Name: _____

SMU
SINGAPORE MANAGEMENT UNIVERSITY
MASTERS

Master of Science in Quantitative Finance

QF624
Machine Learning and Financial Application

INSTRUCTIONS TO STUDENTS

1 The time allowed for this examination paper is **2 hours**.

2 This examination paper contains **forty (40)** multiple-choice questions and **five (5)** short writing questions. You are **REQUIRED** to answer **ALL** questions. No marks will be deducted for wrong answers.

3 Please write the most suitable answer to each question in the exam booklet provided.

4 This is a closed-book examination. **You are allowed to use a scientific/financial calculator. You are allowed to bring ONE (1) double-sided A4-sized help sheet.**

5 There are ten (10) printed pages, including this instruction sheet.

6 You are required to return the full set of question papers at the end of the examination. Please write your name on the top right-hand corner of this instruction sheet.

Part I

This part comprises 40 MCQ questions, each with 1 mark. Please write down the most suitable answer to each question in the answer booklet provided.

1. How does the Mean Squared Error (MSE) change when multiplying all the variable values by a constant?
   A. The MSE remains the same.
   B. The MSE is multiplied by the square of the constant
   C. The MSE is divided by the square of the constant
   D. The MSE is divided by the constant

2. Which of the following is a common method to reduce overfitting?
   A. Increase the amount of training data
   B. Add more features to the model
   C. Increase the model's complexity
   D. Reduce the amount of training data

3. In Multiple Linear Regression, if the coefficient β_i of an input variable x_i is positive, what can be concluded?
   A. Target variable y and input variable x_i are positively correlated
   B. x_i is a significant variable, regardless of P-value
   C. R^2 has been reduced by incorporating x_i
   D. If all other input variables stay constant, we bump x_i up, we will see y going up

4. Which of the following input variables with a numerical value should be considered categorical, when we try to predict a property's value?
   A. Remaining lease (No. of years left)
   B. Floor area in square feet
   C. District code (Singapore has 28 districts, from 1 to 28)
   D. Ceiling height in meters

5. How does bagging mitigate the risk of overfitting?
   A. Reduce both bias and variance
   B. Reduce bias and increase variance a bit
   C. Increase bias a bit and reduce variance
   D. Increase both bias and variance

6. In Multiple Linear Regression, we have an interaction term '`a * b`' in the OLS formula '`smf.ols(y ~ a * b)`'. `a` is a categorical variable with two values 'premium' and 'normal', while `b` is a numerical variable. Which of the following is **FALSE**?
   A. The two linear fits for subgroup 'premium' and subgroup 'normal' will have the same gradient
   B. '`a * b`' becomes three terms '`a + b + a * b`' eventually with three coefficients
   C. In different subgroups indicated by `a`'s value, the relationship between `y` and `b` might be different

D. `a` becomes one dummy variable that maps the two values 'premium' and 'normal' to 0 and 1

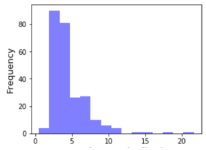
7. How will test data set's R^2 move, as the model complexity increases, from the simplest model that underfits the training set, to the most complex model that overfits the training set?

A. It goes down and then goes up
B. It goes up and then goes down
C. It always goes down
D. It always goes up

8. In which region does overfitting typically occur in the double descent curve?

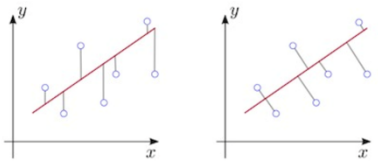
A. The first descent.
B. The second descent
C. The peak between the two descents
D. Overfitting does not occur in the double descent curve

9. For Linear Regression, if this is the distribution of the target variable, what is the first thing we shall try?



A. Take a log transformation of the target variable
B. Take square of the target variable
C. Take cube of the target variable
D. Add an interaction term among the input variables

10. Which plot below describes the residuals of linear regression?



A. Left
B. Right
C. Neither
D. Both

11. A bank wants to ramp up its operation in a country where the official language is Dothraki. The bank is aware that different languages and cultures lead to different financial needs. Unfortunately, only 3% of the existing data set gathered from daily business operations are about Dothraki-speaking people. What should this company **NOT** do? (Note that the bank's models shall serve all kinds of customers well.)

A. Oversample Dothraki-speaking people for the training set
B. Invite more Dothraki-speaking people to try financial products and give feedback
C. Run a survey in that country to understand Dothraki-speaking people's needs
D. Undersample the non-Dothraki-speaking people for the training set

12. There is an input variable 'month' (with 12 values) in the data set. How many dummy variables should be created for Linear Regression?

A. 0 B. 4
C. 11 D. 12

13. Which of the following is **FALSE** about training/validation/test approach?

A. The validation set is never used for training the model
B. The validation set can be used multiple times
C. Multiple models with different parameter values show different performances on the validation set. Based on this performance, we pick the best parameter combination
D. After we run the learned model on the test set, if the performance is bad, we shall retrain the model until the performance on the test set is good

14. What does a higher entropy of a node in a decision tree indicate?

A. The node is very homogeneous
B. The node is very heterogeneous or impure
C. The node contains very few observations
D. The node contains very many observations

15. Based on the confusion matrix below, which of the following is **FALSE**?

		Prediction	
		$\hat{y}$=0	$\hat{y}$=1
Actual	y=0	8	2
	y=1	4	6

A. Accuracy is 70%
B. Recall is 60%
C. Specificity is 80%
D. Precision is 60%

16. Why is cross-entropy loss preferred over mean squared error (MSE) for classification?

A. Cross-entropy penalizes false classifications more heavily
B. Cross-entropy is less computationally expensive
C. Cross-entropy results in larger gradient updates during training
D. Cross-entropy is less sensitive to outliers

17. Which of the following is **TRUE** about Logistic Regression?

A. It can only predict categorical target variable with two categories
B. The mathematical output of Logistic Regression formula can be $(-\infty, \infty)$
C. The target is assumed to follow a Bernoulli distribution
D. For binary classification, the target variable y follows a normal distribution

18. For Logistic Regression, which is **FALSE** about the odds $\frac{p}{1-p}$? (p is a probability)

A. p ranges from 0 to 1

B. $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. The right side is a linear combination of all input variables
C. If p_1 is the probability of an event for x_i = 1, p_0 is the probability of an event for x_i = 0, then $\frac{p_1}{1-p_1} / \frac{p_0}{1-p_0}$ is the odds ratio of x_i
D. The odds ratio of x_i is equivalent to the coefficient β_i of input variable x_i

19. Which of the following is **FALSE** about imbalanced data?
A. Predicting all observations to have the majority *y* label might lead to high accuracy
B. By reducing the dimension of input variables, we can reduce the majority *y* label and improve the balance
C. One could undersample the majority *y* label at the risk of losing information
D. One could oversample the minority *y* label at the risk of overfitting

20. Which of the following is **FALSE** about ROC and AUC?
A. No model could achieve an AUC smaller than 0.5
B. The AUC of a random guess model is 0.5
C. ROC is obtained by changing the classification probability threshold from 0 to 1
D. A perfect classification model has an AUC score of 1

21. What is the Gini impurity index for the below node in a decision tree? 'y=0: 3' means there are 3 observations with y being 0.

y=0: 3 | y=1: 3 | y=2: 4

A. $[(0.3)^2 + (0.3)^2 + (0.4)^2]$
B. $1 - [(0.3)^2 + (0.3)^2 + (0.4)^2]$
C. 6/10
D. $-(0.3 \log_2 0.3) - (0.3 \log_2 0.3) - (0.4 \log_2 0.4)$

22. Which of the following is **NOT** an advantage of Decision Tree?
A. No standardization required
B. No need to think about nonlinear and interaction terms in regression
C. The accuracy rate is outstanding compared to other models
D. Easy to be interpreted into rule sets

23. A data set has one input variable 'gender' with two values: male and female. Gloria creates the following two dummy variables. Which of the following is **FALSE**?

