Improving Medical Insurance Cost Prediction Accuracy with Explainable Supervised Machine Learning based Classification Techniques

Appurva Sharma[1], Alok Bansal[2], Subhash Chandra Jat[3]

[1]*MTech Scholar, Department of Computer Science & Engineering, Jaipur Engineering College, Kukas*

[2]*Assistant professor, Department of Computer Science & Engineering, Jaipur Engineering College, Kukas*

[3]*Professor, Department of Computer Science & Engineering, Sri Balaji College of Engineering and Technology Jaipur*

***Abstract*—Health insurance plans help people financially by covering medical bills and reducing the financial burden of disease. Healthcare and health insurance premiums are influenced by a multitude of variables. The right level of coverage and possible advantages may be better identified with the help of early cost predictions for health insurance. In the insurance sector, ML has the potential to increase policy efficiency. Machine learning algorithms are quite good at predicting expensive healthcare costs. Traditional actuarial methods often fall short in capturing complex relationships in the data. Machine learning models, especially ensemble techniques like LightGBM, CatBoost, and Decision Trees, offer improved accuracy and interpretability. The primary objective of this study is to create supervised ML models capable of producing accurate predictions about the cost of health insurance. The dataset, Medicalpremium.csv from Kaggle, was preprocessed through data cleaning, feature scaling using Standard Scaler, and class balancing using Random Over Sampler. Three advanced regression models—LightGBM, CatBoost, and Decision Tree were developed and compared against baseline models like XGBoost and Random Forest. Model performance was assessed using R-square, MAE, RMSE, and MAPE, and hyperparameter tweaking was done via Grid-SearchCV. LightGBM emerged as the best model with an R-square of 98.67%, outperforming CatBoost (97.62%) and Decision Tree (96.18%), as well as traditional models like XGBoost (82.78%) and Random Forest (82.25%). Visual explainability was incorporated through learning curves, actual vs. predicted plots, residuals, Q-Q plots, prediction error plots, and ICE plots. The study concludes that ensemble-based boosting models, especially LightGBM, offer superior accuracy and generalization in predicting medical insurance costs, establishing a reliable methodology for real-world healthcare applications.**

Keywords—Healthcare, Medical Insurance Costs, Machine Learning, LightGBM, CatBoost, Decision Tree, Class Imbalance, Explainable AI, GridSearchCV.

I. INTRODUCTION

Medical care is becoming more expensive, leaving it out of reach for low-income people and workers in developing countries [1]. Effective insurance coverage may shield workers from financial difficulties in such circumstances. The primary obstacle facing the health care sector is the provision of healthcare [2]. To boost employee productivity, the cooperative organizations and businesses provide healthcare insurance coverage [3][4]. In order to determine premiums more efficiently, the health insurance sector has increasingly prioritized the research of actuarial modelling of insurance claims. This is critical for managing current plan members effectively and drawing in new ones [5][6]. Nevertheless, due to the many and intricate factors that impact these costs, it is challenging to develop a reliable model for predicting medical insurance premiums [7][8]. The anticipated expenses of health insurance may be significantly affected by a wide range of factors, such as demographic information, health condition, regional accessibility, lifestyle choices, provider attributes, etc [9][10]. Coverage amount, plan type, deductible, and enrollee's age are just a few of the critical factors that could affect potential medical insurance prices [11]. A highly efficient and open medical insurance system is essential in view of the challenges posed by the

COVID-19 pandemic and the need for universal healthcare coverage [12].

Insurance companies and consumers alike rely on reliable cost projections for health coverage [13]. Insurance companies rely on cost predictions to set fair premiums, manage risks, and ensure financial sustainability, while policyholders benefit from transparent and equitable pricing [14][15]. To use ML for medical insurance premium prediction, one must first compile models that can estimate new clients' rates based on their demographics, health variables, and insurance coverage history [16]. Insurance businesses may enhance the accuracy of policy pricing and risk management by using algorithms such as neural networks, decision trees, or regression [17]. As mentioned by, ML algorithms' opaque nature might sometimes undermine an otherwise impressive healthcare performance [18]. A potential source of bias in predictive analytics is a lack of clarity or understanding about the use of patients' personal and clinical information [19][20][21]. Patients, healthcare administrators, and insurers may all benefit from the new XAi approaches, which provide more clarity on the reasoning behind forecasts and foster openness and acceptance [22][23]. Therefore, issues like accountability and transparency might be resolved with very accurate medical insurance cost predictions, allowing for control over all stakeholders in patient care [12]. Methods for estimating health insurance premiums using supervised ML models are presented in this research. "Supervised learning" is a subfield of ML that trains models with the use of tagged data. Through the use of XAI methods, the main factors influencing the dataset's medical insurance premium prices were uncovered and clarified.

A. *Motivation and Contributions of the Study*

Insurers and policyholders alike have difficulties in obtaining reasonable and accurate cost estimates due to the growing complexity and variable nature of medical insurance premiums, which is the impetus for this research. Traditional models often fall short in capturing nonlinear interactions among demographic and health-related factors. This research is significant as it leverages advanced ML techniques, specifically ensemble boosting models, to enhance prediction accuracy and provide interpretable outcomes. By doing so, it offers a robust and scalable solution for healthcare providers and insurers, contributing to more transparent, data-driven decision-making processes in the medical insurance sector. The study field benefits from the following research contributions provided by the chosen methodology:

- Developed a comprehensive preprocessing workflow combining feature scaling, class balancing, and data cleaning tailored for healthcare cost prediction tasks.
- Implemented and compared multiple advanced supervised machine learning models (LightGBM, CatBoost, Decision Tree) against traditional models like XGBoost and Random Forest under a unified evaluation framework.
- Incorporated visual explainability techniques such as residual analysis, Q-Q plots, prediction error plots, and ICE plots to interpret model behavior and feature influence.
- Utilized GridSearchCV for hyperparameter tuning to optimize model configurations and improve generalization without relying on default settings.
- Conducted thorough exploratory data analysis using advanced visualization tools to reveal insights into the relationships among medical features and their impact on premium pricing.

B. *Justification and Novelty*

The novelty of this study lies in its integrated approach combining advanced ensemble-based ML models with comprehensive preprocessing and explainable AI techniques to predict medical insurance premiums with high accuracy. Unlike conventional methods that often rely on limited models or overlook data imbalances and interpretability, this research justifies the use of LightGBM and CatBoost for their superior learning capabilities and generalization. The research is justified by the growing need for accurate, interpretable, and data-driven solutions in healthcare finance, where even small prediction errors can lead to significant cost implications. By integrating advanced regression techniques with explainable AI components, the study provides not only improved accuracy but also actionable insights, addressing both performance and trustworthiness critical factors in real-world healthcare applications.

C. *Organization of the paper*

Here is how the rest of the paper is structured: Section II delves into the notion of medical insurance and related works. The methods and techniques used are detailed in Section 3. Section 4 presents the experimental and determinant analysis findings. The analysis and recommendations for further research are presented in Section 5.

II. LITERATURE REVIEW

This section reviews existing machine learning algorithms utilized for predicting medical insurance costs. Various problems with medical insurance premiums are the primary focus of the investigation. Several issues plaguing the insurance sector have been the subject of studies that used machine learning approaches. The optimization of benefit amounts utilizing need-based insurance costs is not a focus of recent study on the topic of medical insurance sector benefit amount optimization.

