Ministry of education and science of Ukraine
V. N. Karazin Kharkiv National University

Department of economic theory and economic methods of management

APPROVED BY
Dean of the faculty of economics

Vitalii DIACHEK

"28" 08 2025

WORK PROGRAM OF THE ACADEMIC DISCIPLINE

Intelligent data analysis systems

Level of higher education: Third (Educational and scientific / PhD level)

Field of knowledge: C – Social Sciences, Journalism, Information and International Relations

Specialty: C1 Economics and International Economic Relations

Educational program: Economics

Specialization: —

Course type: mandatory

Faculty: Economical

Academic year: 2025 / 2026

The program is recommended for approval by the academic council of the faculty of economics

Minutes No. 18, August 26, 2025,

PROGRAM DEVELOPERS: Tamara MERKULOVA, professor of department of economic cybernetics and applied economics

Rostyslav LUTSENKO, associate professor of department of economic cybernetics and applied economics

The program was approved at the meeting of the department of economic cybernetics and applied economics

Minutes No. 1 dated August 26, 2025

Head of department of economic cybernetics and applied economics

_____ _____ Tamara MERKULOVA

The program is agreed with the guarantor of the educational-scientific program «Economics»

Guarantor of the educational-professional program «Economics»

Volodymyr SOBOLIEV

The program was approved by the scientific and methodological commission of the faculty of economics

Minutes No. 1 dated August 26, 2025

Head of the scientific and methodological commission of the faculty of economics

Daria ZAHORSKA

INTRODUCTION

Academic discipline program «Intelligent data analysis systems» is developed in accordance with the educational and scientific program Economics for the training of PhD students

Level of higher education: Third (Educational and Scientific)
Field of knowledge: C – Social Sciences, Journalism, Information and International Relations
Specialty: C1 Economics and International Economic Relations

1. Description of the academic discipline

1.1. The purpose of teaching the academic discipline: forming a system of theoretical knowledge and practical skills in modern data analysis methods.

1.2. The main objectives of studying the discipline: mastering the principles, methods and tools of data analysis and methods for solving typical business data analysis problems using machine learning technology.

1.3. Number of credits – 3 credits.

1.4. The total number of hours – 90 hours.

1.5. Characteristics of the academic discipline	
Normative / optional Optional	
Full-time study	Correspondence (distance) form of study
Year of preparation	
1st	1st
Semester	
2nd	2nd
Lectures	
16 hours	6 hours
Laboratory classes	
14 hours	2 hours
Independent work, including	
60 hours	82 hours
Individual tasks	
5 hours	

1.6. The list of competencies that this discipline forms:
GC02. Ability to search, process, and analyze information from various sources.
PC03. Ability to use modern methodologies, methods and tools of empirical and theoretical research in the field of economics, computer modeling methods, modern digital technologies, databases and other electronic resources, specialized software in scientific and scientific and pedagogical activities.

PC06. The ability to justify and prepare economic decisions based on an understanding of the patterns of development of socio-economic systems and processes using mathematical methods and models.

1.7. The list of learning outcomes that this discipline forms:
PLO03. Develop and research fundamental and applied models of socio-economic processes and systems, effectively use them to obtain new knowledge and/or create innovative products in economics and related interdisciplinary areas.
PLO04. Apply modern tools and technologies for searching, processing and analyzing information, in particular, statistical methods for analyzing large data sets and/or complex structures, specialized software and information systems.
PLO09. Formulate and test hypotheses; use appropriate evidence to substantiate conclusions, in particular, the results of theoretical analysis, empirical research and mathematical and/or computer modeling, available literary data

1.8. Prerequisites: basic computer science course

2. Thematic plan of the academic discipline

Section 1. Basic concepts of intelligent data analysis systems

This section covers the fundamental concepts and basic approaches of data mining as a component of modern information technologies, machine learning, and artificial intelligence. It explains the role of data analysis in managerial decision-making, identifies the main types of data mining tasks, and highlights the importance of data preparation in building accurate and robust models.

