Project Report

on

Health Insurance Cost Prediction Using IBM Auto AI Service

Prepared By

Prof. Swati L. Kariya
Assistant Professor,
Dept. of CSE/ IT
SVMIT, Bharuch

1. Introduction

1.1 Overview

The project titled as ***"Health Insurance Cost Prediction Using IBM Auto AI Service"*** is mainly developed here as a part of **IBM Gurucool program,** of 6 days Training and Project Build-a-thon, exclusively for faculty.
The project aims to predict the Health Insurance Cost of a person by considering various parameters like age, sex, habit of smoking, BMI, gender, and region. Its basically an application which uses IBM AutoAI Service and Machine learning to determine health insurance cost based on lifestyle and body parameters.

1.2 Purpose

In this project, we have studied the effects of age, smoking, BMI, gender, and region to determine how much of a difference these factors can make on your insurance premium. By using our application, customers see the radical difference their lifestyle choices make on their insurance charges. By leveraging artificial intelligence (AI) and machine learning, it help customers understand just how much smoking increases their premium by predicting how much they will have to pay within seconds.

To build this project IBM AutoAI has been used .A model is created from a data set that includes the age, gender, BMI, number of children, smoking preferences, region, and charges to predict the health insurance premium cost that an individual pays.

Following Services have been Used of IBM Cloud:

1. IBM Watson Studio
2. IBM Watson Machine Learning
3. Node-RED
4. IBM Cloud Object Storage

2. Literature Survey

2.1 Existing problem

Rising health care costs are a major public health issue. Thus, accurately predicting future costs and understanding which factors contribute to increases in health care expenditures are important. The objective of this project was to predict person' healthcare costs development in the subsequent year and to identify factors contributing to this prediction, with a particular focus on the role of pharmacotherapy.

One of the key components to restrain the rise in health care costs is access to an accurate medical price prediction system. That is, if patients have accurate information on medical pricing such as, a certain medical procedure costs dollar amount X at hospital A, the same procedure comes with a price tag of Y at hospital B then they have the opportunity to choose the provider that costs them less. so a unique system need to be developed to determine the prediction of health insurance of a person.

