DOI: https://doi.org/10.15276/aait.07.2024.16
UDC 378.147:004.6:331.5

A cluster approach to matching the competences of data specialists with skills in demand on the labour market

Vitaliy M. Kobets[1)]
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@kse.org.ua. Scopus Author ID: 56006224700
Oleksii V. Gulin[1)]
ORCID: https://orcid.org/0009-0002-9043-5376; gulinleshka@gmail.com
Pavlo S. Nosov[2)]
ORCID: https://orcid.org/0000-0002-5067-9766; pason@ukr.net. Scopus Author ID: 57211927353
[1)] Kherson State University, 27, Universitetska Str. Kherson, 73003, Ukraine
[2)] Kherson State Maritime Academy, 20, Ushakov Ave. Kherson, 73000, Ukraine

ABSTRACT

This paper addresses the challenge of aligning the competences of data specialists with skills in demand on the labour market in the rapidly evolving field of data science. Using an open dataset of 3,744 IT job postings, the study applies K-means clustering to identify key skill groupings for data specialist positions. The optimal number of clusters is determined using the elbow method, resulting in four distinct clusters: Data Analyst & Engineer, Data Platform Engineer, Data Science & Engineering Specialist, and Cloud Data Engineer.
The research methodology employs unsupervised learning techniques, specifically K-means clustering, to analyze the distribution of skills across job postings. The clusters are visualized using t-distributed Stochastic Neighbor Embedding (t-SNE), providing insights into the relationships between different skill sets. The study reveals that job titles do not always unambiguously define the required skills, emphasizing the importance of focusing on specific skill sets rather than job titles alone. To bridge the gap between specific subject competences academic programs and industry requirements, the paper proposes a novel approach for comparing the proportion of skills in job clusters with the proportion of professional competencies in academic programs. This method is demonstrated using the Information Systems and Technologies Master's program at Kherson State University as a case study. The chi-square test is applied to confirm the statistical similarity between the skill structure of the Data Science & Engineering Specialist cluster and the competency structure of the academic program. The findings highlight the importance of continuous adaptation of profile of academic program to meet evolving industry needs. The proposed approach provides a data-driven framework for universities to align their programs with labor market demands, potentially improving graduate employability in the data science field. The study also underscores the need for personalized learning paths that can be tailored to individual career goals and skill gaps. Future research directions include the development of an artificial intelligence system to form individualized educational trajectories based on the skills required for specific job clusters. This could further enhance the alignment between education and industry needs, preparing students more effectively for the dynamic data science job market.

Keywords: Data specialist skills; labour market analysis; unsupervised learning; clustering; academic program; specific subject competences

***For citation*: Kobets V. M., Gulin O. V., Nosov P. S. "A cluster approach to matching the competences of data specialists with skills in demand on the labour market". *Applied Aspects of Information Technology*. 2024; Vol.7 No.3: 231–241. DOI: https://doi.org/10.15276/aait.07.2024.16**

INTRODUCTION

The dynamic nature of labour market demands, especially in engineering fields, requires continuous skill development for learners to remain employable [1, 2]. Open Educational Resources (OER) are a valuable source of learning materials that can be used for personalized learning [3].

AI and data mining techniques can be used to analyze job vacancies and extract skill requirements for specific occupations [4, 5], [6]. AI-driven recommender systems can match learners with relevant OER content based on extracted job skill requirements and individual learning goals. These systems face challenges such as cold start issues for new users or elements, and the need for accurate content quality assessment. Topic modelling and metadata analysis can be used to categorize and assess the quality of learning resources [7].

The **purpose** of the paper is to develop an approach to match the specific subject competences of data specialists for academic programs with the skills in demand on the labour market, in order to justify the choice of such educational component of an academic program for an applicant, which would allow to prepare the applicant during studies for obtaining the desired data specialist position on the labour market.

The paper is structured as follows: Section 2 reviews the literature on the requirements of employers in the IT industry; Section 3 describes the research methodology using unsupervised learning in the form of clustering; Section 4 defines the academic program corresponding to the cluster of vacancies; the last section concludes.

RELATED WORKS

In order to create an academic program that meets the needs of the labor market, it is important to identify the IT market skills required by industry for computing students in higher education, using the following steps:

1) analyze job postings and advertisements from online job portals to extract skill requirements for IT positions [4, 6];

2) use text mining and natural language processing techniques to identify key skills and competencies mentioned in job postings [9];

3) use AI and machine learning algorithms to map skills to specific job roles and identify emerging trends in skill requirements [9];

4) conduct surveys and interviews with industry practitioners to understand skill gaps and requirements [10, 11];

5) analyze CVs and profiles of successful professionals in the industry to identify common skills [12].

Next, it is important to provide academic educators and learners with the skills needed in the IT marketplace:

1) develop data-driven frameworks that map industry skill requirements to academic curricula [13];

2) create real-time tools that provide curriculum designers with updates on changing skills requirements [9];

3) implement work-based learning programs to expose students to industry-relevant skills [14];

4) revise and update existing courses based on identified skills gaps [11, 15];

5) offer electives, online courses and mini-courses focused on in-demand skills [15];

6) collaborate with industry partners to incorporate practical, industry-relevant projects into coursework [10];

7) provide personalized learning pathways for students based on their individual skill gaps and career goals [15].

