JOURNAL OF
EMERGING INVESTIGATORS

Article

Deep dive into predicting insurance premiums using machine learning

Prithvi Sairaj Krishnan[1], Vivek Shankar[2]
[1] Westwood High School, Austin, Texas
[2] Department of Computer Science, Stanford University, Stanford, California

SUMMARY

We sought to explore prediction of individual insurance premiums to bring transparency and affordability to the opaque healthcare pricing system in the United States. We aimed to develop machine learning models capable of forecasting a patient's out-of-pocket costs for specific medical procedures based on factors like age, insurance plan, provider details, and health conditions. Our primary hypothesis was that machine learning techniques, including both classification and regression models, could predict healthcare expenses and insurance premiums with a coefficient of determination (R^2) greater than 0.9. We hypothesized a positive trend where increasing age, BMI, and the number of pre-existing conditions would correlate with higher healthcare expenses. Furthermore, we expected the models to reveal correlations between certain factors (such as smoking status and geographic region) and increased insurance premiums and out-of-pocket costs. To identify cost predictors we consolidated comprehensive claims data and employed algorithms such as linear regression, decision trees, random forests, and neural networks. Among these models, the multilayer perceptron (MLP) achieved the highest performance, explaining 88% of the variance in costs and reducing the mean absolute error by 9.4% compared to other models. However, we found that limiting input features to basic demographics reduced the model's predictive power. Future studies should focus on enhancing predictions through richer datasets, more advanced neural architectures, and continuous model updates to deliver personalized, accurate cost estimates, ultimately empowering patients and promoting affordability in healthcare decisions.

INTRODUCTION

The rising and unpredictable insurance premiums in the United States present a major financial burden on individuals and families (1). The opacity of pricing within the healthcare system exacerbates this issue, often leaving patients with unanticipated medical bills, as reflected in studies that use diagnostic data to predict costs (1). One way to mitigate these costs is to develop systems that can predict individual insurance premiums and out-of-pocket expenses, allowing patients to plan for medical care more effectively. These predictions, if accurate and personalized, can empower patients to make informed decisions and promote healthcare affordability.

Machine learning models show great promise in predicting insurance premiums based on various factors. These factors typically include demographic attributes (such as age, gender, and geographic region), health-related metrics (such as body mass index (BMI), smoking status, and the number of children), and insurance-related details (such as the type of insurance plan or whether an individual is a smoker) (2). The ability of machine learning to process large datasets with multiple variables makes it well-suited for modeling complex, non-linear relationships between these factors and the cost of medical care.

Our models included linear regression, which is ideal for modeling simple, linear relationships by fitting a straight line that minimizes prediction errors (3). Decision tree regression, on the other hand, handles both linear and non-linear data by recursively splitting the dataset to minimize errors at each step (4). Random forest regression improves accuracy and reduces overfitting by combining multiple decision trees, making it effective for complex high-dimensional datasets (5). Finally, the multilayer perceptron (MLP) is a deep learning model that captures non-linear relationships through multiple layers of interconnected neurons, making it suitable for tasks with intricate dependencies between features (6). These models were evaluated using the mean squared error, which shows how well the model's predictions match the actual values (7).

Several studies explored the use of machine learning in healthcare cost prediction. One of such studies compared the performance of regression techniques such as linear regression, decision trees, and neural networks in predicting insurance premiums (8). Neural networks were found to be most accurate, outperforming simpler models by 10–15% (1). However, many studies were limited in scope, focusing on specific medical settings or using smaller datasets with narrow feature sets. An example of such studies limited in scope used factors such as age, BMI, smoking status, and region to predict health insurance premiums but achieved a moderate accuracy of 75% with the ARIMA model, suggesting that more comprehensive models were needed to capture the full potential of modern machine learning techniques (9).

To address these gaps, recent studies proposed using more advanced models, such as long short-term memory networks (LSTMs) and convolutional neural networks (CNNs) (1). LSTMs are particularly suited for time-series data, enabling models to capture temporal dependencies in patient records, which can provide more accurate predictions of insurance premiums over time. CNNs, while traditionally used for image

https://doi.org/10.59720/24-251

recognition, have potential in modeling structured data with complex interactions, such as medical claims and patient histories. This architecture offers the ability to capture more intricate patterns and dependencies, enhancing the accuracy of predictions.

