FINAL QUESTION PAPER

1. 31. Which of the following data types represents categories without any intrinsic order or ranking? (a) Ordinal (b) Nominal (c) Interval (d) Ratio

2. 81. Which type of feature selection method evaluates the relevance of features based on their intrinsic properties (like correlation or information gain) without involving a machine learning model? a) Wrapper method b) Filter method c) Embedded method d) Hybrid method

3. 35. In the context of "Preparing to Model" in machine learning, why is it crucial to identify the type of data for each feature? (a) It determines the programming language to be used. (b) It dictates the choice of appropriate preprocessing techniques and statistical analyses. (c) It specifies the required database schema for data storage. (d) It has no direct impact on model performance, only on data presentation.

4. 86. The concept of combining multiple simple models to achieve better predictive performance than any single model is known as: a) Hyperparameter tuning b) Ensemble learning c) Regularization d) Feature engineering

5. 29. "Boosting" is an ensemble technique where: (a) Multiple models are trained in parallel, and their predictions are simply averaged. (b) A single model is repeatedly trained on the entire dataset until optimal performance is achieved. (c) Models are trained sequentially, with each new model attempting to correct the errors made by previous ones. (d) The dataset is randomly sampled without replacement for each individual model's training.

6. 75. Among the following feature selection approaches, which one is generally expected to yield the highest predictive accuracy for a specific model, but at the cost of significantly higher computational expense? (a) Filter methods using a simple correlation score. (b) Wrapper methods employing exhaustive search. (c) Embedded methods like Lasso regularization. (d) Hybrid methods with a pre-filtering step.

7. 69. What is a primary drawback of using Wrapper methods for feature selection, especially with a large number of features? (a) They tend to ignore the interaction between features. (b) Their computational cost can be very high due to repeated model training. (c) They are highly susceptible to overfitting the feature selection process to the training data. (d) They often select features that are redundant rather than relevant.

8. 88. Data normalization (e.g., Min-Max scaling) is primarily performed to: a) Convert categorical data into numerical data b) Reduce the number of features in a dataset c) Scale features to a specific range (e.g., 0 to 1) to prevent features with larger values from dominating the learning process d) Remove outliers from the dataset

9. 65. Which of the following is an example of an Embedded method for feature selection? (a) Principal Component Analysis (PCA) (b) Forward Selection with a Logistic Regression model (c) Lasso (L1 regularization) regression (d) Chi-squared test for feature ranking

10. 45. In machine learning, scaling techniques like Min-Max scaling or Standardization are typically applied to which types of data to normalize their ranges? (a) Nominal and Ordinal data (b) Interval and Ratio data (c) Only Nominal data (d) Only Ordinal data

11. 64. In Wrapper feature selection methods, the selection of an optimal feature subset is typically guided by: (a) Predefined statistical thresholds for individual feature scores. (b) The performance of a chosen machine learning algorithm on different feature subsets. (c) The intrinsic properties of the data, such as variance or covariance. (d) Regularization penalties applied during model training.

12. 6. The process of reducing the number of input features to avoid the curse of dimensionality and improve model efficiency is known as: (a) Feature engineering (b) Data augmentation (c) Dimensionality reduction (d) Model selection

13. 21. If a machine learning model is trained and evaluated using only a single train-test split, what is a potential drawback? (a) The model will always perform well on unseen data. (b) The performance estimate might be highly dependent on the specific split of the data. (c) The training process will be significantly slower. (d) It only works for classification problems, not regression.

14. 90. Why is "Summary And Revision" important in the "Preparing to Model and Preprocessing" stage of a machine learning project? a) It helps in quickly deploying the model without further checks. b) It ensures all preprocessing steps are correctly applied and that the data is ready for modeling, preventing errors in later stages. c) It primarily focuses on model selection and hyperparameter tuning. d) It is an optional step that can be skipped if time is short.