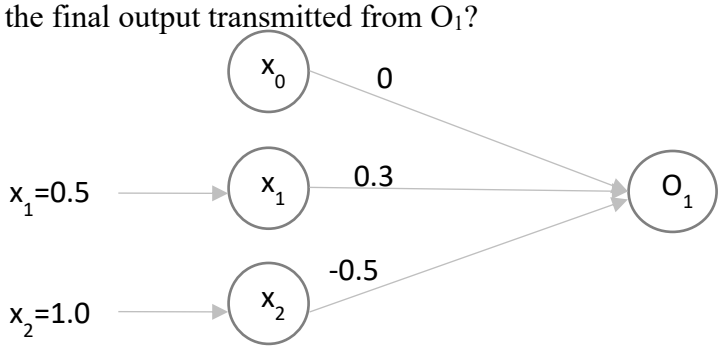
$$\text{is\_male} = \begin{cases} 1 \text{ if gender=male} \\ 0 \text{ if gender=female} \end{cases} \qquad \text{is\_female} = \begin{cases} 1 \text{ if gender=female} \\ 0 \text{ if gender=male} \end{cases}$$

A. is_male = 1 − is_female
B. Including both dummy variables in Linear Regression introduces perfect collinearity
C. is_male and is_female should both be used in a linear regression

D. is_male^2 =is_male

24. Which of the following is **FALSE** about Lasso and Ridge regularization?
   A. Ridge regularization reduces coefficients of all input variables close to zero but never exactly zero
   B. Standardization could be done before Linear Regression with Ridge regularization, to remove the final solution's dependency on the units of input variables (e.g., meter vs. foot)
   C. LASSO could generate sparse solutions with some coefficients becoming exactly zero
   D. Ridge regularization does not restrict the intercept β_0 while LASSO restricts the intercept β_0

25. We have a one-layer ANN below with initial weights randomly selected and the first observation being fed into the network. Assume x_0 is set as 1 and the activation function is sigmoid. What is the final output transmitted from O_1?



   A. $0 \times 1 + 0.3 \times 0.5 - 0.5 \times 1$
   B. $\frac{1}{1+e^{-(0\times1+0.3\times0.5-0.5\times1)}}$
   C. $\frac{1}{1+e^{(0\times1+0.3\times0.5-0.5\times1)}}$
   D. $\max(0 \times 1 + 0.3 \times 0.5 - 0.5 \times 1,\ 0)$

26. Which of the following is **FALSE** about backpropagation in ANN?
   A. It starts by computing the prediction error at output nodes
   B. It uses the prediction error at output nodes to update the weights between the output layer and the last hidden layer
   C. A big learning rate η could result in slow convergence to find the minimum loss function
   D. Each observation will be fed into ANN and complete one round of forward propagation and backpropagation

27. Which of the following is **TRUE** about bootstrap in bagging (**b**ootstrap-**agg**regat**ing**)?
   A. It samples data without replacement
   B. Bootstrap will not sample the first observation more than twice

C. It is usually recommended to bootstrap twice because people only have two boots to bootstrap
D. The observations in bootstrap selection from a data set could be all unique, i.e., no duplicate observations

28. Which of the following is **FALSE** about CART (**C**lassification **A**nd **R**egression **T**ree)?
A. It picks the best input variable with the best split point after trying all input variables and all binary split points
B. The best input variable with the best split point is chosen to maximize reduction of impurity
C. It tends to choose input variables with fewer split points to keep the model simple
D. It can handle classification problems (target y is categorical) and regression problems (target y is numerical)

29. What is one way to mitigate overfitting in decision trees?
A. Building bigger trees
B. Pruning the tree
C. Adding more features to the model
D. Using a smaller training set

30. What is the primary difference between Bagging and Random Forest algorithms?
A. Bagging is a sequential process, while Random Forest is parallel
B. Bagging uses a subset of features at each split while Random Forest uses all features
C. Random Forest uses a subset of features at each split while Bagging uses all features
D. Bagging can handle missing values, but Random Forest cannot

31. Which of the following is **FALSE** about Support Vector Machine?
A. SVM for binary classification is less effective when the data set is noisy, and the two groups based on y = -1 or 1 have overlapping areas
B. Support vectors are the data points that lie closest to the separating hyperplane
C. The goal is to find a hyperplane that separates the two groups based on y=-1 or 1 with maximum margin between the two boundaries where support vectors sit on
D. We can always find a perfect separating hyperplane to fulfil the requirement that no points can violate the two boundaries

32. What does a 'Gaussian Process' represent in Bayesian Optimization?
A. It represents the objective function to be optimized
B. It represents the search space of hyperparameters
C. It represents the prior belief over functions
D. It represents the range of possible outputs for a given input

33. Target (American supermarket) has been reported to guess a customer's pregnancy (before the customer finds out herself) and start to promote pregnancy related products to those customers. What can we learn from this business case?
A. False positive and False negative can have very different business impact
B. Misclassification might lead to pregnancy

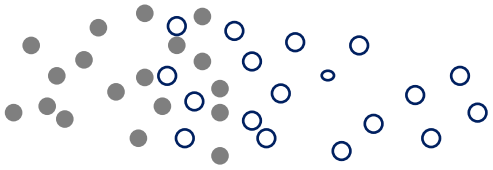
C. Maximizing accuracy rate shall always be the goal of prediction
D. We shall optimize for Precision in this business case

34. For an ordinal input variable 'Satisfaction' with values (bad, average, good), which encoding shall we do before using K-Means with Euclidean Distance?
A. One-hot encoding to create three dummy variables
B. One-hot encoding to create two dummy variables
C. Integer encoding to map 'bad' to 0, 'average' to 1, and 'good' to 2
D. No encoding is needed for ordinal input variables for standard K-Means

35. Which of the following is **FALSE** about Recurrent Neural Networks?
A. Nodes have connections pointing back at themselves
B. It is typically for processing sequential data with the dimension of time
C. Errors are accumulated for each timestep, contributing to the overall loss
D. Long short-term memory can only learn about short-term dependencies

36. Which of the below action describes Clustering?
A. Group customers that are similar to each other based on differences in attributes (age, salary, etc) and ensure that each group of customers is significantly different
B. Separating customers based on gender and marital status into 4 groups, Male/Married, Female/Married, Male/Not Married, Female/Not Married
C. Applying business knowledge of Western vs. Asian differences in banking preference and showing different bank products on online banking app based on the source country of the incoming browsing request
D. None of the above

37. Which linkage is good to detect two clusters with some points from each cluster that are very close in the distance and keep the two clusters separate? Refer to the below graph.



