Problem Statement or Requirement:

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

1.) Identify your problem statement
2.) Tell basic info about the dataset (Total number of rows, columns)
3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal 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.
5.) All the research values (r2_score of the models) should be documented. (You can make tabulation or screenshot of the results.)
6.) Mention your final model, justify why u have chosen the same.

1) Identify you Problem Statement:

Here we going to predict the value so it comes under **regression**

Data are in structure form -> **ML** algo enough

Data comes under **supervised** learning --> due to it label data (charges act as depended variable of independed variable smoker, children, bmi, sex, age)

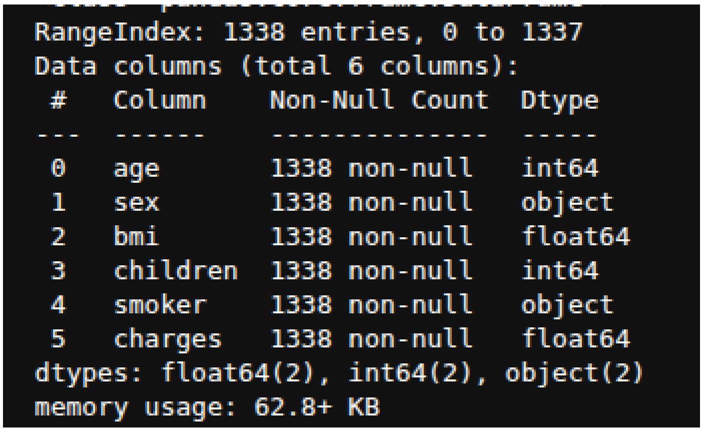
Goal: Predict Insureact charges based independed variable that we have.

2) Basic Info about the dataset:

Depended Variable: charges
Indepeneded Variable: smoker, children, bmi, sex, age

No. Of rows: 1338
No. Of Cols: 6

Data type:



```
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(2)
memory usage: 62.8+ KB
```

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

Here Gender, somker are comes under **nominal** data so we use get_dummies to convert into numerical data for further processing

4) Mode trainging Info:

Linear Regression:
r2_value: 0.067

MultiLinear Regression:
r2_value: 0.75

SVM

s.no	C	Linear	Poly	Rbf	Sigmoid
1	0.1	0.68	0.82	0.83	-45.90
2	1	0.68	0.82	0.83	-45.90
3	10	0.68	0.82	0.84	-5785
4	100	0.68	0.82	0.83	-557843

Decision Tree:

s.no	criterion	splitter	Accuracy
1	squared_error	best	0.76
2	friedman_mse	best	0.75
3	absolute_error	best	0.74
4	poisson	best	0.75
5	squared_error	random	0.72
6	friedman_mse	random	0.76
7	absolute_error	Random	0.75
8	Poission	random	0.74

Random Forest:

s.no	Criterion	n_estimators	Accuracy
1	squared_error	100	0.83
2	absolute_error	100	0.83
3	friedman_mse	100	0.83
4	Poisson	100	0.83
5	Squared_error	50	0.83
6	absolute_error	50	0.83
7	friedman_mse	50	0.83
8	Poisson	50	0.83

Final Model:
SVM [rbf c10] it gives accuracy of 84%