HEALTH INSURTECH AMOUNT PREDICTION

A PROJECT REPORT

Submitted by

MANIKANDAN S (810019205060)

PRAVEEN KUMAR R (810019205077)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

UNIVERSITY COLLEGE OF ENGINEERING, BIT CAMPUS,

TIRUCHIRAPPALLI - 620024

ANNA UNIVERSITY:: CHENNAI - 600 025

MAY 2023

HEALTH INSURTECH AMOUNT PREDICTION

A PROJECT REPORT

Submitted by

MANIKANDAN S (810019205060)

PRAVEEN KUMAR R (810019205077)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

UNIVERSITY COLLEGE OF ENGINEERING, BIT CAMPUS,

TIRUCHIRAPPALLI - 620024

ANNA UNIVERSITY:: CHENNAI - 600 025

MAY 2023

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled **" HEALTH INSURTECH AMOUNT PREDICTION USING MACHINE LEARNING "** is a bonafide work of **"MANIKANDAN S (810019205060), PRAVEEN KUMAR R (810019205077)"** who carried out the project work under my supervision, for the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology certified further that to the best of my knowledge and belief, the work reported here in does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion.

SIGNATURE
Dr. G. ANNAPOORANI,
Assistant Professor,
HEAD OF THE DEPARTMENT
Department of CSE/IT,
University College of
Engineering, BIT Campus,
Tiruchirappalli - 620024

SIGNATURE
Ms. S. CHITRA DEVI
SUPERVISOR
Assistant Professor,
Department of Computer Science
and Engineering,
University College of
Engineering, BIT Campus,
Tiruchirappalli – 620024

Submitted for “IT8811 – Project Work” in B.Tech. Information Technology Degree Jan – May 2023 Examination held on ……………

Internal Examiner **External Examiner**

DECLARATION

We hereby declare that the work entitled **"HEALTH INSURTECH AMOUNT PREDICTION USING MACHINE LEARNING "** is submitted in partial fulfillment of the requirements for the award of the degree in B.Tech(Information Technology), University College of Engineering, BIT Campus, Anna University, Tiruchirappalli. It is a record of my work carried out by me during the academic year 2022- 2023 under the supervision of

Ms. S. CHITRA DEVI, M.E., Assistant Professor, Department of Computer Science and Engineering, Bharathidasan Institute of Technology, Anna University, Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any other degree or diploma, either in this or any other university.

(Signature of the Candidate)	(Signature of the Candidate)
MANIKANDAN S	PRAVEENKUMAR R
(810019205059)	(810019205077)

I certify that the declaration made by the above candidate is true.

(Signature of the Guide)
Ms. S. CHITRA DEVI.M.E.,
Assistant Professor,
Department of Computer Science and Engineering,
University College of Engineering (BIT Campus),
Anna University, Tiruchirappalli.

ACKNOWLEDGEMENT

A truthful heartfelt and deserved acknowledgement comes from one's heart to convey the real influence others have on one's work.

We express our gratitude to our honorable Dean**,** **Dr.T.SENTHILKUMAR, M.E., PhD.,** for giving us chance to complete our education in one of the reputed government institutions running under his leadership.

We express our sincere gratitude to our head of the department **Dr. G. ANNAPOORANI, M.Tech., Ph.D.,** for giving us the provision to do the project.

We are much obliged to our project coordinators **Dr.K.UMAMAHESWARI, M.Tech., Ph.D., and Mr.K.SARAVANA KUMAR**, **M.Tech.,** and our class coordinator **Dr. V.M.PRIYADHARSHINI, M.Tech., Ph.D.,** for giving us the opportunity to do the project, hearty thanks for them.We stand even more thankful to our project supervisor **Ms. S. CHITRA DEVI, M.E.,** for guiding us throughout and giving us the opportunity to present the main project.

We also express our sincere thanks to all other staff members, friends, and our parents for their help and encouragement.

ABSTRACT

Health insurance is a critical aspect of healthcare financing, and its importance has grown significantly over the years. In recent times, predicting health insurance amounts has become a significant challenge due to the complexities involved in the healthcare system. Predicting the health insurance amount accurately is important for insurance companies to remain profitable and for customers to plan their finances effectively. This paper proposes a machine learning approach to predict health insurance amounts based on various factors such as age, sex, gender, BMI, smoking habits, and no. of Children. The proposed model employs a random forest algorithm to predict the insurance amount. The model is trained on a dataset consisting of demographic and medical data of patients. The performance of the proposed model is evaluated on a separate test dataset, and the results show that the model is capable of accurately predicting the insurance amount. The model achieves a mean squared error (MSE) of 0.06, indicating that the model is robust and accurate. The proposed model can be useful for insurance companies to estimate the insurance premium for their customers based on the demographic and medical data. It can also be beneficial for customers to plan their finances effectively and make informed decisions about their health insurance coverage. Further, the proposed model can be extended to include additional factors such as occupation, income, and family history to improve the accuracy of the prediction. Overall, the proposed machine learning model can help insurance companies and customers to estimate the health insurance amount accurately, leading to better financial planning and efficient healthcare financing.

TABLE OF CONTENT

CHAPTER NO. **PAGE NO.**

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

UML	Unified Modelling Language
GUI	Graphical User Interface
UI	User Interface
DFD	Data Flow Diagram
SSD	Solid State Device
AI	Artificial Intelligence
ML	Machine Learning

CHAPTER 1
INTRODUCTION

1.1 Introduction

Health insurance is a critical aspect of healthcare financing, and it has become increasingly important in recent times due to rising healthcare costs. The primary goal of health insurance is to protect individuals and families from the financial burden of medical expenses in case of illness or injury. In many countries, health insurance is mandatory, and it is often provided by private insurance companies or government-sponsored programs. One of the critical challenges in the health insurance industry is predicting the insurance amount accurately. Predicting the insurance amount is essential for insurance companies to remain profitable and for customers to plan their finances effectively. However, predicting the insurance amount is a complex task due to various factors such as age, sex, gender, BMI, smoking habits, and no. of Children.

1.2 Background Understanding

Health Insurtech Amount Prediction System refers to the process of using data analysis techniques to predict the amount of health insurance claims that a particular individual or group is likely to file in the future. This prediction is based on various factors, such as the individual's medical history, demographics, lifestyle habits, and other relevant data. Insurtech, short for "insurance technology," refers to the use of technology to streamline and improve the insurance industry. Insurtech companies often leverage advanced analytics, machine learning, and artificial intelligence to make better predictions, reduce fraud, and enhance the customer experience. Predicting the amount of health insurance claims is important for insurers because it helps them price their policies more accurately and avoid losses. If an insurer underestimates the amount of claims that a policyholder is likely to file, they may end up paying out more in claims than they collected in premiums. On the other hand, if they overestimate the amount of claims, they may charge too much for their policies, leading to lower sales and profits. Insurtech companies use a range of data sources to predict health insurance claims, including medical records, pharmacy data, claims history, socioeconomic data, and behavioral data. They then use machine learning algorithms to analyse this data and make predictions about future claims. Overall, health insurtech amount prediction is a critical area of the insurance industry that is rapidly evolving, as technology continues to advance and more data becomes available for analysis.

1.3 Overview of Chapters:

Chapter 2: This chapter gives detailed view of existing research works, systems, their methodology, algorithms, source of dataset and result.

Chapter 3: The solution provided for this, goal and objective of the Health Insurtech Amount Prediction System, it's features and algorithms used for development and analysis detailed with their definition.

Chapter 4: Overview of the system configuration needed for execution of Health Insurtech Amount Prediction System as a web application, software requirements, and frame works, libraries used for application development are explained.

Chapter 5: Complete information about Dataset collection, steps followed in data pre-processing with algorithm.

Chapter 6: In this chapter, complete detail about feature extraction process for each group of algorithms were given.

