Bronco ID: |0|1|3|4|8|4|6|7|9|

Last Name: Francisco

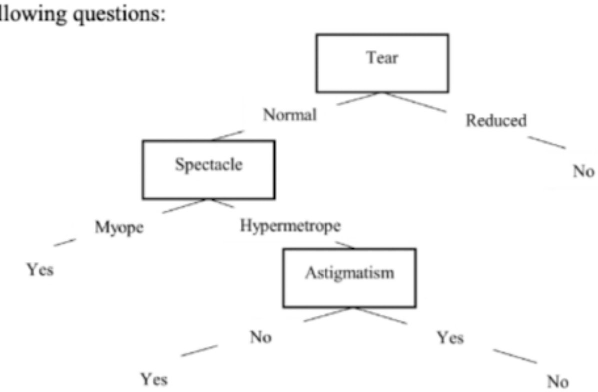
First Name: Serrano

1:

Part a:

A2:

1. [16 points] Considering that ID3 built the decision tree below after analyzing a given training set, answer the following questions:



a) [12 points] What is the accuracy of this model if applied to the test set below? You must **identify each** True Positive, True Negative, False Positive, and False Negative for full credit. For instance: TP = 1,5 | TN = 2,3 …

#	Age	Spectacle	Astigmatism	Tear	Lenses (ground truth)	
1	Young	Hypermetrope	Yes NO	Normal ✓	Yes	
2	Young	Hypermetrope	No yes	Normal ✓	Yes	
3	Young	Myope	No	Reduced	No	— NO
4	Presbyopic	Hypermetrope	No	Reduced	No	— NO
5	Presbyopic	Myope yes	No	Normal ✓	No	
6	Presbyopic	Myope	Yes	Reduced	No	— NO
7	Prepresbyopic	Myope yes	Yes	Normal ✓	Yes	
8	Prepresbyopic	Myope	No	Reduced	No	— NO

	Output:
#1	NO · ×
#2	yes · ✓
#3	NO · ✓
#4	NO · ✓
#5	yes · X
#6	NO · ✓
#7	yes · ✓
#8	NO

TP = 2,7

FP = 5

TN = 3,4,6,8

FN = 1

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FN + FP} = \frac{2+4}{8} = \frac{6}{8} = \boxed{0.75}$$

Part b:

b) Percision $= TP / TP+FP = \frac{2}{2+1} = \frac{2}{3}$

Recall $= TP / TP+FN = \frac{2}{3}$

↳ actually relev. but misclassified

$$F1 = 2 * \left(\frac{\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)}{\frac{2}{3} + \frac{2}{3}} \right)$$

$$= 2 * \left(\frac{\cancel{4}/\cancel{9}}{\frac{4}{3}} \right) \frac{3}{4}$$

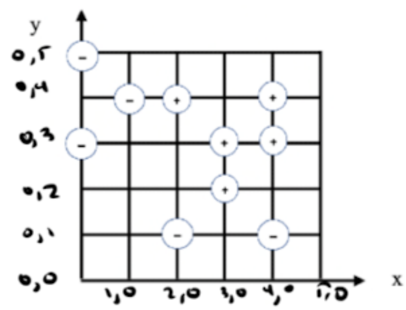
$$= 2 * \left(\frac{1}{3} \right)$$

$$= \frac{2}{3}$$

Recall = percision = F1 score

2:
https://github.com/franserr99/cs4210/blob/main/a2/decision_tree_2.py
3:

3. [32 points] Consider the dataset below to answer the following questions:



a. [4 points] What is the leave-one-out cross-validation error rate (LOO-CV) for **1NN**? Use Euclidean distance as your distance measure and the error rate calculated as:

$$error\ rate = \frac{number\ of\ wrong\ predictions}{total\ number\ of\ predictions}$$

b. [4 points] What is the leave-one-out cross-validation error rate (LOO-CV) for **3NN**?

