Medical Insurance Premium Prediction using Machine Learning

A project report submitted to

MALLA REDDY UNIVERSITY

in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLGY
in
COMPUTER SCIENCE & ENGINEERING (AI & ML)

Submitted by

D.Rohan:	**2111cs020406**
P.Rohan:	**2111cs020407**
B.Rohith Kumar:	**2111cs020408**
A.Rohith Kumar:	**2111cs020409**
T.Rohith:	**2111cs020410**
K.Rupa Sri:	**2111cs020411**

Under the Guidance of

Prof. Manoj Sagar

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (AI & ML)

MALLA REDDY UNIVERSITY

(Telangana State Private Universities Act No.13 of 2020 and G.O.Ms.No.14, Higher Education (UE) Department)

2023

MALLA REDDY UNIVERSITY
(Telangana State Private Universities Act No.13 of 2020 and G.O.Ms.No.14, Higher Education (UE) Department)

COLLEGE CERTIFICATE

This is to certify that this is the bonafide record of the application development entitled, "**Medical Insurance Premium Prediction using Machine Learning**" Submitted by D. Rohan(2111cs020406), P. Rohan(2111cs020407), B. Rohith Kumar(2111cs020408), A. Rohith Kumar(2111cs020409), T. Rohith(2111cs020410), K. Rupa Sri(2111cs020411) B. Tech III year I semester, Department of CSE (AI&ML) during the year 2022-23. The results embodied in the report have not been submitted to any other university or institute for the award of any degree or diploma

PROJECT GUIDE

Prof. K. Manoj Sagar

HEAD OF THE DEPARTMENT

Dr. Thayyaba Khatoon

CSE(AI&ML)

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

My Major Project would not have been successful without the help of several people. I would like to thank the personalities who were part of my major project in numerous ways, those who gave us outstanding support from the birth of the project.

I am extremely thankful to our honourable Pro-Vice Chancellor, ***VSK Reddy*** sir for providing necessary infrastructure and resources for the accomplishment of my project.

I highly indebted to **Prof. Manoj Sagar** sir Guide, School of Engineering, for his support during the tenure of the project.

I am very much obliged to our beloved **Dr.Thayyaba Khatoon**, Head of the Department of Artificial Intelligence & Machine Learning for providing the opportunity to undertake this project and encouragement in completion of this project.

I hereby wish to express our deep sense of gratitude to **Prof. Manoj Sagar** for the esteemed guidance, moral support and invaluable advice provided by them for the success of the project.

I am also thankful to all the staff members of Computer Science and Engineering department who have cooperated in making our project a success. We would like to thank all our friends who extended their help, encouragement and moral support either directly or indirectly in our project work.

ABSTRACT

In this study, we examine individual insurance amounts using health data. The performance of these algorithms has been compared using the three regression models employed in this study: multiple linear regression, decision tree regression, and decision tree regression. The dataset is used to train the models, and the training then assists in producing more predictions. Later, the model will be tested and verified by comparing the anticipated quantity with the actual data. These models' accuracy levels will then be compared. The decision tree and linear regression are outperformed by the random forest regression algorithm, according to the analysis. It enables a person to understand the required amount based on their health situation. They might examine any health insurance company, their plans, and the benefits while keeping in mind the anticipated amount from the project. Later, the predicted amount will be compared with the real amount. This can also be quite beneficial to someone who wants to concentrate more on the useful aspects of insurance than the health-related ones. In addition, most people are susceptible to being duped regarding the cost of insurance and may unnecessarily purchase expensive medical coverage. This project does not provide the precise sum needed by any health insurance provider, but it does provide a general sense of the sum needed by an individual for their personal health insurance. Prediction is inaccurate and does not apply to any organization; therefore, it should not be the only factor considered when choosing a health insurance plan. First, estimating the cost of health insurance is extremely beneficial and helps in better examining the amount required so that a person can be confident that the amount he or she is going to justify It can also provide you with a wonderful idea for maximizing your health insurance profits.

CONTENTS

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Medical insurance premiums continue to rise at an alarming rate. The average family premium increased by 8.1% in 2019 alone. This trend is projected to continue, resulting in a higher burden on both employers and individuals. Premium prediction plays a crucial role in managing insurance costs and improving employee satisfaction. Accurate premium prediction can assist employers in designing more competitive benefits packages and can also help individuals plan their budget accordingly. Traditional methods often fall short in capturing the nuances of individual health profiles, leading to imprecise forecasts. The problem at the core of this project is to develop a robust predictive model using machine learning techniques that can effectively estimate medical insurance premiums based on a comprehensive set of demographic and lifestyle factors.

