2.1.1 Module development M01 Data Analytics for Work

Module Specification:

Data Analytics for Work

Within the Erasmus+ KA2 Capacity Building Project (CBHE)

WORK4CE – Cross-domain competences for healthy and safe work in the 21st century

619034-EPP-1-2020-1-UA-EPPKA2-CBHE-JP

Product Owner: Sergey Subbotin (NU-ZP)

Team Members: UPV/EHU, KNUCA, WUNU, ASCCA, ASOIU

Version 0.1, 01.03.2021

1. Summary

Overall Learning Outcome: The students will be familiarized with methods of data analytics, will receive both theoretical and practical knowledge to use the methods and to develop software for data analytics for various problem specific domain.

Target Group Analysis: The target groups for the module are:

- students of MSc programs in Computer Science,
- practical specialists in programming (needs: data analytics to develop decision making and analytic software; prerequisites: Python/R programming language, Mathematics; previous competences: ability to develop computer programs),
- practical specialists and students in various applied domains (needs: to identify and predict states of multidimensional objects and processes; prerequisites: Python/R programming language, Mathematics; previous competences: ability to use special computer programs for data processing and analysis, ability to solve tasks from applied domain).

Competences & Learning Outcomes: The main competences according to European Qualification Framework (EQF) Level 7 (Master):

- knowledge: highly specialized knowledge in methods and tools of data analytics, which is at the forefront of knowledge in a field of work and study, as the basis for original thinking and/or research, the critical awareness of knowledge issues in a field and at the interface between different fields;

- skills: specialized problem-solving skills on data analytics application required in research and/or innovation in order to develop new knowledge and procedures and to integrate knowledge from different fields;

- responsibility and autonomy: manage and transform work or study contexts that are complex, unpredictable and require new strategic approaches; take responsibility for contributing to professional knowledge and practice.

Selection of Content: The main topics of the module:

- Exploratory Data Analysis (EDA) and Data Visualization
- Unsupervised learning. Cluster Analysis
- Supervised machine learning. Support Vector Machines and Decision Tree for Data Analysis.
- Prediction and Decision Making, and Results Visualization.

Activities and Teaching/Learning Methods: .

For all competences we will use lectures, laboratory works and self-study

The Activity Plan during one semester contains theory classes and laboratory works of 2h each per week (15 weeks), also as Homework and self-study (120 h)

Students read the lecture notes and course books, online tutorials, and prepare templates for laboratory work reports (including program code, written answers on key questions according to the student variant specified by the teacher).

Tailoring & Educational Tracks (Practical, Entrepreneurial, Scientific):

- Practical: to study the software for data analytics, focusing on application existing software tools to solve practical problems.
- Scientific to study the methods of data analytics, focusing on development of software tools to solve practical problems.

Competence Assessment: 50% exam (written), 50% - laboratory works (defended reports).

Curricula Integration: The module may be integrated with Computer Science specialty in one of the courses related to data analysis or computational intelligence.

Quality Evaluation: The learning outcomes will be needed for writing a master's thesis by the students.

Change History & Ownership: Sergey Subbotin..

Table of Content

1. Summary
2. Introduction to the module
3. Module Description
    - 3.1 Overall Learning Outcomes
    - 3.2 Target Group Analysis
    - 3.3 Competences & Learning Outcomes
    - 3.4 Content
    - 3.5 Teaching & Learning Activity Plan
    - 3.6 Teaching & Learning Resources
    - 3.7 Tailoring & Educational Tracks
    - 3.8 Assessment Methods
    - 3.9 Curricula Integration
    - 3.10 Quality Assurance - Evaluation
4. Syllabus/Module Handbook
5. References

2. Introduction to the module

This module is designed to equip students with the knowledge and skills necessary to extract meaningful insights from data, make accurate predictions, and effectively communicate student's findings through visualization. The course covers a range of essential topics, including Exploratory Data Analysis (EDA), Data Visualization, Unsupervised Learning, Supervised Machine Learning, and Results Visualization. These topics are modern tools for successful work of a specialist in many practical applications.

EDA is the first step in any data analysis process. It involves summarizing the main characteristics of a dataset, often using visual methods. R programming is a powerful tool for EDA, offering extensive libraries for data manipulation and visualization. Mastering EDA and R programming will enable you to identify patterns, detect anomalies, test hypotheses, and check assumptions through visual and quantitative methods. These skills are foundational for any data-driven decision-making process.

