Tárgyleírás

Tárgy neve: Introduction to Data Science

Tárgyfelelős neve: Horváth Tomáš

Tárgyfelelős tudományos fokozata: PhD

Tárgyfelelős MAB szerinti akkreditációs státusza: AT

Az oktatás célja angolul / Aim of the subject:

Professional competencies to be achieved by data scientists are as follows:

a) **Knowledge:**
In order to be able to perform their work in an innovative way and do research (when necessary) in data science, they have comprehensive and up-to-date knowledge of general mathematical and computing principles, rules and relationships in areas related to data science and its application fields. They have comprehensive and up-to-date knowledge and understanding of the general theories, contexts, facts, and the related concepts of data science in the areas of data processing and transformation, programming languages for data science, building of statistical and machine learning models, analysis of the results and deployment of models into real life applications as well as ethics and legal aspects of data science. They are familiar with the principles of data processing architectures and pipelines as well as with the methods of describing and designing these architectures. They have extensive knowledge enabling them to perform business analysis, and to establish and run a data-driven enterprise. They have a high level of fluency in the language of data science – including its professional vocabulary and its characteristic features of expression and composition – both in their mother tongue and in English, at least. They are aware of methods and tools for competent and effective networking both in writing and speaking. They know the principles and problems of corporate social responsibility related to data-driven systems.

b) **Abilities:**
They are able to apply their mathematical, data science and informatics skills in a novel way in order to solve tasks in data science research and development. They are able to formalize complex data-driven tasks, to identify and study their theoretical and practical background and then to solve them. They are able to perform design, development, operation, and management tasks when operating complex data processing and management systems, decision support and expert systems. They are able to comprehensively understand, plan, organize, manage and control processes related to data science at management level. They are able to initiate collaboration and work in a team as well as on projects with data scientists or other professionals. They are able to assess the business, market and innovative value of planned or implemented data-driven systems, as well as their compliance with users' and social needs, and to validate data analytics software products. They are able to analyze and apply new problem-solving methods and procedures related to data science. They are able to apply their skills in a diverse, multidisciplinary professional environment. They are familiar with data science professional vocabulary, which enables them to express themselves at a high level, both orally and in writing, in their mother tongue and (at least) in English; i.e. they are able to participate in discussions and debates, to write reports, to work with, understand and utilize scientific and technical literature (e.g. professional books, chapters, articles etc.). They are able to plan and execute quality-management subtasks related to data science. They are able to professionally use scientific and technical information sources to obtain knowledge necessary for solving a problem, and to critically interpret and evaluate it. Under professional guidance, they are able

to carry out scientific research on their own, and to prepare for further studies at postgraduate level.

c) **Attitude:**
They follow professional and technological developments in their field. They are committed to critical feedback and evaluation based on self-examination. They are committed to lifelong learning, and are open to acquiring new IT competencies. They accept and make their co-workers apply the ethical principles of work and organizational culture as well as those of scientific research. They share their knowledge and consider it important to disseminate professional results. They consider it important to propagate and realize environmentally conscious behavior and social responsibility, and they promote them with the help of information technology. They are committed to having quality requirements met and to analyzing them with IT tools. They can be entrusted with developing and operational responsibilities that are in accordance with their professional competencies. are open to proactive collaboration with IT and other professionals.

d) **Autonomy and responsibility:**

They follow professional and technological developments in their field. They are committed to critical feedback and evaluation based on self-examination. They are committed to lifelong learning, and are open to acquiring new IT competencies. They accept and make their co-workers apply the ethical principles of work and organizational culture as well as those of scientific research. They share their knowledge and consider it important to disseminate professional results. They consider it important to propagate and realize environmentally conscious behavior and social responsibility, and they promote them with the help of information technology. They are committed to having quality requirements met and to analyzing them with IT tools. They can be entrusted with developing and operational responsibilities that are in accordance with their professional competencies. are open to proactive collaboration with IT and other professionals.

Az oktatás tartalma angolul / Major topics:

The course navigates through the basic concepts and principles behind the main data science models and techniques. Descriptive techniques such as clustering and frequent pattern mining are explained in more details while, in case of predictive techniques, the focus is put mainly on the concepts of a model, its parameters and hyper-parameters as well as the quality and validation of models including overfitting-underfitting and the bias.-variance trade-offs. Data quality and pre-processing issues related to various data types and modeling problems are also tackled. Finally, basic recommendation techniques and the CRISP-DM methodology are contained in the course as well.

- Clustering: k-means, agglomerative, DBSCAN, cluster validation;
- Frequent Pattern Mining: itemsets, association rules, quality measures;
- Linear Classification and Regression: model, parameters and hyper-parameters, validation, overfitting-underfitting and the bias-variance trade-off;
- Introduction to traditional prediction techniques (as black-box functions);
- data quality and pre-processing: noise, missing values, data transformation, normalization;
- the CRISP-DM process;
- recommendation techniques;

A számonkérés és értékelés rendszere angolul / Requirements and evaluation: exam

Irodalom / Literature:

- Peter Flach (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press.
- Jiawei Han, Micheline Kamber, Jian Pei (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar (2005). Introduction to Data Mining. Addison Wesley.