Review of Computer Engineering Research
2026 Vol. 13, No. 4, pp. 268-290
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i4.5373
© 2026 Conscientia Beam. All Rights Reserved.

A hybrid unsupervised learning framework for Big Five personality-based profiling and mental health risk inference among undergraduate students

Satish N Bhadane[1+]
Kuraparthi TV Reddy[2]
Pradeep Patil[3]
Prateek Verma[4]

[1]*Department of Artificial Intelligence and Data Science, Faculty of Engineering and Technology, Datta Meghe Institute of Higher Education and Research (DU), Wardha, Maharashtra, India.*
[1]*Email: satish.bhadane@gmail.com*
[2]*Department of Artificial Intelligence and Machine Learning, Faculty of Engineering and Technology, Datta Meghe Institute of Higher Education and Research, Wardha, Maharashtra, India.*
[2]*Email: ktvreddy.feat@dmiher.edu.in*
[3]*Department of Psychiatry, Faculty of Engineering and Technology, Datta Meghe Medical College, Nagpur, Maharashtra, India.*
[3]*Email: dr.p.s.patil2012@gmail.com*
[4]*Department of Artificial Intelligence and Machine Learning, School of Engineering, Dayananda Sagar University, Harohalli, Ramanagara, Karnataka, India.*
[4]*Email: prateekverma.aiml@dsu.edu.in*



(+ Corresponding author)

Article History
Received: 19 May 2026
Revised: 28 August 2026
Accepted: 24 September 2026
Published: 20 November 2026

Keywords
Big five personality traits
Educational data mining
Hybrid clustering
Mental health risk inference
Psychological profiling
Unsupervised learning.

ABSTRACT

This study describes the development of a hybrid unsupervised learning approach for Big Five personality profiling and the inference of mental health risk among college students. The goals were to identify meaningful patterns of personality clustering, detect atypical profiles, and relate these patterns to non-clinical measures of stress, anxiety, depression, and psychological well-being in academic settings. Our final design combined DBSCAN for outlier detection and K-means for stable clustering, followed by principal component analysis for visualization. This approach was applied to a real dataset of 1,214 student responses obtained with a well-validated, 50-item instrument measuring the dimensions of Extraversion, Neuroticism, and Agreeableness. The findings showed that the optimal clustering configuration was obtained at K = 2, with a silhouette score of 0.304, yielding three interpretable groups: an atypical/outlier cluster, an emotionally vulnerable cluster, and an adaptive/balanced cluster. High scores on measures of Neuroticism were strongly predictive of negative mental health risk patterns. In contrast, high scores on measures of Extraversion, Agreeableness, and Conscientiousness were associated with high levels of psychological well-being. Statistical validation using Spearman correlation and one-way ANOVA confirmed substantial differences among clusters, especially in measures of anxiety and psychological well-being. Overall, these results support the utility of hybrid unsupervised learning methods for student profiling, early detection of vulnerability patterns, and the design of individualized counseling and academic support programs within non-clinical college settings.

Contribution/Originality: This study contributes to the existing literature on student personality profiling. It uses a new hybrid unsupervised clustering methodology and is one of the few studies investigating mental-health risk. The paper contributes the first clustering-validation analysis. Its primary contribution is finding that neuroticism differentiates vulnerable groups. The study documents interpretable clusters.

© 2026 Conscientia Beam. All Rights Reserved.

1. INTRODUCTION

Personality has become an important focus of educational research because stable individual differences influence how students think, behave, interact, respond to academic demands, and adapt to changing learning environments [1]. In higher education, awareness of personality characteristics is especially useful for understanding variation in social adjustment, emotional stability, academic engagement, resilience, and susceptibility to mental health difficulties [1-3]. Because student well-being is closely connected with academic performance and longer-term psychological development, personality-based analysis can support more individualized counseling, early intervention, and educational support strategies [1, 3, 4].

Among the available personality frameworks, the Big Five model remains one of the most widely accepted and empirically validated approaches for describing stable dimensions of human personality [2, 5, 6]. The five traits, extraversion, neuroticism, agreeableness, conscientiousness, and openness, provide a coherent representation of behavioral, emotional, and cognitive tendencies across populations. In educational and mental health settings, these dimensions are especially relevant because they reflect sociability, emotional regulation, self-discipline, cooperation, adaptability, and psychological well-being. Prior studies have consistently shown that high neuroticism is associated with increased stress, anxiety, and other adverse mental health outcomes. In contrast, extraversion, agreeableness, and conscientiousness are more often associated with coping capacity, adjustment, and psychological well-being [3, 4, 7].

Traditional methods of personality assessment have primarily utilized psychometric instruments, manual interpretation, and descriptive statistical procedures [2, 5, 6]. Although theoretically sound and clinically informative, these methods are relatively ineffective at detecting latent substructures within large student samples. In general, they provide descriptions of personality at the level of the individual or aggregate group, but do not reflect naturally occurring patterns of personality or unusual configurations with potential psychological significance [2, 5, 6]. Moreover, conventional methods of personality analysis are largely disconnected from subsequent interpretations of tendencies toward mental health. They are therefore of limited value for developing data-driven systems for student care.

Recent developments in machine learning methods have provided new opportunities for the analysis of personality and facilitated the automated detection of patterns in highly structured verbal, visual, and behavioral data [8-10]. Applications of supervised and deep learning methods have yielded excellent results in personality prediction, especially under verbal, facial, and multimodal conditions [8-12]. However, the application of supervised methods required labeled data that were costly to obtain, potentially subjective, and difficult to apply with high reliability in clinical and educational settings [13-15]. Patterns of personality and mental health were generally continuous and heterogeneous rather than readily separable into rigid, predefined categories. Consequently, the use of supervised classification methods resulted in considerable oversimplification of the actual diversity of patterns of psychological organization [13-15]. This experience further illustrates the need for unsupervised analysis methods that allow the direct identification of underlying patterns of behavior.

Clustering-based analysis offers a particularly useful alternative in this context [16-19]. Unsupervised clustering can identify natural groupings, detect outliers, and reveal hidden configurations of behavior that may remain undetected in conventional analyses [16-19]. K-means remains one of the most widely applied clustering methods because of its efficiency and interpretability. In contrast, density-based methods are especially effective for identifying atypical observations that should not be forced into centroid-defined groups [16-19]. More recent studies also suggest that hybrid machine learning and clustering-based approaches can provide more robust and informative descriptions of personality patterns than isolated single-method approaches [13, 18, 20, 21].

Despite these advances, important limitations remain in the existing literature. Prior studies have generally focused on one of three directions: supervised personality prediction, multimodal deep learning for personality estimation, or single-method clustering analysis [3, 7, 9, 13, 20]. As a result, the literature remains fragmented.

© 2026 Conscientia Beam. All Rights Reserved.

There has been relatively limited effort to combine hybrid unsupervised clustering, outlier-sensitive personality grouping, statistical validation, and the systematic inference of mental health-related tendencies within a single framework designed specifically for undergraduate students. In particular, the use of Big Five traits to derive psychologically meaningful student clusters and then relate those clusters to stress, anxiety, depression, and well-being through rigorous statistical analysis remains underexplored. Recent interest in AI-based stress analytics further reinforces the need for more integrative approaches that connect behavioral profiling with mental health-related interpretation [22].

To address this gap, the present study proposes a hybrid unsupervised learning framework for Big Five personality-based profiling and mental health risk inference among undergraduate students. We combined density-based and centroid-based clustering methods to detect dominant personality configurations and rare profiles. We then applied correlation- and variance-based statistical techniques to evaluate differences in mental health among clusters. Our results were obtained from a real sample of 1,214 students and therefore represent a highly practical approach to educational and clinical applications.

The main contributions of this work are as follows:

- A hybrid DBSCAN + K-means clustering approach was used to identify dominant and atypical personality profiles from data on the Big Five personality traits.
- An inference model linking personality to mental health risk was applied to relate personality patterns to tendencies toward stress, anxiety, depression, and psychological well-being.
- Results from statistical validation using Spearman correlations, one-way ANOVA, and effect-size analyses confirmed the interpretability and robustness of our results.
- Application of our method to a real student sample of 1,214 subjects confirmed the utility of our approach for practical applications in educational and mental health assessment.

The organization of the remaining sections is as follows. Section 2 reviews related work on personality analysis, machine learning-based personality prediction, clustering methods, and personality-linked mental health studies. Section 3 presents the proposed methodology, including the dataset description, preprocessing, mathematical formulation, clustering framework, and validation strategy. Section 4 presents the experimental results and psychological interpretation of the identified clusters. Section 5 discusses the findings, practical implications, and limitations of the study. Finally, Section 6 concludes the paper and outlines directions for future work.

2. LITERATURE REVIEW

2.1. Big Five Personality Modeling

The Big Five model remains one of the most widely accepted ways of describing stable dimensions of personality because it represents individual differences in behavior in a theoretically well-justified and empirically robust manner. In educational settings, the model is especially relevant because the five personality dimensions are associated with patterns of student engagement, emotional regulation, adaptability, and academic coping. Results of previous studies consistently support associations between high neuroticism and vulnerability to psychological distress and between high scores on the dimensions of extraversion, agreeableness, and conscientiousness and increased resilience and adaptive adjustment. As a consequence, the Big Five model is well suited to student-oriented descriptions of personality and to subsequent interpretation of tendencies toward psychological health or illness.