The adoption of Industry 4.0 technologies in the manufacturing sector (that has a constant need for data analysis and interpretation) has led to significant changes in the skills requirements of the workforce. Several studies have examined the growing skills gap in the manufacturing sector:

The skills gap is particularly pronounced for middle-skill jobs, which typically require less than a bachelor's degree [16]. Employers consistently report difficulties in finding workers with the necessary skills and qualifications, indicating a mismatch between workers' skills and employers' needs. In Europe, around 25% of workers have no or low digital skills, with disadvantaged regions more likely to be negatively affected by automation and AI [17]. Workers will need to develop skills in a range of new digital technologies to fit into Industry 4.0 environments [18]. Several studies in the US highlight the importance of skills that require industry certification, but not necessarily a bachelor's degree [19]. Data science skills are increasingly important in different firms in IT, banks, mobile companies, government organization etc. [20]. There is a growing demand for interdisciplinary talent that has both core manufacturing skills and data science skills. Current manufacturing education and training programs do not fully address the data science skills gap [21].

Existing studies reveal several limitations of existing research [22]:

1) many studies rely on surveys and interviews, which may not fully reflect the comprehensive state of the labour market;

2) most research focuses primarily on employer demand data, neglecting the skills offered by the current and future workforce;

3) there is a lack of comprehensive studies covering different aspects of Industry 4.0, such as cyber manufacturing, cloud computing and sensor engineering;

4) few studies have conducted a data science-specific skills gap analysis of the current manufacturing job market.

The review highlights the need for a more comprehensive, data-driven analysis of both the demand and supply sides of the manufacturing labour market, with a particular focus on data science and Industry 4.0-related skills.

Data science skills in high demand in manufacturing include [23]:

- programming languages (SQL, Python, Java);
- big data processing tools (Apache Hadoop, Apache Spark);
- cloud platforms (AWS and Microsoft Azure);
- machine learning and AI.

Key domain knowledge areas identified as important include artificial intelligence, statistics and algorithms, big data analytics, machine learning,

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

data engineering in different subject domains [24]. Research suggests that there is often a mismatch between the skills/knowledge of the current workforce and what employers require. Education programs and workforce training initiatives need to be updated to address these gaps. There is also an emphasis on the need for continuous learning and upskilling of the workforce to keep pace with technological change [25]. Overall, the literature highlights the critical importance of data science skills and knowledge for the future manufacturing workforce, while also identifying significant gaps that need to be addressed through education, training and recruitment efforts.

Rapid advances in artificial intelligence are shaping key ideas for a new division of labour between humans and machines and automation, including that many blue colors positions are at high risk of automation in the coming decades [26]. Many middle-class and white-collar jobs are at risk. Tasks that follow explicit rules can be automated, while non-routine tasks that are not sufficiently understood to be specified in computer code are more resistant. Recent advances in AI have enabled the automation of some non-routine cognitive tasks, with the main obstacle being insufficient data for pattern recognition. Creativity and innovation are seen as uniquely human capacities that are most resistant to automation, but some argue that it is only a matter of time before AI becomes creative.

The future may require a shift in education from skills to cultivating uniquely human capacities such as judgement, creativity and innovation. Education should focus on sensing, understanding, managing and creating change, as well as collaborating effectively with AI systems. The implications for education include:

1. Cultivating creativity, innovation, judgement and leadership skills that are resilient to automation.
2. Shifting from goal-oriented to purpose-oriented curricula focused on future readiness.
3. Teaching understanding of coding, AI capabilities/limitations and human-AI collaboration.
4. Focus on play and humor in early education to develop creativity.
5. Develop skills to lead human-machine partnerships.

RESEARCH METHODOLOGY

K-means clustering is a supervised learning algorithm designed to cluster data based on its similarity. Unsupervised learning means that there is no specific outcome (number of clusters) to be predicted, and the algorithm tries to find patterns in the data. In the K-means method, we need to specify the number of clusters into which we consider it appropriate to divide the data. The algorithm randomly assigns each observation to a cluster and finds the center of each cluster. The algorithm then iterates through two steps:

1) reassigns data points to the cluster whose center is closest;

2) calculates the new centroid of each cluster.

These two steps are repeated until the variation within the cluster can no longer be reduced beyond a certain threshold. The intra-cluster variation is calculated as the sum of the Euclidean distances between the data points and their respective cluster centroids.

