MD-2068, CHISINAU, 9/7 STUDENTILOR STR, 9/7, PHONE: 022 50-99-63, www.utm.md

Exploratory Data Analysis and Modeling

1. Course/Module information

Faculty	CIM				
Department	Software Engineering and Automation				
Study cycle	Master's Degree				
Study program	**Data Science**				
Year of study	**Semester**	**Evaluation type**	**Formative category**	**Optionality category**	**ECTS credits**
1st Year *(full-time education)*	1	E	F	O	5

2. Estimated total time

Total hours in the curriculum plan	**Including**				
	Auditory hours		**Individual work**		
	Lecture	**Laboratory/ seminar**	**Term paper**	**Study of theoretical material**	**Application development**
Full-time education	20	20	-	110	-

3. Prerequisites for access to the course/module

According to the curriculum plan	Mathematics (Linear Algebra, Differential and Integral Calculus, Probability Theory, Mathematical Statistics), Data Structures and Algorithms, Programming (Python or R), Databases, Web Technologies.
According to competencies	Basic data-related skills and knowledge, fundamental statistical concepts, key concepts from higher mathematics (including linear algebra, differential and integral calculus, etc.), data manipulation skills, understanding graphical representations, critical thinking and logical reasoning, basic programming skills (programming concepts in any language, preferably Python or R), data collection concepts, knowledge of how data is stored and extracted from databases, interest and curiosity for data analysis, understanding of the business context, self-learning skills.

4. Conditions for conducting the educational process

Lecture	A projector and computer are required to present theoretical material in the classroom. Student tardiness, as well as phone conversations during the lecture, will not be tolerated.
Laboratory/ seminar	Students will engage in seminars under the guidance of the professor and assistant and will complete reports according to the methodological instructions. The deadline for submitting individual work is one week from completion. If the work is submitted late, a penalty of 1 point per week of delay will be applied.

5. Specific competencies acquired

Professional competencies	**CPM1** System architecture design and development. **CPM2** Monitoring technological trends. Innovation. Sustainable development. **CPM3** Application development. Component integration. Systems engineering.
Transversal competencies	**CT1.** Autonomy and responsibility **CT2.** Social interaction **CT3.** Personal and professional development

6. Course/Module objectives

General objective	The development of knowledge and the formation of skills necessary for the effective use of data analysis methods and techniques to detect and capture the relationships and patterns hidden behind the data, and to obtain a simplified, clear, and easily interpretable representation of these relationships and patterns.
Specific objectives	To achieve the general objective, it is necessary to develop knowledge and skills in: • Data Profiling and Summarization: Techniques for effective profiling and summarizing of datasets. • Data Visualization: Skills for visually representing data using various charts, graphs, etc. • Pattern and Trend Identification: Identifying patterns, trends, and relationships within data to generate insights. • Handling Missing Data and Outliers: Methods for handling missing data and detecting outliers during exploratory analysis. • Statistical Analysis Techniques: Statistical methods to analyze data and draw conclusions. • Preprocessing Data: Knowledge of data preprocessing techniques to prepare data for analysis. • Domain-Specific EDA: EDA techniques specific to various fields, such as finance, healthcare, or marketing. • Communicating Findings: Skills to effectively communicate insights and findings from EDA to both technical and non-technical audiences. • Collaboration Practices in EDA: Collaborating with colleagues in performing EDA and sharing information for collective understanding. • Ethical and Responsible EDA: Ethical considerations and responsibilities related to performing EDA, including data privacy and protection.

7. Course/Module content

Syllabus of teaching activities	Number of hours
Course topics	
T1 EDA Research Topic, motivation and methods Descriptive statistics and graphical techniques for data exploration. Definitions of basic concepts such as population, sample, and observation. Measures of central tendency (mean, mode, median) and their properties. Measures of variability and dispersion (range, variance, standard deviation, skewness, kurtosis) and their properties. Chebyshev's theorem and the empirical rule of distribution for normal distribution	**2**
T2 Types of Data and graphical methods of exploration. Qualitative data: Nominal vs. ordinal data. Quantitative data: Discrete vs. continuous data. Exploring qualitative data - frequency charts and pie charts. Exploring quantitative data - histograms, quartiles, percentiles, and measures of relative position. Interquartile Range and Box Plots. Time series data - line charts and trends.	**2**
T3. Relationships between two variables. Direction, Type, and Relevance of the Relationship. Correlation, Covariance and Pearson's Correlation Coefficient. Interpretation and Examples. False Correlation.	**2**
T4. Least squares method. Simple linear regression. Sum of squared residuals and stochastic error. Residual minimization problem. Examples.	**2**
T5. Multiple linear regression and fit quality. R-squared, total sum of squares, residual sum of squares and explained sum of squares.	**2**
T6. Hypothesis testing. Null and alternative hypotheses. Type I and Type II errors. Decision rule. Acceptance and rejection regions. T-statistic, critical t-value, and p-value for significance level. Examples.	**2**
T7. Multiple hypothesis testing and regression significance. F-test, critical F-value, degrees of freedom.	**2**

