REGRESSION ALGORITHM – ASSIGNMENT

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.

1.) Identify your problem statement

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

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

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.

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

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

SOLUTION:

1. **Identify your problem statement**

   DOMAIN : STAGE 1: MACHINE LEARNING
   STAGE 2: SUPERVISED LEARNING
   STAGE 3: REGRESSION

This problem falls under **supervised learning** because:

- The dataset includes **labeled outputs** (insurance charges), meaning we know the expected outcomes.
- The model learns from the existing labeled data and **predicts new values** for unseen inputs.

Since we are predicting **continuous numerical values**, this is a **regression** problem within supervised learning.

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

TOTAL NO.OF COLUMNS : 6

TOTAL NO.OF ROWS: 1338

AGE – NUMERICAL

SEX – CATEGORICAL

BMI – NUMERICAL

CHILDREN – NUMERICAL

SMOKER – CATEGORICAL

CHARGES – NUMERICAL . this is the target variable or output or dependent variable.

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

we need pre – processing because dataset contains categorical value.

Convert categorical columns (sex,smoker),into numerical using (nominal method) One-Hot Encoding .

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.

Going to use MULTIPLE LINEAR ALG,SVM,DECISION TREE, RANDOM FOREST

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

MULTIPLE LINEAR REGRESSION: R2 SCORE : 0.78949

SVM (SUPPORT VECTOR MACHINE) REGRESSION ALGORITHM:

SNO	HYPER PARAMETER	LINEAR	RBF(non linear)	POLY	SIGMOID
1.	C10	-0.001617	-0.081969	-0.093116	-0.090783
2	C100	0.543281	-0.1248036	-0.0997617	-0.118145
3	C200	0.595044	-0.1263278	-0.0963778	-0.157549
4	C500	0.627046	-0.124641	-0.0820287	-0.456294
5	C1000	0.634036	-0.117490	-0.055505	-1.66590

6	C2000	0.689326	-0.1077876	-0.0027024	-5.61643
7	C3000	0.75908	-0.0962128	0.048928	-12.01904

SVM Regression use r2 value (linear) and hyper parameter(c3000) = 0.75908

3. DECISION TREE:

SNO	CRETERION	SPLITTER	R VALUE
1	Squared error	best	0.692685
2	Squared error	random	0.753644
3	*friedman_mse*	random	0.68877489
4	*friedman_mse*	best	0.695472
5	*absolute_error*	best	0.695472
6	*absolute_error*	random	0.721762
7	*poisson*	best	0.69756
8	*poisson*	random	0.701292

DECISION TREE Regression use r2 value (Squared Error,random) = 0.753644

4. RANDOM FOREST:

SNO	CRETERION	MAX FEATURES	N_ESTIMATORS	R VALUE
1	Squared error	1.0	10	0.848921
2	Squared error	sqrt	100	0.8720424
3	Squared error	1.0	1000	0.856445
4	Absolute error	1.0	1000	0.8559204
5	Absolute error	1.0	10	0.8474603
6	*friedman_mse*	1.0	10	0.837942
7	*friedman_mse*	1.0	1000	0.85591

8	***poison***	**1.0**	**10**	**0.838798**
9	***poison***	**1.0**	**100**	**0.85157**
10	***poison***	**sqrt**	**100**	**0.8694494**
11	***friedman_mse***	**sqrt**	**100**	**0.870312**
12	***friedman_mse***	**Log2**	**10**	**0.85659**
13	***poison***	**sqrt**	**10**	**0.85747**
14	***poison***	**Log2**	**100**	**0.85661**

RANDOM FOREST REGRESSION : r2 value (squared error,sqrt,100)= 0.8720424

Am going to select randomforest regressor is best model.