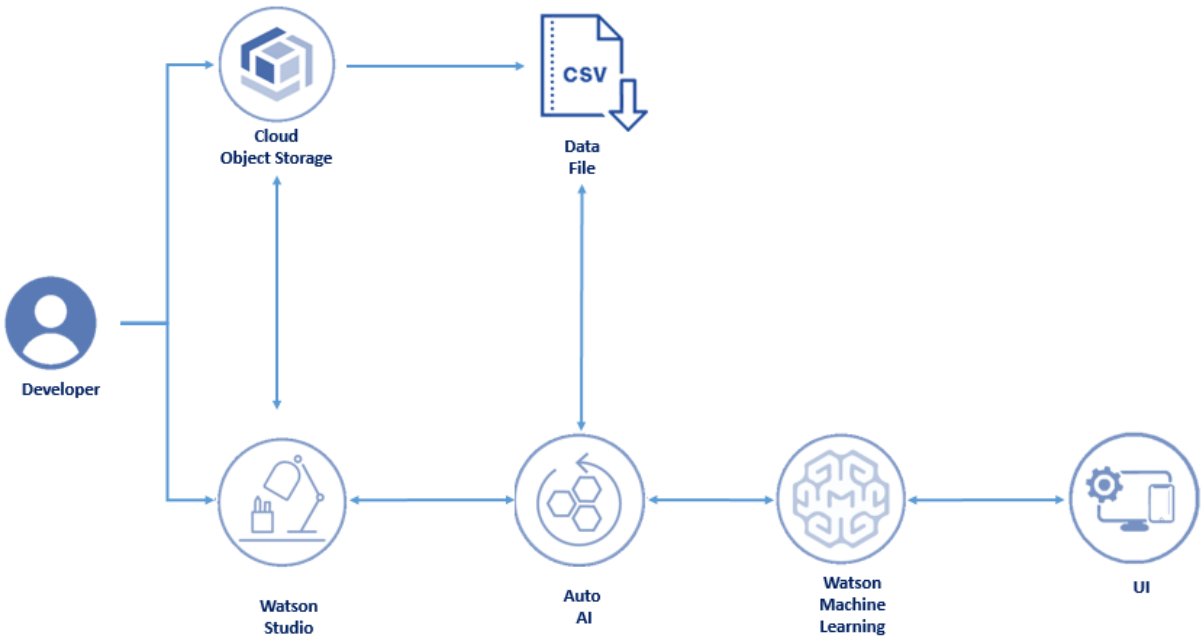
2.2 Proposed Solution

In this project our goal is to predict health insurance based on the data we have in hand. This system is using machine learning and AI techniques with the trained model to get the results of test data. Such a system will be useful for individuals, and government officials and insurance companies alike.

In this project the UI takes various parameters like age, BMI, No of dependents, regions, sex and Smoking habits to determine the insurance cose prediction. User can simply check how the effect of smoking habits implies to the cost of health insurance.

3. Theoretical Analysis

3.1 Block Diagram

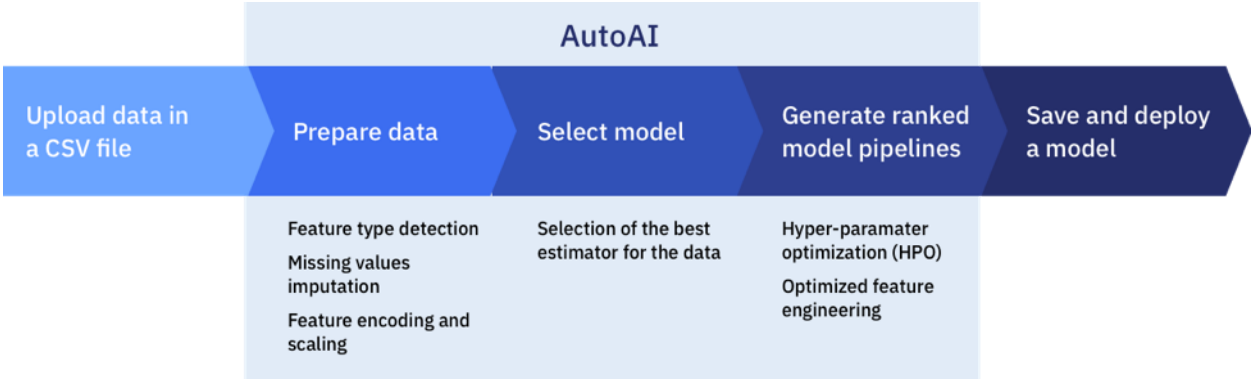


3.2 Hardware/ Software Designing

In this project,Following Services have been Used of IBM Cloud:

1. IBM Watson Studio
2. IBM Watson Machine Learning
3. Node-RED
4. IBM Cloud Object Storage

The Concept of IBM Auto AI is mainly used in this project, which is having following flow to create and deploy the application.



A Dataset is taken from kaggle's health insurance data.

Columns Description :
-**Age:** Age of primary beneficiary
-**Sex:** Primary beneficiary's gender
- **BMI:** Body mass index (providing an understanding of the body, weights that are relatively high or low relative to height)
- **Children:** Number of children covered by health insurance / Number of dependents
-**Smoker:** Smoking (yes, no)
- **Region:** Beneficiary's residential area in the US (northeast, southeast, southwest, northwest)
- **Charges:** Individual medical costs billed by health insurance

