Problem statement

1. Analysis the hospital data and find out the various factor effected to customers for fees charges.

2. Build the machine learning model to predict the hospital fees charges.

nikhil-katwe-copy1

July 22, 2023

1 Import All Libraries

[30]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
df = pd.read_csv(r"C:\Users\admin\Downloads\Health_insurance (1).csv")
```

[3]:
```
df
```

[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 x 7 columns]

[31]:
```
df.shape
```

[31]:
```
(1338, 7)
```

2 data cleaning

[5]:
```
df.isnull().sum()
```

[5]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

[6]:
```
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
```

3 EDA

[7]:
```
df.describe()
```

[7]:

	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

[8]:
```
df1 = df[df['charges'] == df['charges'].max()]
```

[9]:
```
df1
```

[9]:

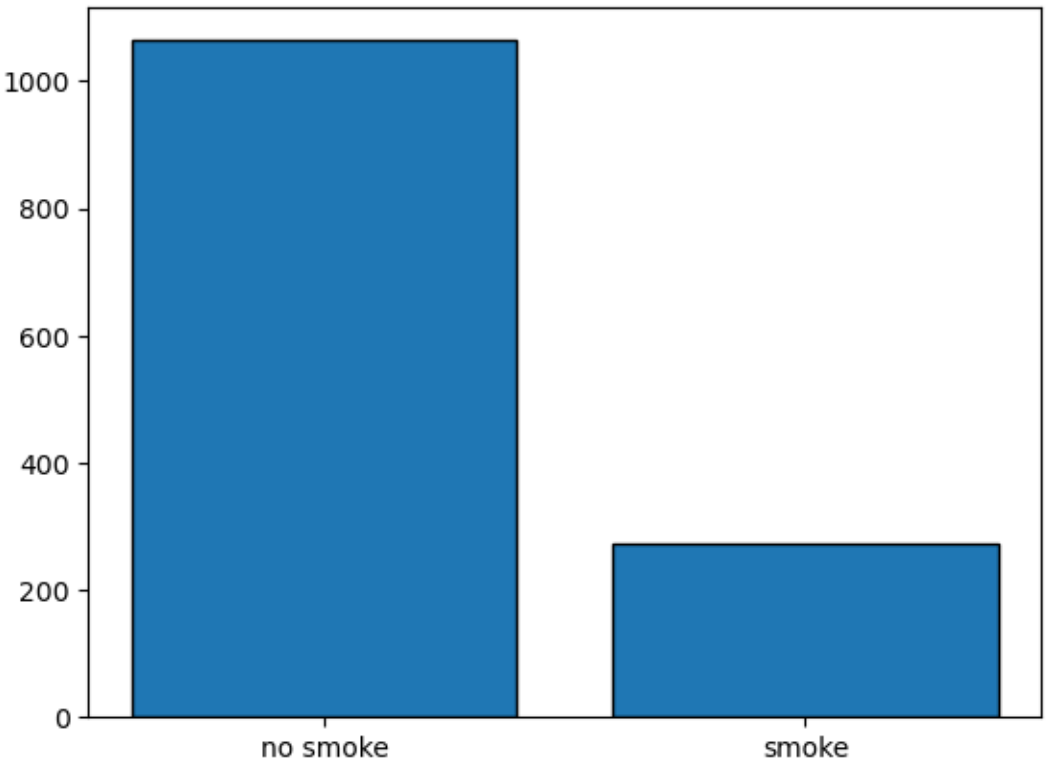
	age	sex	bmi	children	smoker	region	charges
543	54	female	47.41	0	yes	southeast	63770.42801

[10]:
```
df['smoker'].value_counts()
```

[10]:
```
no     1064
yes     274
Name: smoker, dtype: int64
```

[11]:
```
plt.bar(['no smoke','smoke'],df['smoker'].value_counts(),edgecolor = 'k',width
 ↪= 0.8)
```

[11]:
```
<BarContainer object of 2 artists>
```



[12]:
```
df['smoker'] = df['smoker'].map({'yes': 1,'no': 0})
```

[13]:
```
df['sex'] = df['sex'].map({'female': 1,'male': 0})
```

[14]:
```
df.head()
```

[14]:

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	1	southwest	16884.92400
1	18	0	33.770	1	0	southeast	1725.55230
2	28	0	33.000	3	0	southeast	4449.46200
3	33	0	22.705	0	0	northwest	21984.47061

