

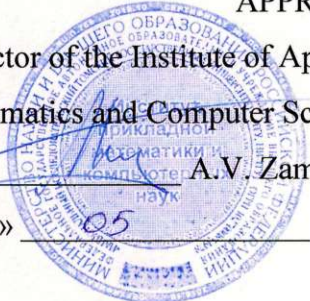
Appendix 1

Ministry of Science and Higher Education of the Russian Federation
NATIONAL RESEARCH
TOMSK STATE UNIVERSITY (NR TSU)

Institute of Applied Mathematics and Computer Science

APPROVE

Director of the Institute of Applied
Mathematics and Computer Science

A.V. Zamyatin

« 16 » 05 2022

Evaluation materials of the current control and intermediate certification in the discipline

(Evaluation tools by discipline)

Applied Machine learning - I

in the major of training

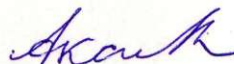
01.04.02 Applied mathematics and informatics

Orientation (profile) of training:

Big Data and Data Science

Tomsk – 2022

ET was implemented:
cand. tech. sciences,
Associate Professor of the Department
of Theoretical Foundations of Informatics



S.V. Aksenov

Reviewer:
cand. tech. sciences,
Associate Professor of the Department
of Theoretical Foundations of Informatics



O.V. Marukhina

Evaluation tools were approved at a meeting of the educational and methodological
commission of the Institute of Applied Mathematics and Computer Science (EMC
IAMCS).

Protocol dated 12.05.2022 № 4

Chairman of the EMC IAMCS,
Dr. tech. Sciences, Professor



S.P. Sushchenko

Evaluation tools (ET) are an element of the system for assessing the formation of competencies among students in general or at a certain stage of its formation.

The ET is developed in accordance with the work program (WP) of the discipline.

1. Competencies and training outcomes, obtained upon the discipline mastery

Competencies	Competence indicator	Code and name of planned training outcomes that characterize the stages of competency formation	Criteria for evaluating training outcomes			
			Excellent	Good	Satisfactory	Unsatisfactory
UK-1. Able to critically analyze problem situations based on a systematic approach and develop an action strategy	IUC-1.1 Identifies a problematic situation and, based on a systematic approach, carries out its multifactor analysis and diagnosis.	OR-1.1.1 The student will: - Know the procedures for identifying, forming and coordinating requirements for the results of analytical work using artificial intelligence and big data technologies	It is awarded to the student if he has deeply and firmly mastered the program material, presents it comprehensively, consistently,	It is awarded to the student if he knows the material firmly, does not allow significant inaccuracies in answering the question,	It is given to a student if he has knowledge only of the basic material, makes inaccuracies, does not use theoretical material	Issued to a student who does not master a significant part of the program material, makes significant mistakes, does not complete practical tasks or with great
	IUC-1.2 Searches, selects and systematizes information to determine alternative strategic solutions in a problem situation.	OR-1.2.1				

	<p>IUC-1.3 Proposes and justifies an action strategy taking into account limitations, risks and possible consequences.</p>	<p>The student will:</p> <ul style="list-style-type: none"> - Know the principles of planning and organizing analytical work using artificial intelligence and big data technologies <p>OR-1.3.1 The student will be able to:</p> <ul style="list-style-type: none"> - Prepare data for analytical work on big data research using artificial intelligence methods <p>OR-1.3.2. The student will be able to:</p> <ul style="list-style-type: none"> - conduct analytical research and develop applications using artificial intelligence and big data technologies in accordance with customer requirements 	<p>clearly and logically, is able to use theoretical knowledge in solving situational problems, and does not have difficulty answering when modifying tasks, competently, using convincing and logical evidence, justifies the decision made, has versatile skills and techniques for performing practical tasks.</p>	<p>correctly applies theoretical principles when answering questions and solving practical problems, and has the necessary skills and techniques for their implementation.</p>	<p>correctly enough to justify decisions, does not demonstrate logical thinking techniques, and has difficulty performing practical work.</p>	<p>difficulty, or makes fundamental errors in completing the tasks provided for in the program.</p>
<p>PC-6. Able to manage the receipt, storage, transmission, and processing of big data.</p>	<p>IPC-6.1 Monitors and evaluates big data processing performance.</p> <hr/> <p>IPC-6.2 Uses methods and tools for receiving, storing, transmitting, and processing big data.</p>		<p>Thoroughly understands modern technologies for constructing and operating artificial intelligence systems based on</p>	<p>Focuses on the basics of modern technologies for constructing and operating intelligent systems, the</p>	<p>Poor understanding of the basics of modern technologies for constructing and operating intelligent</p>	<p>Does not know the basics of modern technologies for constructing and operating intelligent systems, or the principles of functioning of</p>

