Problem Statement:

The client wants to **predict medical insurance charges** based on parameters, age, gender, BMI, smoking status, number of children,

This is a **Supervised Machine Learning problem**, where:

- **Input (Features):** Demographic and lifestyle attributes (e.g., age, sex, BMI, number of children, smoking status).
- **Output (Target): insurance charges** (continuous numeric value).

Since the target variable is continuous, this problem falls under **Regression**.

Dataset:

Row: 1339, Column: 6

Independent columns: 5

age	sex	bmi	children	smoker

Dependent column: 1

Charges

Stage 1: Domain Selection -> ***Machine Learning***

Reason: Dataset is structured numeric data

Stage 2: Learning Selection -> ***Supervised Learning***

Reason: Dataset includes both inputs and outputs, so this would be Supervised Learning

Stage 3: ***Regression***

Reason: Output value is continuous numeric values so this would be Regression

Output is numerical

Pre-Processing method

Features like sex and smoker are ordinal data -> I am going to use “one hot encoding” to keep it simple. Though it is used for nominal data, I see that we can still use it for ordinal as well.

R_Score

Multiple Linear Regression: R_score: 0.79

SVM: -0.08

Decision Tree:

Criterion	Splitter	Max_features	R_score
Squared_error	Random	Log2	0.66
friedman_mse	Random	Log2	0.70

absolute_error	Random	Log2	0.68
poisson	Random	Log2	0.74
Squared_error	best	Log2	0.66

Random Forest:

Criterion	N_estimators	Random_state	R_score
Squared_error	50	0	0.85
Squared_error	100	0	0.85
absolute_error	50	0	0.85
friedman_mse	50	0	0.85

Conclusion:

Random forest is the best model for this dataset