



A DATA-DRIVEN COMPETENCY DETECTION MODEL FOR IMPROVING WORKFORCE PLANNING AND DECISION-MAKING

Llerena Lucrecia*¹, Rodríguez Nancy², Bustamante Jefferson³ and Angel Torres⁴

^{1,2}Software Engineering Program, Quevedo State University, Quevedo, 120301 Ecuador.

^{3,4}Master in data science, Quevedo State University, Quevedo, 120301 Ecuador.

¹<https://orcid.org/0000-0002-4562-6723>, ²<https://orcid.org/0000-0002-0861-4352>

³<https://orcid.org/0009-0001-6541-3258>, ⁴<https://orcid.org/0000-0002-7037-7191>

Email: *lllerena@uteq.edu.ec, nrodriguez@uteq.edu.ec, jefferson.bustamante2016@uteq.edu.ec, atorres@uteq.edu.ec

ARTICLE INFO

Article History

Received: December 14, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Data science,
Workforce analytics,
Competency modeling,
Machine learning,
Decision support.

ABSTRACT

Digital transformation has intensified the need for computational tools capable of identifying emerging workforce competencies across industries. This study proposes a scalable analytical model that integrates data mining, natural language processing (NLP), and machine learning to extract and classify professional skills from large volumes of structured and unstructured labor information. The methodology was implemented through a four-stage pipeline—extraction, preprocessing, modeling, and visualization—and validated using a case study of job-market data collected from official records and digital job platforms between 2020 and 2024. Supervised (Random Forest, SVM) and unsupervised (K-Means, DBSCAN) algorithms were applied to detect semantic patterns and cluster competencies, with Random Forest achieving the highest performance (F1-score = 96.64%). The model identified four dominant competency clusters—technological, managerial, operational, and hybrid—highlighting the growing integration of technical and soft skills required in digitally intensive work environments. The proposed framework contributes a generalizable computational approach for workforce planning, decision support, and operations management. Its modular architecture enables adaptation to multiple industrial contexts and supports continuous monitoring of skill evolution. This makes the model a valuable tool for organizations, policymakers, and researchers seeking data-driven insights into the dynamics of digitally mediated labor markets.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The rapid evolution of digital technologies has transformed job search, recruitment, and talent management processes, with digital platforms and online job portals progressively replacing traditional job-placement mechanisms and introducing new forms of interaction between employers, workers, and automated systems [1]. This shift has generated a data-rich labor environment in which digital labor platforms operate as socio-technical systems, reshaping how skills are signaled, matched, and demanded, and producing significant social and economic impacts that require advanced computational approaches to identify emerging labor-market structures, competency patterns, and inequalities characteristic of the digital society [2-4].

In this context, data science has become a fundamental tool for analyzing large-scale labor information, as text mining, machine learning, and natural language processing (NLP) enable the identification of demand patterns, occupational profiles, and emerging skill requirements [2]. These methodologies support digital transformation by facilitating the adaptation of training and organizational strategies to technological change and by providing analytical insights into how digitalization restructures employment, skills, and labor institutions [5]. Across Latin America, the adoption of data-driven technologies for labor-market analysis remains uneven, particularly in contexts affected by informality, territorial disparities, and gender gaps [6]. Although regional unemployment declined from 6.5% to 6.1% between 2023 and 2024 [7], structural imbalances continue to limit access to quality employment.

While initiatives supported by multilateral organizations such as the Inter-American Development Bank demonstrate the potential of big data to anticipate skill shortages and support more inclusive training policies [8], most advances are concentrated in countries with consolidated digital ecosystems. As a result, emerging economies—such as Ecuador—face significant analytical gaps. In the Ecuadorian context, digitization and automation are reshaping skill demand, particularly in technology-intensive and service-oriented sectors [9]; however, the lack of integrated frameworks linking structured administrative data with unstructured information from digital job portals constrains systematic analysis of evolving skill trends and limits evidence-based planning in education, innovation, and productive development [10].

This study aims to develop and validate a data-science-based analytical model to identify, classify, and cluster professional competencies in the Ecuadorian labor market between 2020 and 2024 by integrating structured administrative records with unstructured data from digital job platforms. The model is implemented through an automated pipeline combining web scraping, natural language processing, and supervised and unsupervised machine learning techniques—including Random Forest, SVM, K-Means, and DBSCAN—with interactive visualization in Streamlit to support scalable labor-demand analysis. Unlike previous approaches based on isolated data sources, the proposed hybrid socio-technical framework integrates heterogeneous information within a unified computational model, enhancing the accuracy, interpretability, and adaptability of labor market analytics for evidence-based decision making in emerging digital economies.

II. RELATED WORK

Over the past decade, data science has significantly transformed how governments and organizations analyze labor markets and design vocational training policies, responding to the growing need to align labor supply with employer demand. In this context, regional competitiveness increasingly depends on the capacity of labor systems to adapt to technological change through predictive and evidence-based strategies [11]. Recent research demonstrates that machine learning and natural language processing (NLP) techniques are effective for extracting and classifying job skills from unstructured sources such as job advertisements and administrative databases [9], [12]. Studies show that tokenization, vectorization, and text-mining approaches enable the detection of skill gaps and the anticipation of demand in emerging sectors, particularly for digital and technical profiles [13], [14]. Beyond skill extraction, a growing body of literature examines platform-mediated labor markets as socio-technical systems shaped by information and communication technologies (ICTs).

Digital platforms restructure how work is matched, monitored, and distributed, influencing how competencies are signaled and evaluated and generating new forms of visibility and precarity [15-17]. These perspectives emphasize that skill demand should be understood not only as an economic variable, but as a process mediated by digital infrastructures. The impact of automation and artificial intelligence on labor-market transformation has also been widely studied. Digitalization alters occupational structures and increases demand for competencies related to analytics, automation, and technological governance [18], while artificial intelligence reshapes career trajectories and requires continuous training in cognitive and digital skills to maintain employability in technology-intensive environments [19]. From a social and institutional perspective, digitalization interacts with demographic and structural factors. Empirical studies highlight persistent inequalities related to gender and family roles, even in contexts with advanced equality frameworks, underscoring the need for inclusive labor policies that address socio-digital barriers [20], [21].

In Latin America, research highlights the challenges associated with automation and competency-based management in public-sector organizations. These studies emphasize that technological adoption must be accompanied by institutional leadership and capacity-building strategies to ensure sustainable digital transformation [22], [23]. In this regard, initiatives promoted by multilateral organizations demonstrate that big data can support the anticipation of skills demand and the design of more inclusive training policies. Finally, broader analyses of digital transformation identify government and economic digitization as key drivers of sustainable development. Evaluating digital maturity through empirically validated models and understanding digitization as a structural factor in economic modernization are essential for strengthening public administration and labor-market resilience [24], [25]. Together, these contributions support the integration of data science into labor-market research as a means to better understand occupational dynamics and the social implications of digitalization and platformization.

