Problem Statement or Requirement:

A client's requirement is, he wants to predict the insurance charges based on several parameters. The Client has provided the dataset of the same.

As a data scientist, you must develop a model which will predict the insurance charges.

1. **Identify your problem statement**

   The objective is to develop a machine learning model to predict **insurance charges** based on several factors such as age, gender, BMI, number of children, and smoking status. The client has provided a dataset, and we need to preprocess it, build multiple models, and determine the best-performing one.

2. **Tell basic info about the dataset (Total number of rows ,columns)**

   The dataset contains:

   - Total Rows = 1338
   - Total columns = 6

   Features:

- `age` (int) → Age of the person
- `sex` (object) → Gender (Male/Female)
- `bmi` (float) → Body Mass Index
- `children` (int) → Number of children
- `smoker` (object) → Smoker status (Yes/No)

Target Variable:

- `charges` (float) → Insurance cost

Preprocessing Steps

Since the dataset contains categorical variables, we need to encode them into numerical values before applying regression models.

1. **Convert categorical variables into numeric values:**
   - `sex`:
     - Male → **1**
     - Female → **0**
   - `smoker`:

- Smoker → **1**
- Non-smoker → **0**

2. **Train-test split :**

Split the dataset into 80% training and 20% testing.

To find the following the machine learning regression method using in r2 value

1.Multiple linear regression(R2 value) = 0.1497

2. Support vector machine:

S.NO	Hyper parameter	RBF(Non linear)	Poly (r_value)	Sigmoid (r_value)
1.	C 10	- 0.0817	- 0.0756	- 0.0819
2.	C 100	- 0.0111	- 0.0819	- 0.1100
3.	C 500	- 0.1192	- 0.0776	- 0.2259
4.	C 1000	- 0.1221	- 0.0774	- 1.0145
5.	C 2000	- 0.1252	- 0.0768	- 5.1526
6.	C 3000	- 0.1234	- 0.0765	- 12.1998

The SVM Regression use R2 Value (Sigmoid and hyper parameter (C 3000) = - 12.1998

3. Decision tree :

S.NO	Criterion	Splitter	R2 value
1.	fried_man mse	best	- 0.5530
2.	fried_man mse	random	- 0.5390
3.	squared_error	best	- 0.5190
4.	squared_error	random	- 0.6842
5.	absolute_error	best	- 0.5787
6.	absolute_error	random	- 0.5811
7.	poisson	best	- 0.6543
8.	poisson	random	- 0.7241

The Decision Tree Regression use R2 value (poisson, random) = - 0.7241

4. Random Forest:

S.NO	Criterion	n_estimaters	R2 value
1.	squared_error	10	- 0.0039
2.	squared_error	50	0.0202
3.	squared-error	100	0.0373
4.	absolute_error	10	0.0444
5.	absolute_error	50	0.0259
6.	absolute_error	100	0.0284
7.	friedman_mse	10	- 0.0163
8.	friedman_mse	50	0.0310
9.	friedman_mse	100	0.0429
10.	poisson	10	- 0.0039
11.	poisson	50	0.0439
12.	poisson	100	0.0234

The Random Forest Regression use R2 value (absolute_error, n_estimaters 10) = 0.0444

Justification for Choosing Multiple Linear Regression

1. **Higher R^2 Score**:
    - The **Multiple Linear Regression** achieved the highest **R^2 score** compared to other models, indicating better predictive performance.
    - Example comparison of R^2 scores (values may vary based on execution):

Model	R2 value
1. MLR	0.1497
2. SVMR	-12.1998
3. Decision Tree	-0.7241
4. Random Forest	0.0444