Co-funded by the Erasmus+ Programme of the European Union

ADSEE project

Applied Data Science Educational Ecosystem

Intellectual Output 5:

Guidelines for improvements in usage of data science methodologies in education

Document Reference

Project Title:	ADSEE (Applied Data Science Educational Ecosystem)
Project Number:	2019-1-HR01-KA203-060984
Project URL:	www.adsee.eu
Output:	Intellectual Output 5
Responsible Partner:	Faculty of Information Studies (SLO)
Output Start Date:	7th December 2021
Output End Date:	31st March 2022
Output Title:	Guidelines for improvements in usage of data science methodologies in education
Output Delivery Date:	March 2022
Status:	Final

Table of contents

Introduction

As a culmination of the project, transferable guidelines on data science in different professional environments, with special emphasis on the non-technical ones, were created (Intellectual Output 5).

In order to create comprehensive guidelines, all previously finalized results of the project were thoroughly examined and the most important lessons-learned were set up in a form of transferable recommendations. Moreover, guidelines actually represent a model for higher educational institutions, which will help them in adopting efficient and innovative approaches when it comes to teaching about data science methods and approach.

Although project partner responsible for Intellectual Output 5 was Faculty of Information Studies (Slovenia), all project partners participated actively in drafting, reviewing and finalizing the guidelines.

The guidelines, which are final result of Intellectual Output 5 of the ADSEE project, primarily respond to two necessities:

a) how to efficiently train individuals in data science

b) how to respond to current challenges and trends represented by data science in business environments.

1. The aim and objective of the ADSEE project

The main objective of the ADSEE project is to deliver useful educational and training program in data science (DS) through: development of educational modules, adaption of contents and methods according to envisaged needs of the target groups, creation of interactive didactic tools and production of guidelines and recommendations on innovative education approaches in DS.

Special attention is paid to data science in non-technical universities and its application in nontechnical business, were previous knowledge in this area is not mandatory. The innovativeness of the project lies in the modular approach allowing tailor-made courses development, according to the participants' specific prior knowledge and competences (or in absence of that knowledge and competences) and in a fully functioning online piloting repository which contributes to the development of participants' new skills and experiences.

The aim of the project is to contribute to the popularization of data science among wider public. The main target groups are higher education institutions and HEIs' employees, students, business/industry sector, institutions (ministries of labor, national employment agencies, employers' associations) and Digital Innovation Hubs (DIH). ADSEE project addresses individuals with in-depth knowledge about data science, those who are attending technical universities or are working in data science related sectors and individuals who know data science exists, are aware of its potential but still without an expertise to make decisions based on data science.

2. Previous project results and accomplishments

In order to achieve the main objective of the project, partners have accomplished a few results before conducting the Intellectual Output 5, all of which are finalized and were crucial for successful creation of guidelines with recommendations.

- Intellectual Output 1: Project partners delivered the report with the analysis of academic supply of the study programs related to data science accessible at HEIs and other education providers in partner countries and, on the other hand, analysis of the current market needs in terms of competences and experience required for data scientist and other occupations using data science approach.

- Intellectual Output 2: Project partners have designed the interactive online repository based on the analysis conducted as part of the Intellectual Output 1, especially using inputs obtained from the target groups. Based on these inputs, project partners developed a fully functioning interactive online repository that covers two main functions. It represents a part of a personalized training mechanism which can contribute to the development of users' new skills and experiences by delivering educational/training material in full-scale training case, filling the gap for increasing demand and limited supply of business efficient approach in training methods. The repository was further used for hosting the learning

materials. To access the repository, users log in using their existing account which guarantees a free access to the newly developed repository.

- Intellectual Output 3: Project partners have created four educational/training modules with a special attention to the non-technical sectors and covering a wide range of different industries and cross-domain topics (Retail and marketing, Culture and Tourism, Privacy and Job Market Signalling). All created modules are driven by industry and interdisciplinary examples and can be used as an improvement of current learning modules, as core part of new learning modules or for the self-study. Thanks to the modular approach used to develop educational and training materials, all modules are transferable and applicable in any study program, since they are structured as a flexible end-to-end business case avoiding a pure data scientific approach. Modules follow "from business problem to business usage" principle in order to fill in the gap between increasing demand and limited supply of practical training methods in data science approach to the business sector.

