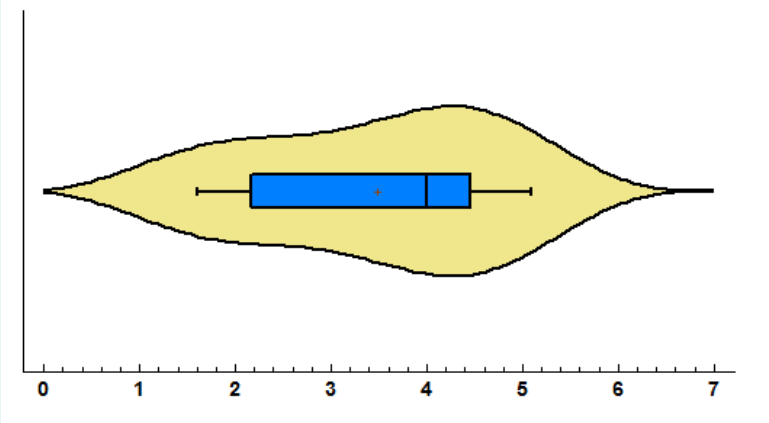
| | |
|---|---|
| **Started on** | Monday, 26 May 2025, 19:24 |
| **State** | Finished |
| **Completed on** | Sunday, 1 June 2025, 15:18 |
| **Time taken** | 5 days 19 hours |
| **Grade** | **5.05** out of 10.00 (**50.5**%) |

**Question 1**

Correct

Mark 0.23 out of 0.23

A violin plot is a visualisation that combines a box-and-whisker plot and a kernel density estimator. From the violin plot we can extract the same information as from the box-and-whisker plot: the median (the vertical segment inside the rectangle); the interquartile range (given by the vertical sides of the rectangle); the whiskers (segments going to either side of the rectangle) give the values $Q_1 - 1.5 \cdot IQR$ и $Q_3 + 1.5 \cdot IQR$, where $IQR = Q_3 - Q_1$ , $Q_1, Q_3$ are the first and third quartiles. Observations outside the interval $(Q_1 - 1.5 \cdot IQR\ , Q_3 + 1.5 \cdot IQR\ )$ are declared outliers.
Below you can see a violin plot for some data set. Choose two correct statements about this data by analysing the graph.




- [ ] The interquartile range is equal to 7
- [x] The median of the data set is 4 ✔
- [ ] The kernel density estimator has more than two modes
- [x] There are outliers in the data ✔

Ваш ответ верный.

The correct answers are:
The median of the data set is 4,

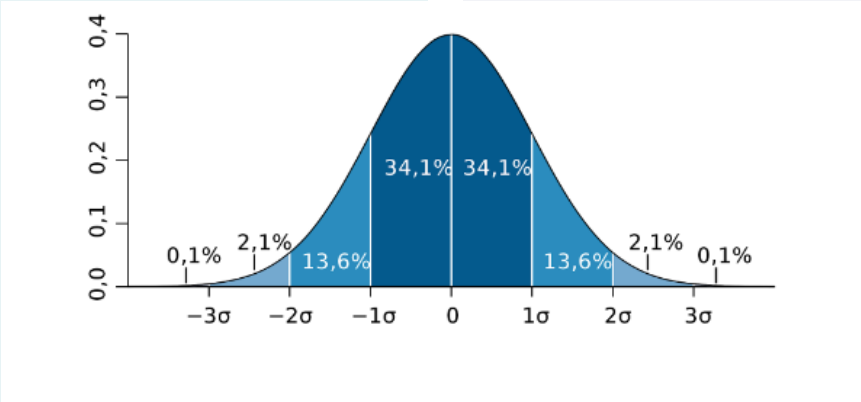
There are outliers in the data

**Question 2**

Correct

Mark 0.23 out of 0.23

Given is the weight distribution of African elephants. Which of the following statements are true if the distribution is normal with a mean of 6 tonnes and a standard deviation of 500 kg?




- [x] 0.1% of African elephants weigh more than 7.5 tonnes ✔
- [ ] The weight of 68.2% of African elephants is between 5.5 and 7 tonnes
- [ ] 13.6 % of African elephants weigh between 5 and 6 tonnes
- [x] 47.7% of African elephants weigh between 6 and 7 tonnes ✔

Ваш ответ верный.

The correct answers are:
0.1% of African elephants weigh more than 7.5 tonnes,

47.7% of African elephants weigh between 6 and 7 tonnes

**Question 3**

Correct

Mark 0.23 out of 0.23

Vasily is trying to send a text message with a weak mobile phone connection. The phone is attempting to send the message until it fails. It is known that the probability of a successful attempt is 0.05 and does not depend on the previous attempts. What is the mathematical expectation of the number of attempts made?

