Development of a Health Insurance Premium Prediction Model using Machine Learning

Tendai Makoni
Department of Mathematics and Computer Science
Great Zimbabwe University
Masvingo, Zimbabwe
tmakoni@gzu.ac.zw
https://orcid.org/0000-0003-0853-7761

Caroline Rukwava
Department of Mathematics and Computer Science
Great Zimbabwe University
Masvingo, Zimbabwe
rukwavacaroline11@gmail.com

Talent Mawere
Department of Mathematics and Computer Science
Great Zimbabwe University
Masvingo, Zimbabwe
tmawere@gzu.ac.zw
https://orcid.org/0000-0003-2411-6780

Peter Tinashe Chinofunga
Department of Mathematics and Computer Science
Great Zimbabwe University
Masvingo, Zimbabwe
chinofunga@gzu.ac.zw

***Abstract*—In Zimbabwe's evolving healthcare landscape, accurately determining health insurance premiums is critical to improving affordability, reducing risk imbalances, and increasing coverage, particularly amid economic constraints and rising health costs. Traditional actuarial models often struggle to represent the complex, non-linear relationships among socioeconomic, health, and lifestyle variables prevalent in the Zimbabwean population. This paper aims to develop a machine learning model that more precisely and rationally predicts health insurance premiums. Five supervised regression algorithms, Linear Regression (LR), LASSO Regression (LASSO), K-Nearest Neighbours (KNN), Random Forest (RF), and Gradient Boosting (GB), are evaluated for their effectiveness using a representative health insurance dataset that includes demographic and health-related attributes relevant to Zimbabwe. Models were assessed based on their Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) values. The results show that ensemble learning methods, particularly Gradient Boosting, significantly outperform traditional linear models, achieving the highest predictive accuracy. Key predictors of premium costs were identified as chronic illnesses, smoking status, and the number of dependents, variables that are particularly pertinent in local risk assessment. This paper advances health insurance analytics in Zimbabwe by providing evidence that machine learning can support more transparent, data-driven, and context-sensitive premium determination. The findings help insurers, policymakers, and healthcare stakeholders aiming to expand coverage and improve trust in private and public insurance schemes.**

Keywords*—Health Insurance, Premium Prediction, Machine Learning, Ensemble Method, Linear Regression (LR), LASSO Regression (LASSO), K-Nearest Neighbours (KNN), Random Forest (RF), and Gradient Boosting (GB)

I. INTRODUCTION

Health insurance plays a crucial role in protecting individuals and households from the high and often unpredictable costs of medical care. By spreading health-related risks across a broad population, health insurance reduces the financial burden of illness, injury, and unexpected medical expenses. [1] defines a health insurance premium as a fixed monthly payment that individuals make to maintain their health coverage, regardless of whether healthcare services are utilised during that period. The regular payment of health insurance premiums ensures continuous access to medical services, providing policyholders with financial security and peace of mind.

Access to affordable health insurance is a key determinant of improved public health outcomes, as it facilitates timely care, reduces healthcare inequalities, and limits out-of-pocket expenditures. [2] explains that health insurance premium pricing is influenced by multiple risk-related factors, including demographic characteristics and health status, which insurers use to estimate future healthcare costs. These factors collectively determine the risk associated with an insured individual and, consequently, influence premium pricing.

In Zimbabwe, demand for health insurance has been steadily increasing, driven by rising healthcare costs and growing public awareness of its benefits [3, 4]. As the insurance market expands, accurate and fair premium determination has become increasingly important for both insurers and policyholders. Insurers require precise pricing mechanisms to ensure financial sustainability, while consumers seek affordable and transparent premiums that reflect their individual risk profiles.

Recent advances in artificial intelligence (AI) and machine learning (ML) have transformed premium prediction and risk assessment in the global insurance industry. Machine learning models enable insurers to analyse large volumes of demographic, health, and behavioural data to generate personalised premium estimates. By identifying complex and non-linear relationships among risk factors, ML techniques improve pricing accuracy and enhance decision-making for both insurers and consumers. [5] highlights that ML models are particularly effective in identifying hidden patterns within large datasets, making them valuable tools for insurance pricing and risk management.

Despite these advancements, traditional actuarial methods and linear regression models remain widely used in Zimbabwe's insurance sector. While these methods are interpretable and straightforward, they often rely on a limited set of variables and assume linear relationships, thereby failing to fully capture individual risk profiles. As a result, premium estimates may be inaccurate, leading to mispricing and inefficiencies in the insurance market.

A. Problem Statement

The Zimbabwean insurance industry faces challenges in adopting artificial intelligence and machine learning due to limited resources, skills shortages, and poor infrastructure [6]. These issues are critical in health insurance, where accurate premium prediction is vital for affordability and sustainability. Relying on outdated actuarial models leads to poor risk assessment, mispriced premiums, and reduced transparency, thereby increasing financial risk and reducing policyholder trust. Rising healthcare costs and consumer expectations have highlighted the limitations of traditional methods. Machine learning provides a data-driven approach to improve premium accuracy and personalisation [7]. However, few studies have applied ML to Zimbabwean health insurance premiums, underscoring the need to develop and evaluate tailored ML models.

B. Aim

This paper aims to develop and evaluate a machine-learning–based model to estimate health insurance premiums in Zimbabwe's health insurance system.