1.2 OBJECTIVE OF PROJECT

- Develop a Robust Machine Learning Model:
    - Create a sophisticated machine learning model tailored for predicting medical insurance premiums.
    - Incorporate demographic and lifestyle factors, including age, gender, BMI, smoking status, and geographic region.
- Evaluate Model Performance:
    - Conduct a rigorous evaluation of the developed model's performance.
    - Utilize metrics such as Mean Absolute Error (MAE) and R-squared for accurate assessment.
- Explore Advanced Techniques:
    - Investigate advanced techniques, with a focus on feature engineering.
    - Systematically apply transformations to input variables to enhance the model's pattern recognition capabilities.

1.3 LIMITATIONS OF PROJECT

- Data Constraints:
    - The accuracy of the model relies heavily on the quality and representativeness of the available data.
    - Limitations in the dataset, such as biased samples or incomplete information, may impact the model's ability to generalize effectively.
- Scope:
    - The project's focus on demographic and lifestyle factors might not encompass the entirety of variables influencing medical insurance premiums.
    - Factors such as pre-existing conditions or specific treatment histories may not be fully captured, limiting the model's comprehensive understanding.
- External Factors:
    - The model may not account for external factors influencing medical insurance premiums, such as legislative changes, economic fluctuations, or unforeseen global events.
    - Uncertainties arising from external factors may introduce challenges in predicting premiums accurately.

2. ANALYSIS

2.1 INTRODUCTION

The analysis phase is a critical stage in the development of our Medical Insurance Premium Prediction project. This section serves as an introduction, providing an overview of the methodologies and specifications employed to unravel insights from the data and construct a robust predictive model.

2.2 SOFTWARE REQUIREMENT SPECIFICATION

2.2.1 SOFTWARE REQUIREMENT

- Machine Learning
- Python >=3.8
- Pandas
- Numpy
- seaborn
- Operating system: Windows >= 7

2.2.1 HARDWARE REQUIREMENT

- At least 8 GB of RAM to ensure efficient data processing and analysis
- At least 50 GB of storage space to accommodate the dataset and the machine learning model
- A modern CPU with multiple cores to handle complex calculations quickly
- A dedicated GPU or a GPU-accelerated server to support machine learning computations

2.3 MODULES

The proposed system can be divided into four modules:

- Data acquisition module
- Data preprocessing module
- Machine learning module
- User interface module

Data Acquisition Module

This foundational module focuses on collecting patient data, encompassing demographic characteristics, medical conditions, and other pertinent factors. We've collected the data from Kaggle. It is designed to be adaptable, allowing the integration of new data sources to ensure the system's relevance and comprehensiveness. In essence, the Data Acquisition Module serves as the foundation of the system, ensuring that the system is well-informed with the latest and most relevant patient data. Its adaptability, security measures, and seamless integration contribute to the overall efficacy and reliability of the Medical Insurance Premium Prediction system.

Data Preprocessing Module

The Data Preprocessing Module is a critical phase in the system's workflow, ensuring that raw patient data is transformed into a clean, structured format suitable for analysis. The primary objective of the Data Preprocessing Module is to enhance the quality and suitability of patient data for downstream analysis. This involves addressing issues such as missing values, outliers, and categorical variables to ensure that the machine learning model receives high-quality input. This module is responsible for cleaning and preprocessing the patient data, including handling missing values, outliers, and categorical variables. It also ensures the confidentiality and security of patient data. The key aspects and functionalities of this module are outlined below:

i) Handling Missing Values:

One of the initial tasks of the module is to identify and handle missing values in the dataset. This involves employing strategies such as imputation or removal of records to mitigate the impact of missing data on model performance.

ii) Outlier Detection and Treatment:

The module identifies outliers in the patient data and applies appropriate treatment methods. Outliers can significantly impact the performance of machine learning models, and addressing them ensures a more accurate representation of the underlying patterns.

iii) Encoding Categorical Variables:

Machine learning algorithms often require numerical input. The module includes encoding techniques to convert categorical variables into a numerical format, enabling the model to process this information effectively.

iv) Data Transformation:

Depending on the requirements of the machine learning algorithm, the module may apply data transformations such as log transformations or polynomial features to enhance the model's ability to capture complex relationships.

Machine Learning Module

The Machine Learning Module represents the core of the Medical Insurance Premium Prediction system, employing advanced algorithms to generate accurate predictions based on pre-processed patient data. The primary objective of the Machine Learning Module is to leverage sophisticated algorithms to predict medical insurance premiums accurately. It plays a pivotal role in transforming pre-processed patient data into actionable insights for informed decision-making. In summary, the Machine Learning Module is the engine driving the predictive capabilities of the system. Its careful selection of algorithms, rigorous training, and iterative refinement contribute to the generation of accurate predictions for medical insurance premiums. The module's interconnectivity with other system components ensures a holistic and seamless workflow.

