Aprendizagem 2024

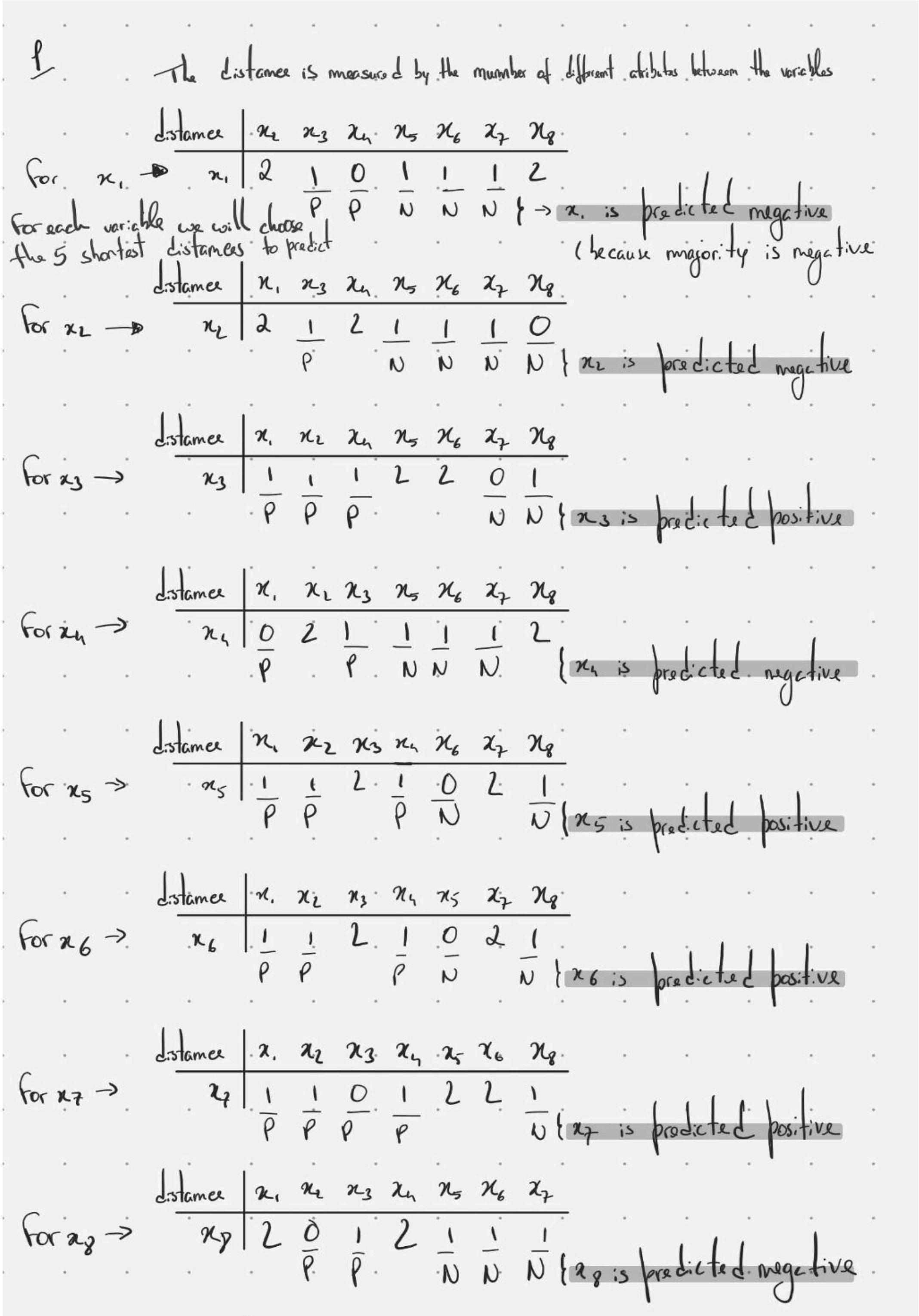
HomeWork II - Group 2

(ist1106078, ist1107095)

Part I: Pen and paper

We collected four positive (P) observations, $\{x_1 = (A, 0), x_2 = (B, 1), x_3 = (A, 1), x_4 = (A, 0)\}$, and four negative (N) observations, $\{x_5 = (B, 0), x_6 = (B, 0), x_7 = (A, 1), x_8 = (B, 1)\}$. Consider the problem of classifying observations as positive or negative.

1) [3.0v] Compute the F1-measure of a kNN with $k = 5$ and Hamming distance using a leave-one-out evaluation schema. Show all calculus.



1

The distance is measured by the number of different atributes between the variables

For each variable we will chose the 5 shortest distances to predict

For x_1 →

distance	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	2	1	0	1	1	1	2
		P	P	N	N	N	

