Capability-Based Workforce Clustering for Agile Cross-Functional Team Formation in Onboarding Programs

Astika Ayuningtyas[1,*], Rindi Nur Wulandari[2]

[1,2]Department of informatics, Adisutjipto of Aerospace Technology, Yogyakarta, Indonesia

ABSTRACT

Agile organizations require the ability to form balanced, cross-functional teams rapidly, particularly during employee onboarding phases when information is limited and decisions must be made quickly. This study proposes a machine-learning-driven approach for workforce segmentation that supports agile squad formation using only onboarding-stage employee data. The objective is to identify interpretable capability-based personas that can inform early staffing decisions without relying on performance outcomes or post-hoc evaluations. A publicly available employee onboarding dataset was analyzed using unsupervised clustering techniques. Feature representation focused on workforce capability proxies and onboarding logistics, including educational attainment, qualification level, trade specialization, geographic information, and joining time. A unified preprocessing pipeline was applied to ensure consistent handling of numeric and categorical attributes, enabling fair comparison across clustering algorithms. Three clustering paradigms—KMeans, Gaussian Mixture Models, and Agglomerative Hierarchical Clustering—were evaluated across a range of cluster counts. Cluster validity was assessed using internal metrics, namely Silhouette Score and Davies–Bouldin Index, complemented by inertia-based elbow analysis for KMeans. The results consistently indicate that the data supports only a small number of meaningful segments, with optimal solutions emerging at low cluster counts. Visualization using two-dimensional principal component analysis further reveals overlapping but structured workforce patterns, suggesting gradual capability transitions rather than sharply separated groups. The findings are discussed in the context of agile workforce management, emphasizing the interpretation of clusters as soft personas rather than rigid categories. These personas can be combined flexibly to form balanced squads, supporting rapid decision-making while preserving adaptability. The proposed framework demonstrates how unsupervised learning can enhance transparency and speed in onboarding-related staffing decisions, offering a practical contribution to agile HR analytics and workforce planning research.

Keywords Agile Workforce Management, Employee Onboarding, Clustering Analysis, Workforce Segmentation, Machine Learning

Submitted: 20 April 2025
Accepted: 5 June 2025
Published: 2 November 2025

Corresponding author
Astika Ayuningtyas,
astika@itda.ac.id

Additional Information and Declarations can be found on page 239

Introduction

Agile principles—originally developed for software engineering—are increasingly adopted across organizational functions, including workforce management and Human Resources (HR). Multiple case studies and review articles document the diffusion of agile values such as rapid iteration, decentralized decision-making, and cross-functional teams into non-software domains, including healthcare operations and enterprise-wide implementations. These studies demonstrate that agile methods can support fast adaptation and iterative deployment of work practices beyond traditional software projects [1], [2], [3]. Empirical evidence from clinical and institutional settings further shows

How to cite this article: A. Ayuningtyas and R. N. Wulandari, "Assessing the Efficacy of Convolutional Neural Networks in Recognizing Handwritten Digits Analysis of the Relationship Between Trading Volume and Bitcoin Price Movements Using Pearson and Spearman Correlation Methods," *Agile manag*, vol. 1, no. 4, pp. 226-241, 2025.

that agile or agile-inspired deployments—such as rapid team restructuring, decentralized "superuser" roles, and iterative process rehearsals—enable organizations to respond effectively to emergent operational demands under severe time pressure [4], [5].

Within such dynamic organizational environments, the ability to form teams rapidly and flexibly during onboarding becomes critical. Rapid team formation allows new employees to be socialized quickly and to rehearse workflows with existing staff, a practice associated with safer and more efficient service delivery in clinical and perioperative contexts. Prior studies report that simulated onboarding and early team practice can reduce operational risk and support end-to-end process readiness before live service delivery [4]. Similarly, technical and research communities report that fast onboarding mechanisms—such as containerized development environments—substantially reduce setup time and accelerate newcomer productivity in highly technical collaborative projects [6].

Despite its importance, onboarding often occurs under conditions of bounded information and acute time pressure. Organizations report wide variation in onboarding practices, with limited empirical evidence identifying which specific onboarding tactics most effectively accelerate newcomer adjustment. As a result, managers are frequently required to make rapid allocation decisions under partial and imperfect information [7]. The COVID-19 pandemic and the rapid transition to remote work further exacerbated these constraints, introducing connectivity, infrastructure, and tooling limitations that reduced opportunities for in-person socialization and informal learning during onboarding [8].

Field reports from staffing and operational contexts indicate that early allocation decisions can have measurable downstream effects. Differences in adherence, coordination, and operational behavior between workforce subgroups—such as agency versus permanent staff—are often first detected through early onboarding data reviews, suggesting that initial team assignments may produce persistent variance in team performance [9]. These findings highlight onboarding as a high-impact decision phase where rapid yet informed team formation is essential.

Onboarding represents a critical early decision point in agile organizations because initial allocations directly shape team balance, coordination pathways, and the early distribution of tacit knowledge. Research on agile team formation and high-performing product teams emphasizes that early team composition and leadership deployment materially influence adaptability, collective intelligence, and the ability to iterate effectively under uncertainty—outcomes that are central to agile performance [2], [3]. Evidence from clinical and emergency reorganizations, including rapid COVID-19 response teams, further illustrates that early allocation choices—such as pairing strategies and advisory role assignments—significantly affect how quickly teams achieve functional competence and interdisciplinary coordination [4], [5].

