Problem Statement:

1. Identify Problem statement: 3 stages
   1. Machine Learning
   2. Supervised Learning (Requirement clear & I/p and O/P are also clear)
   3. Regression – O/P labels are in numerical format

2. Tell basics about dataset:

Dataset have 5 input column (age,sex,children,bmi and smoker) and 1 output (charges). The columns (sex and smoker) are in categorical value, so we have to do some preprocessing.

3.PreProcessing method:

As mentioned above dataset have categorical value as input and it is nominal type , so we converting to 1 and 0. For that we using "One Hot Encoding" in the code.

4. Develop Model:

Developed MultipleLinearRegression, SVM, DecisionTree,RandomFactor to find the best model, whcich shows below.

5. Best Model

As per report of r_score value, Random Forest has chosen as best model and created deployment phase for the respective model. Why, because the r_score value is given as high accuracy for the given dataset, when compared to the other model.

Please find the report below,

Multi Linear Regression:

R_score value is 0.78

1.SVM – Support Vector Machine

Kernel	*C*	*R_score*
linear	0 10 100 1000 10000	-0.01 0.46 0.62 0.76 0.74
rbf	0 10 100 1000 10000	-0.08 -0.03 0.32 0.81 0.870
poly	0 10 100 1000 10000	-0.07 0.03 0.61 0.85 0.85
sigmoid	0 10 100 1000 10000	-0.07 0.03 0.52 0.28 -34.15

Hyper tuning parameter in SVM is kernel=”rbf”,c=10000 for given dataset

2. Decision Tree

criterion	*splitter*	*R_score*
squared_error (default)	best(default)	0.69
	Random	0.74
friedman_mse	best	0.69
	Random	0.68
absolute_error	best	0.67
	Random	0.72
poisson	best	0.72
	Random	0.71

Hyper tuning parameter in Decision tree is criterion =” squared_error”, splitter=random for given dataset

3. Random Forest

criterion	*max_features*	*R_score*
squared_error	sqrt	0.872
	log2	0.866
friedman_mse	sqrt	0.862
	Log2	0.871
absolute_error	Sqrt	0.870
	Log2	0.871
poisson	Sqrt	0.871
	Log2	0.870

Hyper tuning parameter in Decision tree is criterion =”squared_error”, max_features=sqrt for given dataset

By analysing the above hyper tuning report **RandomForest** given **high** accuracy when compared to the other model for the given data set.
So we **saving RandomForest model** for the deployment

4. Gradient Boosting

criterion	*loss*	*R_score*
squared_error (default)	absolute_error	0.80
	Squared_error	0.78
	quantile	0.63
	huber	0.86

friedman_mse	absolute_error	0.83
	Squared_error	0.76
	quantile	0.63
	huber	0.78

XGBoosting Algorithm:

R_score value = 0.866

LGBoosting Algorithm:

R_score value = 0.86