1) Identify your problem statement

Predict insurance charge using age,BMI,children,sex and smoker as input feature

Domain selection: Data is numeric so we use machine learning
Learning selection; Input and output are defined and requirement are clear so we use Supervised learning
3rd stage: output is numerical data so we need to use regression

Its opt for Supervised learning regression in machine learning

2.) Tell basic info about the dataset (Total number of rows, columns)
5 input and 1 output
Rows:1338
Column: 6

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

In this dataset Sex and Smoker input feature has nominal data in categorical side so we need to convert as number using one hot encoding.after that we need to remove the dummies using drop_first parameter because machine learning algorithm wont accept dummy data

4.) Develop a good model with r2_score. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

For checking the good model with r2_score we need use all algorithm for regression in machine learning which one its giving higher score we can deploy it on the model.

In this data set we have more than one input feature so we need to chekc below algorithm

1. Multiple Linear Regression
2. Support Vector Machine Regression
3. Decision Tree Regression
4. Random Forest Regression

5.) All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)

To find the machine learning r2score value with hypertuning parameter

SVMR

R2 score value in SVMR algorithm is 0.761

C is a penalty	***linear***	***Poly***	***rbf***	***sigmoid***
10	**-0.040**	**-0.120**	**-0.095**	**-0.09**
100	**0.521**	**-0.0139**	**-0.155**	**-0.124**
1000	**0.6188**	**-0.092**	**-0.149**	**-1.521**
10000	**0.761**	**0.303**	**-0.055**	**-109**

Decision Tree

R2score value in Decision Tree is 0.908

criterion	**splitter**	**max_features**	**R2 Score**
squared_error	**best**	**sqrt**	**0.662**
squared_error	**random**	**sqrt**	**0.502**

squared_error	best	log2	0.718
squared_error	random	log2	0.700
friedman_mse	best	sqrt	0.605
friedman_mse	random	sqrt	0.671
friedman_mse	best	log2	0.742
friedman_mse	random	log2	0.673
absolute_error	best	sqrt	0.748
absolute_error	random	sqrt	0.464
absolute_error	best	log2	0.663
absolute_error	random	log2	0.720
poisson	best	sqrt	0.694
poisson	random	sqrt	0.659
poisson	best	log2	0.745
poisson	random	log2	0.694

Random Forest

R2 score value in random forest is 0.873

n_estimators	*criterion*	*max_features*	r2score
10	*squared_error*	sqrt	0.849
100	*squared_error*	sqrt	0.869
10	*squared_error*	log2	0.851
100	*squared_error*	log2	0.867
10	*absolute_error*	sqrt	0.861
100	*absolute_error*	sqrt	0.873
10	*absolute_error*	log2	0.864

100	***absolute_error***	**log2**	**0.872**
10	***friedman_mse***	**sqrt**	**0.864**
100	***friedman_mse***	**sqrt**	**0.872**
10	***friedman_mse***	**log2**	**0.849**
100	***friedman_mse***	**log2**	**0.869**
10	***poisson***	**sqrt**	**0.849**
100	***poisson***	**sqrt**	**0.871**
10	***poisson***	**log2**	**0.855**
100	***poisson***	**log2**	**0.868**

6.) Mention your final model, justify why u have chosen the same.

Using r2_score final model will be random forest regression because its giving highest value compare to others.

r2_score=0.873

Typically, model achieves over 90% but in this dataset highest is 87% making it an average model