1. Consider the following dataset:

Name	Debt	Salary	Married	Risk
José	High	High	Yes	High
Ana	Low	High	Yes	Low
João	High	High	No	Low
Maria	High	Low	Yes	High
Rui	Low	Low	Yes	High

Predict Risk for the following instance:

Divide = Low; Salary = High; Married = No

a) based on the K-Nearest-Neighbours Classifier, with K=3

b) based on the Naive-Bayes Classifier, using the Laplace estimation to calculate the conditional probabilities, with laplace correction=1.

2. A classification problem involves **four classes** (1, 2, 3, 4). The training data contain 250 instances of each class, so the total are 1000 cases. Suppose that a particular test based on **Account** attribute divided the dataset into 2 groups of the examples.

1st group (eg Account = yes) contains 600 cases	2nd group (eg Account = no) contains 400 cases
250 Examples of Class 1 **150** Examples of Class 2 **150** Examples of Class 3 **50** Examples of Class 4	**0** Examples of Class 1 **100** Examples of Class 2 **100** Examples of Class 3 **200** Examples of Class 4

a) Develop and present the confusion matrix, assuming the classification is based solely on the Account attribute.

b) Calculate the rate of error of the model. What can be concluded?

c) What is the meaning of the value at the intersection of the lines marked with "Class 2" and "Class 3" and the columns for the prediction of classes?

d) How many "true positive" and "false positive" are in relation to "Classe4"?

e) Calculate the precision and recall measures for classes 1 and 4? What can you conclude?

1a)

Instance: Debt= Low; Salary = High; Married =No

Name	Debt	Salary	Married	Risk	Distance
José	High	High	Yes	High	2
Ana	Low	High	Yes	Low	1
João	High	High	No	Low	1
Maria	High	Low	Yes	High	3
Rui	Low	Low	Yes	High	2

R: The K-Nearest Neighbors Classifier, with **K=3**, classifies the instance **Risk=Low**

1.b)

Priori à Prob → Low Risc = 2/5 (**40%**) High Risc = 3/5 (**60%**)

	Frequency		Probability	
Debt	**Low Risc**	**High Risc**	**Low Risc**	**High Risc**
Low	1	1	1/2	1/3
High	1	2	1/2	2/3
	2	3		

	Frequency		Probability	
Salary	**Low Risc**	**High Risc**	**Low Risc**	**High Risc**
Low	0	2	0	2/3
High	2	1	1	1/3
	2	3		

	Frequency		Probability	
Married	**Low Risc**	**High Risc**	**Low Risc**	**High Risc**
yes	1	3	1/2	1
No	1	0	1/2	0
	2	3		

P(Risk=Low | Debt=High, Salary= High, Married=No) = 2/5 x 1/2 x 1 x 1/2 = 1/10 = 0.1

P(Risk=High | Debt=High, Salary= High, Married=No) = 3/5 x 2/3 x 1/3 x 0 = 0

R: P(Risk=High |...) > P(Risk=Low |...) -> The instance is classified with **Low Risk**

Using the Laplace m-estimate approach for the calculation of conditional probabilities with p = 1.

Priori à Prob → Low Risc = 4/9 High Risc = 5/9

Debt

	Probabilities	
	(Risc,Low)	(Risc,High)
Low	2	2
High	2	3

Salary

	Probabilities	
	(Risc,Low)	(Risc,High)
Low	1	3
High	3	2

Married

	Probabilities	
	(Risc,Low)	(Risc,High)
Yes	2	4
No	2	1

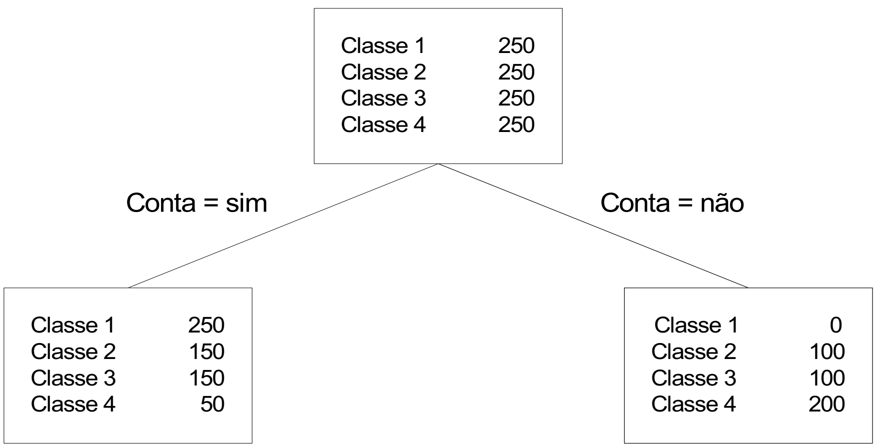
P(Risk=Low | Debt=High, Salary= High, Married=No) = 4/9 x 1/2 x 3/4 x 1/2 = 1/12

P(Risk=High | Debt=High, Salary= High, Married=No) = 5/9 x 3/5 x 2/5 x 1/5 = 2/75

R: P(Risk=Low |...) > P(Risk=High |...) -> The instance is classified with **Low Risk**

2.a) Info(Class) = $-$ 250/1000$\times\log_4$(250/1000) *4 **= 1**

b)



Classe 1 (majority) **Classe4** (majority)

Confusion matrix

	^Classe 1	**^Classe 2**	**^Classe 3**	**^Classe 4**
Classe1	**250**	0	0	0
Classe 2	150	0	0	100
Classe 3	150	0	0	100
Classe 4	50	0	0	**200**

c)

accuracy = (250 + 200) / 1000 = 0.45

Error Rate = 1 - accuracy = 0.55

The model misses more than hits because the error rate (55%) is higher than the hit rate (45%).

d) Means that the classifier totally **predicts** wrong classes 2 and 3.

e)

	Classe 1^	**≠Classe 1^**
Classe 1	250 (TP)	0 (FN)
≠Classe 1	350 (FP)	400 (TN)

	Classe 4^	**≠Classe 4^**
Classe 4	200 (TP)	50 (FN)
≠Classe 4	200 (FP)	550 (TN)

"True positives" Class 1: 250

"False positives" Class1: 350

The success rate of the Account attribute relative to Class 1 is negative, because it misses more than hits the prediction of this class

"True positives" Classe4: 200

"False positives" Classe4: 200

The success rate of the Account attribute regarding Classe4 is annulled by the FP, ie, this model performs a random prediction regarding class4

f)

Class 1

Precision = 250/(250+350) = 42%

Recall = 250/250 = 100%

F1 = 500 / (500 +350) = 59%

Class 4

Precision = 200/(200+200) = 50%

Recall = 200/(200+50) = 80%

F1 = 400 / (400 + 200 + 50) = 62%

Admitting that we have costs associated with false predictions (FP, FN) and we intend to **minimize both costs**, the best measure to evaluate a model is the **F1 measure**, because it is a weighted harmonic mean of precision and recall