The aim of this study was to develop machine learning models that could predict individual insurance premiums based on demographic and health-related factors. Our hypothesis was that advanced machine learning techniques (e.g. the MLP) would outperform traditional models such as linear regression and decision trees in predicting insurance premiums. We expected that factors such as age, BMI, and smoking status would show strong correlations with higher insurance premiums and that these advanced architectures would be better at capturing non-linear and temporal relationships in the data.

Our results show that the MLP achieved the lowest mean squared error (MSE), with an R^2 value of 0.88, outperforming other models. We demonstrated that using larger datasets and more advanced machine learning techniques could provide more accurate and reliable predictions of insurance premiums. These findings have important implications for healthcare cost transparency, as they offer a path toward more personalized and fair insurance pricing.

The features included in the dataset were relevant for predicting insurance premiums. Factors such as age, BMI, smoking habits, and geographic region are known to have a substantial impact on an individual's health risks and associated medical costs. By analyzing the relationships between these features and the target variable 'costs,' we developed predictive models that could forecast insurance premiums based on an individual's demographic and health profile. With a sample size of over 1,300 records capturing pertinent features known to influence insurance premiums, this dataset provided a suitable foundation for developing machine learning models to predict individual insurance premiums accurately. The numerical and categorical data formats aligned well with common regression and classification techniques in supervised learning, enabling the exploration of various modeling techniques to address the research problem effectively.

RESULTS

We evaluated and compared several machine learning techniques to address the research problem of accurately predicting individual insurance premiums based on demographic and health factors. Our methodology involved assessing both linear and non-linear models for this regression task.

We developed predictive models for insurance premiums using both linear and non-linear regression techniques, including multiple linear regression, decision trees, and ensemble methods. The dataset comprised demographic and health-related features such as age, BMI, and the presence of pre-existing conditions. After preprocessing the data by handling missing values, normalizing numerical features, and encoding categorical variables, we trained and tested each model to determine their predictive accuracy. We conducted correlation analyses and statistical tests such as ANOVA to identify key factors influencing insurance costs. These steps provided the foundation for interpreting the relationships observed in the results.

First, our results confirmed a positive correlation between BMI and insurance premiums (R^2 = 0.199, $p < 0.01$), indicating that individuals with higher BMI tended to incur higher insurance premiums (**Figure 1**). Similarly, we found a modest but important correlation between age and insurance premiums (R^2 of 0.156 for smokers, R^2 of 0.404 for non-smokers, $p < 0.05$) (**Figure 2**). These findings supported our hypothesis that increasing age and BMI were associated with higher healthcare expenses.

Comparative analysis

We evaluated the trained models on the test set using mean absolute error (MAE), mean squared error (MSE), and R^2. These metrics measure the agreement between patients' actual insurance premiums and the models' predicted premiums for unseen patients in the test set. The MLP emerged as the top performer, with an MAE of 2,538.19, MSE of 1.640E+07, and an R^2 of 0.88. This meant that, on average, the MLP's predicted premium for unseen patients was $2,538.19 off from the actual insurance premium. This performance also indicated the MLP's superior ability to capture complex non-linear relationships between input features and the target variable. The random forest model also performed well, with an R^2 of 0.88. However, the decision tree and the linear regression models showed higher error rates, which highlighted their limitations (**Table 1**). The results indicated that the MLP achieved lower MSE than both the decision tree (p=0.02) and linear regression models (p=0.01), as determined by ANOVA. Additionally, the random forest model performed better than the decision tree (p=0.03), but the difference was not as pronounced when compared to the MLP (p=0.05).