15. 41. A survey asks respondents to rate their satisfaction level as "Very Unsatisfied", "Unsatisfied", "Neutral", "Satisfied", "Very Satisfied". This collected data is best classified as: (a) Nominal (b) Ordinal (c) Interval (d) Ratio

16. 80. Principal Component Analysis (PCA) is a technique primarily used for which of the following data preprocessing tasks? a) Feature scaling b) Dimensionality reduction c) Handling categorical data d) Outlier removal

17. 26. The F1-score is a harmonic mean of Precision and Recall. It is particularly useful when: (a) Both False Positives and False Negatives are equally costly to the application. (b) Only the total number of correct predictions matters, regardless of class. (c) The model's training speed is the primary concern. (d) The dataset is perfectly balanced across all classes.

18. 49. What is a common consequence of not adequately handling outliers in a dataset before training a linear regression model? a) The model's coefficients might become overly sensitive and skewed. b) The training time of the model will significantly decrease. c) The model will always perfectly fit the training data. d) The model will only learn from the outliers, ignoring other data points.

19. 13. Which of the following data types allows for both meaningful differences between values and a true zero point, enabling ratio comparisons (e.g., height, weight)? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

20. 66. When comparing Filter methods and Wrapper methods for feature selection, which statement regarding computational cost is generally true? (a) Wrapper methods are generally less computationally expensive than Filter methods. (b) Filter methods are generally less computationally expensive than Wrapper methods. (c) Both methods have comparable computational costs. (d) Computational cost depends entirely on the number of samples, not the method.

21. 36. Which of the following operations is permissible and meaningful for Ordinal data? (a) Calculating the average (b) Determining the mode (c) Performing multiplication (d) Comparing ratios

22. 32. Consider a dataset containing 'Educational Qualification' with values like "High School", "Diploma", "Bachelor's", "Master's". This is an example of which data type? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

23. 8. A dataset where the difference between two values is meaningful, but there is no true zero point (e.g., temperature in Celsius), is classified as which type of data? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

24. 61. Which of the following is the primary goal of feature subset selection in machine learning? (a) To increase the interpretability of complex models. (b) To reduce the number of input variables by removing irrelevant or

redundant features. (c) To transform existing features into a new, lower-dimensional space. (d) To impute missing values in the dataset.

25. 2. The primary goal of data preprocessing in machine learning is to: (a) Immediately improve the model's accuracy on the test set (b) Make the data suitable for analysis and improve model performance (c) Automatically select the best machine learning algorithm (d) Visualize the final results of the model

26. 72. How does Lasso (L1 regularization) perform feature selection in a linear model? (a) It adds a penalty to the sum of squared coefficients, which only shrinks their values. (b) It forces some feature coefficients to become exactly zero, effectively removing those features. (c) It transforms features into principal components, reducing the dimensionality. (d) It evaluates features using a statistical test independent of the model's loss function.

27. 44. Which of the following data types allows for ranking and ordering, but the differences between successive ranks are not necessarily equal or meaningful? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

28. 60. A common practice for handling missing values in categorical features, besides mode imputation, is to: a) Convert them to the average numerical value. b) Treat missing as a separate category. c) Replace them with the most common numerical value. d) Interpolate them based on adjacent values.

29. 78. Which strategy for handling missing values involves replacing the missing entry with the most frequently occurring value in that feature? a) Mean imputation b) Median imputation c) Mode imputation d) Regression imputation

30. 24. In a binary classification problem, if a model has a high Recall but low Precision, it indicates that the model: (a) Is very good at identifying positive cases but also incorrectly labels many negative cases as positive. (b) Is very good at identifying positive cases and rarely makes mistakes. (c) Fails to identify most of the actual positive cases. (d) Correctly identifies negative cases but misses many positive cases.

31. 5. K-fold cross-validation is primarily used for which purpose in machine learning? (a) To reduce the dimensionality of the dataset (b) To evaluate the model's performance and generalization ability (c) To handle outliers in the training data (d) To select the most relevant features for the model

32. 19. In the context of data partitioning, the validation set is primarily used for: (a) Training the final machine learning model. (b) Assessing the final, unbiased performance of the chosen model. (c) Tuning hyperparameters and making model selection decisions. (d) Detecting and correcting errors in the raw data.

