Hw 2

Daniel Brigham Évora 99473
Natacha Sousa 107413

I. Pen-and-paper [13v]

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 F1 measure?
KNN, K=5
→ Hamming distance
→ leave one out evaluation scheme

	b_1	b_2	ent cone
x_0	A	0	+
x_1	B	1	+
x_3	A	1	+
x_4	A	0	+
x_5	B	0	−
x_6	B	0	−
x_7	A	1	−
x_8	B	1	−

B 0 +

x_1
$hamm(x_1, x_2) = 2$
$hamm(x_1, x_3) = 1$ +
$hamm(x_1, x_4) = 0$ +
$hamm(x_1, x_5) = 1$ −
$hamm(x_1, x_6) = 1$ −
$hamm(x_1, x_7) = 1$ −
$hamm(x_1, x_8) = 2$
mode(+,+,−,−) = (−) (−) ✗

x_2
$hamm(x_2, x_1) = 2$
$hamm(x_2, x_3) = 1$ +
$hamm(x_2, x_4) = 2$
$hamm(x_2, x_5) = 1$ −
$hamm(x_2, x_6) = 1$ −
$hamm(x_2, x_7) = 1$ −
$hamm(x_2, x_8) = 0$ −
// (−) ✗

x_3
$hamm(x_3, x_1) = 1$ +
$hamm(x_3, x_2) = 1$ +
$hamm(x_3, x_4) = 1$ +
$hamm(x_3, x_5) = 2$
$hamm(x_3, x_6) = 2$
$hamm(x_3, x_7) = 0$ −
$hamm(x_3, x_8) = 1$ −
// (+) ✓

x_4
$hamm(x_4, x_1) = 0$ +
$hamm(x_4, x_2) = 2$
$hamm(x_4, x_3) = 1$ +
$hamm(x_4, x_5) = 1$ −
$hamm(x_4, x_6) = 1$ −
$hamm(x_4, x_7) = 1$ −
$hamm(x_4, x_8) = 2$
// (−) ✗

$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$

$P = \frac{TP}{TP+FP}$ $R = \frac{TP}{TP+FN}$

$P = \frac{1}{1+3} = \frac{1}{4}$

$R = \frac{1}{1+3} = \frac{1}{4}$

$F_1 = \frac{2}{4+4} = \frac{2}{8} = \frac{1}{4} = 0{,}25 = F_1$

x_5
$hamm(x_5, x_1) = 1$ +
$hamm(x_5, x_2) = 1$ +
$hamm(x_5, x_3) = 2$
$hamm(x_5, x_4) = 1$ +
$hamm(x_5, x_6) = 0$ −
$hamm(x_5, x_7) = 2$
$hamm(x_5, x_8) = 1$ −
// (+) ✗

x_6
$hamm(x_6, x_1) = 1$ +
$hamm(x_6, x_2) = 1$ +
$hamm(x_6, x_3) = 2$
$hamm(x_6, x_4) = 1$ +
$hamm(x_6, x_5) = 0$ −
$hamm(x_6, x_7) = 2$
$hamm(x_6, x_8) = 1$ −
// (+) ✗

x_7
$hamm(x_7, x_1) = 1$ +
$hamm(x_7, x_2) = 1$ +
$hamm(x_7, x_3) = 0$ +
$hamm(x_7, x_4) = 1$ +
$hamm(x_7, x_5) = 2$
$hamm(x_7, x_6) = 2$
$hamm(x_7, x_8) = 1$ −
// (+) ✗

x_8
$hamm(x_8, x_1) = 2$
$hamm(x_8, x_2) = 0$ +
$hamm(x_8, x_3) = 1$ +
$hamm(x_8, x_4) = 2$
$hamm(x_8, x_5) = 1$ −
$hamm(x_8, x_6) = 1$ −
$hamm(x_8, x_7) = 1$ −
// (−) ✓

