Low Level Design (LLD)

Insurance Premium Predictions

Revision Number: 1.1

Last date of revision: 12/08/2021

Shrikant Shejwal

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 12th Aug 2021 | 1.1 | Initial HLD-V1.0 | Shrikant V Shejwal |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Contents**

[Document Version Control 2](#_gjdgxs)

[**Abstract** 4](#_30j0zll)

[1](#_1fob9te) Introduction 5

[1.1](#_3znysh7) Why this Low-Level Design Document? 5

[1.2](#_2et92p0) Scope 6

[1.3](#_tyjcwt) Constraints 6

[1.4](#_3dy6vkm) 1.3 Risks 6

[1.5](#_1t3h5sf) 1.4 Out of Scope 7

[2](#_4d34og8) Technical specifications 7

[2.1 Dataset 7](#_2s8eyo1)

[2.2 Logging 7](#_17dp8vu)

[2.3 Database 7](#_3rdcrjn)

[3](#_26in1rg) Technology stack 8

[4](#_lnxbz9) Proposed Solution 9

[5](#_35nkun2) Model training/validation workflow 10

[6](#_1ksv4uv) User I/O workflow 12

[7](#_44sinio) Exceptional scenarios 13

**Abstract**

The main foundational block of health insurance industry is to estimate the future events and measure the associated risk/value of these events, hence it is needless to say that predictive analytics is used widely to determine the risk, insurance premium and enrich overall customer experience. The health insurance industry has always been a slow-moving industry when it comes to adopting the data analytics practices into its business models. With the advent of advanced data analytics technologies, it has become important more than ever to take advantage of such sophisticated analytics to accurately assess and predict the insurance premiums for the insured. Thus, one of the important tasks for health insurance companies is to determine the policy premiums. By using predictive modelling, the insurers can determine the policy premium for the insured based on their behaviors which are indicated by attributes such as age, BMI (Body Mass Index), smoking habits, number of children etcetera. This determination of premiums based on the data collected for an individual helps insurance companies in enhanced pricing, underwriting and risk selection. Additionally, it helps in making better decisions, understanding customer needs and be fair to the customers. Acquiring a comprehensive understanding of customer behaviors and habits from historical data helps insurers to anticipate future behaviors and provide the right insurance product and policy premium.

# Introduction

## Why this Low-Level Design Document?

The purpose of this document is to present a detailed description of the Insurance Premium Prediction. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for both the stakeholders and the developers of the system and will be proposed to the higher management for its approval.

The main objective of the project is to predict the Health Insurance Prediction. The goal of this project is to allows a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part. The main goal of the project is to predict the insurance premium charge based upon other attributes.

An Insurance Premium Prediction contains the information, such as:

✓ age

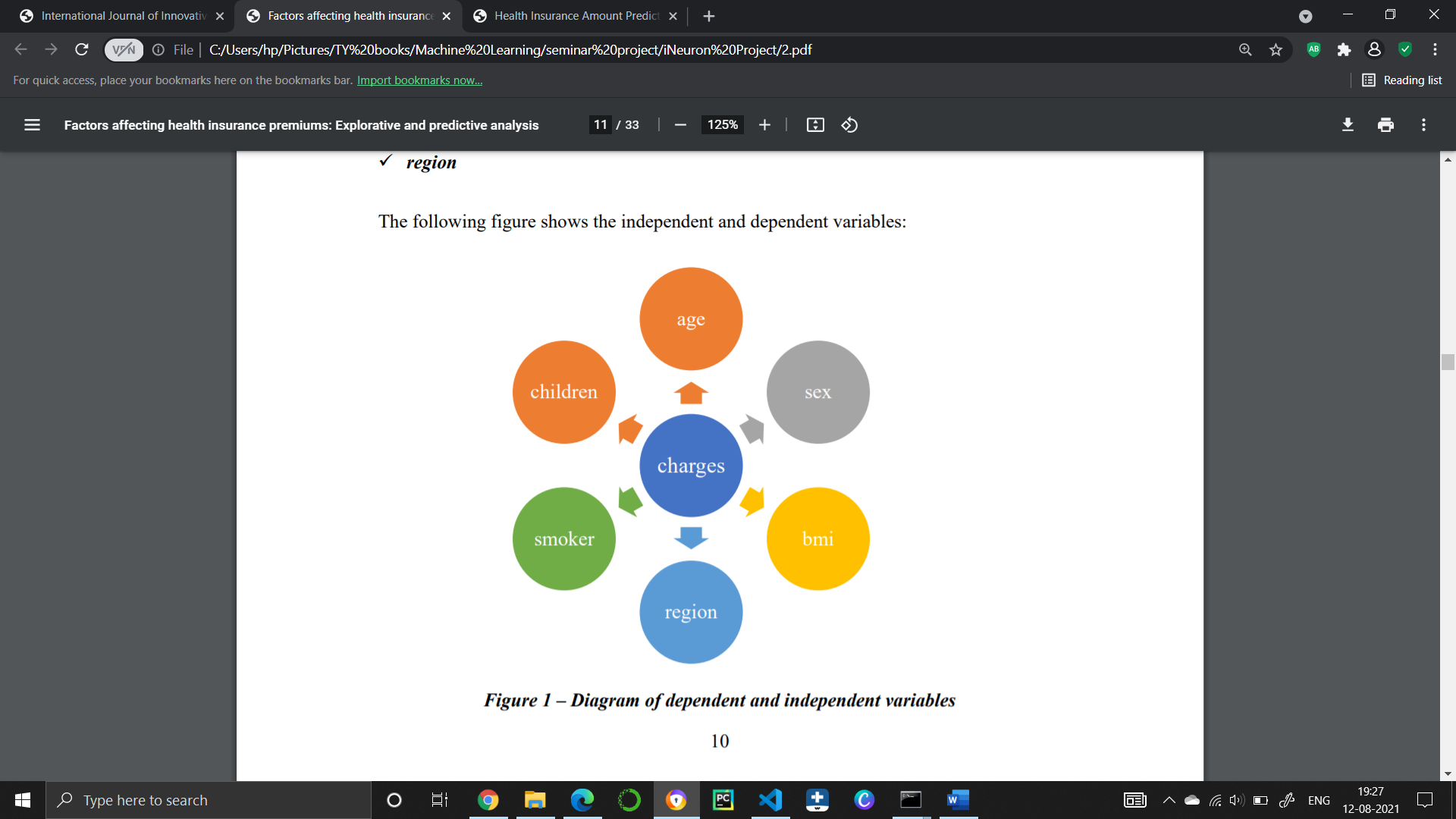
✓ sex

✓ children

✓ bmi

✓ smoker

✓ region



This project shall be delivered in two phases:

Phase 1: All the functionalities with Scikit-learn packages.

Phase2: Integration of UI to all the functionalities.

## Scope

This software system will be a Web application This system will be designed to predict health insurance premium. Premium amount prediction focuses on person’s own health rather than other company’s insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount. This system is designed to predict the health insurance premium based on some information like age, sex, BMI, region, smoker, children etc.

## Constraints

The Insurance Premium Prediction must be user friendly, as automated as possible and users should not be required to know any of the workings.

## Risks

Document specific risks that have been identified or that should be considered.

## Out of Scope

Delineate specific activities, capabilities, and items that are out of scope for the project.

