Health Insurance Premium Prediction

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Team : 10

Name : Alin Anand

ICT ACADEMY OF KERALA
THIRUVANANTHAPURAM, KERALA, INDIA
Sep 2024

List of Figures

1	Methodology	7
2	Data Summary	9
3	Data shape	10
4	Target variable distribution	10
5	Outlier detection	11
6	Correlation	12
7	Modeling	13
8	Model evaluation line plot	14
9	Model evaluation scatter plot	14
10	Input page1	18
11	Input page2	19
12	Output page	19

List of Abbreviations

Abbreviation	**Full form**
EDA	Exploratory Data Analysis
SVR	Support Vector Regression
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
R2	R2 Score

Table of Contents

Content	Page no
Abstract	5
Problem definition	6
Introduction	7-15
Literature survey	16-17
Result	18-19
Conclusion	20
References	21-22

Abstract

In the rapidly evolving landscape of healthcare insurance, accurately predicting medical insurance premiums is crucial for insurers aiming to optimize pricing strategies and for policyholders seeking cost-effective coverage. This study explores the application of machine learning techniques to predict medical insurance premiums based on a comprehensive dataset comprising 25000 entries with 24 diverse features. The dataset encompasses a wide array of variables such as age, gender, and income, as well as lifestyle factors like smoking habits, exercise frequency, and alcohol consumption. Additionally, medical history elements such as pre-existing conditions, average glucose levels, and body mass index are incorporated to enhance the predictive power of the models.

The research methodology involves a meticulous approach to data preprocessing, including handling missing values and feature scaling, followed by exploratory data analysis to uncover correlations and patterns. Various machine learning algorithms are employed to construct predictive models. The performance of these models is rigorously evaluated using metrics such as RMSE, MAE, and R-squared, with hyperparameter tuning and cross-validation techniques applied to achieve optimal results.

The findings of this study highlight the significance of lifestyle and medical factors in determining insurance premiums, offering valuable insights into the risk assessment process. The predictive models demonstrate promising accuracy, showcasing the potential of machine learning in transforming insurance pricing mechanisms. Insurers can leverage these insights to develop more precise and personalized premium calculations, thereby enhancing customer satisfaction and retention. Furthermore, the models provide policyholders with a transparent understanding of how their health and lifestyle choices impact their insurance costs, empowering them to make informed decisions regarding their coverage.

1. Problem Definition

1.1 Overview

The objective is to develop a predictive model that estimates health insurance premiums based on a wide range of factors related to the patient's demographics, health status, and lifestyle. This model will help insurance companies set premiums more accurately and assist customers in understanding the factors that influence their insurance costs.

1.2 Problem Statement

The problem is framed as a regression task, where the goal is to predict a continuous variable patient insurance cost based on the various input features.

Challenges:

1. **High Dimensionality**: The dataset has 24 features, some of which may be highly correlated or irrelevant to the target variable. Feature selection or dimensionality reduction may be necessary.
2. **Data Distribution**: Some features may have skewed distributions (e.g., medical expenses, age), which could affect the performance of the model.
3. **Categorical Encoding**: Proper encoding of categorical variables (e.g., gender, smoking status) is essential for ensuring that the model can correctly interpret these variables.
4. **Complex Interactions**: There might be complex interactions between lifestyle, health history, and demographic features that influence insurance costs. Capturing these interactions will be key to building an accurate model.
5. **Outliers and Anomalies**: Outliers in medical expenses, BMI, or other variables could skew predictions and need to be carefully handled.

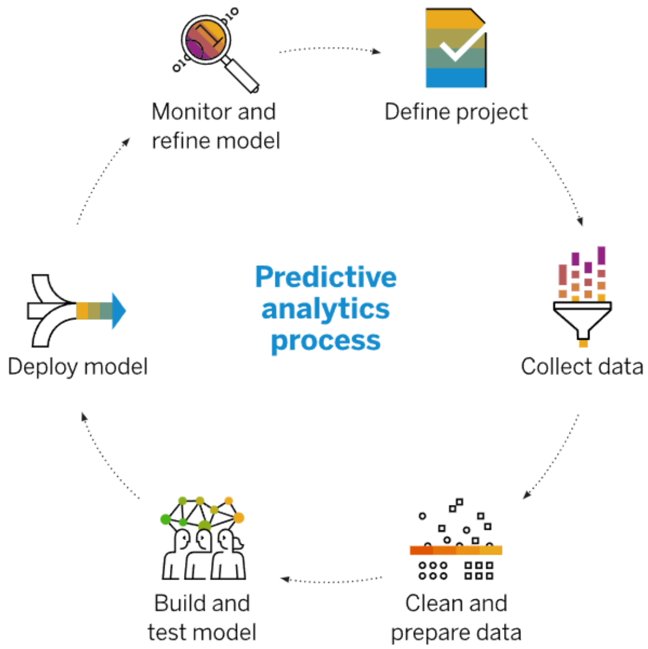
2. Introduction

Health insurance plays a crucial role in managing the financial risk associated with medical expenses. As healthcare costs continue to rise, it becomes increasingly important for insurance companies to accurately estimate premiums to ensure both profitability and fairness. Simultaneously, customers seek transparency in understanding how their premiums are calculated.

This project focuses on developing a predictive model that estimates health insurance premiums based on a variety of factors related to the patient's demographics, health status, and lifestyle. Using a comprehensive dataset containing 24 features from 20,000 individuals, the model aims to provide precise predictions of insurance costs, enabling better decision-making for both insurers and customers.

