Aprendizagem 2024
Homework II - Group 92
(Maria Maló - ist102994,
André Feliciano - ist100286)

Part I: Pen and paper

1. **Leaving x_1 out**

$$\text{hamming } (x_1, x_2) = 2$$
$$\text{hamming } (x_1, x_3) = 1$$
$$\text{hamming } (x_1, x_4) = 0$$
$$\text{hamming } (x_1, x_5) = 1$$
$$\text{hamming } (x_1, x_6) = 1$$
$$\text{hamming } (x_1, x_7) = 1$$
$$\text{hamming } (x_1, x_8) = 2$$

The nearest points are: x_3, x_4, x_5, x_6, and x_7

- x_3, x_4: Positive (P)
- x_5, x_6, x_7: Negative (N)

The majority of these points classify x_1 as Negative, so we'll classify x_1 as Negative.

Leaving x_2 out

$$\text{hamming } (x_2, x_1) = 2$$
$$\text{hamming } (x_2, x_3) = 1$$
$$\text{hamming } (x_2, x_4) = 2$$
$$\text{hamming } (x_2, x_5) = 1$$
$$\text{hamming } (x_2, x_6) = 1$$
$$\text{hamming } (x_2, x_7) = 1$$
$$\text{hamming } (x_2, x_8) = 0$$

The nearest points are: x_3, x_5, x_6, x_7, and x_8

- x_3: Positive (P)
- x_5, x_6, x_7, x_8: Negative (N)

x_2 is classified as Negative.

Leaving x_3 out

$$
\begin{aligned}
\text{hamming }(x_3, x_1) &= 1\\
\text{hamming }(x_3, x_2) &= 1\\
\text{hamming }(x_3, x_4) &= 1\\
\text{hamming }(x_3, x_5) &= 2\\
\text{hamming }(x_3, x_6) &= 2\\
\text{hamming }(x_3, x_7) &= 0\\
\text{hamming }(x_3, x_8) &= 1
\end{aligned}
$$

The nearest points are: x_1, x_2, x_4, x_7, and x_8

- x_1, x_2, x_4: Positive (P)
- x_7, x_8: Negative (N)

Then, x_3 is classified as Positive.

Leaving x_4 out

$$
\begin{aligned}
\text{hamming }(x_4, x_1) &= 0\\
\text{hamming }(x_4, x_2) &= 2\\
\text{hamming }(x_4, x_3) &= 1\\
\text{hamming }(x_4, x_5) &= 1\\
\text{hamming }(x_4, x_6) &= 1\\
\text{hamming }(x_4, x_7) &= 1\\
\text{hamming }(x_4, x_8) &= 2
\end{aligned}
$$

The nearest points are: x_1, x_3, x_5, x_6, and x_7

- x_1, x_3: Positive (P)
- x_5, x_6, x_7: Negative (N)

Then, x_4 is classified as Negative.

Leaving x_5 out

$$\text{hamming}\ (x_5, x_1) = 1$$
$$\text{hamming}\ (x_5, x_2) = 1$$
$$\text{hamming}\ (x_5, x_3) = 2$$
$$\text{hamming}\ (x_5, x_4) = 1$$
$$\text{hamming}\ (x_5, x_6) = 0$$
$$\text{hamming}\ (x_5, x_7) = 2$$
$$\text{hamming}\ (x_5, x_8) = 1$$

The nearest points are: x_1, x_2, x_4, x_6, and x_8

- x_1, x_2, x_4: Positive (P)
- x_6, x_8: Negative (N)

Then, x_5 is classified as Positive.

Leaving x_6 out

$$\text{hamming}\ (x_6, x_1) = 1$$
$$\text{hamming}\ (x_6, x_2) = 1$$
$$\text{hamming}\ (x_6, x_3) = 2$$
$$\text{hamming}\ (x_6, x_4) = 1$$
$$\text{hamming}\ (x_6, x_5) = 0$$
$$\text{hamming}\ (x_6, x_7) = 2$$
$$\text{hamming}\ (x_6, x_8) = 1$$

The nearest points are: x_1, x_2, x_4, x_5, and x_8

- x_1, x_2, x_4: Positive (P)
- x_5, x_8: Negative (N)

Then, x_6 is classified as Positive.

Leaving x_7 out

$$\text{hamming}\ (x_7, x_1) = 1$$
$$\text{hamming}\ (x_7, x_2) = 1$$
$$\text{hamming}\ (x_7, x_3) = 0$$
$$\text{hamming}\ (x_7, x_4) = 1$$
$$\text{hamming}\ (x_7, x_5) = 2$$
$$\text{hamming}\ (x_7, x_6) = 2$$
$$\text{hamming}\ (x_7, x_8) = 1$$

The nearest points are: x_1, x_2, x_3, x_4, and x_8

- x_1, x_2, x_3, x_4: Positive (P)
- x_8: Negative (N)

Then, x_7 is classified as Positive.

