Aprendizagem 2024/25

Homework 2 – Group 98

(João Pinho - ist1102765, Tiago Sousa - ist1110691)

Part I: Pen and paper

1. F1 measure of a kNN with k=5 using Hamming Distance

   The following table shows the Hamming Distances (HD) between any two distinct observations. (the distances between equal elements are zero and were not included, as we're using a leave-one-out schema).

HD (x_i, x_j)	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$x_1 = (A, 0)$	-	2	1	0	1	1	1	2
$x_2 = (B, 1)$	2	-	1	2	1	1	1	0
$x_3 = (A, 1)$	1	1	-	1	2	2	0	1
$x_4 = (A, 0)$	0	2	1	-	1	1	1	2
$x_5 = (B, 0)$	1	1	2	1	-	0	2	1
$x_6 = (B, 0)$	1	1	2	1	0	-	2	1
$x_7 = (A, 1)$	1	1	0	1	2	2	-	1
$x_8 = (B, 1)$	2	0	1	2	1	1	1	-

   We can thereby determine the 5 nearest neighbours (5NN) of each observation (the 5 other observations with the lowest Hamming Distance) and determine its predicted class based on the mode of its neighbours' classes. (Here we denote $y_{out}(x_i)$ as the class of observation x_i and $\widehat{y_{out}}(x_i)$ as the predicted class).

Observation	5NN	$\widehat{y_{out}}$
x_1	x_3, x_4, x_5, x_6, x_7	$\widehat{y_{out}}(x_1) = \text{mode}(P, P, N, N, N) = N$
x_2	x_3, x_5, x_6, x_7, x_8	$\widehat{y_{out}}(x_2) = \text{mode}(P, N, N, N, N) = N$
x_3	x_1, x_2, x_4, x_7, x_8	$\widehat{y_{out}}(x_3) = \text{mode}(P, P, P, N, N) = P$
x_4	x_1, x_3, x_5, x_6, x_7	$\widehat{y_{out}}(x_4) = \text{mode}(P, P, N, N, N) = N$
x_5	x_1, x_2, x_4, x_6, x_8	$\widehat{y_{out}}(x_5) = \text{mode}(P, P, P, N, N) = P$
x_6	x_1, x_2, x_4, x_5, x_8	$\widehat{y_{out}}(x_6) = \text{mode}(P, P, P, N, N) = P$
x_7	x_1, x_2, x_3, x_4, x_8	$\widehat{y_{out}}(x_7) = \text{mode}(P, P, P, P, N) = P$
x_8	x_2, x_3, x_5, x_6, x_7	$\widehat{y_{out}}(x_8) = \text{mode}(P, P, N, N, N) = N$

   Let's compare the actual and predicted classes for each observation.

Observation	y_{out}	$\widehat{y_{out}}$	Outcome
x_1	P	N	FN
x_2	P	N	FN
x_3	P	P	TP
x_4	P	N	FN
x_5	N	P	FP
x_6	N	P	FP
x_7	N	P	FP
x_8	N	N	TN

We can now compute the Recall (R), Precision (P) and F1 measure of this model.

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

$$P = \frac{TP}{TP + FP} = \frac{1}{1+3} = 0.25$$

$$\text{F1} = \frac{2PR}{P+R} = \frac{2 \cdot 0.25 \cdot 0.25}{0.25 + 0.25} = 0.25$$

2. Proposing a new metric to improve the model's performance

   We can start by noting that the second feature (y_2) of each observation has probably less discriminative power than the first feature (y_1), as for each class there are 2 observations with y_2 =0 and 2 others with $y_2 = 1$, while there are 3 positive and just one negative observations with ($y_1 = A$) (and vice-versa with $y_2 = B$). Let's then calculate the Hamming Distance for all pairs of distinct observations considering only the first feature ($\text{HD}|_{y_1}$) .

