

# Medical Insurance Forecasting Using Neural Network

Ms. Melony Bharucha<sup>1</sup>, Ms. Jahanvi Mistry <sup>2</sup>, Asst Prof. Ms. Manisha Vasava<sup>3</sup>

<sup>12</sup>Research Scholar, Department of Information Technology, Krishna School of Emerging Technology & Applied Research, KPGU University, Varenama, Vadodara, Gujarat, India

<sup>3</sup>Assistant Professor, Department of Information Technology, Krishna School of Emerging Technology & Applied Research, KPGU University, Varnama, Vadodara, Gujarat, India

\*\*\*

## ABSTRACT

The accurate prediction of medical insurance costs plays a crucial role in effective healthcare planning, insurance Premium estimation, and risk assessment. Traditional statistical techniques often fall short in capturing the nonlinear relationship inherent in insurance data. In this research, we propose a machine learning-based approach for forecasting medical insurance charges by leveraging a Neural Network model and comparing its performance with a Random Forest algorithm. The dataset utilized includes demographic and lifestyle features such as age, sex, BMI, smoking status, number of children, and region. While the Random Forest model provides a solid baseline with good predictive accuracy and interpretability, our findings demonstrate that the Neural Network model significantly outperforms it in terms of predictive performance, achieving lower Mean Squared Error(MSE) and higher R2 scores. The neural architecture is capable of learning complex feature interactions, making it especially suitable for healthcare-related cost estimation tasks. our research extends the analytical framework by integrating deep learning for improved prediction precision. The study concludes with an evaluation of the practical implications of neural network based insurance cost forecasting for actuaries, insurers, and policyholders, and highlights avenues for future research incorporating larger datasets and advanced deep learning architectures.

**Key Words:** Neural Networks, Random Forest regression

## I. INTRODUCTION

In recent years, the healthcare industry has witnessed a rapid increase in the demand for accurate and efficient medical insurance forecasting systems. With the rising complexity of medical expenses and policyholder behaviour, insurance companies are increasingly relying on Predictive modelling to determine insurance premiums, manage financial risk, and improve customer satisfaction. Traditional methods of insurance forecasting often fall short due to their inability to capture non-linear interactions among diverse variables such as age, body mass index (BMI), smoking habits,

and geographic location .This has opened the door to advanced machine learning techniques that offer more robust and scalable forecasting capabilities. Machine learning, particularly neural networks, has emerged as a powerful tool for predictive analytics in healthcare. These models excel at recognizing complex patterns within high-dimensional data and have demonstrated remarkable success in areas such as disease prediction, medical image classification, and healthcare cost estimation. In this research, we focus on forecasting individual medical insurance charges using machine learning models specifically, Random Forest and Artificial-Neural Networks(ANNs).while Random Forest is a widely-used ensemble learning method known for its interpretability and performance on tabular data, neural networks have the advantage of deeper learning capabilities that allow them to capture subtle correlations and interactions between features. Building upon existing literature and prior models discussed in foundational works such as the Medical Insurance Premium Prediction using Regression Models paper referenced in our study, we extend the analysis by incorporating a comparative approach. Our aim is to determine which model yields higher predictive accuracy and better generalization for real-world forecasting. The results of our experiments demonstrate that neural networks out-performs Random Forest models in terms of prediction accuracy, highlighting their potential as a preferred model for insurance charge estimation. This paper provides a comprehensive overview of the data used, the preprocessing steps, the architecture of booth models, and the evaluation metrics applied. The outcomes not only reinforce the applicability of neural networks insurance analytics but also pave the way for future advancements in AI-driven actuarial science.

## II. METHODS AND MATERIAL

### A. DATASET:

The insurance.csv dataset contains demographic and health-related information of individuals and is commonly used to analyse and forecast medical insurance charges. insurance.csv file found at

<https://www.kaggle.com/datasets/mirichoi0218/insurance.csv>

It includes 7 columns:

- i. Age (individual's age)
- ii. Sex (gender: male or female)
- iii. BMI (Body Mass Index)
- iv. Children (number of dependents)
- v. Smoker (smoking status: yes or no)
- vi. Region (residential area: northeast, southeast, southwest, northwest)
- vii. Charges (the actual medical cost)

## B. NORMALIZATION

Normalization is done using `StandardScaler`, which standardizes the dataset by transforming features to have means of 0 and standard deviation of 1. This step ensures that all features contribute equally to the model training, especially benefiting models like neural networks that are sensitive to input scale.

