Problem Statement for predicting the insurance premium

1. The problem statement is to identify the insurance premium for the customers based on their age, gender, body mass index, no. of children and if the customer is a smoker or a non-smoker.
These criteria have an impact on the premium calculation for the insurance policy.

2. Total number of rows and columns are 1338 rows × 6 columns.

3. The dataset has columns such as Sex and Smoker which are nominal data and get_dummies() method has been used to convert them to categorical data.

4. Data model Tabulation is listed below where the dataset is processed using the Multiple Linear Regression, Support Vector Machine, Decision Tree and Random Forest.

5. Based on the analysis, it is evident that Random Forest algorithm yields the highest R2 Score making it more appropriate to predict this dataset with the criterion as absolute_error, max feature as sqrt or log2 with the n_estimators as 100.

R2 Score for the dataset “Insurance Premium” using various algorithms

1. Multiple Linear Regression:
The R2 score is 0.7894790349867009

2. Support Vector Machine(SVM):
The best R2 score for SVM is 0.84193382759 achieved using linear Poly where C = 3000

Iteration	Penalty Value - C	linear - R2 score	rbf - R2 score	poly - R2 score	sigmoid - R2 score
1	C=0.1	-35396.04768594211	-5440261.70880	-1766046.7478	-1246546.72301874
2	C=1	-332.77071665376997	-54117.2876953	-17454.966238	-12346.3461962857
3	C=10	-1.6415812170241182	-519.116705864	-157.46063728	-110.183700425396
4	C=100	0.0033516801839467	-4.82156703100	-0.3896453881	-0.77891306449674
5	C=1000	0.7372671733693169	0.688275816002	0.81947037838	-0.28834225376504
6	C=2000	0.7637722902026782	0.808554857964	0.84061597474	-0.13559466214339
7	C=3000	0.7646313401248144	0.836065961734	0.84193382759	-0.09891008475063

3. Decision Tree:
The best R2 score for Decision Tree is 0.7466445434644639 achieved using absolute_error as the criterion with the max feature as auto and splitter as random

Use Case	Criterion	Max Features	Splitter	R2 Value
1	absolute_error	auto	best	0.6893033518872613
2	absolute_error	auto	random	0.7349387883843278
3	absolute_error	sqrt	best	0.674523068447676
4	absolute_error	sqrt	random	0.6891483872431307
5	absolute_error	log2	best	0.7340315560069428
6	absolute_error	log2	random	0.6074797925035988
7	poisson	auto	best	0.7262611798891718
8	poisson	auto	random	0.6814529550502325
9	poisson	sqrt	best	0.7364142961841933
10	poisson	sqrt	random	0.6199974631095988
11	poisson	log2	best	0.7333235997692493
12	poisson	log2	random	0.6437429542708426
13	friedman_mse	auto	best	0.7081966858881397
14	friedman_mse	auto	random	0.7045053498470839
15	friedman_mse	sqrt	best	0.7385439221685439
16	friedman_mse	sqrt	random	0.6602615877159796
17	friedman_mse	log2	best	0.74080766177542
18	friedman_mse	log2	random	0.7466445434644639

4. Random Forest:

The best R2 score for Random Forest is 0.8710685856341518 achieved using absolute_error as the criterion with the max feature as sqrt and log2 with the n_estimators as 100.

Use Case	Criterion	Max Features	n_estimators	R2 Value
1	absolute_error	auto	10	0.835063555313752
2	absolute_error	auto	100	0.8520093621081837
3	absolute_error	sqrt	10	0.8574290080917196
4	absolute_error	sqrt	100	0.8710685856341518
5	absolute_error	log2	10	0.8574290080917196
6	absolute_error	log2	100	0.8710685856341518
7	poisson	auto	10	0.8313991040134341
8	poisson	auto	100	0.8526334258892607
9	poisson	sqrt	10	0.8544955286235119
10	poisson	sqrt	100	0.8680156984764337
11	poisson	log2	10	0.8544955286235119
12	poisson	log2	100	0.8680156984764337
13	friedman_mse	auto	10	0.8331662678473348
14	friedman_mse	auto	100	0.8540518935149612
15	friedman_mse	sqrt	10	0.8502777994291519
16	friedman_mse	sqrt	100	0.8710544015500664
17	friedman_mse	log2	10	0.8502777994291519
18	friedman_mse	log2	100	0.8710544015500664