Despite the strategic importance of these early decisions, onboarding allocations are frequently made using heuristics or managerial intuition rather than systematic analysis. Reviews of onboarding practices document heterogeneous adoption of structured onboarding elements, with practitioners often relying on locally defined rules, informal mentorship, or ad-hoc team formation rather than standardized, evidence-based assignment procedures [7], [10]. In emergency or rapid implementations, organizations often prioritize

decentralized and pragmatic decision-making to accelerate deployment, such as empowering staff to independently create and manage digital collaboration spaces, which favors speed but embeds heuristic allocations and limits centralized analytical oversight [1].

Operational constraints further reinforce reliance on intuition during onboarding. Remote onboarding environments, infrastructure limitations, and uneven access to shared knowledge repositories reduce the availability of objective signals for informed decision-making. Under such conditions, managers must allocate personnel quickly despite weak or incomplete information, increasing the likelihood of suboptimal early assignments [8]. These constraints are especially pronounced in agile contexts where delays in team formation can undermine responsiveness and iterative progress.

These patterns motivate the need for complementary data-driven decision support in agile onboarding. Prior work in project management and health informatics highlights how analytics and artificial intelligence can enhance rapid decision-making by surfacing interpretable insights, automating routine analyses, and supporting managers operating under time constraints [11]. Case reports showing that early data analysis can reveal unexpected adherence gaps or operational variance underscore the potential of analytic augmentation to identify allocation risks that heuristic judgment alone may overlook [9]. Accordingly, lightweight and interpretable analytics can augment managerial intuition, preserve agility while reduce the probability of suboptimal early team formation.

Machine learning (ML) has emerged as a key enabler of advanced HR analytics and decision support, offering tools to uncover patterns in workforce data that are not readily apparent through manual analysis [11], [12]. Within this broader landscape, unsupervised learning methods—particularly clustering—are well suited to onboarding contexts because they identify structure in unlabeled data. This characteristic is essential during early onboarding stages, where performance outcomes and supervisory evaluations are not yet available [13], [12].

Unsupervised techniques such as k-means clustering, Gaussian mixture models, hierarchical clustering, density-based methods, and dimensionality reduction approaches can reveal latent groupings of employees based on skills, prior experience, educational background, or learning-profile proxies. These methods operate without requiring historical outcome labels, making them appropriate for day-zero workforce segmentation when only background and contextual data are available [13], [12]. As a result, clustering can support early-stage decision-making under label scarcity.

In practical settings, clustering outputs can inform initial team assignments, suggest complementary pairings (for example, aligning novices with specific mentor profiles), or highlight gaps in team capability coverage that managers may address during early agile iterations [3]. However, successful adoption of ML in HR settings depends not only on technical performance but also on interpretability, transparency, and alignment with managerial workflows. Prior studies show that visualization techniques—such as PCA-based projections combined with cluster labels and descriptive statistics—significantly improve practitioners' ability to interpret clusters and translate them into actionable onboarding interventions [14], [15].

Recent reviews further emphasize that ML systems intended to support agile managerial decision-making must be explainable, human-centered, and

designed to complement rather than replace managerial judgment [11], [16]. When these conditions are met, unsupervised ML can function as a practical decision-support layer that enhances speed and consistency without undermining local contextual knowledge. Collectively, the literature supports framing workforce onboarding in agile organizations as both a time-sensitive management challenge and a suitable application domain for interpretable unsupervised learning techniques that surface structure under uncertainty [11], [3], [14].

Existing research on workforce analytics and HR-oriented machine learning has largely concentrated on performance prediction, retention modeling, and post-hoc evaluation, often relying on outcome labels that become available only after employees have spent significant time in the organization. While these studies provide valuable insights for long-term talent management, they offer limited guidance for early-stage onboarding decisions, where performance data do not yet exist. In addition, prior segmentation studies frequently utilize rich longitudinal or behavioral datasets, leaving a gap in methods that operate exclusively on onboarding-stage attributes. As a result, the specific problem of forming agile, balanced teams rapidly at the point of entry remains underexplored.

Moreover, the literature has paid relatively little attention to agile squad formation as a primary objective of workforce segmentation. Existing work typically frames clustering as a descriptive exercise or as a precursor to prediction, rather than as a direct decision-support mechanism for team composition. Studies that address agile management tend to focus on software development teams or mature project environments, with limited consideration of how agile principles can be operationalized during onboarding using minimal and readily available data. Consequently, organizations lack systematic, data-driven approaches to segment onboarding employees in ways that directly support agile team formation under time and information constraints.

To address these gaps, the objective of this study is to develop a clustering-based workforce segmentation framework that relies solely on onboarding data and is explicitly designed to support agile squad formation. The study contributes by (i) providing a comparative evaluation of multiple clustering algorithms under a unified preprocessing pipeline, (ii) deriving interpretable workforce personas that translate high-dimensional onboarding data into actionable capability profiles, and (iii) demonstrating practical implications for agile HR and onboarding management as a form of decision support rather than automated assignment.

