High Level Design (HLD)

**Medical Insurance Premium Prediction**

Revision Number: 2.0

Last date of revision: 10 March 2024

Siddhi Khetmalis

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 01/03/2024 | 1.0 | Added Introduction and General description | Siddhi Khetmalis |
| 03/03/2024 | 1.1 | Added Design details and abstract | Siddhi Khetmalis |
| 05/03/2024 | 1.2 | Created and organized whole document | Siddhi Khetmalis |
| 06/03/2024 | 1.3 | Updated Design Details | Siddhi Khetmalis |
| 07/03/2024 | 1.4 | Checked and added Performance | Siddhi Khetmalis |
| 08/03/2024 | 1.5 | Added Conclusion and References | Siddhi Khetmalis |
| 09/03/2024 | 2.0 | Updated diagrams in Design Details | Siddhi Khetmalis |

**Table of Content**

Document Version Control ……………………………………………………………. 2

Abstract ………………………………………………………………………………………… 4

Introduction ………………………………………………………………………………….. 5

1. Why this High-Level Document? ……………………………………………. 5

2. Scope …………………………….………………………………………………………. 5

3. Definition ………………………………………………………………………………. 5

General Description ………………………………………………………………………. 6

1. Product Perspective ……………………………………………………….………. 6

2. Problem Statement ………………………………………………………..………. 6

3. Problem Solution ……………………………………………………………………. 6

4. Proposed Methodology …………………………………………………….……. 6

5. Further Improvements ……………………………………………………………. 6

6. Data Required ……………………………………………………...…………………. 6

7. Tools Used ………………………………………………………………………………. 7

8. Constraints …………………………………………………………………..…………. 7

9. Assumptions ……………………………………………………………………………. 7

Design Details ……………………………………………………………..…………………. 8

1. Process Workflow ………………………………………………………..…………. 8

2. Error Handling ………………………………………………………………..………. 9

Performance …………………………………………………………………………….……. 10

1. Reusability ……………………………………………………………..………………. 10

2. Application compatibility ………………………………………..………………. 10

3. Resources Utilization ………………………………………………………….……. 10

4. Deployment ……………………………………………………………………..…..…. 10

Conclusion ………………………………………………………………………….……..……. 11

Reference ………………………………………………………………………………..………. 12

**Abstract**

The medical insurance premium prediction project aims to develop a predictive model that accurately estimates insurance premiums based on various demographic, medical, and lifestyle factors. Using machine learning algorithms such as linear regression, random forest, and XGBoost, we analyze a comprehensive dataset containing information on age, gender, BMI, smoking status, pre-existing conditions, and other relevant variables. Through data preprocessing, feature selection, and model training, we construct predictive models capable of forecasting insurance premiums with high accuracy and reliability.

The project's outcomes facilitate better financial planning for policyholders, assist insurance companies in risk assessment and pricing strategies, and contribute to the enhancement of insurance services through data-driven decision-making. This abstract encapsulates the project's objectives, methods, and potential benefits, reflecting its significance in the domain of medical insurance.

**Introduction**

**1. Why this High-Level Design Document?**

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all the 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 features and the architecture of the project

**2. Scope**

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

**3. Definition**

The terms used in the projects are:

* Apriori – an algorithm for frequent item set mining and association rule learning over relational databases
* Sklearn- open-source data analysis library

**General Description**

**1. Product Perspective**

The medical insurance premium prediction project is positioned within the insurance industry, leveraging advanced analytics to accurately estimate premiums based on individual risk profiles. It aligns with industry trends toward data-driven decision-making, customer-centricity, regulatory compliance, and continuous improvement, aiming to enhance pricing models and customer satisfaction while ensuring ethical and legal standards are met.

**2. Problem Statement**

We aim to develop a system that predicts health insurance costs based on age, gender, lifestyle, and health history. This will assist individuals in budget planning and enable insurance companies to set fair prices efficiently.

**3. Problem Solution**

We've developed a computer program that analyzes information like age, smoking habits, and medical history to estimate insurance costs. By learning from data of insured individuals, our program can predict insurance expenses based on various factors. This helps individuals understand potential insurance costs and plan for medical expenses more effectively.

**4. Further Improvement**

To enhance the medical insurance premium prediction project, consider expanding data coverage, refining feature engineering, exploring model ensembles, tuning hyperparameters, addressing class imbalance, enhancing interpretability, implementing continuous monitoring, improving user interfaces, and ensuring ethical compliance. These improvements aim to boost predictive accuracy, interpretability, and usability while meeting regulatory and ethical standards.

**5. Data Required**

For training the model we need the data that consist of articles- **Age, Sex, Bmi, Children, Smoker, Region, Charges** Data is completely depending upon our problem statement

**6. Tools Used**

* Python programming language and frameworks such as NumPy, Pandas, Matplotlib, Seaborn are used to build the whole model.
* PyCharm and jupyter-notebook is used as IDE.
* For visualization of the plots, Matplotlib and Seaborn are used.
* GitHub is used as version control system.

