Problem Statement :

Predictive study using the breast cancer diagnostic data set

In [19]:

```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")
```

In [20]:

```
df=pd.read_csv(r"C:\Users\91949\Downloads\BreastCancerPrediction.csv")
df
```

Out[20]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothı
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	
...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	
565	926682	M	20.13	28.25	131.20	1261.0	
566	926954	M	16.60	28.08	108.30	858.1	
567	927241	M	20.60	29.33	140.10	1265.0	
568	92751	B	7.76	24.54	47.92	181.0	

569 rows × 33 columns

In [21]:

```
df.head()
```

Out[21]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothnes
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	

5 rows × 33 columns

In [22]:

```
df.tail()
```

Out[22]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothnes
564	926424	M	21.56	22.39	142.00	1479.0	
565	926682	M	20.13	28.25	131.20	1261.0	
566	926954	M	16.60	28.08	108.30	858.1	
567	927241	M	20.60	29.33	140.10	1265.0	
568	92751	B	7.76	24.54	47.92	181.0	

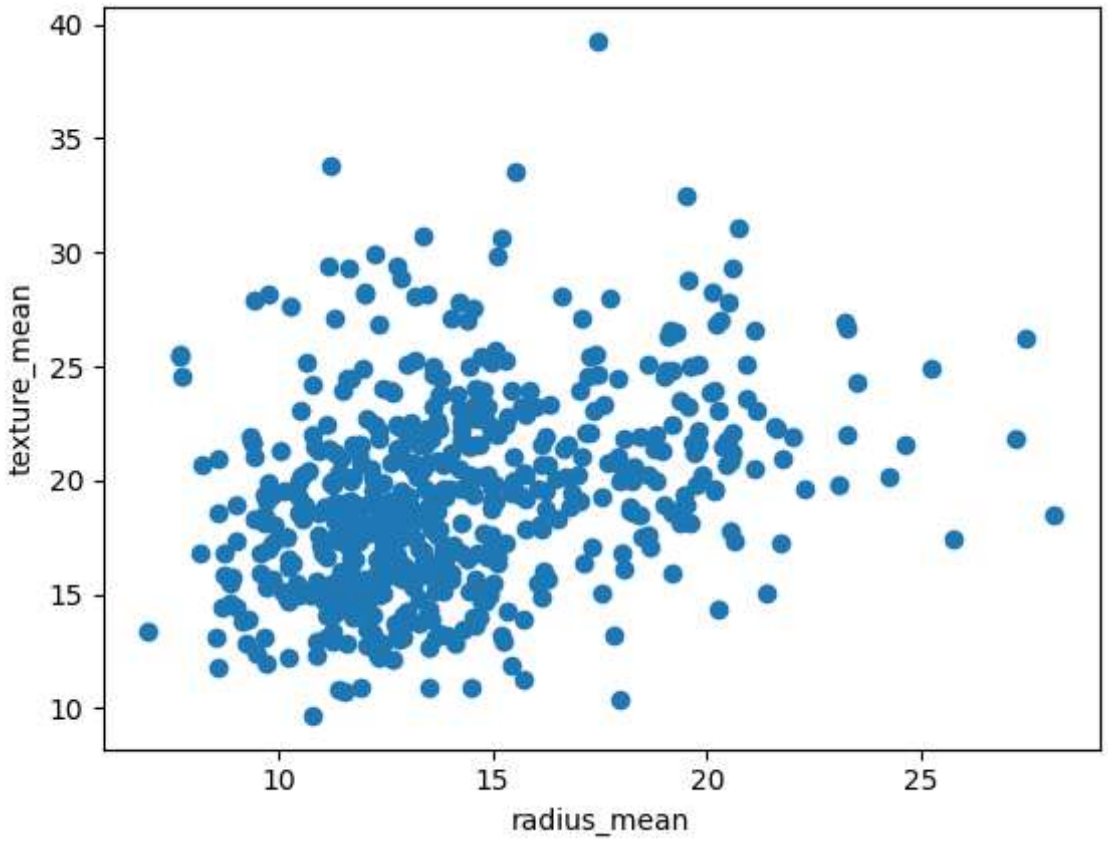
5 rows × 33 columns

In [23]:

```
plt.scatter(df["radius_mean"],df["texture_mean"])
plt.xlabel("radius_mean")
plt.ylabel("texture_mean")
```

