Assignment Machine Learning-Regression:

A client's requirement is to predict the insurance charges based on several parameters. The client has provided the dataset for the same.

1. **Identifying the problem statement**

   Three stages of identification
   Stage 1: Domain identification
   Since the majority of the prediction is based on statistical data the domain could be Machine Learning
   Stage 2: Learning Selection
   The data set has all the data present, has both input and output details, so we can select supervised learning.
   Stage 3: Regression or Classification
   As output is numerical information, we are using Regression here

2. **Basic information about the dataset**
   Input datasets: Age(integer), Sex(Character),BMI(integer),Children(integer),Smoker(Char)
   Output dataset: Charges(integer)

3. **Data Preprocessing:**
   Since we have categorical data as Sex (Male,Female), and Smoker(Yes/No), as these data are nominal data, we need to convert those columns to integer using **One Hot Encoding** technique.

4. **Developing a good model (best r^2 value for all the models)**
   a) Linear Regression: r^2 value= 0.78947
   b) Support Vector Machine: r^2 value=0.85205
   c) Decision Tree: r^2 value=0.75466
   d) Random Forest: r^2 value=0.87106

Hyper Parameter Substitution to find the best r^2 value:

Support Vector Machine

S.No	Hyper Parameter	Linear (r^2 value)	RBF (Non Linear) (r^2 value)	Poly (r^2 value)	Sigmoid (r^2 value)
1	C=0.1	-0.08092	-0.08908	-0.08829	-0.08826
2	C=1	-0.00967	-0.08347	-0.07564	-0.07536
3	C=10	0.46593	-0.03316	0.03906	0.04019
4	C=100	0.63124	0.31379	0.61565	0.53142
5	C=1000	0.76714	0.81149	0.85205	0.28739

Decision Tree

S.No	Criterion	Splitter	Max Features	R^2 Value
1	Squared_error	Best	None	0.69594
2	Squared_error	random	None	0.68107
3	Squared_error	Best	Sqrt	0.74182
4	Squared_error	random	Sqrt	0.61870
5	Squared_error	Best	Log2	0.59929

6	Squared_error	random	Log2	0.54649
7	Friedman_mse	Best	None	0.67712
8	Friedman_mse	random	None	0.69268
9	Friedman_mse	Best	Sqrt	0.72403
10	Friedman_mse	random	Sqrt	0.70404
11	Friedman_mse	Best	Log2	0.70527
12	Friedman_mse	random	Log2	0.71560
13	Absolute_error	best	None	0.65312
14	Absolute_error	random	None	0.73539
15	Absolute_error	best	Sqrt	0.75466
16	Absolute_error	random	Sqrt	0.71993
17	Absolute_error	Best	Log2	0.72675
18	Absolute_error	random	Log2	0.70040
19	Poisson	best	None	0.70614
20	Poisson	random	None	0.74630
21	Poisson	best	Sqrt	0.66786
22	Poisson	random	Sqrt	0.44001
23	Poisson	Best	Log2	0.63493
24	Poisson	random	Log2	0.61147

Random Forest

S.No	Criterion	Max_Features	N_Estimators	R^2 Value
1	Squared_error	None	50	0.84983
2	Squared_error	Sqrt	50	0.86958
3	Squared_error	Log2	50	0.86958
4	Squared_error	None	100	0.85383
5	Squared_error	Sqrt	100	0.87102
6	Squared_error	Log2	100	0.87102
7	Absolute_error	None	50	0.85266
8	Absolute_error	Sqrt	50	0.87081
9	Absolute_error	Log2	50	0.87081
10	Absolute_error	None	100	0.85200
11	Absolute_error	Sqrt	100	0.87106
12	Absolute_error	Log2	100	0.87106
13	Friendman_mse	None	50	0.85007
14	Friendman_mse	Sqrt	50	0.87024
15	Friendman_mse	Log2	50	0.87024
16	Friendman_mse	None	100	0.85405
17	Friendman_mse	Sqrt	100	0.87105
18	Friendman_mse	Log2	100	0.87105
19	Poisson	None	50	0.84910
20	Poisson	Sqrt	50	0.86323
21	Poisson	Log2	50	0.86323
22	Poisson	None	100	0.85263
23	Poisson	Sqrt	100	0.86801
24	Poisson	Log2	100	0.86801

5. **The Final model to find the insurance charges:**
   The best and final model to find the insurance charges using Machine Learning Regression is Random Forest with certain parameters (Criterion=absolute error, Max features=sqrt or log2, n estimators=100) for which r2 value is 0.87106