Literature Review

Agile Management and Workforce Agility

Agile management originated in software engineering as a response to uncertainty, emphasizing iterative development, rapid feedback, decentralized decision-making, and cross-functional collaboration. Foundational reviews describe agile practices as mechanisms for increasing adaptability and shortening response cycles in volatile technical and market environments [1]. Over time, empirical studies have documented the diffusion of agile principles into non-software domains, including enterprise operations and healthcare services. For example, decentralized deployment strategies were used to accelerate organization-wide collaboration platform rollouts during the COVID-

19 response [3], while clinical informatics teams applied rapid-cycle, data-driven interventions to train and scale healthcare workforces during pandemic surges [15]. Case reports further demonstrate that agile-inspired team designs—such as cross-disciplinary structures, iterative learning, and delegated ownership—enable organizations to construct responsive operational systems under acute time pressure [2], [4].

Within this expanded scope, the concept of workforce agility has emerged as a central construct. Workforce agility is commonly characterized by flexibility in reallocating personnel, responsiveness to emergent demands, and cross-functionality achieved through complementary skill composition [2], [3]. These attributes closely align with agile team constructs, which emphasize small, empowered groups capable of rapid role reconfiguration and iterative coordination. Prior theoretical and empirical work links such configurations to improved adaptability and sustained performance in uncertain environments, reinforcing workforce agility as a managerial capability rather than a fixed structural property [2], [3].

Employee Onboarding and Early Workforce Decisions

Research on employee onboarding highlights newcomer socialization, role clarity, and early capability transfer as key objectives that accelerate adjustment and effectiveness [7]. However, empirical reviews consistently note substantial variation in onboarding practices across organizations, driven by time constraints, resource limitations, and heterogeneous institutional norms. These challenges are amplified in remote or crisis contexts, where infrastructure gaps and reduced informal interaction constrain traditional socialization mechanisms [8]. As a result, large-scale operational responses—such as health system surge training—have adopted compressed onboarding approaches that rely on targeted training pathways and visualized workflows to balance speed with safety [15].

Early placement and cohort structuring during onboarding represent critical inflection points for team performance. Evidence from clinical and operational settings indicates that initial pairings, advisory attachments, and cohort composition shape how quickly teams acquire tacit knowledge and establish coordination norms [4], [5], [7]. Small early interventions—such as assigning subject-matter advisors or conducting end-to-end workflow simulations—can materially alter the trajectory of team competence within days [4], [5]. At the same time, studies show that early allocation decisions are often made heuristically, with downstream variance in adherence and coordination revealing the long-term impact of these initial choices [9]. These findings underscore onboarding as a high-leverage but analytically under-supported decision stage.

Human Resource Analytics and Machine Learning

Human resource analytics has evolved from descriptive reporting toward predictive and prescriptive modeling aimed at informing workforce planning, talent development, and organizational design. Recent reviews identify artificial intelligence and machine learning as key enablers of this transition, allowing organizations to move beyond retrospective metrics toward model-driven decision support [16], [18]. Conceptual frameworks for HR analytics emphasize staged processes that integrate data collection, modeling, interpretation, and

reflection while maintaining governance and transparency [16]. This evolution reflects a broader shift toward evidence-informed management in dynamic organizational environments.

Within HR practice, machine learning techniques are applied across a range of tasks, including supervised prediction (e.g., turnover risk), unsupervised segmentation (e.g., skill or role grouping), anomaly detection, and natural language processing of unstructured personnel data [12], [17], [19]. Unsupervised methods are particularly relevant when outcome labels are unavailable at decision time, such as during onboarding. However, the literature also emphasizes ethical and practical constraints associated with ML in HR, including algorithmic bias, data quality limitations, and the need for human-in-the-loop decision framing [20], [21]. Accordingly, reviews recommend that ML systems in HR be designed as transparent decision-support tools that augment, rather than replace, managerial judgment [16], [22].

Clustering and Segmentation in Workforce Analytics

Clustering techniques are widely used to uncover latent structure in unlabeled workforce data. Common methods include K-Means, Gaussian Mixture Models, hierarchical clustering, and density-based approaches, often paired with dimensionality reduction techniques such as Principal Component Analysis (PCA) for visualization [12], [13]. Methodological reviews and applied studies demonstrate that these techniques can produce interpretable summaries of complex, high-dimensional feature spaces, enabling managers to inspect and reason about workforce heterogeneity without relying on predefined categories [12], [13].

In HR and organizational contexts, clustering has been applied to employee segmentation, skill grouping, and role profiling to support targeted learning, mentoring, and team composition decisions [17], [16]. Practical implementations typically combine cluster assignments with descriptive statistics and domain labeling to translate algorithmic outputs into actionable guidance [14], [17]. At the same time, the literature cautions that clustering has inherent limitations when applied to human-centered data: skill boundaries are often fluid, cluster assignments may overlap, and results are sensitive to modeling choices [13], [14]. Consequently, ethical and practical frameworks recommend presenting clustering outputs as supportive artifacts—paired with visualization, uncertainty awareness, and managerial oversight—to ensure responsible and effective use in agile workforce decision-making [22], [20].

Method

Data Source and Ethics

The study uses a structured workforce onboarding dataset where each record corresponds to one newly onboarded employee and includes attributes related to education, skill specialization, location, compensation proxy (stipend), and joining time. The dataset is treated as a cross-sectional snapshot at onboarding, which is appropriate for segmentation aimed at early staffing decisions.