Leaving x_8 out

$$\begin{aligned}
\text{hamming } (x_8, x_1) &= 2 \\
\text{hamming } (x_8, x_2) &= 0 \\
\text{hamming } (x_8, x_3) &= 1 \\
\text{hamming } (x_8, x_4) &= 2 \\
\text{hamming } (x_8, x_5) &= 1 \\
\text{hamming } (x_8, x_6) &= 1 \\
\text{hamming } (x_8, x_7) &= 1
\end{aligned}$$

The nearest points are: x_2, x_3, x_5, x_6, and x_7

- x_2, x_3 : Positive (P)
- x_5, x_6, x_7: Negative (N)

Then, x_8 is classified as Negative.

Comparing the results of our model with the actual ones, we obtained:

- 1 True Positive
- 3 False Positives
- 1 True Negative
- 3 False Negatives

Therefore, one can calculate the F1-measure of this model, knowing the precision and recall values.

Precision

$$\text{p} = \frac{TP}{TP + FP} = \frac{1}{1+3} = \frac{1}{4} = 0.25$$

Recall

$$\text{r} = \frac{TP}{TP + FN} = \frac{1}{1+3} = 0.25$$

F1-Measure

$$F_1 = \frac{2}{r^{-1} + p^{-1}} = \frac{2}{4+4} = \frac{2}{8} = 0.25$$

2. The Hamming distance we are using infers that both the instances x_1 (A or B) and x_2 (0 or 1) have the same weight on determining the classification of our observations. However, this is not entirely true. By observing our data set, one can see that the first instance x_1 seems to have more "power" on the outcome of our classification. When we observe an 'A' it is more likely that we will have a positive evaluation, and 'B' a negative classification (in 3/4 of the observations this happens). Having observed '0' and '1' doesn't seem to interfere with the classification. So we suggest a metric that gives more weight to this first instance (A or B), a weighted type of Hamming distance.

   New Hamming distance:

$$d(x_i, x_j) = w_1 \cdot \mathbb{I}(x_{i1} \neq x_{j1}) + w_2 \cdot \mathbb{I}(x_{i2} \neq x_{j2})$$

$$w_1 = 2 \quad \text{(gives more weight on first instance)}$$

$$w_2 = 1$$

$$\mathbb{I}(x_{i1} \neq x_{j1}) = \begin{cases} 1 & \text{if } x_{i1} \neq x_{j1} \\ 0 & \text{if } x_{i1} = x_{j1} \end{cases}$$

   Moreover, one can reflect on the number of k nearest points used. We believe 5 nearest points is indeed too much. Reducing k from 5 to 3 would help in reducing the effect of noisy or irrelevant neighbours, ensuring that our model is more focused on the most similar points, helping in avoiding the bias introduced by including too many neighbours.

   Keeping all these improvements in mind, we shall redo all our calculations as before.

   Leaving x_1 out

$$\begin{aligned}
\mathrm{d}(x_1, x_2) &= 3 \\
\mathrm{d}(x_1, x_3) &= 1 \\
\mathrm{d}(x_1, x_4) &= 0 \\
\mathrm{d}(x_1, x_5) &= 2 \\
\mathrm{d}(x_1, x_6) &= 2 \\
\mathrm{d}(x_1, x_7) &= 1 \\
\mathrm{d}(x_1, x_8) &= 3
\end{aligned}$$

   The nearest points are: x_3, x_4, and x_7

   - x_3, x_4: Positive (P)
   - x_7: Negative (N)

The majority of these points classify x_1 as Positive, so we'll classify x_1 as Positive.

Leaving x_2 out

$$
\begin{aligned}
\mathrm{d}(x_2, x_1) &= 3 \\
\mathrm{d}(x_2, x_3) &= 2 \\
\mathrm{d}(x_2, x_4) &= 3 \\
\mathrm{d}(x_2, x_5) &= 1 \\
\mathrm{d}(x_2, x_6) &= 1 \\
\mathrm{d}(x_2, x_7) &= 2 \\
\mathrm{d}(x_2, x_8) &= 0
\end{aligned}
$$

The nearest points are: x_5, x_6 and x_8

- x_5, x_6, x_8: Negative (N)

x_2 is classified as Negative.

Leaving x_3 out

$$
\begin{aligned}
\mathrm{d}(x_3, x_1) &= 1 \\
\mathrm{d}(x_3, x_2) &= 2 \\
\mathrm{d}(x_3, x_4) &= 1 \\
\mathrm{d}(x_3, x_5) &= 3 \\
\mathrm{d}(x_3, x_6) &= 3 \\
\mathrm{d}(x_3, x_7) &= 0 \\
\mathrm{d}(x_3, x_8) &= 2
\end{aligned}
$$

The nearest points are: x_1, x_4 and x_7

- x_1, x_4: Positive (P)
- x_7: Negative (N)

Then, x_3 is classified as Positive.