```
4   32    0  28.880         0       0  northwest   3866.85520
```

[15]:
```
df
```

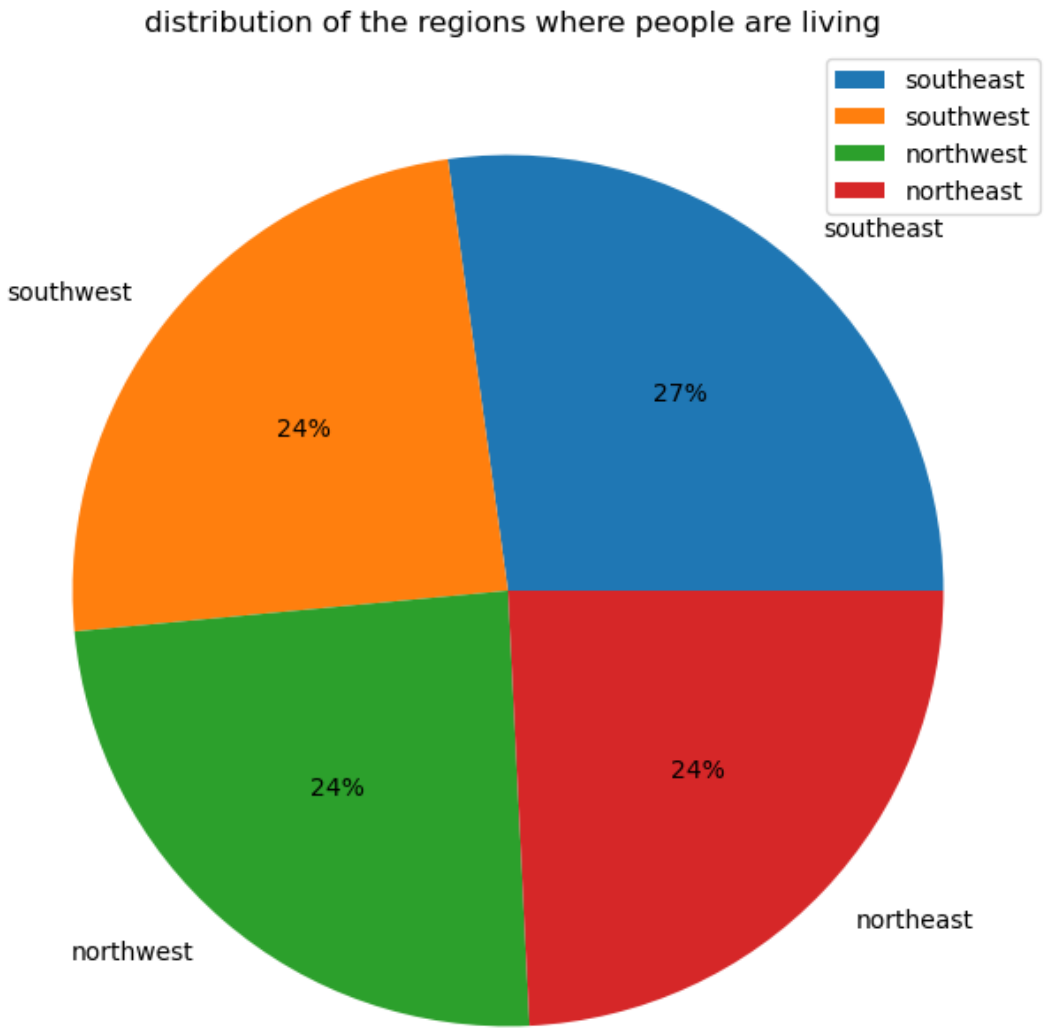
[15]:
```
       age  sex     bmi  children  smoker     region      charges
0       19    1  27.900         0       1  southwest  16884.92400
1       18    0  33.770         1       0  southeast   1725.55230
2       28    0  33.000         3       0  southeast   4449.46200
3       33    0  22.705         0       0  northwest  21984.47061
4       32    0  28.880         0       0  northwest   3866.85520
…    …  …      …       …      …       …          …
1333    50    0  30.970         3       0  northwest  10600.54830
1334    18    1  31.920         0       0  northeast   2205.98080
1335    18    1  36.850         0       0  southeast   1629.83350
1336    21    1  25.800         0       0  southwest   2007.94500
1337    61    1  29.070         0       1  northwest  29141.36030

