Predicting Health Insurance Premiums in Indonesia Using Random Forest: Balancing Accuracy and Interpretability

Sylvia Samuel[1], Hendra Achmadi[2*], Tiurida Lily Anita[3]

[1][2*]Faculty of Business and Economics, Department of Management, Universitas Pelita Harapan, Indonesia

[3]Faculty of Digital Communication and Hotel and Tourism Bina Nusantara University Jakarta, Indonesia

ABSTRACT

This study aims to predict health insurance premiums in Indonesia using the Random Forest algorithm, comparing its performance to traditional logistic regression. A quantitative approach was employed with data from 148 respondents, capturing demographic, socioeconomic, and lifestyle factors. The Random Forest model achieved an accuracy of 63.3%, outperforming logistic regression (55%). Key predictors identified were income, age, and the number of dependents. The study highlights the method's ability to balance predictive accuracy with interpretability, which is crucial in highly regulated industries. The findings offer insurers valuable insights for pricing strategies, customer segmentation, and regulatory compliance by providing transparent, evidence-based approaches to premium determination. The originality of this research lies in its application of machine learning, specifically Random Forest, to health insurance premium prediction in an emerging market, where traditional models often fail to capture complex interactions between variables. This study contributes to both academic literature and industry practice, offering a methodological framework that enhances transparency and fairness in health insurance pricing. By demonstrating how machine learning can support decision-making in a regulatory context, this research provides a novel, data-driven approach to health insurance premium determination in Indonesia.

Keywords: Customer Characteristics; Health Insurance Premiums; Indonesia; Machine Learning; Random Forest

Introduction

Health insurance plays a pivotal role in mitigating financial risks associated with healthcare costs, offering both individuals and households a mechanism to protect themselves from catastrophic expenses. In many emerging economies, including Indonesia, the demand for health insurance has grown significantly over the last two decades, driven by rising incomes, urbanization, and growing awareness of health risks (Oktafia & Taufiq, 2021). However, determining premiums in such heterogeneous environments remains a formidable challenge. Insurers must balance affordability for customers with sustainability for firms, particularly in contexts where regulatory oversight is tightening and consumer trust is fragile. The accurate and transparent determination of premiums is therefore not only a technical task but also a social and managerial priority.

Traditionally, health insurance premium determination has been grounded in actuarial-statistical models, such as generalized linear models (GLMs) and credibility theory. These approaches offer transparency and regulatory acceptability, but are limited in their ability to

capture nonlinear and interactive effects among variables (Charbuty & Abdulazeez, 2021). For example, age, income, and number of dependents may interact in ways that a linear model cannot adequately represent. These limitations motivate the use of machine learning methods that can capture complex, multidimensional patterns.

In recent years, machine learning has been applied in various domains, including marketing, risk management, and healthcare analytics, demonstrating strong potential to enhance decision-making under uncertainty (Duarte et al., 2022; Müller & Guido, 2017). In the insurance sector, ensemble methods such as Random Forest have attracted attention for their ability to enhance predictive performance while reducing the risk of overfitting. Unlike single decision trees, which are prone to variance and instability, Random Forest aggregates multiple trees trained on different subsets of data, producing more robust and generalizable predictions (Breiman, 2016). Importantly, Random Forest also provides estimates of feature importance, which enhances interpretability a critical consideration in regulated industries where decisions must be justified to both regulators and customers.

The relevance of this research is heightened in Indonesia, where the health insurance industry faces unique challenges. First, the market is characterized by significant socioeconomic diversity, requiring premiums to be set for individuals across different income levels, occupations, and regions. Second, regulatory frameworks emphasize fairness and transparency, requiring insurers to demonstrate that premium structures are not arbitrary. Third, consumer awareness of insurance products is growing, and customers increasingly demand value and clarity in the services they purchase. Against this backdrop, a modelling approach that combines predictive accuracy with interpretability can provide a competitive advantage for insurers while addressing regulatory and societal concerns.

Previous studies have contributed to understanding consumer needs and decision-making in insurance contexts. Oktafia & Taufiq, (2021) examined the impact of digital transformation on Sharia insurance in Indonesia, highlighting the strategic importance of adopting data-driven tools. Puntoni et al., (2021) explored the role of artificial intelligence in consumer decision-making, emphasizing the experiential dimension of digital interactions. Pradana & Ha, (2021) applied clustering methods to customer segmentation in retail contexts, while D'Souza & Gurin, (2016) revisited Maslow's hierarchy of needs to explain the prioritization of safety-related products such as insurance. Collectively, these studies underscore the value of data analytics and consumer behavior research but leave a gap in applying advanced predictive models directly to the problem of premium determination.

