International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.118, June 2026

A Stacking based Ensemble Framework for Health Insurance Premium Estimation in Bangladesh

Shreshtha Sayantika Maitra
Department of CSE
Manarat International University
Dhaka, Bangladesh

Jannatul Ferdaous
Department of CSE
Manarat International University
Dhaka, Bangladesh

Md. Zahurul Haque
Department of CSE
Manarat International University
Dhaka, Bangladesh

ABSTRACT

Only 1% of Bangladeshi citizens have access to medical insurance, while in most developed countries, medical insurance coverage is 100%. Medical expenses are increasing worldwide due to inflation, an aging population, and long-term health conditions; for this reason, better health insurance policies should be ensured for the people. The introduction of machine learning algorithms in health insurance improves efficiency by 75% and lower cost by 50%, which plays a vital role in providing better insurance plans to individuals. The paper aims to help insurance companies streamline the process of predicting premium prices and thereby limit medical expenses. This study applies 16 different machine learning models to a new dataset of 300 rows and 24 columns to predict the price. After evaluating the machine learning algorithms using six different evaluation metrics, namely R-square, MAE, MSE, RMSE, RMSNE, and MAPE, it was deduced that a combination of Polynomial, Ridge, and XGBoost algorithms in our stacked model performs the best at predicting results with an accuracy of 89.4%.

Keywords

Machine Learning, Medical Insurance, Medical Expense, Regression, Stacked Model

1. INTRODUCTION

Medical costs has increased over the last few years due to chronic diseases, price surges in medical technologies, and a growing number of elderly [7]. Low income families have been struggling to pay medical expenses [20], especially when faced with unforeseen circumstances. One of the way of solving this is by introducing health insurance policies for underprivileged people [19].

Medical insurance companies collect small payments from individuals based on their annual income and cover a percentage of their medical expenses when the policyholder makes a claim. The insurance companies do so by gathering premiums from a pool of applicants and performing a risk assessment on the individual who submits the claim, depending on which they determine the insured amount. The insurance companies profit by collecting premiums, and the policyholder benefits by getting better access to healthcare, engaging in early treatments, and avoiding large medical charges.

Statistics show that in low-income countries, the average percentage of people with health insurance was 7.9%, in lower-middle-income countries it was 27.3%, and in upper-middle-income countries it was 52.5% [10]. In contrast, health insurance in most developed countries is universal, covering 100% of the population [13]. In Bangladesh, under 1% of citizens have medical health coverage, which is one of the lowest globally [1]. This implies the need for more regulated insurance schemes, so that the majority of the population can profit from health insurance. The government initiated a health protection scheme in some rural areas for the population below the poverty line, which reduced their overall health expenditures [8].

Machine learning technologies can become a cornerstone for modern-day insurance companies in Bangladesh. Incorporating a self-operating system can summarize the contents of large datasets, which consist of features such as health status, lifestyle, and socio-economic factors, to provide more accurate premium pricing. Moreover, a review of previous medical records can be done to detect fraudulent claims and duplicate submissions. In addition, virtual assistance and chatbots can be utilized to enhance the customer experience.

The goal of our study is to help insurance companies automate the cost prediction process, reduce their costs, and processing time. Although medical insurance in Bangladesh operates on a small scale, we anticipate that our study will help develop more suitable insurance schemes, and their benefits will draw in a wider population.

2. LITERATURE REVIEW

Recent studies by Ranawat [18] have shown that 80% of global medical insurance companies have adopted AI-driven technologies into at least one of their operations. Traditional methods, where risk assessment, fraud detection, predictive analysis, claim processing, and customer service were performed manually, involved long processing times and high management costs. Ansel et al. [5] mentioned that in a research of SDP, the introduction of AI in insurance companies reduced the processing time by 75% and the premium cost by approximately 50%. Atikur et al. [16] And Fahad et al. [21] stated that in developing countries such as Bangladesh, digitization has recently been introduced in some medical insurance companies. However,in most companies, scalability and accuracy was hampered by the absence of automated risk assessment technologies.

The absence of any strong policy in terms of protecting the pri-

vacy of the user in Bangladesh was discussed by Shafiqul et al [9]. Therefore, AI-driven platforms that can improve based on real data are hard to implement.