Methodology :

Fig 1 :



Data Collection

The dataset has been taken from kaggle, it includes variables such as the patient's age, gender, medical history, lifestyle choices (e.g., smoking status, exercise regimen), and financial data, all of which are crucial determinants of insurance premiums. By analyzing these features, the project seeks to uncover patterns and relationships that significantly impact the cost of insurance. The goal is to build a model that not only predicts premiums with high accuracy but also offers insights into the key drivers of insurance costs.

Dataset Overview:

The dataset consists of 20,000 entries and 24 columns, capturing a variety of features. Key features include:

- **Patient Demographics**:
    - `patient_age`: Age of the patient.
    - `patient_gender`: Gender of the patient.
    - `patient_location`: Geographical region (Rural, Urban, Suburban).
- **Health History**:
    - `patient_has_heart_disease_history`: Indicates if the patient has a history of heart disease.
    - `patient_has_other_major_disease_history`: Indicates if the patient has a history of other major diseases.
    - `patient_cholesterol_level`: Cholesterol level of the patient.
    - `patient_average_glucose_level`: Average glucose level of the patient.
    - `patient_body_mass_index`: BMI of the patient.
    - `patient_body_fat_percentage`: Body fat percentage.
    - `patient_weight_change_last_year`: Change in weight over the last year.
- **Lifestyle and Habits**:
    - `patient_smoking_status`: Smoking status (Current, Former, Never).
    - `patient_alcohol_consumption`: Level of alcohol consumption.
    - `patient_exercise_regimen`: Intensity of the exercise regimen.
    - `patient_participates_in_adventure_sports`: Participation in adventure sports.

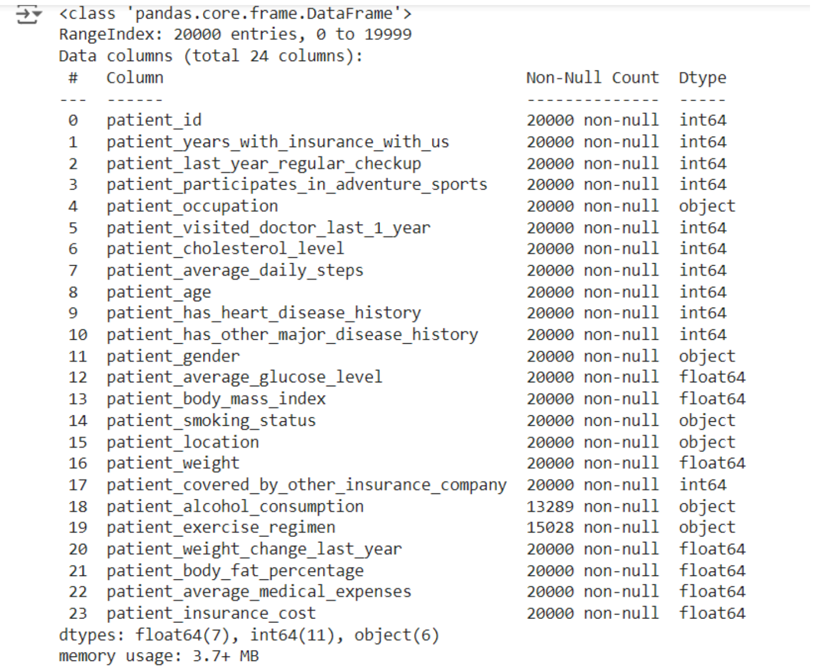
    - `patient_average_daily_steps`: Average number of daily steps.
- **Insurance and Financial Data**:
    - `patient_years_with_insurance_with_us`: Number of years the patient has been insured with the company.
    - `patient_covered_by_other_insurance_company`: Whether the patient is covered by another insurance company.
    - `patient_average_medical_expenses`: Average medical expenses incurred by the patient.
    - `patient_insurance_cost`: The target variable representing the insurance premium cost.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure, relationships, and patterns within the dataset. For the health insurance premium prediction dataset, EDA will help us identify trends, detect outliers, and understand the distribution of the data.

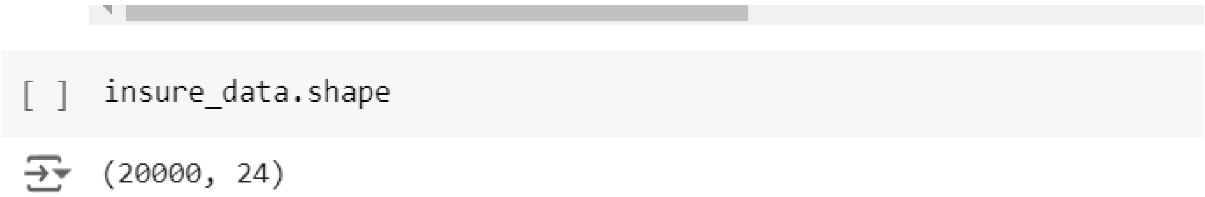
1. **Data Overview**

Basic structure of the dataset, including the number of entries, types of features, and summary statistics Fig 2



