Regression Assignment

Problem Statement or Requirement:

A client's requirement is, he wants to predict the insurance charges based on the several paramet ers. The Client has provided the dataset of the same. As a data scientist, you must develop a mod el which will predict the insurance charges.

1.) Identify your problem statement

Develop a predictive model to estimate insurance charges based on a given dataset containing various parameters. The objective is to create an accurate and reliable machine learning model that can effectively predict insurance charges for individuals. The dataset includes features such as 'age', 'bmi', 'children', 'sex_male', and 'smoker_yes', and the target variable to predict is 'insurance charges'.

2.) Tell basic info about the dataset (Total number of rows, columns)
Insurance dataset is consist of 1338 rows and 6 columns
In sex, male occupies 50.15% and female is 49.85%
Smokers by sex and age
Inthis dataset 79.89% were no smokers and 20.11% are smokers. Here, 8.46% are female and 11. 65% are male were smokers in the given dataset
Lowest and highest insurance charge in the dataset:
Highest insurance charge persons age is 44,gender is female, BMI is 38.06, has no children, smo ker and charges is 48885.13561.
Lowest insurance charge persons age is 18,gender is male, BMI is 53.13, has no children, smoker and charges is 1163.4627.

3.) Mention the pre-processing method if you're doing any (like converting string to number – no minal data)

Dataset Pre-processing:

Encoding categorical variables: The "sex" and "smoker" columns need to be converted into numerical values. This can be done using one-hot encoding or label encoding.

4.) Develop a good model with r2_score. You can use any machine learning algorithm; you can c reate many models. Finally, you have to come up with final model.
Randam Forest :r2_score = 0.875015

Support Vector Machine:r2_score = 0.865103

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

To find the Machine Learning model prediction using r2_score with different method of Algorithms :

Multiple Linear Regression: r2_score = 0.7894790349867009

Support Vector Machine:

kernel	gamma	C	R-squared Score
kernel	auto	10	0.865103
kernel	auto	50	0.825831
kernel	auto	100	0.649008
kernel	scale	100	0.543282
kernel	auto	100	0.543282
kernel	auto	50	0.398786
kernel	scale	50	0.398786
kernel	scale	10	-0.00162
kernel	auto	10	-0.00162
kernel	auto	100	-0.07459
kernel	auto	50	-0.08108
kernel	scale	10	-0.08197
kernel	auto	10	-0.0875
kernel	auto	10	-0.08971
sigmoid	auto	50	-0.08971
sigmoid	auto	100	-0.08971
sigmoid	scale	10	-0.09078
poly	scale	10	-0.09312
sigmoid	scale	50	-0.09852
poly	scale	100	-0.09976
poly	scale	50	-0.10033
rbf	scale	50	-0.11137
sigmoid	scale	100	-0.11815
rbf	scale	100	-0.1248

Randam Forest:

criterion	n_estimators	max_features	R-squared Score
absolute_error	100	sqrt	0.875015
absolute_error	100	log2	0.873655
friedman_mse	100	sqrt	0.87231
squared_error	100	sqrt	0.87171
absolute_error	50	log2	0.871498
friedman_mse	50	sqrt	0.871399
squared_error	100	log2	0.869844
poisson	100	sqrt	0.86959
poisson	100	log2	0.86933
absolute_error	50	sqrt	0.868953
squared_error	50	sqrt	0.868599
friedman_mse	100	log2	0.867662
friedman_mse	10	sqrt	0.867355
friedman_mse	50	log2	0.867097
poisson	50	log2	0.867023
squared_error	50	log2	0.865959
poisson	50	sqrt	0.863952
friedman_mse	10	log2	0.863238
poisson	50		0.858992
squared_error	100		0.857918
squared_error	10	sqrt	0.856981
squared_error	10	log2	0.856855
absolute_error	50		0.855056
poisson	10	log2	0.85452
absolute_error	100		0.853989
absolute_error	10	log2	0.853732
absolute_error	10	sqrt	0.852499
squared_error	50		0.851622
poisson	10	sqrt	0.851561
poisson	100		0.851468

friedman_mse	100		0.850862
friedman_mse	50		0.850525
squared_error	10		0.84765
friedman_mse	10		0.847054
absolute_error	10		0.843257
poisson	10		0.833226

Decision Tree:

criterion	**splitter**	**max_features**	**R-squared Score**
friedman_mse	best	sqrt	0.747527
absolute_error	random		0.74193
poisson	random		0.737802
absolute_error	best	sqrt	0.732132
poisson	best	log2	0.72352
poisson	best		0.719937
squared_error	random		0.71917
absolute_error	best	log2	0.710394
absolute_error	random	log2	0.709487
absolute_error	best		0.700402
poisson	best	sqrt	0.694093
squared_error	best	log2	0.692014
friedman_mse	best		0.687296
friedman_mse	random		0.682968
squared_error	best		0.682307
squared_error	random	log2	0.678822
friedman_mse	best	log2	0.676317
absolute_error	random	sqrt	0.671968
friedman_mse	random	log2	0.66322
squared_error	random	sqrt	0.654428
poisson	random	log2	0.628326
poisson	random	sqrt	0.626985
friedman_mse	random	sqrt	0.563996
squared_error	best	sqrt	0.505659

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

Final Model: The Random Forest Regressor performed better with a higher R-squared score (0.876), indicating a stronger predictive ability. Therefore, the final model chosen is the Random Forest Regressor.

Justification: The Random Forest Regressor is an ensemble learning method that combines the predictions of multiple decision trees. It is robust, handles non-linearity well, and often performs well in practice. The higher R-squared score further supports the choice of the Random Forest model as it indicates a better fit to the data compared to Linear Regression.