

High Level Design (HLD)

Insurance Premium Prediction

Revision 1.0.1

Date: 16/09/2023

Description: Initial revision HLD

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Revision | Description | Author |
| 16/09/2023 | 1.0.1 | Initial revision (1.0.1) | Megha Singhal |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

1. Abstract 4
2. Introduction 5
   1. Why this High Level-Design Document? 5
   2. Scope 5
   3. Definitions 5
3. General Description 6
   1. Product perspective 6
   2. Problem statement 6
   3. Proposed solution 6
   4. Technical requirements 6
   5. Data requirements 7
   6. Tools and technologies used 8
   7. Constraints 8
4. Design Details 9
   1. Process flow 9
   2. Deployment process 10
   3. Event log 10
   4. Error handling 10
   5. Performance 10
   6. Reusability 11
   7. Application compatibility 11
   8. Recourse utilization 11
   9. Deployment 11
   10. User interface 12
5. Conclusion 13
6. References 13

# 1. Abstract

A health insurance policy is a policy that covers or minimises the expenses of losses caused by a variety of hazards. A variety of factors influence the cost of insurance or healthcare. Accurate cost estimates can help health insurers and, increasingly, healthcare delivery organisations to plan for the future and prioritise the allocation of limited care management resources. Furthermore, knowing ahead of time what their probable expenses for the future can assist patients to choose insurance plans with appropriate deductibles and premiums. These elements play a role in the development of insurance policies.

In this study, we have used the USA's medical cost personal dataset from Kaggle, having 1338 entries. Features in the dataset that are used for the prediction of insurance cost include Age, Gender, BMI, Smoking Habit, number of children etc. I used supervised ML models to demonstrate how various regression models, including Linear Regression (LR), Stochastic Gradient Boosting (SGB), Decision Tree Regressor (DTR) and Random Forest Regressor (RFR) can forecast insurance cost.

# 2. Introduction

**2.1 Why this High-Level Design Document?**

The purpose of this High-level document (HLD) is to describe the design of the project in detail which can be used as a reference manual.

The HLD will:

* Present all the design aspects and define them in detail.
* Describe the user interface being implemented.
* Describe the software interfaces.
* Describe the performance requirements.
* Include design features and the architecture of the project.

## 2.2 Scope

The HLD document present the entire structure of the project in parts, such as the data ingestion, data pre-processing, solution development and the deployment part along with their respective architectures. It uses non- technical to mild technical terms which should be understandable to the administrators of the system.

## 2.3 Definitions

|  |  |
| --- | --- |
| **Term** | **Description** |
| **EDA** | Exploratory Data Analysis |
| **Database** | Collection of all the information |
| **IDE** | Integrated Development Environment |
| **AWS** | Amazon web services |
| **API** | Application programming interface |
| **Vs code** | Visual studio code |
| **KPI** | Key performance indicator |
| **PaaS** | Platform as a Service |
|  |  |

## 3. General Description

### 3.1 Product Perspective

The insurance premium predictor is a machine learning based regression model which help us to predict the insurance premium based on variables, which will be used to determine prediction and validation of plans offer by different companies.

### 3.2 Problem Statement

The goal of this project is to give people an estimate of how much they need based on their individual health situation. After that, customers can work with any health insurance carrier and its plans and perks while keeping the projected cost from our study in mind. This can assist a person in concentrating on the health side of an insurance policy rather than the ineffective part.

### 3.3 Proposed Solution

The solution proposed here is a web application, which takes details of the main variables of insurance which contributes to its premium and those details will be taken by a machine learning model in the backend, which predicts the premium in dollars and displays in the front-end page to the user.

### 3.4 Technical Requirements

The solution can be a cloud-based or application hosted on an internal server or even be hosted on a local machine. For accessing this application below are the minimum requirements:

* Good internet connection.
* Web Browser.

For training model, the system requirements are as follows:

* +4 GB RAM preferred
* Operation System: Windows, Linux, Mac
* Visual Studio Code / Jupyter notebook

### 3.5 Data Requirements

Data requirements completely depends on out problem statement.

The dataset used for the project is downloaded from the below mentioned link:

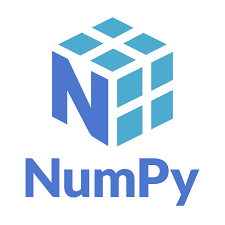
<https://drive.google.com/uc?id=1zZxmaFh07scM6kzJB9PUSGx4JCIfNaTr&export=download>

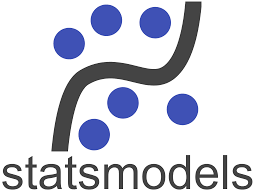
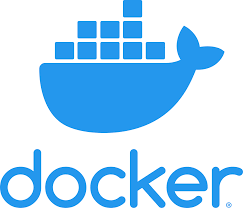
The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, bmi, children and expenses) and 3 nominal features (sex, smoker and region) that were converted into factors with numerical value designated for each level.

