In [1]:

```
import pandas as pd
```

In [2]:

```
df =pd.read_csv("insurance.csv")
```

In [3]:

```
df
```

Out[3]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

In [5]:

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()

column = ['sex','smoker','region']
for i in column :
    df[i] = encoder.fit_transform(df[i])
```

In [6]:

```
df
```

Out[6]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520
...	...	...	...	...	...	...	...
1333	50	1	30.970	3	0	1	10600.54830
1334	18	0	31.920	0	0	0	2205.98080
1335	18	0	36.850	0	0	2	1629.83350
1336	21	0	25.800	0	0	3	2007.94500
1337	61	0	29.070	0	1	1	29141.36030

1338 rows × 7 columns

In [7]:

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
```

In [8]:

```
x = df.drop(['charges'],axis = 1)
y = df['charges']
```

In [9]:

```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = .2,random_
```

In [10]:

```
model = RandomForestRegressor()
cv_scores = cross_val_score(model, x, y, cv=5)
model.fit(x_train,y_train)
y_pred = model.predict(x_test)
```

In [11]:

```
cv_scores
```

Out[11]:

```
array([0.84979061, 0.77394652, 0.87037715, 0.83081739, 0.85143071])
```

```
In [12]:  r2_score(y_pred,y_test)

Out[12]:  0.8553362150794915

In [ ]:
```