IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

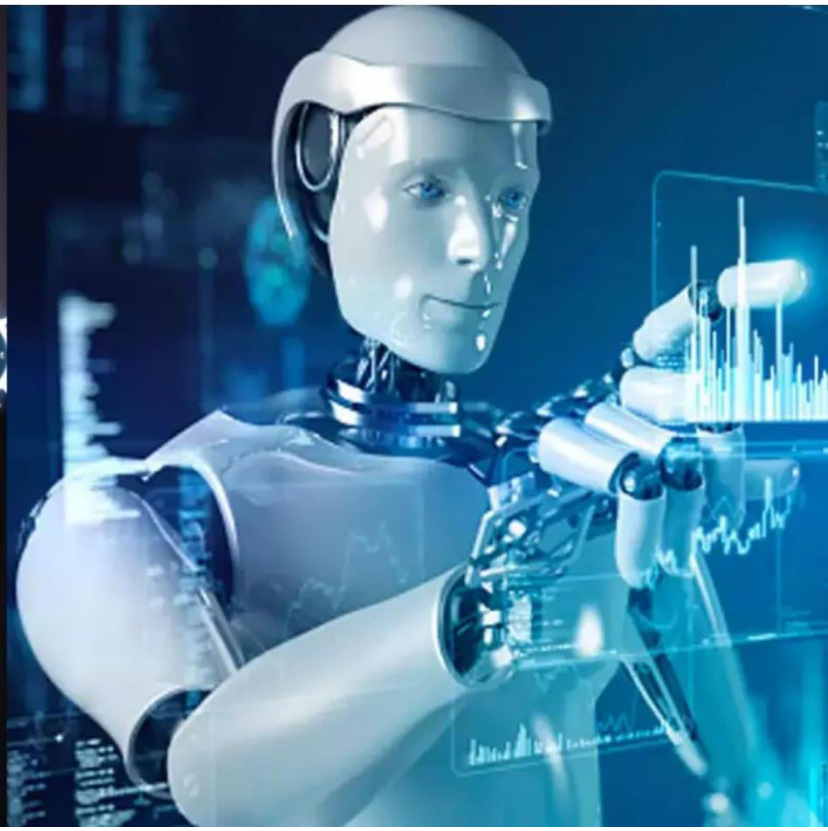
ISSN(O): 2319-8753
ISSN(P): 2347-6710



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 9.935

Volume 15, Issue 10, October 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

AI-Based Health Insurance Premium Prediction and Recommendation System using Random Forest Regression and SHAP Explainable AI

Dr. S. Subashini[1], Dharsan B[2], Abishek D[2]

Associate Professor (Supervisor), Department of Computer Science and Engineering, K.L.N. College of Engineering (An Autonomous Institution, Affiliated to Anna University, Chennai), Tamil Nadu, India[1]

UG Scholars, Department of Computer Science and Engineering, K.L.N. College of Engineering (An Autonomous Institution, Affiliated to Anna University, Chennai), Tamil Nadu, India [2]

ABSTRACT: Health insurance premiums depend on demographic, physical and lifestyle factors, yet most applicants receive a quotation with no understandable reason for the amount and no guidance on which plan suits them. This paper presents an AI-based health insurance premium prediction and recommendation system that turns a premium estimate into a transparent, personalised decision-support experience. User details such as age, gender, Body Mass Index (BMI), smoking status, medical history, region and number of children are cleaned, encoded and reduced to the most influential features before being passed to a Random Forest Regression model that predicts the annual premium. SHAP (SHapley Additive exPlanations) identifies the factors that push the premium up or down, and a Health Risk Score summarises the overall insurance risk. The system then recommends the Top 3 suitable plans, answers insurance queries through an AI chatbot, and generates a downloadable PDF report. The application uses a React.js and Tailwind CSS front end, a Python Flask/FastAPI back end, MongoDB, and Scikit-learn, Pandas, NumPy and SHAP for the machine learning pipeline. The Random Forest model achieves an R^2 above 0.94, and the system delivers prediction, explanation, risk assessment, recommendation, chatbot assistance and reporting on a single platform.