33. 89. A feature selection method that combines a filter approach to pre-select a subset of features, followed by a wrapper approach to fine-tune the selection, is known as a: a) Embedded method b) Hybrid method c) Statistical method d) Brute-force method

34. 79. The Interquartile Range (IQR) method is commonly used to detect which data quality issue? a) Missing values b) Duplicate entries c) Outliers d) Inconsistent data types

35. 57. For decision tree-based models (e.g., Random Forest, Gradient Boosting), how do missing values and outliers typically affect their performance compared to linear models? a) They are generally more sensitive to missing values and outliers than linear models. b) They are generally less sensitive to missing values and outliers than linear models. c) They are equally sensitive to missing values but less sensitive to outliers. d) They cannot handle any missing values or outliers without explicit pre-processing.

36. 58. When the missingness in a dataset is 'Missing Not At Random' (MNAR), which approach might be considered to gain insights, even if it adds complexity? a) Simple mean imputation, as it's quick and easy. b) Removing all rows with missing values, as it ensures data completeness. c) Building a model that predicts the

missingness itself, as the missingness carries information. d) Imputing with a constant value like '0' or '-1'.

37. 9. What is the main objective of feature scaling (e.g., normalization or standardization) during data preprocessing? (a) To make the data easier for humans to interpret (b) To ensure all features contribute equally to the distance calculations in algorithms (c) To reduce the total number of features in the dataset (d) To convert categorical features into numerical ones

38. 34. A machine learning model is being developed to predict house prices. One of the features is 'Number of Bedrooms'. What type of data is 'Number of Bedrooms'? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

39. 25. Accuracy is a common metric, but it can be misleading in cases of imbalanced datasets. Which scenario best illustrates this limitation? (a) A dataset where all classes have an equal number of samples. (b) A classification task where the positive class is extremely rare (e.g., 1% of the data). (c) A regression task where the model predicts continuous values. (d) A dataset with many missing values and outliers.

40. 52. When dealing with categorical features that have missing values, which imputation strategy is generally most appropriate? a) Imputing with the mean of the feature. b) Imputing with the median of the feature. c) Imputing with the mode (most frequent category) of the feature. d) Imputing with a random numerical value.

41. 70. A Hybrid feature selection approach typically aims to combine the advantages of: (a) Dimensionality reduction and data imputation. (b) Supervised and unsupervised learning techniques. (c) Filter methods (computational efficiency) and Wrapper methods (model specific performance). (d) Cross-validation and ensemble learning.

42. 83. During the "Preparing to Model" activity, converting a categorical feature like 'City' into numerical columns such as 'City_NewYork', 'City_London', etc., is an example of which technique? a) Label Encoding b) One-Hot Encoding c) Feature Scaling d) Binning

43. 56. When using K-Nearest Neighbors (KNN) for imputation, the missing value for a data point is estimated by: a) The average of the values from its K nearest neighbors in the feature space. b) The most frequent value across all features in the dataset. c) A randomly selected value from the feature's existing values. d) The overall mean of the feature for the entire dataset.

44. 55. Among the following measures of central tendency, which one is most robust to the presence of outliers in a dataset? a) Mean b) Median c) Mode d) Weighted Mean

45. 22. During the model selection phase, which of the following criteria is generally the most important when comparing different models? (a) The speed at which the model was trained. (b) The complexity of the model's underlying algorithm. (c) The model's performance on the unseen test data or its generalized performance. (d) The number of lines of code required to implement the model.

46. 43. If a feature represents 'Weight in kilograms', what type of data is it? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

47. 3. Data representing a student's grade level (e.g., Freshman, Sophomore, Junior, Senior) is an example of which type of data? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

48. 84. A dataset contains a feature "Number of children". This feature is an example of which type of quantitative data? a) Nominal b) Ordinal c) Interval d) Ratio

49. 82. The main purpose of K-fold cross-validation during the "Learning: Data Partition" activity is to: a) Increase the amount of training data b) Prevent overfitting and provide a more robust estimate of model performance c) Speed up the model training process d) Select the best features for the model

50. 30. Which data quality issue refers to the presence of data points that are significantly different from the majority of the data and can disproportionately influence model training? (a) Missing values. (b) Imbalanced classes. (c) Outliers. (d) Redundant features.

