Insurance Charges Prediction

Problem Statement:

To predict the Insurance charges based on the various parameters as per the given dataset.

About DataSet:

Dataset consists 6 columns and 1338 rows

5 columns are the Inputs (age, sex, bmi, children, smoker)

1 column is the Output (charges)

Domain :

Machine Learning (Since datasets are numbers)

Learning :

Supervised Learning

Requirements are clear

Both I/P's and O/P's are present in the dataset

It is Regression since the O/P values are continuous values.

PreProcessing

Nominal values needs to be converted (One Hot Encoding)

Sex and Smoker inputs are preprocessed like sex_male (1 means Male & 0 means Female) & smoker_yes (1 means smoker & 0 means non smoker)

Various Algorithm Outputs

- Support Vector Machine

SVM				
Hyper Parameter	**Linear**	**RBF**	**POLY**	**Sigmoid**
C10	0.4624	-0.0322	0.03871	0.0393
C100	0.6288	0.32	0.6179	0.5276
C500	0.7631	0.6642	0.8263	0.4446
C1000	0.7649	0.8102	0.8566	0.2874
C2000	0.744	0.8547	0.8605	-0.5939
C3000	0.7414	0.8663	0.8598	-2.1244

- Decision Tree

Decision Tree			
Criterion	**Max_features**	**Splitter**	**R Score**
squared_error	sqrt	best	0.6665
squared_error	sqrt	random	0.6499

squared_error	log2	best	0.7764
squared_error	log2	random	0.6465
friedman_mse	sqrt	best	0.7225
friedman_mse	sqrt	random	0.7233
friedman_mse	log2	best	0.7336
friedman_mse	log2	random	0.6542
absolute_error	sqrt	best	0.6666
absolute_error	sqrt	random	0.6264
absolute_error	log2	best	0.7237
absolute_error	log2	random	0.6742
poisson	sqrt	best	0.6698
poisson	sqrt	random	0.6654
poisson	log2	best	0.7304
poisson	log2	random	0.6803

- Random Forest

Random Forest				
n_estimators	**criterion**	**max_features**	**random_state**	**r_score**
10	*squared_error*	sqrt	0	0.852
50	*squared_error*	sqrt	0	0.8695
100	*squared_error*	sqrt	0	0.871
10	*squared_error*	log2	0	0.852
50	*squared_error*	log2	0	0.8695
100	*squared_error*	log2	0	0.871
10	*absolute_error*	sqrt	0	0.8574
50	*absolute_error*	sqrt	0	0.8708
100	*absolute_error*	sqrt	0	0.871
10	*absolute_error*	log2	0	0.8574
50	*absolute_error*	log2	0	0.8708
100	*absolute_error*	log2	0	0.871
10	*friedman_mse*	sqrt	0	0.8502
50	*friedman_mse*	sqrt	0	0.8702
100	*friedman_mse*	sqrt	0	0.871
10	*friedman_mse*	log2	0	0.8502
50	*friedman_mse*	log2	0	0.8702
100	*friedman_mse*	log2	0	0.871
10	*poisson*	sqrt	0	0.8544
50	*poisson*	sqrt	0	0.8632
100	*poisson*	sqrt	0	0.868
10	*poisson*	log2	0	0.8544
50	*poisson*	log2	0	0.8632
100	*poisson*	log2	0	0.868

Result:

Maximum R^2 value is 0.871 in Random Forest (No of estimators 100 with squared_error, absolute_error & friedman_mse with both log2 & sqrt). We can go with any of these 6 models.