Kashish et al. [3] presented an automated health insurance cost predictor website with an accuracy of 81.3%, which used linear regression to model relationships between features and insurance charges. A Kaggle dataset containing seven features and 1338 entries was used to train the data. Although K-fold was applied to improve accuracy, there was still a chance of decreased predictability capacity on new data, as the model did not evaluate out-of-sample experimentation on test data. Ramya and colleagues [17] evaluated the insurance amount using two machine learning algorithms, Multinomial Logistic Regression(MLR) and Random Forest Classifier, on the same data set. Afterwards F1 score, precision, recall and accuracy were measured to show that the Random Forest Classifier performed better. Sudhir et al. [15] used several regression-based models (simple linear,multiple linear, polynomial,ridge, and lasso regression) to predict the insurance cost, among them polynomial regression generated the best possible result with an accuracy of 83.62% after verification via RMSE and R-square. Since techniques such as SVM, XG-Boost, Decision Tree, and other optimization techniques were not implemented, the study left some modern approaches unexplored. In order to illustrate the intricate non-linear relationship between the premium cost and the features, Narasimhan and colleagues [11] suggested an ANN technique that yielded an accuracy of 92.72%. Machine learning algorithms such as Correlation analysis, PCA, Recursive feature elimination, back-propagation, gradient descent optimization, and hyperparameter tuning were applied for feature engineering and training data. The reliability of the results was questionable, as the model was not tested on real-life data. Mohtaseem et al. presented an insurance price predictor [4] and applied several regression and tree models on a dataset of 2773 rows with seven different features. The models were compared using MSE, MAE, and R-square. Among these algorithms, GTradient Boosting outperformed all models with an R-squared of 0.87, an MAE of 2383.9, and an RMSE of 4453.8.They identified several future goals, among which the two notable observations were the inclusion of dynamic pricing based on market and the addition of explainable AI. Aminul and colleagues [12] employed Linear, Decision Tree, Random Forest, Elastic, Ridge, and Lasso regression to discover that Gradient Boosting had the highest accuracy of 92%. GridSearch was used for hyperparameter tuning, Lazy Predict was used to compare the models, and InterpretML (an XAI tool) was used to describe model decisions. Several investigations were performed by Emon et al. [6], among which Random Forest excelled, similar to the results reported in other research. SHAP was applied to identify important features and gave better interpretability of premium pricing to stakeholders. In the study by Ugochukwu et al [14], three ML algorithms, XG-Boost, GBM, and RF, were used on a dataset of 986 records and 11 features, which was accessible in Kaggle. Four different loss functions (RMSE, MAPE, R-Square, and MAE) were applied to conclude that XGBoost outperformed others. To determine the key factors affecting the premium price, two XAI methods were used: SHAP and ICE. Although both models exhibited similar performance, the ICE model provided a better explanation for determining the cost. Haitham et al. [2] analyzed the premium price using four ML models: AdaBoost, Gradient Boosting, Elastic Net, and Lasso Regression. The algorithms were evaluated using six metrics (MAE, MSE, RMSE, R-square, RMSLE, and MAPE). The research suggests that Gradient Boosting had the best overall result; analysis of the learning curves demonstrated that other algorithms produced a more stable outcome.

In summary, the above studies highlight the importance of ML and AI applications in the field of medical insurance. Some major gaps in this research were the lack of features, out-of-sample experimentation, risk assessment, applications for fraud detection, a modern approach, and dynamic pricing. Validation of data by verifying the predicted values with specialists, and an increase in privacy policies for data ma- nipulation can help increase trust in such applications, which may provide benefits to both consumers and users.

3. METHODOLOGY

This study begins with the collection of raw data, followed by the composition of the data. After having a complete dataset, we preprocessed it for compatibility with model selection. Then several machine learning models, especially regression models, were applied to the processed dataset to compare the performance. In the end, we selected a stacked regression model as it outperformed other models. The process of the workflow of this study is summarized in Fig. 1.

