Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Novel Hybrid Latent Class Analysis –Machine Learning Framework for Multiclass Prediction of Complex Outcomes: Application to Marital Wellbeing among Middle Eastern Families

Iman Hafizi-Rastani
Isfahan University of Medical Sciences
Kamran Mehrabani-Zeinabad
Isfahan University of Medical Sciences
Awat Feizi
awat_feiz@hlth.mui.ac.ir

Isfahan University of Medical Sciences

Research Article

Keywords: machine learning, latent class analysis, hybrid models, random forest, marital satisfaction, psychological distress, personality traits

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11164259/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title Page

A Novel Hybrid Latent Class Analysis –Machine Learning Framework for Multiclass Prediction of Complex Outcomes: Application to Marital Wellbeing among Middle Eastern Families

Iman Hafizi-Rastani [1], Kamran Mehrabani-Zeinabad [2], Awat Feizi [3 *]

[1] Department of Epidemiology and Biostatistics, and Student Research Committee, School of Health, Isfahan University of Medical Sciences, Isfahan, Iran

[2] Department of Epidemiology and Biostatistics, School of Health, Isfahan University of Medical Sciences, Isfahan, Iran

[3] Department of Epidemiology and Biostatistics, School of Health, Pharmaceutical Sciences Research Center and Clinical Toxicology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

*** Corresponding Author**

Awat Feizi,
Professor of Biostatistics, Ph.D.
Department of Epidemiology and Biostatistics, School of Health, Pharmaceutical Sciences Research Center and Clinical Toxicology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
Email: awat_feiz@hlth.mui.ac.ir

Abstract

Background: Traditional statistical profiling of complex human outcomes faces a fundamental theoretical dichotomy: generative models excel at explanatory inference but impose restrictive linear constraints, while discriminative machine learning algorithms capture non-linear relationships but obscure unobserved population heterogeneity. To bridge this divide, we propose a novel, two-stage statistical paradigm that synergistically integrates likelihood-based finite mixture modeling with non-parametric ensemble learning. This architecture reframes the task into the discovery of a multidimensional latent construct, followed by a flexible mapping from observable covariates to latent group memberships using tree-based universal function approximators.

Methods: The methodological framework relies on four sequential principles, instantiated using a Middle Eastern family wellbeing dataset (n=736) as a proof of concept. First, Latent Class Analysis (LCA) serves as the generative engine to model the target outcome as a multinomial latent variable. Second, the framework transitions to a discriminative approach, treating the discovered profiles as the target variables for non-parametric Ordinal Forest and Ordered Forest ensemble algorithms. Third, the principle of minimal sufficient complexity is operationalized via Recursive Feature Elimination (RFE) with cross-validation to satisfy structural risk minimization. Fourth, we deploy a multi-criteria evaluation ontology, utilizing one-vs-rest multiclass AUC for separability, macro-averaged F1-scores for subgroup equity, and the Ranked Probability Score (RPS) as a proper scoring rule for feature importance and calibration.

Results: Empirically, the LCA identified three distinct, meaningful subpopulation profiles with prevalence of 25%, 56%, and 19%, validating the presence of unobserved target heterogeneity. Adhering to structural risk minimization, reduced-feature models achieved multiclass AUC values (0.775–0.793) statistically identical to full-feature architectures, confirming the power of parsimonious representations. Feature importance derived via RPS revealed that psychological constructs (depression, extraversion, and neuroticism) were the primary drivers of profile differentiation, dramatically outperforming demographic or socioeconomic metrics.

Conclusion: We present a theoretically grounded, generalizable statistical framework that reconciles the explanatory depth of generative latent variable modeling with the predictive flexibility of non-parametric machine learning. This paradigm offers a principled solution for multiclass profiling in high-dimensional, heterogeneous data, with broad theoretical applicability extending well beyond the illustrative case of family wellbeing.

Keywords: machine learning, latent class analysis, hybrid models, random forest, marital satisfaction, psychological distress, personality traits

Background

The empirical investigation of complex human outcomes whether in public health, social sciences, behavioral or clinical psychology is persistently constrained by a fundamental theoretical and methodological bottleneck. The target constructs (e.g., well-being, quality of life, psychopathology, relational functioning) are inherently multidimensional, yet conventional analytical frameworks, predominantly rooted in generalized linear models, forcibly collapse them into unidimensional proxies (1). This oversimplification masks latent population heterogeneity and assumes additive, linear relationships among predictors assumptions that are rarely tenable in real-world psychosocial data (2). Moreover, the relational architecture between a high-dimensional set of determinants and the outcome typically involves non-linear interactions, conditional dependencies, and effect modifications that classical parametric models are structurally ill-equipped to capture (3). Consequently, reliance on traditional regression paradigms not only underestimates critical determinants but also yields models with limited generalizability and poor external validity across diverse populations (4).

Building upon the conceptual reorientation proposed by Jo and colleagues (2023) (5) who pioneered the integration of latent variable modeling into supervised learning by demonstrating that prediction targets can be generated through flexible latent variable models and subsequently validated using external criteria we introduce a novel, domain-agnostic hybrid statistical framework that extends their paradigm in two substantive directions. Theoretically, our architecture reframes the predictive task as a principled two-stage pipeline. In the generative stage, Latent Class Analysis (LCA) a likelihood-based finite mixture model serves as a statistical dimensionality-reduction tool that empirically distills the inherent multidimensionality of the response into a discrete, interpretable multinomial latent variable, formally acknowledging and quantifying unobserved heterogeneity (6). In the subsequent discriminative stage, we employ non-parametric ensemble algorithms (Ordinal Forest and Ordered Forest) to approximate the complex, high-dimensional mapping from a comprehensive covariate set to the derived latent class memberships (7). Critically, we extend Jo et al.'s framework in two ways: first, while their empirical illustrations remained largely within binary or limited-category contexts, we generalize the architecture to accommodate multiclass (three or more) latent profiles, enabling a more nuanced stratification that better reflects real-world heterogeneity; second, we embed a principled feature selection mechanism Recursive Feature Elimination (RFE) within the supervised stage, allowing the framework to identify parsimonious predictor subsets that achieve predictive performance

statistically equivalent to full-feature models, thereby adhering to the principle of minimal sufficient complexity [8]. This addresses a critical gap in existing hybrid LCA-ML applications, where the trade-off between model complexity and generalizability has received insufficient methodological attention. To the best of our knowledge, no previous study has evaluated a multidimensional outcome at three distinct latent levels (low, moderate, high) while simultaneously employing systematic feature selection within a unified LCA-ML hybrid framework for predictive profiling. This integrated architecture is inherently portable: it offers a robust, generalizable solution for any empirical context where the outcome is a latent, multidimensional construct and the determinants are numerous, collinear, and interact non-linearly.

To empirically instantiate and validate the practical utility of this portable framework, we deploy it within the domain of family and relational psychology specifically, to model marital satisfaction among Middle Eastern families. Marital satisfaction serves as an exemplary test case precisely because it encapsulates all the statistical complexities our architecture is designed to handle. Defined as a multidimensional evaluative judgment of relationship quality, it is robustly associated with individual well-being, mental health, communication processes, and long-term family functioning (9,10). Its etiology spans multiple distinct domains: relational characteristics, individual personality traits (11,12), psychological vulnerabilities such as perfectionism and distress (13,14), dyadic functioning (15), sociodemographic factors (11), and even neurobiological correlates (15). Moreover, these associations are profoundly moderated by sociocultural context, with cross-national studies revealing substantial variations across gender equality, economic development, and cultural value systems (10,11). Despite this inherent complexity, prior investigations have largely relied on classical regression techniques or, more recently, isolated machine learning applications that nonetheless treat marital satisfaction as a monolithic, undifferentiated outcome thereby overlooking meaningful latent heterogeneity across different levels of relational adjustment (15,16). By applying our multiclass hybrid framework with principled feature selection to this context, we aim to demonstrate its analytical power and provide novel, stratified insights into the determinants of marital wellbeing in a culturally distinct setting.

Methods

Study Design and Participants

This secondary cross-sectional study utilized data from the project entitled “Epidemiology of Emotional and Psychological Problems in Iranian Children in Isfahan (17).

The required sample size was estimated at 1000 participants. To ensure representativeness of the study population, a multistage cluster sampling (four-stage clusters) was adopted. In the first stage, the six educational districts of Isfahan were defined as primary clusters, from which three districts were randomly selected. In the second stage, two schools were randomly chosen within each selected district. Subsequently, one class was randomly selected from each school, covering grade levels from preschool (6-year-olds) to sixth grade (12-year-olds). Parents of all students in the selected classes were invited to participate. Following sample selection, two comprehensive questionnaire batteries were administered to the participating families, including instruments assessing children's general health status, sleep quality, and emotional well-being, as well as measuring parents' psychological problems, personality traits, and family interactions. Of these, 850 questionnaires were returned, and finally 736 families were retained for data analysis. These 736 families either completed 800 questionnaires related to children. Both sets of questionnaires were completed by parents. however, the present study relies exclusively on the parent-reported data.

Written informed consent was obtained from all participants, and the study protocol was approved by the Ethics Committee of the National Institute for Medical Research Development (NIMAD) (Ethics code: IR.NIMAD.REC.1397.462) (17). For this secondary study, separate ethical approval was obtained from the Ethics Committee of Isfahan University of Medical Sciences (code: IR.MUI.DHMT.REC.1404.012).

Study Instruments and Variables Assessment

Marital Satisfaction

Marital satisfaction was assessed using the Revised Dyadic Adjustment Scale (RDAS), developed by Busby et al. in 1995 (18), a 14-item instrument measuring three dimensions: consensus (6 items), satisfaction (4 items), and cohesion (4 items). Items were rated on a 6-point Likert scale (0–5), except item 11 (0–4). Total scores ranged from 0 to 69, with higher scores indicating better marital quality.

