Grădinaru, Giani Ionel; Dinu, Vasile; Rotaru, Cătălin-Laurențiu; Toma, Andreea

Article

The development of educational competences for Romanian students in the context of the evolution of data science and artificial intelligence

Amfiteatru Economic

Provided in Cooperation with:
The Bucharest University of Economic Studies

Suggested Citation: Grădinaru, Giani Ionel; Dinu, Vasile; Rotaru, Cătălin-Laurențiu; Toma, Andreea (2024) : The development of educational competences for Romanian students in the context of the evolution of data science and artificial intelligence, Amfiteatru Economic, ISSN 2247-9104, The Bucharest University of Economic Studies, Bucharest, Vol. 26, Iss. 65, pp. 14-32, https://doi.org/10.24818/EA/2024/65/14

This Version is available at:
https://hdl.handle.net/10419/281807

THE DEVELOPMENT OF EDUCATIONAL COMPETENCES FOR ROMANIAN STUDENTS IN THE CONTEXT OF THE EVOLUTION OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Giani Ionel Grădinaru[1*], Vasile Dinu[2], Cătălin-Laurențiu Rotaru [3] and Andreea Toma[4]

[1)]Bucharest University of Economic Studies, Bucharest, Romania and Institute of National Economy, Romanian Academy, Bucharest, Romania

[2)]Bucharest University of Economic Studies, Bucharest, Romanian and Academy of Scientists, Bucharest, Romania

[3)4)]Bucharest University of Economic Studies, Bucharest

Please cite this article as:
Grădinaru, G.I., Dinu, V., Rotaru, C.L. and Toma, A., 2024. The Development of Educational Competences for Romanian Students in the Context of the Evolution of Data Science and Artificial Intelligence. *Amfiteatru Economic*, 26(65), pp. 14-32.

DOI: https://doi.org/10.24818/EA/2024/65/14

Article History
Received: 17 September 2023
Revised: 6 November 2023
Accepted: 7 December 2023

Abstract

The study explores key academic competencies and professional skills in data science in the context of the development of artificial intelligence, highlighting their importance in the business environment. Using the "2022 Stack Overflow Annual Developer Survey" dataset and machine learning methods such as principal component analysis, K-means clustering, and logistic regression, professional skills in science are analysed the data. The research targets the distribution of jobs in the field, the level of experience, the languages and analysis programs used, the support offered by companies, and the dynamics of data science teams, as well as the impact that artificial intelligence has on the field. With their help, a comprehensive understanding of the impact of academic training on career opportunities in the field of data science is provided, contributing to the development of the profile of the qualified specialist in this field. The research also provides relevant pointers and recommendations for enhancing the skills required in data science in order to outline a skilled profile and fulfil the demands of the business environment in a world dominated by data analytics and artificial intelligence. By including academic skills in the process of

* Corresponding author: **Giani Ionel Grădinaru –** e-mail: giani.gradinaru@csie.ase.ro

training data science specialists, the research brings innovation and highlights the skills needed to be trained in the academic field to facilitate the employment of graduates in specific fields of data science. This aspect is significant because, in practice, it has been observed that most specialists working in data science rely on independent learning rather than skills acquired in the academic field.

Keywords: data science, artificial intelligence, academic skills, professional skills.

JEL Classification: I23, J23, C49, M15.

Introduction

Data science has become one of the most exciting fields in the age of technology and artificial intelligence (AI). With the advancement of technology and the increase in the amount of data available, the job market has experienced a significant increase in the number of data science jobs. Universities have recognised the importance of this field and started building specialised programs to train future data scientists. At the same time, specialised publications constantly and actively write about this topic, highlighting the progress and new trends in the field.

According to Smaldone et al. (2022), the role of a data scientist usually involves the identification of business directions, providing decision support for the requirements and needs of the business environment. Thus, communication skills have a critical role, being necessary to mediate the language used starting from the technical analysis to the information transmitted to the business, management team, whether we refer to an intelligent data visualisation or not. Curiosity and creativity are key skills, because among large data sets, a data science specialist has to search, research, test, and draw various hypotheses that will be the basis of the proposed solutions.

In the context of the development of artificial intelligence and data science, academic skills play an essential role in the advancement of these fields. Both AI and data science are based on the foundation of academic research and education to develop innovative solutions and understand the complexity of these technologies (Leal et al., 2023).

In the context of data science and artificial intelligence, academic skills are critical to collect, process and interpret data. Data scientists must understand the mathematics and statistics behind analytics and machine learning techniques to create accurate and robust models. As stated by Irizarry (2020) developing effective AI algorithms requires a solid knowledge base in programming, mathematics, data analysis, and machine learning. Understanding fundamental AI concepts such as neural networks, deep learning, and machine learning algorithms enable researchers to create and optimise advanced AI models.

In parallel with the technical aspects, academic skills are closely related to the ethical and social aspects of AI and data science (Chen et al., 2020). Professionals in these fields must be aware of the impact that developed technologies can have on society, as a result, it is necessary to consider privacy issues and identify possible risks and biases. They must approach the development of AI technologies responsibly and work to ensure ethical and fair use of these technologies.

Specialist training programs, courses, and academic programs have evolved simultaneously with the labour market, but the branches of the field are diverse. Whether referring to data engineers, who ensure the proper flow of raw data needed for analysis, to data analysts, who deal with identifying data sources, cleansing, and quality assurance of the set, as well as intelligently visualising the results, or to specialists in data science, which builds predictive models, they represent roles that are part of the same field. However, each role requires rigorous training to become a specialist in the analytics process, from data accumulation, cleaning, processing, modelling, interpretation, or even prediction. In this context, most of the time training starts in academic programs (Irizarry, 2020). It is highlighted that the connection between the professional and the academic environment is very important, because in order to be able to occupy the position of specialist in data science in a certain field, it is necessary to prepare and develop the necessary skills.