} → x_1 is predicted negative (because majority is negative

For x_2 →

distance	x_1	x_3	x_4	x_5	x_6	x_7	x_8
x_2	2	1	2	1	1	1	0
		P		N	N	N	N

} x_2 is predicted negative

For x_3 →

distance	x_1	x_2	x_4	x_5	x_6	x_7	x_8
x_3	1	1	1	2	2	0	1
	P	P	P			N	N

} x_3 is predicted positive

For x_4 →

distance	x_1	x_2	x_3	x_5	x_6	x_7	x_8
x_4	0	2	1	1	1	1	2
	P		P	N	N	N	

} x_4 is predicted negative

For x_5 →

distance	x_1	x_2	x_3	x_4	x_6	x_7	x_8
x_5	1	1	2	1	0	2	1
	P	P		P	N		N

} x_5 is predicted positive

For x_6 →

distance	x_1	x_2	x_3	x_4	x_5	x_7	x_8
x_6	1	1	2	1	0	2	1
	P	P		P	N		N

} x_6 is predicted positive

For x_7 →

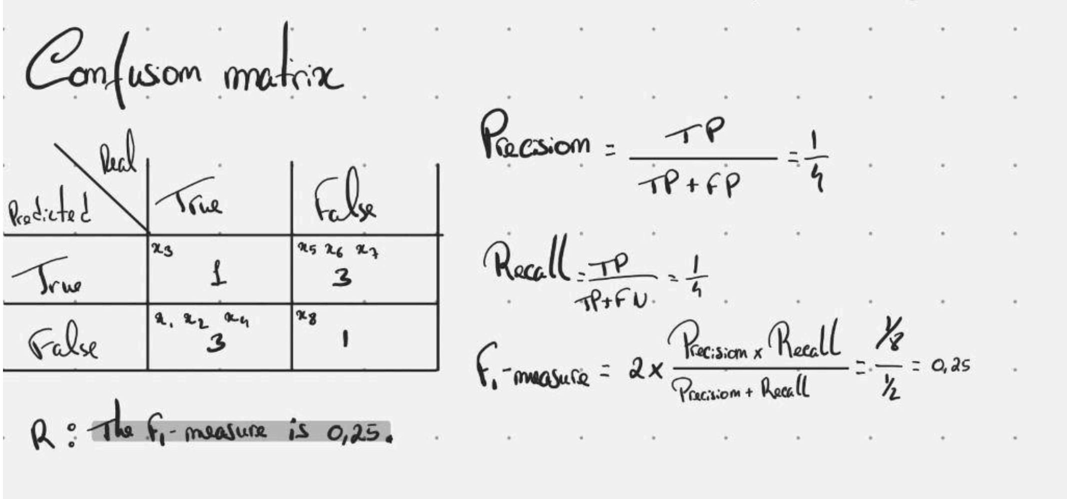
distance	x_1	x_2	x_3	x_4	x_5	x_6	x_8
x_7	1	1	0	1	2	2	1
	P	P	P	P			N

} x_7 is predicted positive

For x_8 →

distance	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_8	2	0	1	2	1	1	1
		P	P		N	N	N

} x_8 is predicted negative



Confusion matrix

Real / Predicted	True	False
True	x_3 1	x_5 x_6 x_7 3
False	x_1 x_2 x_4 3	x_8 1

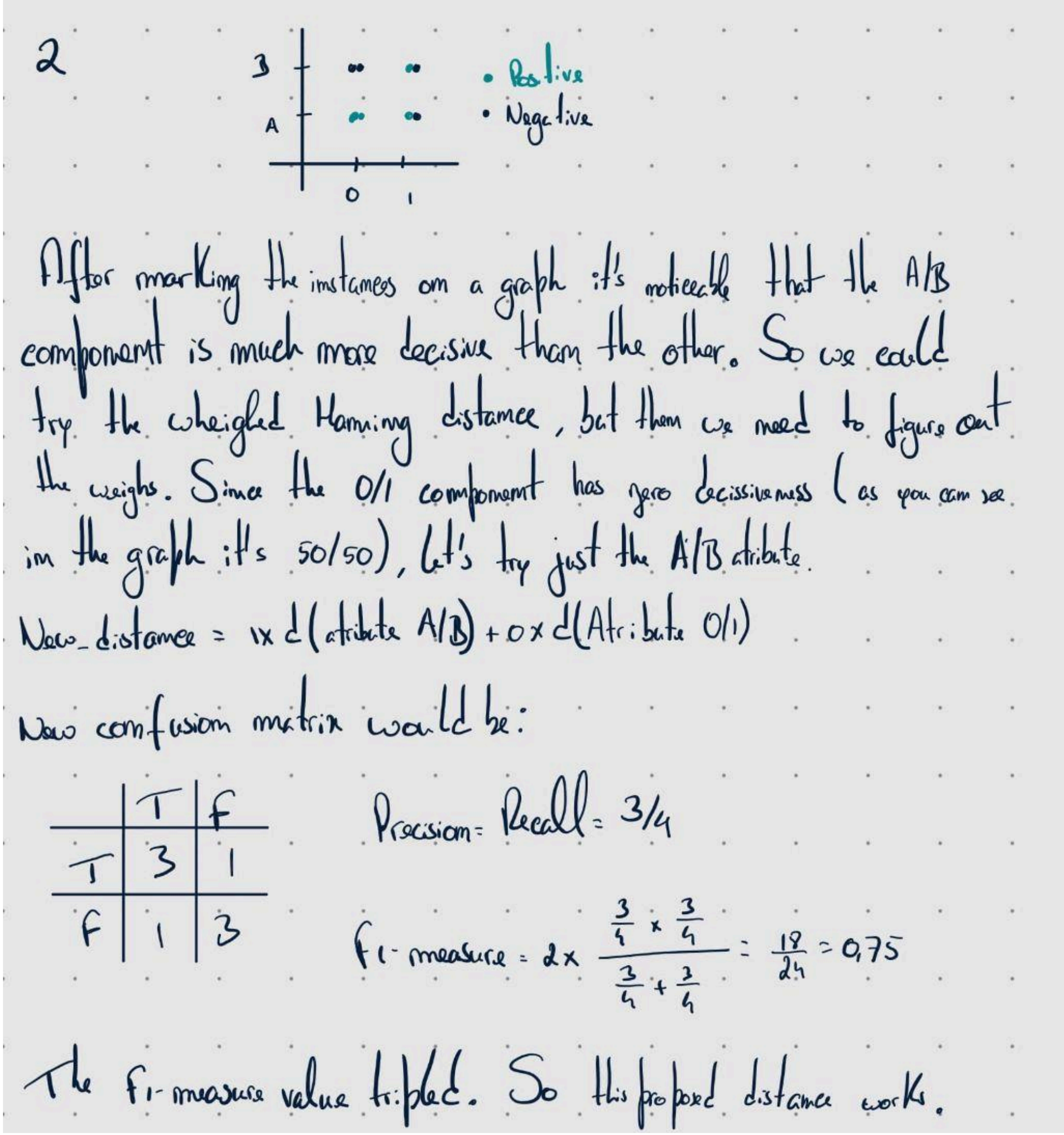
$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{1}{4}$$