	<p>IPK-6.3 Develops proposals to improve the performance of big data processing.</p>		<p>machine learning, the principles of functioning of intelligent systems, and machine learning tools. Able to quickly and efficiently solve practical problems of system modeling using modern software tools for working with data, designing intelligent systems, and programming applications with artificial intelligence.</p>	<p>principles of functioning of algorithms and machine learning tools. Able to solve practical problems using modern software tools for working with data, designing applications with machine learning, while allowing minor errors.</p>	<p>systems, and the principles of functioning of machine learning tools. Uses significant difficulties in solving practical problems, using modern software tools for working with data, and designing applications with machine learning</p>	<p>machine learning tools. Does not know how to use modern software tools for working with data when solving practical problems, and cannot design applications with machine learning.</p>
--	--	--	---	---	---	--

2. Stages of competency formation and types of evaluation tools

№	Stages of competency formation (discipline sections)	Code and name of training outcomes	Type of evaluation tool (tests, assignments, cases, questions, etc.)
1	Section 1. Exploratory data analysis. Fundamentals of neural network computing. Performing a laboratory work No. 1 (Preliminary data analysis)	OR-1.1.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 1
2	Section 2. Classification and regression models. Architecture tuning and tuning algorithms for neural networks of counterpropagation. Performing a laboratory work No. 2 (Building a regressor)	OR-1.2.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 2
3	Section 3. Ensembles of models. Neural network training optimizers. Performing laboratory work No. 3 (Building classifiers)	OR-1.3.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 3
4	Section 4. Working with feature space. Neural networks with feedback. Performing Lab #4 (Sample Balancing and ROC)	OR-1.2.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 4
5	Section 5 Fundamentals of Neural Network Computing. Convolutional neural networks and autoencoders. Performing Lab #5 (Working with Feature Space)	OR-1.2.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 5
6	Section 6. Image processing with convolutional neural networks. Neural networks learning unsupervised and with reinforcement. Performing Lab #6 (Determining Feature Importance)	OR-1.3.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 6
7	Section 7. Autoencoders. Visualization and explainability of neural network models. Performing laboratory work No. 7 (Clustering data and assessing its quality)	OR-1.3.2	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 7
8	Section 8. Analysis of signals and time series. Association storage and memory management in neural network models. Performing laboratory work No. 8 (Natural language processing (text classification)).	OR-1.2.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 8
9	Section 9. Use of generative-competing models. Fundamentals of generative-competing models.	OR-1.1.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 9
10	Section 10. Practical aspects of using reinforcement learning. Reinforcement training.	OR-1.3.1	Survey in the classroom, preparation for laboratory classes, public defense of laboratory work No. 10
11	Intermediate certification (according to the	OR-1.3.2	Public presentation and protection of the

results of laboratory work (min 70%) and the presentation of an individual project - 2-3 min / person)		results of an individual project.
--	--	-----------------------------------

3. Typical control tasks or other materials necessary for the assessment of educational training outcomes

3.1. Typical tasks for conducting ongoing monitoring of progress in the discipline:

1. Preliminary data analysis

Write a Python program that loads a dataset, performs an exploratory analysis of this data, and visualizes a series of dependencies between the features in the variations listed below using the matplotlib and sns libraries.

For your analysis option, you need to look at the last digit of your grade book (or student ID) number and make the following adjustments:

- if the last digit is 0 or 5: dataset – Forest fires (<https://archive.ics.uci.edu/ml/datasets/Forest+Fires>), predicted value – fire area (Area);
- if the last digit is 1 or 6: dataset - Wine Quality (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>) predicted value - quality (Quality), for a dataset with red wine, winequality-red .csv
- if the last digit is 2 or 7: dataset – Wine Quality (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>) predicted value – quality (Quality), for dataset with white wine, winequality-white .csv
- if the last digit is 3 or 8: dataset – Bicycle rental (<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>), predicted value – number of bike rentals per day (Area), day dataset .csv
- if the last digit is 4 or 9: dataset - Bicycle rental (<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>), predicted value - number of bike rentals per hour (Area), hour dataset .csv

1. Get a description of the data set and a list of attributes, get the number of missing values in for each attribute.

2. Handle missing values (delete rows/remove attributes/impute values)

4. Plot a pairplot for the dataset

For each image in the tasks below, add a legend and caption

5. Select at least 2 features with a certain distribution of values, and display this distribution using hist and kdeplot histograms, jointplot

6. Select a part of the features of the dataset and display the correlation between them in the form of a heatmap. After that, build a heat map that will display only high values of direct and inverse correlation.

