1.) Problem statement

Stage 1.Domain--Machine Learning

Stage 2.Learning method--Supervised

Stage 3.Classification/Regression-- Regression

2.) Basic info about the dataset

Total number of rows -1338
columns-6

3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

We are going to use Machine learning domain. ML algorithm we have to pass all input as number. In this problem statement we are going to convert **sex and smoker** column string values as number-nominal data.

4.)To find following the machine learning regression method using in r2 value

1.Multiple Linear Regression(R^2 value)= 0.78947

2.Support Vector Machine:

S.NO	Hyper Parameter	Linear(r value)	RBF(Non Linear) (r value)	POLY (r value)	SIGMOID (r value)
1	C10	0.56651	-0.01810	0.15939	0.07305
2	C100	0.63595	0.39060	0.75081	0.52756
3	C500	0.76514	0.69646	0.85931	0.49063
4	C1000	0.74409	0.82835	0.86058	0.143775
5	C2000	0.74142	0.86073	0.86018	-2.58403
6	C3000	0.74142	0.86853	0.86001	-6.82618

The **SVM Regression** use **R^2 value** (nonlinear (RBF) and hyper parameter (C3000))=0.86853

3.Decision Tree:

S.NO	CRITERION	MAX FEATURES	SPLITTER	R VALUE
1	Mse	auto	best	0.69873
2	Mse	auto	random	0.67270
3	Mse	sqrt	best	0.71419
4	Mse	sqrt	random	0.67173
5	Mse	Log2	best	0.64344
6	Mse	Log2	random	0.59276
7	Mae	auto	best	0.69015
8	Mae	auto	random	0.73781
9	Mae	sqrt	best	0.73385
10	Mae	sqrt	random	0.67806
11	Mae	Log2	best	0.63417
12	Mae	Log2	random	0.66638
13	Friedman_mse	auto	best	0.70494
14	Friedman_mse	auto	random	0.73555
15	**Friedman_mse**	**sqrt**	**best**	**0.74989**
16	Friedman_mse	sqrt	random	0.64825
17	Friedman_mse	Log2	best	0.69597
18	Friedman_mse	Log2	random	0.60541

The **Decision Tree** Regression use **R^2 value** (Friedman,sqrt,best)=0.74989

4. Random Forest

SI.NO	CRETERION	MAX FEATURES	N_ESTIMATORS	R VALUE
1	Mse	auto	10	0.83926
2	Mse	auto	100	0.85553
3	Mse	sqrt	10	0.85191
4	Mse	sqrt	100	0.86861
5	Mse	Log2	10	0.86211
6	Mse	Log2	100	0.87033
7	Mae	auto	10	0.85430
8	Mae	auto	100	0.85197
9	Mae	sqrt	10	0.85755
10	Mae	sqrt	100	0.87327
11	Mae	Log2	10	0.86010
12	Mae	Log2	100	0.86917
13	Friedman_mse	auto	10	0.83611
14	Friedman_mse	auto	100	0.85786
15	Friedman_mse	sqrt	10	0.86210
16	Friedman_mse	sqrt	100	0.86974
17	Friedman_mse	Log2	10	0.85557
18	Friedman_mse	Log2	100	0.86977

The **Random Forest** Regression use **R^2 value** (Mae,Sqrt, 100)=0.87327

4. AdaBoost

SI.NO	LEARNING_RATE	LOSS	N_ESTIMATORS	R VALUE
1	1.0	linear	10	0.84474
2	1.0	linear	50	0.84474
3	1.0	square	10	0.73062
4	1.0	square	50	0.50780
5	1.0	exponential	10	0.82667
6	1.0	exponential	50	0.62928

The **AdaBoost** Regression use **R^2 value** (1.0,linear, 10 and 50)=0.84474

5. XG Boosting

SI.NO	CRETERION	MAX FEATURES	N_ESTIMATORS	R VALUE
1	Friedman_mse	sqrt	10	0.61652
2	Friedman_mse	sqrt	50	0.89039
3	Friedman_mse	sqrt	100	0.89007
4	Friedman_mse	Log2	10	0.61652
5	Friedman_mse	Log2	50	0.89039
6	Friedman_mse	Log2	100	0.89007

The **XG Boosting** Regression use **R^2 value** (Friedman_mse,sqrt and log2, 50)=0.89039

6. Light Gradient Boosting (LightGBM)

SI.NO	Num_Leaves	Learning_Rate	N_ESTIMATORS	R VALUE
1	31	0.1	10	0.78574
2	50	0.1	50	0.87574
3	100	0.1	100	0.86515
4	500	0.1	10	0.78613
5	1000	0.1	50	0.87574
6	10000	0.1	100	0.86515

The **LightGBM** Regression use **R^2 value** (50 and 1000,0.1,50)=0.87574

5.) In the tabulation format, all the research values (r2_score of the models) documented.

6.)Developed a good model with r2_score. I have used "XG Boosting" machine learning algorithm to create final model. We have used all machine learning algorithm to test this dataset. Finally for this dataset " XG Boosting " algorithm only provided almost **0.89039** accuracy.