Data exploration

The histogram of medical costs revealed a heavily right-skewed distribution, with most costs in the lower ranges but a long tail of high-cost cases (**Figure 3**). The relationship between the BMI of the patients and the costs incurred also displayed a positive trend and multiple outliers (**Figure 1**). As BMI increased, costs tended to rise, although the relationship was not strictly linear, and there was considerable variation in

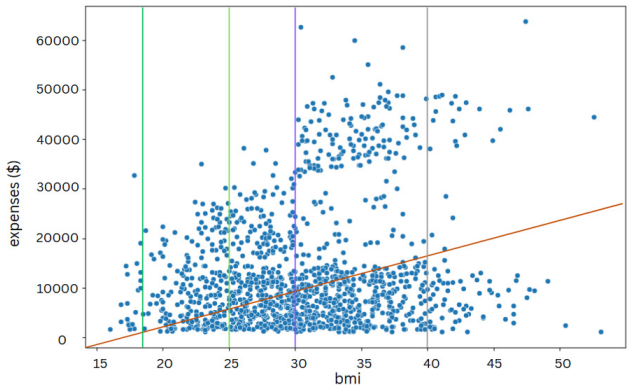

Figure 1: Relationship between insurance premium costs and BMI. BMI and insurance premium costs for all samples (n = 1,338) in the insurance premium prediction dataset. BMI categories included underweight (BMI < 18.5), regular (BMI 18.5–24.9), overweight (BMI 25–29.9), obese (BMI 30–39), and severe obesity (BMI ≥ 40). A linear regression analysis was performed, resulting in the equation y = 394.33x + 1178.18 with an R^2 of 0.199.

https://doi.org/10.59720/24-251

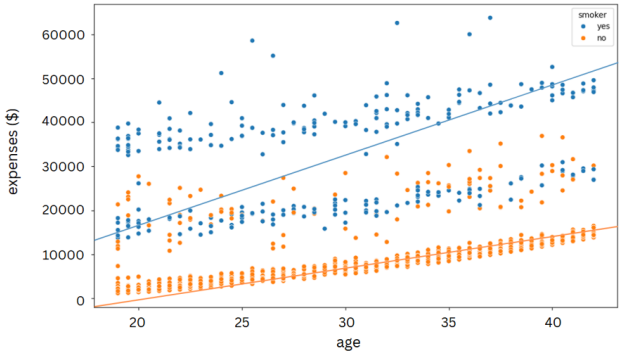

Figure 2: Age vs. insurance costs for smokers and non-smokers. Age and insurance premium costs for all samples (n = 1,338) in the insurance premium prediction dataset, comparing smokers and non-smokers. A linear regression analysis was performed separately for both smokers and non-smokers, resulting in the following equations for smokers: y = 305.24x + 20294.1 with an R^2 of 0.156 and for non-smokers: y = 267.25x - 2091.42 with an R^2 of 0.404.

the data. We observed several outliers with exceptionally high costs across different BMI values. The Pearson correlation coefficient between BMI and costs was approximately 0.199, indicating a weak positive relationship ($p < 0.01$). To investigate potential demographic variations, we created a scatter plot to illustrate insurance costs by region, with data points color-coded by sex (**Figure 4**). While the health charges of patients varied subtly across different regions, they did not change with the sex of patients (**Figure 4**). A 2D grid scatter plot visualized how age and smoking status related to costs incurred, with separate regression lines for smokers and non-smokers (**Figure 2**).

Model optimization

We presented the MAE values for a random forest regressor model as tree depth increased, helping to identify optimal model parameters (**Figure 5**). Similarly, we plotted the MAE of the MLP as maximum training iterations increased, demonstrating an improvement in the model's performance over time (**Figure 6**).

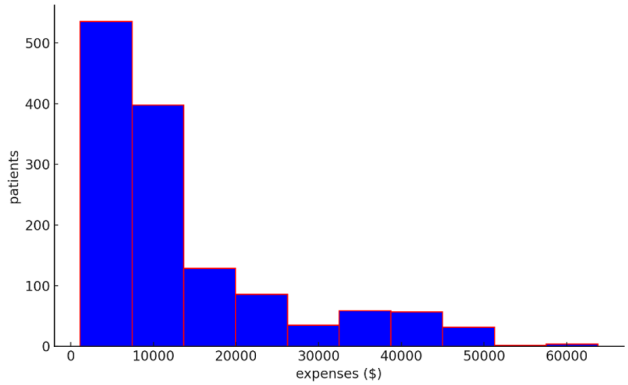

Figure 3: Distribution of insurance premium costs. The right-skewed distribution of insurance premium costs (in dollars) across all samples (n = 1,338) in the insurance premium prediction dataset. The y-axis represents the number of individuals in each bin, with a bin size of $8,500.

