Medical Insurance Price Prediction Using Machine Learning

AL AYACHI Rania
National School Of Applied Sciences Al-Hoceima
rania.alayachi@etu.uae.ac.ma

1-Abstract

The prediction of health insurance costs is a critical task that benefits both providers and consumers by enhancing risk assessment and financial planning. This project applies machine learning techniques to forecast medical insurance premiums using a comprehensive dataset encompassing demographic, lifestyle, and health-related factors. A series of data preprocessing steps, including handling missing values, encoding categorical features, and scaling numerical variables, ensures the dataset's readiness for modeling. Several algorithms, including Linear Regression, Decision Trees, Random Forest, K-Nearest Neighbors, and XGBoost, are implemented and evaluated for predictive accuracy. Random Forest, further optimized using hyperparameter tuning, emerges as the best performer, achieving high R^2 scores for training and testing datasets. This work highlights the potential of machine learning models to provide accurate cost predictions, ultimately fostering a data-driven approach to insurance pricing and risk management.

2. Introduction

This study explores the application of machine learning techniques to predict medical insurance costs, aiming to enhance precision, efficiency, and adaptability in insurance pricing strategies. By leveraging data-driven insights, the research addresses critical challenges faced by stakeholders in the healthcare and insurance sectors, including risk assessment, resource allocation, and the formulation of equitable policies. The intricate nature of medical insurance pricing arises from a variety of factors, such as demographic attributes, lifestyle choices, medical histories, regional differences, and broader economic trends. These complexities often limit the effectiveness of traditional actuarial methods, which struggle to capture dynamic relationships among variables, leading to suboptimal predictions and missed opportunities for proactive risk management.

In the rapidly evolving landscape of healthcare, shaped by technological advancements, demographic shifts, and regulatory changes, the accurate determination of insurance premiums has become a critical focus. Historically, pricing strategies have relied on statistical models and historical data, but these approaches are often inadequate in the face of the growing availability of diverse data sources and the increasing sophistication of machine learning algorithms. These advancements open new avenues for uncovering hidden patterns, extracting actionable insights, and adapting to market dynamics.

Machine learning methods, such as regression models, decision trees, and ensemble techniques, offer the capability to uncover intricate relationships and improve prediction accuracy compared to traditional methods. However, challenges such as balancing accuracy with computational efficiency and ensuring model generalizability remain significant. This research evaluates multiple machine learning algorithms, including Linear Regression, Random Forest, and XGBoost, focusing on their ability to predict insurance costs effectively. By comparing their performance and refining models through hyperparameter tuning, this study identifies key drivers of insurance costs and optimal predictive strategies.

This work underscores the transformative potential of machine learning in reshaping insurance cost prediction, providing insights that enable insurers to develop data-driven, equitable, and adaptive pricing frameworks. As a result, this research contributes to advancing the intersection of predictive analytics and real-world applications in the healthcare sector.

3. Related Work

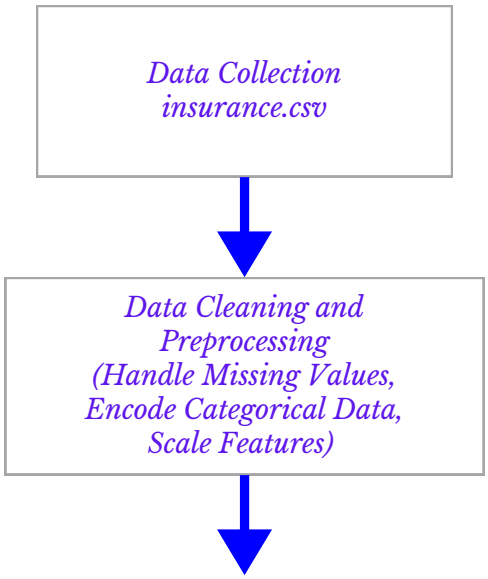
The inspiration for this project originates from three scientific studies that significantly influenced its inception. The first study [1], published in the Journal of Data Analysis and Information Processing, explored various machine learning models, including XGBoost, Lasso, and Ridge regression, for predicting health insurance costs. The research highlighted XGBoost's superior performance, achieving an R^2 score of 86.81% and a Root Mean Squared Error (RMSE) of 4450.4.