This research addresses that gap by applying the Random Forest algorithm to predict health insurance premiums in Indonesia. Specifically, the study examines which customer characteristics, including age, income, dependents, occupation, and education, have the most significant influence on premium categories. While earlier research has focused on segmentation or descriptive analysis, this study advances the literature by offering a predictive modelling framework that is both accurate and interpretable. The inclusion of feature importance analysis enables the study to move beyond prediction to explanation, highlighting not only how premiums can be predicted but also why certain variables are more significant than others.

The importance of this study can be considered from both theoretical and managerial perspectives. Theoretically, it contributes to the modelling literature by situating Random

Forest within the broader discourse of decision sciences and risk theory. By linking predictive analytics with concepts such as adverse selection, risk pooling, and decision-making under uncertainty (Akerlof, 1970; Arrow, 1963; Kahneman et al., 1979). The study demonstrates how machine learning methods can inform longstanding theoretical debates. Managerially, the research equips insurers with evidence-based insights for pricing strategy, customer segmentation, and compliance with fairness regulations. Random Forest is thus positioned not only as a technical solution but as a decision-support system that enhances transparency and accountability.

Despite its potential, the application of machine learning in determining insurance premiums remains underexplored, particularly in emerging markets. Most prior applications of machine learning in insurance have been concentrated in developed economies or in areas such as fraud detection and claims management, rather than premium pricing (Gkikas et al., 2022). Moreover, while advanced models such as gradient boosting and neural networks often achieve high accuracy, their lack of interpretability makes them less attractive in regulated industries where transparency is essential. Random Forest, by balancing accuracy and interpretability, provides a promising alternative that has yet to be systematically evaluated in the Indonesian context. Against this background, this study seeks to answer the following research questions:

1. To what extent can Random Forest predict health insurance premiums in Indonesia with greater accuracy than traditional statistical models?
2. Which customer characteristics are most influential in determining health insurance premiums, and how can these insights inform both theory and practice?
3. How can predictive modelling balance the dual demands of accuracy and interpretability in the context of regulated industries such as insurance?

By addressing these questions, the study helps to close an important gap in the literature and provides practical guidance for industry stakeholders. The remainder of the paper is structured as follows. The next section reviews the relevant literature on determining health insurance premiums, customer characteristics, and machine learning applications in management. The methodology section then explains the data collection, pre-processing, and modelling procedures. Results are presented and discussed in relation to both prior research and theoretical frameworks. The paper concludes with a summary of contributions, limitations, and directions for future research.

Literature Review

Health Insurance Premiums and Customer Characteristics

Health insurance premiums are periodic payments that policyholders are required to make to maintain coverage. They reflect both the risk profile of the insured and the sustainability needs of insurers, creating a delicate balance between affordability and financial soundness (Oktafia & Taufiq, 2021). In emerging markets like Indonesia, premium determination is particularly challenging due to income disparities, diverse demographic structures, and varying access to healthcare services.

Key predictors of premiums often include age, income, dependents, occupation, industry, domicile, and education. Age is a widely acknowledged determinant, as health

insurance premium risks tend to increase over time, leading to higher expected costs for insurers (Maccheroni et al., 2025). Income not only signals purchasing power but also correlates with access to better healthcare and lifestyle choices, which can influence health insurance outcomes. Dependents increase expected claims because family coverage broadens risk exposure. Occupation and industry matter because they shape both income stability and exposure to occupational hazards. Education levels affect health literacy, risk perceptions, and the likelihood of engaging in preventive healthcare. Domicile captures regional differences in access to healthcare infrastructure and costs.

In Indonesia, where direct measures of lifestyle factors may be difficult to obtain, proxies such as household electricity usage, transport expenditure, and ownership of durable goods have been used to approximate socioeconomic status (PERSOLKELLY_Indonesia, 2023). These proxies are valuable for segmentation and affordability modelling, even if they introduce challenges of standardization and causal interpretation. From a behavioral perspective, Maslow's hierarchy of needs situates health protection at the safety level, demonstrating how income and dependents can shape insurance purchasing priorities (D'Souza & Gurin, 2016).

Traditional Approaches to Health Insurance Premium Prediction

Historically, premium determination has been dominated by actuarial and statistical models, particularly generalized linear models (GLMs), survival models, and credibility theory. These models offer interpretability, transparency, and regulatory acceptability. They enable insurers to explain premium structures to regulators and customers, which is vital in industries where fairness is closely monitored (Charbuty & Abdulazeez, 2021).

However, such models rely on linear assumptions and are limited in handling nonlinear relationships or high-dimensional data. For example, the interaction between age and dependents may not be additive but multiplicative, with nonlinear effects on risk. Similarly, the influence of education may vary depending on income and occupation, introducing complexities that linear models struggle to address. These shortcomings often lead to oversimplified predictions, which can potentially result in mispricing, adverse selection, and inefficient risk pooling (Akerlof, 1970).

Machine Learning Approaches

In response to these limitations, researchers have increasingly turned to machine learning. Machine learning methods excel at capturing complex relationships, handling large datasets, and adapting to multidimensional variables (Müller & Guido, 2017). Within machine learning, tree-based models have emerged as particularly useful for classification and prediction tasks in finance, marketing, and healthcare.

