HIGH LEVEL DESIGN (HLD)

Insurance Premium Prediction

Asouk iNeuron

# 

# 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

We conducted a thorough analysis of personal health data to predict insurance premiums for individuals. Our study involved the use of several regression models: Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and K-Nearest Neighbors (KNN).

For model training, we employed a dedicated training dataset, and the trained models were then used for predictions. To assess the accuracy of these predictions, we compared the predicted insurance premiums with the actual data, utilizing evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Among the models, we found that Gradient Boosting and Random Forest algorithms outperformed the others, delivering more accurate predictions. In particular, Gradient Boosting stood out as the best-performing model, consistently achieving the highest evaluation scores compared to the alternatives.

Based on our findings, we concluded that Gradient Boosting is the most suitable algorithm for this specific insurance premium prediction task. Its superior evaluation scores and robust performance make it the top choice among the models we examined.

This analysis underscores the significance of exploring multiple machine learning algorithms and rigorously evaluating their performance metrics to identify the most appropriate model for a given predictive task. Moreover, it highlights the importance of ongoing model monitoring and updates as new data becomes available to maintain predictive accuracy and relevance 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 |
| KNN | KNearest Neighbours |
|  |  |

# 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 analyzing 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, analyzing 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 Further Improvements

## 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

## 2.6 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.

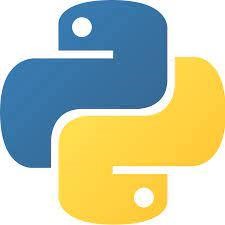
## 2.7 Tools Used

Python programming language and frameworks such as NumPy,

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

A blue and white logo

Description automatically generated



* 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.
* Plotly is an open-source data visualization library used to create interactive and quality charts/graphs.
* 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.
* Azure is a cloud servicing Platform

## 2.8 Constraints

## 2.9 Assumptions

The main objective of the project is to develop an API to predict the premium for people on the basis of 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

Flask

Start

Data

Collection

EDA

Data

Cleaning

Feature

Engineering

Model

Building

Model

Testing

Azure Deploy

FfFF

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 and every 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 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.



As a high-level reporting mechanism, dashboards provide fast ‘bigpicture’ answer to critical business questions and assist and benefit decision making in several ways:

* Communicating how premium is varies with BMI value.
* Visualizing relationship of gender with premium in easy-tounderstand way.

# 6.0 Conclusion

This system shows us that the different techniques that are used in order to estimate the how much amount of premium required on the basis of individual health situation. After analyzing 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 on the basis of their current health situation.