STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 1–21.
Published online in International Academic Press (www.IAPress.org)

Financial risk classification of health insurance premium affordability in Indonesia using random forest: evidence from cross-sectional data

Sylvia Samuel [1,*], Hendra Achmadi [1], Tiurida Lily Anita [2]

[1]*Department of Business and Economics, Faculty of Management, Universitas Pelita Harapan, Tangerang, Banten, 15811, Indonesia*
[2]*Department of Digital Communication and Hotel and Tourism, Bina Nusantara University, Jakarta Barat, Daerah Khusus Ibukota Jakarta, 11530, Indonesia*

Abstract Health insurance premium affordability classification represents a financial risk segmentation problem, particularly in emerging markets such as Indonesia, where incomplete information and reliance on observable socioeconomic indicators complicate classification processes. Conventional classification approaches, especially logistic regression, rely on linear and additive assumptions that may not adequately capture heterogeneous and interaction-based relationships among predictors. This creates a gap in understanding how nonlinear models perform in affordability-based classification under data-constrained conditions and how their outputs can be interpreted within a financial risk framework. Using cross-sectional survey data from 148 respondents in Indonesia, this study evaluates the performance of Random Forest relative to logistic regression and a single decision tree in classifying health insurance premium affordability categories. Model performance is assessed using accuracy, precision, recall, F1-score, and area under the curve, with validation through a 70/30 train–test split and ten-fold cross-validation. The results show that Random Forest achieves the highest performance, with accuracy of 63.3%, precision of 0.62, recall of 0.61, F1-score of 0.61, and AUC of 0.70, outperforming the decision tree and logistic regression models. Feature importance analysis identifies monthly income, age, and number of dependents as the most influential predictors of affordability classification. These findings suggest that affordability segmentation may be shaped by nonlinear relationships and interaction effects among socioeconomic variables. Given the relatively small sample size, the results should be interpreted as exploratory evidence rather than definitive proof of model superiority. The study contributes by illustrating the feasibility of applying ensemble machine learning to affordability-oriented classification in a data-constrained emerging-market context while maintaining interpretability through feature importance analysis.

Keywords Cross-Sectional, Financial Risk Classification, Health Insurance Premium Affordability, Random Forest, Socioeconomic

AMS 2010 subject classifications 62H30, 68T05.

DOI: 10.19139/soic-2310-5070-4005

1. Introduction

Health insurance premium affordability represents a financial classification problem in which insurers and policymakers must distinguish between groups that differ in their capacity to sustain premium payments under conditions of incomplete information. In many practical settings, especially where detailed underwriting and claims data are unavailable, classification decisions rely on observable demographic and socioeconomic characteristics such as income, age, household structure, education, and expenditure patterns [1]. These variables function as proxies for financial capacity and exposure, but their effects are rarely independent or linear. Instead, affordability emerges from the interaction between available resources, household obligations, and expected financial pressures.

*Correspondence to: Sylvia Samuel (Email: sylvia.samuel@uph.edu). Department of Business and Economics, Faculty of Management, Universitas Pelita Harapan, Tangerang, Banten, Indonesia (15811).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

This makes the task of assigning individuals to affordability-relevant premium categories both analytically complex and financially consequential [2]. The effectiveness of classification directly influences how risk groups are formed, how pricing-related categories are structured, and how well limited information can be translated into consistent decision rules.

The importance of this classification problem is particularly evident in Indonesia, where socioeconomic heterogeneity creates substantial variation in income, employment conditions, and household obligations [3]. Because access to detailed administrative insurance data is often limited, affordability classification must rely primarily on cross-sectional socioeconomic indicators. Indonesia therefore provides a relevant setting for examining how observable variables can support affordability classification under incomplete information [4].

Conventional classification models, particularly logistic regression, remain widely used because of their interpretability. However, affordability classification often involves nonlinear and interaction-based relationships among income, age, household size, and expenditure patterns. Although decision trees can capture such relationships, they may be unstable in smaller samples. These limitations motivate the evaluation of more flexible machine-learning approached [5].

Despite increasing interest in machine learning methods in finance and insurance analytics, a specific gap remains in the context of affordability classification in emerging markets. There is limited empirical evidence on how Random Forest performs in classifying health insurance premium affordability as a financial risk problem in Indonesia using cross-sectional socioeconomic data [6]. Existing studies have often focused on different outcomes, such as claims prediction, fraud detection, or insurance uptake, rather than on affordability classification as a forward-looking grouping problem. In addition, much of the available evidence is derived from data-rich environments where detailed administrative records are available, which reduces the relevance of proxy-based classification challenges [7].

Even when machine learning models are applied, the emphasis is frequently placed on predictive performance without sufficient attention to how classification results can be interpreted within a financial risk framework. This creates a gap not only in empirical application but also in the integration of methodological performance with financial interpretation.

In response to this gap, the objective of this study is to classify health insurance premium affordability categories in Indonesia using Random Forest and to compare its performance with logistic regression and decision tree models. The study adopts a cross-sectional classification approach based on survey data, focusing on how observable socioeconomic and demographic variables can be used to distinguish affordability groups [8]. The analysis is explicitly predictive rather than causal and does not attempt to estimate actuarial premiums or infer population-level effects beyond the available data. Instead, it evaluates whether a more flexible ensemble classification method can improve the identification of affordability categories relative to conventional benchmarks [9]. By maintaining a clear alignment between objective, data, and method, the study ensures that its claims remain within the limits of the evidence. The study should be interpreted as an exploratory classification exercise rather than a deployment-ready predictive system. The sample consists of 148 respondents and therefore provides a useful proof-of-concept for affordability classification under limited-data conditions. Consistent with small-sample machine-learning research, the analysis emphasizes comparative model behavior and classification feasibility rather than population-level generalization.

The analysis is guided by one primary research question and two supporting questions. The primary question asks how effectively Random Forest classifies health insurance premium affordability categories in Indonesia relative to benchmark models. This focuses directly on comparative model performance and reflects the core design of the study [10]. The first secondary question examines which variables most strongly influence affordability classification, recognizing that predictive accuracy alone does not explain the structure of classification outcomes. The second secondary question considers what the observed classification pattern reveals about affordability-related financial grouping, including how categories are distinguished and where misclassification occurs [11]. Together, these questions connect model evaluation with substantive interpretation.

The study contributes in three specific ways that are consistent with its design. Methodologically, it compares linear, single-tree, and ensemble classification approaches within the same affordability classification problem under cross-sectional data constraints, providing a controlled evaluation of model performance differences.

Empirically, it offers evidence from Indonesia, a context characterized by socioeconomic heterogeneity and limited access to detailed insurance data, thereby extending the relevance of classification analysis beyond data-rich environments [12]. Practically, it provides an interpretable, data-driven basis for affordability-oriented classification by identifying the variables that contribute most strongly to grouping outcomes and by demonstrating how classification performance varies across models. These contributions remain proportionate to the study's scope and do not extend beyond what the data and methods can support.

The remainder of the article is structured to follow the same analytical logic established in this introduction. The next section reviews the relevant literature and develops the analytical gap in relation to financial risk classification and machine learning approaches. This is followed by a conceptual framework that defines the constructs used in the study. The methodology section then describes the data, variables, preprocessing procedures, model specifications, and evaluation strategy. The results section presents the empirical findings, including descriptive patterns, model comparison, classification error structure, and feature importance [13]. The discussion interprets these findings in relation to financial risk classification and existing research, and the final section concludes by summarizing the study's contribution and limitations.

2. Literature Review## Financial risk classification in insurance

Financial risk classification in insurance is designed to group individuals into categories that reflect differences in expected financial exposure and capacity to sustain insurance participation. Within this framework, classification serves as a bridge between incomplete information and structured decision-making, allowing insurers to assign individuals to premium-relevant groups using observable characteristics. A consistent theme across the literature is that classification relies on proxy variables such as demographic and socioeconomic indicators when direct measures of risk, including claims histories or underwriting records, are unavailable [14]. These proxies simultaneously capture dimensions of financial capacity and exposure, making classification both a statistical and an economic exercise.

The dominant pattern in this literature is that classification frameworks implicitly combine two objectives: differentiating risk and aligning premiums with affordability. However, these objectives are not equivalent. Risk reflects expected healthcare utilization or cost exposure, while affordability reflects the ability to sustain recurring payments under financial constraints. Studies consistently show that these dimensions overlap but are not identical, as individuals with similar risk profiles may differ substantially in their capacity to pay, and vice versa [15]. This creates a structural tension in classification systems, particularly when affordability is treated as a secondary consideration rather than as a core classification outcome.