Data dictionary as follows:

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Description |
| Age | Integer | Input variable |
| Sex | String | Input variable |
| BMI | Decimal | Input variable |
| Children | Integer | Input variable |
| Smoker | String | Input variable |
| Region | String | Input variable |
| Expenses | Decimal | Output variable |

### 3.6 Tools and Technologies used

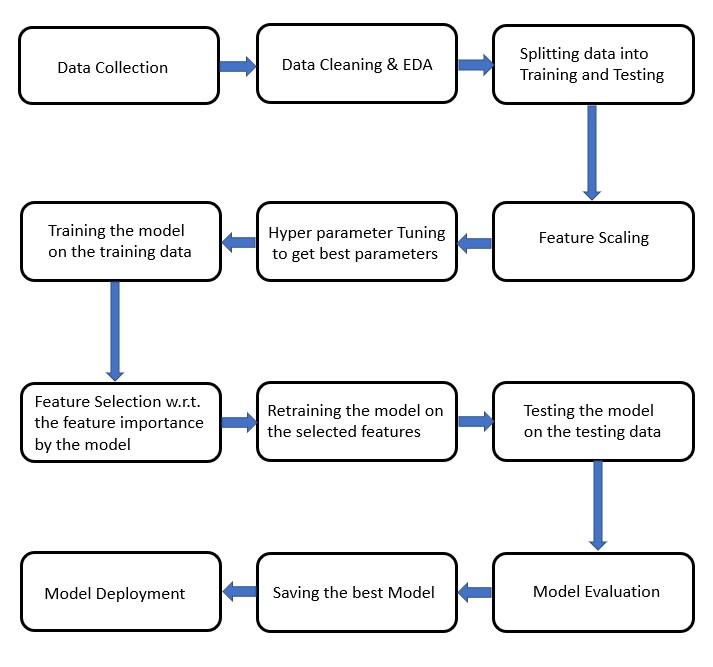
* Jupyter notebook is used for EDA and feature engineering part.
* Visual Studio is used as an IDE.
* For Visualisation of plots, ‘matplotlib’ and ‘seaborn’ python packages are used.
* Numerical computations and statistical analysis are done through numpy, pandas and statsmodels.
* GitHub is used as a version control system.
* Deployed on web using Gunicorn and Heroku.
* Python flask is used for backend development.
* Frontend development is done using html/css.

### 3.7 Constraints

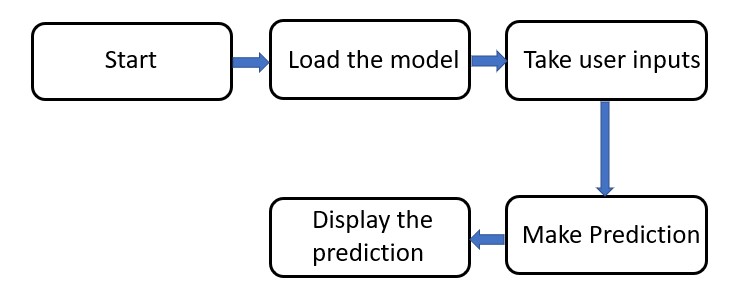
Prediction system must be user friendly, errors free and users should not be required to know about any of the workings.

# 4. Design Details Data Collection

4.1 Process Flow



## 4.2 Deployment Process



## 4.3 Event log

## In this project, I used the “logging” library in both the development and deployment stages, which keeps logging the events at every step into the “.log” files. One of the advantages of event logging is, it makes debugging much easier, like we can directly go to that specific line of code, having errors.

## 4.4 Error handling

## Used exceptions handling to catch the errors, so that they will be recorded in logs and ensures the smooth run, without getting terminated in the middle. Once the run gets completed, we can check the log files for the errors and can take an appropriate debugging action.

## 4.5 Performance

## The ML based Insurance Premium prediction application is used for predicting the Insurance premium based on its age and other factors. So, it should be as accurate as possible, so that it will not mislead the user. Also, the model retraining is very important to keep it relevant if the new factors are added in future or to improve the performance.

## 4.6 Reusability

The code written and the components used have an ability to be reused without any problem.