Chapter 7: This chapter gives the detailed view of planning for Health Insurtech Amount Prediction System development, module split, overview of the working of Health Insurtech Amount Prediction System using System Architecture and technical work flow using Technical Architecture, data flow of the system and use case are described.

Chapter 8: Comparative study and analysis of models which are developed with differed approaches using their performance metrics and statistical metrics used for evaluation of models were given.

Chapter 9: In this chapter, Overview of deployed web application and future work of Health Insurtech Amount Prediction System were described.

Chapter 10: The resources, research articles referred for reference are listed.

CHAPTER 2
LITERATURE SURVEY

2.1 Introduction

The literature survey on health insurtech amount prediction aims to provide an overview of the current research on predictive modelling techniques and their applications in the health insurance industry. This survey includes a comprehensive analysis of the existing literature on the use of machine learning algorithms and statistical models for predicting future health insurance claims. The survey also explores the various data sources used in health insurtech amount prediction and identifies the challenges and limitations of these techniques. Overall, the survey provides a foundation for further research in the field of health insurtech amount prediction.

2.2 Related Work

Keshav Kaushik [1] proposes a machine learning-based framework for predicting health insurance premiums. The authors discuss the importance of accurate premium pricing for the health insurance industry and highlight the challenges associated with traditional premium pricing methods. The proposed framework utilizes regression models and machine learning algorithms, such as linear regression and decision trees, to predict the premiums based on various factors, including demographic information, medical history, and lifestyle choices. The authors also evaluate the performance of the proposed framework using real-world data and compare it to traditional premium pricing methods. The results suggest that the proposed framework provides more accurate and reliable premium pricing predictions compared to traditional methods, making it a promising approach for the health insurance industry.

Angela D. Kafuria [2] proposes a predictive model for computing health insurance premium rates using machine learning algorithms. The author discusses the importance of accurate premium pricing and the limitations of traditional methods. The proposed model utilizes machine learning algorithms such as decision trees and support vector machines to predict the premiums based on various factors such as age, gender, medical history, and lifestyle choices. The author evaluates the performance of the proposed model using real-world data and compares it to traditional premium pricing methods. The results suggest that the proposed model provides more accurate and reliable premium pricing predictions compared to traditional methods, making it a promising approach for the health insurance industry. The paper provides valuable insights into the potential benefits of using machine learning algorithms in health insurance premium pricing and highlights the need for further research in this area.

Shalu Gupta [3] discusses the potential benefits of using big data analytics in the health insurance industry in India. The authors provide an overview of the challenges faced by the health insurance industry and how big data analytics can be used to address these challenges. They also highlight the various sources of data available in the health insurance industry and how this data can be analyzed to improve health insurance operations and customer service. The paper concludes by discussing the potential impact of big data analytics on the health insurance industry in India and the need for further research in this area.

Chengliang Yang [4] proposes machine learning approaches for predicting high cost high need patient expenditures in healthcare. The authors discuss the importance of identifying high-cost patients to help healthcare providers allocate resources and improve patient outcomes. The proposed approaches utilize various machine learning algorithms such as decision trees, random forests, and neural networks to predict the healthcare expenditures of high-cost patients based on factors such as demographic information, medical history, and healthcare utilization. The authors evaluate the performance of the proposed approaches using real-world data and compare them to traditional methods. The results suggest that the proposed approaches provide more accurate and reliable predictions of high-cost patients compared to traditional methods, making them a promising approach for healthcare providers.

Mohamed Hanafy [5] proposes a machine learning-based framework for predicting health insurance costs. The authors discuss the importance of accurate cost prediction in the health insurance industry and the limitations of traditional methods. The proposed framework utilizes various machine learning algorithms, including decision trees, support vector machines, and deep neural networks, to predict the insurance costs based on factors such as age, gender, medical history, and lifestyle choices. The authors evaluate the performance of the proposed framework using real-world data and compare it to traditional methods. The results suggest that the proposed framework provides more accurate and reliable cost predictions compared to traditional methods, making it a promising approach for the health insurance industry. The paper provides valuable insights into the potential benefits of using machine learning algorithms in health insurance cost prediction and highlights the need for further research in this area to fully realize its potential.

Thais Carreira Pfutzenreuter [6] proposes a machine learning-based approach for medical insurance cost prediction. The authors discuss the importance of accurate cost prediction in healthcare management and the limitations of traditional methods. The proposed approach utilizes various machine learning algorithms such as decision trees and random forests to predict medical insurance costs based on factors such as patient demographics, medical history, and lifestyle choices. The authors evaluate the performance of the proposed approach using real-world data and compare it to traditional methods. The results suggest that the proposed approach provides more accurate and reliable cost predictions compared to traditional methods, making it a promising approach for healthcare management. The paper provides valuable insights into the potential benefits of using machine learning algorithms in medical insurance cost prediction and highlights the need for further research.

Nidhi Bhardwaj [7] proposes a machine learning-based approach for predicting health insurance amounts. The authors discuss the importance of accurate prediction of health insurance amounts and the limitations of traditional methods. The proposed approach utilizes various machine learning algorithms such as decision trees, support vector machines, and random forests to predict health insurance amounts based on factors such as age, gender, medical history, and lifestyle choices. The authors evaluate the performance of the proposed approach using real-world data and compare it to traditional methods. The results suggest that the proposed approach provides more accurate and reliable predictions compared to traditional methods, making it a promising approach for the health insurance industry. The paper provides valuable insights into the potential benefits of using machine learning algorithms in health insurance amount prediction and highlights the need for further research in this area to fully realize its potential.

M. Ramya [8] discusses the use of data analytics in predicting potential customers who are likely to purchase insurance. The study involved collecting and analyzing data on various factors such as age, gender, income, and occupation to identify patterns that could be used to predict a customer's likelihood of buying insurance. The results showed that data analytics can be an effective tool for insurance companies to target potential customers and increase their sales. The study used a dataset of customer information from an insurance company and applied various data analysis techniques such as regression analysis and decision tree analysis to identify key factors that influence insurance buying decisions. The researchers also proposed a predictive model that can be used to estimate the probability of a customer purchasing insurance based on their demographic and socio-economic characteristics. The results suggest that data analytics can help insurance companies tailor their marketing strategies to attract and retain customers more effectively. Overall, the study highlights the potential benefits of leveraging data analytics in the insurance industry.

Mukund Kulkarni [9] discusses the use of machine learning techniques in predicting medical insurance costs. The study involved collecting and analyzing data on various factors such as age, gender, BMI, smoking habits, and region to develop a predictive model for medical insurance costs. The researchers used various machine learning algorithms such as linear regression, decision tree, and random forest to predict the costs. The results showed that the random forest algorithm outperformed the other models and could be used to accurately predict medical insurance costs. The study highlights the potential of machine learning in predicting healthcare costs and improving the efficiency of the healthcare industry. The research also suggests that incorporating more variables related to lifestyle, occupation, and medical history could further improve the accuracy of the predictive model. The study's findings could have significant implications for insurers and healthcare providers looking to improve their pricing strategies and better allocate resources to provide quality care.

Ayushi Bharti [10] The results showed that age, BMI, and smoking habits were the most influential variables in determining the cost of medical insurance. In this paper, compared several regression models to predict medical insurance costs and found that a multiple regression model was the most accurate. This model incorporated age, gender, BMI, smoking habits, and region as predictor variables. By analysing a dataset of medical insurance claims, the study demonstrated that regression analysis can be an effective tool for predicting healthcare costs. The findings suggest that insurers and healthcare providers can use this approach to optimize their pricing strategies and better manage healthcare resources.

