= \sqrt{1+0+1+1+1+1}$$

$$= \sqrt{5}$$

$$= 2.2361$$

Euclidean Distance (d_4, d_6) $= \sqrt{\sum_{i=1}^{n} (d_4[i] - d_6[i])$

$$= \sqrt{(0-0)^2 + (0-1)^2 + (1-1)^2 + (0-0)^2 + (3-0)^2 + (1-1)^2 + (1-1)^2}$$

$$= \sqrt{1+0+9+0+0}$$

$$= \sqrt{10}$$

$$= 3.1623$$

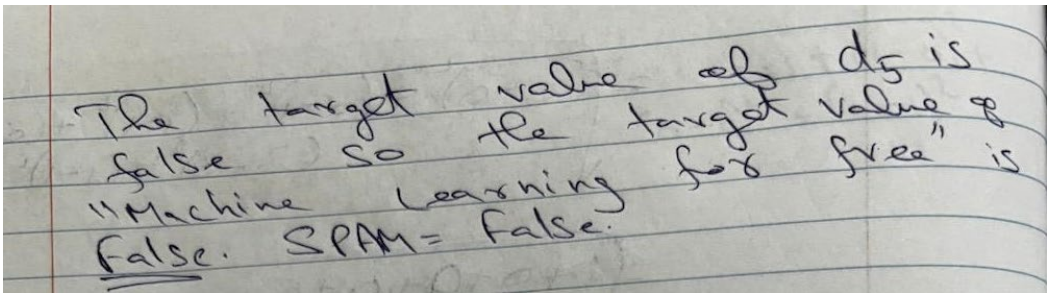
Euclidean Distance $(d_5, d_6) = \sqrt{\sum_{i=1}^{n} (d_5[i] - d_6[i])^2}$

$$= \sqrt{(0-0)^2 + (1-1)^2 + (0-1)^2 + 0^2 + 0^2 + (1-1)^2 + (1-1)^2}$$

$$= 1$$

ID	Euclidean Distance
1	3.6055
2	2.4495
3	2.2361
4	3.1623
5	1

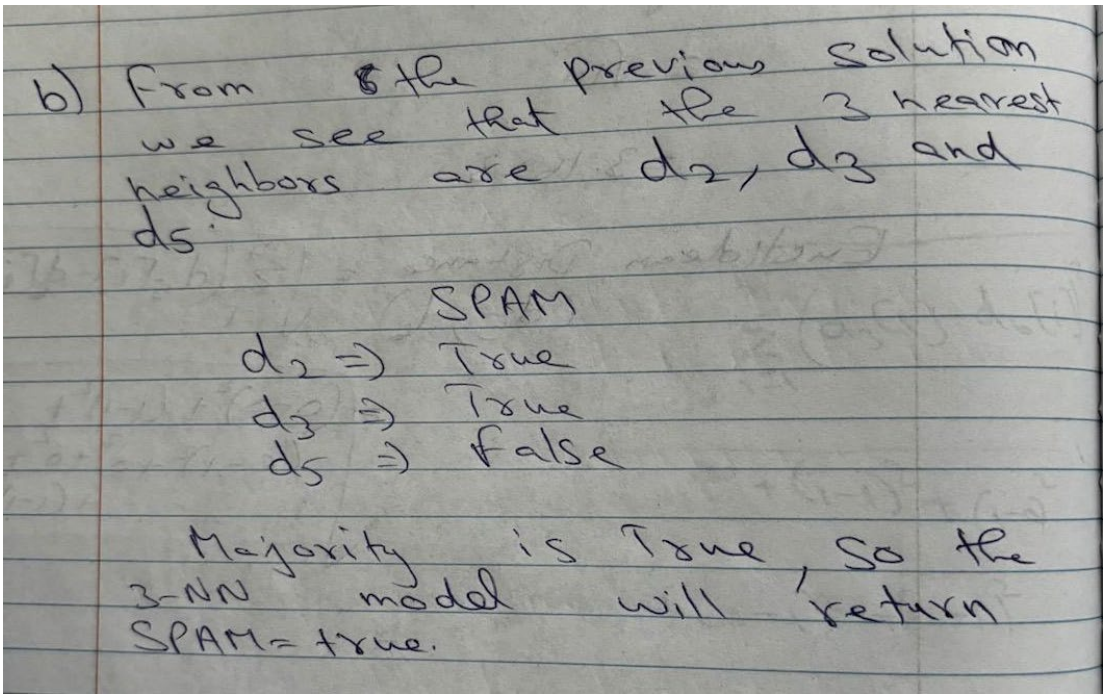
The nearest neighbor is d_5.



Question: b)

What target level would a k-NN model with k _ 3 and using Euclidean distance return for the same query?

Solution:



Question: c)

What target level would a weighted k-NN model with k _ 5 and using a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query, return for the query?

Solution:

c) Weighted k-NN ; k=5

ID	Euclidean Distance
1	3.6055
2	2.4495
3	2.2361
4	3.1623
5	1

Weighted KNN = $\frac{1}{(\text{Euclidean Distance})^2}$

ID	Weighted KNN	
1	$1/(3.6055)^2$	= 0.0769
2	$1/(2.4495)^2$	= 0.1667
3	$1/(2.2361)^2$	= 0.2
4	$1/(3.1623)^2$	= 0.1
5	$1/1^2$	= 1

Total weight of True: $d_1 + d_2 + d_3$

$= 0.0769 + 0.1667 + 0.2$

$= 0.4436$

Weight of False = $d_4 + d_5$

$= 0.1 + 1$

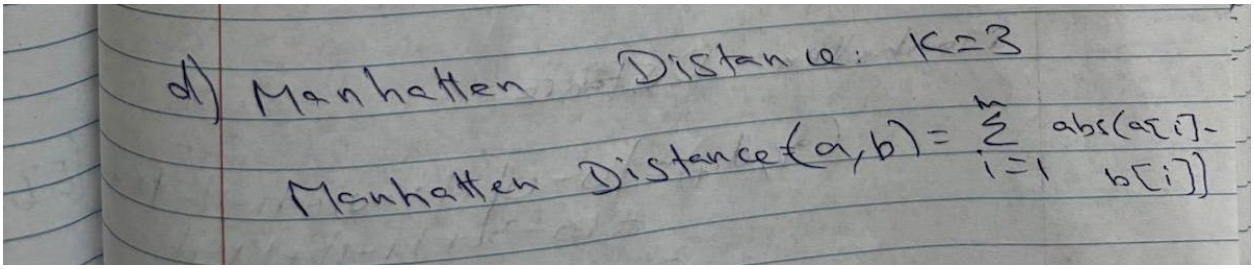
$= 1.1$

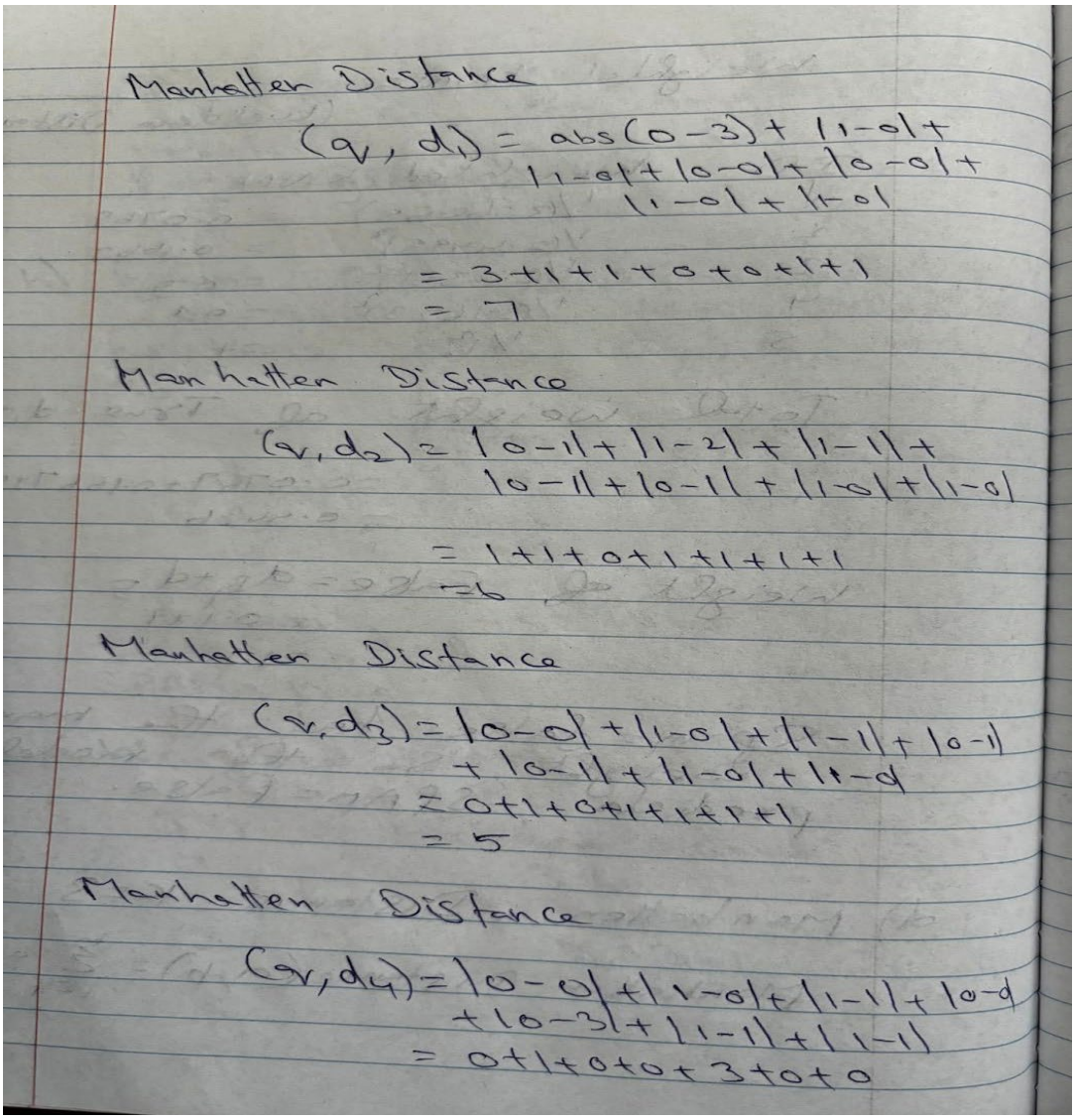
Spam = false has the max. weight, So the model predicts, SPAM = False.