C. Objectives

The objectives of the paper are to:

- Determine the key demographic and health-related variables influencing health insurance premiums in Zimbabwe.
- Apply and compare selected supervised machine learning algorithms for predicting health insurance premiums.
- Evaluate the predictive accuracy and practical applicability of the developed machine learning models within the Zimbabwean healthcare insurance sector.

II. Literature Review

Recent studies have increasingly explored the application of machine learning techniques in health insurance premium prediction, highlighting their superiority over traditional statistical methods. Linear Regression (LR) has historically been used for its simplicity and interpretability; however, several studies have demonstrated its limitations in modelling complex insurance data. [8], using LR to predict health insurance premiums in India, reported an accuracy of 91%. Despite this high accuracy, the authors acknowledged that LR performs poorly when modelling nonlinear relationships and interactions among risk factors, which are common in health insurance datasets.

Comparative studies further illustrate the limitations of traditional regression models relative to advanced machine learning approaches. [9] compared LR, Random Forest (RF), and Gradient Boosting (GB) using United States insurance data and found that Gradient Boosting significantly outperformed the other models in predictive accuracy. Linear Regression performed the worst, particularly at capturing complex patterns in the data. Similarly, [10] evaluated LASSO Regression, Ridge Regression, K-Nearest Neighbours (KNN), and Extreme Gradient Boosting (XGB) using Kaggle health insurance datasets. Their findings indicated that XGB achieved the highest R^2 values and the lowest Root Mean Square Error (RMSE), while KNN demonstrated comparatively poor predictive performance.

Evidence from developed economies further supports the effectiveness of ensemble learning techniques. [11] applied six machine learning algorithms, including Support Vector Machines (SVM), Decision Trees (DT), RF, XGB, and KNN, to health insurance claims data in the United States. Their results showed that XGB and RF consistently achieved the highest predictive accuracy, reinforcing the strength of ensemble methods in handling high-dimensional, non-linear insurance data. The authors further emphasised that ensemble methods are particularly effective at capturing complex interactions among variables that traditional models often overlook.

[12] evaluated machine learning and deep learning models for health insurance premium prediction in Nigeria and found that RF and neural network (NN) models outperformed conventional regression approaches. [13], in a comprehensive paper conducted in Bangladesh, tested nine predictive models, including LR, XGB, RF, and GB. The paper revealed that XGB achieved the best performance, characterised by the lowest prediction errors and the highest R^2 values, followed closely by GB and RF.

[14], using questionnaire-based data from India, found that Gradient Boosting was the most effective model for predicting health insurance premiums, outperforming Random Forest, XGB, and multiple regression. Similarly, [5], using Kaggle insurance datasets, reported that GB achieved the highest R^2 and lowest error metrics when compared to multiple regression and other ensemble models. [10] further noted that while Ridge Regression improved performance relative to KNN, it was still outperformed by LASSO Regression and XGB in terms of accuracy and error reduction.

While prior studies consistently demonstrate the superior predictive performance of machine learning algorithms, particularly ensemble methods such as Gradient Boosting (GB), Random Forest (RF), and XGBoost (XGB), these findings are largely based on datasets from developed countries or public repositories, as summarised in Table 1. Table 1 shows that although Linear Regression (LR) remains widely used due to its interpretability, it is often outperformed by ensemble methods across multiple performance metrics, including RMSE, R^2, and MAE. Evidence from Zimbabwe or similar low- and middle-income countries remains scarce, leaving unanswered questions about which algorithms perform best on locally relevant data and which predictors are most influential in this context. This study addresses this gap by applying and systematically comparing multiple machine learning and regression models to Zimbabwean insurance data, thereby providing context-specific insights into model performance and key premium drivers.

TABLE I. SUMMARY OF STUDIES ON HEALTH INSURANCE COST PREDICTION MODELS

Authors	Application	Evaluation Metrics	LR.	RF	KNN	GB	XGB	Other Models	Best Model
[9]	Predictive modelling of healthcare insurance costs	MSE, R^2	✓	✓		✓	✓		GB
[10]	Analysis and prediction of health insurance costs	R^2, RMSE			✓		✓		XGB
[11]	Prediction of health insurance claims using ML	MSE, RMSE, R^2, MAPE, MAE, Adj. R^2	✓	✓	✓	✓	✓	SVM, DT	XGB
[12]	Predicting health insurance premiums	MAE, MSE, R^2	✓	✓				ANN	RF
[8]	Medical insurance premium prediction	F1 Score, Recall, Accuracy	✓						LR
[15]	Predicting health insurance premiums	Accuracy, R^2, RMSE						ANN	ANN
[5]	Medical insurance price prediction	R^2, MAE, RMSE	✓	✓		✓	✓		RF and DT
[14]	Customisation of health insurance premiums					✓	✓		GB
[13]	Medical insurance cost prediction	MAE, RMSE, R^2	✓	✓		✓	✓		XGB and RF

Note: Linear Regression = LR, Random Forest = RF, K-Nearest Neighbours = KNN, Gradient Boosting = GB, Extreme Gradient Boosting = XGB, Artificial Neural Network = ANN, Decision Tree = DT, Mean Squared Error = MSE, Root Mean Squared Error = RMSE, Mean Absolute Error = MAE, Mean Absolute Percentage Error = MAPE.