Dilip Kumar Sharma [11] aimed to predict health insurance emergencies using multiple linear regression. They collected data from 400 patients and identified the significant factors affecting health insurance emergencies. These factors included age, gender, occupation, monthly income, and the type of health insurance plan. Using multiple linear regression, the authors developed a model that accurately predicted the occurrence of health insurance emergencies based on these factors. They concluded that the model could be useful for insurance companies to identify high-risk customers and develop appropriate risk management strategies.In addition to identifying significant factors, the authors also conducted statistical tests to ensure the validity and reliability of their model. They found that the model had a high level of significance and a low level of multicollinearity, indicating that it was a good fit for the data.

Ghosh Madhumita [12] aimed to predict health insurance premiums using blockchain technology and random forest regression algorithms. They collected data from various sources such as medical records, socio-economic status, and lifestyle habits. The data was processed and analyzed using blockchain technology to ensure privacy and security. The project developed a predictive model using random forest regression algorithms and evaluated its performance using statistical measures. They found that the model had a high level of accuracy in predicting health insurance premiums, and the use of blockchain technology ensured data privacy and security. The study suggests that the integration of blockchain technology and random forest regression algorithms can improve the accuracy of health insurance premium predictions while maintaining data privacy and security. This can benefit both insurance companies and customers by providing more accurate premium estimates and better risk management strategies.

CHAPTER 3

PROPOSED SYSTEM

3.1 Introduction

- The proposed model is to develop a web-based ML system to predict the health insurance amount for the policyholder to file in future.
- It aims to improve the accuracy and efficiency of health insurtech amount prediction by leveraging data analytics and machine learning techniques to provide more accurate and reliable predictions of future health insurance claims.

3.2 Goal:

The main goal of the proposed model is to develop a web-based ML system to enable insurers to make more accurate predictions about the amount of health insurance claims that a policyholder or group is likely to file in the future. This helps insurers to price their policies more accurately and avoid losses, while also ensuring that policyholders are paying a fair price for their coverage. By leveraging advanced analytics and machine learning techniques, insurtech companies can analyse a wide range of data sources and make more precise predictions, which ultimately benefits both insurers and policyholders.

3.3 Objectives:

The objectives of Health Insurtech Amount Prediction may include:

- To reduce the risk of losses for insurers by accurately predicting the amount of health insurance claims that a policyholder or group is likely to file.
- To ensure that policyholders are charged a fair price for their coverage by pricing policies more accurately based on predicted claims.

- To improve the overall customer experience by enabling insurers to offer policies that are tailored to the specific needs and risk profiles of policyholders.
- To reduce fraud by identifying patterns of behavior that may indicate fraudulent activity.
- To identify areas where healthcare costs can be reduced by analyzing data on healthcare utilization and identifying opportunities for preventative care.
- To enable insurers to make data-driven decisions about which policyholders to accept and which to decline based on predicted risk.
- To provide insights to healthcare providers and policymakers on healthcare trends and patterns, which can inform decisions on healthcare policy and resource allocation.

Overall, the objectives of health insurtech amount prediction are to improve the accuracy and efficiency of the insurance industry while also providing better outcomes for policyholders and society as a whole.

3.4 Algorithms:

3.4.1. Random Forest Model:

Random Forest is a supervised learning machine learning algorithm used for both Classification and Regression problems which work based on the concept of ensemble learning technique, a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model**.** Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Decision Tree is also a supervised learning algorithm that has a predefined target variable that is used in classification problems. It works for concepts like categorical and continuous input and output variables for the model.

Random Forest splits the training dataset into n batches of k records and root of no of independent features which contain repeated records and features with different combinations to train the n Decision Trees and make a decision based on the majority. Decision Tree popularly uses two metrics Gini Index, Entropy for interpretation and built tree.

3.4.2. Mean Squared Error (MSE):

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

The formula for MSE is the following.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Where:

- y_i is the i^{th} observed value.
- $\hat{y}_i$ is the corresponding predicted value.
- n = the number of observations.

3.4.3.R-squared value:

R-squared is a statistical measure that represents the goodness of fit of a regression model. The value of R-square lies between 0 to 1. Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value. However, we get R-square equals 0 when the model does not predict any variability in the model and it does not learn any relationship between the dependent and independent variables.

The formula for MSE is the following.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}}$$

The **R^2** is calculated by dividing sum of squares of residuals from the regression model (SSres) by total sum of squares of errors from the average model (given by SStot) and then subtract it from 1.

3.5. Conclusion

In this chapter, objectives and goals of Health insurtech amount prediction system, algorithm and evaluation metrics which are used for measure the performance of models are learned.

CHAPTER 4

REQUIREMENT ANALYSIS

4.1 Introduction

For development of system, knowledge about hardware and software configurations must be known. Following of this chapter, hardware requirements such as CPU, RAM, GPU, Storage, Software requirements such as OS, Programming language, IDE, frameworks such as machine learning and deep learning frameworks, web hosting framework and libraries used for system development will discuss.

4.2 Hardware requirements:

- CPU: Laptop or PC with Intel Core i5 6th generation processor or higher with clock speed 2.5 GHz or above. Equivalent processors in AMD will also be optimal.
- RAM: Minimum 8 GB of RAM is required; 16 GB is recommended.
- GPU: NVIDIA GeForce GTX 960 or higher.
- Storage: SSD is recommended for faster pre-processing of data than HDD.

4.2. Software requirements:

- OS – Windows 7 or higher version but Windows 10 is recommended / Minimum Ubuntu 16.04 is required.
- Python (version: 3.11.0) – Programming Language used for Machine Learning.
- Visual Studio Code – Development environment.
- Spyder – Both deployment and development environment.

4.3. Frameworks:

- Sci-kit Learn – open-source machine learning library developed by the Python community.
- Keras or Tensorflow – Framework developed by Google for machine learning and deep learning.
- Flask – Python framework used for web hosting.

4.4. Libraries:

Pandas:

Pandas is a popular open-source data manipulation library for the Python programming language. It provides data structures for efficiently storing and manipulating large datasets and offers a wide range of tools for data analysis and visualization. Some key features of the Pandas library are:

- Data Structures: Pandas provides two primary data structures for handling data: Series (for one-dimensional data) and DataFrame (for two-dimensional data). Both are built on top of NumPy arrays and provide convenient methods for indexing, merging, and reshaping data.
- Data Cleaning: Pandas provides a set of powerful tools for cleaning and preprocessing data, including methods for handling missing data, removing duplicates, and transforming data.
- Data Analysis: Pandas provides a range of functions for performing common data analysis tasks, such as computing summary statistics, grouping data by categories, and applying functions to data.
- Data Visualization: Pandas integrates with other popular visualization libraries, such as Matplotlib and Seaborn, to enable users to easily create visualizations of their data.

- Overall, Pandas is a widely used and powerful library that makes working with data in Python much easier and more efficient.

Scikit-learn (sklearn):

Scikit-learn (sklearn) is a popular open-source machine learning library for Python. One of the main submodules of scikit-learn is the "ensemble" module, which provides methods for building ensemble models. Ensemble models combine multiple models to improve the overall performance and accuracy of the prediction. The "ensemble" module includes several popular ensemble methods, including:

Random Forest: A decision tree-based ensemble method that builds multiple decision trees and aggregates their predictions to make a final prediction.

Ensemble methods have been shown to be effective in a wide range of machine learning tasks, from classification and regression to clustering and anomaly detection. Scikit-learn's "ensemble" module provides a simple and powerful interface for building ensemble models in Python, making it a popular choice for many data scientists and machine learning practitioners.

Metrics: One of its submodules is the "metrics" module, which provides a range of methods for evaluating the performance of machine learning models. The "metrics" module includes functions for computing metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve. It also provides tools for evaluating clustering models and regression models. These metrics are useful for comparing different models, selecting the best model for a given task, and tuning the hyperparameters of a model. Overall, the "metrics" module in scikit-learn is a valuable tool for evaluating the performance of machine learning models in Python.