Random Forest is an ensemble method that constructs multiple decision trees using bootstrapped samples and random subsets of predictors. By aggregating predictions, it reduces overfitting and improves generalization (Breiman, 2016). Importantly, Random Forest provides feature importance rankings, which enhance the interpretability of the model. This is particularly valuable in regulated industries where decision justification is essential.

Other machine learning approaches, such as gradient boosting machines (GBM), extreme gradient boosting (XGBoost), and deep neural networks, often achieve even higher accuracy but are criticized for their opacity. In contrast, Random Forest strikes a balance between accuracy and interpretability, making it more practical for insurers who must justify their pricing decisions to regulators and customers. Support vector machines (SVM) and logistic regression hybrids have also been explored, but their adoption remains limited due to interpretability concerns (R, 2015).

The debate between accuracy and interpretability has given rise to explainable AI (XAI), which seeks to make machine learning models more transparent. Approaches such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are increasingly applied in financial services. However, in insurance, regulatory requirements often prioritize models that are inherently interpretable, which strengthens the case for Random Forest over black-box alternatives.

Prior Empirical Studies

Machine learning has been applied in insurance primarily for fraud detection, claims management, and predicting customer churn. For example, Gkikas et al., (2022) showed how decision trees combined with genetic algorithms improved consumer behavior classification. Duarte et al., (2022) reviewed applications of machine learning in marketing decision support, noting its value in segmentation and targeting. Puntoni et al., (2021) discussed AI's role in consumer decision-making, emphasizing experiential impacts.

In emerging markets, Oktafia & Taufiq, (2021) examined digital transformation in Sharia insurance, highlighting the need for data-driven strategies. Pradana and Ha (2021) applied K-means clustering to retail customers, demonstrating the utility of data mining, but did not extend it to insurance premiums. Most of these studies either focus on developed economies or use ML for purposes other than health insurance premium prediction. This underscores the scarcity of research applying ML to pricing strategies in health insurance premium in emerging economies.

Gaps in the Literature

Despite the promising potential of machine learning in insurance, several important gaps remain. Much of the existing research is concentrated in developed markets, limiting contextual relevance for emerging economies such as Indonesia, where socioeconomic diversity introduces unique challenges for premium modelling. Moreover, while machine learning has been widely applied in areas such as fraud detection and claims management, its use in premium determination remains limited. Even when advanced algorithms such as gradient boosting or neural networks are applied, their lack of interpretability reduces their practical value in regulated contexts.

This creates an interpretability gap that Random Forest, with its balance of accuracy and explain ability, is well positioned to fill. Additionally, issues of fairness and bias remain underexplored; there is a risk that reliance on proxy variables such as electricity usage or household expenditure could unintentionally reinforce socioeconomic inequalities.

Benchmarking also remains a weakness, as few studies systematically compare Random Forest against both traditional actuarial models and alternative ML methods. Finally, most studies rely on cross-sectional data, leaving a temporal gap in understanding how premiums evolve across the customer lifecycle. These gaps highlight the need for studies that not only test machine learning in health insurance premium prediction, but also emphasize fairness, transparency, and contextual relevance.

Synthesis

Taken together, the literature reveals a trade-off between the transparency of actuarial models and the predictive power of advanced machine learning approaches. Random Forest stands out as a method that balances these demands, offering predictive accuracy along with interpretability through feature importance analysis. However, empirical applications in emerging market health insurance remain scarce, and the integration of fairness and decision-making theories into machine learning based health insurance premium prediction is underdeveloped.

This study addresses these gaps by applying Random Forest to health insurance premium prediction in Indonesia, benchmarking its performance against logistic regression, and interpreting feature importance to identify key customer characteristics. In doing so, it contributes to both theory and practice: theoretically by situating machine learning within decision and risk theory, and managerially by equipping insurers with evidence-based tools for transparent and accurate pricing strategies.

The conceptual of the study highlights the logical flow. Customer characteristics, including demographic (age, dependents), socioeconomic (income, occupation, education), and lifestyle-related proxies (electricity usage, transport expenditure), are treated as independent variables as shown in Figure 1.

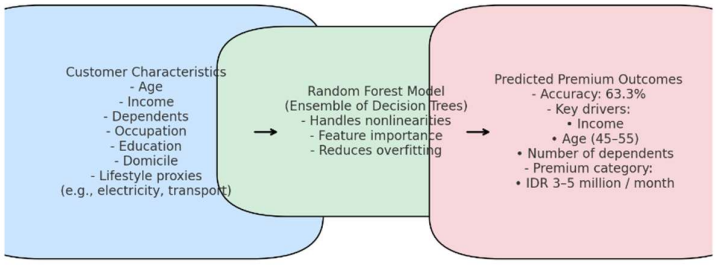

Figure 1. Framework of Health Insurance Premium Prediction

Methodology

This study employs a quantitative research design, utilizing supervised machine learning to predict health insurance premiums based on customer characteristics. The approach integrates predictive modelling with feature importance analysis to balance accuracy and interpretability. Random Forest was chosen as the primary modeling technique due to its robustness against overfitting, ability to handle nonlinear interactions, and provision of variable importance rankings. To ensure methodological rigor, the study also included robustness

checks against baseline models and applied multiple evaluation metrics beyond overall accuracy.

