## Front End Engineering-II /Artificial

## Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Title of the Project

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Ms. Shagun Sharma Kritica, 2210990523 (17)

Kritika Bansal, 2210990524

Lavika Singla, 2210990540

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

## Abstract

Accurately determining medical insurance premiums is a critical yet complex task for insurance companies, as it directly influences their profitability, risk management, and customer satisfaction. Traditional actuarial methods, while foundational, often fall short in capturing the intricate relationships between various demographic and lifestyle factors such as age, gender, BMI, number of children, smoking habits, and geographical region. This project aims to leverage the power of machine learning, specifically a Random Forest regression algorithm, to develop a more sophisticated and precise model for predicting medical insurance premiums.

Our approach begins with a comprehensive exploratory data analysis to understand the distribution and interrelationships of the variables within the dataset. This step is followed by meticulous feature engineering to create more informative and predictive features from the existing data. By converting categorical variables into numerical formats and exploring potential interactions between variables, we enhance the model’s ability to accurately capture the underlying patterns in the data.

The core of our solution involves training a Random Forest regression model, chosen for its robustness and ability to handle complex, non-linear relationships. We split the dataset into training and testing sets to evaluate the model’s performance using metrics such as R-squared, ensuring its accuracy and generalization capabilities.

Upon achieving a reliable and accurate model, we deploy it to provide a practical tool for insurance companies. This tool can assist in making informed decisions regarding premium pricing, ensuring that premiums are both fair and reflective of individual risk profiles.

Ultimately, this project demonstrates the significant potential of machine learning in transforming traditional insurance practices, offering a modern, data-driven approach to premium prediction that benefits both insurers and policyholders. By addressing the limitations of conventional methods and harnessing the capabilities of advanced algorithms, we aim to contribute to a more efficient, equitable, and transparent insurance industry.

## Table of Contents

|  |  |  |
| --- | --- | --- |
| S. No. | Page Title | Page No. |
| 1. | Introduction | 4-7 |
| 2. | Problem Definition and Requirements | 8 |
| 3. | Proposed Design / Methodology | 9-11 |
| 4. | Results | 12-21 |

**1. Introduction**

Health insurance premiums are a significant expense for individuals and families. Accurately predicting these premiums can help consumers make informed decisions and allow insurance providers to optimize their pricing strategies.

The healthcare industry heavily relies on accurate risk assessment and pricing strategies to ensure the financial sustainability of insurance companies while providing affordable coverage to policyholders. One of the critical challenges in this domain is the precise estimation of medical insurance premiums, which are influenced by various demographic and lifestyle factors such as age, gender, BMI, number of children, smoking habits, and geographical region. Inaccurate premium calculations can lead to financial imbalances for insurers and unfair pricing for consumers.

To address this challenge, we propose developing a robust machine learning model to predict medical insurance premiums based on these influential factors. By leveraging historical data and advanced algorithms, our goal is to create a tool that enhances the accuracy, fairness, and transparency of premium determination. This project involves comprehensive data analysis, feature engineering, and model training using a Random Forest regression algorithm to provide a reliable solution for predicting insurance premiums. The resulting model aims to assist insurance companies in making informed decisions, managing risks effectively, and offering equitable insurance products to their customers.

**1.1 Objective**

The primary objective of this project is to develop a predictive model that accurately estimates medical insurance premiums using a dataset containing various demographic and lifestyle factors. The specific goals include:

1.1.1 Conducting thorough exploratory data analysis to understand the distribution and relationships within the data.

1.1.2 Performing feature engineering to enhance the predictive capabilities of the model.

1.1.3 Training and optimizing a Random Forest regression model to predict insurance premiums.

1.1.4 Evaluating the model's performance using appropriate metrics to ensure its accuracy and generalization.

1.1.5 Deploying the model to provide a reliable tool for insurance companies to assess risks and determine fair premiums.

**1.2 Background**

**The determination of medical insurance premiums is a complex process influenced by multiple factors that represent the risk profile of policyholders. Insurance companies traditionally rely on actuarial methods to estimate these premiums, but such approaches can sometimes be limited in their ability to capture intricate patterns and interactions among variables. With the advent of machine learning, there is an opportunity to enhance this process by utilizing advanced algorithms that can analyse large datasets and uncover deeper insights.**

**This project focuses on using a Random Forest regression algorithm, known for its robustness and ability to handle various types of data, to predict insurance premiums. The dataset includes critical factors such as age, gender, BMI, number of children, smoking status, and region, which are known to affect health risks and insurance costs. By building a predictive model based on these features, we aim to improve the precision and fairness of premium calculations, thereby supporting better risk management and pricing strategies in the insurance industry. This initiative not only promises to benefit insurance companies by optimizing their pricing models but also aims to ensure that customers receive fair and transparent premiums based on their individual risk profiles.**

**1.3 Significance:**

**The development of a machine learning model for predicting medical insurance premiums holds significant value for both insurance companies and policyholders.**

**1.3.1. Improved Accuracy and Fairness**

**Traditional methods of determining insurance premiums can be imprecise and may not fully capture the complex interplay between various risk factors. A machine learning model, trained on comprehensive historical data, can significantly enhance the accuracy of premium predictions. This leads to fairer pricing, ensuring that policyholders are charged premiums that truly reflect their individual risk profiles.**

**1.3.2. Enhanced Risk Management**

**Accurate prediction models enable insurance companies to better assess and manage risk. By understanding the risk associated with each policyholder more precisely, insurers can make informed decisions about coverage options and premium levels, ultimately leading to a more stable and sustainable business model.**