III. METHODOLOGY

This methodological design follows a socio-technical perspective, recognizing that digital job platforms not only generate large volumes of textual information but also shape social and economic dynamics in the labor market. Therefore, computational techniques such as NLP and machine learning were used not only to extract and classify skills, but to uncover behavioral patterns associated with digital employment ecosystems.

III.1 DATA COLLECTION AND PREPROCESSING

The methodological process began with the identification and compilation of structured and unstructured data from digital job platforms, which serve as socio-technical environments where employers, workers, and platform algorithms interact. These platforms generate large volumes of textual information that reflect behavioral, economic, and organizational patterns within digital labor ecosystems. Structured data from the National Institute of Statistics and Census (INEC) was incorporated as part of the case study, while unstructured data was extracted from national and international job portals using automated web-scraping procedures.

This integration enabled the construction of a comprehensive dataset suitable for analyzing trends in labor demand across formal and informal employment segments between 2020 and 2024. Data collection followed a semi-automated process designed to ensure the consistency and reliability of the information. Records were validated and standardized according to variables such as occupation, economic sector, educational level, required skills, and geographic location. Cleaning procedures were applied to remove duplicates, correct inconsistencies, and normalize categorical values, ensuring the integrity of the dataset for subsequent analytical stages.

During preprocessing, textual descriptions from job advertisements were transformed into quantifiable representations using natural language processing (NLP) techniques. Tokenization, lemmatization, and TF-IDF weighting were applied, supported by a domain-specific vocabulary of job-related competencies to enhance semantic accuracy and reduce lexical noise. This step ensured the uniform structuring of the textual corpus and strengthened its interpretability for computational modeling.

The resulting dataset was consolidated into an integrated database that served as input for the classification and clustering stages. By combining administrative records with data extracted from digital platforms, the study adopted a hybrid socio-technical approach that enabled the exploration of skill dynamics within digital labor markets. This methodological process ensured the quality, traceability, and timeliness of the information, establishing a robust empirical foundation for the analyses presented in the following sections.

III.2 COMPARATIVE ANALYSIS OF DATA SCIENCE TECHNIQUES

This methodological phase compared the performance of multiple data science techniques to identify the most suitable models for detecting and classifying competencies in job advertisements from digital labor platforms. The comparison followed four criteria commonly applied in computational social science: accuracy, scalability, interpretability, and computational cost. Both supervised and unsupervised approaches were evaluated. Random Forest and SVM were selected for their robustness in handling heterogeneous textual data, while K-Means and DBSCAN were used to identify thematic clusters and emerging skill patterns. This configuration offered complementary perspectives: predictive learning from supervised models and exploratory segmentation from unsupervised ones.

Controlled simulations were conducted on the consolidated dataset generated in the previous phase. Metrics such as accuracy, precision, recall, F1-score, and silhouette coefficient were used to assess model performance. Execution time and resource consumption were also monitored to evaluate operational efficiency in large-scale digital environments. All experiments applied cross-validation and a stratified 70/30 partition to ensure stability and comparability. Hyperparameters were systematically tuned and documented to maintain methodological transparency. The findings highlighted the advantages of combining supervised and unsupervised learning to obtain a comprehensive understanding of skill dynamics in digital employment ecosystems, informing the selection of algorithms detailed in Section 4.2.

III.3 MODEL DESIGN AND IMPLEMENTATION

The analytical model was conceived as a socio-technical system capable of integrating structured and unstructured data to analyze competency patterns within digital job platforms. Its architecture was designed to be modular and scalable, facilitating reproducibility and adaptation to evolving labor market conditions. The model followed a sequential ETL (extract, transform, load) pipeline. During extraction, administrative indicators and scraped job descriptions were merged. The transformation phase applied NLP-based normalization and semantic vectorization to capture relationships between skills while reducing lexical variability. The unified database created during loading served as the foundation for both supervised and unsupervised learning components. Supervised models (Random Forest and SVM) and unsupervised models (K-Means and DBSCAN) were trained using the validated dataset, benefiting from the comparative insights described in Section 3.2.

Evaluation relied on accuracy, precision, F1-score, and cluster coherence metrics, supported by cross-validation to ensure robustness. Hyperparameter tuning followed a documented and systematic process to maintain replicability. To enhance interpretability, an interactive dashboard was developed using Streamlit. This interface enables geographical, sectoral, and thematic exploration of results, revealing distributions of job offers, dominant skills, and emerging hybrid profiles. Visual representations—such as scatter plots, word clouds, and sunburst charts—facilitate the analysis of complex competency structures. The modular architecture allows new data sources to be incorporated and models to be recalibrated as digital platforms evolve. This ensures that the system functions as an adaptive analytical tool for monitoring labor demand trends in digital societies and for supporting longitudinal and cross-platform comparisons.

IV. RESULTS AND DISCUSSIONS

This section presents the results obtained from applying the proposed analytical model. It sets out the findings derived from processing, classifying, and grouping the professional skills identified in the Ecuadorian labor market between 2020 and 2024. The results are organized around a comparison of the performance of the data science techniques used, the validation of the models developed, and the hierarchical characterization of the skills detected.

IV.1 SELECTION OF LABOR DATA ANALYSIS TECHNIQUES

Analysis of the results obtained allowed the identification of the most effective data science techniques for studying the Ecuadorian labor market. The combined use of data mining, natural language processing (NLP), and machine learning models improved the detection of patterns in job offers, facilitating the identification of emerging skills and the classification of professional profiles with high accuracy. The data showed that NLP made it possible to structure textual information from job portals and institutional records, while supervised and unsupervised algorithms helped classify and group skills in a consistent manner. Taken together, these results demonstrated the strength of hybrid methodologies that integrate heterogeneous data sources, thereby expanding the analytical coverage of the Ecuadorian labor market.

IV.1.1 Advances in Data Science Applied to the Labor Market

The results confirmed that data science had established itself as a key tool in labor market analysis. A comparative analysis of recent publications showed a sustained increase in the use of machine learning and natural language processing (NLP) models for the automatic identification of skills. Among the most representative techniques, supervised classification models allowed the detection of regularities in job descriptions, while thematic clustering methods proved useful for segmenting competencies into coherent clusters.

In the Ecuadorian context, the integration of these approaches made it possible to characterize a dynamic labor market in which traditional demands were combined with emerging digital skills. These advances reflected a broader transition toward digitally mediated labor markets, where job platforms increasingly shaped the visibility, categorization, and evolution of required skills.

