REGRESSION ASSIGNMENT

1. Problem statement:-

- **Stage 1: Machine Learning**
- **Stage 2: supervised learning**
- **Stage 3: SVMR**

Details about the dataset:

- The dataset has 1338 rows and 6 columns
- The column sex and smoker are converted into string to nominal data using one hot encoding method
- Pd.get_dummies method is used to convert string data into nominal data

1. Multiple Linear regression = `0.7894790349867009`

2. SVM

kernel	c	r2_score
rbf	1000	0.8102064851758545
linear	1000	0.7649311738597411
linear	10000	0.7414230132360546
poly	1000	0.8566487675946572
poly	1000	0.8591715079473907
rbf	10000	0.8779952401449918
sigmod	1000	0.28747069486976173

3. DECISION TREE

criterion	***splitter***	r2_score
squared_error	***best***	0.6813978163001406
squared_error	random	0.6600493335731996
friedman_mse	random	0.6891555884503506
friedman_mse	best	0.6855447084196503
absolute_error	best	0.6738903616744885
absolute_error	random	0.6638409563638938
poisson	random	0.6837672255430303
poisson	best	0.7242297411765548

4. Random forest = 0.03073750386998919

FINAL MODEL:

I have chosen the Support Vector Regression (SVR) model with an RBF kernel and C = 10000, as it achieved the highest R² score (0.87799) compared to all other algorithms tested. This indicates that the model fits the data well and provides more accurate predictions for insurance charges.