Assignment

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 dataset scientist, you must develop a model which will predict the insurance charges.

1) *Identify your problem statement*

Stage 1: Machine learning

Stage 2: Supervised learning

Stage 3: Regression

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

Total number of rows are 1338 and the total number of columns is 6

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

It's a nominal dataset, because we can compare the age with others.

4)Devolope 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.

My finalized model for this dataset is "Random forest" because the r2 value of this model is the highest one.

5)All the research value (r2_score of the models) should be documented.

1)Decision tree

Criterion	***Splitter***	***R2value***
Squared_error	*best*	*0.65884579*
Friedman_mse	*Best*	*0.6875130*
Absolute_error	*Best*	*0.6856941*
Max_depth=none	*min_samples_split=2*	***0.69820104***
min_samples_leaf=1	*min_weight_fraction_leaf=0.0*	0.667875
max_features=None	*random_state=None*	***0.69805482***
max_leaf_nodes=None	*min_impurity_decrease=0.0*	0.66857183
ccp_alpha=0.0	*monotonic_cst=None*	***0.66857183***
max_depth=3,	*min_impurity_decrease=0.4*	0.70183154

criterion='squared_error'	*random_state=2*	0.7015151
criterion='squared_error'	*min_samples_split=2*	*0.68342783*
max_depth=None	*splitter='best'*	0.6884023
min_impurity_decrease=0.0	*monotonic_cst=None*	0.70506885
ccp_alpha=0.0	*min_samples_leaf=1*	*0.70506885*
max_depth=3	*min_samples_leaf=100*	*0.7313303*
min_weight_fraction_leaf=0	*min_samples_leaf=10*	***0.87467747***
max_leaf_nodes=None	*random_state=0*	0.690923077

2) SVM

Parameter	parameter	R2 value
	Kernel=Linear	-0.0101954
epsilon=0.0	*tol=0.0001*	-0.0834051

epsilon=0.0	max_iter=1000	-0.08340516
max_iter=1000	C=1.0	-0.083405160
epsilon=0.0	C=1.0	-0.083405160
epsilon=0.0	verbose=0	-0.08340516
max_iter=1000	C=1.0	-0.0884616
	Kernal='rbf'	-0.0834051
	Kernal='poly'	-0.0757173
	Kernal='sigmoid'	-0.0754463
	gamma='scale'	0.84752647
C=1.0	epsilon=0.2	-0.08340516
C=100	Epsilon=0.6	-0.08340516
C=1000	Epsilon=0.10	0.810719570

C=0.01	Epsilon=100	-0.08868556
	Gamma='auto'	-0.083405160

3) Random forest

n_estimators=50	*max_features=1.0*	***0.8563508***
ccp_alpha=0.0	*max_features=1.0*	***0.8475264***
min_weight_fraction_leaf=0.0	*max_features=1.0*	0.85041311
min_weight_fraction_leaf=0.50	*max_features=0.20*	0.00320632
criterion='squared_error'	min_samples_split=2	0.85108598
criterion='squared_error'	min_samples_split=7	0.872655517
criterion='squared_error'	min_samples_split=10	0.875290100

bootstrap=True	min_samples_split=10	0.87506315
bootstrap=True	min_samples_split=100	0.86859936
max_depth=None	*ccp_alpha=0.0*	0.85151773
max_depth=None	*ccp_alpha=0.20*	0.852772384
max_depth=None	monotonic_cst=None	0.84871504
n_jobs=None	n_estimators=100	0.852214440
n_jobs=None	n_estimators=1200	0.853187652

6) Mention your final model and justify why you chosen the same.

Random forest is the best method for the given model , because when we combine these two parameter (*criterion=squared_error,min_samples_split=10***) we get the highest r2 value and the r2 value is 0.875290100.**