Health Insurance Cost Prediction Using Machine Learning

Final Report

Yahan Madhuhansa

1. Introduction

In this project, I aim to develop a machine learning model to predict individual medical insurance costs based on various personal and lifestyle factors. By analyzing historical data, I can identify patterns and key factors that influence insurance charges.

Problem Definition

Insurance companies determine premiums based on multiple factors, including age, BMI, smoking habits, and geographical location. However, predicting these costs manually can be inefficient and inaccurate. Machine learning can automate this process, improving efficiency and accuracy.

Objective

The objective of this machine learning model is to predict the medical insurance cost (charges) for an individual based on features such as age, BMI, number of dependents, smoking status, and region. The model will help insurance providers and policyholders understand cost expectations and optimize premium pricing.

2. Dataset Description

Source and Size

The dataset is sourced from GitHub, originally compiled for the book *Machine Learning with R* by Brett Lantz. It contains **1,338** records with **7 features** and **1 target variable** (charges).

Features

- **age**: Age of the primary beneficiary.
- **sex**: Gender of the insurance holder (male or female).
- **bmi**: Body Mass Index, indicating relative weight to height.
- **children**: Number of dependents covered by insurance.
- **smoker**: Whether the beneficiary is a smoker (yes or no).
- **region**: Residential area in the US (northeast, southeast, southwest, northwest).
- **charges** *(Target Variable)*: Individual medical insurance cost.

3. Step of project

Understand the data:

- check the missing values
- understand Object data types
- Removed duplicates
- Label Encode objects columns (sex and smoker)

EDA

- Visualize data to understand relations

Feature engineering

- One hot encoded region column
- Maked 2 new columns (overweight_smoker, normal_Notsmoker)
- Do log transformation for charges

