**Map of test tasks**

**Competence:** PC-2 Capable of using skills of working with big data, planning, forecasting and modeling of transport processes and systems in professional activity

**Indicator:** PC-2.1 Analyzes the volume and quality of data, methods of their processing in the operation of transport systems

**Discipline**: Intelligent Data Analysis in Transportation

**Test Description:**

1. The test consists of 70 tasks, which check the level of mastering the competencies of the student. When testing each student is offered 30 test tasks of 15 open and closed types of different levels of complexity.

2. For the correct answer of the test task the student receives 1 conditional point, for an incorrect answer - 0 points. At the end of the test, the system automatically determines the "earned final score" for the test, according to the evaluation criteria

3 The maximum total sum of points for all correct answers is - 100 points.

4. the test is successfully passed if the student correctly answered 70% of the test tasks (61 points).

5. To pass the test, including the organizational moment, students have no more than 60 minutes. For each test task on average 2 minutes.

6. The learner is given one attempt to pass the computerized test.

The **codifier of the** discipline test is the section of the working program "4. Structure and content of the discipline (module)"

**Set of test tasks**

**Closed-ended tasks**

**Alternative choice tasks**

*Select* ***one*** *correct answer*

**Simple (Level 1)**

1. ANOVA is ...

A) analysis of variance

B) regression analysis

B) factor analysis

1. The procedure for selecting the number and conditions of experiments necessary and sufficient to solve a given problem with the required accuracy is ....

(A) Planning the experiment

B) decision-making

B) risk assessment

1. Exploring the subject area, collecting data, pre-processing the data, analyzing the data, interpreting the patterns found, and using the new knowledge are the steps ...

(A) Data mining

B) making a long-term forecast

B) decision-making

1. Estimating the regression model coefficients and correlation coefficient, testing the hypothesis that the coefficients are significant, and testing the hypothesis that the model is adequate are the main steps ...

A) regression analysis

B) analysis of variance

B) spectral analysis

1. Calculating the correlation matrix for all variables involved in the analysis, extracting factors, selecting factors and rotating factors to create a simplified structure, and interpreting factors are the steps ...

A) factor analysis

B) regression analysis

B) analysis of variance

**Medium-complex (level 2)**

1. Detecting errors, removing errors, and detecting inconsistencies in data are processes...

(A) Data cleansing

B) data transformation

B) data storage

1. A data integration system is a ...

(A) Combining data from different sources and making the data available to users in a unified form

B) sampling of data from different sources

B) sorting data from different sources

1. Data integration creates a new single source of information that is a ...

(A) Information resources of the entire new integrated set of sources

B) a resource that provides a single unified interface

B) a common resource that combines data from a group of sources

1. The content of the data sources should be:

(A) Unchanged or updated

B) not updated

B) changeable or not updatable

1. By analyzing the ratio of mean, mode and median we can conclude that ...

A) data symmetries

B) data cleanliness

B) correlations

1. A method in mathematical statistics aimed at finding relationships in experimental data by examining the significance of differences in mean values is ....

A) analysis of variance

B) factor analysis

B) regression analysis

1. Tasks solved by data mining methods are usually divided into...

(A) Descriptive and predictive

B) statistical and optimization

B) preparation and sorting

1. To process the results of the experiment, the method of...

A) least squares

B) Runge-Kutta

B) Monte Carlo

1. Logistic .... is used in classification tasks

A) regression

B) progression

B) optimization

1. Logistic regression is used for prediction if the response ...

A) binary

B) negative

B) whole

1. A statistical relationship between two random variables is shown by ...

A) correlation coefficient

B) dispersion

B) Fisher's coefficient

1. The adequacy of a regression model can be assessed using ...

A) Fisher's coefficient

B) correlation coefficient

B) dispersions

1. The data under study are measured in nominal scale units. Unable to determine:

A) median

B) fashion

B) frequency distribution

1. The data under study are measured in ordinal scale units. It is impossible to determine:

(A) Average

B) fashion

B) frequency distribution

1. The mode, median, and mean for the set of data under study are the same. This means that the distribution of this random variable ...

A) symmetrically

B) discrete

B) asymmetrically

1. Calculating the adequacy variance requires knowing the ...

A) calculated and experimental values of the response function

B) results of observations in parallel experiments

B) number of degrees of freedom

1. The inversion criterion is used to test the hypothesis that the

A) the presence of a trend

B) equality of variance

B) reliability of data

**Complex (Level 3)**

1. Calculating the variance of reproducibility requires knowing the ...

A) results of observations in parallel experiments

B) the results of observations in one experiment

B) values of factors in parallel experiments

1. The data set under study has fields (columns) with information about the time of fare payment; amount of payment; method of payment (bank card transportation card, other); coordinates of vehicle geopositioning; name of the bank through which the payment was made. The purpose of the analysis is to forecast changes in passenger traffic. Two records (lines) are duplicates if only the fare amount is the same.

