ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

# Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2026-07-27

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Matos, F., Marques, C. M., Vairinhos, V. & Soares, R. (2026). Reskilling and upskilling for a better future: Using a data-based approach. In Jon Hans Coetzer, Lucía Morales, Alexandra Severino, Marybelle Cherfan (Ed.), Selected Papers of the Second UNITAR International Research Conference (2UIRC). (pp. 87-103).: UNITAR.

Further information on publisher's website:
https://unitar.org/sustainable-development-goals/peace/our-portfolio/online-learning-and-education/un-20-our-common-agenda-and-role-higher-education-conflict-peace-and-development-research-conference

Publisher's copyright statement:
This is the peer reviewed version of the following article: Matos, F., Marques, C. M., Vairinhos, V. & Soares, R. (2026). Reskilling and upskilling for a better future: Using a data-based approach. In Jon Hans Coetzer, Lucía Morales, Alexandra Severino, Marybelle Cherfan (Ed.), Selected Papers of the Second UNITAR International Research Conference (2UIRC). (pp. 87-103).: UNITAR.. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

# Reskilling and Upskilling for a Better Future: Using a Data-Based Approach

Florinda Matos[1] | Carolina Marques[2] | Valter Vairinhos[3] | Rui Soares[4]

[1,2] Iscte - University Institute of Lisbon - DINÂMIA'CET, Portugal

[3] ICLab - ICAA - Intellectual Capital Association, Portugal

[4] CENTIMFE - Technological Center for the Mouldmaking, Special Tooling and Plastic Industries, Portugal

**Abstract:**

In today's rapidly changing global economy, the importance of upskilling and reskilling has become increasingly important. Upon industries and technologies major developments and significant transformations, individuals and organisations must adapt to attain and maintain competitive advantage and thus, ensure sustainable growth.

This paper, developed within the context of the *CoVE Future Skills Factory*, a Portuguese Centre of Vocational Excellence, which was created within the scope of the European project *CATALYST - European VET Excellence Centre for Leading Sustainable Systems and Business Transformation*, presents an exploratory case study developed in Portuguese companies. It illustrates the crucial role of a data driven approach to predict criticaly the need of skills, helping to construct educational programs, including higher education ones, to address personal and organisational growth. This informed approach supports the transformation of companies towards new business models that respect the challenges of sustainability within the framework of global instability.

**Keywords:** Reskilling, Upskilling, Data, New Business Models, Sustainability

## 1. Introduction

The *European Green Deal*, introduced by the European Commission in December 2019, aims to transform the European Union (EU) into the first climate-neutral continent by 2050. Achieving this objective requires comprehensive action across all sectors of the European economy, including investment in environmentally friendly technologies, promotion of innovation in industries, adoption of cleaner and more affordable modes of transportation, reduction of carbon emissions in the energy sector, improvement of energy efficiency in buildings, and cooperation with other nations to elevate global environmental standards. This initiative sets an interim goal of reducing greenhouse gas emissions by at least 55% by 2030, compared to 1990 levels, to achieve net-zero emissions by 2050 (European Commission, 2019).

To support this ambitious transition, the European Union has introduced the *Fit for 55* package (European Council, 2025), a collection of proposals to revise and update existing legislation and create new initiatives to align EU policies with these climate objectives (European Union, 2021; European Council, 2025). In recent years, the drive towards carbon neutrality has already led to developing emerging markets and solutions that are both environmentally friendly and economically viable. As a result, an increasing number of governments, regions, cities, and organisations are setting their own carbon neutrality goals, with zero-carbon technologies becoming more competitive in key sectors such as energy and transportation (European Commission, 2019).

The *European Skills Agenda*, launched in 2020 as part of this broader strategy, plays a crucial role in equipping individuals and enterprises with the skills needed to support the green and digital transitions (European Comission, 2020). This initiative focuses on fostering resilience, ensuring equitable access to upskilling opportunities, and promoting sustainable competitiveness. One of its core components, *Action 6*, aims to define and monitor the development of the skills required for the green transition, helping professionals adapt to emerging needs. By providing access to education, training, and lifelong learning, the agenda seeks to strengthen social justice and resilience in the face of future crises (Cedefop, 2012; European Comission, 2020, 2022).

This necessity of a swift transition towards a climate-neutral Europe and widespread adoption of digital technology is revolutionising our work, learning, social participation, and daily routines (European Comission, 2019; European Commission, 2022). Evidently, Europe and all countries, in general, can only seize these chances if their citizens acquire the appropriate skills. The Covid-19 pandemic has had a devastating effect on millions of individuals in the European Union who have either lost their employment or suffered substantial economic loss (Reshma et al., 2022). Many individuals must acquire new skills and transition to new employment opportunities in different areas of the economy. Continued upskilling will be necessary for individuals to maintain their employment in a new work setting (Reshma et al., 2022). If new educational programs were not to be introduced to address these challenges, offering novel training to adapt to the transition to a resource-efficient, circular, digital, and climate-neutral European economy, young people would find entry into the job market exceptionally difficult.

Acquiring new skills and developing existing skills is essential for achieving long-term and sustainable growth, productivity, and innovation. As a result, they play a vital role in determining the competitiveness of organisations, especially for small and medium-sized enterprises (SMEs) that need to re-think and redesign their business models, improving their performance while promoting the common good (European Comission, 2016; European Commission, 2018). This paper explores this topic demonstrating how data analysis can be an excellent tool for making decisions about current and future qualification needs.

## 2. Literature Review

Multiple studies conducted have shown the importance and relevance upskilling and reskilling have in the economy and sustainable growth, especially after the pandemic caused by the COVID-19. Research has emphasised the examination of the connection between the reskilling required by the economy and its growth in the contemporary context, which will also influence new business models for a more effective and efficient utilisation of human resources toward sustainable growth (Reshma et al., 2022).

The European Green Deal, as a catalyst to foster learning on environmental sustainability in the European Union, includes, as one of its core policy actions, the development of a European sustainability competence framework (European Commission, 2019). The mentioned framework aims to contribute and deliver education programmes to help learners into developing skills, attitudes, and a knowledge base in which they act out of empathy responsibility and care for the planet as well as public health (Bianchi et al., 2022).

Upskilling and reskilling are fundamental strategies for maintaining competitiveness, enabling individuals to absorb new competencies or refine current ones to align with the varying profession demands (Hasan et al., 2024). Upskilling and reskilling are sometimes interconnected. However, they denote two separate processes. Upskilling entails enhancing current abilities or acquiring new ones, whereas reskilling requires learning new competencies for a different profession or sector (Kilag et al., 2024).