- 5
- 20 ✔
- 1
- 10

Ваш ответ верный.

The correct answer is:
20

**Question 4**

Incorrect

Mark 0.00 out of 0.23

Suppose that objects in the data have two numerical features. In which case, these objects can be represented on a two-dimensional plane. In our task, we are given 1000 objects, each of which is described by a pair of features $(x_1, x_2)$ uniformly distributed on the unit circle.

Among the statements listed below, find the incorrect ones:

- [ ] The values $x_1$ and $x_2$ are dependent, so when training any machine learning algorithm on our data, one of the features: $x_1$ or $x_2$ can be removed.
- [ ] The values $x_1$ and $x_2$ are independent.
- [ ] The values of $x_1$ and $x_2$ are linearly dependent.
- [x] The value of Pearson correlation coefficient between $x_1$ and $x_2$ is small ✖
- [x] The values of $x_1$ and $x_2$ are dependent, but not linearly dependent. ✖

Ваш ответ неправильный.

The correct answers are:
The values of $x_1$ and $x_2$ are linearly dependent.

,

The values $x_1$ and $x_2$ are independent.

,

The values $x_1$ and $x_2$ are dependent, so when training any machine learning algorithm on our data, one of the features: $x_1$ or $x_2$ can be removed.

**Question 5**

Correct

Mark 0.23 out of 0.23

Given is the singular value decomposition of the matrix X:

$$X = U \cdot \begin{pmatrix} 7 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \cdot V'$$

Find the singular value decomposition $U_{10X} \cdot \Sigma_{10X} \cdot V'_{10X}$ for the matrix $10 \cdot X$. What is the sum of all elements of the matrix $\Sigma_{10X}$?

- 10
- 100 ✔
- depends on the matrix X
- 1000

Ваш ответ верный.

The correct answer is:
100

**Question 6**

Correct

Mark 0.23 out of 0.23

Use machine learning to predict the number of views for each article published on a certain website. You have the following attributes: the name of the author of the article, the rating of the author of the article, the number of articles of this author on the site, the length of the article (number of characters) and several other characteristics of the article. The target variable is used in the algorithm in its original form, without any modification. Which of the following metrics can be used to evaluate the quality of the algorithm in this task?

- [ ] Accuracy
- [x] MSE ✔
- [ ] none of the metrics mentioned
- [ ] ROC-AUC
- [ ] f1-score

Ваш ответ верный.

The correct answer is:
MSE

**Question 7**

Incorrect

Mark 0.00 out of 0.23

Consider a linear regression model in the problem of predicting the target variable by two attributes: $a(x) = w_0 + w_1 x_1 + w_2 x_2$. The loss function is of the form $Q(w) = \sum_{i=1}^{l} (y_i - a(x_i))^2$ where $y_i$ is the value of the target variable at the i-th feature. After evaluating the quality of the algorithm by cross-validation, it was found that the model was overfitted. Which of the following approaches are described correctly and can be undertaken to reduce overfitting?

- [x] Add a regulariser ${w_0}^2 + {w_1}^2 + {w_2}^2$ to the model, as l2 regularisation can reduce overfitting ✗
- [ ] Remove the constant coefficient $w_0$, as it increases the complexity of the model but does not affect the generalisation ability of the model
- [x] Add second-degree polynomials to increase the generalisation ability of the model ✗
- [x] Add a regulariser of the form $[w1 \neq 0] + [w2 \neq 0]$ to the model, since l0-regularisation can reduce overfitting (here $[x] = 1$ if the expression $x$ is true, otherwise 0) ✓
- [x] Add a regulariser $|w_1| + |w_2|$ to the model, since l1-regularisation can reduce overfitting ✓

Ваш ответ неправильный.

The correct answers are:
Add a regulariser $|w_1| + |w_2|$ to the model, since l1-regularisation can reduce overfitting

, Add a regulariser of the form $[w1 \neq 0] + [w2 \neq 0]$ to the model, since l0-regularisation can reduce overfitting (here $[x] = 1$ if the expression $x$ is true, otherwise 0)

**Question 8**

Correct

Mark 0.23 out of 0.23

Given is the following text: "The quick brown foxes are jumping over the lazy dogs" :("
After some processing, the result is:
['the', 'quick', 'brown', 'fox', 'be', 'jump', 'over', 'the', 'lazy', 'dog', '.']

Choose all the steps that have been performed with the source text:

- [x] Lemmatization ✔
- [ ] Vectorization
- [ ] Stemming
- [x] Tokenization ✔

Ваш ответ верный.