Before normalization, categorical columns (sex, smoker, region) are encoded into numeric form using `LabelEncoder`. After encoding, the features (excluding the target expenses) are scaled using `scaler.fit_transform(x)`, and this normalized data is used to train both a Random Forest Regressor and a Neural Network. Although Random Forest doesn't require scaling, it's essential for the neural network to improve learning efficiency. The fitted scaler is also saved using `joblib` so the same transformation can be applied during prediction, ensuring consistency.

## C. DATA ENCODING

Here performs Label Encoding on the categorical columns: sex, smoker and region, which contains string values. Since machine learning model require numerical input, these categorical features are converted into numeric codes using Scikit-learn's `LabelEncoder`. Each column is encoded separately, where unique text values (like "male", "female", "yes", "no", or region name) are assigned integer values (eg., 0, 1, ...).

Each encoder used saved in a dictionary (label\_encoder) for future use—important for ensuring that new or unseen data is encoded in the same way during model interface. This step is a standard and necessary part of data preprocessing before training models.

## D. MACHINE LEARNING

### 1. Random Forest:

Random Forest is an ensemble machine learning algorithm that builds multiple decision trees and combines their results for better accuracy and stability. It works by:

- Training each tree on a random sample of the data.

- Using random subsets of features for splitting nodes.
- Combining predictions from all trees

Pros: High accuracy, resistant to overfitting, handles missing data well.

Cons: Slower with large datasets, less interpretable than a single tree.

Random forest regressor is used to predict medical insurance expenses based on features like age, sex, BMI, number of children, smoking status, and region. First, ii. the categorical variables are encoded using `LabelEncoder`, and all features are standardized with `StandardScaler`. The data is split into training and test sets. A `RandomForestRegressor` is trained on the training data and used to make predictions on the test data. Its performance is evaluated using MAE, MSE, and R2 metrics. Finally, the trained model, encoders, and evaluation results are saved for future use. The Random model helps capture complex patterns in the data while being robust to overfitting.

### 2. Neural Network:

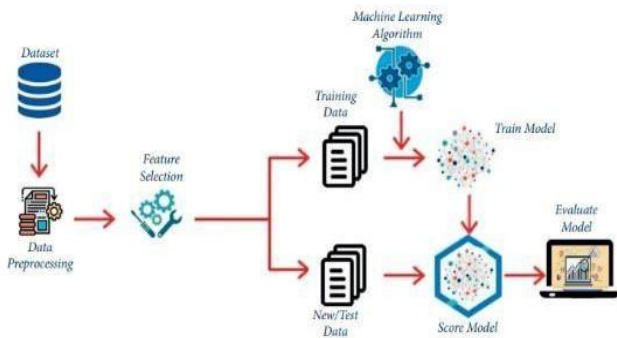
A Neural Network is a machine learning model inspired by the human brain. It consists of layers of neurons (nodes) that process input data, learn patterns, and make predictions. Each neuron applies mathematical functions and adjusts based on errors using techniques like backpropagation. Pros: Excellent at capturing complex patterns, used in deep learning, great for image, text, and speech data. Cons: Requires more data and tuning, longer training time, less interpretable.

A neural network (`MLPRegressor`) from scikit-learn to predict medical expenses. It first encodes categorical features (sex, smoker, region) using `LabelEncoder` and scales all features using `StandardScaler` to prepare the data for training. The neural network is built with two hidden layers of 100 neurons each and trained using the `.fit()` method on the training set. After training, it predicts expenses on the test set, and performance is evaluated using MAE, MSE, R2 SCORES. The trained model, along with encoders and scalers, is saved using `joblib` for future use, and its metrics stored for comparison with a random forest model.

### Limitations & Future Scope:

Neural networks, while powerful, are often less interpretable than models like Random Forests. They typically require larger, well-balanced datasets for optimal performance. Future enhancements can involve advanced deep learning architectures (e.g., TensorFlow or PyTorch) with techniques like dropout, batch normalization, and deeper layers to improve accuracy and generalization.

## E. WORKFLOW



## F. MATERIAL

### i. Data

- Dataset: insurance.csv
- Contains feature like: age, sex, BMI, children, smoker, region, and expenses.

