HIGH LEVEL DESIGN (HLD)

Insurance Premium Prediction

Contributor

Sahil Patel

Taif Qureshi

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| **0.3.03.2023** | V1.0 | Initial HLD- V1.0 | Sahil Patel |

Contents

[Document Version Control 1](#_Toc10259)

[Abstract 3](#_Toc10260)

[**1.0 Introduction** 4](#_Toc10261)

[1.1 Why this High-Level Design Document? 4](#_Toc10262)

[1.2 Scope 4](#_Toc10263)

[1.3 Definitions 5](#_Toc10264)

[**2.0 General Description** 6](#_Toc10265)

[2.1 Product Perspective 6](#_Toc10266)

[2.2 Problem Statement 6](#_Toc10267)

[2.3 Proposed Solution 6](#_Toc10268)

[2.4 Technical Requirements 7](#_Toc10270)

[2.5 Data Requirements 7](#_Toc10271)

[2.6 Tools Used 8](#_Toc10272)

[2.7 Assumptions 8](#_Toc10274)

**[3.0 Design Details](#_Toc10275)** [9](#_Toc10275)

[3.1 Process Flow 9](#_Toc10276)

[3.2 Event Log 9](#_Toc10277)

[**4.0 Performance** 10](#_Toc10278)

[4.1 Reusability 10](#_Toc10279)

[4.2 Application Compatibility 10](#_Toc10280)

[4.3 Deployment 10](#_Toc10281)

[**5.0 Conclusion** 11](#_Toc10283)

# Abstract

We analyse the personal health data to predict insurance premium of individuals. Seven regression models naming Linear Regression,

Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, KNN have been used to compare the performance of these algorithms.

Training dataset was used for training model and that training model helped to come up with some predictions. Then the predicted amount was compared with actual data to test and verify the model accuracy. Later accuracies of all these models were compared. It was gathered that Gradient Boosting and Random Forest algorithms performed better than the remaining models.

Gradient boosting is best suited in this case because it gives best evaluation score comparable to other models.

# 

# 1.0 Introduction

## 1.1 Why this High-Level Design Document?

The purpose of this High-Level document is to add necessary details to current project description to represent a suitable model for coding. This document is used as a reference manual for how the model interact at a high-level.

### The HLD will

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

## 1.2 Scope

The HLD document presents the structure of the system, such as the database architecture, application architecture, and technology architecture. The HLD uses non-technical to middle-technical terms which should be understandable to the administrators of the system.

## 1.3 Definitions

|  |  |
| --- | --- |
| **Term** | **Description** |
| Database | Collection of all the information |
| IDE | Integrated Development Environment |
| VS Code | Visual Studio Code |
| EDA | Exploratory Data Analysis |
| KNN | K-Nearest Neighbour’s |

# 2.0 General Description

## 2.1 Product Perspective

The Insurance premium estimation is a machine learning based predictive model which will help us to predict the premium of the personal for health insurance.

## 2.2 Problem Statement

To develop an API interface to predict the premium of insurance using people individual health data and analysing the following:

* To detect BMI value affects the premium.
* To detect smoking affects the premium of the insurance.
* To create API interface to predict the premium

## 2.3 Proposed Solution

The solution proposed here is an estimating premium of insurance based on people health data and this can be implemented to perform above mention use cases. In first case, analysing how BMI value affect the people health as well as premium of the insurance. In the second case, if model detects the smoking affecting the premium, we will inform that to people. And in the last use case, we will be making an interface to predict the premium.

## 2.4 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 / Jupiter notebook

## 2.5 Data Requirements

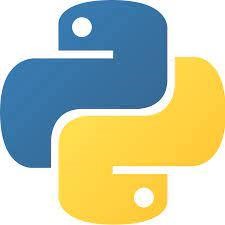
Data requirements completely depends on out problem statement.

* Comma separated values (CSV) file.
* Input file feature/field names and its sequence should be followed as per decided.

## 2.6 Tools Used

Python programming language and frameworks such as NumPy,

Pandas, Scikit-learn, Flask are used to build the whole model.



* Pandas is an open-source Python package that is widely used for data analysis and machine learning tasks.
* NumPy is most commonly used package for scientific computing in Python.
* Scikit-learn is used for a machine learning.
* Flask is used to build API.
* VS Code is used as IDE (Integrated Development Environment)
* GitHub is used as version control system.
* Front end development is done using HTML/CSS.
* Heroku is used for deployment of the model.

## 2.7 Assumptions

The main objective of the project is to develop an API to predict the premium for people based on their health information. Machine learning based regression model is used for predicting above mentioned cases on the input data.

# 3.0 Design Details

## 3.1 Process Flow

Start

Data

Collection

EDA

Data

Cleaning

Feature

Engineering

Model

Building

Model

Testing

Flask Setup

Deployment

## 3.2 Event Log

The system should log every event so that the user will know what process is running internally.

**Initial Step-By-Step Description:**

* The system identifies at what step logging required.
* The system should be able to log each system flow.
* Developer can choose logging method. You can choose database logging.

System should not hang out even after using so many loggings.

# 4.0 Performance

## 4.1 Reusability

The entire solution will be done in modular fashion and will be API oriented. So, in the case of the scaling the application, the components are completely reusable.

## 4.2 Application Compatibility

The interaction with the application is done through the designed user interface, which the end user can access through any web browser.

## 4.3 Deployment



# 5.0 Conclusion

This system shows us that the different techniques that are used to estimate the how much amount of premium required based on individual health situation. After analysing it shows how a smoker and non-smokers affecting the amount of estimate. Also, significant difference between male and female expenses. Accuracy, which plays a key role in prediction-based system. From the results we could see that Gradient Boosting turned out to be best working model for this problem in terms of the accuracy. Our predictions help user to know how much amount premium they need based on their current health situation.