**Syllabus**

*Course Program*

# METHODS OF MULTIDIMENSIONAL ANALYSIS AND BIGDATA IN SOCIOLOGY

*Specialty*
*054 - Sociology*

*Educational program*
*Sociological support of economic activity*

*Level of education*
*Master's level*

*Semester*
*2*

*Institute*
*Institute of Social and Humanitarian Technologies*

*Department*
*Sociology and Public Administration (305)*

*Course type*
*Special (professional), Mandatory*

*Language of instruction*
*English, Ukrainian*

## Lecturers and course developers

**Dina Akramivna Tereshchenko**

**dina.tereshchenko@khpi.edu.ua**

*Doctor of Sciences in Public Administration, Professor, Professor, Professor of Department of sociology and public administration, National Technical University «Kharkiv Polytechnic Institute», Kharkiv (NTU "KhPI").*
*Author of over 200 scientific and educational-methodical publications.*
*Leading lecturer for courses such as "State and Regional Governance," "Administrative Management," and "Public Relations."*
*More about the lecturer on the department's website.*
*http://web.kpi.kharkov.ua/sp/profesors-ko-vikladats-kij-sklad*

## General information

### Summary

*The main tasks of the course are: studying methods of scientific research from the theory of organization of sample observations, processing and analysis of the obtained information, application of multidimensional methods and bigdata for social analysis, identification and recognition of patterns; modeling and forecasting social processes; using information technologies for statistical justification of decision making in sociological support of economic activity.*

### Course objectives and goals

*Mastering the methodological and methodical foundations of using methods of multidimensional analysis and bigdata for studying the nature of social phenomena, for building multidimensional models of existence and functioning of social objects.*

### Format of classes

*Lectures, laboratory classes, consultations, self-study. Final control in the form of an exam.*

### Competencies

*GC05. Ability to estimate and support quality of the performed work.*

*SC02. Ability to detect, diagnose and interpret social problems of Ukrainian society and the global community.*
*SC03. Ability to design and fulfill sociologic research, to develop and substantiate their methodology.*
*SC04. Ability to collect and analyze empirical data with the use of present-day sociologic research methods and digital technologies.*
*SC07. Ability to design and evaluate social projects and programs.*

### Learning outcomes

*PR01. To analyze social phenomena and processes using empirical data and present-day concepts and theories in sociology.*
*PR02. To perform diagnostics and interpretation of social problems of Ukrainian society and the global community, of the causes for their arising and their consequences.*
*PR03. To develop and implement social and interdisciplinary projects with accounting for social, economic, legal, environmental, and other aspects of social life.*
*PR04. To apply scientific knowledge, sociological and statistical methods, digital technologies, specialized software for solving complex tasks in sociology and conterminal knowledge areas.*
*PR05. To carry out search for, to analyze and estimate the needed information in scientific literature, databases and other sources.*

### Student workload

*The total volume of the discipline is 180 hours (6 credits ECTS): lectures - 32 hours, seminar classes - 32 hours, independent work - 116 hours*

### Course prerequisites

*To successfully complete the course, you need to have knowledge and practical skills in the following disciplines: "Mathematical Methods in Sociology", "Workshop on Computer Processing of Sociological Data", "Sociological Support of Economic Activity". "Internet Research of Economic Activity".*

### Features of the course, teaching and learning methods, and technologies

*During the practical classes of the academic discipline, it is envisaged to explain the algorithm of performing practical tasks and their working out. The following methods of learning are used: explanatory-illustrative; reproductive (working out certain algorithms of data analysis); partially-search or heuristic method (when performing individual tasks). The project approach to learning, gamification, attention is focused on the use of information technologies in the organization of sociological research: project and team work, peer-to-peer, cases.*

## Program of the course

### Topics of the lectures

*Topic 1. Basic elements of formalism Analysis of sociological information collected in the course of empirical sociological research is not just a set of technical techniques and methods.*
*Non-unidimensionality of many concepts studied by a sociologist. Its indirect manifestation - violation of transitivity of the order relation. Metric and non-metric MDS. Corresponding stress functions. Implicit comparison of distances between proximity, embedded in the formula of the stress function for metric scaling. The concept of monotonic regression, used in the calculation of the stress function for non-metric scaling. The importance of non-metric scaling for sociology. Formal aspects of the problems of dimensionality of the sought-after Euclidean space and rotation that define its coordinate axes.*

