

**MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE**  
**Volyn National University**  
**named after Lesya Ukrainka**  
**Faculty of Information Technologies and Mathematics**  
**Department of Computer Science and Cyber Security**

**SYLLABUS**  
**of the normative educational component**  
**INTELLIGENT DATA ANALYSIS**  
**for the first (bachelor) level of higher education,**  
**specialty 122 Computer science**  
**educational program**  
**Computer science and information technology**

**Syllabus of the normative educational component** "Intelligent data analysis" of bachelor's training, field of knowledge 12 - Information technologies, specialty 122 - Computer science, educational program "Computer science and information technology"

**Developer:**

Tetyana Mamchych, PhD on Mathematics, Associate Professor,  
Associate Professor of the Department of Computer Sciences and  
Cyber Security

**Agreed**



Guarantor of the educational program Hryshanovych T. O.

The syllabus of the educational component was approved at the meeting of the Department of Computer Sciences and Cyber Security  
protocol No. 2 dated 09/29/2022

Head of the



department Hryshanovych T.O.

## I DESCRIPTION OF THE EDUCATIONAL DISCIPLINE

Indicators	Field of knowledge, specialty, educational program, educational level	Characteristics of the subject
		Normative
Full-time	Field of knowledge 12 Інформаційні технології, спеціальність 122 Комп'ютерні науки, освітньо-професійна програма Комп'ютерні науки та інформаційні технології освітній рівень бакалавр.	Year: 2
Number of hours/кредитів  120/4		Semester: 4
		Lectures: 28 hrs
		Laboratory work: 34 hrs
		Independent work: 50 hrs
		Consultations: 8 hrs
Individual task: none		Form of evaluating: exam

### II Information about the teacher:

Tetyana Ivanivna Mamchych;  
 Academic degree: PhD in Mathematics;  
 Academic title: Associate professor;  
 Position: Associate Professor of the Department of Computer Science and Cyber Security;  
 Contact information: Mamchych.Tetyana@vnu.edu.ua  
 Class days: <http://194.44.187.20/cgi-bin/timetable.cgi>

### III. Description of the educational component

**Abstract of the course.** The syllabus of the educational component is compiled in accordance with the educational and professional program of Bachelors of Computer Science and Information Technologies. The educational component "Intelligent data analysis" consists of the basic methods of processing numerical, ordinal and text data (Data Mining, Text Mining) along with mastering the relevant computer technologies (R program). The list of methods is based on modern statistical approaches, involving some classification, forecasting and pattern recognition algorithms.

**The subject of study** of the educational component is methods and technologies of data analysis.

**The purpose of the educational discipline:** formation of students' knowledge, skills and abilities in data analysis, application of appropriate information technologies for data processing.

#### The list of competencies of a graduate

##### *General competences (GC)*

- GC1. Ability to abstract thinking, analysis and synthesis.
- GC2. Ability to apply knowledge in practical cases.
- GC3. Knowledge and understanding of the subject area and understanding a professional activity.
- GC6. Ability to learn and master modern knowledge.
- GC7. Ability to search, process and analyze information from various sources.
- GC11. Ability to make reasoned decisions.
- GC12. The ability to evaluate and provide the quality of the work performed.
- GC13. The ability to act on the basis of ethical considerations.

### *Special (professional) competences (SC)*

SC1. Ability to mathematically formulate and investigate continuous and discrete mathematical models, justify the choice of methods and approaches for solving theoretical and applied problems in the field of computer science, analysis and interpretation.

SC2. The ability to identify statistical regularities of non-deterministic phenomena, use methods of computational intelligence, in particular statistical, neural network and fuzzy data processing, methods of machine learning and genetic programming, etc.

SC3. The ability to think logically, draw logical conclusions, use formal languages and models of algorithmic calculations, design, develop and analyze algorithms, evaluate their effectiveness and complexity, solvability and unsolvability of algorithmic problems for adequate modeling of subject areas and to develop software and information systems .

SC6. Ability to system thinking, application of system analysis methodology for researching complex problems of various nature, methods of formalization and solving system problems with conflicting goals, uncertainties and risks.

SC7. The ability to apply the theoretical and practical foundations of modeling methodology and technology to study the characteristics and behavior of complex objects and systems, conduct computational experiments with processing and analysis of results.

SC11. Ability to intelligently analyze data based on methods of computational intelligence, including large and poorly structured data, data processing and visualization while solving practical problems.

### *Expected learning outcomes (ELO).*

ELO 3. To use the knowledge of regularities of random phenomena, their properties and operations on them, models of random processes and modern software environments for solving problems of statistical data processing and building predictive models.

ELO 2. Apply methods and algorithms of computational intelligence and intelligent data analysis in tasks of classification, forecasting, cluster analysis, search for associative rules using software tools to support multidimensional data analysis based on DataMining, TextMining, WebMining technologies.

## **IV The structure of the Educational component**

Content module 1. Classical methods of statistics.

Content module 2. Modern computer technologies for data analysis.