Leaving x_4 out

$$
\begin{aligned}
d(x_4, x_1) &= 0 \\
d(x_4, x_2) &= 3 \\
d(x_4, x_3) &= 1 \\
d(x_4, x_5) &= 2 \\
d(x_4, x_6) &= 2 \\
d(x_4, x_7) &= 1 \\
d(x_4, x_8) &= 3
\end{aligned}
$$

The nearest points are: x_1, x_3 and x_7

- x_1, x_3: Positive (P)
- x_7: Negative (N)

Then, x_4 is classified as Positive.

Leaving x_5 out

$$
\begin{aligned}
d(x_5, x_1) &= 2 \\
d(x_5, x_2) &= 1 \\
d(x_5, x_3) &= 3 \\
d(x_5, x_4) &= 2 \\
d(x_5, x_6) &= 0 \\
d(x_5, x_7) &= 3 \\
d(x_5, x_8) &= 1
\end{aligned}
$$

The nearest points are: x_2, x_6, and x_8

- x_2: Positive (P)
- x_6, x_8: Negative (N)

Then, x_5 is classified as Negative.

Leaving x_6 out

$$
\begin{aligned}
d(x_6, x_1) &= 2\\
d(x_6, x_2) &= 1\\
d(x_6, x_3) &= 3\\
d(x_6, x_4) &= 2\\
d(x_6, x_5) &= 0\\
d(x_6, x_7) &= 3\\
d(x_6, x_8) &= 1
\end{aligned}
$$

The nearest points are: x_2, x_5, and x_8

- x_2: Positive (P)
- x_5, x_8: Negative (N)

Then, x_6 is classified as Negative.

Leaving x_7 out

$$
\begin{aligned}
d(x_7, x_1) &= 1\\
d(x_7, x_2) &= 2\\
d(x_7, x_3) &= 0\\
d(x_7, x_4) &= 1\\
d(x_7, x_5) &= 3\\
d(x_7, x_6) &= 3\\
d(x_7, x_8) &= 2
\end{aligned}
$$

The nearest points are: x_1, x_3 and x_4

- x_1, x_3, x_4: Positive (P)

Then, x_7 is classified as Positive.

Leaving x_8 out

$$
\begin{aligned}
d(x_8, x_1) &= 3\\
d(x_8, x_2) &= 0\\
d(x_8, x_3) &= 2\\
d(x_8, x_4) &= 3\\
d(x_8, x_5) &= 1\\
d(x_8, x_6) &= 1\\
d(x_8, x_7) &= 2
\end{aligned}
$$

The nearest points are: x_2, x_5, and x_6

- x_2: Positive (P)
- x_5, x_6: Negative (N)

Then, x_8 is classified as Negative.

Comparing the results of our new model with the actual ones, we obtained:

- 3 True Positive
- 1 False Positives
- 3 True Negative
- 1 False Negatives

Therefore, one can calculate the F1-measure of this model, knowing the precision and recall values.

Precision

$$\text{p} = \frac{TP}{TP + FP} = \frac{3}{3+1} = \frac{3}{4} = 0.75$$

Recall

$$\text{r} = \frac{TP}{TP + FN} = \frac{3}{3+1} = 0.75$$

F1-Measure

$$F_1 = \frac{2}{r^{-1} + p^{-1}} = \frac{2}{\frac{8}{3}} = \frac{6}{8} = 0.75$$

This new Hamming distance, giving more importance to the first variable, and using 3 nearest points instead of 5, seems to enhance our predictions immensely. Our precision, recall and F1-measure enhanced 3 times.

3. This Bayesian classifier is calculated like this:

$$P(C \mid y_1, y_2, y_3) \propto P(y_1, y_2 \mid C)\, P(y_3 \mid C)\, P(C)$$

where $\{y_1, y_2\}$ are dependent, and y_3 is independent, and C is positive or negative.

Using Bayes rule:

$$P(P \mid y_1, y_2, y_3) = \frac{P(y_1, y_2 \mid P)\, P(y_3 \mid P)\, P(P)}{P(y_1, y_2, y_3)}$$

$$P(N \mid y_1, y_2, y_3) = \frac{P(y_1, y_2 \mid N)\, P(y_3 \mid N)\, P(N)}{P(y_1, y_2, y_3)}$$

The prior probabilities are:

$$P(P) = \frac{5}{9}, \quad P(N) = \frac{4}{9}$$

$$P(y_1, y_2 \mid P):$$

$$P(A,0 \mid P) = \frac{2}{5}, \quad P(B,0 \mid P) = \frac{1}{5}, \quad P(A,1 \mid P) = \frac{1}{5}, \quad P(B,1 \mid P) = \frac{1}{5}$$

$$P(y_1, y_2 \mid N):$$

$$P(A,0 \mid N) = 0, \quad P(B,0 \mid N) = \frac{1}{2}, \quad P(A,1 \mid N) = \frac{1}{4}, \quad P(B,1 \mid N) = \frac{1}{4}$$

Gaussian Distribution for y_3

y_3 is normally distributed, so one shall calculate the Gaussian parameters, μ and σ^2, for both the negative and positive observations.