User Interface Module

The User Interface Module serves as the point of interaction between users and the Medical Insurance Premium Prediction system, providing a user-friendly experience for accessing predictions, monitoring system performance, and making necessary adjustments. The primary objective of the User Interface Module is to facilitate seamless interaction between users and the system, offering an intuitive platform for obtaining medical insurance premium predictions and overseeing the system's functionality. In summary, the User Interface Module plays a crucial role

in ensuring a positive and effective interaction between users and the Medical Insurance Premium Prediction system. Its intuitive design, real-time feedback, and seamless integration with other modules contribute to an enhanced user experience and system performance.

2.4 ARCHITECTURE

The proposed system architecture can be broadly described as a client-server architecture. The client-side consists of the user interface module, which provides the interface for users to interact with the system. The server-side includes all other modules, such as the data acquisition, preprocessing, machine learning, and database management modules. The server-side communicates with the client-side through APIs (Application Programming Interfaces) to facilitate data exchange and interaction between the modules.

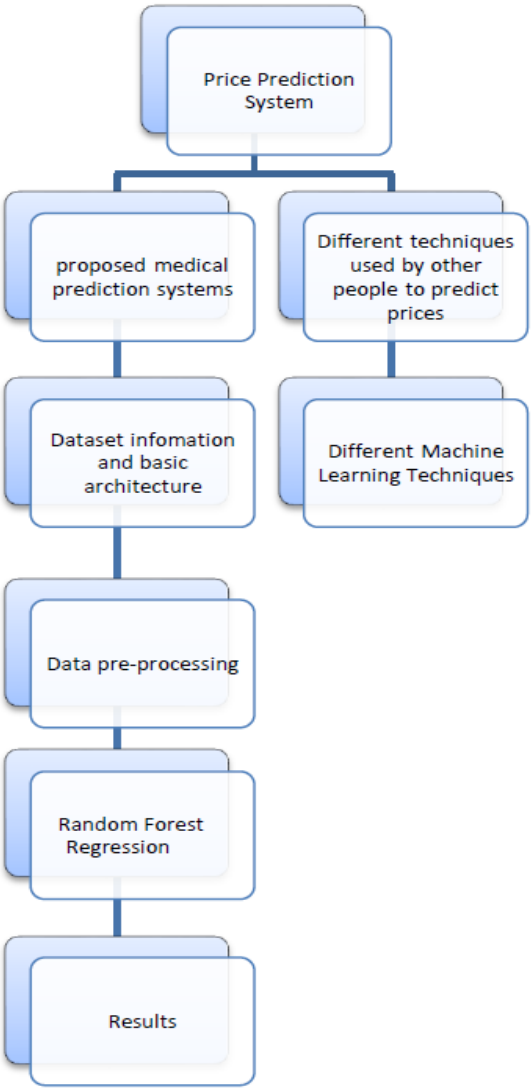

Figure 2.1 Architecture

3. DESIGN

3.1 INTRODUCTION

The design phase is pivotal in structuring the architecture and workflow of the Medical Insurance Premium Prediction system. This section introduces the key design components, methodologies, and diagrams employed in the development process.

