Soru 1

What is the main goal of data mining?

A) To build websites quickly
B) To store large amounts of data
C) To delete old data from databases
D) To write complex software programs
E) To discover useful and previously unknown patterns from large data sets

Answer: E

Which of the following best describes the process of data mining?

A) Collecting and storing large volumes of data
B) Cleaning and transforming raw data
C) Extracting useful patterns from large datasets
D) Visualizing data for presentation purposes
E) Encrypting data to protect it from unauthorized Access

Correct Answer:
C) Extracting useful patterns from large datasets

Which of the following best defines Data Mining?

A) Storing large amounts of data in databases.
B) Cleaning and formatting raw data for reporting.
C) The process of directly visualizing raw data.
D) The non-trivial extraction of implicit, previously unknown and potentially useful information from data.
E) The creation of data through user input forms.

Correct Answer: D

Which of the following statements best describes the relationship between Data Mining and Knowledge Discovery in Databases (KDD)?

A) They are completely different processes with no overlap
B) Data mining is one step in the KDD process
C) KDD is one step in the data mining process
D) They are competing methodologies for analyzing data
E) Data mining focuses on structured data while KDD handles unstructured data

Answer: B

Which of the following is not a predictive data mining task?

A) Classification
B) Regression
C) Deviation Detection
D) Clustering
E) Forecasting future values

Answer: D

Which of the following statements best reflects the need for data mining?

A) We already have too much analyzed data.
B) Data mining helps reduce the amount of data collected.
C) We are data rich, but information poor.
D) Data mining replaces the need for human decision-making.
E) Data mining is only useful for scientific research.

Answer: C

Soru 2

Which of the following is an example of a data mining task?

A) Looking up a phone number in a directory
B) Adding new records to a database
C) Querying the total number of users
D) Sorting data by alphabetical order
E) Discovering that people who buy diapers often buy beer together

Answer: E

Which of the following is not typically considered a kind of data to be mined?

A) Relational databases
B) Transactional data
C) Multimedia data
D) Operating system source code
E) Data streams

Correct Answer:
D) Operating system source code

Which of the following is not an alternative name for Data Mining?

A) Knowledge Discovery in Databases
B) Business Intelligence
C) Data Entry Processing
D) Data/Pattern Analysis
E) Information Harvesting

Correct Answer: C

In the context of data mining, which of the following is NOT mentioned as a challenge in processing modern data?

A) The explosive growth in the volume of collected data
B) The difficulty in analyzing different forms of data
C) The limited computational power of modern systems
D) The gap between being data rich and information poor
E) The need to extract hidden information not readily evident

Answer: C

Which of the following is not a step in the Knowledge Discovery (KDD) process?

A) Data cleaning
B) Data integration
C) Data encryption
D) Pattern evaluation
E) Data selection

Answer: C

Which of the following activities is considered a data mining task?

A) Sorting student records by ID numbers
B) Looking up a phone number in a directory
C) Predicting stock prices using historical data
D) Querying a web search engine for “Amazon”
E) Computing the total sales of a company

Answer C

Which of the following is NOT considered data mining?
A) Identifying buying patterns in a supermarket
B) Sorting students by ID number
C) Predicting future stock prices using historical data
D) Detecting abnormal heart rates
E) Classifying emails as spam or not

Correct answer: B

Which of the following activities falls under the category of Graph Mining rather than Information Network Analysis or Web Mining?
A) Analyzing social networks based on actors and relationships (edges).
B) Discovering web communities and analyzing usage patterns.
C) Finding frequent subgraphs, such as chemical compounds or XML trees.
D) Utilizing PageRank concepts for network analysis.
E) Link mining based on semantic information carried by links.
Correct Answer: C

What is the main objective of clustering in unsupervised learning?
A) Minimize the number of data points in each cluster
B) Maximize the similarity between clusters
C) Maximize similarity within clusters and minimize similarity between clusters
D) Assign predefined labels to data points
E) Reduce the dimensionality of the data
Correct Answer: C)

What is the primary goal of data mining?
A) Data storage
B) Data visualization
C) Discovering paterns in large datasets
D) Data entry
E) Data encryption
Correct Answer: C) Discovering paterns in large datasets

What is the main goal of data mining?

A) To store information securely.
B) To format raw data.
C) To delete unnecessary data.
D) To discover useful patterns and knowledge from large datasets.
E) To build simple databases.

Correct answer: D) To discover useful patterns and knowledge from large datasets.

What is the principle of clustering?
A) Maximizing intra-class similarity & minimizing interclass similarity
B) Assigning unseen records to a class
C) Discovering sequential patterns
D) Detecting outliers
E) Calculating probability scores for items
Correct Answer: A

Which is an example of data mining rather than simple querying?
A) Searching for "Amazon" on Google
B) Looking up a phone number in the directory
C) Finding common item pairs in transactions
D) Asking a friend about trending products
E) Viewing weather updates on a website

Correct answer: C

About the "Mining Methodology" challenges in Data Mining Which of the following is explicitly listed as one of these methodological challenges?
A) Developing privacy-preserving data mining techniques.
B) Ensuring the efficiency and scalability of algorithms on large datasets.
C) Handling noise, uncertainty, and incompleteness of data.
D) Creating effective visualization for data mining results.
E) Mining data from dynamic and networked repositories.
Correct Answer: C

Which of the following tasks involves analyzing and extracting meaningful patterns from networks, such as social or web networks?
A) Regression analysis
B) Graph mining and information network analysis
C) Dimensionality reduction
D) Classification of data points
E) Time series forecasting
Correct answer: B)

Which technique is commonly used for reducing the dimensionality of a dataset in data mining?
A) Clustering
B) Principal Component Analysis (PCA)
C) Classification
D) Association Rule Learning
E) Data Sampling
Correct Answer: B) Principal Component Analysis (PCA)

What is the purpose of classification in data mining?

A) To merge different databases.
B) To count the number of data points.
C) To label data into predefined categories based on patterns.
D) To measure the size of a dataset.
E) To sort files alphabetically.

Correct answer: C) To label data into predefined categories based on patterns.

Question 2: Which of the following is not a typical application of classification methods?
A) Detecting credit card fraud
B) Grouping customers based on purchasing habits
C) Classifying emails as spam or non-spam
D) Categorizing news articles into economy, sports, and health
E) Predicting whether a customer will buy a new product

Correct Answer: B

Which of the following statements is TRUE regarding classification in data mining?
A) Clustering is used to assign predefined class labels to new data.
B) Association rule discovery uses labeled data to train a classifier.
C) Regression is applied to group similar objects based on distance.
D) Data integration removes outliers before testing.
E) A training set is used to build the model and a test set is used to validate it.

Correct Answer: E

Which scenario is a good example of using clustering in data mining?
A) Identifying fraudulent credit card transactions
B) Predicting next week's stock prices
C) Grouping customers based on lifestyle and geography
D) Estimating sales amounts based on ad spending
E) Detecting spam emails
Correct Answer: C)

Which of the following is not a typical Data Mining task?

A) Classification
B) Clustering
C) Regression
D) Association Rule Discovery
E) Data Normalization

Correct Answer: E

Which of the following is not a challenge encountered in the data mining process?

A) Dealing with noisy and incomplete data
B) Integration of multiple disciplines
C) Enhancing knowledge discovery in networked environments
D) Random generation of data
E) Mining knowledge in multi-dimensional data space

Correct Answer: D) Random generation of data

Which of the following best represents a practical example of sequential pattern mining?

