# Health Insurance Prediction

Aman Rai
*Department of Computer Science and Engineering*
*National Institute of Technology, Warangal*
ar24csm1r05@student.nitw.ac.in

Prof. Pisipati Radha Krishna
*Department of Computer Science and Engineering*
*National Institute of Technology, Warangal*
prkrishna@nitw.ac.in

*Abstract* - **This project develops predictive models for insurance charges using machine learning algorithms, leveraging a comprehensive Kaggle dataset. By analysing demographic and health-related variables, it aims to improve policy pricing strategies, optimize risk management, and enhance prediction accuracy, benefiting insurers and customers in understanding financial commitments and managing risks effectively.**

## I. INTRODUCTION

Healthcare has become a fundamental necessity globally, underscored by the COVID-19 pandemic, which highlighted the critical need for health insurance as a financial safeguard. The rising uncertainties in health, coupled with continuously escalating healthcare costs, make health insurance indispensable in modern life. Consequently, predicting health insurance costs has gained significant importance for individuals and insurance companies alike. Accurate predictions help customers understand potential financial commitments and enable insurance providers to design suitable policies while managing risks effectively.

Predicting health insurance costs is a challenging task that can be approached using various methods, with regression techniques often providing reliable results. In the insurance industry, precise and efficient predictive models are essential to analyse potential financial liabilities and facilitate decision-making for selecting optimal policies. Machine learning (ML) methods have proven particularly useful in handling large, complex datasets, making them well-suited for this domain.

This project aims to develop an ML-based predictive model to estimate individual health insurance costs. The objectives include building and evaluating the model, analysing the impact of different features on insurance charges, and comparing multiple algorithms to identify the most suitable approach. The proposed solution aims to enhance decision-making processes for both customers and insurers by streamlining insurance cost predictions.

## II. LITERATURE SURVEY

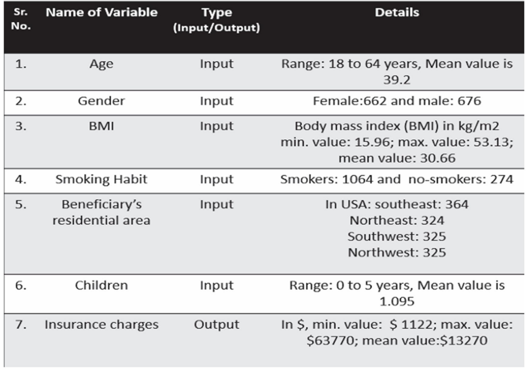
The research paper by Kashish Bhatia and Manish Kumar highlights the potential of machine learning models like linear regression for Insurance Prediction. They achieved an accuracy of over 81% with linear regression. This project aims to build upon their work by incorporating several other regression models to improve accuracy.

Another study of existing machine learning algorithms for handling insurance cost prediction problems highlights the effectiveness of regression-based approaches for modelling such data. However, one of the major hurdles in leveraging medical data lies in the methodological challenges of integrating it for predictive tasks. The demand for more accurate and reliable predictive tools continues to grow, especially for large datasets requiring high precision. Existing research underscores the importance of out-of-sample experiments to validate models, yet many regression studies fail to address this aspect comprehensively.

## III. PROPOSED MODEL

The proposed model leverages advanced machine learning techniques to predict insurance charges accurately. The model selection process involves training and evaluating multiple regression algorithms, including Linear Regression, Decision Tree, Random Forest, XGBoost, and Gradient Boosting. Random Forest, identified as the most effective model during evaluation, is selected due to its robustness in handling non-linear relationships and feature interactions. The model's hyper-parameters are fine-tuned using grid search to optimize performance. To ensure the reliability of predictions, the dataset is split into 80% for training and 20% for testing. Extensive out-of-sample evaluations are conducted to validate the model. The final model achieves a high accuracy of 90.4%, outperforming other algorithms. The proposed system includes a user-friendly interface where users can input height and weight to calculate BMI internally, alongside other features, to predict insurance costs. This model provides a reliable and accurate tool for stakeholders in the insurance domain.

