Problem Statement:

A client's requirement is, he wants to predict the insurance charges based on several parameters. The Client has provided the dataset of the same. we must develop a model which will predict the insurance charges.

1.Identify your problem statement:

Stage 1: I identify that the given problem statement comes under **Machine Learning - Domain.** Clients provide datasets for prediction, the datasets are major in numerical values, so the machine learning domain is suitable for this problem.

Stage 2: Learning Selection

Based on the dataset and requirement from client is clear ,so it comes under '**Supervised Learning'**

Stage 3: dataset contain numerical values ,so its comes under **Regression**

2. basic info about the dataset:

The dataset has clear inputs and output for the model creation.
The dataset contains **6** columns and **1339** rows. The Inputs columns are 'Age','Sex','BMI','Children','Smoker' and the output column is 'Charges'.

3.Pre-processing method of data:

Except 'Sex' and 'Smoker' columns all other columns are in numerical values.so i convert the categorical data into numerical values . It is Nominal data so i used One hot encoding by get_dummies using pandas in python library.

4.Developed Models:

Here i using a machine learning algorithms to develop the varies model for this problems with r2_score results are:

- **Multiple Linear Regression: R^2 value is =0.7894**
- **Support Vector Machine:R^2 value is = 0.8663**
- **Decision Tree: R^2 value=0.78115**
- **Random Forest :R^2 value=0.87197**

5.R^2 value results of all theModels:

- **Multiple Linear Regression:**

  In this MLR based model prediction, the predicted R^2 value is =0.7894

- **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	0.4624	-0.0322	0.0387	0.0393
2	C=100	0.6288	0.3200	0.6179	0.5276
3	C=500	0.7631	0.6642	0.8263	0.4446
4	C=1000	0.7649	0.8102	0.8566	0.2874
5	C=1500	0.7440	0.8427	0.8580	-0.0674
6	C=2000	0.7440	0.8547	0.8605	-0.5939
7	C=3000	0.7414	0.8663	0.8598	-2.1244

In this SVM - Regression model, The predicted high accuracy R^2 value is = 0.8663

Used Parameters :(Kernel: rbf , C=3000)

- **Decision Tree:**

S.NO	CRITERION	MAX Features	SPLITTER	R^2 VALUE
1	squared_error	none	best	0.71540
2	squared_error	none	random	0.73014
3	squared_error	sqrt	random	0.78115

4	**squared_error**	**sqrt**	**best**	**0.72064**
5	**squared_error**	**log2**	**best**	**0.70206**
6	**squared_error**	**log2**	**random**	**0.6573**
7	**friedman_mse**	**none**	**best**	**0.70089**
8	**friedman_mse**	**none**	**random**	**0.74932**
9	**friedman_mse**	**sqrt**	**random**	**0.66154**
10	**friedman_mse**	**sqrt**	**best**	**0.76170**
11	**friedman_mse**	**log2**	**best**	**0.7125**
12	**friedman_mse**	**log2**	**random**	**0.7026**
13	**absolute_error**	**None**	**best**	**0.66141**
14	**absolute_error**	**None**	**random**	**0.78018**
15	**absolute_error**	**sqrt**	**random**	**0.77669**
16	**absolute_error**	**sqrt**	**best**	**0.63591**
17	**absolute_error**	**log2**	**best**	**0.55479**
18	**absolute_error**	**log2**	**random**	**0.581028**
19	**poisson**	**None**	**best**	**0.728146**
20	**poisson**	**None**	**random**	**0.70591**
21	**poisson**	**sqrt**	**random**	**0.67737**
22	**poisson**	**sqrt**	**best**	**0.61713**
23	**poisson**	**log2**	**best**	**0.412009**
24	**poisson**	**log2**	**random**	**0.64301**

Decision Tree - Regression ,acceptable High R^2 value=0.78115

Used Parameters :(CRITERION: squared_error, max_features: sqrt , splitter:random).

- **RandomForest:**

S.NO	CRITERION	MAX Features	N_Estimators	R^2 VALUE
1	squared_error	none	50	0.85091
2	squared_error	none	100	0.85495
3	squared_error	sqrt	50	0.8696
4	squared_error	sqrt	100	0.87081
5	squared_error	log2	50	0.86961
6	squared_error	log2	100	0.87081
7	friedman_mse	none	50	0.85111
8	friedman_mse	none	100	0.85514
9	friedman_mse	sqrt	50	0.87023
10	friedman_mse	sqrt	100	0.87086
11	friedman_mse	log2	50	0.87023
12	friedman_mse	log2	100	0.870861
13	absolute_error	None	50	0.85412
14	absolute_error	None	100	0.8531
15	absolute_error	sqrt	50	0.87168
16	absolute_error	sqrt	100	0.87197
17	absolute_error	log2	50	0.87168
18	absolute_error	log2	100	0.87197
19	poisson	None	50	0.85032
20	poisson	None	100	0.85358
21	poisson	sqrt	50	0.86320
22	poisson	sqrt	100	0.86775
23	poisson	log2	50	0.86320
24	poisson	log2	100	0.86775

Random Forest - Regression ,acceptable High **R^2 value=0.87197**

Used Parameters :(Criterion: absolute_error, max_features: log2 , N_Estimator:100)

6.Final Model:

The final model I choose for this problem (Insurance prediction) is **Random Forest** algorithm based model .Because it gives a high R score value, the accuracy of prediction is higher than the other models. So I created the deployment phase for the Random Forest Model. The parameters used are criterion: **absolute_error, max_features: log2 , n_estimator:100** and the R^2 value is **0.87197**.