The second study [2], available on arXiv, delved into the trade-offs between ensemble models such as Extreme Gradient Boosting (XGBoost), Gradient Boosting Machine (GBM), and Random Forest. While XGBoost demonstrated higher accuracy, Random Forest was noted for its computational efficiency, achieving lower prediction error with reduced resource consumption.

Finally, the third study [3] evaluated the performance of three regression models—Linear Regression, Gradient Boosting, and Support Vector Machine—in predicting medical insurance costs. Using metrics like RMSE, R^2, and K-Fold Cross-validation, the study found Gradient Boosting to be the most effective, with the highest R^2 (0.892) and lowest RMSE (1336.594). K-Fold Cross-validation showed similar results across the models. Exploratory Data Analysis revealed significant differences between smoker and non-smoker groups, with the highest charges linked to the smoker feature, confirming that Gradient Boosting performed the best.

4. objective

The primary objective of this project is to predict medical insurance costs using advanced machine learning techniques. By addressing this as a regression problem, the work aims to provide accurate cost predictions based on demographic, lifestyle, and health-related factors such as age, BMI, smoking status, and region. Furthermore, the project seeks to identify the key variables that most significantly impact insurance premiums. This insight can empower stakeholders, including insurers and policymakers, to develop data-driven pricing strategies and optimize risk management processes.
Through a systematic exploration of machine learning models—such as Linear Regression, Decision Trees, Random Forest, and XGBoost—the project evaluates and compares their performance to identify the most effective approach for this task. Additionally, hyperparameter tuning is employed to optimize the best-performing model, enhancing its predictive accuracy and adaptability. This study contributes to the growing field of predictive analytics in healthcare, offering practical solutions for the dynamic and complex challenges of medical insurance cost estimation. **Fig 1** illustrates the steps of the project.



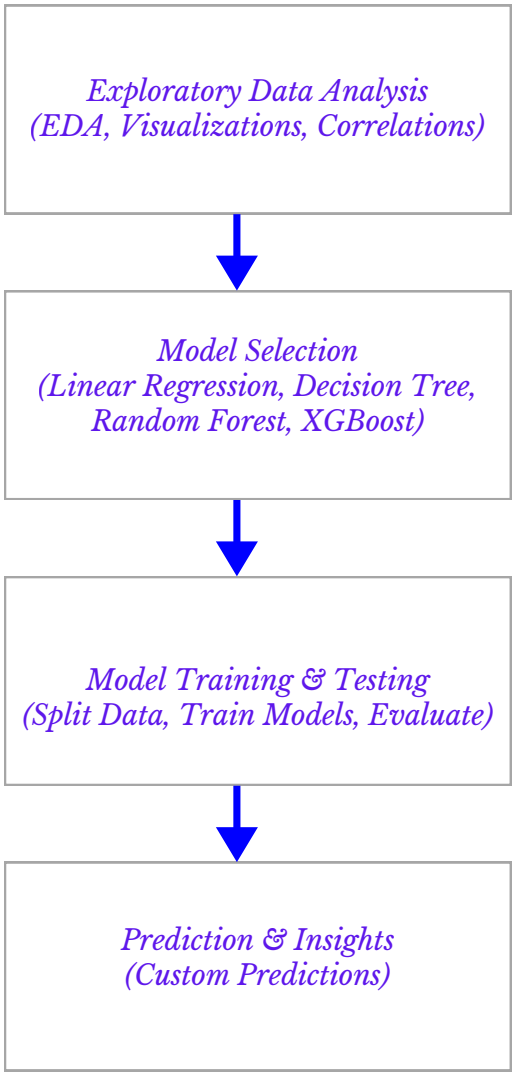


Fig1.*Project Workflow: Medical Insurance Cost Prediction*

5. Methodology

5.1. Data Collection and Description :

The dataset, sourced from Kaggle, is designed to predict medical insurance costs. It includes six attributes—age, sex, BMI, number of children, smoking status, and region—along with the target variable, charges. The dataset consists of 1,338 rows and 7 columns, and it was split into training (80%) and testing (20%) subsets to build and evaluate predictive regression models.

The "charges" attribute, a floating-point value, represents insurance costs based on individuals' demographic and health characteristics. Most individuals in the dataset are aged between 18 and 60, with few having more than three children. BMI values range between 15.96 and 53.13, and the dataset includes four regions: northeast, northwest, southeast, and southwest. The southeast region has the highest proportion of smokers, indicating significant geographic variation in smoking prevalence and its impact on insurance costs.