Unsupervised learning is a type of machine learning that deals with unlabeled data. Cluster analysis, a key technique in unsupervised learning, involves grouping a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. This is crucial for discovering hidden patterns or intrinsic structures in data. Applications of cluster analysis range from market segmentation and image compression to bioinformatics and anomaly detection.

Supervised learning involves training a model on labeled data, which means that each training example is paired with an output label. Support Vector Machines (SVM) and Decision Trees are two powerful supervised learning algorithms. SVMs are effective for high-dimensional spaces and are widely used for classification tasks, while Decision Trees are intuitive and easy to interpret, making them useful for both classification and regression tasks. Learning these algorithms will provide you with robust tools for prediction and decision-making in various fields such as finance, healthcare, and marketing.

Prediction and Decision Making is a critical component of data analysis and machine learning. It involves using data-driven models to forecast future outcomes and make informed decisions based on those predictions. This process integrates statistical analysis, machine learning algorithms, and decision theory to extract actionable insights from data, helping organizations and individuals make better strategic choices.

Results Visualization is a critical step in the data analysis pipeline. It involves presenting data and model outcomes in a way that is understandable and actionable for stakeholders. Effective visualization can highlight key insights, reveal trends, and support compelling narratives. Tools and techniques for results visualization will empower you to communicate complex findings clearly and persuasively.

Generalizing described before, this module will provide students with hands-on experience through practical exercises, real-world datasets, and projects. By the end of the course, student will have a comprehensive understanding of advanced data analysis techniques and machine learning algorithms, and you will be well-prepared to apply these skills to solve complex problems and drive impactful decisions in a career.

3. Module Description

3.1 Overall Learning Outcomes

This chapter summarizes the main learning outcomes and learning goals of the module.

Learning outcomes are defined as statements of what a learner knows, understands and is able to do upon completion of a learning process. In the European Qualification Framework (EQF) [1,2], learning outcomes are therefore defined in terms of knowledge, skills and competence which are understood as follows [3]:

- **Knowledge** means the outcome of the assimilation of information through learning. Knowledge is the body of facts, principles, theories and practices related to a field of work or study. In the context of the European Qualifications Framework, knowledge is described as theoretical and/or factual.
- **Skills** means the ability to apply knowledge and use know-how to complete tasks and solve problems. In the context of the European Qualifications Framework, skills are described as cognitive or practical skills.
- **Competence** means the proven ability to use knowledge, skills and personal, social and methodological abilities in work or study situations and in professional and/or personal development. In the context of the European Qualifications Framework, competence is described in terms of responsibility and autonomy.

Learning Outcomes/Competences need to consider several competence domains [4]:

- **Technical Competence**: This reflects the domain-specific competences, e.g. in engineering or software development. It is beyond tool skills, reflecting the full competence range (knowledge, skills, ability) in order to perform a job in a certain job domain (e.g. engineering competence for developing an engine, business competences for implementing a marketing plan).
  - Statistics and programming foundation. The competences in this area are focused on the knowledge of key statistics concepts and methods essential to finding structure in data and making predictions. Further, the student must have Python or R programming skills and the ability to visualize data, extract insights and communicate the insights in a clear and concise manner.
  - Data preparation. To ensure the student can construct usable data sets, the key competencies required are: – Identifying and collecting the data required – Manipulating, transforming and cleaning the data The student must also demonstrate the ability to deal with data anomalies such as missing values, outliers, unbalanced data and data normalization.
  - Model building. This stage is the core of the data analytic execution, where different algorithms are used to train the data and the best algorithm is selected. The student should know: – Multiple modelling techniques – Model validation and selection techniques What differentiates a data scientist is understanding the use of different methodologies to get insight from the data and translating the insight into business value.
  - Model deployment. An ML model is valuable when it’s integrated into an existing production environment and used to make business decisions. Deploying a validated model and monitoring it to maintain the accuracy of the results is a key skills.
- **Professional Competence**: This covers competences relevant for professional life, e.g. management competences, negotiation and presentation skills, team-related competences, legal topics, but also personal competences, e.g. critical thinking.
  - Skills to present the results of data analysis as a report.

