Machine Learning – Regression Models - Assignment

Problem Statement:

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.

Solution:

3 stages :

1. Domain Selection – Machine learning (data is structured (i.e.)in Excel)
2. Learning – Supervised Learning (Input and Output is clear and present in data itself)
3. Regression/Classification – Regression (Output variable deals with numbers)

No.Of Rows in Data sheet– 1339

No.Of columns – 6 (5 Input and 1 Output)

2 Columns have nominal data. So need to use One-Hot Encoding to convert the string to number.

After converting

Input Fields	Output Field
Age,bmi,children,sex_male,smoker_yes	charges

Algorithms:

1. Multiple Linear Regression

   The R^2 value is 0.7894

2. Support Vector Machine

Kernel	C	R Value
rbf	500	0.6642
rbf	1000	0.8102
rbf	2000	0.8547
linear	500	0.7631
linear	1000	0.7649
linear	2000	0.744
poly	500	0.8263
poly	1000	0.8566
poly	2000	0.8605
sigmoid	500	0.4446
sigmoid	1000	0.2874

Here the R^2 value by using parameters Kernel=poly and c=2000is 0.8605

3. Decision Tree

criterion	splitter	max_features	R Value
squared_error	random	None	0.6905
squared_error	best	None	0.6876
squared_error	random	sqrt	0.7075
squared_error	best	sqrt	0.7324
squared_error	random	log2	0.6066
squared_error	best	log2	0.7288
friedman_mse	random	None	0.6948
friedman_mse	best	None	0.6939
friedman_mse	random	sqrt	0.7093
friedman_mse	best	sqrt	0.7639
friedman_mse	random	log2	0.6384
friedman_mse	best	log2	0.7593
absolute_error	random	None	0.6655
absolute_error	best	None	0.7354
absolute_error	random	sqrt	0.658
absolute_error	best	sqrt	0.7018
absolute_error	random	log2	0.7022
absolute_error	best	log2	0.6635
poisson	random	None	0.6681
poisson	random	sqrt	0.6654
poisson	random	log2	0.6624
poisson	best	log2	0.7629

As there is not much performance these parameters ,tried with other parameter(max_depth)

criterion	splitter	max_depth	R Value
squared_error	random	6	0.827
squared_error	best	6	0.8172
friedman_mse	random	6	0.8586
friedman_mse	best	6	0.8172
absolute_error	random	6	0.8141
absolute_error	best	6	0.8539
poisson	random	6	0.8539
poisson	best	6	0.8226
squared_error	best	2	0.8569
friedman_mse	best	2	0.8569
poisson	best	3	0.8569
squared_error	best	4	0.8837
friedman_mse	best	4	0.8837
poisson	best	4	0.8847

Here the R^2 value for parameters Criterion=poisson, splitter=best, max_depth=4 is 0.8847

4. Random Forest

criterion	n_estimators	random_state	max_features	max_depth	R Value
squared_error	50	0			0.8498
absolute_error	50	0			0.8526
friedman_mse	50	0			0.8511
poisson	50	0			0.8491
squared_error	50		sqrt		0.8711
absolute_error	50		sqrt		0.8694
friedman_mse	50		sqrt		0.8671
poisson	50		sqrt		0.8684
squared_error	50		log2		0.867
absolute_error	50		log2		0.8666
friedman_mse	50		log2		0.8711
poisson	50		log2		0.8648
squared_error	100			4	0.8899
absolute_error	100			4	0.8864
friedman_mse	100			4	0.8882
poisson	100			4	0.8903

Here the R^2 value for parameters Criterion=poisson, n_estimators =100, max_depth=4 is 0.8903

R^2 Values of all Algorithms

S.No	Algorithms	R Value
1	Multiple Linear Regression	0.7894
2	Support Vector Machine	0.8605
3	Decision Tree	0.8847
4	Random Forest	0.8903

By Comparing all the R^2 value, Random Forest model has good performance.

Best Model:

The final best model for this problem statement is **Random Forest with R^2 Value – 0.8903**