

## Regression assignment

### 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  
Output : insurance charges ( premium amount)  
Input : Age ,sex,bmi , children,smoker, charges

This process is called underwriting, based on customer details deciding insurance premium. Usually it will be done by underwriter manually.

Stage1: this problem is having data in numbers ( excel) so domain is ML ( id ML is not at all giving good model then DL )

Stage 2: input and output both the values are available so it is supervised learning

Stage3: Output value is continuous number so it is regression problem.

Many algorithms we can try like

1.Multiple Linear regression 2.SVM 3.Decision tree 4.Random forest

2.) Tell basic info about the dataset (Total number of rows, columns)  
1338 rows × 6 columns

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

Data ->1. Nominal data ( it will be string with no ordering) 2. Ordinal data( this will have certain order)

Here in this data set 2 categorical columns are there

1. Sex
2. Smoker

Sex and smoker both are nominal data so we have to use one hot encoding using `get_dummies()` function

`df = pd.get_dummies(df,drop_first=True)`

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 mode

I have created Multiple Linear Regression , SVM , random forest , decision tree

In that SVM is giving comparatively good result `r2_score = 0.877995 => 88%`

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

1<sup>st</sup> iteration

	model_name	hyperparamter_value	r2score
0	Multiple_LinearRegression	Normal	0.789479
1	SVM	form=Standard,Kernel = rbf,c=100000	0.789479
2	Decision tree	form=normal,criterion =poisson splitter=random	0.759623
3	Random Forest	form=Standard,n_estimators =200	0.853603

	model_name	hyperparamter_value	r2score
3	Random Forest	form=Standard,n_estimators =200	0.853603

2<sup>nd</sup> iteration (other day)

	model_name	hyperparamter_value	r2score
0	Multiple_LinearRegression	Normal	0.789479
1	SVM	form=Standard,Kernel = rbf,c=10000	0.877995
2	Decision tree	form=normal,criterion =poisson splitter=random	0.751467
3	Random Forest	form=Standard,n_estimators =100	0.854955

	model_name	hyperparamter_value	r2score
1	SVM	form=Standard,Kernel = rbf,c=10000	0.877995

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

Tried all possible options choose model which have more accuracy so that it can predict correct value

Tried all possible combination and choosed best model