The Superior University

Project Title

Medical Insurance Prediction

Project Details

1. Course: Artificial Intelligence
2. Instructor: (Your Instructor's Name)
3. Semester: 3rd
4. Section: BSAI-116-3C
5. Submission Date: 02/11/2024
6. Group Members: Individual

Name	Roll No	Email	Contact
Alishba Haroon	BSAI-116		

Abstract

Purpose:

The primary purpose of the Medical Insurance Prediction Model is to estimate the premiums that are to be charged to an individual based on certain demographic, lifestyle, and health factors. Based on historical data, this model enables insurance companies to arrive at fair and accurate pricing strategies while empowering users to understand the factors behind his or her premiums.

Functionality:

This model uses advanced machine learning techniques to compute insurance costs based on inputs such as age, gender, BMI, smoking habits, number of dependents, and residential region. Using algorithms like Linear Regression or Random Forest, the system trains on preprocessed data following missing value handling and

encoding the categorical features. The system can be user-friendly for use by insurance companies and clients alike in search of the transparency of premium calculation.

Importance

This model has great implications on insurance. It promotes data-based decision-making and personalized pricing practices. For the consumers, it provides insight into their lifestyle changes that would minimize paying for premiums. It also tends to make risk assessment a fair and efficient tool and creates trust between the insurers and clients. The model can easily be extended for more comprehensive usage such as health risk analyses and tailored wellness advice.

Table of Contents

Introduction:

Overview of the Project:

This insurance premium prediction model, also known as Medical Insurance, for healthcare insurance aims at trying to estimate insurance premiums, contingent on individual demographic and health factors related to age, BMI, smoker status, and number of dependents. It leverages advanced machine learning techniques to make more accurate predictions that are essential both to the insurance company and clients because it simplifies understanding of a much better system of calculating the premium itself in relation to integrating history through a predictive analytics system.

Explanation:

Medical insurance is one of the most important aspects of modern healthcare systems, which affects both personal financial planning and institutional risk management. The selected topic bridges the gap between healthcare and technology by leveraging predictive analytics to address a real-world problem. The project focuses on developing a user-friendly, efficient model that integrates machine learning techniques and insurance data for practical applications.

Relevance to Concepts of Operating Systems:

The project relates to concepts of operating systems in several ways. The training and prediction processes in the machine learning model are computationally intensive tasks that, in most cases, require efficient resource management by the OS. The use of multithreading and process scheduling optimizes resource utilization for smooth execution of the model. In addition, OS-level storage and memory management can be critical in handling massive datasets and ensuring fast access at both training and inference time. This synergy underlines the importance of OS for developing robust and scalable systems for prediction.

Objectives:

Implement a Predictive Model for Medical Insurance Costs:

To implement and demonstrate the working of a predictive model that estimates medical insurance costs based on factors such as age, sex, BMI, smoking status, and region.

Understand the Relationship Between Features and Insurance Costs:
Analyze and understand how different features, such as age, sex, BMI, and smoking habits, impact medical insurance costs.

Explore Data Preprocessing Techniques:
To explore data preprocessing techniques like encoding categorical variables, handling missing values, and scaling numerical data to prepare the dataset for machine learning models.

Compare the Performance of Different Models:
To evaluate and compare the performance of different regression models (e.g., Decision Tree, Random Forest, Linear Regression) in predicting medical insurance costs and select the most suitable model based on performance metrics like R-squared and Mean Squared Error (MSE).

Analyze Model Accuracy and Make Improvements:
To improve the accuracy of the predictive model through hyperparameter tuning and optimization techniques, aiming to provide reliable cost predictions.

System Requirements

Hardware Requirements:

Processor:

- Intel Core i5

RAM:

- 8 GB of RAM

Storage:

- 167GB storage.

Software Requirements:

Operating System:

- Windows 10

Programming Languages:

- Python 3.12

Integrated Development Environment (IDE):

- Visual Studio Code

Libraries:

- Pandas (for data manipulation and preprocessing)
- NumPy (for numerical operations)
- Scikit-learn (machine learning models such as regression, classification, and model evaluation)
- Matplotlib and Seaborn (for data visualization and plotting graphs)

Methodology

Data Collection & Preprocessing:

- Data Collection: The dataset includes attributes like age, sex, BMI, smoking habits, region, and insurance cost.
- Preprocessing: Missing values were handled, categorical features (sex, smoker, region) were encoded, and numerical features were scaled.

Data Splitting:

- The dataset was split into 80% training and 20% testing sets.

Model Selection & Training:

- Random Forest Regressor was chosen for its ability to handle non-linear relationships and prevent overfitting.
- Hyperparameters such as the number of trees and tree depth were tuned.

Model Evaluation:

- The model was evaluated using R-squared.

Flowchart:

Start --> Data Collection --> Data Preprocessing --> Data Splitting --> Train Model --> Evaluate Model --> Deployment

Techniques Used:

- Random Forest Regressor for prediction.
- Label Encoding for categorical features.
- Feature Scaling to standardize numerical values.
- R-squared, MAE, and RMSE for model evaluation.

Implementation

Step 1: Data Collection and Loading

```
import pandas as pd
df = pd.read_csv('insurance.csv')
print(df.info())
print(df.head())
```

Step 2: Data Preprocessing

```
df['sex'] = df['sex'].replace({'male': 0, 'female': 1})
df['smoker'] = df['smoker'].replace({'yes': 1, 'no': 0})
df['region'] = df['region'].replace({'southeast': 0, 'southwest': 1, 'northeast': 2,
'northwest': 3})
```

Step 3: Splitting Data

```
X = df.drop('charges', axis=1)
y = df['charges']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Step 4: Model Training

```
rf_regressor = RandomForestRegressor(n_estimators=100, random_state=2)
rf_regressor.fit(X_train, Y_train)
```

Step 5: Model Evaluation

```
rf_train_preds = rf_regressor.predict(X_train)
rf_test_preds = rf_regressor.predict(X_test)
rf_r2_train = r2_score(Y_train, rf_train_preds)
rf_r2_test = r2_score(Y_test, rf_test_preds)
print("R-squared value on training data: ", rf_r2_train)
print("R-squared value on test data: ", rf_r2_test)
```

Step 6: Prediction for New Data

```
input_data = (31, 1, 25.74, 0, 1, 0)
```

```
input_data_df = pd.DataFrame([input_data], columns=['age', 'sex', 'bmi', 'children', 'smoker',
'region'])
prediction = regressor.predict(input_data_df)
print('The insurance cost is USD ', prediction[0])
```

Challenges and Solutions

Challenge: Evaluating model performance accurately and identifying areas of improvement.
Solution: Used multiple metrics, including **R-squared**, **MAE**, and **RMSE**, to ensure a comprehensive evaluation. Adjustments were made based on these metrics.

Conclusion

The **Medical Insurance Prediction Project** successfully predicted insurance costs using a Random Forest Regressor with high accuracy. It provided insights into how factors like age, BMI, and smoking habits influence premiums. Key outcomes included mastering data preprocessing, feature engineering, and model evaluation. The project offers a foundation for real-world predictive solutions and potential future deployment.