Wing (2019) states that data science appears to be a vast field defined as a set of fundamental principles that guide the extraction and understanding of data. It is often associated with terms like BD (Big Data), data mining, and AI. To fully understand this field, it is necessary to identify the underlying principles and understand what it offers. This process provides clarity in data science requirements for both academic and professional environments.

The aim of the current research is to expose the vital link between the educational competencies and the professional skills required in the field of data science, in the context of the development of AI with the help of quantitative methods used in machine learning (ML - Machine Learning, a branch of AI), such as component analysis main methods, K-means Clustering (KM) and logistic Regression (RL), and to determine the profile of the specialist who can be considered qualified or unqualified in the field of data science.

In the first phase, the research focuses on the professional skills sought in the field of data science, including the technical and analytical skills needed to become a specialist in this field and to meet the demands of the business environment. It will then identify the typical profile of a data scientist and highlight its importance in the business environment by exploring the role, responsibilities, and added value it brings to decision-making and the development of solutions based on data analysis. Finally, the research focuses on identifying the essential academic skills to become a data science specialist, by analysing academic programs, courses, and relevant study subjects that contribute to the development of these skills, such as programming, data management, and artificial intelligence courses.

The present research aims to explore and highlight the key professional and educational competencies in data science, alongside their importance in the business environment, providing a comprehensive understanding of the impact of academic training on career opportunities in this field, as well as supporting the importance of AI as a key asset in the development of the science profile of the data.

1. Review of the scientific literature

The use of statistical methods and models plays a crucial role in the education sector, by revolutionising the way this field identifies the skills and knowledge that must be introduced into the education curriculum for the training of data science specialists.

Through the use of these advanced techniques, the need to align universities with the demands of the data science market has been highlighted.

Educational and research practices can be shaped by the AI – ML couple in the future, leading to improvements. As stated by Alqahtani et al. (2023) in research, key applications include text generation, data analysis and interpretation, literature review, formatting, and editing. In the education sector, the AI – ML couple can support personalised learning, assessment, specific learning plans, personalised career guidance, and mental health support. Despite the advantages, the responsible use of these technologies must address ethical and algorithmic imbalance issues. According to Chassignol et al. (2018), AI is revolutionising education by providing personalised learning solutions, adapting teaching methods through the emergence of individual guidance and effective assessment solutions. Although AI-powered platforms personalise the content, concerns about AI dominance still linger. However, researchers still draw attention to the importance of maintaining human mentoring and social interaction in education, a fact also emphasised by Tuba et al. (2023), who claims that AI tools do not contribute to the full development of academic skills, not yet having the ability to systematise information. However, AI can influence the future roles of humans and their ways of doing business. Pelău, Ene and Pop (2021) argues that the development of the AI – ML couple represents one of the main paradigms of contemporary society, having a significant impact on individual life and society as a whole, as well as on the economy. The use of AI in people's daily activities and in the interaction between companies and consumers brings many advantages, such as increased efficiency and an engaging interaction. However, there are also concerns about the future development of this field. Amidst its ability to store a large amount of data about individuals' behaviour and process this data quickly, there is a risk that forms of AI will become smarter than humans and influence their decisions.

Particularly interesting for scientific progress and industry evolution is the rapid transition of the AI – ML couple from a narrow area of research in a limited number of academic laboratories to a key topic in the business world. This process is accompanied by the emergence of other new paradigms, such as BD and the Internet of Things (IoT). In addition to these, a new discipline called data science gradually took center stage as the main branch of knowledge, covering all relevant methods and work processes to translate data into effective business solutions. Learning the main principles and capabilities of this discipline becomes a new academic and business challenge. Another cause is represented by the emergence of data specialists (specialist in data science), who have become motivated factors for this transformation (Kordon, 2020).

The term data science was introduced two decades ago, bringing new topics of discussion and analysis in various fields (Cao et al., 2016). According to Grossi et al. (2021), data science is an important sector with applications in business, industry, and education, encompassing areas such as AI and ML, experimental design, and data-driven modelling. Moreover, statistics becomes an essential discipline, providing tools and methods in the process of identifying structure, providing an overview of the field of data science. In this context, obtaining reliable results, developing predictive models, and understanding data structures are conferred by statistics. The idea that statistics, along with ML, plays a fundamental role in data science is reiterated by a group of leaders of the American Statistical Association, which issued a statement on the special powers of these disciplines, highlighting as central elements of the current field. Dyk et al. (2015) says that, data

science is a set of fundamental principles that support and guide the extraction of information and knowledge from data. A concept closely related to this field is data mining, the effective extraction of knowledge from data through technologies that integrate proprietary principles. The applicability of data science is varied, distinguishing itself in business, marketing, finance, sales, and more. Therefore, a successful data scientist must be able to approach business problems from a data-driven perspective and understand the fundamental principles of data analysis (Provost and Fawcett, 2013).

It is necessary to start from the basics of statistics and pay special attention to the evolution of data science, in order to achieve an exhaustive presentation of this field (Donoho et al. 2015). In the context of AI development, data are today considered valuable assets, and data mining has become a very important factor in improving the competitiveness of enterprises. This rapidly growing field encompasses information management, social sciences and computer science. The domain of big data is characterised by the velocity, volume and variety of data, making it valuable for data mining. However, many companies struggle to fully leverage the value of Big Data amid a lack of skilled professionals and uncertainty about how to effectively implement big data analytics to deliver business value. Thus, to maximise the potential of projects, organisations seek data science professionals and business experts who fully understand the company's business model (Xu et al., 2022).