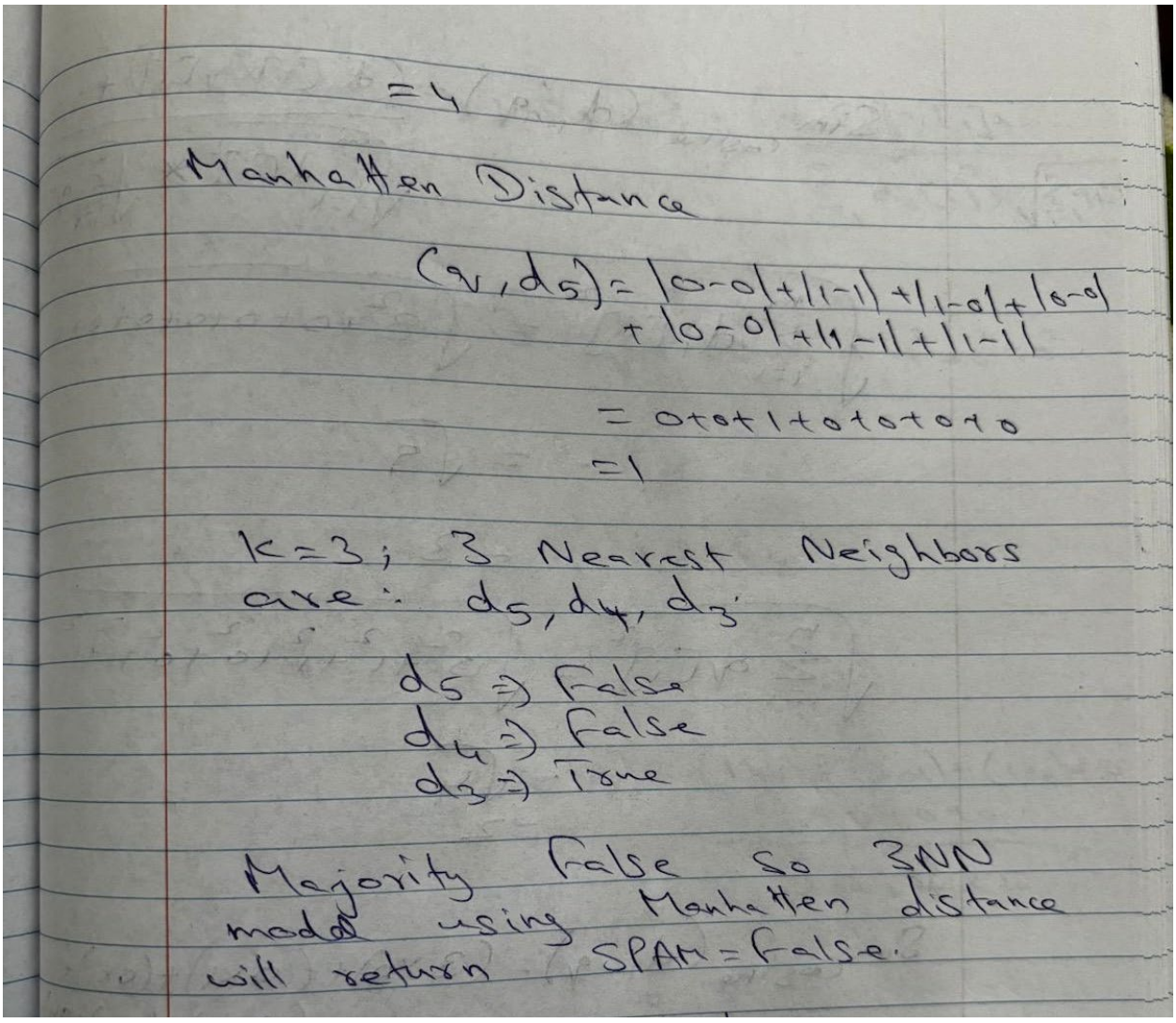
Question: d)

What target level would a k-NN model with k _ 3 and using Manhattan distance return for the same query?

Solution:







Question: e)

There are a lot of zero entries in the spam bag-of-words dataset. This is indicative of sparse data and is typical for text analytics. Cosine similarity is often a good choice when dealing with sparse non-binary data. What target level would a 3-NN model using cosine similarity return for the query?

Solution:

e. Cosine Similarity:

$$Sim_{cosine}(a,b) = \frac{(a[1]\times b[1]) + \cdots (a[n]\times b[n])}{\sqrt{\sum_{i=1}^{n} a[i]^2} \times \sqrt{\sum_{i=1}^{n} b[i]^2}}$$

$$Sim_{cosine}(d_1,q) = \frac{(d_1[1]\times q[1]) + \cdots}{\sqrt{\sum_{i=1}^{m} d_1[i]^2} * \sqrt{\sum_{i=1}^{n} q[i]}}$$

$$\sqrt{\sum_{i=1}^{m} d_1[i]^2} = \sqrt{3^2+0+0+0+0+0+0+0}$$

$$= \sqrt{9}$$

$$= 3$$

$$\sqrt{\sum_{i=1}^{m} q[i]^2} = \sqrt{0^2+1^2+1^2+0^2+0^2+1^2+1^2}$$

$$= \sqrt{4}$$

$$= 2$$

$$Sim_{cosine}(d_1,q) = \frac{(3\times 0)+(0\times 1)+(0\times 1)+(0\times 0)+(0\times 0)+(0\times 1)+(0\times 1)}{3\times 2}$$

$$= \frac{0}{6}$$

$$= 0$$

$$\text{Sim}_{\text{cosine}}(d_2, q) = \frac{(d_2[i]\,q[i] + \cdots}{\sqrt{\sum_{i=1}^{n} d_2[i]^2} \times \sqrt{\sum_{i=1}^{n} q[i]^2}}$$

$$\sqrt{\sum_{i=1}^{m} d_2[i]^2} = \sqrt{1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2}$$

$$= \sqrt{8}$$

$$= 2.8284$$

$$\sqrt{\sum_{i=1}^{m} q[i]^2} = 2$$

$$\text{Sim}_{\text{cosine}}(d_2, q) = \frac{(1\times0) + (2\times1) + (1\times1) + 0 + 0 + 0 + 0}{2.8284 \times 2}$$

$$= \frac{3}{5.6568}$$

$$= 0.5303$$

$$\text{Sim}_{\text{cosine}}(d_3, q) = \frac{(d_3[i]\,q[i] + \cdots\cdots)}{\sqrt{\sum_{i=1}^{m} d_3[i]^2} \times \sqrt{\sum_{i=1}^{n} q[i]^2}}$$

$$\sqrt{\sum_{i=1}^{n} d_3[i]^2} = \sqrt{0+0+1^2+1^2+1^2+0+0}$$

$$= \sqrt{3}$$

$$= 1.732$$

$$\text{Sim}_{\text{cosine}}(d_3, q) = \frac{(0)+(0\times1)+(1\times1)+0+0+0+0}{1.732\times2}$$

$$= \frac{1}{3.4641}$$

$$= 0.2887$$

$$\text{Sim}_{\text{cosine}}(d_4, q) = \frac{(d_4[i]\times q[i]+\dots}{\sqrt{\sum_{i=1}^{n} d_4[i]^2} \times \sqrt{\sum_{i=1}^{n} q[i]^2}}$$

$$\sqrt{\sum_{i=1}^{n} d_4[i]^2} = \sqrt{0+0+1^2+0+3^2+1^2+1^2}$$

$$= \sqrt{12}$$

$$= 3.4641$$

$$\text{Sim}_{\text{cosine}}(d_4, q) = \frac{0+0+1+0+0+1+1}{3.4641 \times 2}$$

$$= \frac{3}{6.9282}$$

$$= 0.4330$$

$$\text{Sim}_{\text{cosine}}(d_5, q) = \frac{(d_5[1] \times q[1] + \cdots}{\sqrt{\sum_{i=1}^{m} d_5[i]^2} \times \sqrt{\sum_{i=1}^{m} q[i]^2}}$$

$$\sqrt{\sum_{i=1}^{m} d_5[i]^2} = \sqrt{0+1+0+0+0+1+1}$$

$$= \sqrt{3}$$

$$= 1.7320$$

$$\text{Sim}_{\text{cosine}}(d_5, q) = \frac{(0)+1+0+0+0+1+1}{1.7320 \times 2}$$

$$= \frac{3}{3.4641}$$

$$= 0.8660$$

In Similary Index, higher the number, the more similar the instances.

ID	Cosine Similary
1	0
2	0.5303
3	0.2887
4	0.4330
5	0.8660

So from the above d_5, d_2 and d_4 are the 3 most similar instances to the query.

$d_5 \Rightarrow$ False
$d_2 \Rightarrow$ True
$d_4 \Rightarrow$ False

Majority False hence the prediction of SPAM = False.

2. (15%) Chapter 5, exercise 3

Question: a)

What value would a 3-nearest neighbor prediction model using Euclidean distance return for the CPI of Russia?

Solution:

② Chapter - 5, Ex - 3

a) Euclidean Distance:

$$\text{Euclidean Distance} = \underset{a}{(\text{Russia}}, \underset{b}{\text{Afganistan}}) = \sqrt{\sum_{i=1}^{m} (a[i]-b[i])^2}$$

$$= \sqrt{(67.62-59.61)^2 + (31.68-23.24)^2 + (10-74.30)^2 + (3.87-4.44)^2 + (12.90-0.40)^2}$$

$$= 66.5354$$

$$\text{Euclidean Distance} = \underset{(\text{Russia, Haiti})}{} \sqrt{\sum_{i=1}^{m} (a[i]-b[i])^2}$$

$$= \sqrt{(67.62-45)^2 + (31.68-47.67)^2 + (10-73.10)^2 + (3.87-0.09)^2 + (12.90-3.40)^2}$$

$$= 69.6670$$

Euclidean Distance
(Russia, Nigeria) =

$$\sqrt{(67.62-51.30)^2+(31.68-38.23)^2+(10-82.60)^2+(3.87-1.07)^2+(12.90-4.10)^2}$$

$$= 75.26808$$

Euclidean Distance
(Russia, Egypt) =

~~$\sqrt{(67.62-51.30)^2+(31.68-}$~~

$$\sqrt{(67.62-70.48)^2+(31.68-26.58)^2+(10-19.60)^2+(3.87-1.86)^2+(12.90-5.30)^2}$$

$$= 13.7168$$

Euclidean Distance (Russia, Agentina) =

$$\sqrt{(67.62-75.77)^2+(31.68-32.30)^2+(10-13.30)^2+(3.87-0.76)^2+(12.90-10.10)^2}$$

$$= 9.7775$$

Euclidean Distance (Russia, china) =

$$\sqrt{(67.62-74.87)^2+(31.68-29.98)^2+(10-13.70)^2+(3.87-1.95)^2+(12.90-6.40)^2}$$

$$= 10.72748$$

Euclidean Distance (Russia, Brazil) =

$$\sqrt{(67.62-73.87)^2+(31.68-42.93)^2+(10-14.50)^2+(3.87-1.43)^2+(12.90-7.20)^2}$$

$$= 14.6801$$

Euclidean Distance (Russia, Israel) =

$$\sqrt{(67.62-81.30)^2+(31.68-28.80)^2+(10-3.60)^2+(3.87-6.77)^2+(12.90-12.50)^2}$$

$$= 15.6514$$

Euclidean Distance (Russia, USA) =

$$\sqrt{(67.62-78.51)^2+(31.68-29.85)^2+(10-6.30)^2+(3.87-4.72)^2+(12.90-13.70)^2}$$

