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.

Problem statement Identification:

- From the Client's requirement it is clear that what are all the input we are going to use and expected output.
- To satisfy this requirement by we can use ML based Model with supervised learning using regression algorithm, since the input and output data are well defined.

Dataset info:

Input data-> Age,sex,bmi,children,smoker
Value to be predicted -> charges

Pre-processing:

Since we have two categorical column Sex and smoker, we are convert it into numerical data using One Hot encoding method.

Model Selection:

R2_value for MLR: 0.7894790349867009

R2_Value for SVM:0.866339395

Column1	kernel	C	gamma	r2_score
0	rbf	10	auto	-0.032273294
1	rbf	10	scale	-0.032273294
2	rbf	100	auto	0.320031783
3	rbf	100	scale	0.320031783
4	rbf	1000	auto	0.810206485
5	rbf	1000	scale	0.810206485
6	rbf	2000	auto	0.854776643
7	rbf	2000	scale	0.854776643
8	rbf	3000	auto	0.866339395
9	rbf	3000	scale	0.866339395
10	poly	10	auto	0.038716223
11	poly	10	scale	0.038716223
12	poly	100	auto	0.617956962
13	poly	100	scale	0.617956962
14	poly	1000	auto	0.856648768
15	poly	1000	scale	0.856648768
16	poly	2000	auto	0.860557926
17	poly	2000	scale	0.860557926

18	poly	3000	auto	0.859893008
19	poly	3000	scale	0.859893008
20	sigmoid	10	auto	0.039307144
21	sigmoid	10	scale	0.039307144
22	sigmoid	100	auto	0.527610355
23	sigmoid	100	scale	0.527610355
24	sigmoid	1000	auto	0.287470695
25	sigmoid	1000	scale	0.287470695
26	sigmoid	2000	auto	-0.593950973
27	sigmoid	2000	scale	-0.593950973
28	sigmoid	3000	auto	-2.124419479
29	sigmoid	3000	scale	-2.124419479
30	linear	10	auto	0.462468414
31	linear	10	scale	0.462468414
32	linear	100	auto	0.628879286
33	linear	100	scale	0.628879286
34	linear	1000	auto	0.764931174
35	linear	1000	scale	0.764931174
36	linear	2000	auto	0.744041831
37	linear	2000	scale	0.744041831
38	linear	3000	auto	0.74142366
39	linear	3000	scale	0.74142366

R2_Value for Decision Tree:0.792153673

Column1	criterion	splitter	max_features	r2_scor
0	mse	random	sqrt	0.675378249
1	mse	random	auto	0.678470739
2	mse	random	log2	0.684439414
3	mse	best	sqrt	0.675292113
4	mse	best	auto	0.69675031
5	mse	best	log2	0.686645289
6	mae	random	sqrt	0.706369085
7	mae	random	auto	0.757211652
8	mae	random	log2	0.792153673
9	mae	best	sqrt	0.658955119
10	mae	best	auto	0.664302723
11	mae	best	log2	0.733328657
12	friedman_mse	random	sqrt	0.698902634
13	friedman_mse	random	auto	0.722049604
14	friedman_mse	random	log2	0.59506613
15	friedman_mse	best	sqrt	0.618744454
16	friedman_mse	best	auto	0.710153088
17	friedman_mse	best	log2	0.701974389

R2_Value for Random Forest: 0.874372

Column1	criterion	estimators	max_features	r2_scor
0	mse	10	sqrt	0.849643
1	mse	10	auto	0.840009
2	mse	10	log2	0.85926
3	mse	100	sqrt	0.869864
4	mse	100	auto	0.858511
5	mse	100	log2	0.870715
6	mae	10	sqrt	0.874334
7	mae	10	auto	0.834414
8	mae	10	log2	0.867458
9	mae	100	sqrt	0.874372
10	mae	100	auto	0.85471
11	mae	100	log2	0.872091
12	friedman_mse	10	sqrt	0.845719
13	friedman_mse	10	auto	0.835182
14	friedman_mse	10	log2	0.852274
15	friedman_mse	100	sqrt	0.870301
16	friedman_mse	100	auto	0.854448
17	friedman_mse	100	log2	0.866721

By considering the result of r2score of all the model we can finalized the Model using Random Forest as a best model.

R2_Value for Random Forest: 0.874372