ASSIGNMENT REGRESSION

1.Problem Statement:

Stage1: Machine Learning (Dataset are numeric)

Stage2: Supervised Learning (Requirements are clear)

Stage3: Regression (Dataset are continuous value)

2.Info about dataset:

Total number of rows:1338

Total number of columns:6

3.Preprocessing method:

Here in dataset contains nominal data (i.e sex column and smoker column) so one hot encoding is proceed by using get_dummies and converted in to numeric values

4. Good model with r^2 value:

- Multiple Linear Regression r^2 value= 0.78
- Support Vector Machine Regression r^2 value=0.75 (kernel=linear, c=1000)
- Decision Tree r^2 value=0.92 (crieterion=friedman_mse, splitter=random, max_features=sqrt)
- Random Forest r^2 value=0.92 (crieterion=absolute_error, max_features=sqrt, n_estimators=10)

5. Research values of all models:

1. MULTIPLE LINEAR REGRESSION: r^2 value= 0.78

2. SUPPORT VECTOR MACHINE:

S.No	Hyperparameter	Linear	rbf	poly	sigmoid
1	C=0.1	-0.07	-0.04	-0.04	-0.04
2	C=10	0.106	-0.02	-0.06	-0.05
3	C=100	0.65	-0.06	-0.07	-0.124
4	C=1000	0.75	-0.08	-0.03	-4.46
5	C=2000	0.64	-0.06	0.076	-16.34
6	C=10000	0.62	0.07	0.48	-367.25

The SVM Regression uses r^2 value (kernel-linear, hyperparameter C=1000) = 0.75

3. DECISION TREE:

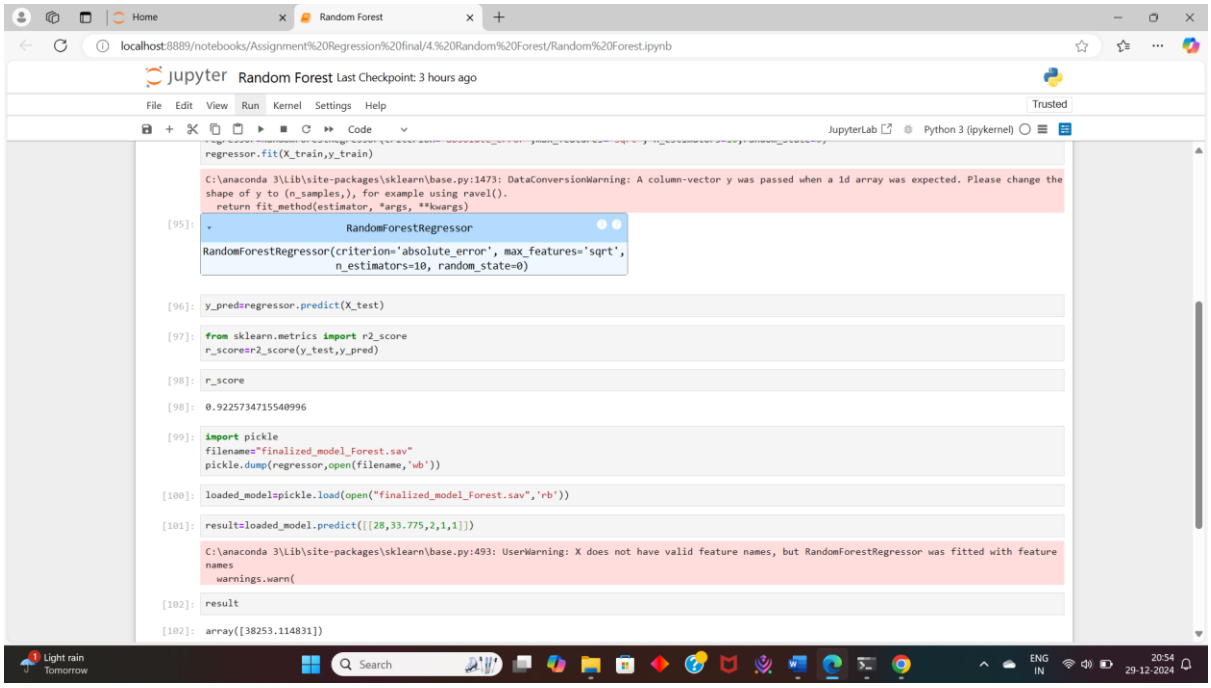
S.No	Crietrion	Splitter	Max_features	R2 Value
1	Squared_error	Best	Sqrt	0.82
2	Squared_error	Best	Log2	0.82
3	Squared_error	Random	Sqrt	0.82
4	Squared_error	Random	Log2	0.64
5	Friedman_mse	Best	Sqrt	0.82
6	Friedman_mse	Best	Log2	0.75
7	Friedman_mse	Random	Sqrt	0.92
8	Friedman_mse	Random	Log2	0.44
9	Absolute_error	Best	Sqrt	0.82
10	Absolute_error	Best	Log2	0.71
11	Absolute_error	Random	Sqrt	0.46
12	Absolute_error	Random	Log2	0.92
13	Poisson	Best	Sqrt	0.70
14	Poisson	Best	Log2	0.81
15	Poisson	Random	Sqrt	0.82
16	Poisson	Random	Log2	0.63

The Decision Tree Regression uses r^2 value (Crietrion= Friedman_mse , Splitter= Random, Max_features = Sqrt) = 0.92

4. RANDOM FOREST:

S.No	Crietrion	Max_features	N_Estimators	R2 Value
1	Squared_error	Sqrt	10	0.86
2	Squared_error	Sqrt	100	0.90
3	Squared_error	Log2	10	0.86
4	Squared_error	Log2	100	0.90
5	Friedman_mse	Sqrt	10	0.86
6	Friedman_mse	Sqrt	100	0.91
7	Friedman_mse	Log2	10	0.86
8	Friedman_mse	Log2	100	0.91
9	Absolute_error	Sqrt	10	0.92
10	Absolute_error	Sqrt	100	0.90
11	Absolute_error	Log2	10	0.92
12	Absolute_error	Log2	100	0.90
13	Poisson	Sqrt	10	0.91
14	Poisson	Sqrt	100	0.91
15	Poisson	Log2	10	0.91
16	Poisson	Log2	100	0.91

The Random Forest Regression uses r^2 value (Crietrion=Absolute_error, Max_features =Sqrt, n_estimators=10) = 0.92



6. Final model:

The final model I have chosen by using algorithm random forest regression with crietrion=absolute_error, max_features=sqrt, n_estimators=10 and found r^2 value is 0.92. Here the model has highest performance metrics and lowest complexity to the model