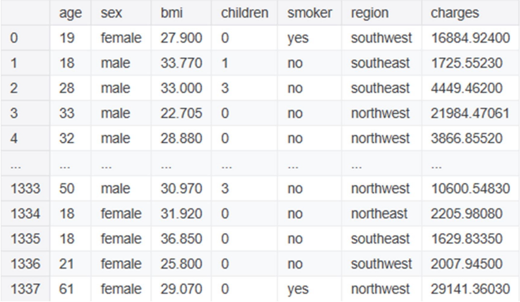
**Dataset**:
We utilize the data from Kaggle named Health Insurance Prediction to solve the insurance prediction task. 1338 observations on insurance costs in four USA regions were gathered. A detailed analysis of the dataset is given below in table 1.



| Sr. No. | Name of Variable | Type (Input/Output) | Details |
|---|---|---|---|
| 1. | Age | Input | Range: 18 to 64 years, Mean value is 39.2 |
| 2. | Gender | Input | Female:662 and male: 676 |
| 3. | BMI | Input | Body mass index (BMI) in kg/m2 min. value: 15.96; max. value: 53.13; mean value: 30.66 |
| 4. | Smoking Habit | Input | Smokers: 1064 and no-smokers: 274 |
| 5. | Beneficiary's residential area | Input | In USA: southeast: 364<br>Northeast: 324<br>Southwest: 325<br>Northwest: 325 |
| 6. | Children | Input | Range: 0 to 5 years, Mean value is 1.095 |
| 7. | Insurance charges | Output | In $, min. value: $ 1122; max. value: $63770; mean value:$13270 |

TABLE I. CHARACTERISTICS OF DATASET USED

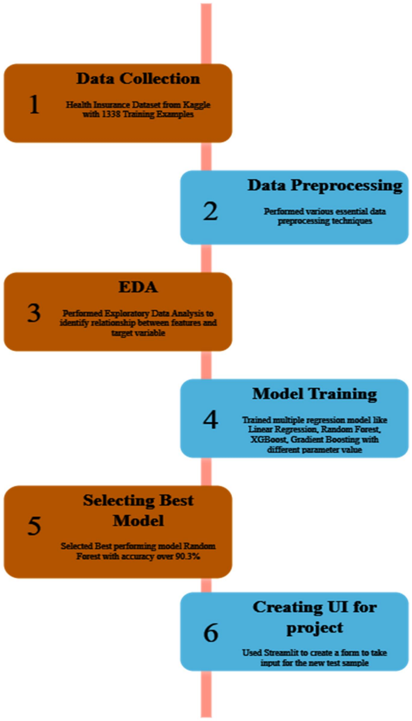
Few entries from the dataset are also shown in figure 1 below. Also, the columns are self-explanatory.



| | 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 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 1333 | 50 | male | 30.970 | 3 | no | northwest | 10600.54830 |
| 1334 | 18 | female | 31.920 | 0 | no | northeast | 2205.98080 |
| 1335 | 18 | female | 36.850 | 0 | no | southeast | 1629.83350 |
| 1336 | 21 | female | 25.800 | 0 | no | southwest | 2007.94500 |
| 1337 | 61 | female | 29.070 | 0 | yes | northwest | 29141.36030 |

Fig.1 Sample of data from the used dataset

**Flow Diagram:**
The project procedure, detailing the steps carried out to ensure the successful completion of the project, is presented in the flow chart below. It encompasses data collection, pre-processing, model selection, training, evaluation, and result interpretation, aiming to achieve accurate predictions and optimize the overall process.


1
Data Collection
Health Insurance Dataset from Kaggle with 1338 Training Examples
2
Data Preprocessing
Performed various essential data preprocessing techniques
3
EDA
Performed Exploratory Data Analysis to identify relationship between features and target variable
4
Model Training
Trained multiple regression model like Linear Regression, Random Forest, XGBoost, Gradient Boosting with different parameter value
5
Selecting Best Model
Selected Best performing model Random Forest with accuracy over 90.3%
6
Creating UI for project
Used Streamlit to create a form to take input for the new test sample

## Exploratory Data Analysis

Exploratory Data Analysis (EDA) serves as a critical step in understanding the dataset and uncovering underlying patterns that influence insurance charges. For this project, the dataset comprises demographic, lifestyle, and health-related features such as age, BMI, smoking status, and region, along with the target variable, insurance charges.

