Machine Learning-Based Regression Framework to Predict Health Insurance Premiums

Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi, Rajani Singh

# Abstract

This paper presents a machine learning framework to predict health insurance premiums based on demographic and lifestyle features. Initially, models such as Linear Regression and Artificial Neural Networks (ANN) were used, with the ANN achieving 92.72% accuracy. To improve performance and robustness, three key extensions were introduced: hyperparameter tuning using RandomizedSearchCV, ensemble learning with VotingRegressor, and explainability using SHAP (SHapley Additive Explanations). The final framework demonstrated enhanced accuracy and interpretability, making it suitable for real-world deployment.

Index Terms—Machine learning, artificial neural network, regression, SHAP, ensemble learning, health insurance, premium prediction.

# I. Introduction

Health insurance pricing is a complex and data-driven process, typically influenced by factors such as age, BMI, smoking habits, and geographic region. Traditional methods for premium determination are often opaque and inflexible. Machine learning (ML) offers the ability to uncover patterns within large datasets and generate accurate, personalized premium predictions. This study presents a machine learning-based regression approach that leverages artificial neural networks and advanced model enhancements to predict insurance premiums and explain underlying decision factors.

# II. Dataset Description

The dataset used for this study was obtained from Kaggle and contains 1,338 entries. Each record includes features such as age, sex, BMI, number of children, smoking status, region, and insurance charges. The target variable, `charges`, represents the premium amount. The dataset contained no missing values, and categorical variables were encoded numerically.

# III. Exploratory Data Analysis

Exploratory data analysis (EDA) revealed strong correlations between smoking status and premium charges. Age and BMI were also positively correlated. Gender and region had relatively minor influence on premiums. These insights guided feature prioritization in model development.

# IV. Modeling Approaches

A baseline model was developed using linear regression. It achieved an R² score of 0.75, with RMSE and MAE values of 0.499 and 0.344, respectively. Although simple and interpretable, the linear model lacked the ability to capture nonlinear relationships in the data.  
  
A multilayer ANN was constructed using the Keras Sequential API. The model consisted of five dense layers with ReLU activations. Trained over 100 epochs using the Adam optimizer, it achieved an R² score of 0.927, RMSE of 0.27, and MAE of 0.143. The ANN significantly outperformed the linear regression model.

# V. Extensions to the Framework

To further optimize performance, RandomizedSearchCV was used to tune hyperparameters for both Random Forest and XGBoost regressors. Parameters such as `n\_estimators`, `max\_depth`, and `learning\_rate` were optimized. This tuning improved accuracy and reduced overfitting.  
  
An ensemble model was created using a VotingRegressor that combined the outputs of tuned Random Forest, XGBoost, and CatBoost regressors. This approach improved the model’s robustness and generalization across test data.  
  
SHAP (SHapley Additive Explanations) was integrated to interpret model predictions. SHAP revealed that smoking, BMI, and age were the most influential features. Both global summary plots and local force plots were used to visualize feature impact.

# VI. Results and Comparison

Performance comparison:  
- Linear Regression: R² = 0.75, RMSE = 0.499, MAE = 0.344  
- ANN: R² = 0.927, RMSE = 0.27, MAE = 0.143  
- Voting Regressor: Best generalization performance post-tuning  
  
The ensemble model combined accuracy with consistency and outperformed individual base models.

# VII. Conclusion

This paper proposes an enhanced regression framework for predicting health insurance premiums. Starting from a baseline ANN model, the framework incorporates hyperparameter tuning, ensemble modeling, and explainable AI. The result is a high-performance and interpretable model that can be applied to real-world insurance pricing applications.

# References

[1] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2009.  
[2] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, 2017.  
[3] Kaggle, "Insurance Premium Prediction Dataset." [Online]. Available: https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction  
[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.  
[5] F. Chollet, "Deep Learning with Python," Manning Publications, 2018.