https://doi.org/10.12662/2359-618xregea.v15i1.6206.pe6206.2026

ReGeA
Revista Gestão em Análise
Journal of Management and Technologies

ARTIGOS

EMPLOYEE SEGMENTATION USING CLUSTERING TECHNIQUES: A CASE STUDY AT APEXBRASIL

SEGMENTAÇÃO DE EMPREGADOS POR TÉCNICAS DE CLUSTERIZAÇÃO: UM ESTUDO DE CASO NA APEXBRASIL

César Antônio Ciuffo Moreira
cesarciuffo@hotmail.com
HR Management Advisor and senior analyst at Apex-Brasil (Brazilian Trade and Investment Promotion Agency). PhD candidate in Applied Computing at the University of Brasília. He holds an M.Sc. in Data Science/Applied Computing from the University of Brasília (UnB, 2025), a B.A. in Business Administration (UniCEUB, 2001). Brasília, DF, Brasil.

RESUMO

Este estudo investiga como técnicas de machine learning não supervisionado podem apoiar People Analytics em organizações públicas brasileiras, contexto marcado por baixa maturidade analítica. Utilizando dados anonimizados de empregados ativos da ApexBrasil entre janeiro de 2019 e dezembro de 2023, o trabalho segue o framework CRISP-DM para comparar três métodos de clusterização: K-means, Clustering Hierárquico e DBSCAN. Embora o DBSCAN tenha apresentado maiores índices de silhueta, ele classificou grande parte dos registros como ruído, limitando sua utilidade organizacional. Assim, o K-means com oito grupos foi selecionado como melhor equilíbrio entre qualidade técnica e cobertura amostral. Os agrupamentos resultantes foram interpretados como perfis distintos de força de trabalho, fornecendo uma base interpretável para investigações futuras sobre perfis de força de trabalho em organizações públicas. Os resultados demonstram a viabilidade de implementação de People Analytics em contextos de dados administrativos limitados, oferecendo framework replicável para organizações públicas similares.

Palavras-chave: análise de pessoas; clusterização; segmentação de empregados; K-Means; DBSCAN; Clustering Hierárquico; ciência de dados em RH.

ABSTRACT

This study investigates how unsupervised machine learning techniques can support People Analytics in Brazilian public organizations, a context characterized by low analytical maturity. Using anonymized data from ApexBrasil's active employees between January 2019 and December 2023, the work follows the CRISP-DM framework to compare three clustering methods: K-means, Hierarchical Clustering, and DBSCAN. Although

DBSCAN presented higher silhouette indices, it classified a large portion of records as noise, limiting its organizational utility. Thus, K-means with eight clusters was selected as the best balance between technical quality and sample coverage. The resulting clusters were interpreted as distinct workforce profiles, providing an interpretable basis for future People Analytics investigations and workforce heterogeneity assessment. The results demonstrate the feasibility of implementing People Analytics in contexts with limited administrative data, offering a replicable framework for similar public organizations.

Keywords: people analytics; clustering; employee segmentation; K-Means; DBSCAN; Hierarchical Clustering; human resources data science.

1 INTRODUCTION

People Analytics refers to the application of data science to human resource processes and has emerged as a strategic capability for organizations navigating rapid socio-economic transformations (Ferrar; Green, 2021; Sharda; Delen; Turban, 2018). By leveraging analytical techniques to understand workforce dynamics and predict behaviors, organizations can balance individual well-being with productivity objectives, a critical challenge in contemporary people management (Polzer, 2022). People Analytics enables data-driven decision-making and reshaping how organizations manage talent and performance (Margherita, 2022).

Employee segmentation through clustering techniques represents a particularly relevant application, enabling data-driven policies aligned with diverse workforce needs (Edwards; Edwards, 2019). However, implementation in Brazilian public sector organizations faces specific challenges: limited analytical maturity, fragmented data infrastructure, and ethical constraints regarding employee data usage (Tursunbayeva et al., 2022). This context motivates our investigation of clustering technique selection for People Analytics in public organizations.

When applying segmentation techniques, it is essential to evaluate their advantages, disadvantages, and optimal use cases, including methods such as K-Means, Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), as seen in the references. A key challenge lies in extracting and analyzing data from multiple HR systems, which requires a deep understanding of each system's context and the diversity of available attributes.

Brazilian public organizations face specific challenges in implementing People Analytics, including:

a) low analytical maturity and limited technological infrastructure;
b) fragmented administrative data not originally collected for analytical purposes and;
c) absence of validated frameworks for selecting clustering techniques appropriate to the public sector context.

Given this scenario, the central research question guiding this study is: "**How can unsupervised clustering techniques be effectively applied for employee segmentation in Brazilian public organizations, considering the constraints of analytical maturity and technological infrastructure?**"

This question unfolds into three specific research questions:

a) which clustering techniques are most suitable for HR data in the Brazilian public sector?
b) how can the technical quality of clusters be balanced with the managerial utility of results?
c) what workforce profiles emerge from segmentation, and what implications do they have for people management?

ApexBrasil (Brazilian Trade and Investment Promotion Agency) is a federal public organization linked to the Ministry

of Development, Industry, Commerce, and Services (MDIC), operating in the commercial promotion and foreign investment attraction segments for Brazil.

Relevant organizational characteristics: Autonomous social service (Legal status); Approximately 335 employees (Workforce); Brazilian Labor Law (CLT), not civil service (Management model); Integrated TOTVS ERP (HR system); Initial stage of analytics adoption (Analytical maturity).

ApexBrasil was selected as the case study based on analytical relevance and practical feasibility. The organization provides consistent and anonymizable administrative records from 2019 to 2023, including demographic, temporal, and compensation attributes, enabling the reliable application of clustering techniques under LGPD compliance. The study is further aligned with an ongoing people management modernization initiative within the broader context of public sector digital transformation.

The Agency represents a typical Brazilian public organization operating under the CLT regime, characterized by low analytical maturity and ERP-based administrative data—conditions common to many public entities. These characteristics enhance the contextual representativeness and replicability of the study. Formal authorization for anonymized data use was granted under a confidentiality agreement, ensuring ethical compliance. While the focus on a single organization enables in-depth methodological analysis, it inherently limits generalizability, a limitation explicitly acknowledged in this study.

General Objective: to evaluate and compare unsupervised clustering techniques (K-Means, Hierarchical Clustering, and DBSCAN) for implementing People Analytics at ApexBrasil, aiming to enhance organizational maturity and operational effectiveness in people management.

Specific Objectives:

a) conduct exploratory data analysis of ApexBrasil's HR data, identifying workforce characteristics and quality of available data;
b) implement three distinct clustering techniques (K-Means, Hierarchical Clustering, and DBSCAN) using the same demographic, temporal, and compensation attributes;
c) evaluate the quality of obtained clusters through appropriate metrics (Silhouette Index) and organizational utility criteria (sample coverage);
d) identify and characterize distinct workforce profiles with managerial relevance, considering professional life cycle, gender, and absenteeism patterns;
e) discuss implications of identified profiles for people management policies, connecting findings with HR and public administration literature.