- Intellectual Output 4: In order to develop the pilot process, the educational modules elaborated in Intellectual Output 3 were integrated into the learning environments of project partners using the developed platform. Moreover, pilot process actually represents phase of the ADSEE project when all previously developed results had to be used in order to successfully conduct the pilots in partners' learning environments. The pilots were conducted by all higher educational institutions which are ADSEE project partners. During the pilot, participants were asked to fill in the survey in order for partners to obtain data for further analysis. Finally, report with students' and lecturers' insights about the pilot and quality of learning materials was prepared, together with a set of recommendations for integrating similar learning process in higher educational institutions.

3. ADSEE project as an answer to data science challenges

Many stakeholders and organizations are seeking to unlock the value that can be provided by data. In order to do so, they hire many data scientists, hoping to immediately drive results. However, many businesses fail to make the best use of their data scientists because they are unable to provide them with the right environment and raw material. Working process of data scientists can be represented with five key steps and each of these steps has its own pain-point, as follows:

- finding the right data
- getting access to the right data
- understanding tables and their purpose
- cleaning the data
- explaining in laypeople's terms how they work links to the organization's performance.

ADSEE project developed ready-to-use educational materials, but also strategies for HEI in order to empower future data scientists and those who are not technically oriented experts, to use the value that can be provided by data in a way to foster business development.

In respect to the steps of data science process and their main pain-points, results of the ADSEE project could help HEI and experts interested into data science to learn how to find the right data. Educational modules prepared as part of the ADSEE project are especially oriented towards finding the data (step 1) and therefore, encourage the "culture of finding the data".

As emphasized, once experts find the right data, they often face with the challenge of getting access to it (step 2). ADSEE educational modules are ready-to-use materials which were prepared in a way that stimulate working with data as well as discussing about data science challenges. Therefore, using the modules could help all interested parties to find useful strategies for accessing the data, but also discussion topics, how-to steps and questions that should be taken into account when approaching the process of getting access to the data.

ADSEE educational modules are shaped in a way to foster dealing with a real data set. In addition to the materials, all interested parties can get engaged with data set and therefore improve their skills related to understanding the data (step 3).

Moreover, beside the theoretical information in prepared materials and prepared data set, using the ADSEE education modules enables learners to practice their data cleaning skills (step 4). Designed modules as nearly "self-service explanatory" and provide learners with a unique possibility to get engaged with a data in a simple try-and-fail sandbox, by working in Jupyter Notebook and editing it. Moreover, it is important to emphasize the value of technical preparation of ADSEE materials in which the servers in the background are prepared so that they do not require any additional technical prerequisite to be able to use, which greatly speeds up the time to results.

Finally, when it comes to explaining the results of data analysis (step 5), ADSEE project with its results responded to one of the most common data science challenge – communicating the results of data anylsis to non-technical stakeloders. For this purpose, these guidelines for improvements in usage of data science methodologies in education were created, gathering conclusions of all project results and project lifetime (with a special emphasis on ADSEE educational modules and its implementation in teaching process during the pilot).

4. The purpose and the intended audience of the guidelines

The purpose of these guidelines is to motivate and help higher educational institutions (HEIs) in adopting efficient and innovative approaches in teaching about data science methods and approach. Guidelines bring tested and reliable methods on how to efficiently train individuals with different professional environments in data science.

A special emphasis in the guidelines, as in the whole ADSEE project, is on non-technical professional environments which could largely benefit out of professionals that are skilled in data science. Consequently, these guidelines are intended and aimed specially towards non-technical HEIs. Since current trends represented by data science in business environments show that data-driven decision making is a successful model in many non-technical industries, in the future HEIs could be faced with a challenge of finding an efficient way of teaching their students and learners about data science methods and approaches. Although students and learners in non-technical fields do not have to be skilled in technical side of data science, they could use data-driven decision-making approach in their professional environment very frequently. Finally, these guidelines could help HEIs in planning their shift towards teaching about data science, although they might have never taught about these topics.

It is becoming increasingly apparent that data scientists need to demonstrate skills necessary to convert data-based scientific inference into accessible and actionable insights for business and upper level management. Today's data scientists need to both straddle the worlds of business boardrooms and IT as well as become a hybrid of them. Therefore, guidelines are developed both for data science studies (the ones that already have implemented data science vertically in their curricula) and for a non-data science studies (studies that have implemented data science as a horizontal element in curricula as well as studies that have not implemented data science in curricula at all).