Topic 1. Data analysis tasks

This topic focuses on the key problems solved using machine learning and data mining methods, including:

- prediction — establishing functional relationships between input variables and continuous output variables, typically implemented through regression tasks (e.g., forecasting sales volumes based on economic factors);
- classification — assigning objects to predefined classes based on discrete output variables (e.g., customer segmentation, credit risk assessment);
- clustering — grouping objects based on similarity to identify homogeneous groups and patterns in large datasets;
- association rule mining — identifying relationships between events and objects, particularly for market basket analysis and behavioral pattern discovery.

The topic also examines the relationships between these tasks and the main machine learning algorithms used to solve them.

Topic 2. CRISP-DM Methodology

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) is presented as a widely accepted methodology for organizing data analysis projects.
Within this topic, the data analysis process is structured as a sequence of logically interconnected stages:

1. business understanding and formulation of research objectives

2. data understanding and initial analysis
3. data preparation and cleaning
4. model building
5. model evaluation
6. deployment of results

This methodology ensures a systematic approach to analysis and alignment between technical solutions and applied research objectives, which is critical in economic and financial applications.

Topic 3. Data preparation

This topic focuses on practical aspects of data preparation that determine the quality and reliability of models. Key stages include:

- encoding variables, including handling numerical and categorical data using basic and advanced encoding techniques;
- calculation of descriptive statistics and analysis of distributions;
- detection and treatment of outliers, considering their impact on model performance;
- data scaling to ensure comparability of variables;
- handling missing values using different methods (deletion, mean/median imputation, most probable values);
- feature selection to reduce model complexity and prevent overfitting;
- cross-validation for evaluating model generalization ability.

It is emphasized that data preparation quality largely determines the accuracy and interpretability of results.

Section 2. Regression analysis

This section examines theoretical foundations and practical methods of regression analysis as one of the main tools for modeling relationships between variables. It describes principles of building regression models, parameter estimation, and model validation using real economic data, as well as the application of modern machine learning methods for modeling complex nonlinear relationships.

Topic 4. Linear Regression Models

This topic covers linear regression as a fundamental method for predicting quantitative dependent variables.

It includes simple (univariate), multiple (multivariate), and polynomial linear regression models, along with their mathematical representation and economic interpretation. Special attention is given to the statistical assumptions of classical linear regression, including homoscedasticity and the absence of autocorrelation, as well as methods for evaluating model quality. Key evaluation metrics include the coefficient of determination (R^2), mean squared error (MSE), and statistical significance tests for both individual parameters and the model as a whole.

The topic is complemented by practical implementation in R, analysis of forecasting results, and detection of overfitting.

Topic 5. Machine learning methods: decision trees and random forest

The topic covers methods for building nonlinear regression and classification models based on decision trees and ensemble approaches. It provides a detailed analysis of the structure of a decision tree as a system of hierarchical «if–then» rules,

the principles of data splitting based on information gain, and the use of entropy to evaluate the quality of splits.

Special attention is given to methods for improving model accuracy using ensemble algorithms, in particular bagging and random forests. The role of bootstrap samples in reducing model variance is explained, along with the mechanism of random feature selection in random forests and its advantages in preventing overfitting.

The topic is complemented by practical examples of building decision trees and random forests in the R environment, analysis of forecasting quality, and comparison with classical regression models.

Section 3. Classification

This section is devoted to the study of classification methods as one of the key areas of data mining. It covers both statistical and algorithmic approaches to building models designed to assign objects to predefined classes. The theoretical foundations of each method are analyzed, along with their advantages, limitations, and areas of application, as well as the practical implementation of these models in the R environment.

Topic 6. Logistic regression

This topic is devoted to logistic regression as a fundamental statistical method for binary and multiclass classification. It examines the logistic function, its properties, and the interpretation of the probability of an object belonging to a particular class. Methods for estimating model parameters, classification quality criteria, error analysis, and the construction of ROC curves are studied. Special attention is given to the application of logistic regression in economic and financial tasks, particularly in risk management and customer behavior prediction.

