ASSIGNMENT- Regression Algorithm

Problem Statement or Requirement:

A client's requirement is, he wants to predict the insurance charges based on several parameters. The Client has provided the dataset of the same. As a data scientist, you must develop a model which will predict the insurance charges.

1.3-**Stage of Problem Identification:**

1) Machine Learning
2) Supervised Learning
3) Regression

2. **Rows and Columns:**

Rows=1338

Columns=6

3. **Pre-processed data**:

Nominal: One Hot Encoding

Sex	Male	Female
Male	1	0
Female	0	1

Smoker	Yes	No
Yes	1	0
No	0	1

4.Implementation

1. Simple and Multiple Linear Algorithm R2 value is 0.78

2.Support Vector Machine:

S.No	Hyper Parameter	Linear R2-Value	RBF R2-Value	Poly R2-Value	Sigmoid R2-Value
1	C10	-113.0	-0.080	-1163348.2	0.013
2	C100	-146.1	-0.021	-32979013.9	-5.48
3	C500	-152.2	-6.231	-29631241.4	-0.54
4	C1000	-156.8	-0.015	-1054390.2	-4.54
5	C2000	-156.6	-0.049	-4507673.5	-16.3
6	C3000	-156.7	-0.065	-2700959.3	-58.5

The SVM Regression algorithm does not support predicting the insurance charges. All Parameters R2 Value gives poor model.

3.Decision Tree:

S.No:	Criterion	Splitter	R2-Value
1	Squared_error	Best	0.69
2	Squared_error	Random	0.71
3	Friedman_mse	Best	0.69
4	Friedman_mse	Random	0.69
5	Absolute_error	Best	0.64
6	Absolute_error	Random	0.77
7	Poisson	Best	0.67
8	Poisson	Random	0.70

The Decision Tree Regression uses R2-value (absolute_error and Random) =0.77.

4.Random Forest:

S.No:		Parameters	R2-Value
1	n_estimators=100	Random_state=0	0.85
2	**Criterion** =Squared_error	Random_state=0	0.85
3	**Criterion** =Friedman_mse	Random_state=0	0.85
4	**Criterion** =absolute_error	Random_state=0	0.85
5	**Criterion** =poisson	Random_state=0	0.85
6	Max_features=sqrt	Random_state=0	0.87
7	Max_features=log2	Random_state=0	0.87
8	Bootstrap=bool	Random_state=0	0.85
9	Oob_score=bool	Random_state=0	0.85

The Random forest Regression uses R2-value (max_features=sqrt,log2 and Random_state=0) =0.87.

5. **Final Model:**

Selected the final model as a Random Forest. It gives a good r2 value nearby 1. compare to other models all values are near by 0. The SVM algorithm is not working perfectly for this problem statement. As a data scientist, predicting the insurance charges using Random Forest Regression algorithm..

The final best model of the Machine Learning Regression algorithm:

The Random forest Regression uses R2-value (max_features=sqrt,log2, and Random_state=0) =0.87.