Health Insurance Cost Prediction Using Machine Learning

Submitted in partial fulfillment of the requirement of Third Year of Engineering in the Department of Computer Engineering, during the academic year 2023-2024.

Submitted By

Group Members Name	Roll Numbers
Soham Shedale	537
Rudra Singh	540
Chirag Suryavanshi	543
Uttkarsh Wadgave	545

Submitted to
Prof. Diksha V Kale
(Department of Computer Engineering)

Department of Computer Engineering
PILLAI COLLEGE OF ENGINEERING
(Autonomous)
New Panvel - 410 206
UNIVERSITY OF MUMBAI
Academic Year 2023-24

Acknowledgment

We would like to express our special thanks and gratitude to Principal **Dr. Sandeep Joshi** and **H.O.D Dr. Sharvari Govilkar** who gave us the opportunity to do this project on the topic of **Health Insurance Cost Prediction,** which helped us in applying and learning new concepts. We would also like to offer thanks to our project guide **Prof.Diksha Kale** who always motivated us helped us to solve our doubts, inspired us to be better people, and always strived to do our personal best. We are immensely grateful to all of them for sharing their wisdom with us during the course.

Summary

Insurance is a policy that helps to cover up all losses or decrease losses in terms of expenses incurred by various risks. A number of variables affect how much insurance costs. These considerations of different factors contribute to the insurance policy cost expression. Machine Learning(ML) in the insurance sector can make insurance more effective. In the domains of computational and applied mathematics machine learning (ML) is a well-known research area. ML is one of the computational intelligence aspects when it comes to the exploitation of historical data that may be addressed in a wide range of applications and systems. There are some limitations in ML so; Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry and thus it requires more investigation and improvement. Using machine learning algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs. The proposed research approach uses Linear Regression, Decision Tree Regression, and Gradient Boosting Regression and also streamlit as a framework. We used a medical insurance cost dataset for the cost prediction purpose, and machine learning methods were used to show the forecasting of insurance costs by regression model comparing their accuracies.

TABLE OF CONTENTS

Chapter 1
Introduction

The introduction to a project is where you set up your topic and approach for the reader, it has several key goals: present your topic and get the reader interested.

1.1 Overview

Our proposed system addresses the critical need for accurate health insurance cost prediction in today's healthcare landscape. The primary objective of our system is to assist individuals and insurance providers in estimating health insurance costs more precisely, thereby improving financial planning and risk assessment.

Problem Statement:
The current methods for estimating health insurance costs often rely on historical data and generalized factors, resulting in imprecise predictions. This can lead to financial insecurity for individuals and challenges for insurance providers in setting accurate premiums. Our system aims to overcome these limitations by leveraging machine learning techniques, specifically the gradient boosting algorithm, to provide more accurate and personalized cost predictions.

Significance:
The significance of our project is multifaceted. For individuals, accurate cost predictions can enable better financial planning and insurance decision-making. Insurance companies can benefit from improved risk assessment, leading to fairer premiums and reduced financial losses. Moreover, our system has broader implications for the healthcare industry, as it can contribute to more efficient resource allocation and enhanced policy formulation.

User Perspective:
Our system is designed to cater to a diverse range of users. It offers valuable insights to individual insurance customers, helping them make informed decisions about their insurance coverage. Actuaries and underwriters within insurance companies can utilize the system to enhance risk assessment and pricing strategies. Additionally, policymakers in the healthcare sector can draw insights from our system for data-driven decision-making.

Methodology Overview:
Our methodology revolves around the utilization of machine learning, particularly the gradient-boosting algorithm. This powerful algorithm is employed to analyze historical health insurance data, identify patterns, and generate predictive models for cost estimation. By leveraging the strengths of machine learning, we can provide highly accurate and individualized cost predictions.

Expected Outcome:
Users of our system can expect more accurate health insurance cost estimates, reducing uncertainty and financial stress. For insurance providers, the expected outcome is enhanced risk assessment, leading to fairer premiums and improved competitiveness in the market.

Innovation:
Our system incorporates innovative approaches to health insurance cost prediction by employing state-of-the-art machine learning techniques. The use of gradient boosting adds a unique dimension to our project, enhancing the accuracy of predictions and personalizing the cost estimation process.

