Status	Finished
Started	Monday, 23 June 2025, 8:17 AM
Completed	Monday, 23 June 2025, 10:44 AM
Duration	2 hours 27 mins

Information

Examination: Machine Learning 382

Total: 100 marks

Duration: 120 + 30 minutes

Examiner: Machingauta MH

Moderator: Tavagwisa C

NB: Questions are given with instructions where necessary.

Question 1

Complete

Marked out of 1.00

Fraud detection, Image classification and Customer retention are applications of which of the following?

a. Unsupervised learning: Regression.

b. Unsupervised learning: Clustering.

c. Supervised learning: Classification.

d. Reinforcement learning.

Question 2

Complete

Marked out of 1.00

Which type of learning is characterized by an agent learning through interactions with an environment and receiving rewards?

a. Semi-supervised learning.

b. Reinforcement learning.

c. Unsupervised learning.

d. Supervised learning.

Question 3

Complete

Marked out of 1.00

Which machine learning approach is based on the assumption that similar data points are more likely to have the same labels?

- a. Classification.
- b. Clustering.
- c. Anomaly detection.
- d. Regression.

Question 4

Complete

Marked out of 1.00

What is the main purpose of the Modeling phase in CRISP-DM?

Select one:

- To visualize the final results
- To define the business problem
- To apply algorithms to data and generate predictive models
- To clean and transform the data
- To evaluate model deployment success

Question 5

Complete

Marked out of 1.00

What is the main focus of the Data Understanding phase in CRISP-DM?

Select one:

- Developing project budget plans
- Defining business success criteria
- Designing user interfaces for deployment
- Deploying machine learning models
- Exploring and assessing data quality and structure

Question 6

Complete

Marked out of 1.00

Which of the following is an appropriate method to evaluate a classification model in the Evaluation phase of CRISP-DM?

Select one:

- Visualizing feature histograms
- Performing time-series decomposition
- Normalizing raw features
- Using precision, recall, and F1-score on a test dataset
- Running a clustering algorithm

Question 7

Complete

Marked out of 1.00

What does the term "feature scaling" refer to in machine learning?

- a. Changing the range of features to a standard scale.
- b. Removing irrelevant features.
- c. Increasing the number of training examples.
- d. Adding more features to the dataset.

Question 8

Complete

Marked out of 1.00

You are working with a dataset that contains categorical variables. These categorical variables need to be converted into numerical representations for further analysis. Which data wrangling technique is appropriate for this task?

- Principal Component Analysis (PCA).
- Normalization.
- Feature scaling.
- One-Hot Encoding.

Question 9

Complete

Marked out of 1.00

What is a common evaluation metric for classification tasks in the presence of class imbalance?

a. R-squared.

b. F1-score.

c. Mean-Squared Error.

d. Accuracy.

Question 10

Complete

Marked out of 1.00

What is the purpose of decomposing a time series?

To identify the best forecasting method for a given dataset.

To separate the time series into its trend, seasonality, and residual components.

To transform the time series into a stationary series.

To identify optimal parameters for a given model.

Question 11

Complete

Marked out of 1.00

What is the primary goal of time series forecasting?

a. To classify data points into different categories.

b. To find the optimal parameters for a given model.

c. To identify patterns in historical data.

d. To predict future values based on past observations.

Question 12

Complete

Marked out of 1.00

What is the primary goal of a Support Vector Machine (SVM) algorithm in machine learning?

- a. To maximize the margin between classes.
- b. To perform unsupervised learning.
- c. To visualize data relationships.
- d. To minimize the number of support vectors.

Question 13

Complete

Marked out of 1.00

Which optimization algorithm is commonly used to update the weights of neural networks during training?

- a. Apriori algorithm.
- b. K-Means.
- c. Decision trees.
- d. Gradient descent.

Question 14

Complete

Marked out of 1.00

What does backpropagation in neural networks mean?

- a. Forward pass of data.
- b. Activation of output layer.
- c. Training process for adjusting weights.
- d. Initialization of weights before training.

Question 15

Complete

Marked out of 8.00

Comparing Classification and Regression

Drag and drop the correct answer into the relevant space in the table below.