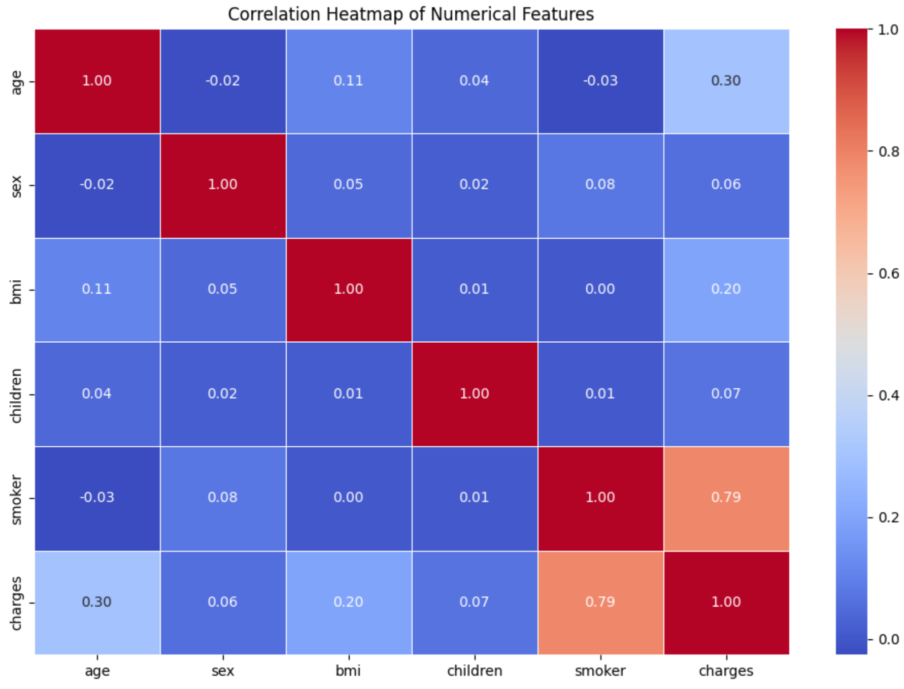
Model selection and training

- Try linear Regression
- Try random forest
- Save the model

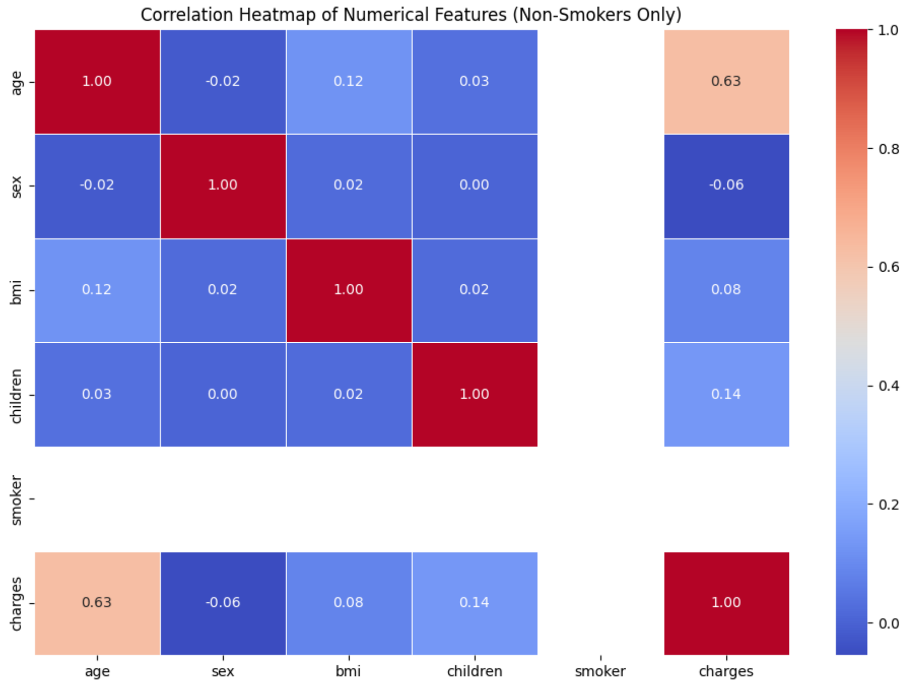
4. Results & Discussion

i. EDA Discussion

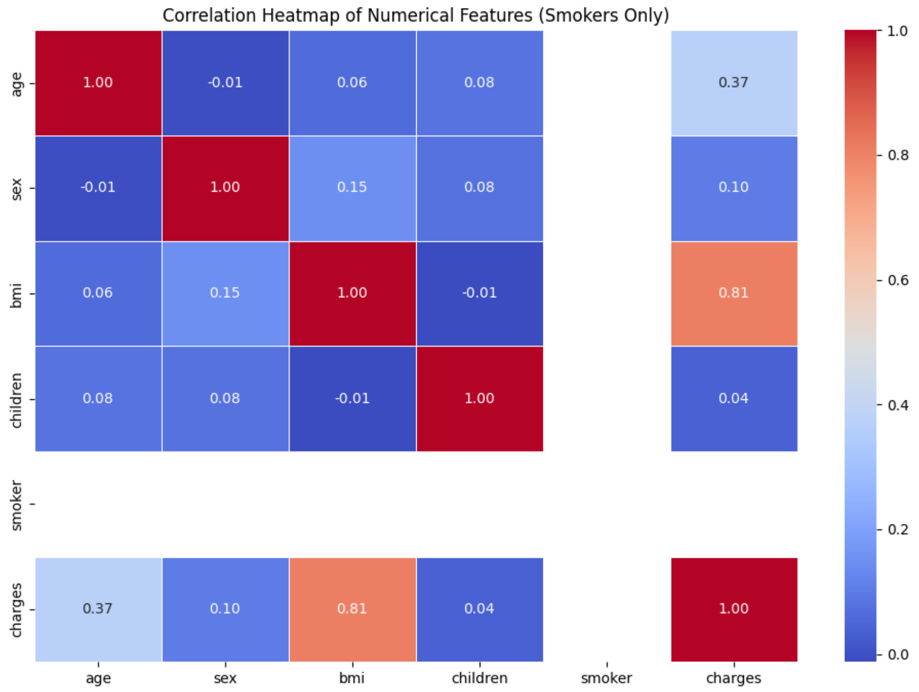
After the EDA I found that smoking has relationship between charges