KEYWORDS: Health Insurance, Premium Prediction, Random Forest Regression, SHAP, Explainable AI, Health Risk Score, Plan Recommendation, AI Chatbot

I. INTRODUCTION

Health insurance is a basic financial safeguard, because the cost of hospitalisation, diagnostics and long-term treatment can quickly exceed household savings. The premium an insurer charges reflects the expected medical cost of the insured person, which depends on age, gender, BMI, smoking habits, existing medical conditions, region and number of dependents. Traditionally, actuaries combine these factors using fixed rate tables, and the applicant sees only the final quotation.

Artificial Intelligence (AI) and Machine Learning (ML) can learn the relationship between personal attributes and the premium directly from historical data, and can estimate a premium instantly for combinations never listed in a rate table. A prediction that cannot be explained, however, is hard to trust, and choosing among the many plans in the market is difficult without expert help. Existing ML work tends to solve only one of these problems at a time.

This paper presents an integrated platform in which a Random Forest model predicts the annual premium, SHAP explains the prediction, a Health Risk Score summarises the risk level, and the Top 3 suitable plans are recommended. An AI chatbot and a downloadable PDF report complete the workflow. The main contributions are:

- A Random Forest Regression model that predicts the annual premium from health, demographic and lifestyle inputs.
- SHAP-based explanations that express each factor's contribution as a monetary amount raising or lowering the premium.
- A Health Risk Score with a risk tier and factor breakdown, and a Top 3 plan recommendation based on the profile and predicted premium.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

- An AI chatbot and PDF report in a single responsive web application.

II. LITERATURE REVIEW

Jyothsna et al. [1] predicted premiums using Multi-Linear, Decision Tree, Random Forest and Gradient Boosting regression on age, sex, BMI, children, smoking status and region, evaluated with MAE, MSE, RMSE and R^2, and added a Telegram chatbot. Their work does not provide plan recommendation, explanation or risk scoring. Kandula et al. [2] compared regression and ensemble methods for medical insurance cost prediction, again without recommendation, risk scoring or chatbot support.

Zhou and Su [3] used collaborative filtering on customer information and online events to recommend insurance products and to reduce cold-start and data-sparsity problems, but without premium prediction or explainability. Breiman [4] introduced Random Forests, which average many decision trees grown on bootstrap samples and handle non-linear interactions well, although an individual prediction remains a black box. Lundberg and Lee [5] proposed SHAP, which assigns each feature an additive Shapley-value contribution to a prediction and has efficient algorithms for tree ensembles.

The literature shows that premium prediction and product recommendation are well studied but separately, and that explainable AI is rarely attached to insurance pricing. The proposed system combines all of them in one workflow.

III. EXISTING AND PROPOSED SYSTEM

A. Existing System

Prediction-oriented systems train regression and ensemble models on demographic and lifestyle attributes and return a number without explanation or plan guidance. Recommendation-oriented systems use content-based filtering, KNN, cosine similarity or collaborative filtering to suggest plans but do not estimate the premium from the user's own profile. Because the two tasks are handled by different tools, the user must combine results manually and has no in-platform help for queries.

B. Proposed System

The proposed system integrates prediction and recommendation on one platform. Inputs are pre-processed, a Random Forest predicts the annual premium, SHAP shows the major influencing factors, and a Health Risk Score is generated. The system recommends the Top 3 plans from the profile and predicted premium, supports insurance queries through a chatbot, and summarises everything in a PDF report. Its advantages are that it is integrated, transparent, personalised, assistive and documented.

IV. METHODOLOGY

A. System Architecture

The architecture is a pipeline from user details to a personalised result (Fig. 1). Inputs are pre-processed and passed through feature selection to the Random Forest model. The prediction is explained by SHAP, a risk score is calculated, and the recommendation module selects the Top 3 plans. The outputs are combined into one result from which the user can consult the chatbot or download the PDF report. MongoDB stores user data, results and plan data.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