The development of data science is the main driver of the new generation of AI – ML tandem specialists. These fields are attracting increasing interest from states, companies, and the educational sector, enjoying important initiatives from various actors such as the United States, China, and the European Union. Wing (2019) show that the field of data science involves a long process, which encompasses both data analysis, the work done before and after it, the issues related to ethics and data privacy, as well as the importance of data collection, processing, storage, management, analysis, and visualisation, alongside the valorisation of all by users, for decision-making in various fields. The life cycle of data is a complex one, consisting of various stages, starting from data generation to their interpretation, with an applicability also at the level of AI techniques. Furthermore, it is necessary to promote ethical responsibility and data protection for each phase of the data life cycle.

The contemporary period is characterised by a rapid increase in the popularity and utility of data science, due to its use in business, industry, and academia (Ley and Bordas, 2018). De Veaux et al. (2016) predict in their studies the need for hundreds of thousands of jobs in the field of data science in the next decade, leading to an increase in study programs in this field within universities. At the level of the business environment, companies from various branches have noticed the need to hire more people in the data science department. The data science job market provides a detailed list of job titles such as Data Science Specialist, Data Analyst, Data Engineer, Statistician, Data and Business Analyst, Business Intelligence Analyst. In this sense, it is very important to clearly define the skills required for this sector, as well as those specific to each function. As for the academic branch, there is a rapid increase in the number of institutions wishing to develop programs to train specialists in this field. In addition to these, there is also a promotion of the field of data science, carried out by various publications, which presents this career as an interesting and viable option for the future (Provost and Fawcett, 2013).

Regarding the role of a data scientist, it was described by Cao (2019) as the most attractive workplace of the 21st century. The role of data analytics, data science and the AI-ML

couple in information and communication technology and in the disciplines of science, engineering, and technology is essential. However, the qualifications and competencies of a data scientist are not yet clearly defined. In this sense, there is a need to establish the duties of a data science specialist who belongs to the new generation, able to transform science, technology, innovation and the current and future economy. While there is a notable increase in the number of data science courses helping to define this new profession, employers in this segment, including entrepreneurs, salespeople, and large enterprises, report the limited availability of qualified data specialists to assist them in strategic development and gives them competitive advantages in the future. Cao (2019) highlighted the discrepancies in existing professional and educational markets, the lack of standardisation and accreditation of data science responsibilities and competencies, and the urgent need to standardise and improve data science qualifications and education.

In the context of the job market, it is essential to define the skills required for Business Data Analytics and Data Science specialist positions. Using a text analysis by Radovilsky et al. (2018) on the Glassdoor.com platform, the most sought-after skills are shown to be Python - 72% (Familiarity with Python syntax and ecosystem), R-64% (Knowledge of statistical programming language R), SQL - 51% (Managing databases using structured queries), Hadoop 39% (Processing Big Data with Hadoop), Java - 33% (Familiarity with Python's syntax and ecosystem), SAS - 30% (Knowledge of SAS statistical programming language), Spark - 27 % (Processing Big Data with Spark), Matlab - 20% (Knowledge of numerical computation and programming in Matlab), Hive - 17% (Processing Big Data with Hive), and Tableau - 14% (Knowledge and concepts of statistical visualisation). As data science professionals are increasingly in demand and experienced ones are still rare, therefore, the idea arises that theoretical training is not enough to excel in this field. In addition to knowledge of programs and algorithms, practical experience in solving real business problems plays a particularly important role (Berthold et al., 2019).

Technology and the field of work have evolved and the curriculum needs to adapt to the new requirements (Hardin et al., 2015). The literature on Big Data skills mainly focuses on the demand side (companies), the supply side (universities and students), and the relationship between them. Following an analysis of recruitment information from companies (to identify the knowledge and skills required for BD positions) and an assessment of university programs (to see if academic training aligns with industry expectations), it was determined that the introduction of a more sustainable and efficient system for training and managing skills in this field. In addition, students' perspectives should be taken into account, to better understand their level of competence and to assess the effectiveness of university programs (Xu et al., 2022).

Currently, according to Bile Hassan et al. (2021), there are more than 530 programs in data science, analytics, and related fields, most of which are master's and certificate programs, offered both online and in traditional institutions. In this sense, an example of good practice is represented by the guide made for academic institutions wishing to develop bachelor's programs in this field, which describes a curriculum for the bachelor's program in the field of data science. The curriculum is detailed and presents courses in related fields, emphasising the importance of collaboration between departments and faculties in the process of developing such an interdisciplinary program. Along with these, there is at the level of the educational offer, and a set of courses that support the training of future specialists in data science, which emphasise disciplines such as exploratory analysis, data

set cleaning, and transformation, the Bayesian statistical model that essentially involves learning from data and probabilities, regression models, dimensionality reduction, ML, model performance, data mining from online platforms, sentiment analysis, relational databases (Hicks and Irizarry, 2018).

2. Research methodology

In accordance with the specialised literature, the research will focus on the following pillars: Data Analysis, Data Collection, Database Administration, Ethics, Prediction, Programming and Project Management. In order to fulfil the objective related to the academic skills in data science that students acquire after graduating from master's programs in data science in Romania, the discipline sheets of the main master's programs in the field of science were taken from the websites of the universities from Romania. Each discipline was included in one of the main pillars of the analysis:

- The Faculty of Mathematics and Informatics (FMI), which includes the master's programs "Data Science for Industry and Society", "Data science", "BD - data science", "Analytics and Technologies", which has three main pillars of interest: Data Collection, Data Analysis, and Programming, complemented by disciplines included in Ethics and Prediction.