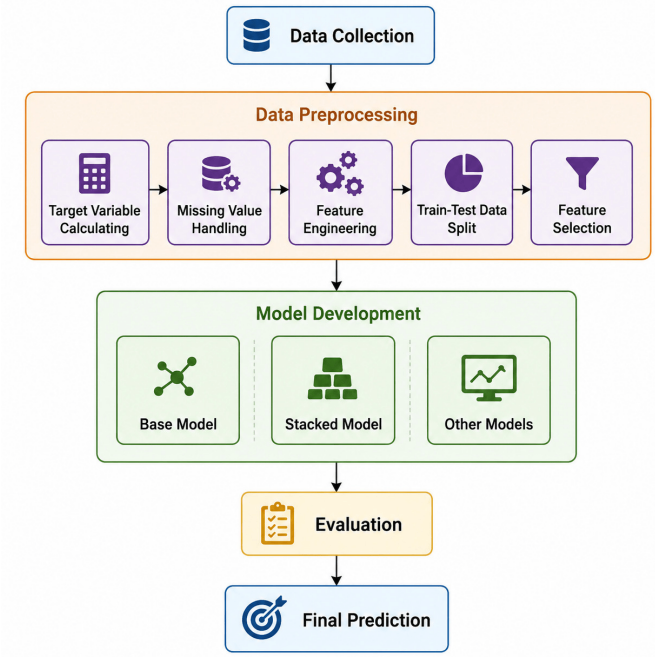

Fig. 1. Workflow of the Study.

3.1 Data Collection & Composition

Our raw dataset contains 24 features with 300 individual tuples. The features can be categorized into demographics, health profile, and policy details. Demographic features are age, gender, division-district, location type, occupation, and annual income. The features representing health profile are pre-existing conditions, family medical history, regular medications, height, weight, blood group, smoker, chew tobacco, physical activity level, and average sleep hours. Policy details can be indicated by coverage type, coverage amount, network hospital preference, critical diseases coverage, maternity coverage, OPD coverage, claim history, and policy tenure. Initially, data were collected from random people living in

Bangladesh and aged between 18 and 80. We got 100 individual real responses from these people, and then we synthesized 200 additional data points based on the real data..

3.2 Data Preprocessing

After getting the raw dataset, we proceed with preprocessing according to the following steps.

3.2.1 Calculating Target Variables. We have calculated two target variables, annual cost for an individual, $y_{\text{cost\_next12m\_bdt}}$, and annual premium to pay, $y_{\text{premium\_calc\_bdt}}$. Firstly, the height is converted from feet to meters to calculate the Body Mass Index (BMI), and then the BMI is normalized. Sum Assured (SA) is calculated. In the next step, different factors assigned to different variables, such as the Age factor (f_{age}), location factor (f_{loc}), claims factor (f_{claims}), BMI (f_{bmi}), smoker (f_{smk}), tobacco (f_{tob}), physical activity (f_{pa}), sleep (f_{sleep}), family history (f_{famhist}), regular medications (f_{regmeds}), division (f_{div}), income (f_{income}), occupation (f_{occ}), network (f_{network}), coverage type (f_{covtype}), pre-existing conditions (f_{cond}), and tenure (f_{tenure}) are all multiplicative and combined in F_{other}.
Composite is calculated prioritizing several factors and multiplying by F_{other}. Gross Need is calculated from this and then adjusted with OPD coverage requirements. Maternity coverage and critical disease coverage are also used to determine the final prediction of the annual insurance cost for an individual. Then the second target variable is calculated by adjusting the first one through multiplication by some constants. After calculating the target variable, the dataset contains 300 rows and 26 columns.