Syllabus of teaching activities	Number of hours
T8. Model specification and predictor selection. OLS assumptions. Influence of omitted variables. Suppressing the intercept. Polynomial regressions. Dummy variables.	**2**
T9. Logistic regression. Understanding logistic regression and the Logit function. Interpreting logistic regression results.	**2**
T10. Time series. Trend, cyclicality, seasonality. AR, MA, ARIMA, SARIMA, VARIMA. Difference and stationarity. Dickey-Fuller (DF) test and Augmented Dickey-Fuller (ADF) test. Autocorrelation function (ACF) and partial autocorrelation function (PACF).	**2**
Total course, hours	**20**
Practical work topics	
Practical work No. 1. Introduction to Using Python in Google Collaboratory - Loading datasets, calculating, and exploring descriptive statistics.	**2**
Practical work No. 2. Data visualization and analysis using graphical methods in Python.	**2**
Practical work No. 3. Correlations between variables in Python. Heatmaps and their interpretation.	**2**
Practical work No. 4. Implementation and interpretation of a simple linear regression model in Python.	**2**
Practical work No. 5. Multivariate regression in Python and evaluating the goodness of fit of the model.	**2**
Practical work No. 6. Hypothesis testing in Python, interpreting statistical significance.	**2**
Practical work No. 7. Multivariate regression in Python and joint significance testing using the F-test.	**2**
Practical work No. 8. Model specification testing and simulating omitted variables.	**2**
Practical work No. 9. Logistic regression in Python. Encoding binary variables.	**2**
Practical work No. 10. Time series modeling in Python: AR, MA, ARIMA. ACF and PACF. DF and ADF tests in Python – implementation and interpretation.	**2**
Total practical work, hours	**20**

8. Using generative AI

Permission to use	The use of generative AI in assignments and projects is permitted, provided that students adhere to the following rules: • Generative AI may be used to generate ideas, text structures, or code, but all generated materials must be reviewed and adjusted by the student to ensure that they meet academic requirements. • Any use of generative AI must be declared in the appendix section of each paper, using the phrase: "During the preparation of this paper, the author used [NAME OF TOOL / SERVICE] for the purpose of [REASON]. After using this tool / service, the author reviewed and edited the content as necessary and assumes full responsibility for the content of the paper."
Restrictions to use	Students ***MUSTN'T consider generative AI as a reliable source of information***, as it does not provide clear references or documented sources. • ***Direct citation of AI-generated content*** in academic papers as if it were a primary source ***isn't permitted.*** • Activities in which the use of **generative AI is prohibited** are specified by the teacher and are usually ***intermediate and final assessments*** or that don't involve professional competence development activities.

9. Bibliographic references

Main	1. Keller Gerald, Statistics for Management and Economics, South-Western College Publishing; 10th edition, 2014 2. Studenmund Arnold H., Unsing Econometrics: A practical Guide, Pearson Education, 2014 3. Educational Materials and Bibliographic Sources on FCIM's ELSE Platform: https://else.fcim.utm.md/enrol/index.php?id=701
Supplementary	1. Wes McKinney, Python for Data Analysis, O'Reilly Media, 3rd Edition, 2022 ISBN: 9781098104009 2. Jake Vanderplas, Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, 2nd Edition, 2023, ISBN-10: 1098121228, ISBN-13: 978-1098121228

10. Evaluation

Periodic		**Current**	**Individual study**	**Project/thesis**	**Exam**
PE 1	**PE 2**				
Full-time education					
15%	15%	15%	15%	-	40%
Minimum performance standards Attendance at lectures; activity and quality of preparation for lectures and practical works; Obtaining a minimum grade of "5" for each assessment and practical work; Demonstrating knowledge of the theoretical content of the course and the Python/R language in the final exam paper.					

11. Evaluation criteria

Activity	**Evaluation components**	**Evaluation method, Evaluation criteria**	**Weight in the final grade of the activity**	**Weight in the course evaluation**
Full-time education				
Periodic evaluation I	Theoretical content, topics 1-4	Test on MOODLE	100%	**15%**
Periodic evaluation II	Theoretical content, topics 6-8	Test on MOODLE	100%	**15%**
Current evaluation	Practical activity	Discussions during seminars	50%	**15%**
		File completed with Reports for each Case Study under discussion	50%	
Individual study	Research on the topic	Presentation/public speech	100%	**15%**
Final evaluation	Theoretical and practical content	Oral exam. Grading according to grading scale	100%	**40%**