|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **WSB University** | | | | | |
| **Field of Study: Management** | | | | | |
| **Module /Item: Practical Aspects of Data Preparation** | | | | | |
| **Training profile: General** | | | | | |
| **Education cycle: II cycle studies** | | | | | |
| **Number of hours per semester** | 1 | | | 2 | |
| I | II | | **III** | IV |
| Fixed studies (w/l/lab/pr/e) |  |  | | **12C/24lab** |  |
| Part-time studies (w/l/lab/pr/e) |  |  | |  |  |
| **Lecturer** | Karol Jędrasiak PhD Eng. | | | | |
| **FORM OF CLASSES** | Classes, laboratories | | | | |
| **OBJECTIVES OF THE SUBJECT MATTER** | * Introduction to the issue of data processing for practical application; * Familiarize yourself with common challenges in data collection and how to solve them; * Familiarize yourself with how to process numeric, text, and time-related data; * Familiarize yourself with how to process hierarchical and relational data; * Familiarize yourself with commonly used methods of extraction of data from heterogeneous sources; * Introduction to automated data processing. | | | | |
| **Material effect** | **Reference to effects** | | | **Description of learning outcomes** | **How to verify the effect** |
| Direction | Area | | Knowledge | |
| DataPrep\_K01 | Z2\_W20 |  | | Knowledge of common data collection challenges and how to solve it | Completion of the oral submission of the draft dossier |
| DataPrep\_K02 | Z2\_W04 | S2A\_W02 | | Knowledge of the use of analytical information in building a company's market advantage | Completion of the oral submission of the draft dossier |
| DataPrep\_K03 | Z2\_W20 |  | | Knowledge of data integration from heterogeneous sources | Completion of the oral submission of the draft dossier |
| Skills | | | | | |
| DataPrep\_A04 | Z2\_U07 | S2A\_U06 | | Ability to normalize data. | Evaluation of the prepared project, oral credit |
| DataPrep\_A05 | Z2\_U07 | S2A\_U06 | | Ability to identify and delete erroneous data and visualize analyzed collections | Evaluation of the prepared project, oral credit |
| DataPrep\_A06 | Z2\_U07 | S2A\_U06 | | Ability to identify relevant data dependencies | Evaluation of the prepared project, oral credit |
| DataPrep\_A07 | Z2\_U07 | S2A\_U06 | | Can use troubleshooting methods | Evaluation of the prepared project, oral credit |
| DataPrep\_A08 | Z2\_U13 | S2A\_U09 | | They can document their work | Evaluation of the prepared project, oral credit |
| Social competence | | | | | |
| DataPrep\_S09 | Z2\_K02 | S2A\_K02 | | The student is aware of the importance of group work roles, assignment of tasks and role of communication in the project team | Understanding the roles and tasks assigned to the team and the degree of completion is part of the documentation |
| **Student effort (1h dyd =45 minutes)\*\*** | | | | | |
| **Full- time**  Participation in lectures =  Participation in classes = 12h  Participation in laboratory = 24h  Preparation to classes = 18h  Preparation to laboratory = 24h  Preparation to lectures =  Preparation to an examination = 22h  Project tasks =  e-learning =  Credit/examination = 4h  others (indicate which) =  **TOTAL: 104h**  **ECTS points: 4**  **Including practical classes: 4** | | | **Part-time**  Participation in lectures =  Participation in classes =  Participation in laboratory =  Preparation to classes =  Preparation to laboratory =  Preparation to lectures =  Preparation to an examination =  Project tasks =  e-learning =  Credit/examination =  others (indicate which) =  **TOTAL:**  **ECTS points:**  **Including practical classes:** | | |
| **DESCRIPTION OF THE SUBJECT MATTER** | The aim of the classes is to familiarize students with the preparation of data for further processing. As part of the classes, students will be familiar with the reasons for the need for pre-processing in order to organize variable types, integrate diverse datasets, work with missing data and data in a variety of formats. Historically and typically used solutions will be presented. Students will then be familiar with the modern approach to designing and modeling hierarchical data and relational databases. As part of the classes, students will learn about good practices of data normalization. Students will practice script design issues that resolve the above problem and identify data typically responding to typical management needs. The final step will be to apply the acquired knowledge in practice in order to implement the project. | | | | |