Data Collection

Data were collected through a structured questionnaire distributed online via Google Forms between September 2024 and March 2025. Respondents were prospective health insurance customers across Indonesia, recruited through university networks, professional associations, and social media platforms. Participation was voluntary, and respondents provided informed consent prior to completing the survey. A total of 202 responses were received. After screening for completeness and consistency, 148 valid cases were retained for analysis. While the dataset used in this study includes 148 valid responses, it should be noted that the relatively small sample size may limit the generalizability of the findings to a broader population. To mitigate this, data augmentation techniques or future studies with larger datasets are recommended.

Instrument and Variables

The questionnaire captured demographic, socioeconomic, and lifestyle-related information commonly used in insurance underwriting. Independent variables included age, gender, marital status, number of dependents, education, occupation, industry, domicile, monthly income, electricity usage capacity, transportation expenditure, and household assets. These variables were chosen because they represent both direct determinants of risk and proxies for socioeconomic status. The dependent variable was the health insurance premium category, measured as the self-reported ability and willingness to pay for monthly premiums, grouped into ranges with a focus on the IDR three to five million bracket, which is a common premium tier in Indonesia.

Data Pre-processing

Pre-processing followed standard steps in supervised machine learning (Tchagna Kouanou et al., 2021). First, incomplete responses were excluded, resulting in 148 complete cases. Second, categorical variables (e.g., occupation, domicile, education) were converted into numerical form through label encoding. Third, continuous variables (e.g., income, age, expenditure) were standardized to reduce scale bias. Finally, the dataset was randomly partitioned into training (70%) and testing (30%) subsets, ensuring that the distribution of premium categories was preserved across sets.

Modelling Approach

The Random Forest algorithm was implemented in Python using the scikit-learn library. The algorithm constructs an ensemble of decision trees by drawing bootstrapped samples from the training set and selecting random subsets of features at each split. Predictions are aggregated across trees through majority voting, which reduces variance and improves

generalization (Breiman, 2016). Key hyperparameters tuned in this study included the number of trees, the maximum depth of the trees, and the minimum number of samples per leaf. Tuning was performed using a grid search approach combined with cross-validation.

Model performance was evaluated using multiple complementary metrics. Accuracy, defined as the proportion of correctly classified observations, provided a straightforward measure of overall performance. However, because class imbalances in premium categories may lead to misleadingly high accuracy, additional metrics were employed. Precision measured the proportion of health insurance premium predictions that were correct, thus reflecting the reliability of positive classifications. Recall, or sensitivity, assessed the model's ability to correctly identify true cases within each premium class, thereby capturing its effectiveness in minimizing false negatives. To balance precision and recall, the F1-score was calculated as the harmonic mean of the two, offering a more holistic measure of classification quality. Finally, the area under the receiver operating characteristic curve (AUC) was examined to assess the model's ability to discriminate between classes across different thresholds. Using this portfolio of evaluation metrics ensured that the assessment of Random Forest's performance was not dependent on a single indicator but reflected a multidimensional view of predictive quality.

To strengthen confidence in the results, robustness checks were conducted. A ten-fold cross-validation approach was employed, in which the training dataset was partitioned into ten folds, with nine folds used for training and one for validation iteratively. This procedure was repeated across all folds, and performance metrics were averaged to mitigate the risk of sampling bias. The comparative evaluation of Random Forest against logistic regression and a single decision tree further allowed the study to demonstrate whether improvements in predictive performance were consistent and statistically meaningful.

Additionally, confusion matrix analysis was conducted to identify misclassification patterns, with particular attention paid to whether errors occurred primarily between adjacent premium categories (e.g., IDR 2–3 million misclassified as IDR 3–5 million), which are more acceptable in practical decision-making than errors across distant categories. To evaluate the relative performance of Random Forest, two benchmark models were used. The first was a logistic regression model, representing the traditional statistical approach commonly applied in actuarial practice. The second was a single decision tree, providing a baseline machine learning model without ensembling. These comparisons highlight whether Random Forest offers improvements in both accuracy and stability.