$$F_1\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{1/8}{1/2} = 0,25$$

R: The F_1- measure is 0,25.

2) [2.5v] Propose a new metric (distance) that improves the latter's performance (i.e., the F1-measure) by three fold.


B
A
0
1
• Positive
• Negative

2

After marking the instances on a graph it's noticeable that the A/B component is much more decisive than the other. So we could try the wheighted Hamming distance, but then we need to figure out the weighs. Since the 0/1 component has zero decissiveness (as you can see in the graph it's 50/50), let's try just the A/B atribute.

New_distance = 1× d(atribute A/B) + 0 × d(Atribute 0/1)

New confusion matrix would be:

	T	F
T	3	1
F	1	3

Precision = Recall = 3/4

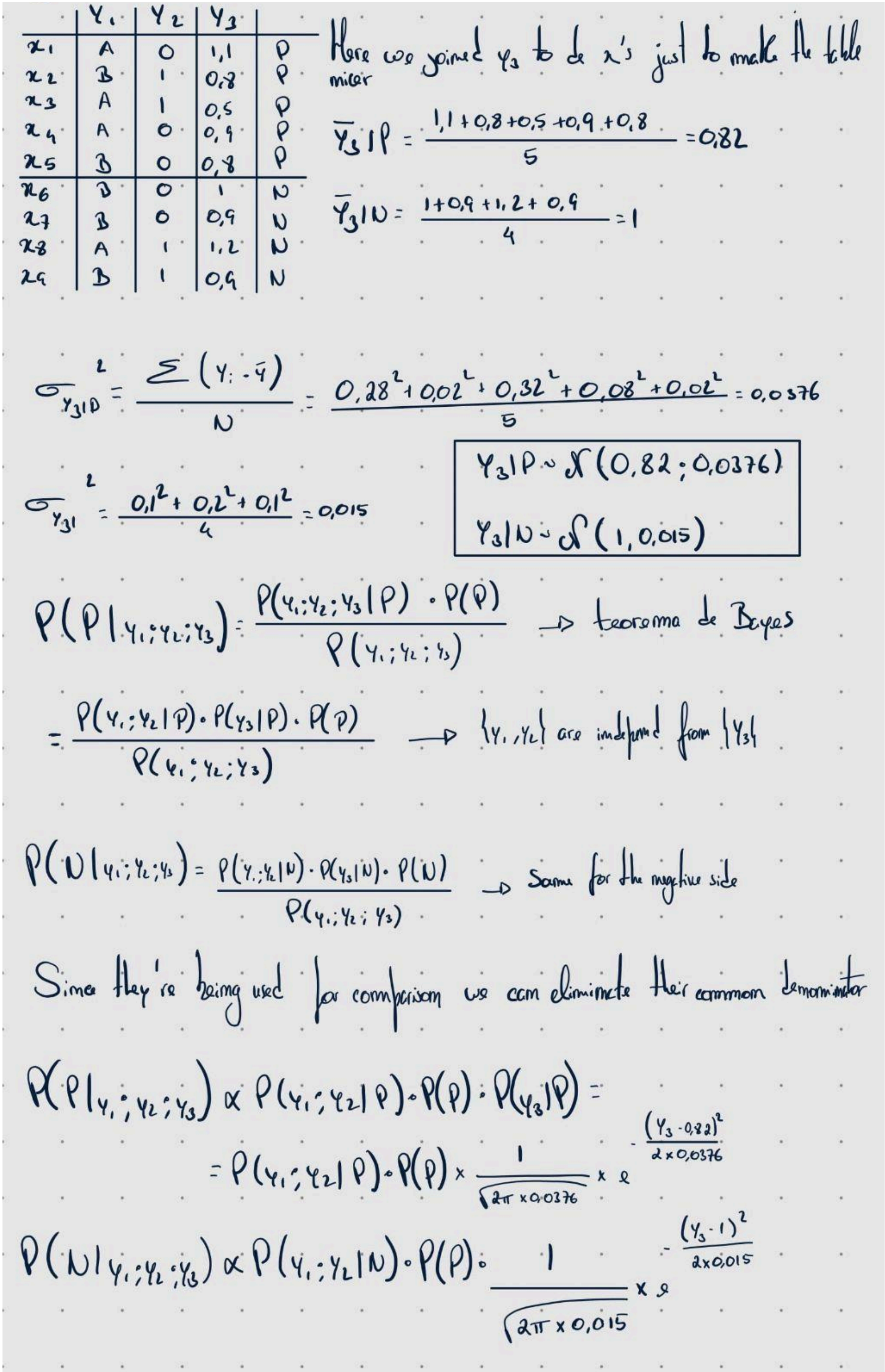
$$F_1\text{- measure} = 2 \times \frac{\frac{3}{4} \times \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{18}{24} = 0,75$$

The F_1-measure value tripled. So this proposed distance works.