- **Global Competence**: This covers all intercultural and international competences, e.g. language skills, knowledge of different markets /countries, but also citizen competences, e.g. ethics, political and social awareness.
  - Leadership (The student must be good problem solvers. They must understand the opportunity before implementing the solution, work in a rigorous and complete manner, and explain their findings. The student needs to understand the concepts of analyzing business risk, making improvements in processes and how systems engineering works)
  - Teamwork
  - Communication

3.2 Target Group Analysis

This chapters lists the target groups with respect to the learners and the teachers addressed by the module. The target groups are:

- Students of MSc programs in Computer Science, The persons who want to work in data analytic area.
- Programmers. The persons who are practical specialists in programming (needs: data analytics to develop decision making and analytic software, prerequisites: Python/R programming language, Mathematics, previous competences: ability to develop computer programs),
- Researchers. The persons who need knowledge and ability in data analytic for conduct research activity.
- Teachers. The persons who need new knowledge in data analytic for teaching students.
- Practical specialists and students in various applied domains (needs: to identify and predict states of multidimensional objects and processes, prerequisites: Python/R programming language, Mathematics, previous competences: ability to use computer programs, ability to solve tasks from applied domain).

3.3 Competences & Learning Outcomes

This chapter contains a more detailed description of the competences delivered by the module. This description for a competence profile which students will gain by attending the module. It is decomposed into competences within a competence breakdown structure.

The competence description should follow EQR/ESG [1,2] and should reach the competence level 7 according to EQF (Master Level). Competences should describe the knowledge, skills and abilities (responsibility and autonomy, in EQF notation).

Competence descriptions are using a certain “wording” which is explained in [3]. This wording (e.g. “students knows”, “student can apply”, “students is able to …”) should be used in the competence description. Each competence element should be “one sentence” which states what a student is competent in after attending the module, e.g.: Apart from technical competences, also professional and global competences (OLOs) need to be included. It is important to describe competences in such a way that allows an assessment of the competences (“keep the examination in mind!”).

- The student knows the programming language R and IDE RStudio
- The student can use dplyr and ggplot2 packages
- The student knows the programming language Python
- The student understands which kind of data visualization he need for exploratory data analysis
- The student can use exploratory graphs
- The students knows Plotting Systems, Graphics Devices and Plotting and Color in R
- The student can provide exploratory data analysis by himself for receiving an answer the research question
- The student can use clustering methods for data analysis

3.4 Content

Exploratory Data Analysis and Data Visualization with R programming (elective)

Topic 1. Introduction to R Programming.

Data types. Reading Data. Subsetting. Vectorized Operations. Control Structures. Functions. Scoping. Coding Standards. Dates and Times. Loop functions (lapply, apply, mapply, tapply). Split data. Debugging.

Topic 2. Exploratory Data Analysis with data. Dplyr package.

Subsetting and Sorting. Summarizing data. Merging data. Editing text variables. dplyr package. Data table package.

Topic 3. Applied Plotting.

Principles of Information Visualization. Plotting Systems in R. Base Plotting System. Graphic Devices. ggplot2.

Unsupervised learning. Cluster Analysis (core)
Topic 1. Basics of unsupervised machine learning.
Topic 2. Methods of centroid-based cluster analysis.
Topic 3. Methods of connectivity-based cluster analysis.

Supervised machine learning. Support Vector Machines and Decision Trees for Data Analysis (core)
Topic 1. Basics of Supervised machine learning.
Topic 2. Support Vector Machines
Topic 3. Decision Trees

Prediction, Decision Making, and Results Visualization (elective)
Topic 1 Prediction
Topic 2 Decision Making,
Topic .3 Results Visualization

3.5 Teaching & Learning Activity Plan

A) Select Teaching/learning methods per competence

For all competences we will use lectures, laboratory works and self-study

B) Define didactic concept: e.g. choose from:

- (Virtual) Lecture,
- Laboratory works

C) Define an Activity Plan, e.g. semester schedule

Activity 1: Theory classes (15 x 2 h = 30 h)

The theory classes are complemented with online tutorials and reading materials (course book).

Activity 2: Laboratory works (15x 2 h = 30 h)

The laboratory works in computer laboratory are complemented with online tutorials and reading materials (course book).