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R^2
Linear Regression	4056.6	3.295E+7	0.744
Random Forest Regressor	2984.54	1.936E+7	0.865
Decision Tree Regressor	3610.56	4.059E+7	0.714
Multilayer Perceptron	2538.19	1.640E+7	0.880

Table 1: Model evaluation metrics. Results of experiments training and evaluating the linear regression, random forest, decision tree, and multilayer perceptron models on the insurance premium prediction dataset. The MLP achieved lower MSE than both the decision tree ($p=0.02$) and linear regression models ($p=0.01$), as determined by ANOVA. Additionally, the random forest model performed better than the decision tree ($p=0.03$), but the difference was not as pronounced when compared to the MLP ($p=0.05$).

Feature importance

To better understand the contribution of individual features to the model predictions, we calculated the feature importance for the random forest model. BMI was the most influential factor, contributing 30.5%, followed by age at 18.5% and smoking status at 10.5% (**Table 2**). In linear regression, the coefficients indicated a similar trend, with BMI and age having the largest positive coefficients. For the MLP, we performed feature elimination, which confirmed that removing BMI increased the model's MSE by 14.2%, further highlighting its predictive value (**Table 3**).

DISCUSSION

We evaluated several machine learning models to predict individual insurance premiums using demographic and health-related factors. The MLP outperformed all others, achieving an average MAE of 2,538.19 and R^2 of 0.88 across training splits, supporting the hypothesis that deep learning models can effectively capture complex, non-linear relationships in healthcare data. These results highlight the strength of deep learning models, especially the MLP, in capturing non-linear relationships between factors like age, BMI, and pre-existing conditions and their impact on insurance costs.

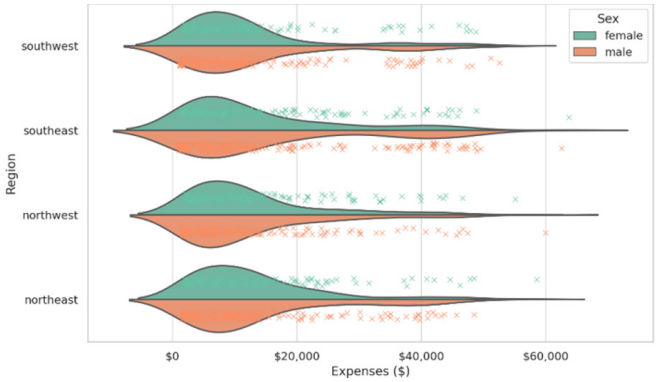

Figure 4: Insurance costs by region and sex. Insurance costs by region in the United States, with data points color-coded by sex (orange for male, green for female), for all samples (n = 1,338) in the insurance premium prediction dataset.

https://doi.org/10.59720/24-251

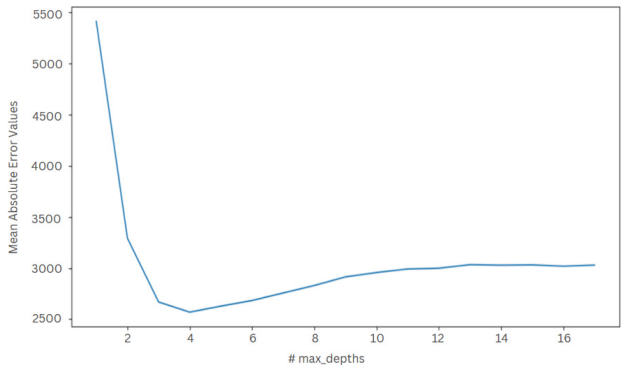

Figure 5: Mean absolute error (MAE) of the random forest regressor. MAE of the random forest regressor model across different maximum depth values. The experiment was conducted by repeatedly training the MLP on the insurance premium prediction dataset using different max_depth settings and measuring the final MAE on the test set after model convergence.