Aspect	Classification	Regression
Machine learning category	Supervised Learning	Unsupervised learning
Goal	Grouping similar customers	Predict continuous numeric values
Example algorithm	K-Means clustering	Linear Regression
Example application	Email spam detection	House price prediction

Logistic Regression

Predict discrete class labels

Supervised Learning

Decision tree regression

Question 16

Complete

Marked out of 5.00

Data Scalars

Data scaling is a technique for transforming the values of variables within a dataset. There are several data scalers provided by the `scikit-learn` library.

For the given scalers, drag and drop the correct description and use case to the correct scaler.

Scaler	Description	Use Case
StandardScaler	Maps data to a uniform or normal distribution via quantiles.	Features vary widely in scale but no strong outliers.
MinMaxScaler	Scales features to a given or specified range.	Need uniform or normal distribution explicitly.
RobustScaler	Centers the data and scales it to unit variance.	Features contain extreme outliers
QuantileTransformer	Scales using the median and IQR (Interquartile Range).	Assumes data is normally distributed.
PowerTransformer	Stabilizes variance and makes data more Gaussian-like.	Want to stabilize variance and make data normal.

Question 17

Complete

Marked out of 6.00

Data Science Workflow

Data Science Workflow defines the phases or steps in a data science project. One of the most commonly used data science workflows is the CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) workflow. List the CRISP-DM workflow steps in their correct order.

Step 1	Business Understanding
Step 2	Data Understanding
Step 3	Data Preparation
Step 4	Modeling
Step 5	Evaluation
Step 6	Deployment

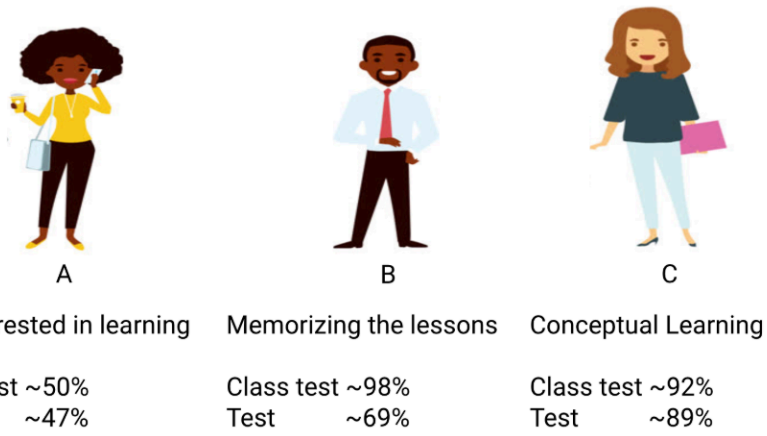
Question 18

Complete

Marked out of 5.00

Model Fitting

Consider the following image showing the learning behaviors of three students (A, B, and C) and their performance in a Class test and final exam (Test). These behaviors can be used to represent a good model fit, underfitting and overfitting in machine learning.



A	B	C
Not interested in learning	Memorizing the lessons	Conceptual Learning
Class test ~50% Test ~47%	Class test ~98% Test ~69%	Class test ~92% Test ~89%

Using this image, answer the following questions.

(a) Overfitting is when the model performs well on training data but poorly on unseen test data, whilst underfitting is when the model is too simple and fails to capture the underlying pattern.

(b) Match each student with the correct machine learning model behavior:

Student	Model Fit Type
A	Underfitting
B	Overfitting
C	Best fit

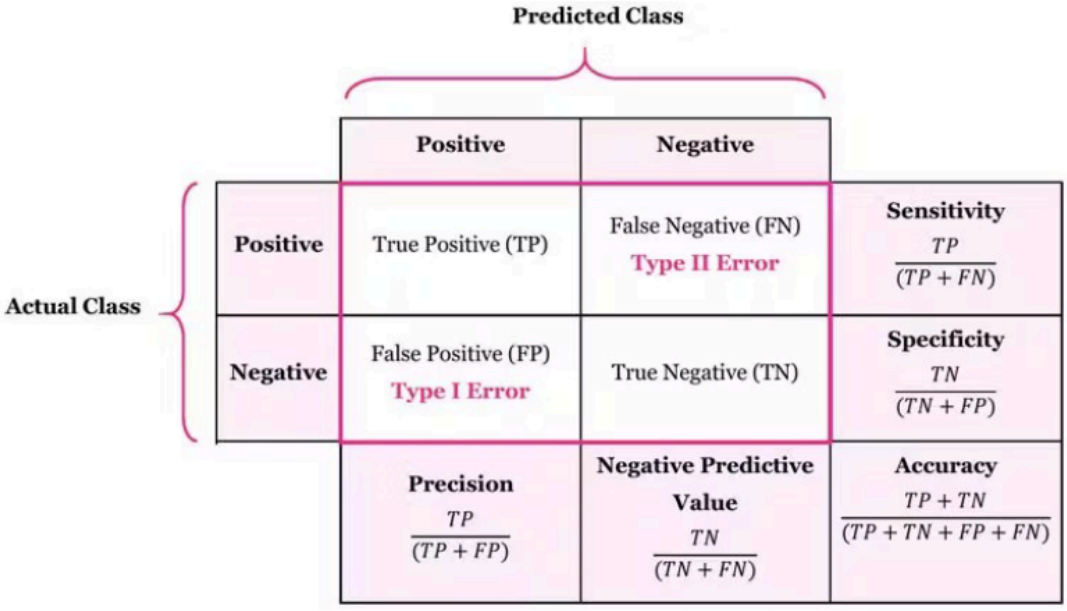
Question 19

Complete

Marked out of 5.00

The Confusion Matrix

The confusion matrix is a foundational tool. Guided by the given diagram depicting the insights that can be gained from a confusion matrix and the given scenario, answer the given questions.



Scenario: In a Covid test of 1000 patients, there were 45 positive tests, of which 30 patients had Covid and 15 were falsely tested positive. Of the 955 negative tests there were 5 that were incorrect, these patients had Covid but were tested negatively.

Using the diagram, find

	Positive	Negative	
Positive	TP = 30	FN = 5	Sensitivity = 0.85
Negative	FP = 15	TN = 950	Specificity = 0.98
	Precision = 0.66	Negative Predictive value = 0.99	Accuracy = 0.98

NB: In the cases where you don't get answers as whole numbers, write your answers correct to 3 decimal places, write the answer in the form 0.123 and not 0,123. In the case where you get an answer as 0.12 write it as 0.120

Question 20

Complete

Marked out of 10.00

Characteristics of Neural Networks

A neural network is a machine learning model inspired by the structure of the human brain. It is made up of layers of neurons that process data.

The input layer receives the raw data and passes it on without computation. The hidden layers perform most of the computation, applying weights and activation functions to the input.

Each connection between neurons has an associated weight, which determines the strength of the connection. Neurons may also have a bias, which shifts the activation value.

The neuron computes the weighted sum of its inputs and passes it through an activation function to produce an output

Information

Pengium Clustering by Physical Characteristics

Data has been collected on physical characteristics of a variety of penguin species.

The dataset consists of 5 columns:

- culmen_length_mm: culmen length (mm)
- culmen_depth_mm: culmen depth (mm)
- flipper_length_mm: flipper length (mm)
- body_mass_g: body mass (g)
- sex: penguin sex

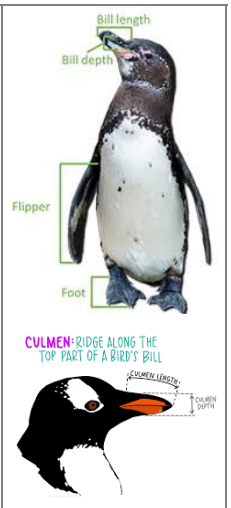
The dataset features:

1. culmen_length_mm: The length of the culmen (mm), which is the top ridge of the bird's ***bill (beak)*** - essentially the distance from the tip of the bill to the base where it meets the face. Think of it like the length of a nose from tip to where it joins the face.

2. culmen_depth_mm: The vertical thickness or depth of the culmen (mm) - measured from the top of the beak to the bottom, essentially how tall the beak is at the base. Similar to the height of a triangle, measured vertically.

3. flipper_length_mm: The length of the penguin's flipper (mm), measured from the shoulder to the tip of the flipper. Flipper size is often used to estimate swimming ability or body size.