**7. Constraints**

The prediction system should give the proper results and be user friendly. Different model to be created for new charges.

**8. Assumptions**

Assumptions for the medical insurance premium prediction project entail stable underwriting, homogeneous risk pools, valid data, linear predictor-premium relationships, independent observations, stable risk pool composition, and regulatory compliance. These assumptions guide model development but require validation for reliability.

**Design Details**

**1. Process Workflow**

For identifying the different types of anomalies, we will use a machine learning model. Below is the process flow diagram.

ML model for Recommendation

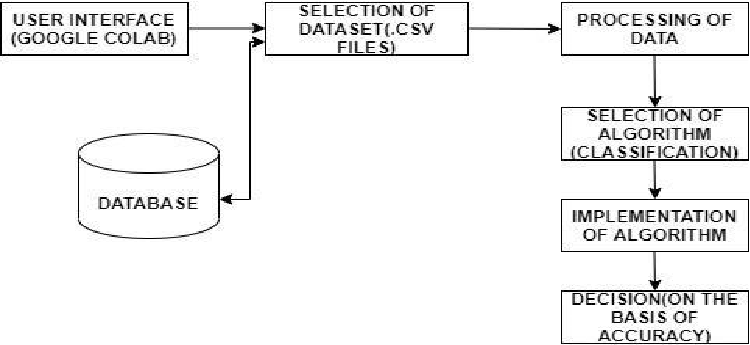
Training/Validation on data

Take the data from sensors

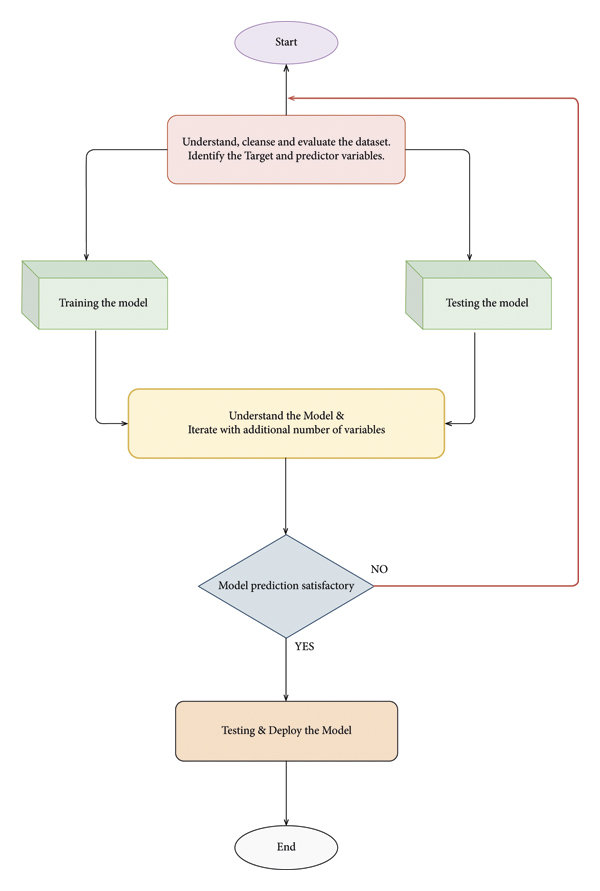
Take the necessary action

Prediction result

Model Training and Evaluation



Deployment Process



**2. Error Handling**

Initially I got an error using Juypter-Notebook while running the prediction system and using dataset too. I was also facing trouble while deploying the model and predicting different accurate charges.

**Performance**

**1. Reusability**

Medical insurance system need not be accurate as possible as it is going to just give prediction charges to customers and also the model gives the customers to spend less insurance.

**2. Application compatibility**

Since we are using python and it is compatible with any platform, we follow Application compatibility

**3. Resource utilization**

At the initial stage, I am using Kaggle as it already has dataset in the files and it need not be uploaded in the files which might consume too much of the space. And the website also uses less of the RAM than any other.

**4. Deployment**

The code is deployed in GitHub.

**Conclusion**

In conclusion, the medical insurance premium prediction project utilizes data-driven approaches to forecast insurance costs based on various factors. Leveraging techniques from machine learning and data analysis, the project aims to provide accurate estimations, aiding insurance companies and policyholders in making informed decisions. By employing robust methodologies, modular code structures, and accessible datasets, the project ensures scalability, reusability, and efficiency. Through continuous refinement and validation, this initiative seeks to contribute to improved risk assessment and financial planning in the healthcare insurance domain.

**References**

1. <https://docs.streamlit.io/en/stable/>
2. <https://numpy.org/doc/>
3. <https://seaborn.pydata.org/examples/regression_marginals.html>
4. <https://seaborn.pydata.org/examples/scatterplot_matrix.html>
5. <https://matplotlib.org/>
6. <https://pandas.pydata.org/docs/>
7. <https://streamlit.io/>
8. **Dataset**

Link: <https://www.kaggle.com/datasets/mirichoi0218/insurance/download?datasetVersionNumber=1>