Attribute	Data Description
Age	The age of individual person
Sex	Sex of the person (Male, Female)
BMI	This is Body Mass Index
Children	Total number of children of the person have
Smoker	Whether the person is a smoker or not
Region	Where the person lives. Considering four regions (Southwest, Southeast, Northeast, Northwest)

Table 1: Overview of the Dataset

This study examines the relationships between "charges" and predictor variables using exploratory data analysis and statistical metrics. Insights from these analyses guided the development of machine learning models to predict insurance costs more accurately.

	age	bmi	children	charges
count	1337.000000	1337.000000	1337.000000	1337.000000
mean	39.222139	30.663452	1.095737	13279.121487
std	14.044333	6.100468	1.205571	12110.359656
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.290000	0.000000	4746.344000
50%	39.000000	30.400000	1.000000	9386.161300
75%	51.000000	34.700000	2.000000	16657.717450
max	64.000000	53.130000	5.000000	63770.428010

5.2 Data cleaning and preprocessing

5.2.1- Data Cleaning

In this project, we carried out a series of data cleaning steps to prepare the dataset for modeling. Initially, we checked for duplicate records and removed any redundancies to ensure data integrity. Following this, we assessed the dataset for missing values, confirming that there were none to address. We identified numerical and categorical columns to better understand the dataset's structure. Furthermore, statistical summaries were computed to examine the distributions and ranges of numerical variables. By addressing these foundational issues, we ensured the dataset was clean and ready for feature engineering and analysis.

5.2.2- Data Processing

In the data processing phase, our primary objective was to transform the raw dataset into a structured format suitable for analysis. This involved encoding categorical variables such as "sex," "region," and "smoker" using Label Encoding to convert them into numerical representations. To ensure uniformity, numerical features like age, BMI, and the number of children were standardized using the StandardScaler. This step normalized the features to improve model performance and convergence. By performing these tasks, we structured the dataset to enhance its suitability for subsequent machine learning tasks.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Table 2: Dataset before encoding categorical variables

	age	sex	bmi	children	smoker	region
0	-1.440418	0	-0.453160	-0.909234	1	3
1	-1.511647	1	0.509422	-0.079442	0	2
2	-0.799350	1	0.383155	1.580143	0	2
3	-0.443201	1	-1.305052	-0.909234	0	1
4	-0.514431	1	-0.292456	-0.909234	0	1

Table 3: Dataset after encoding categorical variables

5.2.3- Feature Scaling

In this project, feature scaling was conducted using the StandardScaler to standardize numerical variables, ensuring they were on a consistent scale. The StandardScaler transforms the data using the formula:

z = (x - mean) / std

Where:

- z: Standardized value.
- x: Original value.
- mean: Mean of the feature values.
- std: Standard deviation of the feature values.

This scaling was applied to numerical features such as age, BMI, and number of children to improve model performance and training consistency.

5.3 Data Analysis & EDA

Exploratory Data Analysis (EDA) is a fundamental approach in data analysis aimed at uncovering general patterns, identifying notable features, and detecting outliers within the dataset. This initial phase is essential for understanding the data and guiding subsequent analyses. Graphical methods, such as histograms and boxplots, are commonly used to visualize data distributions and key characteristics.

Correlation analysis, a key part of EDA, measures the strength of relationships between variables. The correlation coefficient (rrr) quantifies this relationship, with high values indicating strong associations and low values showing weak or no connections.
The heatmap in Figure 2 illustrates correlations between medical charges and factors such as sex, age, BMI, region, number of children, and smoking status. Strong positive correlations are shown in purple, strong negative correlations in green, and weak or no correlations in beige. Smoking status strongly correlates with higher medical charges, highlighting its significant impact on healthcare costs.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Where :

- r is the correlation coefficient
- xi and yi are individual sample points
- $\bar{x}$ and $\bar{y}$ are the respective means of x and y
- n is the total number of samples.