$$h_{\text{m}} = 0.3048\, H_{\text{ft}}, \tag{1}$$

$$\text{BMI} = \min\Big\{\, \max\big(\tfrac{W}{h_{\text{m}}^2}, 10\big), 70 \Big\}, \tag{2}$$

$$\text{SA} = (\text{Coverage Amount (lac)}) \times 100{,}000 \tag{3}$$

$$\begin{aligned} F_{\text{other}} = {} & f_{\text{bmi}}\, f_{\text{smk}}\, f_{\text{tob}}\, f_{\text{pa}}\, f_{\text{sleep}} \\ & \times f_{\text{famhist}}\, f_{\text{regmeds}}\, f_{\text{div}}\, f_{\text{income}}\, f_{\text{occ}} \\ & \times f_{\text{cond}}\, f_{\text{covtype}}\, f_{\text{tenure}}, \end{aligned} \tag{4}$$

$$\text{Composite} = \big(f_{\text{claims}}\big)^{1.30} \big(f_{\text{age}}\big)^{1.20} \big(f_{\text{loc}}\big)^{1.10} F_{\text{other}}, \tag{5}$$

$$\text{Base} = \text{base\_rate\_bdt} \quad (\text{default } 12{,}000 \text{ BDT/yr}), \tag{6}$$

$$\text{GrossNeed} = \text{Base} \times \text{Composite}, \tag{7}$$

$$\text{AdjNeed}_{\text{OPD}} = \begin{cases} \text{GrossNeed}, & \text{if OPD = Yes}, \\ \text{GrossNeed} \times (1 - s), & \text{if OPD = No}, \end{cases} \tag{8}$$

where $s = 0.30$ is the reduction factor

$$\text{Need}_{\text{net}} = \text{AdjNeed}_{\text{OPD}} \times f_{\text{network}} \tag{9}$$

$$\begin{aligned} \text{MaternityEV} = {} & \min\big(0.05 \times \text{SA}, 25000\big) \\ & \times \mathbf{1}\{\text{Maternity = Yes}\} \\ & \times \mathbf{1}\{G = \text{Female}\} \\ & \times \mathbf{1}\{18 \le A \le 45\} \end{aligned} \tag{10}$$

$$\begin{aligned} \text{CriticalEV} = {} & 0.003 \times \text{SA} \times \text{age\_tail}(A) \\ & \times \mathbf{1}\{\text{Critical = Yes}\} \end{aligned} \tag{11}$$

$$\text{PreCap} = \text{Need}_{\text{net}} + \text{MaternityEV} + \text{CriticalEV}, \tag{12}$$

$$y_{\text{cost\_next12m\_bdt}} = \min(\text{PreCap}, \text{SA}), \tag{13}$$

$$\begin{aligned} y_{\text{premium\_calc\_bdt}} = {} & y_{\text{cost\_next12m\_bdt}}\,(1 + e + c) \\ & \times (1 + r + p) \\ & \times (1 + t), \end{aligned} \tag{14}$$

where e = expense load, c = commission load, r = risk margin, p = profit load, t = tax load.

3.2.2 Handling Missing Values. As data were collected from different sources, some values were missing. We use the mean to fill the missing values of numeric columns and the mode for the missing values of categorical columns.

3.2.3 Feature Engineering. Firstly, we use height and weight columns to calculate BMI, and eliminated the previous columns. Then we drop the division-district feature, keeping only the division values, considering that the district has minimal impact on the final cost. We use one-hot encoding to encode categorical variables and finally normalize numeric features (excluding target variables) using Min-Max scaling. After feature engineering, the number of features becomes 131.

3.2.4 Train-Test Data Split. We use 80% of the data for training and 20% of the data for testing, taking random samples from both real and synthetic data.

3.2.5 Feature Selection. We apply permutation importance to rank the important features. After testing with different numbers of top features, we find that using the top 50 features gives the highest model performance, and thus, we select 50 important features for our work.

3.3 Model Selection

We evaluate different regression models on our dataset to predict the target variable (annual insurance cost). Generally, regression indicates the relationship between an independent variable and a dependent variable.

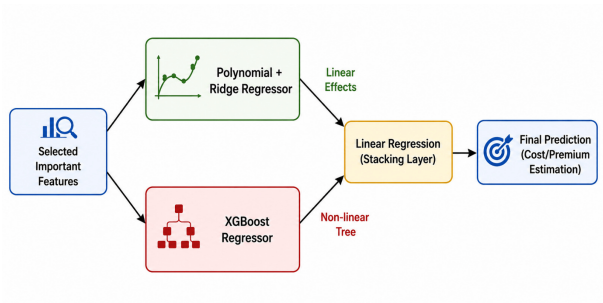

Fig. 2. Proposed Stacked Model.