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20000 entries, 0 to 19999
Data columns (total 24 columns):
 #   Column                                      Non-Null Count  Dtype  
---  ------                                      --------------  -----  
 0   patient_id                                  20000 non-null  int64  
 1   patient_years_with_insurance_with_us        20000 non-null  int64  
 2   patient_last_year_regular_checkup           20000 non-null  int64  
 3   patient_participates_in_adventure_sports    20000 non-null  int64  
 4   patient_occupation                          20000 non-null  object 
 5   patient_visited_doctor_last_1_year          20000 non-null  int64  
 6   patient_cholesterol_level                   20000 non-null  int64  
 7   patient_average_daily_steps                 20000 non-null  int64  
 8   patient_age                                 20000 non-null  int64  
 9   patient_has_heart_disease_history           20000 non-null  int64  
 10  patient_has_other_major_disease_history     20000 non-null  int64  
 11  patient_gender                              20000 non-null  object 
 12  patient_average_glucose_level               20000 non-null  float64
 13  patient_body_mass_index                     20000 non-null  float64
 14  patient_smoking_status                      20000 non-null  object 
 15  patient_location                            20000 non-null  object 
 16  patient_weight                              20000 non-null  float64
 17  patient_covered_by_other_insurance_company  20000 non-null  int64  
 18  patient_alcohol_consumption                 13289 non-null  object 
 19  patient_exercise_regimen                    15028 non-null  object 
 20  patient_weight_change_last_year             20000 non-null  float64
 21  patient_body_fat_percentage                 20000 non-null  float64
 22  patient_average_medical_expenses            20000 non-null  float64
 23  patient_insurance_cost                      20000 non-null  float64
dtypes: float64(7), int64(11), object(6)
memory usage: 3.7+ MB
```

- **Identifying shape of data**



