

Assignment-Regression Algorithm

1) Identify your problem statement:

The problem statement is to predict the insurance charges based on several parameters from the given dataset.

Stage-1 : Input is in Excel format so **Machine learning** should be used.

Stage-2 : Requirement is clear. Input and Output are present in Dataset.

So, it is **Supervised** learning.

Stage-3 : Output is in Numerical format. So, it is **Regression**.

2) Tell basic info about the dataset (Total number of rows, columns):

In this dataset we are having 6 Columns.

- i. Age
- ii. Sex
- iii. BMI
- iv. Children
- v. Smoker
- vi. Charges

And 1338 Rows

3) Pre-processing method:

Columns Sex and Smoker having string data so it should be converted to Numerical data by using **one hot encoding**.

4) Models:

i. **Multi Linear Regression:**

In this Algorithm the obtained R2 Value is **0.78**

ii. Support Vector Machine:

S.No	Hyper Parameter	poly	rbf	sigmoid	Linear
1	C10	0.03	-0.032	0.039	0.462
2	C100	0.61	0.32	0.527	0.62
3	C500	0.82	0.664	0.444	0.763
4	C1000	0.85	0.81	0.287	0.764
5	C2000	0.860	0.85	-0.59	0.744
6	C3000	0.85	0.866	-2.12	0.741

iii. Decision Tree:

criterion	splitter	max_features	R2 Value
squared_error	best	auto	0.716341259
squared_error	best	sqrt	0.746786342
squared_error	best	log2	0.703982637
squared_error	best	none	0.690631844
squared_error	random	auto	0.706875254
squared_error	random	sqrt	0.632354033
squared_error	random	log2	0.681176673
squared_error	random	none	0.694050701
friedman_mse	best	auto	0.701626081
friedman_mse	best	sqrt	0.72505309
friedman_mse	best	log2	0.7291867
friedman_mse	best	none	0.686251784
friedman_mse	random	auto	0.72230874
friedman_mse	random	sqrt	0.622822734
friedman_mse	random	log2	0.651715961
friedman_mse	random	none	0.684456334
absolute_error	best	auto	0.699029123
absolute_error	best	sqrt	0.755677204
absolute_error	best	log2	0.722456836
absolute_error	best	none	0.670637575
absolute_error	random	auto	0.781417787
absolute_error	random	sqrt	0.56912417
absolute_error	random	log2	0.744458911
absolute_error	random	none	0.703884994
poisson	best	auto	0.727006735
poisson	best	sqrt	0.766534229
poisson	best	log2	0.72819623
poisson	best	none	0.711476182
poisson	random	auto	0.663789663
poisson	random	sqrt	0.711604276
poisson	random	log2	0.691153637
poisson	random	none	0.66658979

iv. Random Forest:

criterion	max_features	N_estimators	R2 Value
friedman_mse	Auto	10	0.848252717
friedman_mse	Auto	100	0.85238149
friedman_mse	Sqrt	10	0.862273081
friedman_mse	Sqrt	100	0.865637813
friedman_mse	Log2	10	0.843418825
friedman_mse	Log2	100	0.869705035
absolute_error	Auto	10	0.841988078
absolute_error	Auto	100	0.851471618
absolute_error	Sqrt	10	0.853014663
absolute_error	Sqrt	100	0.870685506
absolute_error	Log2	10	0.85771457
absolute_error	Log2	100	0.871155941

In **Random forest Regression** with the parameters of (absolute error,log2,100)we got a R2 Value of **0.871** which is highest compared to all other Algorithms.