Out[23]:

```
Text(0, 0.5, 'texture_mean')
```



In [24]:

```
from sklearn.cluster import KMeans
km=KMeans()
km
```

Out[24]:

```
KMeans()
```

In [25]:

```
y_predicted=km.fit_predict(df[["radius_mean","texture_mean"]])
y_predicted
```

Out[25]:

```
array([7, 5, 5, 0, 5, 7, 5, 3, 6, 6, 3, 3, 1, 6, 6, 2, 3, 3, 5, 7, 7, 4,
       7, 1, 3, 7, 3, 5, 6, 7, 1, 0, 1, 1, 3, 3, 3, 0, 6, 3, 6, 6, 1, 3,
       6, 5, 0, 0, 4, 6, 6, 7, 0, 5, 3, 0, 5, 3, 0, 4, 4, 0, 6, 4, 6, 6,
       0, 0, 0, 7, 5, 4, 1, 7, 0, 3, 4, 7, 1, 0, 6, 7, 1, 1, 4, 5, 3, 1,
       6, 7, 6, 3, 7, 0, 3, 1, 0, 0, 4, 3, 6, 4, 0, 0, 0, 7, 0, 0, 5, 6,
       0, 6, 3, 0, 4, 6, 4, 7, 3, 5, 4, 5, 5, 4, 7, 7, 6, 5, 7, 1, 4, 3,
       3, 7, 5, 6, 0, 4, 7, 4, 4, 3, 0, 7, 4, 4, 0, 3, 7, 0, 6, 0, 4, 4,
       7, 0, 3, 3, 4, 4, 0, 5, 5, 6, 5, 3, 4, 3, 1, 7, 4, 3, 7, 4, 4, 4,
       0, 3, 6, 4, 5, 1, 3, 4, 3, 4, 5, 0, 0, 7, 6, 6, 0, 2, 6, 7, 6, 5,
       5, 3, 0, 3, 1, 6, 0, 7, 0, 3, 6, 7, 5, 0, 5, 1, 6, 7, 0, 0, 5, 1,
       7, 7, 0, 3, 7, 7, 4, 7, 6, 6, 3, 2, 2, 1, 4, 3, 1, 5, 2, 2, 7, 4,
       0, 6, 1, 0, 0, 4, 6, 4, 1, 0, 5, 7, 5, 7, 1, 7, 3, 2, 1, 3, 3, 3,
       3, 1, 0, 6, 7, 0, 7, 4, 5, 4, 1, 0, 4, 5, 0, 7, 1, 4, 5, 3, 7, 0,
       6, 4, 0, 0, 3, 3, 7, 0, 4, 7, 4, 0, 3, 6, 5, 0, 1, 0, 0, 6, 7, 4,
       4, 4, 0, 7, 4, 4, 0, 0, 4, 5, 0, 0, 4, 5, 4, 5, 4, 0, 7, 0, 3, 3,
       7, 0, 0, 4, 0, 3, 7, 5, 0, 1, 7, 0, 4, 5, 4, 4, 0, 7, 4, 4, 0, 3,
       5, 6, 4, 0, 0, 7, 4, 0, 0, 6, 0, 3, 7, 5, 1, 0, 5, 5, 3, 7, 5, 5,
       7, 7, 0, 2, 7, 0, 4, 4, 6, 0, 7, 6, 4, 7, 4, 1, 4, 0, 3, 5, 0, 7,
       0, 0, 4, 0, 5, 4, 0, 7, 4, 0, 7, 6, 5, 0, 0, 0, 6, 3, 2, 6, 6, 3,
       4, 6, 0, 7, 4, 3, 0, 6, 4, 6, 0, 0, 3, 0, 5, 5, 7, 3, 0, 7, 3, 7,
       0, 1, 7, 0, 5, 6, 1, 7, 3, 5, 6, 1, 2, 7, 0, 2, 2, 6, 6, 2, 1, 1,
       2, 0, 0, 3, 3, 0, 1, 0, 0, 2, 7, 2, 4, 7, 3, 7, 4, 3, 0, 3, 7, 7,
       7, 7, 7, 5, 0, 3, 6, 7, 5, 4, 3, 3, 0, 0, 5, 5, 7, 6, 7, 5, 4, 4,
       0, 0, 7, 6, 4, 7, 3, 7, 3, 0, 5, 5, 0, 7, 4, 5, 0, 0, 4, 4, 0, 4,
       7, 4, 0, 0, 7, 5, 0, 5, 6, 6, 6, 6, 4, 6, 6, 2, 3, 6, 0, 0, 0, 6,
       6, 6, 2, 6, 2, 2, 0, 2, 6, 6, 2, 2, 2, 1, 5, 1, 2, 1, 6])
```