[1338 rows x 7 columns]
```

[16]:
```
df['region'].value_counts()
```

[16]:
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

[17]:
```
plt.figure(figsize = (10,8))
label = ["southeast","southwest","northwest",'northeast']
plt.pie(df['region'].value_counts(),labels = label,autopct='%.f%%')
plt.title(' distribution of the regions where people are living')
plt.legend()
```

[17]:
```
<matplotlib.legend.Legend at 0x1292f7e3dc0>
```



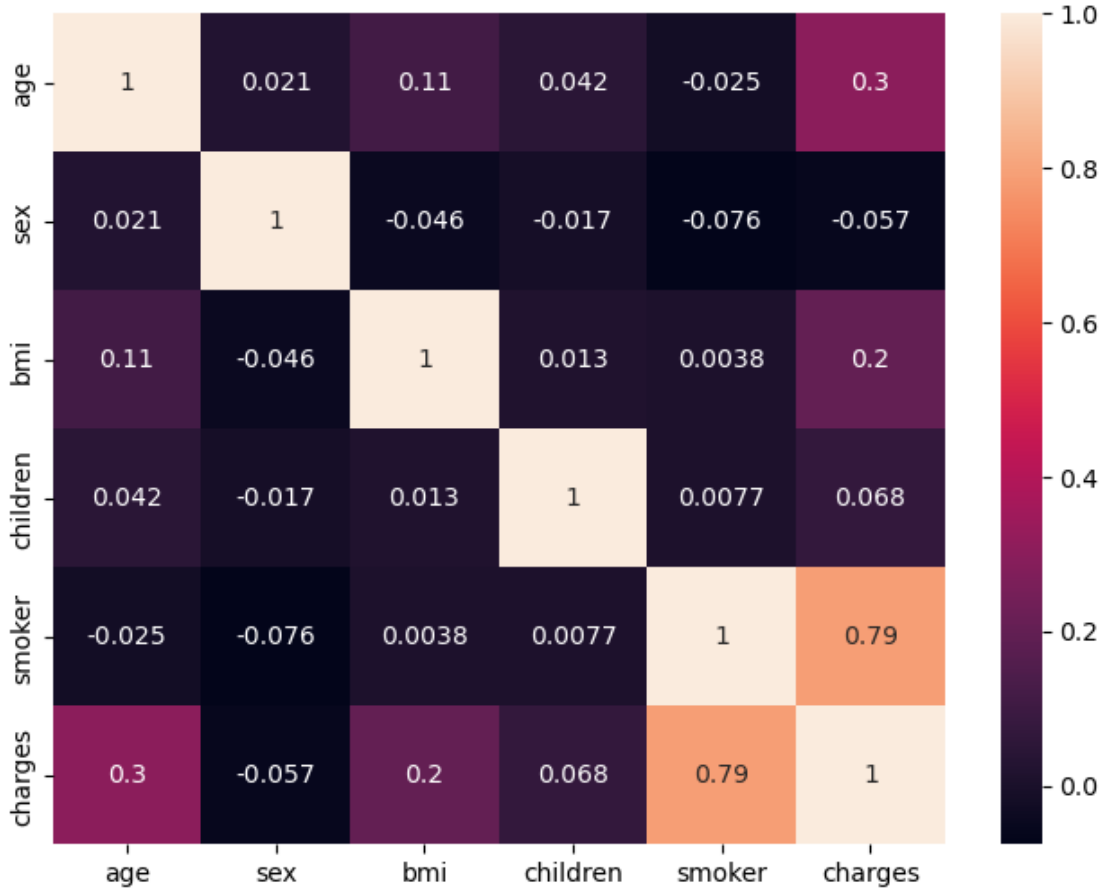
[18]:
```
df.corr()
```

[18]:

	age	sex	bmi	children	smoker	charges
age	1.000000	0.020856	0.109272	0.042469	-0.025019	0.299008
sex	0.020856	1.000000	-0.046371	-0.017163	-0.076185	-0.057292
bmi	0.109272	-0.046371	1.000000	0.012759	0.003750	0.198341
children	0.042469	-0.017163	0.012759	1.000000	0.007673	0.067998
smoker	-0.025019	-0.076185	0.003750	0.007673	1.000000	0.787251
charges	0.299008	-0.057292	0.198341	0.067998	0.787251	1.000000

[19]:
```
plt.figure(figsize = (8,6))
sns.heatmap(df.corr(),annot = True)
```

[19]: <AxesSubplot:>



[20]:
```
x = df.drop(['charges','region'],axis = 1)
y = df['charges']
```

[21]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test = train_test_split(x,y,test_size=0.
 ↪20,random_state=48)
```