# Technical specifications

## 2.1 Dataset

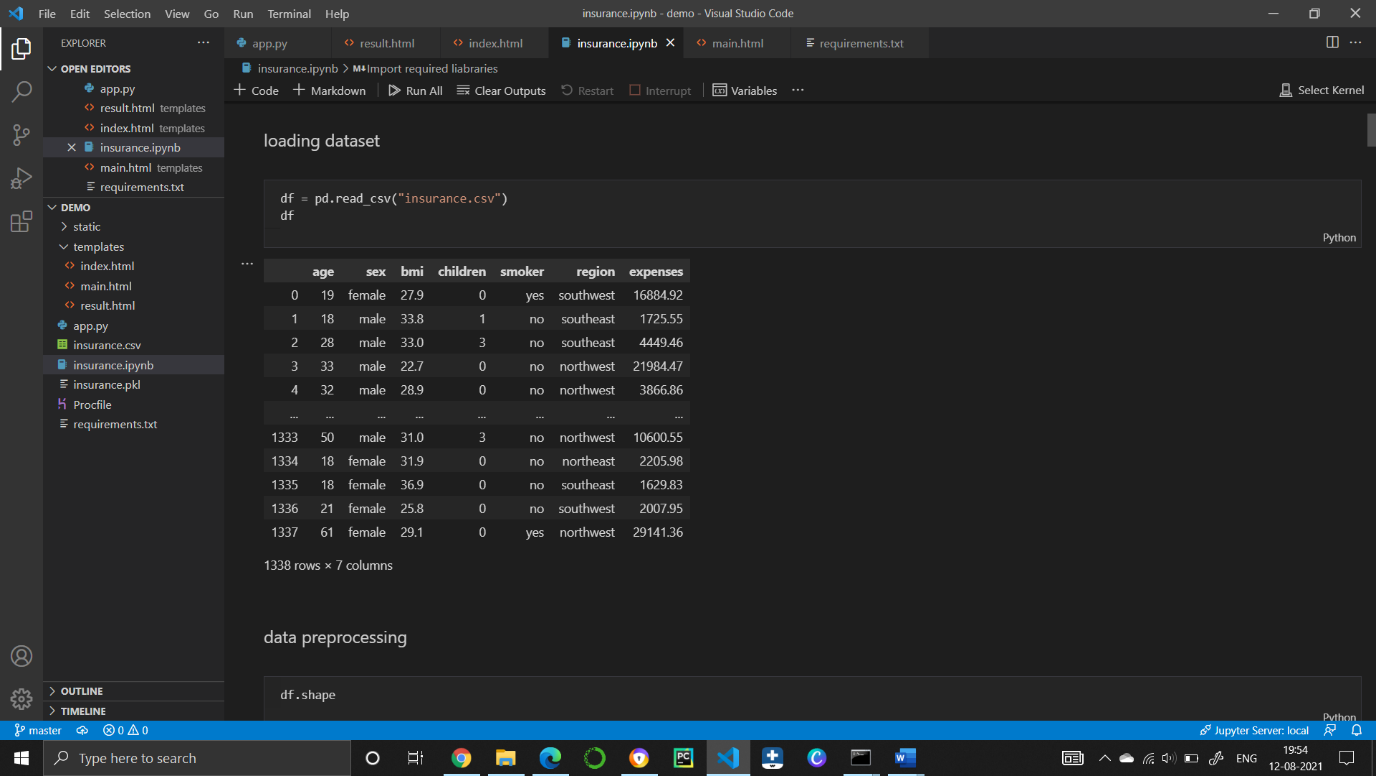
|  |  |
| --- | --- |
| Name | Description |
| Age | Age of the client |
| BMI | Body mass index |
| The number of kids | Number of children the client have |
| Gender | Male / Female |
| Smoker | Weather the client is smoker or not |
| Region | Where the client live southwest, southeast, northwest, northeast. |

## Insurance Premium dataset overview

To create the claim cost model predictor, we obtained the data set through the Kaggle site. The data set includes seven attributes, the data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model. the following table shows the Description of the Dataset.

|  |  |
| --- | --- |
| Kids | Have |
| 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 |
| Expenses (Target Variable) | Medical Cost the client pay |

Some of the records in the dataset are following



## 2.1.2 Input schema

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature name** | **Datatype** | **Size** | **Null/Required** |
| Age | int | 2 | Required |
|  |  |  |  |
|  |  |  |  |

## 2.2 Predicting Disease

* The system displays the form where the all features are available if user can all features correctly then machine can able to predict the health insurance premium.

## 2.3 Logging

We should be able to log every activity done by the user.