The Persian version of RDAS has been psychometrically evaluated in an Iranian population. Confirmatory Factor Analysis (CFA) supported the original three-factor structure, encompassing consensus, satisfaction, and cohesion. Internal consistency was reported as acceptable, with a Cronbach's alpha of 0.79 for the total scale. Concurrent validity was demonstrated through significant positive correlations with established measures of marital happiness and marital satisfaction (e.g., ENRICH). Test–retest reliability further indicated temporal stability (19).

Personality Traits

Personality traits were assessed using the NEO Five-Factor Inventory (NEO-FFI). This comprehensive 60-item instrument measured five domains: neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Each domain consisting of 12 items rated on a 5-point Likert scale (0 = strongly disagree to 4 = strongly agree).

The Persian version of the NEO-FFI has been validated for use in middle-aged general Iranian population. Internal consistency (Cronbach's alpha) for the five personality domains was reported as 0.86 for neuroticism, 0.73 for extraversion, 0.56 for openness, 0.68 for agreeableness, and 0.87 for conscientiousness. Based on the study's empirical data, the alphas were 0.80, 0.74, 0.47, 0.63, and 0.81, respectively, indicating generally the Persian NEO-FFI is a reliable and valid instrument for assessing the Big Five personality traits in middle-aged Iranian adults (20).

Psychological Distress

Psychological distress was assessed using the 21-item version of the Depression Anxiety Stress Scales (DASS-21), a brief form of the original 42-item DASS. This self-report instrument includes three subscales (7 items each) assessing depression, anxiety, and stress. Responses were rated on a 4-point Likert scale (0–3), and subscale were summed and, if needed, multiplied by two for comparability with the full 42-item version

The Persian version of the DASS-21 has been validated for use in Iranian populations. In the validation study CFA supported the three-factor structure corresponding to depression, anxiety, and stress. Internal consistency was high, with Cronbach's alpha coefficients reported as 0.88 for the depression subscale, 0.85 for anxiety, 0.87 for stress, and 0.94 for the total scale, indicating excellent reliability. Concurrent validity was also established through significant correlations with established measures of depression and anxiety, such as the Beck Depression Inventory and the Zung Anxiety Scale (21).

Demographic and Socioeconomic Status

Demographic information of participants including sex, age, education level, marriage duration, number of children along with their genders were collected. Family socioeconomic status was assessed through a self-report survey utilizing the 6-item socioeconomic Status Short-Form Questionnaire (SES-SQ), a culturally validated instrument for the Iranian population. This scale captured various domains, including educational attainment, occupational status, vehicle ownership, access to digital technology (e.g., laptops, tablets), and lifestyle aspects such as leisure and travel. Based on these items, three distinct SES classes (low, middle, and high) were identified (22).

Data analysis

Hybrid Analytic Approach

This study employed a hybrid Latent Class–Machine Learning approach, combining Latent Class Analysis for class identification and supervised machine learning to predict class membership based on predictors. The hybrid model was constructed by using the class membership assignments obtained from LCA as the response variable in the SML model. In this approach, the latent classes represent unobserved subgroups within the data with similar patterns. By treating these classes as the response variable, the supervised learning model can be trained to predict the probability of class membership for new observations based on their features. This method effectively combines unsupervised learning LCA, which uncovers hidden structure in the data, with supervised learning, which leverages these insights for predictive modeling.

Methodology: Multi-Class Ordinal Latent Feature Prediction Framework

The statistical architecture of this study extends the two-stage predictive pipeline pioneered by Jo et al. (2023). While the original framework restricted the supervised learning stage to a binary representation of latent target constructs, the empirical reality of multi-dimensional clinical, health and psychosocial outcomes (such as marital satisfaction or psychopathology profiles) demands a multi-class ordinal/nominal taxonomy. Forcibly collapsing complex human variations into a simplistic unidimensional or binary proxy masks critical population heterogeneity and strips the statistical framework of ordinal nuance.

To overcome these structural limitations, we introduce an integrated, data-driven framework that proceeds in two principal phases: (1) an unsupervised probabilistic measurement stage utilizing Latent Class Analysis (LCA) to identify a discrete, multi-class nominal/ordered partition of the population (6), and (2) a non-parametric supervised predictive stage deploying the Ordinal Forest (OF) ensemble algorithm to map high-dimensional covariate spaces to the multi-class ordinal latent trajectories.

1. Unsupervised Measurement Stage: Probabilistic Multi-Class Latent Partitioning

Let $Y_i = (Y_i1, Y_i2, \dots, Y_id)'$ denote the d-dimensional vector of observed behavioral, clinical, or relational indicator scores for individual i (where i = 1, ..., N). In a finite mixture modeling framework, the observed joint probability density of Y_i is expressed as a convex combination of K mutually exclusive and exhaustive latent classes, corresponding to distinct nominal/ordinal strata of the target construct: $P(Y_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k f_k(Y_i|\theta_k)$

Where π_k represents the latent class mixing proportions (prior probabilities) for class k, subject to the simplex constraints $\pi_k > 0$ and $\sum \pi_k = 1$, and $f_k(Y_i|\theta_k)$ represents the class-specific probability density function, parameterized by the vector θ_k.

Posterior Class Probability and Statistical Ordering

Following maximum likelihood estimation via the Expectation-Maximization (EM) algorithm, the true individual-level assignment remains unobserved. For each subject i, the model yields a conditional posterior vector of membership probabilities across all K classes:

$$p_{ik} = P(C_i = k|Y_i, \hat{\theta}, \hat{\pi}) = [\hat{\pi}_k f_k(Y_i|\hat{\theta}_k)] / \left[\sum_{j=1}^{K} \hat{\pi}_j f_j(Y_i|\hat{\theta}_j)\right]$$

Where, C_i is the latent class variable. Rather than restricting splits to a binary threshold (K=2) as presented by Jo et al. (2023), the K classes are arranged in a strict ordinal sequence based on a monotonic clinical or statistical index (e.g., mean symptom burden, level of relational adjustment). Let k ∈ {1, 2, ..., K} represent this natural ordering, from lowest severity (k=1) to highest severity (k=K).

To bridge the gap between soft clustering and supervised learning while rigorously accounting for measurement uncertainty, we map individual posterior vectors to a hard discrete multi-class ordinal target variable, L_i^{ms}: $L_i^{ms} = arg\ max_{k \in 1,\ldots,K}(p_{ik})$

This discrete variable preserves the topological ordering of the population strata, formally acknowledging and quantifying unobserved group-level heterogeneity.

2. Supervised Stage: Non-Parametric Ordinal Ensemble Mapping

In the subsequent discriminative stage, we construct a high-dimensional mapping from a comprehensive set of baseline predictors and socio-demographic covariates, denoted by the matrix U_i, to the multi-class ordinal target L_i^{ms}.

Classical parametric models, such as ordered logistic regression (proportional odds models), are structurally ill-equipped to handle real-world some clinical and psychosocial data due to the strict violation of the parallel lines assumption, as well as their inability to automatically capture non-linearities, high-dimensional interactions, and complex conditional dependencies. To resolve this, we implement the non-parametric Ordinal Forest (OF) ensemble algorithm (23).

Algorithmic Mechanics of Ordinal Forest

The Ordinal Forest algorithm treats the discrete classes of $L_i^{ms} \in \{1, 2, ..., K\}$ by assuming that the ordinal categories originate from an underlying continuous latent variable, X. The framework operates by establishing optimal latent thresholds $x = (x_1, x_2, \dots, x_{K-1})$ such that the ordinal class is determined by: $L_i^{ms} = k \Leftrightarrow x_{k-1} \leq X_i < x_k$

Where $x_0 = -\infty$ and $x_K = +\infty$. The OF algorithm optimizes the estimation of these thresholds by maximizing an Ordinal Performance Score (OPS), which quantifies the discrimination capability of the trees across the ordinal boundary conditions.

During the forest construction, B bootstrap samples are drawn from the training dataset. For each split in every individual decision tree, instead of using standard nominal Gini impurity or continuous mean squared error, the split criterion evaluates the classification accuracy based on the estimated continuous latent scores derived from the optimized thresholds: \begin{equation}

$$OPS = \frac{1}{N} \times \sum_{i=1}^{N} \prod_{j \neq L_i^{ms}} \mathbf{1}\left(P(L_i^{ms} \mid U_i) > P(j \mid U_i)\right) \times W_{L_i^{ms},j}$$

The components of the OPS formula are defined as follows:

N: The total number of observations (sample size) in the training dataset.

$\boldsymbol{L_i^{ms}}$**:** The true discrete multi-class ordinal target label derived from the unsupervised latent class partitioning stage for individual i.

j: Represents all alternative ordinal classes in the taxonomy that do not equal the true class of individual i ($j \neq L_i^{ms}$).

$\mathbf{1}(\cdot)$**:** An indicator function that evaluates to 1 if the statement inside the parentheses is true (i.e., if the predicted probability of the true class is strictly greater than the predicted probability of alternative class j), and 0 otherwise.

$\boldsymbol{P(k|U_i)}$**:** The conditional class probability predicted by the decision tree ensemble that individual i belongs to class k given their specific baseline covariate profile U_i.

$\boldsymbol{W_{L_i^{ms},j}}$**:** A specialized prioritization and distance weighting function. It penalizes misclassifications more heavily if they occur across more distant ordinal classes (e.g., misclassifying a high-severity case as low-severity) compared to errors involving adjacent classes.