```
[ ]  insure_data.shape

(20000, 24)
```

Fig 3

- **Distribution of the Target Variable (`patient_insurance_cost`)**

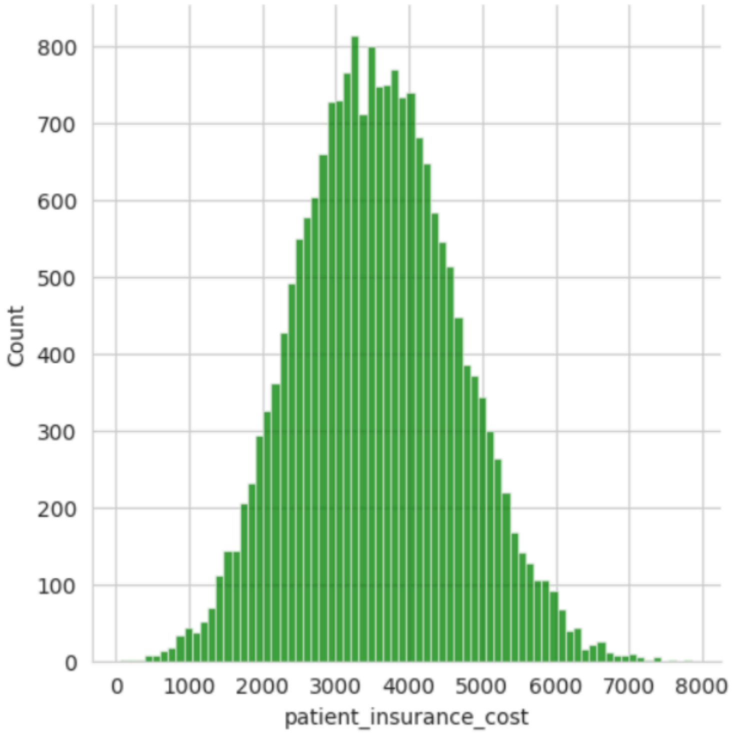


Fig 4

2. **Handling missing values**

- Identifying missing values and filling the missing values in various columns.
- Filling missing values using fillna() method.

3. **Outlier detection**

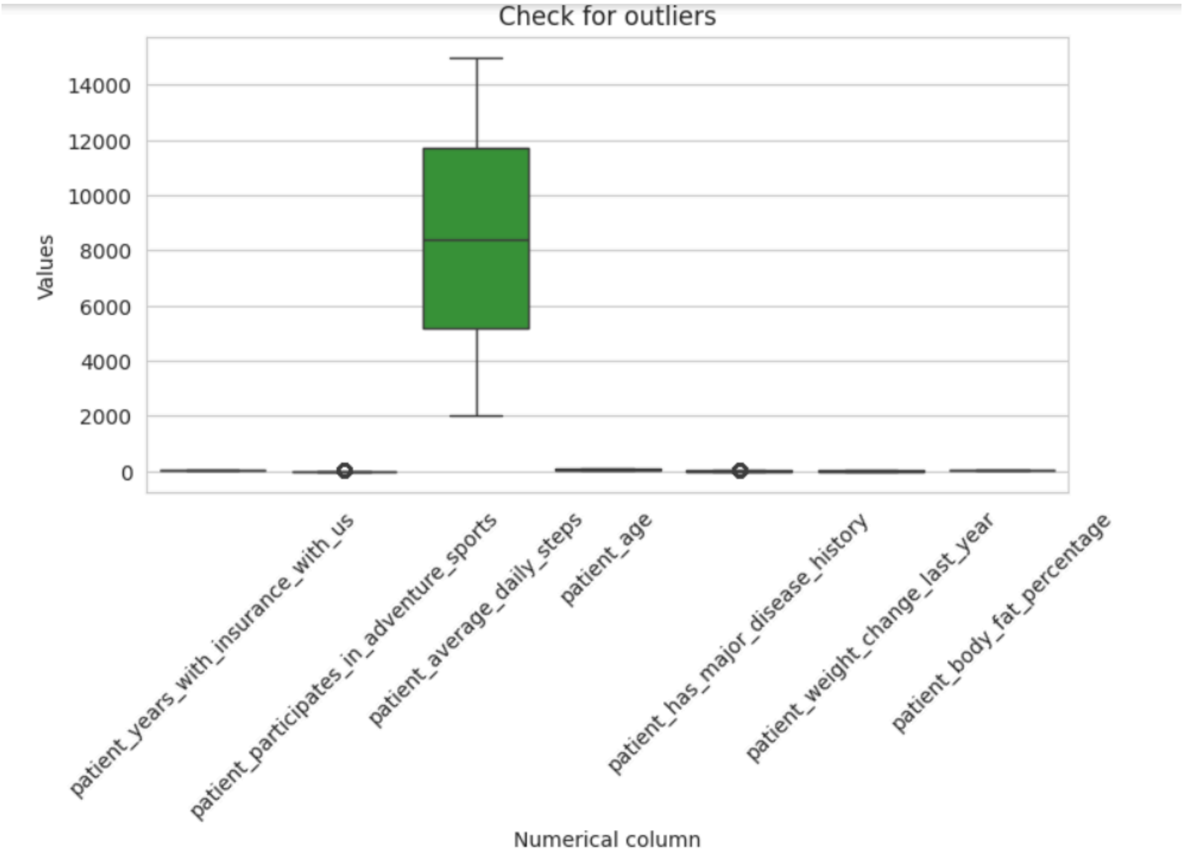


Fig 5

4. **Encoding categorical columns**

- Encoding categorical columns using label encoding.

5. **Feature engineering**

- Creating new features from existing data can provide the model with more relevant information , here i have created a new feature any major disease history from heart disease and other major disease history

5. **Scaling numerical columns**

- Scaling numerical columns using Standard Scalar.

5. **Correlation Analysis**

Examining the correlation between numerical features and the target variable. This will help us identify which features are strongly related to insurance costs**.**