**1.3.3. Increased Transparency**

**Transparency in how premiums are calculated can build trust between insurers and policyholders. A data-driven model provides clear and objective criteria for premium determination, making it easier to explain and justify premiums to customers, thereby enhancing customer satisfaction and trust.**

**1.3.4. Operational Efficiency**

**Automating the premium prediction process with a machine learning model reduces the time and resources required for manual calculations and adjustments. This increases operational efficiency, allowing insurance companies to allocate resources to other critical areas such as customer service and product development.**

**1.3.5. Scalability and Adaptability**

**A machine learning model can be easily updated with new data, allowing insurance companies to adapt to changing trends and emerging risk factors. This ensures that the premium prediction system remains relevant and accurate over time, providing a scalable solution that grows with the business.**

**1.3.6. Policyholder Benefits**

**For policyholders, accurate premium predictions mean that they are less likely to be overcharged for their insurance coverage. Fair and transparent pricing can lead to better customer experiences and increased loyalty, as customers feel they are being treated equitably.**

**1.3.7. Regulatory Compliance**

**In many regions, regulatory bodies require that insurance premiums be calculated based on objective and fair criteria. A machine learning model provides a clear, data-driven basis for premium calculations, helping insurance companies comply with regulatory standards and avoid potential legal issues.**

**In summary, the significance of this project lies in its potential to transform the insurance premium calculation process, making it more accurate, fair, transparent, and efficient. This not only benefits insurance companies by enhancing their risk management and operational efficiency but also provides tangible benefits to policyholders through fairer pricing and increased trust in the insurance system.**

## 2. Problem Definition and Requirements

**2.1 Problem Definition**

The challenge in the medical insurance industry is the accurate prediction of insurance premiums based on various risk factors. Traditional actuarial methods used by insurance companies often do not fully capture the complex interactions between different demographic and lifestyle variables, leading to potential inaccuracies in premium calculations. These inaccuracies can result in financial risks for insurers, unfair pricing for consumers, and a lack of transparency in the insurance process. The primary goal of this project is to develop a robust and accurate machine learning model that can predict medical insurance premiums based on a comprehensive dataset that includes factors such as age, gender, BMI, number of children, smoking status, and region.

**2.2 Requirements**

**Software:** Python, Jupyter Notebook, Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Streamlit, Pickle

**Hardware:** A machine with at least 8GB RAM and a multi-core processor

**Dataset:** Insurance dataset from Kaggle

## 3. Proposed Design / Methodology

**3.1 Data Collection and Preprocessing:** Gather historical data on medical insurance premiums along with demographic and lifestyle factors such as age, gender, BMI, number of children, smoking habits, and region. Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features if necessary.

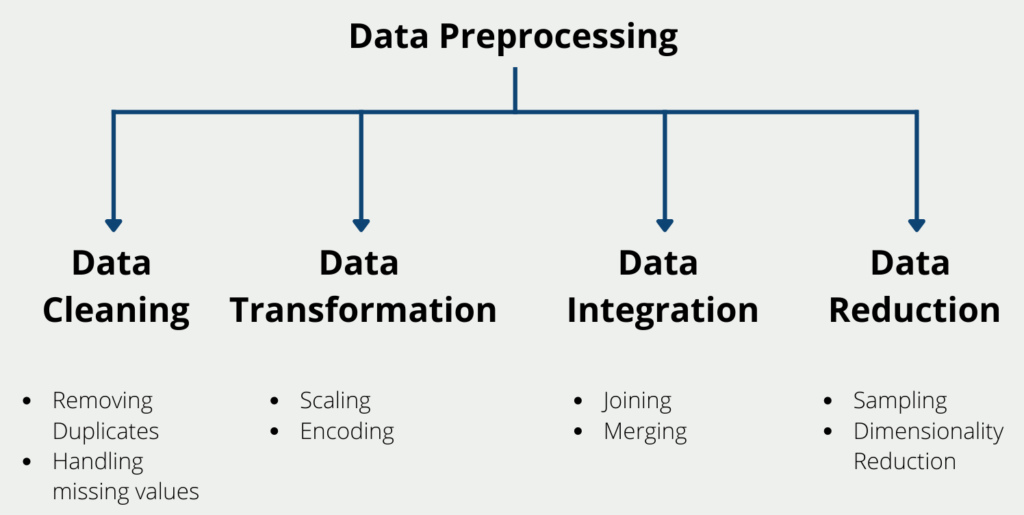


Fig. 1 Methods Of Data Preprocessing

**3.2 Exploratory Data Analysis (EDA):** Perform exploratory data analysis to gain insights into the distribution of variables, identify correlations between features and the target variable (insurance premiums), and understand any patterns or trends in the data.

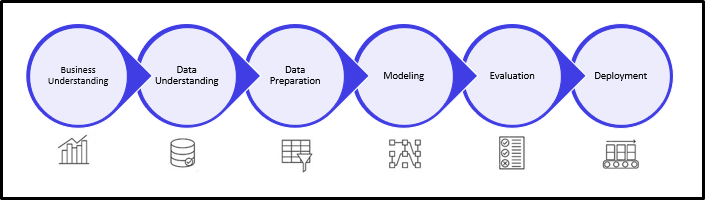


Fig. 2 Steps in EDA

**3.3 Feature Engineering:** Extract meaningful features from the data or create new features that may enhance the model's predictive power. This may involve feature transformation, binning, or interaction terms.