3. Integrated Framework Validation and Performance Metrics

Adhering to the validation standards outlined by Jo et al. (2023), the full sample is partitioned into a training dataset (Data A = 70%) and an independent validation testing dataset (Data B = 30%). Within Data A, a 5-fold cross-validation scheme is executed to optimize hyperparameters and guard against overfitting. Because the target construct is multi-class and ordered, standard binary performance indicators are insufficient. To bridge this gap, we establish a comprehensive multi-criteria evaluation ontology consisting of the following key evaluation indices: We evaluate the framework's external validity and generalizability using two robust ordinal validation indices:

1. Linear Weighted Cohen's Kappa (κ_w): Measures the chance-corrected agreement between the true latent class L_i^{ms} and the predicted class $\hat{L}_i^{ms}$, formally penalizing disagreements proportionally to their ordinal distance: \begin{equation}

$$\kappa_w = 1 - \frac{\sum_{g=1}^{K}\sum_{h=1}^{K} w_{gh} O_{gh}}{\sum_{g=1}^{K}\sum_{h=1}^{K} w_{gh} E_{gh}}$$

The components of the linear weighted kappa formula are defined as follows:

K: The total number of ordered categories or discrete latent classes in the target variable.

g and h: Indices of the confusion matrix representing the true class category (g) and the predicted class category (h), respectively.

$\boldsymbol{w_{gh}}$**:** The weight assigned to the disagreement between class g and class h. Under a linear weighting scheme, it is defined as $w_{gh} = |g - h|$, ensuring that penalties increase linearly with the distance between the observed and predicted categories.

$\boldsymbol{O_{gh}}$**:** The observed proportion or raw frequency of individuals falling into the cell corresponding to true class g and predicted class h.

$\boldsymbol{E_{gh}}$**:** The expected proportion or cell frequency under the hypothesis of complete independence, calculated directly from the row and column marginal totals (24).

2. Ordinal Concordance Index (C-index): Evaluates the probability that for a randomly selected pair of individuals, the model correctly predicts the relative ordinal ranking of their latent class membership based on their covariate profiles U.

3. One-vs-Rest Multiclass AUC: Assesses class separability across the entire taxonomy. By calculating the area under the receiver operating characteristic curve for each latent subgroup against all others combined, it evaluates the model's distinct diagnostic power for specific levels of severity.

4. **Macro-Averaged F1-Scores:** Ensures equitable predictive accuracy and performance balance across all latent subgroups. This index guards against model optimization favoring over-represented classes by averaging performance unweighted by class size, highlighting vulnerabilities in smaller or minority mcl-latent strata.
5. **Ranked Probability Score (RPS):** Acts as a proper scoring rule for ordinal multi-class probabilities, serving as the benchmark for calibration and feature importance. The RPS evaluates the squared error of the cumulative predicted probability distribution relative to the step function of the true category, heavily penalizing models that allocate probability mass far from the actual rank (25).

By linking a probabilistic measurement model to a non-parametric ordinal ensemble machine, this unified architecture preserves multi-dimensional richness, respects the natural hierarchy of human outcomes, and optimizes external validity across diverse populations.

Classes Identified in the Latent Class Analysis

Latent class analysis was employed to categorize participants with similar patterns of marital satisfaction according to different dimensions of marital satisfaction, thereby generating the response variable, which serves as the target outcome within the machine learning framework, where predictive features of outcome are identified.

LCA is a model-based clustering approach used to identify unobserved (latent) subgroups within a population based on individuals' response patterns. Its primary objective is to determine the number and prevalence of latent classes and to characterize the heterogeneity represented by these classes. Furthermore, latent class membership, typically expressed probabilistically, can be incorporated into subsequent analyses as either an outcome variable (as we used it) or a predictor variable (26, 27).

To identify latent classes of marital satisfaction, we conducted LCA using the three dimensions of the RDAS: consensus, satisfaction, and cohesion. For each dimension, participants' raw scores were divided into tertiles (low, medium, high). These tertile-categorized scores served as manifest indicators in the LCA. The analysis yielded distinct latent classes representing different levels of marital satisfaction among participants. Each participant was assigned to the class with the highest posterior probability. Model adequacy was further assessed by examining classification error rates and interpretability of class profiles. Identified classes were subsequently used as categorical outcome in subsequent SML model.

Model estimation in LCA followed the maximum likelihood approach with robust standard errors. Competing models with 2-5 classes were evaluated. The optimal number of latent

classes was selected using a combination of interpretability and statistical criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted Bayesian Information Criterion (SABIC), entropy values, classification error, and likelihood-ratio tests. Lower AIC, BIC, SABIC and classification error, together with higher entropy indicated superior model fit and clearer class separation.

Hybrid LCA-ML Framework for Latent Class Prediction

Two supervised machine learning approach, known as the ordinal forest and ordered forest algorithm, were applied to identify the relationships between the predictors (personality traits, psychological distress, and demographic and socioeconomic status) and the class labels derived from the LCA on RDAS domains.

In SML, models are trained on labeled clinical datasets in which predictor variables are paired with known outcomes. Among common supervised learning algorithms, decision trees provide an intuitive, rule-based structure that recursively partitions the feature space to generate outcome predictions, making them transparent and clinically interpretable (28). Random Forests, an extension of decision trees, aggregate predictions from multiple bootstrapped trees to improve accuracy and robustness while reducing overfitting through ensemble averaging (29). These models are particularly advantageous in medical contexts characterized by complex, nonlinear relationships and heterogeneous patient data, where improved stability and generalization performance are essential (30).

The ordinal forest method, as formalized by Hornung (2020), extends the random-forest paradigm to appropriately handle ordinal response variables by assuming an underlying latent continuous variable that maps onto observed ordered categories via a monotonic step function. Unlike naive approaches that treat ordinal labels as categorical, ordinal forest optimizes a scoring of the ordinal classes: it generates many candidates score sets (via random sampling), fits a regression forest on each, and selects the score set that minimizes out-of-bag error. After finding the best scoring, ordinal forest builds a final forest for prediction and computes a permutation-based variable importance, often using the negative Ranked Probability Score to better respect the ordering of the target. Empirical comparisons show that ordinal forest tends to outperform many competing ordinal-regression techniques, and its importance measure effectively distinguishes predictive covariates from noise (23).

The ordered forest is a machine-learning estimator for ordered categorical (ordinal) outcomes, developed by Lechner & Okasa (2019). Unlike conventional random forests that treat classes as nominal, this method explicitly accounts for the ordering of response categories by estimating conditional choice probabilities in a flexible, non-parametric way. It further

provides tools for estimating marginal effects (i.e., how changes in predictors affect the probability of each outcome) and conducting statistical inference (standard errors) on those probabilities (31).

Dataset Preprocessing

Data preprocessing was initially performed as a preparatory step prior to the final implementation of the machine learning approach. Data cleaning was conducted to ensure completeness and consistency, with implausible or out-of-range values checked against data collection records and corrected where possible. Categorical variables were encoded numerically to allow effective processing by the algorithms. Continuous variables were examined for outliers, which were treated or removed when necessary.

During data cleaning, missing data handling was caried out by imputation using the missForest algorithm. The missForest is a nonparametric algorithm for imputing missing values in datasets that contain both continuous and categorical variables. Using an iterative procedure, the method trains a random forest model for each variable with missing entries, predicting those missing values from the observed ones. This process cycles through all variables repeatedly until the imputation error stabilizes. Its robustness and ability to handle mixed-type data make it a widely used tool for preprocessing in biomedical and machine-learning applications (32). Imputation was performed considering the internal and clinical relationships among the variables. This preprocessing provided a complete dataset suitable for machine learning–based feature selection and subsequent analyses.

Following these steps, feature selection was conducted using the Recursive Feature Elimination (RFE) algorithm. The RFE procedure iteratively removes the least important predictors and retrains the model to identify the subset of features that most effectively predict the target outcome (33). The RFE procedure was based on a random forest algorithm with 5-fold repeated cross-validation (300 repetitions) to reduce sampling bias and enhance model stability, using Mean Decrease Accuracy to rank predictor importance.

To enhance the interpretability and stability of feature importance estimates, we first assessed multicollinearity among all candidate predictors using pairwise correlation matrices. Given that tree-based algorithms, while inherently robust to collinearity in terms of prediction, can still yield unstable importance rankings when highly correlated variables are present particularly when different operationalizations of the same construct (e.g., continuous and categorical versions of a variable) are included we adopted a structured variable selection strategy. Specifically, within clusters of variables exceeding a pre-defined correlation threshold ($|r| > 0.80$), we retained the variable with the strongest theoretical relevance or the highest univariate

association with the outcome. This preselection step effectively reduced feature redundancy. Subsequently, we applied RFE on the remaining set to verify that the excluded features contributed negligibly to predictive performance, which was further confirmed by comparing the full-feature and reduced-feature models (as reported in the Results).

Implementation of Machine Learning Models and Performance Evaluation

Prior to models training, the essential hyperparameters for the ordered forest model were specified for tuning purposes. Hyperparameter optimization was conducted using grid search and 5-fold repeated cross-validation (20 repetitions). The algorithm was evaluated across the predefined hyperparameter combinations, and the weighted Cohen's kappa coefficient was used as the primary criterion for selecting the optimal parameter set. The ordered forest algorithm was subsequently trained using the selected optimal hyperparameters.

In contrast, for implementations of the ordinal forest model that did not require tuning, the algorithm was executed using its default parameter settings (23).

Following this step, to obtain robust results, the models were trained using repeated 5-fold repeated cross-validation (100 repetitions). Also, to handle class imbalance, oversampling in each fold was performed.

Trained models were evaluated on the original (non-overweighted) dataset to predict the latent classes of marital satisfaction. The predicted classes were compared with the original classes, and multiple performance metrics were calculated.

