1. Problem Statement :

Based on the Parameters like Age, Sex, Children,Bmi, Smoker . We need to predict the Insurance charge for the Person.

Stage 1 - Machine Learning (as data deals with numbers)
Stage 2 - Supervised learning (requirement and inputs/outputs are clear)
Stage 3 - Regression (final output is in numerical data)

2. Basic Information about Dataset :

The dataset has 1338 rows and 6 columns

The Input variables are Age, Sex, Bmi, Children, Smoker

The Ouput Variable is Insurance Charges

3. Pre Processing method used :

In the dataset the columns Smoker and Sex are categorical and it is Nominal Type . So One Hot Encoding is used to convert it into numerical

R2 Scores of the Models :

1. **Multiple linear Regression** : **0.789479**

2. Support Vector Machine

Hyper Parameter	linear	poly	rbf	sigmoid
C = 10	0.5665127	0.159392	-0.01810	0.073055
C = 100	0.6359503	0.750819	0.390601	0.52756
C = 500	0.76514321	0.859312	0.696468	0.490635
C = 1000	0.7440909	0.860584	0.828352	0.14377
C = 2000	0.74142388	0.860181	0.860735	-2.58403
C = 3000	0.74142316	0.860011	0.868531	-6.82618

SVM : For 'rbf' and C = 3000 the model gives **0.868531** as R score

3. Decision Tree :

criterion	max_features	splitter	R_Score
squared_error	-	Best	0.7129
friedman_mse	-	Best	0.6909
absolute_error	-	Best	0.6796
poisson	-	Best	0.7258
squared_error	-	random	0.7189

friedman_mse	-	random	0.7431
absolute_error	-	random	0.7551
poisson	-	random	0.6521
squared_error	sqrt	Best	0.7297
friedman_mse	sqrt	Best	0.7332
absolute_error	sqrt	Best	0.7434
poisson	sqrt	Best	0.7150
squared_error	sqrt	random	0.7420
friedman_mse	sqrt	random	0.7413
absolute_error	sqrt	random	0.7178
poisson	sqrt	random	0.6535
squared_error	*log2*	Best	0.7082
friedman_mse	*log2*	Best	0.7427
absolute_error	*log2*	Best	0.6822
poisson	*log2*	Best	0.7634
squared_error	*log2*	random	0.6907
friedman_mse	*log2*	random	0.6582
absolute_error	*log2*	random	0.6826
poisson	*log2*	random	0.6194

The Good R score is obtained for (Poisson,Log2,best) = **0.7634**

4. Random Forest

criterion	max_features	n_estimators	R_Score
squared_error	*log2*	50	0.8667
friedman_mse	*log2*	50	0.8653
absolute_error	*log2*	50	0.8654
poisson	*log2*	50	0.8670
squared_error	*log2*	100	0.8698
friedman_mse	*log2*	100	0.8681
absolute_error	*log2*	100	0.8657
poisson	*log2*	100	0.8687
squared_error	sqrt	50	0.8667
friedman_mse	sqrt	50	0.8653
absolute_error	sqrt	50	0.8653
poisson	sqrt	50	0.8670
squared_error	sqrt	100	0.8698
friedman_mse	sqrt	100	0.8681
absolute_error	sqrt	100	0.8657
poisson	sqrt	100	0.8687

RF The Model performed well in (Squared_error, sqrt, 100) = **0.8698**

The final model that can be chosen is

SVM (‘rbf’ , C = 3000) = **0.868531 or**

RF (Squared_error, sqrt, 100) = **0.8698**