An additional positive observation was acquired, $x_9 = (B, 0)$, and a third variable y_3 was independently monitored, yielding estimates,

$$y_3|P = \{1.1,\ 0.8,\ 0.5,\ 0.9,\ 0.8\} \text{ and } y_3|N = \{1,\ 0.9,\ 1.2,\ 0.9\}.$$

3) [2.5v] Considering the nine training observations, learn a Bayesian classifier assuming: i) y_1 and y_2 are dependent; ii) $\{y_1, y_2\}$ and $\{y_3\}$ variable sets are independent and equally important; and iii) y_3 is normally distributed. Show all parameters.



	Y_1	Y_2	Y_3	
x_1	A	0	1,1	P
x_2	B	1	0,8	P
x_3	A	1	0,5	P
x_4	A	0	0,9	P
x_5	B	0	0,8	P
x_6	B	0	1	N
x_7	B	0	0,9	N
x_8	A	1	1,2	N
x_9	B	1	0,9	N

Here we joined y_3 to de x's just to make the table nicer

$$\overline{Y_3}|P = \frac{1{,}1 + 0{,}8 + 0{,}5 + 0{,}9 + 0{,}8}{5} = 0{,}82$$

$$\overline{Y_3}|N = \frac{1 + 0{,}9 + 1{,}2 + 0{,}9}{4} = 1$$

$$\sigma^2_{Y_3|P} = \frac{\Sigma(Y_i - \bar{Y})}{N} = \frac{0{,}28^2 + 0{,}02^2 + 0{,}32^2 + 0{,}08^2 + 0{,}02^2}{5} = 0{,}0376$$

$$\sigma^2_{Y_3|} = \frac{0{,}1^2 + 0{,}2^2 + 0{,}1^2}{4} = 0{,}015$$

$$\boxed{\begin{array}{l} Y_3|P \sim \mathcal{N}(0{,}82 ; 0{,}0376) \\ Y_3|N \sim \mathcal{N}(1, 0{,}015) \end{array}}$$

$$P(P|Y_1;Y_2;Y_3) = \frac{P(Y_1;Y_2;Y_3|P) \cdot P(P)}{P(Y_1;Y_2;Y_3)}$$ → teorema de Bayes

$$= \frac{P(Y_1;Y_2|P) \cdot P(Y_3|P) \cdot P(P)}{P(Y_1;Y_2;Y_3)}$$ → $\{Y_1, Y_2\}$ are independ from $\{Y_3\}$

$$P(N|Y_1;Y_2;Y_3) = \frac{P(Y_1;Y_2|N) \cdot P(Y_3|N) \cdot P(N)}{P(Y_1;Y_2;Y_3)}$$ → Same for the negative side

Since they're being used for comparison we can eliminate their common denominator

$$P(P|Y_1;Y_2;Y_3) \propto P(Y_1;Y_2|P) \cdot P(P) \cdot P(Y_3|P) =$$

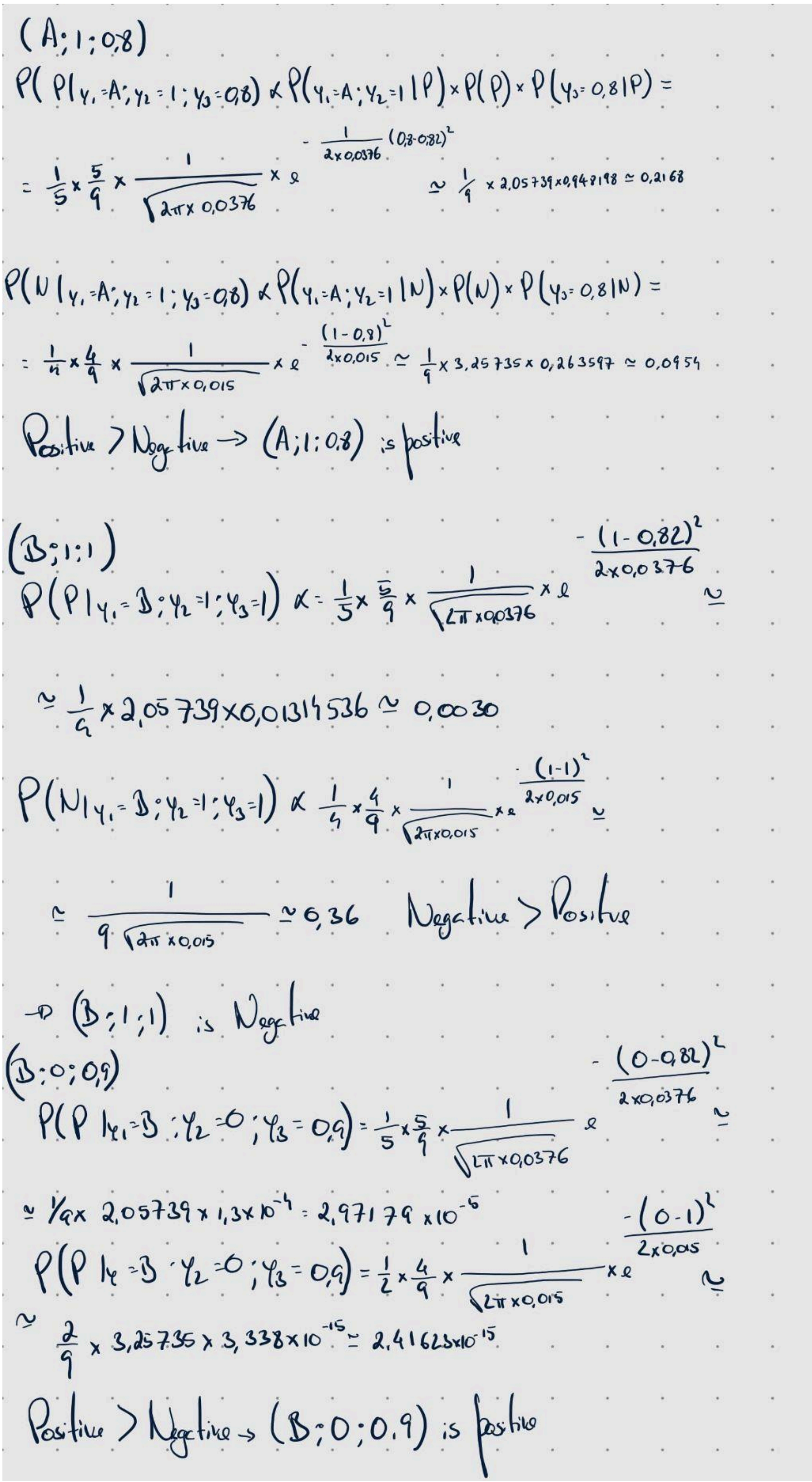
$$= P(Y_1;Y_2|P) \cdot P(P) \times \frac{1}{\sqrt{2\pi \times 0{,}0376}} \times e^{-\frac{(Y_3 - 0{,}82)^2}{2 \times 0{,}0376}}$$