Guidelines represent a clear framework to support the implementation of teaching about data science in HEIs and in different business/industry sectors, containing: recommendations of fields/topics/learning content feasible and appropriate for upgrading by introducing data science, a proposed methodology for the implementation of data science approach, recommendations of appropriate audience (target group) for attending educational program or training that includes data-science approach together with a set of recommendations and advices from HEIs which have successfully implemented data science as a horizontal element in curricula.

Finally, since the approach proposed in the ADSEE project is innovative and affects business and labor market positively suggesting upgrading professionals' skills and knowledge, the long-term intention of these guidelines is to strengthen HEIs through know-how and capacity building of both teaching and technical staff, to help them achieve academic results of high quality and increase visibility of their activities.

5. Guidelines for improvements in usage of data science methodologies in education

4.1. The initial implementation of data science methodologies in education

As part of the ADSEE project, developed ADSEE educational modules were integrated into the learning environments of project partners. This process represented the initial implementation of data science methodologies in HEIs, which implemented data science as a horizontal element in curricula through this process. In total, 212 students from Netherlands, Croatia, Germany and Slovenia participated in the pilot process, which was ran in the following HEIs: University of Amsterdam (Netherlands), Algebra University College (Croatia), University of Siegen (Germany) and Faculty of Information Studies (Slovenia). Students and lecturers who participated in the pilot had different educational backgrounds and professional experiences. Before engaging with data science approach, majority of students (50 %) could only describe in general what the data science is about. In addition, most of the students (57%) were interested in data science above average.

During the pilot, students had the opportunity to gain insights into how data science methods and approach could be used in retail and marketing, culture and tourism, for job market analysis or for privacy protection. According to the survey conducted after the pilot, majority of the students (66 %) pointed out that they are very interested in further implementation of data science methods in their everyday work and that applying data science in their field of expertise could be very useful (53 %).

Finally, majority of the students who were included into the pilot process have recognized the value of data science in a way that majority of them emphasized that:

- decisions based on data science approach can increase business efficiency
- application of data science can assure better organization of the business process
- application of data science enables predicting of the future market trends
- application of data science enables efficient analysis.

These conclusions significantly indicate that data science approach and methods could strongly benefit many businesses and that, therefore, learning about data science should be included in wider circle of educational and study programs (which would enable learners and students to familiarize themselves with data science and learn how to use it in their field of expertise).

However, it is crucial to strategically implement teaching about data science in particular educational and study programs. Firstly, it is important to determine which sectors and businesses could largely benefit from data science approach and further implement teaching about data science in corresponding educational and study programs. Secondly, is it necessary to design learning path in chosen educational and study programs in accordance with students' technical skills and future usage of data science in their field of expertise.

4.2. Implementation of data science in particular fields and study programs

Although digitalized data is exponentially expanding in its penetration across the full range of business sectors, there are few sectors (and therefore educational and study programs) which could largely benefit from implementing data science approach. In general, data science could be beneficial in many different fields, not only computer science related fields and industries, but also social sciences and humanities.

Fields which work with large quantities of data are definitely quite likely to be beneficiaries of data science related approaches. Nevertheless, educational studies and psychology could also benefit from implementation of data-driven decision making, although these fields are part of the social sciences domain. Those who are in charge of documentation and archiving in educational studies and psychology collect and store a huge amount of data, which is used on a daily basis to run the operations. However, deep insight into stored data sets and analyzing the data could provide experts in this field with some indications and conclusions crucial for future business steps and planning next endeavors. For example, trend of decrease in population number in many countries determines decrease in number of pupils and students, which highly affects education sector. However, learners' interests and labor market needs are changing rapidly. Consequently, monitoring these changes, structuring information in some sort of data set, analyzing the data and decision-making based on these analyzes could enable experts in education sector to quickly adapt their supply to market's demands. This is why experts from social studies, i.e. educational studies or psychology could largely benefit from acquiring skills in data-driven decision making.

Moreover, there are a few fields which data science approach could enrich and it is not very abstract how data science could support these industries. For example, data science could be very useful in fintech and e-commerce. By its nature, fintech and e-commerce are industries that use large quantities of data to run the business on a daily basis. Beside gathering and analyzing data, for succeeding in these businesses it is crucial to be able to recognize trends and define future strategies. Although, one may suppose that C-level managers and decision-makers could benefit the most out of data-driven decision making, but the fact is that many experts on lower levels are often faced with planning the strategies and next moves as well as implementing innovations into the business process. This is why business, economy and marketing students could build their capacity by learning about data science approach as well as, eventually, enrich their business portfolio and contribute to business development.

