MEDICAL INSURANCE COST PROJECT

OVERVIEW

This dataset contains medical insurance cost information for 1338 individuals. It includes demographic and health-related variables such as age, sex, BMI, number of children, smoking status, and residential region in the US. The target variable is charges, which represents the medical insurance cost billed to the individual.  
This data set might have some missing values.

PROCESS

• **Data Exploration**:  
 - Used pandas to examine data distribution, missing values, and outliers.  
• **Handling Categorical Data**:  
 - Applied Label Encoding for categorical features.  
 - Considered One-Hot Encoding for future improvements.  
• **Correlation Analysis**:  
 - Generated a heatmap to identify features most strongly influencing house prices.  
• **Feature Visualization**:  
 - Plotted histograms to detect skewness, outliers, and inconsistencies.  
 - Applied transformations where necessary.  
• **Data Splitting:**  
 - Divided dataset into 80% training and 20% testing sets.  
• **Model Selection**:  
 - Started with Linear Regression as a baseline.  
 - Experimented with Random Forest and Gradient Boosting for improved performance.

TOOLS USED

• Python  
• Jupyter Notebook  
• pandas → Data exploration and preprocessing  
• numpy → Numerical operations.  
• matplotlib / seaborn → Data visualization (heatmaps, histograms, etc.).  
• scikit-learn (sklearn) → Machine learning models, Label Encoding, train-test split, cross-validation, evaluation metrics.