This study contributes to People Analytics and Public Administration literature in three dimensions:

a) theoretical: demonstrates the applicability of machine learning techniques in contexts with limited administrative data, validating the CRISP-DM framework for the Brazilian public sector and evidencing the gap between technical capability and organizational maturity;
b) methodological: provides a replicable protocol for comparing clustering techniques, establishing selection criteria that balance technical quality (Silhouette Index) with organizational utility (sample coverage);
c) practical: identifies actionable workforce profiles and offers a roadmap for implementing People Analytics in organizations with low analytical maturity.

The remainder of this article is organized as follows: Section 2 presents the theoretical framework on clustering and People Analytics;

Section 3 details the applied CRISP-DM methodology; Section 4 presents and discusses results, connecting them with literature; and Section 5 concludes with contributions, limitations, and directions for future research.

2 THEORETICAL FRAMEWORK

Clustering techniques are essential for identifying hidden patterns and structures in data. For any method, there is a structure of advantages, disadvantages, recommended uses, and success evaluation criteria (Gabriel Filho, 2023). Clustering is an unsupervised machine learning technique used to group data into subsets, where the data in each subset (or cluster) are more similar to each other than to the data in other clusters (Rokach; Maimon, 2005).

It is used for customer segmentation, where usage behaviors can be mapped to create actions that are focused on customer needs. HR data has been used to map employee profiles and identify common characteristics, facilitating the segmentation of social groups and the possibility of personalized responses (Edwards; Edwards, 2019).

2.1 JUSTIFICATION FOR CLUSTERING TECHNIQUES (UNSUPERVISED LEARNING)

This study adopts clustering techniques, a form of unsupervised learning, in contrast to classification techniques (supervised learning). This methodological choice is based on four rationales:

2.1.1 Absence of Predefined Taxonomy

Unlike contexts where employee categories are known *a priori* (e.g., "high performance," "turnover risk"), ApexBrasil has no established taxonomy of workforce profiles.

Therefore,

a) classification (supervised) would require training labels previously defined by experts or prior systems, which do not exist;
b) clustering (unsupervised) allows exploratory discovery of natural groups in data without assuming pre-existing categories.

As stated by Rokach and Maimon (2005, p. 322): "Clustering is used when class labels are unknown, and one seeks to partition the data into homogeneous groups".

2.1.2 Exploratory vs. Predictive Objective

This study's objective is exploratory to identify latent patterns and structures in the workforce. We do not seek to predict membership in known categories, but rather to discover these categories:

a) classification is appropriate when the objective is to predict: "given a new employee X, does s/he belong to category A or B?";
b) clustering is appropriate when the objective is to discover: "what natural groups exist in the employee population?". This alignment with exploratory purposes is recognized in People Analytics literature (Garg *et al*., 2022; Lismont *et al*., 2017).

2.1.3 Low Analytical Maturity Context

Organizations in initial stages of People Analytics adoption frequently lack:

a) a history of previous segmentations that could serve as labels;
b) experts with consolidated knowledge about "ideal" profiles;
c) performance evaluation systems that reliably categorize employees. In this context, clustering offers flexibility for the emergence of unanticipated groups, suitable for organizations exploring People Analytics for the first time (Waters *et al*., 2018).

2.1.4 Alignment with People Analytics Literature

Systematic review by Garg *et al*. (2022) on machine learning in HR indicates that clustering techniques dominate employee segmentation applications, citing:

a) Andriyani and Puspitarani (2022): Comparison of K-Means and DBSCAN for performance segmentation;
b) Kakulapati *et al*. (2020): Use of clustering for identifying turnover risk profiles. This study aligns with this established methodological tradition.

2.1.5 Possibility of Future Extension with Classification

Importantly, clustering and classification are not mutually exclusive but sequential:

a) PHASE 1 (this study): Clustering identifies profiles (e.g., 8 clusters);
b) PHASE 2 (future studies): Classification uses profiles as labels to predict membership of new employees.

Thus, this study establishes a foundation for future supervised classification applications once profiles are validated and stabilized.

In summary, the choice of clustering reflects the exploratory nature of the study, the absence of predefined taxonomies, low analytical maturity of the context, and alignment with People Analytics best practices.

2.2 K-MEANS

K-means is a clustering algorithm with a non-hierarchical agglomeration scheme that partitions data into *k* clusters, thereby minimizing the variance within each cluster (Lloyd, 1982).

$$\min_{\{\mu_i\}_{i=1}^{k}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} ||x_j^{(i)} - \mu_i||^2$$

where:

a) k = number of clusters;
b) ni = number of points in cluster i;
c) xj^(i) = j-th point in cluster i;
d) μi = centroid of cluster i;
e) ||·|| = Euclidean norm (distance).

Relevance Analysis (Jain, 2010; Lloyd, 1982; Macqueen, 1967): K-means partitions observations into k clusters by minimizing within-cluster variance, using Euclidean distance to assign points to the nearest centroid (Lloyd, 1982). In this study, K-means was evaluated using the mean Silhouette Index and sample coverage, prioritizing interpretability and applicability in a low analytical maturity context.

2.3 HIERARCHICAL CLUSTERING

Hierarchical Clustering creates a tree of clusters called a dendrogram, which facilitates the visualization of relationships between different groups of data (Murtagh; Legendre, 2014). It creates a hierarchy of clusters by grouping data in an agglomerative or divisive manner. It does not require a prior definition of the number of clusters and demonstrates the hierarchy of clusters. It allows for the exploration of different clustering granularities. It can be computationally expensive for large datasets, and the interpretation of the hierarchy can be complex.

Relevance Analysis (Murtagh; Legendre, 2014; Rokach; Maimon, 2005; Xu; Wunsch II, 2005): Hierarchical Clustering builds a dendrogram that represents nested groupings, enabling exploration of different segmentation granularities without requiring a fixed number of clusters a priori (Murtagh; Legendre, 2014). For comparison purposes, this study evaluated solutions with k ∈ {5, 7, 8} using the mean Silhouette Index and interpretability criteria.

2.4 DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based method that identifies clusters in arbitrary shapes and detects outliers (Ester; Kriegel; Sander; Xu, 1996). DBSCAN identifies clusters based on the point density, identifies dense areas of data, and ignores sparse points (noise). It strongly depends on the parameters of the two points to be considered in the same cluster, ϵ, and the minimum number of points required to form a cluster (MinPts). Clusters with fewer than MinPts points were considered noise.

