Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Engenharia de Atributos

january 10th, 2025
Normal Exam
Duration: 2:00h

Observations:
Exam with limited access to student notes (2 A4 sheets)
You can use a calculator to perform the mathematical operations
Any attempt of fraud will result in a grade of 0 points for all the people involved.

1 – Consider the following dataset:

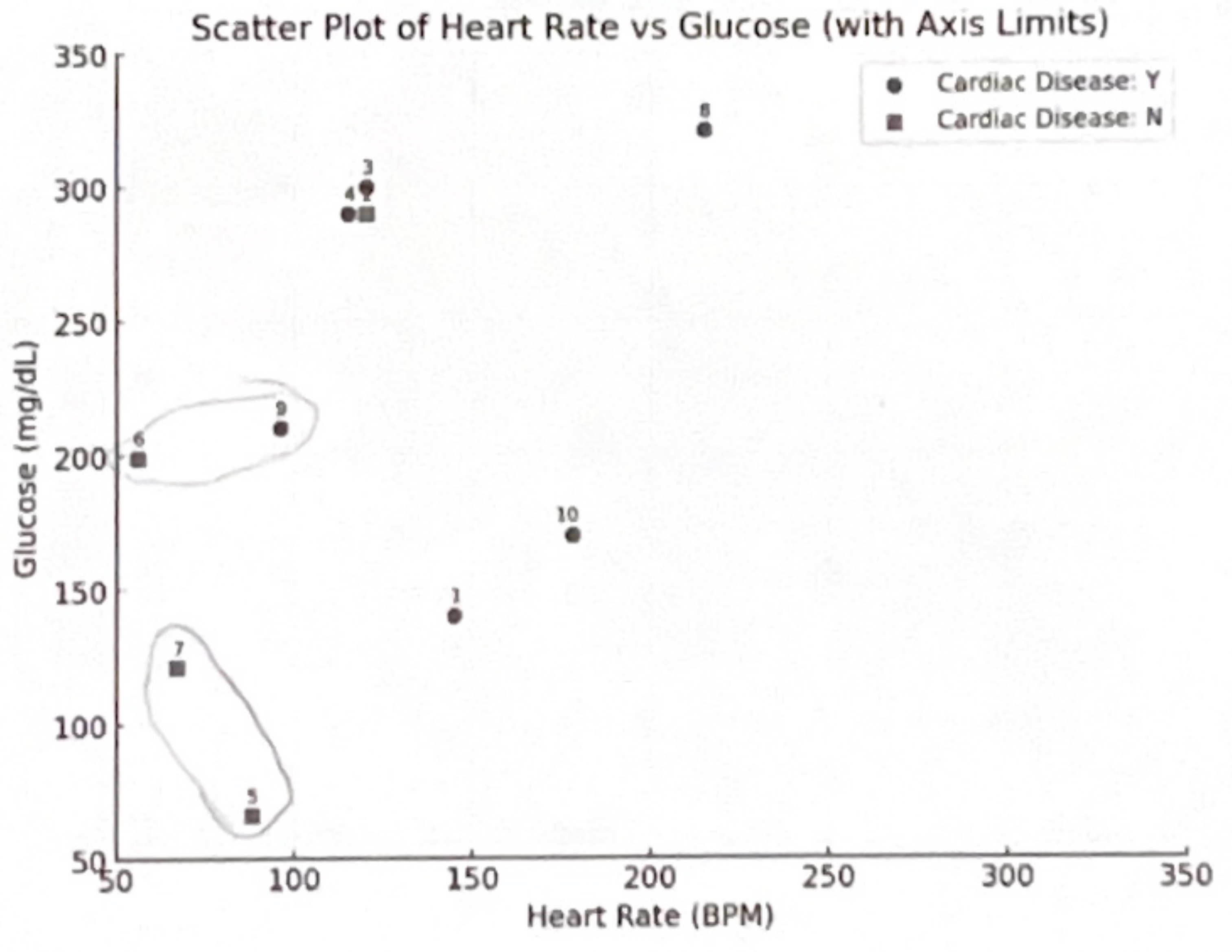
Table 1: Data relative to 10 subjects. It contains 4 independent variables (hypertension, edema, heart rate and glucose) and 1 dependent variable (cardiac disease).

