

# Final exam questions for the course Data Mining

## 1. Introduction to Data Mining

1. Describe what Data Mining is and how it differs from traditional statistical processing.
2. Provide specific examples of how data mining methods are applied in various industries (marketing, healthcare, finance, etc.) and explain the benefits.

## 2. Types of Data and Data Quality

3. **Classification of Data Types.** What are the main data types (nominal, ordinal, interval, ratio), and how do they differ in analysis?
4. **Datasets: Definition and Features.** What is the difference between two-dimensional and multidimensional datasets? What additional challenges arise when working with high-dimensional data?
5. **Data Quality: Main Criteria.** Which aspects (accuracy, completeness, consistency, etc.) define data quality, and why is it important?
6. **Methods of Measuring Data Quality**  
Which metrics or methods can be used to assess data quality in a real project (for example, percentage of missing values, duplicates, etc.)?
7. **Overview of Data Preprocessing Methods**  
Which tasks (cleaning, integration, transformation, reduction) are addressed at the preprocessing stage, and why are these tasks important?

## 3. Data Preprocessing

8. **Data Cleaning Methods**  
Describe several approaches to handling missing and duplicate data (removal, filling with mean values, using reconstruction algorithms).
9. **Transformation and Normalization**  
Explain the difference between normalization (Min-Max, Z-Score) and standardization of data. When and why are they applied?
10. **Discretization**  
What is discretization, and how does it help when using classification and association algorithms?
11. **Binarization**  
In which cases is feature binarization required? Give real-world examples where it is appropriate.
12. **Choosing Optimal Preprocessing Methods**  
Which factors influence the choice of preprocessing methods for a particular project (e.g., data type, algorithms used, quality metrics)?

## 4. Exploratory Data Analysis (EDA)

**13. Purpose of Exploratory Data Analysis**

Describe the main objectives of EDA and how it differs from final modeling stages.

**14. Graphical Visualization Methods**

Name three or four common graphical methods (histograms, boxplots, scatter plots, etc.) and describe how they help understand data structure.

**15. Outlier Detection**

How do visualization and statistical methods (e.g., boxplot) help detect outliers and anomalies in the data?

**16. Correlation Analysis**

How can correlation matrices and scatter plot matrices be used to understand relationships among features?

**17. Conclusions from EDA**

How do EDA results influence the choice of further modeling methods and adjustments to initial hypotheses?

## **5. Classification: Basic Concepts and Methods**

**18. Concept of Supervised Learning**

Explain what is meant by supervised learning. What is the main goal of this approach?

**19. Typical Classification Tasks**

Provide examples of typical classification tasks and explain why they are relevant.

**20. Main Classification Methods**

List the primary classic classification algorithms (decision trees, k-NN, etc.) and briefly characterize their features.

**21. Importance of Training/Test Splits**

How should the data be correctly split into training and testing sets when building classification models?

## **6. Decision Tree Algorithm**

**22. Principle of Decision Tree Construction**

How does a Decision Tree algorithm work? What is the essence of step-by-step splitting by features?

**23. Splitting Criteria**

Explain how the Gini index and entropy are calculated. What is the fundamental difference in how they evaluate “impurity”?

**24. Pruning a Tree**

What is tree pruning (pre-pruning and post-pruning), and how does it affect overfitting?

## **7. Model Evaluation and Performance Metrics**

**25. Basic Performance Metrics**

What do accuracy, precision, recall, and F1-score mean? How are they calculated?

**26. Interpreting the ROC Curve**

How is the ROC curve plotted, and what does the area under the ROC curve (AUC) signify?

**27. Cross-Validation Methods**

What is k-fold cross-validation, and why is cross-validation important for a more accurate model assessment?

**28. Balancing Different Metrics**

How and why does the choice of metric depend on the specifics of the task (e.g., medical diagnostics vs. spam filtering)?

**29. Issues with Metric Selection**

What problems may arise if you choose only one metric (e.g., accuracy) and ignore others?

## **8. Rule-Based Classifier**

**30. Algorithms for Generating Rules**

In general terms, how are rules formed for rule-based classifiers (pattern search, greedy algorithms, etc.)?

**31. Extracting Rules from Decision Trees**

How can a decision tree be converted into a set of rules, and why would we do that?

## **9. Naïve Bayes Classifier**

**32. Bayes' Theorem and Conditional Probability**

How is Bayes' formula used in the Naïve Bayes classifier?

**33. The 'Naïve' Assumption**

What is the "naïve" assumption in this algorithm, and why is it sometimes highly effective?

**34. Different Naïve Bayes Models**

What variants of Naïve Bayes classifiers (Gaussian, Multinomial, etc.) exist, and where are they applied?