- The Faculty of Automation and Computers (FAC), with the "Machine Learning" and "Database Administration" master's programs, which includes most of the pillars. The most subjects focus on Programming and Project Management, followed by Data Collection, Database Administration, Data Analysis, and Ethics.

- The Faculty of Cybernetics, Statistics and Economic Informatics (FCSIE), with the master's program "Applied Statistics and data science", includes most disciplines that belong to the main objective of Data Analysis, as well as Programming and Ethics activities.

- The Faculty of Economics and Business Administration (FEAA) with the Master's Data Mining, as previously presented, focuses on Data Analysis, Data Collection, and Ethics.

From the visualisation of the data for the pillars addressed by each faculty, the assimilated academic skills for each individual pillar were grouped into professional skills, which could also leave their mark on the professional career. Thus, an association of data science roles and academic competencies pursued by each faculty in the list above has been identified.

To meet the research objectives related to professional skills in the field of data science, the 2022 Stack Overflow Annual Developer Survey (Stack Overflow, 2022) data set was used in the current analysis. It is based on the questionnaire applied in 2022 by Stack Overflow, created with the aim of identifying the experience of specialists in the technologies used and their experience in professional environments, globally. After filtering the responses based on the job held by the respondents, records of participants with a data science background were retained, with 11071 responses. In this context, the dataset underpins the analysis of professional data science skills and provides insights for hypotheses such as:

- What is the distribution of jobs in this field?

- How many years of experience do the specialists have?

- What are the programming languages, analysis programs used?
- What is the support offered by companies to employees?
- What are the dynamics of data science teams in companies?

Through the use of descriptive statistics, the information from the data set was synthesised, this being presented in figure no. 1, for which the Power BI visualisation program was used.

Figure no. 1. Data Science – Professional Skills

Source: Authors representation in Power BI based on the "2022 Stack Overflow Annual Developer Survey"

The "2022 Stack Overflow Annual Developer Survey" (Stack Overflow, 2022) provides insight into the data science job market. The main jobs on the data science labour market identified, according to the questionnaire, are: Database administrator, data science specialist or machine learning specialist, Data Engineers and Data/Business Analyst. Of the

total respondents, 70% are full-time employees, and 30% are master's graduates. Hybrid (45%) and remote (39%) work modes predominate, with only 15% working exclusively from the office. Preferences for company size differ: most opt for 20-99 employees, 100-499 employees, and 2-9 employees (48% overall), while 51% go for larger companies. As for experience in data science, those with 1-3 years, 3-5 years, and 10 years of experience stand out. It should be noted that 71% consider their activity in the field of data science to be a hobby, highlighting the passion. Although 80% of employers provide time for learning, 61% of employees use company development resources such as Coursera, Udemy, Pluralsight, Codecademy, and edX. In order to assess how education level contributes to the qualification or non-qualification of a data science specialist for industry, a number of steps were taken. In the first phase, the initial 40 variables were reduced, by applying Principal Component Analysis into two principal components, resulting in two major variables that define the profile of a data scientist. Principal component analysis reduces the number of professional and academic components, being a method of identifying data patterns and expressing data in a way that highlights similarities and differences between them (Jolliffe, 2011).

The name and description of the variables extracted from the survey "2022 Stack Overflow Annual Developer Survey" (Stack Overflow, 2022) are presented in Table no. 1.

Table no. 1. Description of variables

Name of variable	Description
Employment	The main field or branch of work with which the respondent identifies
MainBranch	The respondent's employment status (e.g., full-time, part-time, self-employed, etc.).
Remote Work	Form of remote work or not (e.g., Fully remote, Hybrid, NA).
Coding Activities	The type of coding activities the respondent is involved in (e.g., Hobby, Contribution to open-source projects, etc.).
EdLevel	The highest level of education attained by the respondent.
Buy New Tool	Acquisition of new programming tools.
Years Code ,Years Code Pro	The number of years the respondent has coded, both as a hobby and as a professional.
Organisation Size	Size of the organisation the respondent works for.
Language Want To Work	Programming languages the respondent has worked with and wants to work with.
Database Have Work	Databases the respondent has worked with and wants to work with.
Platform Have Worked	Platforms the respondent has worked with and wants to work with.
OfficeStackAsync	Used and preferred asynchronous collaboration tools.
Age	Age of the respondent.
Work Expirience	The number of years of work experience the respondent has.
Knowledge_3, Knowledge_4, Knowledge_5	Questions related to the respondent's knowledge and expertise in specific areas.
Time searching, Time answering	Time spent searching and responding to programming tasks.
True/False_1, True/False_2, True/False_3	True/False_1, True/False_2, True/False_3: true/false answers to certain statements or questions.

Source: Authors' own processing based on the "2022 Stack Overflow Annual Developer Survey"

For each main component containing the variables in Table no. 1, the KM clustering algorithm was applied to re-code the observations according to the respective component, obtaining distinct groups (clusters). KM clustering is a method of partitioning a data space into K groups, where each group has a specific average value, to be able to define the characteristic of each main component. In KM, each data point in the set is assigned to the group closest to its mean value (Sterling, Anderson and Brodowicz, 2017).

This procedure was repeated for the other two principal components, resulting in a new re-coded database. The KM algorithm was then applied again to this new database to obtain a column representing the profile of a data scientist as skilled or unskilled.

In the final step, logistic regression (RL) was used to predict scenarios and identify the aspects that need to be improved in order to achieve the profile of a qualified data scientist. RL is a widely used statistical model that allows multivariate analysis and modelling of a binary-dependent variable. (Ranganathan, Pramesh and Aggarwal, 2017). As a result of the presentation of the main components and the grouping of their characteristics using KM, it is possible to define, with the help of logistic regression, the academic and professional competencies that should be adjusted to support the career of a future specialist in data science.