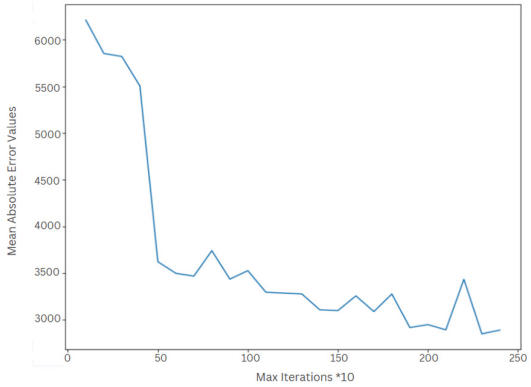

Figure 6: Mean absolute error (MAE) of the multilayer perceptron (MLP). MAE of the MLP across maximum training iterations (scaled by a factor of 10). The experiment was conducted by training the MLP on the insurance premium prediction dataset, using 5-fold cross-validation while monitoring the MAE across training iterations

In our exploratory data analysis, smoking status showed a bimodal distribution, suggesting insurers may use tiered pricing strategies based on smoking-related risk levels. Costs increased with age and were consistently higher for smokers across all age brackets, especially among older adults. Positive correlations between age, BMI, and pre-existing conditions further reinforced these as key contributors to premium pricing, consistent with findings in population-level health cost models (10).

Despite the promising results, there were some limitations. The feature set used in this study was restricted to basic demographic and health information, which may not fully capture the nuances of insurance premiums. For example, the dataset did not include granular clinical details, specific medical treatments, or socioeconomic factors, which could greatly influence medical expenses. Additionally, we observed larger residuals in some cases that suggest the model may have struggled with higher-order feature interactions that were not captured by the available data. These limitations highlight the importance of high-quality, representative data that reflects a broad range of geographic areas, patient profiles, and care settings to improve model generalization. The dataset used was publicly available and not directly sourced from an insurance company, limiting its representativeness. Future work should focus on addressing these limitations through enhanced feature engineering and the inclusion of more detailed medical, socioeconomic, and geographic data. Techniques such as data augmentation and incorporating expert healthcare domain knowledge might improve model performance. Additionally, more advanced architectures like recurrent neural networks (RNNs) or attention mechanisms may better model longitudinal patient data, capturing intricate dependencies across time. Continuously updating the model with fresh data will also ensure that the predictions remain relevant in an ever-evolving healthcare landscape.

In conclusion, we set out to predict individual insurance premiums to bring transparency and affordability to the opaque healthcare pricing system (10). Our findings show that machine learning models, particularly the MLP, can effectively forecast insurance costs, with an R^2 of 0.88, outperforming traditional regression models. Unlike prior studies, which often relied on limited feature sets or simpler models, our research demonstrated that incorporating demographic and health-related factors such as BMI, age, and smoking status enhances predictive accuracy. While our results support the hypothesis that advanced machine learning techniques improve cost prediction, they also highlight the need for richer datasets and model refinements to ensure fairness and generalizability. These insights contribute to ongoing efforts to make healthcare pricing more transparent and equitable, ultimately benefiting both patients and providers.

Feature	Random Forest Importance	XGBoost Importance
BMI	0.305	0.187
Age	0.185	0.171
Smoking Status	0.105	0.142

Table 2: Feature Importance of BMI, Age, and Smoking Status. Feature importance values of BMI, age, and smoking status based on the random forest and XGBoost models. The values reflect each feature's relative contribution to the model's predictive power, with higher values indicating greater importance.

Model	Selected Features	Description
Random Forest	BMI, Age, Smoking Status	All three features retained; most important contributors
XGBoost	BMI, Age, Smoking Status	All three features retained; confirmed as important

Table 3: Feature Elimination Metrics. Features selected through recursive feature elimination with cross-validation for both the random forest and XGBoost models. All three features of BMI, age, and smoking status were retained, indicating their significance in predicting insurance premiums.