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

(2)

weighted distance that only considers the y_1 hamming distance

$1 \times \text{ham}_m(y_1) + 0 \times \text{ham}_m(y_2)$, with $k=3$, lets call it ham_{m_2}

x_1

$\text{ham}_{m_2}(x_1, x_2) = 1$
$\text{ham}_{m_2}(x_1, x_3) = 0$ +
$\text{ham}_{m_2}(x_1, x_4) = 0$ +
$\text{ham}_{m_2}(x_1, x_5) = 1$
$\text{ham}_{m_2}(x_1, x_6) = 1$
$\text{ham}_{m_2}(x_1, x_7) = 0$ −
$\text{ham}_{m_2}(x_1, x_8) = 1$

mode (+,+,−) ⇓ (+) ✓

x_2

$\text{ham}_{m_2}(x_2, x_1) = 1$
$\text{ham}_{m_2}(x_2, x_3) = 1$
$\text{ham}_{m_2}(x_2, x_4) = 1$
$\text{ham}_{m_2}(x_2, x_5) = 0$ −
$\text{ham}_{m_2}(x_2, x_6) = 0$ −
$\text{ham}_{m_2}(x_2, x_7) = 1$
$\text{ham}_{m_2}(x_2, x_8) = 0$ −

// (−) ✗

x_3

$\text{ham}_{m_2}(x_3, x_1) = 0$ +
$\text{ham}_{m_2}(x_3, x_2) = 1$
$\text{ham}_{m_2}(x_3, x_4) = 0$ +
$\text{ham}_{m_2}(x_3, x_5) = 1$
$\text{ham}_{m_2}(x_3, x_6) = 1$
$\text{ham}_{m_2}(x_3, x_7) = 0$ −
$\text{ham}_{m_2}(x_3, x_8) = 1$

// (+) ✓

x_4

$\text{ham}_{m_2}(x_4, x_1) = 0$ +
$\text{ham}_{m_2}(x_4, x_2) = 1$
$\text{ham}_{m_2}(x_4, x_3) = 0$ +
$\text{ham}_{m_2}(x_4, x_5) = 1$
$\text{ham}_{m_2}(x_4, x_6) = 1$
$\text{ham}_{m_2}(x_4, x_7) = 0$ −
$\text{ham}_{m_2}(x_4, x_8) = 1$

// (+) ✓

x_5

$\text{ham}_{m_2}(x_5, x_1) = 1$
$\text{ham}_{m_2}(x_5, x_2) = 0$ +
$\text{ham}_{m_2}(x_5, x_3) = 1$
$\text{ham}_{m_2}(x_5, x_4) = 1$
$\text{ham}_{m_2}(x_5, x_6) = 0$ −
$\text{ham}_{m_2}(x_5, x_7) = 1$
$\text{ham}_{m_2}(x_5, x_8) = 0$ −

// (−) ✓

x_6

$\text{ham}_{m_2}(x_6, x_1) = 1$
$\text{ham}_{m_2}(x_6, x_2) = 0$ +
$\text{ham}_{m_2}(x_6, x_3) = 1$
$\text{ham}_{m_2}(x_6, x_4) = 1$
$\text{ham}_{m_2}(x_6, x_5) = 0$ −
$\text{ham}_{m_2}(x_6, x_7) = 1$
$\text{ham}_{m_2}(x_6, x_8) = 0$ −

// (−) ✓

x_7

$\text{ham}_{m_2}(x_7, x_1) = 0$ +
$\text{ham}_{m_2}(x_7, x_2) = 1$
$\text{ham}_{m_2}(x_7, x_3) = 0$ +
$\text{ham}_{m_2}(x_7, x_4) = 0$ +
$\text{ham}_{m_2}(x_7, x_5) = 1$
$\text{ham}_{m_2}(x_7, x_6) = 1$
$\text{ham}_{m_2}(x_7, x_8) = 1$