A. Single linkage
B. Complete linkage
C. Average linkage
D. Ward's linkage

38. Which of the following is **TRUE** about Hierarchical vs. Partitioning Clustering?
A. Agglomerative clustering is a form of Hierarchical Clustering and uses merging
B. Partitioning Clustering splits the data points into partitions, and eventually merges into one cluster
C. K-Means is a form of Partitioning Clustering, and after the algorithm starts, no. of clusters is indeterministic depending on how the calculation goes
D. Hierarchical Clustering generates a hierarchy called dendrogram, which can visually show the number of clusters in the final clustering solution

39. Which metric can measure the quality of clustering solutions?

A. Average Silhouette score
B. Mean Squared Error
C. Root Mean Squared Error
D. Accuracy rate

40. What does an 'action' represent in the context of reinforcement learning applied to portfolio management?

A. The action represents the choice of stock to buy or sell
B. The action represents the prediction of future stock prices
C. The action represents the communication between the agent and the broker
D. The action represents the calculation of the optimal portfolio weights

Part II

Please write the answers to writing questions in the exam booklet provided.

Question 1 (5 marks)

What is the difference between Supervised Learning and Unsupervised Learning? List one example of Supervised Learning and another of Unsupervised Learning in finance.

Question 2 (10 marks)

We split a data set into training, validation, and test sets and use a validation set to fine-tune the parameters, e.g., regularization's penalty weight. When do we use k-fold cross-validation? Can we use k-fold cross-validation in time series data? If so, state the procedure; if not, state the reason.

Question 3 (10 marks)

When training a supervised machine learning model, proper data processing is often needed. Please state the common steps involved and the order of these steps. Give at least one example (detailed procedure of implementation) for each step in your answer.

Question 4 (15 marks)

Consider a simple neural network used for regression: an input layer, one hidden layer with two neurons using the ReLU activation function, and a single output neuron with no activation function. The network employs Stochastic Gradient Descent (SGD) with a learning rate $\alpha = 0.01$ for weight optimization. The network is trained using a Mean Squared Error (MSE) loss function. Given a minibatch containing a single data point with input $x = [x_1, x_2] = [1, -2]$ and corresponding target output $y = 5$, the initial weights from the input to the hidden layer are $w_{11} = 0.5, w_{12} = -1, w_{21} = 1, w_{22} = 0.5$, and from the hidden layer to the output neuron are $v_1 = 2, v_2 = -3$. Assume biases are zero for simplicity.

Derive the mathematical steps to update the weights w_{ij} and v_i using SGD for this minibatch. Your derivation should include:

1. (**5 marks**) Calculating the output of the network for the given data point.
2. (**5 marks**) Computing the gradient of the MSE loss with respect to each weight.

3. (**5 marks**) Updating the weights based on the computed gradients and the learning rate.

Question 5 (20 marks)

Consider a financial institution that wants to predict whether a customer will default on a loan. The institution has historical data on age, income, loan amount, repayment history, and whether or not the customer defaulted. The institution is considering three machine learning models for this task: linear regression, logistic regression, and decision trees.

1. (**5 marks**) Based on the nature of the problem and the type of data available, which of these three models would you consider most appropriate and why? Can any of these models be completely ruled out for this task?
2. (**5 marks**) How does the requirement for feature scaling differ among these three models? What could be the potential implications if the features are not properly scaled?
3. (**5 marks**) Discuss the role of regularization in linear and logistic regression models. How would it affect the performance of these models? Are there equivalent methods for controlling model complexity in decision trees?
4. (**5 marks**) How would you determine the importance of each feature in making predictions in each of these models? Are there any potential challenges or pitfalls in interpreting feature importance, particularly in the context of logistic regression and decision trees?

Part I

1. Ans: B

Explanation: Recall that the MSE is defined as the average of the squares of the errors, i.e., the average squared difference between the estimated values and the true value. As such, multiplying all the variable values (estimated and true values) by a constant will result in the original MSE being multiplied by the square of that constant.

2. Ans: A

Explanation: Overfitting occurs when a model learns the training data too well and performs poorly on new data. Having more training data allows the model to learn more general patterns and avoid focusing on the specific details of the existing data.

Why other options are incorrect

B: Adding more features can potentially increase the model's complexity, leading to overfitting. A simpler model with fewer features might generalize better.
C: This directly contributes to overfitting. A more complex model has a greater capacity to memorize the training data, resulting in poor generalization.
D: This can lead to underfitting, where the model does not learn the underlying patterns well and performs poorly on both training and new data.

3. Ans: D

Explanation: If the coefficient β_i of an input variable x_i is positive in a MLR model, it indicates that as the value of x_i increases, then the predicted value of the target variable *y* also tends to increase, assuming all other input variables remain constant.

Why other options are incorrect

A: While a positive coefficient suggests a positive association, it doesn't necessarily mean a strong correlation. Multicollinearity (correlation between independent variables) can influence the coefficients.
B: The *p*-value associated with the coefficient β_i determines its statistical significance. A positive coefficient alone doesn't guarantee the variable is statistically significant.
C: Adding a variable with a positive coefficient might increase the R^2 value, especially if it explains a significant portion of the variance in the target variable. R^2 measures the proportion of variance in the dependent variable that is explained by the model.

4. Ans: C

Explanation: District code is a label for one of 28 categories; the numbers carry no quantitative meaning, so it must be treated as categorical.

Why other options are incorrect

A: Remaining lease is a continuous numeric variable (years left), so numeric.
B: Floor area is a continuous measurement in square feet, so numeric.
D: Ceiling height is a continuous measurement in meters, so numeric.

5. Ans: C

Explanation: Bagging, or bootstrap aggregating, is a machine learning technique that helps reduce overfitting by creating multiple models, each trained on a random subset

of the training data and then averaging their predictions. This process primarily reduces the variance of the model, meaning it becomes less sensitive to the specific data it was trained on, thereby improving its ability to generalize to new data. While there might be a slight increase in bias, the overall reduction in variance outweighs this, leading to a more robust and less overfitted model.

Why other options are incorrect
A: Although bagging can improve both bias and variance to some extent, it is primarily known for its variance reduction impact. Sources like and discuss how bagging mainly reduces variance while maintaining similar bias levels. mentions that bagging can slightly increase bias in some cases, but this is not its primary effect. Boosting, another ensemble technique, is more focused on reducing bias.
B: This is the opposite of what bagging does. Bagging aims to reduce variance by averaging multiple models, not increase it. Increasing variance would make the model more sensitive to the specific training data and lead to overfitting (opposite effect).
D: Bagging primarily reduces variance while maintaining or slightly increasing bias. Increasing both bias and variance would worsen model performance, not improve it, which contradicts the purpose of bagging.

