Assignment – Regression Algorithm

Dataset File Name: **insurance_pre.csv**

1. Identify your problem statement:

   3 – Stages:

   Stage -1: Domain → Machine Learning
   Stage -2: Learning → Supervised Learning
   Stage -3: Learning → Supervised Learning (Regression)

2. Tell basic info about the dataset:

   Total Number of rows: 1338
   Total Number of columns: 6

3. Mention the pre-processing method:

   String to Number – Ordinal data – Mapping – Label Encoder

SEX (TEXT)	SEX (NUMERICAL)
Female	0
Male	1

SMOKER (TEXT)	SMOKER (NUMERICAL)
No	0
Yes	1

4. Develop a good model with r2_score:

   - To find the validating parameter r2_score for the same dataset **insurance_pre.csv** using with following machine learning algorithms:

     1. **Multiple Linear Regression**
     2. **Support Vector Machine**
     3. **Decision Tree**
     4. **Random Forest**

- **Multiple Linear Regression:**

r^2_score = **0.789**

- **Support Vector Machine:**

S.NO	Hyper Tuning Parameter			r^2 Value
	kernel	gamma	C' paramter	
1	Linear	scale	5000	0.765
2	rbf	scale	0.01	-0.089
3	poly	scale	5000	0.146
4	sigmoid	scale	0.01	-0.089
5	Linear	auto	5000	0.765
6	rbf	auto	5000	0.125
7	poly	auto	10	0.865
8	sigmoid	auto	5000	-0.089

r^2_Value for SVM is **I. (Kernel "poly", gamma "auto", C=10) = 0.865**

- **Decision Tree:**

S.NO	Hyper Tuning Parameter			R^2 Value
	Criterion	Splitter	Max_Features	
1	squared_error	best	sqrt	0.713
2	friedman_mse	best	sqrt	0.719
3	absolute_error	best	sqrt	0.735
4	poisson	best	sqrt	0.728
5	squared_error	random	sqrt	0.662
6	friedman_mse	random	sqrt	0.695
7	absolute_error	random	sqrt	0.777
8	poisson	random	sqrt	0.743
9	squared_error	best	log2	0.728
10	friedman_mse	best	log2	0.738
11	absolute_error	best	log2	0.750
12	poisson	best	log2	0.752
13	squared_error	random	log2	0.71
14	friedman_mse	random	log2	0.713
15	absolute_error	random	log2	0.717
16	poisson	random	log2	0.683

r^2_ Value for Decision Tree :

(Criterion "absolute_error", Splitter "random", Max_Features "sqrt") = 0.777

- **Random Forest:**

S.NO	Hyper Tuning Parameter			R^2 Value
	Criterion	Max_Features	n_estimators	
1	squared_error	sqrt	100	0.87
2	friedman_mse	sqrt	100	0.871
3	absolute_error	sqrt	100	0.871
4	poisson	sqrt	100	0.868
5	squared_error	log2	100	0.87
6	friedman_mse	log2	100	0.871
7	absolute_error	log2	100	0.871
8	poisson	log2	100	0.868

r^2_Value for Random Forest is **:**

1. **(Criterion "squared_error", Max_features "sqrt", n=100) = 0.870**
2. **(Criterion "friedman_mse", Max_features "sqrt", n=100) = 0.871**
3. **(Criterion "absolute_error", Max_features "sqrt", n=100) = 0.871**
4. **(Criterion "squared_error", Max_features "log2", n=100) = 0.870**
5. **(Criterion "friedman_mse", Max_features "log2", n=100) = 0.871**
6. **(Criterion "absolute_error", Max_features "log2", n=100) = 0.871**

5. **All the research values (r2_score of the models) have been documented.**

6. Mention Your Final Model, Justify why you have chosen this:

Final Model: **Random Forest Algorithm (Regression)**

- For the data set **insurance_pre.csv** as the client requirement is insurance charges prediction. In this particular dataset we have used four different regression algorithms here. Finally we come up with the conclusion to chose the best model **Random Forest Algorithm (Regression)** depending upon the r2_score value. Usually we decide the r2_score value is 1 that is a good model and we can proceed the next step of deployment phase then we send to the end user. The r2_score value for the random forest is **0.871** & this value is almost equal to using all combination of hyper tuning parameters. So this algorithm would be worked perfectly for this data set.