Table 1. Comparison of Modelling Approaches and Robustness Checks

Model	Approach	Key Features	Validation	Accuracy (%)
Logistic Regression (Baseline)	Statistical (GLM-based)	Linear relationships, interpretable coefficients	70/30 train-test split	55.0
Random Forest (Proposed)	Ensemble ML (Decision Trees)	Captures nonlinearities, feature importance, and reduces overfitting	70/30 train-test split + 10-fold cross-validation	63.3

Table 1 presents the comparison between the two modelling approaches. Logistic regression, a GLM-based statistical model, provides interpretability but is limited in capturing

nonlinear relationships. The Random Forest model, in contrast, leverages an ensemble of decision trees to capture complex, multidimensional patterns and generate feature importance rankings that are meaningful for managerial decision-making. The validation strategies also differed: while logistic regression was validated using a simple 70/30 train-test split, Random Forest was evaluated through both a 70/30 split and ten-fold cross-validation to enhance robustness and reduce overfitting. The results indicate that Random Forest achieved higher predictive accuracy (63.3%) compared to logistic regression (55%). This demonstrates the superiority of ensemble methods in predicting health insurance premiums in the Indonesian context, thereby strengthening the study's methodological contribution.

Finally, ethical protocols were observed throughout the study. Participation in the survey was voluntary, with informed consent obtained from all respondents. No personally identifiable information was collected, and all responses were anonymized to ensure confidentiality.

Results and Discussion

Descriptive Statistics

Table 2 presents the demographic and socioeconomic characteristics of the respondents. Out of 148 valid responses, the majority (45.6%) were aged between 18 and 20 years, followed by 34.0% in the 23–26 range and 20.4% aged 21–23. Gender distribution was relatively balanced, with 51.1% female and 48.9% male. Education levels showed a split between secondary school graduates (48.9%) and undergraduates (51.1%), reflecting a young and educated population. The sample also included diverse income levels, occupations, and family statuses, underscoring the heterogeneity of the Indonesian health insurance premium market.

Table 2. Respondent Demographics

Variable	Category	Percentage (%)
Age	18–20	45.6
	21–23	20.4
	23–26	34.0
Gender	Female	51.1
	Male	48.9
Education	Secondary	48.9
	Undergraduate	51.1
Number of Dependents	None	37.2
	1–2	42.0
	3 or more	20.8
Income (Monthly)	< IDR 5 million	39.0
	5–10 million	44.6

	> 10 million	16.4
Premium Category	< IDR 3 million	36.5
	3–5 million	46.0
	> IDR 5 million	17.5

The descriptive analysis highlights a sample dominated by young, educated, and mid-income respondents, which aligns with the target demographic for private health insurance premium in Indonesia. The relatively high share of respondents willing to pay premiums in the IDR 3–5 million range (46%) suggests a growing middle-class segment with increasing awareness of healthcare financing.

Model Performance

The Random Forest model achieved an overall classification accuracy of 63.3% on the test set, outperforming logistic regression (55.0%) and a single decision tree (58.2%). While accuracy provides a general indication of performance, further evaluation using additional metrics is more informative. Table 3 presents the results across accuracy, precision, recall, F1-score, and AUC.

Table 3. Model Performance Comparison

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

The results show that Random Forest not only improves overall accuracy but also demonstrates stronger performance across all key metrics, particularly AUC (0.70), indicating good discrimination between premium categories. Logistic regression, though interpretable, struggled to capture nonlinearities, while the single decision tree overfitted, reducing generalizability.

The confusion matrix from Table 4 further illustrates classification outcomes. Random Forest correctly identified most respondents in the IDR 3–5 million premium category but showed some misclassification between adjacent categories (< IDR 3 million misclassified as 3–5 million, and > IDR 5 million misclassified as 3–5 million). Such errors are less concerning in practice, as they still reflect relatively close premium ranges.

Table 4. Confusion Matrix (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

The matrix highlights Random Forest's particular strength in identifying the middle category (3–5 million), which also represents the largest group in the dataset.

Feature Importance Analysis

Random Forest's feature importance analysis provides insights into which variables most strongly influenced health insurance premium prediction. Figure 1 ranks the top predictors.

Table 5. Feature Importance Rankings

Rank	Variable	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	Transport Expenditure	0.03

The three dominant predictors, such as income, age, and dependents, explain nearly 70% of the variance in premiums. This aligns with established insurance logic: income reflects affordability, age correlates with health risk, and dependents increase expected claims. Secondary predictors such as occupation and education refine segmentation but contribute less overall.

Discussion of Findings

The findings confirm the superiority of Random Forest over traditional statistical models in predicting health insurance premiums. The improvement in accuracy, precision, recall, and AUC demonstrates the value of ensemble methods in capturing nonlinear relationships across demographic and socioeconomic variables.

The dominance of income, age, and dependents is consistent with actuarial theory and prior research. Income directly determines affordability and willingness to pay (Maccheroni et al., 2025), while age reflects an increase in health risks over time (Einav et al., 2018).

Dependents increase risk exposure, leading to higher expected claims. Secondary predictors, such as occupation and education, reflect both health literacy and occupational hazards, aligning with earlier studies that link socioeconomic factors to insurance purchasing behavior (Pradana & Ha, 2021).