In [26]:

```
df["cluster"]=y_predicted
df.head()
```

Out[26]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothnes
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	

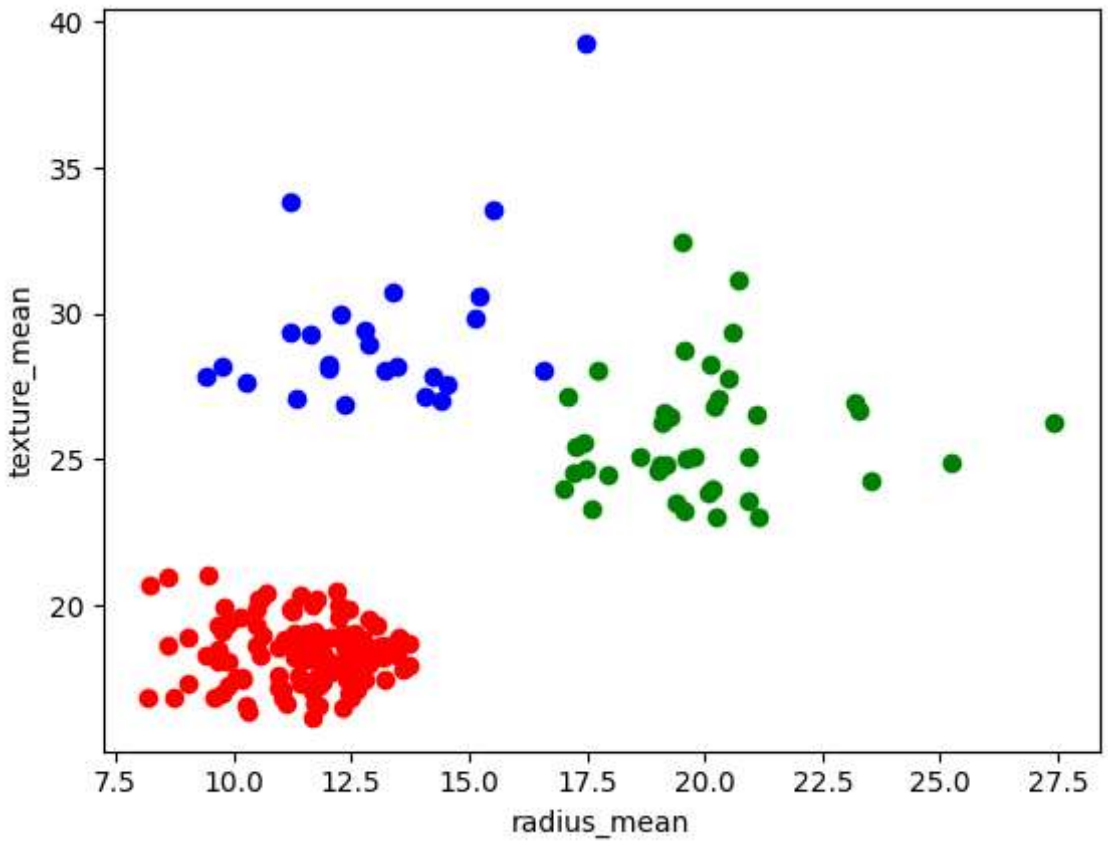
5 rows × 34 columns

In [27]:

```
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["radius_mean"],df1["texture_mean"],color="red")
plt.scatter(df2["radius_mean"],df2["texture_mean"],color="green")
plt.scatter(df3["radius_mean"],df3["texture_mean"],color="blue")
plt.xlabel("radius_mean")
plt.ylabel("texture_mean")
```

Out[27]:

```
Text(0, 0.5, 'texture_mean')
```



In [28]:

```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(df[["texture_mean"]])
df["texture_mean"]=scaler.transform(df[["texture_mean"]])
df.head()
```