A) Grouping customers by age
B) Predicting which links a user will click on based on previous behavior
C) Identifying the most popular survey answers
D) Reporting the stock status of items
E) Classifying customers by income and gender

Correct Answer: B) Predicting which links a user will click on based on previous behavior

Which of the following is used as a similarity measure in clustering algorithms?

A) Gini index
B) Time series graphs
C) Euclidean distance
D) Entropy calculation
E) Decision tree depth

Correct Answer: C) Euclidean distance

A supermarket wants to analyze which products are frequently purchased together to optimize their shelf arrangements. Which data mining technique would be most suitable for this purpose?

A) Regression analysis
B) Classification
C) Sequential pattern discovery
D) Association rule mining
E) Deviation detection
Correct Answer: D) Association rule mining

What type of learning is clustering considered in data mining?
A) Supervised learning
B) Reinforcement learning
C) Semi-supervised learning
D) Unsupervised learning
E) Deep learning
Correct Answer: D)

What is the main purpose of regression in data mining?

A) To divide data into unrelated groups
B) To discover items that are often purchased together
C) To classify emails as spam or not spam
D) To predict a continuous numeric value based on other variables
E) To remove duplicate data entries

Correct Answer: D

Which of the following is not part of user interaction in data mining?

A) Visualization of results
B) Incorporation of background knowledge
C) Building fully automatic decision-making systems
D) Designing interactive mining processes
E) Presenting results in a user-friendly way

Correct Answer: C) Building fully automatic decision-making systems

Which of the following is NOT a commonly used criterion for evaluating the interestingness of a data mining pattern?

A) Predictive accuracy
B) Novelty
C) Coverage
D) Scalability
E) Timeliness

Correct Answer: D) Scalability

Which of the following is an example of anomaly detection?

A) Creating customer groups for marketing
B) Recommending products based on purchase history
C) Detecting unusual credit card transactions
D) Categorizing newspaper articles by topic
E) Summarizing customer satisfaction surveys numerically

Correct Answer: C) Detecting unusual credit card transactions

An online news platform wants to automatically categorize articles as economy, sports, or health-related content.
Which data mining technique should they implement?

A) Classification
B) Clustering
C) Association rule mining
D) Sequential pattern discovery
E) Regression
Correct Answer: A) Classification

Which of the following best describes an outlier in data mining?

A) A data object that is duplicated multiple times in the dataset
B) A data object that does not comply with the general behavior of the data
C) A data object that is missing certain attribute values
D) A data object that occurs most frequently in the dataset
E) A data object that is used as a label in supervised learning

Correct answer: B) A data object that does not comply with the general behavior of the data

Which of the following is not a classification method?

A) Decision trees
B) Naïve Bayesian classification
C) Support vector machines
D) K-means
E) Neural networks

Correct Answer: D) K-means

Which of the following is a descriptive data mining task?

A) Classification
B) Regression
C) Clustering
D) Prediction
E) Deviation Detection

Correct Answer: C) Clustering

Which of the following is a predictive data mining task?

A) Clustering
B) Association Rule Discovery
C) Sequential Pattern Discovery
D) Classification
E) None of the above

Correct Answer: D) Classification

Which of the following attributes cannot be used to compute the arithmetic mean meaningfully?

A) Age of students in a classroom
B) Daily temperature in Celsius
C) Drink size labeled as {1 = Small, 2 = Medium, 3 = Large}
D) Income of employees in dollars
E) Number of items sold in a day

Correct Answer: C) Drink size labeled as {1 = Small, 2 = Medium, 3 = Large}

Which of the following pairs correctly match an attribute type with its valid example?

A) Nominal – Temperature in Celsius
B) Ordinal – Number of children
C) Interval – Calendar dates
D) Ratio – Shirt size (S, M, L)
E) Ordinal – Gender (Male, Female)

Correct Answer: C)

Which of the following is **not** an appropriate use case for Deviation/Anomaly Detection?

A) Identifying unusual patterns in network traffic for security purposes
B) Detecting defective items in a manufacturing process
C) Spotting abnormal spikes in website activity
D) Categorizing news articles by topic
E) Finding irregularities in medical test results

Correct answer: D) Categorizing news articles by topic

Which of the following best describes the purpose of clustering algorithms?

A) Predicting the future using historical data
B) Generating human-interpretable rules
C) Creating classification models using labeled data
D) Grouping data points based on similarity
E) Filling in missing data

Correct Answer: D) Grouping data points based on similarity

Which of the following algorithms can be used to classify emails as spam or not spam?

A) K-means
B) Naïve Bayes
C) Apriori
D) Euclidean Distance
E) Sequential Pattern Mining

Correct Answer: B) Naïve Bayes

What is a test set in classification?
A) A set of records used to find the class attribute
B) A collection of attributes without class labels
C) A dataset used to build the classification model
D) A set of previously unseen records used to determine the accuracy of the model
E) A group of records used for cleaning and preprocessing

Correct answer: D

9. Which of the following similarity measures only considers "presence" values (1) and ignores "absence" matches (0)?

A) Euclidean distance
B) Manhattan distance
C) Simple Matching Coefficient
D) Jaccard Coefficient
E) Cosine Similarity

Correct Answer: D) Jaccard Coefficient

Which of the following statements about attribute types is TRUE?

A) Nominal attributes have a meaningful zero and support ratio comparison.
B) Ordinal attributes support both addition and multiplication operations.
C) Interval attributes allow calculation of ratios between values.
D) Ratio attributes have a true zero and support all basic arithmetic operations.
E) Nominal attributes allow calculation of mean and standard deviation.

Correct Answer: D)

Which of the following statements is true regarding nominal attributes?

A) They represent numerical values with meaningful ratios.
B) Their values can be ranked in a meaningful order.
C) They are numeric and allow for arithmetic operations like addition.
D) They have a true zero point, allowing for meaningful comparisons.
E) They represent categories without any inherent ordering.

Correct Answer: E

Which of the following best describes an interval attribute?

A) Represents categorical data with no inherent order.
B) Has an order, but differences between values are not meaningful.
C) Differences between values are meaningful, but there is no true zero-point.
D) Both differences and ratios between values are meaningful.
E) Can only take a finite set of distinct values.
Correct Answe: C

Which of the following statements about attribute types is TRUE?

A) Nominal attributes have a meaningful order and support arithmetic operations.
B) Interval attributes have a true zero-point, so ratios are meaningful.
C) Ordinal attributes support comparison but not meaningful arithmetic differences.
D) Ratio attributes cannot be used to compute mean or median values.
E) All attribute types support both order and multiplication operations.

Correct Answer: C

Which of the following is an example of a nominal attribute?

A) Temperature (°C)
B) Student grade (A, B, C...)
C) Gender (Male/Female)
D) Age (in years)
E) Height (in meters)

Correct answer: C) Gender (Male/Female)

What is the difference between classification and prediction?
A) Classification predicts continuous values, while prediction uses categorical labels.
B) Classification predicts categorical (discrete, unordered) labels, while prediction models continuous-valued functions.
C) Classification and prediction both work on the same type of data and give different results.
D) Classification is used only for medical diagnosis, while prediction is used only for marketing.
E) Classification models continuous-valued functions, while prediction classifies data based on categories.

Correct answer: B

Question 1:
Which of the following are basic types of attributes in data objects?