However, most applications of the Big Five model have focused primarily on descriptions of personality traits or direct predictions of behavior and have not fully utilized the potential for identifying latent subgroups of students through methods of unsupervised analysis. Although reports of cluster-based personality types have been made, these studies generally remain largely descriptive and fail to achieve full validation of patterns of adjustment associated with various tendencies toward psychological health or illness. Consequently, the present study represents an

© 2026 Conscientia Beam. All Rights Reserved.

appropriate extension of the Big Five conceptual framework. Still, it does not provide sufficient integration of cluster-oriented methods for interpreting student personality and risk for psychological illness.

2.2. Personality Prediction using ML/DL

Machine learning and deep learning methods have greatly expanded the scope of personality prediction, enabling the automatic detection of personality-related patterns in text, images, speech, and multimodal data. Models based on text and social media have provided evidence for inferring personality from linguistic and behavioral data. Methods involving facial images and body language have further demonstrated that deep learning models can identify overt aspects of personality from visual data. Subsequent work with transformer-based architectures, graph-oriented models, and multimodal deep learning has further increased representational richness and prediction accuracy.

Nevertheless, supervised methods of personality prediction remain the major focus of this line of research. These models require labeled data and are primarily designed to maximize prediction accuracy rather than to identify naturally occurring dimensions of personality. In educational datasets, where latent patterns of psychological functioning may be continuous, overlapping, or minimally labeled, this dependence constitutes a major limitation. In addition, most recent studies using multimodal data have primarily focused on the overt assessment of personality rather than on developing student-oriented representations of psychological functioning. Consequently, although the technical performance of personality prediction has substantially improved through the use of supervised and multimodal methods, these approaches have not achieved the major objective of the present study, which is to identify interpretable patterns of student personality and relate them to non-clinical manifestations of risk for mental illness within a single, fully unsupervised framework.

2.3. Clustering in Psychological Data

Clustering offers a meaningful alternative to rigid class-based analyses because psychological data are inherently heterogeneous and often do not fit well into fixed categories. K-means remains widely used for its computational simplicity and ease of interpretation. In contrast, hierarchical methods provide nested patterns of group membership, while density-based methods permit the identification of outliers and unusual cases. In addition, recent work supports the use of hybrid methods and combinations of clustering and classification for personality applications. However, previous studies of clustering in personality research reveal several important limitations.

First, most studies emphasize the performance of clustering or the discovery of personality types but do not extend the results to yield meaningful psychological interpretations. Second, attention to outlier handling is generally inadequate, despite the high clinical importance of unusual personality profiles in student mental health settings. Finally, the results of most studies are not evaluated for psychological significance using correlational analyses, ANOVA, or measures of effect size. Consequently, although clustering is clearly well-suited to applications in personality research, the field still lacks a basis for integrating outlier-sensitive clustering methods, interpretable descriptions of student profiles, and rigorous tests of psychological significance.

2.4. Mental Health Inference Studies

The association between personality and mental health has been well established in both psychological and computational literature. Neuroticism consistently represented the strongest trait-level predictor of stress, anxiety, and depressive symptoms. In contrast, the dimensions of extraversion, agreeableness, and conscientiousness were more strongly associated with positive well-being and adaptive adjustment. These results confirm the utility of using personality as a theoretically meaningful basis for inferences about patterns of mental health risk, especially in settings of higher education where direct clinical labels may not always be available. However, most previous studies of personality and mental health employed conventional statistical or supervised prediction methods.

© 2026 Conscientia Beam. All Rights Reserved.

Very few studies used unsupervised clustering methods to identify subgroups of students with similar personality characteristics and to relate these subgroups to underlying dimensions of mental health risk. This represents an important limitation, because patterns of personality identified by unsupervised clustering methods were sufficient to generate clinically meaningful profiles of mental health risk that were not detectable by conventional statistical methods. Consequently, our results confirm the validity of associations between personality and mental health, but do not support the use of an integrated system of clustering-based methods for interpreting student patterns of mental health risk.

2.5. Comparative Synthesis of Prior Studies

Table 1 shows the comparative synthesis of representative studies related to personality modeling, prediction, clustering, and mental health interpretation.

Table 1. Comparative synthesis of representative studies related to personality modeling, prediction, clustering, and mental health interpretation.

Ref.	Study focus	Data/modality	Main method	Main contribution	Limitation relative to present study
Dalvi-Esfahani, et al. [7]	Personality and psychosocial outcomes	Student/social data	Statistical modeling	Shows moderation effect of personality traits	Does not use clustering or integrated mental health profiling
Romeo, et al. [8]	Apparent personality prediction	Body language / visual	Deep learning	Demonstrates personality estimation from behavioral signals	Supervised and not student-clustering-oriented
Ahmad, et al. [9]	Text-based personality classification	Text	Hybrid deep learning	Improves prediction accuracy from textual inputs	Requires labeled data; no clustering or mental health inference
Xu, et al. [5]	Big Five prediction from facial images	Static facial images	Deep learning	Shows feasibility of visual personality prediction	Supervised and not linked to student mental health profiling
Wang, et al. [13]	Clustering algorithms in personality prediction	Personality data	Comparative clustering	Highlights importance of algorithm choice	Limited psychological interpretation and validation
Gjurković and Šnajder [11]	Personality prediction from Reddit data	Online text	Transformer model	Improves personality prediction from online communication	Focuses on supervised prediction, not subgroup discovery
Chi and Chi [4]	Personality type discovery	Big Five responses	Cluster analysis	Identifies meaningful personality types	Does not extend clusters toward mental health risk interpretation
Zhang, et al. [18]	Hybrid unsupervised personality profiling	Personality data	Hybrid unsupervised learning	Supports value of unsupervised profiling	Limited focus on student mental health and validation depth
Zhou, et al. [20]	Hybrid ML models for personality prediction	Personality data	Clustering + classification	Shows benefit of hybrid models	Emphasis remains predictive rather than interpretive
Sauls, et al. [3]	Personality and mental health outcomes	Psychological data	Statistical analysis	Confirms strong trait–mental health associations	No unsupervised clustering framework
Lukac [10]	Multimodal personality prediction	Speech / multimodal	Deep learning	Expands personality prediction beyond text and images	Focuses on prediction, not student-centered clustering
Aslan, et al. [21]	Multimodal apparent	Visual/Multimodal	Attention-based deep model	Improves multimodal	Apparent personality focus, not student profiling

© 2026 Conscientia Beam. All Rights Reserved.

Ref.	Study focus	Data/modality	Main method	Main contribution	Limitation relative to present study
	personality assessment			personality assessment	
Chen, et al. [23]	Hierarchical personality detection	Personality data	Brain-inspired deep model	Captures complex multi-level patterns	Not focused on clustering or mental health inference
Shen, et al. [24]	Graph-based personality detection	Personality data	Self-supervised graph model	Improves representation learning without standard supervision	Not designed for student subgroup interpretation
Dael, et al. [25]	MBTI personality classification	Structured data	Machine learning	Supports data-driven personality classification	MBTI-focused and not linked to Big Five mental health profiling
Ouarka, et al. [26]	Multimodal personality trait prediction	Multimodal data	Deep multimodal fusion	Shows robustness of multimodal integration	Focuses on prediction rather than interpretable clustering

2.6. Research Gap

The literature reveals four major trends: First, the Big Five remains a powerful basis for personality description and analysis. Second, supervised and deep learning methods have substantially improved personality prediction. Third, clustering methods increasingly represent a valuable means of describing patterns of personality. Finally, personality dimensions are strongly associated with various indices of mental health status. Nevertheless, these approaches remain largely separate. Most studies address only supervised methods of personality prediction, isolated clustering techniques, or conventional relationships between personality and mental health status. Very few studies integrate hybrid methods of unsupervised clustering, sensitive identification of outlier subgroups, statistical validation of results, and inference of risk for nonclinical indices of mental health within a single, student-oriented framework. Here, we extend these efforts and describe a unified hybrid application of DBSCAN and K-means that identifies typical and atypical personality patterns, provides statistical validation of their psychological significance, and interprets the results in terms of the risk of stress, anxiety, and depression, as well as overall indices of well-being. As a result, our work goes well beyond purely descriptive clustering and conventional supervised personality prediction, producing a fully interpretable, student-oriented approach to personality-based profiling and mental health risk prediction.

3. PROPOSED METHODOLOGY

We apply a hybrid unsupervised learning approach to obtain personality-based student profiles and infer non-clinical risk tendencies related to stress, anxiety, depression, and well-being from Big Five trait data. In contrast to previous studies that largely relied on supervised methods for personality prediction or single clustering techniques, our approach integrates density- and centroid-based clustering sequentially. First, we use DBSCAN to identify atypical or sparse personality patterns, then apply K-means clustering to the remaining well-defined observations to obtain stable, easily interpretable student profiles. This approach differs from previous work in two major ways. First, it avoids imposing an artificial structure on irregular patterns of psychological adjustment. Second, it allows the integration of clustering results with statistical criteria and theoretically meaningful interpretations of individual and group risk factors for stress, anxiety, depression, and well-being. Consequently, our method is directed not only toward computational grouping but also toward generating exploratory and psychologically meaningful student profiles for an actual educational setting. Figure 1 presents the methodology for the clustering model using personality traits.

© 2026 Conscientia Beam. All Rights Reserved.

3.1. Reliability of Personality Measurements

The Big Five personality traits were assessed with a standardized 50-item instrument containing both positively worded and reverse-coded items. To obtain estimates of sample-specific internal consistency for the present data, we evaluated reliability using Cronbach's alpha, McDonald's omega, the split-half correlation, the Spearman-Brown coefficient, the Guttman split-half coefficient, and the mean inter-item correlation. The resulting reliability indices are presented in Appendix B to facilitate interpretation and represent estimates of sample-specific internal consistency rather than a revalidation of the original standardized instrument. Reliability coefficients ranged from 0.186 to 0.351 for Cronbach's alpha, from 0.001 to 0.269 for McDonald's omega, and from 0.023 to 0.055 for the mean inter-item correlation. These results indicate low-to-moderate sample-specific internal consistency but confirm the instrument's previously established overall psychometric validity.