*Topic2. Multidimensional unfolding and individual multidimensional scaling*
*Problem statement; the importance of taking into account the specifics of metrics of individual respondents. The way of accounting for such metrics in individual MDS. Type of input and output data, stress functions in individual MDS. One-dimensional unfolding. Justification of the need to move to a space of arbitrary dimensionality for successful completion of the scaling task. The model of the ideal point in the multidimensional case. Non-metric multidimensional unfolding. Type of output data. Stress function.*

Specificity of output data (presence of two types of points, corresponding to objects and respondents respectively). Features of interpretation of results.

### Topic3. Problems of forming output data and interpreting results in multidimensional scaling

The role of a sociologist in obtaining data, output for multidimensional scaling, and interpreting its results. Possible ways of obtaining output data. Direct receipt of proximity from respondents, classification of relevant methods of questioning; problems that arise with this method of data collection. Examples of calculating the proximity matrix based on the analysis of sufficiently reliable data of another kind. Working with MDS statistical programs - the MDS procedure is available in most statistical programs. There is a choice between metric MDS (which allows working with intervals or data on the ratio level), and non-metric MDS (which works with ordinal data). Using formal and informal methods in interpreting the results of multidimensional scaling. The value of the substantive concepts of the researcher in solving the problems of choosing the dimensionality of the Euclidean space and rotating its axes.

### Topic 4. Canonical analysis. General idea of methods based on frequency models General idea of modeling frequency of contingency table.

Substantive understanding of such models, their role for sociologist. Multiplicative and additive frequency models. The role of logarithmization of multiplicative model. The possibility of different understanding of both the meaning of the considered contributions and the “average” level, with which the observed frequencies are compared in the process of their modeling. Canonical correlation analysis - one of the methods of multidimensional data analysis. The need to combine the model, embedded in a specific method of digitization, with the content of the considered problem. An example of such a model - a model used in the scaling method, called the method of sequential splits. Canonical analysis as a method of digitization and a method of measuring the relationship between two nominal features with “common alternatives”. Frequency models corresponding to canonical analysis. Construction of sociological indices using the technique of canonical analysis. Solving the problem of weighting the components of the index feature.

### Topic 5. Loglinear analysis Loglinear analysis - a method of multidimensional statistical analysis for studying contingency tables.

Loglinear analysis allows to statistically test the hypothesis about the system of simultaneous pairwise and multiple relationships in a group of features measured by nominal scales. Multidimensional statistical analysis. Frequency models corresponding to loglinear analysis. Saturated model. The purpose of moving to logarithms of frequencies. The meaning of contributions of different dimensionality. Different understanding of the dependent feature: quantitative feature in variance analysis, quantitative or nominal - in nominal regression and frequency, standing in the cell of multidimensional contingency table, - in loglinear analysis. Different possibilities of searching for combinations of predictor values: testing hypotheses about the presence of multidimensional relationships in loglinear analysis and the possibility of searching for the most effective combinations in the method of sequential splits and regression analysis, predetermined set of combinations of predictor values in variance analysis.

### Topic 6. Causal analysis. Strategy of analysis of the structure of relationships of features

The concept of cause in sociology. The fundamental impossibility of fully formalizing it. The role of statistical methods in studying causal relations. Graph of causal links. Structural coefficients. Input (external, independent) and output (internal, dependent) variables. Rules for reducing causal schemes and forming equations. Repetition of the principles of constructing partial correlation and regression coefficients. The importance for sociologist of studying the relevant relationships. The difference between statistical and causal relationship. The concept of “erroneous” correlation. The main causal schemes that lead to their appearance. The problem of formalization of the task of studying causal and consequential relations in sociology. The concept of the structure of a multidimensional random variable. Formation of generalized indicators based on the analysis of the structure of relationships of features. Complex use of several methods of studying the relationships between features to solve sociological problems (analysis of the structure of a random variable; factor and variance analysis; search for determining combinations of predictor values).

