БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ /
BELARUSIAN STATE UNIVERSITY

УТВЕРЖДАЮ / APPROVED

Ректор Белорусского
государственного университета/
Rector of Belarusian State University

_____А.Д.Король /Andrei D.Karol

27.06.2025

Регистрационный/Registration № 3821/m.

МЕТОДЫ НАХОЖДЕНИЯ И АНАЛИЗА ЗАВИСИМОСТЕЙ В ДАННЫХ/ DATA MINING METHODS

Учебная программа учреждения образования по учебной дисциплине для специальности:
The program of the educational institution of the discipline for the speciality:

7-06-0533-05 Прикладная математика и информатика /
7-06-0533-05 Applied Mathematics and Computer Science
Профилизация / Profilization:
Компьютерный анализ данных / Computer Data Analysis

2025

Учебная программа составлена на основе ОСВО 7-06-0533-05-2023 и учебного плана № М53а-5.3-115/уч. от 11.04.2023.

СОСТАВИТЕЛЬ:

П.А.Пашук, старший преподаватель кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики Белорусского государственного университета, магистр прикладной математики и информационных технологий

РЕЦЕНЗЕНТЫ:

М.С.Абрамович, заведующий НИЛ статистического анализа и моделирования учреждения Белорусского государственного университета «НИИ прикладных проблем математики и информатики», кандидат физико-математических наук, доцент;

А.И.Кишкар, ведущий инженер-программист отдела информационных систем управления бизнес-приложений департамента производства ЗАО «Международный деловой альянс», магистр прикладной математики и информационных технологий

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой теории вероятностей и математической статистики БГУ
(протокол № 15 от 27.05.2025);

Научно-методическим советом БГУ
(протокол № 11 от 26.06.2025)

Заведующий кафедрой А.Ю.Харин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины «Методы нахождения и анализа зависимостей в данных» – изучение современных методов и алгоритмов анализа данных и формирование навыков решения практических задач с использованием современного программного обеспечения.

Задачи учебной дисциплины:

1) изучение особых подходов и специальных методов анализа данных;

2) знакомство студентов с применением специальных методов анализа данных, а также их преимуществом и недостатками;

3) формирование практических навыков решения прикладных задач с использованием современного программного обеспечения.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Учебная дисциплина относится **к модулю** «Специальные методы анализа» компонента учреждения образования.

Учебная программа составлена с учетом межпредметных **связей** и программ по дисциплинам: «Многомерный статистический анализ», «Методы и алгоритмы машинного обучения», «Анализ Интернет-данных», «Математическое и компьютерное прогнозирование».

Требования к компетенциям

Освоение учебной дисциплины «Методы нахождения и анализа зависимостей в данных» должно обеспечить формирование следующих компетенций:

Углубленные профессиональные компетенции (УПК):

Применять знания по современным вероятностным моделям, используемым для анализа сложных данных, применять специальные современные методы анализа сложных данных.

Специализированные компетенции (СК):

Применять методы нахождения и анализа зависимостей в данных;

В результате освоения учебной дисциплины студент должен:

знать:

- основные специальные методы анализа данных;
- особенности методов и алгоритмов анализа данных;

уметь:

- использовать методы и алгоритмы анализа данных;
- выбирать оптимальный метод анализа данных при решении задачи;
- интерпретировать полученные результаты;

иметь навык:

- работы с основными специальными методами анализа данных;
- компьютерной реализации основных методов;
- решения прикладных задач с использованием современного программного обеспечения.

Структура учебной дисциплины

Дисциплина изучается в 3 семестре. В соответствии с учебным планом всего на изучение учебной дисциплины «Методы нахождения и анализа зависимостей в данных» отведено для очной формы получения высшего образования – 198 часов, в том числе 66 аудиторных часов, из них: лекции – 22 часа, практические занятия – 22 часа, семинарские занятия – 22 часа.

Трудоемкость учебной дисциплины составляет 6 зачетных единиц.

Форма промежуточной аттестации – экзамен.

EXPLANATORY NOTE

Aim and tasks of the discipline

Aim of the discipline «Data mining methods» – studying modern methods and algorithms for intelligent data analysis and developing skills in solving practical problems using modern software.

Tasks of the discipline:

1) study of special approaches and methods of data analysis;

2) familiarization of students with the use of specific data analysis methods, as well as their advantages and disadvantages;

3) developing practical skills in solving applied problems using modern software.