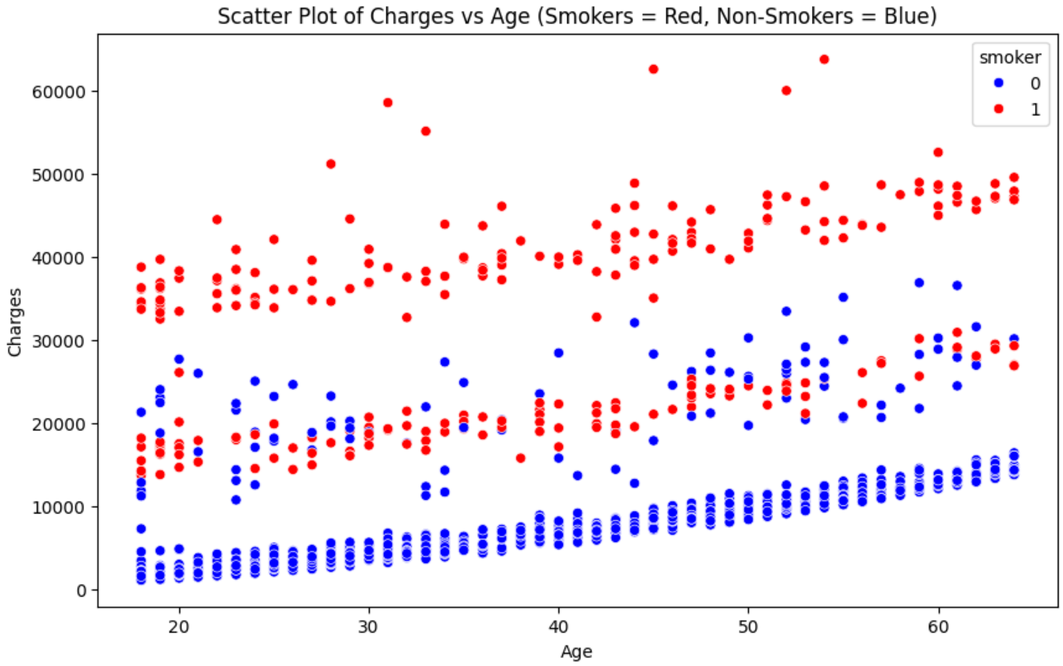
It seem like other attributes have no relationship with charges, but I found that **non-smokers** has a relationship with age


Correlation Heatmap of Numerical Features (Non-Smokers Only)

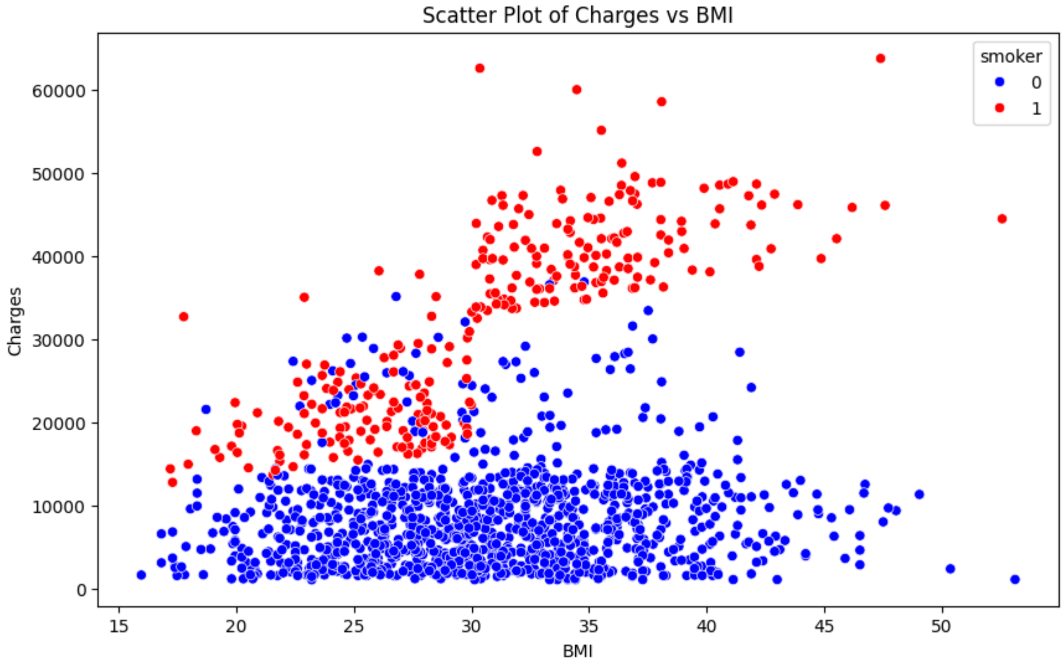
Also I found that smoker has a relationship with BMI


Correlation Heatmap of Numerical Features (Smokers Only)