c. [4 points] What is the leave-one-out cross-validation error rate (LOO-CV) for **9NN**?

d. [5 points] Draw de **decision boundary** learned by the 1NN algorithm.

points:

i=1 0,5 : ⊖
i=2 0,3 : ⊖
i=3 1,4 : ⊖
i=4 2,4 : ⊕
i=5 2,1 : ⊖
i=6 3,2 : ⊕
i=7 3,3 : ⊕
i=8 4,1 : ⊖
i=9 4,3 : ⊕
i=10 4,4 : ⊕

leave i=1 out (i=1 is our test)

distances:

$d_{1,2} = ((0-0)^2+(5-3)^2)^{1/2} = 2$

$d_{1,3} = ((0-1)^2+(5-4)^2)^{1/2} = 1.41$

$d_{1,4} = ((0-2)^2+(5-4)^2)^{1/2} = 2.23$

$d_{1,5} = ((0-2)^2+(5-1)^2)^{1/2} = 4.47$

$d_{1,6} = ((0-3)^2+(5-2)^2)^{1/2} = 4.24$

$d_{1,7} = ((0-3)^2+(5-3)^2)^{1/2} = 3.61$

$d_{1,8} = ((0-4)^2+(5-1)^2)^{1/2} = 5.65$

$d_{1,9} = ((0-4)^2+(5-3)^2)^{1/2} = 4.47$

$d_{1,10} = ((0-4)^2+(5-4)^2)^{1/2} = 4.12$

1NN i=3 is closest ⇒ i=1 classified as ⊖. True label ⊖, no misprediction

3NN: {i = 3, i = 2, i = 4}

i = 3: ⊖
i = 4: ⊕
i = 2: ⊖

we classify as ⊖ true label ⊖, no misprediction

9NN: {all except one left out}

5 ⊖
5 ⊕

overall of set. True label is ⊖ so (4 ⊖, 5 ⊕)

so we classify as ⊕
but true label is ⊖
so misprediction = 1

leave i = 2 out (i = 2 is our test) (0, 3)

distances:

$d_{2,1} = ((0-0)^2 + (3-5)^2)^{1/2} = 2$

$d_{2,3} = ((0-1)^2 + (3-4)^2)^{1/2} = 1.41$

$d_{2,4} = ((0-2)^2 + (3-4)^2)^{1/2} = 2.24$

$d_{2,5} = ((0-2)^2 + (3-1)^2)^{1/2} = 2.83$

$d_{2,6} = ((0-3)^2 + (3-2)^2)^{1/2} = 3.16$

$d_{2,7} = ((0-3)^2 + (3-3)^2)^{1/2} = 3$

$d_{2,8} = ((0-4)^2 + (3-1)^2)^{1/2} = 4.47$

$d_{2,9} = ((0-4)^2 + (3-3)^2)^{1/2} = 4$

$d_{2,10} = ((0-4)^2 + (3-4)^2)^{1/2} = 4.12$

1NN: {i = 3}

i = 3: ⊖, i = 2 true label is ⊕ so ✓ no misprediction

3NN: {i = 3, i = 1, i = 4}

i = 3: ⊖
i = 1: ⊖
i = 4: ⊕

majority vote: ⊖, classification ⊖ no mispre

9NN: same as before. even split on classification means whenever leave out left out the opposite label is what we classify as.
mispredictions = 2

leave $i=3$ out ($i=3$ is our test)

distances:

$$d_{3,1} = ((1-0)^2+(4-5)^2)^{1/2} = 1.41$$

$$d_{3,2} = ((1-0)^2+(4-3)^2)^{1/2} = 1.41$$

$$d_{3,4} = ((1-2)^2+(4-4)^2)^{1/2} = 1$$

$$d_{3,5} = ((1-2)^2+(4-1)^2)^{1/2} = 3.16$$

$$d_{3,6} = ((1-3)^2+(4-2)^2)^{1/2} = 2.83$$

$$d_{3,7} = ((1-3)^2+(4-3)^2)^{1/2} = 2.27$$

$$d_{3,8} = ((1-4)^2+(4-1)^2)^{1/2} = 3$$

$$d_{3,9} = ((1-4)^2+(4-3)^2)^{1/2} = 3.16$$

$$d_{3,10} = ((1-4)^2+(4-4)^2)^{1/2} = 3$$

1NN: $\{i=4\}$

$i=4$: $\oplus$, $i=3$ true label is $\ominus$

misprediction = 1

3NN: $\{i=1, i=2, i=4\}$

$i=4$: $\oplus$

$i=1$: $\ominus$

$i=2$: $\ominus$

classification: $\ominus$, true label $\ominus$

misprediction = 0

9NN: same as vector.

misprediction = 3

leave $i=4$ out ($i=4$ is our test)

distances:

$$d_{4,1} = ((2-0)^2+(4-5)^2)^{1/2} = 2.24$$

$$d_{4,2} = ((2-0)^2+(4-3)^2)^{1/2} = 2.24$$

$$d_{4,3} = ((2-1)^2+(4-4)^2)^{1/2} = 1$$

$$d_{4,5} = ((2-2)^2+(4-1)^2)^{1/2} = 3$$

$$d_{4,6} = ((2-3)^2+(4-2)^2)^{1/2} = 2.24$$

$$d_{4,7} = ((2-3)^2+(4-3)^2)^{1/2} = 1.41$$

$$d_{4,8} = ((2-4)^2+(4-1)^2)^{1/2} = 3.6$$

$$d_{4,9} = ((2-4)^2+(4-3)^2)^{1/2} = 2.24$$

$$d_{4,10} = ((2-4)^2+(4-4)^2)^{1/2} = 2$$

1NN: $\{i=3\}$

$i=3$: $\ominus$, $i=4$ true label is $\oplus$

misprediction = 2

3NN: $\{i=3, i=7, i=10\}$

$i=3$: $\ominus$

$i=7$: $\oplus$

$i=10$: $\oplus$

classification : $\oplus$, true label $\oplus$

misprediction = 0

9NN : same as before.

misprediction = 4

leave $i=5$ out ($i=5$ is our test)

distances:

$$d_{5,1} = \left((2-0)^2+(1-5)^2\right)^{1/2} = 4.47$$

$$d_{5,2} = \left((2-0)^2+(1-3)^2\right)^{1/2} = 2.83$$

$$d_{5,3} = \left((2-1)^2+(1-4)^2\right)^{1/2} = 3.16$$

$$d_{5,4} = \left((2-2)^2+(1-4)^2\right)^{1/2} = 3$$

$$d_{5,6} = \left((2-3)^2+(1-2)^2\right)^{1/2} = 1.41$$

$$d_{5,7} = \left((2-3)^2+(1-3)^2\right)^{1/2} = 2.24$$

$$d_{5,8} = \left((2-4)^2+(1-1)^2\right)^{1/2} = 2$$

$$d_{5,9} = \left((2-4)^2+(1-3)^2\right)^{1/2} = 2.83$$

$$d_{5,10} = \left((2-4)^2+(1-4)^2\right)^{1/2} = 3.61$$

1NN: $\{i=6\}$

$i=6$: $\oplus$, predict $\oplus$

$i=5$ true label = $\ominus$

miss!

misprediction = 3

9NN : same as before.

misprediction = 5

3NN: $\{i=6, i=8, i=7\}$

$i=6$: $\oplus$

$i=7$: $\oplus$

$i=8$: $\ominus$

classification : $\oplus$, true label $\ominus$

misprediction = 1

leave $i=6$ out ($i=$ is our test)

distances:

$d_{,1} = ((3-0)^2+(2-5)^2)^{1/2} = 4.24$

$d_{6,2} = ((3-0)^2+(2-3)^2)^{1/2} = 3.16$

$d_{6,3} = ((3-1)^2+(2-4)^2)^{1/2} = 2.83$

$d_{6,4} = ((3-2)^2+(2-4)^2)^{1/2} = 2.24$

$d_{6,5} = ((3-2)^2+(2-1)^2)^{1/2} = 1.41$

$d_{6,7} = ((3-3)^2+(2-3)^2)^{1/2} = 1$

$d_{6,8} = ((3-4)^2+(2-1)^2)^{1/2} = 1.41$

$d_{6,9} = ((3-4)^2+(2-3)^2)^{1/2} = 1.41$

$d_{6,10} = ((3-4)^2+(2-4)^2)^{1/2} = 2.24$

can pick 2/3 for $i=5,8,9$ bc distances are the same

1NN: $\{i=7\}$

$i=7$: $\oplus$
$i=6$ true label $\ominus$
good classification!

misprediction = 3

3NN: 2/3 are $\ominus$ and 1/3 is $\oplus$, there is already 1 $\oplus$ guaranteed so depending on how you pick you can get a diff classification

we will pick the ones in the majority

$\left.\begin{array}{l} i=7: \oplus \\ i=5: \ominus \\ i=8: \ominus \end{array}\right\} \ominus$ miss!

misclassification = 2

9NN:
same,

misprediction = 6

leave $i=7$ out ($i=7$ is our test)

distances:

$d_{7,1} = ((3-0)^2+(3-5)^2)^{1/2} = 3.61$

$d_{7,2} = ((3-0)^2+(3-3)^2)^{1/2} = 3$

$d_{7,3} = ((3-1)^2+(3-4)^2)^{1/2} = 2.24$

$d_{7,4} = ((3-2)^2+(3-4)^2)^{1/2} = 1.41$

$d_{7,5} = ((3-2)^2+(3-1)^2)^{1/2} = 2.24$

$d_{7,6} = ((3-3)^2+(3-2)^2)^{1/2} = 1$

$d_{7,8} = ((3-4)^2+(3-1)^2)^{1/2} = 2.24$

$d_{7,9} = ((3-4)^2+(3-3)^2)^{1/2} = 1$

$d_{7,10} = ((3-4)^2+(3-4)^2)^{1/2} = 1.41$

1NN : can pick i = 6 or i = 9, both classify as ⊕

true label is ⊕ so no misprediction

mispredictions = 3

3NN: {i = 6, i = 9, i = 4}

i = 6 : ⊕
i = 9 : ⊕
i = 4 : ⊕

classify as ⊕, true label is ⊕

✓

mispredictions = 2

9NN : mispredictions = 7

leave i=8 out (i=8 is our test)

distances:

$d_{8,1} = ((4-0)^2 + (1-5)^2)^{1/2} = 5.66$

$d_{8,2} = ((4-0)^2 + (1-3)^2)^{1/2} = 4.47$

$d_{8,3} = ((4-1)^2 + (1-4)^2)^{1/2} = 4.24$

$d_{8,4} = ((4-2)^2 + (1-4)^2)^{1/2} = 3.61$

$d_{8,5} = ((4-2)^2 + (1-1)^2)^{1/2} = 2$

$d_{8,6} = ((4-3)^2 + (1-2)^2)^{1/2} = 1.41$

$d_{8,7} = ((4-3)^2 + (1-3)^2)^{1/2} = 2.23$

$d_{8,9} = ((4-4)^2 + (1-3)^2)^{1/2} = 2$

$d_{8,10} = ((4-4)^2 + (1-4)^2)^{1/2} = 3$

1NN: i = 6 : ⊕

i = 8 true label: ⊖

another miss

misprediction = 4

3NN : i = 6 : ⊕
i = 5 : ⊖
i = 9 : ⊕

a miss! ⊕

misprediction = 3

9NN : misprediction = 8

leave $i=9$ out ($i=9$ is our test)

distances:

$d_{9,1} = ((4-0)^2 + (3-5)^2)^{1/2} = 4.47$

$d_{9,2} = ((4-0)^2 + (3-3)^2)^{1/2} = 4$

$d_{9,3} = ((4-1)^2 + (3-4)^2)^{1/2} = 3.16$

$d_{9,4} = ((4-2)^2 + (3-4)^2)^{1/2} = 2.24$

$d_{9,5} = ((4-2)^2 + (3-1)^2)^{1/2} = 2.83$

$d_{9,6} = ((4-3)^2 + (3-2)^2)^{1/2} = 1.41$

$d_{9,7} = ((4-3)^2 + (3-3)^2)^{1/2} = 1$

$d_{9,8} = ((4-4)^2 + (3-1)^2)^{1/2} = 2$

$d_{9,10} = ((4-4)^2 + (3-4)^2)^{1/2} = 1$

1NN: $i=7$ or $i=10$
both are ⊕ so classify as ⊕

$i=9$ true label ⊕ ✓
misprediction = 4

3NN: $i=7$: ⊕
$i=10$: ⊕
$i=8$: ⊖
→ ⊕ ✓

mispredictions = 3

9NN: misprediction = 9

leave $i=10$ out ($i=10$ is our test)

distances:

$d_{10,1} = ((4-0)^2 + (4-5)^2)^{1/2} = 4.12$

$d_{10,2} = ((4-0)^2 + (4-3)^2)^{1/2} = 4.12$

$d_{10,3} = ((4-1)^2 + (4-4)^2)^{1/2} = 3$

$d_{10,4} = ((4-2)^2 + (4-4)^2)^{1/2} = 2$

$d_{10,5} = ((4-2)^2 + (4-1)^2)^{1/2} = 3.61$

$d_{10,6} = ((4-3)^2 + (4-2)^2)^{1/2} = 2.24$

$d_{10,7} = ((4-3)^2 + (4-3)^2)^{1/2} = 1.41$

$d_{10,8} = ((4-4)^2 + (4-1)^2)^{1/2} = 3$

$d_{10,9} = ((4-4)^2 + (4-3)^2)^{1/2} = 1$

1NN:

$i=9$ is closest, is $\oplus$

$i=10$ true label is $\oplus$

so no miss

mispredictions $=4$

9NN:

misprediction $=10$

3NN:

$\left.\begin{array}{l} i=9: \oplus \\ i=7: \oplus \\ i=4: \oplus \end{array}\right\rangle \oplus$ ✓