For Positive Observations of y_3:

$$\mu_P = \frac{1.1 + 0.8 + 0.5 + 0.9 + 0.8}{5} = 0.82$$

$$\sigma_P^2 = \frac{\sum_i (x_i - \mu_P)^2}{n-1} = \frac{(1.1 - 0.82)^2 + (0.8 - 0.82)^2 + (0.5 - 0.82)^2 + (0.9 - 0.82)^2 + (0.8 - 0.82)^2}{4} = 0.047$$

For Negative:

$$\mu_N = \frac{1 + 0.9 + 1.2 + 0.9}{4} = 1$$

$$\sigma_N^2 = \frac{\sum_i (x_i - \mu_N)^2}{N-1} = \frac{(1-1)^2 + (0.9-1)^2 + (1.2-1)^2 + (0.9-1)^2}{3} = 0.02$$

The probabilities are then given by:

$$P(y_3 \mid P) = \frac{1}{\sqrt{2\pi \times 0.047}} \exp\left(-\frac{(y_3 - 0.82)^2}{2 \times 0.047}\right)$$

$$P(y_3 \mid N) = \frac{1}{\sqrt{2\pi \times 0.02}} \exp\left(-\frac{(y_3 - 1)^2}{2 \times 0.02}\right)$$

Determining each probability $P(y_1, y_2 \mid C)$, $P(y_3 \mid C)$, $P(C)$ gives us the probability of a new observation being classified as C (positive or negative), and we'll take the highest probability as our classification.

Since $P(y_1, y_2, y_3)$ is the same whether C is negative or positive, it will not influence the classification, so we can just ignore it. We have computed everything needed to classify a new observation.

4. $(A, 1, 0.8)$

If Positive:

$$P(A, 1 \mid P) = \frac{1}{5}$$

$$P(0.8 \mid P) = \frac{1}{\sqrt{2\pi \times 0.047}} \exp\left(-\frac{(0.8 - 0.82)^2}{2 \times 0.047}\right) \approx 1.8324$$

$$P(P) = \frac{5}{9}$$

$$P(P \mid A, 1, 0.8) = \frac{1}{5} \times \frac{5}{9} \times 1.8324 \approx 0.2036$$

If Negative:

$$P(A, 1 \mid N) = \frac{1}{4}$$

$$P(0.8 \mid N) = \frac{1}{\sqrt{2\pi \times 0.02}} \exp\left(-\frac{(0.8 - 1)^2}{2 \times 0.02}\right) \approx 1.038$$

$$P(N) = \frac{4}{9}$$

$$P(N \mid A, 1, 0.8) = 1.038 \times \frac{1}{4} \times \frac{4}{9} = 0.1153$$

Since $P(P \mid A, 1, 0.8) > P(N \mid A, 1, 0.8)$, we classify $(A, 1, 0.8)$ as **Positive**.

—

$(B, 1, 1)$

Positive:

$$P(B, 1 \mid P) = \frac{1}{5}$$

$$P(1 \mid P) = \frac{1}{\sqrt{2\pi \times 0.047}} \exp\left(-\frac{(1 - 0.82)^2}{2 \times 0.047}\right) \approx 1.3036$$

$$P(P) = \frac{5}{9}$$

$$P(P \mid B, 1, 1) = \frac{1}{5} \times \frac{5}{9} \times 1.3036 = 0.1448$$

Negative:

$$P(B, 1 \mid N) = \frac{1}{4}$$

$$P(1 \mid N) = \frac{1}{\sqrt{2\pi \times 0.02}} \exp(0) = 2.8209$$

$$P(N) = \frac{4}{9}$$

$$P(N \mid B, 1, 1) = \frac{1}{4} \times \frac{4}{9} \times 2.8209 = 0.3134$$

$$P(N \mid B, 1, 1) > P(P \mid B, 1, 1) \implies (B, 1, 1) \text{ is classified as } \textbf{Negative}$$

$(B, 0, 0.9)$

Positive:

$$P(B, 0 \mid P) = \frac{1}{5}$$

$$P(P) = \frac{5}{9}$$

$$P(0.9 \mid P) = \frac{1}{\sqrt{2\pi \times 0.047}} \exp\left(-\frac{(0.9 - 0.82)^2}{2 \times 0.047}\right) = 1.7191$$

$$P(P \mid B, 0, 0.9) = \frac{1}{5} \times \frac{5}{9} \times 1.7191 = 0.1910$$

Negative:

$$P(B, 0 \mid N) = \frac{1}{2}$$

$$P(N) = \frac{4}{9}$$

$$P(0.9 \mid N) = \frac{1}{\sqrt{2\pi \times 0.02}} \exp\left(-\frac{(0.9 - 1)^2}{2 \times 0.02}\right) = 2.1970$$

$$P(N \mid B, 0, 0.9) = \frac{1}{2} \times \frac{4}{9} \times 2.1970 = 0.4882$$

$(B, 0, 0.9)$ is classified as **Negative**.

