Assignment-Regression Algorithm

Machine Learning

1. Problem statement

Develop a model to predict insurance charges based on various parameters provided in the dataset.

2. Basic information about Dataset

Google Drive Link Click here .
Git Hub Link: https://raw.githubusercontent.com/RamishaRaniK/dataset/main/insurance_pre.csv

Name of the dataset : *insurance_pre.csv*

- Total number of rows: 1338
- Total number of columns: 6

Description of the Columns

- age: Age of the insured individual
- sex: Gender of the insured individual (male/female)
- bmi: Body Mass Index of the insured individual
- children: Number of children/dependents covered by the insurance
- smoker: Smoking status of the insured individual (smoker/non-smoker)
- charges: Insurance charges associated with the individual

3. Pre-processing method

Convert categorical variables - Encoded *Sex, Smoker* categorical variables using one-hot encoding method.

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.)

Models Explored:

1. Simple Linear Regression (SLR)

```
r value : 0.7894790349867009
```

2. Multiple Linear Regression (MLR)

```
r value : 0.7865108093853883
```

3. Support Vector Machine (SVM)

S.No.	Hyper Parameter	Linear (r value)	Rbf (r value)	Sigmoid (r value)
1	C10	0.43207272266670915	-0.0480855688697257	0.01938608977859435
2	C100	0.6162374776116659	0.2913359565836946	0.5056420492628236
3	C500	0.6803476516742177	0.6397619043375011	0.4638549858721497
4	C1000	0.7594740768777103	0.791561827552486	0.1842218837986186
5	C2000	0.7613141349072211	0.8460208717344976	-0.5786822329194656
6	C3000	0.7612136779520974	0.8609984992147567	-2.0119256717238607

4. Decision Tree (DT)

S.No.	Criterion	Max Features	Splitter	R Value
1	squared_error		best	0.6882356308731509
2	squared_error		random	0.6986954577498701
3	friedman_mse		best	0.6863910695500609
4	friedman_mse		random	0.6978199263516077
5	absolute_error		best	0.7049434380274042
6	absolute_error		random	0.7348819214525363
7	Poisson		best	0.7164050604291204
8	Poisson		random	0.6963996390984526

5. Random Forest (RF)

S.No.	Criterion	N_enstimator	R Value
1	Squared_error	10	0.8489527320435293
2	Squared_error	100	0.8544944797692101
3	Asolute_error	10	0.8418359012447243
4	Asolute_error	100	0.8505347626828681
5	Friedman_mse	10	0.833759102471108
6	Friedman_mse	100	0.8470006451814043
7	Poisson	10	0.8406522584302984
8	Poisson	100	0.8551146432420877

5. All the research values (r2_score of the models) should be documented.

Model Evaluation:

S,No.	Model	R2 Score
1	Simple Linear Regression (SLR)	0.7894790349867009
2	Multiple Linear Regression (MLR)	0.7865108093853883
3	Support Vector Machine (SVM)	0.8609984992147567
4	Decision Tree	0.7348819214525363
5	Random Forest	0.8551146432420877

6. Final Model

Support Vector Machine (SVM) provides high predictive accuracy. Also, it has demonstrated the highest R2 score among the models explored, indicating its superior performance in capturing the relationships within the data. By selecting the Support Vector Machine (SVM) as the final model, we aim to provide the client with the most accurate predictions of insurance charges based on the given parameters.

Conclusion:

Through systematic exploration of various machine learning algorithms and thorough evaluation based on the R2 score, the final model, **Support Vector Machine (SVM)**, is chosen for its superior performance and suitability for the task at hand. This model is expected to provide reliable predictions of insurance charges for the client.