Assignment-Regression Algorithm

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.

1. **Identify your problem statement**
   a. Stage 1 - Machine Learning
   b. Stage 2 - Supervised Learning
   c. Stage 3 – Regression
2. **Tell basic info about the dataset (Total number of rows, columns)**
   a. Total number of rows = 1338
   b. Total number of columns = 6
   c. Input Columns = age, sex, bmi, children, smoker
   d. Output Column = charges
3. **Mention the pre-processing method if you're doing any (like converting string to number – nominal data)**
   a. As sex and smoker fields are Categorical Nominal data, we need to pre-processing the dataset by converting those field values into number using One hot Encoding method.
4. **Regression Result based on Multiple algorithm**
   a. Multiple Linear Regression R2 value = 0.7894790349867009
   b. Support Vector Machine (SVM)

S.No	kernel	C	R2 Value
1	linear	0.1	-0.1220767
2	linear	1	-0.1116613
3	linear	10	-0.0016176
4	linear	100	0.54328182
5	linear	1000	0.63403693
6	poly	0.1	-0.0862525
7	poly	1	-0.0642926
8	poly	10	-0.0931162
9	poly	100	-0.0997617
10	poly	1000	0.05550594
11	rbf	0.1	-0.0895762
12	rbf	1	-0.0884273
13	rbf	10	-0.0819691
14	rbf	100	-0.1248037
15	rbf	1000	0.11749092
16	sigmoid	0.1	-0.0897435
17	sigmoid	1	-0.0899412
18	sigmoid	10	-0.0907832

19	sigmoid	100	-0.1181455
20	sigmoid	1000	-1.6659081

SVM Regression R2 Value = 0.63403693

c. Decision Tree

S.No	Criterion	Splitter	Max Features	R Score
1	squared_error	best	sqrt	0.773231
2	friedman_mse	best	sqrt	0.719123
3	absolute_error	best	sqrt	0.754701
4	poisson	best	sqrt	0.671059
5	squared_error	random	sqrt	0.634898
6	friedman_mse	random	sqrt	0.711312
7	absolute_error	random	sqrt	0.701492
8	poisson	random	sqrt	0.688266
9	squared_error	best	log2	0.726303
10	friedman_mse	best	log2	0.672749
11	absolute_error	best	log2	0.735255
12	poisson	best	log2	0.784228
13	squared_error	random	log2	0.630231
14	friedman_mse	random	log2	0.654595
15	absolute_error	random	log2	0.75119
16	poisson	random	log2	0.655977

Decision Tree Regression R2 Value = 0.784228

d. Random Forest

S.No	Criterion	N Estimators	Max Features	R Score
1	squared_error	100	sqrt	0.87102719
2	friedman_mse	100	sqrt	0.8710544
3	absolute_error	100	sqrt	0.87106859
4	poisson	100	sqrt	0.8680157
5	squared_error	500	sqrt	0.87102589
6	friedman_mse	500	sqrt	0.87109927
7	absolute_error	500	sqrt	0.87220224
8	poisson	500	sqrt	0.87147995
9	squared_error	100	log2	0.87102719
10	friedman_mse	100	log2	0.8710544
11	absolute_error	100	log2	0.87106859
12	poisson	100	log2	0.8680157
13	squared_error	500	log2	0.87102589

14	friedman_mse	500	log2	0.87109927
15	absolute_error	500	log2	0.87220224
16	poisson	500	log2	0.87147995

Random Forest Regression R2 Value = 0.87220224

5. **The final machine learning best method of Regression:**
   Random Forest R2 Value (Criterion=absolute error, Max_Features=sqrt&log2, N_Estimators =500) = **0.87220224**