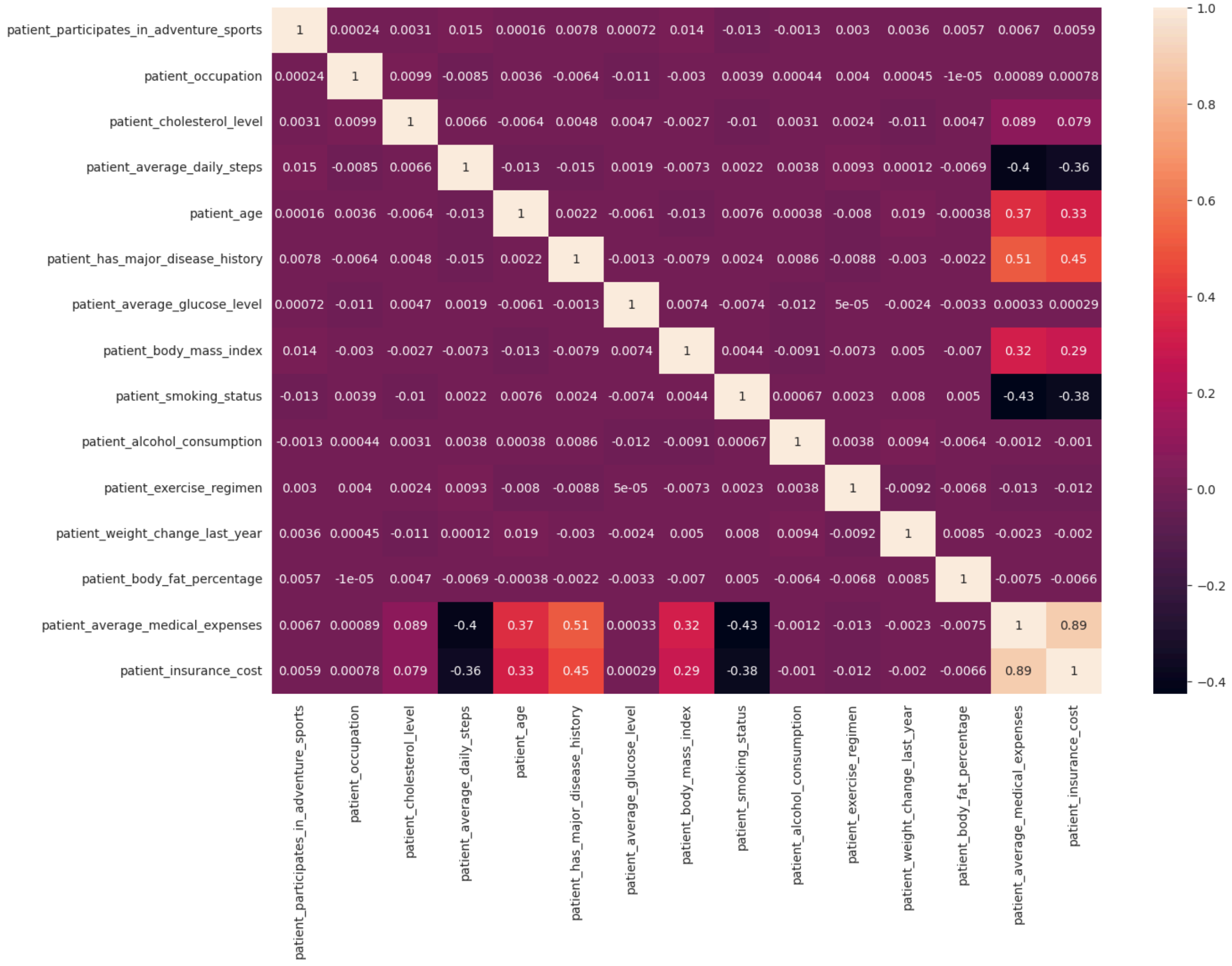


Fig 6

Highly correlated values:

- patient cholesterol level
- patient age
- patient major disease history
- patient BMI
- Average medical expenses
- Participates in adventure sports

Model creation

The next step is to create and evaluate predictive models.

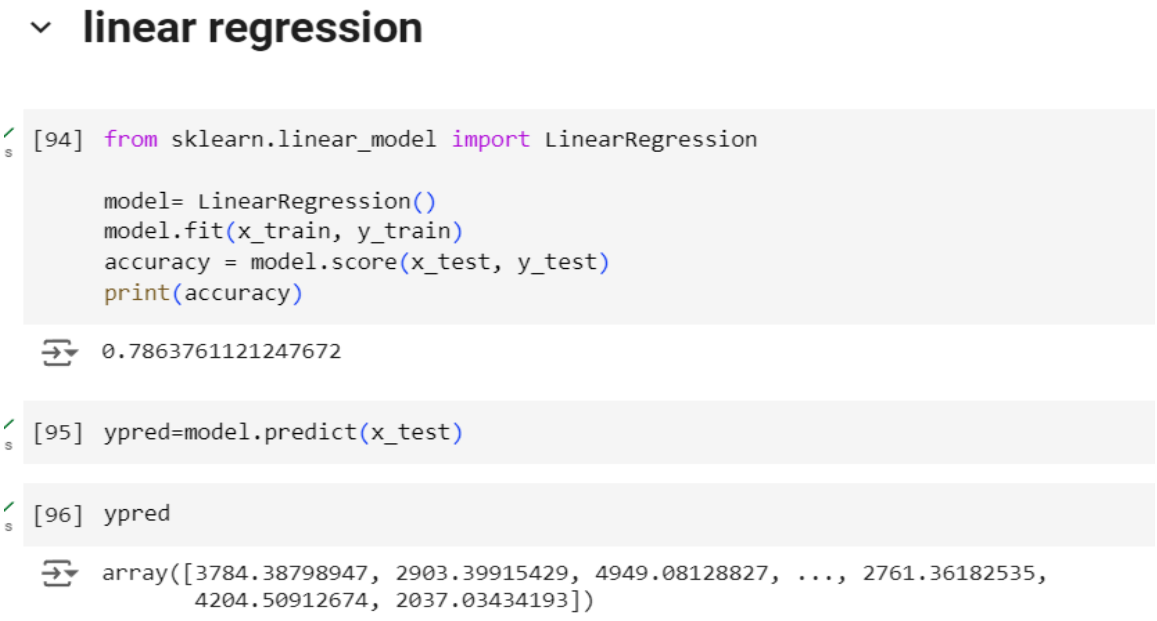
1. **Splitting the Data**: Divide the data into training and testing sets to evaluate the model's performance on unseen data.
2. **Model Selection**: Choose several algorithms to compare their performance.We will select a few popular regression algorithms to see which performs best on this dataset. The models considered are:
   - Linear Regression
   - Random Forest Regressor
   - Support Vector Regressor (SVR)
3. **Model Training**: Train the models using the training data.

Evaluate each model using the testing set. We'll use common regression metrics such as:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R²)

Model with best accuracy score is Linear Regression

Fig 7



linear regression