A) Nominal and Ordinal (doğru)
B) Continuous
C) Single-valued
D) Structured
E) Numerical

Which of the following is the best example of a ratio attribute?

A) Zip code
B) Temperature in Celsius
C) Level of satisfaction (e.g., happy, neutral, sad)
D) Weight in kilograms
E) Academic grade (e.g., A, B, C)

Correct Answer: D

Which type of data is most commonly associated with a sparse matrix representation?

A) Temporal data
B) Sequential data
C) Document (text) data
D) Transaction data
E) Graph data
Correct Answer: D

Consider the properties of attribute values: distinctness, order, addition, and multiplication. Which attribute type supports all four properties?

A) Nominal
B) Ordinal
C) Interval
D) Ratio
E) Binary

Correct Answer: D

Which attribute type allows for a meaningful ordering of values, but not meaningful differences between them?

A) Nominal
B) Ordinal
C) Interval
D) Ratio
E) Binary

Correct answer: B) Ordinal

How do decision trees work?
A) Decision trees perform an attribute test at each internal node, and each branch represents the outcome of that test, eventually providing the class label.
B) Decision trees classify only based on continuous attributes.
C) Decision trees give direct results without analyzing the data.
D) Decision trees make decisions based solely on the majority of examples.
E) Decision trees can handle both categorical and continuous attributes for classification.

Correct answer: A

Question 2:
Which of the following correctly describes the nature of data objects?

A) Data objects consist of attributes and values (doğru)
B) Data objects include records
C) Data objects are normalized
D) Data objects have no relationships
E) Data objects only store values

Question 1
Which attribute type allows ordering but not addition of values?

A) Nominal
B) Ordinal
C) Interval
D) Ratio
E) Binary

Correct Answer: B) Ordinal

Which of the following statements about scatter plots is true?

A) They can show frequency distribution of single variables.
B) They are only suitable for categorical attributes.
C) They are used to explore relationships between two numeric attributes.
D) They replace the need for histograms.
E) They are best used for five-number summaries.

Correct Answer: C) They are used to explore relationships between two numeric attributes.

Which chart shows pairs of values to observe a relationship between two variables?

A) Histogram
B) Boxplot
C) Scatter Plot
D) Quantile Plot
E) Bar Chart

Answer: C) Scatter Plot

Which of the following is a statistical measure used to assess the spread of a data set?

A) Median
B) Mode
C) Variance
D) Mean

Answer: C) Variance

Which of the following is not a method used to measure the central tendency of data?

A) Mean
B) Median
C) Mode
D) Range
E) Variance

Correct answer: D) Range

Soru 1:

Which of the following statements about the arithmetic mean is TRUE?

A) It is always a better central tendency measure than the median, regardless of data distribution.
B) It is resistant to extreme values and outliers.
C) It can be distorted by a small number of very high or low values.
D) It can only be calculated for categorical attributes.
E) It is the average of the first and last values in a dataset.

Correct Answer: C) It can be distorted by a small number of very high or low values.

Question 2
Which of the following is NOT a property of ratio attributes?

A) Distinctness
B) Order
C) Addition
D) Multiplication
E) Zero is arbitrary

Correct Answer: E) Zero is arbitrary

Which of the following graphic displays is best suited to show five-number summaries of a dataset?

A) Scatter plot
B) Bar chart
C) Histogram
D) Boxplot
E) Quantile plot

Correct Answer: D) Boxplot

If a data distribution is positively skewed (right-skewed), which of the following is typically true regarding the relationship between mean, median, and mode?

A) Mean < Median < Mode
B) Mode < Median < Mean
C) Median < Mode < Mean
D) Mode = Median = Mean
E) Median > Mean > Mode

Answer: B)Mode < Median < Mean

What term is used to identify the most frequently occurring value in a data set?

A) Median
B) Mode
C) Mean
D) Interquartile Range

Answer: B) Mode

What is the median?

A) The most frequent value
B) The difference between max and min
C) The middle value
D) The average of all values
E) The square root of variance

Correct answer: C) The middle value

Soru 2:

Which of the following datasets has a standard deviation of zero?

A) {1, 2, 3, 4}
B) {5, 5, 5, 5}
C) {2, 4, 6, 8}
D) {0, 0, 1, 1}
E) {1, 2, 2, 3}

Correct Answer: B) {5, 5, 5, 5}

Which of the following best describes the standard deviation of a dataset?

A) The smallest value in a dataset
B) The average of the absolute differences from the mean
C) The middle value of a sorted dataset
D) The number of observations in a dataset
E) A measure that quantifies the amount of variation or dispersion of a set of data values around the mean, indicating how spread out the data points are

Correct answer: E

Which of the following statistics is sensitive to extreme values (outliers)?

A) Median
B) Mode
C) Mean
D) Interquartile Range
E) None of the above

Correct Answer: C) Mean

Which is not a measure of central tendency?

A) Mean, the average
B) Median, the middle value
C) Mode, most frequent
D) Midrange, average of extremes
E) Standard deviation, which measures spread not center

Which measure is least affected by outliers?

A) Mean
B) Mode
C) Median
D) Standard deviation
E) Range

Correct Answer: C) Median

Which of the following is not a method used to measure similarity between two data objects?

A) Cosine Similarity
B) Jaccard Coefficient
C) Simple Matching Coefficient
D) Euclidean Distance
E) Variance

Correct Answer: E) Variance

Two data points have a Euclidean distance of 0. What does this indicate about the points?
A) They are completely different
B) They lie on the same axis
C) They are identical in all dimensions
D) One of them is missing a value
E) They are from different datasets

Correct Answer: C

What does the five-number summary of a dataset include?

A) Only the maximum and minimum values
B) Only the three quartiles
C) Mean, mode, and standard deviation
D) Frequency, range, and variance
E) The minimum, first quartile (Q1), median (Q2), third quartile (Q3), and the maximum values that provide a quick summary of the distribution of a dataset

Correct answer: E

What is the main purpose of using a box plot in statistical analysis?

A) To display the frequency of data values
B) To show the relationship between two variables
C) To summarize the distribution, central value, and variability of data
D) To calculate the mean and standard deviation
E) To identify the correlation coefficient

Correct Answer: C) To summarize the distribution, central value, and variability of data

How is an outlier commonly detected using IQR?

A) If it’s below Q1
B) If it’s above Q3
C) If it's far from the average
D) If it’s more than max
E) If it’s 1.5×IQR below Q1 or above Q3

Which attribute type supports calculating a meaningful mean?

A) Nominal
B) Ordinal
C) Ratio
D) Binary
E) Categorical

Correct Answer: C) Ratio

Which of the following is more appropriate for measuring similarity between two objects with asymmetric binary attributes?

A) Simple Matching Coefficient
B) Euclidean Distance
C) Pearson Correlation
D) Jaccard Coefficient
E) Cosine Similarity

Correct Answer: D) Jaccard Coefficient

Which of the following is true about the Jaccard similarity coefficient?
A) It is used for numerical data only
B) It compares the frequency of words in documents
C) It measures similarity based on shared and total elements of sets
D) It always gives values greater than 1
E) It ignores all elements that are not common

Correct Answer: C

Question 1:
Which of the following is true regarding the Euclidean distance measure in data similarity?
A) It is used only for categorical data.
B) It is calculated by summing the absolute differences between feature values.
C) It is sensitive to the scale of the data.
D) It cannot be used for data with missing values.
E) It is a type of cosine similarity.