Mohan et al. (2025) aimed to estimate the health insurance claim costs using a ML algorithm with given potential health risk factors, including age, sex, BMI, smoking status, and city. It involves activities such as data pre-processing, feature engineering, model selection along with using measures like R2 and MAE. Due to its extensive capacity to produce a wide range of associations, Random Forest produced the highest predicted accuracy (96.7% in health data modelling) of all the models that were evaluated [24].

Sharma and Jeya (2024) calculates the cost of health insurance by considering a wide range of factors, such as age, sex, BMI, smoking status, and family size. The distinctive feature is the analysis of the robust correlation between these variables and insurance premiums by linear regression. They carefully partitioned data set, with 70% for training, and achieved an impressive 81.3% accuracy. This ethically driven study enhances understanding of the complexities of post-COVID-19 health insurance spending [25].

Vuddanti et al. (2024) utilized the 4968-row USA medical insurance dataset from Kaggle. The dataset's features allow a summary of an individual's personal and medical circumstances. Regression analysis and gradient boosting are the two machine learning methods utilized to build predictive models. The analysis reflected the link between features and costs. After training the model, a 94% accuracy rate has been obtained [26].

Kandula et al. (2024) machine learning algorithms are used to estimate healthcare insurance expenses via the application of computational intelligence. They devised an approach that use many regression models to determine accuracy, as selecting just one would be too challenging when trying to forecast healthcare expenditures. Up to 88.98% of the time, the stochastic gradient boosting method is accurate [27].

Suresh and Shanmugam (2023) studies are conducted on various ML methods to predict health insurance costs using people's personal information and medical records. Results demonstrate that the model produced by TPOT, which is a combination of Gradient Boosting Regressor and LassoLarsCV, achieved 87.45% accuracy on the test set with an RMSE of 0.0686, surpassing the competitors [28].

Thejeshwar et al. (2023) there has been a comparison of the algorithms' efficacy using three different regression models: Linear Regression, SVM, and RF. The models' training on a dataset improved their predictive abilities. They compared the accuracies of various models. This method works wonders in RFR compared to the others since it achieves the performance measure with the highest possible accuracy rate of 87% using a fraction of the computing time [29].

Selvakumar and Panimalar (2023) ran a series of regression analyses on a 24-featured dataset to forecast insurance premiums using R's built-in tools for linear, logistic, decision tree, lasso, ridge, RF, elasticnet, SVM, KNN, and neural network regression. The R-squared score of 0.9533 showed that Random Forest Regression fared better [30].

Existing research on predicting medical insurance costs using machine learning often relies on traditional regression models and lacks systematic comparison with newer ensemble techniques. Many studies fall short in addressing issues like data imbalance, limited feature scaling methods, and minimal application of advanced hyperparameter tuning. Additionally, few incorporate robust visual explainability tools or focus on model interpretability, which are critical for practical deployment in healthcare settings. Table I presents an overview of previous studies that evaluate machine learning approaches for predicting medical insurance costs, providing a comparative analysis of their effectiveness.

TABLE I. SUMMARY OF THE RELATED WORK ON MACHINE LEARNING FOR MEDICAL INSURANCE

Author	Methods	Dataset	Results	Limitation/Future Work
Mohan et al. (2025)	Random Forest Regressor	Health risk factors dataset (Age, Sex, BMI, Smoking Status, City)	Accuracy: 96.7%	May not generalize to different regions, limited variables beyond five features
Sharma and Jeya (2024)	Linear Regression	Custom collected health insurance dataset (Age, Sex, BMI, Smoking, Children)	Accuracy: 81.3%	Limited post-COVID-19 context, dynamic healthcare changes need updated datasets
Vuddanti et al. (2024)	Regression Analysis and Gradient Boosting	USA Medical Insurance Dataset (Kaggle, 4968 rows)	Accuracy: 94%	Structured data only, no external datasets or real-world test cases
Kandula et al., (2024)	Stochastic Gradient Boosting, Multiple Regression Ensemble	Health insurance data	Accuracy: 88.98%	Need for larger and more diverse datasets, scalability not discussed
Suresh and Shanmugam (2023)	TPOT AutoML (LassoLarsCV + Gradient Boosting Regressor)	Health insurance demographic dataset	RMSE: 0.0686, Accuracy: 87.45%	High model complexity, interpretability issues with AutoML pipelines
Thejeshwar et al. (2023)	LinearRegression, SVR, RFR	Health and demographic dataset	Best model: Random Forest Regressor, Accuracy: 87%	future work could expand feature richness
Vijayalakshmi, Selvakumar and Panimalar (2023)	Linear, DT, Lasso, Ridge, RF, ElasticNet, SVR, KNN, Neural Network Regression	24-feature insurance cost dataset	Random Forest: R² 0.9533	Future work: Try ensemble deep learning models for further enhancement

III. METHODOLOGY

Health care predictive modelling is still a hotspot for actuarial research, with many insurance firms looking to make the most of ML strategies to boost efficiency and production. Several systematic procedures were engaged in the suggested technique to guarantee accurate and dependable findings when estimating medical insurance expenses. Initially, a dataset consisting of 986 records and 11 features was sourced from Kaggle. The data preprocessing phase included inspecting the dataset's structure, handling missing values, and standardizing the features using Standard Scaler to ensure uniform distribution. To address class imbalance in the "Premium Price" variable, Random Oversampling was applied, equalizing the number of samples across price categories. The dataset was then split into training and testing sets in a 75:25 ratio. It trained and assessed three ML models: LightGBM, CatBoost, and DT. R², MAE, RMSE, and MAPE were the four measures used to evaluate their performance, which allowed for a thorough comprehension of the prediction accuracy of each model. The general procedure followed in this investigation is shown in Figure 1.

A. Data Collection

The KAGGLE repository served as the source of the medical insurance cost dataset. Age, Diabetes, Blood Pressure Issues, Any Transplants, Any Chronic Diseases, Height, Weight, Known Allergies, History of Cancer in Family, Number of Major Surgeries, and Premium Price are among the 11 attributes/features that affect medical insurance premiums and offer important insights into the connection between specific medical conditions and insurance costs. It includes 986 records.

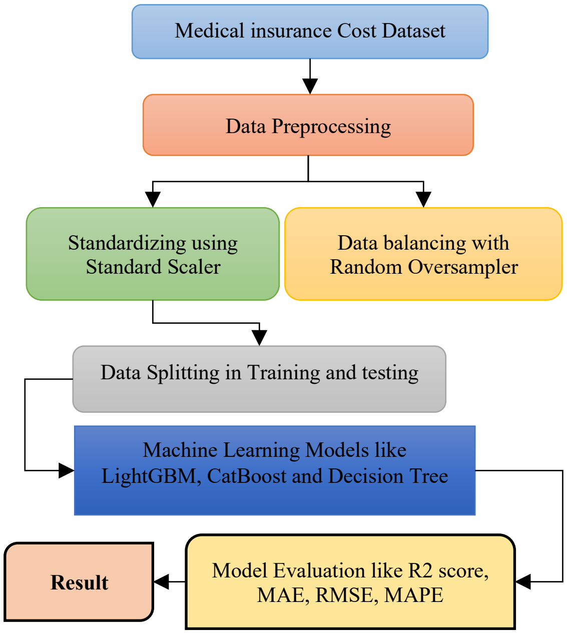
Medical insurance Cost Dataset
Data Preprocessing
Standardizing using Standard Scaler
Data balancing with Random Oversampler
Data Splitting in Training and testing
Machine Learning Models like LightGBM, CatBoost and Decision Tree
Result
Model Evaluation like R2 score, MAE, RMSE, MAPE

Fig. 1. Flowchart for proposed medical insurance cost prediction model using machine learning

Figure 2 displays the correlation heatmap of various features in the dataset, illustrating the strength and direction of linear relationships among pairs of variables. The color gradient ranges from light orange to dark brown, with values closer to 1 indicating strong positive correlations. Notably, "Age" shows a strong correlation with "Premium Price" (0.70) and a moderate correlation with "Number of Major Surgeries" (0.43). Similarly, "Premium Price" is also moderately correlated with "Number of Major Surgeries" (0.26) and "Blood Pressure Problems" (0.24). Most medical conditions, "Diabetes," "Any Chronic Diseases," and "Known Allergies," exhibit weak correlations with the other variables, suggesting minimal linear association. This heatmap helps identify key factors that may influence target variables "Premium Price" for predictive modeling.