### Topic 7. Pattern recognition tasks.

The concept of automatic object classification Classification as one of the fundamental processes in science. Feature space. The task of classification as a search for clusters of points - models of objects in the feature space. Highlighting the task of automatic object classification (synonyms: multidimensional classification, unsupervised pattern recognition, cluster analysis, taxonomy). Classification as one of the fundamental *processes in science. Feature space. The task of classification as a search for clusters of points - models of objects in the feature space. Highlighting the task of automatic object classification (synonyms: multidimensional classification, unsupervised pattern recognition, cluster analysis, taxonomy).*

### *Topic 8. The problem of "matching" content and formalism when using classification algorithms*

*The specificity of solving sociological tasks of building typology using methods of automatic classification. The meaning of contrasting the terms "classification" and "typology". The basis of typology. The role of a priori ideas of the researcher about the types sought in the choice and implementation of the algorithm, interpretation of the results of its application. Identification of the main formal elements of automatic classification algorithms that require matching with the substantive concepts of the sociologist.*

### *Topic 9. Distance functions between objects Axiomatic definition of the distance function and the role of this function in sociology.*

*Examples of unsuitability of Euclidean distance from the point of view of a priori substantive understanding of types of objects. The possibility of using Euclidean distance in the considered examples due to the change of feature space. Modern data analysis is determined by the ways of obtaining values, methods of their processing and depends on the development of mathematical methods and modeling. Distance functions different from Euclidean: weighted Euclidean, city-block, Mahalanobis, Hamming.*

### *Topic 10. Main types of classification procedures.*

*Distances between classes The relevance of studying the essence and methods of multidimensional analysis of sociological information is determined by the specificity of social reality, which always appears as a complex, multifaceted and multivalued phenomenon, which integrates the multidimensionality of society with the multidimensionality of the inner world of an individual. Identification of hierarchical and non-hierarchical classification algorithms. Multidimensional statistical analysis (in a broad sense) - a branch of mathematical statistics that combines methods of studying data that characterize multidimensional objects. Agglomerative and divisive algorithms. The reasons for the need to consider distances between classes in hierarchical procedures. Nearest neighbor algorithm as an example of a classification method that uses such distances.*

### *Topic 11. Hypotheses about the location of objects in the feature space*

*The role of hypotheses about the nature of the location of objects in the choice of classification algorithm. The conditionality of these hypotheses by the a priori ideas of the researcher about the types of objects. The main types of hypotheses: compactness, connectivity (continuity), unimodal distribution. Factor analysis most vividly reflects the features of multidimensional analysis in the part of studying the relationship between features. Cluster analysis reflects these features from the side of object classification. General idea of fuzzy classifications. The role of the membership function in the corresponding algorithms. The expediency of complex use of several classification algorithms in sociological tasks of building typology. Substantive ideas of the sociologist about types and conditions of choosing the step of splitting in the interpretation of results. Adjusting the results of classification to ensure the correspondence of classification and typology.*

### *Topic 12. The concept of interpretation of output data and the main methodological principles of using data analysis methods in sociology*

*Interpretation of output data as one of the main links of "matching" sociology and mathematics. The main factors that determine the interpretation of output data: a priori ideas of the researcher about the way of generating these data (including - about the models of perception by the respondents of the questions, objects, proposed by him, about the probabilistic nature of the data, etc.); the purpose of the study; conceptual ideas of the sociologist about the phenomenon under study; the nature of the model of the phenomenon, "embedded" in the mathematical method, the use of which is planned; consideration of observed variables as indirect indicators of latent factors, which actually interest the researcher, etc. Identification of methodological principles, compliance with which is necessary for the analysis of sociological data to be effective, not to divert the sociologist away from reality: ensuring a certain homogeneity of output data; taking into account the model, "embedded" in each method of data analysis, when choosing the algorithm of analysis, two main principles of interpretation of the results of analysis: the need to harmonize it with the interpretation of output data and filling in the losses that occurred during the transition to formalism; the need for complex use of several methods to solve one problem, etc.*