Given the class imbalance (e.g. latent marital satisfaction class) in the test set, model performance was assessed using metrics that are robust to unequal class distributions. Multiclass area under curve (AUC) served as the primary measure of overall discriminative ability across outcome categories using Receiver Operating Characteristic (ROC) curves generated with the one-vs-rest approach for each class. Performance of the classification models was further examined with agreement- and class-balanced metrics, including the Matthews correlation coefficient (MCC) and balanced accuracy, complemented the evaluation by capturing both global predictive performance and agreement beyond chance. Besides, the geometric mean (G-Mean) was used to assess balanced sensitivity and specificity across classes, and the weighted F1-score was applied to capture the trade-off between precision and recall while accounting for class proportions. Additionally, weighted Cohen's Kappa was reported to account for the ordinal structure of the response variable. Furthermore, SML models were applied to quantify the relative importance of predictors used to classify latent classes of marital satisfaction based on the Ranked Probability Score (RPS).

Figure 1 summarizes the overall data analysis workflow. Following the identification of latent classes, preliminary analyses were conducted using conventional statistical tests (Kruskal–Wallis for continuous and chi-square for categorical variables) to compare baseline characteristics across the LCA-derived classes. All analyses were performed in R (version 4.5.0), and statistical significance was set at $p < 0.05$.

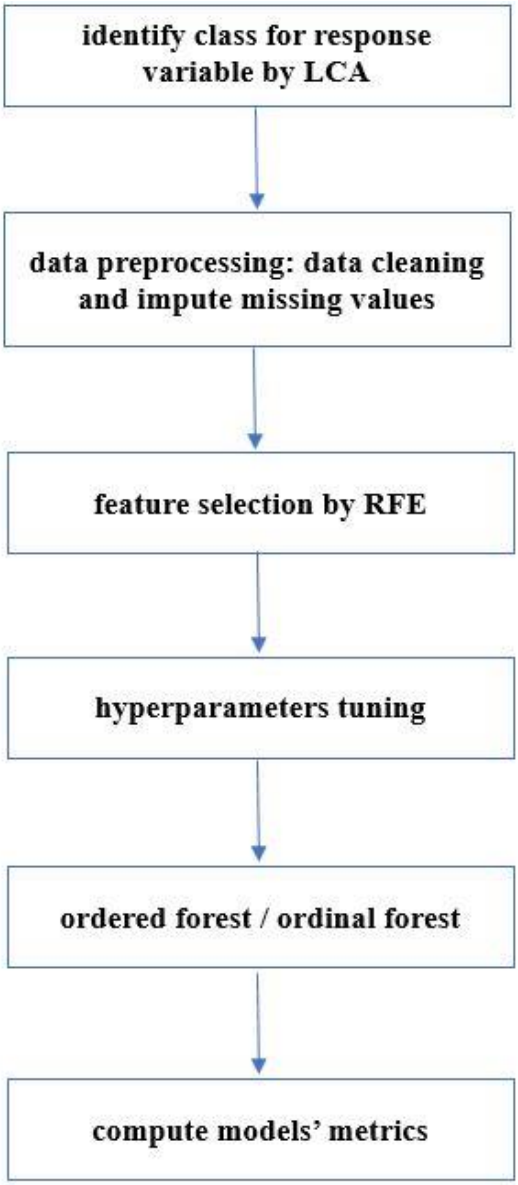


Figure 1: Workflow of the data analysis

Results

Latent Class Analysis

Resulting from latent class analysis and considering profile plot, three-class model demonstrated the most favorable balance between model fit, parsimony and interpretability. It provided the most interpretable and substantively meaningful representation of the underlying latent structure, as indicated by the lowest values of both the Bayesian Information Criterion (BIC = 4388.37) and the sample-size adjusted BIC (SABIC = 4343.92) (Table 1).

Table 1: Fit Indices for Latent Class Models

Number of Class	AIC	BIC	SABIC	Entropy (R^2)	Classification error	LL	L^2	df	p-value
Two-Class	4368.6807	4414.6112	4382.8580	0.7251	0.0753	-2174.3404	84.6743	16	0.0024
Three-Class	4324.0698	4388.3724	4343.9180	0.6514	0.1433	-2148.0349	32.0633	12	0.0014
Four-Class	4321.8294	4404.5042	4347.3485	0.5889	0.2190	-2142.9147	21.8230	8	0.0054
Five-Class	4325.2946	4426.3416	4356.4846	0.5675	0.2503	-2140.6473	0.0017	4	0.0017

AIC: Akaike Information Criterion, BIC (LL): Bayesian Information Criterion based on Log-Likelihood, SABIC: Sample-Size Adjusted Bayesian Information Criterion, LL: Log-Likelihood, L^2: Likelihood-Ratio.

In the final three-class model, the estimated probabilities of class size were 25% for Class 1, 56% for Class 2, and 19% for Class 3. Item-response probabilities for each of the three subdimensions of marital satisfaction were calculated and used to label the each identified class. Class 1, showing higher probabilities of unfavorable responses across the subdimensions, was labeled as “Low Satisfaction”. Class 2, characterized by moderate probabilities of satisfaction and Class 3, with higher probabilities of favorable responses, was labeled as “High Satisfaction” (Table 2).

Based on the tertile distribution of marital satisfaction subdomains presented in Table 2, clear and statistically significant differences were observed across the three latent classes (all $p < 0.001$). In the “Low Satisfaction” class (Class 1), the vast majority of participants fell within the first tertile of all three subdomains: consensus (87.8%), satisfaction (83.0%), and cohesion (89.9%). Conversely, in the “High Satisfaction” class (Class 3), nearly all participants were in

the third tertile for consensus (100%) and cohesion (89.9%), and more than half (57.6%) were in the third tertile for satisfaction. The "Medium Satisfaction" class (Class 2) exhibited an intermediate pattern, with the majority of participants concentrated in the second tertile across all subdomains. Moreover, median scores for each subdomain increased progressively and significantly from the low to the high class (e.g., consensus median: $18 \rightarrow 24 \rightarrow 29$; $p < 0.001$). These findings provide strong empirical support for the labeling of the three classes as low, medium, and high marital satisfaction (Table 2).

Table 2: Subdomain-Specific Tertile Profiles Across the Identified Marital Satisfaction LCA: Validation of Latent Class Separation

Marital satisfaction	Quantile	RDAS Class			P-value [a]
		Low (n = 188)	Medium (n = 390)	High (n = 158)	
Consensus					
	Tertile 1	165 (87.8)	66 (16.9)	0 (0.0)	
	Tertile 2	23 (12.2)	247 (63.4)	0 (0.0)	< 0.001
	Tertile 3	0 (0.0)	77 (19.7)	158 (100.0)	
	Median [IQRb]	18 [15-20]	24 [22-26]	29 [28-30]	< 0.001
Satisfaction					
	Tertile 1	156 (83.0)	56 (14.4)	0 (0.0)	
	Tertile 2	32 (17.0)	249 (63.8)	67 (42.4)	< 0.001
	Tertile 3	0 (0.0)	85 (21.8)	91 (57.6)	
	Median [IQR]	13 [10-15]	17 [16-18]	19 [18-20]	< 0.001
Cohesion					
	Tertile 1	169 (89.9)	75 (19.2)	0 (0.0)	
	Tertile 2	19 (10.1)	190 (48.7)	16 (10.1)	< 0.001
	Tertile 3	0 (0.0)	125 (32.1)	142 (89.9)	
	Median [IQR]	4 [2-6]	8 [7-10]	13 [10-16]	< 0.001

Data presented as frequency (percentage). [a] Resulted from Kruskal-Wallis for continuous and chi-squared for categorical variables; [b] IQR: Inter Quartile Range.

Demographic and other characteristics of study participants are presented in Table 3. Significant differences were observed across latent classes in sex, age, wife's age, marital duration, wife's education, and socioeconomic status (all $P < 0.001$), whereas number of children showed no significant difference ($P = 0.824$). A higher proportion of females was observed in the low marital satisfaction class compared with the other classes. Age, wife's age, and duration of marriage showed statistically significant but small differences across classes. Wife's education demonstrated a shift toward higher educational levels in the high class, while socioeconomic status exhibited a clear gradient, with a greater proportion of participants with high socioeconomic status in the high class.

476 **Table 3: Sociodemographic and psychological characteristics of study participants across**
477 **tertiles of RADS**

Variables set	Variable	Total sample (n=736)	RDAS's Class			
			Low	Medium	High	P-value [a]
Demographic and socioeconomic status	Sex					
	Female	608 (82.6)	174 (92.6)	306 (78.5)	128 (81.0)	< 0.001
	Male	128 (17.4)	14 (7.4)	84 (21.5)	30 (19.0)	< 0.001
	Age	39 [36-42]	39 [36-42]	39 [36-43]	38 [35-42]	< 0.001
	Wife age	42 [39-46]	42 [39-48]	42 [38-45]	41 [38-45]	< 0.001
	Marital duration	15 [13-19]	15 [12-18]	15 [13-19]	14 [12-18]	< 0.001
	Wife education					
	Less than diploma	112 (15.2)	42 (22.3)	54 (13.8)	16 (10.1)	
	Diploma[b]	234 (31.8)	70 (37.3)	115 (29.5)	49 (31.0)	
	Bachelor of Science	260 (35.3)	54 (28.7)	157 (40.3)	49 (31.0)	< 0.001
	Master of Science	72 (9.8)	15 (8.0)	37 (9.5)	20 (12.7)	
	Ph.D	58 (7.9)	7 (3.7)	27 (6.9)	24 (15.2)	
	Child number					
	0	73 (9.9)	19 (10.1)	39 (10.0)	15 (9.5)	
	1	163 (22.1)	46 (24.5)	80 (20.5)	37 (23.4)	0.824
	2	403 (54.8)	98 (52.1)	218 (55.9)	87 (55.1)	
	3	87 (11.8)	21 (11.2)	50 (12.8)	16 (10.1)	
	4	10 (1.4)	4 (2.1)	3 (0.8)	3 (1.9)	
	socioeconomic status					
	Low	158 (21.5)	58 (30.8)	73 (18.7)	27 (17.1)	
	Moderate	407 (55.3)	102 (54.3)	229 (58.7)	76 (48.1)	< 0.001
	High	171 (23.2)	28 (14.9)	88 (22.6)	55 (34.8)	