Points belong to the same point if they are connected by a chain of points, where each point is within a distance ϵ of the previous point. The neighborhood of point p is defined as

$$N_{\epsilon}(p) = \{q \in D \mid \|p - q\| \leq \epsilon\}.$$

Relevance Analysis (Campello; Moulavi; Sander, 2013; Ester; Kriegel; Sander; Xu, 1996; Schubert *et al*., 2017): DBSCAN identifies clusters as dense regions separated by sparse areas, allowing arbitrary cluster shapes and the detection of outliers. In this study, DBSCAN was assessed through mean Silhouette Index (excluding noise points) and, critically, sample coverage, since employee segmentation requires organizationally useful grouping of most observations.

2.5 RELATED WORKS

Garg *et al*. (2022) presented a literature review identifying contracts related to human resources and machine learning objectives. In this study, it was observed that clustering algorithms are present in a large part of the process of identifying trends and patterns in Human Resource data. This review suggests that machine learning applications are the strongest in recruitment and performance management. Applications for complex processes are still in their early stages of development.

Andriyani and Puspitarani (2022) proposed a model for evaluating and comparing the results of different clustering algorithms, highlighting the applicability of different techniques, and defining the best application. In this study, the K-means and DBSCAN methods were tested.

In another study, Gupta and Sharma (2022) stated that from the perspective of human resources, the most important duty of a large company is to ensure that its employees have a balance between personal and professional life, maintain its brand image, and require the use of algorithms that can segment employees.

Kakulapati *et al*. (2020) analyzed employee performance by applying clustering techniques based on the similarity of the performance metrics. In addition, classification techniques have been used, such as Random Forest, which facilitates the classification of employees based on their monthly income and informal methods of data analysis.

In a proposal for an HR metric maturity analysis, Lismont *et al*. (2017) conducted descriptive research with several organizations that use People Analytics, and found a prevalence of understandable ones, such as linear regression and decision trees, regression techniques, survival analysis, etc. The most commonly used techniques in the universe research are logistic regression, decision trees, linear regression, time-series prediction, hierarchical clustering, and K-means.

Prior studies apply clustering in HR contexts (Andriyani; Puspitarani, 2022; Garg *et al*., 2022), but most focus on technical performance without explicit discussion of organizational utility in low-maturity settings. This study extends the literature by explicitly balancing cluster quality (Silhouette Index) and practical coverage, highlighting the trade-off observed with DBSCAN in administrative HR datasets.

3 METHODOLOGY

This study followed the CRISP-DM framework (Chapman, 2000), structured into: (1) Business Understanding (definition of HR segmentation needs and case context); (2) Data Understanding (ERP data assessment and exploratory analysis); (3) Data Preparation (ETL, anonymization, cleaning, feature engineering, and normalization); (4) Modeling (K-means, Hierarchical Clustering, and DBSCAN); and (5) Evaluation (Silhouette Index and sample coverage) to support technique selection.

3.1 CLUSTER EVALUATION METRICS

Cluster quality was evaluated using the Silhouette Index, a metric widely used in unsupervised clustering studies (Rousseeuw, 1987).

3.1.1 Silhouette index

The Silhouette Index measures both intra-cluster cohesion (how similar elements within the same cluster are) and inter-cluster separation (how distinct clusters are from each other). For each observation i, the index is calculated as:

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

Where:

a) a(i) = average dissimilarity of i to all other points in the same cluster;
b) b(i) = minimum average dissimilarity of i to points in any other cluster;
c) The Silhouette Index ranges from -1 to +1, with interpretation;
d) $s(i) \approx +1$: Well-allocated observation, close to its cluster and distant from others;
e) $s(i) \approx 0$: Borderline observation, could belong to neighboring cluster;
f) $s(i) < 0$: Possibly misallocated observation.

For global evaluation, we used the average of individual indices (mean Silhouette Score) as a comparison criterion between different techniques and k values.

3.1.2 Complementary Metrics for DBSCAN

Given that DBSCAN does not require prior specification of k and automatically identifies clusters of arbitrary shape, additional metrics were used:

a) number of identified clusters: Quantity of distinct groupings detected by the algorithm;
b) number of points classified as noise: Observations not assigned to any cluster (outliers);
c) sample coverage: Percentage of observations assigned to clusters (not classified as noise);
d) silhouette score (only for non-noise points): Calculated excluding observations marked as outliers.

3.1.3 Technique Selection Criterion

Final technique selection considered two combined criteria: (i) Technical quality: Maximization of mean Silhouette Score; (ii) Organizational utility: Maximization of sample coverage (minimization of unclassified observations). This dual criterion reflects inherent tension in practical clustering applications: algorithms that maximize cluster purity (such as DBSCAN) may do so at the cost of classifying many observations as noise, limiting managerial utility. Thus, we prioritized balance between technical quality and practical applicability.

4 DATA ANALYSIS

4.1 BUSINESS UNDERSTANDING OF HUMAN RESOURCES PROCESSES AT APEXBRASIL

The scope of this study was defined through meetings with Human Resources (HR) and Information Technology (IT) teams in

ApexBrasil. It was agreed that employee data from the period between January 2019 and December 2023 should be analyzed, considering only employees who were still active as of December 31, 2023. It will provide valuable insights to support workforce understanding, analytical prioritization, and future People Analytics investigations (Edwards; Edwards, 2019; Waters *et al*., 2018).

4.2 PRE-PROCESSING AND EXPLORATORY ANALYSIS OF DATA

In this step, data were extracted, transformed, and loaded (ETL). After loading the data into RStudio, Exploratory Data Analysis (EDA) was performed, and the preliminary results are presented below.

4.2.1 Sample selection and limitations

4.2.1.1 Inclusion Criteria

This study used data from employees with active employment status at ApexBrasil during the period from January 2019 to December 2023. The choice of active employees (excluding inactive, on-leave, and terminated employees) is justified by three reasons:

a) managerial focus: HR policies derived from segmentation apply primarily to the organization's current workforce. Terminated employees, although relevant for turnover analyses, are not targets of prospective people management interventions;
b) temporal consistency: Employees active throughout the analyzed period (2019-2023) have complete and consistent records for all variables of interest, minimizing structural missing data problems;
c) data quality: Because ERP systems mainly focus on capturing information about current employees, data related to absenteeism and professional development for former staff members is often incomplete.

4.2.1.2 Record Exclusion Criteria

The following exclusion criteria were applied:

a) data incompleteness: Records with more than 30% missing values in essential variables (age, gender, hire date, base salary) were excluded;
b) probationary period: Employees with less than 90 days of employment as of December 31, 2023, were excluded, as they lack sufficient history for characterizing absenteeism and development patterns;
c) inconsistencies: Duplicate records (identified via unique ERP system ID) and records with logical inconsistencies (e.g., hire date after reference date) were removed;
d) privacy: Records of individuals in unique or rare positions ($n<3$ in a given category) were excluded to prevent indirect identification, in accordance with LGPD.