**35. Practical Uses of Naïve Bayes**

Give a real-world example of a task (e.g., spam filtering) where Naïve Bayes yields good results.

**36. Limitations and Drawbacks**

In which cases can Naïve Bayes produce incorrect results, and why?

## **10. k-Nearest Neighbors (k-NN) Classification**

**37. Basic Idea of k-NN**

What is the principle behind k-nearest neighbors in a classification task?

**38. Choice of Distance Metric**

Which distance metrics (Euclidean, Manhattan, etc.) are most commonly applied in k-NN, and what are their features?

**39. Selecting the Optimal k**

How does the choice of the k parameter affect classification quality, and how is it selected?

**40. Scaling Issues**

Why is it important to normalize features before using k-NN, and how does this affect results?

**41. Advantages and Disadvantages of k-NN**

Explain in which situations k-NN is effective and where it is not, considering both computational complexity and accuracy.

## **11. Artificial Neural Networks**

**42. Perceptron Model**

How is a perceptron structured in its simplest form, and how is learning carried out by adjusting weights?

**43. Multilayer Neural Network**

What is a multilayer perceptron (MLP), and what role do hidden layers play in learning?

**44. Backpropagation**

Explain how the backpropagation algorithm is implemented and why it is effective for training deep networks.

**45. Role of the Activation Function**

Why are activation functions (Sigmoid, ReLU, Tanh, etc.) needed, and how does their choice affect training?

**46. Trends in Deep Learning**

Name modern directions in deep neural networks (ResNet, Transformers, etc.) and briefly describe one of them.

## **12. Support Vector Machines (SVM)**

**47. Main Idea of SVM**

What is the principle of classification in Support Vector Machines in terms of maximizing the margin?

**48. Kernel Methods (Kernel Trick)**

What is a “kernel” in SVM, and how does it make it possible to solve nonlinear classification problems?

**49. Types of Kernels in SVM.** List common kernel types (linear, polynomial, RBF, etc.) and explain how to choose the right one for a specific task.

**50.** Explain linear and nonlinear SVM (vector machine) algorithms and their differences.

Give an example of their application in real life tasks.

## **13. Ensemble Methods**

**51. The Ensemble Idea**

What are ensemble methods, and why can combining several weak models lead to higher accuracy?

**52. Bagging**

How does the bagging method work, and why is Random Forest considered one of its best examples?

**53. Boosting**

How is boosting fundamentally different from bagging? How are example “weights” updated during training?

**54. Random Forest**

Describe the main ideas behind Random Forest. Which parameters are particularly important in training (number of trees, tree depth, etc.)?

**55. AdaBoost**

How is the final composition formed in AdaBoost, and what is special about the way classification example weights are updated?

## **14. The Class Imbalance Problem**

**56. Class Imbalance**

What is meant by class imbalance, and why does it complicate classification?

**57. Approaches to the Solution**

List several approaches to dealing with imbalanced data (oversampling, undersampling, SMOTE, etc.) and briefly describe them.

**58. Practical Examples**

Give an example from a real field (e.g., fraud detection) where class imbalance is significant. How is the problem solved there?

## **15. Introduction to Clustering**

**59. Unsupervised Learning**

How does clustering differ from classification (supervised learning)? Provide conceptual examples.

**60. Clustering Tasks and Goals**

Why do we need to group objects into clusters? Name at least three practical applications where this is useful.

**61. Different Clustering Approaches**

What are the main types of clustering (partitioning (centroid based), hierarchical, density-based)? How do they differ conceptually?

**62. “Quality” Criteria for Clusters**

Which criteria and metrics (internal, external, silhouette, etc.) are used to determine whether objects are well-separated into clusters?

## **16. K-Means Clustering**

**63. K-Means Algorithm**

How does the k-means algorithm work step by step, from initializing centroids to stopping?

**64. Choosing the Number of Clusters k**

Which methods (elbow method, silhouette score) help determine the optimal number of clusters?

**65. Complexity and Convergence**

What is the computational complexity of k-means, and what factors can impede fast convergence?

**66. Distance Measures in K-Means**

Why is the Euclidean distance often used? Can another metric be applied, and how would it affect the algorithm?

**67. Limitations of K-Means**

In which cases does k-means perform poorly (e.g., non-spherical clusters), and what are the alternatives?

## **17. Hierarchical Clustering**

**68. Agglomerative vs. Divisive**

Describe the differences between agglomerative and divisive hierarchical clustering. How is the merging/splitting strategy chosen?

**69. Linkage Methods**

What is the difference between single, complete, and average linkage? How does this choice affect the shape of clusters?

**70. Dendrogram Construction**

How do you interpret a dendrogram, and at what point do you decide on the number of clusters?