A key limitation emerges from how this dual role is operationalized. Many classification approaches assume that observable variables can be aggregated linearly to approximate both risk and affordability. This assumption simplifies modeling but overlooks the multidimensional and interaction-based nature of financial constraints. In contexts where data are incomplete and proxies dominate; such simplifications can lead to classifications that fail to reflect actual affordability patterns [16]. This limitation suggests that affordability should be treated explicitly as a classification outcome shaped by financial capacity and household exposure, rather than as a residual of risk-based grouping. Addressing this issue requires a closer examination of the determinants of affordability and how they interact in shaping classification outcomes.

2.1. Socioeconomic determinants of premium affordability

The literature on premium affordability consistently identifies socioeconomic variables as the primary determinants of an individual's ability to sustain insurance payments. Income is the most direct indicator of financial capacity, as it defines the resource base available for consumption and risk management. Age is widely used as a proxy for life-cycle position, influencing both expected healthcare needs and financial stability. Household structure, typically measured through the number of dependents, captures the distribution of financial obligations within a household [17]. Additional variables such as education, occupation, and expenditure proxies provide complementary information on economic positioning, consumption behavior, and access to resources.

Across studies, a clear pattern is that these variables are systematically associated with affordability outcomes. However, their effects are not uniform. The influence of income, for example, may vary depending on household size, as higher income may not translate into greater affordability when financial obligations are also higher. Similarly, the impact of age may differ across income groups, and expenditure proxies may capture different aspects of financial pressure depending on context [18]. This indicates that affordability is shaped by the interaction among multiple variables rather than by isolated effects.

Despite this recognition, a key limitation persists in how these determinants are modeled. Many studies rely on approaches that treat variables as independent contributors, implicitly assuming that their effects can be separated and combined additively. This assumption is particularly restrictive when proxy variables are used, as these variables often capture overlapping dimensions of financial behavior. As a result, models may fail to represent the conditional relationships that define affordability in practice [19]. This limitation highlights the need for modeling approaches that can accommodate interaction effects and nonlinear relationships, leading to the examination of conventional statistical classification models and their constraints.

2.2. Conventional statistical classification models

Conventional statistical models, particularly logistic regression and related econometric approaches, have been widely applied to classification problems in insurance and financial analysis. Their primary advantage lies in interpretability, as model coefficients provide direct insight into the relationship between predictors and classification outcomes. This transparency supports hypothesis testing, regulatory compliance, and the communication of results [20]. In relatively stable and homogeneous datasets, these models can provide reliable classification performance.

The dominant pattern in their application is that they perform effectively when relationships between variables and outcomes are approximately linear and additive. However, premium affordability classification often deviates from this structure. When relationships involve interactions such as the combined effect of income and household size or nonlinear thresholds such as affordability limits that change across income brackets linear models may impose constraints that reduce their representational accuracy [21]. Although interaction terms can be introduced, they must be specified explicitly, which becomes impractical when the number of potential interactions is large or when the underlying relationships are not fully known.

Decision tree models offer an alternative by allowing nonlinear partitioning of the data without requiring predefined functional forms. They can capture interaction effects through recursive splitting and provide intuitive classification rules. However, their performance is often limited by instability, as small changes in the data can produce different tree structures. This high variance reduces their reliability, particularly in cross-sectional datasets with limited observations [22]. The combined limitations of linear models and single-tree approaches indicate that while these methods provide useful benchmarks, they may not fully capture the complexity of affordability classification. This motivates the consideration of more flexible approaches that can balance nonlinearity and stability.

2.3. Machine learning approaches in insurance and finance

Machine learning methods have gained prominence in insurance and financial modeling due to their ability to identify complex patterns in data without imposing strict parametric assumptions. Among these methods, tree-based approaches and ensemble techniques are particularly relevant for classification tasks [23]. Decision trees enable flexible partitioning of the data space, allowing models to represent nonlinear relationships and interactions among variables. However, as noted, single-tree models are prone to instability.

A consistent finding across the literature is that ensemble methods, particularly Random Forest, improve upon single-tree models by addressing this instability. Random Forest constructs multiple decision trees using bootstrapped samples and random subsets of predictors, then aggregates their predictions. This process reduces variance and improves generalization, leading to more stable and accurate classification outcomes [24]. Recent insurance-related applications continue to demonstrate the usefulness of feature-importance analysis and explainable machine-learning techniques. Yego [8] found that Random Forest effectively identified dominant socioeconomic predictors of health insurance uptake, while [9] showed that machine-learning models can provide

interpretable insights into insurance-related financial outcomes. These studies suggest that machine-learning approaches can support both prediction and variable interpretation in insurance settings.

Empirical applications in finance and insurance consistently show that Random Forest outperforms both linear models and individual decision trees when data exhibit heterogeneity and complex relationships.

Another important feature of Random Forest is its ability to provide feature importance measures. These measures allow researchers to identify the relative contribution of each variable to the classification outcome, offering insight into the structure of the data. This partially addresses the interpretability challenge often associated with machine learning models [25]. However, a limitation remains in how these models are evaluated. Much of the literature focuses on predictive performance metrics without fully connecting these metrics to the financial meaning of classification outcomes. In the context of affordability, improved accuracy must be interpreted in terms of how well the model captures financial capacity and household exposure [26]. This gap between predictive performance and financial interpretation indicates that machine learning methods need to be assessed not only for accuracy but also for their ability to support meaningful classification within a financial framework.

2.4. Emerging-market and data-constrained evidence

The majority of empirical research on insurance classification and machine learning applications has been conducted in data-rich environments, where detailed administrative records provide direct measures of risk and cost. In such settings, models can rely on comprehensive datasets that reduce the need for proxy variables [27]. In contrast, emerging markets often present data-constrained environments in which researchers must rely on cross-sectional survey data and observable socioeconomic indicators.

A recurring pattern in studies conducted in these contexts is the increased importance of proxy variables, which introduce additional complexity into the classification problem. These proxies often capture multiple dimensions of financial behavior, making it difficult to isolate their individual effects. In addition, smaller sample sizes and heterogeneous populations amplify the challenges associated with model specification and stability [28]. These conditions can weaken the performance of models that depend on strong assumptions or stable relationships.

Despite the relevance of these challenges, literature remains limited in its treatment of affordability classification in emerging markets. Existing studies frequently focus on related outcomes, such as insurance participation or expenditure patterns, rather than on the classification of premium affordability itself. Moreover, when machine learning methods are applied, they are often evaluated in isolation from the financial interpretation of their results. This creates a gap in understanding how classification models perform under realistic data constraints and how their outputs relate to affordability-based financial grouping [29]. Addressing this gap requires empirical analysis that combines model comparison with financial interpretation in a data-constrained setting.

2.5.1. Machine learning and financial inclusion in developing markets. Research in developing and emerging economies increasingly applies machine-learning techniques to financial inclusion, insurance accessibility, and affordability-related decision problems. In these environments, researchers frequently rely on proxy variables because detailed administrative records are unavailable. Existing evidence suggests that machine-learning methods can improve classification performance under heterogeneous socioeconomic conditions, although their effectiveness depends heavily on data quality, sample size, and representativeness. Consequently, applications in data-constrained settings should be interpreted as exploratory and context-specific rather than universally generalizable. This study contributes to that growing literature by examining affordability classification using proxy-based socioeconomic indicators in Indonesia.

2.5. Gap synthesis

Existing research shows that affordability classification depends on observable socioeconomic characteristics and may involve nonlinear interactions among predictors. However, limited evidence examines Random Forest performance in classifying health insurance premium affordability in emerging-market settings using cross-sectional data. In addition, few studies connect predictive performance with affordability-oriented financial interpretation. This study addresses these gaps by comparing Random Forest, logistic regression, and decision tree models for affordability classification in Indonesia while interpreting the results within a financial risk classification framework.

2.6. Conceptual framework

The conceptual framework organizes the classification of health insurance premium affordability as a financial risk problem under conditions of incomplete information. It is constructed in direct response to the gap identified in the literature, which highlights the need to link classification performance with financial interpretation using observable socioeconomic data. Rather than relying on a single formal theory, the framework structures the problem using empirically grounded constructs that can be directly operationalized in a classification model. The framework is therefore intentionally minimal and functional, ensuring that every construct introduced is measured, analyzed, and interpreted within the study. The framework is based on three core components: financial capacity, household exposure, and affordability classification. Socioeconomic and demographic variables serve as observable inputs that define these components. Financial capacity captures the resource side of affordability, while household exposure captures the constraint side. The outcome variable affordability classification represents the grouping of individuals into premium categories based on the interaction between these two constructs [30]. This structure provides a clear and consistent link between variables, constructs, and classification outcomes, ensuring alignment with the research objective. All observed variables are treated as indicators of these constructs and enter the model directly without latent transformation.

2.7. Financial capacity

Financial capacity is defined as the household's observable ability to sustain health insurance premium payments using available economic resources. This construct is operationalized through variables that directly or indirectly represent financial resources and consumption patterns. Income is the central indicator of financial capacity, as it reflects the level of resources available for expenditure and risk management [31]. However, income alone does not fully determine affordability, because households differ in how income is allocated across essential and non-essential expenditures.