This review provided relevant guidance and recommendations regarding the influencing factors and skills required to obtain a qualification in this field. Logistic regression is mainly used because of its function of predicting the outcome of an event using binary values. In the present case, it determines whether the respondent is skilled or unskilled in the labour market. The current research uses the logistic regression function to demonstrate how the identified characteristics influence the predicted outcome, namely whether a respondent is qualified or unqualified to become employed in the business environment in a certain field, similar to Peng, Lee and Ingersoll's (2002) logistic regression study.

The mathematical formula of the model was:

$$P(y=1 \mid x) = 1 / (1 + \exp(- (b0 + b1*x1 + b2*x2))) \quad (1)$$

where:

- P(y=1 | x) – represents the probability that the dependent variable y has the value 1, where 1 is the desired event, given the value of the predictors x;
- b0, b1 și b2 – are the estimated RL coefficients, showing the weights associated with each predictor;
- x1 și x2 – represent the values of the predictors associated with the independent variable x.

3. Results and discussions

Studies that have looked at this topic have focused on discovering academic and professional skills that drastically contributes to employment opportunity. Mainga et al. (2022) addressed the topic of identifying the respondents' skills that make them eligible for the business environment in a certain field. The study was limited only to the descriptive component of the respondents' skills supporting their eligibility to be employed. The current research used the respondents' already acquired skills in the labour market, predicting skilled or unskilled status, as well as how respondents can improve their skills to become data scientists for the business environment.

The academic skills assimilated for each individual topic were grouped following the analysis of the subject sheets, obtaining valuable information. Thus, the academic skills acquired for the study of the subjects included in **the topic of Data Analysis** were: the ability to identify the necessary algorithms and apply the right analysis to deliver the results needed for decision-making purposes to solve the requirements, the ability to approach a hypothesis from multiple perspectives, and to be aware of the limitations of models. After studying the subjects from **the Data Collection topic**, the academic competences undertaken are: the knowledge needed to transform large data sets into complex projects by applying qualitative and quantitative methods, applying exploratory techniques, detecting extreme values or applying the correct principles for the implementation and development of analyses for Big Data systems. For the topic of **Data management**, the main academic skills acquired are to know the principles of database management and the application of the necessary techniques. The forecasting disciplines provide students with the analytical knowledge to develop hypotheses and identify predictions for existing trends. Following the study **Programming courses**, students will acquire skills in using programming languages to solve analysis requirements, and ethics and project management disciplines will provide graduates with the informational foundation needed in the process of project development, customer service delivery, and data science entrepreneurship.

The nature of each institution has also left its mark on the data science programs. Thus, according to the professional skills described in the previous part of the paper, an association of data science roles and academic skills pursued by each faculty was identified. From the list above, the following can be noted:

- The Faculty of Mathematics and Informatics (FMI) prepares specialists whose skills fall within the professional skills for the roles of Data Engineer or Database Administrator, by combining the three: data collection and analysis, alongside programming skills.

- The Faculty of Automation and Computers (FAC), outlines the academic skills covering a wide number of topics in data science, with attention directed to Programming and Project Management, as a result the role that matches the professional skills described being a specialist in data science.

- The Faculty of Cybernetics, Statistics and Economic Informatics (FCSIE) and the Faculty of Economics and Business Administration (FEAA) have as their main objective the outline of academic skills for Data Analysis, an aspect that is also found in the FEEA curriculum.

From the subject sheets of each master's program, the computer programs used in the training of students' skills were also extracted:

- FMI works with R, SAS and Python, and for the previously realised association between the academic skills outlined by the master's programs within this faculty and the professional skills for the roles of DE or Database Administrator, according to the requirements of the labour market, it is necessary to add in the program and the following programs: SQL, Oracle and Microsoft Azure.

- FAC covers a wide range of programs such as Python, SQL, Linux, Matlab, Oracle, and Tableau. As a result, to align the academic skills trained with the professional skills required for the data scientist role, the R program and Cloud technologies could be added.

- FCSIE is the one that covers almost all the technologies required for the role of Data Analyst and Business Analyst by using the following: R, SAS, SPSS, Tableau, Python, Excel, and GeoDa, where the only technology that could be added is SQL.

- For FEEA, where work is mainly done with R, in order to align the academic skills formed with the professional skills needed on the labour market for the role of Data Analyst and Business Analyst, the integration of SAS, SQL, Python, and Tableau technologies would be necessary.

In order to visualise the relationships between variables and the possible existence of some groups of variables, for the database related to professional skills, following the application of the analysis in main components, the major characteristics that a specialist in data science must have been defined. The analysis started by standardising the data, using the Range method. This method was chosen because there were extreme values. In order to reduce the size of the data and to decide how many principal components it would be advisable to keep, the eigenvalue matrix presented in Table no. 2.

Table no. 2. Eigenvalues of the Uncorrected Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	17.69	15.95	0.68	0.68
2	1.74	0.79	0.06	0.74
3	0.94		0.0364	0.78

Source: Authors' own processing using SAS Studio on based on the "2022 Stack Overflow Annual Developer Survey"

The first value in the array explains 17.69 variables, so it would replace 17 variables with just one, thus having a significant reduction in space. The second value also explains almost two variables. Next, two main components were kept in the analysis. Also, the two principal components take up 74.7% of the information.

In order to identify the variables that determine each of the two preserved components (Table no. 3), the two main components were interpreted.