$\text{HD}\|_{y_1}$ (x_i, x_j)	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$x_1 = (A)$	-	1	0	0	1	1	0	1
$x_2 = (B)$	1	-	1	1	0	0	1	0
$x_3 = (A)$	0	1	-	0	1	1	0	1
$x_4 = (A)$	0	1	0	-	1	1	0	1
$x_5 = (B)$	1	0	1	1	-	0	1	0
$x_6 = (B)$	1	0	1	1	0	-	1	0
$x_7 = (A)$	0	1	0	0	1	1	-	1
$x_8 = (B)$	1	0	1	1	0	0	1	-

   As each observation has exactly 3 other observations with $\text{HD}|_{y_1}$= 0, and there are only two possible feature combinations (the remaining obsevations are equally far away), as well as to avoid randomly choosing nearest neighbours, we may consider the 3 nearest neighbours (choose $k = 3$). Let's predict each observation's class:

Observation	3NN	$\widehat{y_{out}}$
x_1	x_3, x_4, x_7	$\widehat{y_{out}}(x_1) = \text{mode}(P, P, N) = P$
x_2	x_5, x_6, x_8	$\widehat{y_{out}}(x_2) = \text{mode}(N, N, N) = N$
x_3	x_1, x_4, x_7	$\widehat{y_{out}}(x_3) = \text{mode}(P, P, N) = P$
x_4	x_1, x_3, x_7	$\widehat{y_{out}}(x_4) = \text{mode}(P, P, N) = P$
x_5	x_2, x_6, x_8	$\widehat{y_{out}}(x_5) = \text{mode}(P, N, N) = N$
x_6	x_2, x_5, x_8	$\widehat{y_{out}}(x_6) = \text{mode}(P, N, N) = N$
x_7	x_1, x_3, x_4	$\widehat{y_{out}}(x_7) = \text{mode}(P, P, P) = P$
x_8	x_2, x_5, x_6	$\widehat{y_{out}}(x_8) = \text{mode}(P, N, N) = N$

Let's compare the actual and predicted classes for each observation with this new model:

Observation	y_{out}	$\widehat{y_{out}}$	Outcome
x_1	P	P	TP
x_2	P	N	FN
x_3	P	P	TP
x_4	P	P	TP
x_5	N	N	TN
x_6	N	N	TN
x_7	N	P	FP
x_8	N	N	TN

Now we can verify that the new model achieved a higher F1 measure.

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

$$P = \frac{TP}{TP + FP} = \frac{3}{3 + 1} = 0.75$$

$$\text{F1} = \frac{2PR}{P + R} = \frac{2 \cdot 0.75 \cdot 0.75}{0.75 + 0.75} = 0.75$$

This kNN model (with (distance, k) = (HD$|_{y_1}$, 3)) improved the previous one's F1 measure by three-fold (0.75 = 0.25 · 3).

3. Bayesian classifier

   To learn a Bayesian classifier, we should calculate the posteriors for each class c, in this exercise positive (P) or negative(N), given each possible set of observations:

$$p(y_{out} = c \mid y_1 = a_1 \wedge y_2 = a_2 \wedge y_3 = a_3) = p(y_1 = a_1 \wedge y_2 = a_2 \wedge y_3 = a_3 \mid y_{out} = c)p(y_{out} = c)$$

Using the assumption that the y_1, y_2 and y_3 variable sets are independent and equally important, we can use Naive Bayes to rewrite this posterior as a product of the likelihoods on the independent variable sets and the prior:

$$p(y_{out} = c \mid y_1 = a_1 \land y_2 = a_2 \land y_3 = a_3) = p(y_1 = a_1 \land y_2 = a_2 \mid y_{out} = c) p(y_3 = a_3 \mid y_{out} = c) p(y_{out} = c)$$

The priors are $p(y_{out} = P) = \frac{5}{9}$ and $p(y_{out} = N) = \frac{4}{9}$.

