# Project Documentation

Project Report on

**Predicting Health Insurance Price using Machine Learning**

Under

**Remote Summer Internship Program 2020**
**by SmartInternz**

Project by :

**Sandhya Ramoliya**
**Vellore Institute of Technology, Bhopal**

# INDEX

# 1. Introduction

## 1.1 Overview

Health Insurance plays a vital role in lessening tremendous expenditures individuals incur in conditions of dubious clinical circumstances. With expanding clinical costs, downturns, new illnesses spreading all finished, ownership of medical coverage strategy makes it simple to battle these troublesome occasions. Hence, a thorough and dependable medical coverage can ease some portion of your pressure related with becoming sick or getting hospitalized. This project is about building a model which will consider data of about 1300 people and predict the health insurance price for new individuals.

## 1.2 Purpose

Health insurance covers hospitalization expenses, day care procedures, domiciliary expenses, and ambulance charges, besides many others. And while determining premiums for their customers, health insurance companies face certain difficulties. Hence, the sole purpose of this project is to consider factors like age, sex, bmi, number of children, region etc of past cases and to predict health insurance price for different people.

Using kaggle dataset and IBM Watson Auto AI, a model was built which basically used Random Forest classification algorithmn to train, test and predict. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called

Bootstrap and Aggregation, commonly known as bagging. The essential thought behind this is to join numerous choice trees in deciding the last yield as opposed to depending on singular choice trees.

## 2. Literature survey

### 2.1 Existing problem

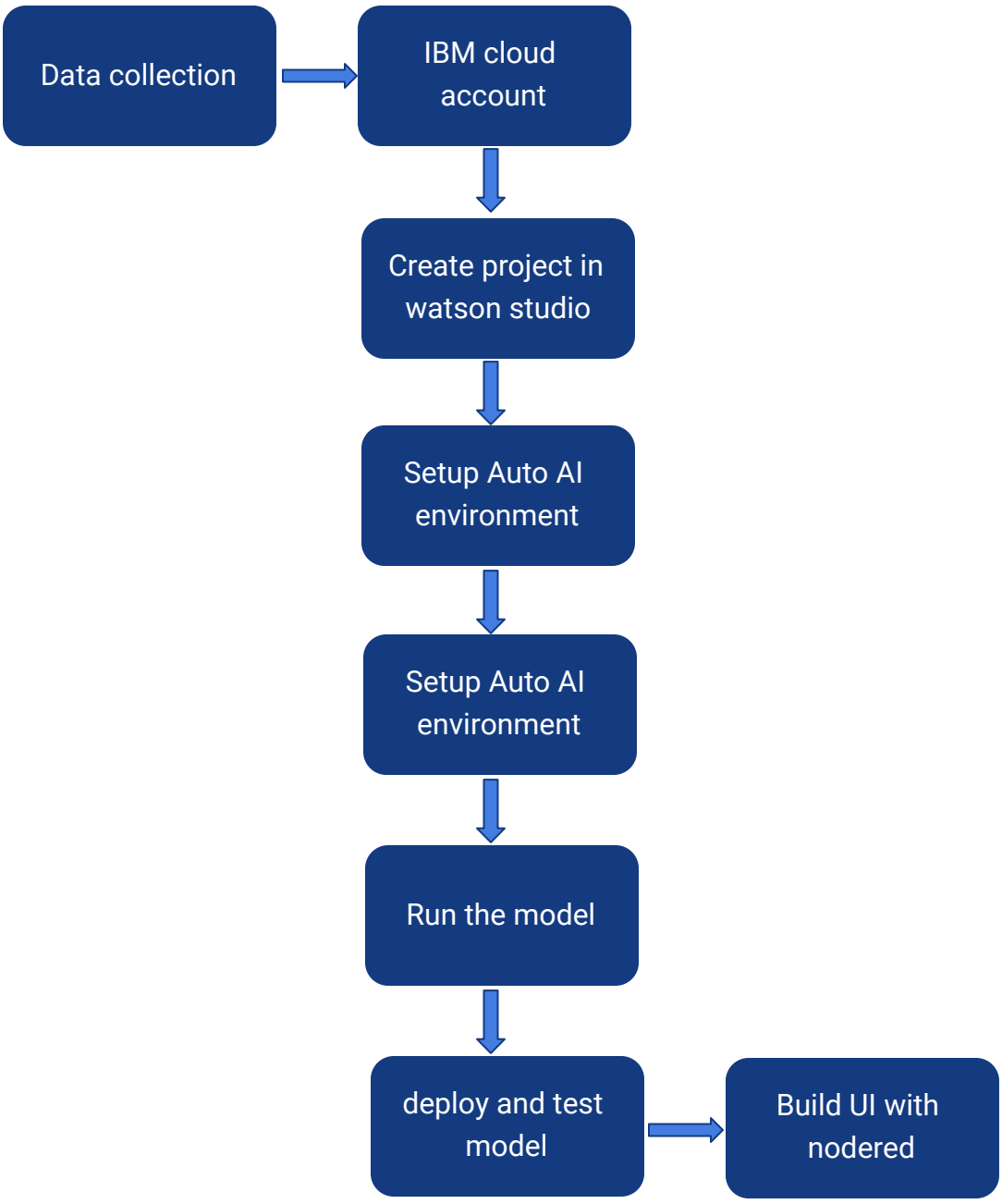
The premium we pay for a health insurance policy depends on various factors like age, gender, lifestyle, medical history etc. And health insurance companies have a tough task at determining premiums for their customers. While the health care law in any country does have some rules for companies to follow to determine premiums, it's really up to the companies on what factor/s they want to hold more weightage. Companies should know the most important factors and how much statistical importance do they hold, which is actually a hefty task.