Answer: C) It is sensitive to the scale of the data.

Which of the following best defines cosine similarity?

A) A method that only works with binary attributes
B) A statistical measure used to evaluate clustering accuracy
C) The ratio of the Euclidean distances between two vectors
D) A dissimilarity measure used exclusively for nominal data
E) A similarity measure that computes the cosine of the angle between two non-zero vectors in a multi-dimensional space

Correct answer: E

Which of the following is a measure of dissimilarity?
A) Euclidean Distance
B) Cosine Similarity
C) Jaccard Similarity
D) Correlation
E) Matching Score

Answer: A) Euclidean Distance

Which of the following is true about the Minkowski distance?

A) It is only used for nominal data
B) It can only be used when r = 2
C) When r = 1, it corresponds to the Manhattan (City Block) distance
D) When r = 0, it becomes the Supremum (L_∞) distance
E) Negative values of r provide more sensitive measurements

Correct answer: C

Which of the following statements about similarity and dissimilarity measures is true?

A) Euclidean distance can only be used for binary attributes.
B) Jaccard coefficient is better suited for symmetric binary attributes.
C) Cosine similarity measures the angle between two vectors and ignores their magnitudes.
D) The value of similarity must always be greater than 1.
E) Dissimilarity measures cannot be used for document data.

Correct Answer: C) Cosine similarity measures the angle between two vectors and ignores their magnitudes.

Question 2:
Which similarity measure is commonly used for text data by comparing the frequency of terms in documents?
A) Jaccard similarity
B) Pearson correlation coefficient
C) Cosine similarity
D) Manhattan distance
E) Hamming distance

Answer: C) Cosine similarity

Which of the following is true about the Euclidean distance in measuring dissimilarity?

A) It is only applicable to categorical data
B) It ignores the magnitude of the data points
C) It is not affected by data scaling or normalization
D) It cannot be used when there are missing values in the dataset
E) It is a geometric distance measure that calculates the straight-line distance between two points in a multidimensional space and is sensitive to the scale of the attributes

Correct answer: E

Which of the following is used to compare two pieces of data?
A) Color
B) Distance
C) Shape
D) Size
E) Name

Answer: B) Distance

Which of the following is not used in calculating the Jaccard similarity for asymmetric binary attributes?

A) f11: The number of attributes where both vectors are 1
B) f10: The number of attributes where the first vector is 1 and the second is 0
C) f01: The number of attributes where the first vector is 0 and the second is 1
D) f00: The number of attributes where both vectors are 0
E) Total number of non-zero attribute matches

Correct answer: D

Which of the following statements correctly describes the difference between Simple
Matching
Coefficient (SMC) and Jaccard Coefficient?

A) SMC considers only 1-1 matches, while Jaccard considers 0-0 matches.
B) Jaccard ignores 0-0 matches and focuses on presence, whereas SMC counts both presence and
absence equally.
C) Both coefficients are suitable for continuous numerical data.
D) SMC is used only in textual data analysis.
E) Jaccard coefficient is always greater than SMC.

Correct Answer: B) Jaccard ignores 0-0 matches and focuses on presence, whereas SMC counts
both presence and absence equally.

Which of the following statements about similarity and dissimilarity measures is correct?

A) Dissimilarities are always in the range [0,1]
B) Similarity measures must obey the triangle inequality
C) Euclidean distance is a similarity measure
D) Similarity between two identical objects is 0
E) Dissimilarity is often referred to as distance

Correct Answer: E) Dissimilarity is often referred to as distance

1. Which of the following is NOT considered a common data quality issue in data mining?
A) Incomplete data
B) Noisy data
C) Redundant data
D) Consistent data
E) Inaccurate data

Correct Answer: D) Consistent data

Which of the following best describes the impact of inconsistent data on the data mining process?

A) It increases the interpretability of data for end-users.
B) It improves the generalization ability of machine learning models.
C) It leads to unreliable mining results due to conflicting or mismatched values.
D) It enhances the integration of data from heterogeneous sources.
E) It reduces the size of the dataset, making processing faster.

Correct Answer: C) It leads to unreliable mining results due to conflicting or mismatched values.

Which of the following statements correctly describes the role of data transformation in preprocessing?

A) It removes duplicate records from the dataset.
B) It integrates data from multiple sources into a single consistent view.
C) It modifies the data to improve its compatibility with specific data mining algorithms.
D) It identifies and replaces missing values with statistical estimates.
E) It reduces the volume of data by eliminating irrelevant attributes.

Correct Answer: C) It modifies the data to improve its compatibility with specific data mining algorithms.

Which of the following best describes the main goals of the data cleaning process in data preprocessing?

A) To create visualizations of the data using charts and graphs.
B) To reduce the number of attributes in a dataset for faster computation.
C) To identify and fix issues such as missing values, noisy data, and outliers that can negatively affect data quality.
D) To format the data into a database-friendly structure for storage.
E) To encrypt sensitive data for secure access and usage.

Correct Answer: C

Let two binary vectors be:
x = [1, 0, 0, 1]
y = [1, 1, 0, 0]

What is the Jaccard similarity coefficient between x and y?

A) 0.25
B) 0.33
C) 0.20
D) 0.50
E) 0.75

Correct Answer: B) 0.33

2. Which of the following is a major task in data preprocessing?
A) Model training
B) Data cleaning
C) Performance evaluation
D) Data visualization
E) Deployment

Correct Answer: B) Data cleaning

Which of the following preprocessing tasks is specifically aimed at reducing data volume without significantly losing analytical value?

A) Data Cleaning
B) Data Transformation
C) Data Integration
D) Data Reduction
E) Data Consistency Checking

Correct Answer: D) Data Reduction

Which of the following is a correct explanation of "noise" in the context of data quality?

A) Duplicate records collected from different data sources
B) Meaningful patterns that deviate from the main data trend
C) Random error or variance in a measured variable
D) Missing values caused by equipment failure
E) Inconsistencies in data due to outdated entries

Correct Answer: C) Random error or variance in a measured variable

Which data preprocessing step is primarily responsible when you need to clean the data by handling missing values, noisy data, and outliers to improve data quality?

A) Data integration
B) Data reduction
C) Data transformation
D) Data cleaning
E) Data discretization

Correct Answer: D

Which of the following is a common data preprocessing method to deal with "missing data", which is one of the factors affecting data quality?

A) Ignoring the missing data completely and continuing the analysis

B) Filling the missing data with random values

C) Filling the missing data with statistical methods such as mean, median or mode

D) Removing all columns with missing data from the data set

Correct Answer: C

Which of the following is NOT a factor that affects data quality?
A) Accuracy
B) Completeness
C) Timeliness
D) Believability
E) Performance

Correct Answer: E) Performance

Which of the following is a data smoothing technique used to handle noisy data?

A) Data encryption
B) File compression
C) Binning which smooths noisy data
D) Removing column names
E) Downloading cleaner data

Correct answer: C

Which of the following is NOT a typical step in the data cleaning process?

A) Handling missing values
B) Removing duplicate records
C) Encrypting sensitive information
D) Correcting inconsistent data formats
E) Detecting outliers

Correct Answer: C) Encrypting sensitive information

Which is a data cleaning step?

A) Model training
B) Data visualization
C) Filling missing data
D) Making predictions
E) Data storage

Correct Answer: C

Which of the following is not a typical task involved in data cleaning?