6. Ans: A
Explanation: When an interaction term $a \cdot b$ is included in an OLS formula, the categorical variable a creates separate intercepts and gradients (slopes) for each category of *a*, 'premium' and 'normal' in this case. If the gradients were the same, it would imply that the relationship between *y* and *b* is identical across both subgroups, which contradicts the purpose of the interaction term.

Why other options are incorrect
B: The interaction term $a \cdot b$ in the formula effectively expands into three separate terms: *a*, *b*, and $a \cdot b$. Each term has its own coefficient in the regression model.
C: The inclusion of the interaction term $a \cdot b$ means that the effect of *b* on *y* can be different depending on whether a is 'premium' or 'normal'.
D: This is how categorical variables with two levels are typically handled in LR models. One level is coded as 0 and the other as 1, creating a dummy variable. This allows the model to distinguish between the two groups in terms of their intercepts.

7. Ans: B
Explanation: As model complexity increases from severe underfitting, test R^2 first improves (better fit) but once you pass the optimal complexity and begin overfitting, test R^2 deteriorates.

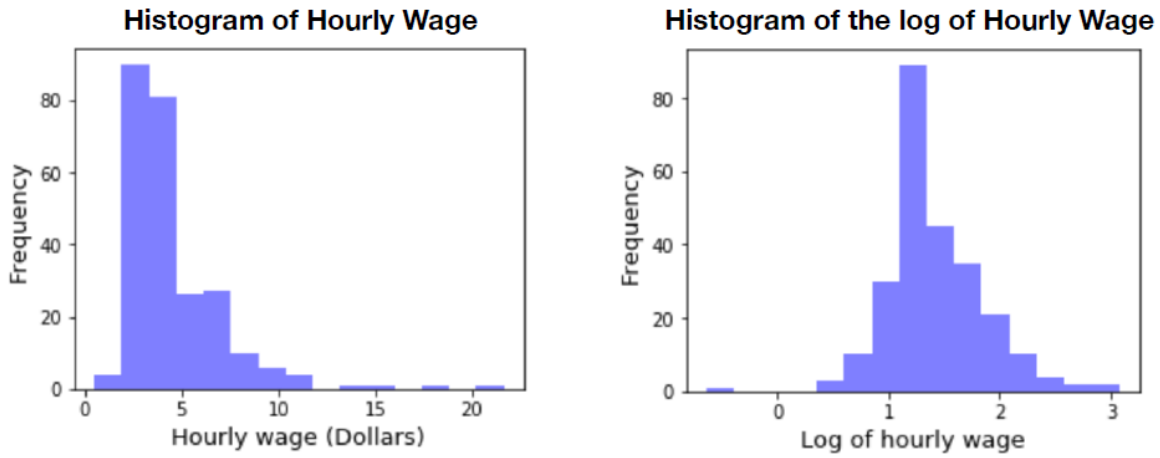
Why other options are incorrect
A (“down then up”): is the reverse of the usual underfit–overfit curve.
C (“always down”) & D (“always up”): ignore the turning point from under- to overfitting.

8. Ans: C
Explanation: When a model initially increases in complexity, its performance improves (first descent) due to reduced underfitting. At the point where the model has enough

capacity to perfectly fit the training data (interpolation threshold), the test error often spikes (peak), indicating overfitting. Beyond the peak, further increasing complexity can surprisingly lead to improved generalization (second descent), as the model starts to learn more meaningful patterns rather than just memorizing the training data.

Why other options are incorrect
A: Overfitting is associated with high model complexity and usually occurs when the model can memorize training data (happens more readily in later stages of 1st descent)
B: The second descent is characterized by improved generalization as the model becomes more robust. While the model may still be complex, it is not overfitting in the same way as at the peak.
D: Overfitting is a key aspect of the double descent phenomenon. The peak in the test error curve during the second descent is a clear indication of overfitting.

9. The two histograms are shown below:



Ans: A
Explanation: When the target variable has a right-skewed distribution (as indicated by the histogram of hourly wage), it's common to transform it using a logarithm to make its distribution more symmetrical. A symmetrical distribution, like the one achieved by taking the log of the target variable, is more suitable for linear regression models.

Why other options are incorrect
B: While this might reduce skewness, it's less likely to result in a symmetrical distribution compared to the log transformation.
C: Like the square, the cube transformation might reduce skewness but is less predictable than a log transformation.
D: Interaction terms are used to model relationships between input variables, not to address skewness in the target variable.

In essence: A log transformation is a well-established and effective method for dealing with right-skewed data, particularly when working with linear regression models.

10. Ans: B

Explanation: A residual plot in linear regression depicts the difference between actual observed values and predicted values (residuals) plotted against the predicted values. If you calculate the residual by taking the vertical difference between the actual *y*-value and the predicted *y*-value along the regression line, the resulting points might not accurately represent the true deviation from the line, especially for outliers. On the other hand, by taking the perpendicular distance from the data point to the regression line, you are measuring the true deviation from the predicted value, which is the standard method for calculating residuals.

Why other options are incorrect
A: While the left plot might show outliers, it doesn't specify how the residuals were calculated. If the residuals were computed using vertical differences, the plot wouldn't accurately represent the true deviations from the line.
C: Both options describe a residual plot with outliers, but only the right option correctly describes how residuals are computed in a standard linear regression.
D: Only the right plot accurately describes the standard method of computing residuals. The left option describes a potential issue with how the residuals were calculated, not the correct representation of a residual plot.

11. Ans: D
Explanation: Under-sampling throws away valuable majority data and risks biasing the overall model.

Why other options are incorrect
A: Oversampling can help the model see enough minority examples.
B: Invitation gathers real-world data and feedback.
C: Running a survey uncovers specific needs of that group.

12. Ans: C
Explanation: Recall that the dummy variable trap in linear regression is a situation where including all possible dummy variables for a categorical feature leads to multicollinearity, making it difficult to estimate unique coefficients for each category. This occurs because the dummy variables become perfectly correlated with each other or with the constant term, rendering the model unstable. To avoid this, one dummy variable (usually the baseline category) should be dropped, effectively creating a reference point for comparison. Here, one dummy variable is removed from the 12, each of which corresponds to a different month of the year.

13. Ans: C
Explanation: While selecting the best performing model on the validation set is a common practice during hyperparameter tuning, it's crucial to understand that this selection is part of the training process and not the final evaluation. A separate, unseen test set is required to assess the generalization performance of the chosen model. If the test set is used for model selection, it introduces bias, and the results won't accurately reflect the model's real-world performance.