On the one hand, upskilling refers to the enhancement of the current skills of the workforce to satisfy present and future demands, in which these skills (often called “soft” skills) can be refined and customised to address skill gaps or weaknesses in the labour market. On the other hand, reskilling is the process of training employees in a new skill to ease their transfer to a different department within the organisation (Hasan et al., 2024). Moreover, upskilling enhances performance in existing jobs and in individual’s competencies to elevate their job performance, whereas reskilling equips workers for new organisational duties and for new positions within the organisation (Hasan et al., 2024).

The significance of upskilling and reskilling in educational leadership and management has collected heightened attention in recent years, as the global labour market evolves swiftly (Kilag et al., 2024). One of the topics on the table for employers, students, educators, and program administrators is what type of skills and competencies the programs develop in students (Brundiers et al., 2021). Nonetheless, the World Economic Forum asserts that technology advancements would need the reskilling of 50% of the global workforce by 2025 (Hasan et al., 2024).

According to Reshma et al. (2022), reskilling and upskilling should be strengthened through investment in workforce development to enhance the skills necessary for maintaining competitiveness in the near future.

Notwithstanding, the arising conjunction of numerous frameworks of competencies in sustainability being proposed, specifically the framework of key competencies in sustainability, transparency and straightforwardness is still yet to be achieved (Brundiers et al., 2021). Nevertheless, to facilitate upskilling, proficiency in technology is essential to ensure a seamless workflow aligned with organisational objectives (Reshma et al., 2022).

According to the OECD in 2021, government policies and incentives are crucial to facilitate these initiatives, to foster collaborations between the commercial and the public sectors, to fund training programs and to guarantee access to education and technology (Hasan et al., 2024). For example, *SkillsFuture* in Singapore have demonstrated its pivotal role by adopting continuous education and improving workforce development (Hasan et al., 2024).

Concomitantly, educational data mining is concerned with developing approaches that extract knowledge from data derived from the educational context (Abu Tair and El-Halees, 2012). As assessed by Hasan et al. (2024), contemporary educational programs and platforms use technology to provide personalised and tailored training resources and to adapt courses to meet the specific need of individuals and organisations. Therefore, the data collected from historical and operational databases that reside within educational institutes, and that can also be collected from e-learning systems, have vast amounts of information that can be used to improve the quality of education programs (Abu Tair and El-Halees, 2012).

A developed and well-trained force that helps in enhancing the employees' capabilities is created by reskilling and upskilling the force by a corporation. Along with increasing the organisation's financial gain, it also improves worker retention and the maintenance of the market share for organisation survival (Reshma et al., 2022).

Thus, both upskilling and reskilling are crucial for maintaining competitiveness and adapting to technological advancements. As the *European Green Deal* promotes environmental sustainability education, it is crucial that governments and organisations set their own carbon neutrality goals as well as invest in policies and incentives within this scope. Simultaneously, educational data mining and technology-driven personalised training improve education quality. In addition, investing in workforce development enhances skills, boosts organisational performance, and ensures market survival, as well as aquiring new skills and developing existing skills are essential for achieving long-term and sustainable growth, productivity, and innovation.

## 3. Methodology

The methodology for the data analysis presented below is suggested as the process input of defining Vocational Education and Training (VET) courses, and illustrates the data-driven character of the process in which the identification, characterisation and skills specification and future VET courses should be based. This means that the available data decisively conditions the suggested methodologies that must always be based on up-to-date evidence. The present work has a pilot study nature that, based on the specific data and resources available within the scope of the project *Future Skills Factory*, a Portuguese Centre of Vocational Excellence (CoVE) created within the scope of the European project *CATALYST*, seeks to identify possible methodologies and tools to be used in studies with more resources. These studies will have to be based on a continuous monitoring of the relevant data for the specification and characterisation of these future skills and courses.

The data available under the *CATALYST* project, with regard to Portuguese participation, covers two components:

1. Interviews - Answers from 14 executives in 14 Portuguese companies during face-to-face interviews. A semi-structured questionnaire was used, covering 36 biographical questions and other questions characterising the respondent's current job and the company's behavior. The data resulting from these interviews is qualitative in the form of texts.

2. Questionnaire - The data collected through a questionnaire gathered information from the heads of 100 companies in Portugal. The resulting data set comprises 100 rows (each row corresponding to one company) and 78 columns. From the 100 answers obtained, 18 were incomplete and have been excluded, leading to a final set of 82 complete answers:

- Questions 1 to 14 refer to data that helps characterise the companies and the respondents.
- Questions 15 to 31 are binary questions referring to the 17 United Nations Sustainable Development Goals (SDGs).
- Questions 32 to 68 refer to questions coded on a 7-point Likert scale concerning the degree to which, in the respondent's opinion, the characteristic in question is present in the company.
- Question 51 - *"The company sets clear strategies and procedures for managing intellectual capital"* - was removed due to the number of missing responses.
- Questions 69 to 78 refer to binary variables that express, in the respondent's opinion, whether the mentioned characteristics are present or not in the company they belong to.

## 4. Data Analysis

### a. Qualitative Data: The Interviews Data Analysis

This research aims to describe the methodological options to examine qualitative and quantitative data gathered in research. The data analysis attempts to define outputs that are going to be relevant to the definition of course content. It is claimed that the CoVE should consider the perceptions collected through face-to-face interviews and dedicated surveys applied to a sample of companies in Portugal. Furthermore, this methodological proposal comprises a multivariate data analysis that will help to identify clusters of subjects related to sustainability issues, as well as segments of companies according to the perceptions of their heads regarding those issues.

Text mining (Kearney and Hvitfeldt, 2022; Michalke, 2012; Stielge and Robinson, 2017; Wickham and Grolemund, 2016) is the general methodology suggested for the analysis of the text resulting from interviews. In addition to general methods of counting words, word sequences, and sentences, a specific text mining method - topic analysis - is proposed to address the specific need for developing VET course content. The topic analysis aims to discover, characterise, and estimate by lists of words the topics or latent themes, implicit or subjacent to the texts resulting from interviews. The topical analysis seeks to identify themes or patterns underlying the interviewee's views regarding the subjects under study in the Portuguese CoVE *Future Skills Factory*. Even though the number of interviews is small, it generates a considerable amount of text and transforms unstructured data into manageable datasets for further multivariate analysis.

Identifying the topics: The text mining

The term *topics* is used to refer to the set or repository of all themes that emerged from the data analysis - either by using automatic means (texts topic analysis) in the case of interviews or by examining the associations among variables representing the numeric answers to questions in questionnaire. In the case of interviews, the topics were obtained by topic analysis (*LDA - Latent Dirichlet Allocation* algorithm) from the texts of 14 executives from 14 Portuguese companies questioned through face-to-face interviews. This specific dataset (interviews) contains 14 rows (corresponding to each of the sampled companies) and the answers (strings or texts) to the 36 questions included in the semi-structured questionnaire form.