Since migration of the population is highly affected by different social and economic changes and situations, traffic is the field that monitors and notes these migrations. Whether daily commutes, temporary or permanent relocations, migrations drive the changes in many other fields. For example, urban planning, construction and supply of different products and services should be adjusted depending on migrations. This is how traffic experts, if they are skilled in deriving conclusions out of analyzed data sets, could accomplish cooperation with experts from other fields, who could benefit out of migration trends preparing their business strategies accordingly. To describe one example, traffic experts could get into the role of advisors who could provide

urban planners and construction experts which changes in urban landscape are expected according to migration trends.

Experts in tourism and culture sectors could adjust services which they provide according to the conclusions that could be made out of analyzing movements of tourists, their spending habits and interests. Not only that existing services could be enriched, but new ones could be tailored and ensure additional profit. For the same reason, health-care experts could benefit from knowing how to tailor new approach to the patients, not only by improving existing procedures, but also trying to design an offer new set of services to the patients.

Therefore, learning about data science methods and approaches should be included in educational and study programs of those fields whose future experts could benefit from acquiring data science skills. After considering which educational and study programs would that be (in accordance with fields that could ensure new business opportunities thanks to the data-driven decision making), it is necessary to design these educational paths in accordance with:

- learners' previous technical background
- level of data science knowledge and skills that should be acquired for successful usage of data science approach in a particular occupation.

4.3. Proposed methodology for the implementation of data science approach on non-technical HEIs

Technically oriented data science studies and educational programs include working on a real data-set which includes analysis of the data set. Primarily, these studies encourage development of technical skills, methods for shaping data into particular data sets and different methods of analyzing specific types of data sets. When setting up the methodology for implementation of data science approach on non-technical HEIs, it is important to realize that majority of the students on these institutions do not have significant technical backgrounds and skills, but also that they are not required to acquire high-level technical skills in order to perform their job successfully in the future.

This is why a recommendation for implementation of data science approach on non-technical HEIs would be to engage students with a real data-set, but only in order for them to gain insight into the process. It should be emphasized that these experts, no matter if they will be decision makers or executives, will not structure data or perform elaborate analysis. Nevertheless, it is recommendable to introduce them with the logic and the process behind data-driven decision making. They could be introduced with the following processes: data collection, exploratory data analysis, creating models and model evaluation.

When it comes to materials which learners will use, project partners have made interesting conclusions based on analysis of the questionnaire conducted after the pilot process. Students who had the opportunity to use ADSEE educational modules and engage themselves with data

science were asked about their opinion on used educational materials and data content prepared on the Jupyter Notebook platform. It is very interesting that nearly third students pointed out that materials and platform are useful but too detailed and another third of the students emphasized that materials and platform are useful but could be more detailed. Since majority of the students did not have significant technical backgrounds and skills before participating in the pilot, these insights are very important for educators and HEIs. This means that educators should examine both backgrounds and interests of learners in a particular group, in order to adjust complexity of teaching to the learners' needs.

The core characteristic of educational and study programs which implemented data science approach would be to help learners to overcome lack of their technical skills and learn how to make conclusions out of analyzed data sets. This means that the main process for these students (future experts) starts in the point in which the process finishes for majority of data scientists and technical experts. This is the point where data analysis is done, but conclusions are yet to be made. This is the reason why non-technical fields/HEIs could benefit strongly from data science – in many cases, conclusions made thanks to data science analysis can upgrade, develop or enhance processes in social sciences or humanities.

However, rather than targeting proficiency in the full range of data science skills, we feel that successful module with implemented data science approach should help learners to overcome lack of technical skills and focus on using data science approach and conclusions based on data analysis in wide range of fields – from setting up business strategies and models to predicting future trends and adapting business plans accordingly.

When designing the learning path, it is important to notice which learning contents could be enhanced with data science approach. Firstly, it highly depends on the learners and their learning context. In many cases learners do not need to know the precise details of analytical methods, but they should be able to interpret the results of data analysis, and be aware of risks and limitations of these methods. It is essential that there is a general understanding of the applicability of data driven approach in the light of traditional research methods.

