1) Problem Identification

Stage 1 : Domain Selection

Input – Number

Domain – Machine Learning

Stage 2 : Learning Selection

Supervised

Stage 3 : Classification or Regression

Here in the dataset the output label contains **numerical values**. So it falls under **Regression**.

2) Basic Information about Dataset

File Name : **insurance_pre.csv**

Total number of Rows : **1338**

Total number of Columns : **6**

Input Columns : **5 [age,sex, bmi,children,smoker]**

Output Columns : **1 [charges]**

3) Pre processing Method

Here in the dataset we have **two columns** as **nominal data [sex, smoker].**

4) Machine Learning Algorithms

For this section refer created python files with many models

5) Research Values (R^2 score of the models)

A. MULTIPLE LINEAR REGRESSION: (R^2 Value) = 0.7894

B. SUPPORT VECTOR MACHINE:

S.NO	HYPER PARAMETER	LINEAR (R^2 Value)	RBF (R^2 Value)	POLY (R^2 Value)	SIGMOID (R^2 Value)
1	C=10	0.5665	-0.0181	0.1593	0.0730
2	C=100	0.6359	0.3906	0.7508	0.5275
3	C=500	0.7651	0.6964	0.8593	0.4906
4	C=1000	0.7440	0.8283	0.8605	0.1437

5	C=2000	0.7414	0.8607	0.8601	-2.5840
6	C=3000	0.7414	0.8685	0.8600	-6.8261
7	C=5000	0.7414	0.8735	0.8588	-17.5541
8	C=10000	0.7414	0.8774	0.8582	-82.1902

In SVM (R^2 Value) = 0.8774 [When HYPER PARAMETER C=10000 and Kernel='RBF']

C. DECISION TREE:

S.NO	CRITERION	MAX FEATURES	SPLITTER	R^2 Value
1	friedman_mse	Auto	best	0.6814
2	friedman_mse	Auto	random	0.7037
3	friedman_mse	Sqrt	best	0.6987
4	friedman_mse	Sqrt	random	0.7017
5	friedman_mse	log2	best	0.6732
6	friedman_mse	log2	random	0.6560
7	squared_error	Auto	best	0.7080
8	squared_error	Auto	random	0.7444
9	squared_error	Sqrt	best	0.7048
10	squared_error	Sqrt	random	0.6661
11	squared_error	log2	best	0.7399
12	squared_error	log2	random	0.7076
13	absolute_error	Auto	best	0.6906
14	absolute_error	Auto	random	0.6881
15	absolute_error	Sqrt	best	0.7675
16	absolute_error	Sqrt	random	0.7196
17	absolute_error	log2	best	0.7146
18	absolute_error	log2	random	0.7104
19	poisson	Auto	best	0.6884
20	poisson	Auto	random	0.6977
21	poisson	Sqrt	best	0.5766
22	poisson	Sqrt	random	0.7360
23	poisson	log2	best	0.5609
24	poisson	log2	random	0.6530

In Decision Tree (R^2 Value) = 0.7675 [When CRITERION=' absolute_error', MAX FEATURES='sqrt' and SPLITTER='best']

D.RANDOM FOREST:

S.NO	CRITERION	MAX FEATURES	N_ESTIMATORS	R^2 Value
1	friedman_mse	None	50	0.8600
2	friedman_mse	None	100	0.8540
3	friedman_mse	Sqrt	50	0.8702
4	friedman_mse	Sqrt	100	0.8678
5	friedman_mse	log2	50	0.8671
6	friedman_mse	log2	100	0.8707
7	squared_error	None	50	0.8491
8	squared_error	None	100	0.8584
9	squared_error	Sqrt	50	0.8708
10	squared_error	Sqrt	100	0.8705
11	squared_error	log2	50	0.8688
12	squared_error	log2	100	0.8704
13	absolute_error	None	50	0.8526
14	absolute_error	None	100	0.8563
15	absolute_error	Sqrt	50	0.8691
16	absolute_error	Sqrt	100	0.8738
17	absolute_error	log2	50	0.8671
18	absolute_error	log2	100	0.8715
19	poisson	None	50	0.8351
20	poisson	None	100	0.8316
21	poisson	Sqrt	50	0.8274
22	poisson	Sqrt	100	0.8216
23	poisson	log2	50	0.8216
24	poisson	log2	100	0.8263

In Random Forest (R^2 Value) = 0.8738 [When CRITERION=' absolute_error', MAX FEATURES='sqrt' and N_ESTIMATORS=100]

6) Final Model

In MLR (R^2 Value) - 0.7894

In **SVM (R^2 Value)** - **0.8774**

In DT (R^2 Value) - 0.7675

In RF (R^2 Value) - 0.8738

So from the above analysis we can choose the **Support Vector Machine (SVM) algorithm** for the **final model.**