$$= 11.7044$$

Euclidean Distance (Russia, Ireland) =

$$\sqrt{(67.62-80.15)^2+(31.68-27.23)^2+(10-3.50)^2+(3.87-0.60)^2+(12.90-11.50)^2}$$

$$= 15.2219$$

Euclidean Distance (Russia, U.K) =

$$\sqrt{(67.62-80.09)^2+(31.68-28.49)^2+(10-4.40)^2+(3.87-2.59)^2+(12.90-13.00)^2}$$

$$= 14.0955$$

Euclidean Distance (Russia, Germany) =

$$\sqrt{(67.62-80.24)^2+(31.68-22.07)^2+(10-3.50)^2+(3.87-1.31)^2+(12.90-12.00)^2}$$

$$= 17.3559$$

Euclidean Distance (Russia, Canada) =

$$\sqrt{(67.62-80.99)^2+(31.68-24.79)^2+(10-4.90)^2+(3.87-1.42)^2+(12.90-14.20)^2}$$

$$= 16.1223$$

Euclidean Distance (Russia, Australia) =

$$\sqrt{(67.62-82.09)^2+(31.68-25.40)^2+(10-4.20)^2+(3.87-1.86)^2+(12.90-11.50)^2}$$

$$= 16.9841$$

Euclidean Distance (Russia, Sweden) =

$$\sqrt{(67.62-81.43)^2+(31.68-22.18)^2+(10-2.40)^2+(3.87-1.27)+(12.90-12.80)^2}$$

$$= 18.5261$$

Euclidean Distance: (Russia, New Zealand) =

$$\sqrt{(67.62-80.67)^2+(31.68-27.81)^2+(10-4.90)^2+(3.87-1.13)^2+(12.90-12.30)^2}$$

$$= 14.804$$

Euclidean Distance order:

ID	Euclidean	CPI
Argentina	9.7775	2.9961
China	10.7248	3.6356
USA	11.7044	7.1357
Egypt	13.7168	2.8622
UK	14.0955	7.7751
Brazil	14.6801	3.7141
Newzealand	14.804	9.4627
Ireland	15.2219	7.5360
Israel	15.6514	5.8069
Canada	16.1283	8.6725
Australia	16.9841	8.8442
Germany	17.3559	8.0441
Sweden	18.5261	9.2985
Afganistan	66.5364	1.5171
Haiti	69.6670	1.7999
Nigeria	75.26808	2.4493

The 3 NN are Argentina China & USA.

CPI value will be the average of the neighbors.

$$\frac{2.9961 + 3.6356 + 7.1357}{3}$$

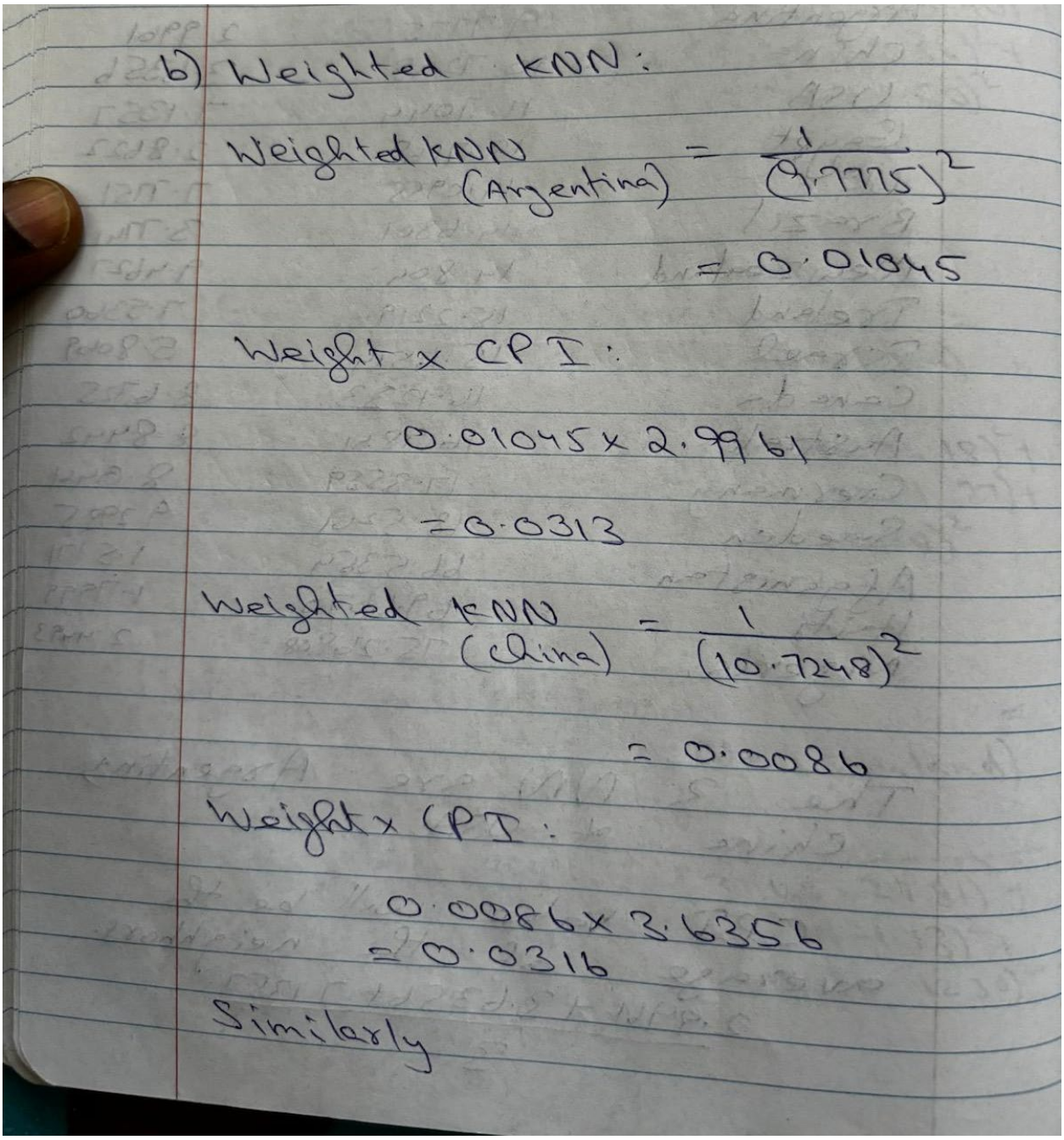
$= 4.5891$

CPI of Russia = 4.5891

Question: b)

What value would a weighted k-NN prediction model return for the CPI of Russia? Use k _ 16 (i.e., the full dataset) and a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query.

Solution:



ID	Euclidean	CPI	Weight	Weight x CPI
Argentina	9.7775	2.9961	0.01045	0.0313
china	10.7248	3.6356	0.0086	0.0316
USA	11.7044	7.1357	0.0072	0.0520
Egypt	13.7168	2.8622	0.0053	0.0152
UK	14.0955	7.7751	0.0050	0.0391
Brazil	14.6801	3.7741	0.0046	0.0172
New Zealand	14.804	9.4627	0.0045	0.043
Ireland	15.2219	7.5360	0.0043	0.0325
Israel	15.6514	5.8069	0.0040	0.0237
Canada	16.1223	8.6725	0.0038	0.0333
Australia	16.9841	8.8442	0.0035	0.0306
Germany	17.3559	8.0461	0.0033	0.0267
Sweden	18.5261	9.2985	0.0029	0.0270
Afganistan	66.5354	1.5171	0.0002	0.0003
Haiti	69.6670	1.7999	0.0002	0.0004
Nigeria	75.26808	2.4493	0.0002	0.0004
			0.06805	0.4043

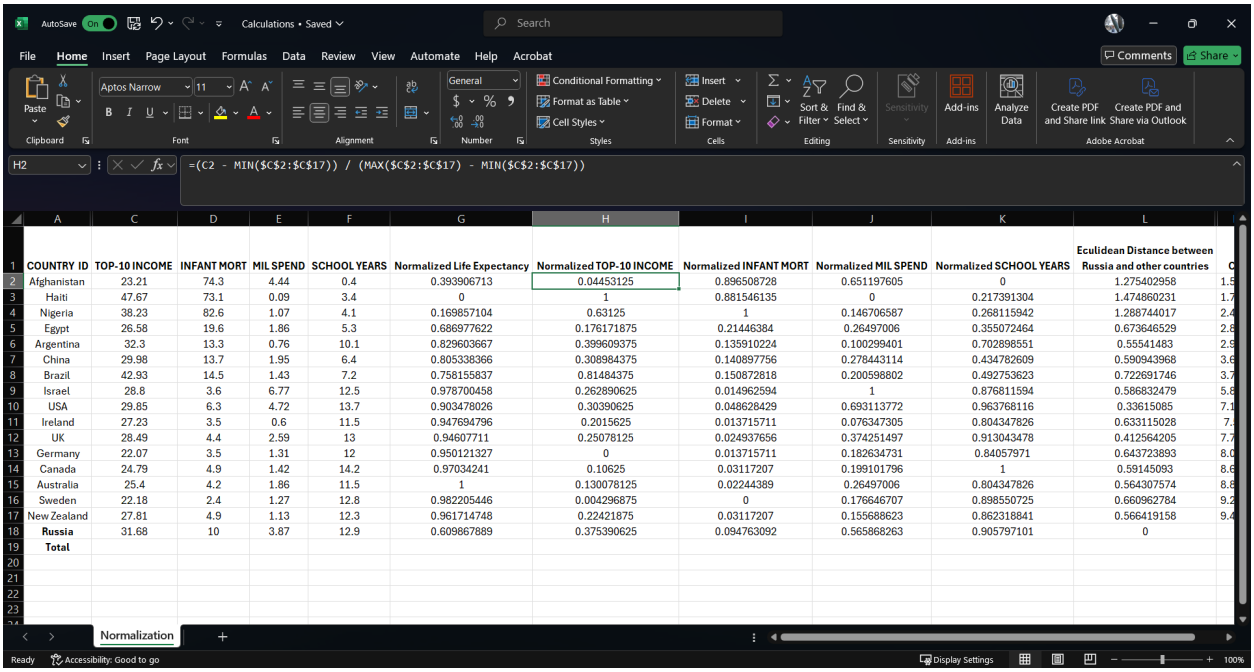
$$\frac{\Sigma \text{Weight} \times \text{CPI}}{\Sigma \text{weight}} = \frac{0.4043}{0.06805}$$