To reduce privacy risk and prevent identity-based clustering, personally identifying fields were excluded from the modelling table. The modelling workflow does not require names or unique employee identifiers; the analysis focuses on capability and onboarding context variables that are relevant to

forming balanced agile squads.

A consistent data dictionary was established by harmonizing column names and resolving formatting inconsistencies across the dataset. Categorical text fields were standardized by trimming whitespace, removing duplicated spacing, and normalizing capitalization to reduce artificial category inflation caused by inconsistent entry formats.

Feature Engineering and Data Preparation

The feature set was designed to represent workforce capability proxies and onboarding logistics, reflecting the goal of rapid segmentation for agile squad formation. Core capability signals include qualification level and trade specialization, complemented by academic percentage fields that provide coarse indicators of prior attainment.

Temporal features were derived from the Date of Joining field. The date string was parsed into a standard datetime format and decomposed into time-based variables such as join month and ISO week number. These derived features allow clustering to reflect cohort and intake-cycle patterns without using personally identifying information.

Numeric fields were converted to consistent numeric types. Education percentages were parsed by removing percent symbols and coercing invalid strings to missing values. Stipend values were cast to numeric with errors converted to missing values. This approach prevents parsing artefacts from propagating into the feature space used by clustering.

The final modelling table separated features into numeric and categorical groups. Numeric features were imputed using the median (robust to skew and outliers) and scaled using standardization (z-score). Categorical features were imputed using the most frequent value and encoded using one-hot encoding with unknown-category handling enabled to support generalization when new categories appear.

Preprocessing Pipeline and Reproducibility Controls

A unified preprocessing pipeline was implemented using a ColumnTransformer architecture, enabling consistent treatment of numeric and categorical subsets. This ensures that the same transformations are applied during both model selection and final training, preventing data leakage or inconsistent representations across experiments.

For numeric variables, the pipeline applies median imputation followed by StandardScaler. StandardScaler centers each numeric feature to zero mean and unit variance, which is important because distance-based clustering methods are sensitive to feature scale. Median imputation was chosen to avoid sensitivity to extreme values and to provide stable behavior on sparse missingness.

For categorical variables, the pipeline applies most-frequent imputation followed by OneHotEncoder with `handle_unknown="ignore"`. This prevents runtime failures if the pipeline encounters a category not present during fitting and ensures that the encoded feature space remains compatible between training and later application.

Clustering Algorithms and Parameter Settings

Three clustering families were evaluated to reduce algorithm-specific bias: centroid-based clustering (KMeans), probabilistic clustering (Gaussian Mixture Models), and hierarchical clustering (Agglomerative). This triangulation provides methodological robustness because each method makes different assumptions about cluster geometry and membership.

KMeans was configured with multiple initializations (`n_init` set to a high value) to reduce sensitivity to local minima. A fixed `random_state` ensured that results are reproducible under the same environment. The algorithm minimizes within-cluster sum of squared distances in the standardized feature space, making scaling and consistent preprocessing essential.

Gaussian Mixture Models were configured with `covariance_type="full"` to allow clusters to take elliptical shapes and to capture correlations among features in the encoded space. Model fitting uses the expectation–maximization (EM) procedure, producing soft membership assignments internally; hard labels are obtained by selecting the maximum posterior probability for evaluation purposes.

Agglomerative clustering used Ward linkage, which merges clusters to minimize the increase in within-cluster variance at each step. Ward linkage is compatible with Euclidean geometry and provides a deterministic alternative to KMeans and GMM that does not require random initialization, enabling a useful comparison across fundamentally different clustering paradigms.

Model Selection Protocol and Internal Validation

Cluster counts were evaluated across a predefined range (k from 2 through 10) to balance interpretability and over-fragmentation risk. For each k, clustering models were fit on the same preprocessed feature matrix, ensuring comparable evaluation conditions.

Internal validation used Silhouette Score and Davies–Bouldin Index, computed from the standardized feature representation. Silhouette Score measures the relative separation between an observation's assigned cluster and its nearest alternative cluster, while Davies–Bouldin evaluates average cluster similarity based on within-cluster dispersion and between-cluster distance. Using both metrics reduces reliance on any single criterion.

For KMeans, inertia (sum of squared errors) was also tracked as a heuristic diagnostic. Inertia decreases monotonically with larger k, so it is used only to identify diminishing returns rather than as an optimization objective. This complements Silhouette and Davies–Bouldin by providing an intuitive view of variance reduction as clusters increase.

A deterministic selection rule was applied to choose a final k for the main segmentation: select k that maximizes Silhouette Score for the primary method (KMeans), and use Davies–Bouldin as a tie-breaker when Silhouette values are effectively equivalent. This rule favors clusters that are simultaneously cohesive and reasonably separated while keeping the decision procedure reproducible.

Persona Extraction and Agile Squad Formation Templates

After selecting the clustering configuration, cluster personas were constructed

using interpretable summaries. For each cluster, the pipeline computes cluster size, categorical modes (e.g., dominant qualification and trade), and numeric means (e.g., central tendencies for education percentages and stipend). These summaries translate high-dimensional encoded clusters into human-readable workforce personas.

To make personas actionable for agile management, the method includes an optional mapping step that relates personas to operational units. Department counts are computed within each cluster, yielding a distribution that can be used to propose staffing templates such as “prioritize personas A and B for teams that frequently draw from department X.” This step is descriptive and intended for decision support rather than causal inference.