A) Handling missing values
B) Removing duplicates
C) Normalizing data types
D) Developing predictive models
E) Identifying outliers
Correct Answer: D) Developing predictive models

In the data preprocessing process, which of the following is one of the methods used to convert categorical data into numerical form? A) Standardization B) One-Hot Encoding C) Min-Max Normalization D) Missing Data Filling Correct Answer: B
Which of the following is NOT a major task in data preprocessing? A) Data Cleaning B) Data Integration C) Data Distribution D) Data Transformation and Discretization E) Data Reduction Correct Answer: C) Data Distribution
Real-world data is often considered "dirty". What does this mean? A) The data is clean and verified B) The data is always collected automatically C) The data contains missing, noisy, or inconsistent values D) The data is encrypted for security E) The data only includes images Correct answer: C
Which method is commonly used to handle missing data in a dataset? A) Ignoring the entire dataset B) Replacing missing values with random numbers C) Filling missing values using mean, median, or mode D) Encrypting the missing data E) Removing all non-missing values Correct Answer: C) Filling missing values using mean, median, or mode
Why are duplicate records removed? A) To free memory B) To increase speed C) To prevent inaccurate analysis D) To generate data E) To simplify charts Correct Answer: C
What is the main purpose of data cleaning in the data mining process? A) To increase the size of the dataset B) To improve the accuracy and quality of data C) To reduce the dimensionality of data D) To visualize trends and patterns E) To eliminate irrelevant or noisy data Correct Answer: B) To improve the accuracy and quality of data

What does data integration mainly involve?

a) Merging all data without changes.
b) Combining only recent data.
c) Cleaning errors but not merging.
d) Combining data from multiple sources and resolving inconsistencies.
e) Deleting duplicate records only.

Answer: d

Which of the following best describes the goal of using ETL (Extract, Transform, Load) processes in data integration?

A) To generate random datasets for machine learning
B) To manually edit data before it enters a database
C) To move data from source systems into a data warehouse by extracting it, transforming it into a suitable format, and loading it into the target system
D) To store unstructured data in spreadsheets
E) To replace all legacy systems with cloud storage

Answer: C

Which of the following is not a primary goal of data integration?

A) Combining data from different sources into a meaningful whole
B) Reducing data redundancy and inconsistencies
C) Directly interpreting the results of data mining
D) Improving data quality
E) Preparing data for advanced analytics ANSWER:C

Which of the following best captures the primary purpose of data integration?

A) To generate separate reports for each data source
B) To merge data from disparate sources while reducing redundancies and resolving inconsistencies
C) To segment data into smaller subsets for specialized analyses
D) To eliminate all data variations by standardizing the user interface
E) To create multiple copies of datasets for backup purposes

Answer: B

Which technique scales features to a fixed range, typically [0, 1]?
A) Standardization
B) Min-max scaling to a 0–1 range
C) Z-score norm
D) Log transform

Answer:
B) Min-max scaling to a 0–1 range

What is the key difference between equal-width and equal-frequency binning in discretization?

A) Equal-width uses class information, while equal-frequency does not.
B) Equal-width bins have the same number of values, while equal-frequency bins have the same interval size.
C) Equal-width bins have the same interval size, while equal-frequency bins have the same number of values.
D) Equal-frequency is supervised, while equal-width is unsupervised.
E) Equal-width is used only for nominal attributes.

Answer : C)

Which is not typically a problem in data integration?

a) Different formats across databases.
b) Duplicate entries in merged data.
c) Automatic consistency without need for transformation or cleaning.
d) Missing attribute values in some sources.
e) Conflicting IDs for the same entity.

Answer: c

What is the main reason organizations invest in data integration tools and technologies?

A) To isolate data into independent silos for department-specific use
B) To improve spreadsheet formatting across departments
C) To streamline access to accurate, consistent, and unified data from multiple sources, enabling better decision-making and operational efficiency
D) To limit data availability to only upper management
E) To eliminate the need for data analysts entirely

Answer: C

In correlation analysis for numeric data, what does a negative correlation coefficient (r<0) between two attributes A and B indicate?

A) A and B are independent of each other
B) As the values of A increase, the values of B also increase
C) As the values of A increase, the values of B tend to decrease
D) A and B are strongly and positively correlated
E) There is no relationship between A and B ANSWER:C

A chi-square test is performed on a contingency table comparing customer preferences and product categories. The test yields a high chi-square value. What does this indicate?

A) The attributes are definitely independent
B) The sample size is too small
C) The attributes are likely unrelated
D) The observed and expected counts differ significantly
E) One of the variables is numeric

Answer: D

Which method converts continuous variables into categories by grouping values into intervals?
A) Label encoding
B) One-hot encoding
C) Binning continuous values into intervals
D) Standard scaling

Answer:
C) Binning continuous values into intervals

In the sorted dataset [5, 10, 11, 13, 15, 35, 50, 55, 72, 89, 204, 215], which values are in the second bin after clustering-based discretization along the two biggest gaps?
A) 5, 10, 11, 13, 15
B) 35, 50, 55, 72, 89
C) 204, 215
D) 15, 35, 50, 55
E) 72, 89, 204
Answer : B) 35, 50, 55, 72, 89

Which of the following is not a method of data normalization?

A) Min-max normalization
B) Z-score normalization
C) Decimal scaling normalization
D) Data compression normalization
E) None of the above

Correct Answer: D)

Which of the following statements about concept hierarchies is TRUE?

A) They can be implicit within the database schema.
B) They may be manually provided by system users.
C) They can be automatically generated using statistical analysis.
D) They may form a total or partial order among attributes.
E) All of the above.

Answer:
E) All of the above.

Which method normalizes data when minimum and maximum values are unknown?

a) Min-max scaling of each value
b) Decimal scaling to reduce large numbers
c) Z-score normalization using mean and standard deviation
d) Clustering to adjust ranges
e) Manual adjustment of values

Answer: c

Which transformation method adjusts attribute values to a specific range and gives equal weight to attributes?

a) Aggregation of multiple attributes
b) Random scaling without standard rules
c) Normalization using min-max or z-score methods to adjust range and balance attribute influence
d) Manual selection of important records
e) Replacing missing values with zeros

Answer: c

Question 1
Which of the following is a dimensionality reduction technique?
A) Data sampling
B) Principal Component Analysis (PCA)
C) Data binning
D) Histogram analysis

Correct Answer: B) Principal Component Analysis (PCA)

What is Data Reduction, and which techniques can be used to reduce the size of data?
A) It is a process done solely by deleting data.
B) It is a process aimed at reducing the size of data by compressing it and selecting important information, helping to optimize data for more efficient analysis.
C) It is a process used to quickly analyze data without losing accuracy.
D) It is a process used to store all the necessary data for analysis.

Answer:
B)

Which of the following statements about data discretization is true?

A) Data discretization converts categorical data into numeric data.
B) Data discretization increases the number of distinct data values.
C) Equal-width binning divides data into intervals of the same size.
D) Clustering-based discretization ignores the closeness of values.
E) Discretization can only be done using supervised learning methods.

Correct Answer: C)

Which of the following is TRUE about the ChiMerge discretization method?

A) It is a supervised, bottom-up method that merges intervals with low χ^2 values.
B) It uses a top-down approach to split intervals based on class entropy.
C) It is unsupervised and relies on correlation measures for merging.
D) It starts by grouping values into predefined equal-width intervals.
E) It merges intervals with the highest χ^2 values to maximize dissimilarity.