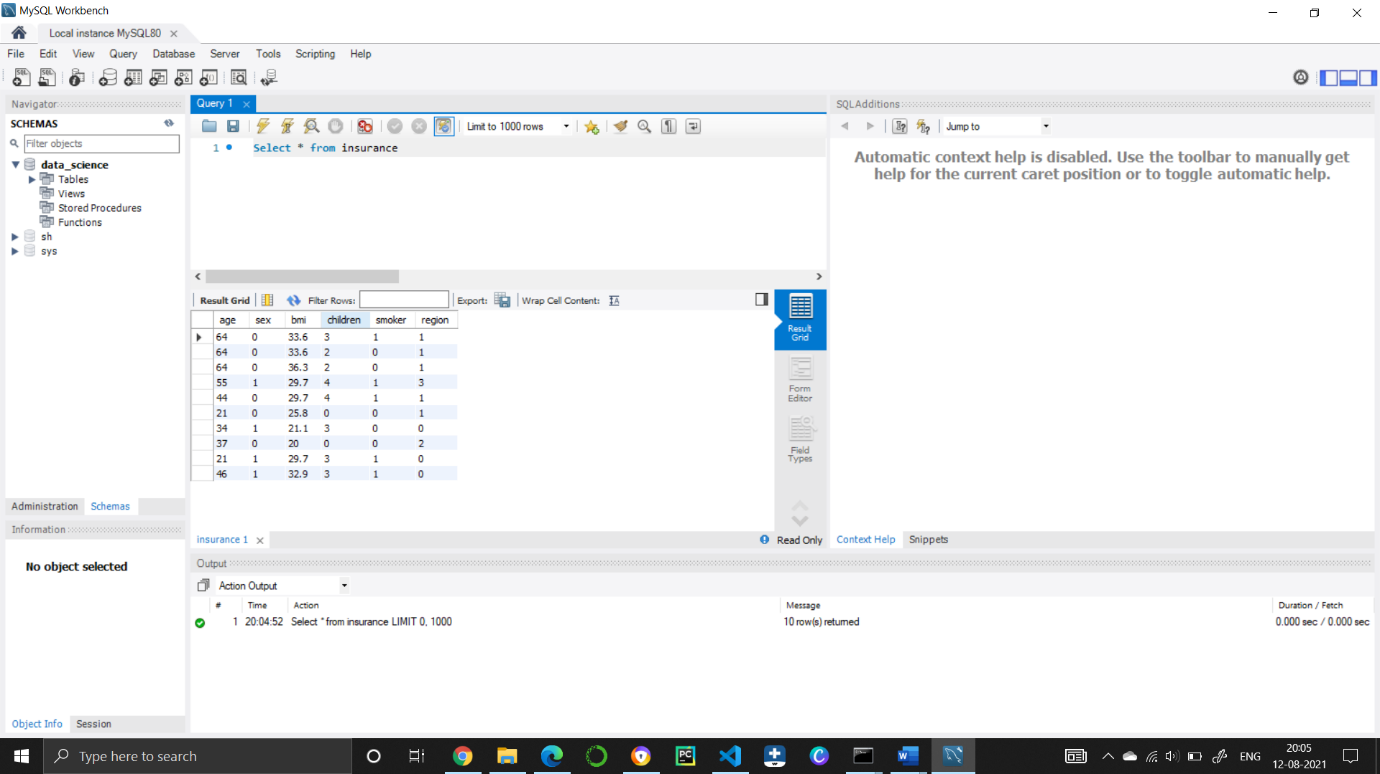
* The System identifies at what step logging required
* The System should be able to log each and every system flow.
* Developers can choose logging methods. You can choose database logging/ File logging as well.
* System should not be hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