3.2. Parameter Selection, Grid Search, and Stability Assessment

To ensure that the clustering solution was not dependent on arbitrary parameter choices, a structured parameter-selection procedure was applied. For K-means clustering, a grid-search strategy was used over values of K ranging from 2 to 10, and clustering quality was evaluated using the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. The optimal number of clusters was selected as K = 2 because it yielded the highest Silhouette Score (0.304344), the highest Calinski–Harabasz value (574.4685), and the lowest Davies–Bouldin value (1.300216) among the tested solutions.

Parameter selection for DBSCAN was performed via a grid search to balance outlier detection and the formation of stable core clusters. The optimal DBSCAN configuration was obtained with $\varepsilon = 1.15$ and MinPts = 5, which produced 2 density-based clusters, identified 90 noise observations (7.41% of the dataset), and yielded a DBSCAN silhouette score of 0.4456. These results indicate a meaningful density-based structure in the standardized Big Five personality data, with 1,124 non-noise observations assigned to clusters.

To test the robustness of the results, we performed repeated-run stability analysis. This yielded an average silhouette score of 0.3043, with a standard deviation of (5.55×10^{-17}) and virtually identical minimum and maximum values, reflecting negligible variation in the K-means solution. In addition, a bootstrap-based stability assessment was performed over 100 iterations. The mean Adjusted Rand Index (ARI) was 0.922, with a standard deviation of 0.080, indicating excellent reproducibility of the major cluster structure across bootstrap samples. These results support the stability of the proposed hybrid clustering framework.

3.3. Data Availability Statement

The dataset used in this study contains questionnaire-based personality responses and associated demographic information collected from undergraduate students. Due to privacy and ethical considerations, raw respondent-level data are not fully publicly available during peer review. De-identified data may be made available by the corresponding author upon reasonable request and subject to institutional approval. Upon publication, the anonymized dataset will be deposited at: https://github.com/satishbhadane-lgtm/Personality_Prediction-/blob/main/Personality_Dataset.csv

3.4. Appendix Note

The survey instrument used in this study is provided in Appendix A, which contains the 50 questionnaire items mapped to the five Big Five personality dimensions, along with their positively and reverse-coded structures. Reliability statistics for the personality measurements are provided in Appendix B.

© 2026 Conscientia Beam. All Rights Reserved.

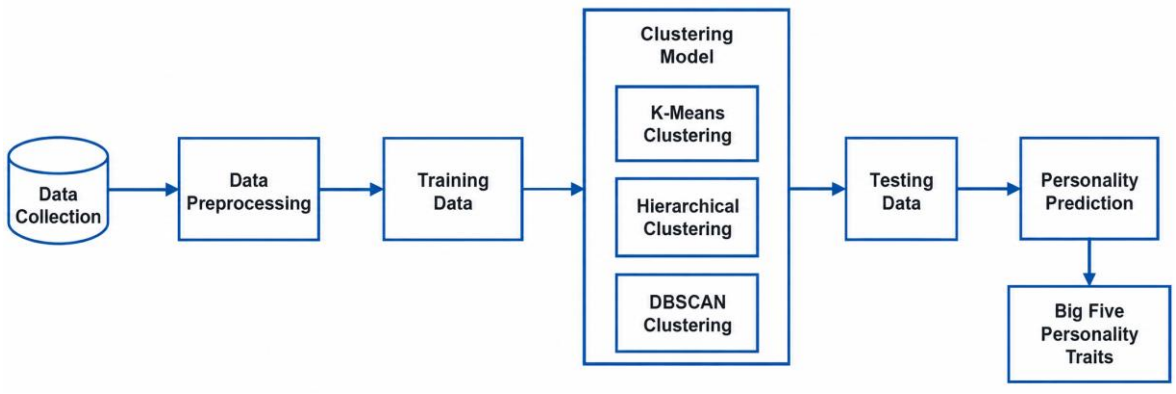


Figure 1. Methodology for clustering model using personality traits.

3.5. Dataset Description

The dataset used in this study contains responses from 1,214 students, collected primarily from undergraduate learners through a structured, Google Forms-based personality assessment. The original data shape is (1,214, 54), comprising 50 questionnaire items mapped to the Big Five personality traits and 4 demographic variables. The five trait dimensions considered are Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness.

In addition, demographic characteristics, including age, gender, and family background, were included to allow for subgroup analyses and interpretation. The instrument included both positively and negatively worded items, for which responses were obtained using a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). Combinations of questionnaire items generated profiles for each subject across the five major dimensions of personality. Subsequent assessment of clustering adequacy yielded positive variances across all dimensions, indicating sufficient variability to support the development of clusters of subjects with similar patterns of personality characteristics. The extent of clustering was quantified using a Hopkins statistic of 1.00.

3.6. Data Preprocessing

The raw questionnaire data were processed to ensure consistency, comparability, and suitability for multivariate analysis. This required three major operations. First, we identified missing or incomplete responses and cleaned the data to produce complete records for subsequent analysis. Valid scores were computed only for complete responses. Second, responses to negatively worded items were converted to ensure a consistent interpretation across all items contributing to each personality dimension. This was necessary because both positive and negative formulations were included for each personality dimension. Third, standardized scores for the resulting aggregated personality dimensions were obtained to prevent distortion of the clustering process caused by imbalances in scale factors. Our procedures required all personality dimensions to be standardized before clustering to avoid subsequent distortion of the cluster analyses caused by imbalances in scale factors. Consequently, the processed data used to generate the cluster analyses represented standardized vectors for all personality dimensions for each student and maintained the full expression of variability in personality dimensions within Euclidean space.

3.7. Mathematical Formulation

3.7.1. Feature Representation

A five-dimensional Big Five trait vector represents each student.

$$\mathbf{x}_i = [E_i, N_i, A_i, C_i, O_i] \quad (1)$$

Where E_i, N_i, A_i, C_i, and O_i denote the Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness scores, respectively, for the i-th student; this representation forms the core feature space for clustering.

© 2026 Conscientia Beam. All Rights Reserved.

3.7.2. Euclidean Distance

The similarity between any two students, xi and xj, is computed using Euclidean distance.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{5}(x_{ik} - x_{jk})^2} \quad (2)$$

This distance measure reflects overall trait-based similarity among student personality profiles and forms the basis for both density and centroid clustering.

3.7.3. DBSCAN Model

In the first stage of clustering, the DBSCAN algorithm is applied to identify dense regions and isolate outlier personality profiles. For any point p, the ε-neighborhood is defined as follows:

$$N_\varepsilon(p) = \{q \in D \mid dist(p,q) \leq \varepsilon\} \quad (3)$$

Here, D denotes the dataset, ε denotes the neighborhood radius, and dist(p,q) denotes the distance between points p and q. A point is considered a core point if its ε-neighborhood contains at least MinPts observations. Points that do not satisfy the density requirement and are not density-connected to a core point are considered noise. In the present work, these points were assigned to cluster -1 and interpreted as psychologically extreme or unusual patterns of behavior rather than measurement errors. This approach prevented aberrant observations from distorting subsequent centroid-based clustering methods.

3.7.4. K-Means Objective Function

Following the removal of noise points identified by DBSCAN, K-Means clustering is applied to the resulting core data. This procedure partitions the data into k clusters by minimizing the within-cluster sum of squares.

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \| x - \mu_i \|^2 \quad (4)$$

Where C_i denotes the i-th cluster, and μ_i represents its centroid, this formulation ensures that cluster members are grouped around representative centroids with minimal intra-cluster variation. The uploaded results indicate that the optimal number of clusters, selected using silhouette analysis, is k=2, with a final silhouette score of 0.304, indicating moderate but meaningful separation among student personality structures.

3.7.5. PCA Transformation

Principal Component Analysis (PCA) projects the standardized trait matrix into a lower-dimensional space for visualization.

$$Z = XW \quad (5)$$

Here, X is the standardized data matrix, and W contains the principal component eigenvectors. PCA is performed only after clustering and provides a two-dimensional display of the first two principal components for visual assessment of the degree of separation between clusters. However, PCA does not provide a basis for clustering and is used solely to facilitate the interpretation and visual appreciation of the results. This approach is described in detail in the submitted methodological notes.

3.8. Hybrid Clustering Framework

The novelty of our approach lies in a hybrid architecture that leverages the complementary advantages of DBSCAN and K-means rather than relying on a single clustering algorithm. The complete framework follows four sequential steps:

1. Apply DBSCAN to detect noise and atypical student profiles
   This stage identifies sparse observations and assigns them to Cluster −1 to avoid bias in centroid estimation.
2. Apply K-means to the remaining core data

© 2026 Conscientia Beam. All Rights Reserved.

Once noisy observations are removed, K-means identifies the dominant personality groups within the student population.

3. Validate cluster quality using silhouette analysis

Different values of k are examined, and the clustering configuration with the highest silhouette score is selected. In this study, the best result was obtained at $k = 2$, with a silhouette score of 0.304.

4. Visualize cluster structure using PCA

The final cluster labels are projected into a two-dimensional PCA space to illustrate compactness, relative separation, and the position of outliers.

The proposed hybrid architecture for clustering is illustrated in Figure 2. Outliers were detected using DBSCAN, and subsequent clustering of the remaining core observations produced stable personality-type configurations. This method was especially appropriate for psychological data because it was sensitive to outliers and robust to irregular patterns of data organization. By contrast, K-means cluster analysis produced stable cluster-organization patterns and interpretable representations of core student populations.