Activity 3: Homework and self-study (120 h)

Students read the lecture notes and course books, online tutorials, and prepare templates for laboratory work reports (including program code, written answers on key questions according to the student variant specified by the teacher).

3.6 Teaching & Learning Resources

The required Literature:

Roger D. Peng (2020): R Programming for Data Science. Retrieved from https://leanpub.com/rprogramming

Roger D. Peng (2020): Exploratory Data Analysis with R. Retrieved from https://bookdown.org/rdpeng/exdata/

Rafael A. Irizarry (2021): Introduction to Data Science. Data Analysis and Prediction Algorithms with R. Retrieved from https://rafalab.github.io/dsbook/

https://scikit-learn.org/stable/modules/clustering.html

https://machinelearningmastery.com/clustering-algorithms-with-python/

The required Data Sources: https://archive.ics.uci.edu/ml/

The required Laboratory Equipment: Personal computer or notebook with access to the Internet for each student.

The required Learning Management System (LMS): moodle.

3.7 Tailoring & Educational Tracks

Educational Tracks:

- Practical: to study the software for data analytics
- Scientific to study the methods of data analytics

3.8 Assessment Methods

Assessment

FORM	**%**	**REMARK**
Written exam	50	Based on theory classes
Laboratory work	50	Based on homework and laboratory classes

3.9 Curricula Integration

NUZP: Integrate to the Master of Science program “Systems of Artificial Intelligence” for the specialty 122 “Computer Science” into the course “Foundations of Computational Intelligence” as a course submodule.

WUNU: Integrate to the Master of Science program “Computer Science” for the specialty 122 “Computer Science” into the course “Analysis and processing of Big Data” as a course submodule.

KNUCA: Integrate to the Master of Science program “Project Management” for the specialty 122 “Computer Science” into the course “Intelligence Data Processing” as a course submodule.

ASOIU: Integrate to the Master of Science program “Information Technologies and Management” for the specialty 122 “Computer Science” into the course “Data processing systems” as a course submodule.

3.10 Quality Assurance - Evaluation

Learning Objectives Assessment: The achievement of learning objectives will be assessed involving exams, tests, presentations, defense of laboratory works.

Criteria for Evaluation: The criteria by which student work will be evaluated include accuracy, completeness, originality, critical thinking, coherence, and relevance to the course material.

Feedback Mechanisms: The feedback will be provided to students on their performance. This will involve written comments at site of the Department of Software Tools, discussion with teachers and administrators at individual and / or group meetings, reviews from peer evaluations.

Assessment Methods: We will use various assessment methods, such as formative assessments (to monitor student progress) and summative assessments (to evaluate student achievement at the end of a unit or course).

Evaluation of Teaching Methods: The effectiveness of teaching methods will be evaluated using student surveys, peer evaluations, classroom observations, and analysis of student performance data.

Quality Assurance Procedures will be used to ensure consistency and fairness in evaluation in form of: Calibration Sessions for Graders (involve bringing together graders or evaluators to review and discuss grading criteria, standards, and sample student work).

Opportunities for Improvement: Students can seek clarification on grading or request re-evaluation of their work if they believe there has been an error or if they wish to improve their grade.