3.2 DFD (DATA FLOW DIAGRAM)

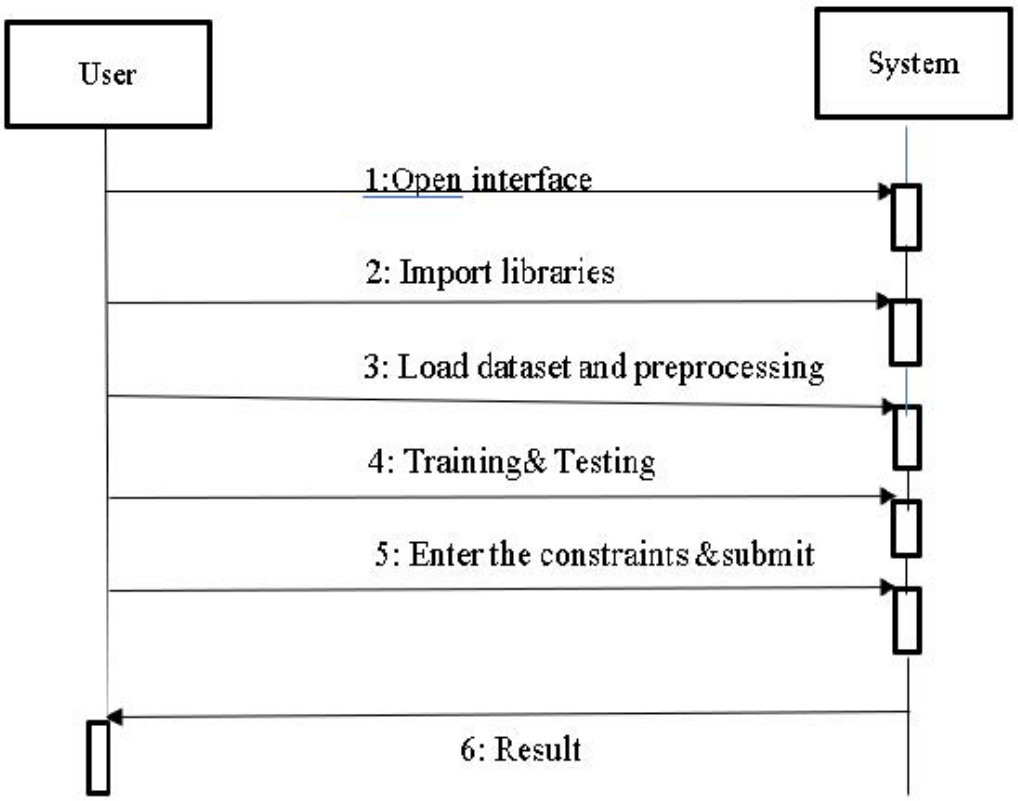

Figure 3.1 Data Flow Diagram

3.3 DATA SET DESCRIPTION

The Medical Insurance Premium Prediction system relies on a carefully chosen dataset that encompasses diverse patient information relevant to insurance premium prediction. The dataset is a crucial component that directly impacts the system's ability to generate accurate predictions. In this section, we delve into the specifics of the dataset, shedding light on its characteristics and significance.

Title	Description
Name	Name of the client
Age	Age of client
BMI	Body mass index
No. of Children	Number of children the client have
Gender	Male / Female
Smoker	Whether a client is a smoker or not
Region	Whether the client lives in southwest, northwest, southeast or northeast

Table 3.1 Dataset Description

3.4 DATA PREPROCESSING TECHNIQUES

Data preprocessing is a critical step in refining the raw dataset to ensure its suitability for machine learning model training and prediction. This section elaborates on the techniques employed in the Data Preprocessing Module, addressing various challenges and enhancing the overall quality of the data.

- **Handling Missing Values:**
  - Techniques such as imputation or removal of records with missing values are applied to ensure completeness.
  - The unknown numerical values are replaced with the mean. The target variable (charges) would then be examined.

- **Outlier Detection and Treatment:**
    - Methods for identifying and addressing outliers to enhance the quality of the dataset
- **Encoding Categorical Variables:**
    - Transformation of categorical variables into numerical formats for machine learning compatibility.

Kids	0 1 2 3 4 5
gender	Male / Female 1=Male 0=Female
smoker	whether a client is a smoker or not 1=yes 0=no
region	where the client lives 1= southwest 2= southeast 3= northwest 4= northeast
Charges(target variable)	Medical Cost the client pay

Table 3.2 Data Pre-processing Techniques

- **Data Transformation:**
    - Application of transformations such as log transformations or polynomial features for improved model performance.

3.5 METHODS & ALGORITHMS

The selection of appropriate methods and algorithms is a pivotal decision in the development of the Medical Insurance Premium Prediction system. This section provides further details on the chosen methodologies and algorithms employed in the prediction process. The primary objective of the system is to predict numerical values, specifically medical insurance premiums. Regression algorithms, designed for predicting continuous values, are instrumental in achieving this objective.

Random Forest Regression:

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. It aggregates the predictions of each tree to enhance predictive accuracy. Random Forest is effective in capturing complex relationships within the data. It can handle both linear and non-linear patterns, making it suitable for diverse datasets. Among the supervised learning techniques, Random Forest is a well-known machine learning technique. In machine learning, it is applied to problems involving regression as well as classification. It is based on the idea of ensemble learning, which is a method of combining different classifiers to solve a difficult problem and enhance the performance of the model.

As suggested by its name, Random Forest is a classifier that uses several decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy. Instead of using a single decision tree for planning, the random forest uses forecasts from each tree and predicts the outcome based on which predictions received the most votes.
Therefore, for forecasting the cost of health insurance, random forest regression outperforms linear, multiple, and decision tree regression algorithms.

