Assignment-Regression Algorithm

Problem Statement or 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. As a data scientist, you must develop a model which will predict the insurance charges.

As a data scientist, you must develop a model which will predict the insurance charges.

1. Identify your problem statement

Name of the Project: Insurance Charge Prediction

Three stages of Problem Identification:

1.) Stage 1-Machine Learning

2.) Stage 2-Supervised

3.) Stage 3-Classification

2. Tell basic info about the dataset (Total number of rows, columns)

- This dataset contains 1338 rows and 6 columns.
- It contains 6 independent variables and 1 dependent variable.

3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

One-Hot Encoding has been done by using get_dummies in pandas.

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.

1. **Multiple Linear Regression: R value: 1.0**

2. **Support Vector Machine**

SL.NO	kernel	R Value
1.	rbf	0.080
2.	linear	**0.095**
3.	poly	0.045
4.	sigmoid	0.064
5.	precomputed	Not supported for this data

The **SUPPORT VECTOR MACHINE** uses R2 value(linear)= **0.095**

3. **Decision Tree**

SL.NO	CRITERION	SPLITTER	R_VALUE
1.	squared_error	best	**0.9998**
2.	friedman_mse	best	0.9997
4.	poisson	best	0.9997
5.	poisson	random	0.9990
6.	squared_error	random	0.9995
7.	friedman_mse	random	0.9997

The **DECISION TREE REGRESSION** uses R2 value(squared_error, best)=**0.9998**

4. **Random Forest**

SL.NO	CRITERION	Parameters	N_estimators	R_Value
1.	Absolute_error	Min_samples_leaf	**100**	**0.9999**
2.	friedman_mse	Min_samples_leaf	**50**	0.9998
3.	squared_error	Min_samples_leaf	**100**	0.9997
4.	poisson	Min_samples_leaf	**50**	0.9998
5.	friedman_mse	Max_depth	**10**	0.9998
6.	poisson	Max_depth	**100**	0.9998
7.	Absolute_error	Max_depth	**50**	0.9998
8.	squared_error	Max_depth	**100**	0.9998

The **RANDOM FOREST** uses R2 value(Absolute_error, Min_samples_leaf)=**0.9999**

5.) All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

MODEL	VALUES
Multiple Linear Regression	**1.0**
Support Vector Machine	**0.095**
Decision Tree	**0.9998**
Random Forest	**0.9999**

6.) Mention your final model, justify why u have chosen the same.

The table above shows that Multiple Linear Regression is the most effective model to fit comparing to other models.