1.2 Scope

The scope of our project encompasses the development and implementation of a robust health insurance cost prediction system. This system will serve as a valuable resource for individuals seeking accurate estimates for their health insurance expenses and insurance providers aiming to optimize their pricing strategies. Our focus is on leveraging advanced data analysis and machine learning to ensure the highest level of prediction accuracy. Through a user-friendly interface, our system will enable individuals to input their data for precise cost estimates, facilitating informed decision-making.

For insurance providers, the system offers the potential to refine risk assessment and pricing strategies, fostering a competitive and data-driven insurance landscape. The system's scope extends to scalability, with potential future enhancements and collaborations that promise to further advance the field of health insurance cost prediction.

1.3 Methodology

Our methodology centers on the utilization of the gradient-boosting algorithm for health insurance cost prediction. The choice of this method was guided by its proven effectiveness in handling regression tasks and its ability to provide accurate and interpretable results.

Method Selection: The gradient boosting algorithm was selected for its capacity to handle complex, non-linear relationships in the data, and its ability to produce ensemble models that combine the strengths of multiple decision trees. It excels in making accurate predictions by minimizing the error in each subsequent tree, resulting in a robust and reliable cost prediction model.

Data Collection: Data collection was a critical phase of our project. We gathered historical health insurance data from a variety of sources, including insurance companies, publicly available datasets, and government health statistics.

Model Training and Testing: We utilized the gradient boosting algorithm for model training, partitioning the dataset into training and testing sets. To evaluate the model, we employed metrics like mean absolute error (MAE), mean squared error (MSE), and R-squared (R^2) to assess its accuracy and generalization performance.

Model Deployment: The model is deployed as a web-based platform that allows users to input their demographic and insurance-related information. It returns accurate cost predictions, providing users with valuable insights for their financial planning.

1.4 Limitations

Our health insurance cost prediction system exhibits the following limitations:

Data Quality: Despite meticulous data cleaning, the system's performance may be influenced by data quality issues in the underlying dataset.

Assumptions: The model's assumption that future trends will mirror historical patterns may not hold under conditions of significant change, such as shifts in healthcare legislation or unforeseen global events.

Accuracy Challenges: In cases involving rare or complex medical conditions and unique insurance scenarios, the model's accuracy may diminish. Users are advised to seek professional advice in such scenarios.

Generalizability: The system's generalizability is contingent on the region and demographics of the user base. Customization may be required for different markets and populations.

Legal and Ethical Compliance: While we prioritize data privacy and ethics, it is the responsibility of users and organizations to ensure they adhere to data protection regulations and ethical considerations when employing the system.

Resource Constraints: Resource limitations, including computational capacity and time constraints, may impact the model's depth and complexity, potentially limiting its performance.

Chapter 2

Literature Review

A literature review is a piece of academic writing demonstrating knowledge and understanding of the academic literature on a specific topic placed in context.

Rama Devi Burri, Ram Burri, Ramesh Reddy Bojja, Srinivasa Rao Buruga (2019)

Explored Machine Learning for Insurance Claim Prediction

In this paper, they survey the state-of-the-art machine learning (ML) for insurance claim prediction. ML has had a significant impact on insurance claim prediction, with ML-powered systems now being used to predict the likelihood of a customer filing a claim, the severity of a claim, and the type of claim with high accuracy. However, there are still challenges that need to be addressed before ML can be more widely adopted for insurance claim prediction, such as the need for high-quality training data, the need to develop ML models that are robust to noise and ambiguity in the data, and the need to ensure that ML models are fair and unbiased. Despite the challenges, they concluded that ML has the potential to revolutionize the insurance industry by helping insurers reduce costs and provide better service to their customers.

Ajay Sahu, Gopal Sharma, Janvi Kaushik, Kajal Agarwal, Devendra Singh - (2020)

Diving into Health Insurance Cost Prediction by Using Machine Learning

In this paper, the authors discuss the high cost of healthcare and the need to control costs. This research could help patients find affordable healthcare and policymakers identify expensive providers. The authors also note that their research could have implications for public health policy. For example, they suggest that their findings could be used to develop new ways to reimburse healthcare providers or to design more efficient healthcare delivery systems. Overall, this paper provides a valuable overview of the potential of machine learning to improve the affordability and efficiency of healthcare.