Index	1	2	3	4	5	6	7	8	9	10
Hipertension	Y	Y	Y	Y	Y	N	Y	Y	N	N
Edema	Y	Y	N	N	Y	N	Y	Y	Y	Y
Heart Rate (BPM)	145	120	120	115	88	56	67	215	96	178
Glucose (mg/dL)	140	290	300	290	66	199	121	321	210	170
Cardiac Disease	Y	N	Y	Y	N	N	N	Y	Y	Y

Y = YES, N = NO

Descriptive statistics of HeartRate and Glucose levels:

	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX
HEART RATE (BPM)	10	120,00	48,93	56,00	90,00	117,50	138,75	215,00
GLUCOSE (MG/DL)	10	210,70	87,24	66,00	147,5	204,50	290,00	321,00



Distance between samples considering the variables Heart Rate and Glucose:

D(i,j)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	0,00	152,07	161,94	152,97	93,41	106,78	80,28	194,06	85,45	44,60
2	152,07	0,00	10,00	5,00	226,27	111,25	177,12	99,93	83,52	133,28
3	161,94	10,00	0,00	11,18	236,18	119,57	186,68	97,29	93,15	142,35
4	152,97	5,00	11,18	0,00	225,62	108,45	175,68	104,69	82,23	135,53
5	93,41	226,27	236,18	225,62	0,00	136,80	58,87	284,88	144,22	137,54
6	106,78	111,25	119,57	108,45	136,80	0,00	78,77	200,41	41,48	125,40
7	80,28	177,12	186,68	175,68	58,87	78,77	0,00	248,81	93,61	121,33
8	194,06	99,93	97,29	104,69	284,88	200,41	248,81	0,00	162,73	155,47
9	85,45	83,52	93,15	82,23	144,22	41,48	93,61	162,73	0,00	91,24
10	44,60	133,28	142,35	135,53	137,54	125,40	121,33	155,47	91,24	0,00

1.1 - Classify each variable as nominal, ordinal or numeric.

1.2 - Using the InterQuartile Range (IQR) method, identify possible outliers in the Heart Rate and Glucose variables. Show the calculations.

1.3 - Which feature (Heart Rate or Glucose) has better discriminative power in identifying cardiac disease? Justify your answer with the Relief score of the variables. Compute the Relief using only the samples with indexes 1, 3 and 7. You can use the following distance table to help your calculus. For each cell, the table presents the Euclidean distance between subjects with index *i* and *j*, represented in the corresponding line and column of the table, respectively.

1.4 - If you were to apply the Neighborhood Cleaning Rule (NCS) to the samples with indexes 5 and 9 of the dataset, considering only the Heart Rate and Glucose variables, which samples would be removed? Justify.

1.5 - Which feature (Hypertension or Edema) has better discriminative power in identifying Cardiac Disease? Justify your answer with the Goodman-Kruskall Lambda.

2 - Consider a dataset D with 100 samples and five numeric variables: F1, F2, F3, F4 and F5. To reduce the dimensionality of the data, a Principal Component Analysis was applied, which generated the following eigenvectors W and eigenvalues λ:

$$
\begin{array}{ccccc}
F_1 & F_2 & F_3 & F_4 & F_5 \\
\end{array}
$$

$$
W = [\begin{array}{ccccc}
-0{,}6 & 0{,}5 & 0{,}6 & 0{,}0 & 0{,}2\ ; \\
-0{,}2 & -0{,}6 & 0{,}5 & -0{,}1 & -0{,}6\ ; \\
0{,}0 & 0{,}1 & 0{,}0 & -1{,}0 & 0{,}1\ ; \\
0{,}7 & 0{,}1 & 0{,}6 & 0{,}0 & 0{,}2\ ; \\
0{,}2 & 0{,}6 & -0{,}1 & 0{,}0 & -0{,}8
\end{array}]
$$

$$
\begin{array}{ccccc}
1{,}8\% & 25{,}6\% & 9{,}1\% & 48{,}6\% & 14{,}6\% \\
\end{array}
$$

$$
\lambda = [\ 0.2 \quad 2.8 \quad 1 \quad 5.3 \quad 1.6\ ]
$$

2.1 - How many principal components are needed to explain 80% of the variance of the dataset? Justify your answer.

2.2 - Project the following example into the two principal components that explain the most variance of the dataset:

S = {F1: 1, F2: 0.5, F3: 0, F4: 1, F5: 0}