COURSE OUTLINE

1. Data about the study programme

1.1 Higher education institution	Transilvania University of Brașov
1.2 Faculty	Mathematics and Computer Science
1.3 Department	Mathematics and Computer Science
1.4 Field of study[1)]	Computer Science
1.5 Study level[2)]	Master
1.6 Study programme/ Qualification	Internet Technologies – taught in English

2. Data about the course

2.1 Name of course			Data Warehousing and Data Mining					
2.2 Course convenor			Assoc. Prof. Alexandra Băicoianu					
2.3 Seminar/ laboratory/ project convenor			Cristina Gavrilă					
2.4 Study year	I	2.5 Semester	II	2.6 Evaluation type	E	2.7 Course status	Content[3)]	AC
							Attendance type[4)]	CPC

3. Total estimated time (hours of teaching activities per semester)

3.1 Number of hours per week	**3**	out of which: 3.2 lecture	**2**	3.3 seminar/ laboratory/ project	**0/1/0**
3.4 Total number of hours in the curriculum	42	out of which: 3.5 lecture	28	3.6 seminar/ laboratory/ project	0/14/9

Time allocation	hours
Study of textbooks, course support, bibliography and notes	14
Additional documentation in libraries, specialized electronic platforms, and field research	20
Preparation of seminars/ laboratories/ projects, homework, papers, portfolios, and essays	70
Tutorial	0
Examinations	4
Other activities.....	

3.7 Total number of hours of student activity	108
3.8 Total number per semester	150
3.9 Number of credits[5)]	6

4. Prerequisites (if applicable)

4.1 curriculum-related	• Programming knowledge, particularly in relevant languages such as Python • Familiarity with basic statistics concepts • Basic knowledge of machine learning (ML) algorithms and how they can be integrated into data mining processes.
4.2 competences-related	-

5. Conditions (if applicable)

5.1 for course development	A classroom with at least 60 seats and a projector.
5.2 for seminar/ laboratory/ project development	Python PyCharm and/or VSCode

6. Specific competences and learning outcomes

Professional competences	**P.C. 1. Specification, design and development of software systems using: procedural languages, object-oriented languages, declarative languages, databases, methodologies and development platforms.** L.O. 1.2. The graduate can frame a problem in a studied theoretical framework L.O. 1.3. The graduate can apply modern programming methods and techniques to solving a wide range of problems. L.O. 1.4. The graduate can provide demonstrations and explanations regarding the validity of the stated IT results. L.O. 1.5. The graduate can apply computer methods and techniques to solve practical problems. L.O. 1.7. The graduate can analyze algorithms that lead to the solution of practical problems. L.O. 1.8. The graduate can perform quantitative evaluations of solutions using Data Mining. **C.P. 3. Deepening the latest methodologies and technologies used in the software industry or with clear prospects of being used soon.** L.O. 3.3. The graduate can make interconnections between different computers fields. L.O. 3.5. The graduate can frame a problem in a studied theoretical framework. L.O. 3.6. The graduate can apply methods and techniques of modern computer science to solving a wide range of problems. **C.P. 4 Establish data processes, administer data collection systems, develop data processing applications, implement data quality processes, perform data mining** L.O. 4.2. The graduate develops and manages methods and strategies used to maximize data quality and statistical efficiency in data collection, to ensure that collected data is optimized for further processing. L.O. 4.4. The graduate applies data quality analysis, validation and verification techniques to verify data quality integrity. L.O. 4.5. The graduate explores large data sets to reveal patterns using statistics, database systems or artificial intelligence and presents the information in an understandable way.
Transversal competences	**T.C. 1. Communication and cooperation in professional contexts** L.O. 1.2. The graduate uses communication and relationship techniques in the virtual environment. L.O. 1.3. The graduate can cooperate and integrate in professional work teams in the educational field and in interdisciplinary teams. L.O. 1.5. The graduate can give presentations and public communications to promote knowledge and professional values. **T.C. 2. Career development and management** L.O. 2.2. The graduate formulates career development objectives and identifies action strategies in this regard. L.O. 2.3. The graduate self-evaluates and reflects on his own career, identifying strategies for regulating and overcoming professional difficulties.