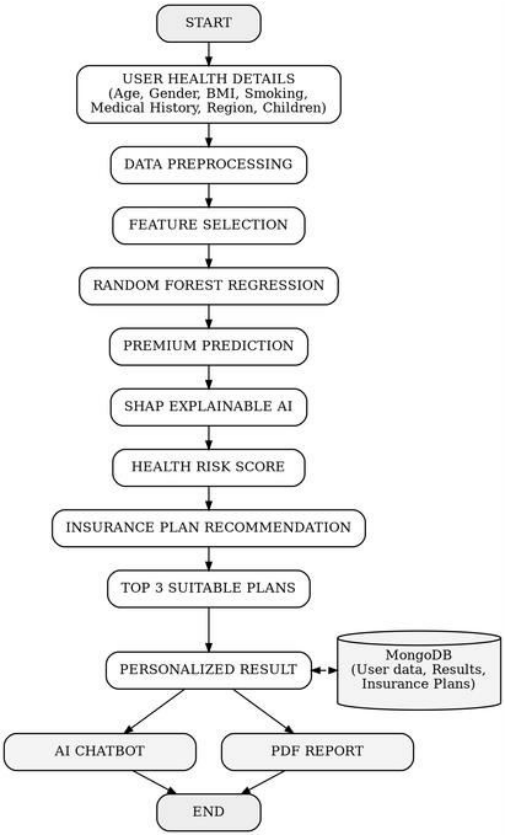


Fig.1. System architecture diagram

B. Data Collection and Pre-processing

The system collects age, gender, BMI, smoking status, medical history, region and number of children. In the interface, height and weight are entered and the BMI and weight category are computed automatically; lifestyle level and sum insured can also be chosen. Incomplete records are removed, categorical attributes such as gender, smoking status and region are encoded numerically, the most influential features are selected, and the data are split 80/20 into training and testing sets (Fig. 2).

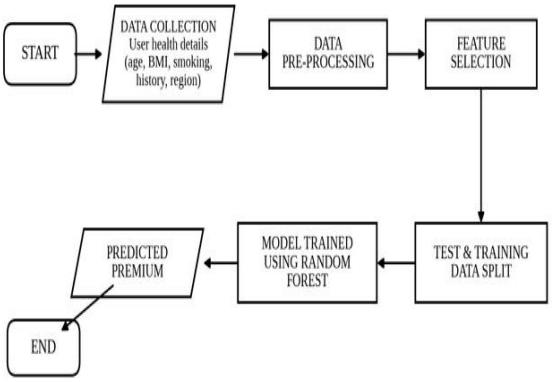


Fig.2. Data flow diagram

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

C. Premium Prediction Using Random Forest

Random Forest Regression builds an ensemble of decision trees on bootstrap samples with random feature subsets and averages their outputs. This captures non-linear effects such as the sharp rise in expected cost for smokers with a high BMI while resisting over-fitting. The model was trained with 200 trees and evaluated on unseen records using MAE, RMSE and R^2. The trained model is saved and loaded once at back-end start-up so each request is answered quickly. The interface presents the annual premium, an expected range and the monthly instalment.

D. SHAP Explainable AI

A SHAP tree explainer, created once at start-up, computes the contribution of each feature to the difference between the average premium and the user's predicted premium. The contributions are shown as bars with their monetary impact and restated in plain-language insights.

E. Health Risk Score

The score (0–100) combines four components: BMI (up to 30 points), smoking risk (30), age factor (20) and medical history (20). It is shown on a gauge with a risk tier and a factor scorecard, together with short advice cards.

F. Plan Recommendation and Chatbot

Plans are ranked by a fit score that rewards a high claim settlement ratio and a large hospital network and penalises the relative gap between the plan premium and the predicted premium. The three highest-ranked plans are displayed with insurer, premium, network hospitals, claim settlement ratio, waiting period, room rent and co-payment terms. The AI chatbot answers questions on premiums, policy features, deductibles and ways to reduce the premium, and the PDF report collects all results.

V. IMPLEMENTATION

The front end uses HTML, CSS, JavaScript, React.js and Tailwind CSS in reusable components (profile form, result cards, SHAP panel, chatbot window). The Python back end (Flask/FastAPI) exposes prediction, explanation, recommendation, chatbot and report features as REST APIs. MongoDB stores user details, predictions and plan data. Model training uses Scikit-learn, Pandas and NumPy, and explanations use the SHAP library. Development used Jupyter Notebook/Google Colab, VS Code, Git and Postman on Windows 10/11, with a minimum configuration of an Intel Core i5 (8th generation) processor, 8 GB RAM and a 256 GB SSD.