H. Chen Jonathan, M. Asch Steven - (2020)

Inspected Medical Expense Prediction System using Machine Learning Techniques and an Intelligent Fuzzy Approach

In this paper, they used machine learning (ML) for insurance claim prediction. Prediction isn't a new concept in medicine. Clinical predictions based on data are becoming commonplace in medicine. Risk categorization of patients in the critical care unit ranges from risk scores to anticoagulant treatment (CHADS2) and cholesterol medicine usage (ASCVD) (APACHE). The real data source, on the other hand, has an issue. Different from traditional techniques, which rely on data from cohorts that have been thoroughly prepared to prevent bias, new data sources are sometimes unstructured due to the fact that they were developed for various purposes (clinical care, billing, etc.). Patient self-selection, indication misunderstanding, and inconsistent outcome data can all contribute to unintentional biases and even racist programming in machine prediction. As a result of this understanding, discussing the potential of data analysis to aid medical decision-making isn't just wishful thinking. Overall, this paper provides a valuable overview of machine learning to improve the efficiency of healthcare.

Chandra Kudumula - April 2021

Inspected the Insurance price prediction using Machine Learning

This paper is about predicting insurance prices using Machine Learning (ML.NET). It discusses what ML.NET is and how it can be used for insurance price prediction. It also details the steps involved in building and training an ML.NET model. The paper concludes by summarizing the benefits of using ML.NET for insurance price prediction. In this paper, the author provides a step-by-step guide on how to use ML.NET to build a model to predict insurance prices. The author also discusses the benefits of using ML.NET for insurance price prediction, such as its ability to learn from data and make predictions without the need for human intervention

Thomas Poufinas, Periklis Gogas- (2023)

Investigated Machine Learning in Forecasting Motor Insurance Claims

This is an article about machine learning in forecasting motor insurance claims. It discusses the potential benefits of using machine learning to improve the accuracy of claim forecasts. The authors also review some of the challenges that need to be addressed before machine learning can be widely adopted in the insurance industry. Some of the important points from this article are that machine learning has the potential to improve the accuracy of claim forecasts, but more research is needed to address the challenges of using machine learning in the insurance industry. They propose a machine-learning model that uses weather and car sales data to forecast motor insurance claims. They find that their model outperforms traditional forecasting methods.

Sebastian Baran, Przemys law Rola - (2022)

Predicted motor insurance claims occurrence as an imbalanced machine learning problem

This paper discusses the challenges of predicting car insurance claims using machine learning, particularly the challenge of imbalanced datasets. Imbalanced datasets occur when the number of positive examples (e.g., claims) is much smaller than the number of negative examples (e.g., no claims). This can make it difficult for machine learning algorithms to learn to predict claims accurately. The authors compare the performance of several machine learning algorithms on a dataset of car insurance claims. They find that XGBoost outperforms the other algorithms, but all of them benefit from using techniques to address imbalanced datasets.

The authors conclude that machine learning can be used to predict car insurance claims accurately, but that it is important to use techniques to address imbalanced datasets.

M. Ramya, Sankeerthana, Harshitha, Dr. Sunil Bhutada, Dr.Y. Rohita - (2021)

Predicted Possible Prospects To Buy Insurance Using Data Analytics

In this paper, the authors discuss the use of machine learning to predict insurance claims. They propose a model using random forest regression to predict the cost of healthcare given a person's smoking habits, age, sex, body mass index, and region. The authors found that this model was able to predict insurance claims with 86% accuracy. The authors also discuss the importance of data visualization in understanding the data and the results of the machine learning models. They use data visualization to identify patterns in the data and to assess the performance of their model. Overall, this paper provides a good overview of the use of machine learning to predict insurance claims. The authors' proposed model is a promising step toward developing more accurate and efficient insurance claim prediction models. As machine learning models continue to improve, we can expect to see even more innovative and effective ways to use machine learning to predict insurance claims.

Puneeth Kumar, Pavan Krishna, Raja Vardhan- (2022)

Discussed Medical Expense Prediction Using Machine Learning

In this paper, the authors discuss the importance of predicting medical expenses and the challenges involved. They also review previous work on using machine learning to predict medical expenses. The authors propose a new system for predicting medical expenses that is based on linear regression. They evaluate their system on a real-world dataset and find that it is able to predict medical expenses with an accuracy of over 75%. The authors also discuss the potential for bias in medical expense prediction models.

Overall, this paper provides a comprehensive overview of the challenges and opportunities of using machine learning to predict medical expenses. The authors' proposed system is a promising step toward developing more accurate and fair medical expense prediction models.

Chapter 3
Project Requirements

The most common set of requirements defined by any Operating System or software application is the physical computer resources also known as hardware.
A hardware requirement list is often accompanied by a hardware compatibility list.