The confusion matrix revealed that Random Forest is particularly effective in predicting the majority premium category (IDR 3–5 million), but less accurate at the margins. This mirrors findings in other emerging market studies, where heterogeneity in high- and low-income groups creates prediction challenges (Oktafia & Taufiq, 2021).

From a theoretical perspective, these results extend decision-making models by demonstrating how predictive algorithms interact with concepts such as risk pooling (Arrow, 1963/updated with Einav et al., 2018) and adverse selection (Akerlof, 1970). Misclassification across adjacent premium categories suggests that while machine learning improves predictive accuracy, it cannot fully eliminate the structural challenges of asymmetric information and heterogeneous risk profiles. Prospect theory (Kahneman & Tversky, 1979) further illuminates customer sensitivity: consumers are more loss-averse to higher premiums, reinforcing the need for transparent justification of pricing.

Managerially, the results equip insurers with tools for more transparent and fair pricing. Feature importance analysis allows insurers to justify premium structures to regulators, demonstrating that premium levels are driven by objective, measurable factors rather than arbitrary decisions. This transparency can enhance trust among customers, while also supporting compliance with evolving regulatory requirements. Furthermore, focusing on the majority premium group (3–5 million) enables insurers to tailor products and marketing strategies to the largest and most price-sensitive segment of the market.

Theoretical and Managerial Contributions

Theoretical Contributions

This study makes significant contributions to the literature on health insurance, machine learning, and decision sciences in several key ways. To begin with, it extends the application of Random Forest to health insurance premium prediction in the context of an emerging market. Much of the existing work on insurance pricing and predictive modelling is concentrated in developed economies, where market structures, regulatory frameworks, and consumer profiles differ considerably. By focusing on Indonesia, this study enriches the literature with empirical insights from a context marked by socioeconomic diversity, rapidly evolving regulatory systems, and expanding middle-class participation in private health insurance.

Additionally, the study provides evidence that ensemble methods outperform traditional actuarial approaches, such as logistic regression, in health insurance premium prediction. While prior studies have established the technical superiority of ensemble models in various domains, few have empirically validated their performance in health insurance pricing. By demonstrating that Random Forest achieves higher accuracy, precision, recall, and AUC than logistic regression and single decision trees, this research strengthens the methodological discourse in management science, confirming that ensemble models can better capture nonlinearities and multidimensionality in customer characteristics.

Beyond this, the research addresses the ongoing debate in the machine learning literature regarding the trade-off between accuracy and interpretability. Black-box algorithms such as gradient boosting or deep neural networks often achieve high predictive accuracy but offer limited transparency, raising concerns in regulated industries where fairness and accountability are paramount. Random Forest strikes a balance by providing not only strong predictive performance but also feature importance rankings that reveal which variables most strongly influence premiums. This dual contribution responds to calls in the literature for predictive methods that are both powerful and explainable, particularly in contexts such as insurance where regulators, customers, and firms demand transparency.

Furthermore, the study integrates decision-making theories into predictive modelling. By interpreting results through the lenses of risk pooling, adverse selection, and prospect theory, the research situates machine learning within broader theoretical traditions (Akerlof, 1970; Einav et al., 2018; Kahneman et al., 1979). This contributes to bridging quantitative modelling with conceptual theories of risk and decision-making, advancing a more holistic view of how predictive analytics informs both individual behavior and institutional strategies. In doing so, the study not only extends the application of Random Forest but also enhances its theoretical grounding in management and economics.

Ultimately, this research contributes to the discourse on fairness and bias in insurance analytics. Although not the primary focus, the study highlights the potential risks of using socioeconomic proxies such as electricity usage or transport expenditure, which could inadvertently reinforce existing inequalities. By pointing out this limitation, the paper sets the stage for future research on fairness-aware machine learning in insurance pricing. This critical perspective strengthens the theoretical contribution by acknowledging that predictive power alone is insufficient if models perpetuate bias.

Managerial Contributions

From a managerial perspective, the findings provide several actionable insights for insurers and policymakers. One important implication is that identifying income, age, and dependents as the most influential determinants of health insurance premiums provides insurers with evidence-based criteria for designing pricing strategies that align with customer characteristics while ensuring risk-adjusted fairness. By understanding which variables matter most, firms can develop targeted marketing campaigns, personalized product offerings, and pricing models that improve both affordability and profitability.

Equally important, the study demonstrates that Random Forest can serve as a decision-support system for managers. Beyond prediction, the algorithm's feature importance analysis provides interpretable outputs that managers can use to justify premium structures to regulators and customers. In a highly regulated sector like insurance, this transparency is crucial for maintaining compliance and preventing disputes. For example, insurers can demonstrate that income and dependents systematically influence premium levels rather than being arbitrary or discriminatory, thereby enhancing trust among stakeholders.

