Regression Baby Step 3 Assignment consolidation

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

To predict the insurance charges by the data having age,bmi,sex, no.of children and smoker

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

1338 rows × 6 columns

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

MLR: The One hot Encoding method is used to convert sex and smoker columns into categorical data
SVM: Standardization to improve the model

4.) Develop a good model with r2_score. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

5.) All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

1. **Multiple Linear Regression:** The best model is 0.78
2. **SVM:**

The best model is created for C3000 and Linear parameter(0.89)

S.No	Hyper tuning parameter	Linear	RBF	Poly	Sigmoid
1	c10	0.76	-0.032	0.038	0.039
2	c100	0.62	0.32	0.61	0.52
3	c1000	0.76	0.81	0.85	0.28

3. **Decision Tree:**

The best model is 0.87

S.No	Criterion	MAX Features	Splitter	R Value
1	Squared_error	None	best	0.69
2	Squared_error	None	random	0.64
3	Squared_error	sqrt	best	0.70
4	Squared_error	sqrt	random	0.71
5	Squared_error	Log2	best	0.70
6	Squared_error	Log2	random	0.65
7	poisson	None	best	0.72
8	poisson	None	random	0.65
9	poisson	sqrt	best	0.71
10	poisson	sqrt	random	0.64
11	poisson	Log2	best	0.63
12	poisson	Log2	random	0.70
13	Friedman_mse	None	best	0.69
14	Friedman_mse	None	random	0.69
15	Friedman_mse	sqrt	best	0.68
16	Friedman_mse	sqrt	random	0.73
17	Friedman_mse	Log2	best	0.70
18	Friedman_mse	Log2	random	0.59

4. Random Forest:

S.No	Criterion	N_criterion	Max_features	R Value
1	Squared_error	10	sqrt	0.84
2	Squared_error	100	sqrt	0.86
3	Squared_error	1000	sqrt	0.87
4	Squared_error	10	None	0.84
5	Squared_error	100	None	0.85
6	Squared_error	1000	None	0.85
7	poisson	10	sqrt	0.85
8	poisson	100	sqrt	0.86
9	poisson	1000	sqrt	0.87
10	poisson	10	None	0.84
11	poisson	100	None	0.85
12	poisson	1000	None	0.85
13	Friedman_mse	10	sqrt	0.85

14	Friedman_mse	100	sqrt	0.86
15	Friedman_mse	1000	sqrt	0.87
16	Friedman_mse	10	None	0.83
17	Friedman_mse	100	None	0.85
18	Friedman_mse	1000	None	0.85

6.) Mention your final model, justify why u have chosen the same.

The best model is Random Forest when the parameters are n_estimators 1000 and max feature as ‘SQRT’. These combinations provides the maximum r value of 0.87.