Let's compute each likelihood (using Bayes' rule) for the categorical variable set y_1, y_2.

$$p(y_1 = A \land y_2 = 0 \mid y_{out} = P) = \frac{2}{5} = 0.4$$

$$p(y_1 = A \land y_2 = 0 \mid y_{out} = N) = \frac{0}{4} = 0$$

$$p(y_1 = A \land y_2 = 1 \mid y_{out} = P) = \frac{1}{5} = 0.2$$

$$p(y_1 = A \land y_2 = 1 \mid y_{out} = N) = \frac{1}{4} = 0.25$$

$$p(y_1 = B \land y_2 = 0 \mid y_{out} = P) = \frac{1}{5} = 0.2$$

$$p(y_1 = B \land y_2 = 0 \mid y_{out} = N) = \frac{2}{4} = 0.5$$

$$p(y_1 = B \land y_2 = 1 \mid y_{out} = P) = \frac{1}{5} = 0.2$$

$$p(y_1 = B \land y_2 = 1 \mid y_{out} = N) = \frac{1}{4} = 0.25$$

In order to compute the posteriors for the y_3 variable, we need to know the likelihood, and for that we may use the assumption that this variable is normally distributed.

$$\mu_{|P} = \frac{1.1 + 0.8 + 0.5 + 0.9 + 0.8}{5} = 0.82$$

$$\mu_{|N} = \frac{1.0 + 0.9 + 1.2 + 0.9}{4} = 1.0$$

$$\sigma_{|P} = \sqrt{\frac{(1.1 - 0.82)^2 + (0.8 - 0.82)^2 + (0.5 - 0.82)^2 + (0.9 - 0.82)^2 + (0.8 - 0.82)^2}{4}} \approx 0.217$$

$$\sigma_{|N} = \sqrt{\frac{(1.0 - 1.0)^2 + (0.9 - 1.0)^2 + (1.2 - 1.0)^2 + (0.9 - 1.0)^2}{3}} \approx 0.141$$

We now obtain the likelihoods using the normal distribution:

$$p(y_3 = x \mid y_{out} = P) = p(x \mid \mu_{|P}, {\sigma_{|P}}^2) = \frac{e^{-\frac{(x-\mu_{|P})^2}{2{\sigma_{|P}}^2}}}{\sqrt{2\pi}\sigma_{|P}} \approx \frac{e^{-\frac{(x-0.82)^2}{0.094}}}{0.544}$$

$$p(y_3 = x \mid y_{out} = N) = p(x \mid \mu_{|N}, {\sigma_{|N}}^2) = \frac{e^{-\frac{(x-\mu_{|N})^2}{2{\sigma_{|N}}^2}}}{\sqrt{2\pi}\sigma_{|N}} \approx \frac{e^{-\frac{(x-1.0)^2}{0.040}}}{0.353}$$

The posteriors may now be computed:

$$p(y_{out} = P \mid y_1 = A \land y_2 = 0 \land y_3 = x) =$$
$$= p(y_1 = A \land y_2 = 0 \mid y_{out} = P)p(y_3 = x \mid y_{out} = P)p(y_{out} = P) =$$
$$= 0.4 \cdot \frac{e^{-\frac{(x-0.82)^2}{0.094}}}{0.544} \cdot \frac{5}{9} \approx 0.408e^{-\frac{(x-0.82)^2}{0.094}}$$
$$p(y_{out} = N \mid y_1 = A \land y_2 = 0 \land y_3 = x) =$$
$$= p(y_1 = A \land y_2 = 0 \mid y_{out} = N)p(y_3 = x \mid y_{out} = N)p(y_{out} = N) =$$
$$= 0 \cdot \frac{e^{-\frac{(x-1.0)^2}{0.040}}}{0.353} \cdot \frac{4}{9} = 0$$

$$p(y_{out} = P \mid y_1 = A \land y_2 = 1 \land y_3 = x) =$$
$$= p(y_1 = A \land y_2 = 1 \mid y_{out} = P)p(y_3 = x \mid y_{out} = P)p(y_{out} = P) =$$
$$= 0.2 \cdot \frac{e^{-\frac{(x-0.82)^2}{0.094}}}{0.544} \cdot \frac{5}{9} \approx 0.204e^{-\frac{(x-0.82)^2}{0.094}}$$
$$p(y_{out} = N \mid y_1 = A \land y_2 = 1 \land y_3 = x) =$$
$$= p(y_1 = A \land y_2 = 1 \mid y_{out} = N)p(y_3 = x \mid y_{out} = N)p(y_{out} = N) =$$
$$= 0.25 \cdot \frac{e^{-\frac{(x-1.0)^2}{0.040}}}{0.353} \cdot \frac{4}{9} = 0.315e^{-\frac{(x-1.0)^2}{0.040}}$$