The validation set is used during the training phase to tune hyperparameters. After each training iteration or after trying different parameter combinations, the model's performance on the validation set is evaluated. The hyperparameters that yield the best performance on the validation set are chosen for the final model. The test set, on the other hand, is held out from the training and validation process. It's used to get an unbiased estimate of how well the final, tuned model generalizes to new, unseen data. If you use the validation set to select the best model and then report its performance, you are essentially overfitting to the validation set. The performance on the validation set will be overly optimistic, and the test set will reveal a lower performance. A typical workflow involves dividing the data into three sets: training, validation, and test. The model is trained on the training set, its hyperparameters are tuned using the validation set, and finally, the performance of the best model is evaluated on the test set to get an unbiased estimate. In some cases, to get a more robust estimate of model performance, cross-validation techniques are used. This involves splitting the training data into multiple folds, training the model on some folds and validating on others, and repeating this process to get an average performance. Even with cross-validation, a separate test set is still needed for final evaluation.

14. Ans: B
Explanation: Here, entropy measures the level of impurity or randomness in a node. A higher entropy means the data within the node is more mixed with different classes, making it less homogeneous. When a node has a high entropy, it means the decision tree needs to further split the data at that node to achieve greater homogeneity.

Why other options are incorrect
A: A very homogeneous node has a low entropy, not a high one. A homogeneous node means most or all data points in the node belong to the same class.
C. The number of observations in a node does not directly determine its entropy. Entropy is a measure of the distribution of classes within the node, regardless of the number of observations.
D. Like option C, the number of observations does not directly impact entropy. A large node with a mixed distribution of classes will have high entropy, while a small, homogeneous node will have low entropy.

15. Ans: C
Explanation: The confusion matrix provided can be interpreted as shown below:

	Predicted $\hat{y} = 0$	Predicted $\hat{y} = 1$
Actual $y = 0$	TN = 8	FP = 2
Actual $y = 1$	FN = 4	TP = 6

Recall the following formulae:
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Recall = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Precision = TP / (TP + FP)

Using these formulae, you can check and verify that all other options are incorrect.

16. Ans: A
Explanation: Cross-entropy is designed to work with probability distributions, penalizing large discrepancies between predicted probabilities and actual labels, especially when

a model makes a confident but incorrect prediction. This makes it suitable for classification tasks where the goal is to predict the correct class with high probability.

Why other options are incorrect
B: While the computations involved in calculating cross-entropy might be slightly different from MSE, the difference is usually negligible and not a significant factor in choosing between them.
C: While cross-entropy can lead to steeper gradients in certain situations, this is not the primary reason for its preference over MSE.
D: This is the opposite. Because cross-entropy is based on probabilities, large errors (representing confident wrong predictions) contribute significantly to the overall loss, making the model more responsive to outliers. MSE, on the other hand, is less sensitive to outliers due to its squared error calculation.

17. Ans: C
Explanation: In binary logistic regression, the target variable represents a binary outcome (e.g., success/failure, 0/1). The Bernoulli distribution is used to model the probability of such binary outcomes.

Why other options are incorrect
A: Logistic regression primarily handles binary classification. It can be extended to handle multiple categories (multinomial or ordinal logistic regression).
B: The logistic function (sigmoid function) maps the linear combination of predictors to a probability between 0 and 1, i.e., the output of the logistic regression formula (the predicted probability) is restricted to the range [0, 1]. The log-odds (logit transformation) can range from (–∞, ∞), but not the direct output probability.
D: For binary classification, the target variable in logistic regression is categorical, not continuous (e.g., target variables in the normal distribution are not categorical)

18. Ans: D
Explanation: The coefficient β_i represents the change in the log-odds (logit) for a one-unit increase in x_i, while the odds ratio is the exponentiated value exp(β_i), representing the multiplicative change in the odds for a one-unit increase in x_i. In logistic regression, the predicted value is not a direct probability but rather the log-odds of the outcome.

The logit (or log-odds) transformation is used to model the relationship between the independent variables and the probability of an event. The logit is the natural logarithm of the odds, where the odds are the ratio of the probability of the event occurring to the probability of it not occurring. The coefficients β_i in a logistic regression model represent the change in the logit for a one-unit change in the corresponding independent variable x_i, holding all other variables constant. To interpret the effect of a predictor variable in terms of odds, we exponentiate the coefficient. This gives us the odds ratio. For example, if $\beta_1 = 0.5$, then the odds ratio is exp(0.5) ≈ 1.65, which means that a one-unit increase in x_i multiplies the odds of the outcome by approximately 1.65.

Example: If you have a logistic regression model with the following equation: logit(p) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, where p is the probability of the event, x_1 and x_2 are the independent

variables, and β_0, β_1, and β_2 are the coefficients. In this case, β_1 represents the change in the logit for a one-unit change in x_1, and $\exp(\beta_1)$ is the corresponding odds ratio.

In essence: While the coefficient β_i in logistic regression is related to the odds ratio, they are not equivalent. The coefficient represents the change in the logit, and the odds ratio is the exponentiated form of the coefficient, representing the change in the odds.

19. Ans: B
Explanation: Dimensionality reduction (DR) of input variables does not inherently affect the class distribution. While certain DR techniques might lead to better separation between classes, the ratio of majority to minority classes remains unchanged unless specifically addressed through other methods like oversampling or under-sampling.

Why other options are incorrect
A: In an imbalanced dataset, a model can achieve high accuracy by simply predicting the majority class for all instances, even if it completely ignores the minority class. This is because accuracy is calculated as the percentage of correct predictions, and a high number of majority class predictions can skew the results.
C: Under-sampling involves removing data points from the majority class to balance the dataset. While this can lead to a more balanced dataset, it also risks losing valuable information from the majority class.
D: Oversampling involves creating copies of data points from the minority class to increase its representation. This can help improve the model's performance on the minority class but might introduce the risk of overfitting, where the model becomes too specialized to the training data and performs poorly on new data.

20. Ans: A
Explanation: An AUC value can be less than 0.5, indicating that the model performs worse than a random guess.

Why the other options are incorrect
B: A random classifier will have an AUC of 0.5 because it is essentially flipping a coin to decide the class of each data point.
C: The ROC curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification probability thresholds.
D: An AUC score of 1 represents perfect classification where the model correctly classifies all positive and negative instances.

21. Ans: B
Explanation: Since the total number of observations is 3 + 3 + 4 = 10, the probability that y takes 0, 1, 2 is correspondingly 0.3, 0.3, 0.4. You can then use the Gini impurity formula $1 - \Sigma^{C}_{i=1} p_i^2$, where p_i is the probability of observing class i and C is the total number of classes, to obtain the desired Gini impurity.

22. Ans: B
Explanation: While decision trees can implicitly capture these relationships, the lack of explicit modeling can lead to overfitting and poor generalization on unseen data.

Decision trees can create very complex structures that perfectly fit the training data but fail to generalize to new data (overfit). This is especially true when dealing with non-linear relationships and interactions, as the tree might learn specific patterns in the training data that don't hold true more broadly. While simple decision trees are easy to interpret, complex ones with many branches can become difficult to understand and explain (lack of interpretability), negating one of their key advantages. Regression models, like linear regression, allow for explicit modeling of non-linear effects using techniques like polynomial features, interactions terms, or non-linear basis functions. This gives more control over the complexity of the model and allows for a better understanding of the underlying relationships. While linear regression struggles with non-linearities, ensemble methods like Random Forests, which are based on decision trees, can effectively capture non-linearities and interactions by combining multiple trees. However, this approach often sacrifices some interpretability. Regression models may require more careful data preparation (e.g., feature engineering) to capture non-linearities, but this can lead to a more robust and generalizable model than relying solely on the implicit learning of decision trees. Decision trees, like CART, can handle non-linear relationships and interactions, but they also have drawbacks like overfitting and difficulty accounting for additive effects. While decision trees are easy to interpret, complex trees can be hard to understand, and in many cases, the interpretation is specific to the training data. Regression models, while sometimes less intuitive, can offer a more structured way to understand the relationships between variables.