Out[28]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothnes
0	842302	M	17.99	0.022658	122.80	1001.0	
1	842517	M	20.57	0.272574	132.90	1326.0	
2	84300903	M	19.69	0.390260	130.00	1203.0	
3	84348301	M	11.42	0.360839	77.58	386.1	
4	84358402	M	20.29	0.156578	135.10	1297.0	

5 rows × 34 columns

In [29]:

```
scaler.fit(df[["radius_mean"]])
df["radius_mean"]=scaler.transform(df[["radius_mean"]])
df.head()
```

Out[29]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothnes
0	842302	M	0.521037	0.022658	122.80	1001.0	
1	842517	M	0.643144	0.272574	132.90	1326.0	
2	84300903	M	0.601496	0.390260	130.00	1203.0	
3	84348301	M	0.210090	0.360839	77.58	386.1	
4	84358402	M	0.629893	0.156578	135.10	1297.0	

5 rows × 34 columns

In [30]:

```
y_predicted=km.fit_predict(df[["radius_mean","texture_mean"]])
y_predicted
```

Out[30]:

```
array([3, 4, 4, 6, 4, 3, 4, 1, 1, 7, 1, 3, 0, 1, 1, 7, 1, 1, 4, 3, 3, 5,
       3, 2, 1, 4, 1, 4, 1, 4, 0, 6, 0, 0, 3, 1, 1, 6, 7, 1, 1, 6, 0, 1,
       1, 4, 5, 6, 5, 1, 6, 3, 6, 4, 1, 6, 4, 1, 6, 5, 5, 6, 1, 5, 7, 1,
       6, 6, 6, 3, 4, 5, 0, 3, 6, 1, 3, 4, 0, 6, 6, 3, 2, 0, 5, 4, 1, 0,
       1, 3, 1, 1, 3, 6, 1, 0, 6, 6, 5, 1, 7, 5, 6, 6, 6, 3, 6, 6, 2, 6,
       6, 1, 1, 6, 5, 6, 5, 3, 1, 4, 5, 4, 2, 3, 3, 3, 7, 4, 3, 0, 5, 1,
       1, 3, 4, 1, 6, 5, 3, 5, 5, 3, 6, 3, 5, 5, 6, 1, 3, 3, 1, 6, 5, 5,
       3, 6, 4, 4, 5, 5, 6, 4, 4, 1, 2, 1, 5, 4, 0, 3, 5, 1, 3, 5, 5, 5,
       6, 1, 1, 3, 2, 0, 1, 5, 1, 5, 4, 6, 6, 3, 1, 1, 6, 7, 1, 3, 1, 4,
       4, 1, 6, 4, 2, 1, 6, 3, 6, 4, 1, 3, 4, 6, 2, 0, 1, 3, 6, 6, 4, 0,
       3, 3, 6, 1, 3, 3, 5, 3, 7, 1, 4, 7, 7, 0, 5, 1, 2, 4, 7, 0, 3, 3,
       6, 1, 0, 6, 3, 3, 7, 5, 0, 6, 4, 4, 4, 3, 0, 3, 1, 7, 0, 0, 4, 1,
       4, 0, 6, 1, 3, 6, 3, 5, 2, 5, 0, 6, 5, 4, 3, 3, 0, 5, 4, 1, 3, 6,
       6, 3, 6, 6, 1, 1, 3, 6, 3, 3, 5, 6, 3, 6, 4, 6, 0, 6, 6, 7, 3, 5,
       3, 3, 6, 3, 3, 5, 6, 6, 5, 4, 6, 6, 5, 4, 3, 4, 5, 6, 3, 6, 1, 1,
       3, 6, 6, 5, 6, 4, 3, 4, 6, 2, 3, 5, 5, 4, 5, 5, 6, 3, 5, 5, 6, 1,
       2, 7, 5, 6, 6, 3, 5, 6, 6, 1, 6, 4, 3, 4, 0, 6, 4, 2, 1, 3, 4, 4,
       3, 3, 6, 7, 3, 6, 5, 5, 1, 6, 3, 1, 5, 3, 5, 0, 5, 5, 1, 2, 6, 3,
       1, 6, 5, 6, 4, 5, 6, 3, 5, 6, 3, 1, 4, 6, 6, 6, 6, 1, 7, 6, 6, 1,
       5, 6, 6, 3, 5, 1, 6, 6, 5, 6, 6, 6, 1, 6, 4, 4, 3, 1, 6, 3, 1, 3,
       6, 0, 3, 6, 4, 7, 0, 3, 1, 4, 6, 0, 7, 3, 6, 7, 7, 7, 7, 7, 0, 2,
       7, 6, 6, 1, 1, 6, 0, 6, 6, 7, 3, 7, 5, 3, 1, 3, 5, 1, 6, 1, 3, 3,
       3, 3, 3, 4, 5, 4, 1, 3, 4, 5, 1, 1, 6, 6, 4, 4, 3, 7, 3, 2, 5, 5,
       6, 6, 3, 1, 5, 3, 1, 3, 1, 6, 4, 4, 6, 3, 5, 2, 6, 1, 5, 5, 1, 5,
       3, 5, 6, 6, 3, 4, 6, 4, 1, 7, 7, 7, 5, 7, 7, 7, 1, 1, 5, 5, 6, 7,
       6, 6, 7, 6, 7, 7, 6, 7, 1, 7, 7, 7, 7, 0, 2, 0, 0, 0, 7])
```