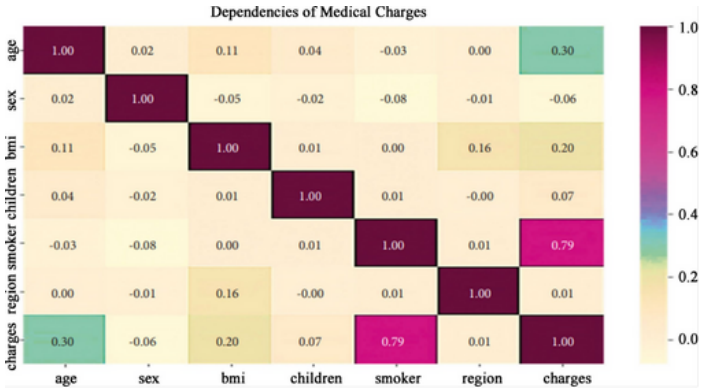
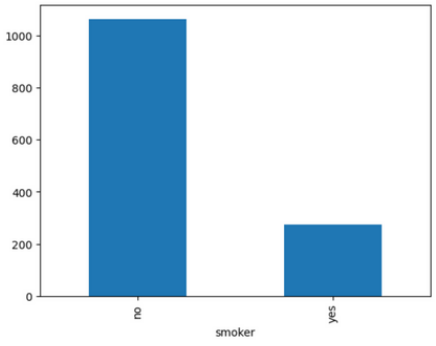
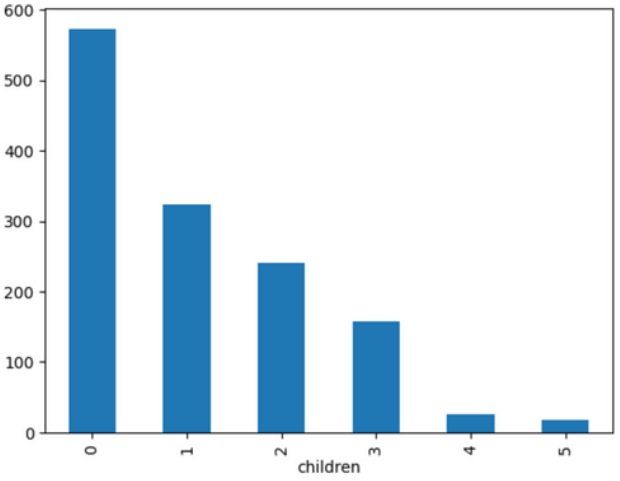
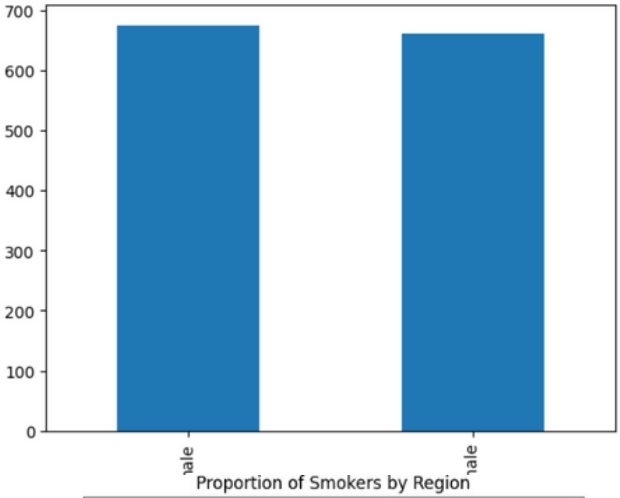