LabelEncoder: LabelEncoder is a utility class in scikit-learn that can be used to encode categorical variables as numeric values. It works by assigning a unique integer to each category in the variable. LabelEncoder is useful when working with machine learning algorithms that require numeric input. It can be applied to both nominal and ordinal variables. LabelEncoder converts string labels to numerical labels in ascending order. It is a simple and effective method for encoding categorical variables, but it may not be appropriate for variables with high cardinality or complex relationships between categories.

Joblib: Joblib is a popular open-source library in Python used for efficient and easy parallel computing of CPU-intensive tasks. It provides tools for caching and parallel execution of functions using multiple CPUs or even multiple machines. The library is especially useful for machine learning tasks that involve large datasets and computationally intensive operations, such as training models and cross-validation. Joblib also provides features for memory management, allowing users to efficiently manage large arrays of data in memory. Overall, Joblib is a powerful tool that can significantly speed up data processing and machine learning workflows in Python.

4.5 Conclusion

The knowledge about system requirements, software, libraries, frameworks are understood.

CHAPTER 5

DATASET AGGREGATION AND ACQUISITION

5.1 Introduction

For all ML or AI-based projects and resources, Dataset is the precise one. The project worth mostly based on Dataset which is used for research or model training. Most of the time, the accuracy of the model increased by train using a large amount of data. Dataset aggregation and acquisition is the process of collecting, cleaning, and organizing datasets from various sources for the purpose of analysis and modeling.

This process is important as it enables the development of accurate predictive models that can inform decisions related to healthcare and insurance policies. However, data aggregation and acquisition can be challenging due to issues such as data quality, privacy concerns, and compatibility between different data sources. To overcome these challenges, advanced data cleaning and transformation techniques, as well as secure data sharing protocols, are often used. Overall, the process of dataset aggregation and acquisition plays a critical role in the development of effective health insurtech solutions. Data preparation consists of two phases,

i. Dataset gathering

ii. Data pre-processing and Transforming

5.2 Dataset Gathering:

Gathering a suitable dataset is a crucial step in any machine learning project. In the context of health insurtech amount prediction, a relevant dataset must be collected. A well-preprocessed dataset is essential to build accurate and reliable models for predicting health insurance costs. The dataset were collected from the Kaggle.com for insurance amount prediction.

Table 1: DataSet

Age	Sex	Bmi	Children	Smoker	Region	Diseases	Amount
19	female	27.9	0	yes	southwest	yes	16884.92
18	male	33.77	1	no	southeast	no	1725.552
28	male	33	3	no	southeast	no	4449.462
33	male	22.705	0	no	northwest	no	21984.47
32	male	28.88	0	no	northwest	no	3866.855
31	female	25.74	0	no	southeast	no	3756.622
46	female	33.44	1	no	southeast	no	8240.59
37	female	27.74	3	no	northwest	no	7281.506
37	male	29.83	2	no	northeast	no	6406.411
60	female	25.84	0	no	northwest	no	28923.14

5.3 Data Pre-processing and Transformation:

Here are some common data pre-processing steps for health insurtech amount prediction:

- Data Cleaning: Remove irrelevant, redundant, and missing data from the dataset.
- Data Transformation: Convert categorical variables to numerical values using techniques like one-hot encoding, label encoding, or binary encoding.
- Feature Scaling: Normalize or standardize numerical features to ensure they are on the same scale and have similar ranges.
- Data Sampling: Select appropriate sampling techniques to address class imbalance or data scarcity issues.
- Feature Selection: Identify and select the most relevant features for the predictive model.
- Outlier Detection: Detect and handle outliers in the dataset using techniques like z-score, interquartile range (IQR), or Mahalanobis distance.

- Data Integration: Combine multiple datasets into a single dataset, ensuring that the data types and formats are compatible.
- Data Reduction: Reduce the dimensionality of the dataset using techniques like PCA or t-SNE to improve model performance and reduce computation time.

These pre-processing steps are crucial for ensuring that the data is suitable for machine learning algorithms and can yield accurate and reliable predictions for health insurtech amount prediction.

5.4 Conclusion

In this chapter, details about dataset background knowledge such as preparation, pre-processing, transformation are discussed.

CHAPTER 6
FEATURE EXTRACTION

6.1 Introduction

The feature extraction process varies for each model of Health Insurtech Amount Prediction System. To predict the insurance amount for an policyholder in Machine Learning Model were used. Machine learning model used to predict the accuracy and expected amount using Random Forest Model.

6.2 Feature extraction

Feature extraction is the process of selecting and transforming relevant features from the available data to represent the characteristics of the problem accurately. In the context of Health Insurtech Amount Prediction System, the following features can be extracted:

- Demographic features: Age, gender, occupation, income, education level, and other demographic information can be used as features as they are known to have an impact on health insurance premiums.
- Health-related features: Pre-existing medical conditions, family medical history, lifestyle habits such as smoking and drinking, and other health-related variables can be used as features. These features provide insight into the health status of the individual and help predict the likelihood of future medical expenses.
- Policy-related features: Type of policy, coverage amount, deductible, co-payment, and other policy-related features can be used as features. These features are crucial in determining the amount of premium and can help predict the cost of health insurance.
- Geographical features: The location of the individual can be used as a feature as healthcare costs vary from region to region. Factors such as availability of healthcare facilities, cost of living, and demographic

characteristics of the region can be used to predict health insurance costs.

- Behavioral features: Data related to the behavior of individuals, such as the frequency of hospital visits, claims history, and other such information, can be used as features. These features provide insight into the utilization of healthcare services and can help predict the likelihood of future medical expenses.
- Social determinants of health: Variables such as income, education, race, ethnicity, and other social determinants of health can be used as features. These factors influence the health of individuals and can have a significant impact on healthcare costs.

Overall, feature extraction is a crucial step in building an accurate prediction model for Health Insurtech Amount Prediction. It involves identifying the relevant features from the available data and transforming them into a suitable format for machine learning algorithms to use.

6.3 Conclusion

In this chapter, what are the features are extracted from dataset, how the dataset in classified and techniques used to extract features for each model are overviewed.

CHAPTER 7
SYSTEM DEVELOPMENT

7.1 Introduction

The detailed explanation of each module such as General Analysis, Learning Required Technology, Dataset Collection, Dataset Pre-processing, Model Development and Implementation, WebApp implementation, Testing Overall system and timeline for each module between 04.02.2023 and 21.04.2023, architectures such as system architecture, and other technical details of the system will discussed in following topics.

7.2. Module Split

The complete system development process is split into 5 modules such as

- General Analysis
- Learning Required Technology
- Dataset Collection
- Dataset Pre-processing
- Model Development and Implementation
- WebApp implementation
- Testing Overall system

7.2.1 Modules:

7.2.1.1. General Analysis

The General Analysis module consists of a Literature Survey (study of existing systems, research works, relevant and similar systems with their approaches and used algorithms, gaining experience and knowledge from them), Requirement Analysis (defining the environments and identifying system needs, finding the required tools, technology, hardware and software with their availability, capabilities, consistency, features, supported platforms and languages then make sure the right fit of them for further development process)

7.2.1.2. Learning Required Technology

After finding the right tools and technology, studying and learning the unknown tools and technologies in-depth which are needed.

7.2.1.3. Dataset Collection

This module consists of planning for collecting data like what data to be collected, how to collect, where to collect, and analysing better choices and fitting them. As per the plan, data must collect.

7.2.1.4. Dataset Pre-processing

In this module, collected data undergoes various cleansing processes and techniques, then transform into the right format and required features are extracted using existing and own defined algorithms.

7.2.1.5. Model Development and Implementation

The development consists of building a model using different algorithms with creative approaches, training the developed model using train dataset then test and evaluate the models to find their performance using evaluation techniques and metrics. The model which performs best is implemented in the application.