IV.1.2 Predominant Models and Techniques

Analysis of the literature and the processed databases determined that Random Forest, Support Vector Machines (SVM), K-Means, and DBSCAN models were the most effective techniques for processing heterogeneous employment data. Their application demonstrated a high capacity to distinguish competency patterns, group professional profiles, and reduce data redundancy. The comparison between techniques revealed significant differences in performance, interpretability, and computational cost, as detailed in Table 1. The results showed that Random Forest and SVM algorithms provided better performance for skill classification tasks, while K-Means and DBSCAN performed more effectively in detecting clustered patterns and reducing noise. These findings supported the selection of a mixed analytical model that combined predictive and exploratory techniques to maximize the accuracy and coverage of labor data. The effectiveness of these models also reflected the increasing complexity of digital labor ecosystems, in which algorithmic mediation and platform-based interactions generated heterogeneous and rapidly evolving data structures.

Table 1: Comparison of classification, prediction, and clustering algorithms according to the literature.

Technique	Type of technique	Advantages	Limitations
Random Forest	Classification	Robustness in the face of unbalanced data	High computational cost
SVM	Prediction	Efficient in small sets	Sensitive to scattered data
Decision Tree Classification	Classification	Interpretability and simple visualization	Prone to overfitting if not pruned properly
Linear Regression	Prediction	Simplicity and computational efficiency	Assumes linear relationship; sensitive to outliers
Logistic Regression	Classification	Good for probabilities and binary classification	Requires independence between explanatory variables
K-Means	Clustering	Easy implementation and scalability	Requires defining the number of clusters a priori
DBSCAN	Clustering	Automatically identifies groups	Sensitive to density parameters

Source: Authors, (2026).

IV.1.3 Methodological Gaps and Observed Trends

Critical examination of the results revealed three substantive aspects in the evolution of job skills analysis:

- Gap in coverage of informal employment. Traditional models showed limitations in capturing skills in unregulated sectors, highlighting the need to incorporate complementary sources of information.
- Limited integration of structured and unstructured sources. The literature reviewed indicated that most studies worked with only one type of data, which restricted the ability to generate comprehensive comparative analyses.
- Limited adaptation to digital evolution. Existing models did not incorporate automatic update mechanisms, which reduced their relevance in contexts of rapid technological change.

Overall, the results reflected a clear trend toward hybrid and modular models capable of combining multiple sources of information with machine learning algorithms. In addition, three transformation patterns in the Ecuadorian labor market were identified:

- A process of hyper-specialization, characterized by a predominance of advanced technical skills.
- Accelerated update cycles, which reduced the effective lifespan of digital skills.
- A growing demand for hybrid profiles that combined technical and soft skills.

These results demonstrated that the application of data science techniques not only optimized the extraction of labor information but also revealed structural market dynamics that had remained hidden in traditional studies. These methodological gaps became more evident as labor systems transitioned toward digitally mediated environments, where skills evolved rapidly, platforms mediated access to employment, and conventional analytical approaches proved increasingly insufficient.

IV.2 COMPARISON OF SELECTED TECHNIQUES

This section presents the comparative results of the performance of data science techniques applied to the analysis of Ecuadorian job offers, based on metrics of accuracy, scalability, interpretability, and computational cost. The findings reflect the effectiveness of the selected models in identifying professional skills and segmenting job profiles derived from INEC data and job portals.

IV.2.1 Comparative Analysis of Selected Techniques

The experimental results revealed that Random Forest and K-Means were the algorithms with the best overall performance, combining high scalability with satisfactory interpretability. In contrast, SVM and DBSCAN performed moderately, standing out for their efficiency in small sets, but showing less robustness in scattered data. These results are summarized in Table 2, which presents the comparative metrics obtained through controlled simulations.

Table 2: Comparison of selected techniques.

Technique	Scalability	Interpretability	Computational Cost
Random Forest	High	Moderate	Medium
SVM	Medium	Low	Medium
K-Means	High	Medium	Low
DBSCAN	Medium	Low	Low

Source: Authors, (2026).

The values recorded reflected an inverse relationship between computational cost and interpretability. Random Forest demonstrated greater stability in heterogeneous contexts, while K-Means achieved remarkable efficiency in thematic clustering tasks. This evidence supported the adoption of a mixed approach that combined supervised and unsupervised models to obtain a comprehensive view of the Ecuadorian labor market.

IV.2.2 Findings by Technique Applied

The individual results by technique showed different levels of performance depending on the type of analytical task. As summarized in Table 3, Random Forest outperformed the other techniques in classification metrics, achieving an F1-score of 0.91 and an AUC of 0.93, which demonstrated strong predictive performance. K-Means achieved an internal cohesion value of 0.70, confirming its effectiveness in segmenting thematic profiles, while DBSCAN excelled in detecting anomalies and nonlinear patterns. SVM, although less robust in the presence of noise, proved useful in smaller and highly homogeneous datasets.

Table 3: Comparison between findings by technique.

Technique	Scalability	Computational Cost
Random Forest	F1-score greater than 0.91; AUC > 0.93; strong performance in binary and multi-class classification	Higher computational cost in large datasets; slower than SVM
SVM	Fast training in datasets with fewer than 10,000 records; competitive accuracy	Sensitive to data dispersion and semantic noise
K-Means	Identification of thematic clusters (technological, managerial, operational, and hybrid); silhouette coefficient > 0.70	Required prior dimensional reduction and manual adjustment of the number of clusters
DBSCAN	Detection of nonlinear clusters and outliers	High sensitivity to density parameters and lower stability in scattered data

Source: Authors, (2026).

IV.2.3 Validation of the Job Competency Analysis Model

The validation of the classification models reaffirmed the superiority of Random Forest, which achieved the highest values in precision and sensitivity. Table 4 presented the performance metrics obtained from the processed dataset.

Table 4: Results of the models applied to the identification of competencies.

Model	Accuracy	Precision	Sensitivity (Recall)	F1-score
Random Forest	0.9667	0.9730	0.9600	0.9664
SVM	0.8383	0.8503	0.8383	0.8419

Source: Authors, (2026).

The results indicated that Random Forest achieved an F1-score of 96.64%, significantly outperforming SVM, which reached an F1-score of 84.19%. These metrics confirmed the suitability of Random Forest for identifying complex patterns in heterogeneous labor data, demonstrating superior generalization capability and lower error rates. Likewise, the unsupervised models K-Means and DBSCAN reinforced the exploratory analysis by identifying four main thematic clusters—technological, managerial, operational, and hybrid—which were validated using a silhouette coefficient of 0.7482. This value demonstrated the internal consistency of the resulting cluster structure.

IV.2.4 Competency Segmentation and Cluster Analysis

Cluster analysis revealed a coherent structure in the organization of professional competencies. The resulting categories combined technical skills—such as programming, data analysis, and visualization—with soft skills including communication, leadership, and customer service. The identified clusters showed the coexistence of technical and soft competencies, indicating the presence of hybrid professional profiles within the analyzed labor market. This pattern was observed consistently across the clustered data, reflecting structural changes in the composition of demanded skills within digitally mediated employment contexts. The segmentation was represented using the elbow method and the silhouette coefficient, as illustrated in Figure 1. The results indicated that the inflection point of the elbow method stabilized at $k = 4$, which coincided with the optimal silhouette coefficient value (≈ 0.7482). This convergence validated the thematic segmentation obtained and confirmed the internal consistency of the clustering model.