// (+)

x_8

$\text{ham}_{m_2}(x_8, x_1) = 1$
$\text{ham}_{m_2}(x_8, x_2) = 0$ +
$\text{ham}_{m_2}(x_8, x_3) = 1$
$\text{ham}_{m_2}(x_8, x_4) = 1$
$\text{ham}_{m_2}(x_8, x_5) = 0$ −
$\text{ham}_{m_2}(x_8, x_6) = 0$ −
$\text{ham}_{m_2}(x_8, x_7) = 1$

// (−)

$$P = \frac{3}{3+1} = \frac{3}{4}$$

$$R = \frac{3}{3+1} = \frac{3}{4}$$

$$F_1 = \frac{2}{\frac{4}{3} + \frac{4}{3}} = \frac{2 \times 3}{8} = \frac{6}{8}$$

$F_1 = \frac{3}{4} = 0{,}75 = 3 \times F_1$ before

F_1 before = 0,25 ✓

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.

(3) $x_9 = (B, 0)$

$y_3|P = \{1.1, 0.8, 0.5, 0.9, 0.8\}$

$y_3|N = \{1, 0.9, 1.2, 0.9\}$

i) y_1, y_2 dependent

Bayesian classifier

ii) $\{y_1, y_2\}$ and $\{y_3\}$ independent, and equally important

y_3 for +:

$(y_3|+)$ $\begin{cases} \mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1,1+0,8+0,5+0,9+0,8}{5} = 0,82 \\ \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2 = 0,0470 \end{cases}$ → parametros

Assuming y_3 → normal distribution: $p(y_3=x|+) = p(y_3=x|\mu=0,82, \sigma^2=0,047) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

y_3 for $\ominus \Rightarrow \begin{cases} \mu = 1 \\ \sigma^2 = 0,02 \end{cases} \Rightarrow p(y_3=x|-) = p(y_3=x|\mu=1, \sigma^2=0,02) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$P(\text{Class}|\, y_1=a, y_2=b, y_3=c) = \frac{P(y_1=a, y_2=b, y_3=c|\text{class})\, P(\text{class})}{P(y_1=a, y_2=b, y_3=c)}$$

$$= \frac{P(y_1=a, y_2=b|\text{class})\, P(y_3=c|\text{class})\, P(\text{class})}{P(y_1=a, y_2=b, y_3=c)}$$

$\text{class} = \{+, -\}$

$y_{new} = \{y_1, y_2, y_3\}$

$$\hat{\text{class}}_{MAP} = \underset{\text{class}}{\text{argmax}} \frac{P(\text{class}|y_{new})\, p(\text{class})}{P(y_{new})}$$

↖ ignore because it's the same...

MAP assumption

(4) $(A \quad 1 \quad 0,8)$

$\frac{1}{\sqrt{2\pi \times 0,047}} e^{-\frac{(x-0,82)^2}{2\times 0,047}} = 1,8324$ (with $x = 0,8$)

$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ → $\frac{1}{5}$ → $\frac{5}{5}$

$$P(+|\, y_1=A, y_2=1, y_3=0,8) = \frac{P(y_1=A, y_2=1|+)\, P(y_3=0,8|+)\, P(+)}{P(y_1=A, y_2=1, y_3=0,8)} = 0,2036$$

$P(y_1=A, y_2=1, y_3=0,8)$ → doesn't matter

$\frac{1}{\sqrt{2\pi \times 0,02}} e^{-\frac{(x-1)^2}{2\times 0,02}} = 1,0378$ (with $x = 0,8$)

$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ → $\frac{1}{2}$ → $\frac{4}{5}$

$$P(-|\, y_1=A, y_2=1, y_3=0,8) = \frac{P(y_1=A, y_2=1|-)\, P(y_3=0,8|-)\, P(-)}{P(y_1=A, y_2=1, y_3=0,8)} = 0,11531$$