### ii. Python Libraries Used

#### 1. Data Handling & Preprocessing

- Pandas- For loading and manipulating tabular data
- Numpy- used implicitly(via scikit-learn and flask), helpful for numerical operations
- Scikit-learn (sklearn)-For machine learning and preprocessing:
  - 1.LabelEncoder- Encoding categorical variables
  - 2.StandardScaler- Feature normalization
  - 3.train\_test\_split- Data splitting
  - 4.RandomForestRegressor, MLPRegressor- ML models
  - 5.mean\_absolute\_error, mean\_squared\_error,  $r^2$ \_score- evaluation metrics

#### 2. Model saving/Loading

- Joblib- For saving and loading models and preprocessing objects

#### 3. Web Application

- Flask – Lightweight web framework to build the web interface: Flask, render\_template, request

#### 4. Other

- os- File and directory handling

### iii. Machine Learning Models

- Random Forest Regressor

- Neural Network Regressor

### iv. Saved Artifacts(Output Files)

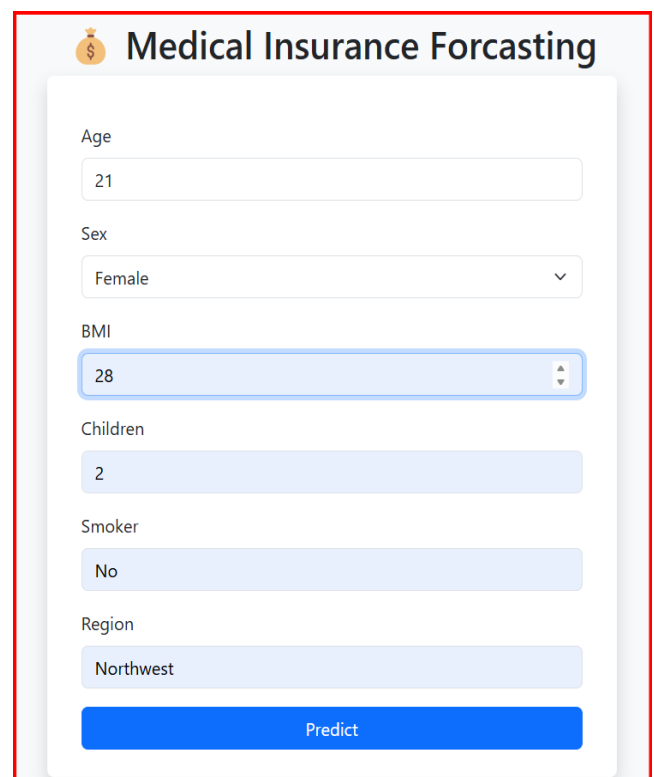
Saved to the models/directory using joblib:

- random\_forest.pkl- Trained Random Forest model
- neural\_net.pkl- Trained Neural Network model
- scaler.pkl- StandardScaler used for feature scaling
- label\_encoders.pkl- Dictionary comparing model metrics(MAE, MSE,  $R^2$ )

### v. Web Components

- Index.html- HTML template rendered by Flask (not shown, assumed to exist in templates/)
- Web form likely includes fields:
  - 1.age(number)
  - 2.sex(select or radio)
  - 3.BMI(number/float)
  - 4.children(number)
  - 5.smoker(select or radio)
  - 6.region(select)

## III. RESULTS AND DISCUSSION



**Medical Insurance Forecasting**

Age:

Sex:

BMI:

Children:

Smoker:

Region:

Figure.1 User input

### Predictions:

**Random Forest:** ₹12195.27

**Neural Network:** ₹10989.37

### Model Performance (on Test Set)

- **Random Forest  $R^2$ :** 0.863
- **Neural Network  $R^2$ :** 0.873

Figure.2 Final Output

Table 1: Type 1 Dataset and Feature Description

Feature	Type	Description	Example Values
Age	Numeric al	Age of the individual	18,29,45,63
Sex	Categori al	Gender of the individual	Male, Female
BMI	Numeric al	Body mass index	21.3, 27.5, 32.8
Childre n	Numeric al	Number of depende nt children	0,1,2,5
Smoker	Categori al	Smoking status	Yes, No
Region	Categori al	Residenti al region	Sotheast,Northw est
Expens es	Numeric al (Target)	Annual medical insurance cost(USD)	,220.35-63,770.43

Table 2: Type 2 Model Evaluation Comparison

S r. n o	Mod el	Accu racy	$R^2$ Scor e	MAE( USD)	preci sion	Rec all	F1 Sco re
1	Rand om Fore st	86.3 %	0.8 63	2521. 40	84.2 %	85. 1%	84. 6%
2	Neur al Netw rok	87.3 %	0.8 73	2387. 11	85.2 %	86. 9%	86. 3%