[22]:
```
X_train
```

[22]:

	age	sex	bmi	children	smoker
655	52	1	25.300	2	1
516	20	0	35.310	1	0
226	28	0	38.060	0	0
149	19	0	28.400	1	0
11	62	1	26.290	0	1
..	…	…	…	…	…
454	32	0	46.530	2	0

```
966   51    0  24.795         2       1
944   62    0  39.930         0       0
347   46    0  33.345         1       0
563   50    0  44.770         1       0

[1070 rows x 5 columns]
```

[23]:
```
X_test
```

[23]:
```
      age  sex     bmi  children  smoker
451    30    0  24.130         1       0
1174   29    0  32.110         2       0
213    34    1  26.730         1       0
174    24    1  33.345         0       0
648    18    0  28.500         0       0
…   …  …    …       …      …
42     41    0  21.780         1       0
782    51    0  35.970         1       0
859    57    0  28.100         0       0
1260   32    1  20.520         0       0
1224   41    0  23.940         1       0

[268 rows x 5 columns]
```

[24]:
```
Y_train
```

[24]:
```
655    24667.41900
516    27724.28875
226     2689.49540
149     1842.51900
11     27808.72510
          …
454     4686.38870
966    23967.38305
944    12982.87470
347     8334.45755
563     9058.73030
Name: charges, Length: 1070, dtype: float64
```

[25]:
```
Y_test
```

[25]:
```
451      4032.24070
1174     4433.91590
213      5002.78270
174      2855.43755
648      1712.22700
           …
```

```
42       6272.47720
782      9386.16130
859     10965.44600
1260     4544.23480
1224     6858.47960
Name: charges, Length: 268, dtype: float64
```

4 Model Building

[26]:
```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor()
model.fit(X_train,Y_train)
```

[26]:
```
RandomForestRegressor()
```

[27]:
```
model.score(X_test,Y_test)
```

[27]:
```
0.8124980456590855
```

[28]:
```
df.head(2)
```

[28]:

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.90	0	1	southwest	16884.9240
1	18	0	33.77	1	0	southeast	1725.5523

5 Prediction

[29]:
```
model.predict([[48,1,37.41,0,1]])
```

[29]:
```
array([46178.6001451])
```

[ ]:
```

```