The correct answers are:
Tokenization,

Lemmatization

**Question 9**

Correct

Mark 0.23 out of 0.23

We are solving a classification problem to identify a person by voice (1 – the voice belongs to the user, 0 – the voice does not belong to the user). Which quality metric should we choose if we want to penalise only incorrect recognition of someone else's voice as the user's voice? (all metrics indicate the quality of the algorithm, i.e. the higher the value of the metric, the higher the quality of the algorithm):

- ( ) TP/(TP+FP)
- ( ) (TP+TN)/(TP+FP+TN+FN)
- (•) TN/(FP+TN) ✔
- ( ) TP/(TP+FN)

Ваш ответ верный.

The correct answer is:
TN/(FP+TN)

**Question 10**

Correct

Mark 0.23 out of 0.23

We are solving a binary classification problem with classes {0, 1}. The algorithm produces some estimate belonging to the segment [0, 1] that the object belongs to class 1. The quality of the algorithm is ROC-AUC=0.5. How does the value of the quality metric change if we square each prediction?

- It depends on the data: may improve or worsen
- It will not change ✔
- It will worsen
- It will improve

Ваш ответ верный.

The correct answer is:
It will not change

**Question 11**

Correct

Mark 0.23 out of 0.23

Select all the correct statements about gradient descent:

- [ ] At each step of the algorithm, the gradient from a single, randomly selected element is considered.
- [x] If you do not make the step length of gradient descent small enough, the algorithm may diverge. ✔
- [ ] Gradient descent is used to find the maximum of a loss function
- [x] Proper selection of the gradient descent step can reduce the number of steps required to find the minimum. ✔

Ваш ответ верный.

Select all the correct statements about gradient descent:

The correct answers are:
If you do not make the step length of gradient descent small enough, the algorithm may diverge.,

Proper selection of the gradient descent step can reduce the number of steps required to find the minimum.

**Question 12**

Incorrect

Mark 0.00 out of 0.23

Which of the following approaches can help reduce overfitting in gradient boosting on decision trees?

- [x] an upper bound on the depth of the tree ✔
- [x] an upper bound on the number of leaves in a tree ✔
- [ ] an upper bound on the absolute value of the predictions in the leaves of a tree in a regression problem
- [x] an upper bound on the number of trees in a composition ✔
- [ ] an upper bound on the minimum number of objects in a leaf

Ваш ответ неправильный.

The correct answers are:
an upper bound on the depth of the tree,

an upper bound on the number of trees in a composition,

an upper bound on the absolute value of the predictions in the leaves of a tree in a regression problem,

an upper bound on the number of leaves in a tree

**Question 13**

Incorrect

Mark 0.00 out of 0.23

Select all the correct statements about the random forest algorithm:

- [x] In a random forest, only a random subset of features is searched at a vertex when selecting the best partition at the vertex ✔
- [x] Object classification is done by voting trees within a random forest. ✔
- [x] In a random forest, each tree is trained on a subsample of the training sample generated in such a way that there are no repeated objects in it (bootstrap) ✖
- [ ] A random forest has a smaller bias than a solver tree of the same depth
- [ ] As the number of trees increases, overfitting does not occur in a random forest

Ваш ответ неправильный.

The correct answers are:
As the number of trees increases, overfitting does not occur in a random forest,

In a random forest, only a random subset of features is searched at a vertex when selecting the best partition at the vertex,

Object classification is done by voting trees within a random forest.

**Question 14**

Incorrect

Mark 0.00 out of 0.23

Select the correct statements about K-means:

- [x] The method is suitable for clusters with complex geometry ✖
- [ ] The method selects the required number of clusters by itself
- [ ] The clustering found by the method depends on the choice of initial position of cluster centres
- [x] The algorithm terminates when there is no change in the intra-cluster distance at some iteration ✔

Ваш ответ неправильный.

The correct answers are: The clustering found by the method depends on the choice of initial position of cluster centres,
The algorithm terminates when there is no change in the intra-cluster distance at some iteration

**Question 15**

Incorrect

Mark 0.00 out of 0.23

The activation function a is used on the hidden layer of the neural network. The output value of some neuron after application of the activation function is equal to "-0.007". Which of the listed activation functions a could have been used in this network?

- (●) Sigmoid ✖
- ( ) Tanh
- ( ) ReLU
- ( ) None of the mentioned

Ваш ответ неправильный.