478
479

480 **Table 3 (continued)**

Variables set	variable	Total sample (n = 736)	RDAS			
			Low	Medium	High	P-value [a]
Personality trait	Neuroticism	20.7 ± 7.8	25.6 ± 7.3	19.9 ± 7.1	16.5 ± 7.4	< 0.001
	Extraversion	28.8 ± 6.3	25.1 ± 6.3	29.3 ± 5.5	32.9 ± 5.5	< 0.001
	Openness	24.7 ± 4.8	24.7 ± 4.8	25.0 ± 4.6	24.1 ± 5.5	0.66
	Agreeableness	32.9 ± 5.3	30.9 ± 4.8	33.1 ± 5.3	35.3 ± 4.8	< 0.001
	Conscientiousness	36.5 ± 5.8	34.3 ± 6.1	36.1 ± 5.6	40.3 ± 4.2	< 0.001
Psychological Distress						
	Depression	4.8 ± 4.5	8.1 ± 5.3	4.2 ± 3.6	2.0 ± 2.8	< 0.001
	Normal	415 (56.4)	44 (23.4)	231 (59.2)	140 (88.6)	< 0.001
	Mild	104 (14.1)	36 (19.1)	64 (16.4)	4 (2.5)	
	Moderate	125 (17)	50 (26.6)	62 (15.9)	13 (8.2)	
	Severe	48 (6.5)	21 (11.2)	26 (6.7)	1 (0.6)	
	Extremely severe	44 (6.0)	37 (19.7)	7 (1.8)	0 (0.0)	
	Anxiety	4.3 ± 4.1	6.5 ± 5.3	3.9 ± 3.4	2.3 ± 2.9	< 0.001
	Normal	398 (54.1)	69 (17.3)	203 (52.1)	126 (79.7)	< 0.001
	Mild	63 (8.6)	12 (6.4)	43 (11.0)	8 (5.1)	
	Moderate	135 (18.3)	34 (18.1)	83 (21.3)	18 (11.4)	
	Severe	56 (7.6)	21 (11.2)	33 (8.5)	2 (1.3)	
	Extremely severe	84 (11.4)	52 (27.7)	28 (7.2)	4 (2.5)	
	Stress	7.7 ± 4.8	10.5 ± 4.9	7.3 ± 4.3	4.9 ± 4.0	< 0.001
	Normal	390 (53)	60 (31.9)	203 (52.1)	127 (80.4)	< 0.001
	Mild	102 (13.9)	27 (14.4)	65 (16.7)	10 (6.3)	
	Moderate	116 (15.8)	34 (18.1)	72 (18.5)	10 (6.3)	
	Severe	89 (12.1)	40 (21.3)	42 (10.8)	7 (4.4)	
	Extremely severe	39 (5.3)	27 (14.4)	8 (2.1)	4 (2.5)	

Continuous and categorical variables are presented as mean ± standard deviation and frequency (percentage) respectively; [a] Resulted from Kruskal-Wallis for continuous and chi-squared for categorical variables. [b]12 years formal education

Machine Learning Algorithms, RFE-Derived Feature Sets, and Model Configurations

To predict the identified latent classes of marital satisfaction (low, medium, high), we employed two machine learning algorithms: Ordered Forest and Ordinal Forest. As shown in Table 4, we considered two different representations of psychological variables: raw continuous scores and categorized (discretized) versions. For each representation, RFE was used to compute variable importance. Based on the importance results obtained from the raw-score dataset (left panel of Table 4), we selected the six most influential features: Depression, Extraversion, Neuroticism, Stress, Conscientiousness, and Anxiety. These six variables formed the reduced predictor set. Consequently, we defined two predictor sets for model development: Set 1 (the full set) included all 15 raw-score psychological variables together with demographic variables (age, wife age, wife education, marital duration, SES categorized, sex, child number); Set 2 (the reduced set) included only the six most important raw-score psychological features listed above. Thus, four models were developed: Ordered Forest with Set 1 (denoted as SML1), Ordered Forest with Set 2 (SML2), Ordinal Forest with Set 1 (SML3), and Ordinal Forest with Set 2 (SML4). This design allowed us to compare the predictive performance of the two algorithms as well as the two feature sets (full 15 variables versus the top-6 selected by RFE) in accurately classifying individuals into marital satisfaction classes. Although Table 4 also presents variable importance from the categorized psychological variables (right panel), the final models were built using the raw-score representation to retain maximum information, and the RFE-based selection was performed on the raw dataset.

Table 4. Feature importance based on Recursive Feature Elimination

Dataset with psychological variables as raw scores		Dataset with categorized psychological variables	
variable	Overall importance [a] (%)	variable	Overall importance [a] (%)
Depression	16.66	Depression	16.12
Extraversion	13.16	Extraversion	13.96
Neuroticism	11.73	Neuroticism	11.81
Stress	10.53	Conscientiousness	10.36
Conscientiousness	10.02	Anxiety	8.58
Anxiety	9.09	Stress	8.16
Wife age	6.77	Wife age	6.68
Agreeableness	6.38	Agreeableness	6.64
Age	6.29	Wife education	6.56
Wife education	6.04	Age	6.51
Openness	5.90	Marriage duration	5.96
Marriage duration	5.60	openness	5.96
SES (categorized)	4.51	SES (categorized)	4.79
Sex	3.55	Sex	3.80
Child number	2.87	Child number	3.29

[a] Overall importance was calculated based on Mean Decrease Accuracy

Supervised Machine Learning Models for Predicting Latent Classes of Marital Satisfaction

Table 5 presents the classification performance of four supervised machine learning models (SML1–SML4) for predicting the three latent classes of marital satisfaction (low, medium, high). Overall, all models achieved moderate discriminative ability, with multiclass AUC values ranging from 0.775 (SML2) to 0.793 (SML1). The highest AUC (0.793) was obtained by Ordered Forest with the full 15-predictor set (SML1), followed closely by Ordinal Forest with the full set (SML3: 0.787). The reduced 6-predictor sets (Set2) yielded slightly lower AUCs in both algorithms (SML2: 0.775; SML4: 0.778), suggesting that the additional predictors contributed marginal discriminatory information.

However, other metrics indicate that overall classification performance was modest. Balanced accuracy ranged from 0.565 (SML4) to 0.596 (SML1), and weighted F1 scores from 0.582 (SML4) to 0.611 (SML1). Weighted Cohen's Kappa values (0.472–0.508) reflect fair to

moderate agreement beyond chance. Notably, MCC was very low for all models (0.053–0.080), implying that while the models performed better than random guessing, they struggled to correctly classify all three classes simultaneously, likely due to class imbalance or overlapping feature distributions. Among all models, SML1 (Ordered Forest with full feature set) showed the best overall performance across most metrics, though the differences between models were relatively small. Therefore, the reduced 6 predictor set (Set2) through ordered or ordinal forest may be considered simpler alternatives without substantial loss in predictive power.

Table 5. Classification Performance Metrics of Machine Learning Models Across Datasets

Algorithm	Data set	Model	Multiclass AUC	MCC	Balanced Accuracy	G-mean	Weighted F1	Weighted Cohen's Kappa
Ordered Forest	Set1	SML1	0.793	0.053	0.596	0.595	0.611	0.494
Ordered Forest	Set2	SML2	0.775	0.068	0.582	0.581	0.589	0.472
Ordinal Forest	Set1	SML3	0.787	0.063	0.581	0.578	0.598	0.508
Ordinal Forest	Set2	SML4	0.778	0.080	0.565	0.561	0.582	0.483

Set 1: 15 predictors; Set 2: 6 predictors; SML: supervised machine learning; MCC: Matthews correlation coefficient, G-mean: Geometric mean.

Figure 2 displays the ROC curves and variable importance results for the four supervised machine learning models (SML1–SML4) in predicting the low, medium, and high marital satisfaction classes. Panel A shows results for the full 15-predictor models (SML1: Ordered Forest; SML3: Ordinal Forest), while panel B shows the reduced 6-predictor models (SML2: Ordered Forest; SML4: Ordinal Forest). For each model, a one-vs-rest ROC analysis was performed separately for each class.

The ROC curves (not shown in detail in the excerpt) generally indicated that the models achieved moderate to good discriminative ability, consistent with the multiclass AUC values reported in Table 5 (0.775–0.793). For the low satisfaction class, all models exhibited relatively higher AUC, suggesting that this class was easier to distinguish from the others. The high satisfaction class also showed acceptable discrimination, whereas the medium satisfaction class proved most challenging, likely due to its intermediate position and greater overlap with the neighboring classes.

