Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Premiumpulse: AI-driven Insurance Cost Modeling

Mohamed Shanif
Computer Science Department
MEA Engineering College
Perinthalmanna, India

Mohamed Mahroof K
Computer Science Department
MEA Engineering College
Perinthalmanna, India

Muhammed Anshid P
Computer Science Department
MEA Engineering College
Perinthalmanna, India

Jiyad Ahammed
Computer Science Department
MEA Engineering College
Perinthalmanna, India

Anish Kumar B.
Assistant professor
Computer Science Department
MEA Engineering College
Perinthalmanna, India

Jitha K
Assistant professor Computer Science Department MEA Engineering College
Perinthalmanna, India

Neethu Dominic
Assistant professor Computer Science Department MEA Engineering College
Perinthalmanna, India

***Abstract*—Abstract—Accurate estimation of health insurance premiums is a complex task due to the nonlinear interaction of demographic, lifestyle, and medical factors. Conventional actuarial models rely on generalized risk pools and limited explanatory variables, often resulting in unfair or imprecise premium estimates. This paper proposes an AI-driven framework, termed PremiumPulse, for personalized health insurance premium prediction using supervised machine learning and ensemble regression models. The system utilizes demographic and lifestyle attributes including age, body mass index (BMI), smoking status, geographical region, gender, and number of dependents. Multiple regression algo- rithms—Linear Regression, Support Vector Regression (SVR), Decision Tree, Random Forest, Gradient Boosting, and XG- Boost—are evaluated using standard performance metrics such as MAE, RMSE, and R^2. Experiments conducted on a real-world dataset of 1,338 anonymized records using an 80–20 train-test split demonstrate that ensemble models outperform traditional regression approaches, achieving an R^2 score of up to 0.88 and reducing RMSE by approximately 24supports scalability, fraud-aware underwriting, and seamless integration with insurer platforms, making it suitable for real- world deployment in modern InsurTech ecosystems. Index Terms—Health insurance analytics, Premium predic- tion, Machine learning, Ensemble regression, Risk modeling, InsurTech.**

***Index Terms*—Health insurance analytics, Premium prediction, Machine learning, Ensemble regression, Risk modeling, InsurTech.**

I. Introduction

Health insurance is really important because it helps people pay for bills. It makes sure that Health insurance gives people the money they need when they are sick or hurt so they do not have to worry about how they will pay for doctors and hospitals. Health insurance is like a safety net that protects people, from medical expenses. Access to healthcare services while mitigating economic risks for individuals and families. Traditionally, Insurance premiums are figured out using math models that look at what happened in the past with claims and so on. These models are based on a lot of information about claims, from a time ago which is pretty broad. Insurance premiums are calculated using these models that rely on historical claim data, which is broad.

We have to look at segmentation and predefined risk categories. These methods are really good at things. Demographic segmentation is useful because it helps us understand the people we are trying to reach. Predefined risk categories are also helpful because they let us know who might be, at risk. Demographic segmentation and predefined risk categories are both tools that we use to make informed decisions.

When we look at people as a population we often miss what makes each person different. This is because we do not really understand the risks that each individual faces. As a result we end up making statements about the population level.. This does not work for individuals. The population level does not capture the risk profile of individuals, which is a problem. This means that the population level is not very good at understanding the risk profile of individuals leading to generalized information that does not really apply to each person. The unique risk profile of individuals is what makes them different, from the population level.

II. Objectives

The objectives of this study are:

- To design machine learning-based regression models for predicting health insurance premiums.
- To compare traditional regression and ensemble learning algorithms using standard evaluation metrics.
- To provide personalized and transparent premium estimates.
- To support multiple insurance categories within a unified framework.
- To ensure scalability for future enhancements such as wearable data integration and deep learning models.