The correct answer is:
Tanh

**Question 16**

Incorrect

Mark 0.00 out of 0.50

The binary classification algorithm produces values $b_i$ , belonging to the segment [0,1]. There are 10.000 observations in total. If we rank them in ascending order of $b_i$ , we will see that observations with $y_i = 1$ occupy places exactly from 6501 to 6600. Find the area under the ROC curve. Round the answer to hundredths.

Answer:

0.83 ✖

The correct answer is: 0.66

**Question 17**

Incorrect

Mark 0.00 out of 0.50

For an interview for the data scientist position at a certain company, candidates either come on foot or arrive by car. We have information about 100 candidates. For these candidates, we also know whether the candidate was hired or not. The data we have is presented in the form of a matrix below:

| | **Candidate hired** | **Candidate not hired** |
|---|---|---|
| Came by car | 20 | 28 |
| Came on foot | 35 | 17 |

Using logistic regression without regularisation, predict the probability of a candidate being accepted for the position depending on whether they came on foot or by car.

What is the probability that a candidate who came on foot will be hired, based on the logistic model prediction? Round your answer to hundredths.

Answer:

0 ✖

The correct answer is: 0.67

**Question 18**

Correct

Mark 0.50 out of 0.50

Given are the following points in two-dimensional space:
$X = [(-1,1),(1,-1),(1,1),(0,0)]$ with corresponding class labels
$y = [1,1,1,-1]$.

Using leave-one-out cross-validation, find the optimal number of neighbours $k \in [1,3]$ in the k-nearest neighbours method.

The Euclidean distance is used as the proximity measure, the quality metric is accuracy.

Answer:

3 ✔

The correct answer is: 3

**Question 19**

Incorrect

Mark 0.00 out of 0.50

Let each object be described by a two-dimensional vector $x = (x_1, x_2)$.

Given are a vector $w = (2,3)$ and a number $w_0 = 7$.

Find the bandwidth between
$< w, x >= w_0 + 1$ and
$< w, x >= w_0 - 1$, where
$< w, x >$ is the scalar product of the vector $w$ and the vector $x$

Round your answer to hundredths.

Answer:

0.66 ✖

The correct answer is: 0.55

**Question 20**

Incorrect

Mark 0.00 out of 0.50

A car insurance company divides drivers into three classes: class A (low risk), class B (medium risk), class C (high risk).

The company assumes that out of all drivers insured by it, 30% belong to class A, 50% to class B, and 20% to class C. The probability that a Class A driver will have at least one car accident during the year is 0.01; for a Class B driver it is 0.03 and for a Class C driver it is 0.1. Mr Jones insures his car with this company and has a car accident within a year. What is the probability that he is a class A driver? Round your answer to hundredths.

Answer:

0.003 ✖

The correct answer is: 0.08

**Information**

The files Data_train.csv and Data_test.csv contain data about cats.

In this task it is proposed to study the behaviour of wild and domestic cats based on several characteristics. There is some basic information about the cats (type, group). The cats had trainers. The trainer provides food for the cat, as well as trains some cats to complete an obstacle course (some do not have such training). Obstacle course performance was scored independently by three judges on a 100-point scale.

Column description:

* type – the type of the cat: wild or domestic.

* group - coded age group of the cat

* education - level of education of the trainer

* meal - type of the cat's diet

* preparation course - whether the cat has been trained in obstacle course (has had special training).

* score-1 - the first judge's score for the cat's obstacle course

* score-2 - the second judge's score for the cat's obstacle course

* score-3 - the third judge's score for the cat's obstacle course

Further on, you can get a maximum of 4 points for the tasks.

Read the data into two pandas dataframes: df_train and df_test.

**Question 21**

Incorrect

Mark 0.00 out of 0.25

Task 1 **(0.25 points).** Fill in the gaps in the column with a unique category (if the column with the gap is categorical), and with an average value (if the column is numeric). Fill in both df_train and df_test at the same time - in the same manner. In your answer, provide the number of different values required to fill in the gaps (it is equal to the number of new unique categories plus the number of average values to fill in the gaps in the numeric columns).

Answer:

32 ✖

The correct answer is: 1

**Question 22**

Correct

Mark 0.30 out of 0.30

Task 2 **(0.3 points).** The judge decides that the cat has passed the obstacle course if they give it more than 50 points. The cat is considered to have passed the obstacle course if all judges gave it more than 50 points. In df_train, create a 'Pass' column and write 1 if the cat passed the obstacle course and 0 otherwise. In your answer, record how many cats from df_train did not pass the obstacle course.

In df_test, the information about the judges' scores is hidden from you, so you don't know if the cat passed the obstacle course or not - this is what you will have to predict in the tasks below.