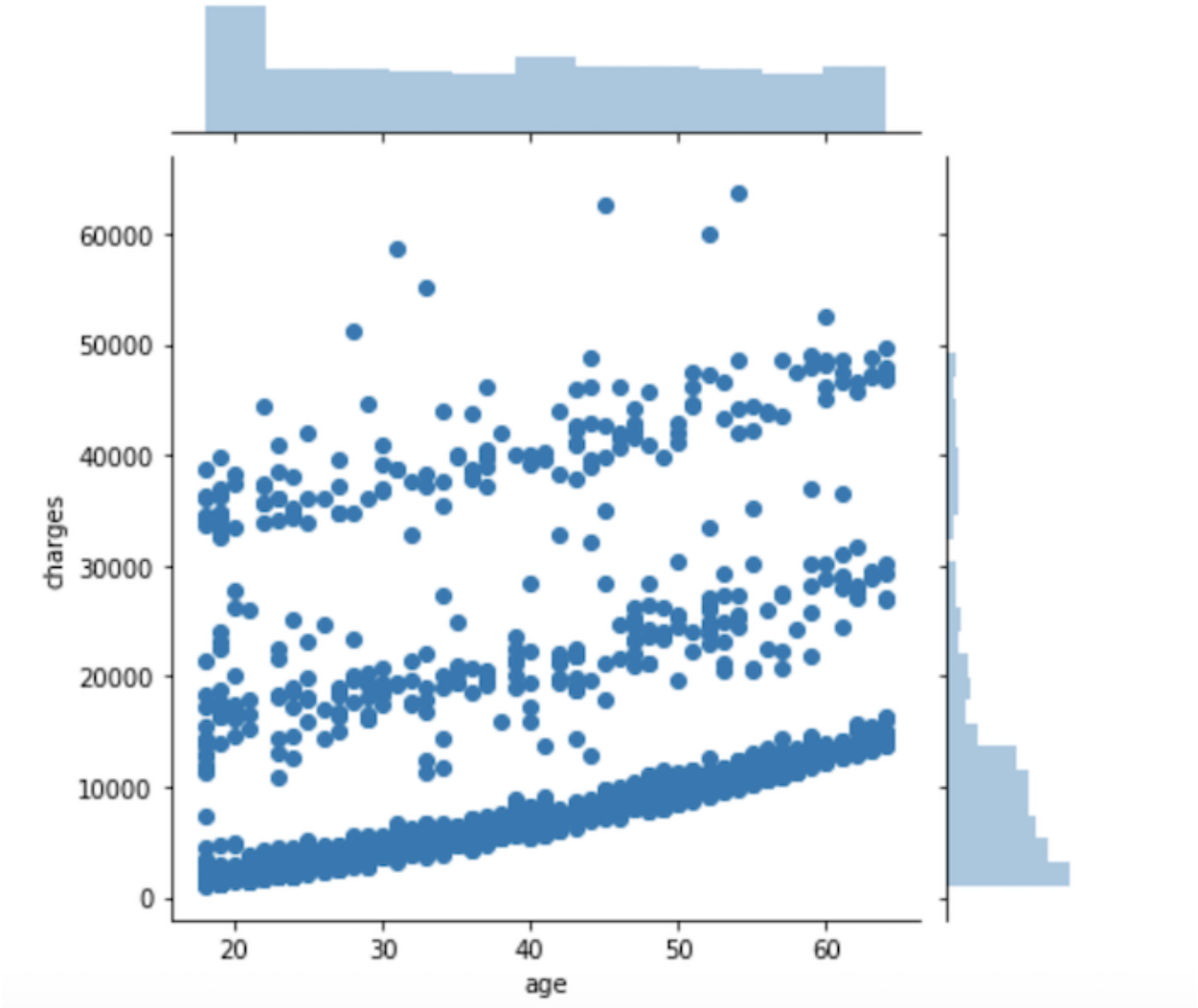
Using IBM Auto AI service the project will simply develop by adding dataset, choose the prediction column according to requirement and input columns. it will simply choose the type of machine learning algorithm best fitted to predict the result.
In this project, the project is of supervised machine learning with regression technique as it is having continuous values in output.