$$p(y_{out} = P \mid y_1 = B \wedge y_2 = 0 \wedge y_3 = x) =$$
$$= p(y_1 = B \wedge y_2 = 0 \mid y_{out} = P) p(y_3 = x \mid y_{out} = P) p(y_{out} = P) =$$
$$= 0.2 \cdot \frac{e^{-\frac{(x-0.82)^2}{0.094}}}{0.544} \cdot \frac{5}{9} \approx 0.204 e^{-\frac{(x-0.82)^2}{0.094}}$$
$$p(y_{out} = N \mid y_1 = B \wedge y_2 = 0 \wedge y_3 = x) =$$
$$= p(y_1 = B \wedge y_2 = 0 \mid y_{out} = N) p(y_3 = x \mid y_{out} = N) p(y_{out} = N) =$$
$$= 0.5 \cdot \frac{e^{-\frac{(x-1.0)^2}{0.040}}}{0.353} \cdot \frac{4}{9} = 0.630 e^{-\frac{(x-1.0)^2}{0.040}}$$

$$p(y_{out} = P \mid y_1 = B \wedge y_2 = 1 \wedge y_3 = x) =$$
$$= p(y_1 = B \wedge y_2 = 1 \mid y_{out} = P) p(y_3 = x \mid y_{out} = P) p(y_{out} = P) =$$
$$= 0.2 \cdot \frac{e^{-\frac{(x-0.82)^2}{0.094}}}{0.544} \cdot \frac{5}{9} \approx 0.204 e^{-\frac{(x-0.82)^2}{0.094}}$$
$$p(y_{out} = N \mid y_1 = B \wedge y_2 = 1 \wedge y_3 = x) =$$
$$= p(y_1 = B \wedge y_2 = 1 \mid y_{out} = N) p(y_3 = x \mid y_{out} = N) p(y_{out} = N) =$$
$$= 0.25 \cdot \frac{e^{-\frac{(x-1.0)^2}{0.040}}}{0.353} \cdot \frac{4}{9} = 0.315 e^{-\frac{(x-1.0)^2}{0.040}}$$

4. Prediction of test observations' classes using a MAP assumption

   Using a Maximum a Posteriori assumption, the predicted class of a new observation will be the one maximising the posterior:

$$\widehat{y_{out}}(x) = \begin{cases} P \text{ if } p(y_{out} = P \mid x) > p(y_{out} = N \mid x) \\ N \text{ if } p(y_{out} = P \mid x) < p(y_{out} = N \mid x) \end{cases} \tag{0}$$

   Therefore:

$$p(y_{out} = P \mid y_1 = A \wedge y_2 = 1 \wedge y_3 = 0.8)) \approx 0.204 e^{-\frac{(0.8-0.82)^2}{0.094}} \approx 0.203$$
$$p(y_{out} = N \mid y_1 = A \wedge y_2 = 1 \wedge y_3 = 0.8)) \approx 0.315 e^{-\frac{(0.8-1.0)^2}{0.040}} \approx 0.116$$
$$\Rightarrow \widehat{y_{out}}(A, 1, 0.8) = P$$

$$p(y_{out} = P \mid y_1 = B \wedge y_2 = 1 \wedge y_3 = 1)) \approx 0.204 e^{-\frac{(1-0.82)^2}{0.094}} \approx 0.145$$
$$p(y_{out} = N \mid y_1 = B \wedge y_2 = 1 \wedge y_3 = 1)) \approx 0.315 e^{-\frac{(1-1.0)^2}{0.040}} \approx 0.315$$
$$\Rightarrow \widehat{y_{out}}(B, 1, 1) = N$$

$$p(y_{out} = P \mid y_1 = B \wedge y_2 = 0 \wedge y_3 = 0.9)) \approx 0.204e^{-\frac{(0.9-0.82)^2}{0.094}} \approx 0.191$$

$$p(y_{out} = N \mid y_1 = B \wedge y_2 = 0 \wedge y_3 = 0.9)) \approx 0.630e^{-\frac{(0.9-1.0)^2}{0.040}} \approx 0.491$$