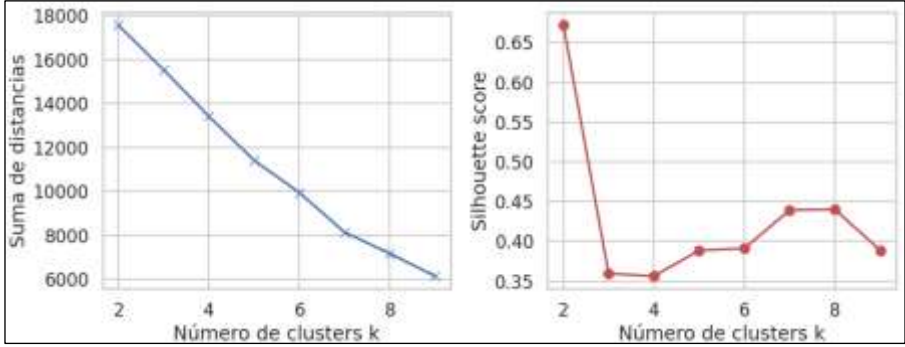


Figure 1: Determination of the optimal number of clusters using the elbow method and silhouette coefficient.
Source: Authors, (2026).

IV.2.5 Classification of Technical and Soft Skills

Semantic classification enabled the clear differentiation between the technical and soft skills most in demand in the Ecuadorian labor market. Skills in Python and R programming, database management, Power BI, and accounting management were identified as highly sought after, along with soft skills such as leadership, teamwork, customer service, and effective communication. The prominence of both technical and soft skills reflected a broader shift observed in digitally mediated labor markets. Digital platforms increasingly prioritized hybrid profiles capable of combining technological proficiency with communication, adaptability, and autonomous problem-solving. This trend was linked to the growing need for workers in platform-based environments to navigate algorithmic interfaces, self-manage tasks, and interact with clients or teams asynchronously.

Therefore, the demand for hybrid competencies was not only an occupational preference but also a structural characteristic of the digital economy, where the boundaries between technical expertise and interpersonal capabilities became progressively blurred. Figure 2 shows the predominance of cross-cutting and operational skills, followed by management and technology skills. These results confirmed the importance of strengthening training programs that integrate both technical and soft components, particularly in sectors undergoing rapid digital transformation, such as technology, commerce, and education.

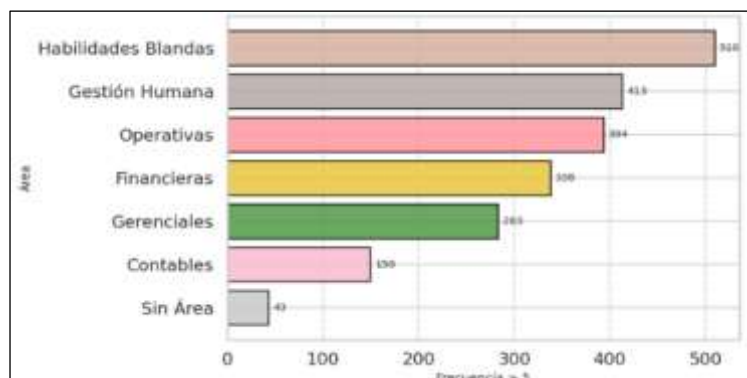


Figure 2: Thematic distribution of standardized competencies extracted from job offers.
Source: Authors, (2026).

IV.2.6 Relative Frequency of Competencies and Territorial Concentration

Frequency analyses generated a word cloud (see Figure 3) that highlighted the most recurrent competencies extracted from job offers. Among the most frequent skills were effective leadership, effective communication, logistics management, financial control, and data analysis, indicating a high demand for multidisciplinary professional profiles.



Figure 3: Word cloud generated from competencies extracted from job offers.
Source: Authors, (2026).

In parallel, the analysis identified a strong geographical concentration of formal employment, as illustrated in Figure 4. Quito (2,328 job offers) and Guayaquil (1,617 job offers) accounted for more than 70% of the total vacancies, followed by Manta (226) and Cuenca (204). The results revealed a pattern of urban concentration accompanied by emerging participation from intermediate cities.

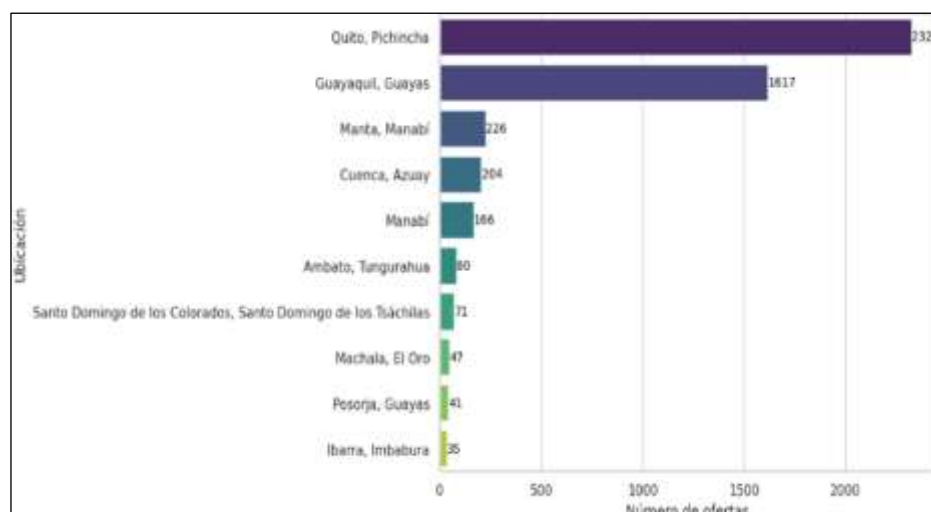


Figure 4: Geographical distribution of job offers in Ecuador.

Source: Authors, (2026).

These geographic disparities reflected structural differences in the spatial distribution of formal employment opportunities. The concentration of vacancies in Quito and Guayaquil was consistent with urban clustering patterns commonly observed in digitally mediated labor markets, where access to connectivity and technological infrastructure was higher than in intermediate or peripheral regions. At the same time, the presence of vacancies in intermediate cities indicated a gradual spatial expansion of formal digital employment.

IV.2.7 Business Concentration in the Labor Market

Figure 5 presented the ten companies with the highest volume of job offers between 2020 and 2024. A high presence of vacancies labeled as Confidential was observed, indicating the widespread use of anonymity practices in digital recruitment processes.