4. Results and Experimental Investigation

Exploratory Data Analysis

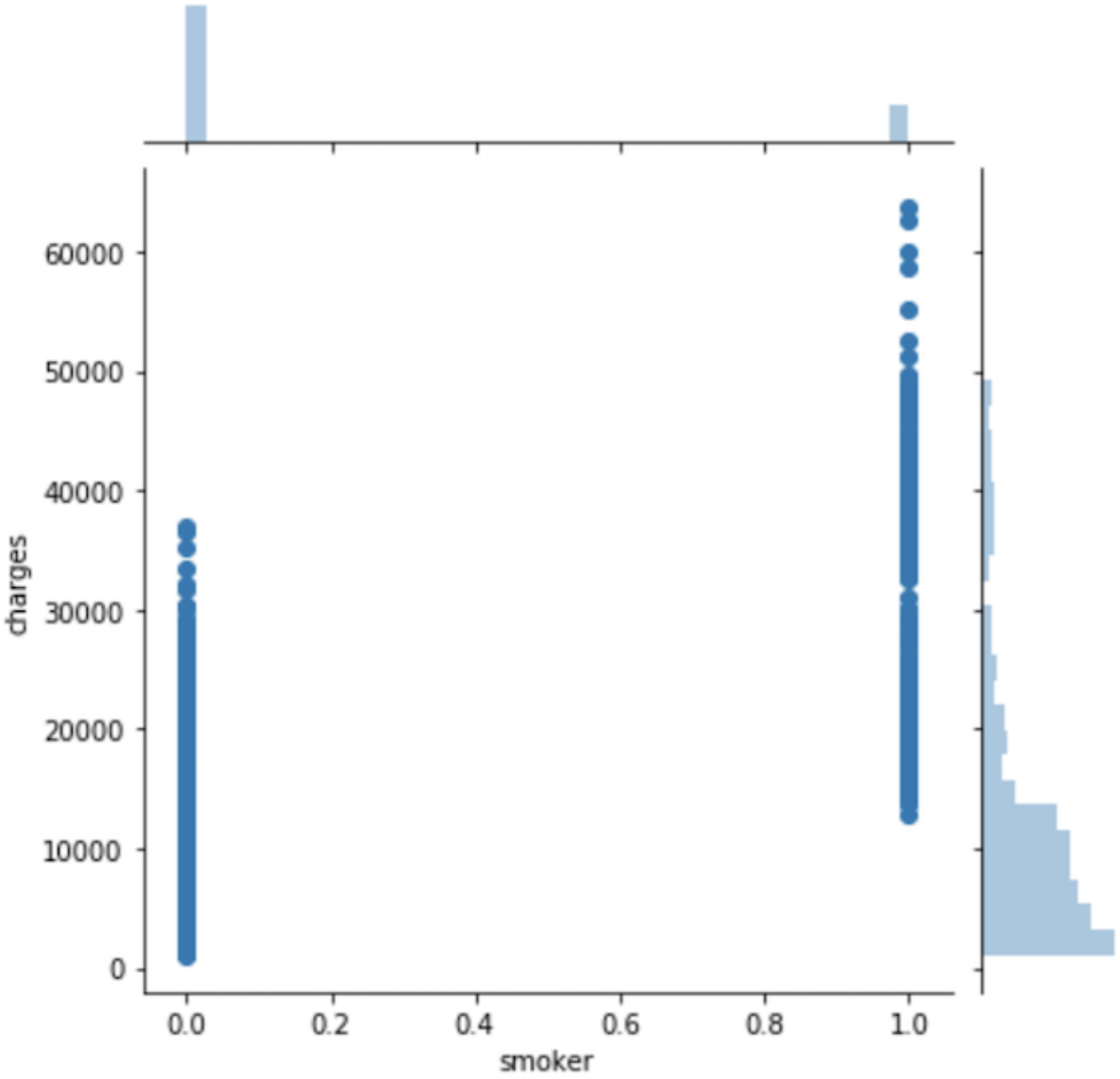
Let's create some simple plots to check out the data.

```
import seaborn as sns
# Correlation betweeen 'charges' and 'age'
sns.jointplot(x=insurance['age'],y=insurance['charges'])
```