In [31]:

```
df["New Cluster"]=y_predicted
df.head()
```

Out[31]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothnes
0	842302	M	0.521037	0.022658	122.80	1001.0	
1	842517	M	0.643144	0.272574	132.90	1326.0	
2	84300903	M	0.601496	0.390260	130.00	1203.0	
3	84348301	M	0.210090	0.360839	77.58	386.1	
4	84358402	M	0.629893	0.156578	135.10	1297.0	

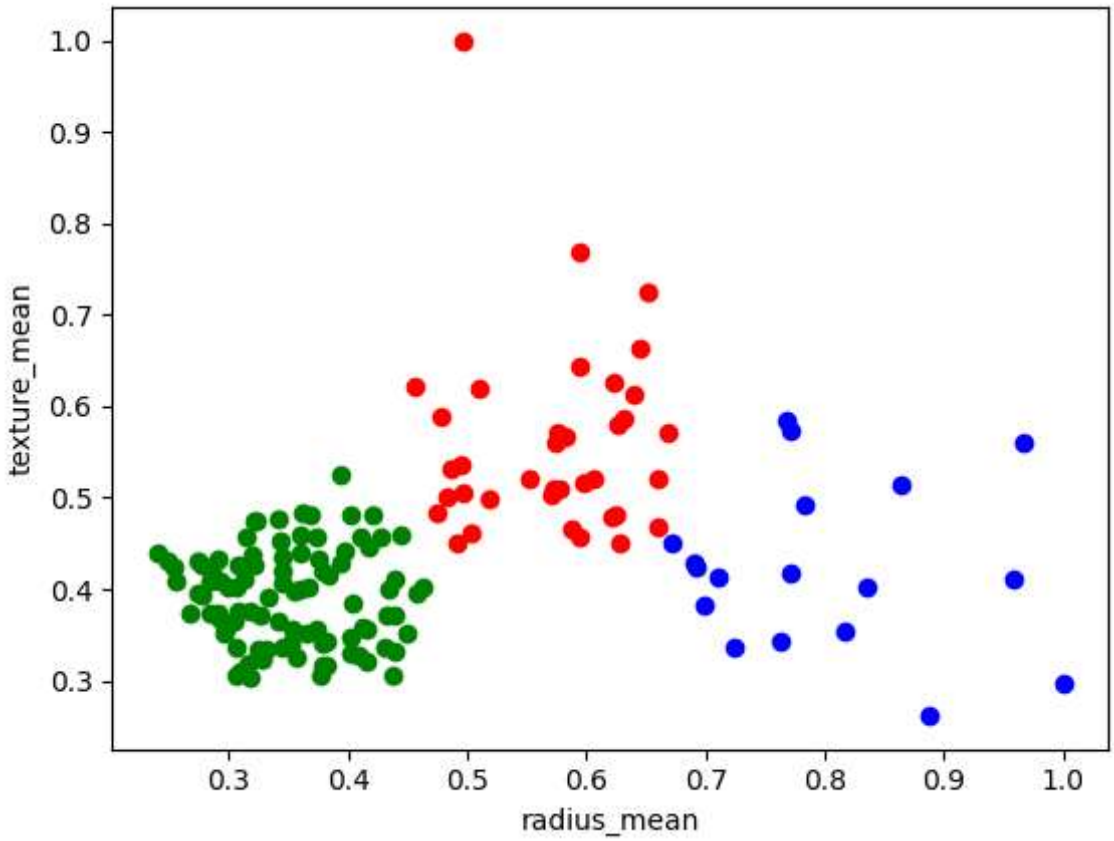
5 rows × 35 columns

In [32]:

```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["radius_mean"],df1["texture_mean"],color="red")
plt.scatter(df2["radius_mean"],df2["texture_mean"],color="green")