5. "I like to run" is composed of the terms 'I', 'like', 'to', 'run'.

$$N_P = 5 \rightarrow \text{"Amazing", "run", "I", "like", "it"}$$

$$N_N = 4 \rightarrow \text{"Too", "tired", "bad", "run"}$$

$$V = 8 \quad \#(\text{"Amazing", "run", "I", "like", "it", "too", "tired", "bad"})$$

Positive:

$$P(\text{"I"} \mid P) = \frac{\text{freq("I")} + 1}{5+8} = \frac{1+1}{5+8} = \frac{2}{13}$$

$$P(\text{"like"} \mid P) = \frac{\text{freq("like")} + 1}{5+8} = \frac{2}{13}$$

$$P(\text{"to"} \mid P) = \frac{0+1}{5+8} = \frac{1}{13}$$

$$P(\text{"run"} \mid P) = \frac{2}{13}$$

Negative:

$$P(\text{"I"} \mid N) = \frac{1}{12}$$

$$P(\text{"like"} \mid N) = \frac{1}{12}$$

$$P(\text{"to"} \mid N) = \frac{1}{12}$$

$$P(\text{"run"} \mid N) = \frac{2}{12} = \frac{1}{6}$$

$$P(P) = \frac{5}{9}, \quad P(N) = \frac{4}{9}$$

$$P(\text{"I like to run"} \mid P) \propto P(\text{"I"} \mid P) \cdot P(\text{"like"} \mid P) \cdot P(\text{"to"} \mid P) \cdot P(\text{"run"} \mid P)$$

$$P(\text{"I like to run"} \mid P) \propto \left(\frac{2}{13}\right)^3 \cdot \frac{1}{13} \approx 2.80 \times 10^{-4}$$

$$P(\text{"I like to run"} \mid N) \propto P(\text{"I"} \mid N) \cdot P(\text{"like"} \mid N) \cdot P(\text{"to"} \mid N) \cdot P(\text{"run"} \mid N)$$

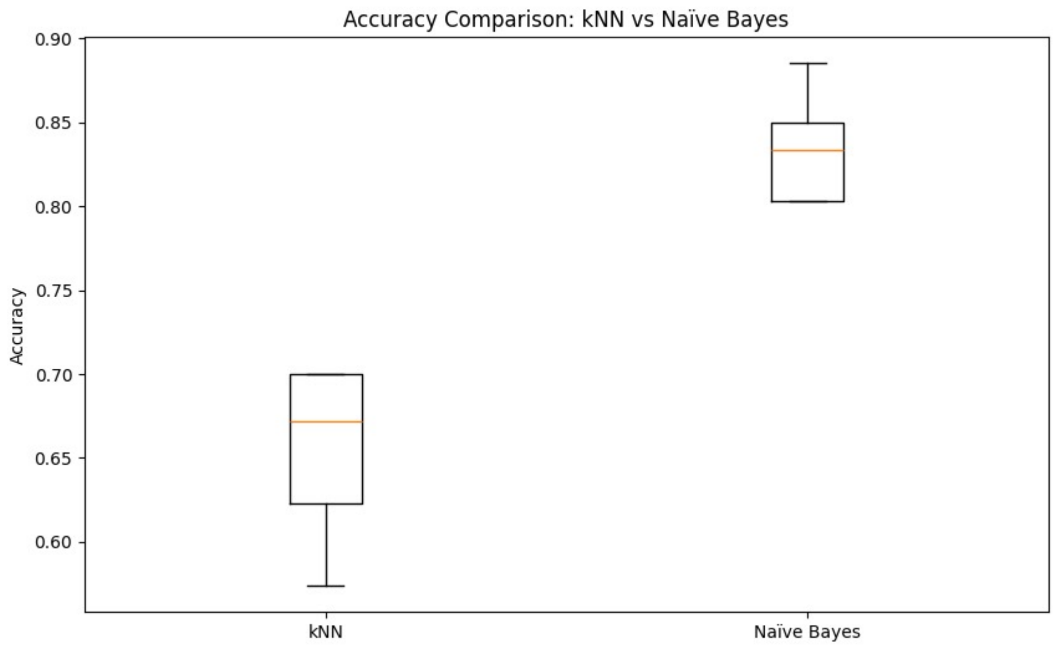
$$P(\text{"I like to run"} \mid N) \propto \left(\frac{1}{12}\right)^3 \cdot \frac{1}{6} \approx 9.65 \times 10^{-5}$$

$$P(\text{"I like to run"} \mid P) > P(\text{"I like to run"} \mid N)$$

So the sentence "I like to run" is classified as **Positive**.