<b>Topics</b>	<b>Total (hrs)</b>	<b>Lectures (hrs)</b>	<b>Lab. Work (hrs)</b>	<b>Selflearning (hrs)</b>	<b>Consult. (hrs)</b>	<b>*Evaluation/ Points</b>
<b>Module 1. Classical methods of statistics</b>						
1. Descriptive statistics. Estimation of parameters.		2	2	2		10
2. Data visualization.		2	2	4	1	10
3. Statistical inference.		2	4	4	1	10
4. Correlation. Regression.		4	4	4	1	10
5. Multivariate statistical methods.		4	4	6	1	10
Total for Module 1	56	14	16	20	4	50
<b>Module 2. Modern computer technologies for data analysis</b>						
6. Time series analysis and		2	4	4	1	10

forecasting						
7. Introduction to machine learning and neural nets.		2	2	8	1	10
8. Text Mining		2	4	6	1	10
9. Introduction to stochastic methods and simulations.		4	4	6		10
10. Some models and algorithms for ordered data and nets.		2	4	6	1	10
Total for Module 2	64	14	18	30	4	50
<b>Final evaluation</b>	<b>Current evaluation (for the semester) (40 points) Module 1 (30 points)+Module 2 (30 points) or Exam (60 points)</b>					
<b>Total (Hours/ Points) 120/ 100</b>	<b>120</b>	<b>28</b>	<b>34</b>	<b>50</b>	<b>8</b>	<b>100</b>

## V Evaluation policy

**Academic Integrity Policy.** Academic integrity is based on condemning the practices of plagiarism (making written works involving external sources of information other than those permitted for use), plagiarism (reproducing published texts of other authors without indicating authorship), fabrication (inventing data or facts used in the educational process). In the case of a violation of academic integrity by a student (copying, plagiarism, fabrication), the work is evaluated as unsatisfactory and must be re-done, and the results of previously passed works are canceled and re-done in the order determined by the teacher. At the same time, the teacher reserves the right to change the assignment.

**Communication policy.** Applicants for higher education must have an activated university mail. All written questions to the lecturers about the course should be sent to the university e-mail, possibly another (additional) source of communication, determined by the lecturer for more prompt communication with students.

**Reorder Policy.** Works that are submitted late without good reason are evaluated at a lower grade. Modules can be rearranged with the permission of the lecturer if there are good reasons (for example, sick leave).

**Assessment Appeals Policy.** Assessment Appeals Policy.

If the student does not agree with the assessment of his knowledge, he can protest the grade issued by the teacher in accordance with the established procedure according to the "Regulations on the procedure and procedures for resolving conflict situations at Lesya Ukrainka Volyn National University"

**Class attendance policy.** For full-time higher education students, attending classes is mandatory. Important reasons for not attending classes are illness, academic mobility, which must be confirmed with relevant documents. The student of higher education must inform the teacher about the absence from the class and the reasons for the absence either personally or through the headmaster. For objective reasons, training can take place online upon agreement with the course leader and the dean of the faculty.

**Recognition of learning outcomes obtained in formal, non-formal education.** During the study of the educational component, it is possible to recognize the learning outcomes obtained in formal, informal and/or informal education. The procedure for recognizing the results of higher education for students of higher education acquired in: formal education (academic mobility of students on the territory of Ukraine or outside of Ukraine, for students who are transferring, renewing from other higher education institutions (domestic or foreign); informal and/or informal education is carried out in accordance "REGULATION on the recognition of learning outcomes obtained in formal, informal and/or informal education at Lesya Ukrainka Volyn National University" (<https://cutt.ly/yNUt5Y4>).

The course "Data analysis and statistical inference with the R program" (30 points), "Data visualization" (10 points) on the Prometheus open course platform can be taken into account during the evaluation.

**Final evaluation**

The form of control is a semester's exam. Evaluation is carried out on a 100-point scale. The assessment includes current control (accrued for the quality performance of laboratory works) and final modular control (accrued for the performance of modular control works and modular test works). The maximum number of points that a student can receive during the current assessment for the semester is 40 points. The final module control for the semester includes grades for all module control works, test tasks and is 60 points.

If at least 75 points are accumulated according to the results of the semester and the student agrees with this result, then the grade for the semester can be issued without taking the exam. Otherwise, the student takes the exam; the maximum number of points that can be obtained on the exam is 60 points, while the points for the final module control are canceled.

The exam is held in written form. The grade for the semester in the case of passing the exam is the sum of the points of the current control and the points obtained during the exam.