51. 42. Which statement is true regarding Quantitative data? (a) It represents categories without any order. (b) It can only be discrete values. (c) It consists of numerical values that can be measured or counted. (d) It is always in string format.

52. 12. In the context of data quality, what does "data remediation" primarily refer to? (a) The process of collecting more data to fill gaps (b) The act of fixing or correcting identified data quality issues (c) The stage of generating new features from existing ones (d) The process of backing up the dataset

53. 10. Which feature selection method evaluates different subsets of features by training and testing a machine learning model for each subset, often leading to better performance but higher computational cost? (a) Filter method (b) Wrapper method (c) Embedded method (d) Hybrid method

54. 47. If the probability of a value being missing is related to the unobserved values themselves (e.g., people with very high income are less likely to report it), this type of missing data is known as: a) Missing Completely At Random (MCAR) b) Missing At Random (MAR) c) Missing Not At Random (MNAR) d) Missing Partially At Random (MPAR)

55. 37. Data representing 'Customer ID' or 'Product Category' (e.g., "Electronics", "Clothing") falls under which broad category of data? (a) Quantitative (b) Numeric (c) Categorical (d) Continuous

56. 28. Which of the following statements about "Bagging" (Bootstrap Aggregating) is true? (a) It trains models sequentially, where each model corrects the errors of the previous one. (b) It trains multiple models independently on different subsets of the training data, then averages their predictions. (c) It assigns higher weights to misclassified samples to focus subsequent models. (d) It is primarily used for deep learning models only.

57. 23. A Confusion Matrix provides a detailed breakdown of a classification model's performance. What does a "False Positive" (FP) represent? (a) The model correctly predicted the positive class. (b) The model incorrectly predicted the positive class when the actual class was negative. (c) The model incorrectly predicted the negative class when the actual class was positive. (d) The model correctly predicted the negative class.

58. 87. In the context of model performance evaluation, a Confusion Matrix is most directly used to assess a model's performance in which type of task? a) Regression b) Clustering c) Classification d) Dimensionality reduction

59. 67. The "Curse of Dimensionality" refers to the phenomenon where: (a) Models become too simple with too few features. (b) Data becomes easier to visualize in high-dimensional spaces. (c) The amount of data required to sufficiently generalize grows exponentially with the number of features. (d) Feature scaling becomes unnecessary in high dimensions.

60. 15. What is the fundamental reason for addressing outliers in the "Preparing to Model" phase? (a) Outliers always represent errors and must be removed (b) Outliers can disproportionately influence model training and lead to biased results (c) Outliers are only relevant for classification tasks, not regression (d) Outliers are only relevant for small datasets

61. 18. The primary reason for partitioning a dataset into training, validation, and test sets is to: (a) Ensure that the model can be trained faster on smaller subsets of data. (b) Prevent the model from memorizing the training data and to assess its generalization ability. (c) Allocate more data for training to achieve higher accuracy. (d) Simplify the process of handling missing values and outliers.

62. 73. Which feature selection approach is generally considered "model agnostic" because it evaluates features independently of a specific machine learning algorithm? (a) Wrapper methods (b) Embedded methods (c) Hybrid methods (d) Filter methods

63. 62. Dimensionality reduction techniques primarily aim to address which of the following issues? (a) Overfitting by increasing model complexity. (b) The "curse of dimensionality" by reducing the number of input variables. (c) Data imbalance by oversampling minority classes. (d) Non-linearity in the data by applying kernel methods.

64. 76. Which of the following is the primary objective of the "Preparing to Model" phase in a machine learning workflow? a) To deploy the final model into production b) To ensure the data is suitable for model training and evaluation c) To determine the optimal hyper-parameters for a chosen model d) To present the model's performance to stakeholders

65. 71. Principal Component Analysis (PCA) is primarily used for: (a) Selecting a subset of original features based on their individual correlation with the target variable. (b) Transforming original features into a new set of orthogonal, uncorrelated components. (c) Iteratively adding or removing features based on a model's performance. (d) Encoding categorical variables into numerical representations.