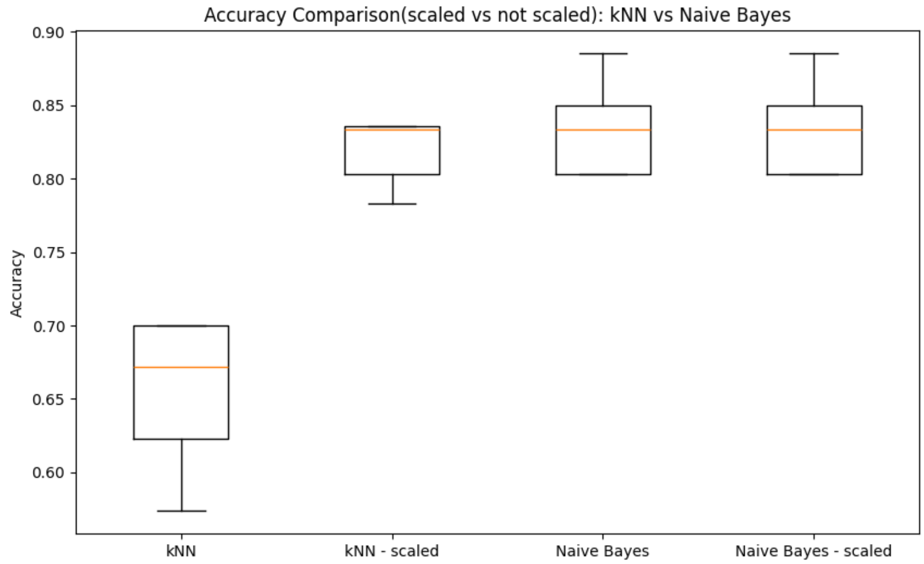
Part II: Programming

1. a)



Both the Interquartile Range and the difference between maximum and minimum values are smaller for the Naïve Bayes model than for kNN model,and this shows that the performance of Naïve Bayes is more stable. It is also true that the accuracy is better on the Naïve Bayes overall. For the kNN model the training data is somewhat small (N=303) and the values from the various classes are very different from one another (for example sex and age, 2 cases for sex, up to 100 for age). Moreover, there are 12 input dimensions, and in such higher dimensionality Naïve-Bayes works better.

b)

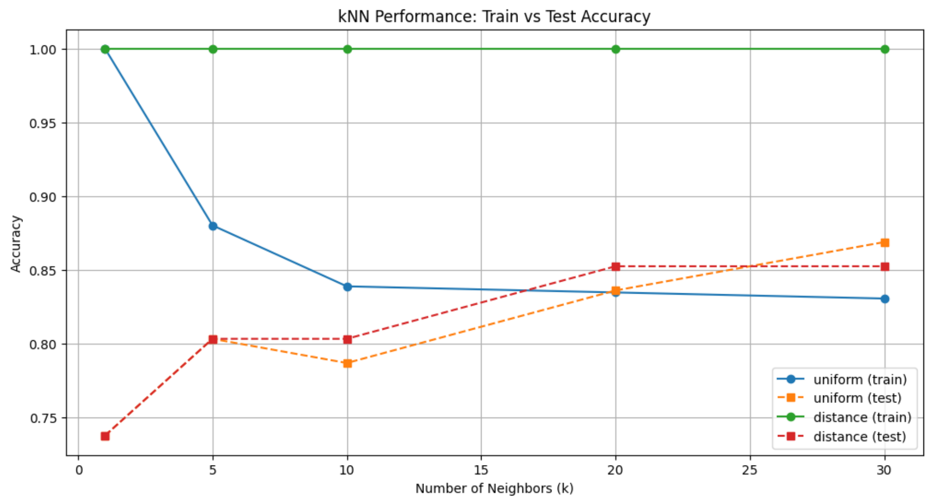


The Min - Max Scaler has a great impact on the performance of the kNN model, whereas it doesn't really impact the performance of the Naïve-Bayes. This comes from the fact that doing a MinMax scaling prevents features with larger scales to dominate the kNN model, on the other side GaussianNB assumes a normal distribution of the parameters and the shape of the distribution doesn't change with the MinMax scaling.

c) After conducting a paired t-test to compare the accuracies of the kNN and Naïve Bayes models, we obtained a p-value of 0.0026, which is lower than the commonly used significance threshold of 0.05. This suggests that the observed performance difference between the two models is statistically significant and unlikely to be due to random variation. Additionally, since Naïve Bayes demonstrated a higher mean accuracy compared to kNN, we conclude that Naïve Bayes outperforms kNN in a statistically meaningful way. As a result, the initial hypothesis that "the kNN model is statistically superior to Naïve Bayes in terms of accuracy" is rejected based on the evidence.