$$= 5.9412$$

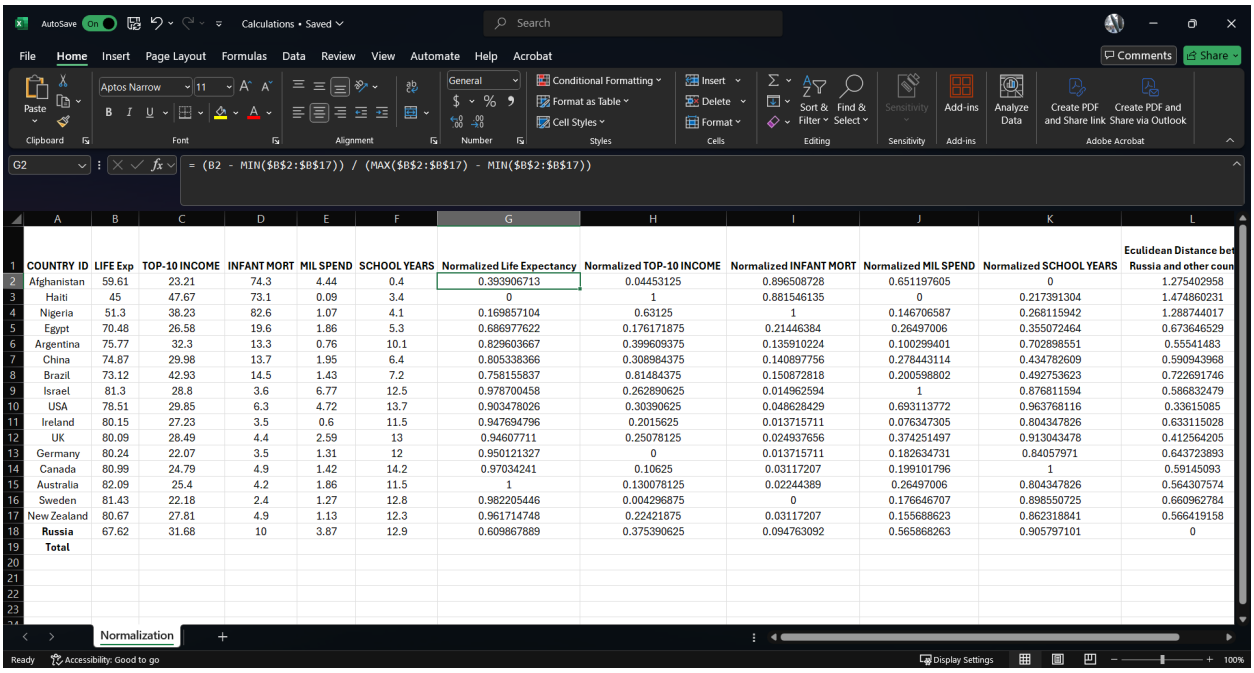
Question: c)

The descriptive features in this dataset are of different types. For example, some are percentages, others are measured in years, and others are measured in counts per 1;000. We should always consider normalizing our data, but it is particularly important to do this when the descriptive features are measured in different units. What value would a 3-nearest neighbor prediction model using Euclidean distance return for the CPI of Russia when the descriptive features have been normalized using range normalization?

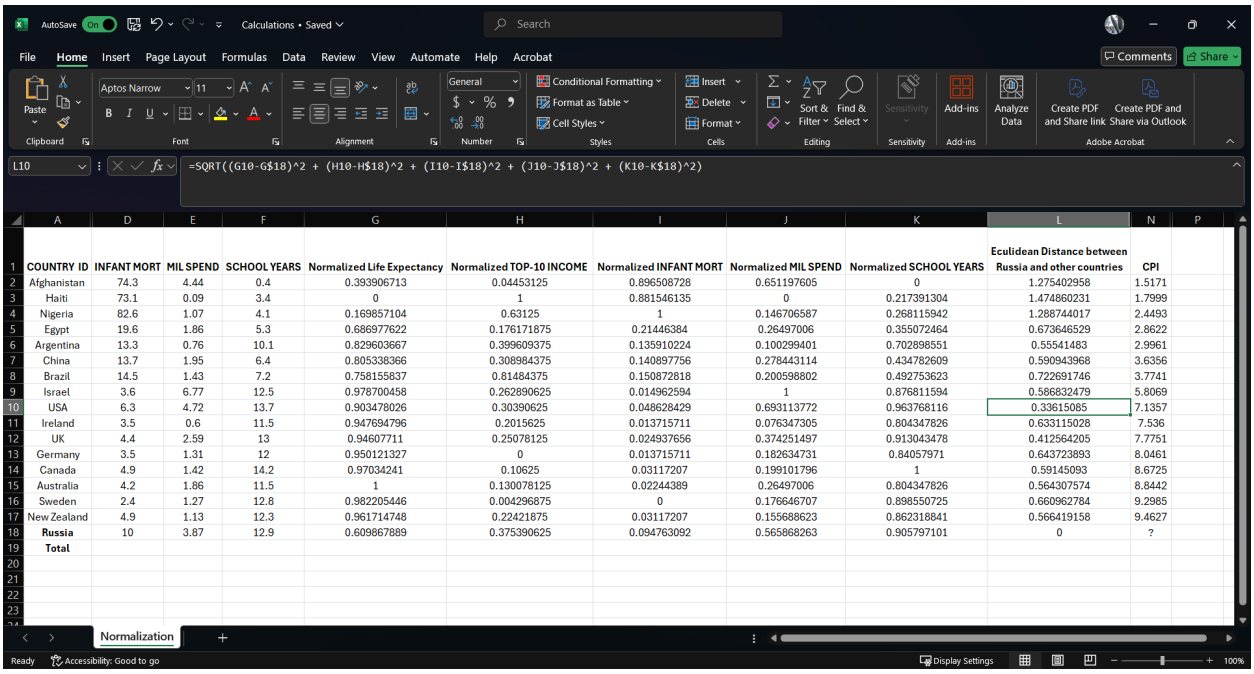
Solution:



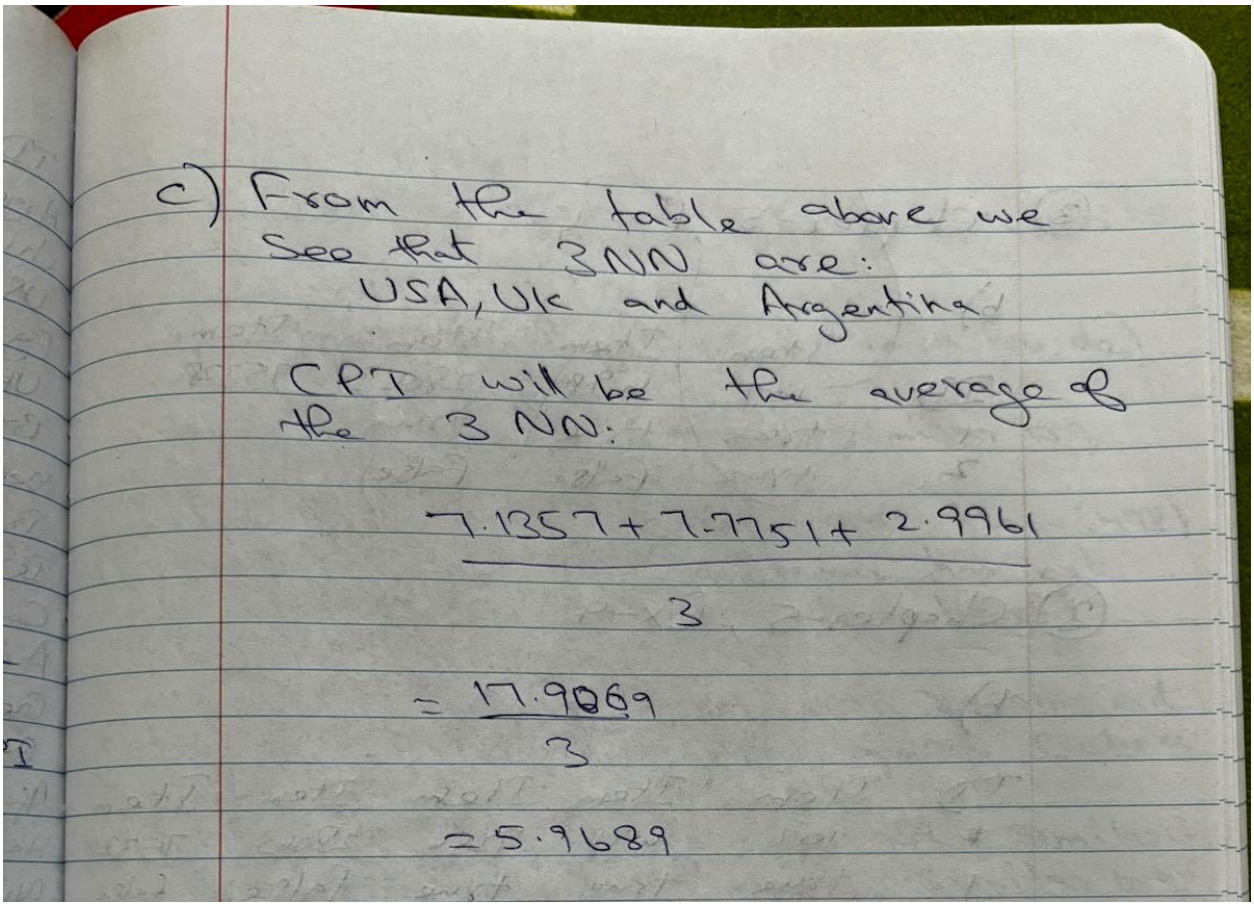
	A	C	D	E	F	G	H	I	J	K	L
1	COUNTRY ID	TOP-10 INCOME	INFANT MORT	MIL SPEND	SCHOOL YEARS	Normalized Life Expectancy	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance between Russia and other countries
2	Afghanistan	23.21	74.3	4.44	0.4	0.393906713	0.04453125	0.896508728	0.651197605	0	1.275402958
3	Haiti	47.67	73.1	0.09	3.4	0	1	0.881546135	0	0.217391304	1.474860231
4	Nigeria	38.23	82.6	1.07	4.1	0.169857104	0.63125	1	0.146706587	0.268115942	1.288744017
5	Egypt	26.58	19.6	1.86	5.3	0.686977622	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529
6	Argentina	32.3	13.3	0.76	10.1	0.829603667	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483
7	China	29.98	13.7	1.95	6.4	0.805338366	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968
8	Brazil	42.93	14.5	1.43	7.2	0.758155837	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746
9	Israel	28.8	3.6	6.77	12.5	0.978700458	0.262890625	0.014962594	1	0.876811594	0.586832479
10	USA	29.85	6.3	4.72	13.7	0.903478026	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085
11	Ireland	27.23	3.5	0.6	11.5	0.947694796	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028
12	UK	28.49	4.4	2.59	13	0.94607711	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205
13	Germany	22.07	3.5	1.31	12	0.950121327	0	0.013715711	0.182634731	0.84057971	0.643723893
14	Canada	24.79	4.9	1.42	14.2	0.97034241	0.10625	0.03117207	0.199101796	1	0.59145093
15	Australia	25.4	4.2	1.86	11.5	1	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574
16	Sweden	22.18	2.4	1.27	12.8	0.982205446	0.004296875	0	0.176646707	0.898550725	0.660962784
17	New Zealand	27.81	4.9	1.13	12.3	0.961714748	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158
18	**Russia**	31.68	10	3.87	12.9	0.609867889	0.375390625	0.094763092	0.565868263	0.905797101	0
19	**Total**										