66. 14. When preparing categorical data for a machine learning model, one-hot encoding is typically used to: (a) Convert ordinal data into a numerical scale (b) Assign a unique numerical identifier to each category (c) Create binary features for each category, avoiding implied order (d) Reduce the number of categories in a feature

67. 63. Which statement accurately describes a characteristic of Filter methods for feature selection? (a) They evaluate feature subsets based on the performance of a specific machine learning model. (b) They are computationally expensive as they involve training a model for each subset. (c) They use statistical measures like correlation or information gain, independent of the learning algorithm. (d) They integrate feature selection directly into the model training process.

68. 40. When preprocessing data for a machine learning task, one-hot encoding is a common technique applied to convert which type of data into a numerical format? (a) Ordinal data (b) Ratio data (c) Nominal data (d) Interval data

69. 51. Winsorization (or capping) is a technique used to handle outliers by: a) Removing the outlier data points entirely from the dataset. b) Replacing outlier values with the mean of the feature. c) Replacing extreme values with values at a specified percentile (e.g., 5th and 95th percentile). d) Transforming the entire feature using a logarithmic function.

70. 16. Which of the following activities is NOT typically considered part of the "Preparing to Model" phase in a machine learning project? (a) Defining the business problem and objectives. (b) Gathering and cleaning the raw data. (c) Deploying the trained model into a production environment. (d) Performing exploratory data analysis to understand data characteristics.

71. 85. Which characteristic is most representative of a "Wrapper" feature selection method? a) It is computationally efficient as it runs independently of the learning algorithm. b) It uses a specific machine learning algorithm to evaluate the performance of feature subsets. c) It combines the strengths of both filter and embedded methods. d) It relies on statistical measures like variance or correlation.

72. 68. Which of the following statistical measures is commonly used in Filter methods to evaluate the relevance of features for classification tasks? (a) Root Mean Squared Error (RMSE) (b) Confusion Matrix score (c) Information Gain or Chi-squared statistic (d) F1-score

73. 74. A key benefit of performing dimensionality reduction or feature subset selection before training a machine learning model is: (a) Guaranteed improvement in model accuracy. (b) Increased complexity of the final model. (c) Reduced training time and potentially better generalization by mitigating overfitting. (d) Conversion of quantitative data into qualitative data.

74. 11. Why is it important to partition data into training, validation, and testing sets? (a) To ensure the model is trained on all available data for maximum accuracy (b) To prevent overfitting and evaluate the model's generalization to unseen data (c) To simplify the data preprocessing steps (d) To speed up the model training process

75. 54. A significant drawback of simply deleting rows (listwise deletion) that contain any missing values is: a) It can introduce bias if missingness is not MCAR. b) It always improves the model's accuracy. c) It makes the dataset larger and more complex. d) It is computationally expensive for large datasets.

76. 46. Which of the following best describes the primary goal of data quality and remediation in machine learning preprocessing? a) To increase the number of features in the dataset. b) To ensure the data is clean, consistent, and suitable for model training. c) To visualize the data in various formats. d) To reduce the dimensionality of the dataset.

77. 17. Before applying a machine learning algorithm, an important step in data preprocessing is feature scaling. This is primarily done to: (a) Reduce the number of features to prevent overfitting. (b) Ensure that all features contribute equally to the distance calculations. (c) Increase the interpretability of the model. (d) Convert categorical features into numerical ones.

78. 7. Which data quality issue is characterized by data points that significantly deviate from other observations in the dataset? (a) Missing values (b) Inconsistent data (c) Outliers (d) Duplicate records

79. 33. Which characteristic primarily distinguishes Interval data from Ratio data? (a) Interval data can be sorted, while Ratio data cannot. (b) Interval data has a meaningful absolute zero point, while Ratio data does not. (c) Interval data allows for addition and subtraction, but ratios are not meaningful, unlike Ratio data. (d) Interval data is qualitative, while Ratio data is quantitative.