Answer:
A) It is a supervised, bottom-up method that merges intervals with low χ^2 values.

Which discretization method merges adjacent intervals based on class similarity and statistical tests?

a) Equal-width binning without class labels
b) Clustering values without supervision
c) ChiMerge method using chi-square and merging intervals recursively
d) Decision tree splitting without checking classes
e) Random grouping of similar data points

Answer: c

Which method uses intervals and class labels to discretize continuous attributes effectively?

a) Random binning without analysis
b) Splitting based on fixed width only
c) Supervised discretization using decision trees and entropy to form pure groups
d) Grouping by highest values only
e) Replacing outliers and noisy data with averages

Answer: c

Which data reduction method selects the most important features?
A) Sampling
B) Sorting
C) Feature selection
D) Indexing

Correct Answer:
C) Feature selection

Data Reduction is most useful in addressing which of the following situations?
A) Detecting errors in a dataset.
B) Accelerating the analysis of large datasets by reducing the amount of data while retaining key features.
C) Increasing data security.
D) Making the dataset more reliable by preserving every single piece of data.

Answer:
B)

Which of the following is a data reduction technique?

A) Encoding
B) Normalization
C) Feature selection
D) Labeling
E) Building decision trees

Correct Answer: C

A data scientist is working with a 100-feature e-commerce dataset containing customer demographics, purchase frequency, and browsing behavior. To improve model efficiency, they want to reduce dimensionality while preserving 95% of the variance.

Which approach is MOST appropriate?
A) Apply PCA and select the top k principal components that explain 95% variance
B) Use correlation matrix to remove features with Pearson coefficient >0.8
C) Delete all categorical features one-hot encoded earlier
D) Randomly discard 50 features
E) None of them

Correct Answer: A

What are the two primary steps in the classification process using Decision Tree Induction, and how do they differ in terms of purpose and dataset usage?

A. Model training and model optimization; both use the same training set to improve accuracy.
B. Feature selection and label encoding; used interchangeably for model testing.
C. Model construction and model usage; construction uses training data to build a tree, usage applies it to test data to evaluate accuracy.
D. Data cleaning and attribute grouping; used for identifying the majority class in a dataset.
E. None of Above
Correct Answer: C

What is the difference between classification and prediction?
A) Classification predicts continuous values, while prediction uses categorical labels.
B) Classification predicts categorical (discrete, unordered) labels, while prediction models continuous-valued functions.
C) Classification and prediction both work on the same type of data and give different results.
D) Classification is used only for medical diagnosis, while prediction is used only for marketing.
E) Classification models continuous-valued functions, while prediction classifies data based on categories.

Correct answer: B

During the execution of the Decision Tree Induction algorithm, what happens if all tuples in the subset D being processed belong to the same class, and how does the algorithm proceed in that case?
A) A new test condition is created to ensure further splits occur, improving model depth.
B) The node is labeled with the majority class and passed to a pruning function to simplify the model.
C) The algorithm stops recursion and returns the node as a leaf node, labeled with the common class value of the tuples.
D) The algorithm chooses the next best attribute regardless of class uniformity to ensure all attributes are used.
E) The node is ignored as it adds no further information to the classification tree.

Correct Answer: C

Which of the following is NOT an advantage of decision tree-based classification according to the lecture?
A) Inexpensive to construct
B) Robust to noise
C) Handles complex relationships among continuous attributes effectively
D) Extremely fast at classifying unknown records
E) Can handle missing data efficiently

Answer: C) Handles complex relationships among continuous attributes effectively

What is the main goal of data reduction?

A) Delete data
B) Make the model complex
C) Reduce computational cost
D) Generate new data
E) Encrypt data

Correct Answer: C

An IoT system generates 10TB of daily temperature readings stored at 1-second intervals. To reduce storage costs while maintaining trend analysis capability:

Which technique is LEAST suitable?
A) Binning: Average readings into 5-minute intervals
B) Clustering: Replace raw data with cluster centroids using k-means
C) Regression: Fit a polynomial curve to original data and store only coefficients
D) sampling: Randomly delete 80% of readings
E) None of them

Correct Answer: D

Which attribute selection measure is most commonly used in decision tree algorithms like ID3, and what is a major drawback of using it?

A. Gini Index
B. Chi-square
C. Information Gain
D. Gain Ratio
E. None of Above
Correct Answer: C

How do decision trees work?
A) Decision trees perform an attribute test at each internal node, and each branch represents the outcome of that test, eventually providing the class label.
B) Decision trees classify only based on continuous attributes.
C) Decision trees give direct results without analyzing the data.
D) Decision trees make decisions based solely on the majority of examples.
E) Decision trees can handle both categorical and continuous attributes for classification.

Correct answer: A

What does the decision tree algorithm do when all instances in a node belong to the same class?
A) Split the node further
B) Assign the majority class label
C) Label the node as a leaf
D) Stop the tree construction
E) Prune the node

Correct Answer: C

What is the primary purpose of the "information gain" measure in decision tree algorithms?
A) To reduce the number of leaf nodes
B) To identify the attribute that best separates data into classes
C) To test the efficiency of classification algorithms
D) To minimize the height of the decision tree
E) To optimize data normalization techniques

Answer: B) To identify the attribute that best separates data into classes

Which of the following is a characteristic of decision tree algorithms?

A) Decision trees only work with continuous data types.
B) In a decision tree, each leaf node represents a class label.
C) Decision tree algorithms work directly on the test data without using training data.
D) Decision tree algorithms can only model linear relationships.

Correct answer:

B) In a decision tree, each leaf node represents a class label.

Which of the following is one of the most commonly used attribute selection measures in decision tree induction?

A) K-Means
B) Gini Index
C) Euclidean Distance
D) Apriori
E) Information Gain
Answer:
B: Gini Index

Which of the following statements best describes the difference between classification and prediction in data mining?

A) Classification assigns a value from a continuous range, while prediction assigns a discrete class.
B) Classification and prediction are the same process under different names.
C) Classification deals with categorical labels, whereas prediction deals with continuous-valued functions.
D) Prediction uses decision trees, but classification does not.
E) Prediction does not require training data, whereas classification does.

Correct Answer: C

Question 1:
What is the fundamental process involved in Classification?

a) Finding hidden patterns.
b) Predicting numerical trends.
c) Building a model from labeled training data to assign predefined categorical labels to new, unseen data instances.
d) Calculating data averages.
e) Organizing data into groups based on similarities.
Answer:
c) Building a model from labeled training data to assign predefined categorical labels to new, unseen data instances.

Question:

Which of the following statements is TRUE regarding classification and prediction in
data mining?

A) Classification models are typically used to predict continuous numerical values.
B) Prediction techniques are mainly applied to generate categorical outcomes.
C) Classification and prediction are unrelated processes used in different data mining domains.
D) Neural networks can be used in both classification and prediction tasks.
E) Decision trees are only applicable in prediction, not classification.

Correct Answer:
D) Neural networks can be used in both classification and prediction tasks.

When building a decision tree from a training dataset, which of the following is the first step?

A) Randomly select an attribute and start branching.
B) Apply a criterion to determine the best splitting attribute in the dataset.
C) Evaluate the model's accuracy on the test dataset.
D) Use all attributes at once to make a final decision in a single step.

Correct answer:

B) Apply a criterion to determine the best splitting attribute in the dataset.

Which of the following is an advantage of decision tree algorithms?