$P(y_1=A, y_2=1, y_3=0,8)$ → doesn't matter

under MAP assumption $\text{class}(x_{new} = (A, 1, 0,8)) = +$

(B 1 1)

$\frac{1}{\sqrt{2\pi \times 0,047}} e^{-\frac{(x-0,82)^2}{2\times 0,047}} = 1,3037$

$P(+ \mid y_1=B, y_2=1, y_3=1) = \frac{P(y_1=B, y_2=1 \mid +)\, P(y_3=1 \mid +)\, P(+)}{P(y_1=B, y_2=1, y_3=1)} = 0,14285556$

$P(y_1=B, y_2=1, y_3=1)$ doesn't matter

$\frac{1}{\sqrt{2\pi \times 0,02}} e^{-\frac{(x-1)^2}{2\times 0,02}} = 2,8209$

$P(- \mid y_1=B, y_2=1, y_3=1) = \frac{P(y_1=B, y_2=1 \mid -)\, P(y_3=1 \mid -)\, P(-)}{P(y_1=B, y_2=1, y_3=1)} = 0,31134333$

$P(y_1=B, y_2=1, y_3=1)$ doesn't matter

under MAP assumption class(x_new = (B, 1, 1)) = −

(B 0 0.9)

$\frac{1}{\sqrt{2\pi \times 0,047}} e^{-\frac{(x-0,82)^2}{2\times 0,047}} = 1,7191$

$P(+ \mid y_1=B, y_2=0, y_3=0,9) = \frac{P(y_1=B, y_2=0 \mid +)\, P(y_3=0,9 \mid +)\, P(+)}{P(y_1=B, y_2=0, y_3=0,9)} = 0,1910111$

$P(y_1=B, y_2=0, y_3=0,9)$ doesn't matter

$\frac{1}{\sqrt{2\pi \times 0,02}} e^{-\frac{(x-1)^2}{2\times 0,02}} = 2,197$

$P(- \mid y_1=B, y_2=0, y_3=0,9) = \frac{P(y_1=B, y_2=0 \mid -)\, P(y_3=0,9 \mid -)\, P(-)}{P(y_1=B, y_2=0, y_3=0,9)} = 0,488222$

$P(y_1=B, y_2=0, y_3=0,9)$ doesn't matter

under MAP assumption class(x_new = (B, 0, 1)) = −

(5)

Am	P
I li	P
Too T	N
Bad R	N

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

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

No conditional assumption !!

$P(+ \mid \text{I like to run}) = \frac{P(+)\, P(\text{I like to run} \mid +)}{P(\text{I like to run})} = \frac{2}{4} \times \left(\frac{0+1}{2+4}\right) = \frac{2}{4} \times \frac{2}{6} = \frac{1}{6} = 0,16667$

$P(\text{I like to run})$ doesn't matter

bigger

$$P(-\mid \text{I like boxes}) = \frac{P(-)\,P(\text{I like boxes}\mid -)}{P(\text{I like boxes})} = \frac{2}{4} \times \left(\frac{0+1}{2+4}\right) = \frac{2}{4} \times \frac{1}{6} = \frac{1}{12} = 0{,}08333$$

($P(-) = \frac{2}{4}$; $P(\text{I like boxes})$ doesn't matter)