III. METHODOLOGY

A quantitative approach was adopted in this study, employing statistical analysis and machine learning techniques to develop a predictive model for health insurance premiums.

A. Dataset Description

The dataset comprises eight variables: Age, Gender, Body Mass Index (BMI), Number of Beneficiaries, Smoking Status, Province (Harare, Bulawayo, Midlands), Number of Chronic Illnesses, and Insurance Charges. Categorical variables were numerically encoded, and data cleaning confirmed that no values were missing. The data were sourced from a licensed private medical insurance provider (hereafter referred to as Insurance Company A) based in Harare, Zimbabwe, and cover the period from March 2023 to March 2025.

The original dataset comprised 812 unique policyholder records before simulation. These records contained anonymised information on policyholder demographics, medical history, and insurance premiums. Two sets of insurance charges were provided: one calculated using Age and Number of Beneficiaries, and a second using all seven explanatory variables.

To improve model stability, generalisability, and robustness for machine learning analysis, the dataset was expanded through simulation to 1,000 policyholders. The simulation used a distribution-based synthetic data generation approach, in which new observations were generated by sampling from the empirical distributions of the original variables, preserving their observed statistical properties and inter-variable relationships. No actual policyholder records were duplicated during this process, ensuring that the expanded dataset reflects realistic yet synthetic policyholder profiles.

All subsequent analyses, including exploratory data analysis (EDA), feature importance assessment, and model training, were conducted in Python to support reproducible and efficient machine learning workflows. Exploratory Data Analysis was performed to examine variable distributions, identify outliers, and assess relationships between predictors and insurance charges.

Ethical considerations and data privacy were strictly observed throughout the study. The dataset was fully anonymised prior to access, with all personally identifiable information removed by the data provider. No names, identification numbers, addresses, or contact details were included. As the study relied exclusively on secondary, anonymised data and posed no risk to individuals, formal institutional ethics approval was not required. The study nevertheless adheres to accepted ethical standards for data protection and responsible research conduct.

B. Machine Learning Algorithms

This paper employed several supervised machine learning algorithms to predict health insurance premiums. Each algorithm was selected for its suitability for regression tasks and proven effectiveness in insurance-related predictive modelling.

1) K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a non-parametric supervised learning algorithm used for regression tasks. The method predicts the value of a continuous response variable by averaging the outcomes of the k nearest observations in the training data, based on a chosen distance metric (e.g., Euclidean distance). KNN makes no assumptions about the underlying distribution of the data, allowing it to capture complex, non-linear relationships [16].

Formally, for a new observation x_0, the predicted value $\hat{y}_0$ is computed as:

$$\hat{y}_0 = \frac{1}{k}\sum_{i\in N_0} y_i, \quad (1)$$

where N_0 denotes the set of indices corresponding to the k nearest neighbours of x_0 and y_i represents the response value of the i-th neighbour.

KNN regression is particularly suitable when the relationship between predictors and the target variable is non-linear or unknown. Hyperparameters, such as the number of neighbours k and the distance metric, were optimised using cross-validation to achieve the best predictive performance for insurance premium estimation.

2) Random Forest Regressor (RF)

Random Forest (RF) is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. According to [17], RF is an additive model that aggregates predictions from a sequence of base learners. Formally, the model can be expressed as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots, \quad (2)$$

where each $f_i(x)$ represents a base learner, typically a decision tree. In RF, each tree is trained independently using a bootstrap sample of the data, and at each split, a random subset of features is considered. This randomness enhances model diversity and improves generalisation. Model ensembling, as applied in RF, is widely recognised for its ability to increase predictive performance compared to single-model approaches [17].

3) Gradient Boosting (GB)

Gradient Boosting (GB) is an ensemble technique that constructs a strong predictive model by sequentially adding weaker learners, usually decision trees. Unlike RF, which builds trees independently, GB builds models iteratively, with each new model attempting to correct the errors of the previous ensemble. The method minimises a specified loss function by fitting new models to the residuals of prior models. Each successive learner improves overall model performance by focusing on observations that were previously poorly predicted. GB is particularly effective for regression problems involving complex, nonlinear relationships, making it suitable for predicting health insurance premiums [18].

4) Linear Regression (LR)

Linear Regression (LR) is a supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Multiple linear regression extends simple linear regression by incorporating several predictors to estimate a single response variable. The general form of the multiple linear regression model is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + e_i, \quad (3)$$

where y is the dependent variable, $X_1, X_2, ..., X_n$ are the independent variables, β_0 is the intercept, and $\beta_1, \beta_2, ..., \beta_n$ are the regression coefficients and e_i is the error term. Although linear regression is easy to interpret and computationally efficient, it assumes linearity and may not adequately capture complex interactions in insurance datasets [19].

5) Least Absolute Shrinkage and Selection Operator (LASSO) regression

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a regularised regression technique that improves model accuracy and interpretability by performing variable selection and coefficient shrinkage. [20] and [21] explain that LASSO regression is particularly useful for high-dimensional datasets and in the presence of multicollinearity. LASSO regression operates by adding a penalty to the residual sum of squares, which constrains the absolute size of regression coefficients. This penalty encourages sparsity by shrinking some coefficients exactly to zero. The LASSO optimisation problem is expressed as:

$$\widehat{\beta}_\tau^{\text{LASSO}} = \arg\min_\beta \{L(\beta)\} \text{subject to} \sum_{j=1}^{k} |\beta_j| \le \tau \quad (4)$$

where τ is a tuning parameter that controls the strength of regularisation.

