A4.2

Data preprocessing

CHAPTER ASSIGNMENT · 17 MARKS

Name _____ Class _____ No. _____ Date _____

Answer every question in the space given. Show your working. You may use the chapter page to help you.

Questions from the Guru Jazzy chapter page, for classroom use.

Exam-style questions

1 Describe one common issue in customer survey data that requires cleaning. [2]

2 Describe how feature selection affects a model's performance. [3]

3 Outline the role of dimensionality reduction for high-dimensional data. [3]

4 Explain why features should be scaled before training. [2]

5 A school district plans to feed raw marks for every test question into a model. Discuss the implications. [4]

6 Compare filter and wrapper methods. [3]

Chapter quiz

Q1 Why does poor data give a poor model?

A	B	C	D
The model learns the errors	Data has no effect	It uses more electricity	Models ignore data

Q2 Which rule flags a value more than 1.5 × IQR beyond the quartiles?

A	B	C	D
The IQR method	Z-score	Hashing	PCA

Q3 Normalization usually...

A	B	C	D
Rescales values to a set range	Removes all outliers	Adds features	Encrypts data

Q4 Standardization gives data...

A	B	C	D
Mean 0 and standard deviation 1	A range of 0 to 1 always	No decimals	A bell shape

Q5 Which feature selection method is usually cheapest?

A	B	C	D
Filter	Wrapper	Embedded	Brute force

Q6 Which considers feature interaction by training on subsets?

A	B	C	D
Wrapper	Filter	Hashing	Spooling

Q7 With a fixed number of records, adding features makes data...

A	B	C	D
Sparser	Denser	Cleaner	Smaller

Q8 Overfitting means a model...

A	B	C	D
Memorizes training data, noise included	Trains too slowly	Has too few inputs	Uses too little memory

---

Q9 Which technique uses class labels to reduce dimensions?

A	B	C	D
LDA	PCA	Z-score	IQR

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/ib-computer-science/a4-2-data-preprocessing/