Problem Statement & Solution

1.Identifying the problem - “Predicting insurance charges”

Stage of Problem Identification

- 1. Machin Learning
- 2. Supervised Learning
- 3. Regression
- 4. Multiple Linear Regression.

2. Basic info about the dataset.

- X = age, sex, bmi, children, smoker
- Y = charges
- The total number of rows = 1338
- The total number of columns = 6 (including the output column)

3. Pre-Processing method

The dataset I received from the client includes both

- numerical values (age, BMI, children, charges) and
- categorical values (sex, smoker).

The categorical data (sex, smoker) is nominal, meaning it does not possess any inherent ranking or order.

#To address this, I utilized one-hot encoding along with the “drop_first=True” function to minimize duplications within the columns.

4. Developing a good model with r^2_score. Below listed algorithm used

- MLR
- SVM
- Decision Tree
- Random Forest

5. All the research values (r2_score of the models)

1. Multiple Linear Regression

r2_score is = 0.7894

2. Support Vector Machine Regression:

S.NO	Hyper Parameter	**Liner** (r value)	Hyper Parameter	**RBF** (r value)
1	C10	0.46246	C10	-0.32273
2	C100	0.62887	C100	0.32003
3	C500	0.76493	C500	0.66429
4	C700	0.76509	C900	0.79459
5	C800	0.76488	C1000	0.81020
6	C1000	0.76493	C10000	0.87799

S.NO	Hyper Parameter	**POLY** (r value)	Hyper Parameter	**SIGMOID** (r value)
1	C10	0.03871	C10	0.03930
2	C100	0.61795	C70	0.48494
3	C500	0.82636	C80	0.50790
4	C900	0.85476	C100	0.52761
5	C1000	0.85664	C150	0.53520
6	C10000	0.85917	C200	0.54555

The good model for SVM is a Poly model with the hyperparameter C set to 10000 resulting in an r2_score is = 0.87799

3. Decision Tree:

S.NO	Criterion	Max Features	Splitter	r Value
1	Squared_error (mse)	Auto	Best	0.6871
2	Squared_error (mse)	Auto	Random	0.7193
3	Squared_error (mse)	Sqrt	Best	0.7308
4	Squared_error (mse)	Sqrt	Random	0.6176
5	Squared_error (mse)	Log2	Best	0.6842
6	Squared_error (mse)	Log2	Random	0.6738
7	Absolute_error (mae)	Auto	Best	0.6884
8	Absolute_error (mae)	Auto	Random	0.7351
9	Absolute_error (mae)	Sqrt	Best	0.7094
10	Absolute_error (mae)	Sqrt	Random	0.6745
11	Absolute_error (mae)	Log2	Best	0.7055
12	Absolute_error (mae)	Log2	Random	0.6533
13	Friedman_mse	Auto	Best	0.6867
14	Friedman_mse	Auto	Random	0.6832
15	Friedman_mse	Sqrt	Best	0.7183
16	Friedman_mse	Sqrt	Random	0.6026
17	Friedman_mse	Log2	Best	0.6550
18	Friedman_mse	Log2	Random	0.6234

The good model for Decision Tree with mean absolute error (MAE) as the evaluation metric and automatic feature selection using the Random algorithm achieved an r2 score of 0.7351

4. Random Forest

When using the Random Forest with n_estimators = 100000, criterion='mse', max_features='auto', and random_state = 0, I obtained an R value of 0.85546

6. Final saved model.

Compared to all the other models, I would recommend using a Polynomial Support Vector Machine (SVM) with the hyperparameter C set to 10000. This particular model achieved an R2 score of 0.8799, making it a strong performer.