C. Model Training and Hyperparameter Tuning

To ensure robust model performance and avoid reliance on default parameter settings, all machine learning models were trained using a systematic hyperparameter optimisation framework. Model tuning was performed using k-fold cross-validation (k = 5), in which the training data were partitioned into folds and model performance was evaluated iteratively across the validation folds. Cross-validation is widely recognised as an effective approach for estimating generalisation performance and reducing overfitting in predictive modelling [22].

Hyperparameter optimisation was implemented using the GridSearchCV utility from the scikit-learn library in Python. Grid search exhaustively evaluates predefined hyperparameter grids and has been shown to provide reliable model selection when combined with cross-validation [23]. For the KNN model, the number of neighbours (k) and distance metrics were tuned. For the RF and GB models, key hyperparameters, including the number of trees, maximum tree depth, and learning rate (for GB), were optimised. For LASSO regression, the regularisation strength (α) was selected through cross-validation to balance model sparsity and predictive accuracy.

The optimal hyperparameter configurations were selected based on cross-validated performance and were then used to train the final models evaluated in the results section. This tuning strategy ensures that the reported model performance reflects well-optimised and fairly compared algorithms rather than suboptimal default configurations.

D. Model Evaluation

Model performance was evaluated using multiple statistical metrics to assess prediction accuracy and generalisability. The metrics included Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-squared (R^2), and Adjusted R-squared (R^2). These measures were computed for both training and testing datasets.

1) Model Accuracy Measures

Root Mean Squared Error (RMSE) is the square root of the average squared prediction error, giving greater weight to larger errors. Mean Squared Error (MSE) calculates the average squared difference between predicted and actual values; lower values indicate better performance. Mean

Absolute Error (MAE) computes the average absolute differences between predicted and actual values, treating all errors equally. Mean Absolute Percentage Error (MAPE) expresses prediction accuracy as a percentage of the actual values, providing insight into relative error. These metrics are widely used in regression evaluation and provide complementary perspectives on the accuracy of predictive models [24, 25]. These four metrics can be expressed collectively as:

$$\left.\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \\ \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \times 100 \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \end{aligned}\right\} \quad (5)$$

where y_i represents the actual value, $\hat{y}_i$ the predicted value and n is the total number of observations. These measures provide a comprehensive assessment of model performance, capturing both absolute and relative prediction errors.

2) R-squared and Adjusted R-squared

The coefficient of determination (R^2) measures the proportion of variance in the dependent variable that the model explains. Lower unexplained variance corresponds to higher R^2 values, indicating better model fit [24, 25, 26]. To account for the number of predictors and prevent overestimation of model performance, the Adjusted R^2 modifies R^2 by penalising the inclusion of irrelevant variables [26]. These two metrics can be expressed together as:

$$\left.\begin{aligned} R^2 &= 1-\frac{\text{RSS}}{\text{TSS}} \\ \text{Adjusted } R^2 &= 1-\frac{(1-R^2)(n-1)}{n-p-1} \end{aligned}\right\}, \quad (6)$$

where RSS is the residual sum of squares, TSS is the total sum of squares, p is the number of predictors in the model, and n is the sample size.

IV. RESULTS

This section presents the outcomes of the predictive modelling, highlighting differences in performance across the applied algorithms, including regression, KNN, and ensemble methods, and identifying the most influential factors affecting health insurance premiums.

A. Summary statistics for continuous variables

Table 2 presents summary statistics for the simulated dataset of 1,000 health insurance policyholders, capturing the central tendency and variability of continuous variables, which are essential for subsequent analysis and modelling.

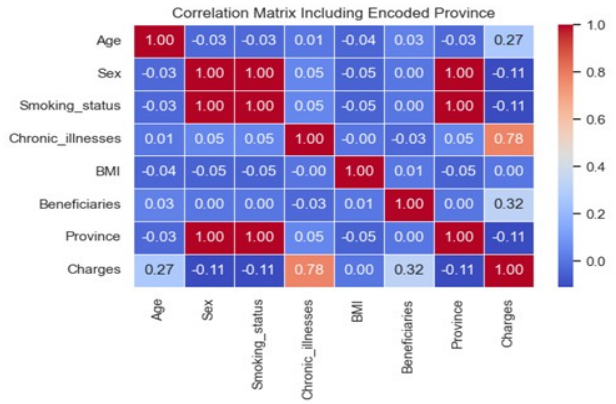
TABLE II. DESCRIPTIVE STATISTICS OF CONTINUOUS VARIABLES

Statistic	Age	Chronic Illnesses	BMI	Beneficiaries	Charges
Mean	50.67	0.80	27.77	2.10	13.16
Standard Deviation	12.95	0.81	7.30	1.04	3.04
Minimum	18.00	0.00	15.00	0.00	6.03
25th Percentile	40.00	0.00	20.00	1.00	10.69
Median (50%)	54.00	1.00	27.90	2.50	12.97
75th Percentile	61.00	1.25	33.90	3.00	15.52
Maximum	70.00	2.00	40.00	3.00	26.92

The average age of policyholders is approximately 51 years, with most clients aged 40 to 61. This suggests higher insurance premiums for older individuals, as age is generally associated with elevated health risks. On average, policyholders report fewer than one chronic illness, though some report up to 2. Body Mass Index (BMI) ranges from 15 to 40, with a mean of 27.77, indicating a predominance of individuals classified as slightly overweight. According to [27], a BMI between 18.5 and 24.9 kg/m^2 is considered normal, whereas values above 25 kg/m^2 are associated with increased health risks. Most policyholders list approximately 2 beneficiaries, while some list none or up to 3. Insurance charges range from \$6.03 to \$26.92, with an average of \$13.16, reflecting the combined effects of policyholder demographics and health profiles.