	A	B	C	D	E	F	G	H	I	J	K	L
1	COUNTRY ID	LIFE Exp	TOP-10 INCOME	INFANT MORT	MIL SPEND	SCHOOL YEARS	Normalized Life Expectancy	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance bet Russia and other coun
2	Afghanistan	59.61	23.21	74.3	4.44	0.4	0.393906713	0.04453125	0.896508728	0.651197605	0	1.275402958
3	Haiti	45	47.67	73.1	0.09	3.4	0	1	0.881546135	0	0.217391304	1.474860231
4	Nigeria	51.3	38.23	82.6	1.07	4.1	0.169857104	0.63125	1	0.146706587	0.268115942	1.288744017
5	Egypt	70.48	26.58	19.6	1.86	5.3	0.686977622	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529
6	Argentina	75.77	32.3	13.3	0.76	10.1	0.829603667	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483
7	China	74.87	29.98	13.7	1.95	6.4	0.805338366	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968
8	Brazil	73.12	42.93	14.5	1.43	7.2	0.758155837	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746
9	Israel	81.3	28.8	3.6	6.77	12.5	0.978700458	0.262890625	0.014962594	1	0.876811594	0.586832479
10	USA	78.51	29.85	6.3	4.72	13.7	0.903478026	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085
11	Ireland	80.15	27.23	3.5	0.6	11.5	0.947694796	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028
12	UK	80.09	28.49	4.4	2.59	13	0.94607711	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205
13	Germany	80.24	22.07	3.5	1.31	12	0.950121327	0	0.013715711	0.182634731	0.84057971	0.643723893
14	Canada	80.99	24.79	4.9	1.42	14.2	0.97034241	0.10625	0.03117207	0.199101796	1	0.59145093
15	Australia	82.09	25.4	4.2	1.86	11.5	1	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574
16	Sweden	81.43	22.18	2.4	1.27	12.8	0.982205446	0.004296875	0	0.176646707	0.898550725	0.660962784
17	New Zealand	80.67	27.81	4.9	1.13	12.3	0.961714748	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158
18	**Russia**	67.62	31.68	10	3.87	12.9	0.609867889	0.375390625	0.094763092	0.565868263	0.905797101	0
19	**Total**											



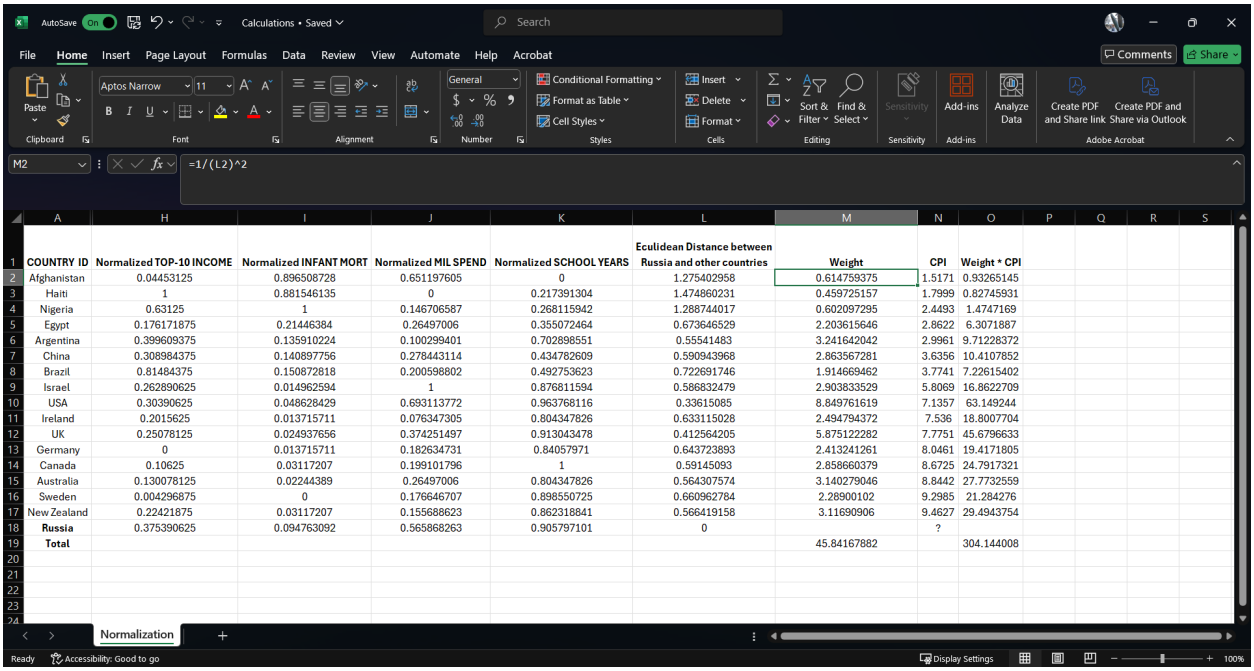
	A	D	E	F	G	H	I	J	K	L	N	P
1	COUNTRY ID	INFANT MORT	MIL SPEND	SCHOOL YEARS	Normalized Life Expectancy	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance between Russia and other countries	CPI	
2	Afghanistan	74.3	4.44	0.4	0.393906713	0.04453125	0.896508728	0.651197605	0	1.275402958	1.5171	
3	Haiti	73.1	0.09	3.4	0	1	0.881546135	0	0.217391304	1.474860231	1.7999	
4	Nigeria	82.6	1.07	4.1	0.169857104	0.63125	1	0.146706587	0.268115942	1.288744017	2.4493	
5	Egypt	19.6	1.86	5.3	0.686977622	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529	2.8622	
6	Argentina	13.3	0.76	10.1	0.829603667	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483	2.9961	
7	China	13.7	1.95	6.4	0.805338366	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968	3.6356	
8	Brazil	14.5	1.43	7.2	0.758155837	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746	3.7741	
9	Israel	3.6	6.77	12.5	0.978700458	0.262890625	0.014962594	1	0.876811594	0.586832479	5.8069	
10	USA	6.3	4.72	13.7	0.903478026	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085	7.1357	
11	Ireland	3.5	0.6	11.5	0.947694796	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028	7.536	
12	UK	4.4	2.59	13	0.94607711	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205	7.7751	
13	Germany	3.5	1.31	12	0.950121327	0	0.013715711	0.182634731	0.84057971	0.643723893	8.0461	
14	Canada	4.9	1.42	14.2	0.97034241	0.10625	0.03117207	0.199101796	1	0.59145093	8.6725	
15	Australia	4.2	1.86	11.5	1	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574	8.8442	
16	Sweden	2.4	1.27	12.8	0.982205446	0.004296875	0	0.176646707	0.898550725	0.660962784	9.2985	
17	New Zealand	4.9	1.13	12.3	0.961714748	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158	9.4627	
18	**Russia**	10	3.87	12.9	0.609867889	0.375390625	0.094763092	0.565868263	0.905797101	0	?	
19	**Total**											



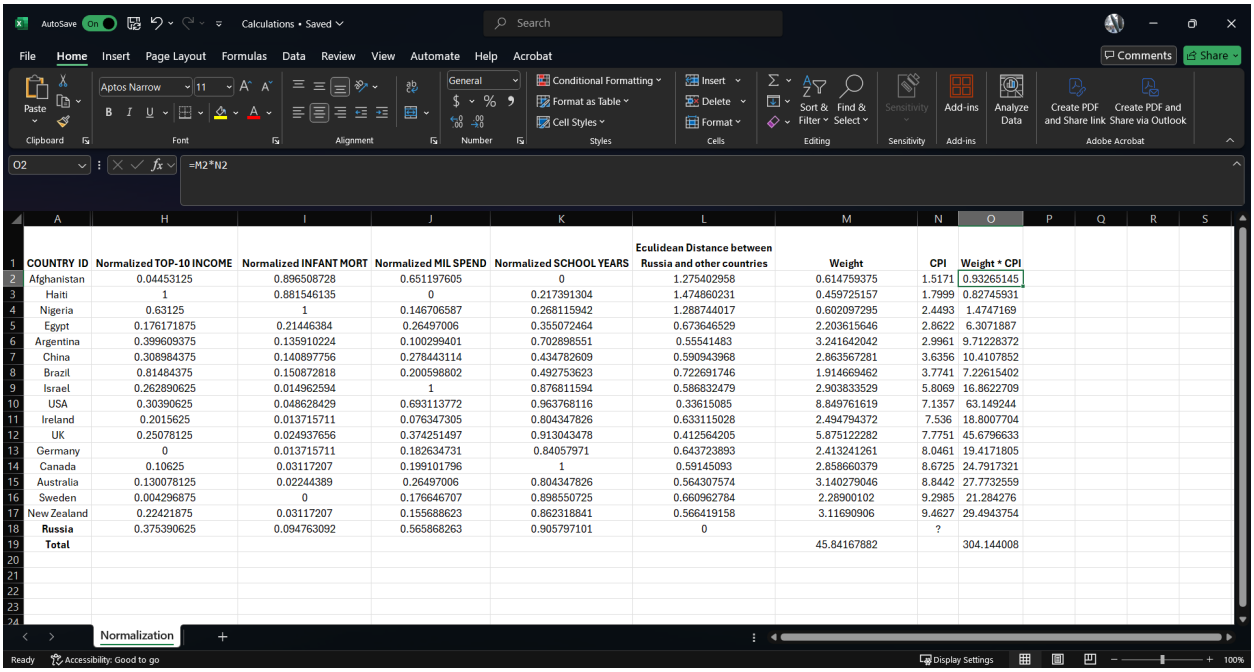
Question: d)

What value would a weighted k-NN prediction model—with k _ 16 (i.e., the full dataset) and using a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query—return for the CPI of Russia when it is applied to the range-normalized data?