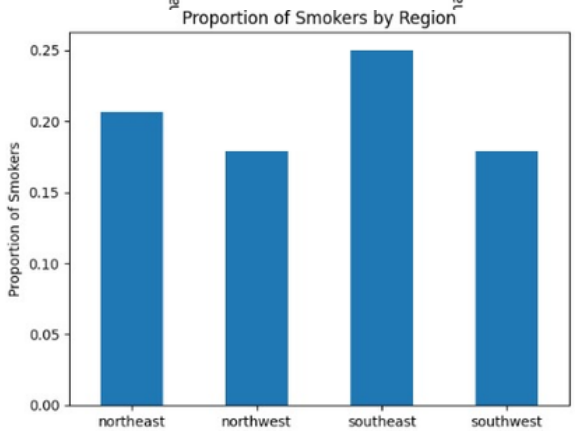
Fig2.correlation matrix with a heatmap

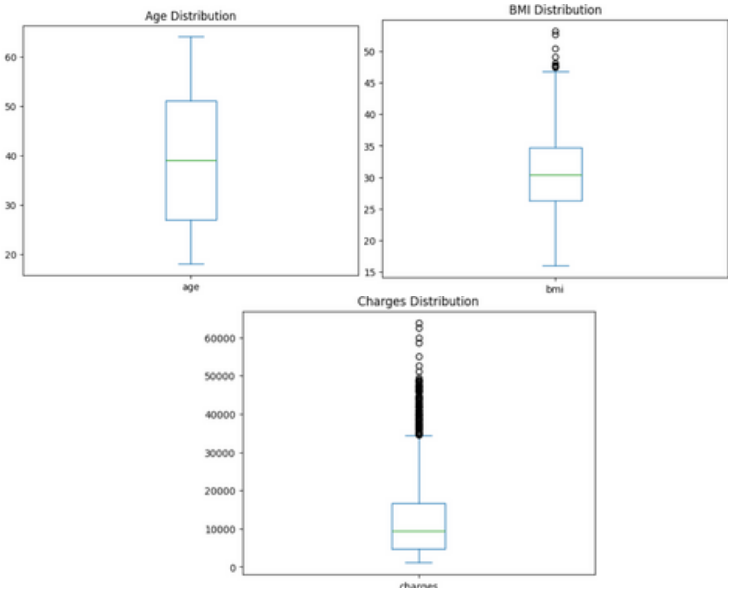
The dataset consists of 1,338 rows and 7 columns, representing insurance costs based on individuals' demographic and health characteristics. The target variable, charges, is a continuous variable representing medical costs. Most individuals in the dataset are aged between 18 and 60, with few having more than three children. BMI values range from 15.96 to 53.13, indicating a wide variation in body mass. The dataset considers four regions: northeast, northwest, southeast, and southwest. Notably, the southeast region has the highest proportion of smokers, reflecting significant geographic variation in smoking prevalence and its impact on insurance costs. Below are some visualizations derived from this data.











5.4 Performance Matrix

To assess the effectiveness of machine learning (ML) models, evaluation metrics are employed to measure their performance. Key metrics such as Root Mean Squared Error (RMSE) and R-squared (R^2) are crucial for comparing different algorithms and understanding their predictive accuracy.

1) Root Mean Squared Error (RMSE)
RMSE is a widely used metric that provides a measure of the differences between predicted and actual values. It is calculated as the square root of the Mean Squared Error (MSE). The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - y_t')^2}{n}}$$

Where:

- y_t' represents the predicted value.
- y_t represents the actual (observed) value.
- n is the total number of observations in the test set.

2) R-squared (R^2)
R-squared, also known as the coefficient of determination, is a statistical metric used to evaluate the proportion of variance in the dependent variable that is predictable from the independent variables. It is computed using the formula:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

Where:

- The unexplained variation refers to the Sum of Squared Residuals (SSR).
- The total variation is represented by the Total Sum of Squares (SST).

The R^2 value ranges from 0 to 1. A value of 1 indicates that the model perfectly predicts the dependent variable, while a value of 0 implies that the model explains none of the variability in the data.

5.5 Model Selection

For predicting medical insurance costs, we implemented various regression algorithms to identify the most suitable one for this task. In the subsequent section, we briefly outline the fundamental concepts of each algorithm used in our study.

5.5.1- .Linear Regression

Linear Regression is a simple yet powerful model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). The mathematical representation is:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n + e$$

Where:
• y is the dependent variable.
• b0 is the y-intercept.
• b1 to bn are the coefficients of the independent variables.
• x1 to xn are the independent variables.
• e is the error term.

Linear Regression	R-squared scor	RMSE
Training set	0.75	6083.22
Testing set	0.74	5957.61

5.5.2- .K-Nearest Neighbors K-NN

K-Nearest Neighbors is a non-parametric classification algorithm that classifies instances based on their similarity to the majority class among their k-nearest neighbors. It does not make any assumptions about the underlying data distribution. Many types of distance can be used, the most commons are: the Euclidean and Manhattan Distances (resp. shown bellow):

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n}|p_i - q_i|$$

The R-squared score for :
Train R^2 score: 81.44
Test R^2 score: 68.87

5.5.3- Support Vector Regression

Support Vector Regression (SVR) is a machine learning algorithm used for regression analysis. SVR Model in Machine Learning aims to find a function that approximates the relationship between the input variables and a continuous target variable while minimizing the prediction error.Unlike Support Vector Machines (SVMs) used for classification tasks, SVR Model seeks a hyperplane that best fits the data points in a continuous space. This is achieved by mapping the input variables to a high-dimensional feature space and finding the hyperplane that maximizes the margin (distance) between the hyperplane and the closest data points, while also minimizing the prediction error.