We will use an open data set of professions and related skills for Data Specialists [27], consisting of 3744 IT professions ('data engineer', 'data analyst', 'data scientist', 'software engineer'). For each of these professions, requirements are formulated in the form of skills (SQL, Python, Scala/Spark, Data Engineering, AWS, Azure, ETL, Data Analysis, Snowflake, Kafka, Java, Data Modeling, Airflow, Data Warehousing, Machine Learning, DevOps, Kubernetes, Docker, Databricks, Git). For each profession, the requirements profile is defined in the form of binary values (1 - if the skill is included in the requirements of this profession; 0 - otherwise). The first values of this correspondence between professions and skills are demonstrated in Fig. 1 using the Python function head (clustered_job_postings).

job_title	SQL	Python	Scala/ Spark
Data Engineer 2	1	0	0
Staff Data Engineer	1	1	0
Senior Data Engineer, Public Company	1	1	0
Senior Data Engineer, Public Company	1	1	0
Senior Staff AI Data Engineer	1	1	0

Data Engineering	AWS	Azure	ETL	Data Analysis	Snow-flake
0	0	1	0	0	0
0	0	0	1	0	1
0	0	0	1	0	1
1	0	0	1	0	1
1	1	1	1	0	1

Fig 1. **Correspondence between professions and skills**

***Source*: compiled by the authors**

This dataset shows which skills are available and which are not for each profession.

Exploratory Data Analysis (EDA)

The data consists of job advertisements for Data Specialists with different skill requirements. In order to determine which list of skills of higher education graduates will better facilitate employment for a particular position, we will identify the main types of positions for Data Specialists by clustering these positions. For K-means clustering, we will use the columns with skills to form clusters using Python. An example of the python code for K-means clustering and finding the cluster centers for Data Specialists positions is shown below (Fig. 2) [28].

```
kmeans = KMeans(n_clusters=6, random_state=42)
kmeans.fit(X_scaled)
y_kmeans = kmeans.predict(X_scaled)
```

Fig. 2. **Distribution of requirements for the positions of data specialists**
***Source*: prepared by the authors [28]**

After clustering, we visualize the results of the K-means clustering in Fig. 3 using the t-SNE method. The data points are colored according to the assigned cluster, and the red dots represent the cluster centers. That is, four clusters were identified within the Data Specialists profession dataset (Fig. 4).

Fig. 3. **Clustering of Data Specialist positions**
***Source*: compiled by the authors**

Let us visualize data clustering by skills using t-SNE method (Fig. 5) [28]. The visualization results are shown in Fig. 2 using t-distributed Stochastic Neighbor Embedding (t-SNE method) as a nonlinear dimensionality reduction method often used to visualize multivariate data in two or three dimensions. It is particularly useful for identifying clusters in data. Thus, the clustering results yielded a distribution of clusters.

Fig. 4. **Visualization of Data Specialist position clusters using t-SNE**
***Source*: compiled by the authors**

```
kmeans = KMeans(n_clusters=4, random_state=20)  # You can change the number
kmeans.fit(X)

# Adding the clustering results to the original DataFrame
df['cluster'] = kmeans.labels_

# Using t-SNE for dimensionality reduction
tsne = TSNE(n_components=2, perplexity=30, random_state=20)
X_tsne = tsne.fit_transform(X)

# Adding t-SNE components to the DataFrame
df['tsne-2d-one'] = X_tsne[:, 0]
df['tsne-2d-two'] = X_tsne[:, 1]
```

Fig 5. **Visualization of Data Specialist position clusters**
***Source*: prepared by the authors [28]**

Distribution of clusters:

• the yellow cluster (cluster 0) is located on the left side of the graph. It contains several dense groups of points;

• the green cluster (cluster 1) is located on the right side of the graph. It contains the most points and is more scattered than the other clusters;

• the blue cluster (cluster 2) is concentrated at the top of the graph. The points are quite compact;

• the purple cluster (cluster 3) is located at the bottom of the graph. The dots form a dense group.

Density of points:

• some clusters have areas of high point density, indicating greater skill similarity between our data in these areas;

• other clusters, such as the green cluster, have more widely distributed points, which may indicate a greater diversity of skills in this cluster.

Features of clusters:

• clusters can correspond to different skill groups or specializations in our data;

• the yellow and blue clusters may represent groups with very specific skills, given their density and compactness;

• the purple cluster can also represent a highly specialized group, but a slightly larger one;

• the green cluster can cover a broader range of skills or a more general specialization.

Finding the optimal number of clusters

To determine the optimal number of clusters, we apply the elbow method. The graph of the elbow method for determining the optimal number of clusters shows that the inertia (the difference between cluster elements) decreases sharply from 2 to 4 clusters, and then decreases less significantly (Fig. 6). This indicates that the optimal number of clusters is in the range of 3 to 4.

Visualization of the elbow method is in Fig. 6.

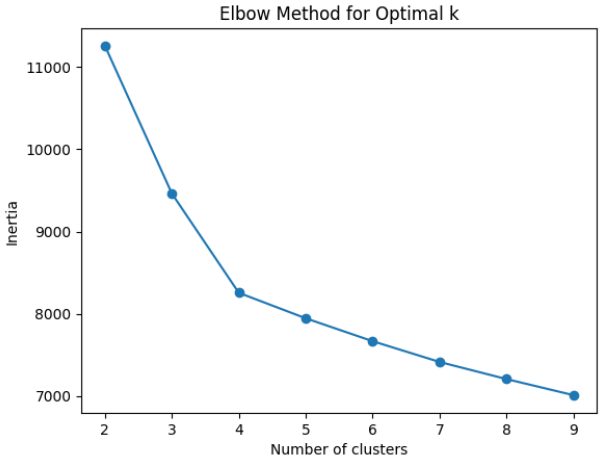

Fig. 6. **Determining the optimal number of clusters using the elbow method**

***Source*: compiled by the authors**

The optimal number of clusters using the elbow method is usually chosen where there is an elbow on the graph, i.e. where the decrease in inertia becomes less significant. In Fig. 6, this is the case for four clusters.