Topic 7. Support vector machines (SVM)

This topic explores support vector machines as a classification tool based on finding the optimal hyperplane that maximizes the distance between classes. Linear and nonlinear SVM models, the role of kernel functions, and regularization parameters are analyzed. The advantages of the method for complex, high-dimensional, and noisy data are explained, along with practical aspects of its application in real-world tasks.

Topic 8. K-Nearest neighbors (KNN)

This topic focuses on the K-nearest neighbors method, a simple and intuitive classification algorithm based on distances between objects. Distance measurement methods, the selection of the optimal K parameter, and the impact of feature scaling on classification results are discussed. Advantages and limitations of the method, particularly its sensitivity to noise and computational complexity for large datasets, are analyzed.

Topic 9. Bayesian classification

This topic studies the principles of building classifiers based on Bayes' theorem. The Naive Bayes classifier, its assumptions about feature independence, and practical applications for text processing, spam filtering, and behavioral data analysis are examined. The effectiveness of Bayesian methods for small training samples and noisy data is analyzed.

Topic 10. Decision trees and random forest

This topic is dedicated to classification algorithms based on decision trees and ensemble methods. The process of building a classification tree using entropy and information gain, recursive partitioning principles, and methods to prevent overfitting are discussed. The random forest algorithm is studied separately as an ensemble of decision trees that enhances prediction accuracy and model robustness against noise and sample variability.

Section 4. Clustering

This section focuses on clustering methods as an important tool for data mining to identify hidden structures in large datasets without predefining classes. Principles of cluster construction, object similarity criteria, and methods for evaluating clustering quality are examined. The role of clustering in customer segmentation, identifying typical behavioral patterns, and supporting managerial decisions is analyzed.

Topic 11. Hierarchical clustering

This topic covers hierarchical clustering methods, which are based on sequential merging or splitting of objects according to their similarity. Agglomerative and divisive approaches, distance computation between objects and clusters, as well as dendrogram construction and interpretation are discussed. The advantages of hierarchical clustering, particularly the ability to visually analyze data structure and the absence of a need to predefine the number of clusters, are analyzed.

Topic 12. K-Means Clustering

This topic focuses on the k-means algorithm, one of the most widely used clustering methods. The working principle, iterative update of cluster centers, and convergence criteria are discussed. The impact of the number of clusters on result quality, methods to determine the optimal k value, and algorithm limitations, such as sensitivity to outliers and initial center placement, are analyzed. Special attention is given to practical applications for economic and financial data analysis.

3. Structure of the academic discipline

Section names	Number of hours											
	full-time education						correspondence education					
	total	including					total	including				
		1	p	lab.	ind.	i. w.		1	p	lab.	ind.	i. w.
1	2	3	4	5	6	7	8	9	10	11	12	13
Section 1. Basic concepts of intelligent data analysis systems												
Total under section 1	22	4		2		14	22	2				20
Section 2. Regression analysis												
Total under section 2	20	4		4		13	20	1		1		18
Section 3. Classification												
Total under section 3	24	4		4		17	24	1				23
Section 4. Clustering												
Total under section 4	24	4		4		16	24	2		1		21
Total hours	**90**	**16**		**14**		**60**	**90**	**6**		**2**		**82**

4. Topics of laboratory classes (classroom)

№	Topic name	Number of hours
1.	Data preparation	2
2.	Linear regression	2
3.	Decision trees and random forest	2
4.	Logistic regression	2
5.	Support Vector Method, K Nearest Neighbors, and Bayesian Classification	2
6.	Hierarchical clustering	2
7.	Clustering based on k-means	2
	Total	**14**

5. Tasks for independent work (MOODLE)

№	Types and content of independent work	Number of hours full-time education
	Section 1	
1.	Collection and preparation of data on the selected topic	14
	Total	*14*
	Section 2	
2.	Data analysis using linear regression	6
3.	Data analysis using decision trees	7
	Total	*13*
	Section 3	
4.	Data analysis using logistic regression	5
5.	Data analysis using support vector, K-nearest neighbors, and Bayesian classification methods	6
6.	Data analysis using decision trees	6
	Total	*17*
	Section 4	
7.	Data analysis using hierarchical clustering	8
8.	Data analysis using k-means	8
	Total	*16*
	Total	**60**

6. Individual tasks (independent work)

Data analysis on a chosen topic. Each student collects data independently on a selected topic, performs preliminary analysis, and prepares the data. According to the structure and scope of the data, research is conducted in the following directions:

- Regression model building and analysis (16 points)
- Classification model building and analysis (16 points)
- Clustering model building and analysis (16 points)

Each assignment includes completing corresponding laboratory work followed by a report-presentation. Based on the results, students prepare and defend their report-presentation.