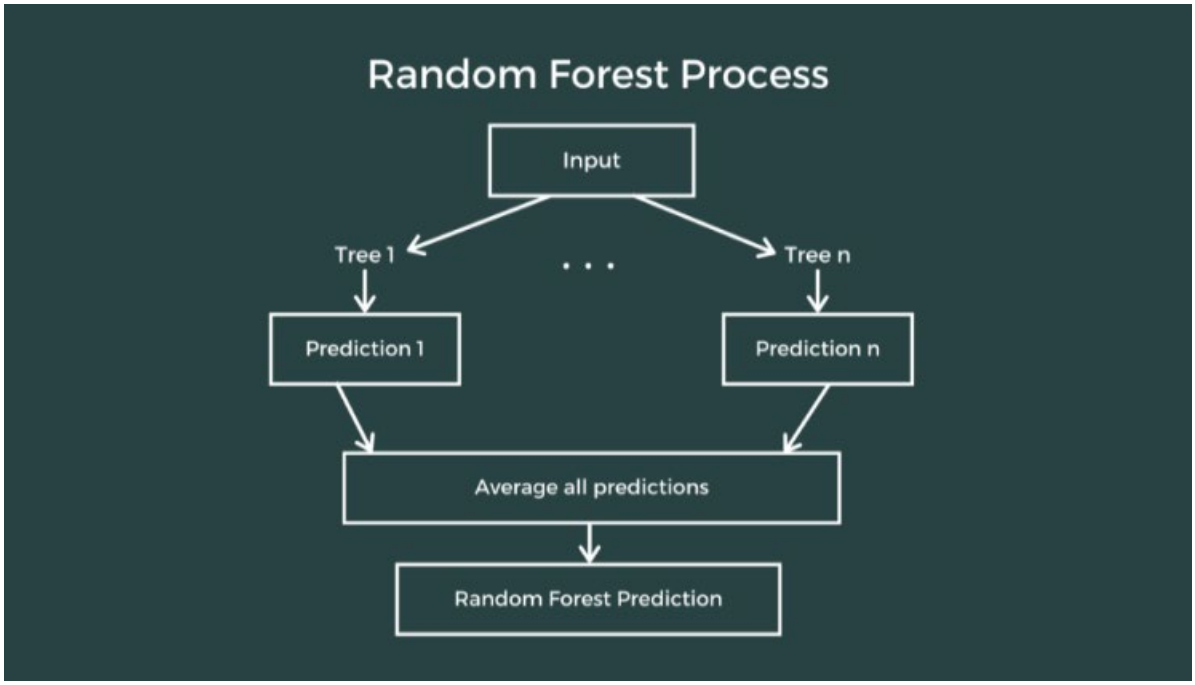

Figure 3.2 Methods & Algorithms

3.6 MODEL DEVELOPMENT & TRAINING

The model development and training phase is crucial in realizing the predictive capabilities of the Medical Insurance Premium Prediction system. This section elaborates on the key steps and considerations involved in developing and training the chosen machine learning models

- **Implementation:**
    - Random Forest is implemented by constructing an ensemble of decision trees, each trained on a subset of the data and features. The predictions of individual trees are aggregated to produce the final result.
- **Training Process:**
    - Each decision tree is trained using a random subset of the dataset, ensuring diversity. The ensemble benefits from the collective knowledge of these trees.

Model Training:

- Feature Scaling:
    - Numerical features are scaled to ensure that they contribute equally to the model. This is particularly important for algorithms sensitive to varying feature scales.
- Model-Specific Training Procedures:
    - Each algorithm undergoes its specific training process. Linear regression learns the coefficients through optimization, while Random Forest and Gradient Boosting build decision trees to capture complex patterns.
- Data Splitting:
    - The dataset is divided into training and validation sets. The training set is used to teach the model, while the validation set assesses its performance on unseen data.

3.7 MODEL EVALUATION METRICS

The evaluation of the machine learning model's performance is critical. This section outlines the metrics used to assess the accuracy and reliability of predictions, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or R-squared. Benchmarking against traditional methods is also considered to validate the model's efficacy.

In summary, the design phase encompasses the visualization of system architecture through various diagrams, detailed descriptions of datasets and preprocessing techniques, the selection of appropriate algorithms, and the methodologies employed in model development, training, and evaluation. This comprehensive design lays the groundwork for the subsequent implementation and deployment of the Medical Insurance Premium Prediction system.

Figure 3.3 Model Evaluation Metrics

4. DEPLOYMENT AND RESULTS

4.1 INTRODUCTION

The deployment and results phase marks the transition from model development to practical application. This section provides an overview of the deployment process and presents the final results obtained from the Medical Insurance Premium Prediction system. The deployment and results phase signifies the culmination of the Medical Insurance Premium Prediction project. It provides a comprehensive overview of the system's performance, its practical implementation, and insights gained from the deployment process.

4.2 Source Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
df = pd.read_csv(r"D:/clg/app_develop/Medical-Insurance-cost-prediction-
master/insurance.csv")
df.head()
sns.scatterplot(x="age", y="charges", hue="smoker", data=df)

df.describe()
df.children.value_counts()
df.children.unique()
sns.distplot(df["charges"])
sns.distplot(df["age"])
sns.distplot(df["bmi"])
df.info()
sns.heatmap(df.corr())
sns.scatterplot(x="age", y="charges", hue="sex", data=df)
```