7.2.1.6. WebApp implementation

Creating interactive GUI and implementing the best model in the back-end of the application then deploying it as a web app using Flask framework are came under this module.

7.2.1.7. Testing Overall system

Testing each and every component, unit, their function and debugging issues if anything raises and evaluating overall system performance using metrics are done in the testing module.

7.2.2 Gantt Chart:

A Gantt chart is a visual project management tool that illustrates the start and finish dates of individual tasks and their dependencies within a project. It helps in planning and scheduling tasks, managing resources, and tracking progress.

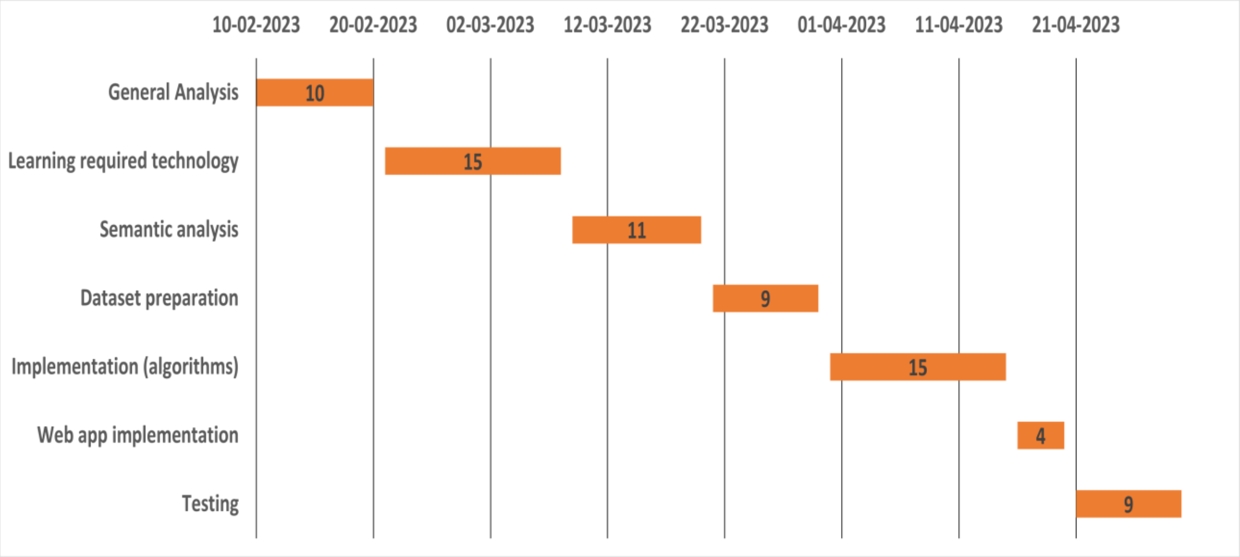


Figure:7.2.2.1 Gantt Chart

7.3 Architecture Diagram:

7.3.1 System architecture:

A system architecture diagram is the diagrammatic representation of an overview of the project. It depicts the methodologies used in the layer and the relationship between layers in the system. The most common architecture pattern is the layered architecture pattern also known as n-tier architecture pattern. The proposed system consists of a database layer, pre-processing layer, core layer, and UI layer. Dataset undergoes several pre-processing techniques in pre-processing layer and is converted to numerical data. Numerical data is then fed to the model in the core layer to depict the result. The result is then displayed in UI layer to the user.

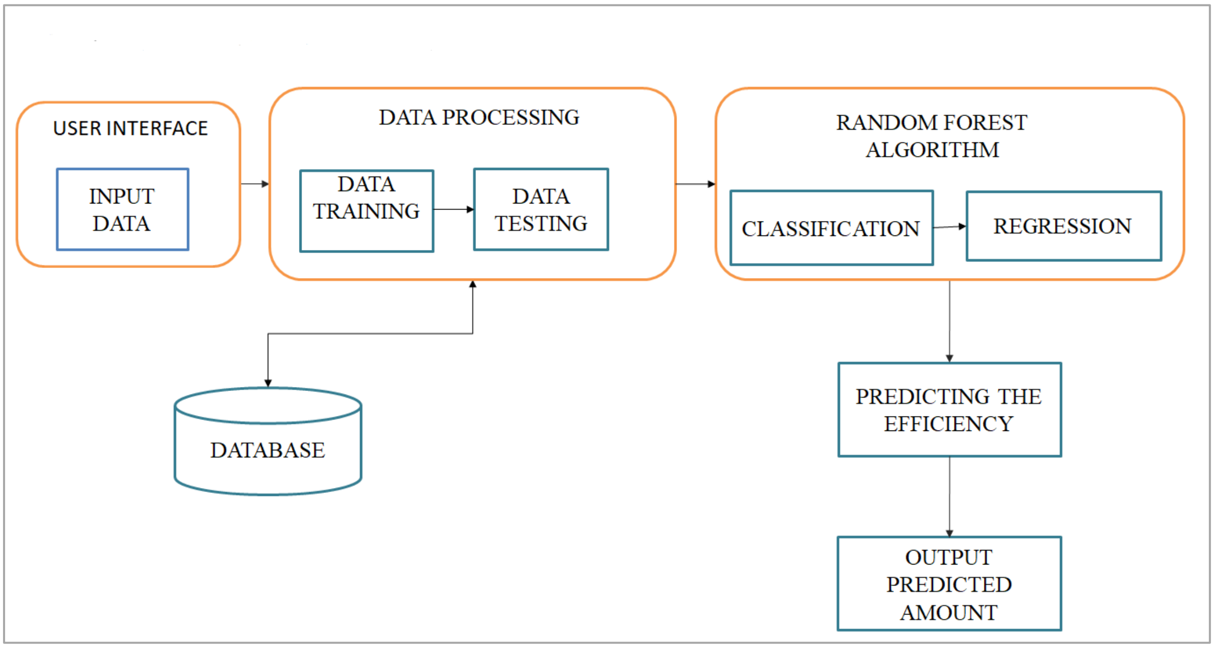


Figure:7.3.1 System Architecture Diagram

The above figure shows the layered architecture of the Health Insurtech Amount Prediction System. Health Insurtech Amount Prediction had four layers

- UI Layer
- Pre-processing Layer
- Core Layer
- Database Layer

UI layer had a user-interactive webpage, used to communicate with the user. Users can give input to the system. After given the output, Health Insurtech Amount Prediction System displays the output through a webpage.

7.4 UML Diagram:

7.4.1 Use case Diagram:

A use case diagram represents the behavior of the system. It expresses how the user can interact with the system, the services provided by the system, and relationship between them. The components are actors represented by stick figures and use cases represented by ellipse, system and lines. Ellipse represents the role of the actors and whole system functions bounded by rectangle bound which portrays complete system functionality.

Following use case diagram shows the actors of Health Insurtech Amount Prediction System and their functions. Health Insurtech Amount Prediction System had two actors

i. User
ii. Developer

User can give input and saw the output. if any queries, can file. System accepts the user input then pre-process, convert and detect the result which is provided as output for user. Database save inputs, outputs and provide saved previous knowledge for the system to detect output. Developer collects new dataset then pre-process, train and test the model, deploy the trained model and update system if required.

User – who interacts with a system for their purpose.

Developer – who develop, maintain and upgrade the system regularly.