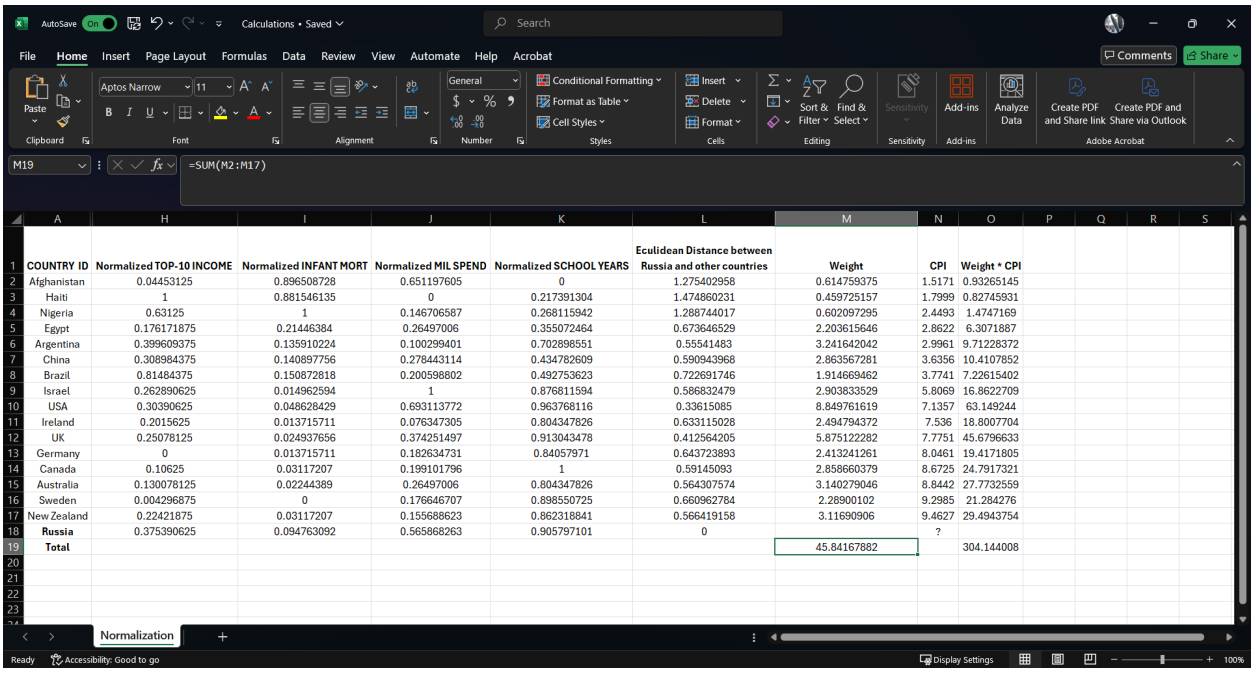
Solution:



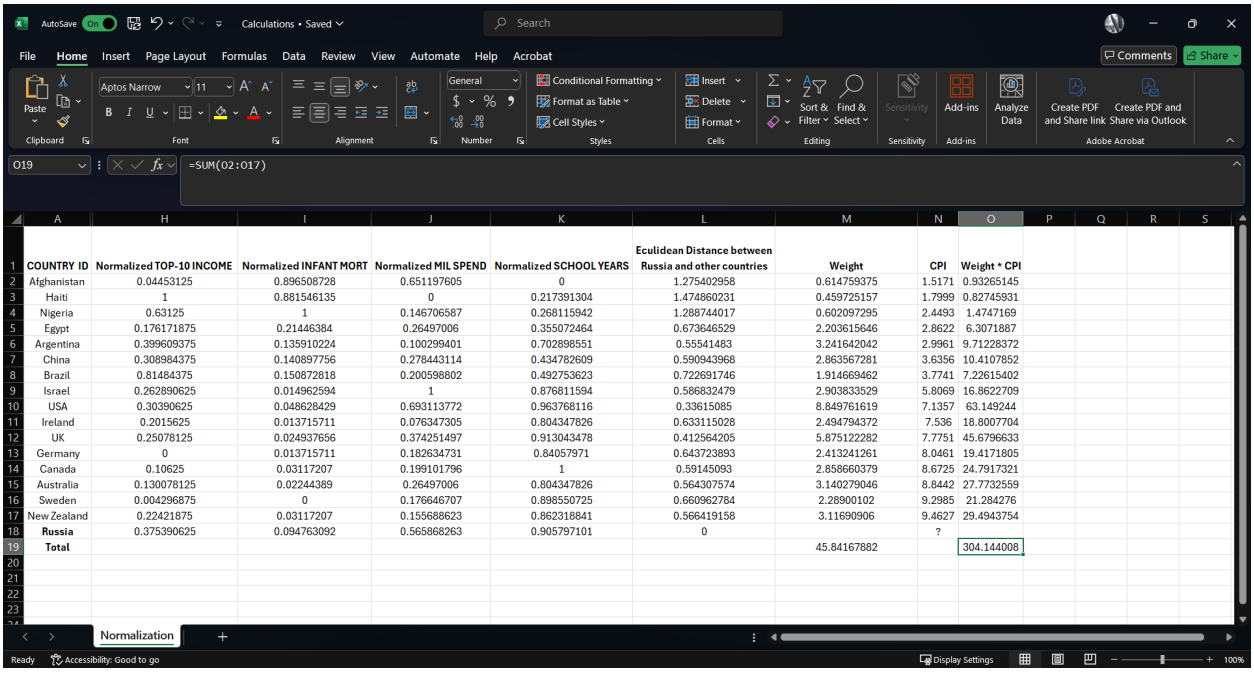
	A	H	I	J	K	L	M	N	O
1	COUNTRY ID	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance between Russia and other countries	Weight	CPI	Weight * CPI
2	Afghanistan	0.04453125	0.896508728	0.651197605	0	1.275402958	0.614759375	1.5171	0.93265145
3	Haiti	1	0.881546135	0	0.217391304	1.474860231	0.459725157	1.7999	0.82745931
4	Nigeria	0.63125	1	0.146706587	0.268115942	1.288744017	0.602097295	2.4493	1.4747169
5	Egypt	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529	2.203615646	2.8622	6.3071887
6	Argentina	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483	3.241642042	2.9961	9.71228372
7	China	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968	2.863567281	3.6356	10.4107852
8	Brazil	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746	1.914669462	3.7741	7.22615402
9	Israel	0.262890625	0.014962594	1	0.876811594	0.586832479	2.903833529	5.8069	16.8622709
10	USA	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085	8.849761619	7.1357	63.149244
11	Ireland	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028	2.494794372	7.536	18.8007704
12	UK	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205	5.875122282	7.7751	45.6796633
13	Germany	0	0.013715711	0.182634731	0.84057971	0.643723893	2.413241261	8.0461	19.4171805
14	Canada	0.10625	0.03117207	0.199101796	1	0.59145093	2.858660379	8.6725	24.7917321
15	Australia	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574	3.140279046	8.8442	27.7732559
16	Sweden	0.004296875	0	0.176646707	0.898550725	0.660962784	2.28900102	9.2985	21.284276
17	New Zealand	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158	3.11690906	9.4627	29.4943754
18	**Russia**	0.375390625	0.094763092	0.565868263	0.905797101	0		?	
19	**Total**						45.84167882		304.144008



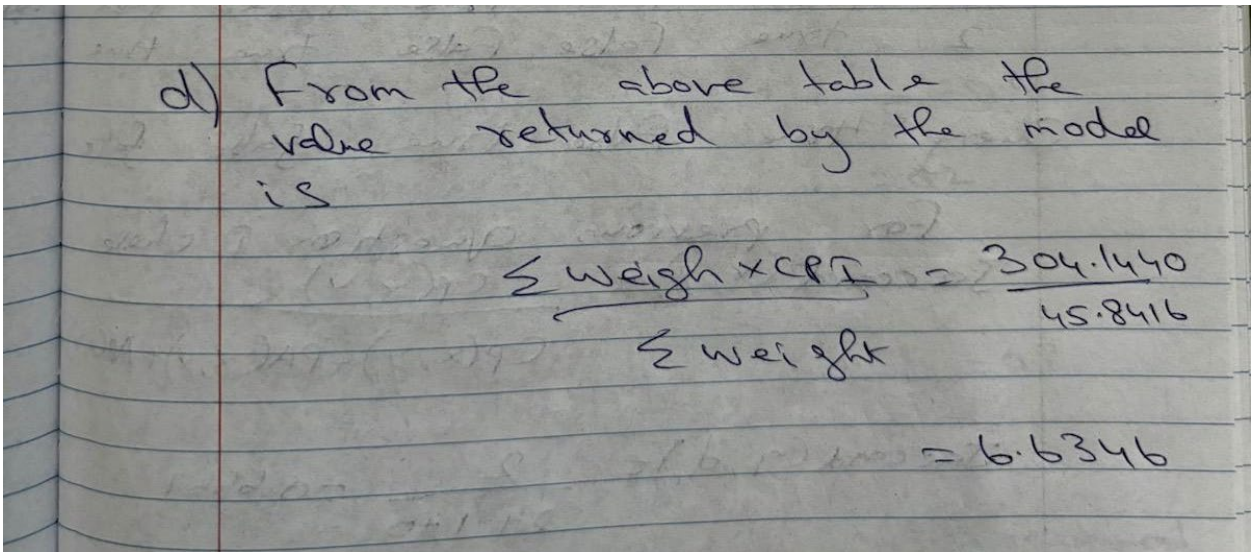
	A	H	I	J	K	L	M	N	O
1	COUNTRY ID	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance between Russia and other countries	Weight	CPI	Weight * CPI
2	Afghanistan	0.04453125	0.896508728	0.651197605	0	1.275402958	0.614759375	1.5171	0.93265145
3	Haiti	1	0.881546135	0	0.217391304	1.474860231	0.459725157	1.7999	0.82745931
4	Nigeria	0.63125	1	0.146706587	0.268115942	1.288744017	0.602097295	2.4493	1.4747169
5	Egypt	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529	2.203615646	2.8622	6.3071887
6	Argentina	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483	3.241642042	2.9961	9.71228372
7	China	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968	2.863567281	3.6356	10.4107852
8	Brazil	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746	1.914669462	3.7741	7.22615402
9	Israel	0.262890625	0.014962594	1	0.876811594	0.586832479	2.903833529	5.8069	16.8622709
10	USA	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085	8.849761619	7.1357	63.149244
11	Ireland	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028	2.494794372	7.536	18.8007704
12	UK	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205	5.875122282	7.7751	45.6796633
13	Germany	0	0.013715711	0.182634731	0.84057971	0.643723893	2.413241261	8.0461	19.4171805
14	Canada	0.10625	0.03117207	0.199101796	1	0.59145093	2.858660379	8.6725	24.7917321
15	Australia	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574	3.140279046	8.8442	27.7732559
16	Sweden	0.004296875	0	0.176646707	0.898550725	0.660962784	2.28900102	9.2985	21.284276
17	New Zealand	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158	3.11690906	9.4627	29.4943754
18	**Russia**	0.375390625	0.094763092	0.565868263	0.905797101	0		?	
19	**Total**						45.84167882		304.144008



COUNTRY ID	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance between Russia and other countries	Weight	CPI	Weight * CPI
Afghanistan	0.04453125	0.896508728	0.651197605	0	1.275402958	0.614759375	1.5171	0.93265145
Haiti	1	0.881546135	0	0.217391304	1.474860231	0.459725157	1.7999	0.82745931
Nigeria	0.63125	1	0.146706587	0.268115942	1.288744017	0.602097295	2.4493	1.4747169
Egypt	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529	2.203615646	2.8622	6.3071887
Argentina	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483	3.241642042	2.9961	9.71228372
China	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968	2.863567281	3.6356	10.4107852
Brazil	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746	1.914669462	3.7741	7.22615402
Israel	0.262890625	0.014962594	1	0.876811594	0.586832479	2.903833529	5.8069	16.8622709
USA	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085	8.849761619	7.1357	63.149244
Ireland	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028	2.494794372	7.536	18.8007704
UK	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205	5.875122282	7.7751	45.6796633
Germany	0	0.013715711	0.182634731	0.84057971	0.643723893	2.413241261	8.0461	19.4171805
Canada	0.10625	0.03117207	0.199101796	1	0.59145093	2.858660379	8.6725	24.7917321
Australia	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574	3.140279046	8.8442	27.7732559
Sweden	0.004296875	0	0.176646707	0.898550725	0.660962784	2.28900102	9.2985	21.284276
New Zealand	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158	3.11690906	9.4627	29.4943754
Russia	0.375390625	0.094763092	0.565868263	0.905797101	0		?	
Total						45.84167882		304.144008