3.3.1 Linear Regression. A linear regression model tries to predict an output variable as a linear function of input variables.

3.3.2 Ridge Regression. A ridge regression model is a kind of linear regression with L2 regularization, which is used to reduce overfitting.

3.3.3 Polynomial Regression. A polynomial regression model is also a variant of linear regression using the powers of the original features.

3.3.4 Support Vector Regression. Support vector regression is a type of model that uses support vector machines to fit a function.

3.3.5 Gradient Boosting. A Gradient boosting is a tree-based model that uses gradient descent on the loss function for iterative predictions.

3.3.6 Decision Tree. A decision tree is also a tree-based model that recursively splits data based on feature value.

3.3.7 Extra Trees Regression. Extra tree regression is an ensemble method combining different decision trees, randomly splitting trees.

3.3.8 AdaBoost. AdaBoost is also an ensemble method starting with weak learners and gives priority to high errors sequentially.

3.3.9 Random Forest. Random forest also uses randomly split decision trees and averages the outputs.

3.3.10 Neural Network (MLP). Multi-layer Perceptron (MLP) is a simple neural network consisting of one or more layers of neurons.

3.3.11 KNN Regressor. K-nearest neighbor regression is a model that predicts an output by averaging the values of K nearest neighbors.

3.3.12 Lasso Regression. Lasso regression is a linear regression model that uses L1 regularization.

3.3.13 ElasticNet. ElasticNet uses both L1 (Lasso) and L2 (Ridge) regularization.

3.3.14 Proposed Stacked Model. Our proposed model is comprised of two base models: Polynomial + Ridge regression and XGBoost regression. Polynomial + ridge regression is used to capture linear effects of the features, and XGBoost is used to capture the non-linear effects using a tree-based structure. A linear regression model acts as the meta model and outputs the final prediction using the combination of two predictions from two base models. We use 5-fold cross-validation to reduce overfitting and fine-tune the stacking combination.
Fig. 2 shows the block diagram of the proposed stacked model, and Table 1 represents the hyperparameters used in our study.

Table 1. Hyperparameter Configuration of the Proposed Stacking Model

Model	Hyperparameter	Value
Polynomial + Ridge (Base Model 1)	Polynomial Degree	2
	Include Bias	False
	Scaler	StandardScaler
	Ridge Alpha	100
XGBoost Regressor (Base Model 2)	Objective	reg:squarederror
	n_estimators	800
	max_depth	3
	learning_rate	0.05
	subsample	0.7
	colsample_bytree	0.7
	random_state	42
Stacking Regressor (Meta Model)	Base Models	Polynomial+Ridge, XGBoost
	Final Estimator	LinearRegression
	Cross Validation (cv)	5
	n_jobs	-1

4. RESULT ANALYSIS

This section represents the performance evaluation of our proposed stacked model. For comparison, we have used the base models individually and other regression models. The results are evaluated using six standard regression metrics: R-squared (R^2 Score), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), RMSNE (Normalized Root Mean Squared Error), and Mean Absolute Percentage Error (MAPE). Our proposed model outperforms in all metrics. Table 2 summarizes the comparison among all the models.

Table 2. Performance Metrics of Different Regression Models

Model	R^2	MAE	MSE	RMSE	RMSNE	MAPE (%)
Proposed Stacked Model	**0.8942**	**0.0428**	**0.0031**	**0.0557**	**0.4357**	**28.67**
Polynomial Regression	0.7961	0.0573	0.0059	0.0773	0.4566	32.61
XGBoost Regression	0.7708	0.0556	0.0067	0.0819	0.3241	24.85
Ridge Regression	0.7214	0.0682	0.0082	0.0903	0.9081	49.14
Linear Regression	0.7174	0.0687	0.0083	0.0909	0.9154	49.35
Gradient Boosting	0.6249	0.0697	0.0109	0.1048	0.4476	32.82
Support Vector Regression	0.6034	0.0869	0.0116	0.1078	0.8439	57.82
Extra Trees Regression	0.5859	0.0829	0.0121	0.1101	0.6617	47.01
AdaBoost	0.5440	0.0921	0.0134	0.1155	1.0976	70.71
Random Forest	0.5438	0.0809	0.0134	0.1156	0.5465	40.67
Decision Tree	0.3236	0.1034	0.0198	0.1407	0.5457	45.19
Neural Network (MLP)	0.3128	0.1095	0.0201	0.1418	1.2515	72.65
KNN Regressor	0.2749	0.1068	0.0212	0.1457	0.7143	51.49
ElasticNet	-0.0164	0.1395	0.0298	0.1725	1.8667	107.17
Lasso Regression	-0.0164	0.1395	0.0297	0.1726	1.8668	107.17
Polynomial + Ridge Regression	-0.1690	0.1348	0.0342	0.1850	1.0439	70.40