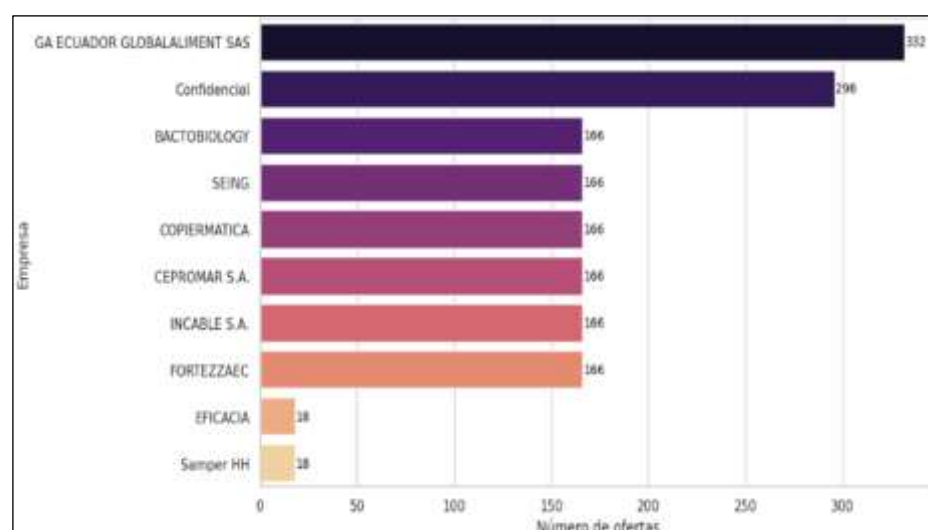


Figure 5: Main companies offering jobs in the Ecuadorian labor market (2020–2024).

Source: Authors, (2026).

The results indicated a high concentration of job offers, in which a limited group of employers generated a substantial proportion of total vacancies. The data showed that large corporations dominated hiring activity during the analyzed period, while medium-sized companies also appeared among the main job offerers, particularly in the logistics and service sectors.

IV.2.8 Model Validation and Comparative Performance

Figures 6 and 7 presented the visual results of the statistical validation performed using confusion matrices and Receiver Operating Characteristic (ROC) curves, respectively. In both analyses, Random Forest exhibited the highest performance, achieving an area under the curve (AUC) value of 0.99, which surpassed the decision tree (AUC = 0.93) and logistic regression (AUC = 0.96).

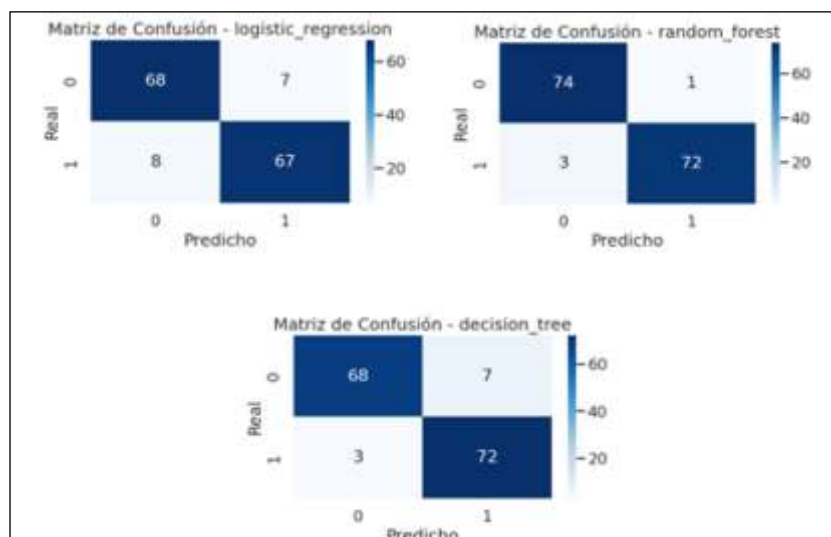


Figure 6: Comparison of confusion matrices between supervised models.
Source: Authors, (2026).

The values shown in Figure 6 indicated that Random Forest achieved 146 correct predictions out of 150 cases, outperforming the decision tree, which recorded 140 correct predictions, and logistic regression, which reached 135 correct predictions in terms of accuracy.

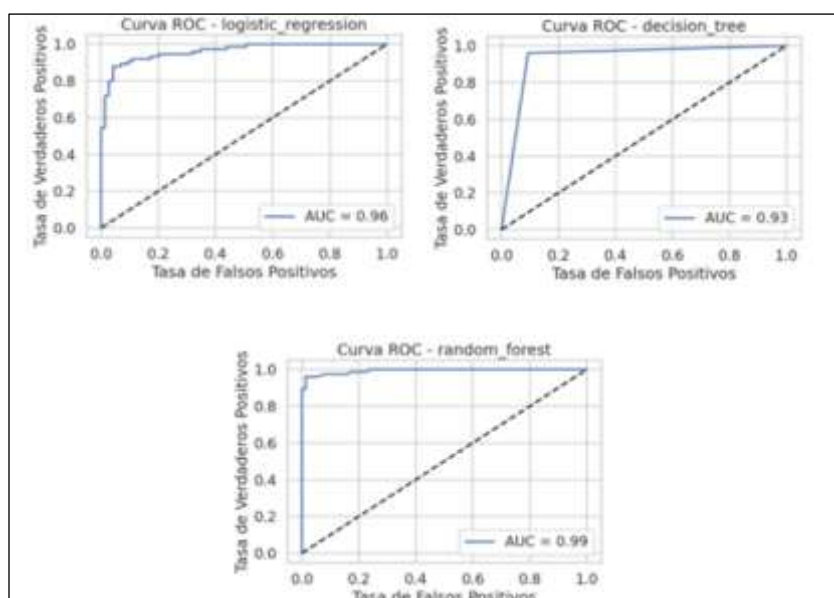


Figure 7: Performance comparison using ROC curves: Logistic Regression, Decision Tree, and Random Forest.
Source: Authors, (2026).

The ROC curves presented in Figure 7 showed consistent performance of the Random Forest model across different classification thresholds, including scenarios characterized by higher semantic variability. These results supported its superior discriminative capacity compared with the other supervised models evaluated.

IV.2.9 Summary of the Comparative Analysis

The results obtained led to the identification of three main findings derived from the comparative analysis. First, the integration of natural language processing with Random Forest provided a favorable balance between accuracy, scalability, and adaptability among the evaluated techniques. Second, the analysis revealed a clear trend toward technical hyper-specialization and the emergence of hybrid professional profiles, accompanied by a strong concentration of job offers in major urban centers. Third, the comparative performance results indicated that hybrid analytical approaches combining semantic analysis with supervised learning achieved higher effectiveness for monitoring professional skill dynamics within the analyzed labor market.

