HIGH LEVEL DESIGN (HLD)

Insurance Premium Prediction

By- Asouk

Date-10.09.23

# 

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 10.09.2023 | 1 | HLD | Asouk |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[Document Version Control 1](#_Toc11178)

[Abstract 3](#_Toc11179)

[1.0 Introduction 4](#_Toc11180)

[1.1 Why this High-Level Design Document? 4](#_Toc11181)

[1.2 Scope 4](#_Toc11182)

[1.3 Definitions 5](#_Toc11183)

[2.0 General Description 6](#_Toc11184)

[2.1 Product Perspective 6](#_Toc11185)

[2.2 Problem Statement 6](#_Toc11186)

[2.3 Proposed Solution 6](#_Toc11187)

[2.4 Further Improvements 7](#_Toc11188)

[2.5 Technical Requirements 7](#_Toc11189)

[2.6 Data Requirements 7](#_Toc11190)

[2.7 Tools Used 8](#_Toc11191)

[2.8 Constraints 9](#_Toc11192)

[2.9 Assumptions 9](#_Toc11193)

[3.0 Design Details 10](#_Toc11194)

[3.1 Process Flow 10](#_Toc11195)

[3.2 Event Log 10](#_Toc11196)

[4.0 Performance 11](#_Toc11197)

[4.1 Reusability 11](#_Toc11198)

[4.2 Application Compatibility 11](#_Toc11199)

[4.3 Deployment 11](#_Toc11200)

[5.0 Dashboards 12](#_Toc11201)

[6.0 Conclusion 13](#_Toc11202)

# Abstract

In our Project, we used personal health data to make predictions about insurance premiums for individuals. We didn't limit ourselves to just one method; instead, we explored the potential of several regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression.

To train these models, we made use of a dedicated training dataset. Once trained, these models were put to the test in making predictions. To gauge how well they performed, we compared their predictions against the actual insurance premium data. This involved crunching the numbers with metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Among the lineup of models we experimented with, both Gradient Boosting and Random Forest algorithms consistently stole the show by delivering the most accurate predictions. Among them, Gradient Boosting stood out as the star performer, consistently racking up the highest evaluation scores compared to its counterparts.

So, based on our rigorous examination, we've concluded that Gradient Boosting is the go-to algorithm for this specific task of predicting insurance premiums. Its remarkable evaluation scores and unwavering performance have earned it the top spot among the models we assessed.

This analysis underscores the significance of casting a wide net when exploring machine learning algorithms and meticulously evaluating their performance metrics to pinpoint the ideal model for a given predictive task. Additionally, it drives home the point that continuous monitoring and updating of the model are crucial as new data streams in, ensuring that the predictions remain accurate and relevant over time.

# 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 information |
| IDE | Integrated Development Environment |
| API | Application Programming Interface |
| KPI | Key Performance Indicator |
| VS CODE | Visual Studio Code |
| EDA | Exploratory Data Analysis |
|  |  |
|  |  |

# 2.0 General Description

## 2.1 Product Perspective

We are using Machine Learning Models to Predict the insurance premium of individuals based on their Sex, BMI, age, and other variables

## 2.2 Problem Statement

To Develop an web Interface or an application to predict health insurance premium of people using Health data and habits of people.

## 2.3 Proposed Solution

The proposed solution involves estimating insurance premiums based on individuals' health data, with applications in three key use cases. Firstly, it can analyze the impact of Body Mass Index (BMI) on both health outcomes and insurance premiums. Secondly, it can detect and inform individuals about how smoking habits affect insurance premiums. Lastly, it enables the creation of a user-friendly interface for real-time premium predictions, allowing users to input their health information and receive personalized estimates, thus serving as a valuable resource for consumers and insurance providers alike.

## 2.5 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
* Azure Account
* Google collab

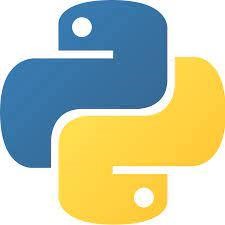
## Data Requirements

It depends on the problem we can download it from the Kaggle or source it from Databases, For this problem we import the dataset from kaggle

## 2.7 Tools Used

A blue and white logo

Description automatically generated



## "Pandas, NumPy, Plotly, Scikit-learn, Flask, VS Code, GitHub, HTML/CSS, Heroku, Azure"

## 2.9 Assumptions

We are assuming the persons are not affected with pre existing Diseases

## 3.1 Process Flow



Data Cleaning

EDA

Data Collection

Feature Engineering

Model Testing

Model Building

## 

Deployment

Azure Deploy

Flask

## 3.2 Event Log

The event logging happens in Azure Machine learning studio in the designated Environment

The system has the capability to identify when logging is necessary, ensuring that each step in the system's processes can be logged. Developers have the flexibility to select their preferred logging method, including the option to choose database logging for comprehensive record-keeping.

# 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 Dashboards

A dashboard is a data visualization and analysis tool that displays on one screen the status of key performance indicators (KPIs) and other important business metrics.



Dashboards serve as a high-level reporting mechanism, offering swift insights into critical business inquiries and enhancing decision-making in multiple ways:

* Dashboards effectively communicate how insurance premiums vary based on BMI values. This visual representation simplifies the understanding of how Body Mass Index influences premium rates.
* Dashboards also provide clear and intuitive visualizations that depict the relationship between gender and insurance premiums. This simplifies the process of understanding how gender factors into premium calculations.

# 6.0 Conclusion

This system illustrates various techniques employed to estimate insurance premiums based on individual health conditions. It delves into the impact of smoking habits and gender on premium estimations, highlighting the substantial differences. Accuracy is a crucial factor in this prediction system.

Upon analysis, Gradient Boosting emerged as the most accurate model for this task. The predictions generated by this system empower users to determine their required premium amount based on their current health status, providing valuable insights for informed decisions.