Objective:

To develop a machine learning model to predict the insurance charges based on parameters such as age, sex, number of children and smoking status.

Dataset Overview:

Basic Information:

- Total Rows: 1,338 entries.
- Total Columns: 6 columns.

Columns in the Dataset:

- Age: Age of the primary beneficiary (integer).
- Sex: Gender of the beneficiary (categorical: male/female).
- BMI: Body mass index (float).
- Children: Number of children/dependents (integer).
- Smoker: Smoking status (categorical: yes/no).
- Charges: Individual medical costs billed by health insurance (float).

Data Preprocessing

- To prepare the dataset for modeling, we need to preprocess it:

1. **Handling Categorical Variables:**
    - Convert the `sex` and `smoker` columns into numerical format:
    - `sex`: Map 'male' to 0 and 'female' to 1.
    - `smoker`: Map 'no' to 0 and 'yes' to 1.
2. **Splitting the Dataset:**
    - Split the dataset into training and testing sets to evaluate the model's performance.

Model	R2 Score
Simple Linear Regression	0.79
Multiple Linear Regression	0.78
Support Vector Machine - Linear	-0.11
Support Vector Machine - RBF	-0.08
Support Vector Machine - Poly	-0.06
Support Vector Machine - Sigmoid	-0.08
Decision Tree – Squared Error / Best	0.688
Decision Tree– Friedman Mse/ Best	0.685
Decision Tree– Absolute Error/ Best	0.668
Decision Tree– Poisson/ Best	0.735
Decision Tree – Squared Error / Random	0.699
Decision Tree– Friedman Mse/ Random	0.71
Decision Tree– Absolute Error/ Random	0.73
Decision Tree– Poisson/ Random	0.70
Random Forest – Squared Error	0.849
Random Forest – Absolute Error	0.849

Random Forest – Friedman MSE	0.849
Random Forest – Poisson	0.849

Final Model Selection

The Chosen Model for this scenario is Random Forest Regressor.

Justification:

The Random Forest achieved the highest R^2 score of **0.849**, indicating that it explains approximately 85% of the variance in insurance charges. This makes it the most accurate model among those tested.