The method treats segmentation as a planning aid, not an evaluation mechanism. Cluster membership is interpreted as a capability cohort indicator that can guide balanced composition (mixing complementary trades and qualification levels) while retaining flexibility for managerial judgment, role requirements, and interpersonal considerations.

Result and Discussion

Cluster Number Determination and Internal Validation

The cluster validation results consistently indicate that meaningful workforce segmentation emerges only at relatively small cluster counts. Across all evaluated algorithms, Silhouette Scores peak in the range of k = 3 to k = 4, after which a sharp decline is observed. This pattern suggests that the onboarding workforce is characterized by a limited number of broad capability groupings rather than many finely separable subgroups.

Among the evaluated methods, KMeans achieves the highest Silhouette Score, reaching its maximum at k = 3–4. Gaussian Mixture Models and Agglomerative Clustering show comparable but slightly lower Silhouette values, following the same downward trend as k increases. This consistency across algorithms strengthens confidence that the observed structure reflects intrinsic properties of the data rather than artefacts of a particular clustering technique.

The Davies–Bouldin Index provides complementary evidence. The index decreases substantially when moving from k = 2 to k = 3 and reaches its lowest values around k = 4, after which it increases again. This behavior indicates that cluster compactness and separation improve initially but degrade as additional clusters are forced, leading to fragmentation of naturally overlapping workforce profiles.

The Elbow Curve (figure 1) based on KMeans inertia further supports this conclusion. A pronounced reduction in within-cluster variance occurs up to approximately k = 4, after which the slope flattens, indicating diminishing returns from additional clusters. Taken together, these three diagnostics converge on k = 3 or k = 4 as the most defensible segmentation choices, balancing numerical validity with interpretability.

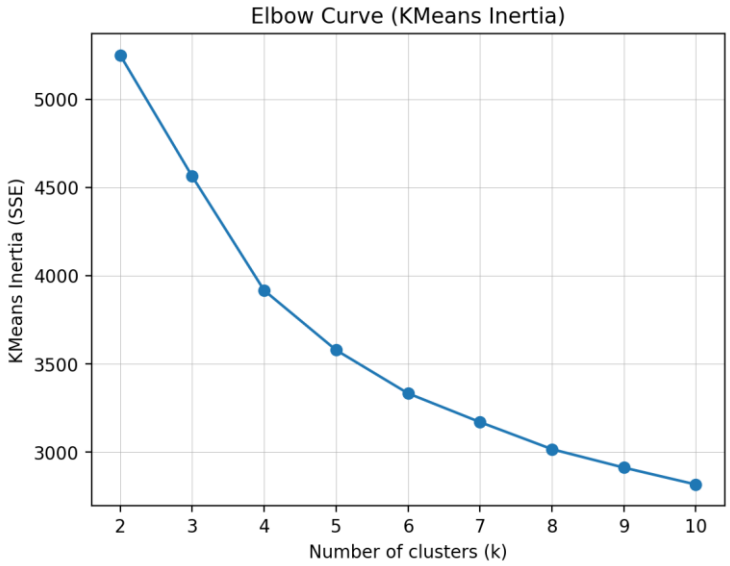

Figure 1 **Elbow Curve**

Comparison of Clustering Algorithms and Structural Patterns

When comparing clustering algorithms, KMeans demonstrates the most stable behavior across validation metrics, as shown in figure 2. Its Silhouette curve exhibits a clear peak at low k values and a smooth decline thereafter, suggesting consistent centroid-based grouping in the standardized feature space. This stability is particularly important for workforce segmentation, where interpretability and reproducibility are critical.

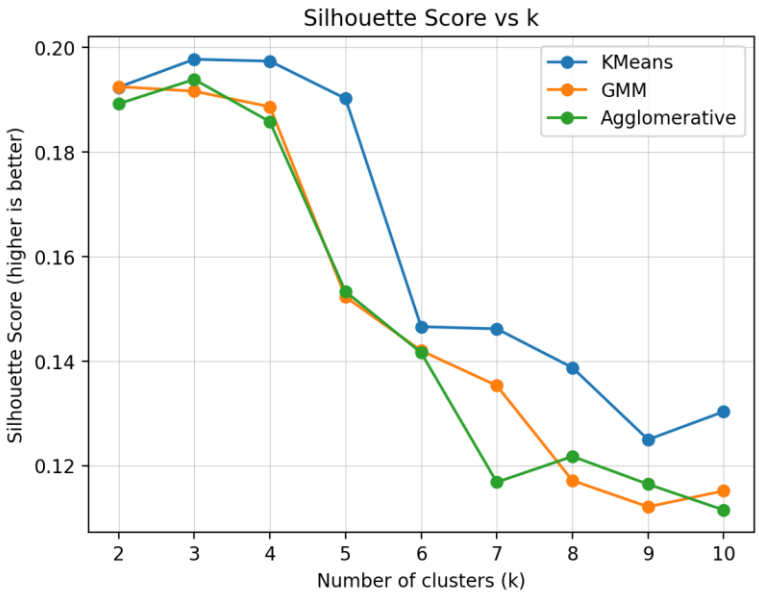

Figure 2 **Silhouette curve across Clusters**

Gaussian Mixture Models produce similar trends but with slightly lower Silhouette Scores and higher Davies–Bouldin values. This indicates that while probabilistic clustering can capture overlapping membership, the added model flexibility does not translate into stronger internal separation for this dataset. The

workforce profiles appear to overlap in ways that are not well described by elliptical Gaussian components alone.