The software for text mining

The text analysis calls for qualitative data analysis software, preferably with modules for performing the usual multivariate data analysis (Cluster Analysis, Biplots, Multidimensional Scaling) text mining and modules for Topic Analysis such as *LDA (Latent Dirichlet Allocation)*. In particular, the languages R and Python are especially useful, not only because they are free of cost, but also because they are sophisticated and up to date, having all the modules suggested above. The packages used in this proposal and those expected to be relevant for the preparation of the final study are:

***Table 1- List of R Packages useful for data analysis tasks, in particular, the main R text mining packages (see CRAN - The Comprehensive R Archive Network - https://cran.r-project.org/)***

| R Package | USED FOR | R Package | USED FOR |
|---|---|---|---|
| ***MASS*** | Data analysis | ***colorspace*** | Graphical text mining |
| ***stats*** | Data analysis | ***tm*** | Text mining |
| ***dplyr*** | Data management | ***quanteda*** | Text mining |
| ***readxl*** | Read EXCEL | ***quanteda.textplots*** | Graphical text mining |
| ***writexl*** | Write EXCEL | ***quanteda.textmodels*** | Statistical models for text mining |
| ***GGally*** | Graphical text mining | ***quanteda.textstats*** | Statistical models for text mining |
| ***FactoMineR*** | Text mining | ***lsa*** | Latent semantic analysis |
| ***factoextra*** | Text mining | ***Lda*** | Latent dirichlet allocation |
| ***dendroextras*** | Graphical text mining | ***topicmodels*** | Topic models for text mining |
| ***dendextend*** | Graphical text mining | ***gg2plot*** | R graphics |

Text mining is the methodology used to analyse the texts of interviews, including modules of NLP (Natural Language Processing) and Topic Analysis – LDA (Latent Dirichlet Allocation), in particular. The logic is as follows: it is assumed that the respondents, in our case, the companies' leaders use terms, expressions, phrases and concepts which express their culture, operational concerns and world views. Through textual analysis of the respective speeches, it is natural to try to detect the topics underlying those speeches, cultures, and concerns. This report suggests using LDA to identify, characterise and estimate the topics underlying the texts resulting from the face-to-face interviews.

The hope is that the topics thus estimated may serve as a slogan, inspire, or suggest to decision-makers the identification of possible relevant contents to include in VET courses. Experiences with texts collected by the Portuguese answers are encouraging and illustrated in the word cloud, present in Figure 1, which represents an analysis that allows a preliminary perception of the frequency and importance with which the various terms occur in the texts (Benoit et al., 2015).




***Figure 1 - Word cloud based on the relative frequency and importance of the terms used in the texts from the Portuguese interviews.***

The subsequent analysis shows the importance of the topics and is illustrated in Figures 2 and 3.