IV.3 DESIGN OF A MODEL FOR DETECTING PROFESSIONAL COMPETENCIES

The analytical model enabled the identification of professional skills in high demand within the Ecuadorian labor market, as well as the hierarchical structure of competencies observed between 2020 and 2024. The results indicated that the integration of structured sources from the National Institute of Statistics and Census (INEC) with unstructured data from digital job portals enhanced the model's predictive performance, achieving accuracy levels above 96%.

IV.3.1 Interactive Visualization and Distribution of Job Offer

The dashboard developed for result visualization integrated geographic, sectoral, and thematic analyses. Figure 8 presented the main interface of the system, which displayed overall dataset metrics, the number of job offers by data source, and dynamic visualizations associated with skills and regions.



Figure 8: Main interface of the dashboard developed in Streamlit.
Source: Authors, (2026).

Figure 8 showed that most job offers originated from private employment portals, with Multitrabajos and BuscoJobs standing out as the sources with the highest volume of vacancies. The indicators revealed a predominance of job offers associated with the commercial, administrative, and technological sectors, together with a clear concentration in the provinces of Pichincha and Guayas.

IV.3.2 Identification of Predominant Skills

The model classified and standardized the competencies identified in thousands of job descriptions, revealing a marked trend toward hybrid skills. Figure 9 presented the main competencies extracted through semantic processing, grouped according to frequency and relevance. The results indicated that the most frequent skills included leadership, customer service, effective communication, logistics management, data analysis, and Python programming, reflecting a convergence between technical and soft competencies within the analyzed dataset.

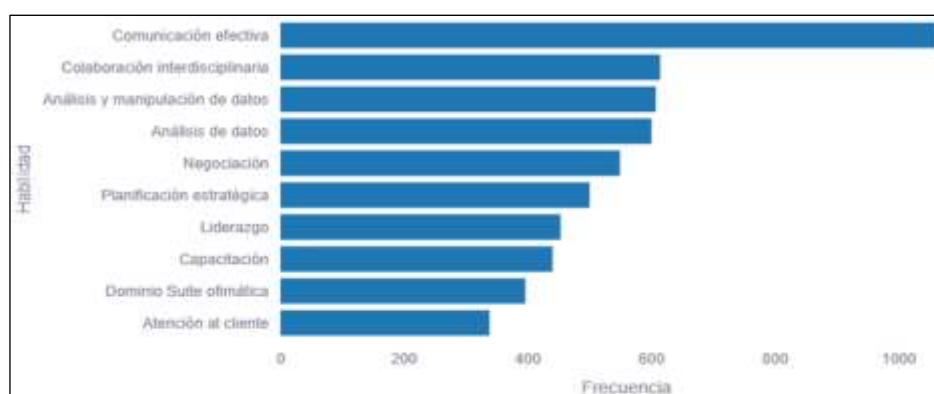


Figure 9: Main skills extracted through semantic processing in Streamlit.
Source: Authors, (2026).

The results showed that job offers across the analyzed sectors frequently required profiles combining technological expertise with management and coordination-related skills. This pattern was observed consistently across all sectors included in the analysis.

IV.3.3 Hierarchical Structure of Competencies

The analysis of semantic hierarchies allowed the representation of the organization of competencies according to levels of generality and frequency. Figure 10 presented the sunburst chart summarizing the relationships between main competencies and sub-competencies derived from the processed textual corpus. The results revealed three dominant hierarchical levels: (i) technological competencies, including data analysis, database management, programming, and Power BI; (ii) management and leadership competencies; and (iii) operational and service-related competencies.

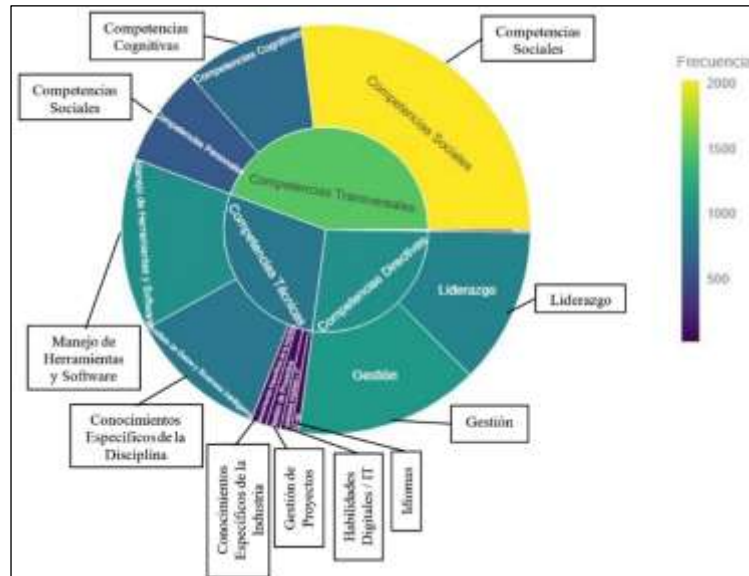


Figure 10: Hierarchy of job skills identified using a sunburst chart.
Source: Authors, (2026).

The technology and commercial sectors exhibited the highest diversity of sub-skills, whereas administrative areas showed lower thematic dispersion. The hierarchical visualization highlighted the interconnection between technical and soft competencies observed in the analyzed data, indicating a structural organization of skills across sectors.

IV.3.4 Model Performance and Consistency of Results

The model's performance showed high accuracy in classifying competencies extracted from unstructured job-related texts. The statistical results indicated an accuracy of 96.67%, a precision of 97.30%, and an F1-score of 96.64%, which validated the robustness of the model under conditions of semantic variability. The consistency observed between the visual outputs and the quantitative performance metrics demonstrated that the model generalized labor demand patterns across different regions of the country.

In addition, the dashboard visualizations enabled the identification of associations between skill types and geographic location. In Quito and Guayaquil, higher demand for technological and administrative profiles was observed, whereas in Manta and Cuenca, operational skills related to logistics and customer service predominated. These spatial patterns were consistently reflected across the visual and quantitative results generated by the system.

IV.3.5 Final Model Architecture

Figure 11 summarized the architecture of the final analytical model, which combined supervised and unsupervised learning techniques within a sequential analysis workflow. The architecture operated through an extraction, transformation, and loading (ETL) process, followed by semantic modeling, statistical validation, and dynamic visualization stages.

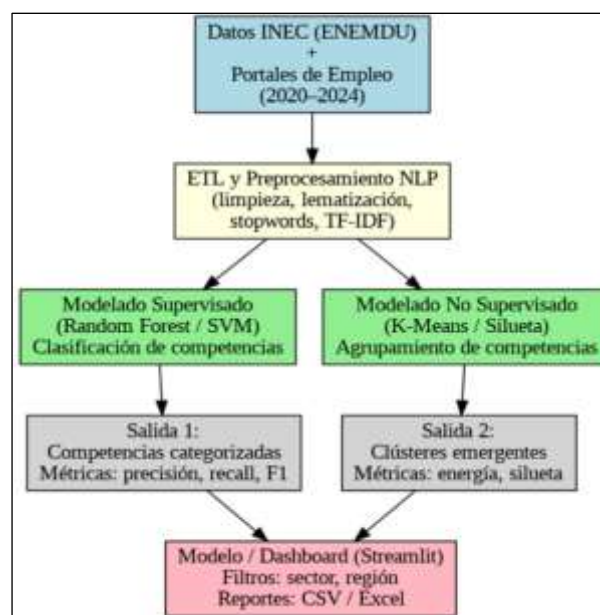


Figure 11: Architecture of the job competency analysis model.
Source: Authors, (2026).