Place of the academic discipline in the system of training a specialist with higher education.

The academic discipline is part of the module «Specific methods of analysis» of higher education institution component.

The curriculum is designed taking into account interdisciplinary connections and programs in disciplines: «Multivariate statistical analysis», «Methods and algorithms of machine learning», «Neural networks in machine learning», «Methods for statistical analysis of complex data», «Data mining methods», «Internet data analysis».

Requirements for competences

Mastering of the academic discipline «Intelligent data analysis» should provide the formation of the following universal and advanced professional competences:

Advanced professional competences:

To apply knowledge of modern probability models for complex data analysis, to apply specific modern methods for complex data analysis.

Specialized competences:

To apply methods of data mining.

As a result of mastering the academic discipline, the student is expected to:

know:

- basic special methods of data analysis;
- features of methods and algorithms of data analysis;

be able to:

- use methods and algorithms of data analysis;
- select the optimal method of data analysis when solving a problem;
- interpret the obtained results;

have skills in:

- working with basic special methods of data analysis;
- computer implementation of basic methods;
- solving applied problems using modern software;

Structure of the academic discipline

The discipline is studied in the 3 semester. In total for the study of the discipline «Data mining methods» is allocated for full-time higher education – 198 hours, including 66 in-class hours, of them: lectures – 22 часов, workshops – 22 hours, seminar classes – 22 hours.

The labour intensity of the discipline is 6 credit units.

Form of certification – exam.

CONTENT OF THE STUDY MATERIAL

Topic 1. Preliminary data analysis.

Testing the sample for homogeneity, randomness. Detecting anomalous observations. Robust estimates of position and scale. Data visualization. Determining the distribution law.

Topic 2. Testing statistical hypotheses.

Principles of testing statistical hypotheses. Testing hypotheses about the equality of means and variances of independent and dependent samples. Robust criteria for testing hypotheses. Nonparametric criteria for testing hypotheses. Multiple hypothesis testing.

Topic 3. Analysis of the relationship and dependence of features/

Sample correlation coefficients. Testing the significance of sample correlation coefficients. Robust estimates of correlation coefficients. Contingency tables.

Topic 4. Multiple linear regression.

Estimation of regression equation coefficients. Model adequacy testing. Residual analysis. Multicollinearity problem.

Topic 5. Other aspects of regression models.

Qualitative predictors. Predictors with only two levels. Predictors with more than two levels. Extensions of the linear model. Nonlinear relationships. Potential problems.

Topic 6. Methods for creating repeated samples.

Cross-validation. Validation sample method. K-fold cross-validation. Bootstrap.

Topic 7. Selection and regularization of linear models.

Optimal subset selection. Stepwise selection. Compression methods. Ridge regression. Lasso regression. Hyperparameter selection.

Topic 8. Dimensionality reduction methods.

Principal Component Regression. Partial Least Squares. High-Dimensional Data. Regression for High-Dimensional Data.

Topic 9. Methods based on decision trees.

Basic concepts. Regression trees. Classification trees. Advantages and disadvantages of decision trees. Bagging. Random forest. Boosting.

Topic 10. Cluster analysis.

Cluster similarity measures. Cluster merging methods. Hierarchical cluster analysis algorithms. Robust k-means method. Visualization.

Topic 11. Time series analysis and forecasting.

Smoothing of time series. Trend and seasonality of time series. Autocorrelation and partial autocorrelation functions. Autoregressive moving average model and estimation of its parameters. Exponential smoothing models with trend and seasonality. Holt-Winters model.

TEACHING AND METHODOLOGICAL MAP OF THE DISCIPLINE

Full-time form of higher education with the use of distance learning technologies (DLT)

Title of section, topic	Title of section, topic	In-class hours					Independent work	Form of control
		Lectures	Workshops	Seminar classes	Laboratory classes	Other		
1	Preliminary data analysis	2		2				Oral test
2	Testing statistical hypotheses	2	2	2				Control work
3	Analysis of the relationship and dependence of features	2	2	2				Individual task, seminar report
4	Multiple linear regression	2	2					Control work, report on the performance of individual task
5	Other aspects of regression models	2	2	2				Control work, individual task, seminar report
6	Methods for creating repeated samples	2	2	2				Control work, report on the performance of individual task
7	Selection and regularization of linear models	2	2	2				Control work, seminar report
8	Dimensionality reduction methods	2	2	2				Control work, report on the performance of individual task
9	Methods based on decision trees	2	2	2				Control work, report on the performance of individual task
10	Cluster analysis	2	4	2				Report on the performance of individual task
11	Time series analysis and forecasting	2	2	4				Control work, individual task, seminar report
TOTAL		**22**	**22**	**22**				

