Medical Premium Insurance Cost Prediction

Final Project Proposal

2024

Team Members:

* PrasannaKumar
* Akhil
* Naga Sai Anudeep
* Manoj

**Proposed Timeline:**

* **Week 1:** Data Collection and Preprocessing
* **Week 2:** Exploratory Data Analysis and Feature Selection
* **Week 3-4:** Model Development and Evaluation
* **Week 5-6:** Fine-tuning and Optimization
* **Week 7-8:** Documentation and Presentation Preparation

**Objectives and Goals:**

The objective of this project is to develop a regression model to estimate future medical costs for individuals based on their demographic and lifestyle factors. By leveraging data such as age, sex, BMI, number of children, smoker status, and regional categorization, the goal is to create a predictive tool that enables users to plan for potential medical expenses. This tool aims to empower individuals to make informed financial decisions regarding healthcare by providing them with personalized estimates of future medical costs.

**Data and Preprocessing**

**Dataset Link :** [**https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction?resource=download**](https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction?resource=download)

**Dataset Attributes:**

* **Age:** The age of the individual.
* **Sex:** The gender of the individual (male or female).
* **BMI (Body Mass Index):** A measure of body fat based on height and weight. It is calculated as weight in kilograms divided by the square of height in meters.
* **Number of Children:** The number of dependent children the individual has.
* **Smoker Status:** Indicates whether the individual is a smoker or a non-smoker.
* **Regional Categorization:** Categorizes the individual's location or region, which may have an impact on medical costs.
* **Medical Expenses:** The target variable representing the medical expenses incurred by the individual.

**Pre-Processing:**

Check the dataset for any missing values in any of the attributes. If missing values are found, decide on an appropriate strategy for handling them, such as imputation (replacing missing values with a suitable estimate) or deletion (removing rows or columns with missing values). Since some of the attributes (such as sex, smoker status, and regional categorization) are categorical, they need to be converted into numerical format for machine learning algorithms to process them effectively. This can be done through techniques like one-hot encoding, where each category is represented by a binary (0 or 1) indicator variable.

The numerical features in the dataset (such as age, BMI, and number of children) may have different scales, which can affect the performance of certain machine learning algorithms. To address this, it's common practice to scale or normalize the numerical features to a similar range. This ensures that all features contribute equally to the model's predictions.

**Methodology:**

We propose using Linear regression as the initial approach for modeling medical costs. Additional algorithms will be explored, and comparisons will be made to determine the most effective method. Model evaluation will be based on R-Squared and adjusted R-Squared values.

**References:**

1. "Predicting Health Care Costs Using Evidence Regression" - This article addresses a similar problem and provides insights into regression-based approaches for estimating healthcare costs.

<https://www.mdpi.com/2504-3900/31/1/74?type=check_update&version=2>

1. "Development of a predictive model for integrated medical and long-term care resource consumption based on health behavior" - This resource demonstrates the use of healthcare data to predict resource consumption for patients with circulatory diseases, serving as a basis for our project.

<https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01874-6>

**Expected Outcomes:**

We will evaluate the project's results by comparing predicted medical costs to actual costs using performance metrics such as mean absolute error and root mean square error. Visual representations, such as scatter plots and regression curves, will be used to analyze the model's accuracy and performance.