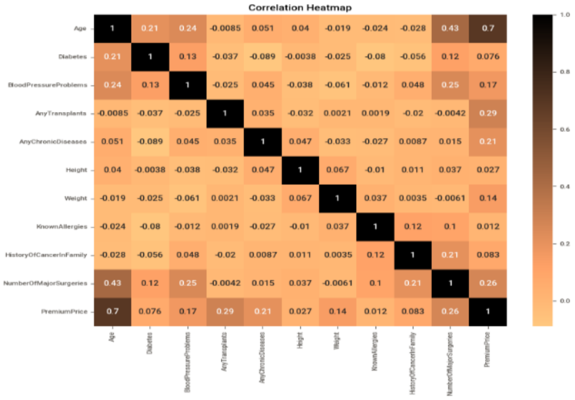

Fig. 2. Correlation Heatmap of Features Related to Insurance Cost Prediction

Figure 3 shows pie charts representing the distribution of patients with respect to transplant history and diabetes. About 94.4% have not had transplants, while 5.6% have. For diabetes, 58% of patients are non-diabetic and 42% are diabetic, indicating moderate class imbalance in these medical features.

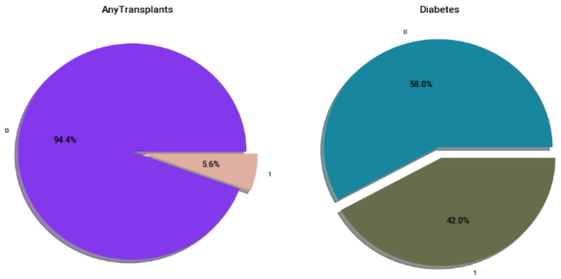

Fig. 3. Distribution of Patients by Transplant History and Diabetes Status

Figure 4 shows bar plots of six medical features, highlighting class distributions. Most patients do not have conditions like transplants, chronic diseases, or a family history of cancer, indicating class imbalance across these variables, which can affect model predictions.

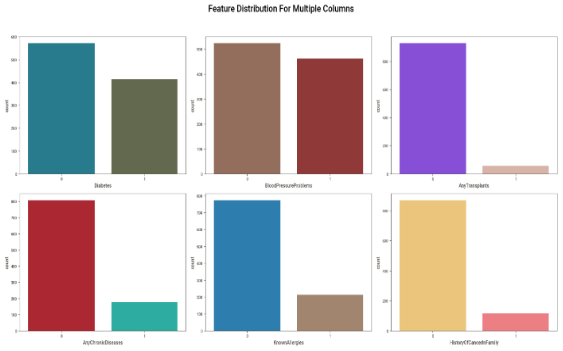

Fig. 4. Feature Distribution for Selected Medical Attributes

Figure 5 presents distributions of numerical features like Age, Height, Weight, Number of Major Surgeries, and Premium Price. Age and Height are fairly normal, while Weight is right-skewed. Premium Price shows outliers, and most patients had no major surgeries, indicating skewed data.

B. Data Preprocessing

This study's data pretreatment included several critical procedures to get the dataset ready for ML modelling. First, the shape of the dataset was obtained to understand the number of records and features. The dataset information was reviewed to check for data types and column names, ensuring consistency and correctness.

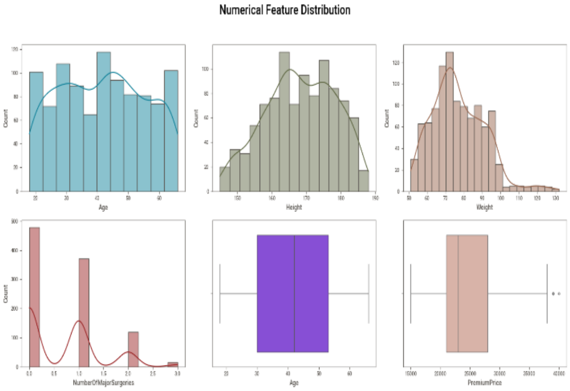

Fig. 5. Numerical Feature Distribution

Descriptive statistics were generated to get an overview of the distribution of each feature. Null values were checked to ensure data completeness and to handle any missing entries appropriately. Every feature was on the same scale since the Standard Scaler changed numerical features to have a mean of 0 and a standard deviation of 1, making them all standardised. To address class imbalance, the Random Over Sampler technique was applied to the "Premium Price" variable, balancing the dataset by increasing the frequency of underrepresented categories. The following key data preprocessing steps are discussed in detail below:

1) Standardizing using Standard Scaler

The data is typically scaled using the Standard Scaler within each component such that the distribution is now centred on 0 and has a standard deviation of 1. The component is first scaled according to Equation (1), following which its mean and standard deviation are calculated:

$$Z_{scaled} = \frac{(X-\mu)}{\sigma} \qquad (1)$$

Where $\mu =$ Mean and $\sigma =$ Standard Deviation

2) Data Balancing with Random Oversampler

A class imbalance occurs when there are more majority classes than minority classes; the ratio in the data is 1: 100, where 100 represents the majority and 1 represents the minority. On the other hand, one method that adds data to a minor class at random without creating changes in class data is called random oversampling. The inaccuracy of class classifications caused by ML is a direct outcome of class imbalance. Synthetic data, a kind of replication, is created using this method to mimic a minority class. As much as the target proportion of duplication is present in each minority data set, the rest is synthetic data.

Figure 6 show the distribution of Premium/Price values for what appears to be an insurance or financial dataset. Figure 6 (blue) displays the original distribution before applying Random Over Sampler, showing a highly imbalanced distribution with significant concentrations at specific price points and many underrepresented values. the distribution after applying Random Over Sampler, which has transformed the data into a perfectly balanced distribution where each Premium/Price value has equal frequency. This comparison effectively demonstrates how Random Over Sampler addresses class imbalance by creating a uniform distribution where previously underrepresented price points now have equal representation in the dataset.

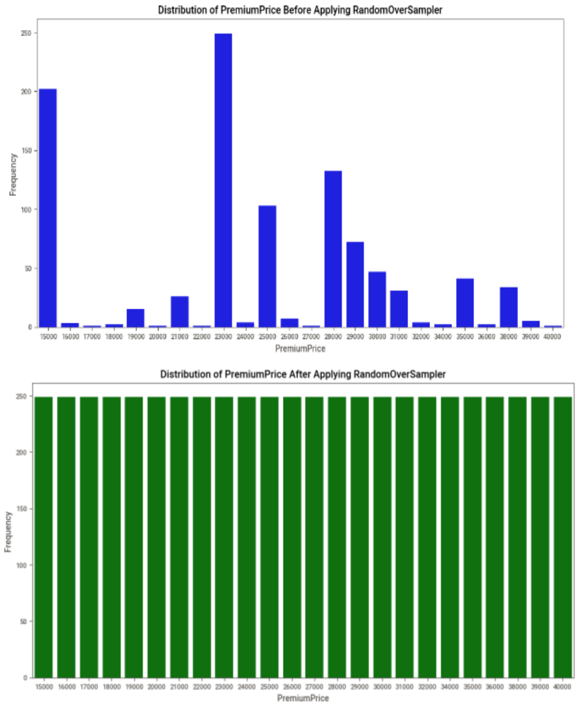

Fig. 6. Distribution of Premium Price Before and After Applying Random Over Sampler

C. Data Splitting

By using the train_test_split function included in the model selection module of the scikit-learn package, x and y may be partitioned into separate sets for testing and training. Partitioning the data in a 75:25 ratio meant that 25% would be used for testing and 75% for training.