## 2.4 Database

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

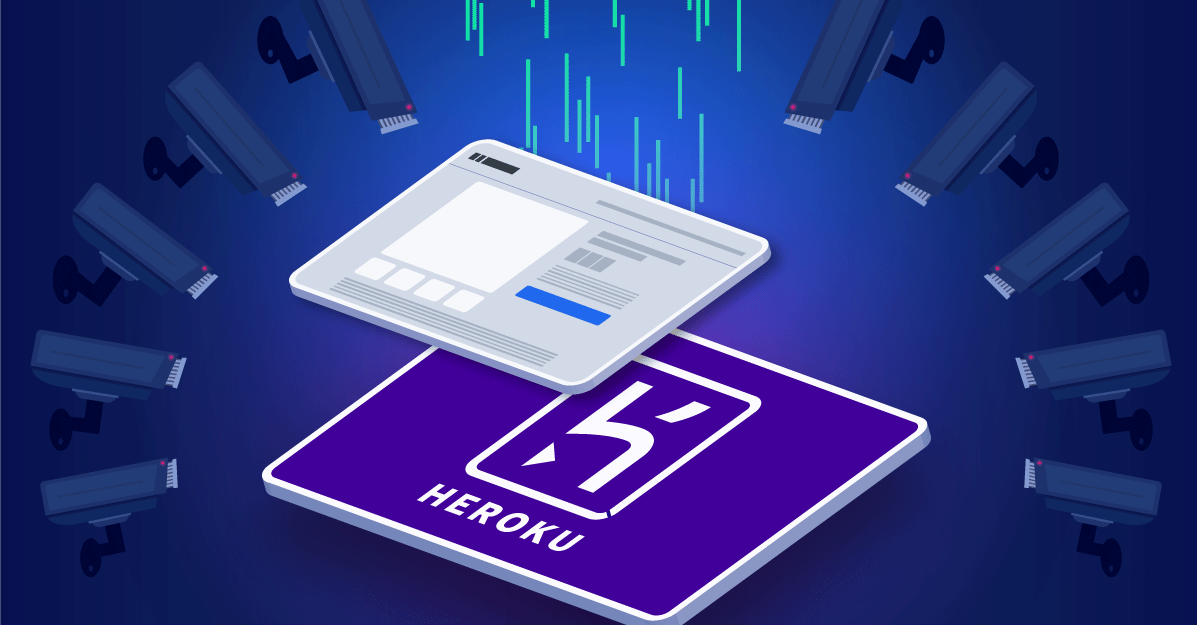
1. The User gives required information.

2. The system stores each and every data given by the user or received on request to the database. Database you can choose your own choice whether MongoDB/ MySQL. Here we use MySQL.



**2.5 Deployment**

1. FLASK



# Technology stack

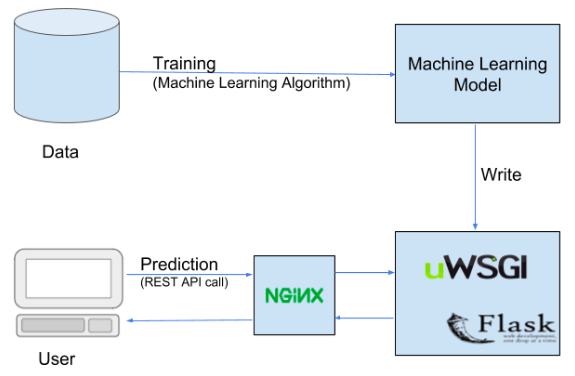
|  |  |
| --- | --- |
| **Front End** | HTML/CSS/JS/React |
| **Backend** | Python Flask |
| **Database** | MySql |
| **Deployment** | Heroku |

# Proposed Solution

Based on the actual research paper, gradient boosting gives better accuracy as compare to other so in this project we use gradient boosting regression algorithm to predict insurance. However, drawing a baseline in the form of some Machine Learning algorithm would be helpful. Why making a baseline model important? Well, to compare the performance of our actual model, let say Gradient Boosting in this case, is very important to ascertain that we are in the right direction as if performance of gradient boosting is not better than the baseline model then there is no point of using gradient boosting.

1. Actual model: Gradient Boosting.

# Model training/validation workflow



# User I/O workflow

Start

Prediction

# Exceptional scenarios

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Exception | Mitigation | Module |
| 12th Aug 2020 | 1.1 | First Draft | Shrikant V Shejwal |
|  |  |  |  |

# Test cases

|  |  |  |  |
| --- | --- | --- | --- |
| Test case | Steps to perform test case | Module | Pass/Fail |
|  |  |  |  |