### 2.2 Proposed solution

While several studies have been conducted in the past about factors influencing health insurance premiums, taking into account several factors like gender, age, region, number of children etc. The effective model to predict accurate premiums was found to have not been existed in the past. Some of the past research has also been carried out considering multiple linear regression for all countries, based on the data set. This research will also concentrate on medical history factors, mortality factors, cultural, social and other health-related factors. Since the results in this dataset are focused on a wide range of data, it will be easier for a company to evaluate the predictive factor that affects to health insurance premium value.

# 3. Theoritical analysis

## 3.1 Block diagram



## 3.2 Project requirements

As we predict the premium for health insurance and build this project, there are several requirements which we need to fulfill.

**Functional requirements :**

- Download health insurance cost prediction dataset (Kaggle)
- Data wrangling
- Create IBM cloud account
- Create appropriate cloud and nodered service
- Train the model on different algorithms
- Select the best model
- Build node red flow for GUI
- Create scoring end point for integrating our model to node red
- Provide the model with input fields
- Get the output as the predicted insurance price

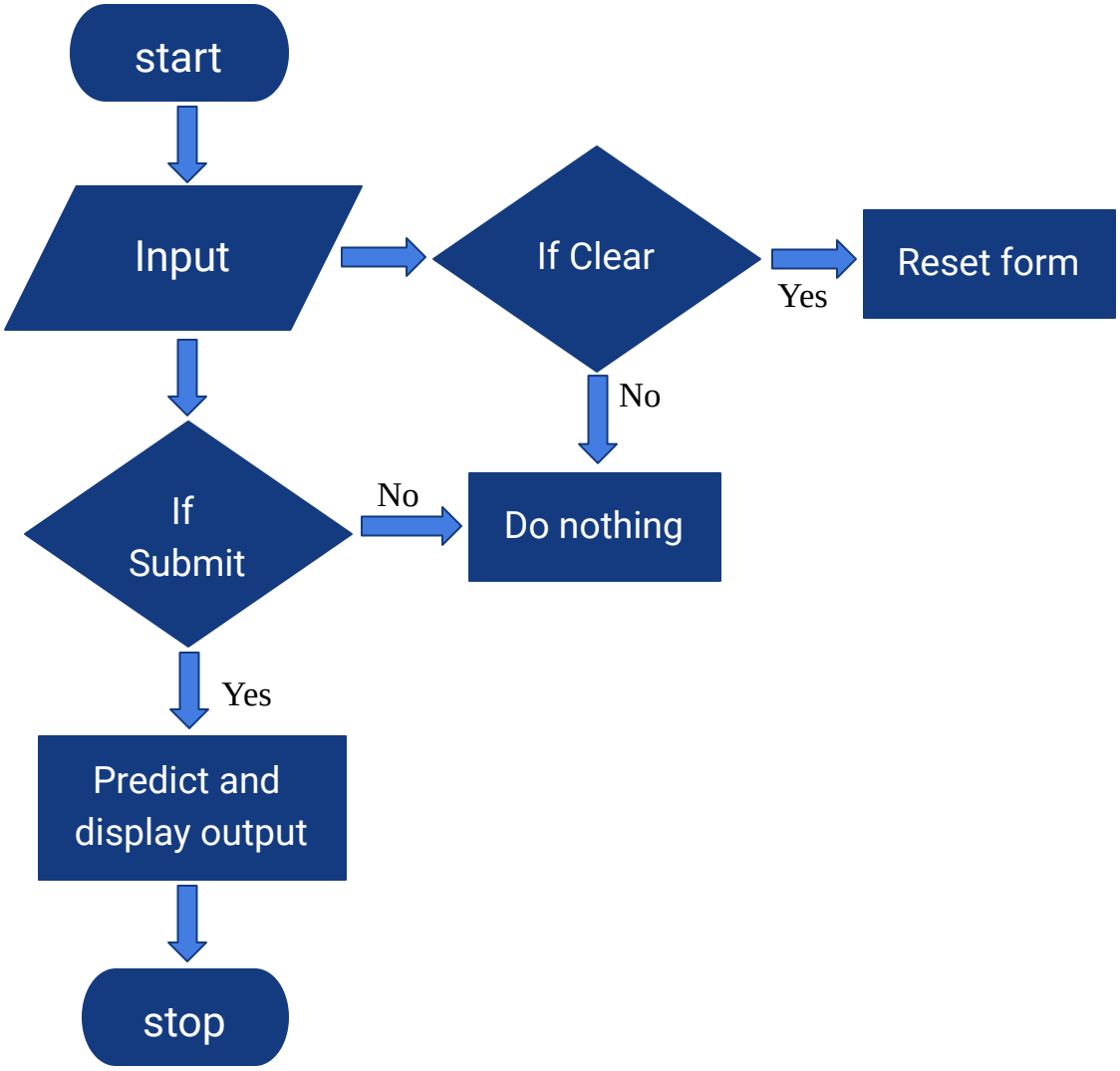
**Technical requirements :**

- A clean dataset
- Integration of GUI and backend trained model

**Software requirements :**

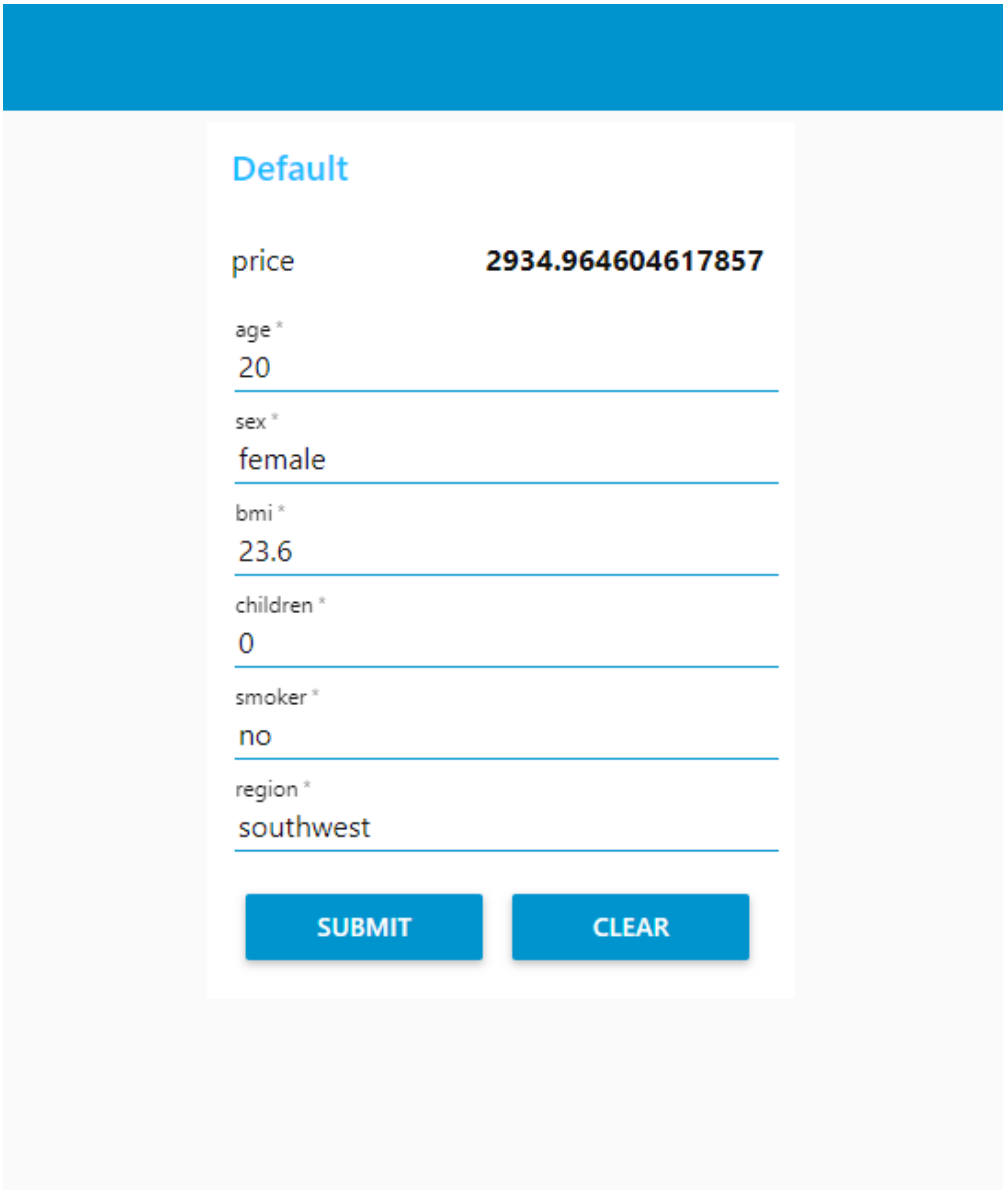
- IBM cloud services
- IBM Watson Studio
- IBM Auto AI experiment
- IBM Node-Red application
- SmartInternz project workspace
- Jupyter Notebook
- Github
- Slack
- Zoho document writer