D. Machine learning Models for Medical Insurance

ML models are able to provide correct predictions in several domains by analyzing data patterns, which improves prediction system efficiency and decision-making. The efficacy is evaluated with the use of three important classifiers: Decision Tree, LightGBM, and CatBoost.

1) Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) was proposed. This approach, which is based on histograms, speeds up training and decreases memory usage by using a maximum depth limit leaf-wise growth strategy for trees [31]. It is necessary to divide the leaves on the same layer at the same time when using the level-wise growth scheme. Even if the leaves on the same layer have varied information gain (IG), that are regarded equally. The anticipated decrease in entropy as a result of partitioning the nodes according to characteristics is represented as information gain. These entropy and IG equations (Equations 2 and 3):

$$\mathrm{IG(B,V)} = E_n(B) - \sum_{v\in(V)} \frac{|B_v|}{B} E_n(B_v) (2)$$

$$E_n(B) = \sum_{\mathrm{d=1}}^{\mathrm{D}} -\mathrm{p_d} \log_2 p_{\mathrm{d}} \qquad (3)$$

An Information Entropy of the set B is denoted by $En(B)$, p_{d} is the B-related category d ratio, D is the total number of classes, ν is the value of attribute V, and a subset of B with an attribute value of ν is called B_v.

Hyperparameters: Hyperparameter tweaking was done using Randomized SearchCV to improve the LightGBM regressor's performance. Using 10-fold cross-validation, it scored 50 different parameter combinations based on their negative mean squared error. The parameters n_estimators, learning_rate, max_depth, num_leaves, min_child_samples, subsample, and colsample_bytree were all part of the search space. The optimal hyperparameters identified were: n_estimators = 300, learning_rate = 0.2, max_depth = 7, num_leaves = 40, min_child_samples = 5, subsample = 0.9, colsample_bytree = 0.8, and verbosity = -1.

2) CatBoost

CatBoost is an effective gradient boosting technique that minimises information loss while working with categorical inputs. As an alternative to more conventional gradient boosting techniques, CatBoost employs ordered boosting to keep targets from leaking throughout training. It works well with tiny datasets and immediately processes categorical characteristics, turning them into numbers during training [32]. This typically involves replacing categorical variables with binary features for each category. Time series analysis and financial analysis are only two of the many fields that have found useful uses for CatBoost. An important benefit is that it helps minimise overfitting by estimating leaf values during tree construction using random permutations. The algorithm builds binary decision trees as base predictors and is designed to provide accurate and reliable predictions even on complex datasets that express in Equation (4):

$$Z = F(x_i) = \sum_{j=1}^{M} \beta_j h(x; b_j) \qquad (4)$$

where the function h$(x; b_j)$ is the Base Learner, x is the Explanatory Variables, β_j is the Expansion Coefficients, and bj is the parameters of the model.

Hyperparameters: To optimize the performance of the CatBoost regressor, hyperparameter tuning was performed using Randomized SearchCV. A total of 50 random combinations were evaluated through 10-fold cross-validation, using negative MSE as the scoring metric. The search explored a range of values for key parameters including iterations, learning rate, depth, L2 regularization, subsample ratio, and random strength. The best set of hyperparameters identified was: iterations = 200, learning_rate = 0.1, depth = 7, l2_leaf_reg = 1, subsample = 1.0, and random_strength = 1. With these parameters, the final CatBoost model was trained and then applied to the training and test datasets for prediction purposes.

3) Decision Tree

A non-parametric learning method for supervised learning, decision trees are utilized in regression and classification. The input properties are represented by a decision tree's nodes. Leaf nodes of the Decision Tree indicate classes, internal nodes indicate input elements, and branching indicates outcomes [33]. Equation (5) is used to determine the entropy of every class. An approximation of the entropy is given by Equation (6) using two characteristics:

$$H(S) = \sum_{i=1}^{c} -T_i \, log_2 T_i \qquad (5)$$

$$H(S,D) = \sum_{c\in D}^{c} X_{(c)}.Y_{(c)} \qquad (6)$$

where D is the dataset, i is the set of classes in D, and T(i) is the probability of each class.

Hyperparameters: To optimize the performance of a Decision Tree Regressor, Randomized SearchCV is employed for hyperparameter tuning. The search explores 100 random combinations of parameters using 10-fold cross-validation, with negative MSE as the evaluation metric. The selected hyperparameters are as follows: criteria = 'absolute_error', max_features = 'auto', min_samples_split = 2, and max_depth = None. Both the accuracy and the generalizability of the model are enhanced by these factors.

E. Model Evaluation

Model performance is evaluated using R², MAE, RMSE, and MAPE. An better fit is shown by a greater value of R², which shows the percentage of variance explained by the model. When comparing errors, MAE is used to assess their magnitude, RMSE is used to reflect their square root, and MAPE is used to convey mistakes as a percentage. The reliability and accuracy of medical insurance expense predictions are guaranteed by these parameters. These measures are calculated as Equations (7) – (10) show:

$$\text{R squared} = \frac{\text{Stated variation}}{\text{Toatal variation}} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|, \quad (8)$$

$$RMSE = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^{\mathbf{2}}} \quad (9)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{y_t}.100\% \quad (10)$$

- n = Total number of observations (or predictions)
- e_t = The discrepancy between the expected and actual values at time t is the error.
- $|e_t|$ = Absolute value of the error at time t.
- y_t = Actual value at time t.
- $\frac{|e_t|}{y_t}$ = Absolute percentage error for each observation.
- $\mathrm{y_i}$ = Actual value for the *i* th observation
- $\hat{\mathrm{y}}_\mathrm{i}$ = Predicted value for the *i* th observation
- $(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^{\mathbf{2}}$ = Squared difference (error) between actual.

The suggested models' abilities to forecast health insurance premiums are assessed using the following equations.

IV. Result Analysis And Discussion

The study enhances medical insurance cost prediction using explainable machine learning models like LightGBM, CatBoost, and Decision Tree. Experiments were conducted on a local system running Windows 11, equipped with an Intel i7 11th-generation processor, 16 GB RAM, and an NVIDIA GTX 1650 graphics card, without the use of external GPUs. The implementation used Python [34] with libraries Pandas, NumPy, and Scikit-learn. Results in Table II show improved prediction across all models.

TABLE II. Model Performance Comparison for Medical Insurance Cost Prediction

Matrix	LightGBM	CatBoost	Decision Tree
R-square	98.67	97.62	96.18
MAE	222.9553	508.2767	148.5943
RMSE	865.2828	1156.8298	1466.7214
MAPE	1.0952	2.1902	0.7812

Figure 7 provide a comparative evaluation of three regression models—LightGBM, CatBoost, and Decision Tree using four key performance metrics: R-square, MAE, RMSE, and MAPE. LightGBM demonstrates the highest R-square of 98.67%, indicating strong predictive accuracy, while the Decision Tree model achieves the lowest MAE of 148.59 and the lowest MAPE of 0.78%, suggesting better performance in minimizing both absolute and percentage errors.