A) not true

B) correct

C) under the given conditions it is impossible to give a precise answer

1. If the statistical value of Fisher's criterion is greater than the critical value, then with respect to the parameter under study, a conclusion is made about statistically significant ....

A) differences between experiments

B) overlap in different experiments

B) differences in a single experiment

**Matching tasks**

*Match the left and right columns.*

**Simple (Level 1)**

|  |  |
| --- | --- |
| 26 Establish a match: |  |
| 1. clustering algorithm | A) K-means |
| 2. classification algorithm | B) "naive Bayes." |
|  | C) method of least squares |

|  |  |
| --- | --- |
| 27 Establish a match: |  |
| 1. regression analysis | A) method of least squares |
| 2. analysis of variance | B) Fisher's criterion |
|  | B) Monte Carlo method |

**Medium-complex (level 2)**

|  |  |
| --- | --- |
| 28 Establish a match: |  |
| 1. logistic regression | A) binary response |
| 2. Normal distribution | B) symmetry |
|  | B) entropy growth |

|  |  |
| --- | --- |
| 29 Establish a match: |  |
| 1. clustering algorithm | A) KNN |
| 2. classification algorithm | B) "naive Bayes." |
|  | B) Monte Carlo method |

|  |  |
| --- | --- |
| 30 Establish a match: |  |
| 1. finding the minimum of a continuous function | (A) Gradient methods |
| 2. search for the minimum of a discrete function | B) simplex method |
|  | C) "naive Bayes" |

|  |  |
| --- | --- |
| 31 Establish consistency: |  |
| 1. genetic algorithms | (A) Optimization |
| 2. "naive Bayes." | B) deep learning |
|  | B) discrete analysis |

|  |  |
| --- | --- |
| 32 Establish a match: |  |
| 1. regression | A) method of least squares |
| 2. clustering | B) machine learning algorithms |
|  | B) Monte Carlo method |

|  |  |
| --- | --- |
| 33 Establish a match: |  |
| 1. nearest neighbor algorithm | A) KNN |
| 2. "To the mean" algorithm | B) K-means |
|  | C) ANOVA |

|  |  |
| --- | --- |
| 34 Establish a match: |  |
| 1 analysis of variance | A) ANOVA |
| 2 clustering | B) K-means |
|  | C) KNN |

**Complex (Level 3)**

|  |  |
| --- | --- |
| 35 Establish a match: |  |
| 1. heuristic algorithms | A) genetic algorithm for classification |
| 2. rapid descent algorithms | B) gradient descent |
|  | B) regression analysis |

**Open-ended tasks**

**Addition tasks**

*Write the missing word.*

**Simple (level 1)**

36 Designation of analysis of variance adopted in international sources and software products: ...(Latin)

37 A term introduced by Pyatetsky-Shapiro, sometimes used to refer to data mining: ... (Latin)

1. Combining data from different sources and providing the data to users in a unified way is an integration system ...

39 A method in mathematical statistics that seeks to find relationships in experimental data by examining the significance of differences in mean values is variance ...

40 The statistical relationship between two random variables is shown by the coefficient of ....

1. The adequacy of a regression model can be assessed using the criterion ...
2. To find confidence bounds to the regression coefficients, we use a distribution of ...

**Medium-complex (level 2)**

1. Detecting errors, removing errors, and detecting data inconsistencies are the processes of cleaning ...
2. To process the results of the experiment, the method of least ...
3. To construct a confidence interval to the regression coefficient, we need to set the confidence ...
4. The K-.... algorithm is used in clustering problems
5. The tasks of investigating the existence of relationships between particular groups of variables fall under the correlation ...
6. Problems in which the variables x1, x2, ..., xn are qualitative in nature, and the degree of their influence on the variable y is investigated and established refer to the variance ...
7. In clustering problems, the K nearest .... algorithm is used
8. The naive Bayes algorithm is based on computing a conditional ...
9. The K-means algorithm requires the formation of a matrix ...
10. To test the hypothesis that there is a trend, the criterion is applied ...
11. Calculating the variance of reproducibility requires knowing the results of observations in parallel ...
12. Calculating the adequacy variance requires knowing the estimated and experimental values of the function ...
13. To test the hypothesis of equality of variances, the .... criterion is applied.
14. To test the hypothesis that the correlation coefficient is significant, the .... criterion is applied.
15. The KNN algorithm belongs to the algorithms of deep ...