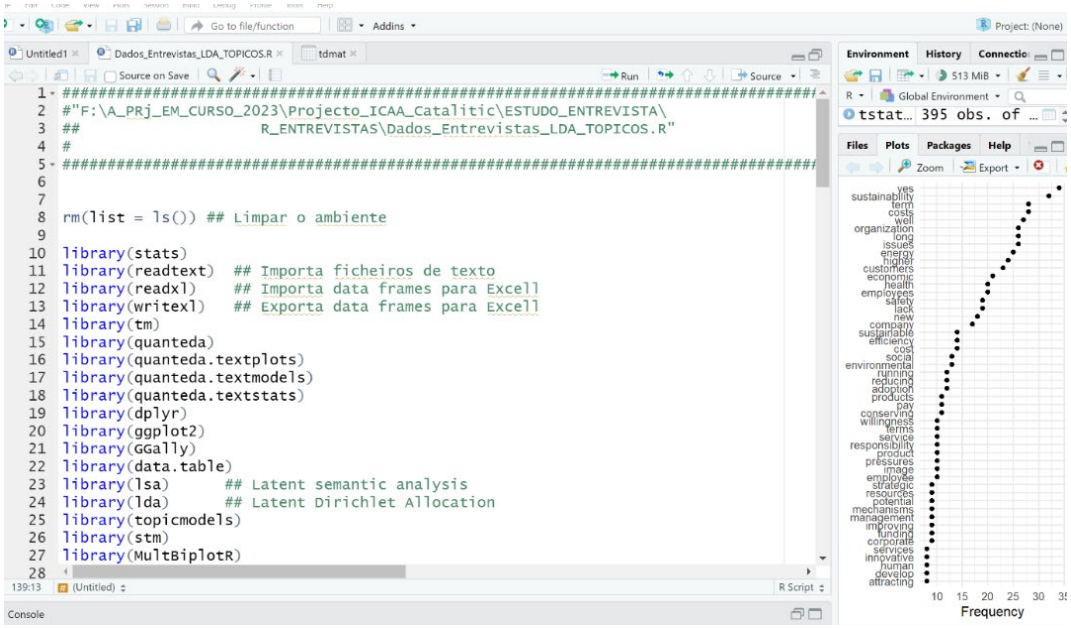



***Figure 2 - Text mining of the interview texts with indication of the packages used in the R script.***

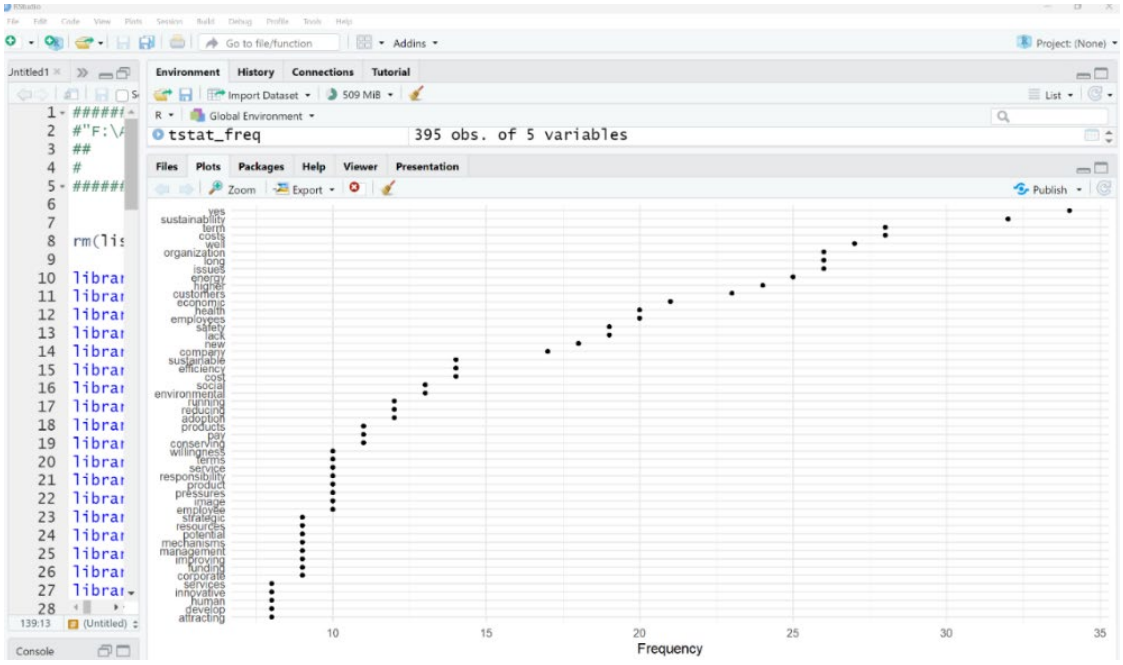
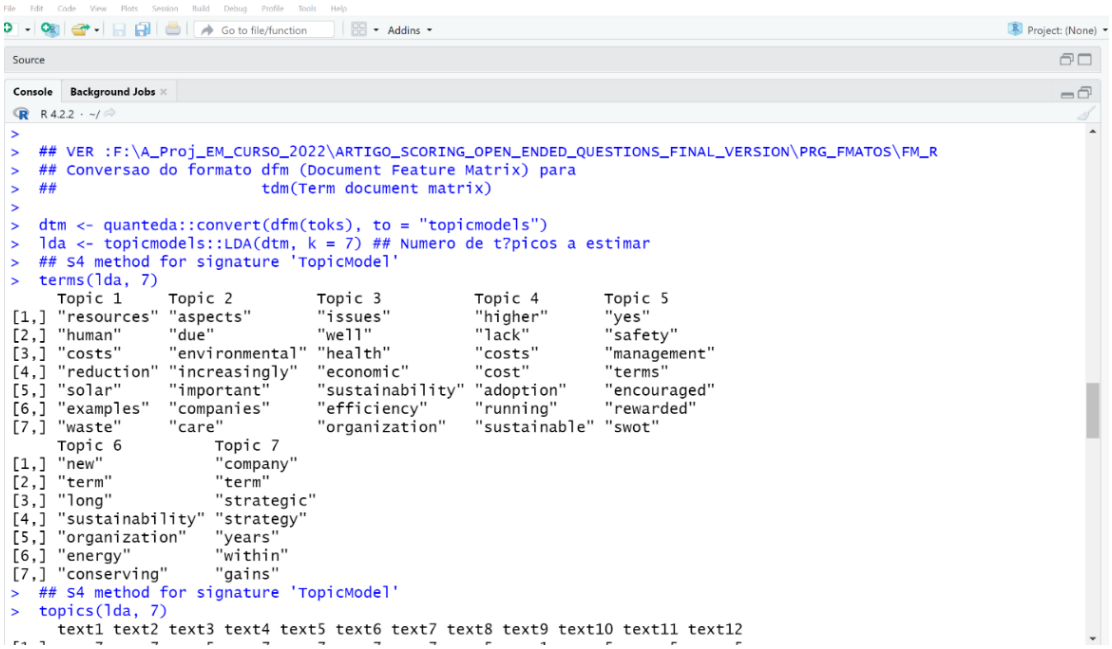



***Figure 3 - Results of the partial analysis of the interview texts, produced with the quanteda R package, with the absolute frequencies of the main terms occurring in the interview texts.***



```
> 
> ## VER :F:\A_Proj_EM_CURSO_2022\ARTIGO_SCORING_OPEN_ENDED_QUESTIONS_FINAL_VERSION\PRG_FMATOS\FM_R
> ## Conversao do formato dfm (Document Feature Matrix) para
> ##                      tdm(Term document matrix)
> 
> dtm <- quanteda::convert(dfm(toks), to = "topicmodels")
> lda <- topicmodels::LDA(dtm, k = 7) ## Numero de t?picos a estimar
> ## S4 method for signature 'TopicModel'
> terms(lda, 7)
     Topic 1     Topic 2         Topic 3          Topic 4       Topic 5
[1,] "resources" "aspects"       "issues"         "higher"      "yes"
[2,] "human"     "due"           "well"           "lack"        "safety"
[3,] "costs"     "environmental" "health"         "costs"       "management"
[4,] "reduction" "increasingly"  "economic"       "cost"        "terms"
[5,] "solar"     "important"     "sustainability" "adoption"    "encouraged"
[6,] "examples"  "companies"     "efficiency"     "running"     "rewarded"
[7,] "waste"     "care"          "organization"   "sustainable" "swot"
     Topic 6          Topic 7
[1,] "new"            "company"
[2,] "term"           "term"
[3,] "long"           "strategic"
[4,] "sustainability" "strategy"
[5,] "organization"   "years"
[6,] "energy"         "within"
[7,] "conserving"     "gains"
> ## S4 method for signature 'TopicModel'
> topics(lda, 7)
     text1 text2 text3 text4 text5 text6 text7 text8 text9 text10 text11 text12
```

***Figure 4 - Partial result of the LDA analysis, using the R Topicmodels package.***

The topics presented in Figure 4 show how different terms cluster together, reflecting underlying or implicit associations between terms in the discourse analysis. From what is verified above, 7 topics were specified and for each topic 7 terms are indicated to characterise the topic. This clustering attempts to reflect the implicit taxonomy of the interviewees' concerns. This method involves a set of steps (Blei, 2012; Grün and Hornik, 2011).

From the outset, the researcher should define the number of topics to be estimated. This number (k = 1, 2,...T) where T is usually a small number - 5, 6,...10 - can be obtained by trial and error, calculating for each possible solution a quality criterion, such as "perplexity", a number that expresses the difficulty to predict results with a specific model (see the LDA handbook of the R topic analysis package - Blei, 2012; Chang, 2009; Grün and Hornik, 2011). Subsquently, given the list of topics and the lists of terms which form the respective estimation, it is now desirable that the user suggests a label for the topic that tries to characterise the meaning of a specific topic, replacing the automatic names Topic 1, ...Topic 7. This is not mandatory but helps a great deal with interpretation.

For example, it is possible to suggest the following labels for the topics estimated from the LDA in Figure 4:

- 1: Human Resources Management
- 2: Environmental Issues
- 3: Short-term Sustainability (since it includes efficiency concerns)
- 4: Costs
- 5: Secure Management
- 6: Long-term Sustainability (the word *sustainability* is associated with the words *long* and *term*)
- 7: Strategic Management

It should be noted that the terms suggested for each topic can be used in a dictionary to identify the occurrence of these topics in additional or supplementary texts. In the context of the present study, this would allow us to know to what extent the topics identified for the study of one country occur or are also relevant for interviews in other countries, as the language used by all is English, thus arriving at common designations and topics across countries, with increased usefulness and general validity of the study, since it would be based on much larger volumes of data.