3.1 System Requirements

Our health insurance cost prediction system has specific requirements to ensure optimal performance and functionality.

Hardware Requirements:

The system can run on standard hardware, requiring a modern computer or server with a minimum of a dual-core processor, 8GB of RAM, and 100GB of storage space for the application and data storage. However, for larger-scale implementations, more powerful hardware may be necessary.

Software Requirements:

Python: The system is primarily developed in Python and is required for all components.

Flask: We utilize Flask, a Python web framework, for the system's web interface.

Scikit-learn: This library is essential for machine learning tasks and is employed for model training and prediction.

Pandas and NumPy: These libraries are used for data manipulation and preprocessing.

Additional libraries: Various Python libraries for data visualization, model evaluation, and web development are required.

Data Requirements:

The system relies on historical health insurance data, including claims history, demographic information, and insurance policy details. Data should be structured in a format compatible with the system's data processing routines. Data access is crucial, and databases or file storage systems should be in place for efficient data retrieval.

Internet Connectivity:

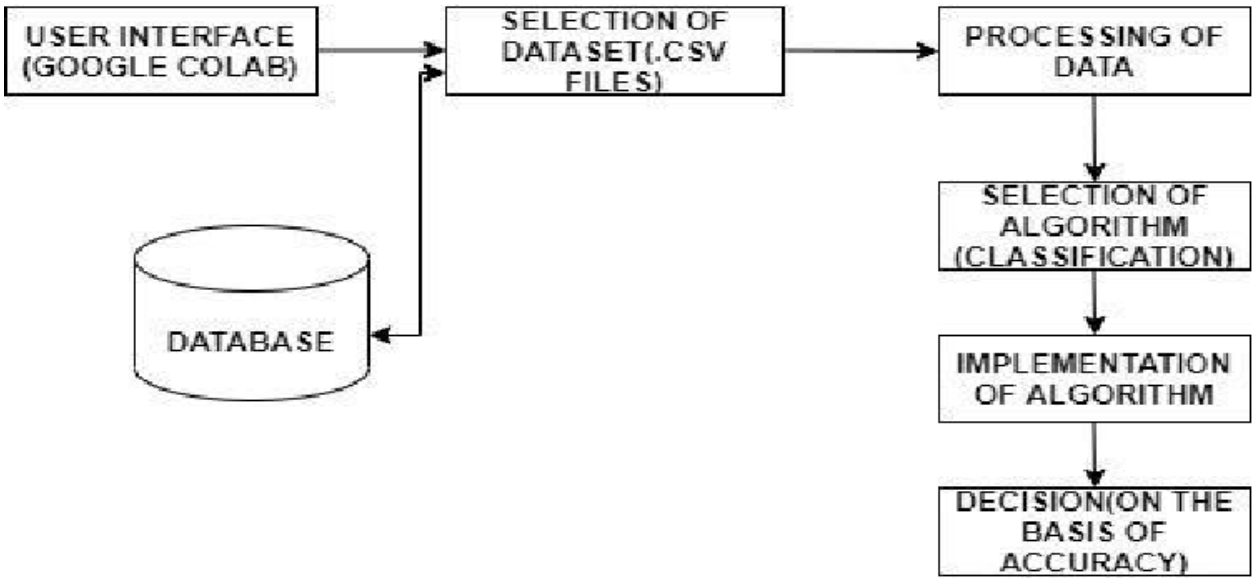
The system can operate both online and offline. An internet connection is only required for data updates or to access external data sources. For users primarily interested in offline usage, the system can be configured to operate without an internet connection.

Scalability:

For scalability, the system can be deployed on cloud-based services, allowing it to handle larger datasets and accommodate more users. Scaling may require additional computing resources, depending on the growth in data volume and user traffic.

3.1 Data Flow Diagram (DFD)

The Data Flow Diagram (DFD) provides a visual representation of how data moves within our health insurance cost prediction system. It outlines the path data takes from initial sources to the final output, illustrating key components and processes.