B. Correlation Matrix of Variables

Figure 1 displays a heat map of the correlation matrix among all variables, including the categorical variable Province.



Note: All categorical variables (Sex, Smoking status, Chronic illnesses and Province) were numerically encoded using one-hot/dummy encoding to enable the calculation of Pearson correlation coefficients with continuous variables.

Fig. 1. Variables correlation matrix.

Chronic Illnesses and Charges exhibit a strong positive correlation (r = 0.78), indicating that individuals with more chronic conditions incur higher healthcare costs [28, 29, 30]. Age and Charges show a moderate positive correlation (r = 0.27), suggesting that older policyholders generally face higher premiums due to increased health risks [31]. Similarly, Beneficiaries and Charges are moderately correlated (r = 0.32), indicating that having more dependents is associated with higher insurance premiums [32].

C. Significance test of variables in the dataset

Table 3 presents the results of the ordinary least squares regression used to assess the significance of variables for the target variable, charges.

TABLE III. SIGNIFICANCE TEST FOR VARIABLES (OLS REGRESSION)

Variable	Coefficient	P-value
Age	0.06	0.001
Sex	0.20	0.001
Smoking Status	10.41	0.001
Province	-0.35	0.001
Chronic Illnesses	3.00	0.001
BMI	0.00	0.250
Beneficiaries	0.99	0.001

An ordinary least squares (OLS) regression was used to assess the significance of each variable in predicting Charges. Table 3 shows that Age, Sex, Smoking Status, Province, Chronic Illnesses, and Beneficiaries are statistically significant ($p < 0.05$), whereas BMI is not statistically significant ($p = 0.25$). Each variable's coefficient indicates the effect size on Charges, and the p-value tests whether this effect is likely due to chance. It is important to note that the OLS p-values presented here are for examining associations between variables, rather than for predictive purposes. Predictive models, including GB, RF, LASSO, and KNN, rely on cross-validated error metrics (such as RMSE, MAE, MSE, and R^2) and feature importance for evaluation rather than on OLS p-values. Furthermore, BMI was retained in the predictive models despite its non-significant p-value, as predictive modelling prioritises improving forecast accuracy over traditional significance tests. This distinction highlights the methodological difference between inferential modelling, which tests hypotheses about relationships, and predictive modelling, which focuses on accurately forecasting future outcomes [33].

D. Feature importance across each model

Table 4 presents the importance (or contribution) of variables, chronic illnesses, Age, Sex, Beneficiaries, BMI, Province, and smoking status across five regression models: RF, GB, LR, LASSO, and KNN. It should be noted that feature importance measures differ by model type and are therefore interpreted within, rather than across, models.

TABLE IV. FEATURE IMPORTANCE OF VARIABLES PER MODEL

Variable	RF	GB	LR	LASSO	KNN
Chronic Illnesses	0.61	0.66	2.44	2.33	0.54
Age	0.12	0.09	0.74	0.64	0.19
Sex	0.02	0.02	0.09	0.00	0.01
Beneficiaries	0.16	0.16	1.04	0.94	0.10
BMI	0.06	0.05	0.05	0.00	0.05
Province	0.02	0.01	0.37	0.27	0.02
Smoking Status	0.02	0.01	1.03	0.96	0.00

Note: Importance values for Linear Regression (LR) and LASSO represent model coefficients, while values for Random Forest (RF) and Gradient Boosting (GB) represent relative Gini/impurity-based feature importance. KNN importance values reflect distance-based contribution measures. Importance values are not directly comparable across different model types but indicate relative ranking within each model.

Chronic Illnesses consistently rank highly, particularly in LR (2.44) and LASSO (2.33), indicating that they are the strongest predictors of premium Charges. Age and Beneficiaries contribute moderately, while Sex, BMI, Province, and Smoking Status generally show little or no importance in some models. To reduce noise and enhance predictive performance, features with zero importance within a given model were excluded from subsequent analyses, consistent with importance-based feature selection strategies outlined by [34].

E. KNN and LASSO Regression Before and After Feature Selection

Performance metrics for KNN and LASSO were evaluated before and after removing unimportant variables. The findings are summarised in Table 5.

TABLE V. KNN AND LASSO REGRESSION BEFORE AND AFTER FEATURE SELECTION

Metrics	KNN Before	KNN After	LASSO Before	LASSO After
R^2	0.4493	0.4411	0.9192	0.9192
MSE	4.39	4.46	0.64	0.64
MAE	1.74	1.75	0.65	0.65
Adjusted R^2	0.4361	0.4297	0.9173	0.9178
RMSE	2.10	2.11	0.80	0.80
MAPE	14.88	15.05	5.27	5.27

After excluding Smoking Status from KNN, the R^2 decreased slightly from 0.4493 to 0.4411, indicating a minor impact. LASSO's R^2 remained unchanged at 0.9192, demonstrating its ability to ignore irrelevant features. This confirms that LASSO simplifies the model without compromising predictive accuracy.