The resulting overall cluster structure was interpreted as follows:

Cluster -1 = Atypical/outlier personality cluster.

Cluster 0 = Emotionally vulnerable cluster.

Cluster 1 = Adaptive, balanced personality cluster.

These descriptive labels reflect theoretical considerations and represent an exploratory approach to cluster interpretation. Consequently, they should not be considered diagnostic classifications.

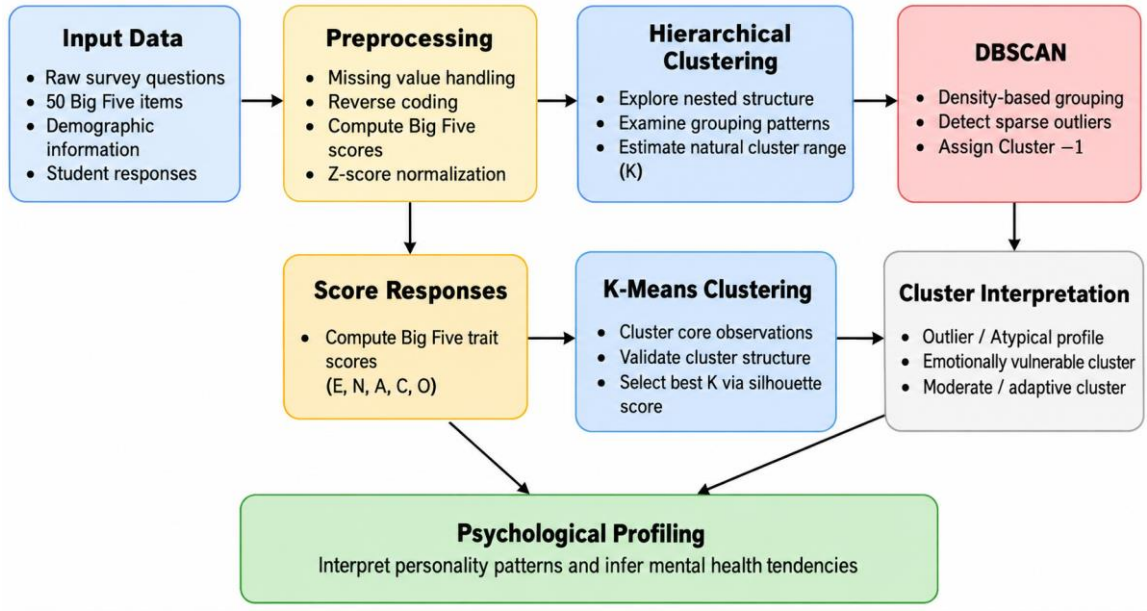


Figure 2. Hybrid clustering for personality prediction.

3.9. Mental Health Inference Model

After the personality clusters are formed, a theory-driven mental health inference layer is applied. In this model, the independent variables (IVs) are the Big Five personality traits.

- Extraversion.
- Neuroticism.
- Agreeableness.
- Conscientiousness.
- Openness.

© 2026 Conscientia Beam. All Rights Reserved.

The dependent variables (DVs) are four inferred mental health indicators.

- Stress.
- Anxiety.
- Depression.
- Psychological Well-Being.

The mental health indices reflect latent, non-clinical psychological tendencies attributable to personality structure, not clinical diagnosis. This is particularly important because the model does not provide a basis for diagnostic interpretation; instead, it describes expected relationships among patterns of emotional experience and levels of well-being associated with various configurations of personality traits. The resulting inferences strongly emphasize the combined influences of personality characteristics and environmental demands on experiences of stress, anxiety, depression, and well-being, and maintain a primarily psychological and exploratory character rather than a clinical one. The overall pattern of inference was based on trait expectations. For example, high levels of neuroticism were associated with relatively unfavorable patterns of experience with stress, anxiety, and depression. In contrast, high levels of extraversion, agreeableness, and conscientiousness were associated with relatively favorable patterns of well-being. Consequently, interpretations of trait clusters primarily reflected an exploratory, non-clinical description of the relationships between personality characteristics and mental health risk, rather than providing a basis for clinical diagnosis.

3.10. Statistical Validation

To ensure that the proposed framework is computationally meaningful and psychologically interpretable, the clustered outputs are validated using correlation analysis, group-difference testing, and effect-size estimation.

3.10.1. Spearman Correlation

Spearman's rank correlation coefficient (ρ) is used to measure monotonic relationships between the Big Five traits and the inferred mental health indicators. This nonparametric method is suitable for personality data obtained with Likert-type scales, for which stringent assumptions of normality are unnecessary. Results confirmed strong positive relationships between Neuroticism and stress, anxiety, and depression, as well as low levels of well-being. In contrast, positive relationships between Extraversion, Agreeableness, and Conscientiousness and well-being were greater in magnitude.

3.10.2. One-Way ANOVA

One-way ANOVA was used to test for significant differences in inferred indices of mental health among the identified personality clusters. This provided a quantitative basis for evaluating the extent to which the clusters differed in inferred patterns of psychological adjustment. Results supported substantial differences among clusters in indices of stress, anxiety, depression, and well-being, with all p-values < 0.001.

3.10.3. Effect Size (η^2) and Pairwise Comparisons

To assess the practical magnitude of cluster-based differences, effect size was measured using eta squared (η^2). This provides information beyond statistical significance by quantifying the proportion of variance in each mental health indicator explained by cluster membership. Anxiety ($\eta^2 = 0.167$) and Well-Being ($\eta^2 = 0.162$) exhibited large effects, while Stress ($\eta^2 = 0.068$) and Depression ($\eta^2 = 0.059$) showed moderate effects. These results indicate that cluster membership is more strongly related to anxiety and well-being than to stress and depression.

To obtain additional estimates of pairwise differences between clusters, we also calculated Welch's independent-samples t-tests, effect sizes, and 95% confidence intervals. This yielded more precise estimates of the magnitude and reliability of differences among pairs of personality clusters.

© 2026 Conscientia Beam. All Rights Reserved.

The overall approach integrates psychometric structure, hybrid unsupervised clustering, dimensionality reduction, exploratory psychological inference, and statistical validation into a unified system for analyzing student personality and mental health. As a result, the methodology is not only technically sound but also highly interpretable and applicable to educational and health-oriented settings.

4. RESULTS AND ANALYSIS

This work reports empirical results from our hybrid clustering approach for Big Five-based student profiling and mental health inference. Our analysis begins with validation of the dataset. It continues with an evaluation of cluster quality, interpretation of cluster distributions, visualization using PCA, and, finally, assessment of trait–mental health associations, effect-size-supported statistical validation, and demographic patterns. The results confirm the dataset's suitability for unsupervised analysis and support the validity of our hybrid DBSCAN + K-means approach in producing psychologically meaningful clusters with clear implications for inferred mental health patterns.

4.1. Dataset Validation

The adequacy of the data for unsupervised clustering was evaluated using analyses of variance and the Hopkins statistic. The resulting dataset represents 1,214 student responses to questionnaire items and demographic characteristics and includes ratings on all five dimensions of the Big Five personality construct. Trait-level variance values were positive for all five personality dimensions (i.e., Extraversion, 0.203; Neuroticism, 0.131; Agreeableness, 0.243; Conscientiousness, 0.157; and Openness, 0.142). These results confirm that sufficient variability exists in the data to justify clustering procedures, rather than merely representing a highly homogeneous sample. Furthermore, the Hopkins statistic value (1.00) was well above the threshold typically associated with strong clustering tendencies. Therefore, both indices of data adequacy supported the appropriateness of applying unsupervised personality classification methods and multivariate pattern analysis. The variance values and the results of the Hopkins statistic used to evaluate data adequacy are presented in Table 2.

Table 2. Trait variance analysis and Hopkins statistic for dataset validation.

Measure	Variables	Value	Interpretation
Variance	Extraversion	0.203	Adequate dispersion
Variance	Neuroticism	0.131	Adequate dispersion
Variance	Agreeableness	0.243	Adequate dispersion
Variance	Conscientiousness	0.157	Adequate dispersion
Variance	Openness	0.142	Adequate dispersion
Hopkins Statistic	Dataset-level	1.00	Very strong clustering tendency

4.2. Clustering Results

The hybrid framework uses DBSCAN to isolate noise points and K-Means to discover stable clusters among the remaining core observations. To determine the appropriate number of K-Means clusters, silhouette analysis was performed across multiple values of K. The results show that the highest silhouette score was obtained at K=2, with a value of 0.304. For higher values of K, the scores decreased progressively, including 0.213 for K=3, 0.155 for K=4, 0.150 for K=5, 0.151 for K=6, 0.145 for K=7, and 0.148 for K=8. This pattern indicates that the two-cluster solution provides the most compact and well-separated grouping among the non-outlier student profiles. Although the silhouette value reflects moderate separation, such values are common in human behavioral and psychological datasets, where trait overlap is naturally expected. The silhouette-based evaluation of K-Means clustering, presented in Table 3, shows that the best clustering quality was achieved at K=2, with a silhouette score of 0.304.

© 2026 Conscientia Beam. All Rights Reserved.

Table 3. Silhouette scores for different values of K in K-Means clustering.

Number of clusters (K)	Silhouette score
2	0.304
3	0.213
4	0.155
5	0.150
6	0.151
7	0.145
8	0.148

As shown in Figure 3, the silhouette score reaches its maximum at K = 2, indicating that the two-cluster solution provides the best separation among core student personality profiles.

