Medical Insurance Cost prediction

-By Subhadeep Jash

Abstract:

The project aims to create a machine learning model that predicts medical insurance costs based on demographic and health-related factors for more accurate pricing and risk assessment.

Introduction:

This report presents a predictive model for estimating medical insurance charges based on factors like smoking habits, BMI, age, number of children, sex, and region of residence. This helps stakeholders make informed decisions about resource allocation, insurance premiums, and healthcare planning.

Methodology:

1.Data Collection:

The project utilized a comprehensive dataset obtained from Kaggle, which includes information on medical insurance costs, BMI, smoking habits, number of children covered by the health insurance, sex, and regions of individuals.

2.Data Preprocessing:

The dataset was pre-processed to handle missing values and outliers, and those were imputed using appropriate techniques to ensure its completeness and suitability for analysis.

3.Feature Engineering:

- The study analyzed variables such as age, BMI, insurance coverage and sex to understand the relationship between these variables and medical insurance costs.
- Results showed that smokers have higher medical expenses, with treatment costs increasing with age.
- Smokers also spend more on smoking-related and age-related illnesses.
- Obese patients with a BMI over 30 are more likely to incur high medical expenses.

4.Encode Categorical Variables:

Categorical variables such as smoking habits, sex, and regions were converted into numerical representations using label encoding. This transformation allows the machine learning algorithms to process the data effectively and make accurate predictions.

5. Model Building and Model Evaluation:

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

- Various machine learning algorithms, including Linear Regression, Lasso Regression, Decision Trees, Random Forests, and Polynomial Regression, were employed to build predictive models.
- The models were trained on training data and evaluated on testing data using metrics like mean squared error, root mean squared error, and R2 value.
- The results indicated that smoking is highly important in predicting medical insurance costs, followed by BMI, age, number of children, region, and sex.
- The Random Forest Regressor and Polynomial Regression models were found to be effective in predicting medical insurance costs, with Polynomial Regression achieving an accuracy of 85%.

Conclusion:

- This analysis demonstrates that BMI, smoking habits, BMI, and age are important factors influencing medical insurance cost charges.
- Obese individuals, smokers, males, and older individuals tend to have higher medical cost charges. These findings emphasize the importance of promoting healthy lifestyles, including weight management, smoking cessation, and preventive care, to reduce healthcare costs.

Limitations:

It is important to note that the analysis is based on a specific dataset and may not be representative of the entire population. Further research is needed to consider factors such as pre-existing medical conditions, socioeconomic status, and geographical location.