Machine Learning - Regression

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.

Dataset

Here I provide the first 10 data's from the dataset

AGE	SEX	BMI	CHILDREN	SMOKER	CHARGES
19	Female	27.9	0	Yes	16884.924
18	Male	33.77	1	No	1725.5523
28	Male	33.7	3	No	4449.462
33	Male	22.705	0	No	21984.47061
32	Male	28.88	0	No	3866.8552
31	Female	25.74	0	No	3756.6216
46	Female	33.44	1	No	8240.5896
37	Male	27.74	3	No	7281.5056
37	Female	29.83	2	No	6406.4107
60	Female	25.84	0	No	28923.13691

Inputs:

- Age
- Sex
- Bmi
- Children
- Smoker

Output:

- Charges

We have to predict the insurance charges with these inputs…

<u>Identifying the problem statement</u>

Stage-1

Domain selection - **Machine Learning**

(Here All the inputs are in Number form)

Stage-2

Learning selection- **Supervised Learning**

(Inputs and output data are given, So the requirement is very clear)

Stage-3

Problem Identification- **Regression**
(The output has Numerical values)

Basic info about the dataset

Total Number of rows and columns
Rows - 1338
Columns - 6

Pre-processing method

In the dataset the column sex and smoker has categorical data so we have to change into numerical value by using one hot encoding method

AGE	SEX_MALE	BMI	CHILDREN	SMOKER_YES	CHARGES
19	FALSE	27.9	0	TRUE	16884.924
18	TRUE	33.77	1	FALSE	1725.5523
28	TRUE	33.7	3	FALSE	4449.462
33	TRUE	22.705	0	FALSE	21984.47061
32	TRUE	28.88	0	FALSE	3866.8552
31	FALSE	25.74	0	FALSE	3756.6216
46	FALSE	33.44	1	FALSE	8240.5896
37	TRUE	27.74	3	FALSE	7281.5056
37	FALSE	29.83	2	FALSE	6406.4107
60	FALSE	25.84	0	FALSE	28923.13691

Here Machine can encoded as True -1 and False - 0.

Machine Learning Regression Algorithms

1. **Multiple Linear Regression**
   R2 score - 0.78

2. **Support Vector Machine**
   It Shows Negative R2 score for this dataset.

3. Decision Tree

CRITERION	SPLITTER	MAX_FEATURES	R2 SCORE
Squared_error	Best	Sqrt	0.79
	Best	Log2	0.70
	Random	Sqrt	0.64
	Random	Log2	0.64
Absolute_error	Best	Sqrt	0.71
	Best	Log2	0.69
	Random	Sqrt	0.59
	Random	Log2	0.69
Friedman_mse	Best	Sqrt	0.73
	Best	Log2	0.69
	Random	Sqrt	0.64
	Random	Log2	0.70
Poisson	Best	Sqrt	0.76
	Best	Log2	0.67
	Random	Sqrt	0.70
	Random	Log2	0.61

4.Random Forest

N_ESTIMATORS	RANDOM_STATE	R2 SCORE
10	0	0.85
50	0	0.84
100	0	0.83
10	None	0.85
50	None	0.85
100	None	0.82

Final Model Selection

ALGORITHM	BEST R2_SCORE
Multiple Linear Regression	0.78
Decision Tree	0.79
Random Forest	0.85

For this dataset Random Forest algorithm gives the highest R2 score as 0.85 , So I choose this as the Best model