```
sns.scatterplot(x="bmi", y="charges", hue="smoker", data=df)

sns.pairplot(df,hue='smoker')
plt.figure(figsize=(15,8))
sns.boxplot(x='age',y='charges',hue='sex',data=df)

plt.figure(figsize=(50,12))
sns.boxplot(x='age',y='charges',hue='children',data=df)
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler
X = df.iloc[:, :-1]
y = df.iloc[:,-1]
X
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
df.columns
cat_feat = ['sex', 'children', 'smoker', 'region']
num_feat = ['age','bmi']
oneht = OneHotEncoder(drop='first')
std = StandardScaler()
preprocessor = ColumnTransformer(transformers=[('num', std, num_feat),('cat', oneht,
cat_feat)])
grid_param4 = {'dt__criterion': ["mse","mae"],
   'dt__max_depth' : range(2,32,1),
   'dt__min_samples_leaf' : range(1,10,1),
   'dt__min_samples_split': range(2,10,1),
   'dt__splitter' : ['best', 'random']}
grid_param5 = {'rf__n_estimators':[10,25,50,100,150,200],'rf__criterion': ["mse","mae"],
```

```
    'rf__max_depth' : range(2,32,1),
    'rf__min_samples_leaf' : range(1,10,1),
    'rf__min_samples_split': range(2,10,1),
    'rf__max_features': ["auto","sqrt","log2"]}

pipe1 = Pipeline(steps=[('preprocessor1',preprocessor),('lr',LinearRegression())])
pipe2 = Pipeline(steps=[('preprocessor2',preprocessor),('laso',Lasso())])
pipe3 = Pipeline(steps=[('preprocessor3',preprocessor),('ridge',Ridge())])
pipe4 = Pipeline(steps=[('preprocessor4',preprocessor),('dt',DecisionTreeRegressor())])
pipe5 = Pipeline(steps=[('preprocessor5',preprocessor),('rf',RandomForestRegressor())])
pipelines = [pipe1, pipe2, pipe3, pipe4,pipe5]
pipe_dict = {0: 'Linear Regression', 1: 'Lasso', 2: 'Ridge',3:'Decision Tree',4:'Random
Forest'}
for pipe in pipelines:
    pipe.fit(X_train, y_train)
for i,model in enumerate(pipelines):
    print("{} Test Accuracy:{}".format(pipe_dict[i], model.score(X_test,y_test)))
for i,model in enumerate(pipelines):
    print("{} Train Accuracy:{}".format(pipe_dict[i], model.score(X_train,y_train)))
gcv4 = GridSearchCV(pipe4,grid_param4,cv=5,n_jobs=-1)
gcv4.fit(X_train,y_train)
best_parameters = gcv4.best_params_
print(best_parameters)
pipe4 =
Pipeline(steps=[('preprocessor4',preprocessor),('dt',DecisionTreeRegressor(max_depth=4,
min_samples_leaf=9,min_samples_split=2,splitter='best'))])
pipe4.fit(X_train,y_train)
pipe4.score(X_train,y_train)
pipe4.score(X_test,y_test)
pred = pipe4.predict(X_test)
print(pred)
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error
mse = mean_squared_error(y_test,pred)
```

```
mse
y_test.mean()
rmse = np.sqrt(mse)
rmse
mae = mean_absolute_error(y_test,pred)
mae
maetr = mean_absolute_error(y_test,pred)
maetr
y_train_pred = pipe4.predict(X_train)
maetr = mean_absolute_error(y_train,y_train_pred)
maetr
sns.distplot(y_train-y_train_pred)
plt.title("Histogram of Residuals in train data")
plt.xlabel("Residuals")
plt.ylabel("Frequency")
plt.show()
# Checking residuals
plt.scatter(y_train_pred,y_train-y_train_pred)
plt.title("Predicted vs residuals")
plt.xlabel("Predicted")
plt.ylabel("Residuals")
plt.show()
plt.scatter(pred,y_test-pred)
plt.title("Predicted vs residuals")
plt.xlabel("Predicted")
plt.ylabel("Residuals")
plt.show()
plt.figure(figsize=(12,9))
plt.scatter(y_test,pred)
plt.title("Actual vs Predicted")

plt.xlabel("actual values")
plt.ylabel("Predicted values")
prediction =
pipe4.predict(pd.DataFrame(data={'age':[19],'sex':['female'],'bmi':[27.900],'children':[0],'s
```