$$\Rightarrow \widehat{y_{out}}(B, 0, 0.9) = N$$

5. Naive bayes classification using Maximum Likelihood assumption:
   According to the naive Bayes algorithm, our prediction $\hat{Z}$ is:

$$\hat{Z} = argmax_C(\prod_{i=1}^{N} P(t_i|C)$$

Our vocabulary has 9 terms, and there are 5 in the positive class and 4 in the negative class. Considering

$$P(t_i|C) = \frac{freq(t_i) + 1}{N_C + V}$$
$$P(P) = P(N) = 0.5$$
$$P('I'|P) = \frac{1+1}{5+9} = \frac{1}{7}$$
$$P('I'|N) = \frac{0+1}{4+9} = \frac{1}{13}$$
$$P('like'|P) = \frac{1+1}{5+9} = \frac{1}{7}$$
$$P('like'|N) = \frac{0+1}{4+9} = \frac{1}{13}$$
$$P('to'|P) = \frac{0+1}{5+9} = \frac{1}{14}$$
$$P('to'|N) = \frac{0+1}{4+9} = \frac{1}{13}$$
$$P('run'|P) = \frac{1+1}{5+9} = \frac{1}{7}$$
$$P('run'|N) = \frac{1+1}{4+9} = \frac{2}{13}$$

we get:

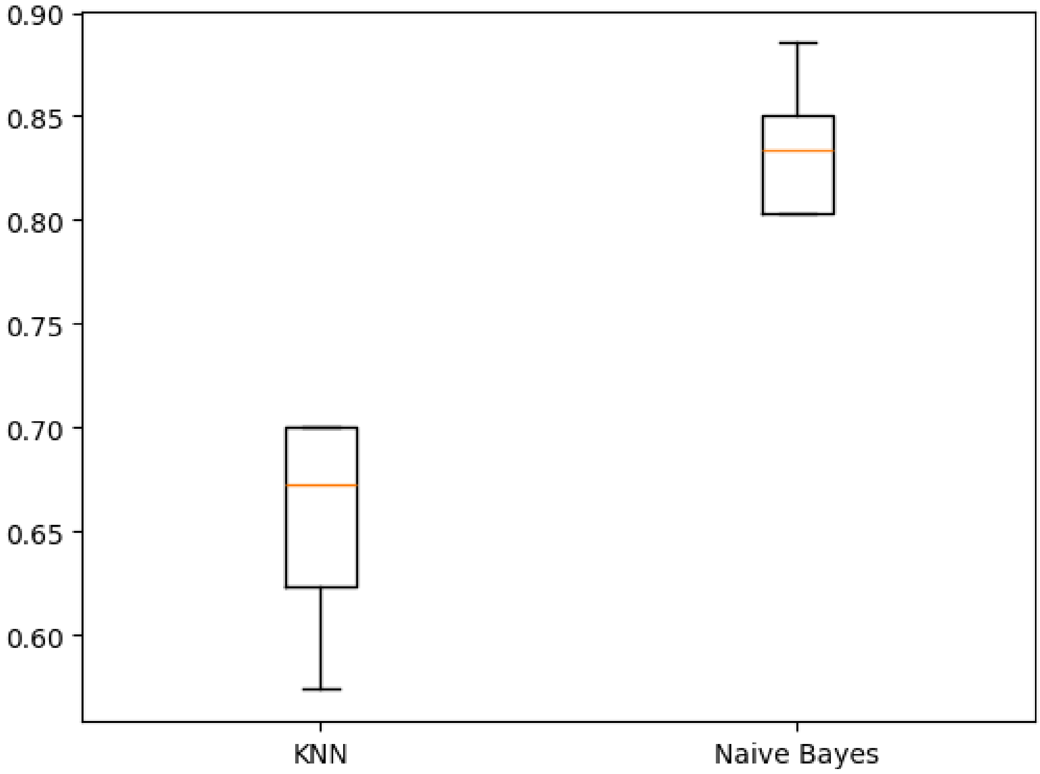
$$\hat{Z} = argmax\left(\prod_{i=1}^{N} P(t_i|P), \prod_{i=1}^{N} P(t_i|N)\right)$$
$$\prod_{i=1}^{N} P(t_i|P) = \frac{1}{7^3} \times \frac{1}{14} = \frac{1}{4802} \approx 2 \times 10^{-4}$$
$$\prod_{i=1}^{N} P(t_i|N) = \frac{1}{13^3} \times \frac{2}{13} = \frac{2}{28561} \approx 7 \times 10^{-5}$$

Therefore, our prediction $\hat{Z}$ is P, and our model will predict the phrase ”I like to run” to have a positive connotation.

