Assignment Regression Algorithm

1.) Identify your problem statement

The user aims to predict insurance charges through an AI application. They possess a dataset featuring fields such as age, sex, BMI, children, smoker status, and charges. Below are three stages of problem identification,

Stage 1: Domain Selection -> Machine Learning
The inputs consist primarily of numerical

Stage 2: Learning Type Selection -> Supervised Learning
Labeled data is available, pairing features with target charges.

Stage 3: Model Type Selection -> Regression
The target variable (charges) is continuous and numerical.

2.) Tell basic info about the dataset

(Total number of rows, columns)

In the provided dataset there are **6 columns** which are spitted as Input and output columns as below,

- **Input Columns (5)** - Age (Numerical), Sex (Category), BMI (Numerical), Children (Numerical), Smoker (Category)
- **Output Column (1)** - Charges (Numerical)

The provided dataset contains **1338 rows** of data

3.) Mention the pre-processing method if you're doing any

(like converting string to number – nominal data)

Preprocessing are required for Column 'Sex' and 'Smoker'. These columns contains categorical values and these values are **Nominal data.** So, **One-Hot Encoding** is done to process these data

4.) Develop a good model with r2_score.

You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

Models are developed using following Algorithms,
01. Multi Linear Regression
02. Support Vector Machine (SVM)
03. Decision Tree
04. Random Forest

5.) All the research values

(r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

I. Multi Linear Regression

R2 Score = 0.7894790349867009

II. Support Vector Machine (SVM)

S.No	Kernel	R2_Score
1	linear	-0.111661287196084
2	poly	-0.0642925840210553
3	rbf	-0.0884273277691388
4	sigmoid	-0.0899412170256757

III. Decision Tree

S.No	Criterion	Splitter	Max_Features	R2_Score
1	squared_error	best	sqrt	0.773436660329055
2	squared_error	best	log2	0.759046446613415
3	squared_error	random	sqrt	0.684721098906109
4	squared_error	random	log2	0.676295192668561
5	friedman_mse	best	sqrt	0.718323162971904
6	friedman_mse	best	log2	0.780149876596692
7	friedman_mse	random	sqrt	0.629822812763356
8	friedman_mse	random	log2	0.71599557550467
9	absolute_error	best	sqrt	0.652457088160599

S.No	Criterion	Splitter	Max_Features	R2_Score
10	absolute_error	best	log2	0.723621364178124
11	absolute_error	random	sqrt	0.586992948234075
12	absolute_error	random	log2	0.496084309447596
13	poisson	best	sqrt	0.75831634265849
14	poisson	best	log2	0.610263482201543
15	poisson	random	sqrt	0.605843278516423
16	poisson	random	log2	0.68550921190886
17	squared_error	best	None	0.696441630216836
18	squared_error	random	None	0.725863068634242
19	friedman_mse	best	None	0.687224340467711
20	friedman_mse	random	None	0.676148963722081
21	absolute_error	best	None	0.662112309722301
22	absolute_error	random	None	0.722937900836069
23	poisson	best	None	0.734022846982422
24	poisson	random	None	0.71740457019699

IV. Random Forest

S.No	Criterion	Max_Features	R2_Score
1	squared_error	sqrt	0.870942133821651
2	squared_error	log2	0.871360447283668
3	friedman_mse	sqrt	0.873203999983993
4	friedman_mse	log2	0.872616778272248
5	absolute_error	sqrt	0.870913889625852
6	absolute_error	log2	0.872179102121558
7	poisson	sqrt	0.869538950394453
8	poisson	log2	0.868527683094692
9	squared_error	None	0.85201334808207
10	friedman_mse	None	0.85071103951953
11	absolute_error	None	0.853582015175819

S.No	Criterion	Max_Features	R2_Score
12	poisson	None	0.851854886135245

6.) Mention your final model, justify why u have chosen the same

Best model is **Random Forest** with Hyper Tuning Parameters **'Criterion = friedman_mse'** and **'Max_feature=sqrt'**

Because it gives the best R2 score of **0.873203999983993** for the created model.