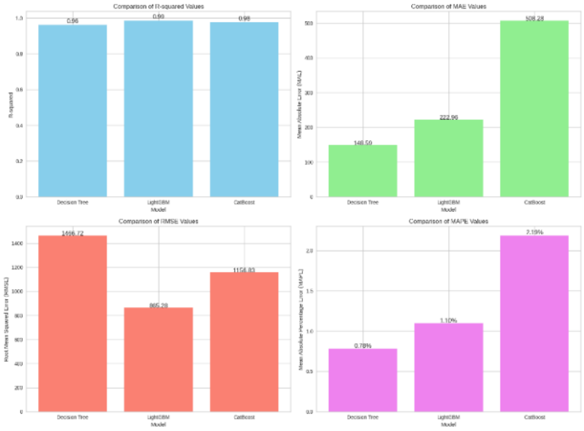

Fig. 7. Model Comparison Medical Insurance Cost for performance matrix

The learning curve of the LightGBM model, shown in Figure 8, shows how the model's performance changes as a function of the number of training examples. You can see the relationship between the amount of training examples and the Negative MSE on the graph. While the training score stays close to zero, the cross-validation score drops with more training occurrences and then stays the same. This proves that the model's generalisability is enhanced with more training data. There seems to be no overfitting as shown by the discrepancy among the training and cross-validation results.

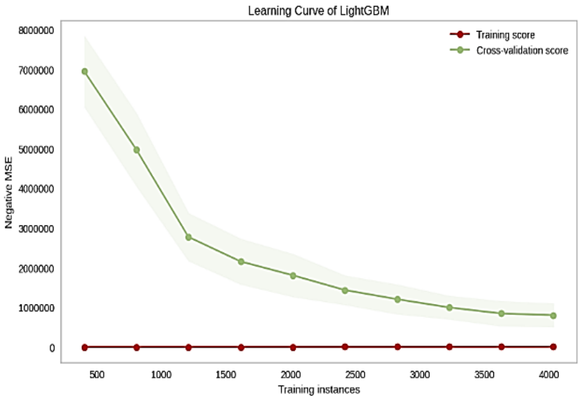

Fig. 8. Learning Curve of LightGBM

Figure 9 shows the CatBoost learning curve, where both training and cross-validation scores improve as training instances increase. The validation error drops sharply at first and then levels off, indicating better generalization with more data. A narrowing gap among the curves suggests reduced overfitting and stable model performance.

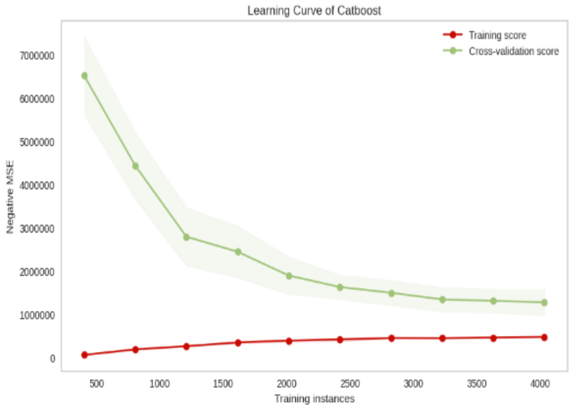

Fig. 9. Learning Curve of CatBoost

The DT Learning Curve for predicting medical insurance costs is shown in Figure 10. The validation error decreases with more data, while the training score remains steady, indicating overfitting. The persistent gap suggests the model could benefit from tuning to improve generalization.

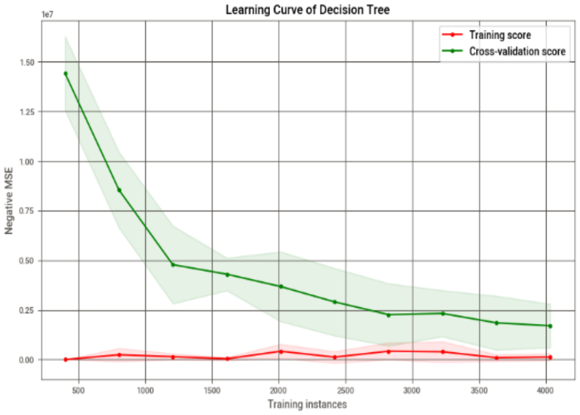

Fig. 10. Learning Curve of Decision Tree

Figure 11 is a scatter plot showing the relationship among the actual values (x-axis) and the anticipated values (y-axis) using a regression model with a R² score of 0.99. Individual forecasts are shown by blue dots, while a perfect prediction is shown by a red diagonal reference line. The close alignment of points around this line, with only minor vertical clustering at certain values, demonstrates the model's excellent predictive accuracy across the range of 15,000 to 40,000 units.

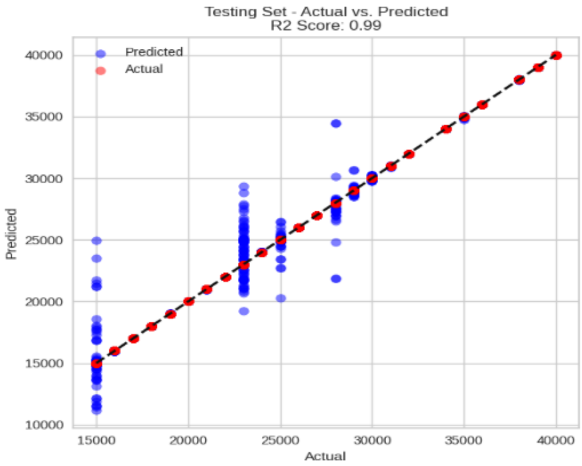

Fig. 11. Actual vs. Predicted for LightGBM

A scatter plot comparing actual and projected medical insurance expenses is shown in Figure 12. Strong predictive performance is shown by the dots aligned along the diagonal line; the model's accuracy is good, as indicated by its R2 score of 0.98.

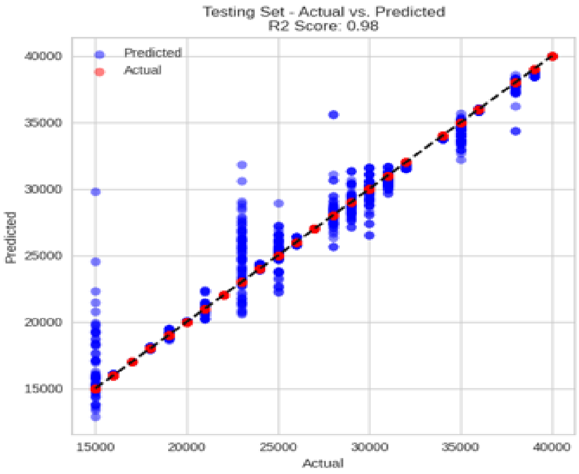

Fig. 12. Actual vs. Predicted Values for CatBoost

Figure 13 presents a scatter plot comparing actual versus predicted insurance costs, highlighting the model's regression performance. The predicted values closely follow the diagonal line, and the R² score of 0.96 indicates a strong correlation and reliable prediction accuracy.