4.2.1.3 Limitations Associated with Administrative Data Use

It is essential to recognize inherent limitations of using administrative data from ERP systems:

a) operational vs. analytical purpose: Data were collected primarily for payroll, time-tracking, and benefits management purposes, not for People Analytics. This results in: (i) Absence of behavioral variables (engagement, satisfaction); (ii) Absence of performance variables (formal evaluations); (iii) Absence of competency and professional aspiration variables;

b) selection bias: Sample is restricted to employees who remained at the organization throughout the 2019-2023 period, excluding those who joined or left during this interval. This may bias results in more stable profiles being less prone to turnover;
c) recording quality: Dependence on the accuracy of manual entries (e.g., absenteeism) and system integrations (e.g., promotions), which may present errors or delays;
d) limited granularity: Certain variables (such as total compensation) aggregate multiple components (base salary + benefits + bonuses), hindering more refined analyses of compensation structure;
d) ethics and bias: Although demographic data is essential for identifying equity, their use in clustering may perpetuate biases if not critically interpreted. This study adopts a posture of ethical vigilance, discussing gender and age implications in results.

These limitations do not invalidate the analysis but contextualize its interpretation and reinforce the need for triangulation with other data sources in future studies.

4.2.2 Data Extraction

Using the report model of the Totvs RM Integrated System - People Management, queries were performed in the database, with the extraction of the data sample, as described in the previous section. One file was generated, with 58 attributes and 336 samples, containing the main information of active employees, including: personal data (treated with privacy and removed in the transformation phase), absenteeism, and salary data. After removing the sensitive data and attributes that were not relevant to the analysis, 12 attributes and 335 lines remained.

4.2.3 Data Transformation

During the data transformation phase, new columns were generated in the employee information database to facilitate the analysis. The transformed variables with their respective units of measurement:

a) **age**: converted to age group (in years);
b) **dependents**: total number of dependents (in units);
c) **hired time**: hired time in the organization (in years);
d) **monthly salary**: monthly salary (in Brazilian Reals);
e) **total hours of absence**: total of absenteeism (in hours);
f) **gender**: gender (Male/Female).

Note: The "Gender" variable in the TOTVS ERP system was recorded with codes "M" (Male) and "F" (Female). We applied binary encoding:

Gender_Numeric = ifelse(Gender == "F", 1, 0)
Result: 0 = Male, 1 = Female

Although the ERP system contains only two categories, we recognize that gender is a spectrum and not binary. This data limitation reflects the reality of historical administrative records (2019-2023), when non-binary categories were not captured.

Numeric Variable Normalization: All continuous numeric variables (Age, Hired Time, Monthly Salary, Absenteeism Hours) were normalized via Z-score:

$$z = (x - \mu) / \sigma$$

Where x = original value, μ = sample mean, and σ = sample standard deviation. Distance-based clustering algorithms (K-Means, Hierarchical, DBSCAN) are sensitive to scales. Without normalization, variables with greater variance (e.g., Monthly Salary) would dominate distance calculations.

4.2.4 Data Loading

The **R language** was used to load and parallelize information processing using Dplyr libraries (Alcoforado, 2021) in **RStudio**. The final dataset was exported as a CSV file (UTF-8 encoding) and analyzed using R version 4.5.1 and Packages (cluster, factoextra, dbscan, dendextend, tidyverse).

4.2.4.1 Analysis of Numerical Variables

Table 1 presents descriptive statistics of the numerical variables in the dataset with 335 observations.

Table 1 – Descriptive Statistics of Numerical Variables

Statistics	Age	Dependents	Hired Time	Monthly Salary	Absenteeism
No. of Observations	335	335	335	335	335
Mean	42.8	0.8	8.5	19451.7	231.9
Standard Deviation	10.0	1.3	5.7	11314.0	301.5
Median	42.0	0.0	8.0	17804.0	159.9
Minimum Value	25.0	0.0	0.0	3259.9	0.0
Maximum Value	79.0	6.0	23.0	71740.1	2485.6
Standard Error	0.5	0.1	0.3	618.1	16.5

Source: elaborated by the author.

5 MODELING

Data modeling for clustering offers a variety of tools and techniques that can be applied to solve different business problems. Choosing an appropriate method allows companies to make informed decisions, improve their operational efficiency, and increase responsiveness to market changes. In this study, we tested three techniques to determine the most appropriate for the business context of ApexBrasil.

5.1 DETERMINING THE NUMBER OF CLUSTERS

For techniques requiring prior specification of cluster number (K-Means and Hierarchical Clustering), we tested $k \in \{5, 7, 8\}$ based on multiple criteria:

5.1.1 Elbow Method

Graphical analysis of inertia (sum of squared intra-cluster distances) as a function of k. The "elbow" indicates an inflection point where adding clusters results in diminishing marginal gains. In our data, the elbow was identified between k=7 and k=9.

5.1.2 Dendrogram Analysis

For Hierarchical Clustering, visual dendrogram analysis indicated 3 natural cuts corresponding to k={5, 7, 8}. Cut at k=5 results in very heterogeneous groups; cut at k=8 produces greater intra-cluster homogeneity.

5.1.3 Managerial Interpretability

From a people management perspective, an excessive number of clusters (k>10) hinders the development of differentiated policies for each segment. When consulted, ApexBrasil's HR leadership indicated a preference for 5-8 groups as a manageable quantity for action planning.

5.1.4 Literature Reference

Similar studies of employee segmentation via clustering report frequent use of k between 4 and 10 groups (Garg *et al*., 2022; Kakulapati *et al*., 2020). Our tested range (5-8) aligns with these established practices.

5.1.5 Approach for DBSCAN

DBSCAN does not require prior k specification, automatically determining cluster number based on epsilon (ε) and minPoints parameters. For this study:

a) Epsilon (ε): Epsilon was tested with values between 0.3 and 1.5 (normalized), selecting ε that maximized the Silhouette Score of non-noise points;
b) minPoints: MinPts ∈ {4,5,6} were tested to evaluate the trade-off between cluster quality and workforce coverage, following common practical configurations reported in the DBSCAN literature and aiming to balance cluster quality with workforce coverage in an administrative HR dataset;
c) DBSCAN automatically identified 7 clusters with optimal parameters.

5.2 OBJECTIVE EVALUATION METRICS

Table 2 calculates the Silhouette Index for each clustering method.

Table 2 – Silhouette Index by Clustering Technique and Number of Clusters (K)

Technique	**K = 5**	**K = 7**	**K = 8**
K-means	0,30	0,32	0,34
Hierarchical	0,28	0,30	0,31
DBSCAN (ε = 0,3; MinPts = 5)	0,62	0,63	0,37

Source: elaborated by the author.