Agglomerative Clustering follows the same general pattern but shows greater sensitivity to increasing k. The rise in Davies–Bouldin Index at moderate cluster counts suggests that hierarchical merging begins to group heterogeneous profiles as k increases. This reinforces the idea that the dataset does not support deep hierarchical subdivision beyond a small number of top-level segments.

Overall, the comparative results suggest that algorithm choice is less critical than cluster count selection in this context. All three methods reveal similar structural limits, indicating that the onboarding workforce naturally forms a small number of broad, partially overlapping capability clusters rather than sharply delineated groups.

PCA-Based Visualization and Cluster Separability

Two-dimensional PCA projections were used to visualize the structure of the clustered feature space, as shown in figure 3, figure 4 and figure 5. Across KMeans, GMM, and Agglomerative clustering at k = 3, the PCA plots reveal dense point clouds with visible sub-structures but substantial overlap between clusters. No algorithm produces clearly isolated “islands” in the projected space.

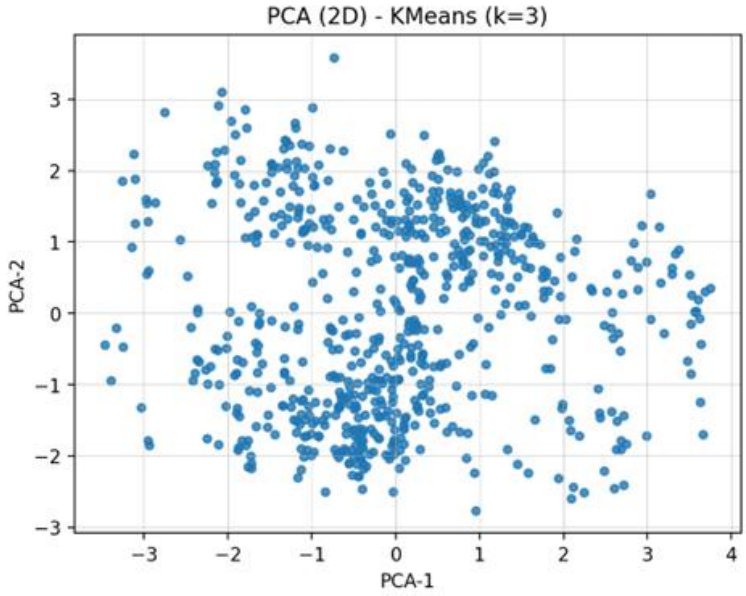

Figure 3 PCA Plots of KMeans Clustering

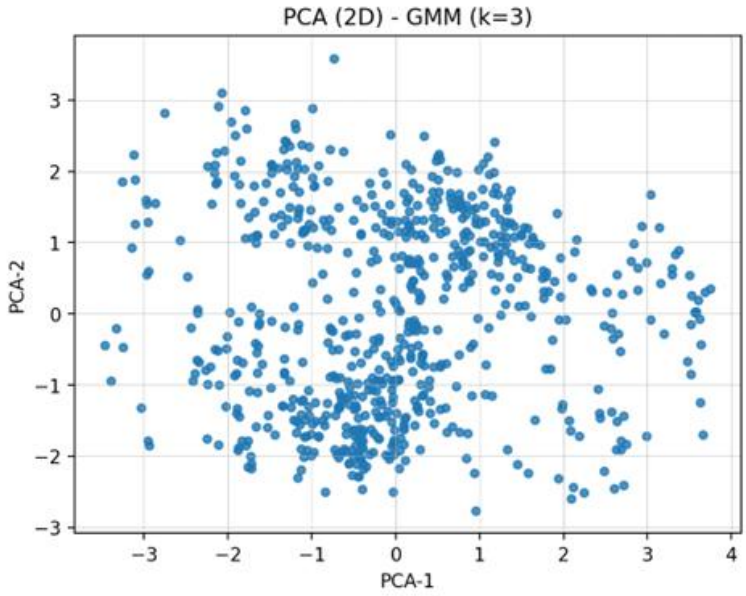

Figure 4 PCA Plots of GMM Clustering

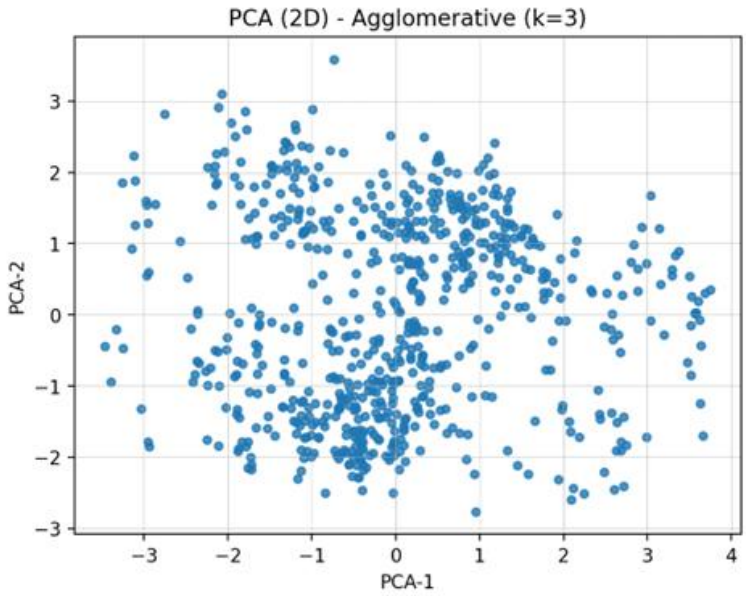

Figure 5 PCA Plots of Agglomerative Clustering

This overlap aligns with the relatively low absolute Silhouette values observed across methods. In practical terms, the clusters represent gradual capability gradients rather than discrete workforce categories. Such patterns are typical in human-resource datasets, where education level, trade specialization, and location interact continuously rather than forming sharply bounded classes.