After preprocessing our dataset, 50 important features were selected to predict the target variable using permutation importance. Fig. 3 shows the top 20 important features selected from our dataset.

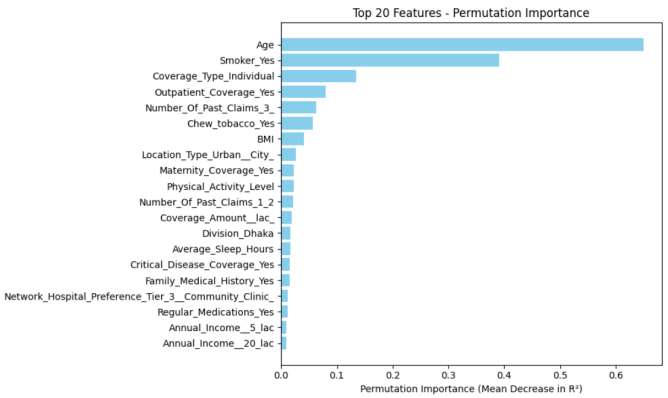

Fig. 3. Visualization of Top 20 Important Features using Permutation Importance.

Fig. 4 presents the plot for true target variable vs. predicted data revealing the visualization of the performance of the model. Most of the data points near the diagonal line indicates that our model is capable of predicting target value with minimal prediction error.
Residual error is the difference between the actual target value and the model's predicted value. Positive value of residual error indicates the underestimation of the model while the negative value represents the overestimation. In our case, as shown in Fig. 5, most of the data points are around the horizontal zero line. Although there is a slight spread at higher predictive values, it highlights the overall high predictive accuracy of the model.

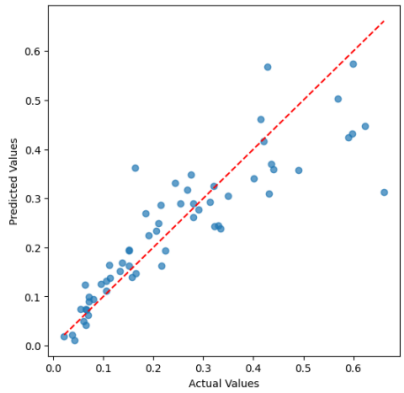

Fig. 4. Predicted vs Actual Target Value Using Stacked Model

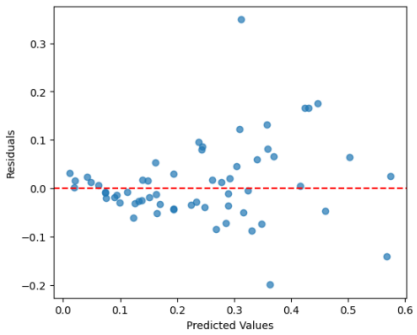

Fig. 5. Analysis of Prediction Residuals Against the Target Variables.

The residual distribution in Fig. 6 depicts the closest representation of a good predictive model having most of the residuals near zero. Slightly right skewed distribution indicates occasional overestimation of the model which may be happened due to presence of noise in the dataset.

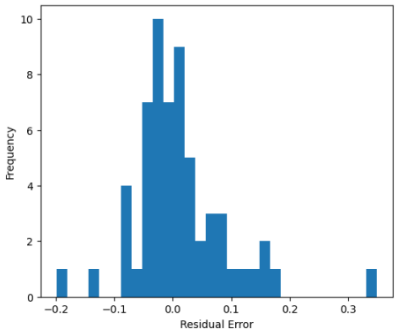

Fig. 6. Residual Distribution Plot Illustrating Error Variance and Model Stability.

