<u>**PROBLEM STATEMENT**</u>

As Data Scientists, we must develop a model to predict insurance charges when age, sex, BMI, children and smoker values are given as input parameters.

<u>**Three stages of Problem Identification**</u>:

Stage 1- Domain Selection- Machine learning

As the input (Insurance Dataset) contains numerical values, we can choose the ***Machine Learning Domain***

Stage 2-Learning – Supervised Learning

As the Requirement is clear (predict insurance for the given input) we can choose ***Supervised Learning***

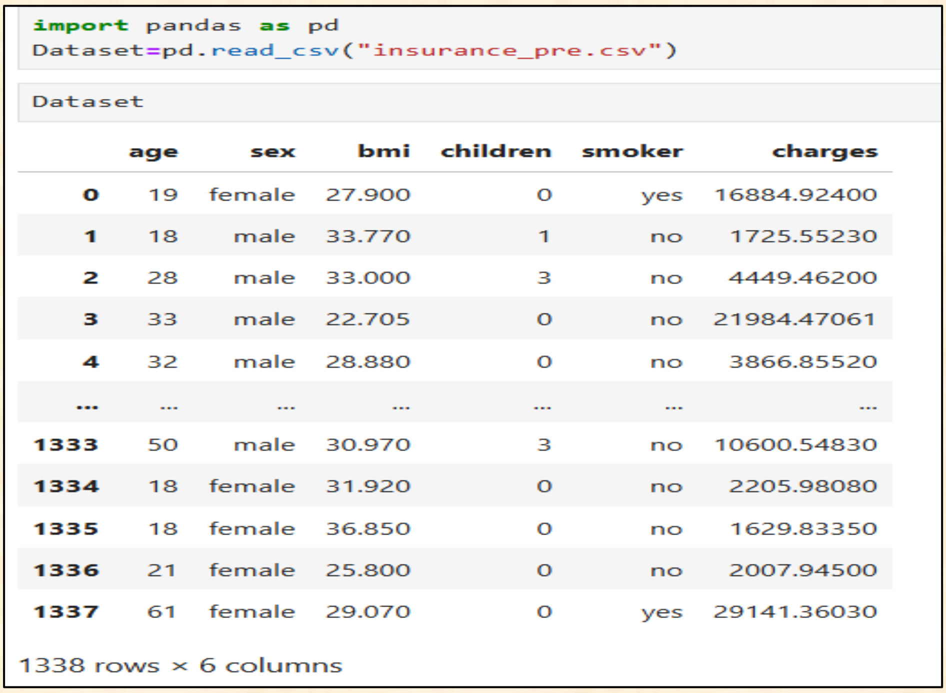
Stage 3 – Regression:

Prediction related to number (continuous value) so we can choose the ***Regression type***

<u>**DATASET INFORMATION**</u>

Total No of rows: 1338 (including column name)

Total No of columns: 6



```
import pandas as pd
Dataset=pd.read_csv("insurance_pre.csv")
```

```
Dataset
```

	age	sex	bmi	children	smoker	charges
0	19	female	27.900	0	yes	16884.92400
1	18	male	33.770	1	no	1725.55230
2	28	male	33.000	3	no	4449.46200
3	33	male	22.705	0	no	21984.47061
4	32	male	28.880	0	no	3866.85520
...	...	...	...	...	...	...
1333	50	male	30.970	3	no	10600.54830
1334	18	female	31.920	0	no	2205.98080
1335	18	female	36.850	0	no	1629.83350
1336	21	female	25.800	0	no	2007.94500
1337	61	female	29.070	0	yes	29141.36030