*Topic 13. Data. Metadata According to GOST, data - presentation of information in a formalized form, suitable for transmission, interpretation and processing.*

*The original concept of data - philosophical, it arises in epistemology when considering the main problem of gnoseology - the cognizability of the world, the search and comprehension of truth. Procedures of verification or falsification of data create information, comprehension of truth creates knowledge. The life cycle of data - a sequence of stages that a specific portion of data goes through from the initial stage of creation or receipt to the moment of archiving or deletion. When collecting data, metadata is generated, which contains any information about the collected data. Overview of the main analytical tools for working with Bigdata in social sciences (Python, R, SAS, etc.). Reading and writing data, file formats. Downloading data from various sources. Interaction with databases. Reading data from Excel. Working with CSV files and data in JSON format. Parsing simple XML data. Reading data from HTML tables. Reading data from SAS file. Interaction with HTML and Web API.*

*Topic 14. Big data. Big data management systems*

*Big data can be of different types. Information obtained as a result of accounting or measurement of any objects or parameters is called master data (MasterData). For example, accounting of quantity, measurement of coordinates and speeds of specific molecules - these are master data. Transactional data (in English literature, the terms TransactionalData, ApplicationSpecificData, OperationalData are used) - these are data that reflect the result of performing any operations. Transactional data describe the interaction of objects with each other or with the surrounding world, which can be obtained by processing master data. Retrospective data (Historicaldata) - these are data provided with a timestamp. Reference data (reference books, NSI, normative-reference information, ReferenceData, LookupData, Dictionaries) - these are basic immutable data, previously known from external sources, such as norms, abbreviations, acronyms, dictionaries, standards. Data format. Structured data have a predefined format. Semi-structured or weakly structured data - these are data that are often collected from different sources.*

*Topic 15. Software platforms and systems for Big Data*

*Currently, a large number of platforms and systems for Big Data are used. Big data processing systems are frameworks, i.e. frameworks, for the use of which it is necessary to combine them with other frameworks, user application software and data storage system. The analytical report BigDataAnalyticsMarketStudy, 2017 Edition presents such a diagram of Big Data infrastructures implemented in enterprises, presented by size of enterprises Distributed data processing is closely related to parallel data processing. However, such processing is always performed using separate machines in a cluster connected to a network. Distributed data processing is a method of executing application programs by a group of systems. The user can work with network services and application processes located in several interconnected subscriber systems. Distributed data processing increases the efficiency of user information needs and ensures the efficiency and effectiveness of decisions.*

*Topic 16. Machine learning using the Scikit-learn library.*

*Types of machine learning. Main machine learning libraries Python (Scikit-learn, Keras, TensorFlow). Creating training sets - preprocessing data. Accuracy and reliability of the model. Choosing the best model. Steps of a typical practical machine learning scenario. Loading a data set. Exploring data using Pandas. Visualization of features using Matplotlib. Splitting data for training and testing. Creating a model. Learning the model. Testing the model. Setting the parameters of the model and evaluating its accuracy. Making predictions based on "live" data that is still unknown to the model. Functionality of the Scikit-Learn library. Classification using K-neighbors. Linear models for regression and classification (linear regression model, logistic regression, etc.). Naive Bayesian classifiers. Decision trees and random forest. Support vector method. Basics of neural networks. Principal component method. Clustering algorithms (clustering by K-means, hierarchical clustering, etc.).*

## Topics of the workshops

*Topic1. Basic elements of formalism*

*Problems of non-unidimensionality of many concepts studied by the sociologist. Features of studying the space of perception of sociological phenomena and processes - the main task of MDS. Kumb's ideas on taking into account the possibility of ordering distances between objects. Vector model or ideal point model as the basis of MDS. Distance function (axiomatic definition). Corresponding stress functions. Space of perception* *by respondents of the objects proposed to them. Formal definition of proximity. Output data for MDS - matrix of proximity between objects. Metric and non-metric MDS. Formal aspects of the problems of dimensionality of the sought-after Euclidean space and rotation that define its axes. Solving practical problems.*

