Medical Insurance Prediction Model Using Linear Regression

Project Synopsis

<Version 1.0>

Industrial Training

(BCA 551)

BACHELOR OF COMPUTER APPLICATIONS

PROJECT GUIDE:

Mr. Ajay Rastogi

SUBMITTED BY:

NAME:-ANKIT SAINI
Enroll_No:- (TCA2201092)
Semester:- 5th
Section:- C

September, 2024

COLLEGE OF COMPUTING SCIENCES & INFORMATION TECHNOLOGY

TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD

Table of Contents

1 Project Title

Medical Insurance Prediction Model Using Linear Regression

2 Domain

Machine Learning, Data Science

3 Problem Statement

Insurance cost estimation is a critical aspect for both providers and individuals, as it helps in understanding and managing the financial aspects related to healthcare. However, determining accurate medical insurance costs based on demographic and lifestyle factors can be challenging. This project aims to develop a machine learning model that can predict medical insurance costs by analyzing factors like age, BMI (Body Mass Index), smoking status, gender, and number of children. The model's goal is to assist users and insurance companies in estimating insurance expenses accurately and efficiently, making the process more accessible and data-driven.

4 Project Description

The Medical Insurance Prediction Model leverages machine learning, specifically a Linear Regression model, to predict medical insurance costs based on a dataset of individual factors. The project uses a dataset obtained from Kaggle, containing data on individuals' ***age, BMI (Body Mass Index), gender, smoking status, number of children, and actual insurance charges****. Initially, the data undergoes preprocessing steps, including handling categorical variables and scaling numerical features. The Linear Regression algorithm is then trained on this processed data to predict the insurance cost. Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to assess model accuracy. This project aims to provide a user-friendly tool that enables users and insurance providers to estimate costs quickly, helping to make informed financial decisions regarding medical insurance.*

4.1 Scope of the Work

- ***Data Processing and Cleaning****: Prepare and clean the dataset by handling categorical variables, normalizing numerical features, and removing any unnecessary columns.*
- ***Model Development****: Implement a Linear Regression model that captures the relationship between the input features (age, BMI, smoking status, etc.) and the target variable (insurance cost).*
- ***Evaluation and Analysis****: Evaluate the model using metrics like MAE, MSE, and RMSE to ensure accuracy and reliability*.
- ***Practical Application****: This model can be integrated into a web-based application or standalone tool that allows users to input their details and receive an estimate of their medical insurance costs.*

- **Future Enhancements**: *The model can be expanded to include more features, such as medical history, location, or lifestyle habits, to improve prediction accuracy. Additionally, the model's functionality could be extended to support integration with other financial or insurance systems.*

4.2 Project Modules

- **Data Collection and Preprocessing** : Clean, encode, and scale features.
- **Model Building** : Train a linear regression model on the processed data.
- **Evaluation** : Analyze model accuracy and visualize predictions against actual values.

5 Implementation Methodology

- The dataset was obtained from Kaggle, containing 1338 records with attributes such as age, sex, Body Mass Index(BMI), children, smoker, and charges.
- Data preprocessing involved encoding categorical features and removing unnecessary columns.
- A linear regression model was trained on a 75:25 train-test split, and the model was evaluated using MAE, MSE, and RMSE metrics.

6 Technologies to be used

6.1 Software Platform

1. Programming Language: Python

2. Libraries and Frameworks:

- **Data Handling:** Pandas (for data manipulation and analysis)

- **Machine Learning:** Scikit-learn (for implementing machine learning algorithms and model evaluation)

- **Data Preprocessing:** NumPy (for numerical operations and data manipulation)

- **Development Environment:**

A) Integrated Development Environment (IDE): jupyter Notebook ,Visual Studio Code
B) Version Control: Git and GitHub (for version control and collaboration)
C) Data Visualization : Matplotlib or Seaborn (for plotting and visualizing data)

6.2 Hardware Platform

A) RAM: Minimum of 8 GB recommended (for handling data and running machine learning models efficiently)

B) Hard Disk: Minimum of 256 GB (for storing datasets, models, and project files)
C) Operating System: Windows 11 or later
D) Editor: Visual Studio Code

6.3 Tools

1) Python

- ***Purpose:*** *Primary programming language for developing the machine learning model.*

2) Pandas

- ***Purpose:*** *Used for data manipulation and preprocessing.*

3) Scikit-learn

- ***Purpose:*** *Provides machine learning algorithms and tools for model training and evaluation.*

4) NumPy

- ***Purpose:*** *Used for numerical operations and data manipulation.*

5) Visual Studio Code (VS Code)

- ***Purpose:*** *Integrated Development Environment (IDE) for writing and debugging code.*

6) Git

- ***Purpose:*** *Version control system for tracking changes in the code*

7) GitHub

- ***Purpose:*** *Platform for hosting repositories and collaborating on code.*

7 Results

- ***Purpose:*** *Our linear regression model performed well on the test data, achieving an MAE (Mean Absolute Error) of 3,534.50, MSE (Mean Squared Error) of 25,306,900.17, and RMSE (Root Mean Squared Error) of 5,030.47. The scatter plot of the predicted versus actual values showed a strong linear relationship, indicating a good fit of the model and reasonable prediction accuracy.*

8 Algorithm & Model

- ***Algorithm***: Linear Regression
- ***Formula:*** $Y=mX+c$
- ***Y:*** Dependent variable
- ***X:*** Independent variables (age, BMI, smoker status)
- ***m:*** Slope y a coefficient of each independent variable
- ***c:*** Y-intercept (constant term)

9 Accuracy

- ***Model Accuracy:*** *0.7064951721664848*

10 Advantages of this Project

- Offers a data-driven approach to estimate insurance costs.
- Helps users and providers understand the factors affecting insurance costs.

11 Future Scope and further enhancement of the Project

- Add additional features like medical history for more accurate predictions.
- Deploy the model in a web application for easier access and usability.

12 Team Details

Project Name & ID	Course Name	Student ID	Student Name	Role	Signature
Medical Insurance Prediction Model Using Linear Regression	BCA	TCA2201092	ANKIT SAINI	*Developer, Tester*	ankitsaini

13 Conclusion

This project effectively utilizes linear regression to predict medical insurance costs, demonstrating a practical application of machine learning for financial estimation in the healthcare sector.

14 References

You Tube, Geeks for Geeks, Kaggle.