To capture this variation, expenditure-related indicators are included as proxies for consumption behavior and financial commitments. These variables provide information on how resources are distributed and therefore refine the measurement of financial capacity. Within the framework, financial capacity is treated as a continuous construct that influences the probability of belonging to different affordability categories. It does not define fixed thresholds but contributes to classification through its interaction with other variables. This ensures that the construct remains consistent with the classification-based design of the study.

2.8. Household exposure

Household exposure is defined as the set of demographic factors that determine the level of financial obligation within a household and thereby constrain the use of available resources. This construct is operationalized through variables such as age and number of dependents. Age represents life-cycle position, which influences both income stability and expected healthcare needs. Different age groups may face different financial conditions, affecting how resources are allocated to insurance. The number of dependents captures the extent of household financial obligations. A higher number of dependents increases the demand on household resources, reducing the share of income that can be allocated to insurance premiums [32]. This variable directly reflects the pressure placed on financial capacity by household structure. Together, age and dependents define the exposure dimension of affordability, indicating how available resources are constrained by demographic conditions. This construct complements financial capacity by representing the demand side of affordability, ensuring that classification reflects both resource availability and financial obligations.

10. *Affordability classification outcome*

Affordability classification is defined as the assignment of individuals to grouped categories of health insurance premium affordability. The dependent variable represents ranges of premium levels that individuals consider financially manageable, rather than actuarial premium calculations. This distinction is critical because the study focuses on classification rather than prediction of exact values.

The classification outcome reflects the combined influence of available economic resources and household obligations. Individuals are grouped into categories based on how their resources and obligations interact to determine affordability [33]. This structure operates the concept of financial risk classification by translating observable characteristics into discrete affordability groups. The outcome variable is therefore fully aligned with the research objective and provides a measurable representation of affordability within a classification framework.

11. *Relationship among constructs*

The framework specifies that affordability classification is determined by the interaction between financial capacity and household exposure. Socioeconomic variables define financial capacity, while demographic variables define household exposure. These constructs are not independent; instead, their effects are conditional on each other. For example, the impact of income on affordability may differ depending on the number of dependents, and the financial implications of age may vary across income levels. This interaction-based structure implies that classification boundaries are not linear [34]. The probability of belonging to a particular affordability category depends on combinations of variables rather than on individual variables in isolation. This relationship is central to the framework because it defines the nature of the classification problem and directly informs the choice of analytical method.

12. *Implications for classification approach*

Given that affordability classification is shaped by nonlinear and interaction-based relationships, the framework supports the use of classification models that can capture such complexity. Linear models, such as logistic regression, assume additive and independent effects, which may limit their ability to represent interaction-driven structures. Decision tree models allow for nonlinear partitioning but may lack stability when used individually. Ensemble methods, particularly Random Forest, are consistent with the framework because they combine multiple decision trees to capture complex relationships while reducing variance [35]. By allowing flexible partitioning of the data and implicit modeling of interactions, Random Forest aligns with the structure of affordability classification defined in this framework. The use of multiple models in the study logistic regression, decision tree, and Random Forest provides a comparative basis for evaluating how well different approaches represent the relationships specified in the framework.

13. *Conceptual model*

The conceptual model integrates the constructs and relationships into a single structure. Socioeconomic variables such as income, education, occupation, and expenditure proxies define financial capacity, while demographic variables such as age and number of dependents define household exposure. These constructs jointly determine affordability classification, which is represented as grouped premium categories. As illustrated in Figure 1, the classification process is implemented through alternative modeling approaches to evaluate how different assumptions affect classification outcomes. The model remains fully consistent with the non-causal design of the study [36]. It does not assume causal relationships between variables but instead represents patterns of association used for classification. All constructs included in the framework are directly operationalized in the methodology and appear in the results, ensuring conceptual consistency across sections.

14. *Hypotheses or analytical expectations*

This study formulates hypotheses consistent with its predictive classification design and the conceptual framework linking financial capacity and household exposure to affordability classification. The hypotheses are derived directly from the literature, which identifies nonlinear relationships among socioeconomic variables, and from the framework, which specifies interaction-based classification logic. Each hypothesis is stated in a testable form and aligned with the variables, models, and evaluation procedures used in the analysis. The hypotheses do not imply causality but define expected classification patterns and model performance under a non-causal, cross-sectional design.

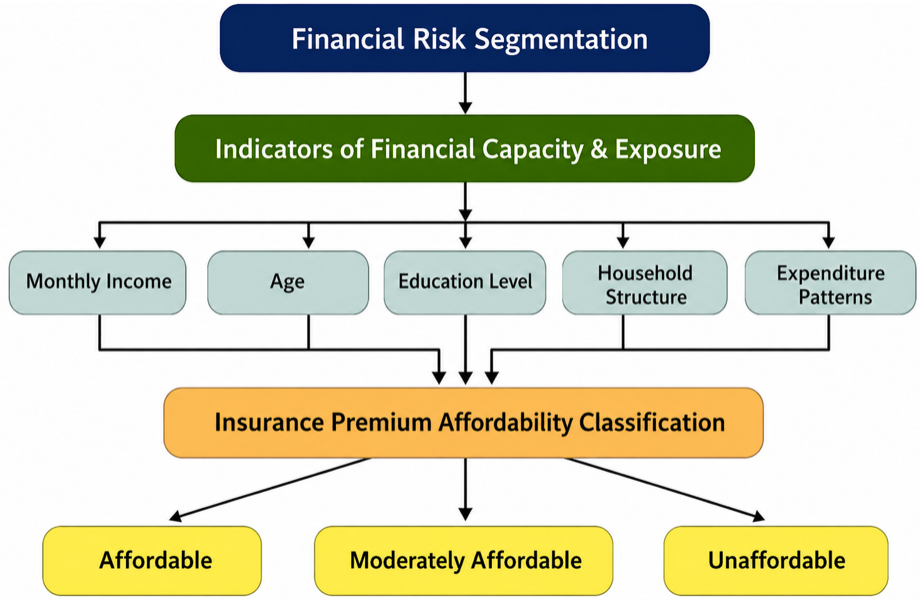

Figure 1. Figure 1. Conceptual framework of health insurance premium affordability classification

2.14.1. Hypothesis 1. H1. Random Forest achieves higher classification performance than logistic regression and decision tree models in classifying health insurance premium affordability categories. This hypothesis follows from the expectation that affordability classification is shaped by nonlinear relationships and interaction effects between financial capacity and household exposure variables. Logistic regression assumes linear and additive effects, while a single decision tree captures nonlinearities but is sensitive to sample variation. Random Forest, as an ensemble method, aggregates multiple trees and reduces variance, allowing it to better approximate complex classification boundaries. This hypothesis is testable using comparative performance metrics, including accuracy, precision, recall, F1-score, and area under the curve, evaluated on out-of-sample data and cross-validation.

2.14.2. Hypothesis 2. H2. Monthly income, age, and number of dependents are the most influential predictors of health insurance premium affordability classification. This hypothesis is derived directly from the conceptual framework, which defines financial capacity and household exposure as the primary constructs shaping affordability. Income operationalizes financial capacity, while age and number of dependents operationalize household exposure. These variables are expected to dominate classification because they represent the core dimensions of resource availability and financial obligation. This hypothesis is testable through feature importance analysis within the Random Forest model, ensuring consistency between predictive modeling and variable interpretation.

2.14.3. Hypothesis 3. H3. Misclassification occurs more frequently between adjacent affordability categories than between non-adjacent categories. This hypothesis reflects the ordered structure of the dependent variable, where affordability categories represent progressively increasing premium levels. Because adjacent categories share similar underlying financial characteristics, classification boundaries between them are expected to be less distinct than those separating distant categories. As a result, classification errors are more likely to occur between neighboring groups. This hypothesis is testable using confusion matrix analysis, where the distribution of misclassified observations across category transitions can be examined.

3. Materials and Methods## Research design

This study employs a quantitative, cross-sectional research design based on supervised multi-class classification. The design is appropriate because the research objective is to assign individuals to predefined health insurance

premium affordability categories using observable socioeconomic and demographic variables. The study is predictive rather than causal, focusing on classification performance and pattern recognition instead of estimating causal effects. This design directly operationalizes the conceptual framework, where financial capacity and household exposure are mapped to affordability categories. A comparative modeling strategy is used to evaluate linear (logistic regression), nonlinear (decision tree), and ensemble (Random Forest) approaches under identical data conditions, ensuring that model performance differences reflect structural assumptions rather than data variation.