The R-squared score for :
Train R^2 score: -10.22
Test R^2 score: -10.27

5.5.4- Decision Tree

The Decision Tree algorithm is a supervised machine learning method employed for classification and regression tasks. It constructs a tree-like structure by recursively partitioning the input space based on feature Conditions (Gini impurity & Information Gain & Entropy).

A Decision Tree is a flowchart-like structure in which each internal node represents a feature (or attribute), each branch represents a decision rule, and each leaf node represents an outcome. The topmost node in a decision tree is known as the root node. It learns to partition based on the attribute value. It partitions recursively in such a manner called recursive partitioning.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

$$Information\ Gain = Entropy_{parent} - Entropy_{children}$$

$$E = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

min_samples_split=5: The minimum number of samples required to split an internal node. It controls the complexity of the tree.

The R-squared score for :
Train R^2 score: 99.9
Test R^2 score: 71.87

5.5.5- XGBRegressor

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.

The R-squared score for :
Train R^2 score: 99.5
Test R^2 score: 80.67

5.5.6- Random Forest Regression

Random Forest Regression is a meta-estimator that fits a number of decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. It's an extension of the bagging method and utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. At each node during the construction of the tree, a random subset of z features is selected out of the total n features. Moreover, using bootstrapping or other methods, it also selects p examples from the m total available examples at each node. The key parameters of a Random Forest include n_estimators, criterion, max_depth, min_samples_split, and min_samples_leaf. The n_estimators parameter specifies the number of trees in the forest. The criterion parameter determines the function to measure the quality of a split. The max_depth parameter controls the maximum depth of the tree. The min_samples_split and min_samples_leaf parameters manage the minimum number of samples required to split an internal node and to be at a leaf node, respectively.

The R-squared score for :
Train R^2 score: 97.67
Test R^2 score: 83.66

5.5.7- Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.
The R-squared score for :
Train R^2 score: 75.08
Test R^2 score: 74.71

5.5.8- Ridge Regression

Ridge regression, also known as Tikhonov regularization, is a linear regression technique that addresses the problem of multicollinearity (correlated independent variables) by adding a penalty term to the loss function. This penalty term, which is proportional to the squared magnitude of the coefficients, shrinks the coefficients, preventing overfitting and improving the model's generalization ability.
Ridge regression modifies the cost function by introducing an L2 regularization term, which helps control the complexity of the model by penalizing large weights.

The Ridge regression optimizes the following cost function:

$$\text{Loss function} = \underset{\mathbf{w}}{\text{minimize}} \left(\sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2 \right)$$

- $\sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$: The ordinary least squares (OLS) loss (difference between predicted and actual values).
- $\|\mathbf{w}\|_2^2 = \sum_{j=1}^{p} w_j^2$: The L2 norm, which is the sum of the squared coefficients.
- λ: The regularization parameter (also called the ridge penalty), which controls the strength of regularization. A higher λ leads to more shrinkage of the coefficients.

The R-squared score for :
Train R^2 score: 75.07
Test R^2 score: 74.71

5.6 Results

5.6.1- Hyperparameters tuning

For the Random Forest model, we used RandomizedSearchCV from Scikit-Learn for hyperparameter tuning. This method randomly samples combinations of parameters from a predefined range, enabling efficient exploration of the hyperparameter space. The selected configuration was determined based on cross-validation performance, ensuring optimal model accuracy.
RandomizedSearchCV was employed to optimize Random Forest parameters, including:

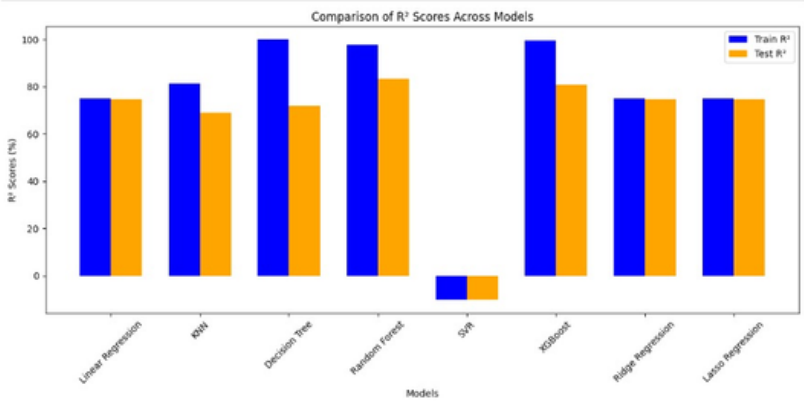
- Number of estimators.
- Tree depth.
- Minimum samples per split and leaf.