## 4.7 Application compatibility

The different components or modules of this project use python version 3.7 as their interface between them. Each component has its own task to perform, and it is the job of the python version to ensure proper transfer of the information.

## 4.8 Resource utilization

In this project, any task may likely to use all the processing power available in the system, until it is accomplished.

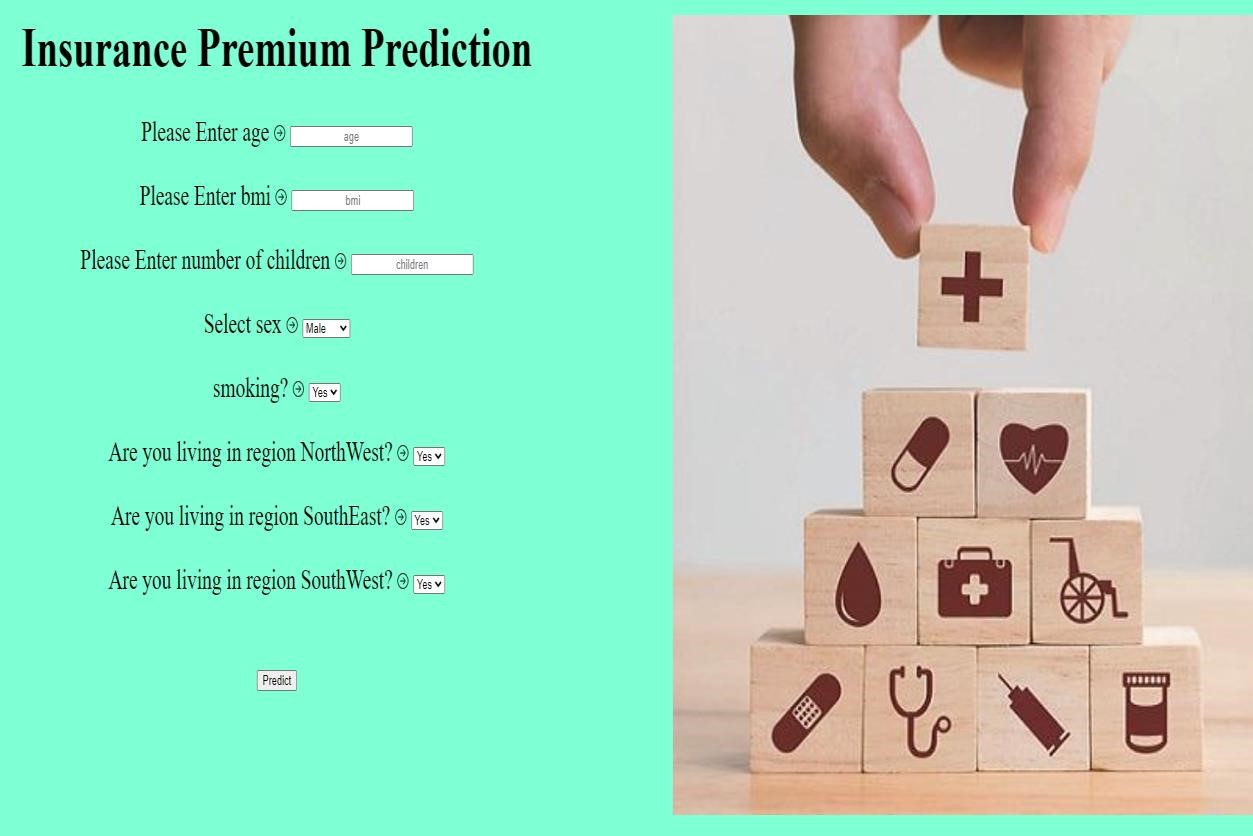
## 4.9 Deployment

Deployment of the application on web using Heroku (PaaS) with Gunicorn (a python WSGI HTTP server) version 20.1.0.

URL:<https://dashboard.heroku.com/apps/insurance-prem-prediction12>

## 4.10 User Interface

Designed user interface using HTML with CSS styling. It looks as per the below image.



# 5. Conclusion

Insurance premium prediction is used to predict the premium of the personal for health insurance based on the given input entities, which enables the user to determine cost and save money on effective plans.

# 6. References

* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
* <https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b>
* <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>
* <https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction>
* <https://www.researchgate.net/publication/361579419_Machine_Learning-Based_Regression_Framework_to_Predict_Health_Insurance_Premiums>
* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>
* <https://matplotlib.org/>
* <https://numpy.org/doc/>
* <https://www.hindawi.com/journals/mpe/2021/1162553/>