Group number: 5
Name of group members: **Mohammad Shaik**
Jagadeesh Siripurapu

Price Prediction for Health Care Insurance

Health care costs are one of the major public health issues in one's personal life. Based on different research studies, BMI, ageing, smoking, and other factors are all related to greater personal medical care costs. Among those different factors mentioned BMI is one of the major which effects towards the price estimation. Thus, accurately predicting future costs and understanding which factors contribute to increases in health care expenditures are important by building one of the Machine Learning Algorithm.

The research of the project is about to find out the areas or factors that are resulting in the increase in Health care costs. The main objective is to **predict a patient's healthcare costs depending upon the features**. To identify factors contributing to this prediction, with a particular focus on the role of pharmacotherapy. Required dataset will be obtained from **Medical Cost Personal Datasets** (USA) which is nothing but the Health Insurance Data with 1338 observations and 7 features with attributes like Age, Sex, BMI, Children, Smoker, Region, Charges. The regression analysis is performed to determine the relationship among two or more exploratory variables with cause-effect relationships and to make predictions for the topic using the relationships [1]. If regression used one independent variable, then it is known as Simple Linear Regression Model, or else if it used more than two or more independent variables then it is known as Multilinear Regression Model [2]. Data Ingestion, Data Wrangling, Data Cleaning & Preprocessing, EDA will be done before Data Modeling depending upon the dataset. Basically, for any Machine Learning problem, multiple models will be built on the top of the dataset and based upon the accuracy of each model and the model with the highest accuracy will be considered and it will be evaluated on the test data. But to decide which model best fits the data, first analysis must be done on the dataset where **"charges" will be the Dependent Variable** which is target feature and **Exploratory Data Analysis (EDA) must be done to find the Independent Variables** by analyzing the correlation or plotting the relationship between each feature with respect to target feature. Hence, the econometric model that will be used for this research is **Multilinear Regression Model** as it has more than **two exploratory variables**. Evaluation of metrics helps to find how well the model generalizes the prediction.

This research aims to fill the gap by exploring the factors that affect price prediction. Pharmacotherapy provides important information for predicting cost increases in the total population. Moreover, its relative importance increases in combination with other features, including health care utilization. The study will examine how society might develop the necessary techniques and ways to minimize the cost of healthcare facilities and people to upcoming health catastrophes.

References:

[1] Kadam V. S., Kanhere S., Mahindrakar S. Regression techniques in machine learning &applications: a review. International Journal for Research in Applied Science and Engineering Technology. 2020;8(10):826–830. doi: 10.22214/ijraset.2020.32019.
[2] Dutta A. K., Ali Aljarallah N., Abirami T., et al. Optimal deep-learning-enabled intelligent decision support system for SARS-CoV-2 classification. Journal of Healthcare Engineering. 2022;2022:14. doi: 10.1155/2022/4130674.4130674
[3] Afifi W. A., El-Bagoury A. H. Optimal multiplicative generalized linear search plan for a discrete randomly located target. Information Sciences Letters . 2021;10(1):153–158.
[4] https://www.kaggle.com/datasets/mirichoi0218/insurance
[5] Suidan R. S., He W., Sun C. C., et al. Impact of body mass index and operative approach on surgical morbidity and costs in women with endometrial carcinoma and hyperplasia. Gynecologic Oncology . 2017;145(1):55–60. doi: 10.1016/j.ygyno.2017.01.025.