MATERIALS AND METHODS

Model performance evaluation

The experiments involved training linear regression, decision tree regression, random forest regression, and MLP regression models on a dataset split into 67% training and 33% testing sets with a random seed of 100 for reproducibility.

https://doi.org/10.59720/24-251

Dataset information

The "Insurance Premium Prediction" dataset, a 2019 dataset obtained from the Kaggle repository, was used for this research project (11). It consisted of numerical data relating to various factors that influenced individual insurance premiums and health insurance premiums. The dataset comprised a total of 1,338 samples, with each sample representing a unique beneficiary record. To facilitate the proper evaluation of the predictive models' performance on unseen data, the dataset was split into training and testing sets using the "train_test_split" function from the Scikit-learn library. The test set size was set to 33% of the total data, and a random seed of 100 was used to ensure the reproducibility of the data-splitting process.

The dataset included seven key features that captured relevant information about the beneficiaries. The age of the primary beneficiary was represented as a numerical variable. The genders of the beneficiaries were encoded as a binary categorical variable, with values indicating male or female. The body mass index (BMI) of the beneficiary was also provided as a numerical variable. Additionally, the dataset included the number of children covered by the health insurance plan as a numerical feature. Two binary categorical variables were present: one indicating whether the beneficiary was a smoker or not, and another that represented the residential region of the beneficiary within the United States. These regions were defined using standard groupings: Northeast (CT, ME, MA, NH, RI, VT, NJ, NY, PA, DE, MD, WI, IL, IN, IA, MI, MN, MO, OH), Southeast (AL, AR, FL, GA, LA, MS, NC, SC, TN, KY, VA, WV), Southwest (AZ, NM, OK, TX, KS, CA, HI), and Northwest (AK, CO, ID, MT, NE, ND, OR, SD, UT, WA, WY, NV). Finally, the individual insurance premiums billed by the health insurance provider served as the target variable to be predicted, represented as a numerical value.

Pre-processing

Before model training, several preprocessing steps were undertaken to prepare the data for machine learning modeling. Firstly, the dataset did not contain any null values, eliminating the need for imputation techniques. The categorical features, namely 'sex', 'smoker', and 'region', were encoded into numerical values using the LabelEncoder class from Scikit-learn. Since the dataset consisted of only numerical and categorical variables, no explicit feature scaling was performed, as most machine learning algorithms could handle these data types without the need for scaling.

Cross-validation

5-fold cross-validation was used to evaluate model generalization and reduce overfitting. This value of K was chosen as a balance between computational efficiency and the need for reliable performance estimates across multiple data splits. In each fold, 67% of the data was used for training, and 33% for testing, ensuring that each data point was used for either training or validation.

Model design and training

The standard, "vanilla" version of linear regression was used in this study, without regularization techniques like L1 or L2 penalties, and no hyperparameters were optimized for this model using grid search. This made linear regression simple and interpretable, but it struggled with capturing non-linear relationships. It was selected as a baseline model to provide a comparison point for evaluating more complex models.

In this study, a maximum depth of 10 was used for decision tree regression, and splits were determined by minimizing mean squared error (MSE) with a minimum of 5 samples per leaf. Grid search was employed to optimize the maximum depth (tested values between 5 and 15) and the minimum samples per leaf (ranging from 2 to 10), with the best configuration being a maximum depth of 10 and a minimum of 5 samples per leaf. This setup helped prevent overfitting while maintaining interpretability. Decision trees were valuable because they provided insights into how individual factors, such as age or BMI, influenced medical costs.

The random forest model in this study used 100 decision trees, with each tree limited to a maximum depth of 10, and the square root of the total number of features was considered at each split. Grid search was used to optimize the number of trees (ranging from 50 to 200), the maximum depth (tested values between 5 and 15), and the number of features considered at each split (square root of the total number of features, log2 of the total number of features, or fixed). The optimal configuration was 100 trees, a maximum depth of 10, and the square root of the total number of features at each split. This combination of trees provided a more robust and generalized model compared to a single decision tree.

The MLP had 3 hidden layers with 128, 64, and 32 neurons, using Rectified Linear Unit (ReLU) activation in the hidden layers and a linear output. The model was trained with a learning rate of 0.001, a batch size of 32, and for 100 epochs using the Adam optimizer. Grid search was employed to optimize the number of neurons in each hidden layer (ranging from 64 to 256 for the first layer, 32 to 64 for the second, and 16 to 32 for the third), the learning rate (ranging from 0.0001 to 0.01), and the batch size (16, 32, 64). The best configuration found was 3 hidden layers with 128, 64, and 32 neurons, a learning rate of 0.001, and a batch size of 32. To prevent overfitting, dropout layers were applied, where 20% of the neurons were randomly dropped during training, and L2 regularization (weight decay) was added to penalize model complexity. The MLP was chosen for its ability to model non-linear dependencies between demographic and health factors, making it ideal for complex healthcare cost predictions.