Three clusters: Identifies separate groups of Data Engineers and other positions, including Data Analysts and Data Scientists.

Four clusters: It shows an even more detailed division between professions, in particular, separates Data Scientists into different clusters.

Recommendations

1. We can use 4 clusters for optimal clustering. This will provide a more detailed view of the distribution of professions and specializations.

2. If a more detailed segmentation is required, 5 clusters can be considered, but keep in mind that the significance of the decrease in inertia decreases after 4 clusters.

According to Elbow Method for Optimal number of clusters the most significant reduction in the coefficient of inertia, which shows the difference between clusters, occurred when the number of clusters was 4 (Fig. 4), further increase in the number of clusters leads to a slight improvement in the quality of clustering, which led to the choice of 4 clusters. Based on the results of the distribution of positions between the clusters, we obtain the following Table 1.

Table 1. **Distribution of positions (professions) between clusters**

Clusters	Data engineer	Data analyst	Data scientist	Software engineer
0	763	42	25	46
1	527	503	22	34
2	894	350	115	56
3	361	0	1	5

***Source*: compiled by the authors**

This table shows that job titles do not always unambiguously define the required skills and competencies of job applicants. And vice versa, job titles with different titles may be similar in terms of the skills they require. This means that a job applicant should first pay attention to the set of skills that he or she must possess in order to be suitable for the desired position.

The distribution of Data Specialists' skills within each cluster is shown in Fig. 7.

Fig. 7. **Distribution of Data Specialists' skills across occupational clusters**

***Source*: compiled by the authors**

Based on the analysis of the distribution of skills, the following names for the respective clusters were derived (Table 2).

Based on the obtained clusters of occupations for Data Specialist, a potential applicant for a relevant vacancy needs to determine the appropriate academic program that would allow him/her to obtain the desired position, taking into account the skills in demand in the labour market within the cluster of relevant positions.

Table 2. **Description of clusters**

Clusters name	Cluster description
Data Analyst & Engineer	SQL skills (21.9 %) and Data Analysis (19.6 %) dominate, emphasizing the focus on data analytics and database work. Data Engineering (9.8%) and Data Modeling (7.7 %) indicate the importance of engineering tasks related to data management. ETL (7.2 %) demonstrates work with data extraction, transformation, and loading processes, which is an important part of both analyst and data engineer responsibilities. The name reflects the integration of analytical and engineering functions in this cluster
Data Platform Engineer	The importance of containerization and orchestration skills: Docker (9.5 %) and Kubernetes (9.4 %). Programming and databases: Python (9.4%) and SQL (9.1 %) are key skills that emphasize the importance of programming and database work. The presence of Kafka (8.8%), Snowflake (8.4 %), Airflow (8.4 %), and ETL (7.2 %) puts the focus on data processing and movement.
Data Science & Engineering Specialist	Python (24.3 %) and SQL (18.4 %) skills dominate, highlighting the key role of programming and data processing. Data Engineering (7.2 %), ETL (5%), and Data Modeling (4.8%) skills are related to data engineering and management. Machine Learning (5.8 %) focuses on data analysis and the application of machine learning models. Data Analysis (6.6%) and Data Warehousing (4.4 %) define analytical skills and data warehousing work that complement the technical aspect.
Cloud Data Engineer	AWS (15.6 %) and Azure (6.8%) together account for over 22% of requirements, indicating a strong focus on cloud platforms. Programming and data management: Python (14.0 %) and SQL (11.3 %) are the second and third most important skills, highlighting the importance of programming and database skills. A variety of technologies for working with data: Snowflake (4.8 %), Kafka (3.7 %), Airflow (2.7 %) indicate the need to work with modern data processing and analysis tools in a cloud environment. The name reflects a combination of cloud technology and data engineering skills.

***Source*: compiled by the authors**

DETERMINATION OF THE ACADEMIC PROGRAM CORRESPONDING TO THE DATA SPECIALIST POSITION CLUSTER

The approach to choosing a position based on a set of skills allows you to build an individual learning path that a job applicant can pursue in higher education institutions, or independently as an informal education on educational platforms (COURSERA. Udemy), whose certificates are recognized by employers, or as part of an internship in the company.

The main idea in choosing an academic program that would allow you to get the desired position in the labour market is to compare the share of skills presented as hard skills requirements in the vacancy cluster with the share of professional competences that are formed within a particular academic program. Next, it is necessary to confirm the reliability of the correspondence of the hard skills of the vacancy cluster to the professional competences of the academic program using the statistical criterion χ^2.

Let us consider the academic program Information Systems and Technologies of the Master's level at Kherson State University (Ukraine) (https://ksu24.kspu.edu/s/m69no) in Table 3.