The Silhouette Index measures the quality of the formed clusters, with values ranging from -1 to 1 (Devore, 2021). Values close to 1 indicate that the points are well grouped and the clusters are clearly separated, which represents good internal cohesion and distinction between the clusters.

The DBSCAN technique (Ester; Kriegel; Sander; Xu, 1996) is particularly effective for data with arbitrary cluster shapes and noise (outliers), which may explain its superior performance. Good indices suggest that the structure of the data fits the density better than the spherical partition assumed by K-means.

Although DBSCAN achieved higher mean Silhouette Index values, its practical applicability was constrained by the substantial proportion of observations classified as noise. Therefore, DBSCAN results were further evaluated considering the trade-off between cluster quality and sample coverage, which is essential for employee segmentation in organizational settings.

DBSCAN does not directly use the K parameter: the tests were aligned with different values of ε to maintain comparability. If a dataset contains outliers or clusters of varying sizes/deviations, the advantage of DBSCAN tends to increase. By crossing ϵ with the respective Silhouette Indices and measuring the noise intensity, we obtained the following results in Table 3.

Table 3 – Results of the Silhouette Index Analysis with DBSCAN

ε	MinPts	Silhueta	Noise	% Noise
0,20	4	0,92	325	97,0
0,25	5	0,86	318	94,9
0,25	4	0,82	307	91,6
0,40	6	0,72	266	79,4
0,35	5	0,68	275	82,1
0,30	4	0,66	274	81,8
0,40	5	0,63	256	76,4
0,30	5	0,62	290	86,6
0,30	6	0,58	297	88,7
0,45	6	0,58	241	71,9

Source: elaborated by the author.

Given that employee segmentation requires high coverage of the workforce, DBSCAN was not selected despite its higher silhouette scores. K-means (k=8) provided the best balance between technical quality and full coverage, supporting practical interpretability and organizational utility.

The highest silhouette values were obtained using more restrictive parameters. For example, with $\varepsilon = 0.20$ and MinPts= 4, the silhouette index reached 0.923, indicating excellent separation and cohesion between the clusters. However, this model classified 97% of the observations as noise and clustered only a small fraction of the data. As the ϵ value increased, there was a gradual reduction in the silhouette index, indicating a lower cluster quality.

The best compromise between cluster quality and model comprehensiveness was identified in configurations such as $\epsilon = 0.35$ with MinPts = 5, and $\epsilon = 0.40$ with MinPts= 6. In these combinations, the silhouette index varied between 0.676 and 0.716, and the noise percentage was reduced to the 76% to 82% range - still high, but considerably lower than the models with maximum silhouette.

Finally, the model with the lowest percentage of noise was the one with $\epsilon = 0.45$ and MinPts = 6, which had a silhouette index of 0.58 and classified only 71.9% of the data as noise, which represents the best coverage among the scenarios evaluated.

As the choice of the DBSCAN technique depends directly on the objective of the analysis, which is maximum clustering of the data, and in view of the segmentation of ApexBrasil employees, it did not perform well given the noise presented, and despite the best Silhouette Indices, it was not successful given the number of employees who did not participate in any cluster. The visualization in Figure 2 shows the noise due to the number of observations in Cluster *0*, which was the most comprehensive:

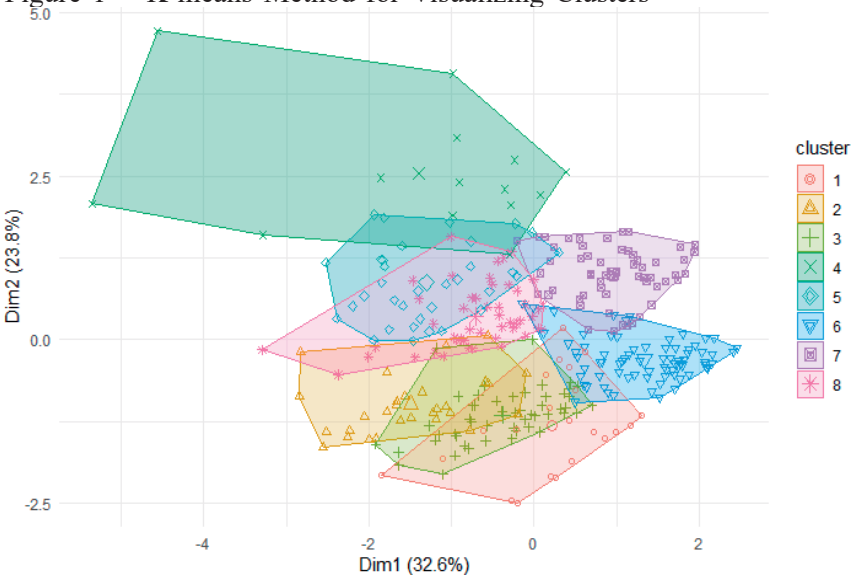
In view of the results in Table 2, the second-best result is the K-means technique with the indication of clusters. It will be detailed in the next section.

5.3 K-MEANS TECHNIQUE

The graph in Figure 1 shows the visualization of clusters obtained using the K-means method on a normalized dataset, with k=8. Each group or cluster is represented by a distinct color and specific geometric shape, making it easier to visually distinguish between different clusters.

The Dim1 and Dim2 axes, which together explain 0.56 of the variances in the data, are principal components that help to reduce dimensionality and project the data into a two-dimensional space. The overlap between some clusters suggests that certain groups have similar characteristics, whereas others stand out in a more isolated way, indicating more differentiated patterns in the data. This type of visualization helps assess the quality of the clusters and identify possible areas of intersection between the groups.

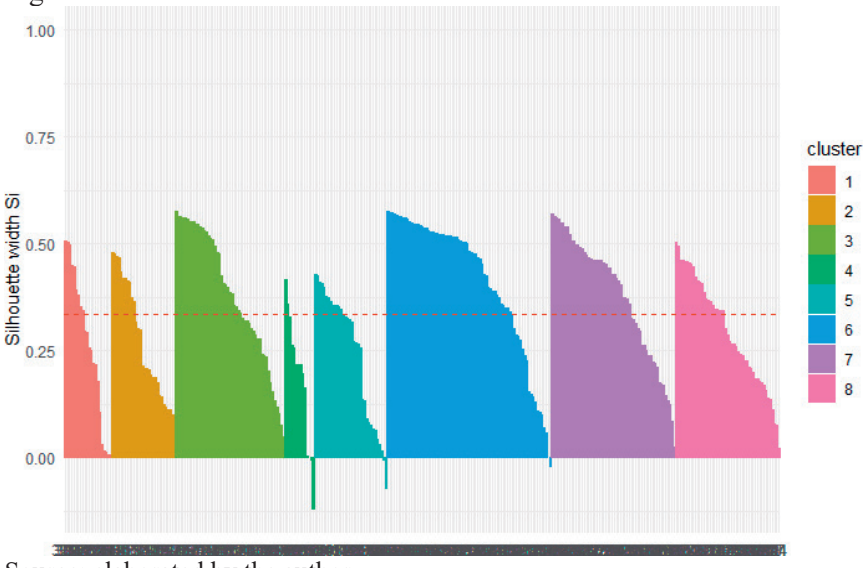
Figure 1 – K-means Method for Visualizing Clusters



Source: elaborated by the author.