**71. Sensitivity to Noise**

How do hierarchical methods react to noisy points? Are there any built-in “protective” mechanisms?

**72. Comparison with K-Means**

In which situations can hierarchical clustering provide a better solution than k-means?

## **18. Density-Based Clustering**

**73. DBSCAN Principle**

How does DBSCAN work? Explain the concepts of “eps,” “minPts,” “core points,” “border points,” and “noise points.”

**74. Noise and Outlier Detection**

Why can DBSCAN separate outliers as individual points, and why is that an advantage over k-means?

**75. Density Reachability**

How is the idea of density reachability and density connectivity formalized?

**76. Choosing DBSCAN Parameters**

How do you choose the eps and minPts parameters in practice? Name at least two methods or heuristics.

**77. Strengths and Weaknesses of DBSCAN**

In which cases is DBSCAN more effective than other methods, and what are its limitations?

## **19. Basic Concepts of Association Analysis**

**78. Frequent Itemsets**

What are frequent itemsets? Provide an example from retail (market basket analysis).

**79. Apriori Algorithm**

How does the Apriori algorithm work? What is the main property it uses to reduce the search space?

**80. Rule Evaluation Metrics**

Explain the meaning of support, confidence, and lift in association rule analysis.

**81. Applications of Association Rules**

Give examples of how association rules are used in industry (recommendation systems, cross-selling, etc.).

**82. Limitations and Extensions**

What are the limitations of the classic Apriori algorithm, and what improvements (FP-Growth, Eclat) exist?

## **20. Summarizing and Additional Questions**

**83. Choosing an Algorithm**

Which key factors (data size, task type, available resources, etc.) should be considered when selecting a data mining algorithm?

**84. Comparison of Supervised and Unsupervised Learning**

Give examples of situations where supervised methods are preferable and where unsupervised methods should be used.

**85. Combining Multiple Techniques**

Is it possible to combine classification and clustering methods in one project? Provide an example of such an approach and explain its benefits.

## **21. Practical tasks**

**86. Practical Assignment:** Design a Decision Tree model based on the following Data Set.  
**Data set:**

| ID | Age   | Income | Student | Credit Rating | Buys Computer |
|----|-------|--------|---------|---------------|---------------|
| 1  | <= 30 | High   | No.     | Fair          | No.           |
| 2  | <= 30 | High   | No.     | Excellent     | No.           |
| 3  | 31–40 | High   | No.     | Fair          | Yes           |
| 4  | > 40  | Medium | No.     | Fair          | Yes           |
| 5  | > 40  | Low    | Yes     | Fair          | Yes           |

87. **Practical assignment:** Perform hierarchical clustering (MIN, MAX, Average, Centroid) based on the following data set (in graphical form) and explain the method of building a model based on hierarchical clustering.

**Data set:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.0 | 1.5 |
| 2  | 1.5 | 1.8 |
| 3  | 5.0 | 8.0 |
| 4  | 8.0 | 8.0 |
| 5  | 1.0 | 0.6 |

88. Design a Random Forest model structure based on the following data set and explain the method of building the model.

**Data set:**

| ID | Feature1 | Feature2 | Label |
|----|----------|----------|-------|
| 1  | 2.5      | 0.5      | 0     |
| 2  | 3.0      | 1.5      | 1     |
| 3  | 2.0      | 2.0      | 0     |
| 4  | 3.5      | 3.0      | 1     |
| 5  | 1.5      | 0.5      | 0     |

89. **Practical Assignment:** Design a decision tree based on the following data set and explain how to build a model based on a decision tree.

**Data set:**

| ID | Gender | Age | Income | Buy |
|----|--------|-----|--------|-----|
| 1  | Male   | 25  | High   | Yes |
| 2  | Female | 30  | Medium | No. |
| 3  | Male   | 35  | Low    | No. |
| 4  | Female | 45  | High   | Yes |
| 5  | Male   | 40  | Medium | Yes |

90. **Practical assignment:** Perform hierarchical clustering (MIN, MAX, Average, Centroid) based on the following data set and explain the method for building the model.

**Data set:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.0 | 2.0 |
| 2  | 2.5 | 4.5 |
| 3  | 3.0 | 6.0 |
| 4  | 8.0 | 8.5 |
| 5  | 1.5 | 0.8 |

91. Design a model structure to build the Random Forest algorithm based on the following dataset and explain the method of building the model.

