INSURANCE VALUE PREDICTION

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

Machine Learning

Supervised Learning

Regression

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

No.Of Rows - 1338

No. Of Columns - 6

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

It has 2 Categorical columns in the dataset. And it is a nominal datatype. so i use the One Hot Encoding algoritham and also use get dummies python libraries. It's helps to convert the categorical columns 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.)

1. MULTIPLE LINEAR REGRESSION(R2 VALUE)=0.789

2. SUPPORT VECTOR MACHINE:

S.NO	HYPER PERAMETER	LINEAR	POLY	RBF	SIGMOID
1	C=100	0.628	0.617	0.320	0.527
2	C=500	0.763	0.826	0.664	0.444
3	C=800	0.764	0.851	0.766	0.297
4	C=1000	0.764	0.856	0.810	0.287
5	C=2000	0.744	0.860	0.854	-0.593
6	C=3000	0.741	0.859	0.866	-2.124
7	C=5000	0.741	0.859	0.874	-7.530
8	C=8000	0.741	0.859	0.878	-21.58

The **SVM REGRESSION** use **R2** Value (RBF and hyper parameter(C=8000))=**0.878**

3. DECISION TREE

S.NO	CRITERION	MAX FEATURES	SPLITTER	R2 VALUE
1	Squared_error(mse)	auto	best	0.722
2	Squared_error(mse)	auto	random	0.750
3	Squared_error(mse)	sqrt	best	0.753
4	Squared_error(mse)	sqrt	random	0.733
5	Squared_error(mse)	log2	best	0.761
6	Squared_error(mse)	log2	random	0.713
7	absolute_error(mae)	auto	best	0.689
8	absolute_error(mae)	auto	Random	0.765
9	absolute_error(mae)	sqrt	best	0.747
10	absolute_error(mae)	sqrt	random	0.719
11	absolute_error(mae)	log2	best	0.643
12	absolute_error(mae)	log2	random	0.644
13	friedman_mse	auto	best	0.692
14	friedman_mse	auto	random	0.728
15	friedman_mse	sqrt	best	0.777
16	friedman_mse	sqrt	random	0.681
17	friedman_mse	log2	best	0.722
18	friedman_mse	log2	random	0.704
19	Poisson	auto	best	0.679

20	Poisson	auto	random	0.712
21	Poisson	sqrt	best	0.701
22	Poisson	sqrt	random	0.650
23	Poisson	log2	best	0.667
24	Poisson	log2	random	0.748

The **DECISION TREE** use $\mathbf{R^2}$ Value (mae, log2 and best)=**0.777**

4. RANDOM FOREST

S.NO	CRITERION	MAX FEATURES	R2 VALUE
1	Squared_error(mse)	auto	0.850
2	Squared_error(mse)	sqrt	0.872
3	Squared_error(mse)	log2	0.872
4	absolute_error(mae)	auto	0.854
5	absolute_error(mae)	sqrt	0.871
6	absolute_error(mae)	log2	0.873
7	friedman_mse	auto	0.853
8	friedman_mse	sqrt	0.871
9	friedman_mse	log2	0.870
10	Poisson	auto	0.832
11	Poisson	sqrt	0.826
12	Poisson	log2	0.831

1	Parameter	n_estimators=100	0.853

The **RANDOM FOREST** use **R2 Value** (mae & log2)=**0.873**

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

The final Machine Learning best method of Regression:

1. Support Vector Machine R2value (RBF and hyper parameter(C=8000)) =**0.878**
2. Random Forest R2value(mae & log2) =**0.873**

Kindly create Repository in the name Regression Assignment. Upload all the ipynb and final document in the pdf Communication is important (How you are representing the document.)