7. Select 3 features (having a certain distribution of values), the target variable (categorical) and display range plots for them (box plot).

8. Visualize some statistics for different attributes using the following tools: violinplot, countplot, FacetGrid, stripplot, swarmplot, catplot, pie.

Write a short conclusion about the most interesting relationships you have found in the data.

2. Building a regressor

Write a Python program that trains three regression models built on a set using three algorithms: a linear regressor, a polynomial regressor, and a random forest regressor.

Select the features used in training and, if necessary, preprocess them. Divide the sample into training and test.

In this work, it is necessary to investigate the operation of the random forest algorithm with different values of hyperparameters and the degree of a polynomial for a polynomial regression model.

For a random forest model, output feature importance values.

Write a short report on the work, including the program with comments, the quality values of the models (determination coefficient, root mean square and mean absolute error).

Choose the best model from the obtained regressors.

For your option, you need to look at the last digit of your grade book (or student ID) number and make the following adjustments:

- if the last digit is 0 or 5: dataset – Forest fires (<https://archive.ics.uci.edu/ml/datasets/Forest+Fires>), predicted value – fire area (Area);

- if the last digit is 1 or 6: dataset - Wine Quality (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>) predicted value - quality (Quality), for a dataset with red wine, winequality-red .csv

- if the last digit is 2 or 7: dataset – Wine Quality (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>) predicted value – quality (Quality), for dataset with white wine, winequality-white .csv

- if the last digit is 3 or 8: dataset – Bicycle rental (<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>), predicted value – number of bike rentals per day (Area), day dataset .csv

- if the last digit is 4 or 9: dataset - Bicycle rental (<https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>), predicted value - number of bike rentals per hour (Area), hour dataset .csv

3. Building classifiers

Write a Python program that loads a dataset and performs the tasks of building classifiers, fitting model hyperparameters, and analyzing the performance of classifying models.

Sample for the Covertype Data Set classifier (<https://archive.ics.uci.edu/ml/datasets/Covertype>).

To do this, you need to look at the last digit of your grade book number (or student ID) and make the following adjustments:

The class label is Cover_Type. Since it is necessary to create binary classifiers, and there are 7 possible classes, we first need to change the value of the Cover_Type label.

To do this, you need to look at the last digit of the number of your record book (or student card) and make the following adjustments: if the last digit is 0 or 5: replace label 0 with class A, labels 1, 2, 3, 4 replace with class B;

if the last digit is 1 or 6: replace label 1 with class A, labels 0, 2, 3, 4 replace with class B;

if the last digit is 2 or 7: replace label 2 with class A, labels 0, 1, 3, 4 replace with class B;

if the last digit is 3 or 8: replace label 3 with class A, labels 0, 1, 2, 4 replace with class B;

if the last digit is 4 or 9: replace label 4 with class A, labels 0, 1, 2, 3 replace with class B.

1. Pre-process the data set.

2. Build classifying models using GridSearch for RandomForest, XGBoost, LogisticRegression and SVC algorithms. Experiment with regularization.

Choose the best combination of algorithm and hyperparameters, draw conclusions and reflect the work done in the report.

4. Sample balancing and ROC

As the data under study, a set from the previous laboratory work (LR No. 3) is taken.

Progress:

1. Change the number of objects of different classes to obtain three balanced samples using different rebalancing methods and perform step 2 from the previous work (Build classifying models using GridSearch for RandomForest, XGBoost, LogisticRegression and SVC algorithms. Conduct experiments with regularization.) for new ones samples.

2. Perform K-block stratified testing for the indicated algorithms with the hyperparameter values obtained earlier for the balanced and initial samples.

3. Visualize the results of the ROC analysis. Calculate the average ROC and standard deviation.

Choose the best combination of algorithm and hyperparameters, draw conclusions and reflect the work done in the report.

5. Working with feature space

1. To complete the task, you need to download a dataset from the <https://archive.ics.uci.edu/ml/datasets.php> repository for any task (regression, clustering or classification), which contains at least 350 objects in the sample and the number Numeric attributes that are subject to analysis are at least 12.