7. Course objectives (resulting from the specific competences to be acquired)

7.1 General course objective	The development of algorithmic thinking and the enhancement of skills to extract valuable information, patterns, and knowledge from large datasets, employing data analysis techniques and algorithms.
7.2 Specific objectives	• Identifying relevant patterns and trends within heterogeneous and large-scale datasets. • Employing machine learning algorithms to make predictions and forecasts based on historical data and identified patterns. • Categorizing or segmenting data into groups, enabling a deeper understanding of each group's characteristics. • Detecting anomalies or unexpected behaviors in the data, highlighting potential issues or opportunities. • Uncovering relationships and trends in the data that can provide novel and actionable insights within the field of application.

8. Content

8.1 Course	Teaching methods	Number of hours	Remarks
Understanding the stages of a data mining process.		4 hours	

Exploring techniques for feature extraction. Gaining knowledge of data cleaning methods, including: Handling missing or incorrect values Normalizing data Examining techniques for dimensionality reduction.	Lectures Presentations Dialogue Problem formulation Case study Examples		
Understanding distance and similarity functions for different types of data. Analyzing the impact of dimensionality and data distribution on distance calculations.		4 hours	
Association Rule Mining: Identifying association rules to uncover meaningful patterns in datasets. Frequent Itemset Mining: Detecting patterns that frequently appear in data. Apriori algorithm: A foundational algorithm for identifying frequent itemsets. Enumeration-Tree algorithms: Efficiently exploring data and recognizing patterns.		6 hours	
Data Clustering Analysis - Selecting relevant features for clustering. - k-Means Algorithm - Kernel k-Means Algorithm - k-Medians Algorithm Evaluating Clustering Quality - Internal validation criteria. - External validation criteria. - Using clustering for anomaly detection or outlier identification.		8 hours	
General Principles of Data Classification Building decision trees for classification tasks. Understanding interpretable rule-based classification models. Applying the Naive Bayes algorithm for probabilistic classification tasks.		6 hours	
		Total 28 hours	

Bibliography

Data Mining Concepts and Techniques, Third Edition, Jiawei Han, University of Illinois at Urbana–Champaign
Micheline Kamber, Jian Pei, Simon Fraser University - https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
Introduction to Data Mining (Second Edition) - https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
Data Mining Practical Machine Learning Tools and Techniques - https://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf

8.2 Seminar/ laboratory/ project	Teaching-learning methods	Number of hours	Remarks
Applying data cleaning and dimensionality reduction on large datasets to optimize the processing workflow and improve the performance of machine learning models.		2 hours	

Introduction to Natural Language Processing (NLP) techniques, including text embedding, to transform textual data into a numeric format suitable for further analysis. The laboratory will focus on the practical application of these techniques, preparing students for their use in real-world data processing and natural language projects.			
Exploring distance and similarity functions in various practical contexts to understand how these functions can be used in data analysis. Applying cosine similarity in information retrieval tasks to identify and evaluate relevant documents or fragments based on vector comparisons of texts. The goal of the laboratory is to provide students with hands-on experience, focusing on the applications of these techniques in real-world scenarios to enhance search performance and textual data processing.	Exercises Dialogue Teamwork Problem-solving Individual study	2 hours	
Applying association rules to discover meaningful patterns in datasets. Implementing the Apriori algorithm to extract frequent itemsets and derive association rules. Exploring Enumeration-Tree algorithms for identifying and recognizing patterns in large datasets. The laboratory will place a strong emphasis on hands-on experience, concrete applications, and real-world scenarios, enabling students to apply these techniques in practical contexts.		4 hours	
Students will explore essential algorithms and validation strategies in a practical laboratory activity focused on clustering analysis. They will learn how to efficiently select features, apply validation criteria to assess the quality of clusters, and use well-known clustering algorithms such as k-Means and k-Medians through hands-on exercises. The laboratory activities will include: - Applying feature selection techniques relevant to the clustering process. - Implementing and testing clustering algorithms on datasets. - Applying internal and external validation criteria to assess the performance and consistency of the resulting clusters. - Hands-on exercises for anomaly detection using clustering, with real-world examples. A real-world clustering example for anomaly detection will be presented at the end of the session, applying the learned techniques to data from real domains.		4 hours	