## b. Quantitative Data: The Questionnaire Data Analysis

The quantitative data gathered through the dedicated survey that was applied to a sample of companies in Portugal allowed for various multivariate data analysis techniques, namely:

- Biplots: in which the identifiers of the variables and the identifiers of the companies appear simultaneously in the same dispersion graph, allowing visual inspection of "interesting" associations.
- Multidimensional scaling and cluster analysis techniques: which allow to detect and examine associations between variables and relate the respective meanings that were the subject of the choices made by respondents in a sample of around 100 companies.

Beforehand, it is important to clarify the data structure and methodological options that were made to obtain a more reliable analysis. The data collected through a questionnaire gathered information from the heads of 100 companies in Portugal. The data set comprises 100 rows (each row corresponding to one company) and 78 columns / questions.

- Questions 1 to 14 refer to data that helps characterise the companies and the respondents. For these questions, there are only 81 valid answers / rows, 18 answers with missing values were eliminated.
- Questions 15 to 31 (SDG1, SDG2,...SDG17) are binary questions resulting from the manual recoding of the original question number 14, which referred to the 17 United Nations Sustainable Development Goals (SDGs).
- Questions 32 to 68 refer to questions coded on a 7-point Likert scale concerning the degree to which, in the respondent's opinion, the characteristic in question is present in the company. Although qualitative, the meaning of these variables allows them to be treated, without major issues, as ordinal variables using methods reserved for quantitative variables.
- One of these questions - [51] *"The company sets clear strategies and procedures for managing intellectual capital"* - was eliminated due to the number of missing values. Its inclusion would unnecessarily reduce the valid cases to only 30 answers or rows, so the final number of questions considered in the analysis in the Portuguese case was 36.
- Questions 69 to 78 refer to binary variables that express, in the opinion of the respondent, if the mentioned characteristics are present or not in the company they belong to.
- The total number of cases available for analysis was 100.

Nonetheless, it is assumed in this analysis that the data is observational (that available given the constraints), so the analysis must be performed from the perspective of a case study.

The next step of the data analysis, i.e., after the preparation of the dataset, is to apply statistical methods to discover and characterise patterns of associations and clusters of variables and companies. In Figures 5, 6, 7 and 8 are illustrated the outputs of biplots, cluster analysis and multidimensional scaling of the Portuguese data. A biplot (Gabriel, 1971) plots simultaneously information on the observations (companies) and the variables (related to sustainability) in a multidimensional dataset. Given the objectives of the study (related to the specification of VET course contents), the most important are patterns of association and clustering of questions/variables, which are especially insightful to suggest topics or themes to consider in the definition of courses content. For this reason, biplots and cluster analysis are very useful.

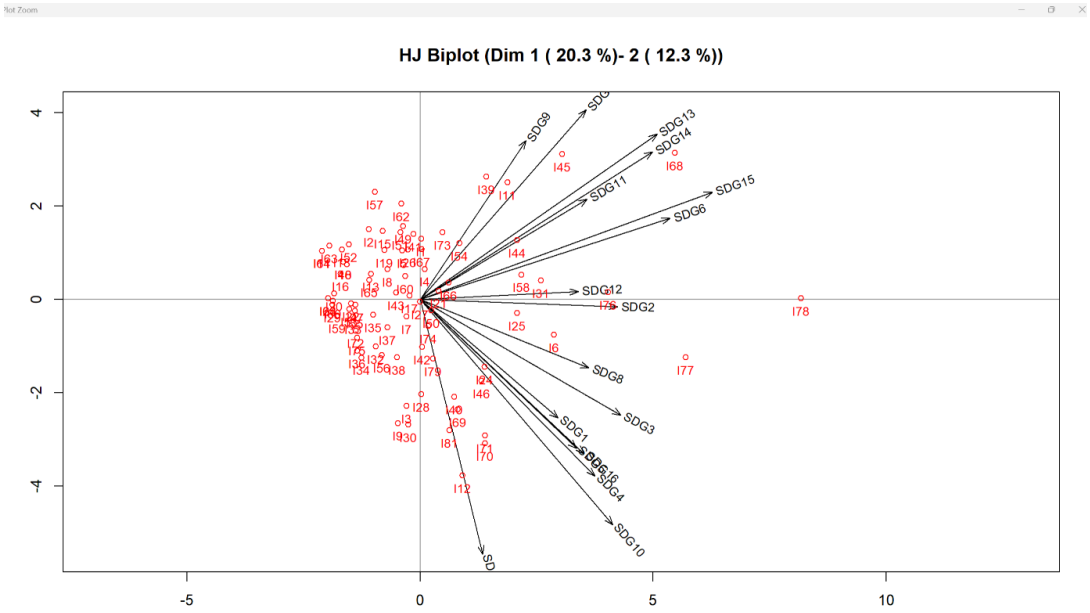



**Figure 5 - Biplot with the variables representing the 17 United Nations SDGs, produced with the packages *stats* and *ggplot2* (Wickham, 2009).**

The biplot displays in red, in each point, one of the 100 companies that answered the questionnaire, and the arrows, in black, display the SDGs. For example, quadrant II (top right) shows an association of SDGs related to innovation, climate, or water, while quadrant IV (bottom right) regards poverty and immaterial goals. Therefore, the clustering indicates how the interviewees are associated with the SDGs and how they responded to the questionnaire. In other words, the angles between variables represent correlations, which means that it is possible to detect by visual inspection relevant associations between these variables, for example {SDG11, SDG13, SDG14, SDG15}, which would possibly correspond to a *topic* considering the texts describing the meanings of these variables. Recall, for example:

- SDG11 - Sustainable Cities and Communities
- SD13 - Climate Actions
- SD14 - Life Below Water
- SD15 - Life On Land

These associations are confirmed in the following tree, obtained by cluster analysis. It is important to remember that the similarity between members of a cluster is bigger than the similarity between those members and the members of different clusters.