Table no. 3. Eigenvectors

The name of the variables	Principal Component 1	Principal Component 2
MainBranch	0.11	-0.08
Employment	0.13	-0.32
Remote Work	0.21	0.03
Coding Activities	0.22	0.07
EdLevel	0.17	0.07
Work Experience	0.12	0.55
Learn Code	0.14	-0.11
Learn Code Courses	0.17	0.08
Years Code	0.2	0.32
Organisation Size	0.2	-0.13

The name of the variables	Principal Component 1	Principal Component 2
Buy New Tool	0.22	0.08
Language Want To Work	0.2	0.04
Database Have Worked	0.21	0.18
Platform Have Worked	0.18	0.05
Knowledge_3	0.23	-0.11
Knowledge_4	0.23	-0.09
Knowledge_5	0.23	-0.07
Time Searching	0.21	-0.23
True/Fals_1	0.22	-0.16
True/Fals_2	0.22	-0.2
True/Fals_3	0.22	-0.15
Time Answering	0.21	-0.21
Years Code Pro	0.17	0.38
OfficeStackAsync	0.2	0.04
Age	0.19	0.25

Source: Authors' own processing using SAS Studio on based on the "2022 Stack Overflow Annual Developer Survey"

The first component is positively determined by the variables: Remote Work, Coding Activities, Organisation Size, Buy New Tool, Language Want To Work, Database used, Can I find updated information within my organisation to help me- I do my job, I know which system or which resource the user needs to find information and answers to the questions I have, I know which system or which resource the user needs to find information and answers to the questions I have, Time Searching, Time Answering, Are you involved in supporting new hires during their onboarding?, Do you use employer-provided learning resources?, Does your employer give you time to learn new skills?, variables related to the technical professional skills of respondents in the data science labour market.

The second component is positively determined by the variables: Years Code, Years Code Pro, Age, Work Experience, variables related to respondents' experience and negatively by the variables: Employment, Time Searching, Time Answering, Use learning resources provided by the employer?, which refer to their development. Thus, the secondary component will represent the respondents' experience and development in the data science labour market.

By applying principal component analysis, redundancy was removed and two principal components were retained, along with significant variables for each.

In the continuation of the research, the Cluster analysis was carried out for each component in order to identify the groups of respondents. Thus, following the Cluster Observation algorithm, they resulted, according to Table no. 4, the following:

Table no. 4. Cluster History

Cluster Number	First Principal Component R-Square	Second Principal Component R-Square
11070	1	1
11069	1	1
11068	1	1
...	...	...
5	0.58	0.70
4	0.56	0.67
3	0.54	0.61
2	0.52	0.53
1	0.00	0.00

Source: Authors' own processing using SAS Studio on based on the "2022 Stack Overflow Annual Developer Survey"

Table no. 4 shows that for the first component the greatest loss of homogeneity, where R-Square (0.54)>0.5, is if the variables were divided into three clusters, and for the second component, where R-Square (0.53)>0.5, is if we split the variables into two clusters.

To define the 3 clusters of respondents within the first component, the KM Clustering analysis was applied, resulting in the first cluster with 3779 respondents, the second 2789, and the last including 4503 respondents.

For the characterisation of each cluster, Table no. 5 in which only significant variables with R-Square > 0.5 were taken into account.

Table no. 5. Cluster Means

The name of the variables	Cluester no. 1	Cluester no. 2	Cluester no. 3
Knowledge_3	0.58	0.36	0.99
Knowledge_4	0.63	0.39	0.99
Knowledge_5	0.66	0.47	0.99
Time Searching	0.27	0.25	0.99
Time Answering	0.32	0.21	0.99
True/Fals_1	0.33	0.40	0.99
True/Fals_2	0.29	0.34	0.99
True/Fals_3	0.41	0.39	0.99

Source: Authors' own processing using SAS Studio on based on the "2022 Stack Overflow Annual Developer Survey"

Cluster 1 represents the most experienced data science professionals. They quickly find up-to-date information in the company, efficiently use systems to solve tasks, and have a short average time to find answers. However, they engage less in the onboarding process and use fewer learning resources because they have already developed professional skills.

Cluster 2 represents specialists with medium experience in data science. They have difficulty finding up-to-date information in employers' systems and do not always know where to look for answers, but they find solutions quickly. Companies' appreciation of development time and resources is average.

Cluster 3 represents early-career data scientists. They find up-to-date information and know where to look, but the process of finding the information they need takes the longest. They invest a lot of time in answering their work, especially during the onboarding period, and make heavy use of learning resources provided by employers to develop their professional skills.

For the second component, the KM Clustering analysis was applied again, resulting in the first cluster with 4555 respondents, and the second with 6516. To characterise each cluster, table no. 6, in which only significant variables with R-Square > 0.5 were taken into account.

Table no. 6. Cluster Means

The name of the variables	Cluester no. 1	Cluester no. 2
Time Searching	0.99	0.25
Time Answering	0.99	0.27
True/Fals_2	0.98	0.31

Source: Authors' own processing using SAS Studio on based on the "2022 Stack Overflow Annual Developer Survey"

Cluster 1 includes respondents who have a long time in the process of searching for an answer and who use the learning resources provided by employers the most. It is the group of data science specialists in the process of developing professional skills.

Cluster 2 includes the respondents for whom finding solutions in their work takes the least time and they use the company's learning resources in a small to medium percentage. Thus, this is the cluster with specialists who have completed the process of developing professional skills.

Also, for the two components, the KM Clustering method was applied, with the aim of observing which cluster of the two each model belongs to. This was done to construct a new binary variable representing the profile of the data scientist and to be further used as a dependent variable for predictive purposes.

Logistic regression is constructed in order to model and predict the binary dependent variable created above (the profile of the qualified/unqualified specialist in data science), according to the independent variables, the two components identified in the previous analyses: the first related to the technical professional skills of the respondents on the data science job market and the second, which represents the respondents' experience and development on the data science job market.