The modular structure allowed algorithms to be updated and new data sources to be incorporated without reconfiguring the overall system architecture, ensuring scalable, transparent, and replicable operation. The resulting configuration supported continuous monitoring of professional competencies based on the analyzed datasets and facilitated the systematic analysis of labor demand patterns within the Ecuadorian context.

IV.4 DISCUSSION

The results of this study demonstrate that applying data science techniques to the analysis of the Ecuadorian labor market is an effective approach for detecting patterns of professional demand and anticipating emerging skills trends. Beyond the technical accuracy achieved—exceeding 96%—the findings show that computational models offer a meaningful way to understand how labor behaviors are increasingly shaped by digital platforms and socio-technical ecosystems. Integrating structured and unstructured sources thus provides not only richer predictive capacity but also insights into how digital employment infrastructures structure access to work opportunities. The hierarchical analysis of competencies revealed a clear trend toward skill hybridization, where technical competencies related to programming, data analytics, and digital management are complemented by soft skills such as leadership, effective communication, and problem solving.

This phenomenon, observed globally in digitally mediated labor markets, reflects the transition toward interdisciplinary profiles capable of operating in environments characterized by automation, virtual collaboration, and accelerated technological cycles. These findings are consistent with research in Latin America [26], [27], which documents a sustained rise in demand for professionals who combine advanced digital abilities with cross-cutting competencies needed for platform-based and remote work environments. Likewise, the results coincide with findings by [28], who emphasize that integrating supervised and unsupervised learning enables deeper and more consistent identification of semantic patterns than traditional statistical methods. This study expands that perspective by incorporating large-scale web scraping and text-mining techniques, enabling the analysis of both formal and informal segments of the labor market.

Such a hybrid methodological approach contributes to addressing a documented gap in the regional literature, revealing how platform-mediated employment and uneven digital adoption reshape skill requirements in heterogeneous work contexts [29]. This broader analytical scope also exposes structural inequalities in skill demand and training gaps, which become more visible when analyzing labor markets through digital traces. The empirical evidence obtained has relevant theoretical and practical implications. Academically, it offers a replicable methodological framework for studies on employability and digital skills, reinforcing the potential of data science for building evidence-based labor observatories. In practical terms, the model provides useful inputs for public policy, curriculum design, and business strategies in human-talent management. When combined with an interactive dashboard, the system enables continuous monitoring of the evolution of labor demand by sector, region, and specialization level.

This capability is crucial for digital-transformation agendas and can support governments in designing e-employment policies, anticipating future skill shortages, and reducing territorial gaps in digital labor inclusion. Nevertheless, some limitations must be acknowledged. Reliance on digital sources may underrepresent sectors characterized by low digitization or high informality, reflecting the broader issue of digital inequality in emerging economies. Moreover, variability in the language of job advertisements can introduce semantic noise, despite the model's strong generalization metrics. Future research could expand temporal coverage, incorporate deep learning architectures to enhance contextual understanding of job descriptions, and integrate socioeconomic, gender, and territorial variables to provide a more comprehensive view of inequality in digital labor markets. Such improvements would strengthen both the predictive capacity and social relevance of the system, enabling its evolution into a full digital labor observatory capable of capturing transformations in work, skills, and employment in digitally evolving societies.

V. CONCLUSIONS

This study demonstrated that the application of data science techniques constitutes an effective approach for analyzing and understanding the Ecuadorian labor market in a context increasingly mediated by digital technologies. By integrating structured sources, such as INEC records, with unstructured data from digital job platforms, the proposed model accurately identified professional competencies, revealed training gaps across productive sectors, and captured patterns of hybrid skill demand combining technical and soft skills. The high performance achieved—exceeding 96% accuracy—confirmed the value of combining supervised and unsupervised learning within a modular and scalable analytical framework supported by natural language processing and interactive visualizations. Although reliance on digital job sources may underrepresent informal or low-digitization sectors and introduce semantic variability, the model showed consistent and robust performance. Overall, this research contributes a replicable methodological approach to digital labor analytics and provides a foundation for evidence-based decision making, workforce planning, and future extensions aimed at monitoring emerging skills and addressing territorial and socio-digital inequalities in evolving labor ecosystems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Methodology: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Investigation: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Discussion of results: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Writing – Original Draft: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Writing – Review and Editing: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Resources: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Supervision: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