From an experience point of view, if the target learners have not had any experience with data science, we should depict the power of data science through the lens of tangible topics (like weather, or medical applications) by highlighting the analytical power and the potential of predictive analytics. Additionally, for more advanced learners, we should cover topics like collecting and visualizing the data, creating simple analytical models and the basic evaluation of those models. As learners gain experience, we can go for real samples from their expertise area and start showing them how to build and evaluate more complex models and analyses.

4.4. Recommendations from HEIs which have successfully implemented data science as a horizontal element in curricula

As a final result, based on the experience, HEIs which have successfully implemented data science as a horizontal element in curricula, have prepared a set of recommendations. These HEIs are: University of Amsterdam (Netherlands), Algebra University College (Croatia), German National Library and Leibniz Information Centre for Science and Technology (Germany) and Faculty of Information Studies (Slovenia).

Recommendations are prepared in a way to bring important advices for preparation of educational programs/trainings that introduce data science approach and methods, bearing in mind that learners have non-technical backgrounds. Additionally, recommendations describe effective ways of converting data-based scientific inference into accessible and actionable insights for professionals/experts in non-technical fields.

Since preparing and conducting classes and lessons is a process, recommendations are designed as simple ready-to-implement steps/phases, a recipe for successful implementation of data science as a horizontal element in curricula.

Phase 1 – Preparation

- For the beginning – know your audience. This will largely help you to successfully prepare and conduct all next phases. It is crucial to examine the background and expertise of your learners. This will help you to understand how they could benefit from data science approach. Later on, you can approach them individually introducing them with strategies that could be essential to them. This is the ground for successful tailor-made learning process.

- Together with backgrounds and expertise, it is important to examine learners' previous knowledge about data science as well as their preferences for learning about the topic. Of course, the lecturer is the one who structures the learning outcomes, but it would be useful to know about knowledge starting point of your audience if you would like to assure smooth learning process implementation.

- Once these steps are done, you can proceed with the final step of the preparation phase – personalization. It is one of the most important factors in learning. As in any educational program, the recommendation is to think about learners' goals, context and interest. Finally, you as a lecturer will be able to align the program, content and assessment strategies according to the audience.

Phase 2 – Diving into data science

- Once you set the ground for successful tailor-made learning process, we recommend to start with topics that learners will find immediately helpful. These could be: reading code, version control, visualizations, automated reporting as well as the basics of managing your data and code, processes and conventions.
- Introduce learners with data visualization methods – this could help learners to understand the data set as well as possibilities and power of the visualization.
- Show case methods with short recipes that are applicable to their field. This approach highlights the relevance of the data science approach and motivates the learners.

Phase 3 – Analysis – potentials and concerns

- The course (teaching and learning process) should cover different steps of data science approaches. Present the process and let learners to engage themselves with all data science process phases: data collection, accessing data, exploratory data analysis, model creation, and evaluation, understanding the data and cleaning the data. Level of engaging with each part of the process will be determined by learners' technical skills and their future intention to use data science methods and approach.
- The last part of the data science process, explaining and decision making should be part of another phase, in order to devote enough time and attention to this matter.
- It is important to show risks and limitations of data analysis methods in use and also the ethical and privacy concerns. This can be done by maximizing the openness and transparency of software, algorithms and data pipelines by publishing these elements under an open license.

Phase 4 – Data science decision making path

- At least one real, tangible data-set should be included in the training, what the audience is well aware of and capable to understand.
- Oftentimes non-technical learners do not need to know much about the software technology behind data science. However, it is critical to develop and deploy such user interfaces, which are intuitive, self-explanatory, and make the user understand how certain results from the analysis are produced.
- Demonstrate areas in which (managerial) decision making is evidently suboptimal. Introduce tenets of evidence-based decision making. Have students come up improved evidence generation methods through data science methods.

Phase 5 – Data science in business challenges

- Using an EC digital playground accessible to all learners with relevant OER content will allow for a base level to be defined for learners' support. Look at the features of providers of training such as Datacamp.com for an estimate of the requirements for the digital playground.
- Ask experts in the field to suggest questions that may be resolved using data science methods. This could help you to prepare a themed-workshop, during which students should resolve a real business challenge.
- Organize hackathons in which teams of data science and non-technical students need to collaborate to solve the problem using scientific inference.

References

- Top 5 challenges of data scientists: https://www.castordoc.com/blog/top-5-challenges-of-data-scientists