

SCHOOL OF SCIENCE AND TECHNOLOGY
Semester 1, 2026

FINAL EXAMINATION PAPER

ITC357: PRACTICAL DATA SCIENCE

Important Instructions to Students:

1. Read the questions, instructions carefully and answer the questions.
2. You are allowed an **extra 10 minutes reading time** (during this time, you are not allowed to write with pen) and **180 minutes (3 Hours) to complete this paper**.
3. Silent, non-programmable calculators are allowed, but not supplied.
4. Write all your answers in the answer booklet provided. If you need extra sheets, please ask from the supervisor. Write your name, course code and student identification number in the answer booklet and in the attendance sheet.
5. This paper contains four (4) sections. All questions are compulsory to answer.
6. Write all your answers legibly. The marker has the right to award zero for answers that are not legible.
7. This is a **Closed Book Examination**. The University views plagiarism, cheating and dishonest practice as serious offences and if found, offenders are penalized.

Examination Paper Content (Format):

<u>Section</u>	<u>Question</u>	<u>Content</u>	<u>Marks</u>	<u>Suggested Time</u>
A	1 -5	True/False	10	18 minutes
B	6- 10	Multiple Choice	10	18 minutes
C	11 -20	Short Answer	20	36 minutes
D	21 -24	Case Studies	60	108 minutes
		TOTAL	100	180 Minutes

Section A: True/False (10 Marks)

1. Data science combines statistics, computing, and domain knowledge to extract insights from data.
(True / False)
2. In a dataset, the observational unit refers to the variable being measured.
(True / False)
3. Scatterplots are useful for visualizing the relationship between two quantitative variables.
(True / False)
4. If the mean is larger than the median, the distribution is typically left-skewed.
(True / False)
5. Random sampling reduces bias and improves the representativeness of a sample.
(True / False)

Section B: Multiple Choice (10 Marks)

6. Which of the following best describes the goal of data science?
A. Store large datasets
B. Extract insights and knowledge from data
C. Replace statistical analysis
D. Eliminate uncertainty in decision making
7. Which term refers to the entire group of individuals that a study aims to understand?
A. Sample
B. Variable
C. Population
D. Observation
8. Which visualization is most appropriate for showing the distribution of a single quantitative variable?
A. Histogram
B. Scatterplot
C. Pie chart
D. Line graph
9. What does a boxplot primarily help identify?
A. Relationships between variables
B. Data types
C. Distribution characteristics and outliers
D. Data cleaning errors
10. Which database operation is used to combine rows from two tables based on a shared key?
A. Filter
B. Join
C. Sort
D. Aggregate

Section C: Short Answer Questions (20 Marks)

Answer the following questions in brief. (2 marks each).

11. Describe the main objective of data science?
12. Describe key steps in data wrangling.
13. Explain the four types of table joins and their business implications using an example?
14. What is the purpose of hypothesis testing?
15. Describe machine learning.
16. What is the difference between regression and classification?
17. Explain how a decision tree makes predictions.
18. What is unsupervised learning?
19. What is clustering?
20. What is an advantage of random forests over single decision trees?

Section D: Case Studies (60 Marks)

21. Case Study: Exploratory Data Visualisation (FSC Context)

(15 marks)

The Fiji Sugar Corporation (FSC) is analysing sugarcane production data from **4,000 farmers**. The dataset includes the following fields:

- farmer_id, region, farm_size_ha, tonnes_delivered, sugar_content, and burnt_flag (Burnt / Non-burnt cane)

FSC wants to understand:

- How tonnes delivered is distributed across farmers
- Whether farm size influences tonnes delivered
- How production differs across regions
- Whether burnt cane affects production patterns

Tasks (3 marks each):

- FSC wants to visualise the distribution of *tonnes_delivered across farmers*. Which type of graph should be used and what insights this graph would provide.
- FSC wants to analyse the relationship between *farm_size_ha* (farm size) and *tonnes_delivered* (production). Which visualisation should be used, and What would a positive relationship look like in this graph?
- FSC wants to compare *tonnes_delivered* across different regions. Which visualisation is most appropriate, and what insights can this visualisation provide?
- FSC suspects that some farmers have unusually high or low **tonnes_delivered** values. Explain how visualisation can help identify these anomalies and which graph is most suitable.
- FSC creates a visualisation where:
 - x-axis = **farm_size_ha**
 - y-axis = **tonnes_delivered**
 - bubble size = **sugar_content**
 - colour = **region**

What type of visualisation is this, and explain what additional insights this visualisation provides compared to a simple scatterplot.