### *Topic2. Multidimensional unfolding and individual multidimensional scaling*

*Setting the task of the importance of taking into account the specifics of metrics of individual respondents. Type of input and output data, stress function in individual MDS. One-dimensional unfolding. Justification of the need to move to a space of arbitrary dimensionality for successful completion of the scaling task. Non-metric multidimensional unfolding. Features of interpretation of results. Method of accounting for such metrics in individual MDS. Ideal point model in the multidimensional case. Stress function. Specificity of output data (presence of two types of points corresponding to objects and respondents respectively). Solving practical problems.*

### *Topic3. Problems of forming output data and interpreting results in multidimensional scaling*

*The role of the sociologist in obtaining data, output for multidimensional scaling and interpretation of its results. Classification of relevant methods of questioning; problems that arise with such a method of data collection. Examples of calculating the proximity matrix based on the analysis of sufficiently reliable data of another kind. Use of formal and informal methods in the interpretation of the results of multidimensional scaling. The value of substantive concepts of the researcher in solving the problems of choosing the dimensionality of the Euclidean space and turning its axes. Possible ways of obtaining output data. Problems of applying statistical methods in sociology. Basic functions and procedures of data analysis. The value of substantive concepts of the researcher in solving the problems of choosing the dimensionality of the Euclidean space and turning its axes. Creating multidimensional tables using secondary variables. General characteristics of modern software tools for analyzing sociological data. Solving practical problems. .Content of the workshop, if necessary. Content of the workshop, if necessary.*

### *Topic 4. Canonical analysis.*

*General idea of methods based on frequency models General idea of modeling frequencies of contingency table. Multiplicative and additive frequency models. The role of logarithmization of the multiplicative model. The main task of canonical analysis. Principles of their derivation based on the analysis of the contingency table. Frequency models corresponding to canonical analysis. The relationship of canonical correlation coefficients with the “chi-square” criterion. General idea of digitizing values of nominal features. Canonical analysis as a method of digitization and a method of measuring the relationship between two nominal features with “common alternatives”. The concept of relationship between two groups of features. Sequence of canonical correlation coefficients. Principles of obtaining canonical correlation coefficients based on the analysis of the contingency table. Using canonical correlation in the analysis of contingency tables. The need to combine the model embedded in a particular method of digitization. Building sociological indices using the technique of canonical analysis. Solving the problem of weighting the components of the index of features. Solving practical problems.*

### *Topic 5. Log-linear analysis*

*The reasons for the difference between the real and the uniform distribution. Frequency models corresponding to log-linear analysis. Saturated model. The purpose of moving to logarithms of frequencies. Hypotheses about the relationship of features. Their role in building frequency models. Calculation of log-linear model coefficients for the two-dimensional case. Dominance ratios. Interpretation of coefficients through dominance ratios (for a model of arbitrary dimensionality). Comparison of log-linear analysis with nominal regression and variance analysis, as well as with the method of sequential splits. The comparison is made at the substantive level. Different understanding of the dependent feature: quantitative feature in variance analysis, quantitative or nominal - in nominal regression and frequency, which stands in the cell of a multidimensional contingency table, - in log-linear analysis. The impossibility of obtaining new knowledge based on the analysis of the uniform distribution (the essence of data analysis - the study of changes, comparison of indicators of different kinds). The meaning of contributions of different dimensionality. The role of the “chi-square” criterion when using log-linear analysis. Dominance ratios. Interpretation of coefficients through dominance ratios (for a model of arbitrary dimensionality). Different possibilities of searching for combinations of predictor values: testing hypotheses about the presence of multidimensional relationships in log-linear analysis and the possibility of finding the most effective combinations in the* *method of sequential splits and regression analysis, a predetermined set of combinations of predictor values in variance analysis. Solving practical problems.*