2. Build PCA, T-SNE, SOM models for the selected data.

3. For PCA, obtain the values of the sample variance, the percentage of the variance of the components, the cumulative percentage of the variance, the formula for obtaining the principal components, the first 20 vectors of new features of the principal components for the sample under study. For SOM, the spread over clusters, as well as the first 20 vectors of new features.

6. Determining the importance of features

As the data under study, a set from the previous laboratory work (LR No. 4) is taken.

1. Visualize feature importance values as a contribution to the information gain used in splitting for models using trees for balanced and original samples.

2. Visualize the values of the logistic regression coefficients.

3. Get the Shapley values and visualize them using the shap library tools.

4. Display the importance of features using the dalex library tools.

5. Select the most important features that affect the target variable, draw conclusions from the results of the study of importance and reflect the work done in the report.

7. Clustering data and assessing its quality

Write a Python program that analyzes clustering algorithms. For the data set, choose one of the ones you looked at in previous labs. For clustering, select two arbitrary numerical features. If necessary, bring signs to a standard scale. Use the following clustering algorithms: K-means, agglomerative clustering, and DBScan. Conduct an experiment to identify the optimal number of clusters, for each result derive a quality metric (for example, the silhouette coefficient, etc.), show clusters and centroids for each algorithm and their hyperparameters on the graph. Choose the best combination of algorithm and hyperparameters, draw conclusions and reflect the work done in the report.

8. Natural language processing (text classification)

Collect articles from an e-newspaper (of your choice) related to five topics (eg travel, food, cars, health, culture, etc.). Write a PythonSpark program that loads a dataset according to your variant and builds classifiers for review articles based on four classification algorithms using any three vectorization algorithms (Bag-of-Words, TF-IDF, ...) of texts.

Write a short conclusion about the most interesting results and errors that occurred during the work.

Topics of individual projects:

To strengthen the studied material, it is planned to carry out an individual project within hours of independent work. The project can be completed both individually and in a mini-group (2-3 people), provided that the amount of work will also be increased. At the end of the semester, a mini-presentation on the results of the work is presented for each project.

The theme of the individual project is related to the theme of the master's degree student. The purpose of the work is the use of deep learning methods in their scientific work.

Topics of surveys in the classroom:

Linked to the material of previous lectures, as well as the personal experience of students. Students can offer options for solving the problem set by the teacher, as well as solution tools.

Sample questions:

1. Which of the following neural network models is best suited for predicting time sequences?

a) Single-Layer Perceptron	b) CNN
c) LSTM	d) Multi-layer Perceptron

2. What is the name of several examples from the training set that are used to simultaneously calculate the gradient and weights of the network?

3. Why do models based on convolutional neural networks show the best performance in classifying objects in images compared to other models?

a) They are highly optimized for handling vectors with numeric rather than categorical features.	b) They have a wide range of feature space transformation tools that can be varied by the developer in the model.
c) They take into account the correlation of adjacent components of the vector	d) They use a significantly larger number of adjustable parameters compared to other models

3.2. Typical tasks for conducting intermediate certification in the discipline.

Credit is awarded based on the presentation and defense of an individual project.

The student makes a presentation and also demonstrates the program code. Questions based on the results can be asked by all students in the group, not just the teacher.

Grading Criteria Table

Mark	Criteria
Passed	The student worked actively during the semester, made a presentation of an individual project, attended lectures, and completed laboratory work on time.
Not passed	The student did not work during the semester, did not make a presentation of an individual project, did not attend lectures, laboratory work was not passed or was passed with less than 10 points.

4. Methodological materials that determine the procedures for evaluating training outcomes

4.1. Methodological materials for assessing the current control of progress in the discipline.

The current control of laboratory work is carried out in the form of checking the fulfillment of laboratory work tasks. The current control of progress on theoretical material is carried out in the form of tests.

The assessment of current control is carried out on the basis of the assessment of competencies corresponding to the current section of the discipline, according to the table in section 1.

4.2. Methodological materials for conducting intermediate certification in the discipline.

The final grade for the subject (exam) is set as follows:

"Excellent" - the student completed all laboratory work, there are no unsatisfactory marks for control work, the average (rounded) mark for control work is "excellent";

"Good" - the student completed all laboratory work, there are no unsatisfactory marks for control work, the average (rounded) mark for control work is "good";

"Satisfactory" - the student completed all laboratory work, there are no unsatisfactory marks for control work, the average (rounded) mark for control work is "satisfactory";

"Unsatisfactory" - the student did not pass laboratory work or passed at least one control work as "unsatisfactory".