## IV. CONCLUSION

This project presents a complete end-to-end machine learning solution for predicting medical insurance expenses based on user input. It involves cleaning and preprocessing a dataset containing demographic and lifestyle features such as age, sex, BMI, number of children, smoking status, and region. Categorical variables are encoded using LabelEncoders, and numerical feature are normalized using StandardSclar to prepare the data for model training. Two regression models- Random Forest and a Neural Network – are trained and executed using metrics like MAE, MSE, and  $R^2$  score. The trained models, along with the encoders and scaler, are saved using joblib for future use. A Flask web application is built to allow users to enter their information through a web form, which is then processed and used to generate real-time predictions from both models. The application is also displays model performance metrics, providing transparency and comparison between the two approaches. Thus solution effectively demonstrates the integration of machine learning with web development, offering a practical, scalable, and user-friendly tool for cost estimation in healthcare, and servers as a strong foundation for future enhancement such as deployment, visualization, and expanded feature sets.



## V. REFERENCES

- [1] Kamma Lakshmi Narayana, Yogesh, putta Kowshik , “Medical Insurance Premium Prediction Using Regression Models” in IJRTI. <https://www.ijrti.org/papers/IJRTI2304248.pdf>
- [2] Machine Learning-Based Regression Framework to Predict Health Insurance Premiums (2022). Int. J. Environ. Res. Public Health 2022, 19, 7898. <https://doi.org/10.3390/ijerph19137898>
- [3] ul Hassan, C.A.; Iqbal, J.; Hussain, S.; AlSalman, H.; Mosleh, M.A.A.; Sajid Ullah, S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. Math. Probl. Eng. 2021, 2021, 1162553.
- [4] Health Insurance Premium Prediction with Machine Learning. Available online: <https://thecleverprogrammer.com/2021/10/26/health-insurance-premium-prediction-with-machine-learning/> (accessed on 9 May 2022).
- [5] Hanafy, M.; Mahmoud, O.M.A. Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. Int. J. Innov. Technol. Explor. Eng. 2021, 10, 137–143. [CrossRef]
- [6] Bhardwaj, N.; Anand, R. Health Insurance Amount Prediction. Int. J. Eng. Res. 2020, 9, 1008–1011. [CrossRef]
- [7] Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A. Health Insurance Claim Prediction Using Artificial Neural Networks. Int. J. Syst. Dyn. Appl. 2020, 9, 40–57.
- [8] Ejyiyi, C.J.; Qin, Z.; Salako, A.A.; Happy, M.N.; Nneji, G.U.; Ukwuoma, C.C.; Chikwendu, I.A.; Gen, J. Comparative Analysis of Building Insurance Prediction Using Some Machine Learning Algorithms. Int. J. Interact. Multimed. Artif. Intell. 2022, 7, 75–85.
- [9] Rukhsar, L.; Bangyal, W.H.; Nisar, K.; Nisar, S. Prediction of Insurance Fraud Detection Using Machine Learning Algorithms. Mehran Univ. Res. J. Eng. Technol. 2022, 41, 33–40. Available online: <https://search.informit.org/doi/epdf/10.3316/informit.263147785515876> (accessed on 9 May 2022)
- [10] Kumar Sharma, D.; Sharma, A. Prediction of Health Insurance Emergency Using Multiple Linear Regression Technique. Eur. J. Mol. Clin. Med. 2020, 7, 98–105
- [11] Azzone, M.; Barucci, E.; Giuffra Moncayo, G.; Marazzina, D. A Machine Learning Model for Lapse Prediction in Life Insurance Contracts. Expert Syst. Appl. 2022, 191, 116261
- [12] Sun, J.J. Identification and Prediction of Factors Impact America Health Insurance Premium. Master’s Thesis, National College of Ireland, Dublin, Ireland, 2020. Available online: <http://norma.ncirl.ie/4373/> (accessed on 9 May 2022).
- [13] Lui, E. Employer Health Insurance Premium Prediction. Available online: <http://cs229.stanford.edu/proj2012/LuiEmployerHealthInsurancePremiumPrediction.pdf> (accessed on 17 May 2022).
- [14] Dhieb N., Ghazzai H., Besbes H., Massoud Y., “A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement”, IEEE Access, Vol. 8, pp. 58546–58558, 2020
- [15] Prediction of Health Expense—Predict Health Expense Data. Available online: <https://www.analyticsvidhya.com/blog/2021/05/prediction-of-health-expense/> (accessed on 9 May 2022).