Hope Artificial Intelligence

Assignment-Regression Algorithm

Dataset : Insurance_pre

1. Identifying Problem statement

   For the given dataset and the client's requirement the following stages are followed.

   Stage 1 : Domain selection
   - Machine Learning

   Stage 2 : Selecting the Learning algorithm
   - Supervised Learning algorithm

   Stage3 : The type of supervised learning
   - Regression Learning method

2. Information about the dataset:

   The given dataset has 6 columns. In that
   Input column's are : age, sex, bmi, children, smoker_yes [5 columns]
   Output column : charges [1 column]

3. Data preprocessing columns:
   In the given dataset 2 string columns are : sex and smoker_yes
   These are converted into **nominal data**

4. Developing the model using the following algorithms:
   - Multiple Linear Regression method
   - Support Vector Machine Regression method
   - Decision Tree Regression method
   - Random Forest Regression method

   Among developing these models the best final model with a better R-Score Value is using a Support Vector Machine Regression method using hyper-tuning parameter C=10000 and kernel='rbf'. **R^2-score =0.87799**

5. **Research Values** :

Dataset = insurance_pre

Method 1: Multiple Linear Regression

R^2-SCORE VALUE= 0.78947

Method 2 : Support Vector Machine Regression

Using Hypertuning Parameter

S.no	**Hypertuning Parameter**	**Kernel=linear R-value**	**Kernel=rbf R-value**	**Kernel=poly R- value**	**Kernel=sigmoid R-value**
1	Default C=1.0	-0.0101	-0.08338	-0.07569	-0.07542
2	C=10	0.46246	-0.03227	0.038716	0.039307
3	C=100	0.62887	0.32003	0.617956	0.52761
4	C=1000	0.76493	0.8102	0.85664	0.28747
5	C=10000	0.74142	0.87799	0.85917	-34.1515
6	C=100000	0.741418	0.87249	0.85778	-3465.953

Best R2-Score Value = Kernel='rbf' , C=10000, R^2–Value = 0.87799

Method 3: Decision Tree Regression

S.No	**criterion**	**splitter**	**max_features**	**R_Score value**
1.	***squared_error***	*best*	*None*	0.68415
2.		random	None	0.66479
3.		Best	Sqrt	0.69676

4.		Random	sqrt	0.624433
5.		best	Log2	0.63420
6.		Random	Log2	0.77535
7.	***friedman_mse***	Best	None	0.68751
8.		Random	None	0.68961
9.		Best	Sqrt	0.66582
10.		Random	sqrt	0.65176
11.		Best	Log2	0.72305
12.		Random	Log2	0.67170
13.	***absolute_error***	Best	None	0.69266
14.		Random	None	0.71183
15.		Best	Sqrt	0.6827
16.		Random	Sqrt	0.60791
17.		Best	Log2	0.75664
18.		Random	Log2	0.66556
19.	***poisson***	Best	None	0.72404
20.		Random	None	0.66327
21.		Best	Sqrt	0.72776
22.		Random	Sqrt	0.64671
23.		Best	Log2	0.57555
24.		Random	Log2	0.75316

Best Method for Decision Tree Regressor – criterion = squared_error, splitter= random, max- features = log2

R^2-Value = 0.77535

Method-4 : Random Forest Regressor

S.No	**n_estimators**	**criterion**	**max_features**	**R-Value**
1	100	***squared_error***	None or 1.0	0.855493
2	50			0.8545
3	10			0.83615
4	100		sqrt	0.8697
5	50			0.8668
6	10			0.8481
7	100		Log2	0.8695
8	50			0.86591
9	10			0.8583
10	100	***absolute_error***	None / 1.0	0.85779
11	50			0.85001
12	10			0.83403
13	100		Sqrt	0.87233
14	50			0.8666
15	10			0.8582
16	100		Log2	0.8705
17	50			0.867483
18	10			0.85544

19	100	***friedman_mse***	None / 1.0	0.854966
20	50			0.84954
21	10			0.84530
22	100		Sqrt	0.87055
23	50			0.86587
24	10			0.86005
25	100		Log2	0.86969
26	50			0.86985
27	10			0.85598
28	100	***Poisson***	None / 1.0	0.85398
29	50			0.85036
30	10			0.84006
31	100		Sqrt	0.870639
32	50			0.8675
33	10			0.84676
34	100		Log2	0.86933
35	50			0.86431
36	10			0.850343

Best Method for Random Forest Regressor is n_estimators=100, criterion='absolute_error', max_features=log2. R^2_Value= 0.87233

6. Final best model for this given dataset and as per the clients requirement to find the charges for the customers is using

    - Support Vector Machine Regression method (C=10000 and kernel='rbf'). In this model we get a **R^2-score=0.87799**

    Or

    - Random Forest Regressor (n_estimators=100, criterion='absolute_error', max_features=$\log_2$). **R^2_Value= 0.87233**