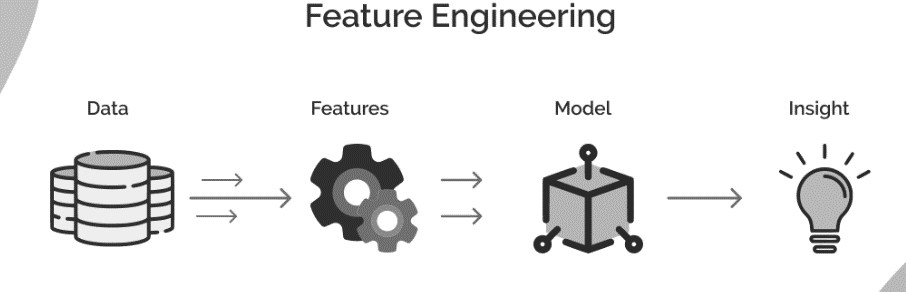


Fig. 3 Steps in Feature Engineering

**3.4 Model Training:** Split the data into training and testing sets. Train a Random Forest regression model using the training data.

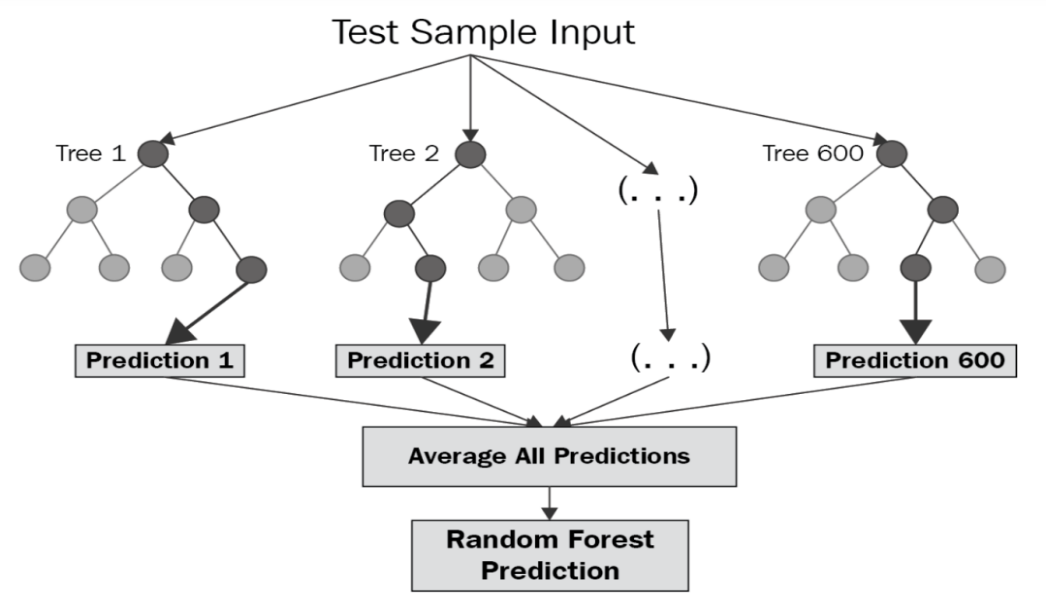


Fig. 4 Random Forest Model Training

**3.5 Model Evaluation:** Evaluate the trained model using appropriate metrics such as R-squared (coefficient of determination) to assess its predictive performance on unseen data. Conduct cross-validation to ensure the model's robustness and generalization ability.

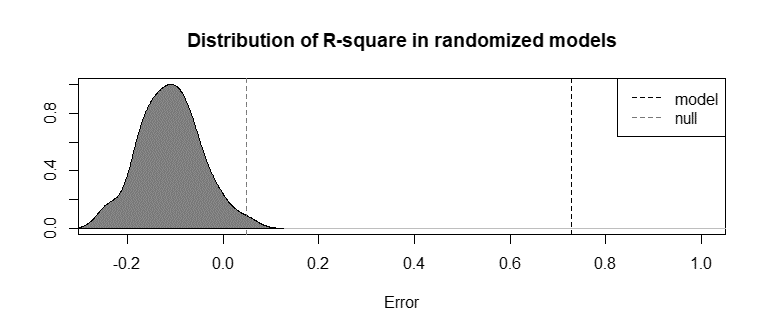


Fig. 5 Model Evaluation Using R-Squared

**3.6 Prediction and Deployment:** Use the trained model to make predictions on new data points. Deploy the model into production either as a standalone application or as part of a larger system where it can be utilized to predict insurance premiums in real-time.