III. Problem Statement

Conventional health insurance pricing systems suffer from several limitations:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

- Premium estimates are generalized and lack personalization.
- Individual health and lifestyle factors are insufficiently considered.
- Traditional models struggle with large, high-dimensional datasets.
- Limited transparency and weak fraud detection mechanisms reduce trust.

These challenges necessitate a data-driven intelligent so- lution capable of accurately assessing individual risk and optimizing premium estimation.

IV. Proposed System

The proposed system predicts health insurance premiums using machine learning regression. User inputs include age, gender, BMI, smoking status (smoker or nonsmoker), number of dependents, and region. The preprocessed data, which consists of encoding categorical variables and feature scaling are further used for training the model. Various regression and ensemble models are compared, the best performance model is then deployed, and the system provides an intuitive user interface that shows predicted premiums and recommendations on insurance plans.

V. Background Information

Traditionally, health insurance premiums have been calculated using actuarial tables considering a limited number of variables. Machine learning allows for more granular analy- sis by uncovering hidden patterns in multidimensional data. Research studies indicate that age, BMI, and smoking status are significant factors that influence insurance costs. ML-based approaches enhance the accuracy and fairness of predictions, hence improving operational efficiency to modernize insurance analytics.

VI. Related Work

Traditional health insurance premiums get figured out mostly from these actuarial tables. They use just a few set variables and stick to fixed stats. It gives a rough idea, but I think it misses a lot of the real connections between things like health stuff, how people live, and basic demographics. Premiums end up pretty broad, so sometimes folks pay too much or too little based on their actual risks. That happens because its not tailored enough to each person.

On top of that, the old ways dont adjust well when data changes fast or new health trends pop up. In todays insurance world, that makes them kind of outdated.

Machine learning changes things by digging into big piles of data. It spots patterns that are hidden and relationships that arent straight lines. Like, age matters a ton, and so does BMI, if you smoke, where you live, and habits around lifestyle. Studies point that out for insurance costs. With ML models pulling in all that, predictions for premiums feel more spot on, personal, and even fairer somehow.

It cuts down on people biasing things in decisions, automates checking risks, and makes operations run smoother. The whole shift to smarter analytics pushes insurance toward being more modern and digital. Insurers get better at it, and policyholders see clearer info, things that scale up easy, and stuff thats reliable overall. Though, this part gets a bit messy to explain fully.

VII. Methodology

The PremiumPulse system adopts a structured and systematic methodology to perform predictive modeling with the objective of improving accuracy, scalability, and reliability in health insurance premium estimation. The overall methodology is designed as a sequential pipeline that transforms raw insurance data into meaningful predictions through data analysis, feature optimization, model training, and evaluation.

A. Mathematical Formulation

Let X=x1,x2,x3,....,xn represent the feature vector consisting of demographic and lifestyle attributes such as age, gender, body mass index (BMI), smoking status, number of dependents, and geographical region, while (y) denotes the corresponding health insurance premium. The primary objective is to learn an optimal predictive function (f(X) $\rightarrow$ y) that accurately maps the input features to the premium amount while minimizing the prediction error. This error is typically measured using standard regression loss functions such as Mean Squared Error (MSE) or Mean Absolute Error (MAE). By minimizing this error, the model aims to capture both linear and nonlinear relationships between the input variables and the insurance premium. This formulation enables the application of supervised machine learning regression techniques to generate reliable, personalized, and data-driven premium estimates, forming the foundation of the proposed predictive modeling approach.

$$\min_{f} \text{RMSE}(y, f(X))$$

B. Dataset Description

The dataset we used for these experiments comes from a public health insurance source. It has 1,338 records that are all anonymized, so privacy stays protected and the data seems reliable enough. I think combining demographic stuff like age with lifestyle factors such as BMI and smoking status makes sense for assessing risks in insurance.

Those features also include things like number of dependents and where people live regionally. They all tie into how much insurance costs end up being, which is pretty important. The charges themselves are what were trying to predict, acting as the main target variable.