Table 3. **Educational components of the Master's degree program in IST (OC – obligatory/mandatory component/course, EC – elective component/course)**

No.	Disciplines of the academic program	ECTS
OC 3	Formal methods of software engineering	3
OC 4	Management of financial instrument development technology	3
OC 5	Digital currencies and blockchain technologies	3
OC 6	Time series forecasting models for business analytics	6,5
OC 7	Modelling and design of information systems	3
EC	EC 7, EC 8, EC 9, EC 10	3*4
OC 4	Management of financial instrument development technology	3

***Source*: compiled by the authors**

Let us compare the proportion of skills of cluster 2 Data Science & Engineering Specialist with the professional competences corresponding to this study program. To determine the proportion of professional competences, we will use the number of credits for the educational components. The comparison results in Fig. 8.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

№	Cluster 2	Number	Skill share, SS	ECTS	Competences share, CS	Average, Av	χ^2_{fact}	Courses
1	Python	1409	24.3%	5	13.5%	18.9	1.54	ОС 3, ЕС 7.1, ЕС 8.4
2	SQL	1067	18.4%	4	10.8%	14.6	0.99	ОС 4, ЕС 8.3
3	Data Engineering	416	7.2%	6	16.2%	11.7	1.74	ОС 3, ЕС 7.1, ЕС 8.1, ЕС 9.2
4	Data Analysis	384	6.6%	5.5	14.9%	10.7	1.59	ОС 6, ЕС 7.4, ЕС 9.1.
5	Machine Learning	335	5.8%	5.5	14.9%	10.3	1.99	ОС 6, ЕС 7.4, ЕС 10.1.
6	ETL	290	5.0%	4.5	12.2%	8.6	1.49	ОС 6, ЕС 9.1
7	Data Modelling	278	4.8%	4.5	12.2%	8.5	1.60	ОС 7, ОС 5
8	Java	272	4.7%	0	0.0%	2.4	2.35	
9	Data Warehousing	254	4.4%	1	2.7%	3.6	0.20	ЕС 8.1
10	Databricks	208	3.6%	0	0.0%	1.8	1.80	
11	Snowflake	193	3.3%	0	0.0%	1.7	1.65	
12	Azure	137	2.4%	0	0.0%	1.2	1.20	
13	Airflow	115	2.0%	0	0.0%	1.0	1.00	
14	Git	110	1.9%	1	2.7%	2.3	0.07	ОС 9
15	Kafka	112	1.9%	0	0.0%	1.0	0.95	
16	DevOps	104	1.8%	0	0.0%	0.9	0.90	
17	Kubernetes	61	1.1%	0	0.0%	0.6	0.55	
18	Docker	57	1.0%	0	0.0%	0.5	0.50	

Fig. 8. **Calculation of shares of skills and competences**

***Source*: compiled by the authors**

Let us compare the proportion of skills of cluster 2 Data Science & Engineering Specialist with the professional competences corresponding to this study program. To determine the proportion of professional competences, we will use the number of credits for the educational components. The comparison results in Fig. 8.

To test the hypothesis that the actual and expected distributions are not randomly similar, we apply the χ^2 test. Steps to calculate χ^2 criterion.

1. Calculation of the actual χ^2 for each skill category using the formula:

$$\chi^2 = \frac{(SS-Av)^2}{Av} = 22.1, \quad (1)$$

where SS is skill share, CS is competences share of specific subjects, $Av = \frac{SS+CS}{2}$ is average (expected) value between skill and competences share for each category in Fig. 8.

2. Compare the actual value with the critical value: $\chi^2_{cr} = CHISQ.INV.RT(5\%; 18-1) = 27.587$. Since $\chi^2 = 22.1 < 27.587 = \chi^2_{cr}$ we cannot reject the null hypothesis. This means that the difference between the actual and expected distribution of skill shares is not significant, and the distance between the shares can be explained by random fluctuations, e.g., these shares are same.

This means that the distribution of specific subject competencies of the Information systems and technology academic program meets the requirements for Data Science & Engineering Specialist cluster and this academic program can be chosen to develop the appropriate skills to obtain the desired position in the labor market.

CONCLUSIONS

The optimal number of clusters was determined based on the open data on data specialist positions, the clustering method and the elbow method: Data Analyst & Engineer, Data Platform Engineer, Data Science & Engineering Specialist, Cloud Data Engineer. Each of these clusters has a unique combination of 17 to 20 skills for data specialists. In general, clustering allows us to identify the most sought-after groups of data specialists with the most in-demand skills. Based on a comparison of the shares of skills in the profile of specialists in each cluster and the shares of professional competencies in academic programs, an approach to identifying similarities between the requirements of the labor market and the academic program for training data specialists is proposed. An example of applying this approach to the Data Science & Engineering Specialist cluster and the master's degree program Information Systems and Technologies at Kherson State University is presented. Using a statistical criterion, the similarity between the structure of the cluster by skills and the structure of the academic program by professional competences is confirmed. The application of the proposed approach will be useful for higher education institutions when reviewing educational programmes, as the design of the programme will take into account not only the names of labour market positions that graduates can work in, but also the professional competencies in clusters of positions, training in which will allow the graduate to have a better understanding of future requirements from potential employers.

In the future, it is planned to study the formation of an individual educational trajectory for applicants based on the skills they need to develop for employment or advanced training for the labor market, using an artificial intelligence system.