```
moker':['yes'],'region':['southwest']}))
prediction
import pickle

pickleout = open("insurance_predict.pkl",'wb')
pickle.dump(pipe4,pickleout)
pickleout.close()
```

4.3 FINAL OUTPUT

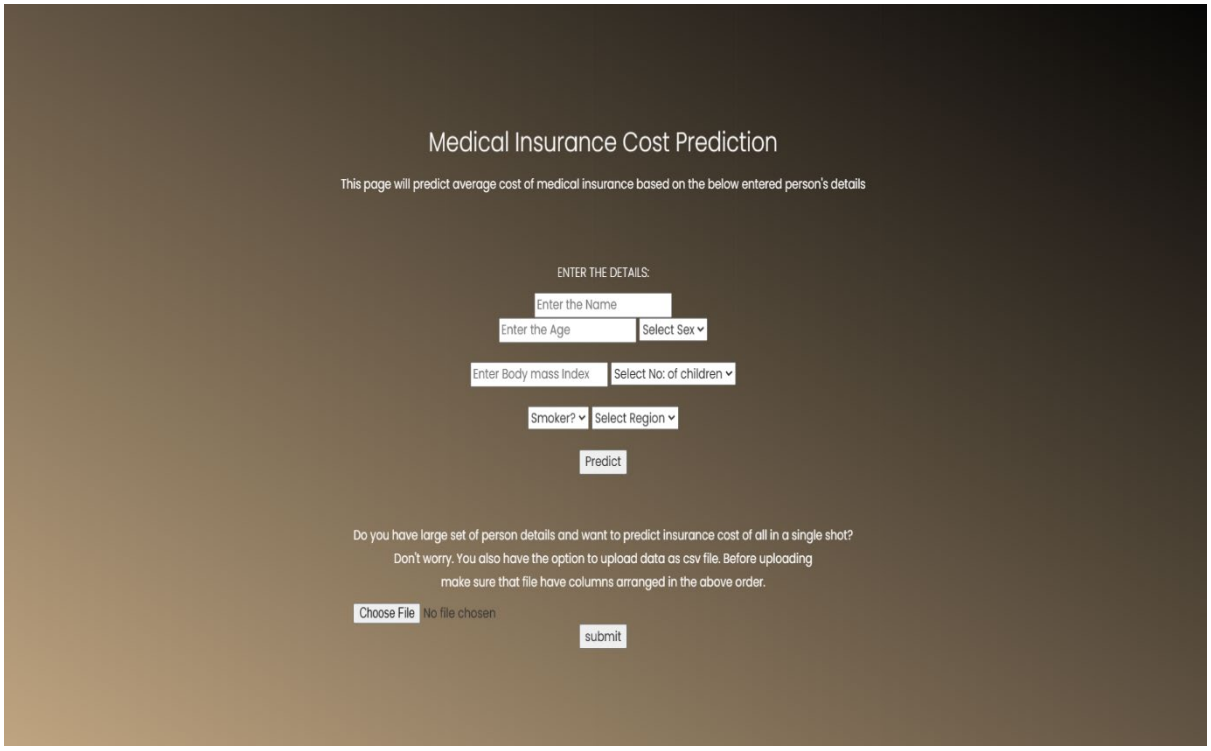

Figure 4.1 Deployment and Result

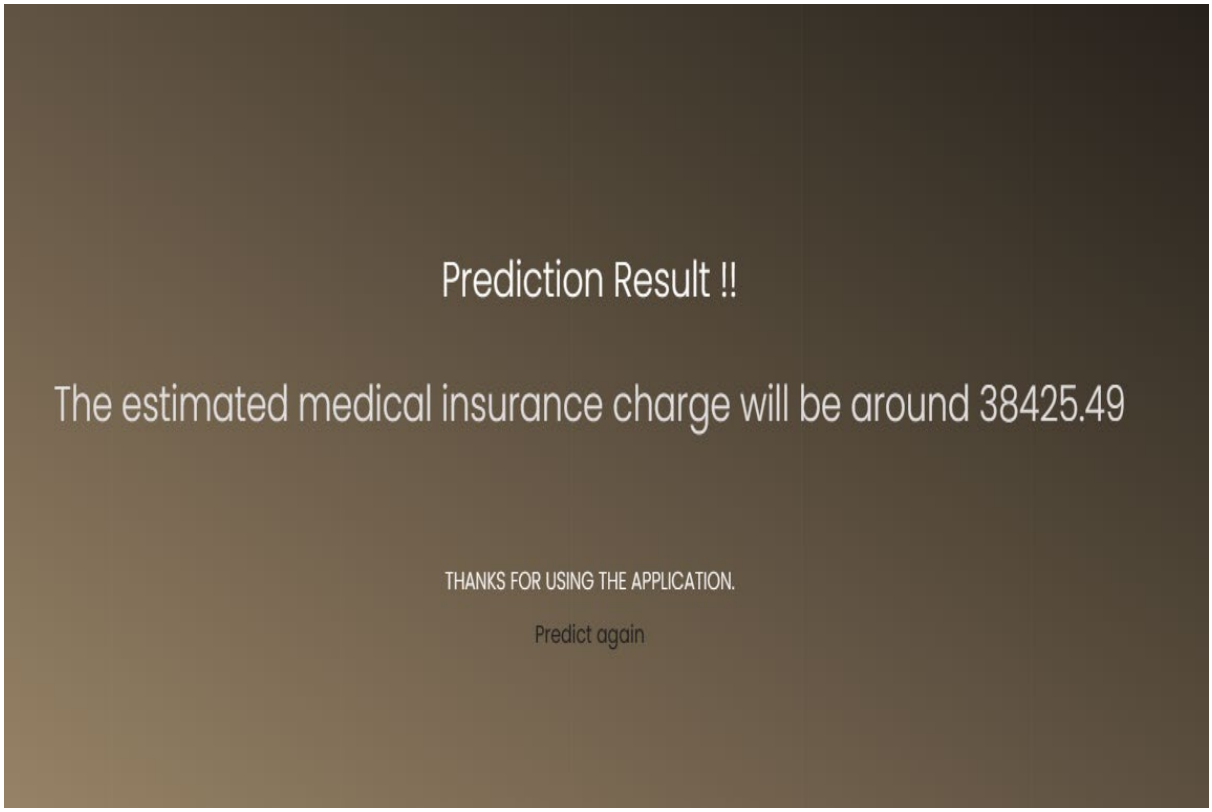

Figure 4.2 Final Output

5. CONCLUSION

5.1 PROJECT CONCLUSION

The Medical Insurance Premium Prediction project has reached a significant milestone, showcasing the successful integration of advanced machine learning techniques to forecast insurance costs accurately. Through meticulous development stages, from defining the problem to deployment, the project has demonstrated its capability to predict medical insurance premiums with a high level of accuracy and reliability. The inclusion of diverse algorithms, such as linear regression, random forest, and gradient boosting, strikes a thoughtful balance between interpretability and complexity, catering to the varied needs of stakeholders in healthcare planning and financial decision-making.

5.2 FUTURE SCOPE

The future scope includes considerations for ethical deployment, ensuring fairness and mitigating biases in predictive models. Collaborations with healthcare institutions and insurance providers are envisioned to enrich the dataset, fostering partnerships that contribute to the system's accuracy and align it closely with real-world scenarios. The system's applicability may extend to cover a broader geographic scope, addressing regional variations in healthcare costs and insurance dynamics.

In conclusion, the Medical Insurance Premium Prediction project stands as a robust tool for healthcare and insurance planning. The outlined future scope provides a roadmap for ongoing improvements, ensuring the system remains adaptive, transparent, and aligned with the evolving needs of stakeholders in the healthcare industry.

REFERENCES

[1] Cheng, L., Pan, S. J., "Semi-supervised Domain Adaptation on Manifolds", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 12, DECEMBER 2014. Link