Another managerial contribution lies in demonstrating how firms can strike a balance between accuracy and interpretability in strategic decision-making. While deep learning models may offer slightly higher accuracy, their lack of explainability poses risks in customer-

facing industries. Random Forest offers a middle ground, allowing insurers to benefit from machine learning without sacrificing transparency. Managers can thus integrate advanced analytics into their decision processes while maintaining accountability to customers, regulators, and shareholders.

Additionally, the results have direct implications for regulatory compliance and pricing fairness. In Indonesia, as in many emerging markets, regulators are increasingly concerned with ensuring that premiums are fair and accessible across socioeconomic groups. By adopting models that identify key variables and explain premium outcomes, insurers can demonstrate compliance with fairness standards. Moreover, the identification of income and dependents as central drivers allows firms to avoid hidden biases that might arise from overreliance on less transparent proxy variables.

Beyond compliance, the findings also contribute to strategic positioning in competitive insurance markets. As the middle class in Indonesia grows, competition among private insurers is intensifying. Firms that adopt data-driven approaches to health insurance premium prediction gain an advantage in product differentiation and customer targeting. The use of Random Forest not only improves predictive accuracy but also enhances firms' ability to understand their customer base. This deeper customer insight supports innovation in product design, such as tiered premium structures or bundled services, that better meet the needs of diverse customer groups.

Lastly, the study underscores the importance of digital transformation in the insurance industry. As insurers increasingly rely on data-driven strategies, the adoption of machine learning models such as Random Forest is no longer optional but essential. Managers must therefore invest in building organizational capabilities for data collection, pre-processing, and modelling. The findings of this study provide a roadmap for how insurers can integrate machine learning into their operations, starting with health insurance premium prediction and extending to other areas, such as claims processing, fraud detection, and customer retention.

Limitations and Future Research

Every empirical study has its boundaries, and this research is no exception. To begin with, the dataset comprised 148 valid responses collected from prospective health insurance customers in Indonesia. While sufficient for modelling, the relatively modest sample size may restrict the generalizability of the findings to the broader population. A larger and more diverse dataset would provide greater external validity, enabling more nuanced modeling across different demographic and socioeconomic groups.

Moreover, the research design employed a cross-sectional approach, capturing customer characteristics and premium preferences at a single point in time. This limits the ability to analyze how premiums evolve in response to changes in income, family composition, or health risks throughout the customer lifecycle. A longitudinal study design would enable researchers to track premium dynamics more accurately, offering richer insights into long-term insurance behavior.

Beyond this, the modelling scope focused primarily on Random Forest, benchmarked against logistic regression and a single decision tree. While Random Forest demonstrated clear

advantages, the exclusion of other advanced algorithms such as gradient boosting machines, extreme gradient boosting (XGBoost), or deep learning models restricts the breadth of methodological comparisons. Future research could expand the analysis by systematically benchmarking multiple algorithms to identify the optimal balance of accuracy, interpretability, and computational efficiency.

At the same time, although Random Forest provides variable importance rankings, this study did not incorporate fairness-aware machine learning techniques. Insurance models that rely heavily on socioeconomic proxies risk reinforcing existing inequalities if fairness metrics are not explicitly considered. Future studies should therefore explore fairness-adjusted modelling frameworks, ensuring that predictive gains do not come at the expense of equity in premium determination.

A further limitation lies in the reliance on self-reported survey data, which is susceptible to biases such as social desirability or inaccurate recall. For example, income or expenditure data may not be precisely reported by respondents, introducing potential noise into the model. Future research could address this limitation by triangulating survey data with actual financial or administrative records from insurers, thereby improving data accuracy and reliability.

One of the key limitations of this study is the sample size of 148 respondents. While this sample size is sufficient for the scope of this research, it may restrict the ability to generalize the findings to the broader population. Future research should aim to collect a larger, more diverse sample to enhance external validity. Additionally, data augmentation techniques could be employed to bolster the robustness of the findings

Finally, the study was limited to the health insurance premium sector within Indonesia. While the findings are relevant to this context, their applicability to other forms of insurance, such as life, property, or vehicle insurance, remains uncertain. Replicating the study in different insurance domains and across multiple countries would strengthen confidence in the robustness of Random Forest and its utility as a general-purpose tool in insurance pricing.

Looking forward, future research should build on these limitations by adopting larger and more diverse samples, integrating longitudinal designs, and experimenting with alternative algorithms to deepen methodological insights. Moving ahead, scholars should also prioritize fairness-aware analytics and cross-country comparisons, exploring how cultural, economic, and regulatory contexts shape the predictive power of machine learning in insurance. By extending this line of inquiry, future studies will not only improve the robustness of predictive modelling but also ensure that machine learning contributes to fair, transparent, and sustainable practices in global insurance markets.

