Medical Insurance Cost Prediction using EDA and Regression

Domain: Healthcare Analytics

Tools Used: Python, Pandas, NumPy, Matplotlib, Seaborn, Jupyter Notebook

Objective:

To analyze an insurance dataset and build insights that help in predicting the medical insurance cost based on several factors such as age, BMI, smoking habits, number of children, sex, and region.

Problem Statement:

ABC Insurance, a health insurance agency, wants to understand the key drivers behind medical insurance premiums. Their dataset includes details about individuals, such as age, sex, BMI, number of children, smoking status, and region. The goal is to identify patterns that affect insurance costs and explore how various features contribute to higher or lower premiums.

Dataset Description:

- age: Age of the policyholder
- sex: Gender of the policyholder (male/female)
- bmi: Body Mass Index, indicates health status
- children: Number of children covered by insurance
- smoker: Indicates if the person is a smoker
- region: Region in the U.S. (northeast, northwest, southeast, southwest)
- charges: Medical insurance premium charged to the individual

Steps Followed:

1. Data Import and Library Setup:

Imported necessary libraries and loaded the dataset.

2. Data Inspection:

- Verified dataset shape and data types
- Checked for null or missing values
- Result: No missing values were found; data types were appropriate.

3. Exploratory Data Analysis (EDA):

- Univariate Analysis:

  Found an equal distribution of genders and more non-smokers than smokers.

- Bivariate Analysis:
  - Age and BMI correlated with charges.
  - Smokers had significantly higher charges.
  - Region showed slight variations.
- Correlation Heatmap:

  Age and smoking status had strong correlation with charges.

4. Key Observations:

- Smokers are charged nearly 3x-4x more than non-smokers.
- Premiums increase with age and high BMI.
- Children and sex have less impact on charges.
- No missing values; no imputation required.

5. Modeling (Optional if done):

A regression model was trained with good R^2 score, confirming reliable predictions.

Conclusion:

The analysis identified age, BMI, and smoking habits as key factors in premium costs. Insights can help ABC Insurance improve pricing strategies and promote healthy behaviors.