Regression Assignment

Usually a proper way to approach a dataset is by identifying the three stages of AI Prediction

1.Identify the Problem statement

There are three stages of AI Prediction

Stage 1:

Domain Selection: Machine Learning

Stage 2:

Learning: Supervised

Stage 3:

Classification: Regression

2. Information about the dataset

Insurance_pre.csv contains 1338 rows × 6 columns

3. Prepocessing Method

The Insurance_pre.csv has Nominal Data

4. r2 values for each algorithm and their respectives max_features,C parameter

1.Multiple Linear Regression r2 value ( ) is 0.7894790349867009

2.Support Vector Machine

serial.No	Hyper parameter	Linear R value	RBF(Non Linear R value)	Poly (r value)	Sigmoid (r value)
1	C10	0.4624684142339 68	-0.03227329390671052	0.038716222760231345	0.039307143782743447
2	C100	0.6288792857320359	0.3200317832050831	0.6179569624059795	0.5276103546510407
3	C500	0.7631057975975 37	0.6642984645143137	0.826368354126896	0.44460610338694795
4	C1000	0.7649311738608445	0.8102064851758545	0.8566487675946551	0.28747069486976695

5	C2000	0.7440418308107487	0.854776642539298	0.860557927577388	-0.5939509731283508
6	C3000	0.7414236599249815	0.8663393953081686	0.8598930084494408	-2.1244194786689863

Support Vector Machine best r2 value (RBF(Non Linear), Hyperparameter (C3000)) is 0.8663393953081686

3. Descision Tree

Serial.No	*criterion*	Max_features	Splitter	R_score Value
1	Mse	auto	best	0.687194563268898
2	Mse	auto	random	0.7058331807027536
3	Mse	sqrt	best	0.5780449056811394
4	Mse	Sqrt	random	0.671774985077185
5	Mse	Log2	best	0.6799833114025472
6	Mse	Log2	random	0.6801072400316461
7	Mae	auto	best	0.6648432062010856
8	Mae	auto	random	0.7500551403583591
9	Mae	sqrt	best	0.7339690095918607
10	Mae	sqrt	random	0.6941169283379627
11	Mae	Log2	best	0.5658811496163172
12	Mae	Log2	random	0.7028754036814628
13	***friedman_mse***	auto	best	0.7188171695197177

14	***friedman_mse***	auto	random	0.6783110980709615
15	***friedman_mse***	sqrt	best	0.7055531104846458
16	***friedman_mse***	sqrt	random	0.5952600931249437
17	***friedman_mse***	Log2	best	0.7963033250752416
18	***friedman_mse***	Log2	random	0.6226609825934093
19	***poisson***	auto	best	0.6792193646949658
20	***poisson***	auto	random	0.6982855340536165
21	***poisson***	sqrt	best	0.6106388856607337
22	***poisson***	sqrt	random	0.6542909381256115
23	***poisson***	Log2	best	0.5227100972953048
24	***poisson***	Log2	random	0.6467486767283614

Descision Tree best r2 value (*criterion* (***friedman_mse)***, Max_features (Log2), splitter (Best))

is 0.7963033250752416

4.Random Forest

Serial.No	*criterion*	Max_features	**n_estimators**	R_score Value
1	Mse	auto	10	0.8446897393177992
2	Mse	auto	100	0.8582064953815349
3	Mse	sqrt	10	0.8535877981749798

4	Mse	sqrt	100	0.8722476146127074
5	Mse	Log2	10	0.8584178716708083
6	Mse	Log2	100	0.8706412205393324
7	Mae	auto	10	0.8205612752559868
8	Mae	auto	100	0.8501514295 69389
9	Mae	sqrt	10	0.8500750883670041
10	Mae	sqrt	100	0.8717456643495153
11	Mae	Log2	10	0.8543698071729846
12	Mae	Log2	100	0.8736539550478307

Random Forest best r2 value (*criterion* (***mae***), Max_features (Log2), N_estimator is 100) is 0.8736539550478307

The final Machine Learning Best Model of Regression for insurance dataset :

From implementing algorithm in order to find the best model .I found that both Random Forest and Support Vector Machine has best R_score values

1.Random Forest best r2 value (*criterion* (***mae***), Max_features (Log2), N_estimator is 100) is 0.8736539550478307

or

2.Support Vector Machine best r2 value (RBF(Non Linear) and HyperParameter(C3000)) is 0.8663393953081686