These measurements help distinguish between different species of penguins in the dataset.



Note:

- The data (sex) column contains some **NA or invalid** values. You may need to remove these.
- There are huge and unrealistic outliers in the flipper length variable. You may want to remove those points.

Requirements

Based on the physical characteristics, determine the number of species that can be identified using **K-Means algorithm**.

Attachments

1. The dataset is provided in the attached csv file: penguins
2. The template Notebook is provided as a ***zip file*** with additional instructions: 07.CAS.001

Question 21

Complete

Marked out of 35.00

Notebook Upload

Upload your updated Jupyter notebook for this problem.

Marks are assigned as indicated in the supplied notebook.

- Make sure that all cells run without errors.
- ***No marks will be assigned to any cells after an error.***
- Use the following file name convention for the notebook: "**07.CAS.001_Surname_Firstname_StudentID.ipynb**", where you must change the filename to reflect your details:
  - You ***MUST*** replace **Surname** with your surname
  - You ***MUST*** replace **Firstname** with your firstname
  - You ***MUST*** replace **StudentID** with your student id.
  - Do NOT change the part of the filename prefix: "**07.CAS.001_**"
- Do ***NOT*** edit or delete the marks cells.
- You will need to ***insert new cells*** in each section where indicated.

Note: If you do not fully comply with these instructions, the notebook will NOT be marked or you may lose marks. It remains the responsibility of the candidate to ensure they have uploaded the correct file.

Upload the file by dropping the file into the download section below:

07.CAS.001_van Zyl_Dean_600367.zip

Information

Cafe Sales Forecasting

You are provided with a time series dataset that contains monthly cafe' sales data from Australia (in billions AUD) from **April 1982** onwards.

Requirements

You are required to build a time series model and generate a forecast for the next 48 months.

Attachments

1. The dataset is provided in the attached csv file: auscafe_simulated
2. The template Notebook is provided in the zip file with additional instructions: 06.CAS.002

Question 22

Complete

Marked out of 30.00

Notebook Upload

Upload your updated Jupyter notebook for this problem.

Marks are assigned as indicated in the supplied notebook.

- Make sure that all cells run without errors.
- ***No marks will be assigned to any cells after an error.***
- Use the following file name convention for the notebook: "**06.CAS.002_Surname_Firstname_StudentID.ipynb**", where you must change the filename to reflect your details:
  - You MUST replace **Surname** with your surname
  - You MUST replace **Firstname** with your firstname
  - You MUST replace **StudentID** with your student id.
  - Do NOT change the part of the filename prefix: "**06.CAS.002_**"
- Do ***NOT*** edit or delete the marks cells.
- You will need to ***insert new cells*** in each section where indicated.

Note: If you do not fully comply with these instructions, the notebook will NOT be marked or you may lose marks. It remains the responsibility of the candidate to ensure they have uploaded the correct file.

Upload the file by dropping the file into the download section below:

06.CAS.002_van Zyl_Dean_600367.ipynb

Question 23

Complete

Marked out of 12.00

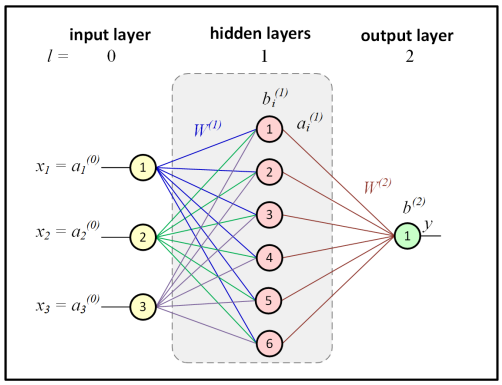
Backpropagation Algorithm

A hospital wants to use an artificial neural network (ANN) to *predict the probability that a patient will be readmitted within 30 days* of being discharged. This aims to help allocate resources and reduce preventable readmissions.

They collect three key features for each discharged patient:

- Length of hospital stay (in days), x_1
- Number of prior admissions, x_2
- Patient age (in years), x_3

The proposed ANN network includes a ***single hidden layer*** with ***6 neurons*** and uses a ***sigmoid activation function*** in the output layer.



Note: For this question you do ***not*** need to submit any code/notebooks.