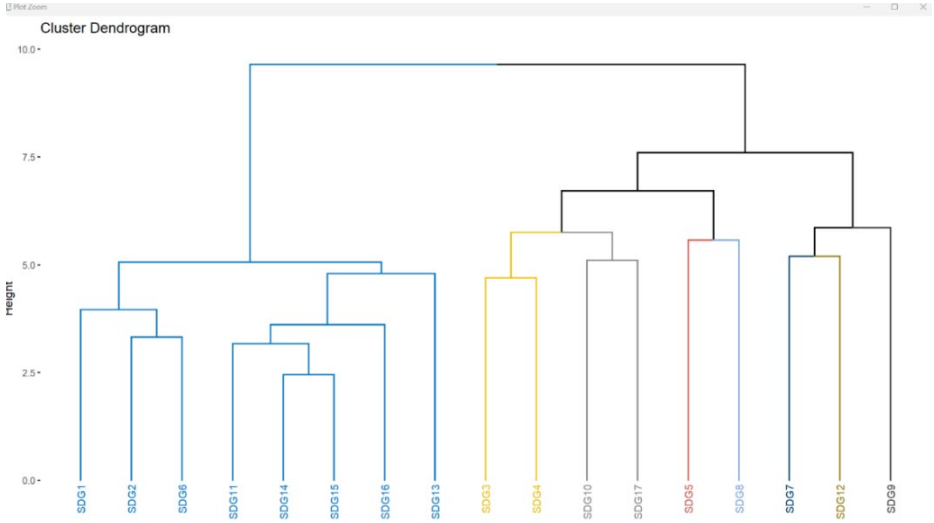



***Figure 6 - Result of the cluster analysis of the binary variables SDG.***

Figure 6 shows the groupings of SDGs that can be considered as possible course topics after examining these variables' textual descriptions. Those topics may be characterised by the suggestive labels proposed by the users and the lists of terms resulting from the study of the variables' metadata, such as the description of their meanings. Note that the grouping {SDG11, SDG14, SDG15 and SDG16} already appeared in the biplot of Figure 5.

Figure 7 shows a cluster analysis illustration of the columns expressed in likert scales. The labels used (see the inverted tree leaves) are inadequate in relation to their fulfilment and compromise the usefulness of the analysis, which will have to be repeated using much shorter symbols for those variables. The associations to be detected will serve to interpret the resulting *topics*.

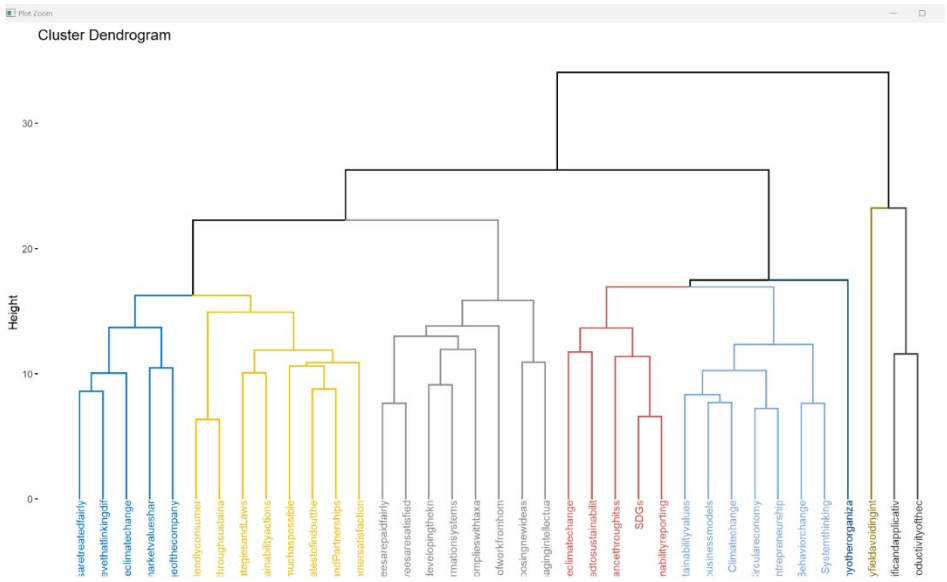



***Figure 7 - Cluster analysis of questions 32 to 68 - excluding question 51 - expressed in the Likert scale. The tree was cut in k=7 groups of variables.***

Once again, the analysis of variable associations suggested by this cluster analysis can serve as a starting point for defining topics to be considered in course content specification. In Figure 8, the binary variables 69, 70...76 form in a biplot clusters that, interpreted with the texts of the characteristics that describe them, can suggest topics to be used in the definition of courses.

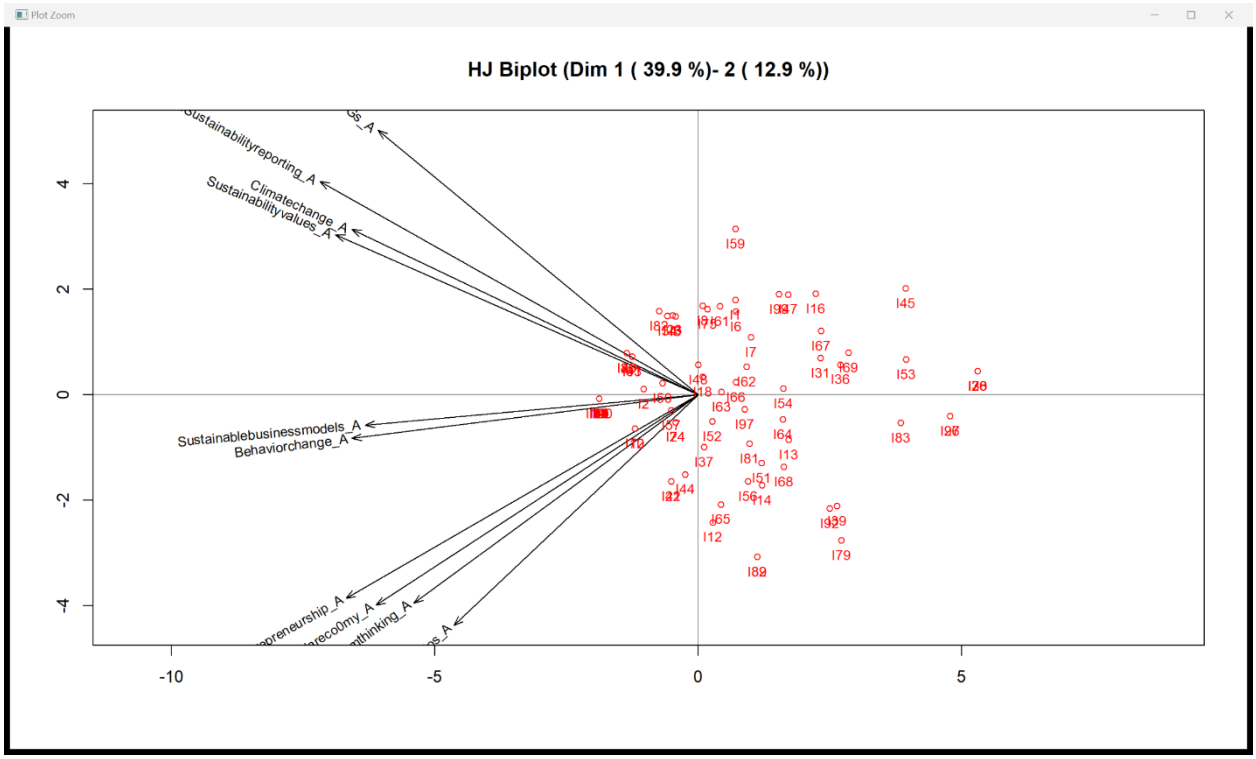



***Figure 8 - Biplots containing the binary variables 69, 70...76 represented by arrows whose angles are the respective correlations and by a cloud of red dots representing the 100 companies in the sample.***

For example, it is visible the group formed by the variables {Sustainable business models, Behaviour change} or the group {Sustainability reporting, Climate Change, Sustainability Values}. These two discovered variable associations may or may not be considered relevant for defining topics to include in courses and adding to the *topics* "repository" (see Figure 1).