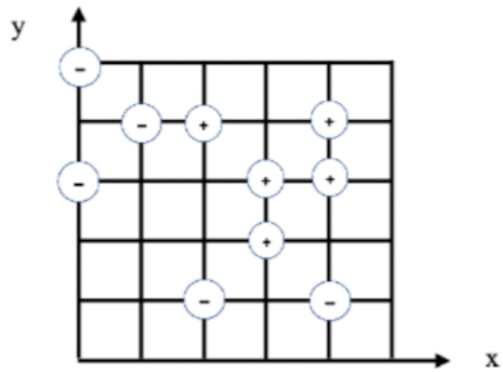
mispredictions $=3$

1NN error rate $= \frac{4}{10} = 0.4$

b.) 3NN error rate $= \frac{3}{10} = 0.3$

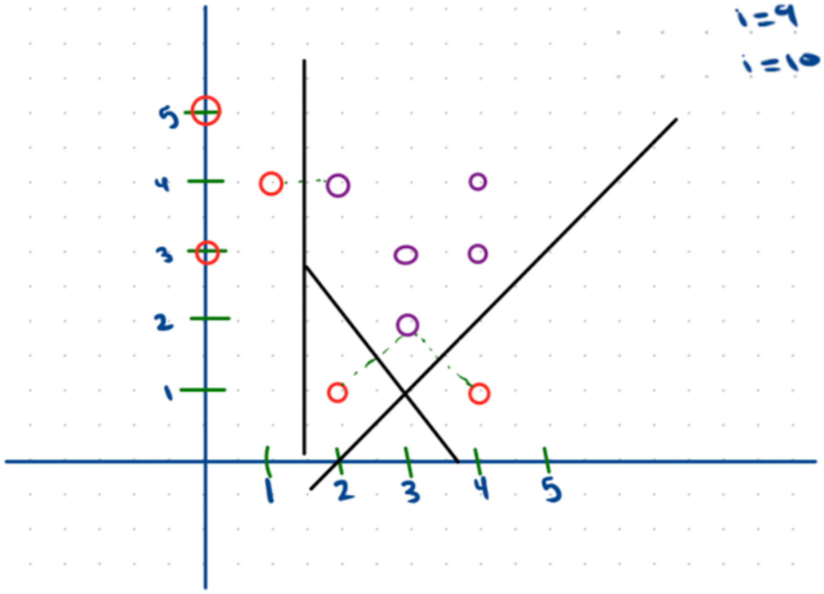
c.) 9NN error rate $= \frac{10}{10}$

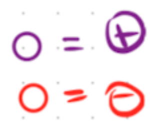
Part d:

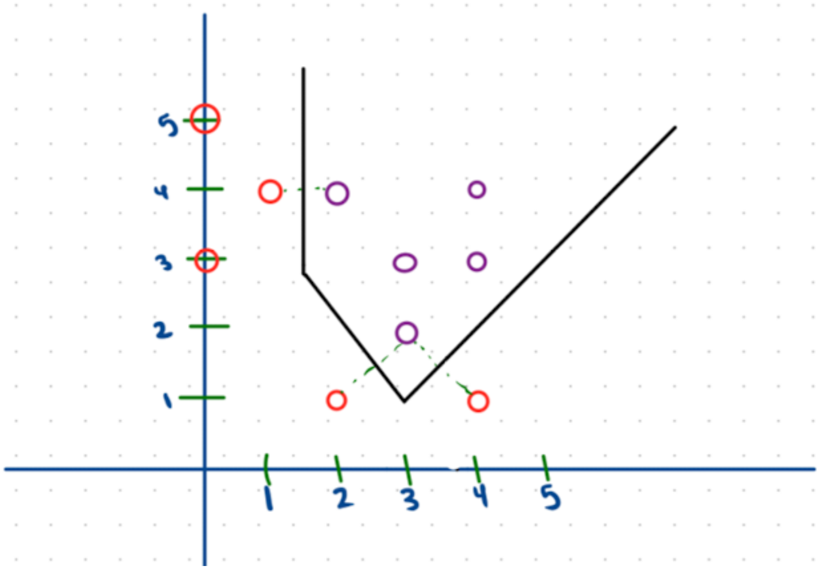


points:

i=1	0,5 :	⊖
i=2	0,3 :	⊖
i=3	1,4 :	⊖
i=4	2,4 :	⊕
i=5	2,1 :	⊖
i=6	3,2 :	⊕
i=7	3,3 :	⊕
i=8	4,1 :	⊖
i=9	4,3 :	⊕
i=10	4,4 :	⊕







Part e: https://github.com/franserr99/cs4210/blob/main/a2/knn.py

4:

4. [12 points] Find the class of instance #10 below following the 3NN strategy. Use Euclidean distance as your distance measure. You must **show all your calculations** for full credit.

ID	Red	Green	Blue	Class
#1	220	20	60	1
#2	255	99	21	1
#3	250	128	14	1
#4	144	238	144	2
#5	107	142	35	2
#6	46	139	87	2
#7	64	224	208	3
#8	176	224	23	3
#9	100	149	237	3
#10	154	205	50	?

