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

Mathematics & Statistics for Data Science, Advanced track - II

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.P.V.02

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:

– GPC-1 – the ability to solve actual problems of fundamental and applied mathematics;
– GPC-2 – the ability to improve and implement new mathematical methods for solving applied problems.

The results of mastering the discipline are the following indicators of the achievement of competencies:

IOPC-1.3 Demonstrates the skills of using the basic concepts, facts, principles of mathematics, computer science and natural sciences to solve practical problems related to applied mathematics and computer science.

IOPC-2.1 Uses the results of applied mathematics to adapt new methods for solving problems in the field of his professional interests.

IOPC-2.2 Implements and improves new methods, solving applied problems in the field of professional activity.

IOPC-2.3. Carries out a qualitative and quantitative analysis of the resulting solution in order to build the best option.

2. Tasks of mastering the discipline

- To learn how to solve problems of statistical data analysis, starting from stating the initial problems of the corresponding subject area in terms of applied statistics, the choice of solution methods and quality criteria for the solutions developed, and ending with the interpretation of the findings in subject area terms.
- To study the main methods of statistical data analysis.
- To acquire skills required to use statistical data processing software.

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

Second semester, exam

5. Entrance requirements for mastering the discipline

Successful mastering of the discipline requires competencies formed during the development of educational programs of the previous level of education, knowledge of the basics of mathematical analysis, linear algebra, optimization methods, probability theory and mathematical statistics, as well as the basics of programming and of Statistical Methods for Machine Learning – I of this program.

.

6. Implementation language

English.

7. Scope of discipline

The total labor intensity of the discipline is 3 credits, 108 hours, of which:
- lectures: 10 hours
- laboratory: 20 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. Multiple regression.

Basic concepts and tasks of regression analysis, General problem statement of multiple regression. Least squares method for regression parameters estimating. Gauss-Markov theorem. Estimation of variances. Checking the quality of a multiple regression model. Nonlinear models and linearization.

Topic 2. Additional issues of regression analysis.

Dummy variables. The case of shifted noise. The case of correlated observations Heteroscedasticity. Multicollinearity.

Topic 3. Tasks of classification.

Basic concepts and tasks of classification. Binary classification and logistic regression. quality metrics. ROC analysis.

9. Ongoing evaluation

The ongoing evaluation is carried out by monitoring attendance, performing practical work and is recorded in the form of a checkpoint at least once a semester.

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

Final assessment in the second semester is carried out as a test. The test consists of 10-15 questions. The duration of the exam is 30 minutes.

An approximate list of theoretical questions and topics for preparing for the exam:

1. Nonlinear models and linearization.
2. The case of shifted noise.
3. The case of correlated homoscedastic observations.
4. The case of uncorrelated heteroscedastic observations.
5. Multicollinearity.
6. Dummy variables.
7. Classification Problem Statement.
8. Logistic Regression.
9. Quality metrics of a binary classifier.
10. ROC analysis.

Examples of tasks for laboratory work on statistical analysis

Practical work. Data preprocessing

Exercise.

Import the given data set.

1. Build graphs to visualize data and their relations.

2. Check the relations of features with each other and their influence on the dependent target variable.
3. Build and analyze a multiple regression model of the target variable from all the presented quantitative and ordinal factors.
4. Carry out processing and coding of categorical factors.
5. Build and analyze a multiple regression model on all the proposed features.
6. Remove insignificant factors. Build the final model.
7. Check the residuals of the model for normality.
8. Set a new observation with your own feature values and build a target variable forecast for it.

Practical work. Logistic Regression.

Exercise

Generate observations related by one-way logistic regression.

1. Set the sample size $n = 20:50$.
2. Form the values of the factor *x* as an integer uniformly distributed random variable in the interval $[a,b]$.
3. Specify normally distributed noise $\varepsilon \sim N(0,\sigma)$.
4. Define the regression model
$$\Pi(x) = \frac{e^{\theta_0+\theta_1 x+\varepsilon}}{1+e^{\theta_0+\theta_1 x+\varepsilon}}.$$
5. The value of the binary dependent variable is determined as
$$y_i = \begin{cases} 0, & \Pi(x_i) < \frac{1}{2}; \\ 1, & \Pi(x_i) \geq \frac{1}{2}. \end{cases}$$
Set all parameters yourself, depending on the scatterplot.
6. Estimate the model parameters.
7. Check the quality of the model.

The test is graded “passed” or “not passed”.

For a 10 questions test. For each question, depending on its complexity, you can get from 1 to 3 points. Max 20.

passed	from 11 to 20
not passed	from 0 to 10

To complete the course successfully it is necessary to score more than 10 points in a test and complete all the lab works throughout the semester.

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=00000

b) Assessment materials of the ongoing evaluation and intermediate certification in the discipline.

c) Guidelines for practical work.

12. List of educational literature and Internet resources

a) basic literature

1. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) 1st ed. 2013, Corr. 7th printing 2017 Edition.
2. https://book.stat420.org/applied_statistics.pdf
3. http://thuvienso.bvu.edu.vn/bitstream/TVDHBRVT/15780/1/Applied-Statistics.pdf
4. http://wpage.unina.it/cafiero/books/stat.pdf
5. https://www.researchgate.net/publication/242692234_Statistical_foundations_of_machine_learning_the_handbook

б) additional literature:

6. https://bookdown.org/ndphillips/YaRrr/
7. https://mml-book.github.io/book/mml-book.pdf

13. List of information technologies

a) licensed and freely distributed software:

- Microsoft Office Standard 2013 Russian: software package. Includes applications: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);
- publicly available cloud technologies (Google Docs, Yandex disk, etc.)
- R The R Foundation, USA freeware.
- RStudio RStudio, PBC, USA freeware.
- JASP University of Amsterdam, Netherlands freeware.

14. Logistics

Halls for lectures.

Classrooms for seminars, individual and group work, ongoing evaluation and intermediate certification.

Classrooms for independent work, equipped with computer technology and access to the Internet, to the electronic information and educational environment and to information reference systems.

Halls for lectures and seminars, individual and group consultations, ongoing evaluation and intermediate certification in a mixed format (“Aktru”).

15. Authors information

Tatiana Valerievna Kabanova, PhD, Associate Professor, Department of Probability Theory and Mathematical Statistics, Institute of Applied Mathematics and Computer Science TSU.