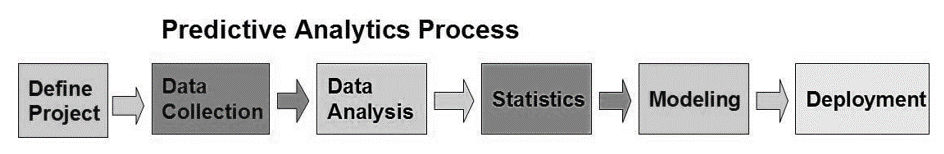


Fig. 6 Predictive Analytic Process

**4. Results**

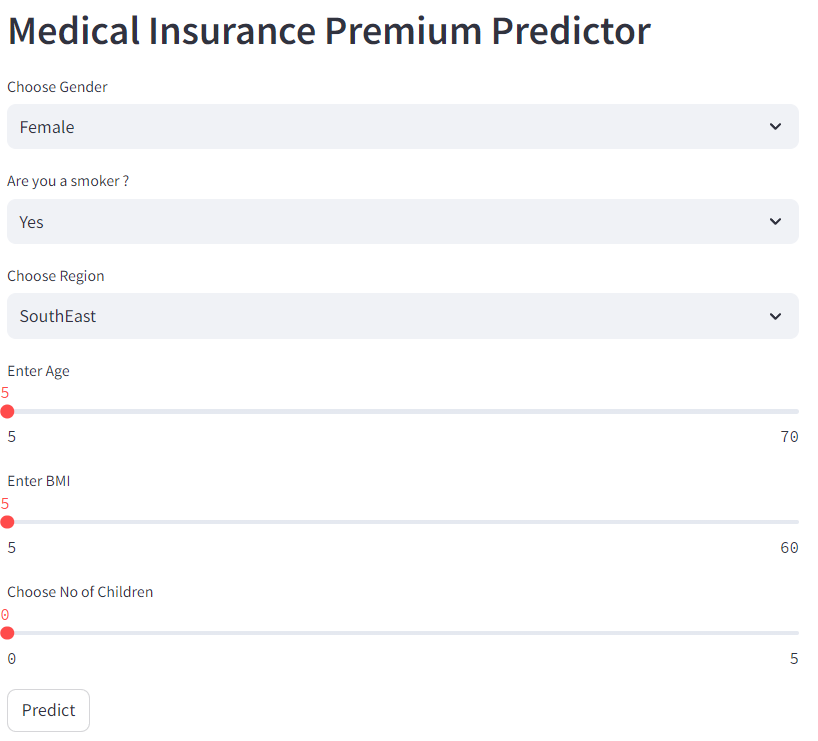
****

Fig. 7 Web Interface Of The Project

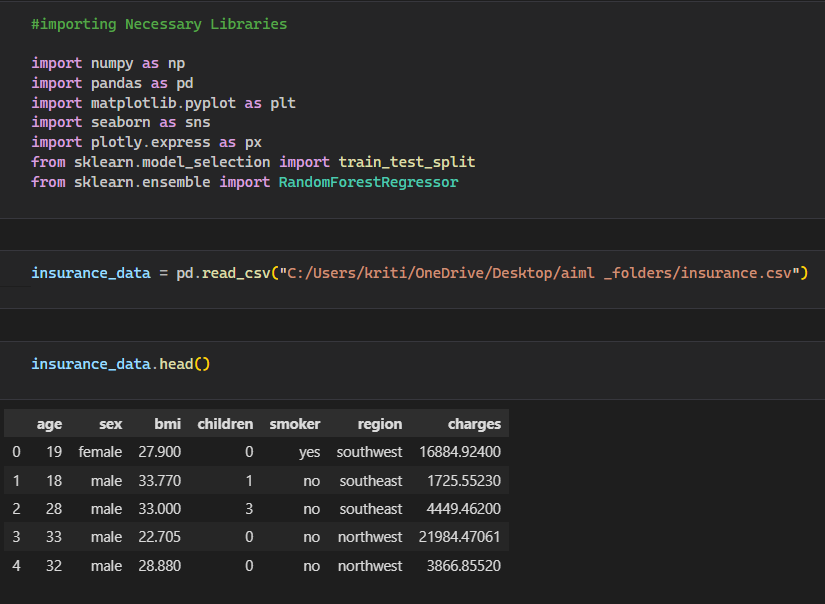
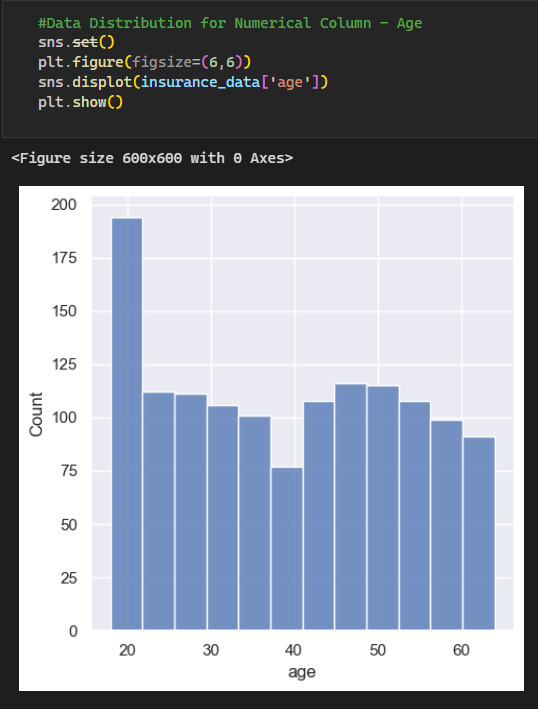
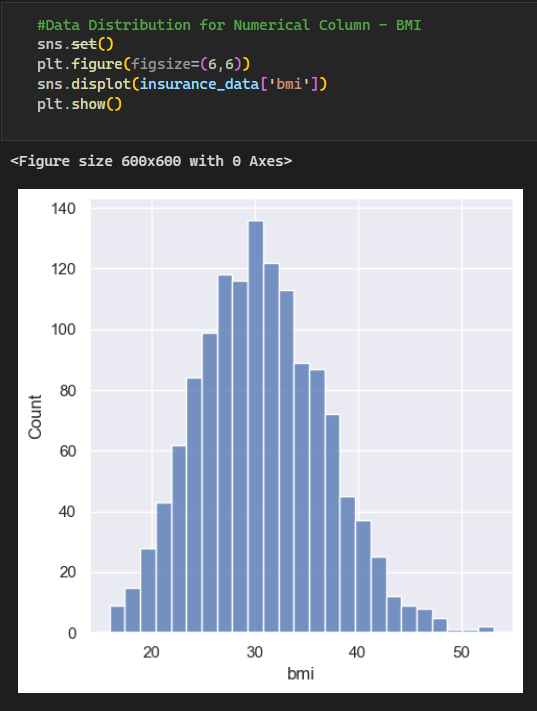
****