### *Topic 6. Causal analysis.*

*Strategy of analysis of the structure of relationships of features Graph of causal relationships. Repetition of the principles of constructing partial correlation and regression coefficients. The importance for the sociologist of studying the relevant relationships. The concept of "spurious" correlation. The main causal schemes that lead to their appearance. Calculation of covariances (correlations) between any two features based on the graph of relationships. Structural equations. Calculation of structural coefficients. Their relationship with partial regression coefficients. The main theorem of causal analysis. Its role in the study of statistical dependencies. The concept of the structure of a multidimensional random variable. Formation of generalized indicators based on the analysis of the structure of relationships of features. The role of statistical methods in the study of causal relations. Structural coefficients. Input (external, independent) and output (internal, dependent) variables. Rules for reducing causal schemes and forming equations. The difference between statistical and causal relationship. The study of statistical relationships based on causal schemes as the main task of causal analysis. The concept of auxiliary theory of measurements of Blalock. Causal analysis as a conceptual approach to the study of social phenomena. The problem of formalization of the task of studying causal relationships in sociology. Complex use of several methods of studying the relationships between features to solve sociological problems (analysis of the structure of a random variable; factor and variance analysis; search for determining combinations of predictor values). Solving practical problems.*

### *Topic 7. Pattern recognition tasks.*

*The concept of automatic classification of objects Classification as one of the fundamental processes in science. General idea of pattern recognition tasks (synonyms: image, class, cluster, taxon; ambiguity of interpretation of terms in the literature). Identification of tasks: search for classes, description of classes, determination of the most effective system of features. Identification of the task of automatic classification of objects (synonyms: multidimensional classification, unsupervised pattern recognition, cluster analysis, taxonomy). Feature space. The task of classification as a search for clusters of points - models of objects in the feature space. The role of the presence or absence of a training sample. Solving practical problems.*

### *Topic 8. The problem of "stitching" content and formalism when using classification algorithms*

*The specificity of solving sociological tasks of building a typology using methods of automatic classification. The meaning of contrasting the terms "classification" and "typology". Identification of the main formal elements of automatic classification algorithms that require stitching with the substantive concepts of the sociologist. The basis of typology. The role of the researcher's prior ideas about the types sought in the selection and implementation of the algorithm, interpretation of the results of its application. Solving practical problems.*

### *Topic 9. Functions of distance between objects*

*The role of hypotheses about the nature of the location of objects in the choice of classification algorithm. The main types of hypotheses: compactness, connectivity (continuity), unimodal distribution. Examples of sociological tasks of building a typology for which each hypothesis would be reasonable. Examples of algorithms that search for patterns of points in the feature space that correspond to each of the hypotheses: Forel algorithm (compactness hypothesis), nearest neighbor algorithm (connectivity hypothesis), algorithm based on the selection of local maxima of the membership function (unimodal distribution hypothesis). The role of the membership function in the corresponding algorithms. Substantive ideas of the sociologist about the types and conditions of choosing the step of splitting in the interpretation of the results. Adjustment of the results of classification in order to ensure the correspondence of classification and typology. Solving practical problems.*

### *Topic 10. The main types of classification procedures.*

*Distances between classes Identification of hierarchical and non-hierarchical classification algorithms. Agglomerative and divisive algorithms. Optimization of splitting in the sense of maximizing the pre-selected quality functional as one of the main elements of formalism in non-hierarchical classification algorithms. The main substantive meaning of optimization. The meaning of measuring the proximity between classes in* *such cases. Methods of measuring the total estimates of proximity to each other of objects within classes. Solving practical problems.*

*Topic 11. Hypotheses about the location of objects in the feature space*
*The role of hypotheses about the nature of the location of objects in the choice of classification algorithm. Examples of sociological tasks of building a typology for which each hypothesis would be reasonable. General idea of fuzzy classifications. The role of the membership function in the corresponding algorithms. Substantive ideas of the sociologist about the types and conditions of choosing the step of splitting in the interpretation of the results. Solving practical problems.*