Approval of the final text: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

VII. REFERENCES

- [1] L. Bareño Pinzón, P. Orjuela Garzón, and F. Pinzón Galeano, “Análisis de los beneficios de las nuevas tecnologías implementadas al proceso de reclutamiento de personal a partir del Covid19 en grandes empresas colombianas en el 2023,” *Universitaria Agustiniiana*, Bogotá D. C., 2023. Accessed: Aug. 05, 2025. [Online]. Available: <http://repositorio.uniagustiniana.edu.co/handle/123456789/2496>
- [2] A. Taboada Villamarín, “Big data en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático,” *Revista CENTRA de Ciencias Sociales*, vol. 3, no. 1, Jan. 2024, doi: 10.54790/rccs.51.
- [3] F. Da Silva and N. Georgina, “La era de las plataformas digitales y el desarrollo de los mercados de datos en un contexto de libre competencia,” *Documentos de Proyectos (LC/TS.2021/173)* Dec 2021, Santiago, Comisión Económica para América Latina y el Caribe (CEPAL). Accessed: Sep. 08, 2024. [Online]. Available: <https://www.cepal.org/es/publicaciones/47540-la-era-plataformas-digitales-desarrollo-mercados-datos-un-contexto-libre>
- [4] M. Jiang and P. Jia, “Does the level of digitalized service drive the global export of digital service trade? Evidence from global perspective,” *Telematics and Informatics*, vol. 72, p. 101853, Aug. 2022, doi: 10.1016/j.tele.2022.101853.
- [5] A. Alibasic, H. Upadhyay, M. C. E. Simsekler, T. Kurfess, W. L. Woon, and M. A. Omar, “Evaluation of the trends in jobs and skill-sets using data analytics: a case study,” *J Big Data*, vol. 9, no. 1, p. 32, Dec. 2022, doi: 10.1186/s40537-022-00576-5.
- [6] R. Rodríguez, “Brecha digital y transformación social: el impacto de las nuevas tecnologías en América Latina y el Caribe,” *Revistas UPR*, vol. 5, 2024.
- [7] ILO, “Progress in the labour market in Latin America and the Caribbean remains insufficient,” International Labour Organization. Accessed: Jan. 05, 2025. [Online]. Available: https://www.ilo.org/resource/news/insufficient-labour-market-progress-latin-america-caribbean-2024-overview-report?utm_source=chatgpt.com
- [8] IDB, “Using Big Data and its Analytical Techniques for Public Policy Design and Implementation in Latin America and the Caribbean,” Inter-American Development Bank. Accessed: Jan. 09, 2025. [Online]. Available: <https://publications.iadb.org/publications/english/document/Using-Big-Data-and-its-Analytical-Techniques-for-Public-Policy-Design-and-Implementation-in-Latin-America-and-the-Caribbean.pdf>
- [9] M. Villamar Choez, G. Mite Vivero, M. Vera Franco, and G. Figueroa Morán, “Impacto de TIC en el mercado laboral aplicado a egresados, carrera tecnologías de la información,” *Journal TechInnovation*, vol. 3, no. 0, pp. 40–46, Dec. 2024.
- [10] INEC, “Metodología del Registro Estadístico de Empleo en la Seguridad Social (REESS),” Instituto Nacional de Estadística y Censos.
- [11] A. Kouskoura, E. Kalliontzi, D. Skalkos, and I. Bakouros, “Evaluating Experts’ Perceptions on Regional Competitiveness Based on the Ten Key Factors of Assessment,” *Sustainability*, vol. 16, no. 14, p. 5944, Jul. 2024, doi: 10.3390/su16145944.
- [12] L. Echeverría and G. Rucci, “¿Qué suma la ciencia de datos a la identificación y anticipación de la demanda de habilidades?,” Washington, D. C., Nov. 2022. doi: 10.18235/0004554.
- [13] O. Grljević and T. Kecojević, “An NLP approach to skill analysis in ICT job advertisements from a gender perspective,” *Strategic Management*, no. 00, pp. 94–94, 2025, doi: 10.5937/StraMan2500008G.
- [14] V. Escudero, H. Liepman, and A. Podjanin, *Big Data Applications in Labor Economics, Part B*. Emerald Publishing Limited, 2024. doi: 10.1108/S0147-9121202552B.
- [15] O. Kassi and V. Lehdonvirta, “Online labour index: Measuring the online gig economy for policy and research,” *Technol Forecast Soc Change*, vol. 137, pp. 241–248, Dec. 2018, doi: 10.1016/j.techfore.2018.07.056.
- [16] C. Muntaner, “Digital Platforms, Gig Economy, Precarious Employment, and the Invisible Hand of Social Class,” *International Journal of Health Services*, vol. 48, no. 4, pp. 597–600, Oct. 2018, doi: 10.1177/0020731418801413.
- [17] K. M. Kuhn, J. Meijerink, and A. Keegan, “Human Resource Management and the Gig Economy: Challenges and Opportunities at the Intersection between Organizational HR Decision-Makers and Digital Labor Platforms,” 2021, pp. 1–46. doi: 10.1108/S0742-730120210000039001.
- [18] D. Autor and A. Salomons, “Is automation labor displacing? Productivity growth and the labor share,” *Brookings Pap Econ Act*, vol. 53, no. 2, pp. 1–58, 2018, Accessed: Mar. 07, 2025. [Online]. Available: https://www.brookings.edu/wp-content/uploads/2018/03/1_autorsalomons.pdf
- [19] A. Henry and R. Essah, “Artificial Intelligence and Workforce Transformation: An Analysis of Employment Disruption and Adaptation Strategies,” *International Journal of Latest Technology in Engineering Management & Applied Science*, vol. 14, no. 8, pp. 493–501, Sep. 2025, doi: 10.51583/IJLTEMAS.2025.1408000059.
- [20] H. Kleven, C. Landais, and J. E. Sogaard, “Children and Gender Inequality: Evidence from Denmark,” *Am Econ J Appl Econ*, vol. 11, no. 4, pp. 181–209, Oct. 2019, doi: 10.1257/app.20180010.
- [21] E. Struffolino and M. Raitano, “Early-Career Complexity Before and After Labour-Market Deregulation in Italy: Heterogeneity by Gender and Socio-economic Status Across Cohorts,” *Soc Indic Res*, vol. 151, no. 1, pp. 231–257, Aug. 2020, doi: 10.1007/s11205-020-02373-0.
- [22] J. A. Toledo Íñiguez and D. A. Angulo Vélez, “Automatización de procesos administrativos en instituciones públicas en Ecuador,” *Revista Científica Multidisciplinar G-nerando*, vol. 6, no. 1, Mar. 2025, doi: 10.60100/rcmg.v6i1.547.
- [23] M. G. Mayor Ravines, J. G. Mayor Gamero, and J. V. Mayuri Barrón, “Transformación digital en la administración pública: una revisión sistemática de marcos, desafíos y resultados (2013–2023),” *Nexus Research Journal*, vol. 4, no. 2, pp. 35–63, Aug. 2025, doi: 10.62943/nrj.v4n2.2025.325.
- [24] A. Aristovnik, D. Ravšelj, and E. Murko, “Decoding the Digital Landscape: An Empirically Validated Model for Assessing Digitalisation across Public Administration Levels,” *Adm Sci*, vol. 14, no. 3, p. 41, Feb. 2024, doi: 10.3390/admsci14030041.

- [25] T. Bielialov, I. Kalina, V. Goi, O. Kravchenko, N. Shyshpanova, and A. Negoda, "Global Experience of Digitalization of Economic Processes in the Context of Transformation," *Journal of Law and Sustainable Development*, vol. 11, no. 3, p. e814, Aug. 2023, doi: 10.55908/sdgs.v11i3.814.
- [26] A. Zuñiga-Collazos, J. M. Velásquez Orozco, and A. Rojas-Ospina, "Digital Competences and Their Impact on Employability in the Tourism Sector—An Applied Study," *Sustainability*, vol. 17, no. 13, p. 6133, Jul. 2025, doi: 10.3390/su17136133.
- [27] C. Buenadicha Sánchez *et al.*, "21st Century Skills: Transversal Skills Development in Latin America and the Caribbean," Washington, D. C., Mar. 2019. doi: 10.18235/0001574.
- [28] C. Quan, M. Wang, and F. Ren, "An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature," *PLoS One*, vol. 9, no. 7, p. e102039, Jul. 2014, doi: 10.1371/journal.pone.0102039.
- [29] J. Schedlbauer, G. Raptis, and B. Ludwig, "Medical informatics labor market analysis using web crawling, web scraping, and text mining," *Int J Med Inform*, vol. 150, p. 104453, Jun. 2021, doi: 10.1016/j.ijmedinf.2021.104453.