F. Models' Performance Metrics

Table 6 summarises performance metrics (R^2, Adjusted R^2, MSE, RMSE, MAE, and MAPE) for five regression models used to predict premium Charges.

TABLE VI. MODEL PERFORMANCE METRICS

Model	R^2	Adjusted R^2	MSE	RMSE	MAE	MAPE
RF	0.9480	0.9468	0.41	0.64	0.34	2.92
GB	0.9685	0.9678	0.25	0.50	0.29	2.44
LR	0.9272	0.9254	0.58	0.76	0.61	4.85
LASSO	0.9192	0.9178	0.64	0.80	0.65	5.27
KNN	0.4411	0.4297	4.46	2.11	1.75	15.05

GB achieved the highest R^2 (0.9685) and the lowest error metrics, closely followed by RF. LR and LASSO performed reasonably well but were outperformed by ensemble models. KNN exhibited the weakest performance, with the lowest R^2 and the highest error rate, indicating that it is less suitable for this dataset. These results align with previous studies, which show that ensemble methods, such as bagging and boosting, generally outperform individual regression models because of their superior ability to capture complex, non-linear relationships [16, 18].

G. Gradient Boosting Errors in Prediction

Table 7 compares actual insurance charges with those predicted by the Gradient Boosting algorithm, a powerful machine learning technique known for its high accuracy in regression tasks. The Error column shows the difference between expected and actual charges; negative values indicate underprediction, while positive values indicate overprediction. The observations shown in Table 7 are illustrative examples and are not intended as a comprehensive evaluation of all prediction errors.

TABLE VII. GRADIENT BOOSTING PREDICTION ERRORS

Age	Sex	Smoking	Chronic Illnesses	BMI	Beneficiaries	Province	Original Charges	Predicted Charges	Error
18	0	0	0	32.00	3	2	9.26	10.66	-1.40
26	0	0	0	28.00	0	2	6.65	7.77	-1.12
28	0	0	0	20.80	0	3	6.03	6.78	-0.75
40	1	0	0	40.00	1	0	8.70	9.29	-0.60
26	0	0	2	40.00	2	4	12.51	13.09	-0.59
52	0	0	2	37.00	0	1	15.09	14.57	0.51
47	1	0	2	22.40	2	4	14.10	13.63	0.47
40	0	0	0	34.00	3	2	9.83	10.27	-0.44
59	1	0	1	17.60	0	4	11.79	12.22	-0.43
60	1	0	2	17.00	3	1	18.71	18.28	0.43

The GB model demonstrates strong predictive performance, as evidenced by the relatively small errors between actual and predicted insurance charges. In several instances, the predicted charges exceed the company's original charges, suggesting it may be undercharging certain policyholders, which could result in financial losses. Conversely, when the original charges exceed expectations, the company may be overcharging, which could negatively impact customer satisfaction.

Using the GB model enables the company to align premiums more accurately with individual risk factors, ensuring fair pricing and maintaining profitability. Accurate premium pricing is essential to ensure that policyholders pay amounts commensurate with their risk levels: higher-risk individuals pay more and lower-risk individuals pay less, reflecting the fundamental principles of actuarial risk assessment [35]. Setting premiums too low can threaten an insurer's financial stability, whereas excessively high premiums may drive customers to competitors offering more reasonable rates. These observations highlight the critical importance of precise pricing strategies in health insurance.

The findings confirm that machine learning models outperform traditional linear regression approaches in predicting health insurance premiums. The superior performance of GB aligns with the existing literature, which highlights the effectiveness of ensemble learning in handling complex insurance datasets. Identifying smoking status and chronic illness as key predictors underscores the significant role of lifestyle and health risk factors in determining healthcare costs.

A numerical residual analysis was conducted to assess systematic bias. Across all 1,000 simulated policyholders, residuals had a mean of -0.02 and a standard deviation of 0.78, indicating that predictions were, on average, very close to actual charges. Examination of residual variance across predicted charge ranges showed values ranging from 0.72 to 0.85, suggesting relatively stable error variance and no evidence of heteroscedasticity. Residuals were also summarised across key groups; for example, policyholders with 0, 1-2, or 3+ chronic illnesses had mean residuals of –0.05, –0.01, and 0.03, respectively, and no subgroup showed systematic over- or underprediction. These findings confirm that the Gradient Boosting model provides robust and

unbiased predictions, supporting its suitability for accurate and fair health insurance premium estimation.

These results suggest that adopting machine learning techniques can improve pricing accuracy, enhance fairness, and support data-driven decision-making in Zimbabwe's health insurance industry. However, practical implementation requires investment in data infrastructure, technical skills, and regulatory support.

V. Limitations and Future Research Directions

Despite the strong predictive performance achieved in this study, several limitations should be acknowledged and offer opportunities for future research.

A. Data Limitations and Generalizability

This study's main limitation is the data source. The dataset from a single private insurer in Zimbabwe was expanded via simulation to increase the sample size for machine learning. While this enhances model stability and comparison, relying on a single insurer may limit how well the findings apply to others or to the broader Zimbabwean population, given differences in underwriting, risk policies, and client demographics that affect model transferability.

