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

Work program of the 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

Form of study
full-time

Qualification
Master

Year of admission
2023

Code of discipline in the curriculum: B1.O.01

AGREED:
Head of EP
_____ A.V. Zamyatin

Chairman of the EMC
_____ _S.P. Sushchenko

Tomsk – 2023

1. Purpose and planned results of mastering the discipline

The purpose of mastering the discipline is the formation of the following competencies:
- UK-1 - the ability to carry out a critical analysis of problem situations based on a systematic approach, to develop an action strategy;
- GPC-1 - the ability to solve actual problems of fundamental and applied mathematics.
The results of mastering the discipline are the following indicators of the achievement of competencies:
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.
IOPC-1.1 Analyzes problems in the field of fundamental and applied mathematics.

2. Tasks of mastering the discipline

– to study the main models and methods of data development;

– learn how to apply these models and methods, as well as the software in which they are implemented;

– to gain experience in analyzing real data using the methods studied.

3. The place of discipline in the structure of the educational program

Discipline belongs to the mandatory part of the educational program.

4. Semester of mastering and form of intermediate certification in the discipline

First semester, exam.

5. Entrance requirements for mastering the discipline

Successful mastering of the discipline requires competencies formed in the course of mastering educational programs of the previous level of education.

6. Implementation language

English.

7. Scope of discipline

The total labor intensity of the discipline is 6 credits, 216 hours, of which:
- lectures: 20 hours
- laboratory: 44 hours
  including practical training: 0 h.
The volume of independent work of the student is determined by the curriculum.

8. The content of the discipline, structured by topics

Topic 1. Main problems of building systems Brief content of the topic. Relevance, basic terminology and development trends. Main tasks, stages and classification of data analysis methods. Data preprocessing. Classification. Regression. Association, serial association, anomalies and visualization. High performance data processing. Software environments for data mining.

Topic 2. Data preprocessing. Classification Brief content of the topic. Basic methods and data preprocessing. Feature space optimization without feature space transformation. Controlled non-parametric neural network classification. Classification by the method of support vector machines. decision trees. Section 3. Regression. Association, serial association, anomalies and visualization. The concept of regression. The main stages of regression analysis. Description of the association algorithm. Algorithms of the "Apriori" family. GSP algorithm. Anomaly detection and visualization methods.

Topic 3. Regression. Association, serial association, anomalies and visualization Brief content of the topic. The concept of regression. The main stages of regression analysis. Description of the association algorithm. Algorithms of the "Apriori" family. GSP algorithm. Anomaly detection and visualization methods.

Topic 4. High performance data processing Brief content of the topic. Principles of organizing high-performance computing. SMP systems. Models of parallel computing MPMD, SPMD. Software environments for data mining.

9. Ongoing evaluation

The ongoing evaluation is carried out on the basis of attendance control, preparation and defense of abstracts, work on a group project and is recorded in the form of a control point at least once a semester.

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.

10. The procedure for conducting and criteria for evaluating the intermediate certification

The exam in the first semester is held in writing by tickets. The ticket contains two theoretical questions. The duration of the exam is 1.5 hours.

An approximate list of theoretical questions:

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.

11. Educational and methodological support

a) Electronic training course on the discipline at the electronic university "Moodle" - **https://moodle.tsu.ru/course/view.php?id=22102**

b) Evaluation materials of the current control and intermediate certification in the discipline.

12. List of educational literature and Internet resources

a) main literature:

– Zamyatin A.V. Introduction to Data Mining: Study Guide /A. V. Zamyatin. - Tomsk: Publishing House of Tomsk State University, 2016. - 118 p. – URL: http://vital.lib.tsu.ru/vital/access/manager/Repository/vtls:000529594.-1

– Pocket Data Mining electronic resource : Big Data on Small Devices / by Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes. - Cham:: Springer International Publishing:: Imprint: Springer, , 2014. - 108 p. – URL: http://dx.doi.org/10.1007/978-3-319-02711.-1

- Principles of Data Mining electronic resource /by Max Bramer. - London: Springer London: Imprint: Springer, 2013. - 440 p. – URL: http://dx.doi.org/10.1007/978-1-4471-4884-5

b) additional literature:

–Principles of Data Mining electronic resource /by Max Bramer.Bramer, Max. London :: Springer London :: Imprint: Springer, 2013, XIV, 440 p. [Electronic resource]. – Access mode: http://dx.doi.org/10.1007/978-1-4471-4884-5.

– Pocket Data Mining electronic resource : Big Data on Small Devices / /by Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes.Gaber, Mohamed Medhat. Cham :: Springer International Publishing :: Imprint: Springer, 2014. IX, 108 p. [Electronic resource]. – Access mode: http://dx.doi.org/10.1007/978-3-319-02711-1.

– Mirkin B. G. Introduction to data analysis: textbook and workshop for undergraduate and graduate studies: [for university students studying in engineering, natural science and economic areas and specialties] /B. G. Mirkin; "Higher School of Economics" National Research University. - Moscow: Yurayt, 2015. - 173 p.

– Kulaichev A.P. Methods and means of complex data analysis: textbook. - Moscow: Forum [and others], 2014. - 511 p.

13. List of information technologies

a) licensed and freely distributed software:

– C, C++, C#, Python, R-Studio, Rapid Miner, MS Azure.;

b) information reference systems:

– Data Mining for Service electronic. Berlin, Heidelberg, Imprint: Springer, Springer eBooks VIII, 291 p. 2014 (edited by Katsutoshi Yada) [Electronic resource]. – Access mode: http://dx.doi.org/10.1007/978-3-642-45252-9

– Data Mining for Geoinformatics electronic resource : Methods and Applications / /edited by Guido Cervone, Jessica Lin, Nigel Waters. New York, NY : : Springer New York : : Imprint: Springer, , 2014, 166 p. [Electronic resource]. – Access mode: http://dx.doi.org/10.1007/978-1-4614-7669-6

c) professional databases (if any):

– University Information System RUSSIA – https://uisrussia.msu.ru/

– Unified Interdepartmental Information and Statistical System (EMISS) – https://www.fedstat.ru/

14. Logistics

The material and technical support of the discipline requires the availability of computer equipment with the appropriate software installed and other equipment that supports presentations, the construction of project documentation, and access to the Internet.

15. Authors information

Zamyatin Alexander Vladimirovich, Doctor of Technical Sciences, Professor, Head of the Department of Theoretical Foundations of Informatics, TSU, Director of the IAMCS.