4.1 Comprehensive Evaluation and Scenario Analysis

To further validate the robustness and generalizability of the proposed stacked model, extensive evaluations were conducted under various scenarios and on an external dataset.

4.1.1 Cross-Dataset Generalization: . The proposed framework was evaluated on the standard US Medical Insurance Cost dataset (Kaggle). Despite differences in demographic and policy features, the stacked model (Polynomial+Ridge, XGBoost, and Linear Meta-model) achieved an R^2 of 0.86, outperforming standalone Random Forest (0.81), XGBoost (0.83), and MLP (0.79). This confirms the model's adaptability to different insurance structures.

4.1.2 Robustness Across Splitting Scenarios:. To ensure the 89.4% accuracy was not an artifact of the 80-20 split, we tested 70-30 and 90-10 splits. The proposed model maintained R^2 scores of 0.876 and 0.902, respectively, showing minimal variance compared to baseline models like Decision Tree and MLP, which suffered significant performance drops in the 70-30 scenario.

5. CONCLUSION

Medical insurance cost prediction using machine learning is becoming increasingly popular worldwide, especially in developing countries like Bangladesh. This paper introduces a completely new and enriched dataset consisting of 24 features and 300 data points of the people of Bangladesh. As our research utilizes a wider range of features compared to the previous studies, the predicted cost becomes more insightful.

The process of calculating two target values is shown, considering the importance factors of different features of the dataset. Finally, a stacked model is proposed and compared with the other 15 models for predicting the target variable. Six evaluation metrics are used to show the performance of these models, and in all cases, the proposed stacked model with an R^2 score of 89.4% outperformed the others. Although there were some limitations, this study makes a significant contribution to the medical insurance sector in Bangladesh.

6. LIMITATION & FUTURE WORK

Although the proposed model achieved high standards, there is scope for exploration and development in several directions. Future work will focus on utilizing a larger dataset consisting of real-life data to evaluate the accuracy of our model. Additionally, integration of technologies like XAI to interpret results can be added to improve the reliability of the predicted results.

Subsequent investigations will focus on creating Web interface that can be incorporated in the company website for faster processing and automated prediction. Moreover, features such as automating the process of risk assessment and detection of fraud can also benefit the insurance companies.

Another potential direction could be verifying the predicted costs with experts for better clarification of the results produced by the model. Further studies need to be carried out on approaches that aid in risk assessment and dynamic pricing of premiums.

7. REFERENCES

[1] Md Fuad Al Fidah, Syeda Sumaiya Efa, and Md Ziaul Islam. Willingness-to-pay for community-based health insurance among formal and informal doctors in dhaka, bangladesh: a comparative cross-sectional study. *BMJ Public Health*, 3(2), 2025.

[2] Haitham M Alzoubi, Nizar Sahawneh, Ahmad Qasim Al-Hamad, Umar Malik, Ameer Majid, and Ayesha Atta. Analysis of cost prediction in medical insurance using modern re-

gression models. In *2022 International Conference on Cyber Resilience (ICCR)*, pages 1–10. IEEE, 2022.

[3] Kashish Bhatia, Shabeg Singh Gill, Navneet Kamboj, Manish Kumar, and Rajesh Kumar Bhatia. Health insurance cost prediction using machine learning. In *2022 3rd International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE, 2022.

[4] Md Mohtaseem Billa and Tapsi Nagpal. Medical insurance price prediction using machine learning. *Journal of Electrical Systems*, 20(7):2270–2279, 2024.

[5] Ansel Durant, Farren McClure, Maheshwari Karunakaran, and Liam Anderson. Artificial intelligence is transforming the insurance industry: Introducing innovative methods that revolutionize the buying process for customers. *Journal of Transformative Global Research*, 12(9):105–113, 2022.

[6] Shahriar Emon, Md Rakib Hossain, SM Mahedy Hasan, Azmain Yakin Srizon, Farzana Akter, Md Farukuzzaman Faruk, and Md Shakib Hossain. Prediction of medical insurance costs: A shap-enhanced predictive analysis for transparency and interpretability. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE, 2023.