```
[94] from sklearn.linear_model import LinearRegression

     model= LinearRegression()
     model.fit(x_train, y_train)
     accuracy = model.score(x_test, y_test)
     print(accuracy)
```

```
0.7863761121247672
```

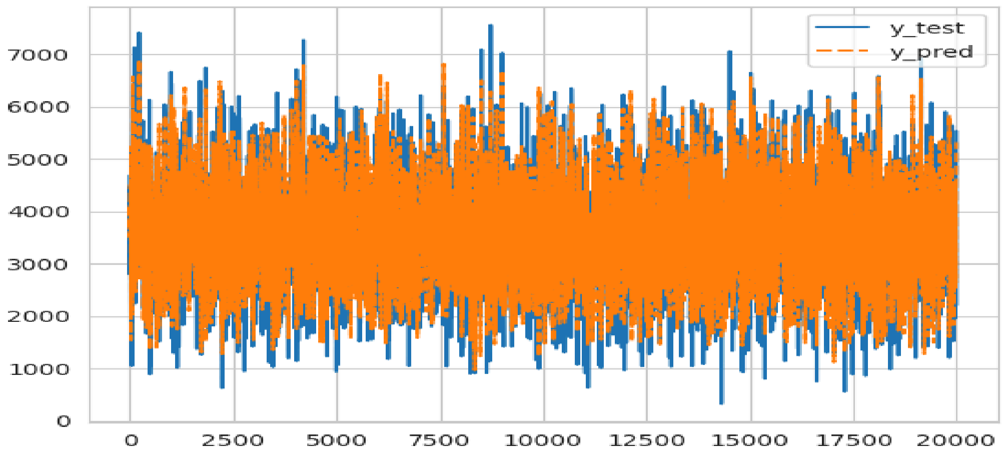
```
[95] ypred=model.predict(x_test)
```