Importantly, the similarity of PCA plots across algorithms suggests that differences in clustering outcomes are primarily quantitative rather than qualitative. Regardless of method, the workforce appears organized around a small number of central tendencies, with individuals distributed along transitions between these tendencies rather than concentrated in isolated regions.

From a managerial perspective, this visualization supports interpreting clusters as soft personas rather than rigid classifications. Employees near cluster boundaries may reasonably fit into multiple squad compositions, reinforcing the need for human judgment and contextual adjustment when applying clustering outputs to agile staffing decisions.

Implications for Agile Workforce Segmentation

The combined results indicate that agile workforce segmentation during onboarding should prioritize simplicity and flexibility over excessive granularity. The convergence of Silhouette, Davies–Bouldin, and Elbow analyses at low k values implies that attempting to define many distinct personas would artificially divide naturally overlapping capability profiles.

Selecting k = 3 provides a parsimonious representation of workforce diversity while remaining interpretable for operational use. Although k = 4 shows marginally better compactness by Davies–Bouldin Index, the additional complexity may not yield proportional managerial benefit, especially in fast-paced onboarding contexts where rapid decision-making is required.

The observed overlap in PCA space reinforces agile principles: squads should be formed by combining complementary capability profiles, not by isolating narrowly defined roles. The clustering results support using personas as guides for balancing skills and qualifications across squads rather than as deterministic assignment rules.

Finally, the results highlight the value of unsupervised learning as a decision-support mechanism rather than an evaluative tool. The segmentation reveals structural patterns in onboarding data that can accelerate squad formation and capacity planning, while preserving adaptability—an essential characteristic of agile workforce management.

Conclusion

This study demonstrates that unsupervised machine learning can be effectively applied to workforce onboarding data to support agile squad formation through data-driven segmentation. By evaluating multiple clustering algorithms and internal validation criteria, the analysis shows that the onboarding workforce is best represented by a small number of broad capability-based personas rather than many finely separated groups. The convergence of Silhouette, Davies–Bouldin, and Elbow diagnostics indicates that meaningful structure exists primarily at low cluster counts, highlighting the inherently overlapping nature of employee capability profiles at the point of entry. From an agile management perspective, these findings reinforce the value of using clustering as a decision-support tool rather than a deterministic assignment mechanism. The identified personas provide a practical abstraction that can accelerate early staffing, balance cross-functional squads, and improve transparency in onboarding decisions, while still allowing flexibility and managerial judgment. Overall, the proposed approach offers a lightweight, reproducible, and interpretable framework for agile workforce segmentation that can be readily adapted to other onboarding contexts with similar data characteristics.

Declarations

Author Contributions

Conceptualization: A.A. and R.N.W.; Methodology: R.N.W.; Software: A.A.; Validation: A.A. and R.N.W.; Formal Analysis: A.A. and R.N.W.; Investigation: A.A.; Resources: R.N.W.; Data Curation: R.N.W.; Writing Original Draft Preparation: A.A. and R.N.W.; Writing Review and Editing: R.N.W. and A.A.; Visualization: A.A.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Mehta, T. Yates, P. Smith, D. Henderson, G. Winteringham, and Á. Burns, “Rapid implementation of Microsoft Teams in response to COVID-19: One acute healthcare organisation’s experience,” *BMJ Health & Care Informatics*, vol. 27, no. 3, p. e100209, 2020, doi: 10.1136/bmjhci-2020-100209.

[2] M. Govindaraju and P. Kumar, “Building and leading high-performing product teams in fast-paced and innovative environments,” *IJRMEET*, vol. 13, no. 3, pp. 1–12, 2025, doi: 10.63345/ijrmeet.org.v13.i3.25.

[3] D. A. Putri Zainal, R. Razali, and Z. Mansor, “Team formation for agile software development: A review,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 2, pp. 555–561, 2020, doi: 10.18517/ijaseit.10.2.10191.

[4] M. J. Ferguson, C. Sampson, J. Duff, and T. Green, “Integrated simulations to build teamwork, safety culture and efficient clinical services: A case study,” *J. Perioperative Nursing*, vol. 35, no. 2, pp. 1–9, 2022, doi: 10.26550/2209-1092.1168.

[5] A. Kangas-Dick et al., “Rapid adoption of an interdisciplinary care team model for surgical residents managing coronavirus disease-19,” *J. Laparoendoscopic & Adv. Surg. Tech.*, vol. 31, no. 5, pp. 541 - 545, 2021, doi: 10.1089/lap.2021.0120.