Answer:

145 ✔

The correct answer is: 145

**Question 23**

Incorrect

Mark 0.00 out of 0.25

Task 3 (each point is **0.25** points, **1.25** points maximum).

Complete this task using df_train data.

1) Among all wild cats, find the proportion of cats that passed the obstacle course. Calculate the same proportion for domestic cats. In your answer, give the modulus of the difference of these fractions. Round your answer to hundredths.

Answer:

0.03 ✖

The correct answer is: 0.02

**Question 24**

Correct

Mark 0.25 out of 0.25

2) How many cats among those who did not pass the obstacle course had trainers with a "high school" level of education?

Answer:

35 ✔

The correct answer is: 35

**Question 25**

Correct

Mark 0.25 out of 0.25

3) How many wild cats among those who have passed the obstacle course have not had special training course?

Answer:

152 ✔

The correct answer is: 152

**Question 26**

Correct

Mark 0.25 out of 0.25

4) What is the median of the scores given by the first judge?

Answer:

66 ✔

The correct answer is: 66

**Question 27**

Correct

Mark 0.25 out of 0.25

5) Find the interquartile range of the third judge's score (third quartile minus first quartile) for domestic cats that have not received special training.

Comment: To calculate the quartiles of a discrete distribution, use lower interpolation. This means that if the quartile you are looking for lies between the two dimensions i and j, the quartile value is i.

Answer:

20 ✔

The correct answer is: 20

**Question 28**

Incorrect

Mark 0.00 out of 0.30

Task 4 **(0.7 points)**.

a) **(0.3 points)**. Further on, use only categorical columns. Encode them using One-hot encoding taking into account that we do not want to get multicollinearity in the new data. How many numeric columns did we get from the original categorical columns? Encode both df_train and df_test.

Answer:

18 ✖

The correct answer is: 13

**Question 29**

Correct

Mark 0.40 out of 0.40

b) **(0.4 points).** Let us try to predict from the cat's characteristics (former categorical and now numerical columns) whether it passed the obstacle course or not.

Form the object-attribute matrix X and the response vector y from df_train.

Train a decision tree (DecisionTreeClassifier from the sklearn.tree library) of depth 5 with an entropy-based informativeness criterion on the cross-validation training data coded in (a) with three folds, quality metric is roc-auc.

What is roc-auc averaged over folds? Round your answer to the nearest tenth.

Comment: leave other hyperparameters of the tree default (splitter='best', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0).

Answer:

0.7

The correct answer is: 0.7

**Question 30**

Correct

Mark 0.25 out of 0.25

Task 5 (**1.5 points** maximum).

a) **(0.25 points)**. Find the depth of the decision tree (max_depth) by searching the depth from 2 to 20 in steps of 1 and using grid search (GridSearchCV from sklearn.model_selection library) with three folds and quality metric - roc-auc. In your answer, write the best among the searched values of max_depth.

Comment: leave other hyperparameters of the tree default (splitter='best', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0).

Answer:

2

The correct answer is: 2

**Question 31**

Correct

Mark 0.50 out of 0.50

b) **(0.5 points)**. Add a new trait cat_bio to the data, containing pairs of values from the type column and the group column as values. For example, if a cat has type='wild' and group='group B', cat_bio will contain the string '(wild, group B)'. Apply OneHotEncoding (given that we don't want to get multicollinearity in the new data) to the columns 'cat_bio', 'education', 'meal', 'preparation course', and then train a decision tree of depth 5 with an entropy-based informativeness criterion on the resulting post-encoding data.

What is the roc-auc equal to? Round your answer to hundredths.

Comment: leave the other hyperparameters of the tree default (splitter='best', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0).

Answer:

0.68

The correct answer is: 0.68

**Question 32**

Incorrect

Mark 0.00 out of 0.75

c) (**0.75 points**). Now you can use any machine learning model to solve the problem. You can also do any other feature processing. Your task is to get the best quality (ROC_AUC).

The quality is checked on the test data.

- ROC_AUC (greater) 0.7 - 0.25 points

- ROC_AUC (greater) 0.74 - 0.75 points.

Submit the file result.txt: the file should have one column with the predicted values of the target variable for the test sample, without index and header. Attached is an example file for submitting the results.

Attention! Only the result of the last submission will be considered! Before completing the test, make sure that you sent the most accurate prediction last.

**Answer:** (penalty regime: 0 %)

```
1
```

result.txt

Your code failed one or more hidden tests.
Your code must pass all tests to earn any marks. Try again.

Incorrect

Marks for this submission: 0.00/0.75.

Contact site support

You are logged in as Хромов Даниил Максимович KHROMOV DANIIL MAKSIMOVICH (Log out)