## 4. Flow chart


start
Input
If Clear
Yes
Reset form
No
If
Submit
No
Do nothing
Yes
Predict and
display output
stop

## 5. Result

The user friendly Graphical User interface is shown in Figure below. This GUI is connected to the trained machine learning model present in the backend (IBM Watson notebook). The company has to fill in the inputs accordingly and click on the “Submit" button present at the end of the form. On clicking the “Submit” button, the predicted premium price will be displayed at the price label on top, based on the inputs provided as shown in Figure.



## 6. Advantages and Disadvantages

### Advantages

- Easy for users to interact with the model via the UI.
- User-friendly.
- IBM helps to process the data easily and improves the performance.
- Easy to build and deploy.
- Doesn’t require much storage space.
- Data can be analyzed.
- Factors affecting premium can be analyzed.

### Disadvantages

- Error in data can result in wrong prediction
- Accuracy is not 100%
- Error may occur due to inappropriate analysis of data

## 7. Applications

There are over 20 general insurance companies in India which offer health insurance policies. Each of these companies offers more than one type of health insurance policies in India. In other words, there are many types of health insurance policies available in the market – family health insurance, individual health insurance, critical illness insurance, group health insurance, etc. This project can be useful to all of these companies. Changing lifestyle habits, increase in pollution levels, and many other factors have a severe impact on an individual’s health. This may cause various health conditions and medical diseases.. This project/idea is useful for Insurance companies as they consider age, lifestyle choices, family medical history, and several other

factors when determining premium rates for individual life insurance policies. Not only companies, but customers will be benefitted by knowing anaccurate premium cost. It can be used by researchers to make meaningful research out of it and thus, bring something that will help recognizing right factors which impact premiums.

## 8. Conclusions

Thus, we have developed a model that will predict the health insurance premium price of a specific individual based on the inputs provided. Various factors have a significant impact on the premium such as age, bmi, gender, region and many more. Companies can interact with the system via a simple Graphical user interface which is in the form of a form with input spaces which the user needs to fill the inputs into and then press the “submit” button.

## 9. Future Scope

As future scope, we can connect the model to the database which can predict the price of not only health insurance but also of various other types of existing insurances. This will help us analyze how other factors affect different types of insurances.

## 10. References and links


- Dataset : https://www.kaggle.com/annetxu/health-insurance-cost-prediction
- Web app url : https://node-red-nchao.eu-gb.mybluemix.net/ui/#!/0?socketid=KWzVLEiF7gUq6FALAAAF
- Demostration link : https://youtu.be/gHvSr1E04vQ
- https://youtu.be/apFbFikesjA
- https://developer.ibm.com/tutorials/
- https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/