Healthcare Insurance Analysis - Project Summary

Problem Statement:

The rising cost of healthcare is a major concern, making it crucial to predict future healthcare expenses and understand their key contributing factors. Insurance companies can leverage this analysis to make informed decisions regarding pricing and risk assessment.

Objective:

- Predict healthcare costs based on patient demographics and health-related attributes.
- Identify the most significant factors affecting healthcare expenses.
- Analyze the interdependencies between these factors.
- Utilize various machine learning models to optimize cost prediction.

Dataset & Features:

The dataset includes attributes such as:

- **Age** (Numerical)
- **BMI (Body Mass Index)** (Numerical)
- **Number of Children** (Numerical)
- **Smoker Status** (Categorical)
- **Region** (Categorical)
- **Sex** (Categorical)
- **Medical Charges** (Target Variable)

Methodology:

1. **Data Preprocessing:**
    - Handle missing values (if any).
    - Convert categorical variables into numerical using one-hot encoding.
    - Scale numerical features using StandardScaler.
    - Split data into training and testing sets.
2. **Exploratory Data Analysis (EDA):**
    - Visualized distributions and relationships between features.
    - Analyzed the impact of smoking, BMI, and age on healthcare costs.
    - Used correlation heatmaps and boxplots to gain insights.
3. **Model Selection & Training:**
    - Implemented various regression models:
        - Linear Regression
        - Decision Tree Regression

        - Random Forest Regression
        - Gradient Boosting Regression
        - XGBoost
    - Evaluated models using MAE, RMSE, and R^2 score.
4. **Feature Importance Analysis:**
    - Identified the most influential factors affecting healthcare costs.
    - Smoking status, BMI, and age were found to be the top predictors.

Results & Insights:

- **XGBoost and Random Forest** performed the best in terms of RMSE and R^2 score.
- **Smoking status** had the highest impact on healthcare costs, followed by BMI and age.
- **Insurance companies** can use these insights to adjust premium calculations based on high-risk factors.

Conclusion & Recommendations:

- Healthcare costs are significantly influenced by lifestyle choices (e.g., smoking).
- Insurance providers can use this model to predict costs and offer personalized plans.
- Further optimization through hyperparameter tuning and ensemble methods could improve accuracy.