INFORMATION AND METHODOLOGICAL PART

List of basic literature

1. Труш, Н. Н. Введение в компьютерный и интеллектуальный анализ данных: учеб. материалы / Н.Н. Труш. - Минск: БГУ, 2022 - 69 с. - https://elib.bsu.by/handle/123456789/277034.
2. Якимов, А. И. Интеллектуальный анализ данных для имитационного моделирования производственных систем / А. И. Якимов, Е. А. Якимов, Е. М. Борчик ; М-во образования Республики Беларусь, М-во науки и высшего образования РФ, МОУ ВО "Белорусско-Российский ун-т". - Могилев : Белорусско-Российский ун-т, 2021. - 183 с.
3. Нилд, Т. Математика для Data Science : управляем данными с помощью линейной алгебры, теории вероятностей и статистики / Т. Нилд ; [пер. с англ. А. Гаврилов ; науч. ред. К. Кноп, Р. Чебыкин]. - Астана : Спринт Бук, 2025. – 349 с.

List of additional literature

1. Джеймс, Г. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани. – М.: ДМК Пресс, 2016. – 450 с.
2. Буяльская, Ю.В. Введение в компьютерный и интеллектуальный анализ данных: метод. указания / Ю.В. Буяльская, В.В. Казаченок – Минск: БГУ, 2016. – 46 с.
3. Степанов, Р.Г. Технология DATA MINING: Интеллектуальный анализ данных / Р.Г. Степанов – Казань, Издательство Казанского госуниверситета, 2008 – 58 с.
4. Мусаев, А.А. Интеллектуальный анализ данных: учебное пособие / А.А. Мусаев. – СПб.: СПбГТИ(ТУ), 2018. – 176 с.
5. Рафалович, В. Data mining, или Интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович – Москва: Литагент И-Трейд, 2014. – 96 с.
6. Петрунин, Ю.Ю Информационные технологии анализа данных. Data analysis. Изд.4 / Ю.Ю. Петрунин – М.: Издательство МГУ, 2023. – 296 с.
7. Зарова, Е.В. Методы Data mining в обработке и анализе статистических данных (решения в R) / Е.В. Зарова – М.: Инфра-М, 2021. – 232 с.
8. Rutkowski L., Jaworski M., Dudo P. Stream Data Mining: Algorithms and Their Probabilistik Properties. – Springer. 2020. – 340 p.
9. Ghavami P. Big Data Analitics Methods. Analitics Techniques in Data Mining. Deep Learning and Natural Language Processing. – De Gruyter Pub., 2019. – 254 p.
10. Rutkowski L., Jaworski M., Dudo P. Stream Data Mining: Algorithms and Their Probabilistic Properties. Springer, 2020. 340 p.

11. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAD. СПб.: БХВ Петербург, 2007. – 384 с.
12. Han J., Kamber M., Morgan Kaufman. Data Mining: concepts and techniques. – Morgan Kaufman Pub., 2012. – p.324.
13. Hand J., Mannila H., Smyth P. Pricinciples of Data Mining. – MIT Press, 2001. – p.261.
14. Дюк, В. А. Логический анализ данных : учебное пособие / В. А. Дюк. - Санкт-Петербург ; Москва ; Краснодар : Лань, 2020. - 77 с. - https://e.lanbook.com/book/126935

List of recommended diagnostic tools and methodology for final mark formation

The object of diagnostics of students' competences is the knowledge and skills acquired as a result of studying the academic discipline. Identification of students' learning achievements is carried out by means of current and interim certification.
The following means of current certification can be used to diagnose competences: control work, report on the performance of individual task, seminar reports, oral test.

The form of interim certification in the discipline "Data mining methods" in accordance with the curriculum is exam.

A rating system of the student knowledge is used for the final mark formation, which makes it possible to trace and evaluate the dynamics within the process of achieving learning objectives. The rating system stipulates the use of weighting coefficients for current and interim certification of students in the academic discipline.