23. Ans: C
Explanation: When using dummy variables for a binary categorical variable like gender, only one dummy variable is needed to represent the categories. Including both would lead to perfect multicollinearity.

Dummy variables are used to represent categorical data in a way that can be used in statistical models. For a binary variable (like male/female), you create one dummy variable (e.g., is_male), where a value of 1 indicates the category (male) and 0 indicates the other category (female). Perfect multicollinearity occurs when one independent variable can be perfectly predicted from another. In this case, if you know if someone is male (is_male = 1), you automatically know they are not female (is_female = 0), and vice versa. Therefore, is_female = 1 - is_male. This means is_female is completely redundant and including both is_male and is_female creates perfect multicollinearity, making the model unstable. As to why only one is needed, the coefficient for the single dummy variable (e.g., is_male) represents the difference in the outcome variable between males and females. If you add the second dummy variable, it essentially duplicates this information, leading to the problem described above.

24. Ans: A
Explanation: Ridge regression shrinks coefficients but doesn't reduce them to zero, whereas Lasso regression can do so.

25. Ans: B

Explanation: Recall that the activation function in the output layer transforms the weighted sum of inputs into the final output of the network, which can be a prediction, a classification label, or a probability distribution. For the case of the artificial neural network (ANN) provided, the activation function is sigmoid in nature.

26. Ans: C
Explanation: A large learning rate can lead to rapid oscillations and overshooting the optimal minimum, hindering convergence rather than slowing it down.

Why the other options are incorrect
A: Backpropagation begins by calculating the difference between the network's predicted output and the actual target output.
B: The error is propagated backward through the network, allowing the weights to be adjusted to minimize the overall error.
D: In training a neural network, each data point is typically processed through a forward pass to generate a prediction, followed by a backward pass to update the weights based on the prediction error.

27. Ans: D
Explanation: It's mathematically possible (though unlikely) that all draws happen to pick each observation exactly once.

Why other options are incorrect
A: Bagging uses sampling with replacement.
B: Any given observation can be sampled many times, without a strict upper bound of 2.
C: Joke

28. Ans: C
Explanation: CART aims to minimize impurity at each split, not necessarily to keep the model simple by limiting split points. While a simpler model might be desirable in some cases, the algorithm prioritizes finding the best split, regardless of the number of resulting split points.

Why other options are incorrect
A: This is the core mechanism of CART. At each node, it evaluates all possible splits on all features and selects the one that best improves the model's performance.
B: CART uses a metric like Gini impurity or entropy to measure the "impurity" of the data at each split. The goal is to choose the split that minimizes this impurity, effectively separating the data into more homogeneous groups.
D: CART is designed to handle both classification tasks (where the target variable is categorical) and regression tasks (where the target variable is numerical). It uses the same basic algorithm for both types of problems, adjusting the prediction method based on the type of target variable.

29. Ans: B
Explanation: Pruning a decision tree involves removing unnecessary branches or nodes from the tree, which helps to reduce its complexity and prevent overfitting by removing

parts of the tree that might be overly specific to the training data and not generalize well to new data.

Why other options are incorrect
A: Building larger trees can worsen overfitting as they become more complex and capture finer details of the training data, leading to poor performance on unseen data.
C: Adding more features can potentially increase the model's capacity to overfit the training data, as it has more info to learn from, not necessarily improving generalization.
D: While a smaller training set can lead to underfitting, not overfitting, using a smaller training set with a large model is still susceptible to overfitting. Pruning is a better approach to address overfitting in conjunction with using a suitable training set size.

30. Ans: C
Explanation: Bagging (bootstrap aggregating) involves training multiple decision trees on different random subsets of the training data, where each tree considers all the features available for splitting at each node. Random Forest builds upon bagging by adding an extra layer of randomness. Besides randomly sampling data for each tree, it also randomly selects a subset of features to consider for splitting at each node. This random feature selection helps to reduce the correlation between individual trees, improving the overall model's performance. This also explains why option B is incorrect.

Why other options are incorrect
A: Both bagging and Random Forest can be implemented in parallel. The key difference lies in how the trees are constructed, not their execution.
D: Both bagging and Random Forest can handle missing values. Various strategies, such as imputing missing values with the mean or median or ignoring missing values during the splitting process, can be employed.

31. Ans: D
Explanation: In non-linearly separable data you cannot find a perfect hyperplane without misclassifications (unless you use slack variables or kernels).

Why other options are incorrect
A: Noisy, overlapping classes make hard-margin SVM less effective.
B: Support vectors are the points closest to the decision boundary.
C: SVM maximizes the margin between classes.

32. Ans: C
Explanation: The 'Gaussian process' (GP) represents the initial, uncertain knowledge about the objective function before any observations are made. This prior belief is then updated with observed data to form a posterior distribution, which provides a more informed belief about the function.

In Bayesian optimization, the GP is initialized with a prior distribution over the possible functions that could be the objective function. This prior reflects the initial assumptions or beliefs about the function's behavior, such as smoothness or typical value ranges. The GP acts as a surrogate model, meaning it's a probabilistic approximation of the true,

unknown objective function. It provides predictions and uncertainty estimates for any input value, allowing the algorithm to intelligently choose the next point to evaluate. As Bayesian optimization evaluates the objective function at specific points, the GP's prior is updated using Bayes' theorem to form a posterior distribution. This posterior distribution reflects the updated knowledge about the objective function after considering the observed data. The posterior distribution from the GP is then used to define an acquisition function, which guides the selection of the next point to evaluate. This function aims to balance exploration (trying new, uncertain areas) and exploitation (improving already promising regions).

33. Ans: A
Explanation: In this case, a false positive (predicting pregnancy when the customer isn't) is more damaging than a false negative (failing to predict pregnancy; would mean missed revenue opportunities and potentially frustrated customers). A false positive might lead to awkward interactions, embarrassment for the customer, and a perception that Target is invading privacy. False negatives (missing a pregnancy), on the other hand, might mean lost sales, but the consequences aren't as severe as alienating customers.

Why other options are incorrect
B: Misclassification (predicting pregnancy when the customer isn't) doesn't cause pregnancy. This option misinterprets the situation.
C: While maximizing accuracy is generally a good goal, in this specific case, the impact of different types of errors is more important than the overall accuracy rate. For example, a high-accuracy model that produces mostly false positives could be worse than a less accurate model that minimizes false positives.
D: While optimizing for precision (reducing false positives) would be beneficial in this scenario, it's not the only takeaway. The key lesson: understanding the consequences of different types of errors. Optimizing precision is a strategy, not the primary takeaway.