Building Decision Trees: - Creating and training a decision tree model on classification datasets using specific algorithms. - Visualizing decision trees to understand how decisions are made at each node. - Validating the model's performance using evaluation techniques such as cross-validation and performance metrics like accuracy and the confusion matrix. Applying the Naive Bayes Algorithm: - Implementing and applying the Naive Bayes model on probabilistic datasets to perform classifications based on conditional probabilities. - Comparing the performance of the Naive Bayes algorithm with other classification algorithms, such as decision trees, using performance metrics (e.g., accuracy, precision, recall, F1 score). Classification Model Evaluation Exercises: - Evaluating and comparing the results obtained for each algorithm (decision trees, Naive Bayes, etc.) based on clear criteria (e.g., accuracy, F1 score, etc.). - Analyzing classification errors and identifying potential improvements for each model, including discussions on overfitting. The goal of the laboratory activity is to enable students to apply theoretical knowledge in a practical environment to build, evaluate, and compare different classification models on real datasets or data obtained from various simulations.		2 hours	
		Total 14 hours	

Bibliography
Data Mining Concepts and Techniques, Third Edition, Jiawei Han, University of Illinois at Urbana–Champaign
Micheline Kamber, Jian Pei, Simon Fraser University - https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
Introduction to Data Mining (Second Edition) - https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
Data Mining Practical Machine Learning Tools and Techniques - https://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf

9. Correlation of course content with the demands of the labour market (epistemic communities, professional associations, potential employers in the field of study)

Correlation applies in the Partnership Agreements and Internship Contracts concluded with the socio-economic partners of the Faculty/University.

10. Evaluation

Activity type	10.1 Evaluation criteria	10.2 Evaluation methods	10.3 Percentage of the final grade
10.4 Course 10.5 Seminar/ laboratory/ project	Developing the competencies targeted by the course content. Achieving the educational objectives outlined in the course syllabus.	The final grade for this course can be obtained by choosing one of the following options: Continuous assessment of projects launched for each module (4 modules): The projects are developed in teams of 2 students. Prerequisites condition: The arithmetic average of the grades obtained for all projects must be ≥ 5. **OR** Final Research Project (FRP) evaluation: The project consists of a paper developed in teams of 2 students, focusing on research directions in the fields of data mining and data warehousing. Research topics will be established in collaboration with the course coordinator, and at least 3 progress meetings are required for each topic, during which the stages and potential adjustments of the approach will be discussed. The project aims to develop an original study with a significant practical component and includes the completion of a concise report, containing conclusions, analyses, and comments, reflecting the processes and results obtained throughout the research. Prerequisites condition: FRP ≥ 5.	100%
10.6 Minimal performance standard			
Data cleaning and text embedding, along with working with and understanding basic classification algorithms (including using frameworks like WEKA), are fundamental elements that must be known to meet the minimum performance standard in this course. These skills provide the essential foundation for developing efficient projects and applications in the fields of data mining and data warehousing.			

This course outline was certified in the Department Board meeting on 26/09/2024 and approved in the Faculty Board meeting on 26/09/2024.

Assoc. Prof. Ion Gabriel Stan Dean	Assoc. Prof. Nicușor Minculete Head of Department
Assoc. Prof. Alexandra Băicoianu Course holder	Cristina Gavrilă Holder of seminar/ laboratory/ project

Note:

1) Field of study – select one of the following options: Bachelor / Master / Doctorat (to be filled in according to the forceful classification list for study programmes);

2) Study level – choose from among: Bachelor / Master / Doctorat;

3) Course status (content) – for the Bachelor level, select one of the following options: FC (fundamental course) / DC (course in the study domain)/ SC (speciality course)/ CC (complementary course); for the Master level, select one of the following options: PC (proficiency course)/ SC (synthesis course)/ AC (advanced course);

4) Course status (attendance type) – select one of the following options: CPC (compulsory course)/ EC (elective course)/ NCPC (non-compulsory course);

5) One credit is the equivalent of 25 study hours (teaching activities and individual study).