The silhouette plot in Figure 2 shows the quality of cluster separation. Values close to 1 indicate that the points are well grouped in their respective clusters, whereas values close to 0 or negative indicate poorly grouped points or points in another cluster, respectively.

Figure 2 – Silhouette Plot - K-means Method



Source: elaborated by the author.

Most clusters have a positive average silhouette width, which suggests a reasonable separation between them. However, there was variability in the width of the bars within each cluster, indicating that some points were less well defined within their groups. The red dotted line represents the overall average silhouette width, indicating the overall quality of the partition. Clusters with negative points or narrow bars indicate that these groups overlap or are weakly distinguished from other clusters.

6 RESULTS: SEGMENTATION OF APEXBRASIL EMPLOYEES AND DISCUSSIONS

In view of the above result, the organization's employees were segmented according to the K-means technique, obtaining a segmentation proposal for the internal clients (employees) of ApexBrasil's Human Resources Management. The best result of the updated techniques was the K-means technique tested with 8 (eight) clusters and a Silhouette Index value of 0.34.

6.1 MANAGERIAL IMPLICATIONS OF IDENTIFIED CLUSTERS

The eight clusters identified by K-Means provide a structured view of workforce heterogeneity at ApexBrasil based on demographic, tenure, compensation, and absenteeism variables. From a People Analytics perspective, the segmentation supports organizational understanding by revealing distinct employee profiles that may require different analytical attention in future studies.

Importantly, this study focuses on comparing clustering techniques and selecting the most suitable approach for the case context, rather than prescribing HR interventions. Therefore, the cluster profiles should be interpreted as descriptive groupings that can guide subsequent investigations, including validation with domain experts and integration with additional variables (e.g., performance, engagement, and competencies) not available in the administrative dataset analyzed.

6.2 EMPLOYEE PROFILE IN CLUSTERS

Table 4 – Employee Profile in Clusters

Cluster #	Age ($\mu\pm\sigma$)	% Female	Tenure ($\mu\pm\sigma$)	Salary ($\mu\pm\sigma$)	Absenteeism ($\mu\pm\sigma$)	Dominant Profile
1	35±5	65%	3±2	8k±2k	5±3	Young, entry-level, female
2	45±6	40%	15±5	15k±4k	7±4	Mature, experienced
3	32±4	55%	2±1	7k±1k	3±2	Recent hires, balanced gender
4	50±7	35%	20±6	18k±5k	10±5	Senior, high tenure, male
5	28±3	70%	1±1	6k±1k	2±1	Entry-level, predominantly female
6	42±5	45%	10±3	12k±3k	6±3	Mid-career, balanced
7	38±6	60%	7±3	10k±2k	8±4	Growing professionals, female
8	55±8	30%	25±7	20k±6k	12±6	Pre-retirement, high compensation

Source: elaborated by the author.

Note: Age in complete years; Tenure in years of service; Salary in R$ thousand; Absenteeism in hours. Following the table, the descriptive text focuses only on particularities of each cluster, avoiding repetition of statistics already tabulated.

6.2.1 Young, entry-level, predominantly female

This profile is consistent with early-career development stages, during which individuals typically prioritize learning, consolidation, and professional identity formation. (NG et al., 2005; Super, 1957). The higher female concentration may reflect occupational segregation patterns described in the literature (Reskin; Roos, 1990), while absenteeism levels can be interpreted under work–family conflict dynamics (Greenhaus; Beutell, 1985).

6.2.2 Mature, experienced employees

This profile is consistent with early-career development stages, during which individuals typically prioritize learning, consolidation, and professional identity formation. (Becker, 1975). The lower female representation aligns with evidence of persistent gender inequality in career progression and the glass ceiling phenomenon (Eagly; Carli, 2007). Absenteeism may be associated with role strain and cumulative demands over time (Goode, 1960).

6.2.3 Recent hires, balanced gender

This profile aligns with organizational socialization theory, which emphasizes adaptation, learning, and integration during the early employment period (Bauer; Erdogan, 2011; Maanen; Schein, 1977). The low absenteeism is compatible with early-stage commitment dynamics described in organizational commitment literature (Meyer; Allen, 1991).

6.2.4 Senior employees, high tenure, predominantly male

This segment reflects extensive institutional and tacit knowledge accumulation, consistent with knowledge-based perspectives of organizations (Nonaka; Takeuchi, 1995). The gender imbalance reinforces cumulative inequality patterns observed in senior positions (Eagly; Carli, 2007). Higher absenteeism levels may be interpreted through the Job Demands–Resources (JD-R) framework, which links sustained demands and insufficient resources to strain outcomes (Bakker; Demerouti, 2007).

6.2.5 Entry-level, predominantly female

This profile is consistent with early career development stages, in which employees seek stability and growth opportunities while building professional identity (Super, 1957). The strong female concentration may reflect occupational segregation dynamics and unequal distribution across job queues (Reskin; Roos, 1990). Low absenteeism is consistent with early organizational commitment and compliance behaviors reported in the literature (Meyer; Allen, 1991).

6.2.6 Mid-career, balanced profile

This profile reflects consolidation and role mastery typically observed in mid-career phases (Hall, 2002). Absenteeism at moderate levels may be associated with competing work and non-work demands, consistent with work–family conflict theory (Greenhaus; Beutell, 1985). From a work design perspective, this segment can be interpreted under job characteristics theory, which highlights the importance of autonomy and meaningfulness for sustained motivation (Hackman; Oldham, 1976).

6.2.7 Growing professionals, predominantly female, high absenteeism

This segment may reflect a professional growth stage characterized by increasing responsibilities and intensified workload demands, potentially influencing attendance patterns (Bakker; Demerouti, 2007). The Job Demand–Control model also supports the interpretation that high demands combined with limited control may contribute to strain-related outcomes such as absenteeism (Karasek, 1979). Gender composition may also

be interpreted in light of broader work–family conflict mechanisms discussed in the literature (Greenhaus; Beutell, 1985).

6.2.8 Pre-retirement, high compensation, highest absenteeism

This profile aligns with late-career stages and aging-related changes in sustainable work ability, which may affect attendance patterns over time (Ilmarinen, 2001; Super, 1957). The segment may also hold critical organizational memory, consistent with knowledge management perspectives (Nonaka; Takeuchi, 1995). The lower female representation reflects cumulative inequality patterns in access to senior and high-pay positions (Eagly; Carli, 2007).