The results obtained in the prediction of the binary dependent variable (skilled/unskilled data scientist profile) were:

- For category 1 – specialists with experience in the labour market and developed professional technical skills, from the independent variable component 1 and for category 1 – specialists in the process of development, from the independent variable component 2, the logistic regression model predicted the value 0 – qualified profile.

- For the same category 1 - specialists with experience on the labor market and developed professional technical skills, if category 2 is set for the second component - specialists who have completed the development process, the logistic regression model predicted the value 1 - unqualified profile for a specialist from data science.

- For category 2 – specialists with an average level of knowledge, from the independent variable component 1 and for category 1 – specialists in the process of learning, from the independent variable component 2, the model framed the profile of the specialist as a qualified one.

- For category 3 – specialists at the beginning of their career, from the independent variable component 1 and for category 1 – professional skills in the process of development, from the independent variable component 2, the logistic regression model predicted the value 0 – qualified profile.

To validate the logistic regression model, the accuracy of the model was calculated for the analysed data set. The accuracy was 100% for the model included in the above analysis, which shows that the model correctly predicted the output values based on the input data.

Following the logistic regression model, the profile of the data science specialist was predicted according to the characteristics of the two independent variables, and to be a qualified specialist the following were identified: even though the level of experience is very important, to be a specialist qualified, continuous and active development is necessary, as a result the time and resources that employers provide are important. For an average level of experience, by appreciating career development opportunities and interest in learning, employers are looking for such a profile. For specialists at the beginning of their career, the development of professional skills occupies a large part of the current activity. Although the level of experience is not as high, openness to learning is recommended for this category of employees.

By applying the logistic regression model, the main conclusion was that regardless of the experience level of a data scientist, it is the continuous and active development of professional skills that is recommended and attracts the attention of employers. Thus, both the learning opportunities that companies offer to their employees, but also the academic skills acquired by specialists, have a critical role for their profile.

Conclusions

The need to know the profile of the data science specialist has become increasingly important, both for outlining the academic program and for identifying the professional skills needed in the labour market. Thus, this research represents a useful informational support in order to make decisions for adapting academic programs to current requirements and understanding the key skills to practice in one of the roles under the umbrella term data science.

Considering the fact that data science focuses on the knowledge of computer programs, algorithms, and the accumulation of experience in their use, the first part of the research had as its main objectives the identification of the professional skills needed on the labour market in this field, but also the outline of the specialists' profile. The level of experience, technical professional skills, and interest in learning outline the profile of the data scientist. Although the years of professional experience and the openness to the development of skills will place a specialist in the category of those qualified and sought after on the labour market, a significant role for their profile is played by the academic skills acquired by specialists, an idea reiterated by Irizarry (2020).

Completion of a bachelor's or master's program may be considered mandatory to practice data science. In this sense, the educational plans for master's programs in Romania were analysed. One of the main results was to identify the association of the academic skills acquired by graduating from these programs and the roles in the data science labour market. Thus, FMI outlines the academic skills required for the roles of Data Engineers or Database Administrator, FAC outlines the direction towards the role of Data Science Specialist, and FEAA and FCSIE teach and train the skills required for the role of Data Analyst and Business Analyst.

Technical professional skills are specific to each role. On the one hand, knowledge in SQL, Python or Shell Scripting is not required for a Database Administrator, Python, Java, or C++ are sought for DE. On the other hand, for a DA or BA the requirements included are R, SAS, Python, SQL, Tableau or Power BI and for the Data Science Specialist role skills in Python, AI, SQL or R are required. Equally significant, there are also communication skills, time management, teamwork, and analytical thinking required for all roles. The profile of the data scientist is defined by the level of experience, professional technical skills, and interest in learning.

In conclusion, academic skills and professional skills depend on each other. By aligning study programs with the labour market, data science is a field that will have qualified and trained specialists.

The research is limited by the lack of data for the year 2023, as this is a year in which AI has developed at a rapid pace, with data science participating in the development of chatbot-like search engines and helping to open up new opportunities for data scientists. In the future, it is necessary to study the feasibility of introducing chatbot-type AI in the field of data science in academic institutions, even more so in the context of increasing labour supply. As a result, there is a need to identify the main characteristics that a specialist in data science should possess, with the aim of introducing them into academic programs.

References

Alqahtani, T., Badreldin, H.A., Alrashed, M., Alshaya, AI., Alghamdi, S.S., Bin Saleh, K., Alowais, S.A., Alshaya, O.A., Rahman, I., Yami, M.S. and Albekairy, A.M., 2023. The Emergent Role of Artificial Intelligence, Natural Learning Processing, and Large Language Models in Higher Education and Research. *Research in Social and Administrative Pharmacy*, 19(8), pp. 1236-1242. https://doi.org/10.1016/j.sapharm.2023.05.016.

Berthold, M.R., 2019. What Does It Take to Be a Successful Specialist in Data Science? *Harvard Data Science Review*, 1(2). https://doi.org/10.1162/99608f92.e0eaabfc.

Bile Hassan, I., Ghanem, T., Jacobson, D., Jin, S., Johnson, K., Sulieman, D. and Wei, W., 2021. Data Science Curriculum Design: A Case Study. In*: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. Virtual Event USA: ACM. pp. 529-534. https://doi.org/10.1145/3408877.3432443.

Cao, L., 2019. Data Science: Profession and Education. *IEEE Intelligent Systems*, 34(5), pp. 35-44. https://doi.org/10.1109/MIS.2019.2936705.