plt.scatter(df3["radius_mean"],df3["texture_mean"],color="blue")
plt.xlabel("radius_mean")
plt.ylabel("texture_mean")
```

Out[32]:

```
Text(0, 0.5, 'texture_mean')
```



In [33]:

```
km.cluster_centers_
```

Out[33]:

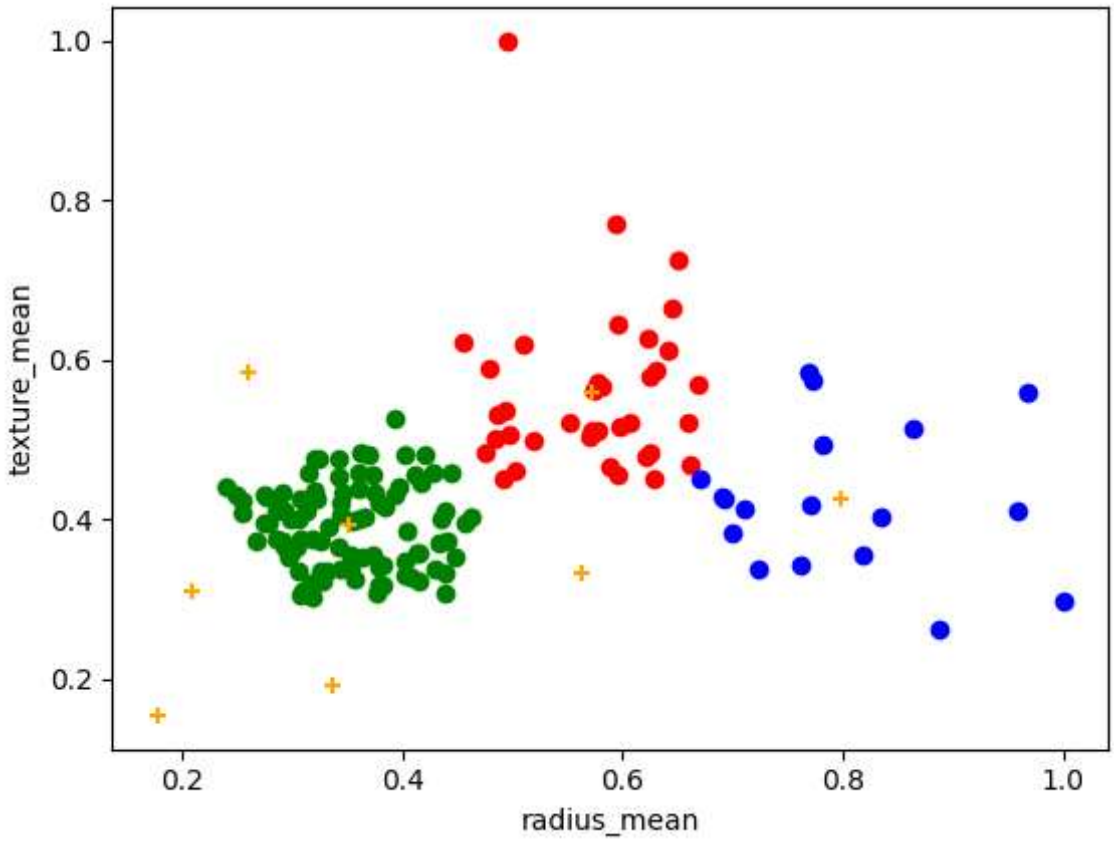
```
array([[0.57132058, 0.55893025],
       [0.35173159, 0.39188367],
       [0.79840767, 0.42469846],
       [0.33570532, 0.19063107],
       [0.56287997, 0.33184226],
       [0.17750575, 0.15412045],
       [0.20867092, 0.3094643 ],
       [0.2590623 , 0.58293879]])
```