In simpler terms: Target's case highlights the importance of considering the impact of errors in predictive models, not just the overall accuracy. False positives and false negatives can have different, sometimes drastic, consequences for businesses.

34. Ans: C
Explanation: Integer encoding preserves the ranking and gives meaningful distances for *k*-means with Euclidean distance.

Why other options are incorrect
A and B (one-hot): ignore the inherent ordering of “bad < average < good.”
D: you must convert strings to numbers before *k*-means.

35. Ans: D
Explanation: LSTM units are specifically designed to capture long-term dependencies.

Why other options are incorrect
A: RNNs have feedback loops.
B: They handle sequential/time-series data.

C: Errors accumulate through time steps via backpropagation through time.

36. Ans: A
Explanation: Clustering is a technique in data analysis that aims to identify groups of data points that are like each other within a group, but distinct from other groups.

Why other options are incorrect
B: Separating customers based on gender and marital status into 4 groups, Male/Married, Female/Married, Male/Not Married, Female/Not Married: This describes a form of segmentation based on demographic factors, not clustering. While segmentation can involve grouping customers, it doesn't necessarily use the clustering algorithm to achieve this.
C: Applying business knowledge of Western vs. Asian differences in banking preference and showing different bank products on online banking app based on the source country of the incoming browsing request: This describes a form of targeted marketing based on cultural differences, not clustering. While clustering can be used to inform marketing strategies, this specific action is more about direct application of cultural insights rather than using a clustering algorithm to group customers.

37. Ans: B
Explanation: This method calculates the distance between clusters by considering the maximum pairwise distance between any points in the two clusters. This means that even if a few points from each cluster are very close, the clusters will only be merged when all pairs of points have a large distance, effectively preventing premature merging of clusters with close outliers.

Why other options are incorrect
A: This uses the minimum pairwise distance between points in the clusters, which can lead to clusters being merged prematurely even if a few outlier points are very close.
C: This method calculates the average distance between all pairs of points from the two clusters. While it is more robust than single linkage, it can still be susceptible to outliers if they significantly influence the average distance.
D: This method minimizes the variance within the combined clusters. It is not specifically designed to handle outliers and may not be ideal for clusters with close points from different clusters.

38. Ans: D
Explanation: This method builds a hierarchy of clusters by merging or splitting them, visually represented as a dendrogram. The dendrogram clearly shows how clusters are formed and at which level of the hierarchy they are located, allowing for visual determination of the optimal number of clusters.

Why other options are incorrect
A: This statement is true, but it doesn't explain the key characteristic of hierarchical clustering regarding the dendrogram representation.

B: This is the opposite of how partitioning clustering works. It starts with a pre-defined number of clusters and iteratively assigns data points to them. It doesn't involve merging partitions into one cluster.
C: While *k*-means is a partitioning clustering algorithm, the number of clusters is not inherently indeterministic. The user specifies the desired number of clusters (*k*) at the start of the algorithm. The outcome depends on the initial cluster assignments and the optimization process, but the final number remains fixed by the initial parameter.

39. Ans: A
Explanation: This score evaluates how well each data point is clustered with similar points and separated from other clusters, providing a measure of clustering quality.

Why other options are incorrect
B: MSE is typically used in regression to measure the difference between predicted and actual values. It's not designed for evaluating clustering quality, as clustering is an unsupervised learning task where there are no "true" labels to compare against.
C: RMSE is the square root of MSE. While it's used in regression, like MSE, it's not suitable for evaluating clustering due to the lack of ground truth labels in clustering.
D. Accuracy rate is primarily used in supervised learning tasks to measure the percentage of correctly classified instances. It's not applicable to clustering as clustering is an unsupervised task where there are no predefined classes.

40. Ans: D
Explanation: In portfolio management RL, an action is the allocation vector (optimal weights) the agent chooses.

Why other options are incorrect
A ("choice of stock to buy or sell"): is more akin to a single trading action but doesn't capture the multi-asset portfolio context.
B (predicting prices): isn't an "action" but a model output or intermediate step.
C (communication): isn't an RL concept.

Part II
Question 1 (5 marks)
(a) Supervised learning

- Learns a mapping from input *X* to a known target *y*, using labeled examples.
- Objective is to minimize prediction error on *y* (e.g. classification or regression).
- Example in finance (credit-default prediction): train a classifier on borrower features (income, balance, repayment history) to predict default vs. no-default.

Unsupervised learning

- Finds structure or patterns in *X* alone, without any labels.
- Objectives include clustering, density estimation, dimensionality reduction.
- Example in finance (equity-style clustering): group stocks into clusters (e.g. "value" vs. "growth") based on fundamental ratios using *k*-means.

Question 2 (10 marks)

Use cases for *k*-fold CV

- When data are i.i.d. and limited in size, to get a more robust estimate of out-of-sample performance than a single train/validation split.
- Helps detect overfitting and variance of the model by averaging metrics over *k* different train/validation splits.

Time-series data?

- Standard random *k*-fold CV is not appropriate for time series, because it would violate temporal order and leak future information into training.
- Instead apply rolling (walk-forward) cross-validation:
  1. Sort observations by time.
  2. Choose an expanding or sliding window scheme:
     - Expanding window
       - Fold 1: train on time 1, ..., T_1, validate on $T_1 + 1$, ..., T_2
       - Fold 2: train on time 1, ..., T_2, validate on $T_2 + 1$, ..., T_3
       - ... and so on.
     - Sliding window: Fix window size W; for each fold, train on $[t, t + W - 1]$, validate on $[t + W, t + W + H - 1]$, then roll forward by H.
  3. Average the validation scores across folds.
- This preserves causality and prevents peeking into "future" data.

<u>Question 3 (10 marks)</u>

1. Data cleaning

- Purpose: remove or correct errors, handle missing values, detect outliers.
- Example
  - Use pandas to identify NaNs

```python
df.isna().sum()
```

  - Impute a numeric column ("loan_amount") with the median

```python
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='median')
df['loan_amount'] = imputer.fit_transform(df[['loan_amount']])
```

2. Data integration (if multiple sources)

- Purpose: merge datasets on keys, reconcile schema differences.
- Example (merge transaction history with customer profiles on "customer_id")

```python
df = pd.merge(transactions, profiles, on='customer_id', how='left')
```

3. Feature encoding

- Purpose: convert categorical or ordinal variables into numeric form.
- Example (one-hot encode a “property_type” column)

```python
df = pd.get_dummies(df, columns=['property_type'], drop_first=True)
```

4. Feature scaling/normalization

- Purpose: put numeric features on comparable scale for distance-based models & gradient-based optimization.
- Example (standardize “floor_area” and “ceiling_height”)

```python
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df[['floor_area','ceiling_height']] =
scaler.fit_transform(df[['floor_area','ceiling_height']])
```

5. Feature engineering

- Purpose: create new predictors or transform existing to capture domain knowledge.
- Example (create a rolling 30-day volatility of daily returns)

```python
df['volatility_30d'] = df['daily_return'].rolling(window=30).std()
```