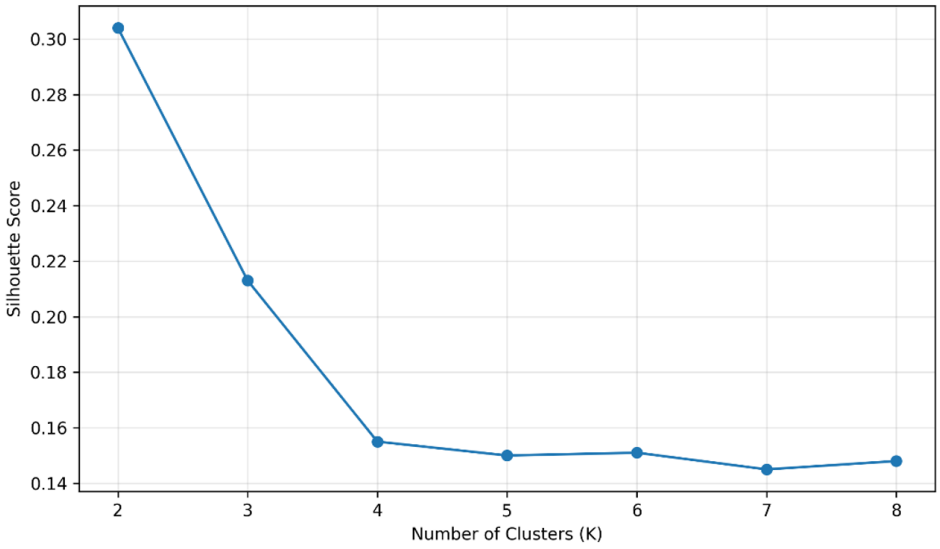


Figure 3. Silhouette score variation across different cluster counts.

4.3. Cluster Distribution

The final cluster solution, following application of the hybrid method, included three labels: clusters 0, 1, and -1. The two core clusters accounted for 342 and 790 students, respectively, while 82 cases were identified as outliers and classified into cluster -1. These cases reflect students with relatively unusual or poorly integrated personality patterns relative to the predominant cluster configuration. Rather than being considered merely artifacts of technical error, these cases were retained because they may represent relatively uncommon but potentially important personality configurations. Consequently, the use of the hybrid method provided improved estimates of cluster centroids and maintained the opportunity for cautious interpretation of atypical cases. The resulting hybrid model produced three cluster labels, with the greatest number of students assigned to cluster 1 (see Table 4).

Table 4. Final cluster-wise distribution of students in the hybrid clustering framework.

Cluster label	Number of students	Interpretation
-1	82	Atypical / Outlier
0	342	Emotionally Vulnerable
1	790	Adaptive / Balanced

4.4. PCA Visualization

To depict the overall pattern of clustering, we applied principal component analysis to reduce the high-dimensional space of Big Five personality traits to two principal components, PC1 and PC2. These axes reflect linear

© 2026 Conscientia Beam. All Rights Reserved.

combinations of standardized trait scores and account for the maximum amount of variance in two dimensions. Most importantly, the use of principal component analysis was solely for data visualization and did not influence cluster assignment, which was based on the original configuration of trait scores using DBSCAN and K-means clustering methods.

The PCA map shows that the three hybrid cluster labels occupy distinguishable regions of the reduced space. Cluster −1 appears more dispersed and less cohesive, Cluster 0 occupies a neighboring region with moderate separation, and Cluster 1 forms a relatively denser and more compact grouping. These patterns support the visual interpretability of the proposed clustering solution, although the observed separation should be understood as exploratory rather than as evidence of sharply distinct psychological classes. The representation of individual students in the PCA space also helps illustrate the relative compactness of Cluster 1, the intermediate positioning of Cluster 0, and the dispersed structure of Cluster −1. Figure 4 illustrates the PCA-based visualization of hybrid clustering.

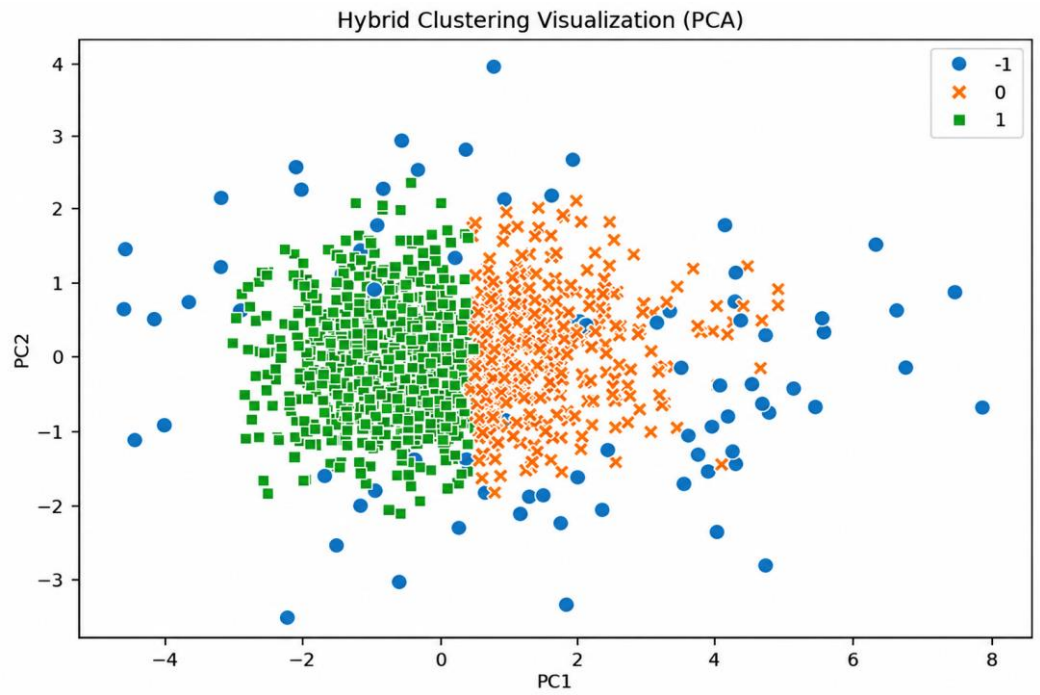


Figure 4. PCA-based visualization of hybrid clustering results in the PC1 - PC2 space.

4.5. Cluster Interpretation

The psychological interpretation of the clusters should be understood as exploratory and theory-informed rather than diagnostic. The final hybrid clustering structure is not treated as a direct measure of mental health status, but rather as a data-driven grouping of personality configurations that may be associated with different patterns of non-clinical mental health risk.

Cluster −1 is interpreted as an atypical or outlier group. These students show relatively unusual combinations of personality traits and appear more sparsely distributed in the reduced PCA space. Rather than being dismissed as technical noise, these cases may represent less common personality configurations that warrant cautious interpretation.

Cluster 0 is interpreted as an emotionally vulnerable group. This cluster is characterized by comparatively higher neuroticism, together with less favorable inferred patterns of stress, anxiety, depression, and psychological well-being. These findings suggest that students in this group may show greater non-clinical vulnerability to emotional distress, although this should not be interpreted as a clinical diagnosis.

© 2026 Conscientia Beam. All Rights Reserved.

Cluster 1 is interpreted as an adaptive/balanced group. This cluster is characterized by lower neuroticism and comparatively more favorable inferred mental health-related patterns. The results suggest that this group reflects a more adaptive personality configuration within the present sample.

The psychological meaning of the final clusters is summarized in Table 5, where Cluster 0 reflects emotional vulnerability, Cluster 1 represents adaptive functioning, and Cluster −1 captures atypical personality profiles.

Table 5. Psychological interpretation of the final hybrid personality clusters.

Cluster	Psychological Interpretation
-1	Atypical/Outlier personality cluster characterized by irregular or non-normative combinations of Big Five traits; represents psychologically meaningful exceptions and distinct adaptation patterns.
0	Emotionally Vulnerable cluster characterized by relatively higher Neuroticism and elevated inferred stress, anxiety, and depressive tendency, with comparatively lower psychological well-being.
1	Adaptive/Balanced cluster characterized by lower Neuroticism and more balanced trait structure, associated with lower psychological burden and better emotional adjustment.

4.6. Correlation Analysis

The relationship between Big Five personality traits and inferred mental health indicators was evaluated using Spearman's rank correlation. Among the five traits, Neuroticism emerged as the strongest predictor of negative mental health burden. It showed a very strong positive correlation with stress (ρ=0.905), an extremely strong positive correlation with anxiety (ρ=0.985), and a very strong positive correlation with depression (ρ=0.895). It also showed a negative association with well-being (ρ = -0.218). These findings indicate that emotional instability is the dominant driver of mental health vulnerability in the present student population.

In contrast, Extraversion showed its strongest contribution in relation to psychological well-being, with a strong positive correlation (ρ=0.625). It also had a weak positive association with anxiety (ρ=0.220) and a weak negative association with depression (ρ=−0.099). Agreeableness and Conscientiousness also displayed positive associations with well-being, with correlation coefficients of 0.596 and 0.496, respectively. These findings suggest that positive, socially adaptive, and self-regulated personality traits serve a protective function, whereas Neuroticism acts as the primary vulnerability dimension. The detailed Spearman correlation results between the Big Five traits and inferred mental health indicators are presented in Table 6.

Table 6. Spearman correlation between big five traits and inferred mental health indicators.