3NN, Euclidean distance $= \left(\sum_{i=1}^{n} (q_i - p_i)^2 \right)^{1/2}$

#1 $d_1 = \left((154-220)^2 + (205-20)^2 + (50-60)^2\right)^{1/2} = 196.67$

#2 $d_2 = \left((154-255)^2 + (205-99)^2 + (50-21)^2\right)^{1/2} = 149.26$

#3 $d_3 = \left((154-250)^2 + (205-128)^2 + (50-14)^2\right)^{1/2} = 128.22$

#4 $d_4 = \left((154-144)^2 + (205-238)^2 + (50-144)^2\right)^{1/2} = 100.12$ *

#5 $d_5 = \left((154-107)^2 + (205-142)^2 + (50-35)^2\right)^{1/2} = 80.01$ *

#6 $d_6 = \left((154-46)^2 + (205-139)^2 + (50-87)^2\right)^{1/2} = 131.87$

#7 $d_7 = \left((154-64)^2 + (205-224)^2 + (50-208)^2\right)^{1/2} = 187.83$

#8 $d_8 = \left((154-176)^2 + (205-224)^2 + (50-23)^2\right)^{1/2} = 39.67$ *

#9 $d_9 = \left((154-100)^2 + (205-149)^2 + (50-237)^2\right)^{1/2} = 202.536$

#4, 5, 8 are selected

#4 → class 2
#5 → class 2
#8 → class 3

⇒ #10 is classified as class 2 by majority vote

5:Part a:

5. [25 points] Use the dataset below to answer the next questions:

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

a) [10 points] Classify the instance ‹D15, Sunny, Mild, Normal, Weak› following the Naïve Bayes strategy. **Show all your calculations** until the final normalized probability values.

b) [15 points] Complete the Python program (naïve_bayes.py) that will read the file weather_training.csv (training set) and output the classification of each test instance from the file weather_test (test set) **if the classification confidence is >= 0.75**. Sample of output:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis	Confidence
D15	Sunny	Hot	High	Weak	No	0.86
D16	Sunny	Mild	High	Weak	Yes	0.78

a.) P(class = yes | outlook = sunny, Temp = mild, Humidity = normal, wind = Weak) =

$= P(class = yes) * \prod_i P(A_i = x_i \mid class = yes)$

$= \frac{9}{14}$ (P(outlook = sunny | class = yes) * (P(temp = mild | class = yes)) *

(P(humidity = normal | class = yes) * P(wind = weak | class = yes))

$= \frac{9}{14}\left(\frac{2}{9} * \frac{4}{9} * \frac{6}{9} * \frac{6}{9}\right)$

$= \frac{2592}{91854} = 0.0282$

P(class = No | outlook = sunny, Temp = mild, Humidity = normal, wind = Weak) =

$= P(class = no) * \prod_i P(A_i = x_i \mid class = no)$

$= \frac{5}{14}$ (P(outlook = sunny | class = no) * (P(temp = mild | class = no)) *

(P(humidity = normal | class = no) * P(wind = weak | class = no))

$= \frac{\cancel{5}}{14}\left(\frac{3}{\cancel{5}}\right)\left(\frac{2}{5}\right)\left(\frac{1}{5}\right)\left(\frac{2}{5}\right) = \frac{3 \cdot 2 \cdot 2}{14 \cdot 25 \cdot 5} = \frac{12}{1750} = 0.00685714$

now normalize the values:

P(class = yes | outlook = sunny, Temp = mild, Humidity = normal, wind = Weak)

$$= \frac{0.0282}{0.0282 + 0.00685714}$$

$$= 0.804$$

P(class = no | outlook = sunny, Temp = mild, Humidity = normal, wind = Weak) =

$$\frac{.00685714}{.0282 + .00685714} = 0.1955$$

$$= 0.196$$

∴ instance #15 classified as <u>yes</u>.

Part b:
https://github.com/franserr99/cs4210/blob/main/a2/naive_bayes.py