Part 2: Programming

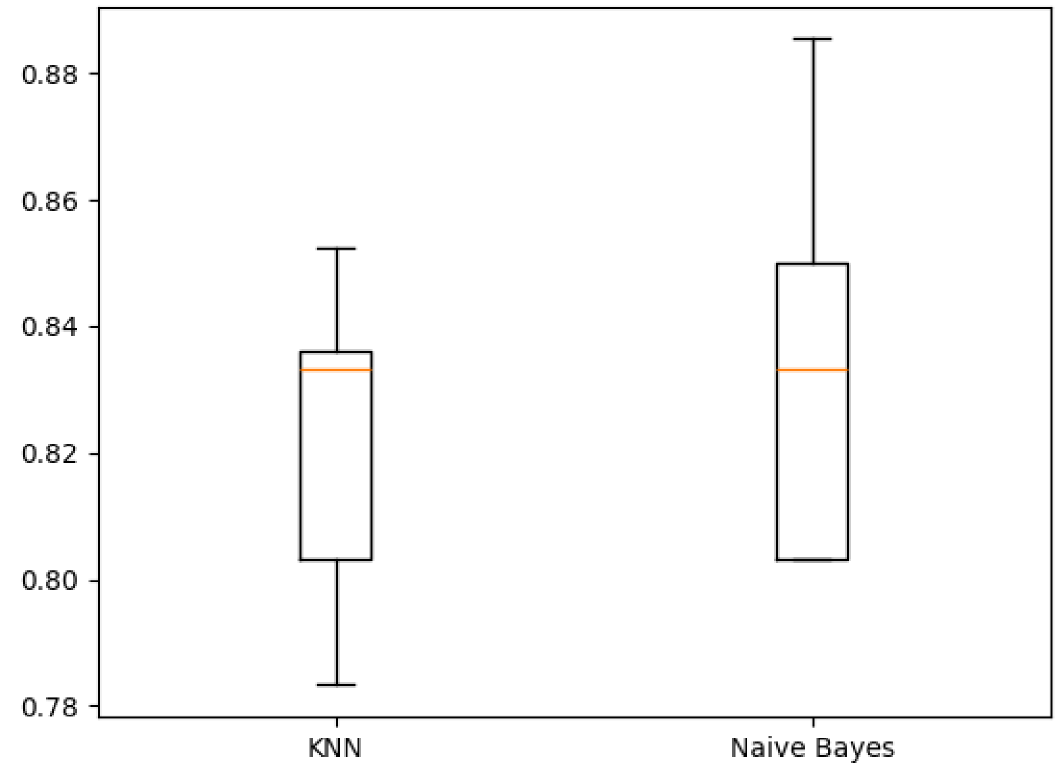
1. kNN vs Naive Bayes with Gaussian assumption

a) Boxplots of models’ accuracy



The Gaussian approach seems to be more stable, as its values span a smaller range.
This should be because of the kNN’s high sensitivity to local noise, making it harder to generalize through different folds of the dataset. The NB is less sensitive to the training instances, as it’s decision is based on the aggregate probabilities of the dataset, making it less affected by an outlier. This could also indicate that the dataset has significant feature independence, rather than defined ’clusters’, where neighbours are closely related.

b) Scaled data analysis



Scaling the data before training the models had different effects on each model.

On one side, the Naive Bayes model was not affected by this change, keeping its accuracy and stability from the previous exercise. As it is a probability-based model, this shouldn't surprise us, as the relative frequencies on which the probabilities are calculated remained the same.