| **Prerequisites** | This course requires basic knowledge of the ways in which data is collected and the purposes of its processing, the basics of programming and data analysis | | | | |
| **ITEM CONTENT** | * Direct content: * Modern data applications; * Common challenges in data collection; * Ways to solve challenges in data collection; * Processing of numeric data; * Processing of text data; * Processing of time-related data; * Processing hierarchical and relational data; * Data normalization; * Introduction to data mining as a process; * Introduction to automated data processing; * Introduction to data completeness analysis; * Processes extract data from different data sources; * Data integration processes from different data sources; * Implementation of the project. | | | | |
| **Literature**  **Mandatory** | Pyle, Dorian. Data preparation for data mining. morgan kaufmann, 1999. https://pdfs.semanticscholar.org/470a/828d5e3962f2917a0092cc6ba46ccfe41a2a.pdfC. Zhang, Q. Young, Data Preparation for Data Mining. <https://www.researchgate.net/publication/220355854_Data_Preparation_for_Data_Mining>Ratanamahatana, C., and D. Gunopulos. Feature selection for the Naive Bayesian classifier using decision trees. Applied Artificial Intelligence 17(5-6):475-487. https://www.researchgate.net/publication/220355799\_Feature\_Selection\_for\_the\_Naive\_Bayesian\_Classifier\_Using\_Decision\_TreesBatista, G., and M. Monard. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence 17(5–6):519-533. https://www.researchgate.net/publication/220355769\_An\_Analysis\_of\_Four\_Missing\_Data\_Treatment\_Methods\_for\_Supervised\_Learning | | | | |
| **Literature**  **Supplementary** | Rattenbury, Tye, et al. Principles of data wrangling: Practical techniques for data preparation. " O'Reilly Media, Inc.," 2017.Castanedo, Federico. Data Preparation in the big data era. O'Reilly Media, Incorporated, 2015.Chen, Zhengxin. Intelligent Data Warehousing: From data preparation to data mining. CRC press, 2001.Abdullah, N., M. Liquie' re, and S. A. Cerri. 2003. GAsRule for knowledge discovery. Applied Artificial Intelligence 17(5-6):399-417.Hruschka, E., Jr., E. Hruschka, and N. Ebecken. A feature selection Bayesian approach for extracting classification rules with a clustering genetic algorithm. Applied Artificial Intelligence 17(5-6):489-506.Saravanan, M., P. Reghu Raj, and S. Raman. Summarization and categorization of text data in high level data cleaning for information retrieval. Applied Artificial Intelligence 17(5–6):461-474.Tseng, S., K. Wang, and C. Lee. A pre-processing method to deal with missing values by integrating clustering and regression techniques. Applied Artificial Intelligence 17(5-6):535-544.Tuv, E., and G. Runger. 2003. Pre-processing of high-dimensional categorical predictors in classification settings. Applied Artificial Intelligence 17(5-6):419-429.Yan, X., C. Zhang, and S. Zhang. Towards databases mining: Pre-processing collected data. Applied Artificial Intelligence 17(5–6):545-561.Yang, Q., T. Li, and K. Wang. 2003. Web-log cleaning for constructing sequential classifiers. Applied Artificial Intelligence 17(5–6):431-441.Zhang, C., and S. Zhang. 2002. Association Rules Mining: Models and Algorithms. In Lecture Notes in Artificial Intelligence, volume 2307, page 243, Springer-Verlag.Zhang, H., and C. Ling. 2003. Numeric mapping and learnability of Na€ve Bayes. Applied Artificial Intelligence 17(5-6):507-518.Zhang, Z., C. Zhang, and S. Zhang. 2003. An agent-based hybrid framework for database mining. Applied Artificial Intelligence 17(5–6):383-398. | | | | |
| **TEACHING METHODS** | In directform:  Slide show, traditional array method, presentation of materials in a computer environment | | | | |
| **SCIENTIFIC AIDS** | * Room with computer station * Whiteboard classes | | | | |
| **METHOD OF CREDIT** | Presentation in the project group with documentation | | | | |
| **FORM AND CONDITIONS OF CREDIT** | The condition for obtaining credit is to obtain a positive assessment from all forms of credit provided for in the curriculum, considering the quantitative evaluation criteria set out in the Framework System of Student Assessments at the WSB Academy. | | | | |