80. 77. In the context of data types, which of the following best describes Ordinal data? a) Data with no inherent order, only distinct categories b) Data with a meaningful zero point, allowing ratio comparisons c) Data that can be ordered or ranked, but with unequal intervals between values d) Data where differences between values are meaningful, but without a true zero

81. 59. What is a key consideration when deciding whether to remove or impute outliers? a) If the outlier is a genuine observation or a data error. b) The computational resources available for processing. c) The aesthetic appeal of the data visualization. d) The number of features in the dataset.

82. 50. The Interquartile Range (IQR) method is commonly used to detect outliers. A data point is often considered an outlier if it falls below Q1 - 1.5 * IQR or above what threshold? a) Q2 + 1.5 * IQR b) Q3 + 1.5 * IQR c) Q1 + 3 * IQR d) Q3 - 3 * IQR

83. 27. Ensemble methods like Bagging and Boosting are techniques for "Performance Improvement". What is their fundamental approach? (a) To reduce the dimensionality of the dataset by selecting the most important features. (b) To train a single, very complex model with millions of parameters. (c) To combine multiple simpler models to achieve better overall predictive performance. (d) To apply advanced regularization techniques to a

single model.

84. 53. Which of the following is NOT a common reason for the presence of outliers in a dataset? a) Measurement errors or data entry mistakes. b) Natural variation within the data. c) Intentional data manipulation by an adversary. d) Errors during data preprocessing steps like standardization.

85. 38. The temperature in degrees Celsius (e.g., 0 degrees Celsius, 10 degrees Celsius, 20 degrees Celsius) is an example of which data type? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

86. 1. Which of the following activities is typically performed during the "Preparing to Model" stage in a machine learning pipeline? (a) Training the final model (b) Data cleaning and preprocessing (c) Deploying the model to production (d) Monitoring model performance

87. 39. For which data type is a true or absolute zero point meaningful, allowing for the calculation of ratios? (a) Nominal (b) Ordinal (c) Interval (d) Ratio

88. 48. For a numerical feature with a highly skewed distribution containing missing values, which imputation method is generally preferred to minimize the impact on the distribution? a) Mean imputation b) Mode imputation c) Median imputation d) Zero imputation

89. 4. Which of the following is a common strategy for handling missing numerical values in a dataset? (a) Deleting all rows containing any missing value (b) Imputing with the mean or median of the feature (c) Converting all missing values to zero (d) Replacing missing values with a randomly generated number

90. 20. K-fold cross-validation is a technique primarily used to: (a) Speed up the initial data loading and preprocessing steps. (b) Obtain a more robust and less biased estimate of a model's performance. (c) Select the optimal features from a large set of available features. (d) Combine the predictions of multiple models to improve overall accuracy.

ANSWER KEY

1. (b)
2. (b)
3. (b)
4. (b)
5. (c)
6. (b)
7. (b)
8. (c)
9. (c)
10. (b)
11. (b)
12. (c)
13. (b)
14. (b)
15. (b)
16. (b)
17. (a)
18. (a)
19. (d)
20. (b)
21. (b)
22. (b)
23. (c)
24. (b)
25. (b)
26. (b)
27. (b)

28. (b)

29. (c)

30. (a)

31. (b)

32. (c)

33. (b)

34. (c)

35. (b)

36. (c)

37. (b)

38. (d)

39. (b)

40. (c)

41. (c)

42. (b)

43. (a)

44. (b)

45. (c)

46. (d)

47. (b)

48. (d)

49. (b)

50. (c)

51. (c)

52. (b)

53. (b)

54. (c)

55. (c)

56. (b)

57. (b)

58. (c)

59. (c)

60. (b)

61. (b)

62. (d)

63. (b)

64. (b)

65. (b)

66. (c)

67. (c)

68. (c)

69. (c)

70. (c)

71. (b)

72. (c)

73. (c)

74. (b)

75. (a)

76. (b)

77. (b)

78. (c)

79. (c)

80. (c)

81. (a)

82. (b)

83. (c)

84. (d)

85. (c)

86. (b)

87. (d)

88. (c)

89. (b)

90. (b)