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

_____ A. V. Zamyatin

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

(Evaluation tools by discipline)

Introduction to Data Science & Data Mining - I

in the major of training
01.04.02 Applied mathematics and informatics

Orientation (profile) of training:
Big Data and Data Science

Tomsk – 2025

ET was implemented:
Dr. tech. sciences, professor,
Head of the Department of Theoretical Foundations of Informatics   A.V. Zamyatin

Reviewer:
Dr. tech. sciences, professor,
Head of the Department of Applied Informatics   S.P. Sushchenko

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 22.05.2025 № 1

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 carry out a critical analysis of problem situations based on a systematic approach, develop an action strategy	IUK-1.1 Identifies a problem situation, on the basis of a systematic approach, carries out its multifactorial analysis and diagnostics. IUK-1.2 Carries out the search, selection and systematization of information to determine alternative options for strategic solutions in a problem situation. IUK-1.3 Suggests and justifies the strategy of action, taking into account the limitations, risks and possible consequences.	MR-1.1.1. The student will be able to: - find and use sources of additional information to improve the level of general and professional knowledge; - to select and process information on the chosen research topic; correctly quote and make references to the sources used in written works; - be able to apply natural science and mathematical knowledge to solve scientific and engineering problems in the field of informatics and computer technology.	Demonstration of a high level of knowledge; Has a well-formed understanding of the main methods of scientific and practical search in data mining and other areas using information technology; the ability to formulate a scientific and practical problem and apply existing methods of data mining to solve it.	In general, successful, but containing some gaps in knowledge about the main methods of scientific and practical search in data mining and other areas using information technology. the ability to formulate a scientific and practical problem and apply existing methods of data mining to solve it.	Fragmentary, incomplete knowledge without gross errors of knowledge about the main methods of scientific and practical search in data mining and other areas using information technology.	Has no idea about modern methods of data mining.

GPC-1. Able to solve actual problems of fundamental and applied mathematics.	IGPC -1.1 Analyzes problems in the field of fundamental and applied mathematics.	MR-1.1.2. The student will be able to: - develop means of implementing information technologies (methodological, informational, mathematical, algorithmic, technical and software) - to conduct experimental studies according to a given methodology and analyze the results. - perform processing and analysis of data obtained in theoretical and experimental studies.	Demonstration of a high level of knowledge; Has a well-formed understanding of the existing methods and approaches to data mining of various nature.	In general, successful, but containing some gaps in knowledge about existing methods and approaches to data mining of various nature.	Fragmentary, incomplete knowledge without gross errors of knowledge about existing methods and approaches to data mining of various nature.	Has no idea about existing methods and approaches to data mining.

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. Basic problems of building systems	MR-1.1.1, MR-1.2.1	Tasks for current control, questions for interim certification
2	Section 2. Data preprocessing. Classification.	MR-1.1.1, MR-1.2.1, MR-1.2.2	Tasks for current control, questions for interim certification
3	Section 3. Regression. Association, serial association, anomalies and visualization.	MR-1.1.1, MR-1.2.1, MR-1.2.2, MR-1.3.1;	Tasks for current control, questions for interim certification
4	Section 4. High performance data processing. Software environments for data mining.	MR-1.1.1, MR-1.2.1, MR-1.2.2, MR-1.3.1;	Tasks for current control, questions for interim certification

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

3.1. Typical tasks for conducting ongoing monitoring of progress in the discipline Abstract (on an agreed topic). The abstract must be accompanied by a presentation. Topic examples:

Modern neural networks in data processing (images, videos, technological signals, music, etc.);

Modern classification algorithms (images, texts, etc.);

Intelligent data processing in ... (industry, medicine, business, entertainment, leisure, etc.);

Extraction of knowledge from texts;

Anomaly detection;

Varieties of convolutional neural networks;

Intelligent algorithms in the early diagnosis of diseases; Intelligent algorithms in personalized medicine;

Intelligent algorithms in robotics, transport systems, etc.;

Intelligent algorithms in banking/insurance/…;