COUNTRY ID	Normalized TOP-10 INCOME	Normalized INFANT MORT	Normalized MIL SPEND	Normalized SCHOOL YEARS	Eculidean Distance between Russia and other countries	Weight	CPI	Weight * CPI
Afghanistan	0.04453125	0.896508728	0.651197605	0	1.275402958	0.614759375	1.5171	0.93265145
Haiti	1	0.881546135	0	0.217391304	1.474860231	0.459725157	1.7999	0.82745931
Nigeria	0.63125	1	0.146706587	0.268115942	1.288744017	0.602097295	2.4493	1.4747169
Egypt	0.176171875	0.21446384	0.26497006	0.355072464	0.673646529	2.203615646	2.8622	6.3071887
Argentina	0.399609375	0.135910224	0.100299401	0.702898551	0.55541483	3.241642042	2.9961	9.71228372
China	0.308984375	0.140897756	0.278443114	0.434782609	0.590943968	2.863567281	3.6356	10.4107852
Brazil	0.81484375	0.150872818	0.200598802	0.492753623	0.722691746	1.914669462	3.7741	7.22615402
Israel	0.262890625	0.014962594	1	0.876811594	0.586832479	2.903833529	5.8069	16.8622709
USA	0.30390625	0.048628429	0.693113772	0.963768116	0.33615085	8.849761619	7.1357	63.149244
Ireland	0.2015625	0.013715711	0.076347305	0.804347826	0.633115028	2.494794372	7.536	18.8007704
UK	0.25078125	0.024937656	0.374251497	0.913043478	0.412564205	5.875122282	7.7751	45.6796633
Germany	0	0.013715711	0.182634731	0.84057971	0.643723893	2.413241261	8.0461	19.4171805
Canada	0.10625	0.03117207	0.199101796	1	0.59145093	2.858660379	8.6725	24.7917321
Australia	0.130078125	0.02244389	0.26497006	0.804347826	0.564307574	3.140279046	8.8442	27.7732559
Sweden	0.004296875	0	0.176646707	0.898550725	0.660962784	2.28900102	9.2985	21.284276
New Zealand	0.22421875	0.03117207	0.155688623	0.862318841	0.566419158	3.11690906	9.4627	29.4943754
Russia	0.375390625	0.094763092	0.565868263	0.905797101	0		?	
Total						45.84167882		304.144008



Question: e)

The actual 2011 CPI for Russia was 2:4488. Which of the predictions made was the most accurate? Why do you think this was?

Solution:

The most accurate prediction is the **3-Nearest Neighbor without normalization** (4.59), though it still overestimates the actual CPI for Russia. The normalization process seems to have increased the predicted CPI in both models.

3. (15%) Chapter 5, exercise 4

Question: a)

The company has decided to use a similarity-based model to implement the recommender system. Which of the following three similarity indexes do you think the system should be based on?

Solution:

For a large data set considering co-absences are not meaningful, as there will be more sparse data. Jaccard similarity index is ideal as it ignores co-absences. **Jaccard Similarity** is likely the best choice because:

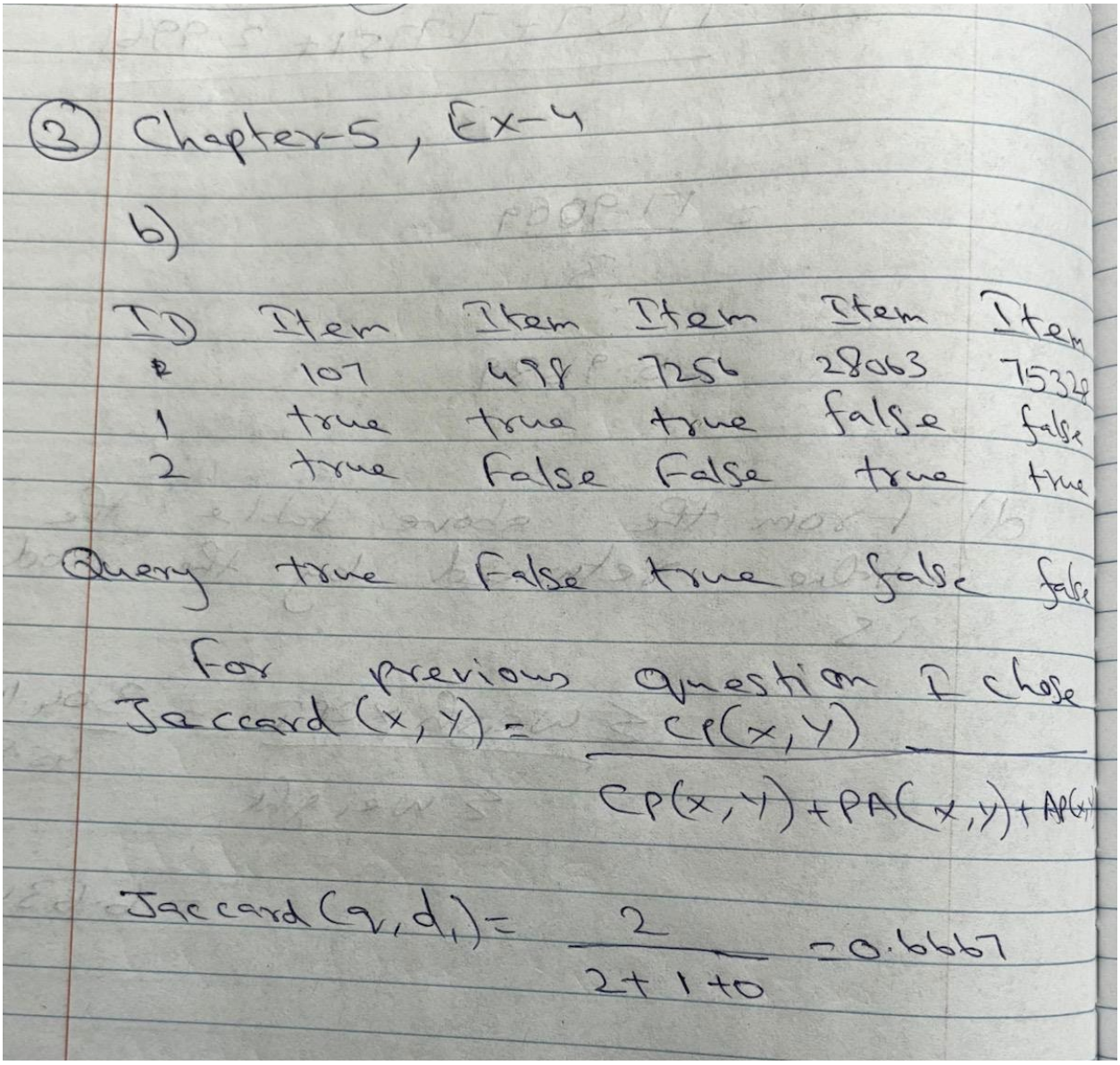
- The data is binary (either bought or not bought).

- Jaccard focuses on the overlap in purchases between two customers, which is important for making recommendations based on shared interests.
- It only considers the items both customers have bought (true-true matches), which aligns well with the goal of recommending similar items.

Question: b)

What items will the system recommend to the following customer? Assume that the recommender system uses the similarity index you chose in the first part of this question and is trained on the sample dataset listed above. Also assume that the system generates recommendations for query customers by finding the customer most similar to them in the dataset and then recommending the items that this similar customer has bought but that the query customer has not bought.

Solution:



(3) Chapter-5, Ex-4

b)

ID	Item	Item	Item	Item	Item
*	107	498	7256	28063	75328
1	true	true	true	false	false
2	true	False	False	true	true
Query	true	False	true	false	false

For previous question I chose

$$Jaccard(x,y) = \frac{cp(x,y)}{cp(x,y) + pA(x,y) + Ap(x,y)}$$

$$Jaccard(q,d_1) = \frac{2}{2+1+0} = 0.6667$$

$Jaccard(q, d_2) = \frac{1}{1+2+1} = \frac{1}{4} = 0.25$

$Jaccard(q, d_1) > Jaccard(q, d_2)$

So d_1 is more similar to q.

There is only one item (498) the query customer has not bought and d_1 has bought.

So the system will recommend item 498 to the query customer.

For this example in all similarity metrics d_1 is more similar to q.

$Russell\text{-}Rao(q, d_1) = \frac{2}{5} = 0.4$

$Russell\ Rao(q, d_2) = \frac{1}{5} = 02$

$(q, d_1) > (q, d_2)$

$Sokal\text{-}Michener(q, d_1) = \frac{4}{5} = 08$

$Sokal\ Michener(q, d_2) = \frac{2}{5} = 0.4$

So the system will recommend 498 regardless of similarity metrics.

4. (15%) Chapter 5, exercise 5

Question: a)

A good measure of distance between two instances with categorical features is the overlap metric (also known as the hamming distance), which simply counts the number of descriptive features that have different values. Using this measure of distance, compute the distances between the mystery animal and each of the animals in the animal dataset.

Solution:

The overlap metric between the query instance and each instance in the dataset by counting the number of feature values that are different.

ID	CLASS	Overlap Metric
1	Mammal	6
2	Amphibian	1
3	Mammal	6
4	Bird	2

Question b):

If you used a 1-NN model, what class would be assigned to the mystery animal?

Solution:

D2 is the first nearest neighbor, so the mystery animal would be Amphibian.

Question c):

If you used a 4-NN model, what class would be assigned to the mystery animal? Would this be a good value for k for this dataset?