22. Case Study: Data Management Scenario

(15 marks)

FSC stores data in separate tables:

- **farmers.csv:** farmer_id, region
- **cane_deliveries.csv:** farmer_id, tonnes

Issues:

- Some farmers have no deliveries
- Some deliveries have no matching farmer
- Data stored in separate systems

Tasks (3 marks each):

- a) Which join should be used for **valid analysis**?
- b) Which join identifies **farmers with no deliveries**?
- c) Which join identifies **data errors (unknown farmers)**?
- d) Which join gives a **complete dataset for auditing**?
- e) Why can using the wrong join mislead FSC decisions?

23. Case Study: Machine Learning and Prediction

(15 marks)

Refer to the three figures below to answer the questions. Each question is worth 1.5 marks.

Fonterra is a large New Zealand-based dairy cooperative owned by thousands of dairy farmers. It is one of the world's biggest exporters of dairy products, producing items such as milk powder, cheese, butter, and other dairy ingredients for global markets. Fonterra relies heavily on data analytics, forecasting, logistics, supply chain management, and quality monitoring to support its dairy operations and business decisions.

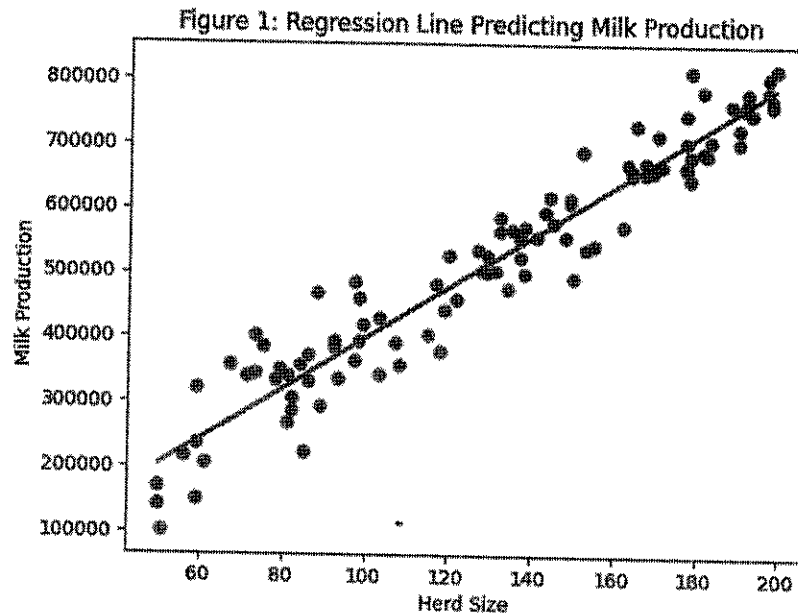
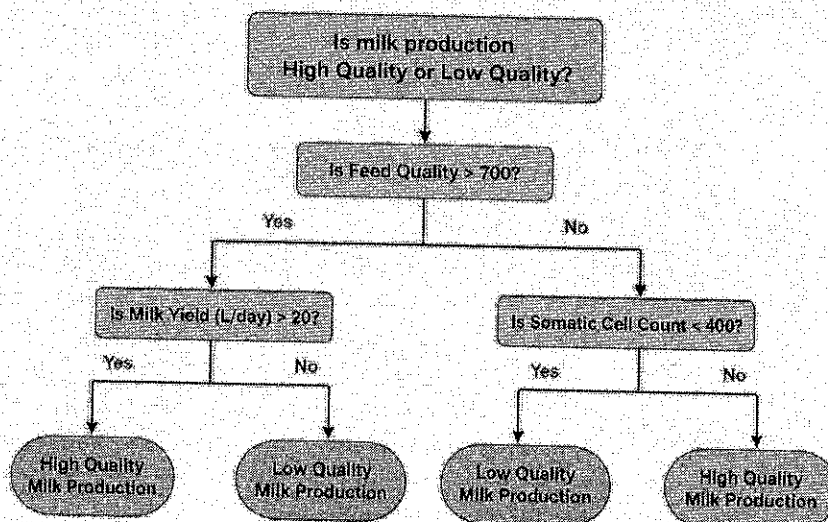
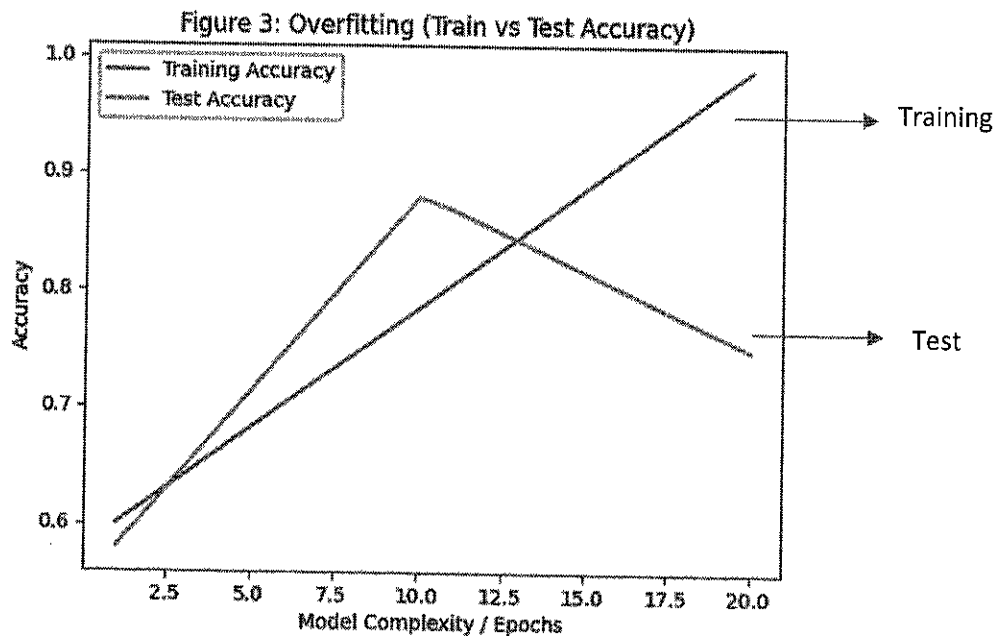


