projects-based-on-knn-1

February 21, 2024

[ ]:
```
K NEAREST NEIGHBORS
```

[ ]:
```
PROJECT-1
```

[ ]:
```
Cancer Analysis using KNN
```

[68]:
```
import pandas as pd
df2=pd.read_csv(r"C:\Users\user\Downloads\cancerKNNAlgorithmDataset.csv")
```

[69]:
```
df2
```

[69]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	\
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	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	0.11840	0.27760	0.30010	0.14710	
1	0.08474	0.07864	0.08690	0.07017	
2	0.10960	0.15990	0.19740	0.12790	
3	0.14250	0.28390	0.24140	0.10520	
4	0.10030	0.13280	0.19800	0.10430	
..	…	…	…	…	
564	0.11100	0.11590	0.24390	0.13890	
565	0.09780	0.10340	0.14400	0.09791	
566	0.08455	0.10230	0.09251	0.05302	
567	0.11780	0.27700	0.35140	0.15200	
568	0.05263	0.04362	0.00000	0.00000	

```
     …  texture_worst  perimeter_worst  area_worst  smoothness_worst  \
0    …          17.33           184.60      2019.0           0.16220
1    …          23.41           158.80      1956.0           0.12380
2    …          25.53           152.50      1709.0           0.14440
3    …          26.50            98.87       567.7           0.20980
4    …          16.67           152.20      1575.0           0.13740
..   …            …                …           …                …
564  …          26.40           166.10      2027.0           0.14100
565  …          38.25           155.00      1731.0           0.11660
566  …          34.12           126.70      1124.0           0.11390
567  …          39.42           184.60      1821.0           0.16500
568  …          30.37            59.16       268.6           0.08996

     compactness_worst  concavity_worst  concave points_worst  symmetry_worst  \
0              0.66560           0.7119                0.2654          0.4601
1              0.18660           0.2416                0.1860          0.2750
2              0.42450           0.4504                0.2430          0.3613
3              0.86630           0.6869                0.2575          0.6638
4              0.20500           0.4000                0.1625          0.2364
..                 …                …                     …               …
564            0.21130           0.4107                0.2216          0.2060
565            0.19220           0.3215                0.1628          0.2572
566            0.30940           0.3403                0.1418          0.2218
567            0.86810           0.9387                0.2650          0.4087
568            0.06444           0.0000                0.0000          0.2871

     fractal_dimension_worst  Unnamed: 32
0                    0.11890          NaN
1                    0.08902          NaN
2                    0.08758          NaN
3                    0.17300          NaN
4                    0.07678          NaN
..                       …            …
564                  0.07115          NaN
565                  0.06637          NaN
566                  0.07820          NaN
567                  0.12400          NaN
568                  0.07039          NaN

[569 rows x 33 columns]
```

[70]:
```
df2.head()
```

[70]:
```
         id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
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

   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017
2          0.10960           0.15990          0.1974              0.12790
3          0.14250           0.28390          0.2414              0.10520
4          0.10030           0.13280          0.1980              0.10430

   …  texture_worst  perimeter_worst  area_worst  smoothness_worst  \
0  …          17.33           184.60      2019.0            0.1622
1  …          23.41           158.80      1956.0            0.1238
2  …          25.53           152.50      1709.0            0.1444
3  …          26.50            98.87       567.7            0.2098
4  …          16.67           152.20      1575.0            0.1374

   compactness_worst  concavity_worst  concave points_worst  symmetry_worst  \
0             0.6656           0.7119                0.2654          0.4601
1             0.1866           0.2416                0.1860          0.2750
2             0.4245           0.4504                0.2430          0.3613
3             0.8663           0.6869                0.2575          0.6638
4             0.2050           0.4000                0.1625          0.2364

   fractal_dimension_worst  Unnamed: 32
0                  0.11890          NaN
1                  0.08902          NaN
2                  0.08758          NaN
3                  0.17300          NaN
4                  0.07678          NaN

[5 rows x 33 columns]
```

[71]:
```
df2.drop(['Unnamed: 32','id'],axis=1,inplace=True)
```

[72]:
```
df2
```

[72]:
```
    diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0           M        17.99         10.38          122.80     1001.0
1           M        20.57         17.77          132.90     1326.0
2           M        19.69         21.25          130.00     1203.0
3           M        11.42         20.38           77.58      386.1
4           M        20.29         14.34          135.10     1297.0
..        …           …             …               …          …
564         M        21.56         22.39          142.00     1479.0
565         M        20.13         28.25          131.20     1261.0
566         M        16.60         28.08          108.30      858.1
```

```
567         M        20.60         29.33          140.10     1265.0
568         B         7.76         24.54           47.92      181.0

     smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0            0.11840           0.27760         0.30010              0.14710
1            0.08474           0.07864         0.08690              0.07017
2            0.10960           0.15990         0.19740              0.12790
3            0.14250           0.28390         0.24140              0.10520
4            0.10030           0.13280         0.19800              0.10430
..               …                 …               …                    …
564          0.11100           0.11590         0.24390              0.13890
565          0.09780           0.10340         0.14400              0.09791
566          0.08455           0.10230         0.09251              0.05302
567          0.11780           0.27700         0.35140              0.15200
568          0.05263           0.04362         0.00000              0.00000

     symmetry_mean  …  radius_worst  texture_worst  perimeter_worst  \
0           0.2419  …        25.380          17.33           184.60
1           0.1812  …        24.990          23.41           158.80
2           0.2069  …        23.570          25.53           152.50
3           0.2597  …        14.910          26.50            98.87
4           0.1809  …        22.540          16.67           152.20
..             …  …             …              …                …
564         0.1726  …        25.450          26.40           166.10
565         0.1752  …        23.690          38.25           155.00
566         0.1590  …        18.980          34.12           126.70
567         0.2397  …        25.740          39.42           184.60
568         0.1587  …         9.456          30.37            59.16

     area_worst  smoothness_worst  compactness_worst  concavity_worst  \
0        2019.0           0.16220            0.66560           0.7119
1        1956.0           0.12380            0.18660           0.2416
2        1709.0           0.14440            0.42450           0.4504
3         567.7           0.20980            0.86630           0.6869
4        1575.0           0.13740            0.20500           0.4000
..          …                 …                  …                …
564      2027.0           0.14100            0.21130           0.4107
565      1731.0           0.11660            0.19220           0.3215
566      1124.0           0.11390            0.30940           0.3403
567      1821.0           0.16500            0.86810           0.9387
568       268.6           0.08996            0.06444           0.0000

     concave points_worst  symmetry_worst  fractal_dimension_worst
0                  0.2654          0.4601                  0.11890
1                  0.1860          0.2750                  0.08902
2                  0.2430          0.3613                  0.08758
3                  0.2575          0.6638                  0.17300
```