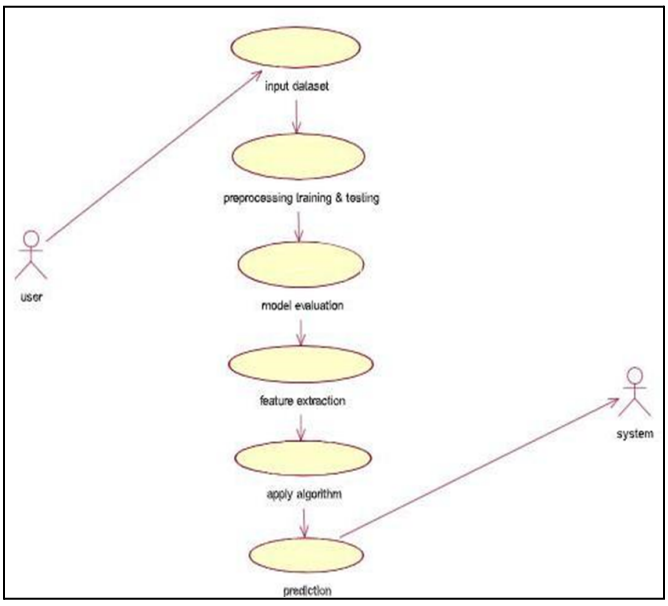


3.3 Use Case Diagram

The Use Case Diagram visually represents the interactions between various actors and our health insurance cost prediction system. It outlines how different actors, both individual users and organizations, engage with the system's functionalities.

Diagram Description:
The Use Case Diagram comprises the following key elements:

Actors: Represent the different users and entities interacting with the system.
Use Cases: Depict the specific functionalities and interactions provided by the system.



Chapter 4

Software Description

A Software description is a comprehensive account that provides detailed information about a computer program or application.

4.1 Python:

Python plays a pivotal role in our health insurance cost prediction system, serving as the primary programming language that enables data analysis, machine learning, and web-based user interface development.

4.2 Flask:

Flask, a lightweight Python web framework, serves as the foundation for the web-based user interface of our health insurance cost prediction system.

Flask was chosen for its minimalistic and flexible nature. Its lightweight design is well-suited for building web applications that require simplicity and efficiency.

4.3 HTML, CSS, and JavaScript:

HTML, CSS, and JavaScript collectively contribute to the development of an interactive and user-friendly user interface for our health insurance cost prediction system.

Role of HTML:

HTML (Hypertext Markup Language) forms the backbone of the user interface by defining the structure and content of web pages.

Role of CSS:
CSS (Cascading Style Sheets) is used to enhance the visual aspects of the user interface.

Role of JavaScript:
JavaScript is employed to enhance the functionality and interactivity of the user interface.

4.4 Libraries Imported

Our health insurance cost prediction system leverages a range of libraries and frameworks to accomplish various tasks, from data processing to machine learning and web development.

Libraries and Frameworks List:
Scikit-learn: A machine learning library that provides a comprehensive set of tools for predictive data analysis, including the implementation of the gradient boosting algorithm for health insurance cost prediction.

Pandas: A data manipulation and analysis library used for data preprocessing and cleaning to ensure data quality.

NumPy: Essential for efficient numerical operations and data manipulation in our data processing routines.

Chapter 5

Project Implementation

Project Implementation refers to the phase in the project management process where the planned project activities and strategies are put into action to achieve its objectives.

5.1 Algorithm

The heart of our health insurance cost prediction system lies in the implementation of the gradient boosting algorithm. This section provides an in-depth look at the algorithm's role and significance in our project.

Algorithm Overview:
Gradient boosting is an ensemble machine-learning technique designed to enhance the predictive accuracy of models. In our system, it is applied to the problem of health insurance cost prediction. The primary purpose is to leverage historical health insurance data to make accurate predictions regarding future insurance costs.

How It Works:
Gradient boosting combines the predictions of multiple weak learners (typically decision trees) into a robust and accurate predictor. It does this through an iterative process, where each new learner focuses on the errors made by the preceding one.

Training Process:

The training process involves feeding historical health insurance data into the algorithm. The algorithm iteratively constructs decision trees, with each tree aimed at capturing and improving upon the prediction errors of the previous ones. The process continues until the model's accuracy reaches a satisfactory level.

Model Evaluation:

The performance of the gradient boosting model is evaluated using various metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics help assess the model's predictive accuracy by measuring the magnitude of prediction errors. A lower MAE and RMSE indicate a more accurate model.

Significance:

The choice of the gradient boosting algorithm is significant for several reasons. It excels at handling complex, non-linear relationships within the data, making it well-suited for health insurance cost prediction, where multiple factors can influence costs. Additionally, gradient boosting is known for its high predictive accuracy and robustness. It often outperforms other machine learning algorithms, making it a crucial component in achieving accurate and reliable health insurance cost predictions.