Trait	Mental health indicator	Spearman Rho (ρ)	p-value	Interpretation
Extraversion	Stress	0.011	0.690	Not significant
Extraversion	Anxiety	0.220	< 0.001	Weak positive
Extraversion	Depression	−0.099	< 0.001	Weak negative
Extraversion	Well-Being	0.625	< 0.001	Strong positive
Neuroticism	Stress	0.905	< 0.001	Very strong positive
Neuroticism	Anxiety	0.985	< 0.001	Extremely strong positive
Neuroticism	Depression	0.895	< 0.001	Very strong positive
Neuroticism	Well-Being	−0.218	< 0.001	Weak negative
Agreeableness	Stress	0.156	< 0.001	Weak positive
Agreeableness	Anxiety	0.179	< 0.001	Weak positive
Agreeableness	Depression	0.043	0.139	Not significant
Agreeableness	Well-Being	0.596	< 0.001	Strong positive
Conscientiousness	Stress	0.024	0.409	Not significant
Conscientiousness	Anxiety	0.322	< 0.001	Moderate positive
Conscientiousness	Depression	0.195	< 0.001	Weak–moderate positive
Conscientiousness	Well-Being	0.496	< 0.001	Moderate positive
Openness	Stress	0.130	< 0.001	Weak positive
Openness	Anxiety	0.234	< 0.001	Weak–moderate positive
Openness	Depression	0.100	< 0.001	Weak positive
Openness	Well-Being	0.314	< 0.001	Moderate positive

© 2026 Conscientia Beam. All Rights Reserved.

4.7. ANOVA and Effect Size

To evaluate whether the inferred mental health indicators differed significantly across personality clusters, a one-way ANOVA was performed for stress, anxiety, depression, and psychological well-being. All four indicators showed statistically significant differences across clusters, with $p < 0.001$ in each case. However, the magnitude of these differences varied by indicator, as shown by eta squared (η^2).

Anxiety showed the greatest effect size ($\eta^2 = 0.167$), followed closely by well-being ($\eta^2 = 0.162$), indicating that patterns of personality-cluster membership strongly influence these outcomes. Stress and depression demonstrated moderate effect sizes, reflecting the important, but not exclusive, contributions of personality to these outcomes. In general, these results support the view that stable patterns of personality organization more directly shape anxiety and well-being. In contrast, stress and depression reflect interactions between personality-based predispositions and environmental influences. The ANOVA results provided in Table 7 confirmed that all four measures of mental health differed significantly across the identified personality clusters, with the greatest differences observed for anxiety and well-being. Table 8 presents a pairwise effect-size analysis of mental health indicators across personality clusters.

Table 7. ANOVA and effect size results for mental health indicators across personality clusters.

Mental health indicator	F-statistic	p-value	Eta Squared (η^2)	Interpretation
Stress	44.055	< 0.001	0.068	Moderate effect
Anxiety	121.011	< 0.001	0.167	Large effect
Depression	37.855	< 0.001	0.059	Moderate effect
Well-Being	117.297	< 0.001	0.162	Large effect

Table 8. Pairwise effect size analysis of mental health indicators across personality clusters.

Mental health indicator	Cluster comparison	Mean difference	t	p	Cohen's d	95% CI of d	Effect size	Interpretation
Stress	-1 vs 0	-0.188	-0.908	0.366	-0.160	(-0.401, 0.081)	Negligible	No significant difference
Stress	-1 vs 1	0.386	1.904	0.060	0.399	(0.171, 0.627)	Small	Borderline difference
Stress	0 vs 1	0.575	9.601	<0.001	0.657	(0.528, 0.787)	Medium	Significant difference
Anxiety	-1 vs 0	-0.197	-0.938	0.351	-0.171	(-0.412, 0.070)	Negligible	No significant difference
Anxiety	-1 vs 1	0.692	3.365	0.001	0.756	(0.526, 0.987)	Medium	Significant difference
Anxiety	0 vs 1	0.888	15.811	<0.001	1.101	(0.967, 1.236)	Large	Very strong difference
Depression	-1 vs 0	-0.230	-1.079	0.283	-0.195	(-0.436, 0.047)	Negligible	No significant difference
Depression	-1 vs 1	0.311	1.489	0.140	0.317	(0.089, 0.545)	Small	Not statistically significant
Depression	0 vs 1	0.540	9.134	<0.001	0.620	(0.491, 0.750)	Medium	Significant difference
Well-Being	-1 vs 0	0.162	0.727	0.469	0.144	(-0.097, 0.385)	Negligible	No significant difference
Well-Being	-1 vs 1	0.972	4.396	<0.001	1.014	(0.782, 1.246)	Large	Significant difference
Well-Being	0 vs 1	0.809	15.783	<0.001	1.031	(0.897, 1.165)	Large	Very strong difference

To further quantify the practical significance of the ANOVA findings, pairwise comparisons between personality clusters were performed using Welch's independent-samples (t) test, Cohen's (d), and 95% confidence intervals. The largest differences were observed between Cluster 0 and Cluster 1, particularly for Anxiety (d = 1.101, 95% CI: 0.967 to 1.236) and Well-Being (d = 1.031, 95% CI: 0.897 to 1.165), both of which showed large effects. Stress (d = 0.657)

© 2026 Conscientia Beam. All Rights Reserved.

and Depression (d = 0.620) also differed significantly between these two principal clusters, with medium effect sizes. In contrast, comparisons involving Cluster −1 generally yielded negligible to small effects. They were often statistically non-significant, supporting the interpretation of this group as an atypical or outlier cluster rather than a stable core personality profile.

As shown in Figure 5, the three personality clusters differed substantially in their mental health profiles, with the emotionally vulnerable cluster exhibiting greater distress, anxiety, and depression, and the adaptive cluster exhibiting substantially less psychological morbidity and more stable well-being.

The three personality clusters differed substantially in their patterns of mental health. The vulnerable cluster experienced higher levels of stress, anxiety, and depression than the other two clusters. In contrast, the adaptive cluster had relatively low levels of psychological distress and high levels of stable well-being.

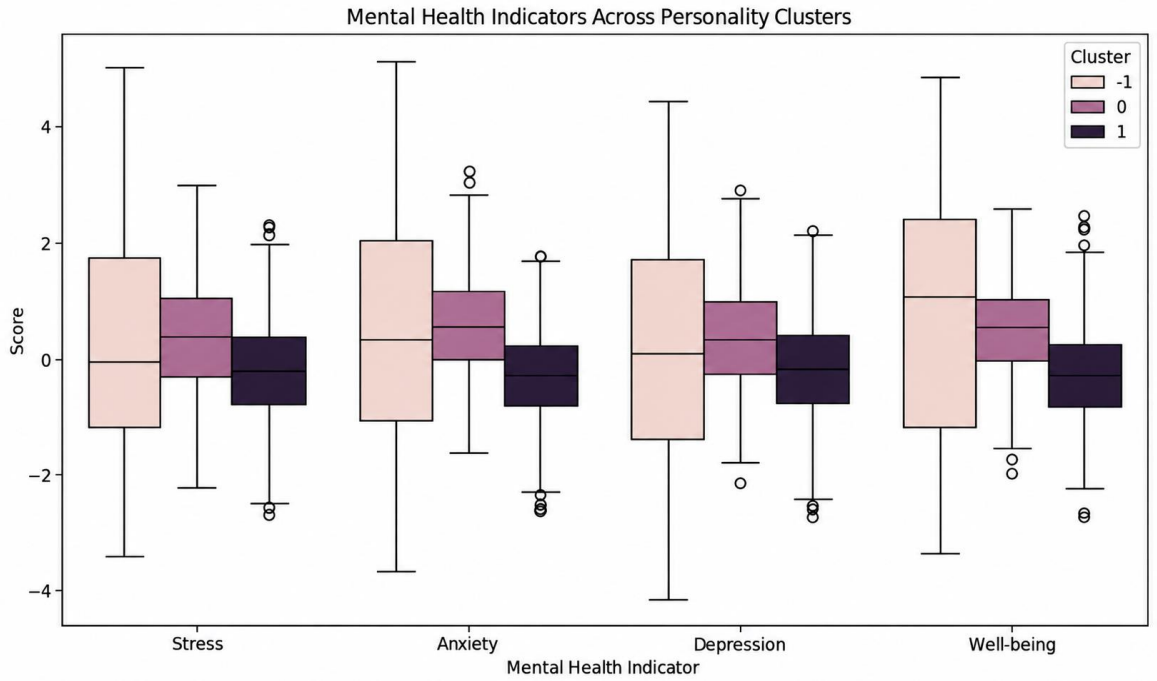


Figure 5. Cluster-wise variation in stress, anxiety, depression and psychological well-being.

4.8. Demographic Analysis

The influence of demographic factors was further examined to determine whether personality variation and cluster membership differed by age, gender, and family background. For age, Spearman analysis showed very weak, statistically nonsignificant associations with all five personality traits: Extraversion (ρ=0.029, p=0.3207), Neuroticism (ρ=−0.018, p=0.5335), Agreeableness (ρ=−0.019, p=0.5035), Conscientiousness (ρ=0.024, p=0.3947), and Openness (ρ=0.040, p=0.1687). This indicates that, within the relatively narrow student age range, personality variation is not meaningfully explained by age.

Gender showed a stronger influence. Significant differences were observed for Extraversion (p=0.0419), Agreeableness (p=0.0125), and Openness (p<0.001), while Neuroticism was not significant (p=0.4226), and Conscientiousness was borderline (p=0.0525). In addition, chi-square analysis showed that cluster membership itself was significantly associated with gender (χ^2=13.67, p=0.0011), indicating that the distribution of adaptive and vulnerable personality profiles differs across male and female students.

Family background showed partial significance. Trait-wise differences were significant for Neuroticism (p=0.0334), Agreeableness (p=0.0345), and Openness (p=0.0319), but not for Extraversion (p=0.7057) or Conscientiousness (p=0.4434). However, family background was not significantly associated with overall cluster

© 2026 Conscientia Beam. All Rights Reserved.

membership in the chi-square analysis ($\chi 2 = 5.03$, $p = 0.0809$). This suggests that some individual traits vary across family settings, but the broader personality cluster structure remains relatively stable across this demographic factor.