1. The coefficients of the regression model are found using the method of least .....
2. In analysis, it is not possible to determine the mean if the data under study are measured in units of ... scale
3. Deviations from expected values that may indicate errors, outliers, or unusual events are called ....
4. To test the hypothesis that two sets of data belong to the same population, the criterion is applied...
5. The probability density plot of the normal distribution is symmetric about a line passing through the mathematical ...
6. Student's distribution is used to construct confidence limits ....
7. The Student's coefficient depends on the confidence level and the number of degrees ...
8. The conditional probability is computed in an algorithm naive ...
9. A training dataset is needed to implement deep learning algorithms ...
10. The centroid matrix is needed to implement the K- ... clustering algorithm.

**Complex (Level 3)**

1. An excel function that returns the arithmetic mean of all cells that match multiple conditions: ... (Cyrillic)
2. An excel function that applies criteria to cells in multiple ranges and calculates the number of matches to all criteria:... (Cyrillic)
3. The high-level programming language in which most neural networks are written is called ...(Latin).

**Record card of test tasks (variant 1)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competence | PC-2 Capable of applying modeling, analysis, forecasting and evaluation methods in organizing transport and logistics activities | | | |
| Indicator | PC-2.1 Analyze the volume and quality of data, methods of their processing in the operation of transport systems | | | |
| Discipline | Intelligent data analysis in transportation | | | |
| Level of mastery | Test tasks | | | Total |
| Closed | | Open type |
| Alternative choice | Matching/sequencing | In addition |
| 1.1.1 (20%) | 5 | 2 | 7 | 14 |
| 1.1.2 (70%) | 17 | 7 | 24 | 48 |
| 1.1.3 (10%) | 3 | 1 | 4 | 8 |
| Bottom line: | 25 pcs. | 10 pcs. | 35 pcs. | 70 pcs. |