Fig. 8 Libraries And Dataset

****

Fig. 9 Histogram showing No. Of Smokers

****

****

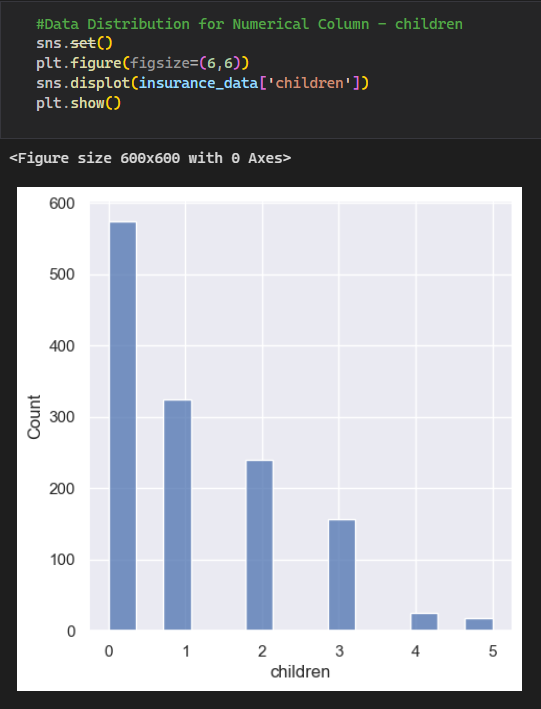
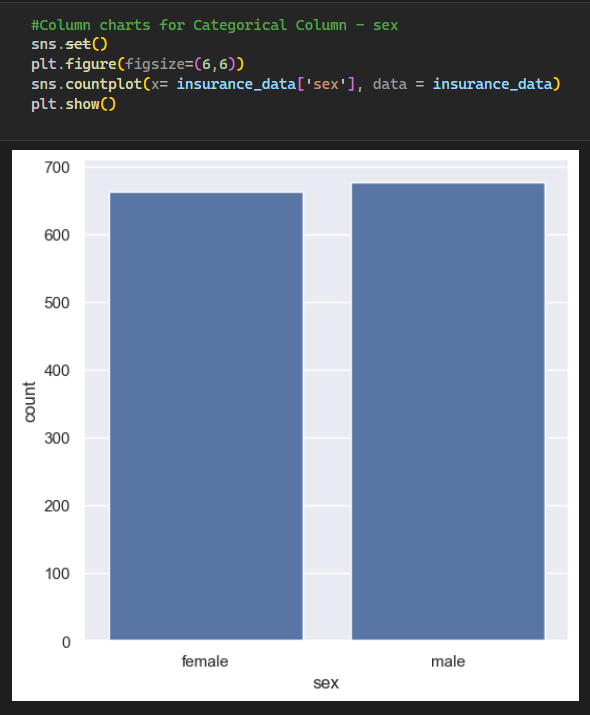
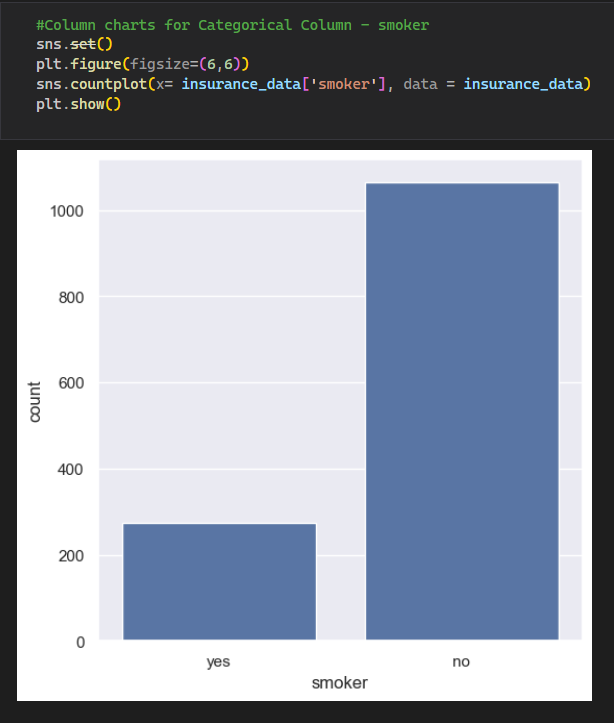
****