The demographic effects on personality traits and ultimate cluster classifications are summarized in Table 9. Age was not significant, gender effects were substantial, and family-type effects were partial.

Table 9. Demographic analysis of personality traits and cluster membership.

Demographic variable	Statistical test	Target variable	Result	p-value	Interpretation
Age	Spearman correlation	Extraversion	($\rho = 0.029$)	0.3207	Not significant
Age	Spearman correlation	Neuroticism	($\rho = -0.018$)	0.5335	Not significant
Age	Spearman correlation	Agreeableness	($\rho = -0.019$)	0.5035	Not significant
Age	Spearman correlation	Conscientiousness	($\rho = 0.024$)	0.3947	Not significant
Age	Spearman correlation	Openness	($\rho = 0.040$)	0.1687	Not significant
Gender	Mann–Whitney test	Extraversion	Significant	0.0419	Gender difference exists
Gender	Mann–Whitney test	Neuroticism	Not significant	0.4226	No meaningful difference
Gender	Mann–Whitney test	Agreeableness	Significant	0.0125	Gender difference exists
Gender	Mann–Whitney test	Conscientiousness	Borderline	0.0525	Weak trend
Gender	Mann–Whitney test	Openness	Significant	0.0000	Strong gender difference
Family type	Mann–Whitney test	Extraversion	Not significant	0.7057	No meaningful difference
Family type	Mann–Whitney test	Neuroticism	Significant	0.0334	Difference exists
Family type	Mann–Whitney test	Agreeableness	Significant	0.0345	Difference exists
Family type	Mann–Whitney test	Conscientiousness	Not significant	0.4434	No meaningful difference
Family type	Mann–Whitney test	Openness	Significant	0.0319	Difference exists
Gender	Chi-square test	Cluster membership	($\chi 2 = 13.67$)	0.0011	Significant association
Family type	Chi-square test	Cluster membership	($\chi 2 = 5.03$)	0.0809	Not significant

Overall, our results confirm that the hybrid approach we report yields statistically meaningful and psychologically interpretable personality clusters. Our results support three major conclusions: First, our dataset exhibited strong tendencies toward clustering. Second, applying our hybrid method—DBSCAN followed by K-means—yielded patterns of dominant and atypical personality configurations among students. Third, these personality patterns were closely associated with inferences concerning tendencies toward mental health, particularly anxiety and psychological well-being.

5. DISCUSSION, PRACTICAL IMPLICATIONS, AND LIMITATIONS

The results of the present study support the usefulness of personality-based clustering as an exploratory framework for understanding variation in student mental health-related tendencies. Among the Big Five dimensions, neuroticism showed the strongest association with higher inferred stress, anxiety, and depression, and lower well-being. This pattern is consistent with the broader personality literature, in which neuroticism is regularly associated with emotional instability, negative affect, and vulnerability to psychological distress. In the present study, the emotionally vulnerable cluster was characterized by comparatively higher neuroticism and less favorable mental health-related indicators, suggesting that neuroticism may be an important differentiating trait in personality-based student profiling.

At the same time, our interpretation of these results must remain cautious. The present analysis does not confirm that personality traits produce mental health outcomes or provide evidence of a disorder or diagnosis. Rather, it

© 2026 Conscientia Beam. All Rights Reserved.

supports the view that certain patterns of personality are associated with different non-clinical estimates of risk for mental health problems within the student sample. Traits of extraversion, agreeableness, and conscientiousness were associated with greater overall estimates of psychological well-being. They were largely consistent with previous reports of the adaptive and socially competent aspects of these personality traits. However, these associations should be considered primarily exploratory and descriptive.

Our results also confirm the usefulness of unsupervised methods for psychological data because patterns of personality in student samples are typically overlapping rather than clearly separable. The combined application of DBSCAN and K-means methods was especially effective in retaining profiles of atypical personality and producing clearly interpretable estimates of dominant personality patterns. This contrasts with most previous studies, which largely focused on the accuracy of supervised text- or image-based predictions or on models that integrate multiple information modalities. In contrast, our present approach provided an extensive basis for student-oriented personality profiles and non-clinical estimates of risk for mental health problems. Our results were generally consistent with those of previous studies that demonstrated strong relations between neuroticism and negative mental health outcomes, as well as strong relations between other personality dimensions and well-being.

In addition, they confirmed the capacity of advanced machine learning methods to represent complex patterns in personality. However, previous work was not directed toward discovering unsupervised subgroups, and there was general agreement that the magnitude and stability of personality-related predictions varied according to modes of representation and conditions of experience. Consequently, the present approach represents an additional contribution of an exploratory, rather than fully predictive, nature.

These findings are consistent with studies reporting that personality traits, particularly neuroticism, are strongly associated with mental health-related outcomes [3] and with work showing that machine learning methods can capture complex personality patterns from diverse data sources [10, 21]. At the same time, recent prediction-oriented studies based on multimodal and graph-based architectures have focused primarily on improving representational performance rather than deriving interpretable student subgroups [23, 24, 26]. This contrast highlights that the present work contributes less as a high-accuracy prediction model and more as an interpretable exploratory framework for student profiling and mental health risk inference.

6. CONCLUSION AND FUTURE WORK

This study presented a hybrid unsupervised learning framework for Big Five personality-based student profiling and non-clinical mental health risk inference using a real dataset of 1,214 undergraduate student responses. The proposed framework combined DBSCAN and K-means to identify dominant and atypical personality patterns, followed by statistical validation and psychological interpretation. The results showed that the dataset exhibited a strong clustering tendency and that the optimal core grouping was obtained at $K = 2$, with a silhouette score of 0.304. The final hybrid structure yielded three interpretable profiles: an atypical or outlier group, an emotionally vulnerable group, and an adaptive or balanced group. Correlation analysis further indicated that neuroticism was most strongly associated with adverse mental health-related tendencies, whereas extraversion, agreeableness, and conscientiousness were more positively associated with psychological well-being. In addition, ANOVA and effect size analyses showed that anxiety and well-being were the mental health indicators most clearly differentiated across the identified personality clusters.

These results should be considered exploratory rather than causal. The present work does not confirm that personality traits directly produce health outcomes or provide clinical diagnoses. Rather, the resulting framework represents a data-driven approach to identifying student patterns of personality associated with greater or lesser non-clinical risks for stress, anxiety, depression, and well-being. Consequently, the utility of the framework lies in its role as a supportive tool for educational settings. From an implementation perspective, the framework could support a tiered system of student care.

© 2026 Conscientia Beam. All Rights Reserved.

First, data would be obtained from institutional student health or counseling systems. Second, application of the hybrid clustering method would identify broad patterns of personality. Third, these patterns would guide counselors, faculty mentors, or staff who provide academic support in implementing preventive care programs, resilience-enhancing activities, and individualized educational counseling. Finally, the resulting approach represents an ancillary method of early detection, not an alternative to clinical assessment or professional care for mental health disorders. Future work will extend the utility of the present approach by applying deep learning methods to obtain more complete representations of personality.

Additional improvements will be achieved by integrating multiple modes of behavioral information to extend the interpretation of questionnaire-based results. Finally, the use of longitudinal methods will permit evaluation of the extent to which patterns of risk associated with personality change over time and under conditions of academic stress. These efforts will further increase the effectiveness, flexibility, and clinical relevance of personality-based methods of student care.

Funding: This research received no specific funding.
Institutional Review Board Statement: The Ethical Committee of the Faculty of Engineering and Technology, Datta Meghe Institute of Higher Education and Research (DU), Sawangi, Wardha 442001, Maharashtra, India, has granted approval for this study on 21 May 2025 (Ref. No. DMIHER/(DU)/IEC/2025/835).
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Disclosure of AI Use: The authors used OpenAI's ChatGPT to edit and refine the wording of the Introduction. All outputs were reviewed and verified by the authors.

REFERENCES

[1] World Health Organization, *Mental health: Strengthening our response.* Geneva, Switzerland: World Health Organization, 2021.

[2] B. W. Roberts and H. J. Yoon, "Personality psychology," *Annual Review of Psychology*, vol. 73, no. 2022, pp. 489-516, 2022. https://doi.org/10.1146/annurev-psych-020821-114927

[3] M. M. Sauls, T. R. Johnson, and H. Q. Nguyen, "Exploring the impact of personality traits on mental health outcomes," *Journal of Personality Research*, vol. 45, no. 2, pp. 145–160, 2020.

[4] J. Chi and Y. N. Chi, "Cluster analysis of personality types using respondents' big five personality traits," *International Journal of Data Science*, vol. 4, no. 2, pp. 116-135, 2023. https://doi.org/10.18517/ijods.4.2.116-135.2023

[5] J. Xu, W. Tian, G. Lv, S. Liu, and Y. Fan, "Prediction of the big five personality traits using static facial images of college students with different academic backgrounds," *IEEE Access*, vol. 9, pp. 76822-76832, 2021.

[6] A. Alhothali, A. l-Salman, and H. S. Al-Khalifa, "Multilingual personality prediction using BERT-based architectures," *Computational Intelligence*, vol. 39, no. 2, pp. 514–533, 2023.

