Mini Project: Medical Insurance Cost Prediction

Objective:

Can you accurately predict medical insurance costs based on patient data? In this project, you'll build various regression models and compare their performance to recommend the best one to the stakeholders.

Dataset:

Use the below dataset:

Health_Insurance.csv

Below is the Data dictionary:

Column	Description
age	Age of the primary beneficiary
sex	Gender of the insurance policyholder (female/male)
bmi	Body Mass Index (kg/m²)
children	Number of dependents covered under the insurance
smoker	Smoking status (yes/no)
region	Residential area in the US (northeast, southeast, southwest, northwest)
charges	Individual medical costs billed by health insurance

Tasks:

1. Import and Load the Data

- Import required libraries (pandas, numpy, matplotlib, seaborn, sklearn, etc.)
- Load the dataset and explore the structure using .head(), .info() and .describe()

2. Exploratory Data Analysis (EDA)

- Visualize the distribution of each feature
- Understand correlations (especially with the target variable)

- Check for variables distributions.
- Summarize insights from EDA

3. Missing Values & Outlier Treatment

- Check for missing values and treat them if any
- Check if there are anu outliers.

4. Feature Engineering & Preprocessing

- Encode categorical variables (sex, smoker, region)
- Feature scaling for numerical values (StandardScaler / MinMaxScaler)
- Check for skewness and treat it if required.

5. Model Building: Try Multiple Regressors

Use all the regression-based models to train and test the data:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- SVR
- KNN
- Ensemble Learning methods

6. Model Evaluation & Overfitting Check

- Use metrics:
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - R^2 Score
    - Adjusted R^2 Score
- Compare performance on both datasets (Training and testing) to detect overfitting

7. Hyperparameter Tuning

- Use GridSearchCV or RandomizedSearchCV to optimize the best-performing models
- Document best parameters and improvement in performance

8. Model Comparison Table

Create a comparison table like below:

Model	**Train RMSE**	**Test RMSE**	**Train R^2**	**Test R^2**	**Overfitting (Y/N)**
Linear Regression					
Decision Tree					
Random Forest					
Gradient Boosting					
Best Model (e.g. XGB)					

- Create a simple UI where user inputs age, sex, BMI, smoker, etc. and receives a predicted insurance charge