$$P(N|Y_1;Y_2;Y_3) \propto P(Y_1;Y_2|N) \cdot P(P) \cdot \frac{1}{\sqrt{2\pi \times 0{,}015}} \times e^{-\frac{(Y_3 - 1)^2}{2 \times 0{,}015}}$$

Consider now three testing observations,

$$\{(A,\ 1,\ 0.8), (B,\ 1,\ 1), (B,\ 0,\ 0.9)\}.$$

4) [2.5v] Under a MAP assumption, classify each testing observation showing all your calculus.



$(A;1;0{,}8)$

$P(P|y_1=A;y_2=1;y_3=0{,}8) \propto P(y_1=A;y_2=1|P)\times P(P)\times P(y_3=0{,}8|P) =$

$= \frac{1}{5}\times\frac{5}{9}\times\frac{1}{\sqrt{2\pi\times 0{,}0376}}\times e^{-\frac{1}{2\times 0{,}0376}(0{,}8-0{,}82)^2} \simeq \frac{1}{9}\times 2{,}05739\times 0{,}948198 \simeq 0{,}2168$

$P(N|y_1=A;y_2=1;y_3=0{,}8) \propto P(y_1=A;y_2=1|N)\times P(N)\times P(y_3=0{,}8|N) =$

$= \frac{1}{4}\times\frac{4}{9}\times\frac{1}{\sqrt{2\pi\times 0{,}015}}\times e^{-\frac{(1-0{,}8)^2}{2\times 0{,}015}} \simeq \frac{1}{9}\times 3{,}25735\times 0{,}263597 \simeq 0{,}0954$

Positive > Negative → $(A;1:0.8)$ is positive

$(B;1:1)$

$P(P|y_1=B;y_2=1;y_3=1) \propto = \frac{1}{5}\times\frac{5}{9}\times\frac{1}{\sqrt{2\pi\times 0{,}0376}}\times e^{-\frac{(1-0{,}82)^2}{2\times 0{,}0376}} \simeq$

$\simeq \frac{1}{9}\times 2{,}05739\times 0{,}01314536 \simeq 0{,}0030$

$P(N|y_1=B;y_2=1;y_3=1) \propto \frac{1}{4}\times\frac{4}{9}\times\frac{1}{\sqrt{2\pi\times 0{,}015}}\times e^{-\frac{(1-1)^2}{2\times 0{,}015}} \simeq$

$\simeq \frac{1}{9\sqrt{2\pi\times 0{,}015}} \simeq 0{,}36$ Negative > Positive

→ $(B;1;1)$ is Negative

$(B;0;0{,}9)$

$P(P|y_1=B;y_2=0;y_3=0{,}9) = \frac{1}{5}\times\frac{5}{9}\times\frac{1}{\sqrt{2\pi\times 0{,}0376}}\, e^{-\frac{(0-0{,}82)^2}{2\times 0{,}0376}} \simeq$

$\simeq 1/9\times 2{,}05739\times 1{,}3\times 10^{-4} = 2{,}97179\times 10^{-5}$

$P(P|y_1=B;y_2=0;y_3=0{,}9) = \frac{1}{2}\times\frac{4}{9}\times\frac{1}{\sqrt{2\pi\times 0{,}015}}\times e^{-\frac{(0-1)^2}{2\times 0{,}05}} \simeq$

$\simeq \frac{2}{9}\times 3{,}25735\times 3{,}338\times 10^{-15} = 2{,}41623\times 10^{-15}$

