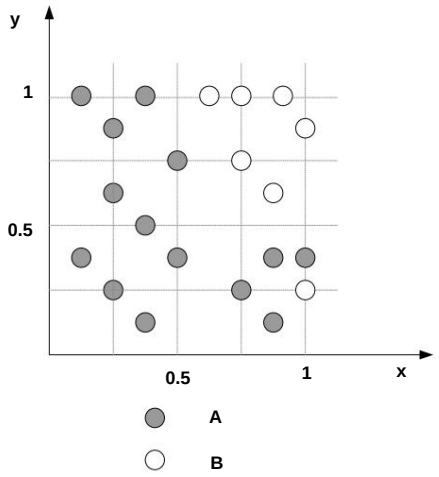
4 values

1. Consider the training set represented in the figure below, where the x and y values are restricted to the range [0.0, 1.0]. In this set, although the attributes are originally continuous, all values were discretized using the value 0.5 as a limit.



a) Determine the uncertainty associated with this training set

b) Present the decision tree generated with this training set, using the Gain of Information

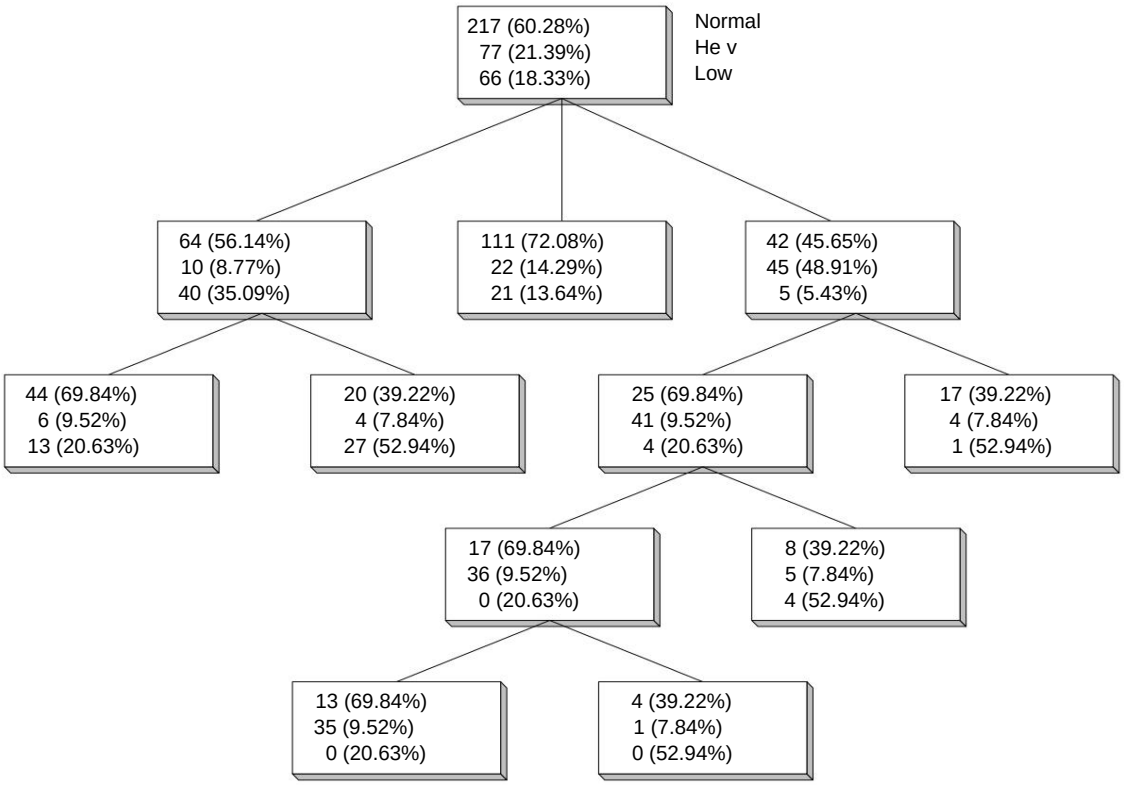
c) What are the main advantages and disadvantages of tree-based classification algorithms of decision.

d) Indicate advantages and disadvantages of attribute discretization when applied in construction of a classification model

e) Briefly present two processes for estimating the error of a classifier in a set of data

5 values

2. Consider the following decision tree obtained with the C5.0 algorithm applied to a set of data



a) Build the confusion matrix relative to this tree

b) Calculate the Accuracy rate of the model, what can you conclude?

c) What is the meaning of the value at the intersection of the line identified with the text **"Low"** with the column identified with the text **"^Elev" ?**

d) Which of the predictions: **^Normal, ^Elev, ^Low** presents the best True Positive rate compared to for False Positives, justify by presenting the calculations. What can you conclude?

e) Present the final form of the decision tree above, after applying the pruning algorithm given in classes that considers the following rule for calculating error estimates:

estimate (errors, #examples) = (2 * errors + 1) / (1 + #examples)

Justify your answer by presenting the calculations

isep instituto superior de engenharia do porto

Department of Computer Engineering

Knowledge Discovery

Master's degree Computer Engineering

Época Resource Exam

July 21, 2022

Duration 2 hours

5 values

3. Consider the following set of 1-dimensional points: {6, 12, 18, 24, 30, 42, 48}

a) For each of the following initial sets of centroids:

i. 18 and 45

ii. 15 and 40

b) Create two clusters by assigning each point to the closest centroid and then calculate the sum of the squared errors for each of the cluster sets. Show both clusters and the sum of the squared errors.

c) Both sets of centroids present stable solutions, that is, if the k-means algorithm on this set of points with the same initial centroids would obtain the same clusters? Justify your answer without resorting to calculations.

d) What clusters are produced by the single-link algorithm, minimum distance?

e) What definition of clustering do the partitions obtained with the K-means algorithm correspond to? (consider the partition with the lowest sum of the squared errors) and with the single-link algorithm?

f) What is the characteristic of the K-means algorithm that explains the previous behavior?

4 values

4. Consider the following sales contingency table for the three items A, B and C.

			A	
			0	1
C = 0	**B**		20	15
		1 0	0	25
C = 1	**B**	1	15	10
		0	5	10

a) Indicate the Support, Trust and Interest of the rules:

i. BC® A

ii. B ® AC

iii. What can you conclude from the items in these rules?

b) Assume that the rule **{a, b}** ® **{c, d} is found in the final set** of rules presented by the Apriori algorithm and the rule **{c, d}** ® **{a, b} is not found.**

For each of the following rules:

i. {a, b, c} ®{d}

ii. {a} ®{b, c, d}

iii. {b, c, d}®{a}

iv. {c} ®{a, b, d}

Indicate, justifying whether:

a) the rule is in the final set

b) the rule is not in the final set

c) there is a possibility of the rule appearing in the final set of rules

2 values

5. Consider the following data mining problem. Given a database constituted

through the tables:

Customer (IdCli, City, Postal Code)

Store (StoreId, Name, City, Postal Code)

Purchase (IdCli, IdLoja, Item, Date)

The aim is to understand the evolution of purchases by customers who live in the city with the

CodPostal A, who made purchases in store B, which is in the city with CodPostal C.

a) Of the algorithms studied, indicate the most appropriate one to solve this problem

b) Define the parameters necessary for the algorithm as well as the data you would use to extract

the intended patterns, and the results that will be provided by the algorithm

c) Will any adaptation to the indicated algorithm be necessary? Justify.

END