3.1. Study context and data source

The empirical setting is Indonesia, characterized by heterogeneous socioeconomic conditions and limited access to administrative insurance data. Data was collected through a structured online survey conducted between September 2024 and March 2025. The survey captured demographic, socioeconomic, and expenditure-related information relevant to affordability classification. This data source is appropriate because the study models perceived affordability, which depends on self-reported financial conditions rather than insurer-determined premiums. The use of survey data is therefore consistent with the classification objective under incomplete information conditions. Survey recruitment was conducted through online distribution channels, including personal networks, university-affiliated networks, and social-media sharing. Consequently, the sample is more representative of digitally connected respondents than of the Indonesian population.

3.2. Sample and sampling

A total of 202 responses were initially obtained. Data screening was conducted to ensure completeness, logical consistency, and plausibility. Observations with missing key variables, contradictory responses, or implausible values were excluded, resulting in a final sample of 148 observations. The study uses non-probability convenience sampling, which is appropriate for exploratory modeling in data-constrained environments but limits statistical generalization. The retained sample exhibits variation across income, age, household structure, and other relevant variables, allowing meaningful classification analysis. The limitation of external validity is explicitly acknowledged, and conclusions are restricted to the observed data context. Because recruitment occurred through online convenience sampling, younger and digitally active individuals were more likely to participate. As a result, the sample should not be interpreted as nationally representative. The findings are most appropriately viewed as evidence from an exploratory affordability-classification sample rather than as estimates of population-level affordability behavior.

3.3. Variable definition and operationalization

All variables are defined to match the conceptual framework without deviation. Socioeconomic variables operationalize financial capacity, while demographic variables operationalize household exposure. The dependent variable represents affordability classification. Table 1 presents the definitions, coding schemes, corresponding constructs, and analytical roles of all variables used in the study.

Table 1. Variable definitions, coding, and analytical role

Variable	**Type**	**Measurement/Coding**	**Construct**	**Analytical role**
Premium affordability category	Categorical	$<$ IDR 3M; IDR 3–5M; $>$ IDR 5M	Affordability classification	Outcome
Age	Continuous	Years (standardized)	Household exposure	Predictor
Gender	Categorical	Binary encoding	Control	Predictor
Marital status	Categorical	Encoded categories	Household exposure	Predictor

Variable	Type	Measurement/Coding	Construct	Analytical role
Number of dependents	Ordinal	None; 1–2; $\geq$3	Household exposure	Predictor
Education	Categorical	Encoded levels	Financial capacity proxy	Predictor
Occupation	Categorical	Encoded categories	Financial capacity proxy	Predictor
Industry	Categorical	Encoded categories	Financial capacity proxy	Predictor
Domicile	Categorical	Encoded regions	Contextual proxy	Predictor
Monthly income	Ordinal	Ordered income (encoded numerically and standardized)	Financial capacity	Predictor
Transportation expenditure	Continuous	Monthly value (standardized)	Financial capacity proxy	Predictor
Electricity usage	Continuous	Monthly value (standardized)	Financial capacity proxy	Predictor
Asset proxy	Ordinal	Ownership indicators	Financial capacity proxy	Predictor

Affordability category construction. The dependent variable was derived from a survey question asking respondents to indicate the monthly health insurance premium level that they believed would be financially manageable for their household under current conditions. Respondents selected one of three affordability ranges: less than IDR 3 million, IDR 3–5 million, or greater than IDR 5 million. The measure therefore captures perceived affordability thresholds rather than observed premium payments or actuarially determined affordability. While perceived affordability provides useful information regarding financial tolerance and willingness to allocate resources toward insurance, it may also contain subjective judgment and reporting bias. Consequently, the classification outcome should be interpreted as perceived affordability segmentation rather than direct measurement of objective payment capacity.

All constructs introduced in the framework are fully operationalized, and no additional variables are introduced later in the analysis.

3.4. Data preprocessing

Data preprocessing ensures consistency across models. Missing or inconsistent observations were removed during screening. Categorical variables were converted into numerical forms using label encoding. Label encoding was applied to maintain a parsimonious feature space given the relatively small sample size. Although one-hot encoding is commonly recommended for logistic regression, preliminary assessment indicated that expanding multiple categorical variables into numerous dummy variables would substantially increase model dimensionality relative to the available observations. To ensure consistent preprocessing across models, the same encoded dataset was used for logistic regression, decision tree, and Random Forest estimation.

Continuous variables, including income and expenditure proxies, were standardized to ensure comparability across scales. The dataset was split into training (70%) and testing (30%) subsets using stratified sampling to preserve the distribution of affordability categories across both datasets. This separation ensures that model evaluation reflects out-of-sample predictive performance rather than in-sample fit, strengthening internal validity.

3.5. Model specification

3.6.1. Logistic regression. Logistic regression is used as the linear benchmark model. It estimates class membership probabilities under the assumption of linear relationships between predictors and the log-odds of classification. This model provides interpretability but is constrained in capturing nonlinear interactions.

3.6.2. Decision tree. The decision tree model is used as a nonlinear, non-ensemble benchmark. It partitions the data into homogeneous groups using recursive splitting. While it captures nonlinear relationships, it is sensitive to sample variation and may produce unstable classification boundaries.

3.6.3. Random forest. Random Forest is the primary model. It constructs multiple decision trees using bootstrapped samples and random feature selection. Predictions are aggregated through majority voting. This approach reduces variance and allows flexible modeling of interaction effects, consistent with the conceptual framework.

3.6. Hyperparameter tuning and validation

Model parameters are optimized using grid search combined with ten-fold cross-validation. In this procedure, the dataset is partitioned into ten folds, with each fold used once as a validation set. Performance is averaged across folds to reduce dependence on a single data split. Model stability is assessed by examining variation in performance metrics across folds and alternative random splits. This approach ensures robustness and reduces overfitting risk. The final Random Forest specification selected through grid search consisted of 100 trees, maximum depth 5, minimum samples split 4, and minimum samples leaf 2. These parameters produced the highest average cross-validation performance and were subsequently used for test-set evaluation.

3.7. Evaluation metrics

Model performance is evaluated by using multiple metrics to capture different aspects of classification quality. Accuracy measures overall correctness. Precision evaluates the correctness of predicted class assignments. Recall measures the proportion of actual cases correctly identified. The F1-score balances precision and recall. The area under the ROC curve evaluates discriminatory ability. For multi-class classification, AUC was calculated using a one-versus-rest framework and reported as the macro-average across classes. Because AUC values near 0.70 indicate moderate rather than strong discriminatory performance, the metric is interpreted cautiously and in conjunction with accuracy, precision, recall, and F1-score.

The use of multiple metrics is necessary because single measures may not fully capture performance in multi-class classification.

3.8. Statistical and analytical integrity

The analytical strategy is designed to ensure methodological rigor. Observations are independent and non-duplicated. The class structure is clearly defined and consistently applied across models. Preprocessing steps are identical for all models, ensuring comparability. Class distribution was examined prior to model estimation and did not indicate severe imbalance; therefore, no resampling techniques were applied. Out-of-sample testing and cross-validation address overfitting. Although the sample size is relatively small for machine-learning applications, model evaluation relies on repeated validation rather than single estimates. Robustness checks are conducted using alternative data splits and parameter settings. The analysis does not assume causal relationships and therefore avoids violations of causal inference assumptions.

10. *Validity, reliability, and ethics*

Internal validity is supported through train–test separation and cross-validation. Reliability is ensured by consistent preprocessing and repeated evaluation procedures. External validity is limited due to convenience sampling and self-reported data, and findings are interpreted within this constraint. Ethical standards are maintained through voluntary participation and informed consent. No personally identifiable information was collected, and all data was anonymized prior to analysis. The study involves minimal risk, as it uses non-sensitive socioeconomic information.

11. *Methodological limitations*

The study is subject to several limitations. The sample size is relatively small, which may affect generalizability. Convenience sampling introduces potential selection bias. Self-reported variables may contain measurement error.

The dependent variable reflects perceived affordability rather than actual premiums, limiting interpretation as pricing outcomes. These limitations are consistent with the study's exploratory classification design and do not invalidate the analytical approach but define its scope.

4. Results

This section presents empirical findings based on the analytical procedures described in Section Results are organized by analytical stage and aligned with the stated hypotheses. All reported values are derived from the test dataset and cross-validation procedures, with consistent variable definitions and classification categories.

4.1. Sample characteristics

The final analytical sample consists of 148 observations after data screening. As shown in Table 2, the distribution of respondents across demographic and socioeconomic variables, as well as premium affordability categories, exhibits variation across all predictor variables and outcome categories.