7. Teaching methods

Forms and methods: lectures, practical problem-solving classes, computer modeling, laboratory work, presentations and report discussions, independent work, and study of educational

materials and literature. Learning is based on the use of information technology and computer modeling, modern information search technologies, and development of skills in working with large datasets.

The programming language used is R, with the R Studio development environment. Default libraries: library(base), library(datasets), library(graphics), library(grDevices), library(methods), library(stats), library(utils). Additional libraries: data preparation: library(dplyr), library(ggplot2), library(psych), library(caTools). Regression: library(rpart), library(randomForest). Classification: library(ROCR), library(ElemStatLearn), library(e1071), library(class). Clustering: library(cluster).

8. Assessment methods

Assessment methods include ongoing monitoring, completion of independent assignments, and a control test. Final assessment is conducted as a written test, which includes practical tasks (2 tasks, 40 points).

9. Grading scheme

Current control, independent work, individual tasks			Examination paper (written)	Total
Independent work (laboratory work)	Test work provided for in the curriculum	Total		
48	12	60	40	100

Evaluation criteria:

1) Control work (12 points) consists of 2 practical tasks:
   1. Data preparation
   2. Linear regression

   Each task requires completing a lab exercise and preparing a combined report.
2) Ongoing assessment includes 6 lab assignments (individual tasks, each worth 8 points).

Evaluation		Evaluation criteria
Test work	Laboratory work	Failure to meet the deadlines for submitting tasks for verification may result in a score reduction of up to 30% of the total points.
12-11	8-7	The task was completed in full and clearly, in compliance with all requirements.
10-9	6-5	The task was completed in full, minor errors were made.
8-6	4-3	The task was not completed completely, significant errors were made.
5-0	3-0	Less than 50% of the task was presented, gross errors were made.

3) Final Test (40 points)

   Consists of 2 practical tasks.

   Each task report must include:
   - A script with loaded and error-free exam data
   - A report file in MS Word format «Surname.docx»

Number of points	Evaluation criteria
36-40	The higher education applicant correctly chooses the method of solving the problem, possesses versatile abilities, skills and techniques for solving

	problems. The task is completed without errors. The report contains detailed author's comments. The script code is clearly structured
30-35	The higher education applicant correctly applies theoretical knowledge and provisions when solving a practical problem, possesses the necessary skills and abilities to work with programs. Completed the task with some minor errors. The report contains concise author's comments. The script code is clearly structured
24-29	The higher education applicant correctly applies theoretical knowledge and provisions when solving a practical problem, possesses the necessary skills and abilities to work with programs. Completed the task with some minor errors. The report does not contain author's comments. The script code is not clearly structured
18-23	A higher education student made a significant mistake while solving a practical problem. He is not sufficiently fluent in the skills and abilities to work with programs. The report does not contain author's comments. The script code is not clearly structured.
12-17	A higher education student made a significant mistake while solving a practical problem. The report was not generated. The script code contains minor errors
6-11	A higher education student made a significant error while solving a practical problem. The report was not generated. The script code contains significant errors
0-5	The higher education applicant is unable to apply knowledge in practice. Did not solve the task at all or made gross errors. The report was not generated. The script code contains significant errors

Grading Scale

Total points for all types of learning activities during the semester	Evaluation
	for a four-level rating scale
90–100	excellent
70–89	good
50–69	satisfactorily
1–49	unsatisfactorily

10. Recommended Literature

Core literature:

1. Kononova, K. (2020). Machine learning: Methods and models: Textbook for bachelor's, master's, and PhD students in specialty 051 "Economics". V. N. Karazin Kharkiv National University.