[6] B. Rottler, M. Böhler, G. Duckeck, A. M. Lory, C. Mitterer, and J. Schovancová,

"Bringing the ATLAS HammerCloud setup to the next level with containerization," *EPJ Web Conf.*, vol. 295, no. 8, p. 04011, 2024, doi: 10.1051/epjconf/202429504011.

[7] H. J. Klein, B. Polin, and K. L. Sutton, "Specific onboarding practices for the socialization of new employees," *Int. J. Selection Assessment*, vol. 23, no. 3, pp. 263–283, 2015, doi: 10.1111/ijsa.12113.

[8] J. Oberholzer, C. Schultz, and K. F. Lessing, "Contributing and constraining factors regarding the implementation of human resource management onboarding during the COVID-19 pandemic," *Athens J. Bus. Econ.*, vol. 11, no. 3, pp. 275-304, 2025, doi: 10.30958/ajbe.11-3-3.

[9] L. A. Smith and B. Savage, "Agency nurse usage of infusion interoperability," *Comput. Informatics Nursing*, vol. 42, no. 2, p. e01206, 2024, doi: 10.1097/CIN.0000000000001206.

[10] N. Svetozarovová, T. Kincl, J. Coculʼová, and A. Burdová, "Implementation of trends in human resources management as a precondition for business performance," *J. Manage. Bus. Res. Practice*, vol. 13, no. 2, pp. 1-11, 2021, doi: 10.54933/jmbrp-2021-13-2-7.

[11] I. A. Adeniran, E. E. Agu, C. P. Efunniyi, O. S. Osundare, and H. O. Iriogbe, "The future of project management in the digital age: Trends, challenges, and opportunities," *Eng. Sci. Technol. J.*, vol. 5, no. 8, pp. 2632-2648, 2024, doi: 10.51594/estj.v5i8.1516.

[12] M. Kagzi, S. Khanra, and S. K. Paul, "Machine learning for sustainable development: Leveraging technology for a greener future," *J. Syst. Inf. Technol.*, vol. 25, no. 4, pp. 440–479, 2023, doi: 10.1108/JSIT-11-2022-0266.

[13] D. Ghosh, S. Chakraborty, H. Kodamana, and S. Chakraborty, "Application of machine learning in understanding plant virus pathogenesis: Trends and perspectives," *Virology J.*, vol. 19, no. 42, pp. 1–11, 2022, doi: 10.1186/s12985-022-01767-5.

[14] R. E. Yuliana, D. M. Ulya, and M. Jamhuri, "Mall customer segmentation using K-means clustering optimized by the elbow method," *Jurnal Riset Mahasiswa Matematika*, vol. 4, no. 5, pp. 273–286, 2025, doi: 10.18860/jrmm.v4i5.33389.

[15] C. Lin et al., "Clinical informatics accelerates health system adaptation to the COVID-19 pandemic: Examples from Colorado," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 12, pp. 1955–1963, 2020, doi: 10.1093/jamia/ocaa171.

[16] W. Cho, S. Choi, and H. Choi, "Human resources analytics for public personnel management: Concepts, cases, and caveats," *Administrative Sciences*, vol. 13, no. 2, p. 41, 2023, doi: 10.3390/admsci13020041.

[17] L. Ayanponle, C. G. Okatta, and D. Ajiga, "AI-powered HR analytics: Transforming workforce optimization and decision-making," *Int. J. Sci. Res. Archive*, vol. 5, no. 2, pp. 338–346, 2022, doi: 10.30574/ijsra.2022.5.2.0057.

[18] N. Wang, X. Zhang, S. Li, and G. Xue, "Applications of artificial intelligence in enterprise human resource management," *Inf. Resources Manage. J.*, vol. 38, no. 1, pp. 1–19, 2025, doi: 10.4018/IRMJ.389707.

[19] E. J. Patel, K. Modi, and M. H. Bhavsar, "Employee performance evaluation using machine learning," *Int. J. Adv. Eng. Manage.*, vol. 6, no. 11, pp. 160–164, 2024, doi: 10.35629/5252-0611160164.

[20] I. Fetahović, E. Mekić, K. Kuk, B. Popović, and E. Ć. Dolićanin, “Ethical challenges in open learning analytics,” *Sci. Publ. State Univ. Novi Pazar, Ser. A*, vol. 15, no. 2, pp. 73–85, 2023, doi: 10.46793/SPSUNP2302.073F.

[21] R. Tiwari, “Ethical and societal implications of AI and machine learning,” *Int. J. Sci. Res. Eng. Manage.*, vol. 7, no. 1, pp. 1–10, 2023, doi: 10.55041/ijsrem17519.

[22] B. Theiling et al., “Science autonomy for ocean worlds astrobiology: A perspective,” *Astrobiology*, vol. 22, no. 8, pp. 901-913, 2022, doi: 10.1089/ast.2021.0062.

[23] A. Kumar Agrawal, “Awareness of side effects of application of HR analytics,” *Int. J. Sci. Res. Eng. Manage.*, vol. 9, no. 5, pp. 1–4, 2025, doi: 10.55041/ijsrem49074.

[24] M. Hoche, O. Mineeva, G. Rätsch, E. Vayena, and A. Blasimme, “What makes clinical machine learning fair? A practical ethics framework,” *PLOS Digital Health*, vol. 4, no. 3, p. e0000728, 2025, doi: 10.1371/journal.pdig.0000728.