PROJECT DOCUMENTATION

1. Title of the project:

Medical Insurance Cost Prediction Using Machine Learning

2. Objective:

The main objective of this project is to build a machine learning model that accurately predicts the medical insurance cost of an individual. The model aims to support insurance companies and individuals in estimating potential insurance costs with higher precision.

3. Dataset Used:

Source: Kaggle

URL: https://www.kaggle.com/datasets/mirichoi0218/insurance

Attributes Included:

- age: Age of the individual
- sex: Gender of the individual
- bmi: Body Mass Index
- children: Number of children covered by health insurance
- smoker: Smoking status
- region: Residential area (in the U.S.)
- charges: Medical insurance cost (target variable)

4. Model Chosen:

Multiple regression algorithms were implemented and compared to identify the most effective model for this prediction task.

Models Evaluated:

a. Linear Regression
b. Random Forest Regressor
c. **Gradient Boosting Regressor** (Best Performing)

Final Model Selected: Gradient Boosting Regressor
This model outperformed others in terms of predictive accuracy and generalization capability.

5. **Performance Metrics:**
   The following regression metrics were used to assess model performance:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score (Coefficient of Determination)

Performance metrics	**Linear Regression**	**Random Forest Regressor**	**Gradient Boosting Regressor**
MAE	4267.213	2717.018	2393.6435
MSE	38337035.486	24288047.435	19869359.123
RMSE	6191.690	4928.290	4457.505
R^2 Score	0.7447	0.8382	0.8676

6. **Challenges & Learnings:**

Challenges Faced:

- Managing categorical variables through encoding techniques
- Avoiding overfitting while using high-capacity models like Random Forest and Gradient Boosting

Learnings:

- Importance of feature engineering and preprocessing (like one-hot encoding)
- Evaluating multiple regression models to find the best fit
- Understanding model performance through various error metrics
- Simpler models like Linear Regression were easier to interpret but lacked accuracy, whereas ensemble methods performed better.