**Record card of test tasks (variant 2)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Competence | PC-2 Capable of using skills of working with big data, planning, forecasting and modeling of transport processes and systems in professional activity | | | |
| Indicator | PC-2.1 Analyze the volume and quality of data, methods of their processing in the operation of transport systems | | | |
| Discipline | Mathematical methods in transportation research | | | |
| Level of mastery | Test tasks | | | |
| Closed | | | Open type |
| Alternative choice | Matching/Sequencing | In addition | |
| 1.1.1 | 1. ANOVA is ...   A) analysis of variance  B) regression analysis  B) factor analysis   1. The procedure for selecting the number and conditions of experiments necessary and sufficient to solve a given problem with the required accuracy is ....   (A) Planning the experiment  B) decision-making  B) risk assessment   1. Exploring the subject area, collecting data, pre-processing the data, analyzing the data, interpreting the patterns found, and using the new knowledge are the steps ...   A) data mining  B) making a long-term forecast  B) decision-making   1. Estimating the regression model coefficients and correlation coefficient, testing the hypothesis that the coefficients are significant, and testing the hypothesis that the model is adequate are the main steps ...   A) regression analysis  B) analysis of variance  B) spectral analysis   1. Calculating the correlation matrix for all variables involved in the analysis, extracting factors, selecting factors and rotating factors to create a simplified structure, and interpreting factors are the steps ...   A) factor analysis  B) regression analysis  B) analysis of variance | |  |  | | --- | --- | | 26 Establish a match: | | | 1. clustering algorithm | A) K-means | | 2. classification algorithm | B) "naive Bayes." | |  | C) method of least squares |  |  |  | | --- | --- | | 27 Establish a match: | | | 1. regression analysis | A) method of least squares | | 2. analysis of variance | B) Fisher's criterion | |  | B) Monte Carlo method | | 36Definition of analysis of variance as used in international sources and software products: ...  37Term introduced by Pyatetsky-Shapiro, sometimes used to refer to data mining: ....   1. Combining data from different sources and providing the data to users in a unified way is an integration system ...   39 A method in mathematical statistics that seeks to find relationships in experimental data by examining the significance of differences in mean values is variance ...  40 The statistical relationship between two random variables is shown by the coefficient of ....   1. The adequacy of a regression model can be assessed using the criterion ... 2. To find confidence bounds to the regression coefficients, we use a distribution of ... | |
| 1.1.2 | 6 Detecting errors, removing errors, and detecting inconsistencies in data are processes...  (A) Data cleansing  B) data transformation  B) data storage  7The data integration system is a ...  (A) Combining data from different sources and making the data available to users in a unified form  B) sampling of data from different sources  B) sorting data from different sources  8Data integration creates a new single source of information that is a ...  (A) Information resources of the entire new integrated set of sources  B) a resource that provides a single unified interface  B) a common resource that combines data from a group of sources  9The content of the data sources should be:  (A) Unchanged or updated  B) not updated  B) changeable or not updatable  10Analyzing the relationship of mean, mode, and median we can conclude that ...  A) data symmetries  B) data cleanliness  B) correlations  11A method in mathematical statistics aimed at finding relationships in experimental data by examining the significance of differences in mean values is ....  A) analysis of variance  B) factor analysis  B) regression analysis  12 Tasks solved by data mining methods are usually divided into....  (A) Descriptive and predictive  B) statistical and optimization  B) preparation and sorting  13The method used to process the results of an experiment is...  A) least squares  B) Runge-Kutta  B) Monte Carlo  14The logistic .... is used in classification tasks  A) regression  B) progression  B) optimization  15Logistic regression is used for prediction if the response ...  A) binary  B) negative  B) whole  16The statistical relationship between two random variables is shown by ....  A) correlation coefficient  B) dispersion  B) Fisher's coefficient  17The adequacy of a regression model can be assessed using ....  A) Fisher's coefficient  B) correlation coefficient  B) dispersions  18The data under investigation are measured in nominal scale units. It is not possible to determine:  A) median  B) fashion  B) frequency distribution  19The data under study are measured in ordinal scale units. It is not possible to determine:  (A) Average  B) fashion  B) frequency distribution  20Mode , median, and mean for the data set under study are the same. This means that.  A) the distribution of this random variable is symmetric  B) the distribution of this random variable is discrete  C) the distribution of this random variable is asymmetric  21Calculating the variance of adequacy requires knowing the ...  A) calculated and experimental values of the response function  B) results of observations in parallel experiments  B) number of degrees of freedom  22The inversion criterion is used to test the hypothesis that the  A) the presence of a trend  B) equality of variance  B) reliability of data | |  |  | | --- | --- | | 28 Establish a match: | | | 1. logistic regression | A) binary response | | 2. Normal distribution | B) symmetry | |  | B) entropy growth |  |  |  | | --- | --- | | 29 Establish consistency: | | | 1. clustering algorithm | A) KNN | | 2. classification algorithm | B) "naive Bayes." | |  | B) Monte Carlo method |  |  |  |  | | --- | --- | --- | | 30 Establish a match: | | | | 1. finding the minimum of a continuous function | (A) Gradient methods | | 2. search for the minimum of a discrete function | B) simplex method | |  | C) "naive Bayes" |  |  |  | | --- | --- | | 31 Establish a match: | | | 1. genetic algorithms | (A) Optimization | | 2. "naive Bayes." | B) deep learning | |  | B) discrete analysis |  |  |  | | --- | --- | | 32 Establish a match: | | | 1. regression | A) method of least squares | | 2. clustering | B) machine learning algorithms | |  | B) Monte Carlo method |  |  |  | | --- | --- | | 33 Establish a match: | | | 1. nearest-neighbor algorithm | A) KNN | | 2. "To the mean" algorithm | B) K-means | |  | C) ANOVA |  |  |  | | --- | --- | | 34 Establish a match: | | | 1 analysis of variance | A) ANOVA | | 2 clustering | B) K-means | |  | C) KNN | | 1. Error detection, error removal, and data inconsistency detection are the processes of cleaning ... 2. To process the results of the experiment, the method of least ... 3. To construct a confidence interval to the regression coefficient, we need to set the confidence ... 4. The K-.... algorithm is used in clustering problems 5. The tasks of investigating the existence of relationships between particular groups of variables fall under the correlation ... 6. Problems in which the variables x1, x2, ..., xn are qualitative in nature, and the degree of their influence on the variable y is investigated and established refer to the variance ... 7. In clustering problems, the K nearest .... algorithm is used 8. The naive Bayes algorithm is based on computing a conditional ... 9. The K-means algorithm requires the formation of a matrix ... 10. To test the hypothesis that there is a trend, the criterion is applied ... 11. Calculating the variance of reproducibility requires knowing the results of observations in parallel ... 12. Calculating the adequacy variance requires knowing the estimated and experimental values of the function ... 13. To test the hypothesis about the adequacy of the regression model, the criterion .... is applied. 14. To test the hypothesis that the correlation coefficient is significant, the .... criterion is applied. 15. The KNN algorithm belongs to the deep ... algorithms ... 16. The coefficients of the regression model are found using the method of least ..... 17. In analysis, it is not possible to determine the mean if the data under study are measured in units of ... scale 18. Deviations from expected values that may indicate errors, outliers, or unusual events are called .... 19. To test the hypothesis that two sets of data belong to the same population, the criterion is applied... 20. The probability density plot of the normal distribution is symmetric about a line passing through the mathematical ... 21. Student's distribution is used to construct confidence limits .... 22. The Student's coefficient depends on the confidence level and the number of degrees ... 23. The conditional probability is computed in an algorithm naive ... 24. A training dataset is needed to implement deep learning algorithms ... 25. The centroid matrix is needed to implement the K- ... clustering algorithm. | |
| 1.1.3 | 1. Calculating the variance of reproducibility requires knowing the ...   A) results of observations in parallel experiments  B) the results of observations in one experiment  B) values of factors in parallel experiments   1. The dataset under study has fields (columns) with information about the time of fare payment; payment amount; payment method; vehicle geopositioning coordinates. The purpose of the analysis is to forecast changes in passenger flow. Two records (rows) are duplicates if only the fare amount is the same.   A) not true  B) correct  C) under the given conditions it is impossible to give a precise answer   1. If the statistical value of Fisher's criterion is greater than the critical value, then the conclusion about the studied parameter is made about statistically ....   A) a significant difference between experiments  B) significant overlap across experiments  B) a significant difference in a single experiment | |  |  | | --- | --- | | 35 Establish a match: | | | 1. heuristic algorithms | A) genetic algorithm for classification | | 2. rapid descent algorithms | B) gradient descent | |  | B) regression analysis | | 1. An excel function that returns the arithmetic mean of all cells that match multiple conditions: ... (Cyrillic) 2. An excel function that applies criteria to cells in multiple ranges and calculates the number of matches to all criteria:... (Cyrillic) 3. The high-level programming language in which most neural networks are written is called ... | |
| Bottom line: | 25 pcs. | 10 pcs. | 35 pcs. | |