Fig. 11 Graphs showing Data Distribution For Numerical Column

****

****

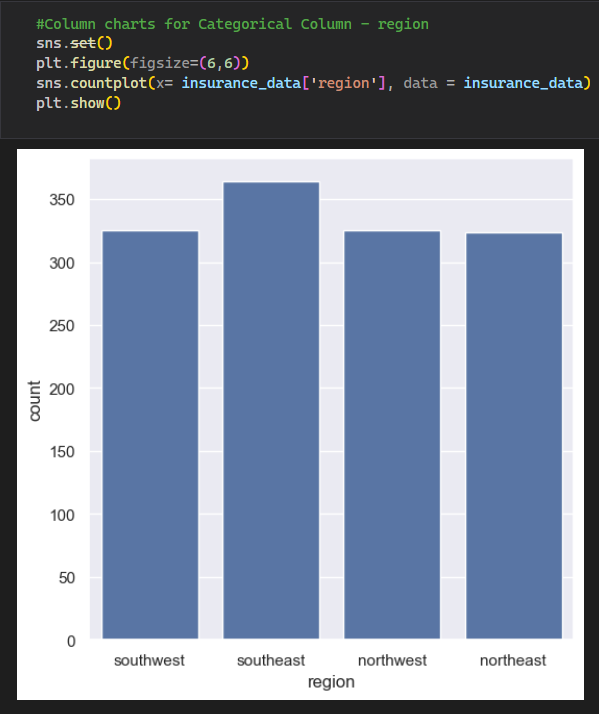
****

Fig. 12 Graphs showing Column Charts For Categorical Column

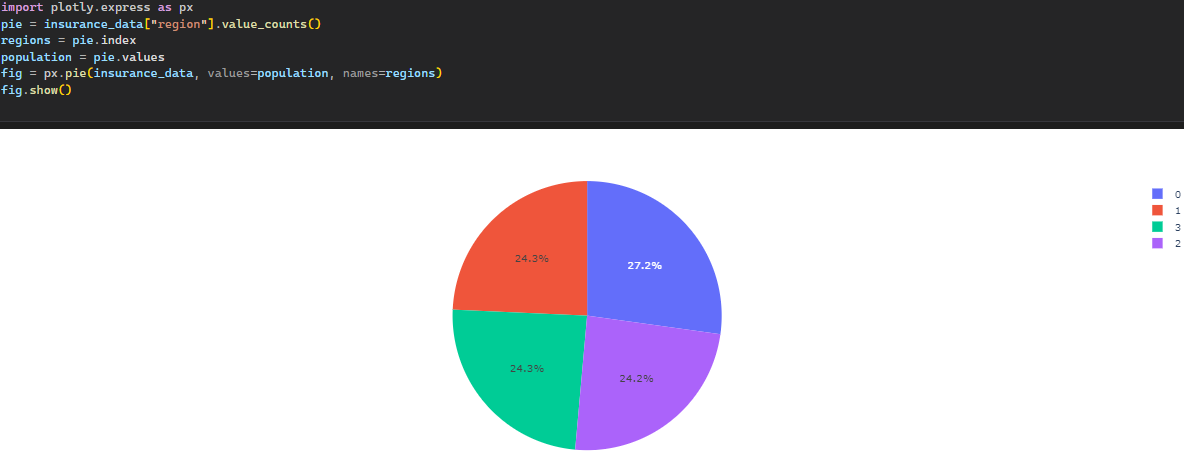
****

Fig. 13 Pie Chart Showing Percentage Of Population In Different Regions

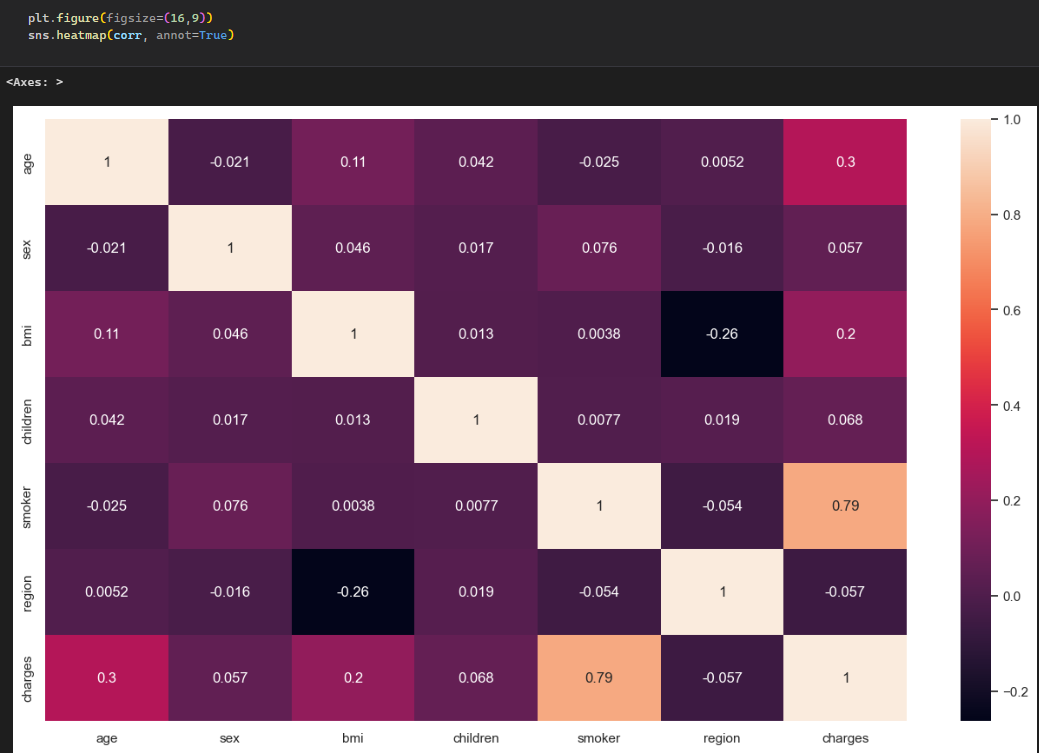
****

Fig. 14 Heatmap Showing Correlation Coefficient between Two Variables

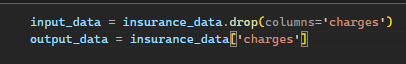
****

Fig. 15 Separating Input Data And Output Data

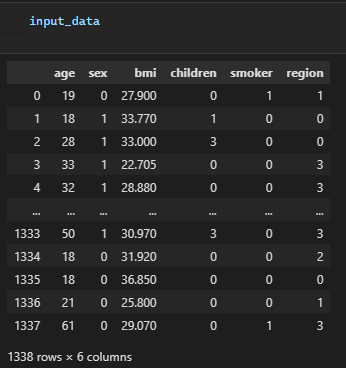