A) They require data normalization
B) They can only work with numerical data
C) Their results are easy for humans to interpret
D) They require a lot of data preprocessing
E) They can handle both numerical and categorical data

Answer:
C) Their results are easy for humans to interpret

In decision tree induction, which of the following is NOT a typical criterion or method used to evaluate the best attribute to split the data?

A) Information Gain
B) Gini Index
C) Misclassification Error
D) Euclidean Distance
E) Gain Ratio

Correct Answer: D

Question 2:
In the context of Decision Tree Induction as described, what does an internal node in the decision tree represent?

a) A final class label assigned after traversing the entire path of the tree structure.
b) The root point where the dataset is initially split for the very first time during training.
c) A decision-making point that evaluates a specific attribute by applying a test, and based on the outcome, directs the data down different branches of the tree.
d) The result that is obtained after testing an attribute, usually appearing at the leaf nodes.
e) A placeholder used to store temporary training information before pruning occurs.

Answer:
c) A decision-making point that evaluates a specific attribute by applying a test, and based on the outcome, directs the data down different branches of the tree.

Question:
Which of the following is TRUE about decision trees in data mining?

A) A decision tree classifies data instances by randomly assigning them to leaf nodes.
B) Every leaf node in a decision tree contains a test condition for an attribute.
C) The root node of a decision tree is always a classification result.
D) Each inner node in a decision tree represents a test on a specific attribute.
E) Branches in a decision tree connect only leaf nodes to the root.

Correct Answer:
D) Each inner node in a decision tree represents a test on a specific attribute.

Question 1:
Which of the following statements about overfitting in decision tree models is TRUE?

A) Overfitting occurs when the decision tree is too shallow and cannot capture the training data properly.
B) Overfitting means the model performs better on unseen data than on training data.
C) Overfitting occurs when a model fits the training data too well, including noise, and performs poorly on unseen data.
D) Overfitting can be completely avoided by using more attributes during training.

Correct answer: C

Which of the following best defines what sets Decision Trees apart from many other classification algorithms?

a) They are always the most accurate, outperforming Random Forests and Gradient Boosting.
b) They only use linear decision boundaries, limiting complex pattern modeling.
c) They are easy to interpret and visualize, clearly showing classification rules.
d) They are the most scalable, even faster than optimized SVMs.
e) They require feature scaling to perform well.

Answer: C

Which of the following is not a commonly used attribute selection measure in decision tree induction?

A) Information Gain
B) Gain Ratio
C) Gini Index
D) Mean Squared Error
E) Misclassification Error

Answer: D) Mean Squared Error

Which of the following is true about attribute selection measures in decision tree learning?

A) They randomly select an attribute to split the data.
B) The goal is to maximize partition impurity.
C) Information gain, gain ratio, and Gini index are common measures for splitting.
D) The attribute with the lowest score is chosen for splitting.
E) The splitting criterion doesn’t affect partition purity.

Answer:C

What is the goal of bagging?

a) Build one strong model.
b) Merge all data into one tree.
c) Train multiple models and combine their predictions for better stability.
d) Pick only the best attributes.
e) Use fewer data points for faster results.

Answer: c

2. What is the primary reason for using Information Gain in decision tree algorithms like ID3?

A) To reduce the size of the dataset
B) To identify the attribute that results in the largest decrease in impurity
C) To ensure the dataset is balanced before training
D) To normalize the attribute values
E) To cluster the data before splitting
Correct Answer: B)

According to the Minimum Description Length (MDL) principle, which tree model is preferred?

A) The tree with the lowest number of leaves, regardless of classification accuracy.
B) The tree that has the fewest misclassification errors, even if it is very large and complex.
C) The tree with the smallest combined cost of encoding the tree and its classification errors.
D) The tree that only uses attributes with high information gain.

Correct answer: C

In Decision Tree Induction, the process of finding the "best"
attribute to split on at each node primarily involves:

a) Maximizing the number of branches from the split.
b) Minimizing Gini impurity or maximizing information gain.
c) Selecting the most frequent attribute.
d) Creating equal-sized child nodes.
e) Picking the attribute with most unique values.

Answer: B

Which of the following is not a type of attribute used to determine test conditions in decision tree induction?

A) Binary
B) Nominal
C) Ordinal
D) Continuous
E) Relational

Answer: E) Relational

Which statement correctly describes the Gini Index in decision trees?

A) It measures correlation between attributes and the target.
B) It selects the attribute that maximizes child nodes' impurity.
C) Gain is the difference between parent impurity and child impurity.
D) It increases when a node is split, indicating a better split.
E) It’s only used in binary classification tasks.
Answer: C

What does a rule-based classifier use?

a) Random guesses.
b) Only decision trees.
c) If-then rules and conditions to predict the class label.
d) Nearest neighbor distance.
e) Linear functions only.

Answer: c

2. In decision tree learning, what is overfitting most likely caused by?

A) Using too small of a dataset
B) Using a very shallow tree
C) Pruning the tree too early
D) A tree that is too simple to capture the data patterns
E) A tree that is too complex and fits noise in the training data

Correct Answer: E

In decision tree learning, why do some algorithms such as C4.5 use Gain Ratio instead of directly using Information Gain for attribute selection?

A) Because Gain Ratio further increases the bias towards attributes with many distinct values.
B) Because Information Gain tends to favor attributes with many unique values, and Gain Ratio normalizes this bias.
C) Because Gain Ratio completely ignores the impurity of partitions.
D) Because Gain Ratio only applies to continuous-valued attributes.
E) Because Information Gain cannot handle missing values in data.

Correct Answer: B

When building a decision tree, how is the best split point determined for a continuous attribute?

A. By randomly selecting a split value between the minimum and maximum of the attribute.
B. By checking only the median value and comparing its information gain.
C. By sorting the values, calculating all possible midpoints between consecutive values,
computing the information gain for each, and choosing the one with the highest gain.
D. By splitting the data into equal-sized bins and choosing the first bin boundary as the split point.
E. Based on the number of missing values only.
Correct Answer: C

In Decision Tree algorithms, which of the following is commonly used to select the best attribute for splitting the data?

A) Random Sampling
B) Entropy and Information Gain
C) Backpropagation
D) Cross-validation
E) Gradient Descent

Answer: B) Entropy and Information Gain

What does a Random Forest classifier use at each split to build better and more diverse trees?

a) Full dataset without any change.
b) Random samples but fixed attribute set.
c) Random selection of attributes and random linear combinations to reduce correlation among trees.
d) Only the top attributes based on frequency.
e) Combining identical trees for better accuracy.

Answer C

Which of the following metrics is commonly used to split nodes in a classification decision tree?

A) R-squared
B) Mean squared error
C) Gini index
D) Pearson correlation
E) Z-score

Correct Answer : C) Gini index

Which of the following statements about pruning in decision trees is TRUE?

A) Pruning reduces the training time of the decision tree.
B) Pruning increases the risk of overfitting to the training data.
C) Pruning is used to remove branches that do not contribute to improved accuracy on unseen data.
D) Pruning is only applied after testing the model on the test set.
E) Pruning prevents the use of categorical attributes in the tree.

Correct Answer: C) Pruning is used to remove branches that do not contribute to improved accuracy on unseen data.

Which of the following conditions indicates the best stopping criterion during the growth phase of a decision tree (pre-pruning stage)?

A) The entropy of the node increases after a potential split.
B) The Gini index of the node is greater than 0.5.
C) All records at the node share the same class label.
D) There are more attributes than training records.
E) Information Gain becomes negative after a split.