VI. RESULTS AND DISCUSSION

The system was executed end to end through its web interface to verify each feature. The Random Forest model achieves an R^2 above 0.94, indicating that it explains most of the variation in the premium. The sample applicant has a BMI of 23.0 kg/m² (Normal); four preset profiles (Young & Healthy, Balanced Adult, Active Smoker, Senior with Hypertension) allow quick testing.

A. Premium Prediction and Recommendation

After prediction, a summary strip shows the annual premium with its range, the monthly instalment, the risk score with its tier and the best matching plan (Fig. 3). For the sample profile the best fit is the Optima Secure Policy of HDFC ERGO General Insurance at ₹6,622.24 per year, followed by the Care Supreme Health Plan of Care Health Insurance at ₹6,081.65 and the Medicare Premier of Tata AIG General Insurance at ₹6,419.52.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

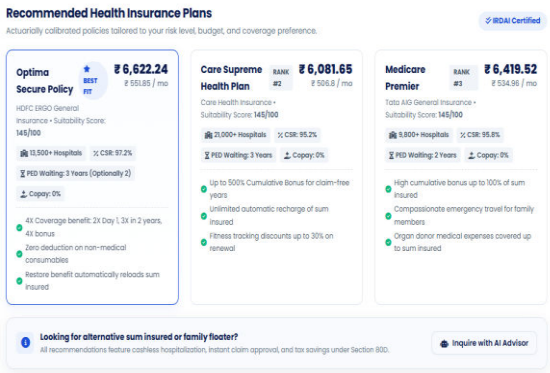


Fig.3. Predicted premium summary and Top 3 recommended plans

B. SHAP Explanation

For the sample profile (Fig. 4), the population-mean base value of ₹20,709 is reduced by ₹13,951 to give a predicted premium of ₹6,757. Smoking status, medical history and BMI have the largest influence, while age, dependents, region and gender have a smaller effect. Being tobacco-free reduces the premium by ₹5,875, a clean medical record by ₹4,636, a healthy BMI of 23.0 by ₹2,001 and an age of 32 years by ₹736.

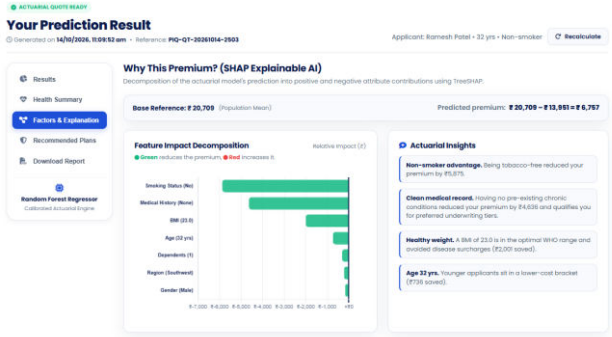


Fig.4. SHAP factor contributions

C. Health Risk Assessment

The sample applicant obtains an overall score of 11 out of 100 (Low Risk tier). The scorecard (Fig. 5) attributes 4 of 30 points to BMI, 0 of 30 to smoking, 7 of 20 to age and 0 of 20 to medical history. Cards beneath give advice on BMI optimisation, preventive care and the no-claim bonus.

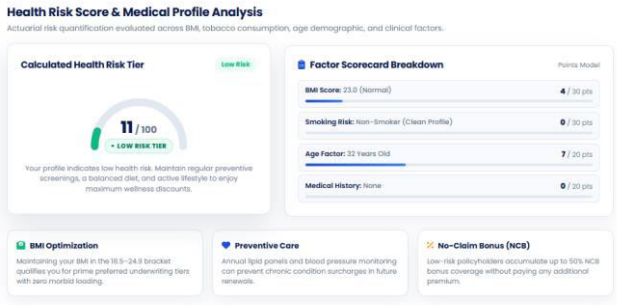


Fig.5. Health risk assessment and risk score

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

D. AI Chatbot and PDF Report

The chatbot (Fig. 6) answers questions on premiums, policy features, deductibles and premium reduction, with quick buttons such as "Why is my premium high?" and "Which plan suits me?". The PDF tab collects the prediction, risk assessment, explanations and recommendations into a downloadable document.