****

Fig. 16 Showing Input Data

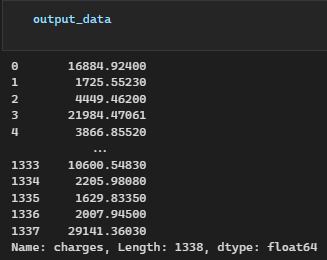
****

Fig. 17 Showing Output Data

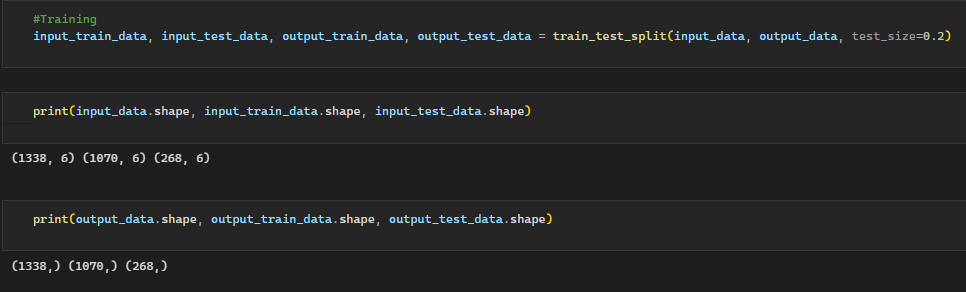
****

Fig. 18 Splitting The Model

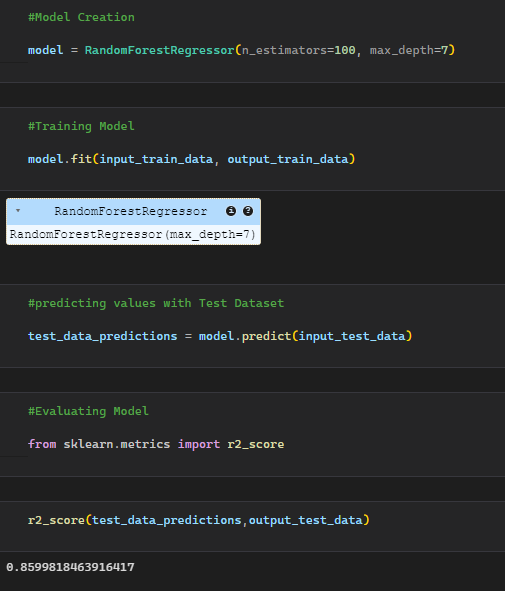
****

Fig. 19 Training The Model And Evaluating The Metrics

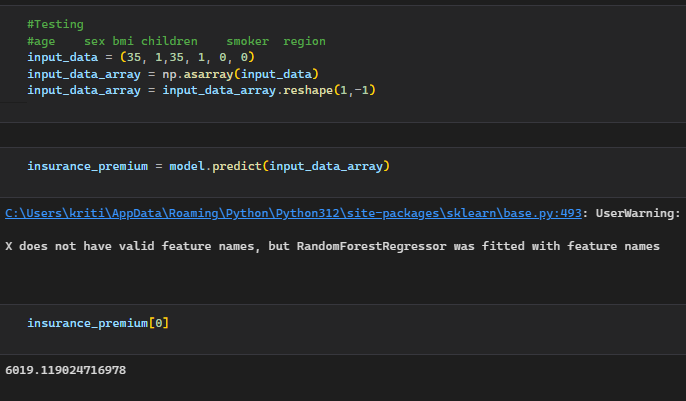
****

Fig. 20 Testing The Model With User Input

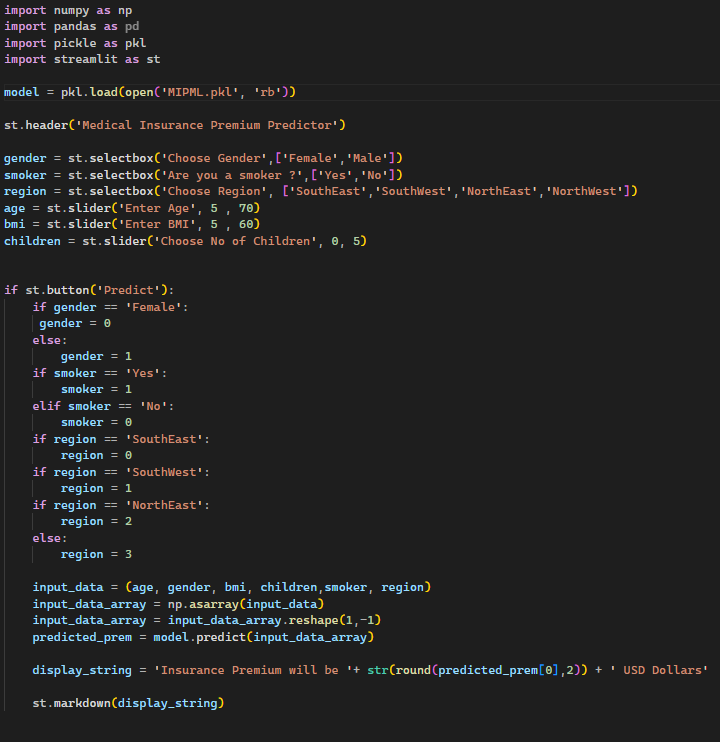
****

Fig. 21 Python File To Build Web Interface

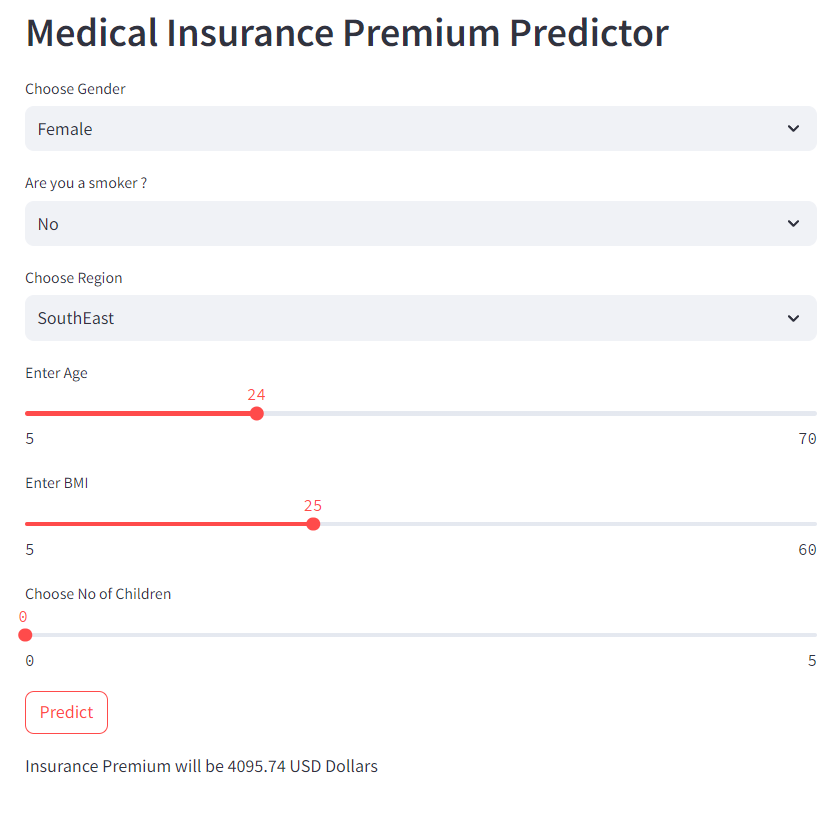
****

Fig. 22 Model Predicting The Value According To The User Input