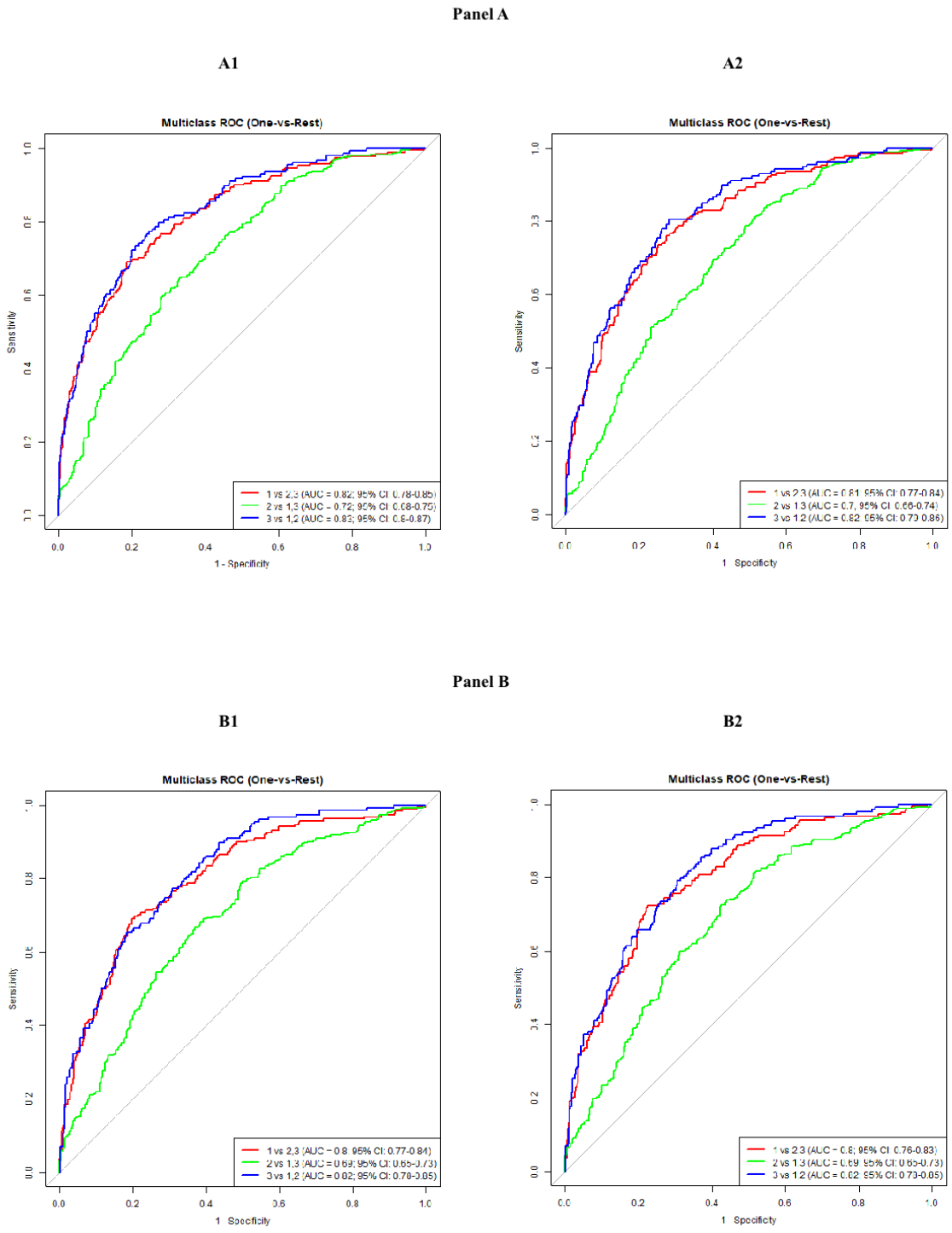


Figure 2: Receiver Operating Characteristic Analysis (One-vs-Rest) for the Latent Marital Satisfaction Class

Panel A presents results obtained using the full 15-feature dataset and includes: (A1) the Ordered Forest model (SML1) and (A2) the Ordinal Forest model (SML3). Panel B presents results obtained using the reduced feature dataset and includes: (B1) the Ordered Forest model (SML2) and (B2) the

Ordinal Forest model (SML4); 1: low satisfaction (class 1), 2: moderate satisfaction (class 2), 3: high satisfaction (class 3)

The bar plots in Figure 3 present raw variable importance values based on the RPS for each model. Variables are ordered according to their aggregated importance ranks, calculated by averaging model specific ranks. Across all four models, Depression consistently emerged as the most important predictor, followed by Extraversion, Neuroticism, and Stress. This pattern aligns with the RFE results in Table 4. Notably, the importance rankings were highly similar between the full (SML1, SML3) and reduced (SML2, SML4) models, confirming that the six selected features captured the majority of predictive information. Moreover, the two algorithms (Ordered Forest vs. Ordinal Forest) produced comparable importance patterns, supporting the robustness of the feature selection. Overall, Figure 3 reinforces that a small set of psychological variables particularly Depression and personality traits is sufficient to achieve performance close to that of the full model, which has practical implications for screening and intervention in marital satisfaction research.

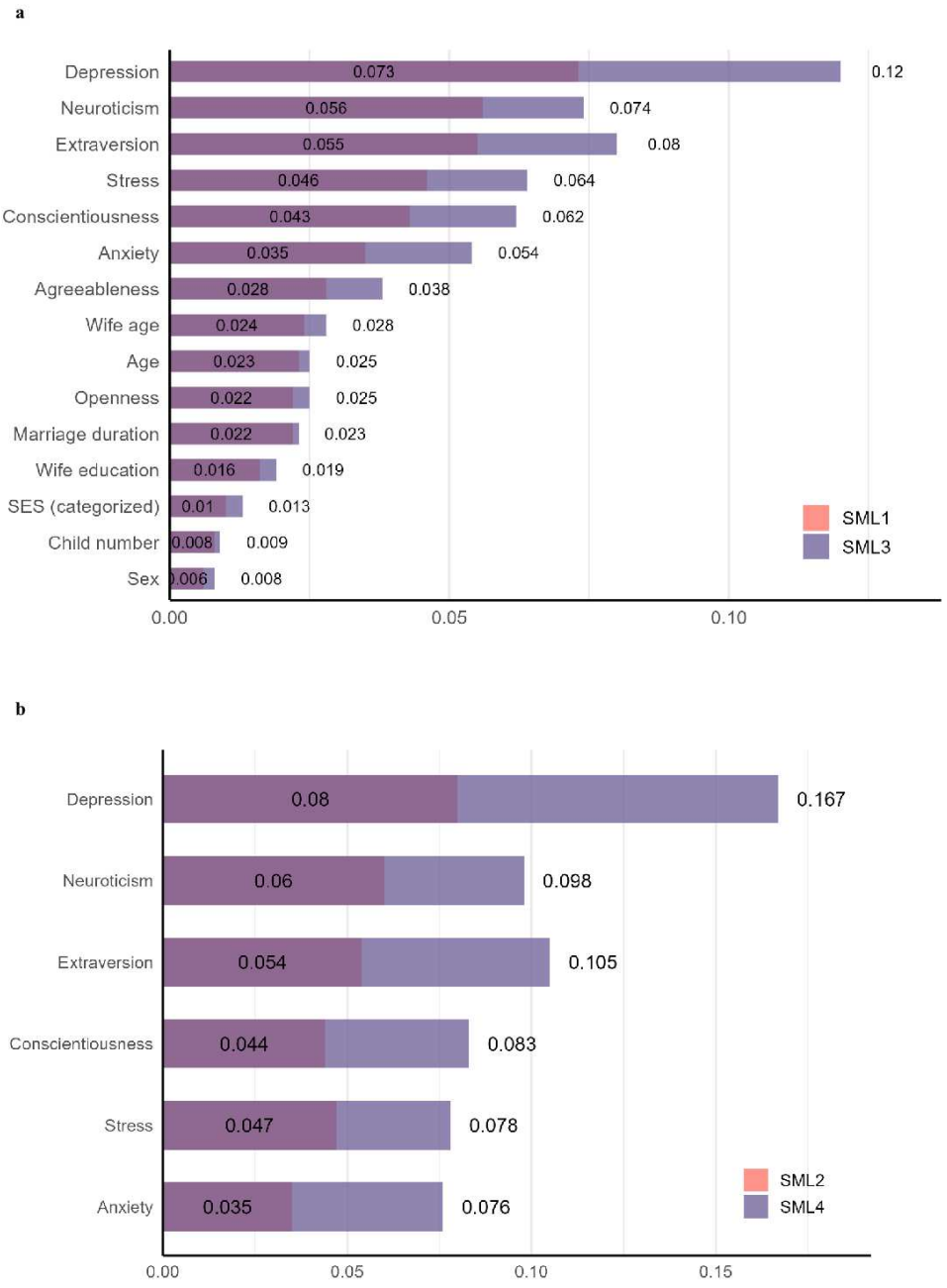


Figure 3: Variable Importance Rank

a) SML1: Ordered Forest model based on the full 15-predictor set; SML3: Ordinal Forest model based on the full 15-predictor set

b) SML2: Ordered Forest model based on the reduced 6-predictor set; SML4: Ordinal Forest model based on the reduced 6-predictor set

bars show raw importance values based on the Ranked Probability Score for each model. Variables are ordered according to their aggregated importance ranks, calculated by averaging model-specific ranks.

Discussion

This study establishes a unified, two-stage statistical learning paradigm that resolves the persistent theoretical dichotomy between generative measurement models and discriminative ensemble learning. Rather than treating complex human outcomes as directly observable or rigidly continuous univariate targets a practice that ignores measurement error and population-level heterogeneity our architecture treats the response space as an unobserved, multi-dimensional latent construct. By utilizing Latent Class Analysis (LCA) as a likelihood-based generative engine on the Revised Dyadic Adjustment Scale (RDAS) indicators (n=736), we unmasked a discrete, three-tier ordinal subpopulation topology (low, medium, and high satisfaction), effectively executing a probabilistic data reduction that retains structural richness. By routing these derived latent target variables (L_i^{ms}) into non-parametric Ordinal and Ordered Forest ensemble machines, we circumvented the restrictive parallel lines and linearity assumptions that structurally cripple classical parametric alternatives like proportional odds models. This integrated framework optimizes the mapping of high-dimensional, non-linear interactions across a mixed covariate space (demographic, socioeconomic, and psychological distress/personality profiles). Consequently, this approach demonstrates that combining a probabilistic measurement phase with a non-parametric machine learning stage allows researchers to preserve the interpretable taxonomies of latent variable modeling while capitalizing on the superior predictive generalizability, implicit interaction capture, and flexible function approximation of ensemble methods.