In [34]:

```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["radius_mean"],df1["texture_mean"],color="red")
plt.scatter(df2["radius_mean"],df2["texture_mean"],color="green")
plt.scatter(df3["radius_mean"],df3["texture_mean"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",marker
plt.xlabel("radius_mean")
plt.ylabel("texture_mean")
```

Out[34]:

```
Text(0, 0.5, 'texture_mean')
```



In [35]:

```
k_rng=range(1,10)
sse=[]
```

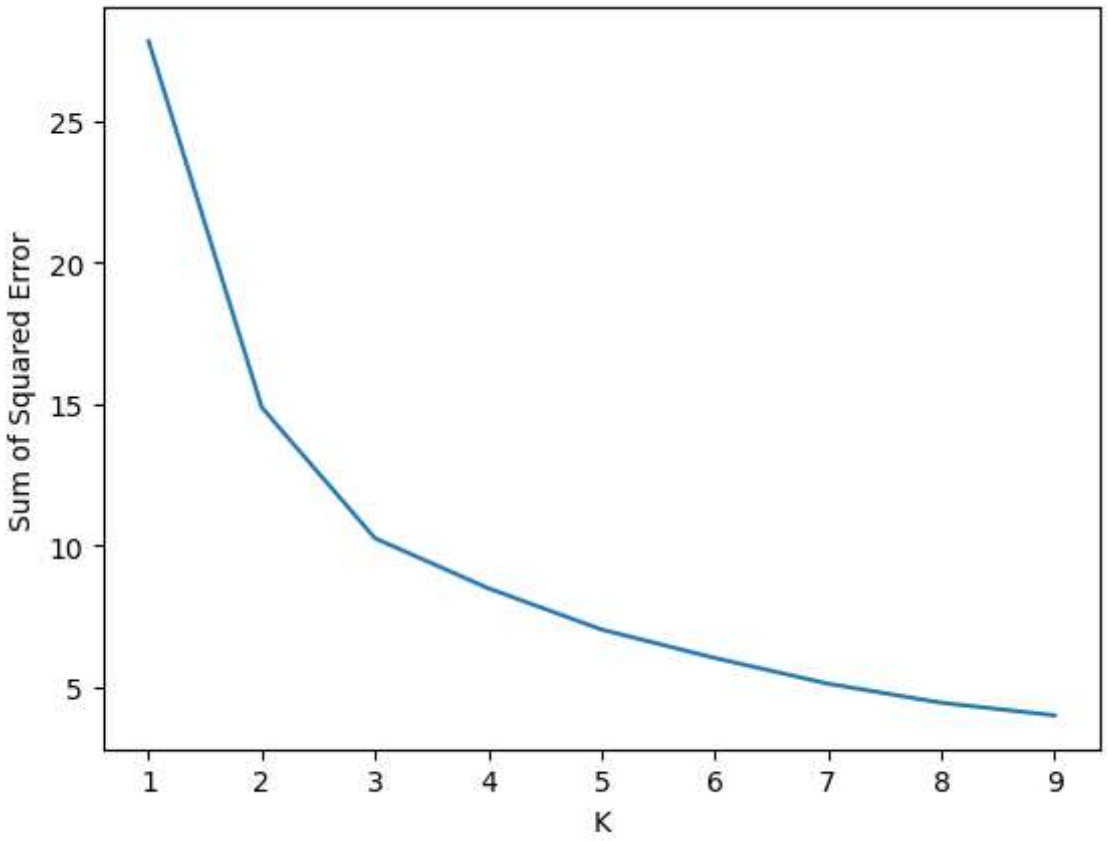
In [36]:

```
for k in k_rng:
 km=KMeans(n_clusters=k)
 km.fit(df[["radius_mean","texture_mean"]])
 sse.append(km.inertia_)
print(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

```
[27.817507595043075, 14.872296449956036, 10.252751496105198, 8.4874074509
59015, 7.030487343041076, 6.025921805510964, 5.11711415242544, 4.44443596
08281525, 3.9986180020558613]
```

Out[36]:

```
Text(0, 0.5, 'Sum of Squared Error')
```



conclusion:

we conclude the for the given data set we can do prediction by various models, but accuracy from that models are not good as k-means clustering ,so we prefer k-means clustering for this dataset.