## c. Preliminary Proposal of Courses

Assuming the availability of a joint list or repository of topics, either identified and estimated automatically or manually ("by hand"), based on previous analyses, which includes both topics identified automatically by LDA and those identified manually by interpreting the variable associations suggested by the methods used to analyse the questionnaire responses, let the total number of retained topics, regardless of their origin, be denoted as T. These topics, validated by decision makers, can now be used to specify potential courses. A course may be built from a single topic, two topics, three topics, or all T topics.

Theoretically, the total number of courses would be the number of parts it is possible to count in a set with T objects. That is: $2^T$. For example: if filtering reduces the number of admissible topics to T=10, then the total number of courses would be $2^{10} = 1024$, which is excessive for the purposes of the Project. If T = 7, the total number of possible courses would be $2^7 = 128$. From these considerations, it is possible to suggest the following methodology to identify courses:

1. Identify and characterise T (t = 1, 2,...T) "topics" or themes to be integrated into the courses and that can result from LDA-type automatic analysis or associations of variables detected by the analysis of the answers to the questionnaire. Each course can correspond to one or more topics. Thus, the total number of possible courses based on T (t = 1, 2,...T) topics would be 2T (binomial formula).
2. There can be $t$ courses corresponding to only one topic, $0.5*t*(t-1)$ courses with 2 topics, $(1/6)*t(t-1)(t-2)$ courses ,…$t$ courses with $(t-1)$ topics and 1 course with $t$ topics. All together gives $2^t$.

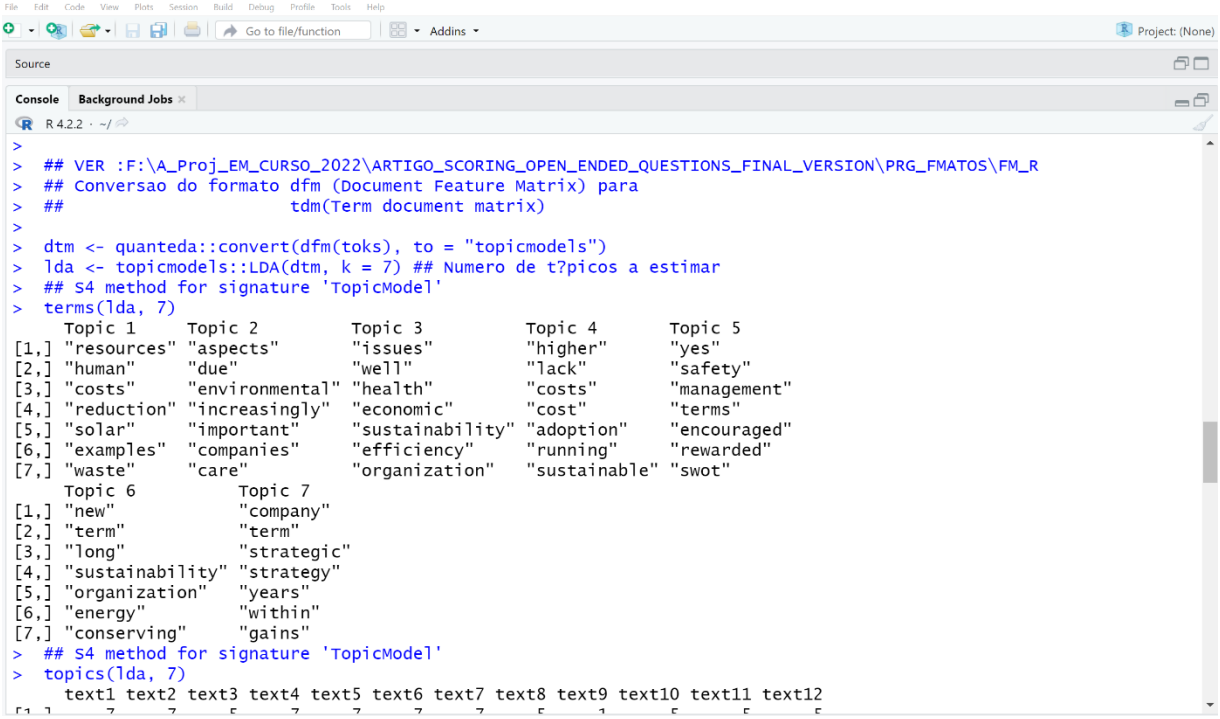



***Figure 9 - Topics estimated automatically using the LDA algorithm.***

As an example (see Figure 9), assume that the lists of terms automatically generated by LDA can be interpreted in such a way that it makes sense to assign, in the case of T = 7 topics the following labels or names to the topics corresponding to those 7 lists:

- Topic 1: HRM - Human Resources Management
- Topic 2: EI - Environmental Issues
- Topic 3: STS - Short-Term Sustainability
- Topic 4: COST - Costs
- Topic 5: SMG - Security Management
- Topic 6: LTEC - Long-Term Energy Consumption
- Topic 7: STMG - Strategic Management

The definition of admissible courses content can be made in a matrix (for k = 7, for instance).

Each row in the following matrix is an admissible choice of topics or themes to be tentatively included in the content of each course.

***Table 2 – This table illustrates tentative course contents using the topics identified through a data mining process. Tables like this one can be used as imput for the evaluation process and comparison of tentative course content.***

| COURSE | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 |
|---|---|---|---|---|---|---|---|
| **Course 1** | x | | | | | | |
| **Course 2** | | | | | | | |
| **Course 3** | x | | | x | | | |
| **Course 4** | | | | | | | |
| **Course 5** | | x | x | | | | x |
| **.....** | | | | | | | |
| **Course 128** | x | x | x | x | x | x | x |

As an example, three possible courses contents may eventually be specified from a table like this:

- Course 1 – HRM - Human Resources Management
- Course 3 – HRM + COST – Costs Considerations in Human Resources Management
- Course 5 – EI + STS - Environment and Short-Term Sustainability

Each of these courses, before being eventually proposed, should be analysed considering the following factors:

1. Homogeneity
2. Content Complexity
3. Duration
4. Requirements of the Students
5. Purpose
6. Cost

The final decision is the result of a specific study of each course, allowing the attribution of a particular priority. Decision analysis methodologies should be used for this purpose.

## 5. Research Findings & Actionable Recommendations

Despite the results obtained (see section 4c), which allowed to identify important topics and themes for the development of courses, in future work data collection, although it is possible to continue using quantitative questionnaires and interviews - aimed at randomly selected samples or following a specific plan - they must include data collection based on observatory monitoring literature on important activities and/or professions, considering specialised professional networks (e.g. LinkedIn).