This study found that depressive symptoms were the most important predictor of marital satisfaction. Anxiety and stress also contributed to distinguishing marital satisfaction levels. Individuals with higher psychological distress reported substantially lower marital satisfaction. Accordingly, Rakhshani et al. found that stress and anxiety are inversely associated with marital satisfaction, while Shi and Whisman highlighted the complex interplay between stress, depression, and marital satisfaction within dyadic processes (34). Dong et al. reported that parenting-related stress is associated with increased marital conflict, which in turn contributes to lower marital satisfaction, highlighting the mediating role of dyadic strain in family contexts (35). Yang et al. indicated that depressive symptoms are associated with reduced marital satisfaction and relational difficulties. Moreover, the relationship between psychological distress and marital satisfaction is reciprocal, meaning that each may influence the other over time (36). These patterns are consistent with the vulnerability–stress–adaptation framework,

which posits that enduring vulnerabilities interact with stress processes to shape marital outcomes (37).

Our results also demonstrated that personality traits emerged as key differentiators of marital satisfaction. Lower neuroticism and higher extraversion and conscientiousness were associated with greater satisfaction, while agreeableness and openness had less consistent effects. Consistent with our findings, Sayehmiri et al., O'Meara et al. and Munini et al. reported that neuroticism is negatively associated with marital satisfaction, potentially due to heightened emotional reactivity and negative affect, whereas conscientiousness and extraversion were positively associated with satisfaction, reflecting greater self-regulation, more adaptive problem-solving, and more positive emotional expression (38,39). Couples high in conscientiousness also report higher satisfaction, highlighting the predictive value of stable personality traits across life domains, including marriage (38). Regarding agreeableness, extraversion, and higher openness, O'Meara and South found that these traits are positively associated with marital satisfaction (39). Collectively, stable personality dispositions appear to shape marital processes and may buffer emotional distress.

Collectively, stable personality traits appear to play an important role in marital satisfaction and are closely linked to individuals' vulnerability to psychological distress. Personality traits are closely associated with emotional functioning, and evidence from the Five Factor Model literature consistently indicates that neuroticism is linked to higher levels of depression, anxiety, and stress, whereas extraversion, conscientiousness, and agreeableness are generally linked to more adaptive emotional and interpersonal functioning (40, 41). These emotional conditions, in turn, are important for marital satisfaction. Elevated psychological distress can undermine relationship functioning through increased negative affect and maladaptive interaction patterns, ultimately lowering marital satisfaction (42). In addition, personality research suggests that personality traits are associated with relationship outcomes and emotional well-being, indicating that psychological distress may represent one pathway through which personality characteristics contribute to variations in marital satisfaction (43).

Our further findings showed that although demographic and socioeconomic factors were related to marital satisfaction, their associations were weaker compared to those of psychological and personality variables. Similar to our study, Dashineau et al., Zhang & Liu, and Rashidi et al. showed that structural and socioeconomic characteristics, such as education level, perceived SES, and social support, contribute to marital satisfaction primarily through psychological and relational processes. Additionally, higher socioeconomic status and greater educational resources were associated with higher marital satisfaction, likely because they

facilitate coping, communication, and relational problem-solving (44-46). This evidence suggests that SES exerts both direct and indirect effects through these psychological and relational mechanisms, with lower SES linked to greater marital distress via stress-related processes. Moreover, the influence of these structural conditions appears to operate largely through psychological pathways and is shaped by marital and family context (46, 47).

From analytical point of view, the supervised machine learning models demonstrated reliable performance, indicating their utility in predicting latent class membership. Both machine learning algorithms exhibited comparable performance and similar patterns of variable importance. Importantly, reduced models including six key predictors performed similarly to full models with 15 variables, suggesting that a limited set of primarily psychological features captured much of the variance in marital satisfaction.

Strengths and limitations

It is worth noting that this study offers a meaningful methodological and conceptual contribution by integrating person-centered and predictive analytic frameworks. By combining latent class analysis with supervised machine learning, it introduces a novel hybrid approach that strengthens the precision of predictive modeling. It further derived empirically grounded levels of marital satisfaction and is among the first to operationalize marital satisfaction across three distinct levels.

The study benefits from the integration of cross-domain and multidimensional predictors, a relatively large sample size, and culturally contextualized evidence from an Iranian population, thereby extending relational research beyond predominantly Western contexts, alongside the use of evaluation metrics tailored to imbalanced data.

Several limitations should be acknowledged. The cross-sectional design does not allow for causal inference or the examination of temporal ordering. Moreover, the reliance on self-reported measures may introduce shared method variance and potential response bias. Another potential limitation is bias arising from the imputation process, as imputed variables in the dataset were used to predict missing values in other variables. While the simultaneous inclusion of psychological distress and personality traits improved the comprehensiveness of the model, their potential overlap may complicate the interpretation of their independent contributions to marital satisfaction. Finally, although highly accurate predictive machine learning models were employed, the modest agreement observed may reflect a limitation of the study, as the complex and multidimensional nature of marital satisfaction may be difficult to adequately capture when operationalized as a categorical construct.

Conclusions

Beyond the specific empirical context of this study, the proposed hybrid framework offers a broadly generalizable analytical strategy applicable across diverse disciplines ranging from public health and clinical psychology to education, organizational behavior, and social sciences wherever multidimensional outcomes and complex psychosocial or behavioral constructs are under investigation. Its core methodological value lies in its dual capacity: first, to uncover latent heterogeneity within populations through LCA, thereby moving beyond one-size-fits-all approaches; and second, to harness the predictive power of ML algorithms for accurately modeling the intricate, nonlinear interplay among a wide array of determinants. This synergy makes the framework particularly well-suited for early risk stratification, population profiling, and the design of theory-driven, tailored interventions that address the specific characteristics of each identified subgroup. Moreover, its demonstrated ability to maintain high predictive accuracy with parsimonious feature sets underscores its scalability and practical feasibility in real-world settings, where cost-effective and interpretable screening tools are increasingly in demand.

We operationalized this framework to identify three latent marital satisfaction profiles. The findings demonstrated that psychological factors, particularly depression, and key personality traits (neuroticism, extraversion, and conscientiousness) consistently outperformed demographic and socioeconomic factors as determinants of profile membership.

Overall, our findings validate the proposed hybrid LCA–ML pipeline as a robust, interpretable, and versatile decision-support framework. By seamlessly integrating person-centered latent structure discovery with algorithmic prediction, it provides a principled methodological pathway for uncovering hidden population subtypes and forecasting complex outcomes under the influence of multiple interacting risk and protective factors thereby offering substantial translational value across a broad spectrum of research and applied domains.

List of Abbreviations

AIC: Akaike Information Criterion

AUC: Area Under Curve

BIC: Bayesian Information Criterion

CFA: Confirmatory Factor Analysis

DASS: Depression Anxiety Stress Scales

FFM: Five-Factor Model

G-Mean: Geometric Mean

IQR: Inter Quartile Range

LCA: Latent Class Analysis

LL: Log-Likelihood

MCC: Matthews Correlation Coefficient

ML: Machine Learning

NEO-FFI: NEO Five-Factor Inventory

NIMAD: National Institute for Medical Research Development

RDAS: Revised Dyadic Adjustment Scale

RFE: Recursive Feature Elimination

ROC: Receiver Operating Characteristic

RPS: Ranked Probability Score

SABIC: Sample-Size Adjusted Bayesian Information Criterion

SES: Socioeconomic Status

SES-SQ: Socioeconomic Status Short-Form Questionnaire

SML: Supervised Machine Learning

VSA: Vulnerability-Stress-Adaptation

Declarations

Ethics approval and consent to participate: This study is a secondary analysis of a project that initially received ethical approval from the National Institute for Medical Research Development (NIMAD) (code: IR.NIMAD.REC.1397.462). For the purpose of this specific secondary study, separate ethical approval was granted by the Ethics Committee of Isfahan University of Medical Sciences (code: IR.MUI.DHMT.REC.1404.012). The study was conducted in accordance with ethical standards, and written informed consent was obtained from all participants before their involvement in the original study.

Consent for publication: Not applicable

Availability of data and materials: The datasets generated and/or analyzed during the current study are not publicly available due to containing information that could compromise the privacy of research participants but are available from the corresponding author on reasonable request.

Competing interests: The authors have no conflicts of interest to declare.

Funding: Not applicable

Authors' contributions: IHR: Data Cleaning, Programming, Software, Formal Analysis, Writing – Original Draft. KMZ: Programming, Software, Review & Editing. AF: Conceptualization, Data Visualization, Verification, Supervision, Writing - Review & Editing.

Acknowledgements: We sincerely thank the families who participated in this study and patiently completed the questionnaires with care and attention. We are also grateful to the Isfahan Department of Education for granting permission to conduct the study in local schools. Additionally, we acknowledge the partial financial support provided by the National Institute for Medical Research Development (NIMAD) for the primary project, on which the current secondary analyses are based.