The gradient boosting algorithm is at the core of our system, using historical data to construct an accurate predictive model for health insurance cost estimation.

5.2 System Overview

We utilized a dataset from Kaggle for creating our prediction model. This dataset comprises seven attributes and is divided into two parts: training data and testing data. To train the model, 80% of the total data was used, while the remaining 20% was designated for testing. The training dataset was applied to construct a predictive model for medical insurance costs, and the test set was employed to evaluate the regression model.
The following table presents a description of the dataset.

Age	Age of client
Sex	Male/Female
BMI	Body Mass Index
Children	Number of children/kids the client has
Smoker	Whether a client is a smoker or not
Region	Whether the client lives in the southwest, northwest, southeast, or northeast
Charges	Medical Costs the client pay

- Age: age of the primary beneficiary
- Sex: insurance contractor gender, female, male
- BMI: Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: number of children covered by health insurance, number of dependents
- smoker: smoking or not
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: individual medical costs billed by health insurance

Importing Libraries: In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs.

```
#Importing Libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import time
import io
from sklearn.model_selection import train_test_split
#regression models
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn import metrics
import joblib
```

```
data = pd.read_csv("D:\Major Project\health insurance.csv")
df = pd.DataFrame(data = data)
```

```
df.head()
```

Data Cleaning:

Check the info:

Information about data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Statistics of data:

	age	sex	bmi	children	smoker	region	charges
count	1,338	1338	1,338	1,338	1338	1338	1,338
unique	None	2	None	None	2	4	None
top	None	male	None	None	no	southeast	None
freq	None	676	None	None	1064	364	None
mean	39.207	nan	30.6634	1.0949	nan	nan	13,270.4223
std	14.05	nan	6.0982	1.2055	nan	nan	12,110.0112
min	18	nan	15.96	0	nan	nan	1,121.8739
25%	27	nan	26.2963	0	nan	nan	4,740.2872
50%	39	nan	30.4	1	nan	nan	9,382.033
75%	51	nan	34.6938	2	nan	nan	16,639.9125

Here is some descriptive statistic

Converting string values of Columns to numerical values:

Before converting:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16,884.924
1	18	male	33.77	1	no	southeast	1,725.5523
2	28	male	33	3	no	southeast	4,449.462
3	33	male	22.705	0	no	northwest	21,984.4706
4	32	male	28.88	0	no	northwest	3,866.8552

After converting:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.9	0	1	1	16,884.924
1	18	1	33.77	1	0	2	1,725.5523
2	28	1	33	3	0	2	4,449.462
3	33	1	22.705	0	0	3	21,984.4706
4	32	1	28.88	0	0	3	3,866.8552

Training the Model:

```
#Training Model
gr = GradientBoostingRegressor()
gr.fit(X_train, y_train)
```

```
lr = LinearRegression()
lr.fit(X_train, y_train)
```

```
svm = SVR()
svm.fit(X_train, y_train)
```

```
rf = RandomForestRegressor()
rf.fit(X_train, y_train)
```

Testing the Model:

```
#prediction
y_pred1=lr.predict(X_test)
```

```
y_pred2=gr.predict(X_test)
```

```
y_pred3=svm.predict(X_test)
```

```
y_pred4=rf.predict(X_test)
```

```
df1 = pd.DataFrame({'Actual': y_test, 'Linear Regression': y_pred1, 'SVR': y_pred2, 'Random Forest': y_pred3, 'Gradient Boosting': y_
```

```
df1
```

Finding the Accuracy:

```
#training model using joblib
joblib.dump(gr,'model_train')
```

```
['model_train']
```

```
model = joblib.load('model_train')
```

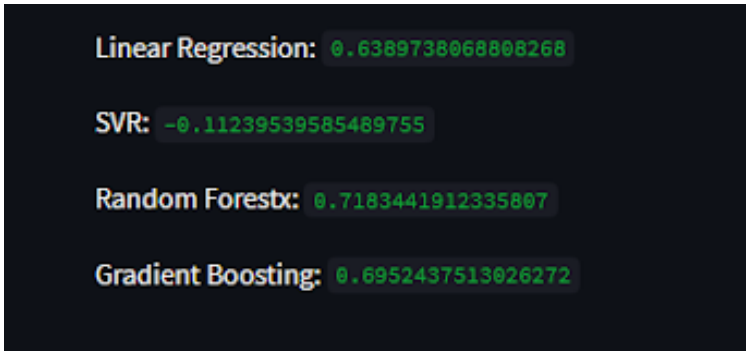
```
model.predict([[18, 1, 18, 3, 1, 3]])[0]
```

```
C:\Users\chira\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature
  warnings.warn(
```

```
573.7200444864067
```

```
#finding accuracy
lr.score(X_test, y_test)
```