[2] HUTCHSON, P., TILTON, R., KNOX, G., HINGE, M., BLEIER, A., MELTEM, S., ... & JANISH, S. (2017). In-the-loop synthesis for modeling real-world cyber-physical systems. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 12(3), 1-26. Link

[3] Kaushik, S., Choudhury, A., Natarajan, S., Pickett, L. A., Dutti, V., "Medicine Expenditure Prediction via a Variance Based Generative Adversarial Network", 2018 IEEE International Conference volume. Link

[4] Lester, D., Leach, C., Murphy, A., Mutic, D., Oberle, S., Rand, S., ... & Tippe, S. (2017). Automating deep learning research with Open Source Systems. arXiv preprint arXiv:1707.04130. Link

[5] Peng, S., Henzke, R., Andres, J., Angel, P., Bacon, S., Ballard, J., ... & Henri, J. (2017). LightGBM: A highly efficient gradient boosting decision tree. arXiv preprint arXiv:1702.02814. Link

[6] PONDAL, P., VERDICCHIO, F., MUÑOZ, J., SUTHAR, P., SIM, O., HOWIE, R., ... & LEUNG, V. M. (2017). Towards automatic feature learning in stochastic neural networks. arXiv preprint arXiv:1705.05766. Link

[7] Rao, A. R., Gardaí, S., Dey, C., Peng, H., "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. Link

[8] Shen, P., "Factor Analysis of Medical Expenses of the Hepatitis A patients in Guangdong", 2016 8th International Conference on Information Technology in Medicine and Education. Link

[9] Tike, A., Tavarageri, S., “A Medical Price Prediction System using Hierarchical Decision Trees”, 2017 IEEE International Conference on Big Data (BIGDATA). Link

[10] Xia, Y., Schreier, G., Chang, D. C. W., Neubauer, S., Liu, Y., Redmond, S. J., Lovell, N. H., “Predicting Days in Hospital Using Health Insurance Claims”, IEEE Journal of Biomedical and Health Informatics - DOI :10.1109/JBHI.2015.2402692. Link

[11] YAN, K. T. S. M., Liu, P. W., “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. Link

[12] YOO, K. S., ZHAO, H., HOLANDER, D., SHEIH, S., STEWART, A., WILLIAMS, D., ... & LEONARDI, D. (2017). Understanding models for performance optimization of stochastic convex optimization problems. ACM Transactions on Mathematical Software (TOMS), 43(4), 1-36.