*Topic 12. The concept of interpretation of output data and the main methodological principles of using data analysis methods in sociology*
*Interpretation of output data as one of the main links of "stitching" sociology and mathematics. Identification of methodological principles, compliance with which is necessary for the analysis of sociological data to be effective, not to divert the sociologist away from reality: ensuring a certain homogeneity of output data; taking into account the model, "embedded" in each method of data analysis, when choosing an analysis algorithm, two main principles of interpretation of analysis results: the need to harmonize it with the interpretation of output data and filling in the losses that occurred during the transition to formalism; the need for a comprehensive use of several methods to solve one problem, etc. Solving practical problems.*

*Topic 13. Data. Metadata*
*Data creation (DataGeneration/DataCapture). Data maintenance (DataMaintenance). Data synthesis (DataSynthesis). Data usage (DataUsage). Data publication (DataPublication). Data archiving (DataArchival). Data purging (DataPurging) Solving practical problems.*

*Topic 14. Big data. Big data management systems Distributed file systems. Distributed frameworks. Benchmarking. Server programming. Planning. Deployment systems. Solving practical problems.*

*Topic 15. Software platforms and systems for Big Data*
*Data stream management systems. Big data storage systems. Big data platforms. Real-time data processing. Big data management systems. Analytical platforms. Solving practical problems.*

*Topic 16. Machine learning using the Scikit-learn library.*
*Steps of a typical practical machine learning scenario. Loading a dataset. Exploring data using Pandas. Visualizing features using Matplotlib. Setting model parameters and evaluating its accuracy. Functionality of the Scikit-Learn library. Classification using K-neighbors. Linear models for regression and classification (linear regression model, logistic regression, etc.). Decision trees and random forest. Basics of neural networks. Clustering algorithms (K-means clustering, hierarchical clustering, etc.). Solving practical problems.*

### Topics of the laboratory classes

*Topics of laboratory work Laboratory work is not provided.*

### Self-study

*Independent work Independent work for the course consists of independent study by students of topics and questions that are not taught in class, performing individual tasks. Students are also recommended additional materials (videos, articles) for self-study and analysis.*

## Course materials and recommended reading

*1. Trevor F Cox Multidimensional Scaling, Second Edition // Chapman & Hall/CRC. 2000. - DOI:10.1201/9781420036121*
*2. Pedro Delicado, Cristian Pachon-Garcia Multidimensional Scaling for Big Data // 2020 (v1). - https://arxiv.org/abs/2007.11919*

*3. Susan S. Schiffman Introduction to Multidimensional Scaling: Theory, Methods and Applications // Emerald Publishing; F First Edition (October 28, 1981). - https://www.amazon.com/Introduction-Multidimensional-Scaling-Methods-Applications/dp/0126243506*

## Assessment and grading

### Criteria for assessment of student performance, and the final score structure

*100% of the final grade consists of the results of the assessment in the form of an exam (20%) and the current assessment (80%). Exam: performance of a calculation task and oral report. Current assessment: 16 online tests on topics (48%), two individual tasks (22%) and two calculation tasks (10%)*

### Grading scale

| *Total points* | *National* | *ECTS* |
|---|---|---|
| *90–100* | *Excellent* | *A* |
| *82–89* | *Good* | *B* |
| *75–81* | *Good* | *C* |
| *64–74* | *Satisfactory* | *D* |
| *60–63* | *Satisfactory* | *E* |
| *35–59* | *Unsatisfactory (requires additional learning)* | *FX* |
| *1–34* | *Unsatisfactory (requires repetition of the course)* | *F* |

## Norms of academic integrity and course policy

*The student must adhere to the Code of Ethics of Academic Relations and Integrity of NTU "KhPI": to demonstrate discipline, good manners, kindness, honesty, and responsibility. Conflict situations should be openly discussed in academic groups with a lecturer, and if it is impossible to resolve the conflict, they should be brought to the attention of the Institute's management.*
*Regulatory and legal documents related to the implementation of the principles of academic integrity at NTU "KhPI" are available on the website:*[*http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/*](http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/)

## Approval

| | | |
|---|---|---|
| *Approved by* | *Date, signature*<br>*30.06.2023* | *Head of the department*<br>*Vladimir MOROZ* |
| | *Date, signature*<br>*30.06.2023* | *Guarantor of the educational program*<br>*Yuri KALAGIN* |