[7] M. Dalvi-Esfahani, A. Niknafs, Z. Alaedini, H. Barati Ahmadabadi, D. J. Kuss, and T. Ramayah, "Social media addiction and empathy: Moderating impact of personality traits among high school students," *Telematics and Informatics*, vol. 57, p. 101516, 2021. https://doi.org/10.1016/j.tele.2020.101516

[8] M. Romeo, D. Hernández García, T. Han, A. Cangelosi, and K. Jokinen, "Predicting apparent personality from body language: Benchmarking deep learning architectures for adaptive social human–robot interaction," *Advanced Robotics*, vol. 35, no. 19, pp. 1167-1179, 2021. https://doi.org/10.1080/01691864.2021.1974941

[9] H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan, and A. H. Mosavi, "A hybrid deep learning technique for personality trait classification from text," *IEEE Access*, vol. 9, pp. 146214-146232, 2021. https://doi.org/10.1109/ACCESS.2021.3121791

© 2026 Conscientia Beam. All Rights Reserved.

[10] M. Lukac, "Speech-based personality prediction using deep learning with acoustic and linguistic embeddings," *Scientific Reports*, vol. 14, no. 1, p. 30149, 2024. https://doi.org/10.1038/s41598-024-81047-0

[11] M. Gjurković and J. Šnajder, "Reddit personality prediction using transformer-based models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[12] A. Naz, H. U. Khan, A. Bukhari, B. Alshemaimri, A. Daud, and M. Ramzan, "Machine and deep learning for personality traits detection: A comprehensive survey and open research challenges," *Artificial Intelligence Review*, vol. 58, no. 8, p. 239, 2025. https://doi.org/10.1007/s10462-025-11245-3

[13] S. Wang, Y. Zhang, and Y. Liu, "Clustering algorithms in personality prediction: A comparative study," *Expert Systems with Applications*, vol. 179, p. 115074, 2021.

[14] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010. https://doi.org/10.1016/j.patrec.2009.09.011

[15] Hartono *et al.*, "An adaptive clustering framework for personality prediction using enhanced seed optimization," *Decision Analytics Journal*, vol. 16, p. 100630, 2025. https://doi.org/10.1016/j.dajour.2025.100630

[16] M. Müller, L. Fuchs, and L. Krieger, "Hierarchical clustering and its application in data analytics," *Data Science Review*, vol. 3, no. 2, pp. 45–56, 2021.

[17] G. Barbieri Lizama and H. D. Calderón-Vilca, "Model for automatic detection of the Big Five personality traits through facial images," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 14, pp. 60–67, 2022.

[18] Y. Zhang, H. Li, and W. Zhao, "A hybrid approach for personality profiling using unsupervised learning techniques," *Journal of Artificial Intelligence*, vol. 17, no. 1, pp. 89–102, 2023.

[19] M. Ashawa, J. D. Bryan, and N. Owoh, "Personality prediction model: An enhanced machine learning approach," *Electronics*, vol. 14, no. 13, p. 2558, 2025. https://doi.org/10.3390/electronics14132558

[20] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, "Attention-based BiLSTM models for personality recognition from user-generated content," *Information Sciences*, vol. 596, pp. 460–471, 2022. https://doi.org/10.1016/j.ins.2022.03.038

[21] S. Aslan, U. Güdükbay, and H. Dibeklioğlu, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image and Vision Computing*, vol. 110, p. 104163, 2021. https://doi.org/10.1016/j.imavis.2021.104163

[22] M. Ghonge, V. K. Shukla, N. Pradeep, and R. K. Solanki, "Recent advances in multimodal deep learning for stress prediction: Toward cycle-aware and gender-sensitive health analytics," *EPJ Web of Conferences*, vol. 341, p. 01059, 2025. https://doi.org/10.1051/epjconf/202534101059

[23] G. Chen, L. Shen, X. Cai, I. Razzak, and S. Jameel, "HIPPD: Brain-inspired hierarchical information processing for personality detection," *Proceedings 17th Asian Conference Machine Learning (ACML), Proceedings of Machine Learning Research*, vol. 304, 2025.

[24] L. Shen *et al.*, "LL4G: Self-supervised dynamic optimization for graph-based personality detection," presented at the 2025 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2025.

[25] F. A. Dael, S. Maidanova, I. Shayea, G. Abitova, and A. Seraly, "A data-driven approach to MBTI personality classification: Insights from machine learning models," in *Proceedings. 2024 IEEE 16th International Conference Communication Systems and Network Technologies (CICN)*, 2024, pp. 1297-1303.

[26] A. Ouarka, T. Ait Baha, Y. Es-Saady, and M. El Hajji, "A deep multimodal fusion method for personality traits prediction," *Multimedia Tools and Applications*, vol. 84, no. 25, pp. 29665-29687, 2025. https://doi.org/10.1007/s11042-024-20356-y

© 2026 Conscientia Beam. All Rights Reserved.

Appendix A.

1. Questionnaire Description

The survey instrument used in this study was designed to assess the Big Five personality dimensions among undergraduate students. The questionnaire presented in Table A1 consisted of 50 personality-related statements, with each statement rated on a 5-point Likert scale, where:

Table A1. Questionnaire Description.

Response	Score
Strongly disagree	1
Disagree	2
Neutral	3
Agree	4
Strongly agree	5

The questionnaire included both positive and negative items. Responses to negative items were reversed during data processing to yield final scores for personality traits. Participants were 1,214 undergraduate students with diverse academic backgrounds. Participation was voluntary, anonymous, and in accordance with institutional guidelines for ethical research.

2. Big Five personality trait mapping

Table A2 presents the Big Five Personality trait mapping.

Table A2. Big five personality trait mapping.

Personality trait	Positive items	Negative items	Total items
Extraversion	Q1, Q6, Q18, Q25, Q28	Q9, Q21, Q35, Q40, Q44	10
Neuroticism	Q2, Q23, Q31, Q38	Q10, Q16, Q19, Q29, Q32, Q41	10
Agreeableness	Q14, Q22, Q30, Q47, Q50	Q4, Q17, Q24, Q37, Q42	10
Conscientiousness	Q7, Q13, Q34, Q36, Q45, Q48	Q3, Q11, Q27, Q49	10
Openness	Q5, Q15, Q26, Q33, Q39	Q8, Q12, Q20, Q43, Q46	10

3. Personality Trait Scoring

The total score for each of the five personality dimensions for each participant was computed as the sum of responses to both positively and negatively worded items.

Positive responses contributed directly to the corresponding trait score. Negative responses were converted to positive values before summation. Final scores reflected the relative position of each participant on each of the five personality dimensions.

The five personality dimensions derived from the questionnaire were:

- Extraversion.
- Neuroticism.
- Agreeableness.
- Conscientiousness.
- Openness.

These standardized personality scores were subsequently used as input variables for the proposed hybrid clustering framework.

4. Demographic Information

In addition to the personality questionnaire, the following demographic variables were collected. Table A3 presents the demographic information.

© 2026 Conscientia Beam. All Rights Reserved.

Table A3. Demographic Information.

Variables	Description
Age	Participant age (Years)
Gender	Male / Female / Other (If applicable)
Family type	Rural / Urban

These demographic variables were used for descriptive statistical analysis and cluster profiling.

Before analysis, the collected responses underwent the following preprocessing steps.

1. Verification of questionnaire completeness.
2. Reverse coding of negatively worded items.
3. Computation of Big Five personality trait scores.
4. Standardization of personality trait scores using Z-score normalization.
5. Hybrid clustering using DBSCAN and K-Means algorithms.
6. Statistical validation using Spearman correlation, one-way ANOVA, and effect size (η^2) analysis.

The questionnaire was administered electronically with Google Forms over the academic year. Participants were informed of the aims of the study and provided informed consent to participate. Participation was entirely voluntary, and no information that could identify individual participants was obtained. All results were subsequently anonymized for statistical analysis.

Appendix B. Reliability Statistics for Personality Measurements.

The personality characteristics of undergraduate students were obtained with the five-factor personality questionnaire developed by Dr. R.N. Singh (Prasad Psycho Corporation, India). This well-standardized instrument consists of 50 items representing the five dimensions of personality: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness. The instrument had previously been fully validated by the original author for reliability using split-half and test-retest methods. Here, we strictly followed the official scoring procedures, including reversing the scores for items worded negatively. Evaluation of internal consistency for the present sample (N = 1,214) was conducted using Cronbach's alpha, McDonald's omega, split-half correlation, Spearman-Brown coefficient, Guttman split-half coefficient, and mean inter-item correlation. Results provided indices of the internal consistency of the present dataset and are reported for completeness. However, because the primary goal of the present work was the development of a hybrid method for personality profiling and mental health assessment rather than the validation of the personality instrument, our reliability estimates reflect measures specific to the present sample. In contrast, validation of the standardized instrument is fully supported by the original validation procedures described in the instrument manual. Table B1 presents the reliability statistics for personality measurements.

Table B1. Reliability statistics for personality measurements.

Personality Trait	Items	Cronbach's α	McDonald's ω	Split-Half Correlation	Spearman–Brown	Guttman Split-Half	Mean Inter-Item Correlation
Extraversion	10	0.294	0.244	0.237	0.383	0.383	0.046
Neuroticism	10	0.186	0.041	0.262	0.415	0.415	0.023
Agreeableness	10	0.351	0.001	0.244	0.392	0.392	0.055
Conscientiousness	10	0.237	0.269	0.270	0.425	0.425	0.043
Openness	10	0.311	0.001	0.233	0.378	0.378	0.039

Note: α = Cronbach's Alpha; ω = McDonald's Omega; MIIC = Mean Inter-Item Correlation. Reliability coefficients represent sample-specific estimates of internal consistency for the present study (N = 1,214). The Big Five Personality Scale employed in this research is a standardized instrument previously validated by Dr. R. N. Singh using Split-Half and Test–Retest reliability methods.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.

© 2026 Conscientia Beam. All Rights Reserved.