2. Wei Fang. Data Mining and Machine Learning with Applications" — MDPI AG, 2024. https://freecomputerbooks.com/Data-Mining-and-Machine-Learning-with-Applications.html

3. Havrylenko, S. Yu. (2024). Machine learning: Lecture notes [Electronic resource]. National Technical University "Kharkiv Polytechnic Institute". URL: https://repository.kpi.kharkov.ua/handle/KhPI-Press/80167.

4. Nikolaieva O.H., Lutsenko R.R. *Economic-Mathematical Methods and Models, Chapter 2: Econometric Methods and Models*. Kharkiv: V.N. Karazin KhNU, 2024. 108 p.

Supplementary Literature:

5. Kateryna Kononova, Rostyslav Lutsenko. COVID-19 Measures Sentiment Analysis Based on a Social Network Dataset. Proceedings of the Workshop on the XIII International Scientific Practical Conference Modern problems of social and economic systems modelling (MPSESM-W 2021). Kharkiv, Ukraine, April 9, 2021. P 8-17. http://ceur-ws.org/Vol-2927/paper2.pdf
6. Danich V., Lutsenko R. Virtual assets of the distributed register. Bulletin of V. N. Karazin Kharkiv National University Economic Series. 2023. № 104. С. 5 –10. DOI: https://doi.org/10.26565/2311-2379-2023-104-01.
7. Olena Nikolaieva, Anzhela Petrova, Rostyslav Lutsenko. (2020). Forecasting of the stock rate of leading world companies using econometric methods and dcf analysis. International Journal of Innovative Technologies in Economy, (2(29), 33-41. https://doi.org/10.31435/rsglobal_ijite/31052020/7067
8. Rokach, L., Oded, M. (2005) Clustering methods. Data mining and knowledge discovery handbook. Springer US. https://download.e-bookshelf.de/download/0000/0002/17/L-G-0000000217-0002331690.pdf
9. Hurianova, L. S., & Lutsenko, R. R. (2024). Models for analyzing cryptocurrency market dynamics considering stakeholders' behavioral metrics based on social media data. Business Inform, (9), 129–138. DOI: https://doi.org/10.32983/2222-4459-2024-9-129-138
10. Merkulova, T. V. (2024). Behavioral economics and machine learning methods in managing a hybrid investment portfolio with virtual assets. Business Inform, (12), 270–276. DOI: https://doi.org/10.32983/2222-4459-2024-12-270-276
11. Merkulova, T., & Kosiashvili, D. (2024). Analysis of poverty indicators using the PPI methodology. Bulletin of V. N. Karazin Kharkiv National University. Economic Series, (106), 5–14. DOI: https://doi.org/10.26565/2311-2379-2024-106-01
12. Merkulova, T. V., & Nikolaieva, O. H. (2022). Cluster analysis of tax indicators in European countries. Bulletin of V. N. Karazin Kharkiv National University. Economic Series, (102), 69–81. DOI: https://doi.org/10.26565/2311-2379-2022-102-08
13. Merkulova, T., Bohdanova, H. (2021) Determinants of social trust: Analysis using machine learning methods. Machine Learning Methods and Models, Predictive Analytics and Applications. CEUR Workshop Proceedings. Volume 2927, 2021, Pages 108-124. Scopus URL: https://ceur-ws.org/Vol-2927/

Links to information resources on the Internet, video lectures, other methodological support

14. http://mymagictools.blogspot.com/2015/07/r.html?view=classic
15. http://re9ulus.github.io/2015/12/07/trees_in_r/
16. https://ranalytics.github.io/data-mining/105-Cohonen-Maps.html
17. https://clarkdatalabs.github.io/soms/SOM_NBA
18. https://www.shanelynn.ie/self-organising-maps-for-customer-segmentation-using-r/
19. https://ranalytics.github.io/data-mining/105-Cohonen-Maps.html
20. https://cran.r-project.org/web/packages/
21. https://freecomputerbooks.com