[7] Faizan Fazal, Tayyaba Saleem, Mohammad Ebad Ur Rehman, Tehseen Haider, Abdul Rauf Khalid, Usama Tanveer, Haris Mustafa, Junaid Tanveer, and Arooba Noor. The rising cost of healthcare and its contribution to the worsening disease burden in developing countries. *Annals of medicine and surgery*, 82, 2022.

[8] Md Zahid Hasan, Sayem Ahmed, Gazi Golam Mehdi, Mohammad Wahid Ahmed, Shams El Arifeen, and Mahbub Elahi Chowdhury. The effectiveness of a government-sponsored health protection scheme in reducing financial risks for the below-poverty-line population in bangladesh. *Health Policy and Planning*, 39(3):281–298, 2024.

[9] S Hassan, M Dhali, F Zaman, and M Tanveer. Big data and predictive analytics in healthcare in bangladesh: regulatory challenges. heliyon, 7 (6), e07179, 2021.

[10] Brady Hooley, Doris Osei Afriyie, Günther Fink, and Fabrizio Tediosi. Health insurance coverage in low-income and middle-income countries: progress made to date and related changes in private and public health expenditure. *BMJ global health*, 7(5), 2022.

[11] Mylib In. Predicting health insurance premiums using machine learning: A novel regressionbased model for enhanced accuracy and personalization. *World Journal of Advanced Research and Reviews*, 2023.

[12] Md Aminul Islam, Anindya Nag, Pretam Chandra, Bhupesh Kumar Mishra, SM Firoz Ahmed Fahim, and Md Mozammel Hoque. Healthcare cost patterns and prediction: investigating personal datasets using data analytics. In *International Conference on Signal and Data Processing*, pages 341–359. Springer, 2023.

[13] Enos Mirembe Masereka, Linda Grace Alanyo, Antony Ikiriza, Maureen Andinda, Pardon Akugizibwe, and Emmanuel Kimera. Perspective chapter: Public health insurance in developing countries. In *Health Insurance Across Worldwide Health Systems*. IntechOpen, 2024.

[14] Ugochukwu Orji and Elochukwu Ukwandu. Machine learning for an explainable cost prediction of medical insurance. *Machine learning with applications*, 15:100516, 2024.

[15] Sudhir Panda, Biswajit Purkayastha, Dolly Das, Manomita Chakraborty, and Saroj Kumar Biswas. Health insurance cost prediction using regression models. In *2022 International conference on machine learning, big data, cloud and parallel computing (COM-IT-CON)*, volume 1, pages 168–173. IEEE, 2022.

[16] Atikur Rahman and Amirul Al Rafi. Life insurance underwriting in bangladesh: A comprehensive analysis of practices, challenges, and opportunities. *European Journal of Business and Management Research*, 10(4):177–185, 2025.

[17] D Ramya, J Deepa, et al. Health insurance cost prediction using machine learning algorithms. In *2022 International Conference on Edge Computing and Applications (ICECAA)*, pages 1381–1384. IEEE, 2022.

[18] Chetan Prakash Ranawat. Ai-driven operational efficiency optimization in insurance: A technical implementation guide. *International Journal for Multidisciplinary Research (IJFMR)*, 22, 2024.

[19] Ileana Vilcu, Lilli Probst, Bayarsaikhan Dorjsuren, and Inke Mathauer. Subsidized health insurance coverage of people in the informal sector and vulnerable population groups: trends in institutional design in asia. *International Journal for Equity in Health*, 15(1):165, 2016.

[20] Runar Vilhjalmsson. Family income and insufficient medical care: A prospective study of alternative explanations. *Scandinavian Journal of Public Health*, 49(8):875–883, 2021.

[21] Fahad Zeya, Nargis Sultana, Kazi Saifur Rahman, and Shakil Ahmad. Digital transformation adoption in the insurance sector of bangladesh: A quantitative study from the perspective of insurer. In *2023 4th IEEE global conference for advancement in technology (GCAT)*, pages 1–5. IEEE, 2023.