Figure 2. Decision Tree for Milk quality



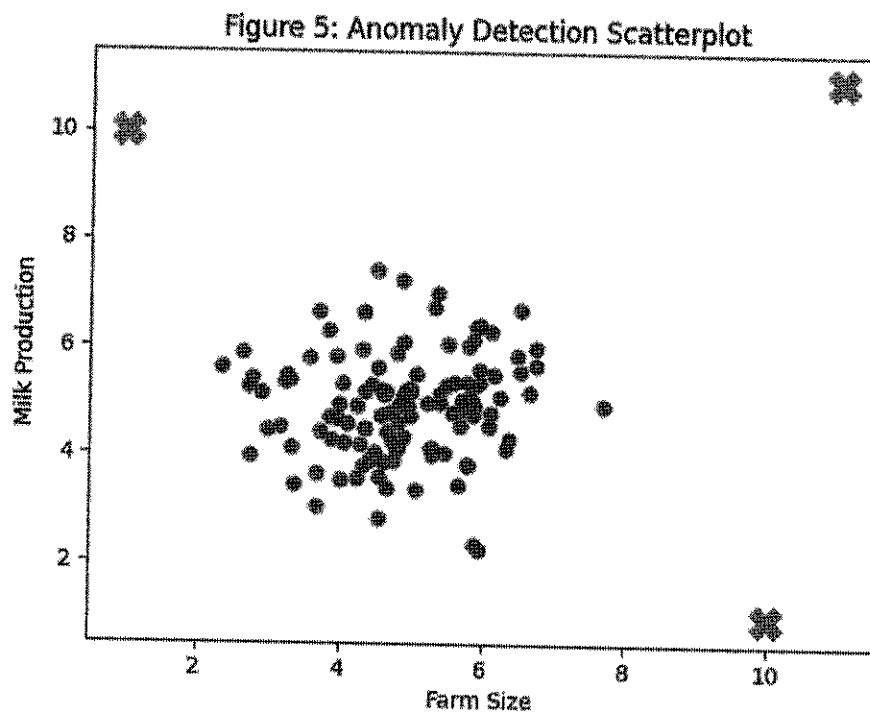
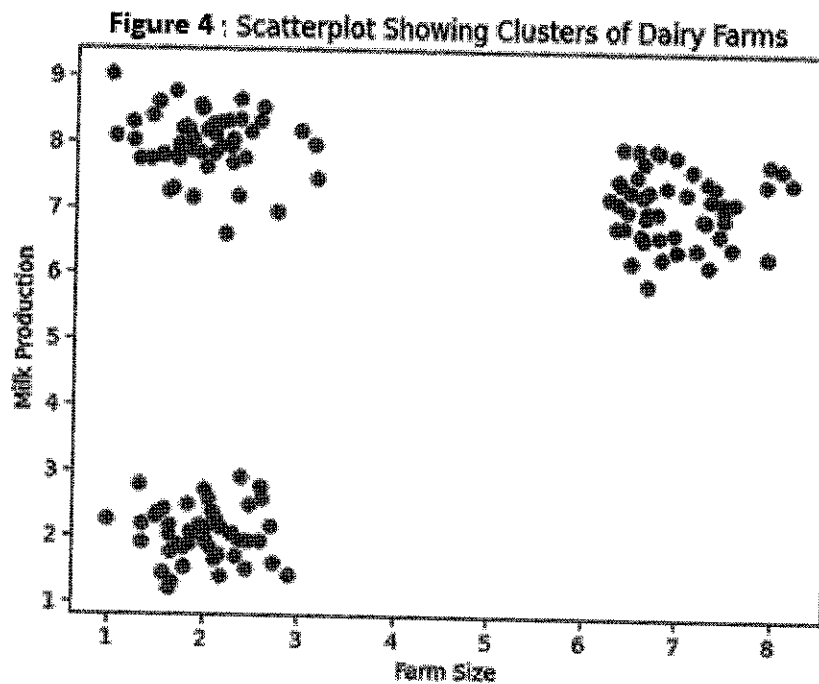


- a) Fonterra uses a system that predicts milk production based on herd size, weather, and feed quality. This is an example of _____.
- b) Fonterra trains a model using historical milk production data where the correct production values are already known. This is _____.
- c) Fonterra wants to identify groups of similar farms without predefined labels. Which method is most appropriate?
- d) (Regression) Refer to Figure 1. What is the primary purpose of the regression line?
- e) Refer to Figure 2. What is the decision tree predicting?
- f) Refer to Figure 3. A model performs very well on training data but poorly on test data. This indicates _____.
- g) Explain unsupervised learning using a Fonterra example.
- h) Refer to Figure 1. Explain regression.
- i) Refer to Figure 2. Explain how a decision tree works.
- j) Explain difference between regression and classification.

24. Case Study: Supervised Learning and Modelling

(15 marks)

Refer to the three figures below to answer the questions. Each question is worth 1.5 marks.



a) Fonterra wants to group farms with similar production characteristics without predefined labels. This is an example of _____.

- b) Refer to Figure 5. Which farms are most likely anomalies: *Farms inside dense clusters* or *Farms far away from other farms*?
- c) Refer to Figure 4. What do dense groups of farms represent?
- d) Explain unsupervised learning using a Fonterra example.
- e) Refer to Figure 4. Explain how clusters appear in scatterplots.
- f) Explain why clustering is useful for Fonterra.
- g) Refer to Figure 5. Explain anomaly detection.
- h) Refer to Figure 4. What indicates the presence of clusters: *Dense groups of nearby points* or *Random noise*?
- i) Why would Fonterra use clustering: *Predict exact future values* or *Discover natural groupings of farms*?
- j) If two farms appear very close together in a scatterplot, what does this suggest: *they are highly similar* or *they are unrelated*?

THE END