Future research could address this limitation by using multi-institutional datasets from several insurers operating across different provinces. Such an approach would enhance model robustness and improve external validity.

B. Feature Set Constraints

A limitation concerns the scope of explanatory variables. While key predictors such as age, illnesses, smoking, and beneficiary status were available, other variables used in insurance models, such as claims history, clinical codes, occupation, income, and lifestyle details, were not accessible.

The absence of these variables may constrain the predictive capacity of the models and partially explain the relatively weaker contribution of some predictors, such as BMI. Future studies should seek to incorporate more granular clinical and behavioural data to improve the accuracy of premium estimates.

C. Model Interpretability versus Predictive Performance

This study highlights the trade-off between predictive performance and interpretability. Ensemble methods, such as GB and RF, outperform traditional regression in accuracy and error reduction but are less interpretable than linear regression, which offers transparent, coefficient-based explanations.

For regulatory compliance, actuarial justification, or customer-facing pricing explanations, model interpretability is a critical consideration. While GB delivers superior predictive performance, its "black-box" nature may limit its direct applicability in highly regulated insurance environments without additional interpretability tools.

D. Directions for Future Research

Future research should explore integrating explainable artificial intelligence (XAI) techniques, such as Shapley Additive Explanations (SHAP), to enhance the interpretability of high-performing ensemble models. XAI methods can provide insight into individual-level premium drivers while retaining predictive accuracy.

Future studies could explore temporal modelling using longitudinal claims data to track changes in risk profiles. Using more features, multi-insurer datasets, and comparing traditional actuarial models with machine learning would bolster evidence-based premium setting in Zimbabwe's health insurance.

VI. Conclusion and Recommendations

This paper analysed and predicted health insurance premiums using regression and machine learning. Key predictors included Age, Chronic Illnesses, BMI, Beneficiaries, and Charges. Correlation analysis showed that chronic illnesses were strongly associated with Charges. In contrast, Age and Beneficiaries were positively associated, indicating that older policyholders, those with more chronic conditions, or those with more dependents tend to have higher premiums.

OLS regression confirmed that most variables, including Age, Sex, Smoking Status, Province, Chronic Illnesses, and Beneficiaries, significantly influenced insurance charges. In contrast, BMI was not statistically significant but was retained for prediction. Feature importance across models (RF, GB, LR, LASSO, KNN) ranked Chronic Illnesses as the most influential, followed by Age and Beneficiaries. Ensemble methods, particularly GB and RF, outperformed regression models, with GB achieving the highest accuracy and lowest error rates. KNN was less effective, showing its limitations.

Health insurers should adopt Gradient Boosting and Random Forest models for premium estimation due to their high accuracy and ability to capture complex relationships in data. They should focus on key predictors such as chronic illnesses, Age, and Beneficiaries, and regularly update models with new data. Accurate, personalised pricing ensures profitability and fairness by avoiding undercharging high-risk clients or overcharging low-risk clients. Including additional health or lifestyle factors beyond BMI could further improve model precision.

Acknowledgment

The authors thank Insurance Company A for providing access to the anonymized dataset used in this study. We also acknowledge the support of colleagues and reviewers who provided valuable insights during the development of the predictive modelling framework.

References

[1] E. Walker, "How are health insurance premiums calculated?," PeopleKeep, 2024. [Online]. Available: https://www.peoplekeep.com/blog/how-are-health-insurance-premiums-calculated.

[2] L. N. Srinivasagopalan, "Applying reinforcement learning to optimise healthcare insurance premium pricing," *Front. Health Inform.*, vol. 13, no. 2, pp. 11282–11293, 2024.

[3] Statista, "Health insurance – Africa: Market forecast," Statista Market Forecast, 2024. [Online]. Available: https://www.statista.com/outlook/fmo/insurances/non-lifeinsurances/health-insurance/africa. Accessed: May 21, 2025.

[4] T. Chipunza and S. Nhamo, "Potential demand for National Health Insurance in Zimbabwe: Evidence from selected urban informal sector clusters in Harare," *PLoS ONE*, vol. 18, no. 5, p. e0286374, May 2023, doi: 10.1371/journal.pone.0286374

[5] M. M. Billa and T. Nagpal, "Medical insurance price prediction using machine learning," *J. Electr. Syst.*, vol. 20, no. 7s, pp. 2270–2279, 2024.

[6] J. Moyo, N. Watyoka, and F. Chari, “Challenges in the adoption of artificial intelligence and machine learning in Zimbabwe’s insurance industry,” in *Proc. 1st Zimbabwe Conf. Inf. Commun. Technol. (ZCICT)*, 2022.

[7] Lotus Labs, “Insurance premium prediction app: Journey from data to prediction,” Medium, 2024. [Online]. Available: https://lotuslabs.medium.com/insurance-premium-predictionapp-journey-from-data-to-prediction-c2b3b45f43e7. Accessed: May 21, 2025.

[8] G. Reddy and N. Madhuri, “Medical health insurance price prediction,” *Int. J. Novel Res. Dev. (IJNRD)*, vol. 9, no. 4, pp. e592–e603, 2024.