Table 2. Descriptive statistics of respondents and premium affordability categories (n = 148)

Variable	**Category**	**Frequency**	**Percentage (%)**
Age	18–20	67	45.6
	21–23	30	20.4
	23–26	51	34.0
Gender	Male	72	48.9
	Female	76	51.1
Education	Secondary	72	48.9
	Undergraduate	76	51.1
Dependents	None	55	37.2
	1–2	62	42.0
	$\geq$3	31	20.8
Income	$<$ IDR 5 million	58	39.0
	IDR 5–10 million	66	44.6
	$>$ IDR 10 million	24	16.4
Premium affordability	$<$ IDR 3 million	54	36.5
	IDR 3–5 million	68	46.0
	$>$ IDR 5 million	26	17.5

The distribution reflects variation across all predictor variables and the dependent variable categories. Following the 70/30 partition, the training dataset contained 104 observations, and the test dataset contained 44 observations, corresponding to the predefined 70/30 partition of the 148 valid observations. The distribution of affordability categories remained broadly comparable across both subsets, reducing the likelihood that performance differences were driven by class composition. Figure 2 illustrates the distribution of the key demographic and socioeconomic characteristics of the respondents.

4.2. Model performance comparison (H1)

Model performance is evaluated using the test dataset. As shown in Table 3, classification metrics are reported for logistic regression, decision tree, and Random Forest models.

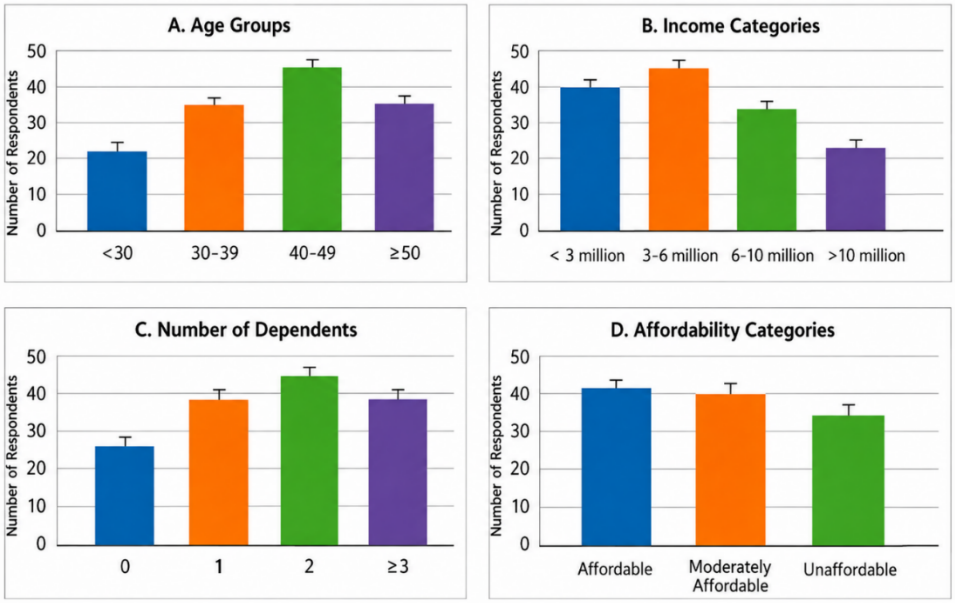

Figure 2. Figure 2. Distribution of key demographic and socioeconomic variables

Table 3. Comparative classification performance of models

Model	Accuracy (%)	Precision	Recall	F1-score	AUC
Logistic regression	55.0	0.54	0.52	0.53	0.60
Decision tree	58.2	0.56	0.55	0.55	0.63
Random Forest	63.3	0.62	0.61	0.61	0.70

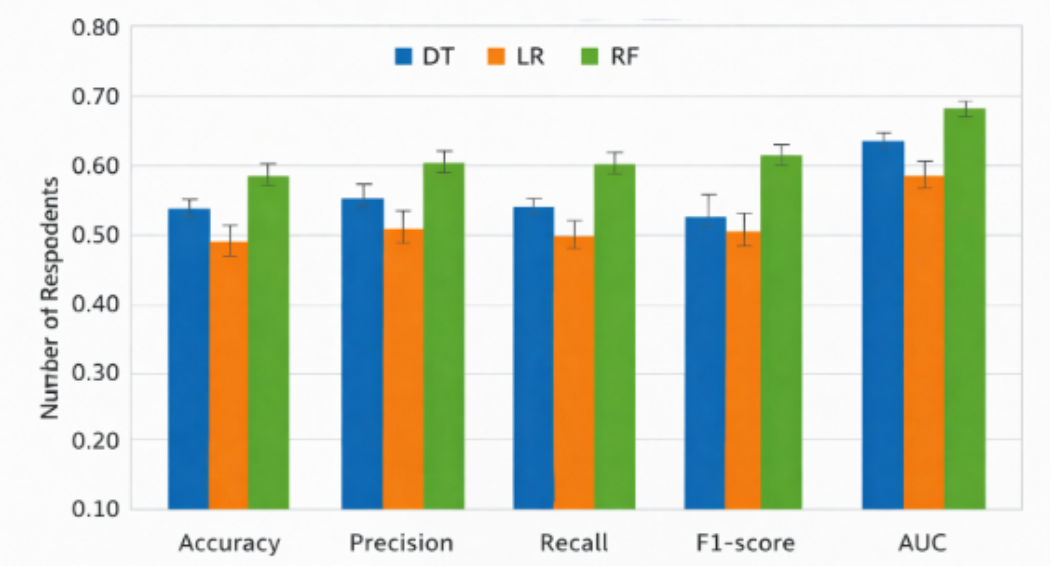

Figure 3. Figure 3. Model performance comparison across evaluation metrics

While Random Forest achieved the highest performance among the evaluated models, the observed differences should be interpreted cautiously given the small sample size. The results indicate relative improvement within the analysed dataset rather than definitive evidence of superiority across broader populations. To assess variability in model performance, cross-validation results are also reported in Table 4. The standard deviations indicate that performance differences between models remained relatively stable across validation folds. Random Forest records the highest value across all reported metrics, as illustrated in Figure 3. Logistic regression shows the lowest values, while the decision tree produces intermediate results. These results indicate that the ensemble model provides improved classification performance relative to the benchmark models in this sample.

4.3. Cross-validation and model stability

Ten-fold cross-validation results are presented in Table 4 to assess stability across data partitions. For the Random Forest model, average accuracy across folds ranged from approximately 60.2% to 64.3%, indicating modest but reasonably stable variation under repeated validation.

Table 4. Ten-fold cross-validation summary by model

Model	Mean accuracy (%)	Std. deviation	Mean precision	Mean recall	Mean F1-score
Logistic regression	54.8	2.1	0.53	0.52	0.52
Decision tree	57.6	2.5	0.55	0.54	0.54
Random Forest	62.5	1.8	0.61	0.60	0.60

Performance values remain within narrow ranges across folds for all models. Random Forest shows the lowest standard deviation among models. The observed standard deviations indicate limited variation across folds, supporting model stability.

4.4. Classification error structure (H3)

The classification error structure is evaluated using the confusion matrix of the Random Forest model. Table 5 indicates the distribution of predicted versus actual categories. Percentages shown in Table 5 are row-normalized and therefore sum to 100% within each actual affordability category.

Table 5. Confusion matrix for random forest (% of cases)

Actual \ Predicted	$<$ 3 million	3–5 million	$>$ 5 million
$<$ 3 million	71.0	24.5	4.5
3–5 million	18.2	68.3	13.5
$>$ 5 million	10.7	26.8	62.5

Transportation expenditure received a relatively low importance score. One possible explanation is that transportation spending exhibited less variation across respondents than income and household-structure variables. In addition, transportation expenses may reflect lifestyle preferences or geographical circumstances that are only indirectly related to perceived insurance affordability. Consequently, the variable contributed less to class separation than direct indicators of financial resources and obligations.

Correct classification rates exceed 60% for all categories, as illustrated by the confusion-matrix heatmap in Figure 4. Misclassified observations are distributed across adjacent and non-adjacent categories, with higher proportions observed in transitions between neighboring categories. Feature importance scores were normalized to sum to 1.00, allowing direct comparison of the relative contribution of each predictor to the Random Forest classification process.

