REGRESSION ASSIGNMENT

1(a) **INPUT**:

Age,Sex,Children,BMI,Smoker.

(b) OUTPUT:

Charges.

STAGE 1

Machine Learning

STAGE 2

Supervised Learning (Because input &output is clear)

STAGE 3

Regression (output is numerical)

2) BASIC INFORMATION ABOUT DATASET

No of Columns:6(age,sex,bmi,children,smoker,charges)

No of Rows:1339

3) FINDING CATEGORICAL DATA

Smoker,Sex – Nominal Data (Because we can't compare it)

4) **FINDING MODEL**

a) Multiple Linear Regeession:

Before Standardisation: r2_score=0.78947

After Standardization: r2_score=0.78947

b) Support Vector Machine:

Before Standardisation:

Kernel	**R2_score**
linear	-0.01116
rbf	-0.08842
sigmoid	-0.08994
poly	-0.06429

After Standardisation:

Kernel	**R2_score**
linear	-0.01012
rbf	-0.08338
sigmoid	-0.07542
poly	-0.07542

c) Decision Tree:

Before Standardisation:

Criterion	Splitter	R2_score
squared_error	best	0.69475
friedman_mse	best	0.69000
absolute_error	best	0.67017
poisson	best	0.72562
squared_error	random	0.70674
friedman_mse	random	0.73290

absolute_error	random	0.74475
poisson	random	0.67233

After Standardisation:

Criterion	Splitter	R2_score
squared_error	best	0.70029
friedman_mse	best	0.69164
absolute_error	best	0.67107
poisson	best	0.73331
squared_error	random	0.63009
friedman_mse	random	0.75793
absolute_error	random	0.69089
poisson	random	0.74346

d) Random Forest

Before Standardisation:

n_estimator	random_state	max_features	R2_score
50	0	sqrt	0.86958
50	0	log2	0.86958
100	0	sqrt	0.87102
100	0	log2	0.87102

After Standardisation:

n_estimator	random_state	max_features	R2_score
50	0	sqrt	0.86961
50	0	log2	0.86961
100	0	sqrt	0.8701
100	0	log2	0.8701

5) **FINAL MODEL**

My final model is RANDOM FOREST.

R2_SCORE=0.87102

Because, comparing Random Forest with other algorithm (Multiple Linear, Support Vector Machine, Decision Tree) the r2_score value is high.