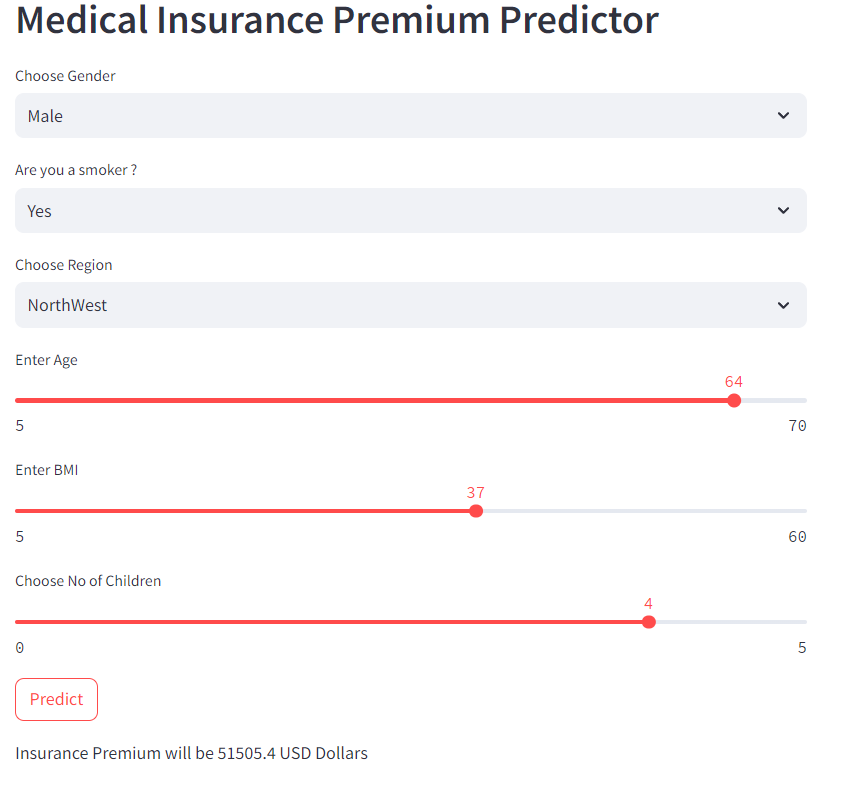
****

Fig. 22 Model Predicting The Value According To The User Input

**References**

1. [**https://www.kaggle.com/**](https://www.kaggle.com/)
2. [**https://www.geeksforgeeks.org/**](https://www.geeksforgeeks.org/)
3. [**https://www.datacamp.com/**](https://www.datacamp.com/)
4. [**https://www.jmlr.org/**](https://www.jmlr.org/)
5. [**https://www.coursera.org/**](https://www.coursera.org/)
6. [**https://scholarspace.manoa.hawaii.edu/**](https://scholarspace.manoa.hawaii.edu/)
7. [**https://www.youtube.com/**](https://www.youtube.com/)