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**
   We have to predict the insurance charges based on given data
   1. Stage 1: Machine Learning
   2. Stage 2: Supervised
   3. Stage 3: Regression
2.) **Tell basic info about the dataset (Total number of rows, columns)**
   Columns: 6
   Rows: 1338
3.) **Mention the pre-processing method if you're doing any (like converting string to number – nominal data)**
   We are doing pre-processing method for smoker column and sex column by using One Hot Encoding to convert Nominal data to numerical value.
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.)**

Multiple Linear Regression r_ score value = **0.78**

SVM- While using SVM below are the r_score values

Kernel	gamma	C	r_score
Linear		default	-0.1
Linear		10.00	0.46
Linear		100.00	0.62
Linear		1000.00	0.76
Poly	Scale	1.00	-0.07
Poly	Scale	10.00	0.03
Poly	Scale	1000.00	0.85
rbf	Scale	1.00	-0.08
rbf	Scale	10.00	-0.03
rbf	Scale	1000.00	0.81
rbf	auto	1.00	-0.83

rbf	auto	100.00	0.32
rbf	auto	1000.00	0.81
sigmoid	auto	1.00	-0.07
sigmoid	auto	100.00	0.52
sigmoid	auto	1000.00	0.28
sigmoid	Scale	1.00	-0.07
sigmoid	Scale	100.00	0.52
sigmoid	Scale	1000.00	0.28

By using kernel=poly, gamma=scale, C=1000 r_score value = **0.85**

Decision Tree- While using decision tree below are the r_score values

Criterion	splitter	r_score
squared_error	best	0.68
squared_error	random	0.72
friedman_mse	random	0.72
friedman_mse	best	0.69
absolute_error	best	0.69
absolute_error	random	0.71
poisson	random	0.72
poisson	best	0.73

By using Criterion=poison, splitter=best r_score value **= 0.73**

Random Forest- while using random forest below are the r_score values

n_estimators	Criterion	r_score
100	squared_error	0.85
200	squared_error	0.85
500	squared_error	0.85
100	absolute_error	0.85
200	absolute_error	0.85
500	absolute_error	0.85
1000	absolute_error	0.85
100	friedman_mse	0.85
200	friedman_mse	0.85
500	friedman_mse	0.85

1000	friedman_mse	0.85
100	poisson	0.84
200	poisson	0.85
500	poisson	0.85
1000	poisson	0.85

Most of the method in Random forest given the same value so here r_score value = **0.85**

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

Here both the SVM and Random forest given the same answers, however in SVM we have increased the C value but in Random forest by using default value itself we get the same answer and I would like to choose with Random Forest