This means that analysis methods not mentioned in this paper - such as text mining for big data and machine learning using LLM (Large Language Models) - will play a very important role.

## 6. Policy Implications

The transition to a more data-driven approach in analysing and defining vocational courses and specific qualification areas for different levels of education has significant policy implications. By taking advantage of data analysis methodologies, policymakers can improve decision-making processes, optimise the distribution of resources and improve the alignment between vocational education and labour market requirements.

It's important to recognise that plenty of data is available and can be used for approaches that facilitates policymakers' decisions. The data-driven approach makes it possible to identify trends and skills gaps in the labour market. It also makes it possible to identify gaps between training offers at different levels of education and qualification, employment data, business needs, public education, and employment policies**:**

1. The specified analysis can, for example, ensure that graduates have the right skills to meet the demands of the labour market.
2. Based on predictive models, this type of forecasting can provide a global view of the labour market and anticipate future changes in skill requirements, as well as potential labor shortages or oversupply.
3. This forecast can highlight regional disparities, enabling policymakers to design policies tailored to the specific needs of each region or sector, thereby optimising funding allocation.
4. The data makes it possible to strengthen industry-education co-operation and the public system. The industry's need for competencies makes it possible to design apprenticeship curricula that responds to these needs and public policies that foster this collaboration.
5. Data-based analysis provides information that can be used to define personalised learning paths, tailored to needs, which maximises employment possibilities.
6. These methodologies make it possible to identify gaps and reinforce lifelong learning strategies (constantly updated as labour markets change).

7. Data analysis can help to identify disparities in access to teaching and learning systems and can, therefore, help to increase equity and inclusion in access to education.

## 7. Conclusion

Upskilling and reskilling strategies are crucial for empowering workers in meeting the demands of the dynamic labor market and driving productivity while also fostering innovation. As such, policymakers have a decisive role in enabling these efforts to progress, by creating frameworks promoting lifelong learning, by bridging skill gaps and by ensuring equitable access to training and development opportunities.

The investment in workforce development, public-private sectors partnerships, reskilling programs for job transition, integration of technology into education and training, focus on soft skills development, localised and customised training solutions are all strategies that can be adopted to aim at upskilling and reskilling initiatives.

Integrating predictive, data-driven methodologies into the design of courses and training areas has the potential to transform workforce development strategies. However, due to the lack of universally accepted models, further research with larger datasets is needed. This exploratory research, which cannot be generalised, is the first stage of a more comprehensive, ongoing research project.

## 8. Acknowledgements

This paper is an output of the Erasmus+ Project *CATALYST - Leading Sustainable Systems and Business Transformation*. CATALYST has received funding from the ERASMUS-EDU-2021-PEX-COVE (reference number 101056114).

## 9. References

Abu Tair, M.M., El-Halees, A.M., 2012. Mining Educational Data to Improve Students' Performance: A Case Study 2.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A., Lowe, W., 2015. quanteda: Quantitative Analysis of Textual Data. CRAN: Contributed Packages. https://doi.org/10.32614/CRAN.package.quanteda

Bianchi, G., Pisiotis, U., Cabrera, M., Punie, Y., Bacigalupo, M., 2022. The European sustainability competence framework. https://doi.org/10.2760/13286

Blei, D.M., 2012. Probabilistic topic models. Commun ACM 55, 77–84. https://doi.org/10.1145/2133806.2133826

Brundiers, K., Barth, M., Cebrián, G., Cohen, M., Diaz, L., Doucette-Remington, S., Dripps, W., Habron, G., Harré, N., Jarchow, M., Losch, K., Michel, J., Mochizuki, Y., Rieckmann, M., Parnell, R., Walker, P., Zint, M., 2021. Key competencies in sustainability in higher education—toward an agreed-upon reference framework. Sustain Sci 16, 13–29. https://doi.org/10.1007/s11625-020-00838-2

Cedefop, 2012. Green skills and environmental awareness in vocational education and training Synthesis report, 24. https://doi.org/10.2801/78825

Chang, J., 2009. lda: Collapsed Gibbs Sampling Methods for Topic Models. CRAN: Contributed Packages. https://doi.org/10.32614/CRAN.package.lda

European Comission, 2022. Green Skills and Knowledge - Labelling ESCO. Brussels.

European Comission, 2020. European Skills Agenda for sustainable competitiveness, social fairness and resilience. Brussels.

European Comission, 2016. EntreComp : the entrepreneurship competence framework. Publications Office.

European Commission, 2019. The European Green Deal. Brussels.

European Commission, 2018. EntreComp: The European Entrepreneurship Competence Framework, Publications Office. Luxembourg. https://doi.org/10.2767/88978

European Council, 2025. Fit for 55 [WWW Document]. https://www.consilium.europa.eu/en/policies/fit-for-55/#what.

European Union, 2021. REGULATION (EU) 2021/1119 European Climate Law, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R1119&from=EN.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453–467. https://doi.org/10.1093/biomet/58.3.453

Grün, B., Hornik, K., 2011. topicmodels : An *R* Package for Fitting Topic Models. J Stat Softw 40. https://doi.org/10.18637/jss.v040.i13

Hasan, M., Haque, Md.A., Nishat, S.S., Hossain, Md.M., 2024. Upskilling and Reskilling in a Rapidly Changing Job Market: Strategies for Organizations to Maintain Workforce Agility and Adaptability. European Journal of Business and Management Research 9, 118–126. https://doi.org/10.24018/ejbmr.2024.9.6.2502

Kearney, M.W., Hvitfeldt, E., 2022. Textfeatures: Extracts Features from Text (0.3.3).

Kilag, O.K.T., Padilla, K.S., Yorong, F.S., Merabedes, J.G.A., 2024. Importance of Upskilling and Reskilling in Educational Leadership and Management. European Journal of Learning on History and Social Sciences 1.

Michalke, M., 2012. koRpus: Text Analysis with Emphasis on POS Tagging, Readability, and Lexical Diversity. CRAN: Contributed Packages. https://doi.org/10.32614/CRAN.package.koRpus

Reshma, D., Ph, S., Rajan, R., Yaseen, S., Ravichandran, R., 2022. The Importance Of Reskilling And Upskilling In Post Covid Economy-A Critical Study The Importance Of Reskilling And Upskilling In Post Covid Economy-A Critical Study. https://doi.org/10.17605/OSF.IO/U26BW

Stielge, J., Robinson, D., 2017. Text Mining with R. https://www.tidytextmining.com/.

Wickham, H., 2009. ggplot2. Elegant Graphics for Data Analysis.

Wickham, H., Grolemund, G., 2016. R for Data Science.