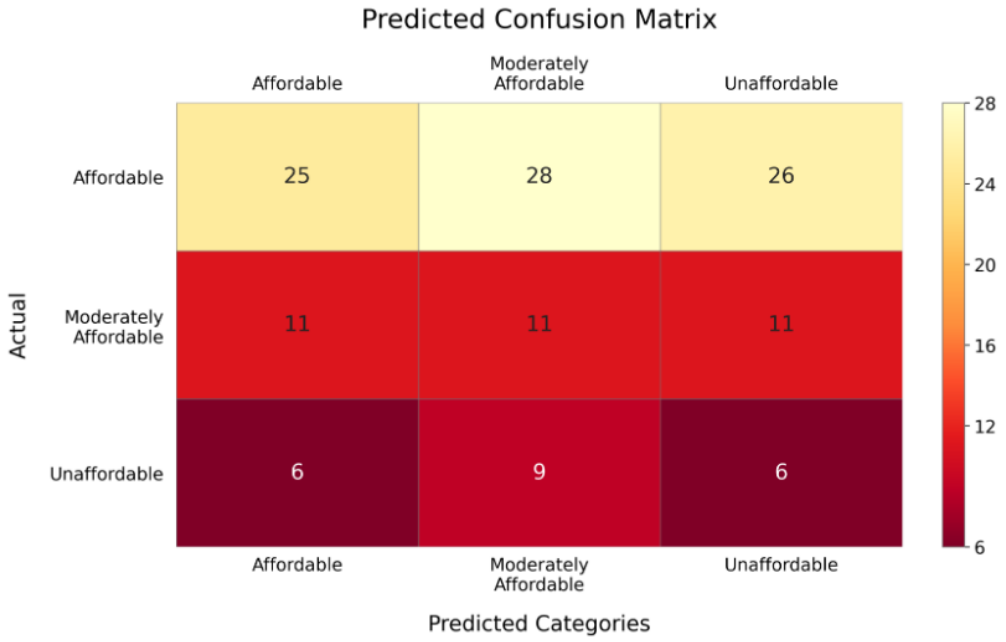

Figure 4. Figure 4. Heatmap of random forest confusion matrix

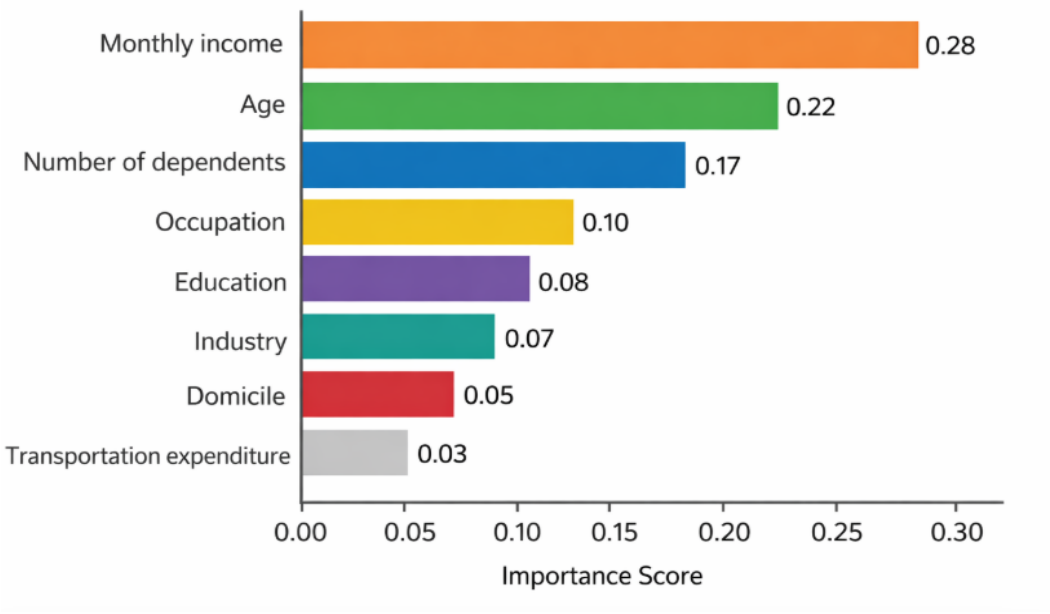

Figure 5. Figure 5. Feature importance of predictors in random forest classification

4.5. Feature importance (H2)

Feature importance scores derived from the Random Forest model are presented in Table 6.

Table 6. Feature importance rankings from Random Forest

Rank	Predictor	Importance score
1	Monthly income	0.28
2	Age	0.22
3	Number of dependents	0.17
4	Occupation	0.10
5	Education	0.08
6	Industry	0.07
7	Domicile	0.05
8	Transportation expenditure	0.03

As shown in Table 6 and visually illustrated in Figure 5, the importance scores vary across predictors, with monthly income, age, and number of dependents making the largest relative contributions to the Random Forest classification.

4.6. Robustness checks

Robustness tests are conducted using alternative specifications. Table 7 summarizes the results across different model conditions.

Table 7. Robustness checks under alternative specifications

Check	Description	Main result	Consistency with baseline
Alternative split	Different random train–test split	Random Forest highest accuracy	Yes
Reduced model	Core predictors only	Lower accuracy than full model	Yes
Parameter variation	Alternative RF hyperparameters	Minor metric variation	Yes

Across all specifications, model ranking and performance patterns remain unchanged, with Random Forest consistently achieving the highest classification performance.

5. Discussion## Answer to the primary research question

The results establish that Random Forest achieves higher classification performance than logistic regression and decision tree models in classifying health insurance premium affordability categories using the cross-sectional socioeconomic dataset analysed in this study. This performance advantage is consistent across all evaluation metrics and remains stable under cross-validation, indicating that the relative superiority of Random Forest appears to be associated with the characteristics of the analysed dataset rather than driven by a specific sample partition [37]. The primary research question is therefore resolved in favour of the Random Forest model as the most effective classifier among the evaluated approaches.

5.1. Interpretation of model performance under nonlinear structure

The observed performance difference reflects the ability of Random Forest to represent nonlinear relationships and interaction effects among predictors that define affordability classification. Logistic regression imposes linear and additive constraints, which restrict its capacity to model interaction-based structures unless explicitly specified.

The decision tree model relaxes these assumptions but remains sensitive to sample-specific partitioning, resulting in higher variance and less stable generalization. Random Forest mitigates these limitations through ensemble aggregation, which reduces variance while allowing multiple interaction patterns to be captured simultaneously [38]. The improvement in AUC indicates that the model not only increases classification accuracy but also enhances the consistency of category separation across probability thresholds. This indicates that affordability classification in the observed data is governed by interaction-sensitive structures that are not adequately represented by linear or single-tree models. A plausible example involves the interaction between income and dependents. In a linear specification, the effect of income is assumed to be constant across households. In contrast, Random Forest can classify respondents differently when similar income levels are combined with different numbers of dependents. This flexibility allows affordability thresholds to vary across household contexts and may explain part of the observed performance advantage of the ensemble model.

3. ***Interpretation of key predictors within the classification framework***

5.3.1. Income as the primary financial capacity constraint. The prominence of income confirms that affordability classification is fundamentally anchored in available economic resources. Income directly determines the feasible allocation of resources to recurring premium payments, making it the principal variable defining classification boundaries [39]. The model's reliance on income indicates that differences in financial capacity dominate the segmentation process, even when additional socioeconomic variables are included.

5.3.2. Age as a life-cycle adjustment mechanism. Age contributes to classification through its association with life-cycle variation in financial behavior and resource allocation. The importance of age suggests that affordability is influenced by temporal positioning within the income expenditure trajectory, where financial obligations and consumption priorities evolve across stages of life [40]. This indicates that affordability classification reflects not only static income levels, but also dynamic allocation patterns linked to demographic structure.

5.3.3. Number of dependents as a household-level constraint. The number of dependents represents household exposure by capturing the distribution of financial resources across multiple individuals. Its influence indicates that affordability classification incorporates intra-household constraints, where increasing dependency ratios reduce the margin available for premium expenditure. This demonstrates that classification is jointly determined by individual financial capacity and household composition, rather than by isolated variables.

5.2. Interpretation of classification error structure

The concentration of misclassification between adjacent affordability categories indicates that classification boundaries are narrow and locally overlapping. Individuals positioned near category thresholds exhibit similar socioeconomic profiles, resulting in higher sensitivity to small variations in predictors. This pattern reflects a continuous underlying affordability gradient rather than sharply separated groups. The lower incidence of misclassification between non-adjacent categories indicates stronger separation at the extremes, where differences in economic resources and household obligations are more pronounced [41]. This structure supports the interpretation that affordability categories are ordinal in nature, even when modeled as discrete classes, and that classification uncertainty is concentrated at intermediate thresholds. The present findings are broadly consistent with previous studies reporting stronger predictive performance of Random Forest in insurance-related classification tasks. For example, studies examining insurance uptake and medical expenditure prediction have similarly found that ensemble methods outperform conventional statistical models when predictor relationships are heterogeneous and potentially nonlinear. The current study extends this evidence to affordability classification in an emerging-market setting using cross-sectional survey data.

5.3. Positioning within existing literature and contribution

The performance hierarchy observed in this study is broadly consistent with prior evidence. showing that ensemble learning methods outperform traditional econometric models in complex classification settings. Existing literature has demonstrated that Random Forest improves predictive accuracy by capturing nonlinear interactions and reducing model variance, particularly in heterogeneous datasets.