Correct Answer: C

How are nominal and ordinal attributes handled in decision trees?

A. They are grouped randomly.
B. Nominals are split by frequency, ordinals into two equal groups.
C. Nominals are split without order, ordinals maintain order and are split into ordered groups.
D. They are converted to numerical format.
E. They are ignored.

Correct Answer: C

When building a decision tree using algorithms like ID3, a common strategy is to choose the attribute that provides the most significant reduction in uncertainty about the class labels after the split. Which metric quantifies this reduction, essentially measuring the gain in information obtained by using a particular attribute to partition the data?
A) Gini Impurity
B) Classification Error
C) Information Gain
D) Entropy
E) Pruning Factor
Answer: The correct answer is C) Information Gain.

What does boosting improve?

a) Single weak model.
b) Only training speed.
c) Misclassified instances and model accuracy through multiple rounds.
d) Tree depth and size.
e) Data storage only.

Answer: c

In decision trees, what does entropy measure?

A. The probability of misclassification
B. The number of features in a dataset
C. The amount of randomness or impurity in the dataset
D. The depth of the tree
E. The distance between data points

Correct Answer: C. The amount of randomness or impurity in the dataset

How does the Gain Ratio improve upon the Information Gain measure?

A) By choosing attributes with the smallest number of values
B) By removing the need to calculate entropy
C) By normalizing Information Gain to penalize attributes with many distinct values
D) By preferring numerical attributes over categorical ones
E) By eliminating the need for attribute selection

Correct Answer: C) By normalizing Information Gain to penalize attributes with many distinct values

What does "overfitting" mean in the context of data mining?
A) When a model is too simple and has high training and test errors
B) When a model fits the training data very well but generalizes to unseen data poorly
C) When the training error is larger than the test error
D) When a model is balanced between complexity and simplicity
E) When preprocessing takes too much time
Correct Answer: B) When a model fits the training data very well but generalizes to unseen data poorly

Which of the following is a key reason why decision tree induction is popular in classification tasks?

A) It always produces the most accurate model compared to other methods
B) It requires no preprocessing of the data
C) It offers fast learning speed and easy-to-understand classification rules(Correct Answer is C)
D) It is only suitable for small datasets and cannot scale
E) It guarantees 100% accuracy on unseen data

Question 1:
Which criterion is most commonly used to split data at a node in decision trees?

A) Calculating the average
B) Entropy and Information Gain
C) Finding the median
D) Random selection
E)Principle Component Analysis
Correct Answer: B) Entropy and Information

Which principle suggests preferring a simpler model when two models have similar generalization errors?

A) Bias-Variance Tradeoff
B) Occam’s Razor
C) Overfitting Principle
D) Cross-Validation
E) Regularization
Answer: B

According to Han and Kamber, which attribute selection measure is biased toward multivalued attributes in decision tree induction?

A. Gini Index
B. Gain Ratio
C. Chi-Square
D. Information Gain
E. None of the above
Answer D

Which of the following best describes overfitting in decision trees?

A) The model has high error rates on both the training and test datasets.
B) The model fits the training data very well but performs poorly on the test data.
C) The model shows low performance on the training data but performs excellently on the test data.
D) The model performs the classification task randomly.
E) All nodes of the model are split based on the same attribute.
CORRECT ANSWER:C

Which of the following factors can cause overfitting in a model?
A) Too many training examples
B) Noise in the data
C) Using simple models
D) Underfitting
E) Having too few features
Correct Answer: B) Noise in the data

What is the main goal of post-pruning in decision tree induction?

A) To reduce training time by stopping early during tree construction
B) To increase the depth of the decision tree for better accuracy
C) To simplify the tree by removing nodes that do not improve generalization
D) To assign random class labels to leaf nodes for variety in predictions
E) To make the decision tree perfectly fit the training data
Correct Answer:
C) To simplify the tree by removing nodes that do not improve generalization

Question 2:
Which of the following is not an advantage of decision trees?

A) Easy to interpret
B) Good data visualization
C) Not prone to overfitting
D) Can work with both numerical and categorical data
E) Requires little data preprocessing
Correct Answer: C) Not prone to overfitting

What does "training error" measure?

A) Model performance on unseen data
B) Model’s fit on the training data
C) Error on the validation set
D) Generalization ability of the model
E) Computational complexity of the model
Answer: B

In the context of decision tree pruning as described by Han and Kamber, reduced-error pruning works by:

A. Replacing each node with the most frequent class and keeping it if accuracy improves on the validation set
B. Removing all leaf nodes and testing again on the training set
C. Splitting each node recursively until no further information gain is achieved
D. Assigning weights to leaves based on entropy
E. Calculating Gini Index on each branch after pruning
Answer A

Which of the following best explains the Minimum Description Length (MDL) principle in the context of decision tree learning?

A) MDL aims to select the model that has the highest number of attributes to avoid underfitting.
B) MDL favors models with the lowest training error, regardless of model size.
C) MDL chooses the model that minimizes the total cost of encoding the model and the errors it makes.
D) MDL always prefers simpler models, even if their classification performance is significantly worse.
E) MDL only considers the number of leaf nodes in a decision tree to determine its quality.

Correct Answer: C

In the context of encoding a decision tree model, what is the cost of encoding each internal node if there are m attributes?
A) $\log_2 k$ bits
B) $\log_2 n$ bits
C) m bits
D) $\log_2 m$ bits
E) m + k bits
Answer: D

Which of the following attribute selection measures is based on the χ^2 (chi-square) test?
a) Information Gain
b) Gini Index
c) CHAID
d) Gain Ratio
e) MDL
Correct Answer: c

1. What is the main difference between classification and prediction in data mining?
A) Classification uses continuous values, while prediction uses categories.
B) Classification is only used in image recognition.
C) Classification predicts categorical labels, while prediction models continuous-valued functions.
D) Prediction is used for decision trees only.
E) Classification always produces numeric outputs.
Correct Answer: C

What is a decision tree and how does it work? A decision tree is:

A) A data storage system used for organizing large datasets
B) A neural network model used for deep learning tasks
C) A model that classifies data by testing attributes at each node and assigning a label at the leaves
D) A linear regression model used for predicting continuous values
E) A clustering algorithm used to group similar data points

Correct Answer: C

Which of the following is a key reason for the popularity of decision tree induction in classification tasks?
A) Decision trees always produce the most accurate model
B) They require a lot of computational resources
C) They are slow to classify new records
D) They can be easily converted into simple and understandable classification rules
E) They only work with categorical attributes
Answer: D

Which pruning technique involves growing the tree fully and then trimming it?
a) Pre-pruning
b) Prune-first method
c) Random pruning
d) Split-and-stop
e) Post-pruning
Correct Answer: e

2. In a decision tree, what does each internal node represent?
A) A final decision or class label
B) A mathematical formula
C) A test on an attribute
D) A random guess
E) A path to the root
Correct Answer: C

What does it mean for a decision tree to overfit the data, and how can overfitting be prevented?
A) Overfitting means the model is too simple; it can be fixed by removing pruning
B) Overfitting occurs when the model performs well on both training and test data
C) Overfitting happens when a model memorizes training data and fails on new data; pruning methods can prevent it
D) Overfitting is useful because it reduces the error on new data
E) Overfitting happens when the model ignores the training data and fails to recognize patterns

Correct Answer: C