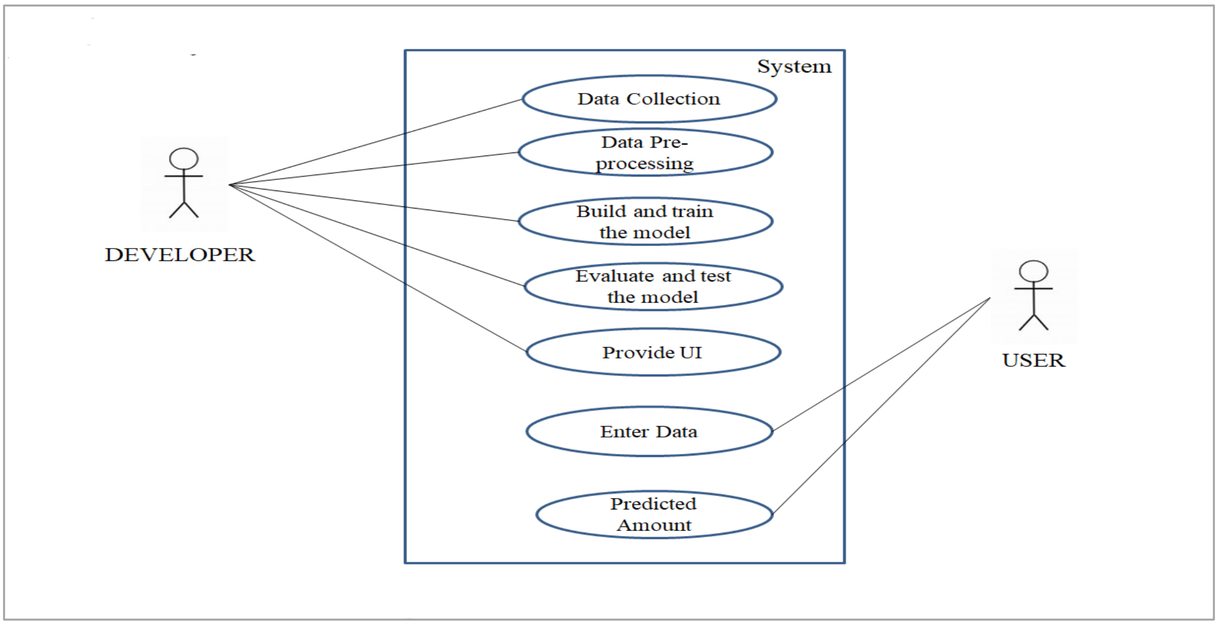


Figure:7.4.1 Use Case Diagram

The above figure shows the Use case diagram of the Health Insurtech Amount Prediction System.

7.4.2 Data Flow Diagram:

Data Flow Diagram (DFD) is the representation of information flows in the system. It shows how data enters, what process takes place and where is data stored in that system. It is also known as data flow graph. It is classified into three different levels based on increasing information and functionality of the system by,

- DFD Level 0
- DFD Level 1
- DFD Level 2

Dataflow Diagram Level 0:

Zeroth level DFD shows the overall data flow of the Health Insurtech Amount Prediction System model. Data undergoes the subsequence of processes such as pre-processing, feature extraction, and predicting the output by the model with help of a database.

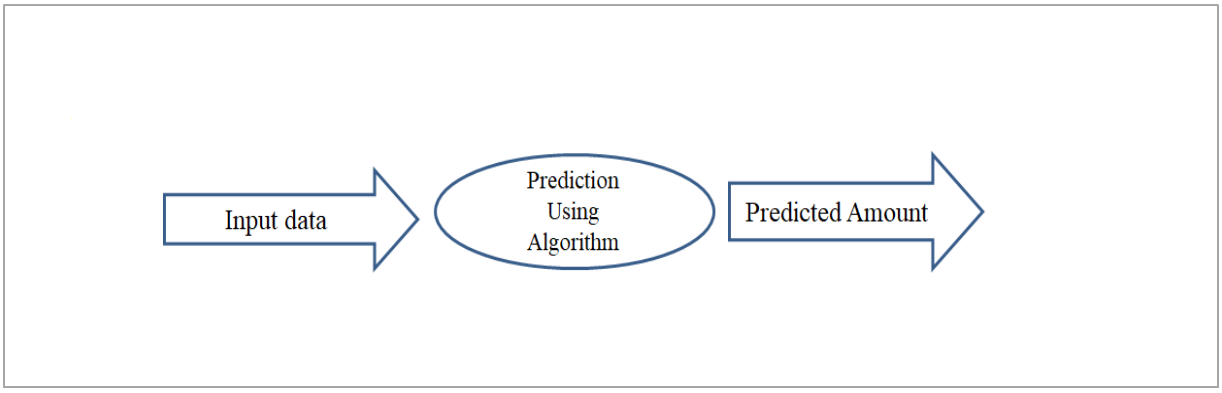


Figure:7.4.2.1 0th DFD Diagram

The above figure shows the 0th level DFD Diagram of the Health Insurtech Amount Prediction System.

Dataflow Diagram Level 1:

Level one DFD shows the detailed view of dataflow in Health Insurtech Amount Prediction System model with their techniques. Till feature extraction, all the details are same as level zero DFD. Data undergoes the subsequence of processes such as pre-processing, feature extraction, and predicting the output by the model with help of a database.

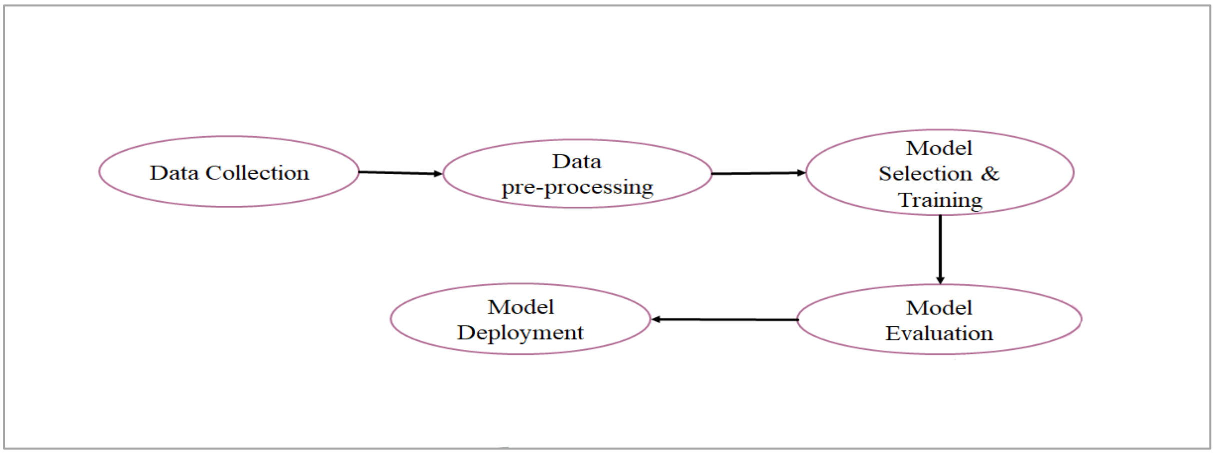


Figure:7.4.2.2 1st DFD Diagram

The above figure shows the 1st level DFD Diagram of the Health Insurtech Amount Prediction System.

Dataflow Diagram Level 2:

Level two DFD gives the complete and detailed view of full Health Insurtech Amount Prediction System as a web application by dividing them into client and server sides. In client side, get input from the user and pass to server by request. Health Insurtech Amount Prediction System took input from the client-end and gives output by passing through various processes then give output.