REFERENCES

1. Tavakoli, M., Kismihok, G. & Mol, S. T. "Labour market information driven, personalised, OER recommendation system for lifelong learners". *Proceedings of the 12th International Conference On Computer Supported Education*. 2020. https://www.scopus.com/record/display.uri?eid=2-s2.0-85091752092&origin=resultslist DOI: https://doi.org/10.48550/arXiv.2005.07465.

2. Wang, F., Jiang, Z., Li, X. & Li, G. "Cognitive factors of the transfer of empirical engineering knowledge: A behavioural and fNIRS study". *Advanced Engineering Informatics*. 2021; 47: 101207, https://www.scopus.com/record/display.uri?eid=2-s2.0-85097584314&origin=resultslist.
DOI: https://doi.org/10.1016/j.aei.2020.101207.

3. Tavakoli, M., Faraji, A., Mol, S. T. & Kismihók, G. "OER recommendations to support career development". *2020 IEEE Frontiers in Education Conference*. Uppsala, Sweden. 2020. p.1–5. https://www.scopus.com/record/display.uri?eid=2-s2.0-85098561976&origin=resultslist.
DOI: https://doi.org/10.1109/FIE44824.2020.9274175.

4. Colombo, E., Mercorio, F. & Mezzanzanica, M. "Applying machine learning tools on web vacancies for labour market and skill analysis. In terminator or the jetsons?". *The Economics and Policy Implications of Artificial Intelligence*. 2018.

5. Li, X., Jiang, Z., Guan, Y., Li, G. & Wang, F. "Fostering the transfer of empirical engineering knowledge under technological paradigm shift: An experimental study in conceptual design". *Advanced Engineering Informatics*. 2019; 41: 100927, https://www.scopus.com/record/display.uri?eid=2-s2.0-85065895120&origin=resultslist. DOI: https://doi.org/10.1016/j.aei.2019.100927.

6. Ermolayev, V., Suarez-Figueroa, M. C. & Molchanovskyi, O. "Architecting data science education". *Proceedings of 14th International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*. 2018; 2104: 734–746. – Available from: https://ceur-ws.org/Vol-2104/paper_266.pdf. – [Accessed: February 2024], https://www.scopus.com/record/display.uri?eid=2-s2.0-85048355807&origin=resultslist.

7. Molavi, M., Tavakoli, M. & Kismihók, G. "Extracting topics from open educational resources". *Lecture Notes in Computer Science*. 2020; 12315: 455–460, https://www.scopus.com/record/display.uri?eid=2-s2.0-85091181893&origin=resultslist. DOI: https://doi.org/10.1007/978-3-030-57717-9_44.

8. Kobets, V. & Osypova, N. V. "Identification of factors for providing the higher education quality assurance for students". *International Journal for Quality Research*. 2023; 17 (1): 195–208, https://www.scopus.com/record/display.uri?eid=2-s2.0-85150269094&origin=resultslist.
DOI: https://doi.org/10.24874/ijqr17.01-12.

9. Ketamo, H., Moisio, A., Passi-Rauste, A. & Alamäki, A. "Mapping the future curriculum: Adopting artificial intelligence and analytics in forecasting competence needs". *Proceedings of the 10th European Conference on Intangibles and Intellectual Capital*. Chieti-Pescara. Italy. 2019. p. 144–153, https://www.scopus.com/record/display.uri?eid=2-s2.0-85070005046&origin=resultslist

10. Metrôlho, J.C., Ribeiro, F.R. & Batista, R. "Prepare students for software industry a case study on an agile full stack project". *The Seventeenth International Conference on Software Engineering Advances*. 2022. p.75–80. http://hdl.handle.net/10400.11/8167.

11. Kobets, V., Yatsenko, V. & Buiak, L. "Bridging business analysts competence gaps: labor market needs versus education standards". *The Information and Communication Technologies in Education, Research, and Industrial Applications*. 2021; 1308: 22–45, https://www.scopus.com/record/display.uri?eid=2-s2.0-85111787730&origin=resultslist. DOI: https://doi.org/10.1007/978-3-030-77592-6_2.

12. Kara, A., Daniş, F. S., Orman, G. K., Turhan, S. N. & Özlü, A. "Job recommendation based on extracted skill embeddings". *Lecture notes in networks and systems*. 2022; 544: 497–507, https://www.scopus.com/record/display.uri?eid=2-s2.0-85138240004&origin=resultslist.
DOI: https://doi.org/10.1007/978-3-031-16075-2_35.

13. Ramakrishnan, M., Gregor, S., Shrestha, A. & Soar, J. "Achieving Industry-aligned Education through Digital-Commons: A Case Study". *Journal of Computer Information Systems.* 2022; 63 (4): 950–964, https://www.scopus.com/record/display.uri?eid=2-s2.0-85137825210&origin=resultslist.
DOI: https://doi.org/10.1080/08874417.2022.2115955.