Both sentences can visualize like this

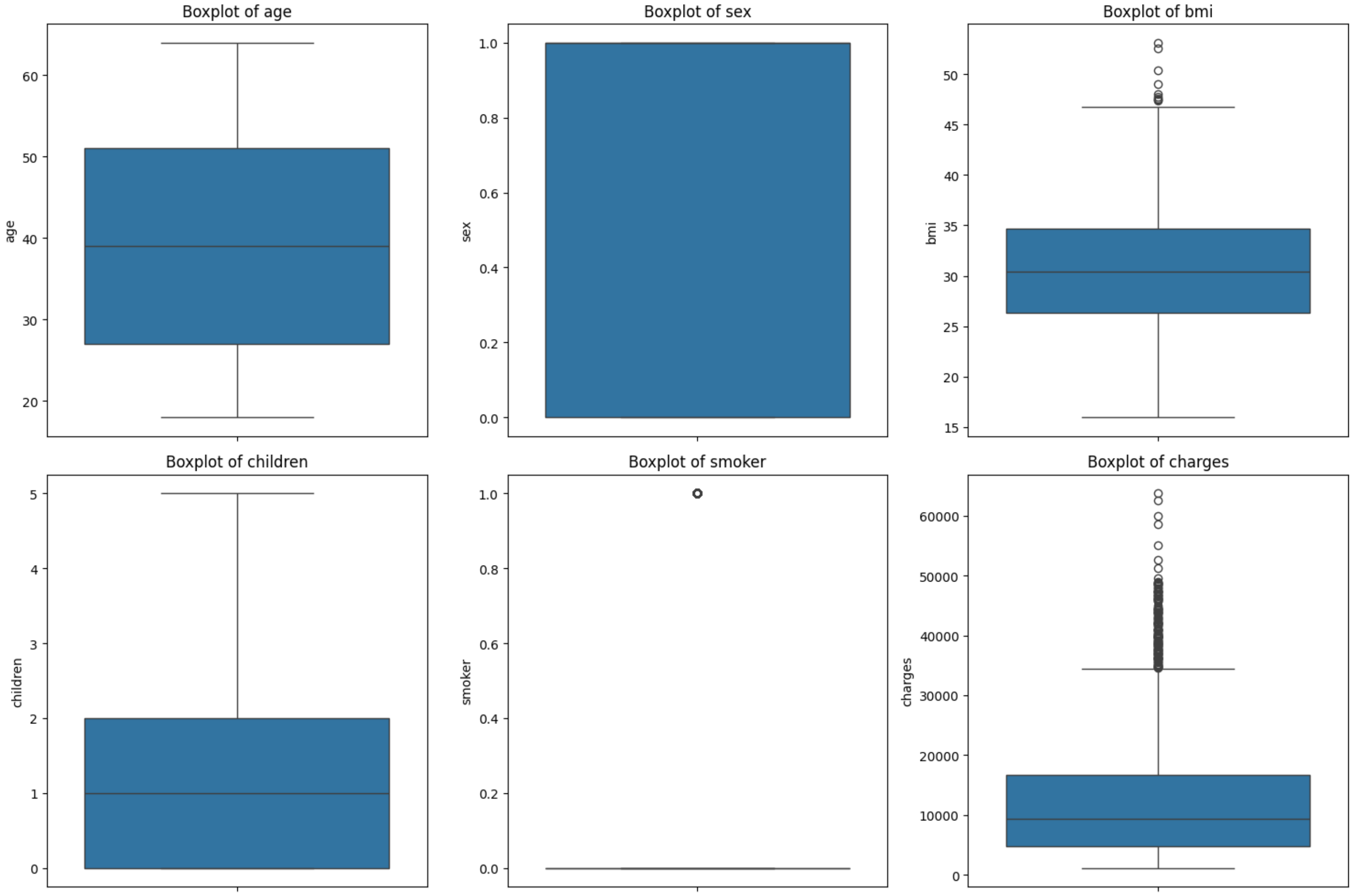


Scatter Plot of Charges vs Age (Smokers = Red, Non-Smokers = Blue)