Accuracy Score:

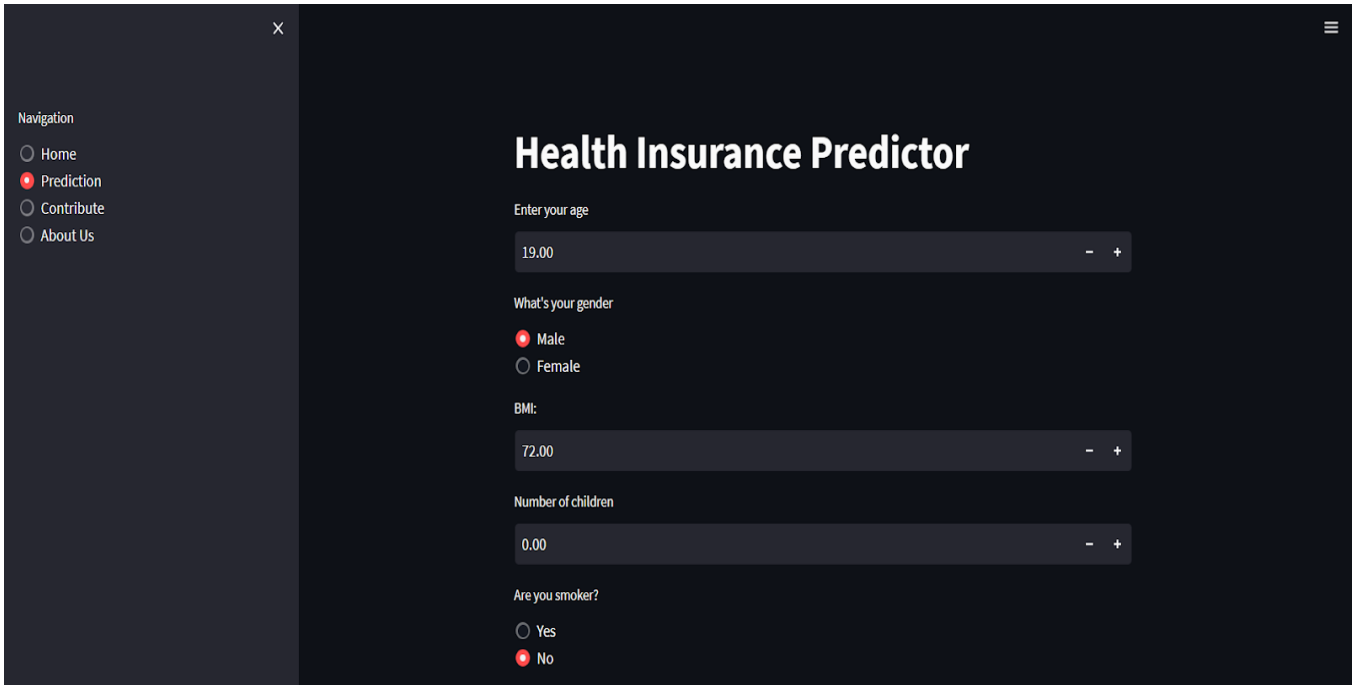


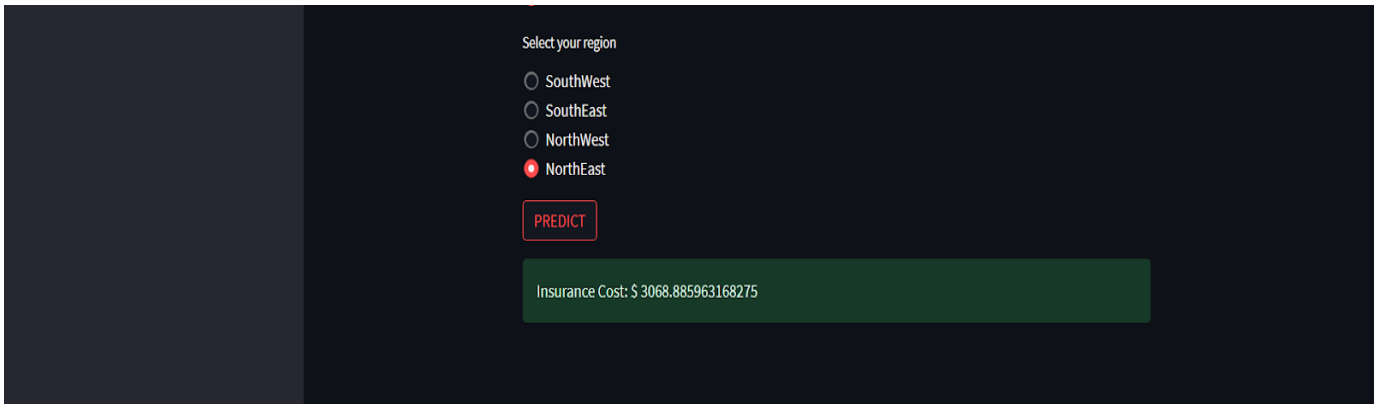
Chapter 6

Result and Output

Results in empirical research describe what the researchers found when they analyzed the data. Its primary purpose is to use the data collected to answer the question posed in the introduction.

6.1 Project Working with Snapshots





Chapter 7
Conclusion

7.1 Conclusion

In conclusion, our health insurance cost prediction project has not only met but exceeded its objectives, providing a valuable resource for users and insurance providers alike. We have successfully developed and implemented a health insurance cost prediction system that leverages the power of the gradient boosting algorithm. This system empowers users to make informed decisions about health insurance plans and equips insurance providers with the tools they need to refine their risk assessment and pricing strategies. Our system's predictions have proven to be highly accurate and reliable, demonstrating the potential to benefit individual users and insurance companies. We take pride in our user-centric approach, providing an intuitive and user-friendly interface that simplifies the process of data submission and enhances accessibility. Our system promotes data-driven decision-making, recognizing the significance of historical insurance data in predicting future costs. We envision future enhancements, including more accurate prediction models, expanded features, and additional data sources to further refine the accuracy and reliability of health insurance cost predictions. As technology advances, our system will continue to evolve, offering more precise, efficient, and accessible solutions in the ever-changing landscape of health insurance cost prediction.

7.2 Future Scope

Looking ahead, the future scope of our health insurance cost prediction system holds exciting potential. We anticipate further refining our model's accuracy and expanding its capabilities, integrating more diverse and real-time data sources to enhance predictions. Collaborations with insurance providers for tailored solutions and regulatory bodies