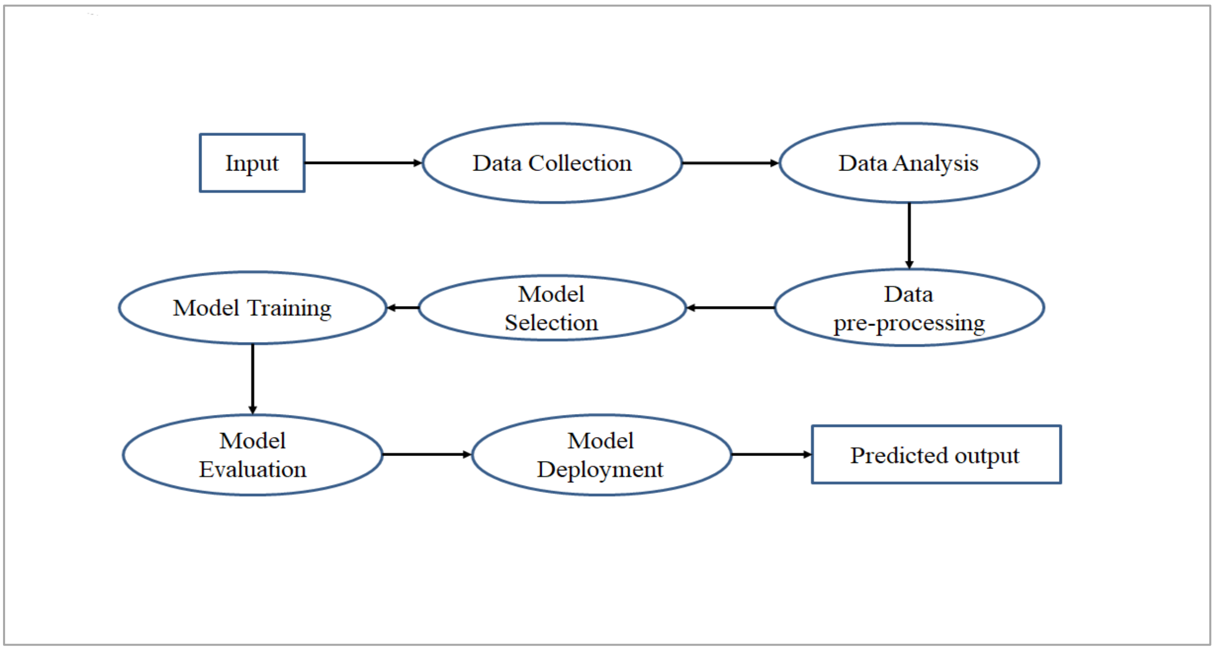


Figure:7.4.2.3 2nd DFD Diagram

The above figure shows the 2nd level DFD Diagram of the Health Insurtech Amount Prediction System.

7.5 Conclusion

Working of the system, their data flow and overview of development process are seen in this chapter.

CHAPTER 8

MODEL EVALUATION

8.1 Introduction for Model Testing and Evaluation:

After feature extraction, models were building various techniques, algorithms and approaches, respective developed models trained and tested with their data for evaluating their performance by evaluation metrics.

8.2 Machine Learning with model evaluation techniques:

Model evaluation is a crucial step in health insurtech amount prediction as it helps to determine the performance and accuracy of the model. The following are some common evaluation metrics used in machine learning For Health Insurtech Amount Prediction System:

1. Mean squared error (MSE): MSE measures the average squared difference between the predicted and actual values. It is widely used to evaluate regression models in health insurtech amount prediction.

2. Root mean squared error (RMSE): RMSE is the square root of the MSE and provides a more interpretable measure of the error in health insurtech amount prediction.

3. Mean absolute error (MAE): MAE measures the absolute difference between the predicted and actual values and is another common evaluation metric used in health insurtech amount prediction.

4. R-squared (R2): R2 measures the proportion of the variance in the target variable that can be explained by the model. A high R2 value indicates a good fit of the model to the data in health insurtech amount prediction.

5. Cross-validation: Cross-validation is a technique used to assess the performance of the model on unseen data. It involves dividing the data into

several subsets, training the model on one subset, and evaluating it on the remaining subset.

6. Receiver operating characteristic (ROC) curve: ROC curve is a graphical representation of the performance of the binary classification model in health insurtech amount prediction. It plots the true positive rate against the false positive rate at different classification thresholds.

7. Precision and recall: Precision and recall are evaluation metrics used for classification models in health insurtech amount prediction. Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives.

8.3 Conclusion

Model evaluation allows us to ensure that our predictive model is robust and reliable and can make accurate predictions for health insurance amounts, which can ultimately help insurance companies better manage risk and improve customer satisfaction.

CHAPTER 9

CONCLUSION

9.1 Introduction

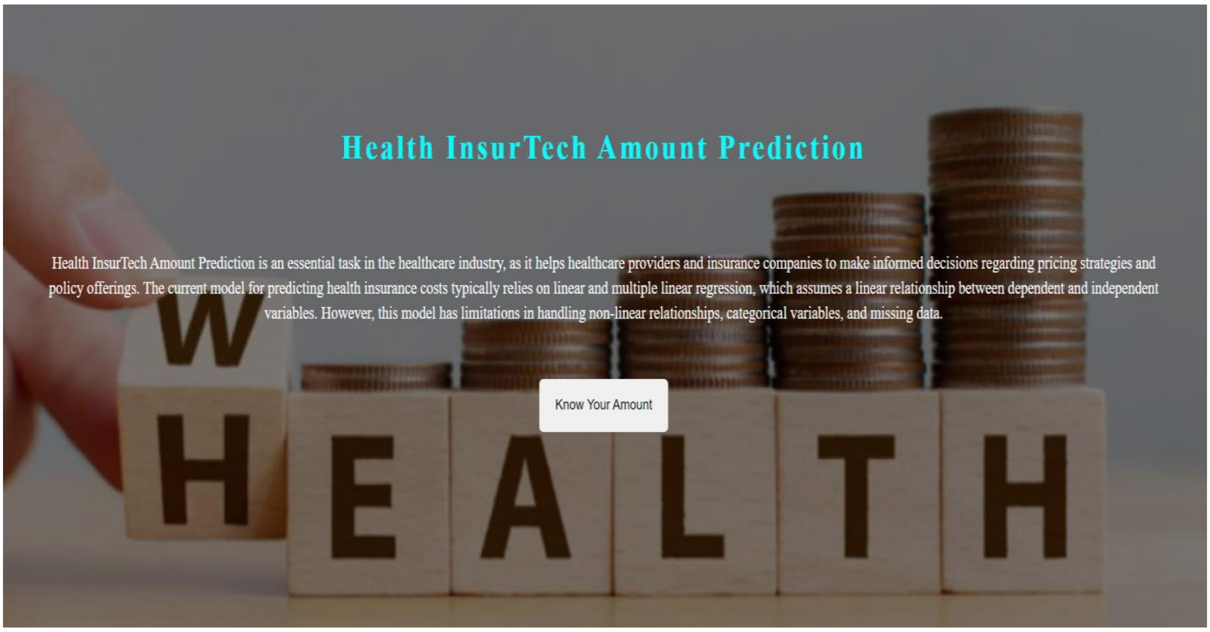
Health Insurtech Amount Prediction System model deployed as a web application using Flask web hosting framework, overview of the UI features and conclusion with advantages, disadvantages and future work of Health Insurtech Amount Prediction System will be described.

9.2 Web Application:

For users to interact with the model, Health Insurtech Amount Prediction System was developed as a web application.

Flask is used to deploy the Health Insurtech Amount Prediction System. It is a popular Python framework used for deploying and hosting web applications using Python language.