Through EDA, we aim to identify relationships between these features and the target variable, assess data distribution, and detect potential anomalies or outliers. By leveraging visualizations like scatter plots, box plots, and correlation heatmap, we found deeper insights into how individual factors, such as smoking habits or BMI, significantly impact insurance costs. This foundational analysis sets the stage for building predictive models by ensuring a thorough understanding of the data's structure and relationships.

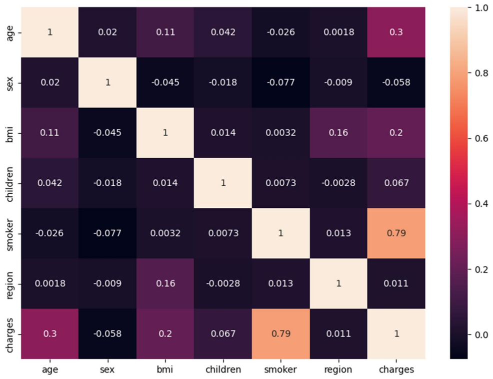


Fig 2. Correlation Heatmap

Key insights from the data analysis include:

**1. Age and Charges:** Positive correlation indicating higher charges for older individuals.

**2. BMI and Charges:** Higher BMI generally results in increased charges, especially for smokers.

**3. Smoker Effect:** Smoking significantly impacts insurance costs, with smokers paying more.
**4. Region:** Minimal influence of the region on charges.

**Training Set:** Allocate 80% of the data to the training set.

**Test Set:** Allocate the remaining 20% of the data to the test set.

## IV. MODELS USED

For the proposed model trained multiple machine learning regression models to identify the most effective technique for achieving optimal results.

**Linear Regression**:
Linear Regression serves as the baseline model to predict insurance charges by assuming a linear relationship between the features and the target variable. It provides interpretable results, highlighting how each feature contributes to the predictions.

**Decision Tree:**
Decision Tree Regression splits the dataset into smaller subsets based on feature values, creating a tree-like structure. It is effective for capturing non-linear relationships and provides an intuitive model that explains how decisions impact predictions.

**Random Forest Regression:**
This ensemble learning method builds multiple decision trees and combines their predictions for robust and non-linear modelling. It is highly effective in capturing feature interactions and handling missing or imbalanced data.

**Gradient Boosting:**
Gradient Boosting iteratively improves the model by reducing errors through boosting weak learners. It is particularly suited for capturing complex relationships and delivers superior predictive performance compared to other methods.

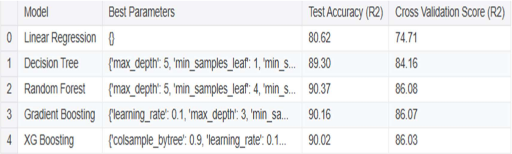
**XGBoost (Extreme Gradient Boosting):**
XGBoost is an advanced implementation of Gradient Boosting that optimizes speed and performance using regularization and parallel processing. It is highly effective for handling large datasets and capturing complex, non-linear relationships

## V. RESULT

After training the above listed models on the given dataset on different parameters with 80:20 data split for training and testing, the following result was achieved

### *A. Model Performance*

The table below summarizes the accuracy of individual models and the ensemble system on the test dataset



| | Model | Best Parameters | Test Accuracy (R2) | Cross Validation Score (R2) |
|---|---|---|---|---|
| 0 | Linear Regression | {} | 80.62 | 74.71 |
| 1 | Decision Tree | {'max_depth': 5, 'min_samples_leaf': 1, 'min_s... | 89.30 | 84.16 |
| 2 | Random Forest | {'max_depth': 5, 'min_samples_leaf': 4, 'min_s... | 90.37 | 86.08 |
| 3 | Gradient Boosting | {'learning_rate': 0.1, 'max_depth': 3, 'min_sa... | 90.16 | 86.07 |
| 4 | XG Boosting | {'colsample_bytree': 0.9, 'learning_rate': 0.1... | 90.02 | 86.03 |