2. We have decided to keep the MinMax scaling given the previous positive results in 1.b) showing a significant improvement over the performance of the kNN model.

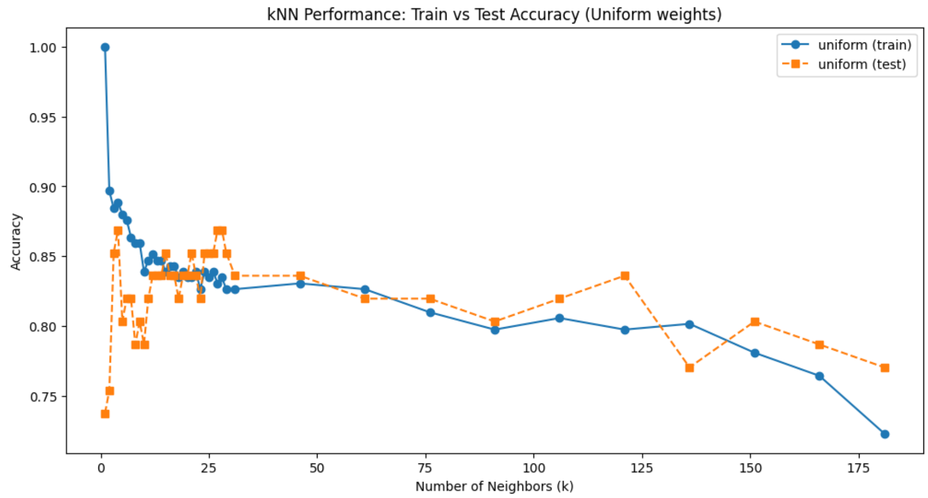
   a)

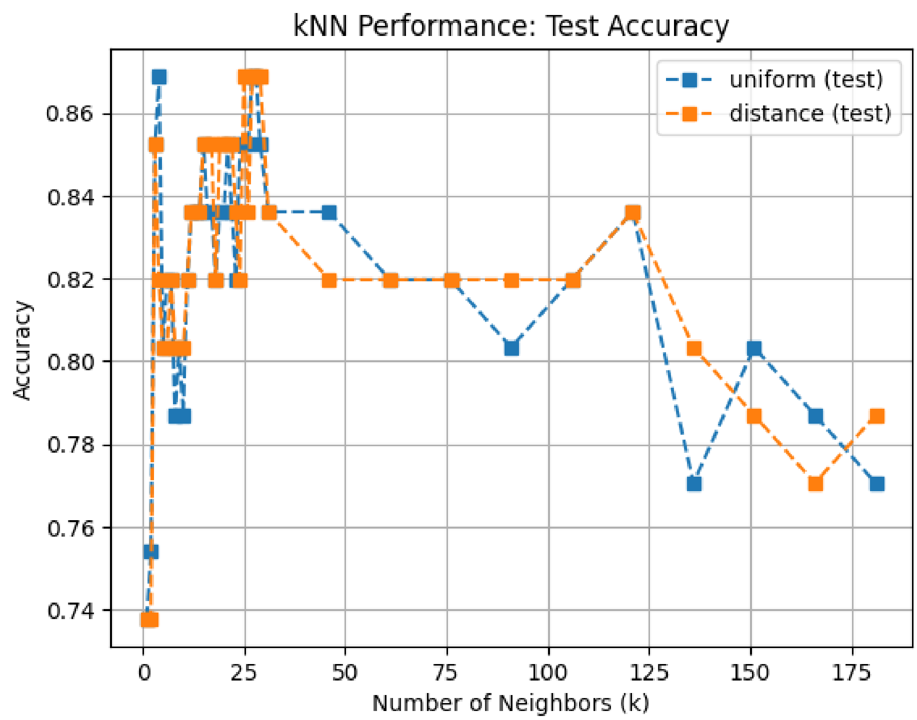


In this graph we can see the problem of overfitting for k=1 with very high train accuracies, but lower test accuracies. Then an improvement in test accuracy up to k=30. We should also note that the distance weights option has a 1.0 accuracy for the training whereas the training accuracy for the uniform weights gets down as k increases.

   b)

