ASSIGNMENT ON MACHINE LEARNING REGRESSION

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 several datasets for the same.

PROBLEM IDENTIFICATION:

STAGE 1: DOMAIN SELECTION (Machine Learning)

STAGE 2: SUPERVISED LEARNING (inputs and outputs are very clear)

STAGE 3: REGRESSION (we are going to predict the charges for insurance which is in the form of numbers, so it comes under the regression)

INFORMATION ABOUT THE GIVEN DATASET:

There are 6 columns and 1339 rows.

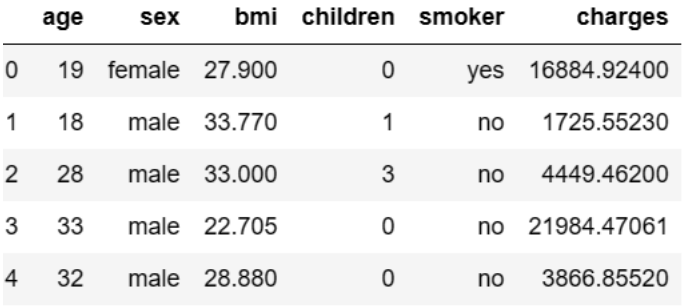
The given columns are (Age, Sex, Bmi, Children, Smoker, Charges).

We can take **Age, Sex, Bmi, Children, Smoker as an input** and take **Charges as an output.**

DATASET PRE_PRCOCESSING:

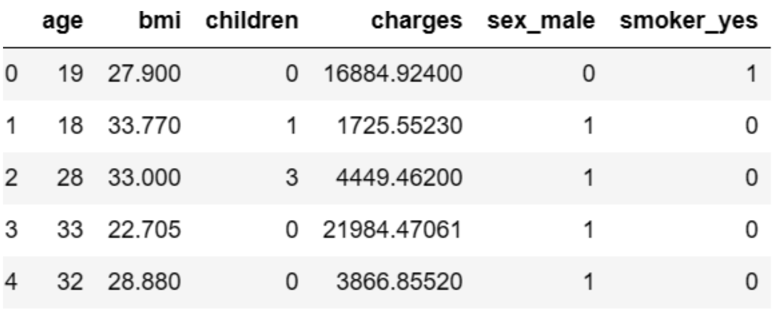
The given dataset having the column like Sex and Smoker provided the categorical values like (male/ female and yes/ no). Machine Learning Regression can't handle categorical data. We have to convert that categorical data into numerical values.

Before Pre_Processing: Example



	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

After Pre_Processing:



	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

DEVELOPING THE MODEL BY USING VARIOUS MACHINE LEARNING REGRESSION ALGORITHMS

MULTIPLE LINEAR REGRESSION:

S.NO	HYPER TUNNING PARAMETER	R_SCORE
1	-	0.7894
2	fit_intercept=True	0.7894
3	copy_X=True	0.7894

MAXIMUM R_SCORE

BY USING MULTIPLE LINEAR REGRESSION IS = 0.7894

SUPPORT VECTOR MACHINE:

S.NO	KERNEL	HYPER TUNNING PARAMETER	R_SCORE
1	Rbf	-	-0.0833
2	Rbf	C=10	-0.0322
3	Rbf	C=50	0.1478
4	Rbf	C=100	0.3200

MAXIMUM R_SCORE

BY USING SUPPORT VECTOR MACHINE IS = 0.3200

DECISION TREE:

S.NO	HYPER TUNNING PARAMETER	R_SCORE

1	Criterion=’squared_error’
	Splitter=’best’
		0.6872

2	Criterion=’friedman_mse’
	Splitter=’best’
		0.6845

3	Criterion='friedman_mse'
	splitter="random"
		0.7151

MAXIMUM R_SCORE

BY USING DECISION TREE IS = 0.7151

RANDOM FOREST:

S.NO	HYPER TUNNING PARAMETER	R_SCORE

1	n_estimators=50
	random_state=0
		0.8498

2	n_estimators=100
	random_state=0
		0.8539

3	n_estimators=10
	random_state=0
		0.8331

MAXIMUM R_SCORE

BY USING RANDOM FOREST REGRESSION IS = 0.8539

FINALIZED MODEL FOR THE GIVEN PROBLEM STATEMENT is RANDOM FOREST

We have created the different regression algorithms for the given problem statement. From all those alogorithms, finally I have selected the Random_forest algorithm as a better model.

Because comparatively it gives the highest accuracy for the given problem statement. I tabulated the R_score value for different parameters for our reference.