Positive > Negative → $(B;0;0.9)$ is positive

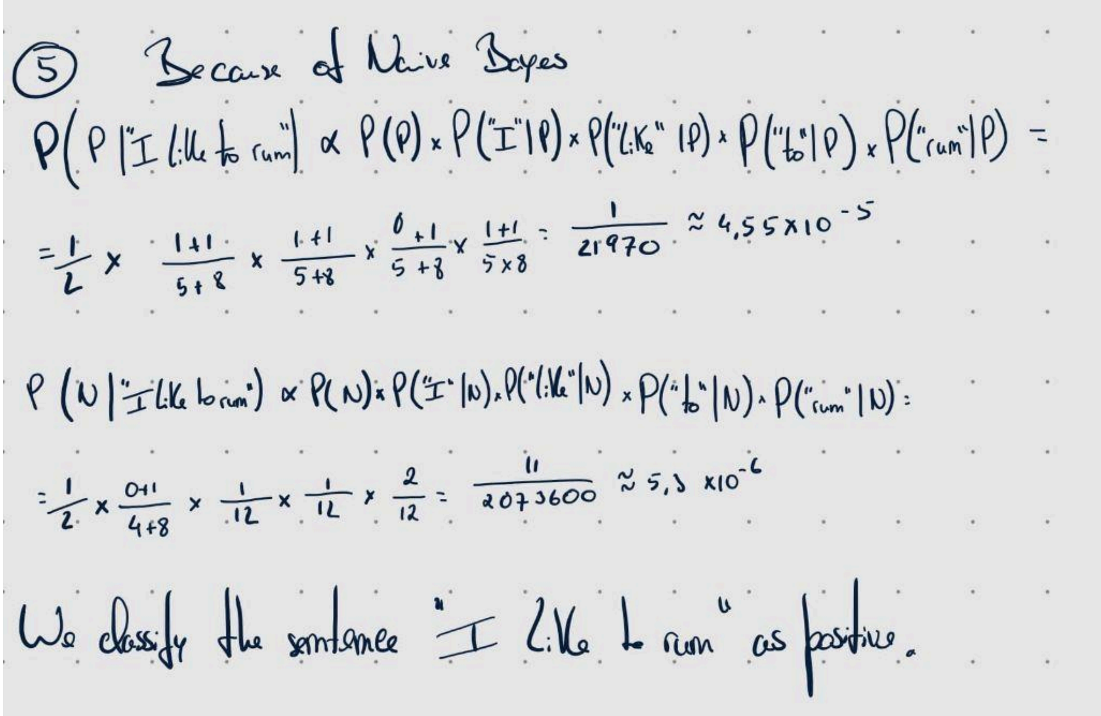
At last, consider only the following sentences and their respective connotations,

$$\{(\text{"}Amazing\ run\text{"},\ P),\ (\text{"}I\ like\ it\text{"},\ P),\ (\text{"}Too\ tired\text{"},\ N),\ (\text{"}Bad\ run\text{"},\ N)\}.$$

5) [2.5v] Using a naïve Bayes under a ML assumption, classify the new sentence "*I like to run*". For the likelihoods calculation consider the following formula,

$$p(t_i|c) = (freq(t_i) + 1)/(N_c + V),$$

where t_i represents a certain term i, V the number of unique terms in the vocabulary, and N_c the total number of terms in class c. Show all calculus.



(5) Because of Naive Bayes

$$P\left(P \,|\, \text{"I like to run"}\right) \propto P(P) \times P(\text{"I"}|P) \times P(\text{"like"}|P) \times P(\text{"to"}|P) \times P(\text{"run"}|P) =$$

$$= \frac{1}{2} \times \frac{1+1}{5+8} \times \frac{1+1}{5+8} \times \frac{0+1}{5+8} \times \frac{1+1}{5\times 8} = \frac{1}{21970} \approx 4{,}55 \times 10^{-5}$$

$$P\left(N \,|\, \text{"I like to run"}\right) \propto P(N) \times P(\text{"I"}|N) \cdot P(\text{"like"}|N) \times P(\text{"to"}|N) \cdot P(\text{"run"}|N) =$$

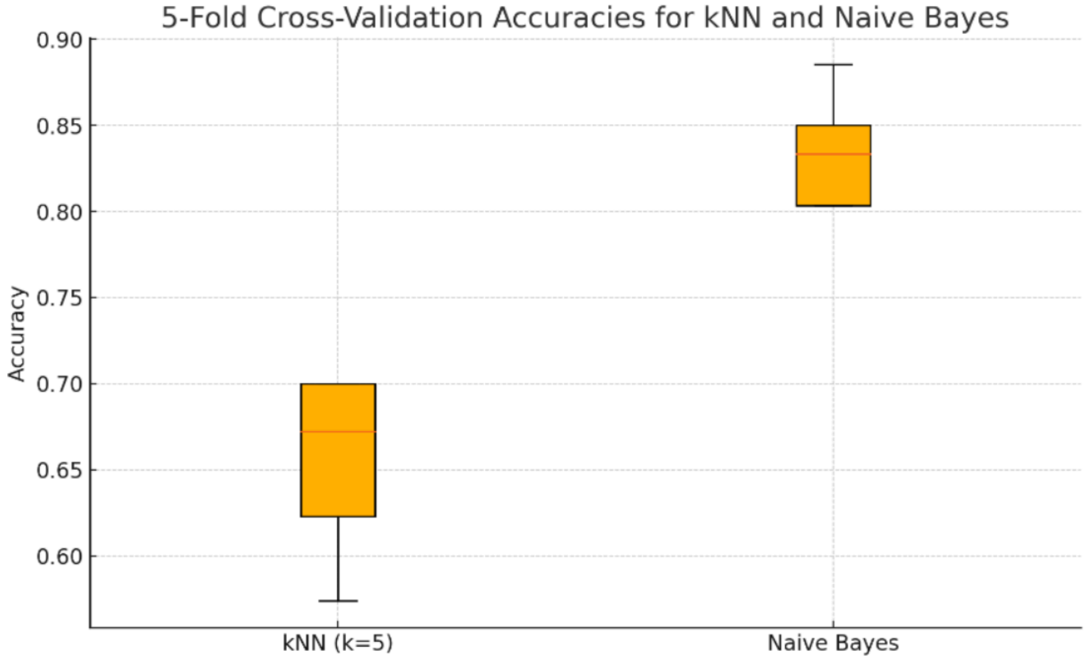
$$= \frac{1}{2} \times \frac{0+1}{4+8} \times \frac{1}{12} \times \frac{1}{12} \times \frac{2}{12} = \frac{11}{2073600} \approx 5{,}3 \times 10^{-6}$$

We classify the sentence "I like to run" as positive.