Conclusion

This study demonstrates the potential of the Random Forest algorithm in predicting health insurance premiums in Indonesia, showing its superiority over traditional models like logistic regression. The research not only improves the accuracy of premium prediction (achieving 63.3%) but also enhances interpretability, which is essential in the regulated insurance industry. By identifying income, age, and dependents as the most significant predictors, the study provides actionable insights for insurers to develop transparent and fair pricing strategies.

Theoretically, this work extends the application of machine learning to health insurance pricing in emerging markets. It bridges the gap between predictive analytics and established decision-making theories such as risk pooling and adverse selection, offering a more nuanced understanding of premium determination. Managerially, the findings highlight the practical value of Random Forest for insurers, allowing them to justify premium structures to regulators and customers, thus fostering trust and regulatory compliance. This study also stresses the importance of adopting data-driven approaches to remain competitive in Indonesia's growing health insurance market.

Despite its contributions, the study has limitations, particularly the sample size of 148 respondents and the cross-sectional nature of the data. Future research should focus on expanding the dataset, integrating longitudinal designs to track premium dynamics over time, and exploring alternative algorithms to enhance predictive accuracy. Additionally, incorporating fairness-aware machine learning models would ensure that predictive gains do not inadvertently perpetuate socioeconomic inequalities. By addressing these limitations, future studies will improve the robustness of health insurance premium prediction and its applicability to other forms of insurance.

REFERENCES

Akerlof, G. A. (1970). The Market for " Lemons ": Quality Uncertainty and the Market Mechanism Author (s): George A . Akerlof Stable URL : https://www.jstor.org/stable/1879431. *The Quarterly Journal of Economics*, *84*(3), 488–500.

Arrow, K. J. (1963). Uncertainty and The Welfare Economics of Medical Care. *American Economic Review*, *5*, i–viii. https://doi.org/10.1257/aer.102.7.i

Breiman, L. (2016). RANDOM FORESTS. *International Journal of Advanced Computer Science and Applications*, *7*(6), 1–33. https://doi.org/10.14569/ijacsa.2016.070603

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

D'Souza, J., & Gurin, M. (2016). The universal significance of Maslow's concept of self-actualization. *Humanistic Psychologist*, *44*(2), 210–214. https://doi.org/10.1037/hum0000027

Duarte, V., Zuniga-Jara, S., & Contreras, S. (2022). Machine Learning and Marketing: A Literature Review. *SSRN Electronic Journal*, 1–19. https://doi.org/10.2139/ssrn.4006436

Einav, O., Glass, G., & Einav, H. (2018). Systems and methods for modular storage and management. *US Patent 10,065,799*. https://patents.google.com/patent/US10065799B2/en

Gkikas, D. C., Theodoridis, P. K., & Beligiannis, G. N. (2022). Enhanced Marketing Decision Making for Consumer Behaviour Classification Using Binary Decision Trees and a Genetic Algorithm Wrapper. *Informatics*, *9*(2). https://doi.org/10.3390/informatics9020045

Kahneman, D., Tversky, A., & Tversky ', A. (1979). Prospect Theory: An Analysis of Decision

under Risk E C O N OMETRICA I C I VOLUME 47 MARCH, 1979 NUMBER 2 PROSPECT THEORY: AN ANALYSIS OF DECISION UNDER RISK. *Source: Econometrica*, *47*(2), 263–292. http://www.jstor.org/stable/1914185%5Cnhttp://www.jstor.org/%5Cnhttp://www.jstor.org/action/showPublisher?publisherCode=econosoc.%5Cnhttp://www.jstor.org

Maccheroni, F., Marinacci, M., Wang, R., & Wu, Q. (2025). Risk Aversion and Insurance Propensity. In *American Economic Review* (Vol. 115, Issue 5). https://doi.org/10.1257/aer.20231529

Müller, A. C., & Guido, S. (2017). Introduction to Machine Learning with Python: a guide for data scientist. In *O'Reilly Media, Inc.* http://kukuruku.co/hub/python/introduction-to-machine-learning-with-python-andscikit-learn

Oktafia, R., & Taufiq, M. (2021). Sharia Insurance Company Business Management Model in the Digital 4.0 Era. *IQTISHADIA Jurnal Ekonomi & Perbankan Syariah*, *8*(2), 222–233. https://doi.org/10.19105/iqtishadia.v8i2.4914

PERSOLKELLY_Indonesia. (2023). *SALARY GUIDE 2023*.

Pradana, M. G., & Ha, H. T. (2021). Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering. *Journal of Applied Data Sciences*, *2*(1), 19–25. https://doi.org/10.47738/jads.v2i1.18

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

R, P. T. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. *Ijarcce*, *4*(1), 196–199. https://doi.org/10.17148/ijarcce.2015.4142

Tchagna Kouanou, A., Mih Attia, T., Feudjio, C., Djeumo, A. F., Ngo Mouelas, A., Nzogang, M. P., Tchito Tchapga, C., & Tchiotsop, D. (2021). An Overview of Supervised Machine Learning Methods and Data Analysis for COVID-19 Detection. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/4733167