The new sentence under MAP and naïve Bayes assumption is classified as positive

$$\text{class}(\text{I like it}) = +$$

TÉCNICO LISBOA

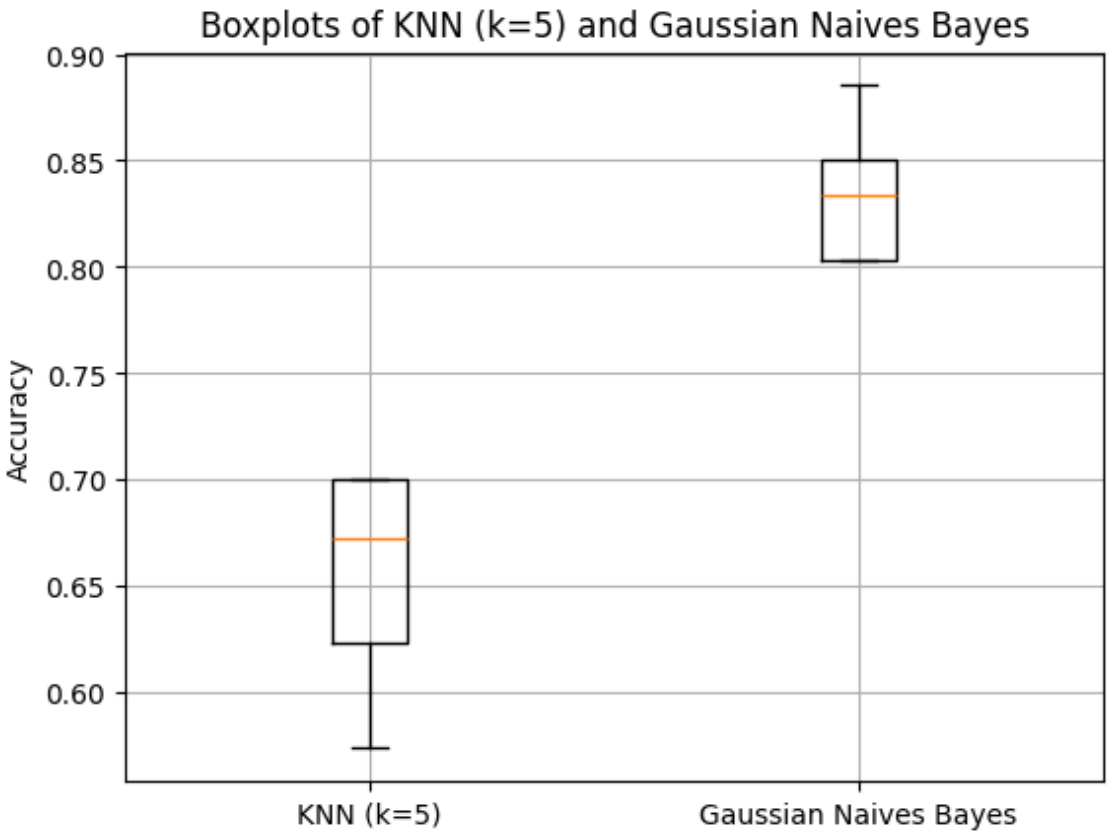
Homework 2
Trabalho realizado por:
Daniel Évora, 99473
Natacha Sousa, 107413

Parte II

1. a)

```python
import pandas as pd
from sklearn.model_selection import StratifiedKFold, cross_val_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
import matplotlib.pyplot as plt

# Load the CSV dataset using pandas (correct file path)
file_path = 'heart-disease.csv'
heart_data = pd.read_csv(file_path)

# Separate features (X) and target (y)
X = heart_data.drop(columns=['target'])
y = heart_data['target']

# Define classifiers
knn = KNeighborsClassifier(n_neighbors=5)
gnb = GaussianNB()

# Set up 5-fold stratified cross-validation with shuffling
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=0)

# Perform cross-validation for KNN and GaussianNB
knn_scores = cross_val_score(knn, X, y, cv=cv, scoring='accuracy')
gnb_scores = cross_val_score(gnb, X, y, cv=cv, scoring='accuracy')

plt.boxplot([knn_scores, gnb_scores])
plt.xticks([1, 2], ['KNN (k=5)', 'Gaussian Naives Bayes'])
plt.grid(True)
plt.title('Boxplots of KNN (k=5) and Gaussian Naives Bayes')
plt.ylabel('Accuracy')
```

Text(0, 0.5, 'Accuracy')