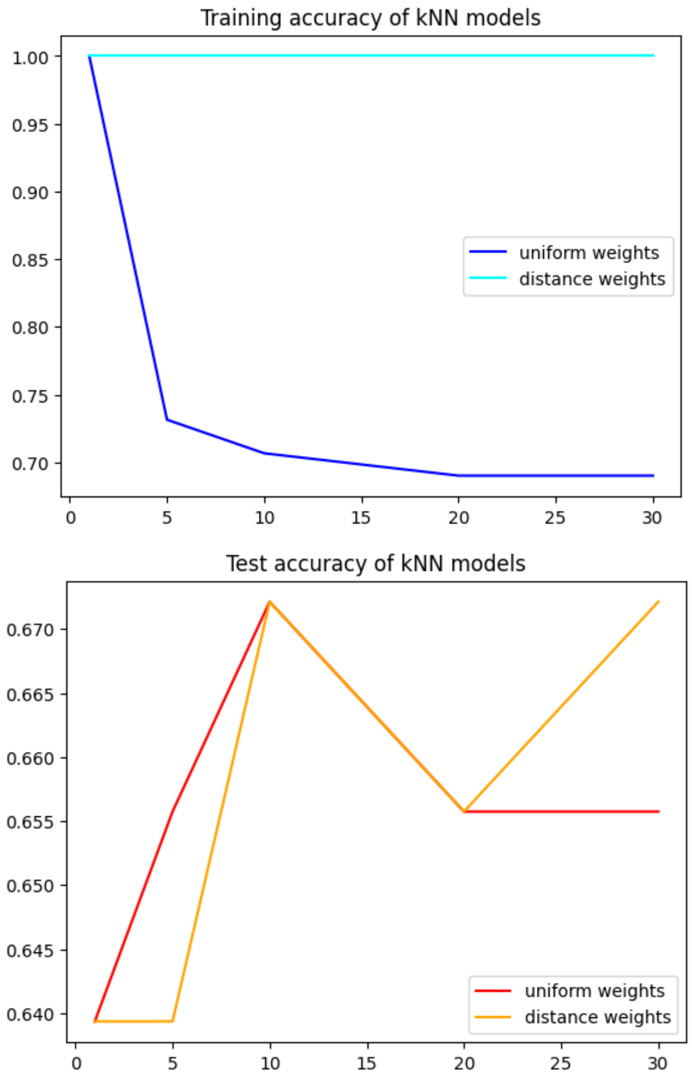
On the other side, the kNN was positively affected by this change, improving on both accuracy and stability. This change is due to the fact that kNNs are very susceptible to distances. Before, different features had different scales, so some features would bias the model, increasing the distance significantly. This change normalizes scales across features, allowing better distance calculations and better generalization abilities.

c) Hypothesis testing

Using scipy, the p-value reached was 0.507. This value is not significant, so we reject the hypothesis that the kNN model is statistically superior to the Naive Bayes regarding accuracy.

2. Training and testing multiple kNN models

a) Plot of accuracies



b) Analysis of the results

Increasing the neighbours of a kNN doesn't directly correlate with a better generalization ability.

Regarding the distance weighted model, it was always able to correctly classify every instance of the training set, but only 64 to 67 percent of the test set. Increasing the number of neighbours shows that the optimal amount is 10 or 30, while the worse is from 1 to 5 neighbours.

The uniform weights model's performance on the training set was inversely proportional to k, decreasing every time k increased. However, on the test dataset the accuracy improved from 1 to 10 neighbours and fell down with k ¿ 10.

It seems that increasing the number of neighbours of a kNN will improve the generalization ability of a model only up to a certain threshold value, after which it won't improve anymore. In our case, that value was $k = 10$.

3. Difficulties of the Naive Bayes model

   The biggest difficulty of the naive Bayes model with the heart-disease.csv dataset is likely its assumption of conditional independence between features, that for health and heart disease shouldn't be the case. For example, the age of a patient influences any doctor on how to evaluate other symptoms, as they are heavily correlated. The model oversimplifies these relations, making it yield sub-optimal results.

   Another difficulty is the small sample size, as the dataset only has around 300 instances. This can lead the model to make uninformed predictions when a new instance is not represented by the dataset. Additionally, there's the risk of overfitting, where the model learns the small dataset's characteristics instead of actual characteristics of heart disease.