Here we see that as Age goes up Charges for health insurance also trends up

```
# Correlation betweeen 'charges' and 'smoker'
sns.jointplot(x=insurance['smoker'],y=insurance['charges'])
```



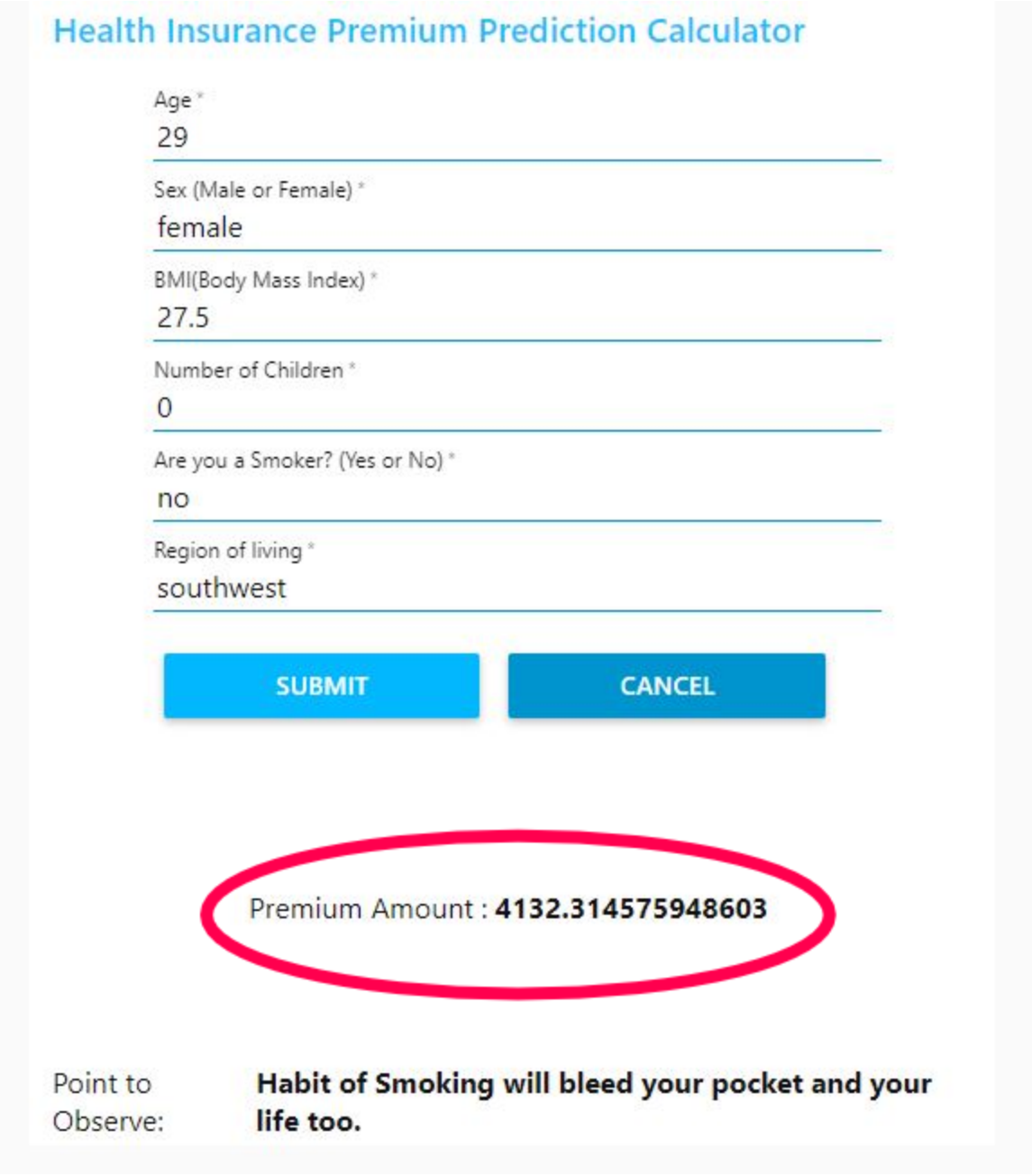
0: Non-smoker, 1: Smoker

Here we see that charges for smokers are higher than non-smokers

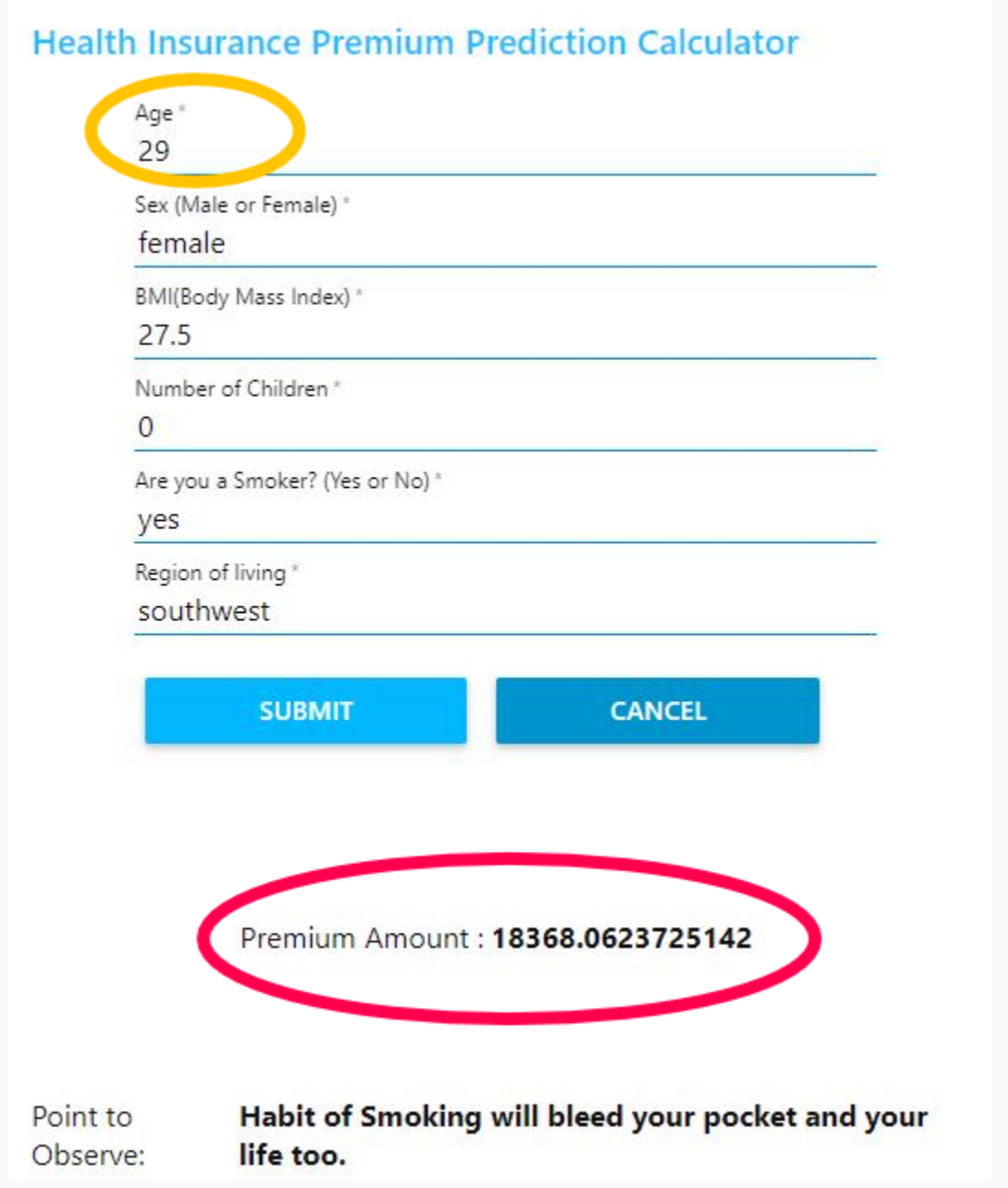
Sample Output Screenshots :