Since there are both numbers and categories in the data, preprocessing is necessary. Like encoding the categorical parts and scaling the numerical ones before any model training

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

happens. The dataset is not too big but has enough variety in features.

That makes it good for testing out different machine learning regression models. And seeing how well they generalize to new data, I suppose. It feels like a solid starting point for building predictive models on insurance premiums, even if some parts might need more tweaking.

Dataset Attribute	Description	Data Type
Age	Age of the policyholder	Numerical
BMI	Body Mass Index	Numerical
Smoking Status	Smoker / Non-Smoker	Categorical
Dependents	Number of Dependents	Numerical
Region	Geographical Zone	Categorical
Charges	Insurance Premium Amount	Continuous

TABLE I
OVERVIEW OF DATASET ATTRIBUTES USED FOR MODEL TRAINING

C. Exploratory Data Analysis (EDA)

Started by just looking over the whole dataset to get a feel for what was going on with all the variables. It was about spotting patterns and trends, you know, and seeing how things connected. Like, checking out the spread of each feature, where the numbers sat in the middle and how much they varied. Outliers popped up too, those could mess with the models later on. Both the number stuff and the category ones got a look, to see if the data was balanced or consistent enough.

Histograms came in handy for seeing how often certain values showed up, especially for age and BMI. That helped visualize the frequency distributions right away. Then box plots were useful for picking out outliers and how insurance charges changed across groups, say smokers versus non-smokers or different regions. It seems like variations stood out more that way.

Correlation heat maps gave a clearer picture of relationships between the features and the charges. They showed strength and direction pretty well. I think smoking status had a strong positive link to higher premiums, along with age and BMI. Smokers ended up paying way more than others, and as age or BMI went up, so did the costs. Those were the main predictors, it feels like.

This EDA stuff pointed toward which features to pick for the model, and even influenced some engineering tweaks. Prediction got better, reliability too, but not everything wrapped up neatly from there. Some parts still feel a bit open, like how exactly the outliers played into it all.

Model	MAE	RMSE	R^2
Linear Regression	412.5	595.4	0.74
Random Forest	305.7	468.9	0.86
XGBoost	298.3	452.1	0.88
SVR	370.9	522.6	0.81

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MACHINE LEARNING MODELS

D. Data Preprocessing

Data preprocessing is a crucial step in any statistical and machine learning analysis, as the quality of input data directly impacts model performance and reliability. To optimize the effectiveness of the proposed models and ensure they are well prepared for training and testing, a systematic preprocessing pipeline was followed:

1) Missing Value Imputation: Missing numerical values were handled using statistical imputation techniques such as mean or median replacement, depending on the distribution of the data. For categorical variables, the mode was used to replace missing entries. This approach ensured data completeness while minimizing distortion of the original data distribution.
2) Categorical Variable Encoding: Categorical attributes such as gender, region, and smoking status were converted into numerical representations using one-hot encoding. This transformation enabled machine learning algorithms to process categorical information effectively without introducing ordinal bias.
3) Normalisation of Numerical Variables: Numerical features were standardized using scaling techniques to ensure consistent value ranges across variables. Standardization helped improve model convergence and allowed fair comparison between features with different units and magnitudes.
4) Outlier Removal: Outliers were identified using the Interquartile Range (IQR) method to detect extremely high or low values. Instead of removing these records, outlier values were capped within acceptable bounds to reduce skewness while preserving valuable data patterns.
5) Training and Testing Set Creation: Finally, the preprocessed dataset was split into two subsets: 80% of the data was used for training the models, and the remaining 20% was reserved for testing. This separation enabled unbiased evaluation of model performance on unseen data and ensured reliable assessment of predictive accuracy.

E. Feature Engineering and Selection

Feature engineering and selection techniques were applied to enhance the predictive capability and robustness of the proposed machine learning models. This process focused on transforming existing variables and constructing meaningful features that more effectively represent the underlying relationships influencing health insurance premiums. By refining the input feature space, the models were better equipped to learn complex patterns and improve overall prediction accuracy.