1. Understanding the Problem

(a) Which of the following best describes the nature and suitability of this ANN?

- The ANN is being used for a classification problem because the output is either 0 and 1.
- The network is unsuitable for this task since sigmoid functions cannot be used in regression problems.
- The ANN is suitable for regression because it outputs a continuous probability of readmission.
- An ANN requires more than one hidden layer to solve this type of problem.

(b) In the patient readmission prediction model using an artificial neural network (ANN), why is a sigmoid activation function appropriate in the output layer?

- It maps the output to a value between 0 and 1, ideal for predicting probabilities.
- It enables the model to predict values outside the 0 to 1 range.
- It simplifies the backpropagation process by linearizing the error.
- It increases training speed compared to other activation functions.

(c) The artificial neural network outputs a value of 0.73 for a particular patient. What does this value represent, and how would the model typically be evaluated?

- The customer is certain to be readmitted in 0.73 months; the model is evaluated using classification accuracy.
- The model is incorrectly configured, as regression problems require ReLU in the output layer.
- The output is a probability (e.g. 73% chance of readmission); evaluation would typically use regression metrics like MSE or MAE.
- The output is a raw score that needs to be normalized post-training.

2. Constructing the Neural Network

Consider the data for a particular patient is:

- Length of hospital stay (in days: $x_1 = 8$
- Number of prior admissions: $x_2 = 3$
- Patient age (in years): $x_3 = 68$
- Probability of readmission: $y = 0.84$

Construct the ANN based on the following weight and bias matrices:

You may find the following python code useful:

```
W1 = np.array([[0.014, 0.017, 0.12],
               [0.005, 0.023, 0.322],
               [0.034, 0.024, 0.044],
               [0.018, 0.161, 0.053],
               [0.014, 0.053, 0.041],
               [0.034, 0.036, 0.081],
              ])
b1 = np.array([[0.03],
               [0.05],
               [0.04],
               [0.02],
               [0.06],
               [0.02],
              ])
W2 = np.array([[0.05, 0.06, 0.07, 0.04, 0.07,
0.36]])
b2 = np.array([[0.07]])
```

- **Hidden Layer:** Weight Matrix $W^{(1)}$ and Bias Vector $b^{(1)}$

$$W^{(1)} = \begin{bmatrix} 0.014 & 0.017 & 0.120 \\ 0.005 & 0.023 & 0.322 \\ 0.034 & 0.024 & 0.044 \\ 0.018 & 0.161 & 0.053 \\ 0.014 & 0.053 & 0.041 \\ 0.034 & 0.036 & 0.081 \end{bmatrix}, \qquad b^{(1)} = \begin{bmatrix} 0.03 \\ 0.05 \\ 0.04 \\ 0.02 \\ 0.06 \\ 0.02 \end{bmatrix}$$

- **Output Layer:** Weight Matrix $W^{(2)}$ and Bias Vector $b^{(2)}$

$$W^{(2)} = [\,0.05 \quad 0.06 \quad 0.07 \quad 0.04 \quad 0.07 \quad 0.36\,], \qquad b^{(2)} = [\,0.07\,]$$

After a forward pass of the neural network algorithm, calculate the following:

- **Hidden Layer activations**: (*Fill in the missing values*)

$a^{(1)} =$

0.99976437
1.0
0.966946
0.98595023
0.95766971
0.99728975

- **Output Layer activation**:

$y = a^{(2)} =$ 0.67110204

- **Output Loss:**

$\mathcal{L} =$ 0.01426326

---

3. Backpropagation Training Step

Implement a backpropagation algorithm to update the weights and biases of all the layers of the network. Use a network training rate of $\eta = 0.22$.

Execute a *single step* of the backpropagation algorithm and provide the following:

- **Output Layer Error Rate**:

$\delta^{(2)} =$ -0.16889796

- **Updated Hidden Layer Weights**: (*Fill in the missing values*)

$W^{(1)} =$

0.01400077	0.01700029	
0.005		
	0.0240550	
		0.05330902
	0.0530698	
0.03406384	0.05131964	

Perform another forward pass of the neural network, and calculate the new network output with the updated weights and biases:

$y = a^{(2)} =$ 0.8363552