Project (on an agreed topic). Implement a small data mining project using the RapidMiner environment or one of the programming languages (for example, Python, R), with the possible use of public databases (or data from other sources). Stages of project implementation: Search and preparation of a data set; Development of technical specifications; Pilot implementation of one model, choice of metric and accuracy assessment (fixing the obtained accuracy at this stage); Implementation of all points of the technical task, setting the parameters of the models, assessing the accuracy (the accuracy obtained at this stage should be greater than at the previous one): Preparation of a report (with a description of the subject area, selected algorithms and model parameters), presentations, public defense of the project;

Each student implements an individual or group project as a sequence of laboratory work:

Laboratory work №1. Individual task on the topic "Analysis of the subject area, formulation of the goals and objectives of the study. Extraction and primary storage of data. The purpose of the work is to teach students to solve the problems of analyzing the subject area, its

adaptation for data analysis methods, taking into account the fundamental features of the subject area.

Laboratory work №2. Individual task on the topic "Data pre-processing: cleaning, integration, transformation". The purpose of the work is to teach students to solve problems of data preprocessing, which involves a laborious cleaning procedure (elimination of contradictions, random emissions and interference, omissions), integration (combining data from several possible sources in one storage), transformation (may include data aggregation and compression, discretization attributes and dimensionality reduction, etc.).

Laboratory work №3. Individual task on the topic "Meaningful data analysis by Data Mining methods". The purpose of the work is to teach students to reasonably apply the basic methods of data mining, taking into account the peculiarities of both the theoretical construction of the applied methods and the chosen subject area.

Laboratory work №4. Individual task on the topic "Visualization and interpretation of the results." The purpose of the work is to teach students to visualize and interpret the results obtained in a form suitable for making managerial decisions.

Examples of topics for self-study: □
- Neural network methods of data analysis, convolution neural networks, deep learning. □
- Methods of intellectual analysis of media (social media data mining). □
- Methods of machine learning in the problems of financial analytics. □
- Methods of machine learning in problems of early medical diagnostics. □
- Combining models in data analysis, boosting. □
- Independent component analysis method. □
- High-dimensional data visualization methods.

3.2. Typical tasks for conducting intermediate certification in the discipline Questions for the exam:

1. Basic concepts, terminology;
2. Data Mining / Data Science;
3. Big Data (basic concepts and properties);
4. Deduction and induction;
5. Data mining in business application examples;
6. Data mining in solving complex applied problems;
7. Data mining in the early diagnosis of dangerous diseases;
8. Data mining in industrial predictive analytics;
9. Main tasks and classification of data analysis methods;
10. Fundamentals of machine learning;
11. Pre-processing of data;
12. Optimization of feature space;
13. Statement of the problem of classification;
14. Controlled non-parametric classification;
15. Controlled non-parametric neural network classification;
16. Classification according to the method of support vector machines;
17. Decision trees;
18. Uncontrolled classification (clustering);
19. Regression (the concept of regression, the main stages of regression analysis, methods for restoring regression);

20. Association;

21. Sequential association (algorithms of the Apriori family, GSP algorithm);

22. Multilevel machine learning (bootstrapping, bagging, staking, boosting);

23. Anomaly detection;

24. Visualization in Data Mining;

25. Activation functions;

26. Main types of artificial neural networks;

27. Convolutional neural networks;

28. Deep learning environments and frameworks;

29. Basic tasks of text processing;

30. Stages of text pre-processing;

31. Classification quality metrics;

32. Hypothesis A/B, Kappa-index of agreement, ROC-curve;

33. Metric of the quality of the time series forecast;

34. Clustering quality metrics;

35. Principles of high performance computing;

36. Features of building a computing cluster;

37. Environments and tools for high performance computing;

38. Data mining tools.

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.

Rating system for assessing the current progress of students

Table - Scoring for control elements

Elements of learning activity	Maximum score since the beginning of the semester	Competence to be assessed
Essay on topic with presentation	20	UK-1,GPK-1.
Project implementation	40	UK-1,GPK-1.
Poll in class	10	UK-1,GPK-1.
Exam	30	UK-1,GPK-1.

4.2. Methodological materials for conducting intermediate certification in the discipline. The sum of points scored by the student during the semester is translated into the assessment of the intermediate assessment of the student's progress according to the scale below.

Recalculation of points into grades for intermediate certification

Points on the checkpoint date	Grade
≥ 90% of the maximum score	5
From 70% to 89% of the maximum points	4

From 60% to 69% of the maximum points	3
< 60% of the maximum score	2