REGRESSION ASSIGNMENT

Requirement or Problem statement:

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, i develop a model which will predict the insurance charges.

1.Identification of statement

- In this case, i am clearly identify the problem statement.
- I am going to predict the insurance charges based on the several parameters given by the client.
- As a data scientist, i am going create good model and predict the insurance charges

2.Basics information about the dataset

Input data					Output data
Age	**Sex**	**Bmi**	**children**	**smoker**	**Insurance charges**
19	Male	27.9	0	yes	16884.92
18	female	33.77	1	no	1725.552
17	male	33	2	no	4449.45
Sample data for analysis purpose					

- In the dataset, i have 6 row and 1339 columns.
- The categories are Age, Sex, Bmi, Children, Smoker and Insurance charges.

3. Pre-Processing method

- In the dataset, two categories are categorical data. So i need to change the data as a meaningful number.
- In this case, i am going to use nominal data to expanse the column and change the values to meaningful number.
- In smoker category- In this case, yes or no response are considered nominal data. yes or no is not express any form of rank or order.
- So i am going to use nominal data to change the smoke category for meaning data (sting to number)

4. R Squared value comparison

1. Multiple linear regression = 0.789

2. Support Vector Machine

s.no	parameter	R2 score
1	rbf	-0.081
2	Rbf(c=1000)	0.0823
3	poly	-0.062
4	Poly(C=1000)	0.860
5	sigmoid	-0.072
6	Sigmoid(C=1000)	0.143

3.decision tree

s.no	criterion	Splitter	R2 score
1	Squared_error	best	0.709
2	Squared_error	Random	0.691
3	Friedman_mse	best	0.692
4	Friedman_mse	Random	0.689
5	Absolute_error	best	0.678
6	Absolute_error	Random	0.705
7	poisson	best	0.729
8	poisson	Random	0.682

4. Random Forest = 0.855 (n_estimators=100)

5. final model

My final model is Random Forest. Because comparing all the r2_score value with respective algorithm, Random Forest r2_score value is good. So finally select the Random Forest algorithm is my final model for deployment process.