Node-red URL : **https://node-red-vvgho-2020-10-18.eu-gb.mybluemix.net/ui**

UI Screenshot1 :

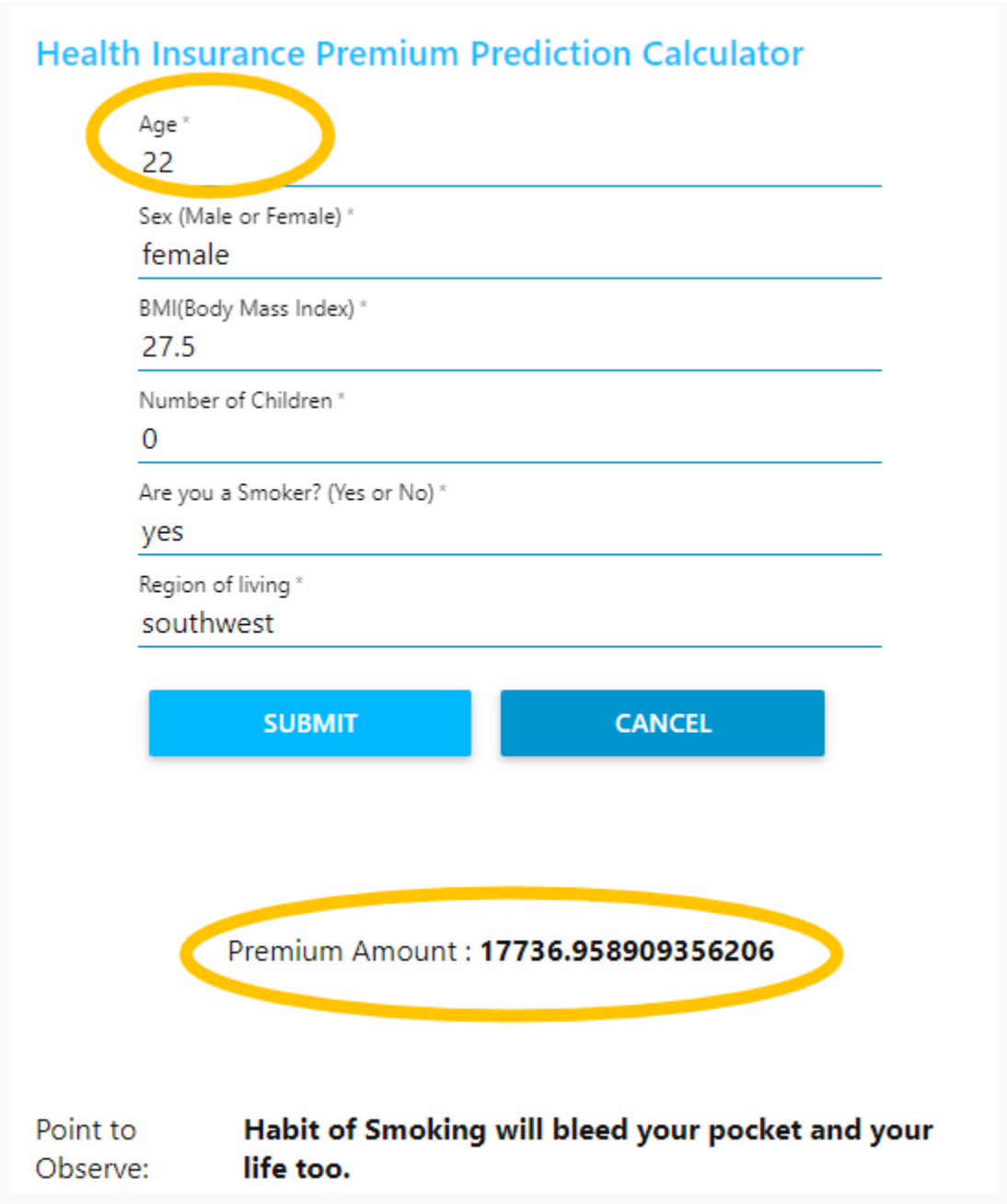


UI Screenshot 2 :



The above two screenshot is having only the difference in smoking habit. we can clearly observe from premium amount value that if a person is a smoker the amount he has to pay is more than the non-smoker.

UI Screenshot 3 :



The above two screenshot is having only the difference in Age. We can clearly observe from premium amount value that if a person is taking insurance at younger age the premium amout will be less compare to taking insurance at older age.

5. Conclusion and future scope

Conclusion:

In this project, we have successfully implemented various IBM Cloud services. IBM watson studio is a very powerful and flexible architecture using which it is very easy to prepare data and build models at scale across any cloud. We can easily integrate IBM Auto AI and Machine learning instance to our project as well as it is easy to build and deploy model using NODE-RED service of IBM cloud. Without going in depth knowledge of algorithms it can choose the best fitted algorithm by itself for best outcome and prediction results.

Future scope:

As possible future work for this project, we can add new features to the dataset we are already using for more accurate predictions. Another addition in the future work can be, making the system even more scalable. Right now we are using few thousands of records to train and test the algorithm. In future, we can try to scale the algorithm for a larger dataset having at least a million records and see the results for it.

6. Bibiliography

- https://www.kaggle.com/mirichoi0218/insurance

