Regression Model

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.

Sample from the provided dataset –

age	sex	bmi	children	smoker	charges
19	female	27.9	0	yes	16884.92
18	male	33.77	1	no	1725.552
28	male	33	3	no	4449.462
33	male	22.705	0	no	21984.47
32	male	28.88	0	no	3866.855
31	female	25.74	0	no	3756.622

Inference – The data provided has four inputs dominated by numerical, hence the first decision goes to Machine Learning – Regression. Since we have more than one input that should be considered to derive a model, we can use either Multiple Linear Regression, Support Vector Machine (SVM) or Decision Tree algorithms to find a better model.

Multiple Linear Regression –

Changing nominal columns into numerical so that it can used for computation.

Parameters	**Weight**	**Bias**	**R^2 Score**
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)	array([[260.1423112 , 315.22441969, 545.72248029, 71.76915955,23252.13608407]])	array([-12013.76012735])	0.7565108093853883
LinearRegression(copy_X=True, fit_intercept=False, n_jobs=None, normalize=False)	[[199.98499634 34.80896984 342.96418146 -714.04772677 22616.98650384]]	0.0	0.7865108093853883
LinearRegression(copy_X=True, fit_intercept=False, n_jobs=None, normalize=True)	[[199.98499634 34.80896984 342.96418146 -714.04772677 22616.98650384]]	0.0	0.7584513449219452
LinearRegression(copy_X=False, fit_intercept=False, n_jobs=None, normalize=False)	[[199.98499634 34.80896984 342.96418146 -714.04772677 22616.98650384]]	0.0	0.7584513449219452

Multiple Linear Regression has a maximum R^2 Score of 78%. Hyper-tuning parameter of LinearRegression can optimise up to 78%. The default LinearRegression algorithm provides an accuracy of 75% only.

Support Vector Machine – Regression –

Parameters	R - Square Value
SVR()	-0.098510883
SVR(kernel = 'linear')	-0.14846453125202963
SVR(kernel = 'rbf')	-0.09851088349819759
SVR(kernel = 'sigmoid')	-0.0987122675639922
SVR(C=0.1)	-0.0986921139451229
SVR(C=1)	-0.09851088349819759
SVR(C=5)	-0.09770963326717874
SVR(C=0.001)	-0.09871206601107008
SVR(gamma='auto')	-0.09851088349819759

The R^2 Score obtained as a result of using multiple hyper-tuning parameters is not near to 1, hence **SVM – Regression** would not be considered to produce a stable model with the current data set behaviour and requirement.

Decision Tree –

Parameters	R – Squared Value
DecisionTreeRegressor()	0.698846999057871
DecisionTreeRegressor(splitter='random')	0.748204080083073
DecisionTreeRegressor(splitter='random', max_features="auto")	0.6950333262258626
DecisionTreeRegressor(splitter='random', max_features="sqrt")	0.6540393957314437
DecisionTreeRegressor(splitter='random', max_features="log2")	0.6305238584914996

Hyper-tuning the parameters of **Decision Tree** algorithm with different values to discover an accurate model. The best R^2 value obtained is 74%.

Considering the different algorithms that has been included in this **Regression** example, it has been found that **Multiple Linear Regression** provides the accuracy value of **78%**, hence the model is saved and can be used for future use.