PROBLEM STATEMENT OR REQUIREMENT

A client's requirement is, he wants to predict the insurance charges based on the 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.

STEPS TAKEN

1. Identifying the problem statement

The client has data regarding to insurance, and wants to predict their insurance charge.

2. Basic info about the dataset (Total number of rows, columns)

The client has shared the dataset with 1338 rows of data containing columns: age, sex, bmi, children, smoker and charges.

3. The pre-processing method (like converting string to number – nominal data)

There is no missing data. There is seemingly no wrong data. The categorical data belonging to 'sex' and 'smoker' is converted into numerical values.

4. Developing a good model with good r2_score.

1. **Linear Regression: 0.6089** (Seems like 'smoker' feature is significantly contributing to the insurance)
2. **Multiple Linear Regression R^2 value = 0.7535**
3. **Support Vector Machine**

#	Hyper parameter(c=)	R^2 value(kernel=)			
		Linear	RBF	Poly	Sigmoid
1	10	-0.0836	-0.1854	-0.1880	-0.1873
2	100	0.4916	-0.2053	-0.1921	-0.2125
3	500	0.5760	-0.2071	-0.1715	-0.4971
4	1000	0.5895	-0.2021	-0.1433	-1.5452
5	2000	0.6171	-0.1928	-0.0895	-5.0716
6	3000	**0.6439**	-0.1825	-0.0376	-10.941

4. **Decision Tree**

#	Criterion	Max Features	Splitter	R^2 value
1	squared_error	None	best	0.7539
2			random	0.7325
3		Sqrt	best	0.6939
4			random	0.7233
5		Log2	best	0.7450
6			random	0.5041
7	absolute_error	None	best	0.7535
8			random	**0.7671**
9		Sqrt	best	0.7177
10			random	0.6736
11		Log2	best	0.7059
12			random	0.6975
13	friedman_mse	None	best	0.7506
14			random	0.7389
15		Sqrt	best	0.6815

16			random	0.6712
17		Log2	best	0.7168
18			random	0.7088

5. Random Forest

#	criterion	max_features	n_estimators	R^2 value
1	poisson	log2	50	0.8417
2			75	0.8381
3			100	0.8412
4		sqrt	50	0.8414
5			75	0.8415
6			100	**0.8450**
7	squared_error	log2	50	0.8412
8			75	0.8416
9			100	0.8391
10		sqrt	50	0.8399
11			75	0.8409
12			100	0.8433
13	absolute_error	log2	50	0.8390
14			75	0.8386
15			100	0.8404
16		sqrt	50	0.8379
17			75	0.8384
18			100	0.8401
19	friedman_mse	log2	50	0.8377
20			75	0.8392
21			100	0.8429
22		sqrt	50	0.8421
23			75	0.8397
24			100	0.8418

5. Justifying why I have chosen the Random Forest Model

The Random Forest model has provided with the so far best value '0.8450', when the hyper parameter which are criterion, max_features, n_estimators are tuned to specific values.

Linear Regression	Multiple Linear Regression	Support Vector Machine	Decision Tree	Random Forest
0.6089	**0.7535**	**0.6439**	**0.7671**	**0.8450**