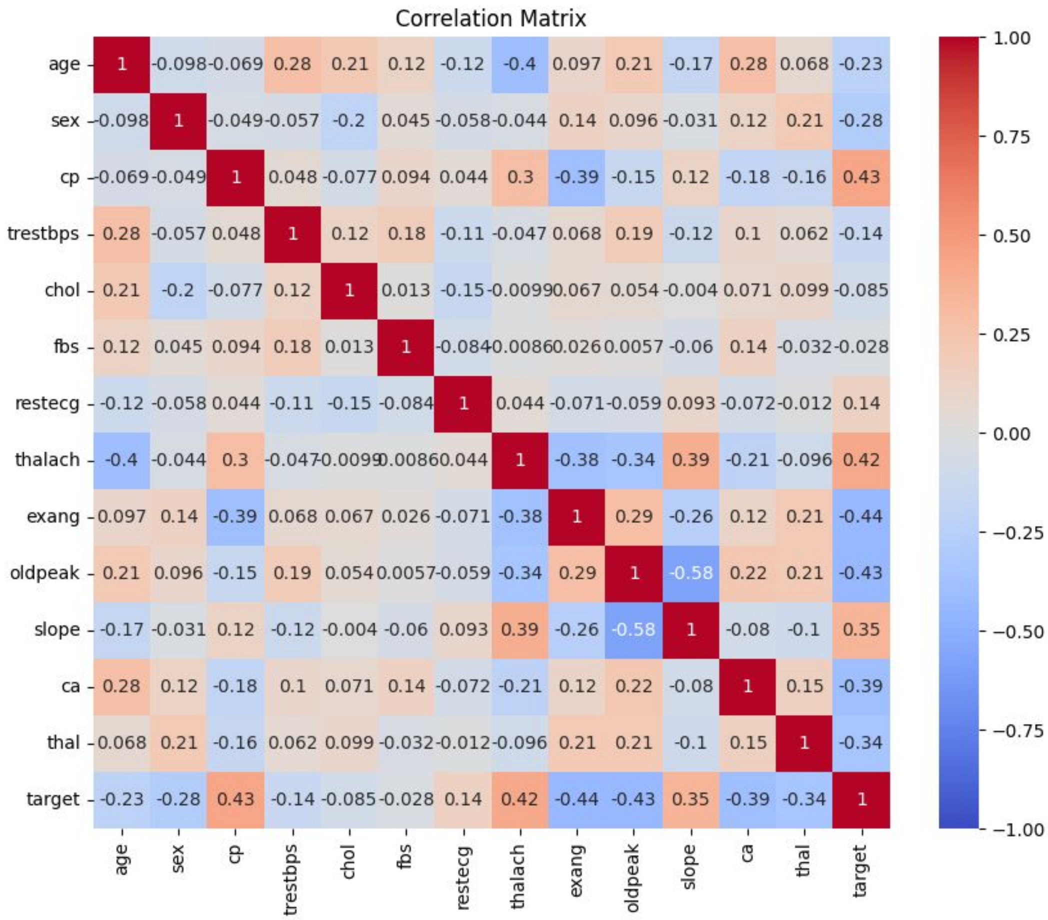
It is interesting to notice that using the distance weights we always get a training accuracy of 1.00. The same does not apply to the uniform weights where there is a tendency to reduce the accuracy based on the number k of Nearest neighbours considered in the training, this may be due to underfitting, and given we have no distinguishing factor between the closest neighbours and the neighbours more far away. When it comes to the testing accuracy, we can see that changing the number k of Nearest neighbours for a value of 3 or 4 there is a small peak, this first peak is in both cases a bit volatile (volatility that is probably related to overfitting) and then another peak at around 25-30, this second peak in accuracy is more stable with the change of the number k. After that the training accuracy decreases as k increases, probably





due to underfitting. When we increase the number of neighbours, k, we firstly start from an overfitting model, at small k, in particular k = 1, where our test accuracy is very low, and the training accuracy very high. Lastly, at high k the training and test accuracies from the uniform weigths and the test accuracy from the distance weights start decreasing, as the model starts to become too generalized, underfitting the data. We can also note that although the training accuracy is very different for the different weights, the difference in the testing accuracy is not as big.

3. One of the problems in the application of Naïve Bayes model is that we assume that the variables are independent, which is generally not the case for health related data. In order to illustrate this we can make a correlation matrix of the variables at hand. From the correlation matrix below we can, for example, see that there is a clear negative correlation between the oldpeak and the slope classes.



Another problem that we may have in using the Naïve Bayes model with Gaussian Assumption is that we assume that there is a Bayesian/Normal distribution for the continuous data, this may lead to a problem given that the majority of the input variables in our data set are not continuous data, rather closer to binary or very small discrete classes, as we can see in the table below of the unique values for each class, with most of them being below 5 different values.

Class	Unique_Values
age	41
sex	2
cp	4
trestbps	49
chol	152
fbs	2
restecg	3
thalach	91
exang	2
oldpeak	40
slope	3
ca	5
thal	4
target	2