Cao, L., 2016. Data Science and Analytics: A New Era. *International Journal of Data Science and Analytics*, 1, pp. 1-2. https://doi.org/10.1007/s41060-016-0006-1.

Chassignol, M., Khoroshavin, A., Klimova, A. and Bilyatdinova, A., 2018. Artificial Intelligence Trends In Education: A Narrative Overview. *Procedia Computer Science*, 136, pp. 16-24. https://doi.org/10.1016/j.procs.2018.08.233.

Chen, X., Xie, H., Zou, D. and Hwang, G.-J., 2020. Application and Theory Gaps During the Rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, article no. 100002. https://doi.org/10.1016/j.caeai.2020.100002.

De Veaux, R.D., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R.J., Sondjaja, M., Tiruviluamala, N., Uhlig, P.X., Washington, T.M., Wesley, C.L., White, D. and Ye, P., 2017. Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application*, 4(1), pp. 15-30. https://doi.org/10.1146/annurev-statistics-060116-053930.

Donoho, D., 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4), pp. 745-766, https://doi.org/10.1080/10618600.2017.1384734.

Dyk, D.V., Fuentes, M., Jordan, M.I., Newton, M., Ray, B.K., Lang, D.T. and Wickham, H., 2015. *ASA Statement on the Role of Statistics in Data Science*. [online] Available at: <http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science> [Accessed 15 August 2023].

Grossi, V., Giannotti, F., Pedreschi, D., Manghi, P., Pagano, P. and Assante, M., 2021. Data Science: A Game Changer for Science and Innovation. *International Journal of Data Science and Analytics*, 11, pp. 263-278. https://doi.org/10.1007/s41060-020-00240-2.

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D. and Ward, M.D., 2015. Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*, 69(4), pp. 343-353. https://doi.org/10.1080/00031305.2015.1077729.

Hick, S.C. and Irizarry, R.A., 2018. A Guide to Teaching Data Science. *The American Statistician*, 72(4), pp. 382-391. https://doi.org/10.1080/00031305.2017.1356747.

Irizarry, R.A., 2020. The Role of Academia in Data Science Education. *Harvard Data Science Review*, 2(1). https://doi.org/10.1162/99608f92.dd363929.

Jolliffe, I., 2011. Principal Component Analysis'. Berlin: Springer. pp. 1094-1096, https://doi.org/10.1007/978-3-642-04898-2_455.

Kordon, A.K., 2020. *Data Science Based on Artificial Intelligence.* Cham: Springer. https://doi.org/10.1007/978-3-030-36375-8_1.

Leal, W., Eustachio, J.H.P.P., Nita, D.A.C., Dinis, M.A.P., Salvia, A.L., Cotton, D.R.E., Frizzo, K., Trevisan, L.V. and Dibbern, T., 2023. Using Data Science for Sustainable Development in Higher Education. *Sustainable Development*, https://doi.org/10.1002/sd.2638.

Ley, C. and Bordas, S.P.A., 2018. What makes Data Science different? A Discussion Involving Statistics2.0 and Computational Sciences. *International Journal of Data Science and Analytics*, 6(3), pp. 167-175. https://doi.org/10.1007/s41060-017-0090-x.

Mainga, W., Murphy-Braynen, M.B., Moxey, R. and Quddus, S.A., 2022. Graduate Employability of Business Students. *Administrative Sciences*, 12(3), article no. 72. https://doi.org/10.3390/admsci12030072.

Pelau, C., Ene, I. and Pop, M.I., 2021. The Impact of Artificial Intelligence on Consumers' Identity and Human Skills. *Amfiteatru Economic*, 23(56), pp. 33-45, https://doi.org/10.24818/EA/2021/56/33.

Peng, C.-Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), pp. 8-9, https://doi.org/10.1080/00220670209598786.

Provost, E. and Fawcett, T., 2013. Data Science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), pp. 51-59, https://doi.org/10.1089/big.2013.1508.

Radovilsky, Z., Vishwanath, H., Anuja, A. and Uma, U., 2018. Skills Requirements of Business Data Analytics and Data Science Jobs: A Comparative Analysis. *Journal of Supply Chain and Operations Management*, 16(1), pp. 82-101.

Ranganathan, P., Pramesh, C.S. and Aggarwal, R., 2017. Common Pitfalls in Statistical Analysis: Logistic Regression. *Perspective in Clinique Research*, 8(3), pp. 148-151. https://doi.org/10.4103/picr.picr_87_17.

Smaldone, F., Ippolito, A., Lagger, J., and Pellicano, M., 2022. Employability Skills: Profiling Data Scientists in the Digital Labour Market. *European Management Journal*, 40(5), pp. 671-684. https://doi.org/10.1016/j.emj.2022.05.005.

Stack Overflow, 2022. *Stack Overflow Annual Developer Survey*. [online] Available at: <https://survey.stackoverflow.co/2022> [Accessed 8 December 2023].

Sterling, T., Anderson, M. and Brodowicz, M., 2017. *High Performance Computing: Modern Systems and Practices*. New York: Elsevier.

Tuba, L. and Süheyla, A., 2023. The impact of Artificial Intelligence in academia: Views of Turkish academics on ChatGPT, *Heliyon*, 9(9), article no. 19688. https://doi.org/10.1016/j.heliyon.2023.e19688.

Xu, A., Wu, Y., Meng, F., Xu, S. and Zhu, Y., 2022. Knowledge and Skill Sets for Big Data Professions: Analysis of Recruitment Information Based on The Latent Dirichlet Allocation Model. *Amfiteatru Economic*, 24(60), pp. 464-484. https://doi.org/10.24818/EA/2022/60/464.

Wing, J.M., 2019. The Data Life Cycle. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.e26845b4.