[9] B. A. A. Mahathir et al., “Predictive Modelling of Healthcare Insurance Costs Using Machine Learning,” *Preprints*, 2025, Art. no. 2025021873.

[10] G. K. Patra et al., “An analysis and prediction of health insurance costs using machine learning–based regressor techniques,” *J. Data Anal. Inf. Process.*, vol. 12, no. 4, pp. 581–596, 2024.

[11] A. Alam and V. R. Prybutok, “Use of responsible artificial intelligence to predict health insurance claims in the USA using machine learning algorithms,” *Explor. Digit. Health Technol.*, vol. 2, pp. 30–45, 2024, doi: 10.37349/edht.2024.00009.

[12] Z. Asimiyu, “Predicting health insurance premiums using machine learning: A novel regression based model for enhanced accuracy and personalisation,” Unpublished manuscript, ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/388643113_Predicting_Health_Insurance_Premiums_Using_Machine_Learning_A_Novel_Regression-Based_Model_for_Enhanced_Accuracy_and_Personalization. Accessed: May 21, 2025.

[13] S. Hossen, “Medical insurance cost prediction using machine learning,” Master’s thesis, 2023. [Online]. Available: https://www.researchgate.net/publication/374553777_Medical_Insurance_Cost_Prediction_Using_Machine_Learning.

[14] M. Kapse, V. Sharma, R. Vidhale, and V. Vellanki, “Customisation of health insurance premiums using machine learning and explainable AI,” *Int. J. Inf. Manage. Data Insights*, vol. 5, no. 1, p. 100328, 2025.

[15] L. N. Srinivasagopalan, “Predicting health insurance premiums using machine learning: A novel regression-based model for enhanced accuracy and personalization,” World Journal of Advanced Research and Reviews, vol. 19, no. 01, pp. 1580–1592, 2023, doi:10.30574/wjarr.2023.19.1.1355

[16] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, “Enhancing K Nearest neighbour algorithm: A comprehensive review and performance analysis of modifications,” *J. Big Data*, vol. 11, p. 113, 2024.

[17] N. S. O’Connell et al., “A comparison of random forest variable selection methods for regression modelling of continuous outcomes,” *Brief. Bioinform.*, vol. 26, no. 2, pp. 1–15, 2025.

[18] L. W. Rizkallah, “Enhancing the performance of gradient boosting trees on regression problems,” *J. Big Data*, vol. 12, p. 35, 2025.

[19] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 6th ed., Wiley, 2021.

[20] W. Chen, Q. Liu, H. Li, and J. Zou, “The prediction performance analysis of the LASSO model with convex and non convex sparse regularisation,” *Algorithms*, vol. 18, no. 195, pp. 1–21, 2025.

[21] J. Bhattacharyya, “LASSO Regression – A Procedural Improvement,” ResearchGate, 2025. [Online]. Available: https://www.researchgate.net/publication/389661114_LASSO_Regression_-A_Procedural_Improvement. Accessed: May 21, 2025.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[23] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] F. Che Rose, N. Rosili, and M. F. Marsani, “Comparison of machine learning model performance for predicting the climate variables in Johor Bahru, Malaysia,” *Sci. Rep.*, vol. 15, p. 23465, 2025, doi: 10.1038/s41598-025-08033-y.

[25] H. Khoshvaght, R. R. Permala, A. Razmjou, and M. Khiadani, “A critical review of selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction,” *J. Environ. Chem. Eng.*, vol. 13, no. 6, p. 119675, 2025.

[26] J. Gao, “R Squared (R^2) – How much variation is explained?,” *Res. Methods Med. Health Sci.*, vol. 5, no. 4, pp. 104–109, 2024.

[27] Z. Zheng et al., “Health insurance purchase intentions in the past decade: a systematic review and future research directions,” *BMC Health Serv. Res.*, vol. 25, p. 788, 2025.

[28] R. A. Glynn *et al.*, “Multimorbidity and healthcare costs: Evidence from a national study,” *PubMed*, 2025. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38090275/

[29] Centers for Disease Control and Prevention, “Chronic disease facts & statistics,” 2025. [Online]. Available: https://www.cdc.gov/chronic-disease/data-research/facts-stats/index.html

[30] OECD, *Health at a Glance 2025: Chronic conditions*, 2025. [Online]. Available: https://www.oecd.org/en/publications/2025/11/health-at-a-glance-2025_a894f72e/full-report/chronic-conditions_e0110c98.html

[31] J. Zhou, M. Li, and Y. Lv, “Social medical insurance system and self rated health: Medical service utilisation as the mechanism of action,” *Front. Public Health*, vol. 13, p. 1581130, 2025.

[32] X. Wang and S. Tuo, “Predictive modelling of personal medical insurance costs: Analysing key factors and interactions,” *J. Comput. Electron. Inf. Manage.*, vol. 15, no. 2, pp. 12–22, 2024.

[33] G. S. Collins et al., “Evaluation of clinical prediction models (Part 1): from development to external validation,” *BMJ*, vol. 384, p. e074819, 2024.

[34] H. Wang et al., “Feature selection strategies: a comparative analysis of SHAP value and importance based methods,” *J. Big Data*, vol. 11, p. 44, 2024.

[35] J. Piontkowski, “Pricing German health insurance products with only a few insured persons,” *Eur. Actuar. J.*, vol. 15, pp. 831–857, 2025.