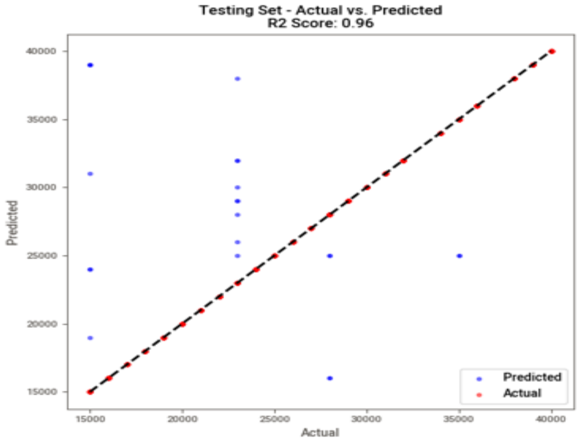

Fig. 13. Actual vs. Predicted Values for Decision Tree

Figure 14 presents two diagnostic plots for evaluating a logistic regression model: the left panel shows "Residuals for Logistic/GLM Model" with orange data points displaying several diagonal patterns plotted against fitted values, with a horizontal dashed line at zero indicating perfect prediction; the right panel displays a "Q-Q plot" with blue points comparing theoretical quantiles against sample quantiles, where points should ideally follow the red diagonal reference line to confirm normality of residuals, though some deviation appears at the tails suggesting potential departure from normal distribution.

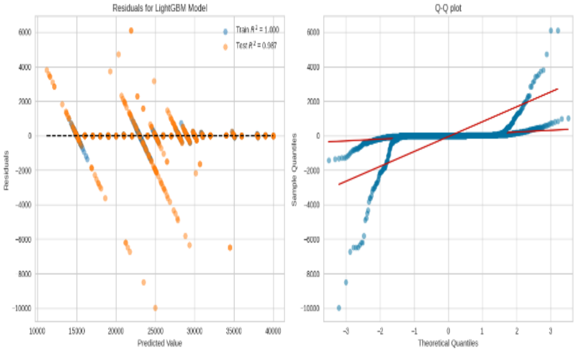

Fig. 14. Residuals and plot + Q-Q plot for LightGBM

Figure 15 provides a diagnostic assessment of the CatBoost model through residual and Q-Q plots. The distribution of residuals about zero is shown in the residual plot (left), which aids in assessing homoscedasticity and identifying bias or trends in forecasts. Departures from the red line in the Q-Q plot (right) show that the errors do not follow a normal distribution, which is based on theory.

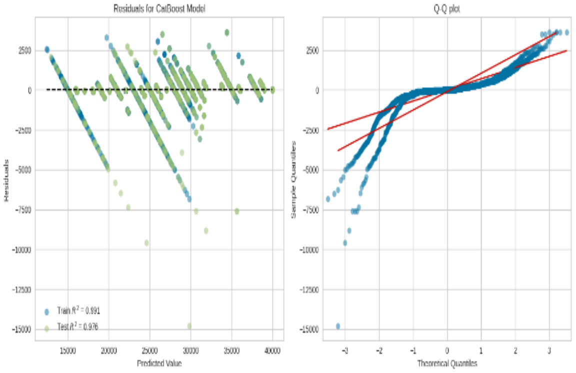

Fig. 15. Residual and Q-Q Plot for CatBoost Model

Figure 16 illustrates the residual plot and Q-Q plot for the Decision Tree model. The residual plot (left) displays the distribution of prediction errors, showing a pattern of non-uniform variance and indicating potential overfitting. The residuals do not seem to be normally distributed, as shown by the large dispersion around the predicted quantiles in the Q-Q plot (right), which evaluates residual normality.

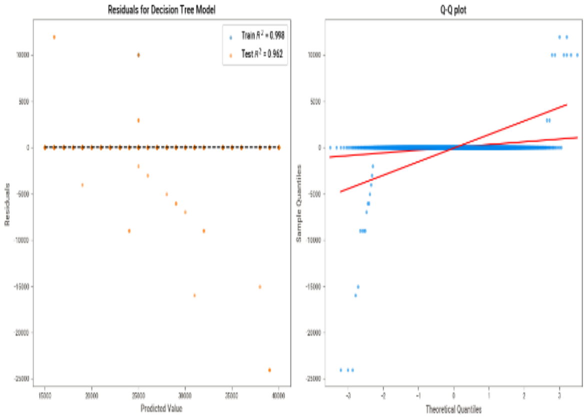

Fig. 16. Residual and Q-Q Plot for Decision Tree Model

Figure 17 shows a scatter plot comparing actual values (x-axis) with predicted values (y-axis), achieving an impressive R² of 0.987. Blue data points follow both the black dashed best-fit line and gray identity line closely across the 15,000-40,000 range, with some vertical clustering at specific values. There is very little discrepancy between the actual and anticipated numbers, which is evidence of the LGBM Regressor model's high prediction accuracy.

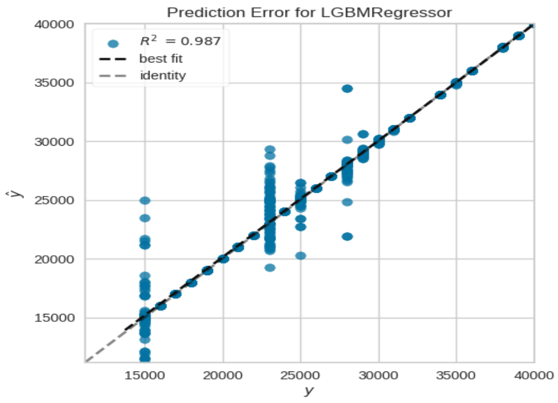

Fig. 17. Prediction error plot for LightGBM Model

Figure 18 presents the prediction error plot for the DTR, showing the relationship among actual values (y-axis) and forecasted values (x-axis). A high R² value of 0.962 and data points that are near the identity line (dashed) show that the model is doing well, with predictions matching up well with actual outputs.

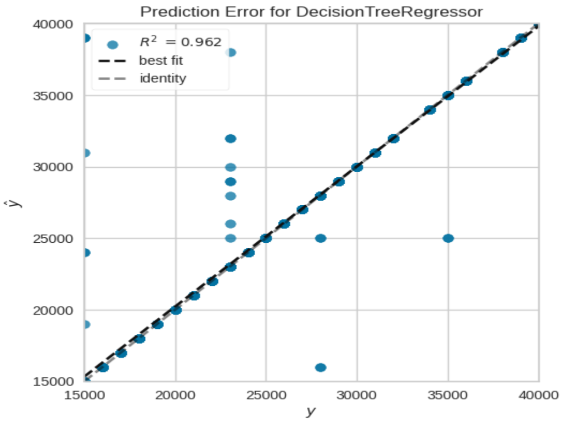

Fig. 18. Prediction Error Plot for Decision Tree Regressor

Figure 19 presents Individual Conditional Expectation (ICE) plots for the LightGBM model across nine key features. Each subplot illustrates the feature value on the x-axis and the corresponding model prediction on the y-axis, with multiple green lines representing individual instances. These visualizations reveal how predictions vary with changes in feature values, highlighting linear trends, threshold effects, and complex non-linear interactions.

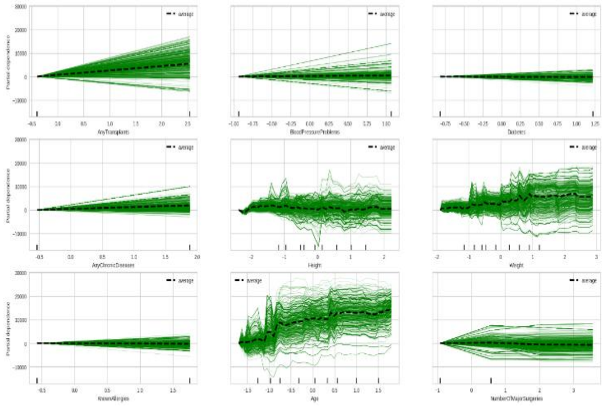

Fig. 19. ICE plot for LightGBM

Figure 20 presents Individual Conditional Expectation (ICE) plots for key features in the CatBoost model, showing how changes in each feature avgDT Crossing, Blood Pressure At Admission, and age impact predictions while keeping others constant. These plots reveal both the direction and variation of feature influence across individual instances.