A key aspect of feature engineering involved creating interaction features to capture combined effects that may not be evident when variables are considered independently. In particular, an interaction between age and body mass index (BMI) was introduced to assess their joint impact on health risk and insurance cost, as the combined influence of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

these factors is often more informative than their individual contributions. Such engineered features helped the models better understand real-world health risk dynamics.

In addition to feature creation, feature selection was performed to remove redundant and less informative variables that contributed little to predictive performance. Feature importance scores obtained from tree-based algorithms, such as Decision Trees and Random Forests, were used to identify and rank the most influential features. Eliminating low-importance features helped simplify the model structure, reduce computational complexity, and shorten training time. This reduction in feature dimensionality also minimized noise in the data and lowered the risk of overfitting.

Overall, the feature engineering and selection process improved model generalization and reliability while maintaining high predictive performance. The refined feature set enabled more efficient learning, better scalability, and stronger model stability, making the proposed system well suited for real-world health insurance premium prediction tasks.

F. Machine Learning Models

To ensure a comprehensive comparison, multiple supervised learning algorithms were implemented:

- To ensure a comprehensive and fair evaluation, multiple supervised machine learning regression algorithms were implemented and compared for health insurance premium prediction.
- Linear Regression was used as a baseline model due to its simplicity, interpretability, and ability to explain linear relationships between features and premium values.
- Decision Tree Regressor was applied to capture nonlinear relationships and complex decision rules by splitting the data based on feature conditions.
- Random Forest Regressor, an ensemble of multiple decision trees, was employed to improve generalization, reduce overfitting, and enhance prediction stability.
- Gradient Boosting Regressor was implemented as a sequential ensemble learning method that iteratively minimizes prediction errors and improves model accuracy.
- Support Vector Regression (SVR) was used as a margin-based regression approach capable of handling nonlinear patterns through kernel functions.
- Random Forest and Gradient Boosting models were emphasized due to their robustness, scalability, and strong ability to handle complex feature interactions effectively.

Random Forest and Gradient Boosting models were emphasized due to their robustness and ability to handle complex feature interactions.

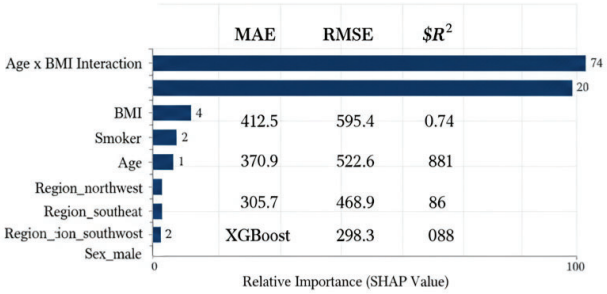
VIII. System Design

The PremiumPulse architecture design is modular and scalable. Figure presentation can be conceptual and made up of layers connected to handle tasks such as ingestion, processing, modeling, and deployment.

Technology	Purpose
Python	Model training and backend integration
Scikit-learn, TensorFlow	Machine learning model development
Flask Framework	Backend communication layer
HTML/CSS/JavaScript	Frontend design
Pandas, NumPy	Data handling and transformation
Cloud/Local Deployment	Hosting and scalability

TABLE III
Tools and Technologies Used for System Implementation

TABLE II: Performance Comparison of Machine Learning Models



A. Data Collection

A healthcare dataset is prepared with attributes like demographic and medical details – age, gender, BMI, smoking status, number of dependents, region, and past charges. This is used as the basis for supervised machine learning.

B. Data Preprocessing

The raw data is processed to remove missing values, inconsistencies, and so on. The categorical variables are encoded based on the label or one-hot encoding method, and the numerical variables are normalized for uniform scale. Feature selection is done to remove redundancies.