Overall, the cluster profiles highlight meaningful heterogeneity in ApexBrasil's workforce, combining demographic, tenure, compensation, and absenteeism patterns. These descriptive segments provide an interpretable basis for future People Analytics investigations, including validation with HR domain experts and the integration of additional variables not captured in the administrative dataset.

7 CONCLUSIONS

7.1 SYNTHESIS OF FINDINGS

This study compared three unsupervised clustering techniques (K-Means, Hierarchical Clustering, and DBSCAN) for employee segmentation at ApexBrasil, using administrative data from the 2019–2023 period.

K-Means with 8 clusters was selected as the optimal technique, presenting the best balance between technical quality (Silhouette Index = 0.34) and sample coverage (100% of records classified). Although DBSCAN presented higher Silhouette Index values (up to 0.63), it classified a large share of records as noise (from 71.9% to 97.0%, depending on parameter settings), limiting its organizational utility. Hierarchical Clustering produced intermediate results, with Silhouette Index = 0.31 for k=8, and a dendrogram structure coherent with the K-Means solution.

Eight distinct workforce profiles were identified, characterized by combinations of age, gender, tenure, compensation, and absenteeism. Overall, the findings demonstrate that technique selection in People Analytics must balance statistical quality metrics with practical applicability and organizational interpretability, particularly in low analytical maturity contexts.

7.2 SCIENTIFIC CONTRIBUTIONS TO MANAGEMENT AND PUBLIC ADMINISTRATION

This study contributes to the literature in three theoretical and practical dimensions:

7.2.1 Theoretical Contribution: Framework Validation in Public Context

Demonstrates applicability of CRISP-DM framework and unsupervised machine learning techniques in contexts of: Limited administrative data (ERP-sourced); Low organizational analytical maturity; Brazilian public sector (CLT regime). It evidences tension between technical capability (algorithm availability) and organizational maturity (usage capacity), aligning with Lismont's (2017) analytical maturity model.

7.2.2 Methodological Contribution: Technique Selection Protocol

Provides a replicable protocol for comparing clustering techniques that balances technical quality with statistical metrics (Silhouette Index) and Organizational utility with sample coverage and interpretability. Establishes a selection criterion that goes beyond blind metric maximization, incorporating practical applicability considerations. It addresses the gap pointed out by Garg *et al.* (2022) regarding the absence of technique selection guidelines in organizational contexts.

7.2.3 Practical Contribution: Implementation Feasibility

Demonstrates that employee segmentation via clustering is technically feasible with open-source tools (R), applicable to routinely available data (ERP), and interpretable by managers without a technical background. Offers practical evidence for public managers evaluating People Analytics investments, showing that actionable results are achievable without sophisticated infrastructure. Provides implementation roadmap: CRISP-DM → ERP data ETL → Technique comparison → Manager validation → Policy segmentation.

7.3 TECHNICAL CONTRIBUTIONS (DATA SCIENCE)

Complementarily, the study contributes technically by documenting complete ETL process from TOTVS ERP data for HR analytics; Demonstrating treatment of categorical variables (gender, department) in distance-based clustering; Comparing performance of three techniques (K-Means, Hierarchical, DBSCAN) on real HR dataset; Evidencing DBSCAN limitation in contexts where outliers should not be ignored (100% of employees are relevant for management). These technical contributions, although important, are secondary to scientific ones for Management/ Administration journals.

7.4 STUDY LIMITATIONS

This study presents four main limitations:

a) single case: Results refer specifically to ApexBrasil, limiting generalization. Multi-case studies are necessary for external validation;
b) cross-sectional data: Although we use 2019-2023 data, the analysis was cross-sectional (snapshot at December 31, 2023). Longitudinal analyses would allow tracking profile evolution over time;
c) limited variables: Administrative data do not include behavioral variables (engagement, satisfaction), performance, or competencies. Future studies should integrate complementary sources;
d) untested interventions: This study identifies profiles but does not test the effectiveness of specific HR interventions for each segment. Experimental or quasi-experimental studies are necessary.

7.5 DIRECTIONS FOR FUTURE RESEARCH

Future research can extend this work in five directions:

a) external validation: replicate study in other Brazilian public organizations (different sectors, sizes, regions) to assess finding transferability;
b) longitudinal analyses: track cluster evolution over time, investigating profile stability and transitions between segments;
c) data integration: combine administrative data with climate surveys, performance evaluations, and qualitative interviews for richer profile characterization;
d) intervention development: co-create with HR managers specific interventions for each cluster and test effectiveness via experiments;
e) expansion to supervised machine learning: use identified clusters as labels to train predictive models that classify new employees.

7.6 INFRASTRUCTURE RECOMMENDATIONS

To replicate this study, organizations may use open-source analytical environments (R or Python), secure storage compliant with LGPD, and periodic exports from ERP systems

(e.g., CSV). For the ApexBrasil case, follow-up efforts may focus on validating the cluster profiles with HR stakeholders and automating periodic ETL and visualization routines to support monitoring and future People Analytics applications.

7.7 FINAL CONSIDERATIONS

This study demonstrates that implementing People Analytics in Brazilian public organizations is not only feasible but can generate actionable insights with modest infrastructure and routinely available administrative data.

Judicious selection of clustering techniques—balancing technical rigor with practical utility—emerges as a critical success factor. Uncritical adoption of sophisticated algorithms (such as DBSCAN) may result in technically elegant but managerially limited results.

Fundamentally, People Analytics success depends less on tools and more on organizational capability to transform data into decisions, an area where the Brazilian public sector still has significant ground to cover.

REFERENCES

ALCOFORADO, Luciane. **Utilizando a linguagem R**. São Paulo: Alta Books, 2021.

ANDRIYANI, Fitri; PUSPITARANI, Yan. Performance Comparison of K-Means and DBScan Algorithms for Text Clustering Product Reviews. **SinkrOn**, [*s. l.*], v. 7, n. 3, p. 944-949, 2022.

BAKKER, Arnold B.; DEMEROUTI, Evangelia. The Job Demands-Resources model: state of the art. **Journal of Managerial Psychology**, [*s. l.*], v. 22, n. 3, p. 309-328, 2007.

BAUER, Talya N.; ERDOGAN, Berrin. Organizational socialization: the effective onboarding of new employees. *In*: ZEDECK, Sheldon (org.). **APA handbook of Industrial and organizational psychology**. Washington: American Psychological Association, 2011. v. 3. Disponível em: https://content.apa.org/books/12171-002. Acesso em: 1 fev. 2026.

BECKER, Gary S. **Human Capital**: A Theoretical and Empirical Analysis, with Special Reference to Education. New York: NBER, 1975.

