**Insurance Premium Prediction**

**Objective**

The objective of this project is to analyze the insurance dataset to identify key factors influencing medical insurance premiums and to build a predictive model that can estimate the cost of insurance based on various input features such as age, gender, BMI, smoking status, and region.

By understanding these factors, insurance companies can optimize premium pricing and enhance their business operations.

**Dataset Description**

The dataset (insurance.csv) contains the following columns:

* **age**: Age of the individual.
* **sex**: Gender (male or female).
* **bmi**: Body Mass Index, a measure of body fat based on height and weight.
* **children**: Number of dependents.
* **smoker**: Smoking status (yes or no).
* **region**: Geographic region (northeast, northwest, southeast, southwest).
* **charges**: The cost of medical insurance (target variable).

**Steps Followed**

1. **Importing Libraries and Loading Data**:
   * Libraries such as Pandas, NumPy, Matplotlib, and Seaborn were used for data analysis and visualization.
   * The dataset was loaded into a Pandas DataFrame for analysis.
2. **Exploratory Data Analysis (EDA)**:
   * **Summary Statistics**: Basic statistics of the dataset were analyzed to understand data distribution.
   * **Missing Values**: Missing values were checked and handled using forward fill.
   * **Correlation Analysis**: A heatmap was used to visualize relationships between numerical variables and the target column (charges).
   * **Categorical Analysis**: Count plots and box plots were used to examine the impact of categorical variables (e.g., smoker, region) on charges.
   * **Scatter Plots**: Relationships between numerical variables (e.g., age, BMI) and charges were visualized.
3. **Data Preprocessing**:
   * Categorical variables (e.g., sex, smoker, region) were encoded using LabelEncoder to convert them into numeric values.
   * The dataset was split into features (X) and the target (y), with further splitting into training and test sets using an 80-20 split.
4. **Model Training and Evaluation**:
   * Four regression models were trained and evaluated:
     + **Linear Regression**
     + **Decision Tree Regressor**
     + **Random Forest Regressor**
     + **Gradient Boosting Regressor**
   * 5-fold cross-validation was used to evaluate each model's performance based on:
     + Mean Absolute Error (MAE)
     + Mean Squared Error (MSE)
     + R² Score
   * The best-performing model was identified for further tuning and deployment.
5. **Model Selection**:
   * The model with the best combination of low error (MAE, MSE) and high R² score was selected.
   * Random Forest and Gradient Boosting models performed well due to their ability to handle complex relationships and interactions between features.
6. **Deployment**:
   * The best model was saved using joblib for potential deployment.

**Key Findings**

1. **Impact of Smoking**:
   * Smoking significantly increases insurance premiums. Smokers have much higher charges than non-smokers across all age groups.
2. **Age and Charges**:
   * Older individuals tend to have higher premiums due to increased health risks associated with aging.
3. **BMI and Charges**:
   * Higher BMI values are associated with increased insurance costs, particularly for individuals categorized as overweight or obese.
4. **Regional Differences**:
   * Premiums vary slightly across regions, but the variation is less significant compared to other factors.
5. **Gender**:
   * No significant difference was observed between male and female insurance charges.

**Conclusion**

This project successfully identified key factors influencing insurance premiums and built predictive models to estimate charges accurately. Insights gained can help insurance companies refine their pricing strategies and develop personalized insurance products.