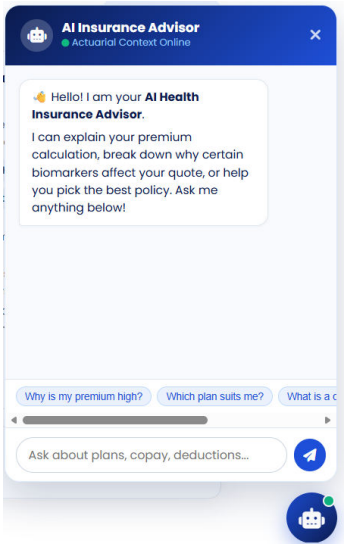


Fig.6. AI insurance chatbot

VII. SYSTEM TESTING

Unit tests check pre-processing, BMI calculation, risk-score logic and that exactly three plans are returned. Functional testing (Table I) uses the preset profiles and invalid inputs. Integration testing confirms that a change in any input, such as smoking status, is reflected consistently in the premium, risk score, explanation and recommended plans, and that results are stored in MongoDB. Performance, usability and compatibility testing confirm fast responses, readable results and correct behaviour on desktop and mobile browsers.

TABLE I. FUNCTIONAL TEST CASES

ID	Test Scenario	Expected Behaviour
TC1	Young & Healthy preset (25 yrs)	Low premium; Low Risk tier
TC2	Balanced Adult preset (38 yrs)	Moderate premium and risk score
TC3	Active Smoker preset (42 yrs)	Higher premium and risk; smoking among major SHAP factors
TC4	Senior with Hypertension (55 yrs)	Higher risk from age and history
TC5	Age outside 18–85 or empty height/weight	Validation message; no prediction
TC6	Question typed in chatbot	Relevant answer displayed
TC7	Download PDF report	PDF holds premium, risk, explanation, plans

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 10, October 2026

|DOI: 10.15680/IJIRSET.2026.1510072|

VIII. LIMITATIONS

Prediction accuracy depends on the quality of user-supplied data, and the model may generalise poorly to regions, age groups or insurers that are under-represented in the training data. The system handles sensitive health information and therefore needs strong data protection. Recommended plans may not reflect the latest terms or prices unless the plan database is updated regularly. The premium is an estimate learned from historical data, not a binding insurer quotation.

IX. CONCLUSION AND FUTURE WORK

The proposed system uses Random Forest Regression to predict health insurance premiums and combines the prediction with SHAP explanations, a Health Risk Score, Top 3 plan recommendation, an AI chatbot and PDF reporting on a single platform, moving the user from an unexplained number to an informed choice of plan.
Future work includes training on larger and more diverse datasets, connecting to live insurer data, stronger encryption, authentication and consent management, comparing Random Forest with gradient boosting and XGBoost under cross-validation, what-if analysis, a multilingual chatbot, family and group policies, wearable-data integration and a mobile application.

ACKNOWLEDGMENT
The authors thank the Principal, Dr. A.V. Ram Prasad, the Head of the Department, Dr. A. Askarunisa, the project guide Dr. S. Subashini, and the Mini Project Coordinator, N. Nazumunisha, of K.L.N. College of Engineering for their guidance and support.

REFERENCES

[1] C. Jyothsna, K. Srinivas, B. Bhargavi, A. E. Sravanth, A. T. Kumar and J. N. V. R. Swarup Kumar, “Health Insurance Premium Prediction using XGBoost Regressor,” IEEE, 2022, pp. 1645–1652.
[2] A. R. Kandula, S. Kalyanapu, S. N. Rayapalli, K. R. Veerabathina, V. Modugumudi and S. R. Kanikella, “Medical Insurance Predictive Modelling: An Analysis of Machine Learning Methods,” IEEE IATMSI, 2024, pp. 1–5.
[3] Z. Zhou and M. Su, “A Recommendation Algorithm of Insurance’s Products Based on Optimal Collaborative Filtering,” IEEE ICMLCA, 2021, pp. 281–284.
[4] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[5] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems 30, 2017.
[6] Scikit-learn Documentation, “Machine Learning and Random Forest Regression,” https://scikit-learn.org.
[7] SHAP Documentation, “Explainable Artificial Intelligence and Model Interpretation,” https://shap.readthedocs.io.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details