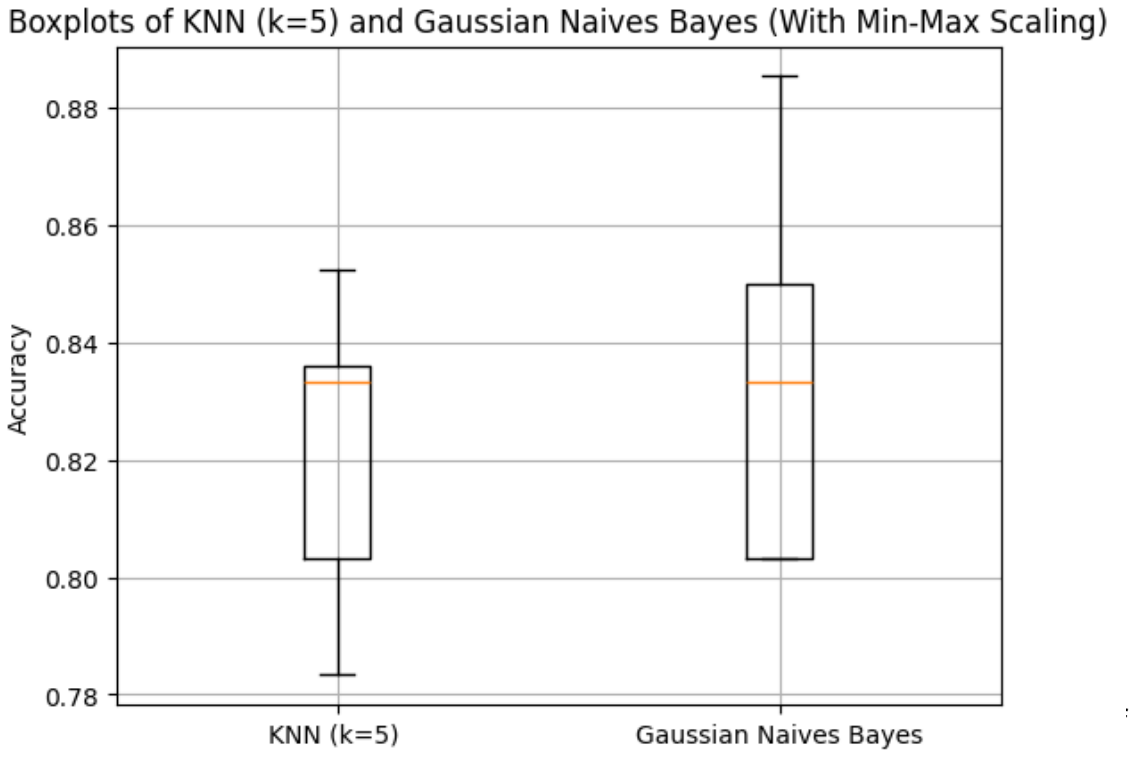
Trabalho realizado por:
Daniel Évora, 99473
Natacha Sousa, 107413

It is clear from the boxplot that the accuracy of Gaussian Naive Bayes has a minimum value that is higher than all of the accuracy parameters for K-Nearest Neighbors (K=5). This indicates that Gaussian Naive Bayes is the better estimator in this comparison. This outcome is expected, as KNN does not involve variable normalization, meaning that variables with different scales can disproportionately influence the distance calculations, causing them to have a greater impact on the model's performance.

b)

```
from sklearn.preprocessing import MinMaxScaler

# Aplicar Min-Max Scaling aos dados
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)  # Assumindo que 'X' já foi definido com seus dados

# Perform cross-validation for KNN and GaussianNB with scaled data
knn_scaled_scores = cross_val_score(knn, X_scaled, y, cv=cv, scoring='accuracy')
gnb_scaled_scores = cross_val_score(gnb, X_scaled, y, cv=cv, scoring='accuracy')


plt.boxplot([knn_scaled_scores, gnb_scaled_scores])

plt.xticks([1, 2], ['KNN (k=5) ', 'Gaussian Naives Bayes'])
plt.grid(True)
plt.title('Boxplots of KNN (k=5) and Gaussian Naives Bayes (With Min-Max Scaling)')
plt.ylabel('Accuracy')
```