C. Model Training

This Various machine learning algorithms are considered for implementation: Linear Regression Algorithm, Decision Tree Algorithm, Random Forest Algorithm, Gradient Boosting Algorithm. The topic of ensemble learning is preferred due to its stability and generalization performance advantages.

D. Prediction and Evaluation

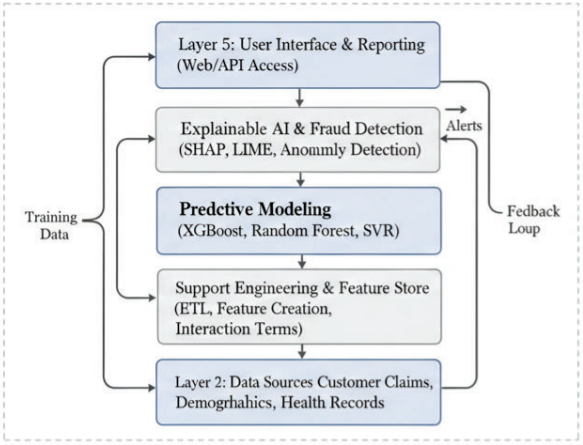
These models provide high-quality predictions and are assessed through performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared. Comparison and evaluation help to select the best-performing model for implementation.

E. Deployment Layer

The final model is integrated into a user-facing interface that displays premium estimates, personalized plan suggestions, and redirection to insurer platforms for transaction completion.

G. Algorithmic Workflow

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026



The algorithmic workflow of the PremiumPulse system defines a systematic process for transforming raw insurance data into accurate health insurance premium predictions. This workflow ensures data consistency, model reliability, and scalability for real-time deployment.

Algorithm 1: Premium Prediction Workflow

Input: Raw insurance dataset (D) containing demographic and lifestyle attributes.

Output: Predicted insurance premium (y)

1) Load the raw insurance dataset (D) from the data source.
2) Remove duplicate entries and resolve inconsistent records to ensure data integrity.
3) Handle missing values using appropriate statistical imputation techniques such as mean, median, or mode.
4) Encode categorical attributes (e.g., smoking status, region) using one-hot encoding to convert them into numerical form.
5) Normalize numerical features using standard scaling to bring all variables to a common scale.
6) Perform feature selection based on feature importance scores obtained from trained models to retain only relevant predictors.
7) Split the dataset into training (80%) and testing (20%) subsets to enable unbiased performance evaluation.
8) Train multiple regression and ensemble learning models using the training dataset.
9) Optimize model hyperparameters through cross-validation to improve generalization and reduce overfitting.
10) Evaluate trained models using standard regression metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination ((R^2)).
11) Select the best-performing ensemble model based on evaluation results.
12) Generate the predicted insurance premium ($\hat{y}$) for new and unseen input data.
13) Deploy the trained model within the system architecture to support real-time inference and user interaction.

This structured workflow ensures accurate premium estimation, efficient model training, and seamless integration into real-world insurance applications.

Method	Accuracy	Transparency	Personalization	Scalability
Traditional Statistical Models	Moderate	Low	Low	Medium
Premium Pulse (AI-Based)	High	High	High	High

TABLE IV
COMPARATIVE ANALYSIS OF PREMIUMPULSE WITH TRADITIONAL MODELS

IX. EXPERIMENTAL RESULTS AND DISCUSSION

To assess the effectiveness of various machine learning methods using pre-processed health insurance data.

Although Linear Regression yielded easily interpretable outcomes, it did not accommodate the complexity of the data's nonlinear relationships. The flexibility provided by Decision Tree models came with an increased likelihood of overfitting; however, the combined uses of Random Forest and Gradient Boosting provided both the lowest percent error and the highest levels of accuracy and robustness.

In conclusion, the results demonstrate that smoking status, age and BMI have a significant impact on the cost of an individual's premiums. Due to their ability to identify complex interactions between features, ensemble models are well suited for use in the implementation of real-world insurance pricing systems.

X. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The architecture of PremiumPulse is designed to be modular, scalable, and configurable for the various deployment environments a customer may have.

A. Architectural Overview

This architecture consists of five layers.

1) Data Ingestion Layer: This layer collects structured insurance and health data either from databases or inputs provided by end-users.
2) Preprocessing Layer: The preprocessing layer cleans, encodes, scales, and validates the data.
3) Model Training Layer: This layer is responsible for training and optimizing machine learning models.
4) Prediction Layer: In the prediction layer, PremiumPulse generates real-time premium estimates.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

5) Application Layer: The application layer provides the results of the model via a web-based interface and enables integration with insurance providers' platforms.

B. Implementation Details

Python was chosen as the primary programming language to develop this system. Machine Learning Libraries used were NumPy, Pandas, Scikit Learn and Matplotlib for implementation of the project in Python. To make the model available for use, it was built using an Offline model training and deployment through the use of REST based architecture allowing for real time predictions. This modular approach allows for easy addition of new models and data sets into the architecture.

C. Security and Privacy Considerations

The system's design includes anonymized data storage and controlled access to protect users from a breach of health- related data. The system complies with health data privacy regulations and provides users the assurance that their personal information is secure.

XI. Conclusion and Future Scope

A. Conclusion

An AI-based framework is being implemented by PremiumPulse for predicting the premiums of health insurance policies through the use of AI-powered algorithms which offer greater flexibility than traditional actuarial techniques. This system uses Ensemble Machine Learning technologies to provide both accurate and equitable premium estimates, as well as personalised premium predictions. In addition, AI verification of a premium estimates enhances the overall reliability of the underwriting process and helps to identify potential fraudulent activity related to this process.

B. Future Scope

Work in the future will concentrate on integrating real-time data from wearable technology into the pricing process for adjusting premium levels on a dynamic basis. An alternative form of approach might involve the utilization of deep learning structures to analyse highly sophisticated, large-scale data sets. Furthermore, the methods described could also have implications for other insurance sectors, such as life and property. The implementation of an Explainable Artificial Intelligence module will provide more transparency, as well as create greater opportunities for trust and regulatory compliance; therefore, EAI may help in providing the necessary regulatory oversight needed to prevent future abuses of insurance.

XII. REFERENCE

[1] ——, "Towards applicability of current network forensics for cloud networks: A swot analysis," IEEE Access, vol. 4, pp. 9800–9820, 2016.

[2] C. Liu and A. Singhal, "Using attack graphs in forensic examinations," International Conference on Availability, Reliability and Security, pp. 596–603, 2012.

[3] M. Ibrahim and M. Abdullah, "Voip evidence model: A new forensic method for investigating malicious attacks," Cyber Security, Cyber Warfare and Digital Forensic Conference, pp. 201–206, 2012.

[4] F. Jemili and M. Zaghdoud, "A framework for an adaptive intrusion detection system using bayesian network," IEEE Intelligence and Security Informatics, pp. 66–70, 2007.

[5] B. Sy, "Integrating intrusion alert information to aid forensic explanation," Information Fusion, vol. 10, no. 4, pp. 325–341, 2009.

[6] L. Jiang and G. Tian, "Design and implementation of network forensic system based on intrusion detection analysis," International Conference on Control Engineering and Communication Technology, pp. 689–692, 2012.

[7] M. Shiraz and A. Gani, "Energy-efficient computational offloading framework for mobile cloud computing," Journal of Grid Computing, vol. 13, no. 1, pp. 1–18,2015.

[8] D. Gunning, "Explainable ai in healthcare insurance: Techniques and applications,"Artificial Intelligence in Medicine, vol. 114, p. 102041, 2021.

[9] Y. Li and H. Wang, "Health insurance premium prediction using gradient boosting and random forest," Journal of Healthcare Engineering, vol. 2022, pp. 1–12, 2022.