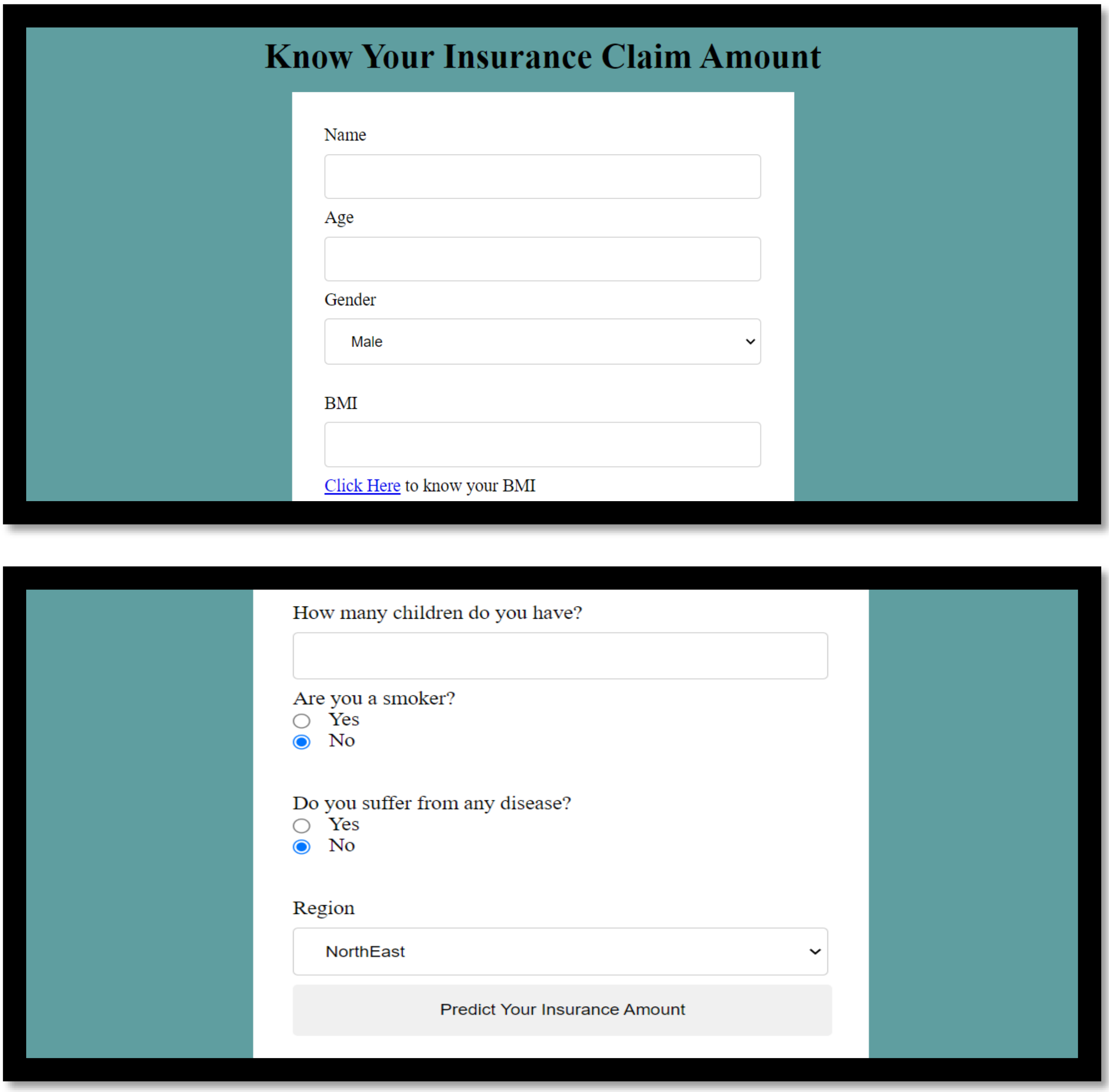
Following figs show the user interface of Health Insurtech Amount Prediction System with its interactive components,



Health InsurTech Amount Prediction

Health InsurTech Amount Prediction is an essential task in the healthcare industry, as it helps healthcare providers and insurance companies to make informed decisions regarding pricing strategies and policy offerings. The current model for predicting health insurance costs typically relies on linear and multiple linear regression, which assumes a linear relationship between dependent and independent variables. However, this model has limitations in handling non-linear relationships, categorical variables, and missing data.

FIGURE:9.2.1 GUI Image 1



Know Your Insurance Claim Amount

Name

Age

Gender

Male

BMI

Click Here to know your BMI

How many children do you have?

Are you a smoker?
○ Yes
● No

Do you suffer from any disease?
○ Yes
● No

Region

NorthEast

FIGURE:9.2.2 GUI Image 2

9.3 Conclusion

Calculation of health insurance charges in the traditional process are a hefty task for the insurance companies. Following human intervention in the process may sometime produce faulty or inaccurate results and also when the data increases the time taken for calculation by human's increases.

In scenarios like these the implementation of Machine Learning models can be very beneficial to the company and the policyholders. In this project, Machine Learning model are used to predict the amount of health insurance based on specific attribute values present in the dataset.

The results obtained are R2 of 0.88, and an accuracy of 88 percent, using the Random Forest model. Based on the model's configuration parameters which are tuned during the training phase, on the basis of performance the different proposed models are arranged.

9.3.1 Future Work:

- Add the particular diseases in the dataset sources.
- Incorporating additional data sources.
- One direction is to explore the use of more advanced machine learning algorithms and techniques, such as deep learning and ensemble methods, to improve the accuracy and robustness of the prediction models.

CHAPTER 10

REFERENCES

1. Keshav Kaushik , Akashdeep Bhardwaj , Ashutosh Dhar Dwivedi , and Rajani Singh, ” Machine Learning-Based Regression Framework to Predict Health Insurance Premiums”, IJERPH, p 4, 2022.

2. Angela D. Kafuria, “Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms”, IJC vol 44, p 4, 2022.

3. Shalu Gupta, Dr. Pooja Tripathi, “An Emerging trend of Big Data Analytics with Health Insurance in India”, ICICCS, p 5, 2016.

4. Chengliang Yang, Chris Delcher , Elizabeth Shenkman and Sanjay Ranka, “Machine learning approaches for predicting high cost high need patient expenditures in health care”, IWCBBE, p 5, 2017.

5. Mohamed hanafy, Omar M. A. Mahmoud, “Predict Health Insurance Cost by using Machine Learning and DNN Regression Models”, IJITEE, p 6, 2021

6. Thais Carreira Pfutzenreuter, Edson Pinheiro de Lima,”Machine Learning In Healthcare Management For Medical Insurance Cost Prediction”, p6,2021.

7. Nidhi Bhardwaj , Rishabh Anand, “Health Insurance Amount Prediction”,IJERT, p 7, 2020.

8. M. Ramya, A. Sankeerthana, S. Harshitha, Dr. Sunil Bhutada, Dr.Y. Rohita,”Predicting Possible Prospects To Buy Insurance Using Data Analytics”, IJCRT, p 8,2021.

9. Mukund Kulkarni , Dhammadeep D. Meshram , Bhagyesh Patil , Rahul More , Mridul Sharma , Pravin Patange, “Medical Insurance Cost Prediction using Machine Learning”, ILRASET, p 9,2022.

10. Ayushi Bharti, Lokesh Malik, “Regression Analysis And Prediction Of Medical Insurance Cost “, IJCRT, p 9, 2022.

11. Dilip Kumar Sharma, Ashish Sharma,”Prediction of Health Insurance Emergency using Multiple Linear Regression Technique”, EJMCM, p 10, 2020.

12. Ghosh Madhumita, Ravi Gor, “Health Insurance Premium Prediction Using Blockchain Technology And Random Forest Regression Algorithms”,IJEST, p10, 2022.

13. Omar, T.; Zohdy, M.; Rrushi, J. Clustering Application for Data-Driven “Prediction of Health Insurance Premiums for People of Different Ages”, (ICCE), 2021.

14. Sailaja, N.V.; Karakavalasa, M.; Katkam, M.; Devipriya, M.; Sreeja, M.; Vasundhara, D.N. “Hybrid Regression Model for Medical Insurance Cost Prediction and Recommendation”, (ICISSGT), 2021.23.

15. Dutta, K.; Chandra, S.; Gourisaria, M.K.; GM, H. A “Data Mining Based Target Regression-Oriented Approach to Modelling of Health Insurance Claims”, (ICCMC), 2021.