5.6.2- Model Performance

The following table summarizes the performance of each tested algorithm, evaluated using the R^2 metric for both training and testing datasets:

Model	Train R^2 Score (%)	Test R^2 Score (%)
Linear Regression	75.08	74.71
KNN	81.44	68.87
Decision Tree	99.9	71.29
Random Forest	97.7	83.53
SVR	-10.22	-10.27
XGBoost	99.5	80.67
Ridge Regression	75.07	74.71
Lasso Regression	75.08	74.71

To complement this table, a visualization of the training and testing R^2 scores for all models was created, providing a clear comparative view of their performance. This graphical representation highlights Random Forest's superior performance while illustrating the disparities between training and testing accuracies for other models.



5.6.3- Algorithm Selection

After testing multiple algorithms, Random Forest was selected as the most suitable model for predicting medical insurance costs. It achieved the highest test R^2 score of 83.53%, indicating excellent predictive accuracy and generalization performance. While Decision Tree and XGBoost also performed well, Random Forest demonstrated a better balance between training and testing performance, minimizing overfitting compared to Decision Tree.

Other models, such as Linear Regression, Ridge, and Lasso, showed consistent but lower performance, with R^2 scores around 74.71% on the test set. These models are simpler and computationally less intensive but failed to capture complex relationships in the data as effectively as ensemble methods.

KNN showed reasonable training performance (81.44%) but had a notable drop in testing accuracy (68.87%), suggesting overfitting. SVR, on the other hand, failed to perform well, with negative R^2 scores indicating poor predictions.

5.8 Limitations

5.8.1- Algorithm-Specific Limitations

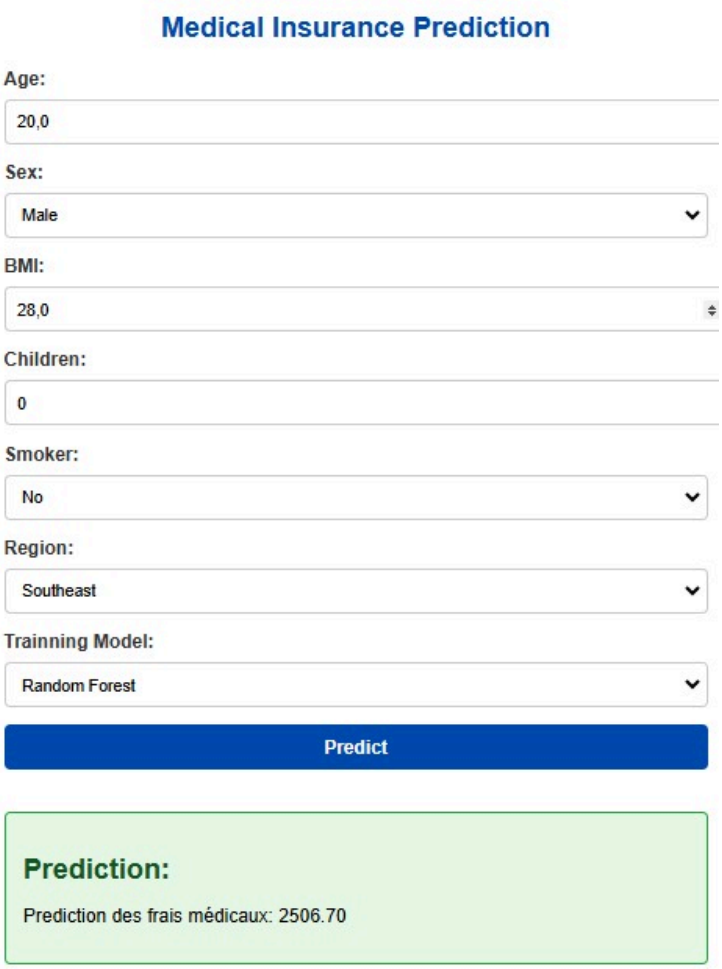
1. Linear Regression:
   - Assumes a linear relationship between features and the target variable, which may not hold true for complex datasets.
   - Sensitive to multicollinearity, leading to unstable coefficient estimates.
2. K-Nearest Neighbors (KNN):
   - Struggles with high-dimensional datasets due to the curse of dimensionality.
   - Sensitive to the choice of (number of neighbors) and the distance metric.
   - Computationally expensive for large datasets, as it requires computing distances for all points.
3. Decision Tree:
   - Prone to overfitting, especially with small datasets, as it tends to create overly complex trees.
   - Sensitive to slight variations in the data, leading to instability.
4. Random Forest:
   - While robust, Random Forest models can be computationally expensive due to the aggregation of multiple trees.
   - May still overfit if the number of trees or maximum depth is not appropriately tuned.
5. Support Vector Regressor (SVR):
   - Poor performance in this project, likely due to inappropriate kernel choice or hyperparameters for this dataset.
   - Not scalable to large datasets as it requires solving a quadratic optimization problem.
6. XGBoost:
   - Computationally intensive, particularly during hyperparameter tuning.
   - Prone to overfitting without careful tuning of regularization parameters.
7. Ridge and Lasso Regression:
   - Limited in capturing complex nonlinear relationships due to their linear nature.
   - Ridge does not perform feature selection, while Lasso may exclude important correlated features.