6. Feature selection / Dimensionality reduction

- Purpose: remove redundant or irrelevant features to reduce overfitting and improve interpretability.
- Example
    - Use variance threshold to drop near-constant features

```python
from sklearn.feature_selection import VarianceThreshold
sel = VarianceThreshold(threshold=0.01)
df_reduced = sel.fit_transform(df)
```

    - Or apply PCA to compress correlated predictors

```python
from sklearn.decomposition import PCA
pca = PCA(n_components=5)
df_pca = pca.fit_transform(df_scaled)
```

7. Data splitting

- Purpose: partition into training, validation, and test sets to tune hyperparameters and evaluate final performance without leakage.
- Example (stratified random split for a classification target “default”)

```python
from sklearn.model_selection import train_test_split
X_temp, X_test, y_temp, y_test = train_test_split(X, y, test_size=0.2, stratify=y)
X_train, X_val, y_train, y_val = train_test_split(X_temp, y_temp, test_size=0.25, stratify=y_temp)
# results: 60% train, 20% val, 20% test
```

Note: in practice you’d wrap all of the above in a Pipeline (e.g. sklearn.pipeline.Pipeline) so that each transformation is safely applied to train, then identically to validation/test without leakage.

Question 4 (15 marks)

Network architecture

- Input $\mathbf{x} = [x_1, x_2] = [1, -2]$
- Hidden layer: 2 neurons, ReLU activation
  - Weights into neuron 1: $w_{11} = 0.5$ (from x_1), $w_{12} = -1$ (from x_2)
  - Weights into neuron 2: $w_{21} = 1$ (from x_1), $w_{22} = 0.5$ (from x_2)
- Output neuron: linear, weights $v_1 = 2$ (from hidden_1), $v_2 = -3$ (from hidden_2)
- No biases, learning rate $\alpha = 0.01$, loss $= \frac{1}{2}(\hat{y} - y)^2$

1. **Forward pass (compute network output $\hat{y}$)**
   - **Hidden pre-activations**
     - $a_1 = w_{11} \cdot x_1 + w_{12} \cdot x_2 = 0.5 \cdot 1 + (-1) \cdot (-2) = 0.5 + 2 = 2.5$
     - $a_2 = w_{21} \cdot x_1 + w_{22} \cdot x_2 = 1 \cdot 1 + 0.5 \cdot (-2) = 1 - 1 = 0$
   - **Hidden activations (ReLU)**
     - $h_1 = \text{ReLU}(a_1) = \max(0, 2.5) = 2.5$
     - $h_2 = \text{ReLU}(a_2) = \max(0, 0) = 0$
   - **Output pre-activation & prediction**
     - $\hat{y} = v_1 \cdot h_1 + v_2 \cdot h_2 = 2 \cdot 2.5 + (-3) \cdot 0 = 5.0$
2. **Gradients of MSE loss**
   - **Loss**: $L = \frac{1}{2}(\hat{y} - y)^2 = \frac{1}{2}(5.0 - 5)^2 = 0$
   - **Error signal at output:** $\delta\_\text{output} = \partial L/\partial \hat{y} = \hat{y} - y = 5.0 - 5 = 0$
   - Because $\delta\_\text{output} = 0$, all downstream gradients will be zero. But to show the general recipe, the **gradients w.r.t. output weights** are computed as $\partial L/\partial v_i = \delta\_\text{output} \cdot h_i \Rightarrow \partial L/\partial v_1 = 0 \cdot 2.5 = 0 \Rightarrow \partial L/\partial v_2 = 0 \cdot 0 = 0$
   - **Backpropagate into hidden**
     - $\delta\_\text{hidden}_i = (\delta\_\text{output} \cdot v_i) \cdot \text{ReLU}'(a_i)$
     - $\text{ReLU}'(a_1) = 1$ (since $a_1 > 0$), $\text{ReLU}'(a_2) = 0$ (since $a_2 = 0$)
     - $\Rightarrow \delta\_\text{hidden}_1 = 0 \cdot 2 \cdot 1 = 0 \Rightarrow \delta\_\text{hidden}_2 = 0 \cdot (-3) \cdot 0 = 0$
   - **Gradients w.r.t. input-to-hidden weights**
     - $\partial L/\partial w_{ij} = \delta\_\text{hidden}_i \cdot x_j \Rightarrow$ all $\partial L/\partial w_{ij} = 0 \cdot x_j = 0$

3. **SGD weight updates**
    - **General rule**: $w \leftarrow w - \alpha \cdot (\partial L/\partial w)$
    - Since every gradient is zero, **no change** occurs:
        - $v_1 \leftarrow 2 - 0.01 \cdot 0 = 2$
        - $v_2 \leftarrow -3 - 0.01 \cdot 0 = -3$
        - $w_{11} \leftarrow 0.5 - 0.01 \cdot 0 = 0.5$
        - $w_{12} \leftarrow -1 - 0.01 \cdot 0 = -1$
        - $w_{21} \leftarrow 1 - 0.01 \cdot 0 = 1$
        - $w_{22} \leftarrow 0.5 - 0.01 \cdot 0 = 0.5$

Remark: Because the network already predicts $\hat{y} = 5$ exactly, the MSE loss gradient is zero and SGD makes no update for this minibatch.

Question 5 (20 marks)
Task: predict default (yes/no) from age, income, loan amount, repayment history

1. Model choice
- **Logistic regression**
  - Directly models probability of a binary outcome, interpretable odds ratios.
  - Appropriate for classification with continuous & categorical features.
- **Decision trees**
  - Non-parametric, handles non-linear interactions & missing values naturally.
  - Provides intuitive "if–then" rules.
- **Linear regression**
  - Predicts continuous targets; not suitable for binary default without manual thresholding and violation of homoscedasticity assumptions.
  - **Rule out** linear regression as a primary model.

2. Feature scaling requirements
- **Linear & logistic regression**
  - Sensitive to feature scale: unscaled features can skew gradient descent convergence and coefficient magnitudes.
  - **Implications if unscaled:** large-scale features dominate regularization penalty, slow training.
- **Decision trees**: **Invariant** to monotonic feature transformations; splitting based on thresholds, so scaling not required.

3. Regularization & complexity control
- **Linear & logistic regression**
  - L1 (Lasso) or L2 (Ridge) penalties shrink coefficients to prevent overfitting.
  - Strength λ tuned via CV; L1 yields sparse models.
- **Decision trees**
  - Control complexity via max depth, min samples per leaf, or pruning.
  - No direct "penalty" term but hyperparameters limit tree growth.

4. Feature-importance determination & caveats
- **Logistic regression**

- Importance via absolute magnitude of standardized coefficients.
- **Pitfall:** multicollinearity can inflate variances and mislead about true importance.

- **Decision trees**
  - Importance via impurity-reduction (e.g. Gini gain) or permutation importance.
  - **Pitfall:** features with many categories or continuous splits may appear artificially more important.
- **General caution:** correlation among predictors can distort single-feature attributions; consider SHAP or partial-dependence for more robust insights