Text(0, 0.5, 'Accuracy')



the data and do not depend on the magnitude of the features. However, for K-Nearest Neighbors (KNN), the accuracy results show a significant improvement after normalization. This is because the Min-Max scaling ensures that no variable, due to its magnitude, computes a disproportionately large distance compared to other variables or instances. As a result, normalization helps KNN treat all features more equitably in its distance-based calculations, leading to better performance.

c)

```
from scipy import stats


# Perform a paired t-test (one-tailed)
t_stat, p_value = stats.ttest_rel(knn_scores, gnb_scores, alternative='greater')

# Adjust p-value for one-tailed test
p_value_one_tailed = p_value / 2 if t_stat > 0 else 1 - (p_value / 2)

# Print the results
print(f"T-statistic: {t_stat}")
print(f"One-tailed P-value: {p_value}")
print(f" P-value one tailed: {p_value_one_tailed}")
```

T-statistic: -0.7270523395133756

One-tailed P-value: 0.7462688051215336

P-value one tailed: 0.6268655974392332

With a t-statistic of -0.727 and p-values above 0.05, both one-tailed tests indicate that there is not enough evidence to reject the null hypothesis. This means that the difference between the observed mean and the expected mean (null hypothesis) is not statistically significant.

TÉCNICO LISBOA

Homework 2

Trabalho realizado por:
Daniel Évora, 99473
Natacha Sousa, 107413

2. a)

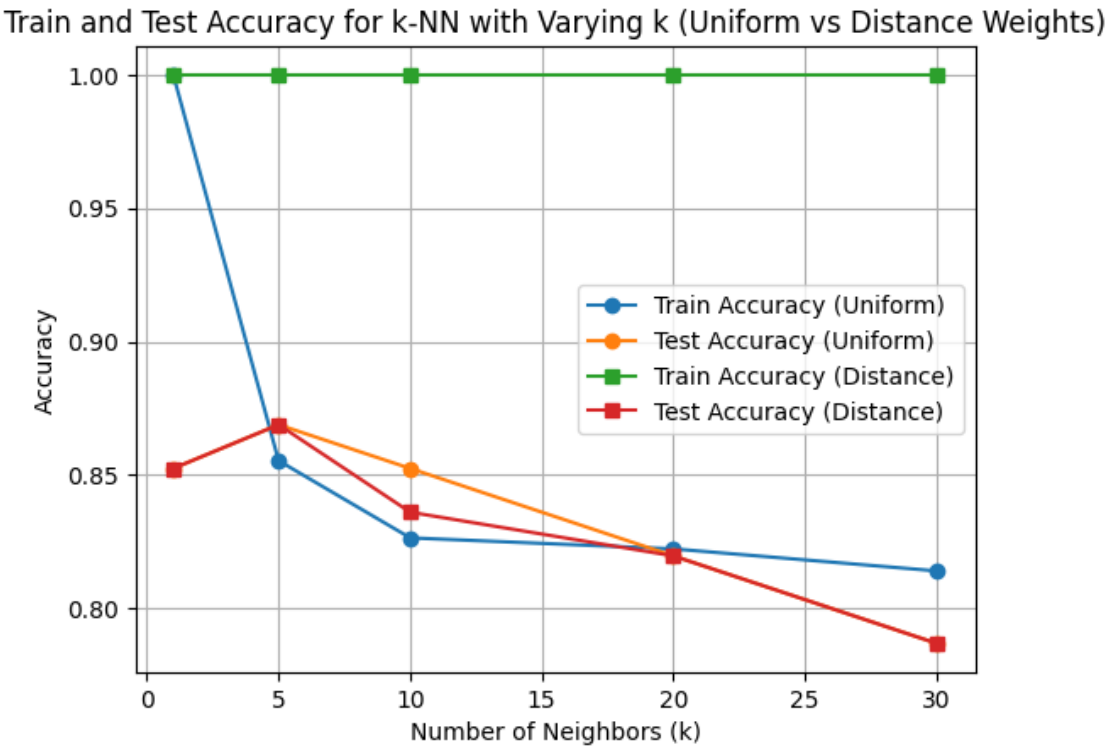
```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt

# Load the CSV dataset using pandas
file_path = 'heart-disease.csv'
heart_data = pd.read_csv(file_path)

# Separate features (X) and target (y)
X = heart_data.drop(columns=['target'])
y = heart_data['target']

# Apply Min-Max scaling to the data
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)

# Perform an 80-20 train-test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=0)

# List of k values to try
k_values = [1, 5, 10, 20, 30]

# Initialize lists to store accuracies for plotting
train_accuracy_uniform = []
test_accuracy_uniform = []
train_accuracy_distance = []
test_accuracy_distance = []
```