References

1. Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. Psychol Rev. 2003 Apr;110(2):203-219. doi: 10.1037/0033-295X.110.2.203. PMID: 12747522.
2. Guastello SJ. Nonlinear dynamics in psychology. Discrete Dynamics in Nature and Society. 2001;6(1):11-29. doi:10.1155/S102602260100002X
3. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018 Apr;15(4):233-234. doi: 10.1038/nmeth.4642. Epub 2018 Apr 3. PMID: 30100822; PMCID: PMC6082636.
4. Wan B, Caffo B, Vedula SS. A unified framework on generalizability of clinical prediction models. Front Artif Intell. 2022;5:872720. doi:10.3389/frai.2022.872720
5. Jo B, Hastie TJ, Li Z, Youngstrom EA, Findling RL, Horwitz SM. Reorienting latent variable modeling for supervised learning. Multivariate behavioral research. 2023;58(6):1057-71.doi: https://doi.org/10.1080/00273171.2023.2182753
6. Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. Prev Sci. 2013;14(2):157-168. doi:10.1007/s11121-011-0201-1
7. Hornung R. Ordinal forests. J Classification. 2020;37(1):4-17. doi:10.1007/s00357-018-9302-x
8. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013. doi:10.1007/978-1-4614-6849-3
9. Regan A, Walsh LC, Horton C, Rodriguez A, Kaufman VA. Contextualizing marital dissatisfaction: examining profiles of discordant spouses across life domains. Front Psychol. 2025;16:1458129. doi:10.3389/fpsyg.2025.1458129
10. Sorokowski P, Randall AK, Groyecka A, Frackowiak T, Cantarero K, Hilpert P, et al. Corrigendum: Marital satisfaction, sex, age, marriage duration, religion, number of children, economic status, education, and collectivistic values: Data from 33 countries. Frontiers in psychology. 2017;8:1728.doi: https://doi.org/10.3389/fpsyg.2017.01728
11. Grau I, Miketta L, Ebbeler C, Banse R. Cultural differences in correlations with marital satisfaction: The Vulnerability-Stress-Adaptation Model in 65 countries. Journal of Cross-Cultural Psychology. 2025;56(5):528-57.doi: https://doi.org/10.1177/00220221251323216
12. Sayehmiri K, Kareem KI, Abdi K, Dalvand S, Gheshlagh RG. The relationship between personality traits and marital satisfaction: a systematic review and meta-analysis. BMC psychology. 2020;8(1):15.doi: https://doi.org/10.1186/s40359-020-0383-z
13. Hilpert P, Vowels MR, Mestdagh M, Sels L. Emotion dynamic patterns between intimate relationship partners predict their separation two years later: A machine learning approach. PLoS One. 2023;18(7):e0288048.doi: https://doi.org/10.1371/journal.pone.0288048
14. Ma S-S, Zhang J-T, Song K-R, Zhao R, Fang R-H, Wang L-B, et al. Connectome-based prediction of marital quality in husbands' processing of spousal interactions. Social Cognitive and Affective Neuroscience. 2022;17(12):1055-67.doi: https://doi.org/10.1093/scan/nsac034
15. Dias Martins C, Vergara RC, Khoury B. A machine learning approach for investigating variable importance in relationship and sexual satisfaction: The role of interpersonal mindfulness and psychological safety. Journal of Marital and Family Therapy. 2025;51(2):e70026.doi: https://doi.org/10.1111/jmft.70026

16. Regan A, Walsh LC, Horton C, Rodriguez A, Kaufman VA. Contextualizing marital dissatisfaction: examining profiles of discordant spouses across life domains. Front Psychol. 2025;16:1458129. doi:10.3389/fpsyg.2025.1458129
17. Feizi A, Parsaei R, Heidari Z, Haghighatdoost F, Najmi B. Epidemiology of emotional and psychological problems in Iranian children: Protocol of a cross-sectional study in Isfahan. Journal of Research in Medical Sciences. 2023;28(1):19.doi: https://doi.org/10.4103/jrms.jrms_877_22
18. Busby DM, Christensen C, Crane DR, Larson JH. A revision of the Dyadic Adjustment Scale for use with distressed and nondistressed couples: Construct hierarchy and multidimensional scales. Journal of Marital and family Therapy. 1995;21(3):289-308.doi: https://doi.org/10.1111/j.1752-0606.1995.tb00163.x
19. Isanezhad O, Ahmadi S-A, Bahrami F, Baghban-Cichani I, Farajzadegan Z, Etemadi O. Factor structure and reliability of the Revised Dyadic Adjustment Scale (RDAS) in Iranian population. Iranian journal of psychiatry and behavioral sciences. 2012;6(2):55.doi:
20. Joshanloo M DF, Bakhshi A, Nazemi S, Ghafari Z. Construct validity of NEO personality inventory revised in Iran. Iran J Psychiatry Clin Psychol 2010;3:220-30.doi:
21. Sahebi A, Asghari MJ, Salari RS. Validation of depression anxiety and stress scale (DASS-21) for an Iranian population. 2005.doi:
22. Roohafza H, Feizi A, Gharipour M, Khani A, Dianatkhah M, Sarrafzadegan N, et al. Development and validation of a socioeconomic status short-form questionnaire (SES-SQ). ARYA atherosclerosis. 2021;17(4):1.doi: https://doi.org/10.22122/arya.v17i0.2355
23. Hornung R. Ordinal forests. Journal of Classification. 2020;37(1):4-17.doi: https://doi.org/10.1007/s00357-018-9302-x
24. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213-220.
25. Epstein ES. A scoring system for probability forecasts of ranked categories. J Appl Meteorol Climatol. 1969;8(6):985-987.
26. Lezhnina O, Kismihók G. Latent class cluster analysis: selecting the number of clusters. MethodsX. 2022;9:101747.doi: https://doi.org/10.1016/j.mex.2022.101747
27. Weller BE, Bowen NK, Faubert SJ. Latent class analysis: a guide to best practice. Journal of black psychology. 2020;46(4):287-311.doi: https://doi.org/10.1177/0095798420930932
28. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees: Chapman and Hall/CRC; 2017.
29. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.doi: https://doi.org/10.1023/a:1010933404324
30. Verma AA, Murray J, Greiner R, Cohen JP, Shojania KG, Ghassemi M, et al. Implementing machine learning in medicine. Cmaj. 2021;193(34):E1351-E7.doi: https://doi.org/10.1503/cmaj.202434
31. Lechner M, Okasa G. ORF: Ordered Random Forests. CRAN R package version 012019.

32. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.doi: https://doi.org/10.1093/bioinformatics/btr597
33. Priyatno AM, Widiyaningtyas T. A systematic literature review: recursive feature elimination algorithms. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). 2024;9(2):196-207.doi: https://doi.org/10.33480/jitk.v9i2.5015
34. Shi Y, Whisman MA. Marital satisfaction as a potential moderator of the association between stress and depression. Journal of affective disorders. 2023;327:155-8.doi: https://doi.org/10.1016/j.jad.2023.01.093
35. Dong S, Dong Q, Chen H. Mothers' parenting stress, depression, marital conflict, and marital satisfaction: The moderating effect of fathers' empathy tendency. Journal of affective disorders. 2022;299:682-90.doi: https://doi.org/10.1016/j.jad.2021.12.079
36. Yang L, Yang Z, Yang J. The effect of marital satisfaction on the self-assessed depression of husbands and wives: investigating the moderating effects of the number of children and neurotic personality. BMC psychology. 2023;11(1):163.doi: https://doi.org/10.1186/s40359-023-01200-8
37. Ross JM, Nguyen TP, Karney BR, Bradbury TN. Three tests of the Vulnerability-Stress-Adaptation Model: Independent prediction, mediation, and generalizability. Frontiers in Psychology. 2022;13:921485.doi: https://doi.org/10.3389/fpsyg.2022.921485
38. Munini DN, Asatsa S, Macharia E. The Big Five Personality Traits and Marital Satisfaction Among Couples in Kilungu Deanary in Catholic Diocese of Machakos in Makueni County, Kenya. International Journal of Research and Innovation in Social Science (IJRISS). 2024:476-86.doi: https://doi.org/10.47772/ijriss.2024.809043
39. O'Meara MS, South SC. Big Five personality domains and relationship satisfaction: Direct effects and correlated change over time. Journal of personality. 2019;87(6):1206-20.doi: https://doi.org/10.1111/jopy.12468
40. Soto CJ. How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. Psychological Science. 2019;30(5):711-27.doi: https://doi.org/10.1177/0956797619831612
41. Jeronimus BF KR, Riese H, Ormel J. Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 participants. Psychological Medicine. 2016;46(14):2883–906.doi: https://doi.org/10.1017/s0033291716001653
42. Karney BR BT. The longitudinal course of marital quality and stability: A review of theory, methods, and research. Psychological Bulletin. 1995;118(1):3.doi: https://doi.org/10.1037//0033-2909.118.1.3

43. Malouff JM TE, Schutte NS, Bhullar N, Rooke SE. The five-factor model of personality and relationship satisfaction of intimate partners: A meta-analysis. Journal of research in personality. 2010;44(1):124–27.doi: https://doi.org/10.1016/j.jrp.2009.09.004
44. Dashineau SC, Reed P, Aiken H, Depoy M, South SC. A Meta-Analysis of the Association Between Socioeconomic Status and Marital Satisfaction. Personality and Social Psychology Bulletin. 2024:01461672251410276.doi:
45. Zhang C, Liang Y. The impact of education level on marital satisfaction: Evidence from China. Social Sciences & Humanities Open. 2023;7(1):100487.doi: https://doi.org/10.1016/j.ssaho.2023.100487
46. Rashidi Fakari F, Doulabi MA, Mahmoodi Z. Predict marital satisfaction based on the variables of socioeconomic status (SES) and social support, mediated by mental health, in women of reproductive age: Path analysis model. Brain and behavior. 2022;12(3):e2482.doi: https://doi.org/10.1002/brb3.2482
47. Dobrowolska M, Groyecka-Bernard A, Sorokowski P, Randall AK, Hilpert P, Ahmadi K, et al. Global perspective on marital satisfaction. Sustainability. 2020;12(21):8817.doi: https://doi.org/10.3390/su12218817