**Data set:**



| ID | Feature1 | Feature2 | Label |
|----|----------|----------|-------|
| 1  | 1.0      | 3.5      | 0     |
| 2  | 2.0      | 2.0      | 1     |
| 3  | 3.5      | 1.5      | 1     |
| 4  | 4.0      | 3.0      | 0     |
| 5  | 5.0      | 2.5      | 1     |

92. **Practical Assignment:** Design a decision tree based on the following data set and explain the method of building the model

**Data set:**

| ID | Job type   | Experience | Salary | Promotion |
|----|------------|------------|--------|-----------|
| 1  | Manager    | 5          | High   | Yes       |
| 2  | Engineer   | 3          | Medium | No.       |
| 3  | Technician | 2          | Low    | No.       |
| 4  | Engineer   | 8          | High   | Yes       |
| 5  | Manager    | 10         | High   | Yes       |

93. **Practical Assignment:** Perform hierarchical clustering based on the following dataset and explain the method of building the model.

**Data set:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.2 | 0.9 |
| 2  | 2.8 | 3.4 |
| 3  | 3.0 | 5.0 |
| 4  | 6.0 | 7.5 |
| 5  | 0.5 | 1.0 |

94. **Practical assignment:** Build a decision tree model based on the following data set and explain the method for building the model.

**Data set:**

| ID | Exam1 | Exam2 | Passed |
|----|-------|-------|--------|
| 1  | 25    | 20000 | No.    |
| 2  | 34    | 45000 | Yes    |
| 3  | 29    | 32000 | No.    |
| 4  | 45    | 50000 | Yes    |
| 5  | 39    | 30000 | No.    |

95. **Practical assignment:** Design a Random Forest algorithm based on the following data set and explain the method for building the model.

**Data set:**

| ID | Exam1 | Exam2 | Passed |
|----|-------|-------|--------|
|----|-------|-------|--------|

|   |    |    |     |
|---|----|----|-----|
| 1 | 85 | 78 | Yes |
| 2 | 52 | 61 | No. |
| 3 | 89 | 92 | Yes |
| 4 | 55 | 45 | No. |
| 5 | 72 | 81 | Yes |

96. **Practical Assignment:** Perform hierarchical clustering (MIN, MAX, Average, Centroid) based on the following Dataset and explain the method of building the model.

**Data set:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.2 | 2.3 |
| 2  | 1.5 | 2.7 |
| 3  | 5.0 | 7.0 |
| 4  | 7.8 | 8.1 |
| 5  | 2.0 | 2.9 |

97. **Practical Assignment:** Design a Random Forest algorithm based on the following Dataset and explain the method of building the model.

**Data set:**

| ID | Feature1 | Feature1 | Feature1 | Target |
|----|----------|----------|----------|--------|
| 1  | 3.0      | 2.5      | 1.8      | 0      |
| 2  | 2.8      | 3.5      | 1.5      | 1      |
| 3  | 4.0      | 4.0      | 2.5      | 1      |
| 4  | 3.2      | 2.8      | 1.0      | 0      |
| 5  | 4.5      | 4.1      | 2.8      | 1      |

98. **Practical assignment:** Construct a distance matrix for hierarchical clustering based on the following Dataset, indicate the method of linkage (single, complete) and explain the method of building the model.

**Dataset:**

| ID | X | Y |
|----|---|---|
| 1  | 2 | 4 |
| 2  | 6 | 8 |
| 3  | 4 | 2 |
| 4  | 8 | 6 |
| 5  | 5 | 7 |

99. **Practical Assignment:** Design a model to build the Random Forest algorithm based on the following dataset and explain the method of building the model.

**Data set:**

| ID | Feature1 | Feature2 | Label |
|----|----------|----------|-------|
| 1  | 1.5      | 0.5      | 0     |
| 2  | 3.0      | 1.5      | 1     |

|   |     |     |   |
|---|-----|-----|---|
| 3 | 2.0 | 2.0 | 0 |
| 4 | 4.0 | 3.0 | 1 |
| 5 | 5.0 | 4.5 | 1 |

100. **Practical Assignment:** Perform hierarchical clustering based on the following dataset and explain the method of building the model.

**Data set:**

| ID | X   | Y   |
|----|-----|-----|
| 1  | 1.2 | 0.9 |
| 2  | 2.8 | 3.4 |
| 3  | 3.0 | 5.0 |
| 4  | 6.0 | 7.5 |
| 5  | 0.5 | 1.0 |

-----

### **Important Announcement for Students!**

The use of mobile phones, tablets, laptops, cheat sheets, or any other auxiliary devices during the final examination is strictly prohibited. Any student found using or attempting to use such devices will be immediately disqualified from the exam.

To ensure success, please prepare thoroughly for the exam in advance.

For clarification of any exam-related questions, a consultation session will be held on **January 13, 2025, at 1:00 PM** in room **F-704**. All students are welcome to attend.