Regression Algorithm - Assignment

1. **Identification of problem statement**

   The company wants to predict the insurance charges based on the several parameters given in the sheet.

2. **Info about the dataset**

   The dataset for this problem statement contains age of the customer, sex of the customer ,their bmi, how many children they have and whether they are smokers or not. Based on these details, the company wants to predict the insurance charges.

3. **The pre-processing method**

   The dataset has nominal data for sex and smoker columns. So to convert the nominal data, “one-hot encoding” technique has been introduced after importing the dataset using panda library. Here, the **get_dummies** function is used to convert the nominal data.

Algorithm R2 Value:

1. **Multiple Linear Regression -** 0.73
2. **Support Vector Machine**

Kernel	Gamma	C value = 10	C value = 100	C value = 500	C value =1000	C value =1500	C value =2000	C value =2500	C value =3000
linear	scale	0.46	0.62	0.76	0.76	0.74	0.74	0.74	0.74
linear	auto	0.46	0.62	0.76	0.76	0.74	0.74	0.74	0.74
poly	scale	0.038	0.61	0.82	0.85	0.85	0.86	0.85	0.85
poly	auto	0.038	0.61	0.82	0.85	0.85	0.86	0.85	0.85
rbf	scale	-0.032	0.32	0.66	0.81	0.84	0.85	0.86	0.86
rbf	auto	-0.032	0.32	0.66	0.81	0.84	0.85	0.86	0.86
sigmoid	scale	0.039	0.52	0.44	0.28	-0.067	-0.59	-1.28	-2.12

sigmoid	auto	0.039	0.52	0.44	0.28	-0.06	-0.59	-1.28	-2.12

3. Decision Tree:

S.No	Criterion	Splitter	max_features	R2 Value
1	squared_error	best	sqrt	0.67
2	squared_error	best	log2	0.68
3	squared_error	best	None	0.68
4	squared_error	random	sqrt	0.64
5	squared_error	random	log2	0.70
6	squared_error	random	None	0.73
7	friedman_mse	best	sqrt	0.76
8	friedman_mse	best	log2	0.64
9	friedman_mse	best	None	0.69
10	friedman_mse	random	sqrt	0.63
11	friedman_mse	random	log2	0.67
12	friedman_mse	random	None	0.72
13	absolute_error	best	sqrt	0.74
14	absolute_error	best	log2	0.74
15	absolute_error	best	None	0.68
16	absolute_error	random	sqrt	0.64
17	absolute_error	random	log2	0.73
18	absolute_error	random	None	0.72

19	poisson	best	sqrt	0.70
20	poisson	best	log2	0.74
23	poisson	best	None	0.72
21	poisson	random	sqrt	0.66
22	poisson	random	log2	0.65
24	poisson	random	None	0.72

4. Random Forest

S.No	n_estimators	Criterion	max_features	R2 Value
1	50	squared_error	sqrt	0.86
2	100	squared_error	sqrt	0.87
3	50	squared_error	log2	0.86
4	100	squared_error	log2	0.87
5	50	squared_error	None	0.84
6	100	squared_error	None	0.85
7	50	friedman_mse	sqrt	0.87
8	100	friedman_mse	sqrt	0.87
9	50	friedman_mse	log2	0.87
10	100	friedman_mse	log2	0.87
11	50	friedman_mse	None	0.85
12	100	friedman_mse	None	0.85
13	50	absolute_error	sqrt	0.87

14	100	absolute_error	sqrt	0.87
15	50	absolute_error	log2	0.87
16	100	absolute_error	log2	0.87
17	50	absolute_error	None	0.85
18	100	absolute_error	None	0.85
19	50	poisson	sqrt	0.86
20	100	poisson	sqrt	0.86
21	50	poisson	log2	0.86
22	100	poisson	log2	0.86
23	50	poisson	None	0.84
24	100	poisson	None	0.85

Final Model:

The final model to predict the insurance charges would be the **"Random Forest"** Algorithm.
The parameter for this algorithm is

n_estimators	Criterion	max_features	R2 Value
100	squared_error	sqrt	0.87

When compared to the other algorithm models, the r2 value is high - 87%