14. Kobets, V., Tsiuriuta, N., Lytvynenko, V. & Mykhaylova, V. "Web-service management system for job search using competence-based approach". *Proceedings of 16th International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications.* 2020; 2732: 290–302, https://www.scopus.com/record/display.uri?eid=2-s2.0-85096124292&origin=resultslist. – Available from: https://ceur-ws.org/Vol-2732/20200290.pdf. – [Accessed: March 2024],

15. Wu, W. "Investigating internship experiences of data science students for curriculum enhancement". *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education* 2022; 1: 505–511, https://www.scopus.com/record/display.uri?eid=2-s2.0-85134433374&origin=resultslist. DOI: https://doi.org/10.1145/3502718.3524741.

16. Kravtsov, H. & Kobets, V. "Evolutionary revision model for improvement of computer science curriculum". *Proceedings of the Information and Communication Technologies in Education, Research, and Industrial Applications.* 2019; 1007: 127–147, https://www.scopus.com/record/display.uri?eid=2-s2.0-85063463525&origin=resultslist. DOI: https://doi.org/10.1007/978-3-030-13929-2_7.

17. Berger, T. & Frey, C. B. "Digitalization, jobs, and convergence in Europe: Strategies for closing the skills gap". *Oxford Martin School*. 2016. – Available from: https://oms-www.files.svdcdn.com/production/downloads/reports/SCALE_Digitalisation_Final.pdf. – [Accessed: Sep 2024].

18. Moldovan, L. "State-of-the-art analysis on the knowledge and skills gaps on the topic of Industry 4.0 and the requirements for work-based learning". *Procedia Manufacturing.* 2019; 32: 294–301, https://www.scopus.com/record/display.uri?eid=2-s2.0-85065666632&origin=resultslist.
DOI: https://doi.org/10.1016/j.promfg.2019.02.217.

19. Short, M. N. & Keller-Bell, Y. "Essential Skills for the 21st Century Workforce". *Research Anthology on Developing Critical Thinking Skills in Students*. 2021. p. 97–110. DOI: https://doi.org/10.4018/978-1-7998-3022-1.ch006.

20. Kobets, V. Yatsenko, V. & Buiak, L. "Identifying the gaps in the preparing of a business analyst between the requirements of the labor market and the standards of study programs: Case of Ukraine". *Proceedings of 16th International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications.* 2020; 2732: 499–514, https://www.scopus.com/record/display.uri?eid=2-s2.0-85096149927&origin=resultslist. – Available from: https://ceur-ws.org/Vol-2732/20200499.pdf. – [Accessed: March 2024].

21. Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P. & Sanghvi, S. "Jobs lost, jobs gained: Workforce transitions in a time of automation". *McKinsey Global Institute*. 2017. – Available from: https://www.mckinsey.com/~/media/BAB489A30B724BECB5DEDC41E9BB9FAC.ashx. – [Accessed: March 2024].

22. Agarwal, A. & Ojha, R. "Prioritising implications of Industry-4.0 on the sustainable development goals: A perspective from the analytical hierarchy process in manufacturing operations". *Journal of Cleaner Production.* 2024; 444: 141189, https://www.scopus.com/record/display.uri?eid=2-s2.0-85185199842&origin=resultslist. DOI: https://doi.org/10.1016/j.jclepro.2024.141189.

23. Li, G., Yuan, C., Kamarthi, S., Moghaddam, M. & Jin, X. "Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis". *J. of Manufact. Systems.* 2021; 60: 692–706, https://www.scopus.com/record/display.uri?eid=2-s2.0-85111559556&origin=resultslist.
DOI: https://doi.org/10.1016/j.jmsy.2021.07.007.

24. Novak, O. & Kobets, V. "Artificial intelligence impact on food security of states in the world". *Proceedings of International conference on Information and Communication Technologies in Education, Research, and Industrial Applications.* 2023; 1980: 240–251, https://www.scopus.com/record/display.uri?eid=2-s2.0-85180631318&origin=resultslist.

DOI: https://doi.org/10.1007/978-3-031-48325-7_18.

25. Vista, A. "Data-Driven identification of skills for the future: 21st-Century Skills for the 21st-Century Workforce". *SAGE Open.* 2020; 10 (2): 215824402091590. DOI: https://doi.org/10.1177/ 2158244020915904.

26. Kobets, V., Tsiuriuta, N., Lytvynenko, V., Novikov, M. & Chizhik, S. "Recruitment web-service management system using competence-based approach for manufacturing enterprises". *Proceedings of Advances in Design, Simulation and Manufacturing II.* 2020. p. 138–148, https://www.scopus.com/record/display.uri?eid=2-s2.0-85067038865&origin=resultslist.
DOI: https://doi.org/10.1007/978-3-030-22365-6_14.

27. Asaniczka. "Linkedin Data engineer job postings [Data set]." *Kaggle*. 2023. DOI: https://doi.org/10.34740/KAGGLE/DSV/7292935.

28. "Clustering of Data Specialists positions". – Available from: https://github.com/AlexSorrow/clustering/blob/main. – [Accessed: March 2024].