However, the present study extends this body of work in two specific ways. First, machine learning applies to affordability classification rather than to claims prediction or risk estimation, thereby addressing a forward-looking segmentation problem [42]. Second, it suggests that these performance advantages may persist in data-constrained environments where proxy variables replace detailed administrative records. This establishes that the relevance of machine learning approaches is not limited to data-rich contexts but extends to settings characterized by incomplete information. In terms of variable importance, the results reinforce established findings regarding the role of income and demographic factors while contributing a classification-based interpretation of how these variables jointly structure affordability segmentation [43]. This positions the study within both financial modeling and insurance analytics literature while addressing a previously underexplored application.

5.4. Theoretical implications for financial risk classification

The findings indicate that affordability-based risk classification is shaped by interaction-dependent relationships between financial capacity and household exposure. This challenges strictly additive representations of risk segmentation and supports the use of flexible modeling frameworks when analyzing proxy-based socioeconomic data. The study contributes to theory by demonstrating that machine learning models can be integrated into financial classification contexts without abandoning interpretability. Feature importance measures provide a direct link between predictive outcomes and economic constructs, allowing classification results to remain analytically meaningful within an insurance framework. This reinforces the feasibility of combining predictive modeling with conceptually grounded financial interpretation.

5.5. Practical implications for insurance and policy contexts

The results provide an operational basis for affordability classification using observable socioeconomic variables in contexts where direct underwriting data is unavailable. While Random Forest achieved the highest predictive performance among the evaluated models, the observed accuracy of 63.3% should be interpreted as moderate rather than high. Consequently, the model may serve as a useful decision-support tool but should not be viewed as a substitute for comprehensive underwriting or policymaking assessments. Classification models can support segmentation processes by identifying groups with distinct affordability profiles, thereby informing product design, targeting, or policy analysis.

For policymakers, the findings indicate that affordability assessment can be structured through data-driven classification rather than relying solely on predefined thresholds. This may improve alignment between policy interventions and actual household conditions. However, such applications should be interpreted cautiously because the classifications reflect statistical grouping rather than normative pricing rules.

5.6. Limitations of the study

The findings are subject to important limitations. First, the sample size of 148 observations is relatively small for machine-learning applications and may limit model stability despite the use of cross-validation and robustness checks. Second, convenience sampling and online survey recruitment introduce selection bias toward younger and digitally connected respondents. Third, the dependent variable measures perceived affordability rather than observed payment behavior, meaning that the results reflect affordability perceptions and preferences rather than objective affordability. Fourth, potentially relevant variables such as household savings, debt obligations, and existing insurance coverage were unavailable, which may introduce unobserved heterogeneity into the classification process.

5.7. Directions for future research

Future research may extend this analysis by using larger and more representative samples to improve generalizability. The integration of administrative or longitudinal data would allow examination of dynamic affordability patterns and model stability over time.

Further work may also compare alternative machine learning approaches to assess robustness across model classes. In addition, the incorporation of fairness-aware modeling frameworks would enable analysis of distributional effects in affordability classification, particularly in relation to socioeconomic inequality.

10. ***Transition to conclusion***

Taken together, the findings demonstrate that affordability classification is governed by interaction-based socioeconomic relationships that are more effectively captured by ensemble learning approaches, while remaining interpretable within a financial risk framework. The following section synthesizes these insights into a final statement of contribution and provides a proportionate conclusion consistent with the study's evidential scope.

6. Conclusion

Health insurance premium affordability classification represents a critical financial risk segmentation problem, particularly in emerging market contexts such as Indonesia, where incomplete information and reliance on observable socioeconomic indicators complicate classification processes. Under these conditions, insurers and policymakers must depend on proxy variables to group individuals according to their capacity to sustain premium payments, making the accuracy and structure of classification models a central concern for financial decision-making. This study finds that Random Forest achieved the highest classification performance among the evaluated models within the analyzed sample. However, given the limited sample size and exploratory design, the observed performance differences should be interpreted as preliminary evidence rather than definitive confirmation of model superiority. The results further indicate that monthly income, age, and number of dependents are the most influential predictors, reflecting the combined influence of economic resources and household obligations in shaping affordability categories.

The study demonstrates by demonstrating that ensemble classification methods can improve affordability-based financial risk grouping while remaining interpretable through structured variable importance analysis. By applying Random Forest within a classification framework grounded in financial capacity and household exposure, the study provides evidence that nonlinear and interaction-sensitive modeling approaches are better suited to affordability segmentation under data-constrained conditions. This contribution is methodological and empirical, showing how machine learning can be integrated into financial classification problems without exceeding the limits of cross-sectional evidence.

Overall, the study suggests that ensemble machine-learning methods may provide a useful exploratory approach to affordability segmentation when only proxy-based socioeconomic data are available. Additional validation using larger, more representative, and longitudinal datasets is required before stronger conclusions can be drawn regarding broader applicability.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors. The authors would like to thank all respondents who participated in the survey and contributed to the data collection process.

REFERENCES

1. M. Geruso, and T. J. Layton, *Selection in health insurance markets and its policy remedies*, The Journal of Economic Perspectives, vol. 31, no. 4, pp. 23–50, 2017. https://doi.org/10.1257/jep.31.4.23
2. M. Fujii, H. Sakaji, S. Masuyama, and H. Sasaki, *Extraction and classification of risk-related sentences from securities reports*, International Journal of Information Management Data Insights, vol. 2, no. 2, pp. 1–13, 2022. https://doi.org/10.1016/j.jjimei.2022.100096

3. U. Temursho, M. Weitzel, and R. Garaffa, *Projection of household-level consumption expenditures in a macro-micro consistent framework*, Structural Change and Economic Dynamics, vol. 73, pp. 112–135, 2025. https://doi.org/10.1016/j.strueco.2024.12.015
4. V. Venugopal, A. Dongre, and P. Ponnusamy, *Constructing practical and realistic asset-based socioeconomic status assessment scale using principal component analysis for urban population of Puducherry, India*, Indian Journal of Community Medicine, vol. 46, no. 3, pp. 1–5, 2021. https://doi.org/10.4103/ijcm.ijcm_925_20
5. P. Panjee, and S. Amornsawadwatana, *A generalized linear model and machine learning approach for predicting the frequency and severity of cargo insurance in Thailand's border trade context*, Risks, vol. 12, no. 2, pp. 1–33, 2024. https://doi.org/10.3390/risks12020025
6. K. I. Jones, and S. Sah, *The Implementation of Machine Learning in The Insurance Industry with Big Data Analytics*, International Journal of Data Informatics and Intelligent Computing, vol. 2, no. 2, pp. 21–38, 2023. https://doi.org/10.59461/ijdiic.v2i2.47
7. P. R. A. Puchakayala, *Data quality management for effective machine learning and AI modelling, best practices and emerging trends*, International Research Journal of Innovations in Engineering and Technology, vol. 06, no. 12, pp. 327–340, 2022. https://doi.org/10.47001/irjiet/2022.612062
8. N. K. K. Yego, J. Nkurunziza, and J. Kasozi, *Predicting health insurance uptake in Kenya using Random Forest: An analysis of socio-economic and demographic factors*, PLoS ONE, vol. 18, no. 11, pp. 1–19, 2023. https://doi.org/10.1371/journal.pone.0294166
9. U. Orji, and E. Ukwandu, *Machine learning for an explainable cost prediction of medical insurance*, Machine Learning With Applications, vol. 15, pp. 1–20, 2023. https://doi.org/10.1016/j.mlwa.2023.100516
10. R. Gonzalez, A. Saha, C. J. V. Campbell, P. Nejat, C. Lokker, and A. P. Norgan, *Seeing the random forest through the decision trees. Supporting learning health systems from histopathology with machine learning models: Challenges and opportunities*, Journal of Pathology Informatics, vol. 15, pp. 1–8, 2024. https://doi.org/10.1016/j.jpi.2023.100347
11. C. S. A. Yong, P. Symonds, G. Petrou, and Z. Chalabi, *Assessing the association between socio-demographic factors and home energy efficiency in England and Wales: A machine learning approach*, Energy and Buildings, vol. 353, pp. 1–25, 2025. https://doi.org/10.1016/j.enbuild.2025.116832
12. A. Dzelihodzic and D. Donko, *Comparison of ensemble classification techniques and single classifiers performance for customer credit assessment*, Modeling of Artificial Intelligence, vol. 11, no. 3, pp. 140–150, 2016. https://doi.org/10.13187/mai.2016.11.140
13. S. Weinzierl, S. Zilker, S. Dunzer, and M. Matzner, *Machine learning in business process management: A systematic literature review*, Expert Systems with Applications, vol. 253, pp. 1–43, 2024. https://doi.org/10.1016/j.eswa.2024.124181
14. H. Chenet, J. Ryan-Collins, and F. Van Lerven, *Finance, climate-change and radical uncertainty: Towards a precautionary approach to financial policy*, Ecological Economics, vol. 183, pp. 1–14, 2021. https://doi.org/10.1016/j.ecolecon.2021.106957
15. R. De Costa, I. Ferrara, M. Toplak, A. Alam, R. Bowie, and A. Burnett, *Behavioural insights and environmental sustainability: Key findings and policy implications from a systematic review*, Journal of Environmental Management, vol. 390, pp. 1–19, 2025. https://doi.org/10.1016/j.jenvman.2025.126118
16. C. Ling, Z. Chen, J. Yang, and T. Yang, *Evaluating inclusive transit-oriented development with location affordability and its influencing factors*, Transportation Research Part A: Policy and Practice, vol. 201, pp. 1–24, 2025. https://doi.org/10.1016/j.tra.2025.104672
17. A. M. Dg. M. Nasir, S. N. Shamsuddin, and N. Ismail, *Predicting Life Insurance Ownership: the Role of Socioeconomic Factors*, Journal of Information System and Technology Management, vol. 10, no. 38, pp. 252–270, 2025. https://doi.org/10.35631/jistm.1038017
18. B. Bergougui, O. Ben-Salha, and S. M. Murshed, *How income inequality drives climate inequality in the world's largest emitting nations: Novel evidence from multivariate quantile-on-quantile and wavelet quantile methods*, Journal of Cleaner Production, vol. 513, pp. 1–25, 2025. https://doi.org/10.1016/j.jclepro.2025.145666
19. M. Aghaabbasi, and S. Chalermpong, *Machine learning techniques for evaluating the nonlinear link between built-environment characteristics and travel behaviors: A systematic review*, Travel Behaviour and Society, vol. 33, p. 100640, 2023. https://doi.org/10.1016/j.tbs.2023.100640
20. B. Barabino, and R. Ventura, *Comparing fare evasion severity by econometric and artificial intelligence models: An Italian case study*, Transport Policy, vol. 178, pp. 1–17, 2025. https://doi.org/10.1016/j.tranpol.2025.103971
21. W. R. Mgomezulu, P. Thangata, B. Mkandawire, and N. Amoah, *Advancing predictive analytics in child malnutrition: Machine, ensemble and deep learning models with balanced class distribution for early detection of stunting and wasting*, Human Nutrition & Metabolism, vol. 42, pp. 1–18, 2025. https://doi.org/10.1016/j.hnm.2025.200340
22. Y. Syahra, Y. F. Br. Tarigan, K. Andriani, H. W. N. S, and R. Setik, *Decision Trees in Predicting Loan Default Risk in Customer Relationships within the Financial Sector*, SinkrOn, vol. 9, no. 2, pp. 734–745, 2025. https://doi.org/10.33395/sinkron.v9i2.14672
23. I. Ismail, P. J. A. Stam, F. R. M. Portrait, A. Van Witteloostuijn, and X. Koolman, *Addressing unanticipated interactions in risk equalization: A machine learning approach to modeling medical expenditure risk*, Economic Modelling, vol. 130, pp. 1–10, 2023. https://doi.org/10.1016/j.econmod.2023.106564
24. Z. M. Omer, and H. Shareef, *Comparison of decision tree based ensemble methods for prediction of photovoltaic maximum current*, Energy Conversion and Management X, vol. 16, pp. 1–17, 2022. https://doi.org/10.1016/j.ecmx.2022.100333
25. X. Yuan, S. Liu, W. Feng, and G. Dauphin, *Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm*, Remote Sensing, vol. 15, no. 21, pp. 1–20, 2023. https://doi.org/10.3390/rs15215203
26. J. Wang, Z. Qin, J. Hsu, and B. Zhou, *A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure*, Healthcare Analytics, vol. 5, pp. 1–10, 2024. https://doi.org/10.1016/j.health.2024.100312

