Insurance Charge Predictions

1. **Identify your problem statement**

   Given the dataset predict insurance charges using the inputs (age, sex, bmi, children, smoker

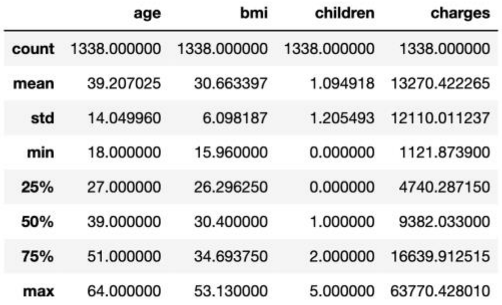
2. **Tell basic info about the dataset (Total number of rows, columns)**

   No of rows and columns
   1338 rows × 6 columns

   Data Types:

```
age            int64
sex           object
bmi          float64
children       int64
smoker        object
charges      float64
```

   Statistics:



	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

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

   Converted sex and smoker columns to nominal data using pandas library.

4. **Develop 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.**

   Developed models using SLR, MLR, SVM, Decision Tree and Random Forest

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

   Please find the details in next section.

6. **Mention your final model, justify why u have chosen the same.**

   Random Forest provides good model with R2 score as 87%

r2_score of the models:

To find following the Machine Learning Regression method using r2 value

1. **Multiple Linear Regression**

   R2 value = 0.78

2. **Support Vector Machine (SVM)**

   R2 values for different Kernel and C,

#	Hyper Parameter	Linear	Poly	RBF (Default)	Sigmoid	Precomputed
1	C=0.10	-0.1220	-0.0862	-0.0895	-0.0899	Must be square matrix
2	C=1.0	-0.1116	-0.0642	-0.0884	-0.0899	
3	C=10	-0.0016	-0.0931	-0.0884	-0.0907	
4	C=100	0.5432	-0.0997	-0.1248	-0.1181	
5	C=1000	0.6340	-0.0555	-0.1174	-1.6659	
6	C=2000	0.6893	-0.0027	-0.1174	-5.6164	
7	C=3000	0.7590	0.04892	-0.0962	-12.0190	

The **SVM Regression** use R2 value (Linear & C3000) = 0.7590

3. **Decision Tree**

#	Criterion	Max Features	Splitter	R Value
1	friedman_mse	log2	best	0.7128
2	friedman_mse	log2	random	0.6336
3	friedman_mse	sqrt	random	0.6756
4	friedman_mse	sqrt	best	0.7593
5	squared_error	log2	best	0.7115
6	squared_error	log2	random	0.6580
7	squared_error	sqrt	random	0.6691
8	squared_error	sqrt	best	0.7412
9	absolute_error	log2	best	0.6827
10	absolute_error	log2	random	0.6796
11	absolute_error	sqrt	random	0.5808
12	absolute_error	sqrt	best	0.7536
13	poisson	log2	best	0.7050
14	poisson	log2	random	0.6413
15	poisson	sqrt	random	0.6973
16	poisson	sqrt	best	0.6607

The **Decision Tree** use R2 value (friedman_mse, sqrt, best) = 0.7593

4. Random Forest

n_estimators=100, random_state=0

#	Criterion	Max Features	R Value
1	squared_error	log2	0.8710
2	squared_error	Sqrt	0.8710
11	absolute_error	Log2	0.8710
12	absolute_error	Sqrt	0.8710
13	poisson	sqrt	0.8680
14	poisson	log2	0.8680
15	friedman_mse	log2	0.8710
16	friedman_mse	sqrt	0.8710

The **Random Forest** use R2 value for multiple combinations = 0.8710