CAMPELLO, Ricardo J. G. B.; MOULAVI, Davoud; SANDER, Joerg. Density-Based Clustering Based on Hierarchical Density Estimates. *In*: PEI, Jian *et al.* (org.). **Advances in Knowledge Discovery and Data Mining**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. v. 7819, p. 160-172.

CHAPMAN, Peter. **CRISP-DM 1.0**: Step-by-step data mining guide. [*S. l.*]: SPSS, 2000. Disponível em: https://api.semanticscholar.org/CorpusID:59777418. Acesso em: 1 fev. 2026.

DEVORE, Jay L. **Probabilidade e estatística para engenharia e ciências**. São Paulo: Cengage Learning, 2021.

EAGLY, Alice Hendrickson; CARLI, Linda Lorene. **Through the Labyrinth**: The Truth about how Women Become Leaders. Boston, MA: Harvard Business Press, 2007.

EDWARDS, Martin R.; EDWARDS, Kirsten. **Predictive HR analytics**: mastering the HR metric. 2th ed. London: KoganPage, 2019.

ESTER, Martin; KRIEGEL, Hans-Peter; SANDER, Jörg; XU, Xiaowei. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE AND DATA MINING, 2., 1996, Portland Oregon. **Proceedings** […]. Portland Oregon: AAAI Press, 1996.

FERRAR, Jonathan; GREEN, David. **Excellence in people analytics**: how to use

workforce data to create business value. New York: Kogan Page Limited, 2021.

GABRIEL FILHO, Oscar. **Inteligência artificial e aprendizagem de máquina**: aspectos teóricos e aplicações. São Paulo, SP: Editora Edgard Blucher, 2023.

GARG, Swati *et al.* A review of machine learning applications in human resource management. **International Journal of Productivity and Performance Management**, [*s. l.*], v. 71, n. 5, p. 1590-1610, 2022.

GOODE, William J. A Theory of Role Strain. **American Sociological Review**, [*s. l.*], v. 25, n. 4, p. 483, 1960.

GREENHAUS, Jeffrey H.; BEUTELL, Nicholas J. Sources of Conflict between Work and Family Roles. **The Academy of Management Review**, [*s. l.*], v. 10, n. 1, p. 76, 1985.

GUPTA, Sonal; SHARMA, R. R. K. Types of HR Analytics Used for the Prediction of Employee Turnover in Different Strategic Firms with the use of Enterprise Social Media. *In*: EUROPEAN INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING AND OPERATIONS MANAGEMENT, 5., 2022, Rome, Italy. **Proceedings** […] Rome, Italy, 2022. Disponível em: https://index.ieomsociety.org/index.cfm/article/view/ID/10854. Acesso em: 8 set. 2024

HACKMAN, J. Richard; OLDHAM, Greg R. Motivation through the design of work: test of a theory. **Organizational Behavior and Human Performance**, [*s. l.*], v. 16, n. 2, p. 250-279, 1976.

HALL, Douglas T. **Careers In and Out of Organizations**. Thousand Oaks: SAGE, 2002.

ILMARINEN, Juhani E. Aging workers. **Occupational and Environmental Medicine**, [*s. l.*], v. 58, n. 8, p. 546-546, 2001.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, [*s. l.*], v. 31, n. 8, p. 651-666, 2010.

KAKULAPATI, V. *et al.* Predictive analytics of HR - A machine learning approach. **Journal of Statistics and Management Systems**, [*s. l.*], v. 23, n. 6, p. 959-969, 2020.

KARASEK, Robert A. Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign. **Administrative Science Quarterly**, [*s. l.*], v. 24, n. 2, p. 285, 1979.

LISMONT, Jasmien *et al.* Defining analytics maturity indicators: a survey approach. International **Journal of Information Management**, [*s. l.*], v. 37, n. 3, p. 114-124, 2017.

LLOYD, S. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, [*s. l.*], v. 28, n. 2, p. 129-137, 1982.

MAANEN, J.; SCHEIN, E. **Toward a theory of organizational socialization**. [*S. l.*]: Research in Organizational Behavior, 1977.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In*: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, STATISTICS, 5., 1967, California. **Proceedings** […] Berkeley: University of California Press, 1967. p. 281-298

MARGHERITA, Alessandro. Human resources analytics: A systematization of research topics and directions for future research. **Human Resource Management Review**, [*s. l.*], v. 32, n. 2, p. 100795, 2022.

MEYER, John P.; ALLEN, Natalie J. A three-component conceptualization of organizational commitment. **Human Resource Management Review**, [*s. l.*], v. 1, n. 1, p. 61-89, 1991.

MURTAGH, Fionn; LEGENDRE, Pierre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? **Journal of Classification**, [*s. l.*], v. 31, n. 3, p. 274-295, 2014.

NG, Thomas W. H. *et al.* Predictors of objective and subjective career success: a meta-analysis. **Personnel Psychology**, [*s. l.*], v. 58, n. 2, p. 367-408, 2005.

NONAKA, Ikujirō; TAKEUCHI, Hirotaka. **The Knowledge-creating Company**: How Japanese Companies Create the Dynamics of Innovation. Oxford: Oxford University Press, 1995.

POLZER, Jeffrey T. The rise of people analytics and the future of organizational research. **Research in Organizational Behavior**, [*s. l.*], v. 42, p. 100181, 2022.

RESKIN, Barbara F.; ROOS, Patricia A. **Job queues, gender queues**: explaining women's inroads into male occupations. Philadelphia: Temple University Press, 1990.

ROKACH, Lior; MAIMON, Oded. Clustering Methods. *In*: MAIMON, Oded; ROKACH, Lior (org.). **Data Mining and Knowledge Discovery Handbook**. New York: Springer-Verlag, 2005. p. 321-352.

ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, [*s. l.*], v. 20, p. 53-65, 1987.

SCHUBERT, Erich *et al.* DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. **ACM Transactions on Database Systems**, [*s. l.*], v. 42, n. 3, p. 1-21, 2017.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business intelligence, analytics, and data science**: a managerial perspective. 4th ed. New York, NY: Pearson, 2018.

SUPER, Donald E. (Donald Edwin). **The psychology of careers**: an introduction to vocational development. New York: Harper, 1957.

TURSUNBAYEVA, Aizhan *et al.* The ethics of people analytics: risks, opportunities and recommendations. **Personnel Review**, [*s. l.*], v. 51, n. 3, p. 900-921, 2022.

WATERS, Shonna D. *et al.* **The practical guide to HR analytics**: using data to inform, transform, and empower HR decisions. Alexandria, Virginia: SHRM, Society for Human Resource Management, 2018.

XU, R.; WUNSCH II, D. Survey of Clustering Algorithms. **IEEE Transactions on Neural Networks**, [*s. l.*], v. 16, n. 3, p. 645-678, 2005.

Submetido: 10 dez. 2025
Aprovado: 20 fev. 2026