27. C. Rodrigues, M. Veloso, A. Alves, and C. L. Bento, *Socioeconomic and functional zoning characterization in a city: a clustering approach*, Cities, vol. 163, pp. 1–14, 2025. https://doi.org/10.1016/j.cities.2025.106023
28. E. Vecchi, G. Berra, S. Albrecht, P. Gagliardini, and I. Horenko, *Entropic approximate learning for financial decision-making in the small data regime*, Research in International Business and Finance, vol. 65, pp. 1–31, 2023. https://doi.org/10.1016/j.ribaf.2023.101958
29. F. B. Mushonga, and S. Mishi, *The Impact of Insurtech on insurance Inclusion: A Systematic Literature review*, Journal of Risk and Financial Management, vol. 19, no. 2, pp. 1–20, 2026. https://doi.org/10.3390/jrfm19020122
30. S. Shouri, M. De La Sen, and M. E. Gordji, *Designing a smart health insurance pricing system: integrating XGBOOST and repeated NASH equilibrium in a sustainable, Data-Driven framework*, Information, vol. 16, no. 9, pp. 1–18, 2025. https://doi.org/10.3390/info16090733
31. S. Syed, A. Acquaye, M. M. Khalfan, T. Obuobisa-Darko, and F. A. Yamoah, *Decoding sustainable consumption behavior: A systematic review of theories and models and provision of a guidance framework*, Resources, Conservation & Recycling Advances, vol. 23, pp. 1–26, 2024. https://doi.org/10.1016/j.rcradv.2024.200232
32. C. K. Chikeya, and L. Ntsalaze, *Determinants of Household Debt: A Systematic Review of the literature*, Economies, vol. 13, no. 3, pp. 1–76, 2025. https://doi.org/10.3390/economies13030076
33. N. N. Ridzuan, M. Masri, M. Anshari, N. L. Fitriyani, and M. Syafrudin, *AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility*, Information, vol. 15, no. 8, pp. 1–30, 2024. https://doi.org/10.3390/info15080432
34. X. Chen, G. Hu, and H. Wen, *Investigating the factors influencing household financial vulnerability in China: An exploration based on the Shapley Additive Explanations approach*, Sustainability, vol. 17, no. 12, pp. 1–28, 2025. https://doi.org/10.3390/su17125523
35. N. M. Latha, K. Geetha, and S. Damodharan, *Overview of regression models and how to determine the best model for data*, Journal of Scientific Research and Reports, vol. 30, no. 10, pp. 250–266, 2024. https://doi.org/10.9734/jsrr/2024/v30i102452
36. P. Kumar, R. Pillai, N. Kumar, and M. I. Tabash, *The interplay of skills, digital financial literacy, capability, and autonomy in financial decision making and well-being*, Borsa Istanbul Review, vol. 23, no. 1, pp. 169–183, 2022. https://doi.org/10.1016/j.bir.2022.09.012
37. J. M. Baladjay, N. Riva, L. A. Santos, D. M. Cortez, C. Centeno, and A. A. R. Sison, *Performance evaluation of random forest algorithm for automating classification of mathematics question items*, World Journal of Advanced Research and Reviews, vol. 18, no. 2, pp. 034–043, 2023. https://doi.org/10.30574/wjarr.2023.18.2.0762
38. Y. M. Abbas, A. A. Abadel, and Y. R. Alharbi, *AI-driven prediction of carbonation depth in recycled aggregate concrete: Influential parameters and nonlinear interactions from machine learning perspectives*, Case Studies in Construction Materials, vol. 23, pp. 1–25, 2025. https://doi.org/10.1016/j.cscm.2025.e05251
39. L. M. Hüsken, J. H. Slinger, H. S. I. Vreugdenhil, and M. A. Altamirano, *Money talks. A systems perspective on funding and financing barriers to nature-based solutions*, Nature-Based Solutions, vol. 6, pp. 1–18, 2024. https://doi.org/10.1016/j.nbsj.2024.100200
40. M. D. Walugembe, O. K. Osunsan, D. Kisuule, and N. Kiwabudde, *Age-Related Effects on Financial Literacy, Personal Financial Management, and Utilization of Financial Products Among Y-Save Multi-Purpose Cooperative Members in Uganda*, International Journal of Management Studies and Social Science Research, vol. 07, no. 02, pp. 104–119, 2025. https://doi.org/10.56293/ijmsssr.2025.5515
41. M. Maghsoudi, N. Mohammadi, and M. Bakhtiari, *Artificial intelligence and sustainable development: Public concerns and governance in developed and developing nations*, Cleaner Environmental Systems, vol. 19, pp. 1–20, 2025. https://doi.org/10.1016/j.cesys.2025.100340
42. D. Pattnaik, S. Ray, and R. Raman, *Applications of artificial intelligence and machine learning in the financial services industry: A bibliometric review*, Heliyon, vol. 10, no. 1, pp. 1–19, 2024. https://doi.org/10.1016/j.heliyon.2023.e23492
43. I. Katnic, M. Katnic, M. Orlandic, M. Radunovic, and I. Mugosa, *Understanding the Role of Financial Literacy in Enhancing economic stability and resilience in Montenegro: A Data-Driven Approach*, Sustainability, vol. 16, no. 24, pp. 1–25, 2024. https://doi.org/10.3390/su162411065