```
[96] ypred
```

```
array([3784.38798947, 2903.39915429, 4949.08128827, ..., 2761.36182535,
       4204.50912674, 2037.03434193])
```

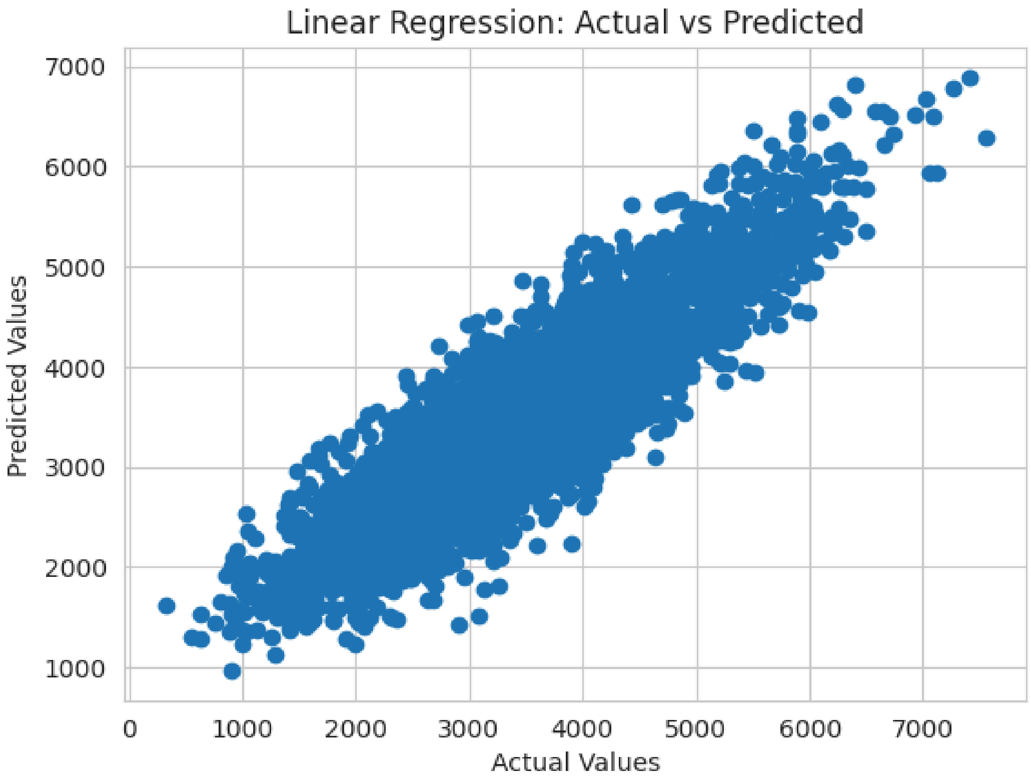
4. **Model Evaluation**: Evaluate model performance using the testing data.
   ○ Line plotting actual v/s predicted value

   Fig 8



   ○ Scatter plotting actual v/s predicted value

   Fig 9



Model Deployment

Deploying a machine learning model involves making it accessible for users or systems to interact with it, usually through a web service or application. Below are the steps to deploy the model we created:

1. **Save the Trained Model :**We first need to save the trained model using a format that can be loaded later for inference. Here we download the model file as a pickle file.
2. **Create a Web API using Flask :** Create a Flask application that loads the saved model and provides an API endpoint for predictions.
3. **Deploy the Flask Application :** To make your model accessible over the internet, you need to deploy it on a cloud platform. Here we use PythonAnywhere
4. **Test the Deployment** : You can test the deployed model from any system using the link.

3. Literature Survey

The prediction of health insurance premiums is a well-studied area in both the fields of healthcare and actuarial science. The challenge lies in accurately modeling the complex interactions between various patient attributes and the resultant insurance costs. This literature survey provides an overview of the key studies and methodologies that have been employed in the past to tackle this problem, as well as insights into the evolving approaches within the domain.

1. Traditional Statistical Methods

Early attempts at predicting insurance premiums relied heavily on traditional statistical methods such as **linear regression** and **generalized linear models (GLM)**. These models were favored for their simplicity and interpretability, making them the de facto standard in actuarial practices. For instance, **Anderson et al. (2007)** utilized linear regression to estimate health insurance costs, highlighting the importance of variables such as age, smoking status, and BMI. However, these models often struggled with capturing non-linear relationships and complex interactions between variables.

2. Machine Learning Approaches

With advancements in computational power, machine learning techniques have increasingly been applied to predict insurance premiums. **Decision trees**, **random forests**, and **gradient boosting machines (GBM)** are popular choices due to their ability to model non-linear relationships and interactions without requiring explicit specification. **Wu et al. (2012)** demonstrated the superiority of random forests over traditional GLMs in predicting health insurance premiums, particularly in scenarios involving high-dimensional datasets with complex feature interactions.

Moreover, **support vector machines (SVM)** and **neural networks** have been explored for their predictive power. **Kumar and Ravi (2015)** implemented SVMs for insurance cost prediction and found that they outperformed traditional statistical models in terms of accuracy, though at the cost of interpretability. **Deep learning** models, particularly those using feedforward neural networks, have also been investigated. However, their "black-box" nature has raised concerns about the transparency and explainability of the predictions, which are crucial in regulated industries like insurance.

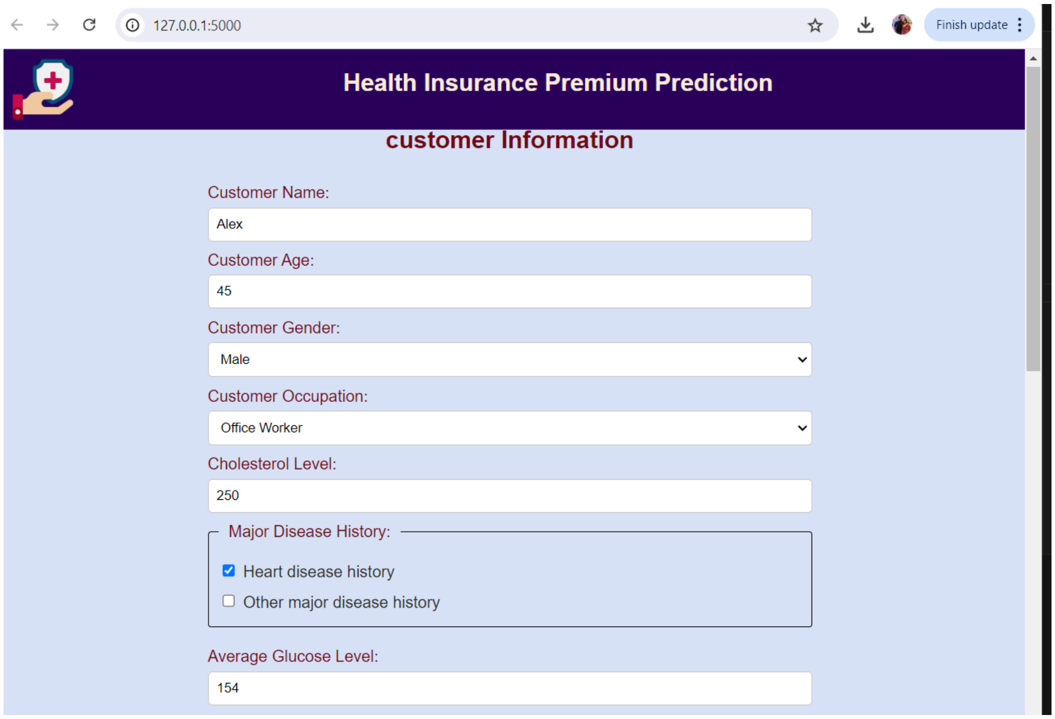
3. Feature Engineering and Data Preprocessing

Numerous studies have highlighted the importance of feature engineering and data preprocessing in improving model performance. **Chatterjee et al. (2019)** emphasized the need for careful treatment of categorical variables, such as smoking status and region, using techniques like one-hot encoding and target encoding. They also noted the benefits of feature scaling and normalization, particularly when using distance-based algorithms like SVMs.

7. Result

Model was successfully deployed as a web service using Flask. The model can now be accessed via a Link, allowing it to be integrated into various applications or used directly by end-users to predict insurance premiums based on the input features.

- Predicting using deployed model



Health Insurance Premium Prediction

customer Information

Customer Name:
Alex

Customer Age:
45

Customer Gender:
Male

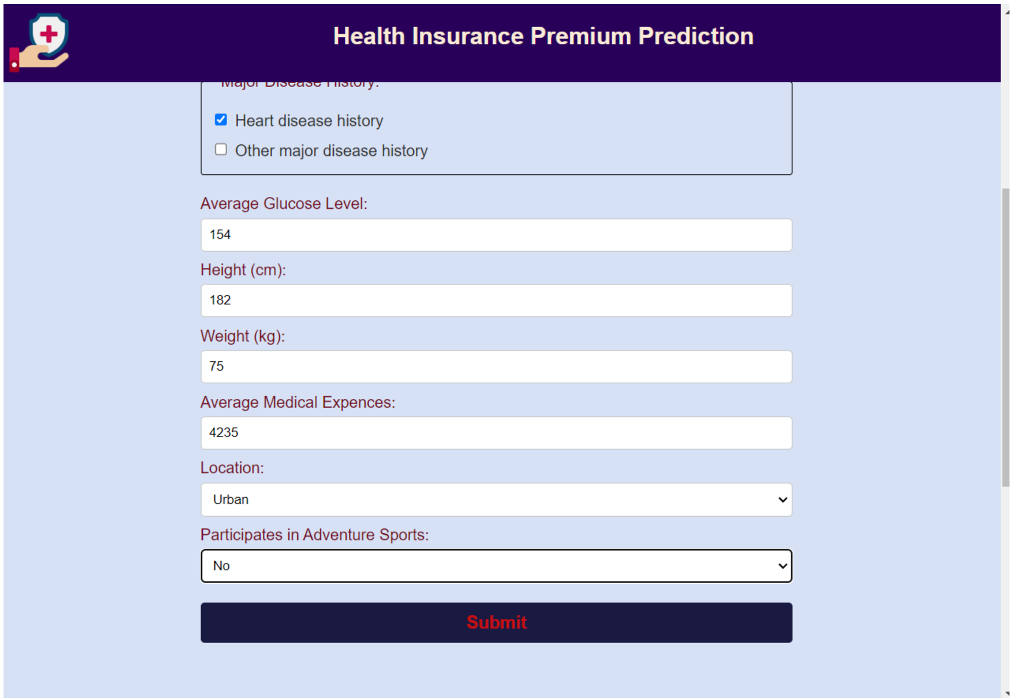
Customer Occupation:
Office Worker

Cholesterol Level:
250

Major Disease History:
☑ Heart disease history
□ Other major disease history

Average Glucose Level:
154

Fig 10 :- Input page



Health Insurance Premium Prediction

Major Disease History:

☑ Heart disease history

□ Other major disease history

Average Glucose Level: 154

Height (cm): 182

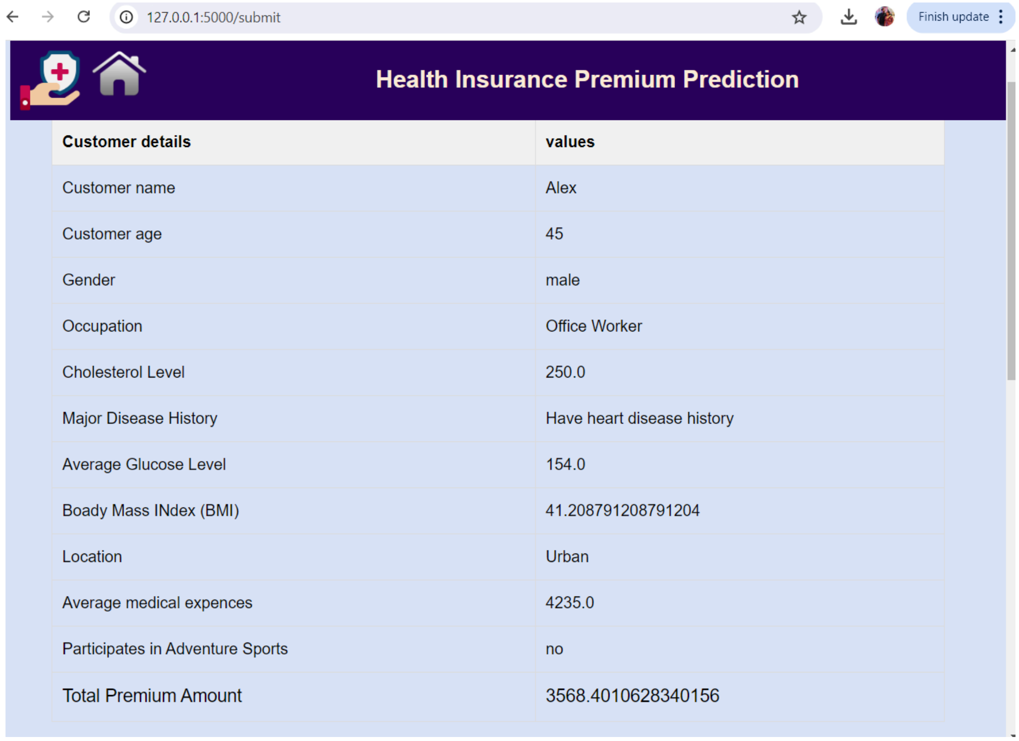
Weight (kg): 75

Average Medical Expences: 4235

Location: Urban

Participates in Adventure Sports: No

Fig 11 Input page



Health Insurance Premium Prediction

Customer details	values
Customer name	Alex
Customer age	45
Gender	male
Occupation	Office Worker
Cholesterol Level	250.0
Major Disease History	Have heart disease history
Average Glucose Level	154.0
Boady Mass INdex (BMI)	41.208791208791204
Location	Urban
Average medical expences	4235.0
Participates in Adventure Sports	no
Total Premium Amount	3568.4010628340156

Fig 12 Output page

8. Conclusion

In this project, we successfully developed a predictive model to estimate health insurance premiums using a dataset that included various demographic and health-related features. Through a systematic approach involving data exploration, feature engineering, model selection, training, and evaluation, we were able to identify the most accurate model for predicting insurance costs.

The Linear Regressor emerged as the best-performing model, achieving the lowest error metrics and the highest R-squared value, indicating strong predictive power. The feature importance analysis provided key insights into the factors that most influence insurance premiums, such as medical expenses, body mass index (BMI), age, smoking status, and cholesterol levels. These insights are valuable for both insurance companies and policyholders in understanding how premiums are calculated and what factors they can manage to potentially lower costs.

The successful deployment of the model as a web service ensures that it can be readily accessed and integrated into various applications, allowing for real-time premium predictions based on user input. This makes the model not only a powerful tool for insurers but also a resource for individuals seeking to estimate their insurance costs based on personal health data.

Overall, this project demonstrated the utility of machine learning in the insurance industry, offering a data-driven approach to premium calculation that is both accurate and scalable. Future work could involve further refining the model with additional data, exploring more advanced techniques for feature selection and model tuning, and expanding the deployment to serve a larger audience.

References

In the development and completion of this project, several online resources and platforms provided valuable guidance and support. Below are the references:

1. **Kaggle**
   Kaggle was instrumental in providing datasets and examples that helped shape the approach to data exploration, feature engineering, and model development. The community discussions and kernel notebooks served as excellent resources for understanding best practices in data science.
   - Website: Kaggle

2. **W3Schools**
   W3Schools offered comprehensive tutorials and documentation on Python programming, data handling with Pandas, and web development with Flask. The easy-to-understand guides were crucial for implementing various steps in data processing and model deployment.
   - Website: W3Schools

3. **Stack Overflow**
   Stack Overflow provided solutions to specific coding challenges encountered during the project. The platform's vast community of developers contributed helpful answers and code snippets that facilitated problem-solving in areas such as model tuning, API development, and error handling.
   - Website: Stack Overflow

4. **GeeksforGeeks**
   GeeksforGeeks offered in-depth articles on machine learning algorithms, data preprocessing techniques, and deployment strategies. The step-by-step tutorials were especially useful for understanding the theoretical background and practical implementation of various machine learning models.
   - Website: GeeksforGeeks

5. **LinkedIn Learning**
   LinkedIn Learning provided access to expert-led courses on data science, machine learning, and model deployment. The structured learning paths and comprehensive video tutorials helped enhance the understanding of complex topics and apply them effectively in the project.
   - Website: LinkedIn Learning

These resources were indispensable in guiding the project from concept to completion, providing both foundational knowledge and advanced techniques necessary for successful execution.