4. Syllabus/Module Handbook

Entry for the Syllabus/Module Handbook

<table>
<tr><td colspan="7"><b>Data Analytics for Work (M1 DA4W)</b></td></tr>
<tr><td colspan="2"><b>Module Owner</b>
Sergey Subbotin</td><td><b>Workload</b>
180 h</td><td><b>Credits</b>
6 ECTS</td><td><b>Semester</b>
<i>2</i></td><td><b>Frequency</b>
<i>spring semester</i></td><td><b>Duration</b>
1 Semester</td></tr>
<tr><td><b>1</b></td><td><b>Course Title</b>
Data Analytics for Work</td><td colspan="2"><b>Contact hours</b>
4 hours per week / 60 h in total</td><td><b>Self-Study</b>
120 h</td><td colspan="2"><b>Planned Group Size</b>
25 students</td></tr>
<tr><td><b>2</b></td><td colspan="6"><b>Course Description</b>
Data analytics intend to analyze data of different types and scales to identify patterns (dependencies) and generate information to provide and automate decision-making. Revealing dependencies allows enterprises to gain value from their data as modern data analytics methods can uncover hidden data models through predictive, self-learning, and adaptive capabilities. The module intends to give students a set of theoretical methods and practical tools making possible to automatize data analytics in various domain specific problems.</td></tr>
<tr><td><b>3</b></td><td colspan="6"><b>Course Structure</b>
<i>1. Exploratory Data Analysis (EDA) and Data Visualization (elective)</i>
1.1. Introduction to R Programming.
1.2. Exploratory Data Analysis with Dplyr package.
1.3. Applied Plotting.
<i>2. Unsupervised learning. Cluster Analysis (core)</i>
2.1 Basics of unsupervised machine learning.
2.2 Methods of centroid-based cluster analysis.
2.3 Methods of connectivity-based cluster analysis.
<i>3. Supervised machine learning. Support Vector Machines and Decision Trees for Data Analysis. (core)</i>
3.1 Basics of Supervised machine learning.
3.2 Support Vector Machines
3.3 Decision Trees
<i>4. Prediction and Decision Making, and Results Visualization (elective)</i>
4.1 Prediction
4.2 Decision Making,
4.3 Results Visualization</td></tr>
<tr><td><b>4</b></td><td colspan="6"><b>Application Focus</b>
Students will consider different practical cases for which they will use data analysis tools and methods. Students will be encouraged to use that for the solution of real problems for companies or other stakeholders.</td></tr>
<tr><td><b>5</b></td><td colspan="6"><b>Scientific Focus</b>
Literature review and analysis. Deductive own research based on the literature. Scientific reflection and discussion in the teams.</td></tr>
<tr><td><b>6</b></td><td colspan="6"><b>Parameters</b>
• ECTS: 6
• Hours of study in total: 180
• Weekly hours per semester: 4
- Contact hours: 60
- Self-Study hours: 120
• Course characteristics: elective
• Course frequency: every year - spring semester
• Maximal capacity: 25 students
• Course admittance prerequisites: none
• Skills trained in this course: theoretical, practical and scientific skills and competences
• Assessment of the course: set of conducted laboratory works (defended reports – 50%) and written exam (50%)
• Teaching staff: teachers from Open Community of Practice</td></tr>
</table>

7	**Learning outcomes** 6.1 Knowledge • explain methods and tools for data analysis • explain methods and tools for data visualization 6.2 Skills • conduct exploratory data analysis and visualization • apply R programming language for data visualization • apply Python programming language and libraries for data analysis 6.3 Competence – ability & attitude • Students train to develop and discuss concepts in teams • They can present their results to companies and discuss in a professional context • Students work and set up a data analytic and visualization project for their respective case study
8	**Teaching and training methods** • lectures introducing concepts, methods and tools, own literature reading • group work in the case study project to practice concepts and methods, to develop skills and to work on case studies • presentations to communicate results and do a scientific discussion and reflection
9	**Curricula Integration** None
10	**References** *Roger D. Peng (2020): R Programming for Data Science. Retrieved from* https://leanpub.com/rprogramming *Roger D. Peng (2020): Exploratory Data Analysis with R. Retrieved from* https://bookdown.org/rdpeng/exdata/ *Rafael A. Irizarry (2021): Introduction to Data Science. Data Analysis and Prediction Algorithms with R. Retrieved from* https://rafalab.github.io/dsbook/ https://scikit-learn.org/stable/modules/clustering.html https://machinelearningmastery.com/clustering-algorithms-with-python/

5. References

[1] EU: The European Qualifications Framework: supporting learning, work and cross-border mobility, Luxembourg: Publications Office of the European Union, 2018

[2] EU: Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG), https://enqa.eu/index.php/home/esg/, Brussels, Belgium, 2015

[3] Gruen, G.; Tritscher-Archan, S.; Weiß, S.: Guidelines for the Description of Learning Outcomes, ZOOM partnership (www.zoom-eqf.eu), 2009

[4] Rajala, S.A.: Beyond 2020: Preparing Engineers for the Future. Proceedings of the IEEE, Vol. 100, pp. 1376-1383, DOI 10.1109/JPROC.2012.2190169, 2012

[5] European Institute of Innovation and Technology (EIT), “Quality for learning” EIT Quality Assurance and Learning Enhancement Model, https://eit.europa.eu/sites/default/files/eit_label_handbook.pdf, 2016