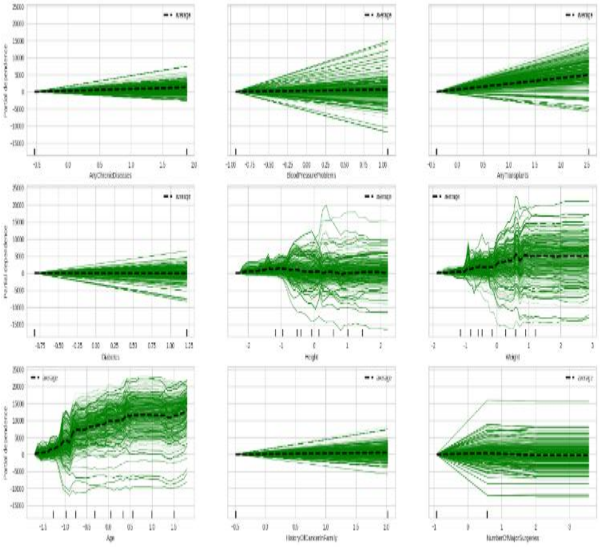

Fig. 20. ICE plot for CatBoost

Figure 21 displays ICE plots for the Decision Tree model, illustrating how individual predictions change as specific feature values vary. Each green line represents a single instance, while the dashed black line indicates the average trend. These plots provide insight into feature-level interactions and non-linear effects in the model's decision process.

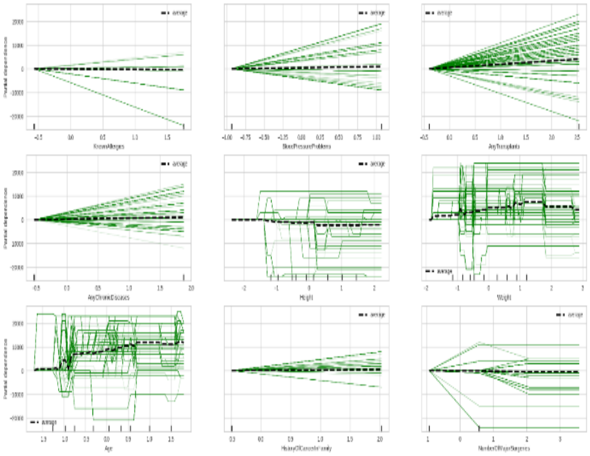

Fig. 21. ICE plots for the Decision Tree Model

A. *Comparison and Discussion*

Table III provides a performance evaluation of supervised ML models for predicting medical insurance costs reveals that the proposed regression models—LightGBM, CatBoost, and Decision Tree—significantly outperform the existing models, XGBoost and Random Forest. LightGBM emerges as the most accurate model with the highest R-square value of 98.67% and low error rates (MAE: 222.95, RMSE: 865.28, MAPE: 1.09%), demonstrating its strong generalization and predictive capability. CatBoost also performs well with a high R-square of 97.62%, though it has slightly higher error metrics compared to LightGBM. The Decision Tree model shows the lowest MAE (148.59) and MAPE (0.78%), indicating excellent precision in individual predictions, although its RMSE (1466.72) suggests some deviation in larger estimates. In contrast, XGBoost and Random Forest, the existing models, lag behind with significantly lower R-square values (82.78% and 82.25%, respectively) and much higher error values across all metrics, indicating weaker predictive performance. Overall, the suggested models and LightGBM in particular offer a better and more reliable way to forecast how much health insurance will cost.

TABLE III. PERFORMANCE EVALUATION OF SUPERVISED MACHINE LEARNING MODELS FOR MEDICAL INSURANCE

MODELS	R-square	MAE	RMSE	MAPE
LightGBM	98.67	222.9553	865.2828	1.0952
CatBoost	97.62	508.2767	1156.8298	2.1902
Decision Tree	96.18	148.5943	1466.7214	0.7812
XGBoost	82.78	1326.1650	2740.2238	5.5419
Random Forest	82.25	1287.8923	2782.1299	5.3162

The proposed advantage of this work lies in its comprehensive analysis and demonstration of the superior performance of ensemble-based boosting algorithms particularly LightGBM and CatBoost over traditional regression methods such as Decision Tree, XGBoost, and RF in forecasting medical insurance premium costs. The suggested models improve accuracy, resilience, and generalizability by attaining much higher R-square values and reduced error metrics. LightGBM, in particular, not only delivers the highest overall predictive performance but also demonstrates strong consistency across large-scale predictions, making it highly effective for real-world insurance cost estimation tasks. This highlights the potential of leveraging advanced boosting techniques to build more reliable and precise predictive models in the healthcare and insurance domains.

V. CONCLUSION AND FUTURE SCOPE

There is an increasing need for accurate estimates of medical insurance costs, making health insurance rate forecasting essential for customers and insurance firms alike. This research delves into the use of regression methods to forecast health insurance premiums via the integration of deliberate preprocessing, feature standardization, and class balancing techniques. The research shows that ensemble-based boosting models are good at forecasting the prices of medical insurance premiums, and that of the methods tested, LightGBM is the most reliable and accurate. Achieving an exceptional R-square accuracy of 98.67%, LightGBM not only demonstrated superior predictive performance but also maintained low error rates across MAE, RMSE, and MAPE metrics. Compared to traditional models like DT, XGBoost, and Random Forest, which showed significantly lower accuracy and higher errors, LightGBM's results underscore its capability for precise and efficient cost estimation. There are a few caveats to this research, albeit the encouraging findings. Since variables beyond the control of the model could affect the prices of medical insurance, its performance may fluctuate when applied to other datasets or in different geographical locations. The research also doesn't look at hybrid methods or deep learning strategies, which can improve prediction accuracy even more, instead concentrating on conventional supervised learning models. Future work could involve testing models on diverse datasets such as the HealthCare Cost or Insurance Dataset and implementing feature reduction techniques like PCA or RFE to improve model efficiency. Exploring advanced models such as DNNs or RNNs, along with incorporating additional factors like regional variations, could further enhance prediction accuracy and generalization.

REFERENCES

[1] S. Pandya, "A Machine and Deep Learning Framework for Robust Health Insurance Fraud Detection and Prevention," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 3, pp. 1332–1342, Jul. 2023, doi: 10.48175/IJARSCT-14000U.

[2] A. Shiwlani, S. Kumar, S. Kumar, S. U. Hasan, and M. H. A. Shah, "Transforming Healthcare Economics: Machine Learning Impact on Cost Effectiveness and Value-Based Care," *Pakistan J. Life Soc. Sci.*, vol. 22, no. 2, 2024, doi: 10.57239/PJLSS-2024-22.2.001494.

[3] I. Matloob, S. A. Khan, F. Hussain, W. Haider, R. Rukaiya, and F. Khalique, "Need-based and optimized health insurance package using clustering algorithm," *Appl. Sci.*, 2021, doi: 10.3390/app11188478.