TÉCNICO LISBOA

Homework 2

Trabalho realizado por:
Daniel Évora, 99473
Natacha Sousa, 107413

```python
# Loop through each value of k and train two classifiers (uniform and distance)
for k in k_values:
    # Train k-NN with uniform weights
    knn_uniform = KNeighborsClassifier(n_neighbors=k, weights='uniform')
    knn_uniform.fit(X_train, y_train)
    train_accuracy_uniform.append(knn_uniform.score(X_train, y_train))
    test_accuracy_uniform.append(knn_uniform.score(X_test, y_test))

    # Train k-NN with distance weights
    knn_distance = KNeighborsClassifier(n_neighbors=k, weights='distance')
    knn_distance.fit(X_train, y_train)
    train_accuracy_distance.append(knn_distance.score(X_train, y_train))
    test_accuracy_distance.append(knn_distance.score(X_test, y_test))

# Plot for uniform weights
plt.plot(k_values, train_accuracy_uniform, marker='o', label='Train Accuracy (Uniform)')
plt.plot(k_values, test_accuracy_uniform, marker='o', label='Test Accuracy (Uniform)')

# Plot for distance weights
plt.plot(k_values, train_accuracy_distance, marker='s', label='Train Accuracy (Distance)')
plt.plot(k_values, test_accuracy_distance, marker='s', label='Test Accuracy (Distance)')

# Customize the plot
plt.title('Train and Test Accuracy for k-NN with Varying k (Uniform vs Distance Weights)')
plt.xlabel('Number of Neighbors (k)')
plt.ylabel('Accuracy')
plt.legend()
plt.grid(True)
plt.show()
```



b)

Increasing the number of neighbors (k) in a k-NN model can reduce its generalization ability. While a moderate k can improve stability, too high a k includes distant neighbors, which can introduce noise and irrelevant information. This leads to poorer predictions, as the model becomes less sensitive to local patterns. In the extreme, where k equals the total training samples, the model simply predicts the majority class, ignoring data structure. Therefore, choosing the right k is key for optimal performance.

This phenomenon can be visualized through performance plots, which often show a decreasing trend in accuracy or other performance metrics as k increases beyond an optimal point. Thus, selecting an appropriate k is crucial for balancing bias and variance and ensuring the model generalizes well.

3.

Violation of the Independence Assumption: Naive Bayes assumes that all features are independent of each other, but in the heart-disease dataset, many features (like age, cholesterol, and blood pressure) are likely correlated. This leads to inaccurate probability estimates and underfitting, as the model oversimplifies the relationships between features.

Naive Bayes requires a large, unbiased dataset for reliable predictions. With small datasets, such as 300 observations, the model may struggle to estimate probabilities accurately, leading to poor performance. Additionally, the dataset must be representative of diverse characteristics to avoid biases. If the observations are not random or come from similar individuals, the model's predictions may not generalize well to new data, making it sensitive to dataset size and composition.