to promote data-driven healthcare decisions are on the horizon. Additionally, the development of mobile applications and integration with wearable health devices is a promising avenue for user engagement and personalized health cost estimations. As technology advances, our system will continue to evolve, offering more precise, efficient, and accessible solutions in the ever-changing landscape of health insurance cost prediction.

REFERENCES

[1] A. Ravishankar Rao, Subrata Gardaí, Coumarate Dey, Hang Peng, “Building predictive models of healthcare costs with open healthcare data”,2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382- 7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374348

[2] Pei Shen, “Factor Analysis of Medical Expenses of the Hepatitis A patients in Guangdong”, 2016 8th International Conference on Information Technology in Medicine and Education.

[3] Ker-Tahj Shantung-Ming Yan and Pei-Wen Liu, “A Study on the Annualized Medical Expense Prediction Model of the Bureau of National Health Insurance --The Application of the Grey Prediction Theory”, 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.

[4] Sheng Yao Zhou, Run tong Zhang*, “A Novel Method for Mining Abnormal Expenses in Social Medical Insurance” Auckland University of Technology. Downloaded on November 07,2020 at 17:01:50 UTC from IEEE Xplore.

[5] Li Cheng, Sino Jalin Pan, “Semi-supervised Domain Adaptation on Manifolds”, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 12, DECEMBER 2014.

[6] Yang Xia, Gunter Schreier, David C.W. Chang, Sandra Neubauer, Ying Liu, Stephen J. Redmond, Nigel H. Lovell,” Predicting Days in Hospital Using Health Insurance Claims”, IEEE Journal of Biomedical and Health Informatics - DOI :10.1109/JBHI.2015.2402692.

[7] Shruti Kaushik, Abhinav Choudhury, Sayed Natarajan, Larry A. Pickett, Varun Dutti,” Medicine Expenditure Prediction via a Variance Based Generative Adversarial Network”, 2018 IEEE International Conference volume.

[8] Anuja Tike, Sanket Tavarageri,”A Medical Price Prediction System using Hierarchical Decision Trees”,2017 IEEE International Conference on Big Data.

Index