1338 rows × 6 columns

Column Name: age, sex, bmi,children,smoker,charges

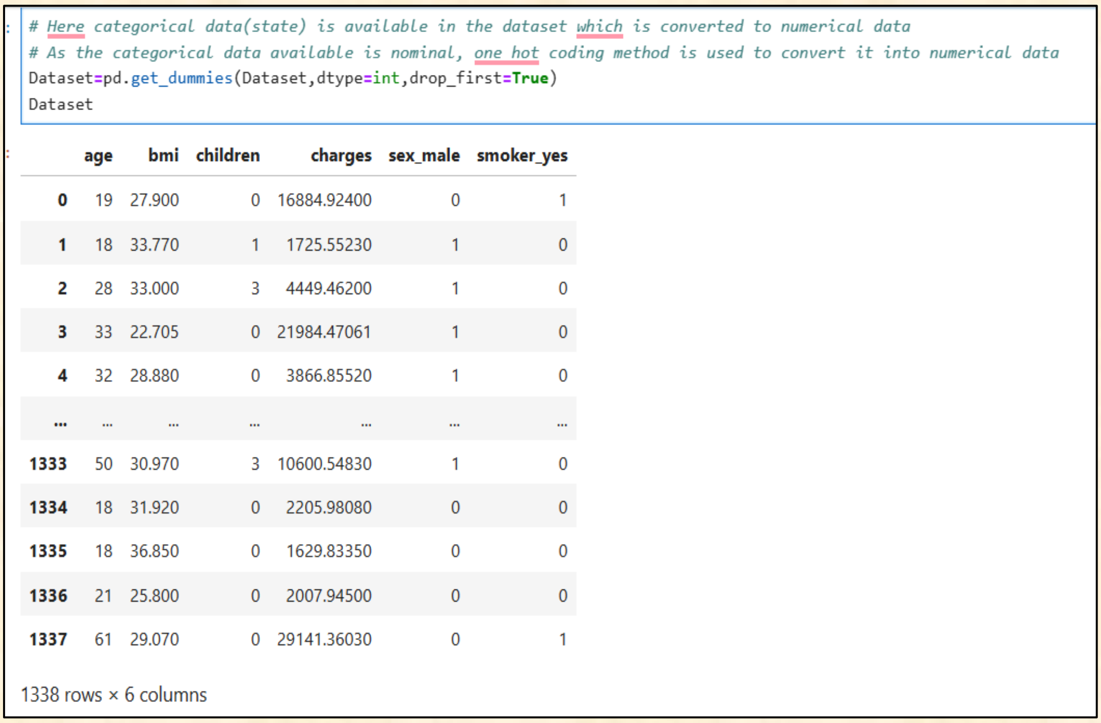
Input Variables: age,sex,bmi,children,smoker

Output Variable: Charges

PREPROCESSING METHOD:

Since the Dataset has categorical columns sex and smoker, it should be converted to numbers

Here the categorical data is nominal, so one hot encoding method is used to convert the categorical data (string) to numerical data (numbers 1 or 0)



```
# Here categorical data(state) is available in the dataset which is converted to numerical data
# As the categorical data available is nominal, one hot coding method is used to convert it into numerical data
Dataset=pd.get_dummies(Dataset,dtype=int,drop_first=True)
Dataset
```

```
      age     bmi  children      charges  sex_male  smoker_yes
   0   19  27.900         0  16884.92400         0           1
   1   18  33.770         1   1725.55230         1           0
   2   28  33.000         3   4449.46200         1           0
   3   33  22.705         0  21984.47061         1           0
   4   32  28.880         0   3866.85520         1           0
 ...  ...     ...       ...          ...       ...         ...
1333   50  30.970         3  10600.54830         1           0
1334   18  31.920         0   2205.98080         0           0
1335   18  36.850         0   1629.83350         0           0
1336   21  25.800         0   2007.94500         0           0
1337   61  29.070         0  29141.36030         0           1

1338 rows × 6 columns
```

1. **MULTIPLE LINEAR REGRESSION:**

   The R^2 value of ***Multiple Linear Regression*** is **0.7894**

2. **SUPPORT VECTOR MACHINE:**

S.NO	HYPERTUNING PARAMETER			KERNEL TYPE			
	C	Max_iter	EPSILON	LINEAR (r score)	RBF (r score)	POLY (r score)	SIGMOID (r score)
1	1.0	-1	0.1	-0.0101	-0.0833	-0.0756	0.0393
2	10	-1	0.5	-0.0016	-0.0322	0.0387	0.0393
3	100	-1	1.0	0.6288	0.3200	0.6179	0.5270
4	500	1000	0.1	0.7622	0.6642	0.8260	0.4446
5	1000	2000	0.5	0.7646	0.8102	0.8566	0.2874
6	2000	4000	1.0	0.7438	0.8547	0.8604	-0.5939
7	3000	6000	0.1	0.7422	0.8663	0.8599	-2.12441
8	4000	8000	0.5	0.7410	0.8717	0.8604	-5.5103
9	6000	12000	1.0	0.7426	0.8767	0.8592	-12.9896
10	10000	20000	0.1	0.7433	0.8779	0.8590	-34.1515