The final mark formation in the course of control measures for current certification (approximate weighting coefficients determining the contribution of current certification to the mark for passing interim certification) includes:

- performance of control work – 30 %;
- seminar reports – 30 %;
- report on the performance of individual tasks – 40 %.

The final mark for the discipline is calculated on the basis of the mark of current certification (rating system of knowledge) — 60 % and exam mark — 40 %.

Approximate list of workshop

Class № 1. Testing statistical hypotheses.
Class № 2. Analysis of the relationship and dependence of features.
Class № 3. Multiple linear regression.
Class № 4. Other aspects of regression models.
Class № 5. Methods for creating repeated samples.
Class № 6. Selection and regularization of linear models.
Class № 7. Dimensionality reduction methods.
Class № 8. Methods based on decision trees.
Class № 9. Iterative methods of cluster analysis.
Class № 10. Hierarchical methods of cluster analysis.

Class № 11. Time series analysis and forecasting.

Approximate list of seminar classes

Class № 1. Preliminary data analysis.
Class № 2. Testing statistical hypotheses.
Class № 3. Analysis of the relationship and dependence of features.
Class № 4. Other aspects of regression models.
Class № 5. Methods for creating repeated samples.
Class № 6. Selection and regularization of linear models.
Class № 7. Dimensionality reduction methods.
Class № 8. Methods based on decision trees.
Class № 9. Cluster analysis.
Class № 10. Stationary Time Series Analysis and Forecasting.
Class № 11. Exponential Smoothing Methods.

Description of innovative approaches and methods for teaching the discipline

When organizing the educational process, a practice-based approach is used, which entails the following:

- mastering the educational content through solving practical tasks;
- acquiring skills for effective performance in various types of professional activities;
- orientation towards idea generation, implementation of students' group projects, development of business culture;
- use of evaluation procedures, assessment methods, indicating the formation of professional competences.

Methodological recommendations for the organization of independent work

Independent work for the purpose of studying the material of the academic discipline involves working with recommended educational literature and Internet resources. Theoretical information is consolidated by completing laboratory assignments, during which one should be guided by the methodological developments posted in the electronic library of the university and on the educational portal. Additional assignments (tests, assignments for independent completion) may also be offered for self-assessment and deeper assimilation of the material received.

Sample list of questions for the exam

1. Testing the sample for homogeneity, randomness. Detecting anomalous observations. Robust estimates of position and scale.

2. Data visualization. Determining the distribution law
3. Principles of testing statistical hypotheses. Testing hypotheses about the equality of means and variances of independent and dependent samples.
4. Robust criteria for testing hypotheses.
5. Nonparametric criteria for testing hypotheses.
6. Multiple hypothesis testing.
7. Sample correlation coefficients. Testing the significance of sample correlation coefficients.
8. Robust estimates of correlation coefficients.
9. Contingency tables.
10. Estimation of regression equation coefficients. Model adequacy testing.
11. Residual analysis. Multicollinearity problem.
12. Qualitative predictors. Predictors with only two levels.
13. Predictors with more than two levels.
14. Extensions of the linear model. Nonlinear relationships.
15. Cross-validation. Validation sample method.
16. K-fold cross-validation. Bootstrap.
17. Optimal subset selection. Stepwise selection. Compression methods.
18. Ridge regression. Lasso regression. Hyperparameter selection.
19. Principal Component Regression.
20. Partial Least Squares.
21. High-Dimensional Data. Regression for High-Dimensional Data.
22. Basic concepts. Regression trees. Classification trees. Advantages and disadvantages of decision trees.
23. Bagging. Random forest. Boosting.
24. Cluster similarity measures. Cluster merging methods. Hierarchical cluster analysis algorithms. Robust k-means method. Visualization.
25. Smoothing of time series. Trend and seasonality of time series.
26. Autocorrelation and partial autocorrelation functions.
27. Autoregressive moving average model and estimation of its parameters.
28. Exponential smoothing models with trend and seasonality.
29. Holt-Winters model.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Учебная дисциплина не требует согласования			

Заведующий кафедрой теории
вероятностей и математической статистики,
доктор физ.-мат. наук, профессор А.Ю.Харин

27.05.2025

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ УО

на _____/_____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 202_ г.)

Заведующий кафедрой
_____ _____ _____

УТВЕРЖДАЮ
Декан факультета
_____ _____ _____