Solution:

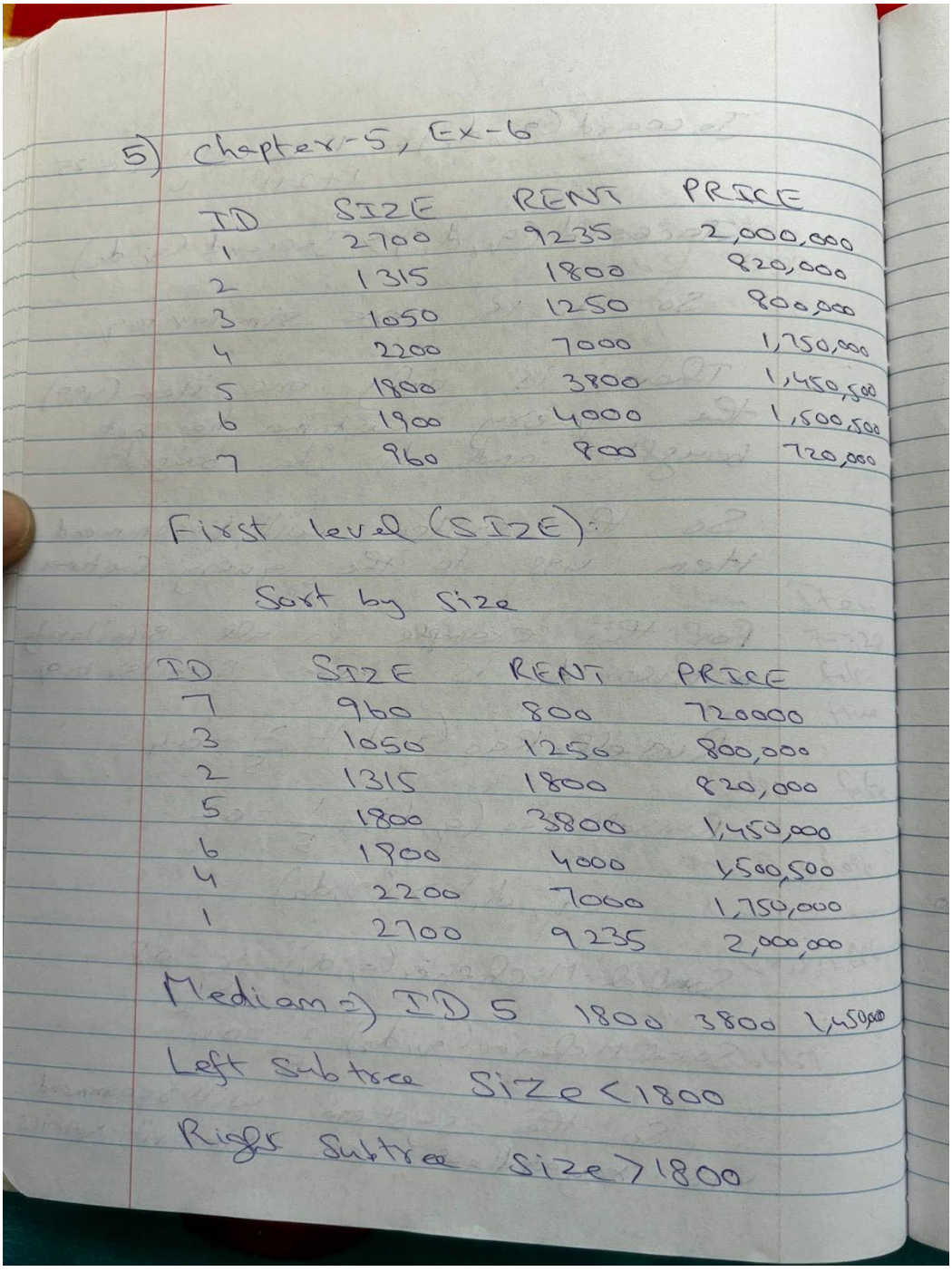
4-NN covers all the instances in the dataset as there are only 4 records. As a result, any query would be assigned the majority class in the dataset, that is mammal in this case. So, for this dataset, a 4-NN model would underfit the dataset. 4 is not a good value of k for this dataset.

5. (15%) Chapter 5, exercise 6

Question a):

Create a k-d tree for this dataset. Assume the following order over the features: RENT then SIZE.

Solution:



5) Chapter-5, Ex-6

ID	SIZE	RENT	PRICE
1	2700	9235	2,000,000
2	1315	1800	820,000
3	1050	1250	800,000
4	2200	7000	1,750,000
5	1800	3800	1,450,500
6	1900	4000	1,500,500
7	960	800	720,000

First level (SIZE):

Sort by size

ID	SIZE	RENT	PRICE
7	960	800	720000
3	1050	1250	800,000
2	1315	1800	820,000
5	1800	3800	1,450,000
6	1900	4000	1,500,500
4	2200	7000	1,750,000
1	2700	9235	2,000,000

Median) ID 5 1800 3800 1,450,000

Left Subtree Size < 1800

Right Subtree Size > 1800

Left Subtree:

ID	Size	Rent	Price
7	960	800	720000
3	1050	1250	800000
2	1315	1800	820000

Sort by RENT.

Median:

3 1050 1250 800000

Left of Left Subtree:

Size < 1050

7 960 800 720000

This is the leaf node.

Right of lft Subtree:

Size > 1050

2 1315 1800 820000

Leaf node.

Right Subtree: (>1800)

6	1900	4000	1500500
4	2200	7000	1750000
1	2700	9235	2000000

Mediam:

4 2200 7000 1750000

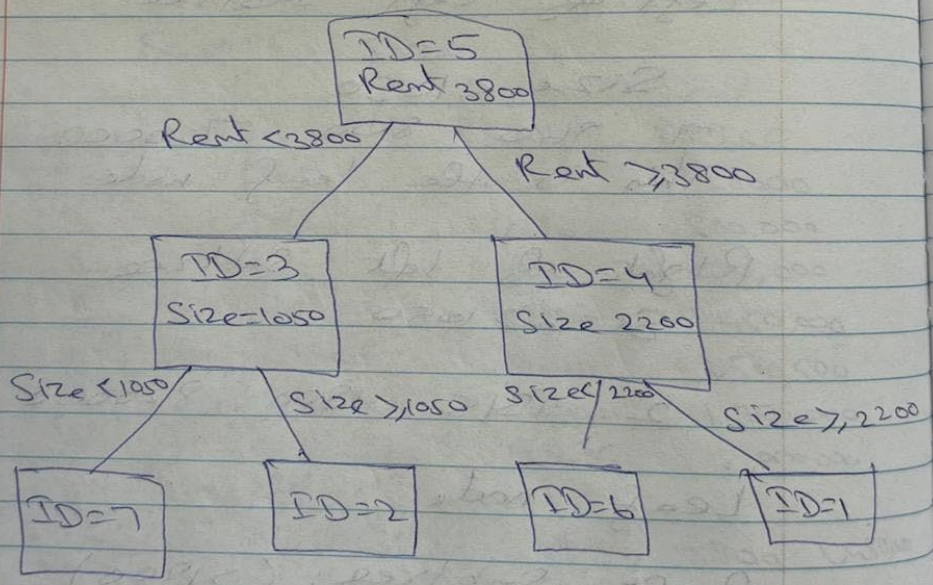
Left of Right Subtree (<2200)

6 1900 4000 1500500

Leaf node.

Right of Right Subtree (>2200)

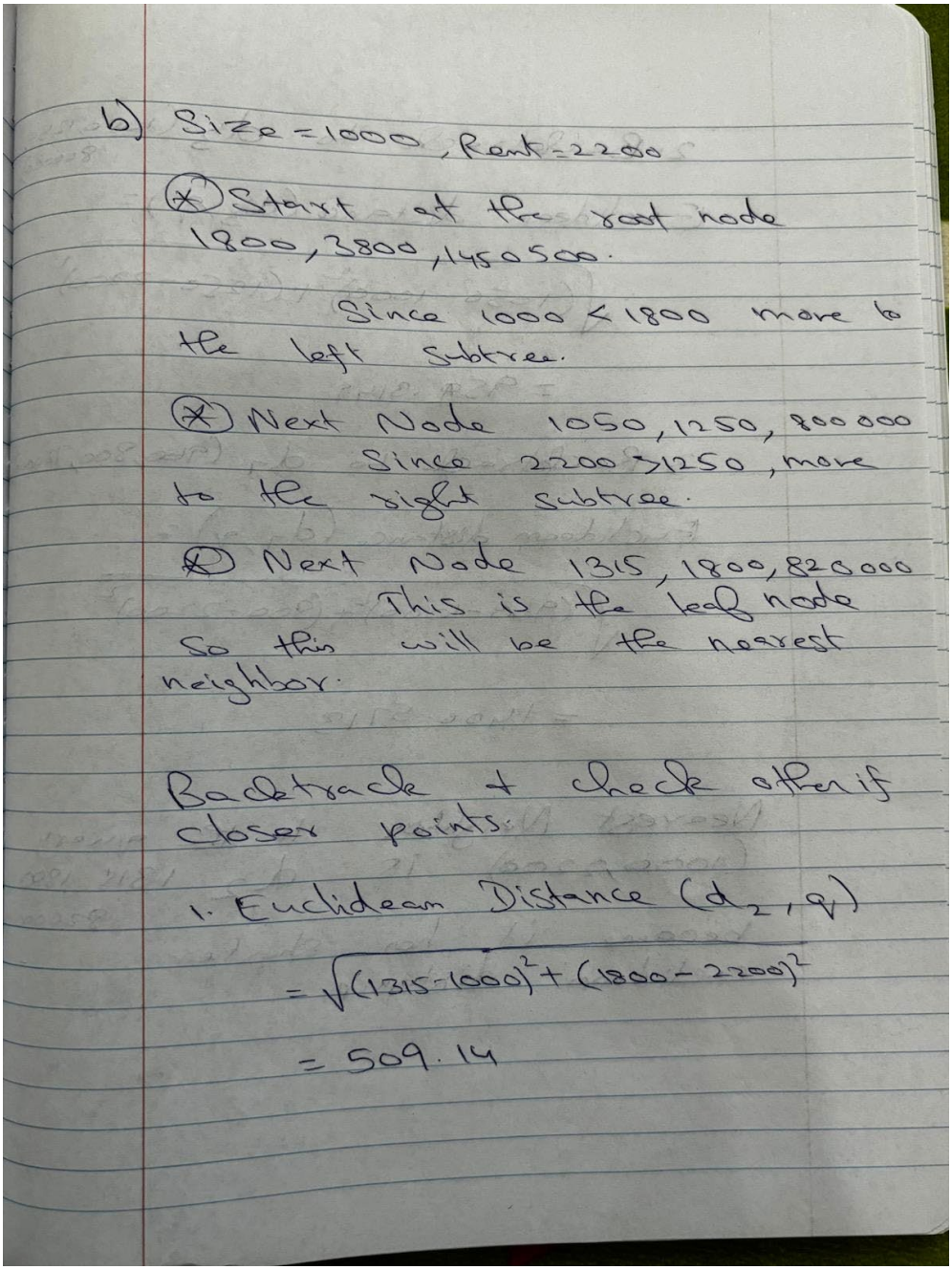
1 2700 9235 2000000



Question b):

Using the k-d tree that you created in the first part of this question, find the nearest neighbor to the following query: SIZE = 1;000, RENT = 2;200.

Solution:



2. Backtrack to d_3 (1050, 1250, 800000)

Euclidean distance $(d_3, q) =$

$$\sqrt{(1050-1000)^2 + (1250-2200)^2}$$

$$= 959.3149$$

3. Backtrack to d_7 (960, 800, 720000)

Euclidean distance $(d_7, q) =$

$$\sqrt{(960-1000)^2 + (800-2200)^2}$$

$$= 1400.5713$$

Nearest Neighbor to the query (1000, 2200) is d_2 1315 1800 8200000 because it has shorter euclidean distance.