To Find the Best Regression Algorithm for the given problem statement

To Predict the Insurance Charges based on the given parameters from the Dataset

1) PROBLEM STAEMENT IDENTIFICATION:

STEP-1: Domain Selection- Machine Learning

STEP-2: Learning Selection- Supervised Learning

STEP-3: Regression

2) Basic information about the Dataset:

Total Number of Rows in the given Dataset=1338

Total Number of Columns in the given Dataset=6

3) Pre-Processing Method:

In the given dataset two columns Gender and Smoker consists of categorical value, hence converted to Nominal values (True or False) by **One Hot Encoding** method using **get_dummies** function .

4) To find the best model in Machine Learning Regression Algorithm using r2 value:

1.MULTIPLE LINEAR REGRESSION (R^2 Value)=0.7894

2.SUPPORT VECTOR MACHINE:

S.NO	HYPER PARAMETER	LINEAR (r Value)	RBF (NON LINEAR) (r Value)	POLY (r Value)	SIGMOID (r Value)
1	C10	0.4624	-0.0322	0.0387	0.0393
2	C100	0.6288	0.3200	0.6179	0.5276
3	C500	0.7631	0.6642	0.8263	0.4446
4	C1000	0.7649	0.8102	0.8566	0.2874
5	C2000	0.7440	0.8547	0.8605	-0.5939
6	C3000	0.7414	0.8663	0.8598	-2.2144

The SVM Regression use R^2 Value(Non-Linear “rbf” and Hyper Parameter C=3000)=0.8663

3.DECISION TREE:

S.NO	CRITERION	MAX FEATURES	SPLITTER	R VALUE
1	Squared_error	None	Best	0.7058
2	Squared_error	None	Random	0.7482
3	Squared_error	Sqrt	Best	0.7300
4	Squared_error	Sqrt	Random	0.5946
5	Squared_error	Log2	Best	0.5061
6	Squared_error	Log2	Random	0.6598
7	Absolute_error	None	Best	0.6904
8	Absolute_error	None	Random	0.7491
9	Absolute_error	Sqrt	Best	0.6587
10	Absolute_error	Sqrt	Random	0.7139
11	Absolute_error	Log2	Best	0.7203
12	Absolute_error	Log2	Random	0.7290
13	Friedman_mse	None	Best	0.6874
14	Friedman_mse	None	Random	0.6640
15	Friedman_mse	Sqrt	Best	0.7217
16	Friedman_mse	Sqrt	Random	0.7315
17	Friedman_mse	Log2	Best	0.7586
18	Friedman_mse	Log2	Random	0.6654
19	Poisson	None	Best	0.7158
20	Poisson	None	Random	0.7356
21	Poisson	Sqrt	Best	0.6999
22	Poisson	Sqrt	Random	0.6718
23	Poisson	Log2	Best	0.7292
24	Poisson	Log2	Random	0.7186

The Decision Tree Regression use R^2Value (Criterion=Friedman_mse, Splitter=Best, Max_features=log2)=0.7586

4.Random Forest: n_estimators=100

S.NO	CRITERION	MAX FEATURES	R VALUE
1	Squared_error	None	0.8538
2	Squared_error	Sqrt	0.8710
3	Squared_error	Log2	0.8710
4	Absolute_error	None	0.8520
5	Absolute_error	Sqrt	0.8710
6	Absolute_error	Log2	0.8710
7	Friedman_mse	None	0.8540
8	Friedman_mse	Sqrt	0.8710
9	Friedman_mse	Log2	0.8710
10	Poisson	None	0.8526
11	Poisson	Sqrt	0.8680
12	Poisson	Log2	0.8680

The Random Forest Regression use R^2Value(Criterion=Squared_error,**Absolute_error,Friedman_mse** and Max_features=**Sqrt,log2**)=0.8710

Conclusion:

Hence for the given problem statement of predicting Insurance charges from the given dataset **Random Forest algorithm** works better than other algorithms with $\boldsymbol{R^2}$**Value=87%** which is higher than other R^2Value of Multiple Linear Regression, Support Vector Machine and Decision Tree.