Part II: Programming and critical analysis

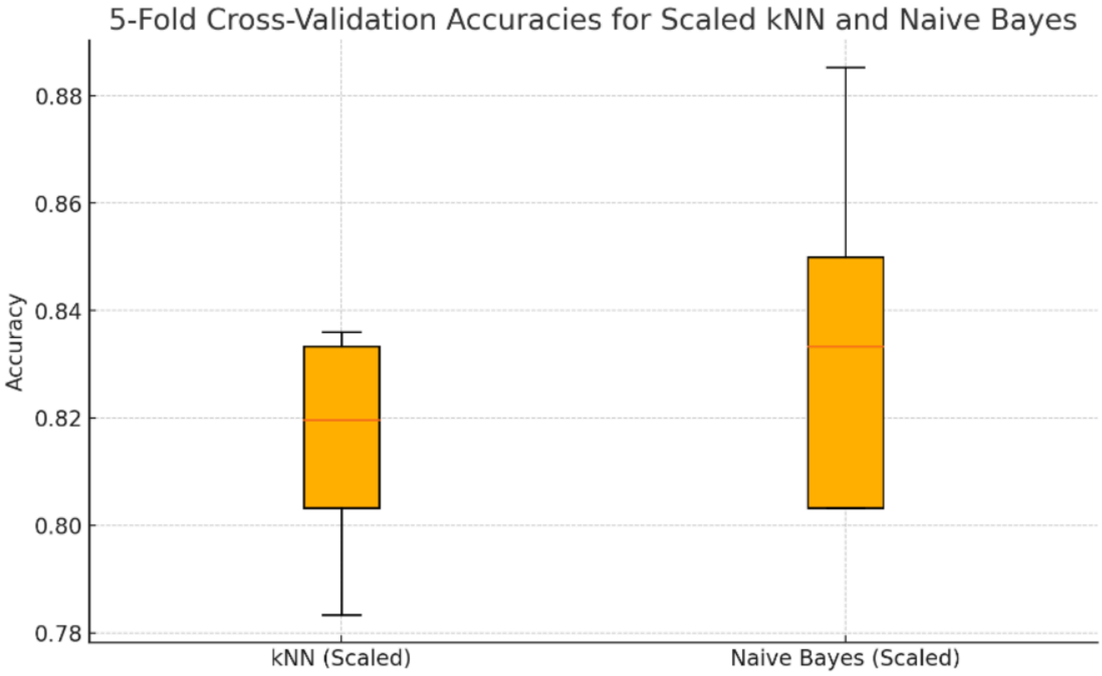
1) Compare the performance of a kNN with $k = 5$ and a naïve Bayes with Gaussian assumption (consider all remaining parameters as default):

    a. [1.0v] Plot two boxplots with the fold accuracies for each classifier. Is there one more stable than the other regarding performance? Why do you think that is the case? Explain.



For k-NN, the accuracy scores range from about 0.57 to 0.70, with a noticeable variance between folds. For the Naive Bayes the accuracy is much higher and more stable, ranging from 0.80 to 0.88 across the folds. The Naive Bayes model seems to be more stable than the k-NN model, as evidenced by the smaller range of accuracy values and a higher median performance. One possible reason for this difference is that k-NN's performance is more sensitive to the distribution of the data and the distance between points in feature space, while Naive Bayes assumes independence between features, which can sometimes generalize better across folds when the data aligns with those assumptions.

b. [1.0v] Report the accuracy of both models, this time scaling the data with a Min-Max scaler before training the models. Explain the impact that this preprocessing step has on the performance of each model, providing an explanation for the results.



After scaling the data using a Min-Max scaler the accuracy for k-NN (Scaled) significantly improved compared to the unscaled data. The accuracy now ranges from approximately 0.78 to 0.84, with a median close to 0.82 and for Naive Bayes (Scaled) there was little change in performance after scaling. The accuracies remain within the range of 0.80 to 0.88.

Scaling had a positive effect on the k-NN model's performance, as k-NN is a distance-based classifier and sensitive to the feature scale. When features are on different scales, distance computations can be skewed, leading to poorer performance. Naive Bayes, on the other hand, is based on probabilities and assumptions of feature independence, so scaling had little to no effect on its performance.

c. [1.0v] Using `scipy`, test the hypothesis “the kNN model is statistically superior to naïve Bayes regarding accuracy”, asserting whether it is true.

accuracy k-NN: [0.8197, 0.8033, 0.8197, 0.7667, 0.8000]

accuracy Naive Bayes: [0.8033, 0.8852, 0.7869, 0.8333, 0.7167]