Scatter Plot of Charges vs BMI

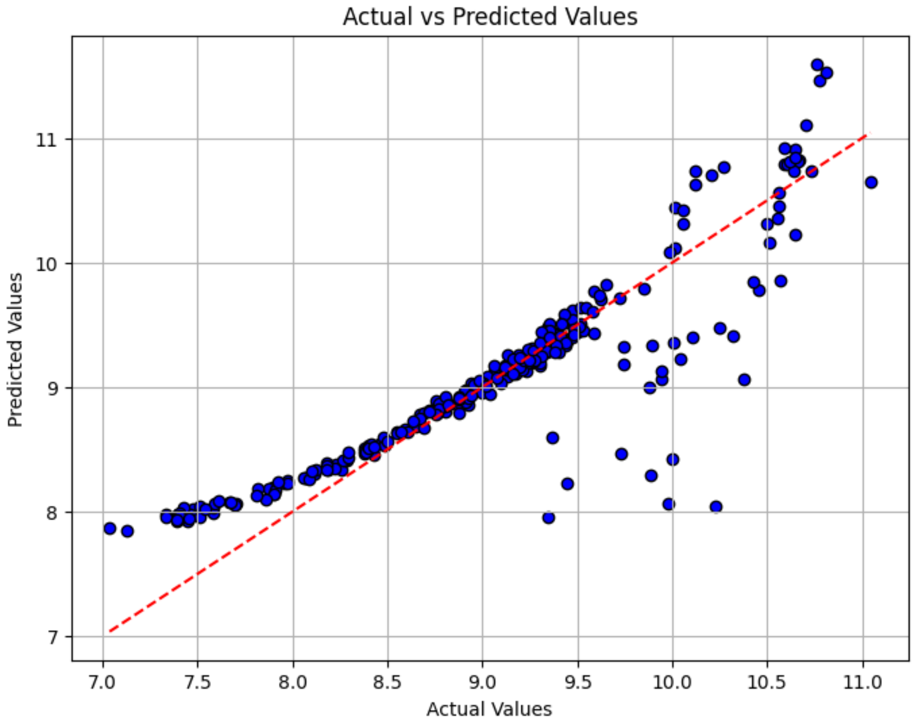
Also there is outliers show in boxplot but in case these outliers is useful to make new features in features Engineering. Those features are overweight_smoker and normal_Notsmoker



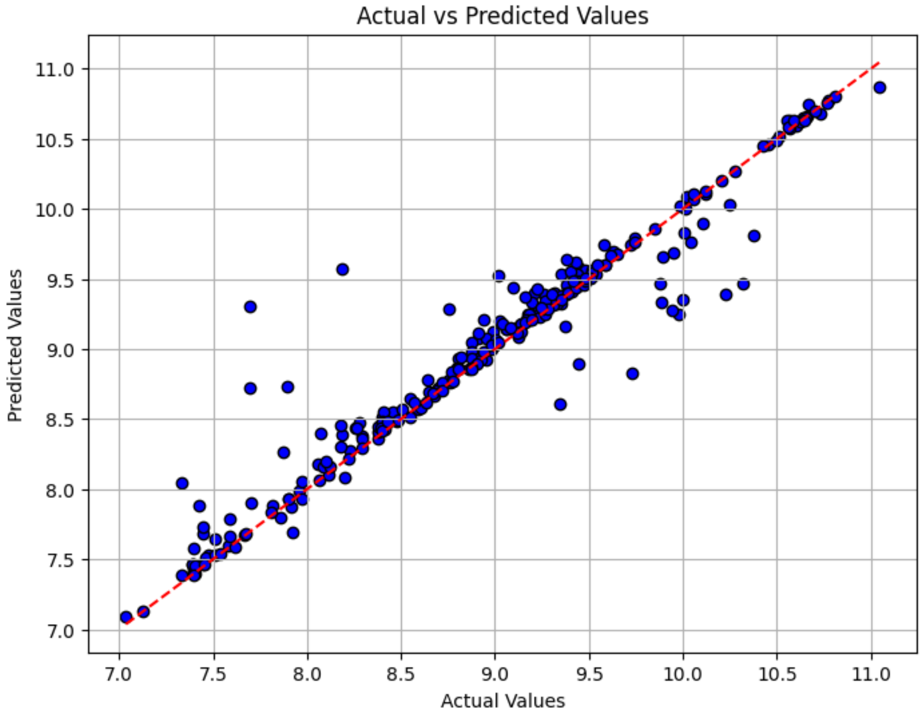
ii. Model performance

Model	R2
Linear registration	0.8435740541402428
Random forests	0.8407455282540128