Support Vector machine has highest $\boldsymbol{R^2}$ **value 0.8779** for kernel_type=”rbf”, c=10000,max_iter=20000,epsilon=0.1

3. DECISION TREE:

S.No	CRITERION	MAX FEATURES	SPLITTER	Random_state	Max_depth	R SCORE
1	Squared_error	None	best	0	4	0.8837
2	Squared_error	None	random	None	6	0.8683
3	Squared_error	Sqrt	best	0	4	0.8475
4	Squared_error	Sqrt	random	None	6	0.8382
5	Squared_error	log2	best	0	4	0.8475
6	Squared_error	log2	random	None	6	0.7980
7	friedman_mse	None	best	0	4	0.8837
8	friedman_mse	None	random	None	6	0.8490
9	friedman_mse	Sqrt	best	0	4	0.8475
10	friedman_mse	Sqrt	random	None	6	0.7757
11	friedman_mse	log2	best	0	4	0.8475
12	friedman_mse	log2	random	None	6	0.7880
13	absolute_error	None	best	0	4	0.8823
14	absolute_error	None	random	None	6	0.8562
15	absolute_error	Sqrt	best	0	4	0.8444
16	absolute_error	Sqrt	random	None	6	0.7591
17	absolute_error	log2	best	0	4	0.8444
18	absolute_error	log2	random	None	6	0.7890
19	Poisson	None	best	0	4	0.8847
20	Poisson	None	random	None	6	0.8857
21	Poisson	Sqrt	best	0	4	0.8383
22	Poisson	Sqrt	random	None	6	0.8096
23	Poisson	log2	best	0	4	0.8383
24	Poisson	log2	random	None	6	0.7081

Decision Tree has highest $\boldsymbol{R^2}$ **value 0.8857** for max_feature=None, criterion=poisson, splitter='random', random_state=None and max_depth=6

4. RANDOM FOREST:

S.No	CRITERION	MAX FEATURES	n_estimators	Max_depth	Random_state	R SCORE
1	Squared_error	None	100	4	0	0.8897
2	Squared_error	None	50	5	None	0.8844
3	Squared_error	Sqrt	100	4	0	0.8565
4	Squared_error	Sqrt	50	5	None	0.8765
5	Squared_error	log2	100	4	0	0.8565
6	Squared_error	log2	50	5	None	0.8850
7	friedman_mse	None	100	4	0	0.8897
8	friedman_mse	None	50	5	None	0.8827
9	friedman_mse	Sqrt	100	4	0	0.8565
10	friedman_mse	Sqrt	50	5	None	0.8817
11	friedman_mse	log2	100	4	0	0.8565
12	friedman_mse	log2	50	5	None	0.8727
13	absolute_error	None	100	4	0	0.8858
14	absolute_error	None	50	5	None	0.8881
15	absolute_error	Sqrt	100	4	0	0.8395
16	absolute_error	Sqrt	50	5	None	0.8712
17	absolute_error	log2	100	4	0	0.8395
18	absolute_error	log2	50	5	None	0.8681
19	Poisson	None	100	4	0	0.8885
20	Poisson	None	50	5	None	0.8830
21	Poisson	Sqrt	100	4	0	0.8567
22	Poisson	Sqrt	50	5	None	0.8761
23	Poisson	log2	100	4	0	0.8567
24	Poisson	log2	50	5	None	0.8812

Random Forest Tree has highest R^2 **value 0.8897** for max_feature=None, max_depth=4, criterion= squared_error or friedman_mse, n_estimators=100, random_state=0

MODEL SELECTION:

Out of all the above models***, Random Forest*** is selected as the ***Best model*** as it has highest R^2 **value 0.8897,** so it is saved for the *deployment phase*