5.8.2- General Limitations

- Data Size: The dataset comprises only 1338 records, limiting the model's ability to generalize to unseen data.
- Feature Limitations: The dataset lacks critical factors such as pre-existing conditions or lifestyle details, which are key predictors of medical costs.
- Overfitting Risk: Some models, such as Decision Tree and XGBoost, demonstrated signs of overfitting, which can degrade performance on new data.

5.9 Visual Deployment

Additionally, a user-friendly Streamlit web interface was developed to provide stakeholders with a hands-on experience of the model's predictions. This interactive platform allows users to input data and observe the real-time outcomes, enhancing the practical utility of the project. The complete source code, along with relevant links, is provided to ensure transparency and enable further exploration of the project's implementation.



5.10 Extensions and Future Work

- Larger Dataset: Incorporating more data points would improve model training and generalization.
- Additional Features: Including variables such as medical history, diet, and exercise habits could enhance prediction accuracy.
- Deep Learning Models: Exploring neural networks may capture complex nonlinear relationships better.
- Optimization Techniques: Using advanced optimization methods like Bayesian Optimization for hyperparameter tuning could improve model performance.

6. Conclusion and Future Work

This study explored multiple machine learning algorithms to predict medical insurance costs, evaluating their performance through training and testing R^2 scores. Among the tested models, Random Forest emerged as the best performer with a test R^2 score of 83.53%. Its ensemble-based approach effectively captured complex relationships within the data while maintaining generalization. XGBoost also performed well, achieving a test R^2 score of 80.67%, but it demonstrated slight overfitting compared to Random Forest. Simpler models like Linear Regression, Ridge, and Lasso showed consistent but lower predictive accuracy.

Despite the promising results, the study had certain limitations. The dataset's small size and lack of critical features, such as lifestyle or medical history, constrained the models' ability to achieve higher accuracy. Additionally, models like Decision Tree and XGBoost showed susceptibility to overfitting without careful hyperparameter tuning.

Future work should focus on expanding the dataset and incorporating additional predictors to enhance the models' performance. Exploring advanced optimization techniques like Bayesian Optimization and leveraging deep learning methods may also yield better results for more complex datasets. By addressing these limitations, the predictive power and applicability of machine learning in healthcare cost forecasting can be further improved.

References

[1] Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Venkata Nagesh Boddapati, Manikanth Sarisa, Mohit Surender Reddy ‘’An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques ‘’ Journal of Data Analysis and Information Processing, 2024, 12, 581-596

[2] Machine Learning For An Explainable Cost Prediction of Medical Insurance
Ugochukwu Orji, Elochukwu Ukwandu
arxiv

[3] Performance Evaluation of Regression Models in Predicting the Cost of Medical Insurance
Jonelle Angelo S. Cenita, Paul Richie F. Asuncion, Jayson M. Victoriano
arxiv