**Evaluation criteria**

**Criteria for evaluating test tasks**

Evaluation criteria: correct fulfillment of one test task is evaluated by 1 conditional point, incorrect fulfillment - 0 points.

The maximum total score for all correct answers is the highest score of 100 points.

**Scale for evaluating the results of computerized testing of students** (recommended)

|  |  |  |
| --- | --- | --- |
| Assessment | Percentage of correct answers | Points |
| "satisfactory." | 70-79% | 61-75 points |
| "good." | 80-90% | 76-90 points |
| "excellent." | 91-100% | 91-100 points |

**Answer Keys**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No. of tests** | **Number and option of the correct answer** |  |  | **36** | ANOVA |
| **1** | А) |  |  | **37** | data mining |
| **2** | А) |  |  | **38** | data |
| **3** | А) |  |  | **39** | analysis |
| **4** | А) |  |  | **40** | correlations |
| **5** | А) |  |  | **41** | Fisher |
| **6** | А) |  |  | **42** | Student's t-test |
| **7** | А) |  |  | **43** | data |
| **8** | А) |  |  | **44** | squares |
| **9** | А) |  |  | **45** | likelihood |
| **10** | А) |  |  | **46** | means, averages |
| **11** | А) |  |  | **47** | analyzing |
| **12** | А) |  |  | **48** | analyzing |
| **13** | А) |  |  | **49** | neighbors |
| **14** | А) |  |  | **50** | probabilities |
| **15** | А) |  |  | **51** | centroids |
| **16** | А) |  |  | **52** | inversions |
| **17** | А) |  |  | **53** | experiments |
| **18** | А) |  |  | **54** | response |
| **19** | А) |  |  | **55** | Fisher |
| **20** | А) |  |  | **56** | Student's t-test |
| **21** | А) |  |  | **57** | trainings |
| **22** | А) |  |  | **58** | squares |
| **23** | 1A, 2B |  |  | **59** | ordinal |
| **24** | 1A, 2B |  |  | **60** | anomaly, anomaly |
| **25** | 1A, 2B |  |  | **61** | Wilcoxon, Wilcoxon. |
| **26** | 1A, 2B |  |  | **62** | expectations |
| **27** | 1A, 2B |  |  | **63** | boundaries, intervals |
| **28** | 1A, 2B |  |  | **64** | liberties |
| **29** | 1A, 2B |  |  | **65** | Bayes |
| **30** | 1A, 2B |  |  | **66** | trainings |
| **31** | 1A, 2B |  |  | **67** | means, averages |
| **32** | 1A, 2B |  |  | **68** | Srznaczeslimn, srznaczeslimn. |
| **33** | 1A, 2B |  |  | **69** | CHOTESLIMN, CHOTESLIMN. |
| **34** | 1A, 2B |  |  | **70** | Python, PYTHON, python. |
| **35** | 1A, 2B |  |  |  |  |