[4] M. Shah, P. Shah, and S. Patil, "Secure and Efficient Fraud Detection Using Federated Learning and Distributed Search Databases," in *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*, 2025, pp. 1–6. doi: 10.1109/ICAIC63015.2025.10849280.

[5] G. Olaoye, "Comparative Study of Machine Learning Models for Predicting Health Insurance Costs," no. Ml, pp. 1–17, 2025.

[6] V. Kolluri, "AI for Personalized Medicine: Analyzing How AI Contributes to Tailoring Medical Treatment to the Individual Characteristics ff Each Patient," *IJRAR - Int. J. Res. Anal. Rev. (IJRAR), E-ISSN 2349-5138*, 2023.

[7] J. K. Chaudhary, S. Tyagi, H. P. Sharma, S. V. Akram, D. R. Sisodia, and D. Kapila, "Machine Learning Model-Based Financial Market Sentiment Prediction and Application," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, May 2023, pp. 1456–1459. doi: 10.1109/ICACITE57410.2023.10183344.

[8] Y. Han and R. Tarafdar, "Algorithms for the Majority Problem," in *Proceedings of the 2017 International Conference on Foundations of Computer Science*, 2017.

[9] I. Duncan, M. Loginov, and M. Ludkovski, "Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs," *North Am. Actuar. J.*, 2016, doi: 10.1080/10920277.2015.1110491.

[10] S. Pandya, "A Systematic Review of Blockchain Technology Use in Protecting and Maintaining Electronic Health Records," *Int.*

J. Res. Anal. Rev., vol. 8, no. 4, 2021.

[11] A. Garodia, S. Pahune, and N. Rewatkar, "Harnessing AI and Quantum Computing for Enhancing Supply Chain and Healthcare Operations: A Comprehensive Survey," *Int. Res. J. Eng. Technol.*, vol. 12, no. 3, pp. 723–729, 2025.

[12] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," *Mach. Learn. with Appl.*, 2024, doi: 10.1016/j.mlwa.2023.100516.

[13] V. Pillai, "Anomaly Detection Device for Financial and Insurance Data," *J. AI-Assisted Sci. Discov.*, vol. 4, no. 2, pp. 144–183, 2024.

[14] A. Immadisetty, "Real-Time Fraud Detection Using Streaming Data in Financial Transactions," *J. Recent TRENDS Comput. Sci. Eng.*, vol. 13, no. 1, pp. 66–76, Feb. 2025, doi: 10.70589/JRTCSE.2025.13.1.9.

[15] N. Malali, "Predictive AI for Identifying Lapse Risk in Life Insurance Policies: Using Machine Learning to Foresee and Mitigate Policyholder Attrition," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 5, no. 5, pp. 1–11, 2025.

[16] A. C. Reddy, M. T. Chowdary, and P. Renukadevi, "Medical insurance premium prediction using machine learning," in *AIP Conference Proceedings*, 2024, p. 020099. doi: 10.1063/5.0217289.

[17] M. Kulkarni, D. D. Meshram, B. Patil, R. More, M. Sharma, and P. Patange, "Medical Insurance Cost Prediction using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, 2022, doi: 10.22214/ijraset.2022.47923.

[18] N. Shakhovska, N. Melnykova, V. Chopiyak, and M. Gregus Ml, "An ensemble methods for medical insurance costs prediction task," *Comput. Mater. Contin.*, 2022, doi: 10.32604/cmc.2022.019882.

[19] A. A. Hira Zainab, Ali Raza A Khan, Muhammad Ismaeel Khan, "Innovative AI Solutions for Mental Health: Bridging Detection and Therapy," *Glob. J. Emerg. AI Comput.*, vol. 1, no. 1, pp. 51–58, 2025.

[20] S. S. S. Neeli, "The Convergence of AI and Database Administration in Revolutionizing Healthcare," *ESP Int. J. Adv. Comput. Technol.*, vol. 2, no. 4, pp. 150–153, 2024, doi: 10.56472/25838628/IJACT-V2I4P119.

[21] A. Polleri, R. Kumar, M. M. Bron, G. Chen, S. Agrawal, and R. S. Buchheim, "Identifying a classification hierarchy using a trained machine learning pipeline," 2022

[22] S. R. Thota and S. Arora, "Neurosymbolic AI for Explainable Recommendations in Frontend UI Design - Bridging the Gap between Data-Driven and Rule-Based Approaches," vol. 11, no. 5, pp. 766–775, 2024.

[23] V. Pillai, "System And Method For Intelligent Detection And Notification Of Anomalies In Financial And Insurance Data Using Machine Learning," 202421099024, 2025

[24] S. Mohan, S. Sharma, S. Agrawal, and S. Kamatchi, "Optimization of Insurance Claim Cost Prediction Through Health Data and Machine Learning," in *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, Jan. 2025, pp. 1–7. doi: 10.1109/ICAECT63952.2025.10958928.

[25] A. Sharma and R. Jeya, "Prediction of Insurance Cost through ML Structured Algorithm," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, IEEE, Feb. 2024, pp. 495–500. doi: 10.1109/IC2PCT60090.2024.10486304.

[26] S. Vuddanti, V. G. S. K. R. K. Jamili, S. R. Bommareddy, V. Rotta, and V. Pagadala, "Machine Learning Insights into Personalized Insurance Pricing," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, IEEE, Jul. 2024, pp. 923–927. doi: 10.1109/ICSCSS60660.2024.10625562.

[27] A. R. Kandula, S. Kalyanapu, S. N. Rayapalli, K. Rao Veerabathina, V. Modugumudi, and S. R. Kanikella, "Medical Insurance Predictive Modelling: An Analysis of Machine Learning Methods," in *2024 IEEE International*

Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), IEEE, Mar. 2024, pp. 1–5. doi: 10.1109/IATMSI60426.2024.10502643.

[28] R. Mary Chittilappilly, S. Suresh, and S. Shanmugam, “A Comparative Analysis of Optimizing Medical Insurance Prediction Using Genetic Algorithm and Other Machine Learning Algorithms,” in *Proceedings of the 2nd IEEE International Conference on Advances in Computing, Communication and Applied Informatics, ACCAI 2023*, 2023. doi: 10.1109/ACCAI58221.2023.10199979.

[29] T. T, S. H. T, V. K. V, and K. R, “Medical Insurance Cost Analysis and Prediction using Machine Learning,” in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, Mar. 2023, pp. 113–117. doi: 10.1109/ICIDCA56705.2023.10100057.

[30] V. Vijayalakshmi, A. Selvakumar, and K. Panimalar, “Implementation of Medical Insurance Price Prediction System using Regression Algorithms,” in *Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023*, 2023. doi: 10.1109/ICSSIT55814.2023.10060926.

[31] R. Szczepanek, “Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost,” *Hydrology*, vol. 9, no. 12, p. 226, Dec. 2022, doi: 10.3390/hydrology9120226.

[32] H. S. Chandu, “Efficient Machine Learning Approaches for Energy Optimization in Smart Grid Systems,” *Ijsart*, vol. 10, no. 9, 2024.

[33] M. M. Rahman *et al.*, “Empowering early detection: A web-based machine learning approach for PCOS prediction,” *Informatics Med. Unlocked*, vol. 47, no. January, pp. 1–16, 2024, doi: 10.1016/j.imu.2024.101500.

[34] V. S. Thokala, “Integrating Machine Learning into Web Applications for Personalized Content Delivery using Python,” *Int. J. Curr. Eng. Technol.*, vol. 11, no. 06, 2021, doi: https://doi.org/10.14741/ijcet/v.11.6.9.