Insurance Prediction Details

1.) Identify your problem statement

Using the Machine Learning-> Supervised learning-> Regression

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

Number of rows: 1338
Number of columns: 6

```
In [2]: #To slice the data set like row cunt,column count
        row_count=len(data.index)
        print("Number of rows:",row_count)
        column_count= data.shape[1]
        print("Number of columns:",column_count)

        Number of rows: 1338
        Number of columns: 6
```

3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

Given data set column Sex & Smoker having the nominal data so converting string to number using the pandas library get_dummies function.

```
In [38]: #Update the categorical column data using the one hot encode
         data= pd.get_dummies(data,drop_first=True)
         data
```

Out[38]:

	age	bmi	children	charges	sex_male	smoker_yes
0	19	27.900	0	16884.92400	0	1
1	18	33.770	1	1725.55230	1	0
2	28	33.000	3	4449.46200	1	0
3	33	22.705	0	21984.47061	1	0
4	32	28.880	0	3866.85520	1	0

4.) Develop a good model with r2_score.You can use any machine learning algorithm

Algorithm Name	R2 Value
MLR	0.78
SVM	0.87
DT	0.76
RF	0.85

Comparing all the algorithm the high r2 value is for SVM r2= 0.87

5.) All the research values (r2_score of the models) below:

a) MLR r2 vale for the dataset -> **0.78**

b) SVM (Support Vector Machine)

Sno	Hyper Parameter C value	linear r value	rbf (default) r value	poly	sigmoid
1	10	0.46	-0.032	0.038	0.039
2	100	0.62	0.32	0.61	0.52
3	1000	0.76	0.81	0.85	0.28
4	2000	0.74	0.85	0.86	-0.59
5	3000	0.74	0.86	0.85	-2.13
6	4000	0.74	0.87	0.86	-5.51

Conclusion:

The r2 value for the hyper parameter of C=4000 rbf is 0.87

c) Decision Tree

SNo	criterion	max_features	splitter	R2 value
1	squared_error (default)	None(default)	Base(default)	0.69
	squared_error (default)	Sqrt	base	0.74
2	squared_error (default)	Log2	base	0.73
3	squared_error (default)	auto	base	0.70
4	squared_error (default)	Auto	random	0.69
5	squared_error (default)	none	random	0.69
6	squared_error (default)	sqrt	random	0.66
7	squared_error (default)	Log2	random	0.60
8	mae	Auto	Base	0.67
9	Mae	Log2	Base	0.57
10	Mae	sqrt	Base	0.75
11	Mae	none	Base	0.66

12	Mae	None	random	0.68
13	Mae	Log2	Random	0.71
14	Mae	Sqrt	Random	0.74
15	Mae	Auto	Randam	0.76
16	friedman_mse	None	random	0.68
17	friedman_mse	Auto	random	0.644
18	friedman_mse	Sqrt	random	0.611
19	friedman_mse	Log2	random	0.55
20	friedman_mse	Auto	base	0.68
21	friedman_mse	None	base	0.69
22	friedman_mse	Log2	base	0.74
23	friedman_mse	Sqrt	base	0.64

Conclusion:

The r2 value of the decision tree regression use (mae,random,auto) -> 0.76

d) Random Forest

Sno	n_estimators	criterion	R2value
1	100	squared_error	0.85
2	50	squared_error	0.84
3	200	squared_error	0.85
4	100	friedman_mse	0.84
5	50	friedman_mse	0.85
6	200	friedman_mse	0.85

Conclusion:

The r2 value of the Random Forest use (n_estimators =100, squared_error**) -> 0.85**