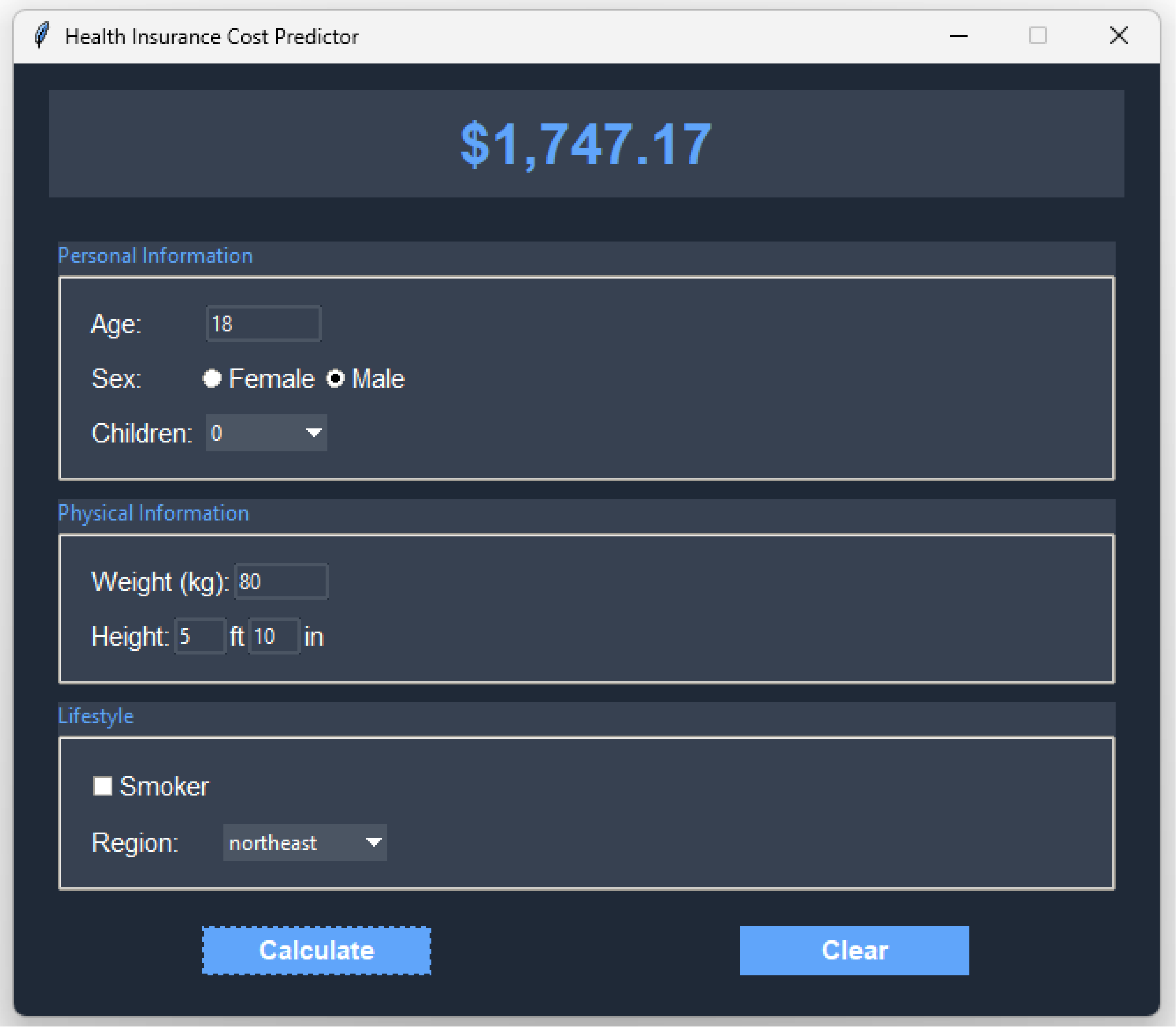
Linear regression



Random forest



iii. Final User Interface using Tkinter



$1,747.17

Age: 18

Sex: ● Female ◉ Male

Children: 0

Weight (kg): 80

Height: 5 ft 10 in

□ Smoker

Region: northeast

Conclusion

This project successfully developed a machine learning model to predict medical insurance costs using demographic and lifestyle factors like age, BMI, and smoking status. Exploratory data analysis (EDA) revealed that **smoking** was the most influential factor, with smokers incurring significantly higher charges. Feature engineering, including log transformation of the target variable (charges) and creation of derived features (e.g., overweight_smoker), improved model performance. The tested algorithms, **Linear Regression** and **Random Forest** (R2: **0.84** , 0.84). The project culminated in a **Tkinter-based UI**, enabling users to input their details and receive instant cost estimates. Future enhancements could integrate more granular health data or deploy the model as a web application for broader accessibility