Appendix

A. Source code

```
import pandas as pd
insurance = pd.read_csv("insurance.csv")

# Replacing string values to numbers
insurance['sex'] = insurance['sex'].apply({'male':0,      'female':1}.get)

insurance['smoker'] = insurance['smoker'].apply({'yes':1, 'no':0}.get)

insurance['region'] = insurance['region'].apply({'southwest':1, 'southeast':2,
'northwest':3, 'northeast':4}.get)

import seaborn as sns
# Correlation betweeen 'charges' and 'age'
sns.jointplot(x=insurance['age'],y=insurance['charges'])

# Correlation betweeen 'charges' and 'smoker'
sns.jointplot(x=insurance['smoker'],y=insurance['charges'])

# features
X = insurance[['age', 'sex', 'bmi', 'children','smoker','region']]
# predicted variable
y = insurance['charges']

# importing train_test_split model
from sklearn.model_selection import train_test_split

# splitting train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4)

len(X_test) # 402
len(X_train) # 936
len(insurance) # 1338
```

```
# importing the model
from sklearn.linear_model import LinearRegression
model = LinearRegression()
# Fit linear model by passing training dataset
model.fit(X_train,y_train)

# Predicting the target variable for test datset
predictions = model.predict(X_test)

import matplotlib.pyplot as pltplt.scatter(y_test,predictions)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')

# Predict charges for new customer : Name- Rahul
data = {'age' : 40,
        'sex' : 1,
        'bmi' : 45.50,
        'children' : 4,
        'smoker' : 1,
        'region' : 3}
index = [1]
Rahul_df = pd.DataFrame(data,index)
Rahul_df

prediction_Rahul = model.predict(Rahul_df)
print("Medical Insurance cost for Rahul is : ",prediction_Rahul)
```