### *B. Observations*

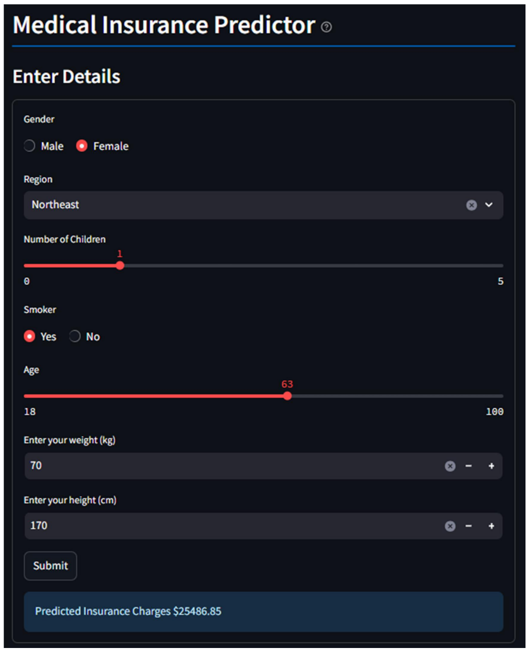
- **Linear Regression:** The accuracy was the least compared to all other models for both testing dataset and cross validation score was also low compared to others.
- **Decision Tree:** Achieved high accuracy, leveraging its ability to model complex patterns
- **Random Forest:** Delivered robust results by aggregating multiple decision trees, effectively minimizing overfitting and emerged as the best model among all.
- **Gradient Boosting:** Performed really well and the accuracy score was second best after the random forest.
- **XGBoost:** Enhanced prediction stability and confidence by combining predictions, which led to a significant improvement in handling challenging cases.

It is evident from the table that the Random Forest Algorithm has worked best on testing dataset; hence we select the Random Forest as our final model.

**User Interface:**

The project incorporates an intuitive System Interaction Framework, developed using Streamlit, to facilitate user interaction. The user interface (UI) was developed to allow users to interact with the system. Since many users are unfamiliar with BMI, the UI provides input fields for height and weight. The system internally calculates the BMI based on these values and uses it, along with gender, smoking habit, region, age, number of children to predict the insurance cost.

The user interface features both dark and light modes, and it also validates the entered details, generating notifications if any information is in an incorrect format.



Medical Insurance Predictor

Enter Details

Gender

○ Male ● Female

Region

Northeast

Number of Children

1

0 5

Smoker

● Yes ○ No

Age

63

18 100

Enter your weight (kg)

70

Enter your height (cm)

170

Submit

Predicted Insurance Charges $25486.85

## VI. CONCLUSION

In conclusion, the developed project demonstrates the power of machine learning in predicting insurance charges with high accuracy. This project analysed various regression-based model on the USA's medical cost personal dataset available on Kaggle. This explores the correlation of various features: Age, BMI, Smoking habits and obesity with overall charges of insurance. Random Forest emerged as the most effective model for this task. Then we also developed a website to easily interact with the system and predict the total cost of insurance. With Random Forest we achieved a testing accuracy of 90.37%.

## VII. FUTURE WORK

**Real-Time Prediction**: Real-time insurance cost predictions by incorporating live data feeds, allowing insurers to provide dynamic pricing updates based on changing customer profiles or market conditions.

**Cross-Industry Data Integration**: Combine healthcare, financial, and lifestyle data to improve predictions, as health and lifestyle factors are often linked to insurance charges.

**Deep Learning Models**: Explore neural networks for better capturing complex feature interactions and patterns.

## VIII. REFERENCES

i. https://www.kaggle.com/code/amanraiji786/insurance-prediction

ii. Machine Learning Resources: Andrew Ng's Coursera1 Course, ISLR textbook.

iii. Bhatia, K., & Kumar, M., "Health insurance cost prediction using machine learning," *2022 4th International Conference on Smart Electronics and Communication (ICOSEC)*, 2022, pp 124-129, doi: 10.1109/ICOSEC55071.2022.9824201.

iv. M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation," AMIA Annual Symposium Proceedings, vol. 2017, p. 1312, 2017.