Computing resources

The models were trained and tested using Python (version 3.x) with the Scikit-learn, NumPy, and Matplotlib libraries. The computations were performed on a local machine with an Intel i7 processor and 16 GB of RAM.

Statistical tests

Pearson correlation coefficients were used to measure the strength of the linear relationships between input features (age, BMI, and number of pre-existing conditions) and insurance premiums. Analysis of Variance (ANOVA) was conducted to compare the performance of different models (linear regression, decision tree, random forest, and multilayer perceptron) based on their mean squared error (MSE) and mean absolute error (MAE). Before applying ANOVA, Shapiro-Wilks tests were used to assess the normality of the data, as ANOVA assumes that the residuals are normally distributed.

Post-hoc tests were conducted using Tukey's honestly significant difference (HSD) test to further investigate pairwise comparisons between models, allowing us to identify specific

https://doi.org/10.59720/24-251

differences between models' performance metrics. Additionally, residual analysis was conducted to examine the predictive errors, ensuring that no systematic bias existed in the models' predictions.

Performance metrics

Several evaluation metrics were used to comprehensively assess the predictive models' performance. Models were evaluated using MAE (the average magnitude of errors without considering directions), MSE (the average squared difference between predicted and actual values), and R^2 (the proportion of variance explained by input features). K-fold cross-validation assessed generalization performance. Techniques like regularization and dropout were employed to prevent overfitting. The best-performing model was selected based on MSE, providing a robust solution for predicting insurance premiums. These performance metrics demonstrated how we evaluated our models' performance and highlighted the importance of evaluation, especially with models of high impact.

Code

Our code developed during this study can be downloaded from the following Github link: https://github.com/prithsk/Insurance-Premium-Predictor.git.

Received: June 16, 2024
Accepted: February 5, 2024
Published: August 6, 2025

REFERENCES

1. Kaushik, Keshav, et al. "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums." *International Journal of Environmental Research and Public Health*, vol. 19, no. 13, 28 June 2022, https://doi.org/10.3390/ijerph19137898.
2. Ash, Arlene, et al. "Using Diagnoses to Describe Populations and Predict Costs." *Health Care Financing Review*, vol. 21, no. 3, 20 Mar. 2000, https://pmc.ncbi.nlm.nih.gov/articles/PMC4194673/.
3. "LinearRegression." Scikit. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. Accessed 4 Aug. 2024.
4. "DecisionTreeRegressor." Scikit. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html. Accessed 4 Aug. 2024.
5. "RandomForestRegressor." Scikit. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Accessed 4 Aug. 2024.
6. "Classification Using Sklearn Multi-Layer Perceptron." GeeksforGeeks. https://www.geeksforgeeks.org/classification-using-sklearn-multi-layer-perceptron/. Accessed 4 Aug. 2024.
7. "3.4. Metrics and Scoring: Quantifying the Quality of Predictions." Scikit. https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed 4 Aug. 2024.
8. Kaushik, Shruti, et al. "AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures." *Frontiers*, vol. 3, 18 Mar. 2020, https://doi.org/10.3389/fdata.2020.00004.
9. Yang, Chengliang, et al. "Machine learning approaches for predicting high cost high need patient expenditures in health care." *BioMedical Engineering OnLine*, vol. 17, no. 1, 23 July 2018, pp. 1–20. https://doi.org/10.1186/s12938-018-0568-3.
10. Drewe-Boss, Philipp, et al. "Deep Learning for Prediction of Population Health Costs" *BMC Medical Informatics and Decision Making*, vol. 22, no. 32, 03 Feb 2022, https://doi.org/10.1186/s12911-021-01743-z.
11. Nursnaaz. "Insurance Premium Prediction." Kaggle. https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction. Accessed 17 July 2024.