Conflicts of Interest: The authors declare that there is no conflict of interest

Received 22.07.2924
Received after revision 04.09.2024
Accepted 18.09.2024

DOI: https://doi.org/10.15276/aait.07.2024.16
УДК 378.147:004.6:331.5

Кластерний підхід увідповіднення компетентностей фахівців з даних до затребуваних навичок на ринку праці

Кобець Віталій Миколайович[1)]
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@kse.org.ua. Scopus Author ID: 56006224700
Гулін Олексій Вадимович[1)]
ORCID: https://orcid.org/0009-0002-9043-5376; gulinleshka@gmail.com
Носов Павло Сергійович[2)]
ORCID: https://orcid.org/0000-0002-5067-9766; pason@ukr.net. Scopus Author ID: 57211927353
[1)] Херсонський державний університет, вул. Університетська, 27. Херсон, 73003, Україна
[2)] Херсонська державна морська академія, проспект Ушакова, 20. Херсон, 73000, Україна

АНОТАЦІЯ

Стаття присвячена проблемі увідповіднення компетентностей фахівців з даних із навичками, затребуваними на ринку праці в галузі науки про дані, що стрімко розвивається. Використовуючи відкритий набір даних з 3744 вакансій у сфері IT, у дослідженні застосовано кластеризацію за методом K-середніх для визначення ключових груп навичок для позицій спеціалістів з обробки даних. Оптимальна кількість кластерів визначається за допомогою методу ліктя, в результаті чого отримано чотири окремі кластери: аналітик та інженер даних, інженер платформи даних, спеціаліст з науки про дані та інженерії, а також інженер хмарних даних.

Методологія дослідження використовує методи навчання без учителя, зокрема кластеризацію за методом K-середніх, для аналізу розподілу навичок у вакансіях. Кластери візуалізуються за допомогою t-розподіленого включення стохастичних сусідів (t-SNE), що дає змогу зрозуміти взаємозв'язок між різними наборами навичок. Дослідження показує, що назви посад не завжди однозначно визначають необхідні навички, що підкреслює важливість фокусування уваги на конкретних наборах навичок, а не лише на назвах посад. Для подолання розриву між фаховими компетентностями освітніх програм і вимогами галузі в роботі запропоновано новий підхід до порівняння частки навичок у кластерах робочих місць із часткою фахових компетентностей в освітніх програмах. Цей метод продемонстровано на прикладі магістерської програми «Інформаційні системи та технології» Херсонського державного університету. Тест ксі-квадрат застосовано для підтвердження статистичної подібності між структурою навичок кластеру «Data Science & Engineering Specialist» та структурою компетентностей освітньої програми. Отримані результати підкреслюють важливість постійної адаптації профайлу освітніх програм до потреб галузі, що постійно змінюються. Запропонований підхід надає закладам вищої освіти основу, що ґрунтується на даних, для узгодження їхніх програм з потребами ринку праці, покращуючи можливості працевлаштування випускників у галузі науки про дані. Дослідження також підкреслює необхідність персоналізованих навчальних траєкторій, які можна адаптувати до індивідуальних кар'єрних цілей та прогалин у навичках. Майбутні напрямки досліджень включають розробку системи штучного інтелекту для

формування індивідуальних освітніх траєкторій на основі навичок, необхідних для конкретних кластерів робочих місць. Це може сприяти подальшому узгодженню між освітою та потребами галузі, що дозволить ефективніше готувати здобувачів до динамічного ринку праці в галузі науки про дані.

Ключові слова: навички спеціаліста з даних; аналіз ринку праці; метод навчання без вчителя; кластеризація; освітня програма, фахові компетентності

ABOUT THE AUTHORS

Vitaliy M. Kobets - Doctor of Economic Science, Professor, Department of Computer Science and Software Engineering. Kherson State University. 27, Universitetska Str. Kherson, 73003, Ukraine
ORCID: https://orcid.org/0000-0002-4386-4103; vkobets@kse.org.ua. Scopus Author ID: 56006224700
***Research field*:** Data Science in Economics; Evolutionary Microeconomics; Robo-Advisor

Кобець Віталій Миколайович - доктор економічних наук, професор кафедри Комп'ютерних наук та програмної інженерії. Херсонський державний ун-т, вул. Університетська, 27. Херсон, 73003, Україна

Oleksii V. Gulin - student of the master's degree program on Information systems and technology. Kherson State University. 27, Universitetska Str. Kherson, 73003, Ukraine
ORCID: https://orcid.org/0009-0002-9043-5376; gulinleshka@gmail.com
***Research field*:** Business analytics; data science; machine learning

Гулін Олексій Вадимович - студент магістерської освітньої програми Інформаційні системи та технології кафедри Комп'ютерних наук та програмної інженерії. Херсонський державний ун-т, вул. Університетська, 27. Херсон, 73003, Україна

Pavlo S. Nosov - Candidate of Technical Sciences, Associate Professor, Head of the Department of Shipboard Computer Systems and Networks. Kherson State Maritime Academy, 20, Ushakov Ave. Kherson, 73000, Ukraine
ORCID: https://orcid.org/0000-0002-5067-9766; pason@ukr.net. Scopus Author ID: 57211927353
***Research field*:** Behavior Models; Human Factor; Navigation

Носов Павло Сергійович - кандидат технічних наук, доцент, завідувач кафедри Суднових комп'ютерних систем та мереж. Херсонська державна морська академія, проспект Ушакова, 20. Херсон, 73000, Україна