### **Exam questions**

1. Types of variables(indicators) for statistical analysis. Examples.
2. Statistical distribution of data.
3. Visualization of statistical distribution. Histogram, cumulative chart, graph, frequency polygon, pie chart, box-whisker chart.
4. Numerical characteristics of the level and spread of statistical indicators: mean, mode, median, variance, standard deviation, standard error, quantiles, quartiles, deciles, decile coefficient, coefficient of variation, range of variation.
5. Point and interval estimations. Dependence of accuracy and reliability of confidence intervals.
6. Basic concepts of statistical inference: testing of statistical hypotheses, errors of the type I and II, significance level, power of a test.
7. Tests of Pearson  $\chi^2$ , Student, Fisher, Shapiro, Kolmogorov-Smirnov, ANOVA, MANOVA methods.
8. Correlation analysis. Correlation coefficients for two indicators.
9. Pair and multiple regression. The least square method.
10. Analysis and forecasting of dynamic series, including with "seasonal changes".
11. Cluster analysis.
12. Factor analysis.
13. Discriminant analysis.
14. Concepts of machine learning: pattern recognition, regression and classification, support vector algorithms (SVM). Neural networks (NN). Bootstrap statistical procedures.
15. Concepts of stochastic computing algorithms: simulations, Monte Carlo method, random descent method for optimization, ant algorithm (Ant Colony Optimization), bee algorithm (Artificial Bee Colony Algorithm), genetic algorithms.
16. Text data analysis (Text Mining), content analysis of a given corpus of data. Modern capabilities of the R program for searching and analyzing textual data.
17. Some methods of analyzing ordinal data: finding collective decisions (social choice), the Kemeny-Snell metric.
18. Graph models for network data. Balance in models of sign graphs. Harari status measure in directed graph models for hierarchical structures. Markov chain models for information networks.
19. Computing with R program: calculations in interactive mode, vectors and matrices, data frame, writing and running of scripts, input and output of data, import and export.
20. Using packages in R. Statistical functions for descriptive statistics, data visualization, hypothesis testing, multivariate methods, Text Mining, SVM, NN.

The exam is conducted in a written form. The written report should contain answers for theoretical questions and the results of practical tasks (copying executed commands from the console).

100 minutes are allocated for calculations, preparation of answers and preparation of the report.

The exam's list of questions contains 5 theoretical questions (these can also be tasks for calculations without computer programs) and 5 practical tasks for calculations using the R program. For calculations each student is offered an individual set of data.

### **An example of an examination ticket:**

All questions are worth 6 points.

Theoretical questions (short written answers):

1. For a given set of discrete data, find the mean and median without using computer programs.
2. The method of least squares. Regression line.
3. The method of support vectors for the classification of multidimensional data.
4. Factor analysis: purpose and conditions of applicability.
5. Calculate the result of collective choice for a set of individual profiles (ordinal data).

Practical tasks (performed with the R program):

6. For a given set of data, calculate the mean and median.
7. For the given set of data create a frequency histogram, add a density line.
8. For a given set of textual data, find the frequency matrix for the indicated lexemes (terms).
9. Calculate the coefficients of correlation for two given documents regarding the use of a given set of terms (data from task 8).
10. Construct a linear regression for the given data set.

### Grading scale (national and ECTS)

A scale for evaluating the knowledge of education seekers on educational components, where the form of control is an exam

Score in points	Linguistic score	Оцінка за шкалою ECTS	
			пояснення
90–100	Excellent	A	Outstanding performance without errors
82–89	Good	B	Above the average standard but with minor errors
75–81	Good	C	Generally sound work with some errors
67–74	Satisfactorily	D	Fair but with significant shortcomings
60–66	Satisfactorily	E	Performance meets the minimum criteria
1–59	Unsatisfactorily	Fx	Fail. Reexamination is required

### Recommended literature and Internet resources

1. Maiboroda R.E. "Computer statistics". VOC "Kyiv University", 2019. 589 p. (in ukr.)  
<http://probability.univ.kiev.ua/userfiles/mre/cscolor.pdf>
2. Maiboroda R. Multivariate descriptive statistics. Methodical recommendations for the course. KNU named after Taras Shevchenko, 2020. (in ukr.)  
<https://probability.knu.ua/userfiles/mre/descrtasks.pdf>
3. Lande D.V., Subach I.Yu., Boyarynova Yu.E. Fundamentals of the theory and practice of intelligent data analysis in the field of cyber security: a study guide. K.: ISZZI KPI named after Igor Sikorskyi, 2018. 300 p. (in ukr.)
4. TextMining with R:  
<http://www.slideshare.net/whitish/textmining-with-r>
5. A. Agresti. Statistical methods for social sciences. Boston: Pearson. 2018.
6. The R Project for Statistical Computing <https://www.r-project.org/>

7. Sergejev-Gorchynskiy O.O., Ishchenko G.V. Intelligent data analysis: computer workshop. KPI named after Igor Sikorskyi, 2018. 73 p. (in ukr.)  
[https://ela.kpi.ua/bitstream/123456789/24971/1/Komp\\_prakt.pdf](https://ela.kpi.ua/bitstream/123456789/24971/1/Komp_prakt.pdf)
8. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning.  
<https://hastie.su.domains/ElemStatLearn/>
9. Data Science Full Course - Learn Data Science in 10 Hours | Data Science For Beginners | Edureka  
<https://www.youtube.com/watch?v=-ETQ97mXXF0>
10. Statistics - A Full University Course on Data Science Basics <https://www.youtube.com/watch?v=xxpc-HPKN28>
11. T. Mamchych. Intelligent data analysis. Part 1. Applied statistics. Lecture Notes. Lesya Ukrainka Volyn National University. 2022. (in ukr.)