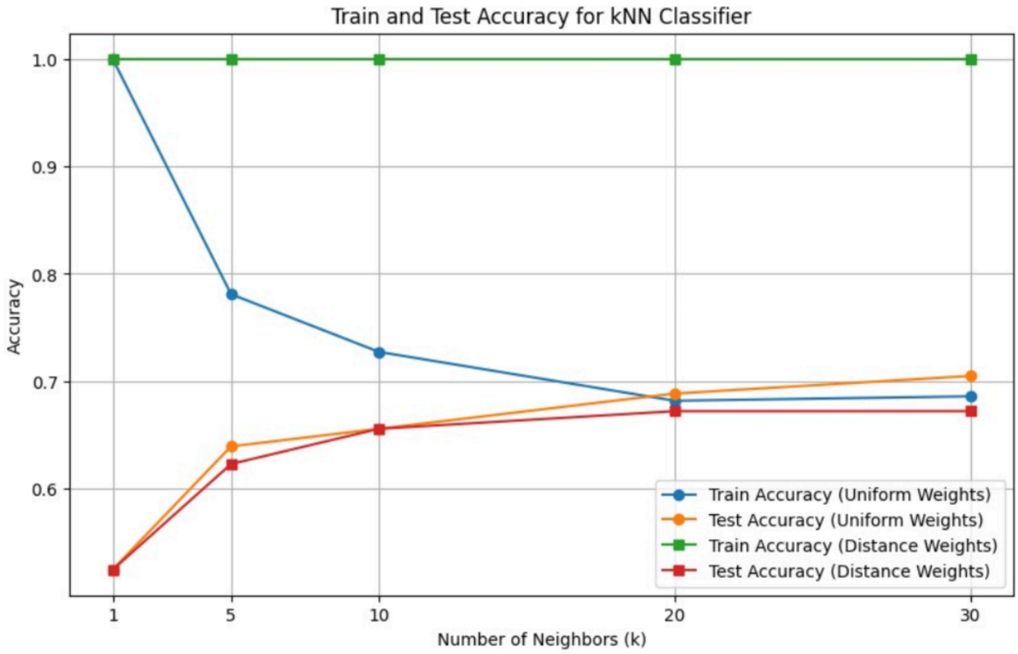
statistics t: -0.1098

p-value: 0.9153

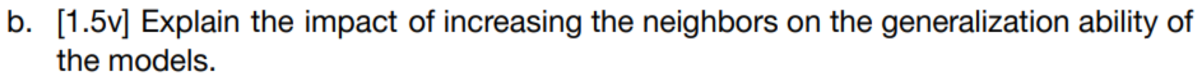
The p-value 0.9153 is much higher than the acceptable significance level 0.05, which means that we cannot reject the null hypothesis. This indicates that there is no statistically significant difference in accuracy between the k-NN model with k = 5 and the Naive Bayes model for this dataset.

2) Using a 80-20 train-test split, vary the number of neighbors of a kNN classifier using $k = \{1, 5, 10, 20, 30\}$. Additionally, for each k, train one classifier using uniform weights and distance weights.

a. [1.0v] Plot the train and test accuracy for each model.



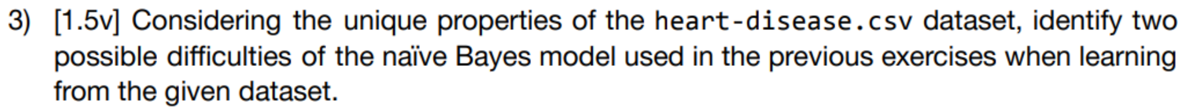
The graph shows the train and test accuracy for different values of k (number of neighbors) using uniform weights and distance weights. For uniform weights, the train accuracy decreases as k increases, starting high at k=1 and continuing to decrease up to k=30. The test accuracy initially increases from k=1 to k=5, but then begins to decrease as k increases further. For distance weights, the train accuracy also decreases as k increases, but the decrease is less pronounced compared to uniform weights. The test accuracy follows a similar pattern to uniform weights, increasing initially from k=1 to k=5 and then decreasing with larger values of k.



b. [1.5v] Explain the impact of increasing the neighbors on the generalization ability of the models.

Increasing the number of neighbors k in a k-NN classifier impacts the model's generalization ability. With fewer neighbors small k, the model can be very sensitive to noise in the training data, leading to overfitting. This means the model performs well on training data but poorly on unseen test data because it memorizes the training data rather than generalizing from it. As more neighbors are considered larger k, the predictions become smoother because each decision is based on a larger subset of points, which tends to average out noise or outliers. However, this can also lead to underfitting if too many neighbors are used, as distinct categories may start getting averaged together, resulting in poorer performance on both train and test sets. Using uniform weights means all neighbors contribute equally regardless of their distance from the point being classified, whereas distance weights give more influence to closer points, which can help mitigate some overfitting issues seen with smaller values of k when using uniform weights since closer points are likely more similar than distant ones.

Overall, there is a balance that needs to be struck between too few and too many neighbors, depending on the specific dataset characteristics such as its dimensionality and noise level, among other factors, for optimal generalization ability.



3) [1.5v] Considering the unique properties of the `heart-disease.csv` dataset, identify two possible difficulties of the naïve Bayes model used in the previous exercises when learning from the given dataset.

Some of the difficulties for the Naïve Bayes Model with the Heart-Disease Dataset Feature Independence Assumption are that Naïve Bayes assumes that all features are independent given the target label. However, in datasets like heart disease, many features are correlated for example blood pressure and cholesterol levels or age and maximum heart rate. This violation of the independence assumption can lead to poor model performance, as Naïve Bayes cannot properly model the interactions between related features. Other difficulty is Continuous Data Handling. Naïve Bayes with Gaussian assumptions handles continuous features by fitting a normal distribution to each feature. However, in datasets like heart-disease, some features may not follow a Gaussian distribution for example cholesterol or oldpeak. This mismatch between the actual feature distribution and the assumed Gaussian distribution can affect the accuracy of the model, as the probability estimates become less reliable. These challenges can limit Naïve Bayes' ability to perform well compared to more flexible models like kNN, which do not make strong distributional assumptions.