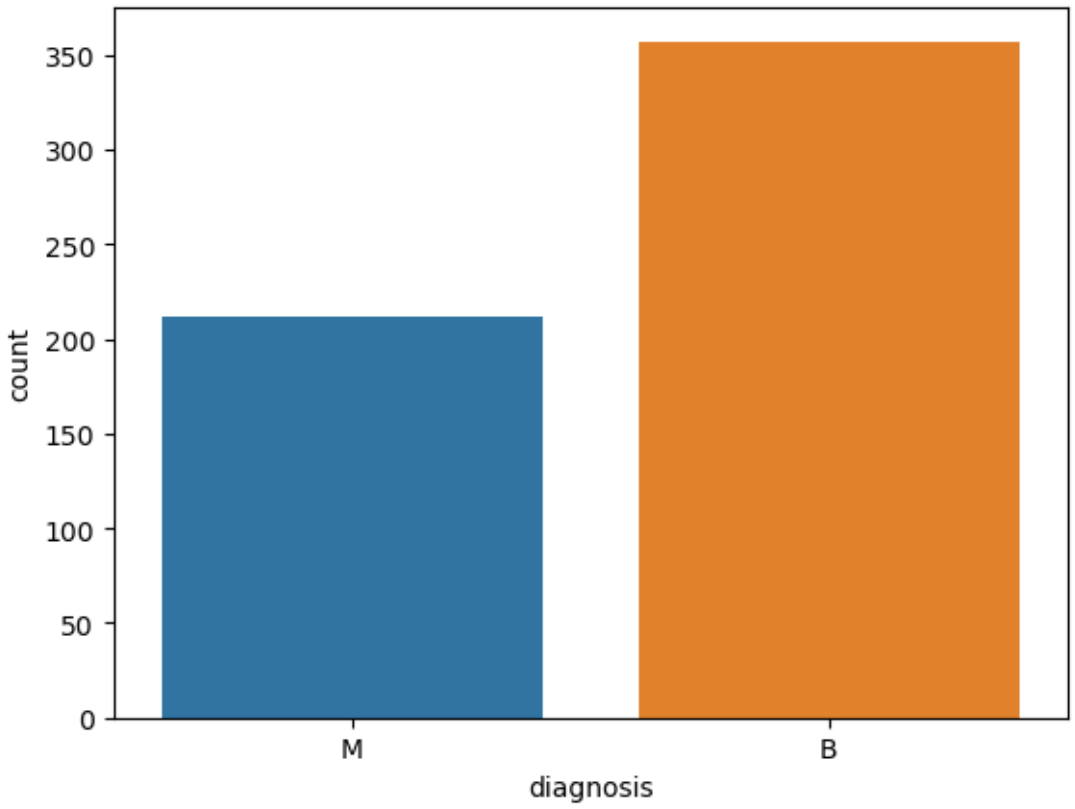
```
4                 0.1625          0.2364                0.07678
..                   …               …                      …
564               0.2216          0.2060                0.07115
565               0.1628          0.2572                0.06637
566               0.1418          0.2218                0.07820
567               0.2650          0.4087                0.12400
568               0.0000          0.2871                0.07039

[569 rows x 31 columns]
```

[73]:
```
# no of values in each class
import matplotlib.pyplot as plt
import seaborn as sns
print(df2['diagnosis'].value_counts())

# plot class distribution
sns.countplot(x='diagnosis',data=df2)
plt.show()
```

```
diagnosis
B    357
M    212
Name: count, dtype: int64
```



[8]:
```
from sklearn.preprocessing import StandardScaler
```

[9]:
```
scalar=StandardScaler()
```

[10]:
```
df2.replace({'M':0,'B':1},inplace=True)
```

[11]:
```
df2
```

[11]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	\
0	0	17.99	10.38	122.80	1001.0	
1	0	20.57	17.77	132.90	1326.0	
2	0	19.69	21.25	130.00	1203.0	
3	0	11.42	20.38	77.58	386.1	
4	0	20.29	14.34	135.10	1297.0	
..	…	…	…	…	…	
564	0	21.56	22.39	142.00	1479.0	
565	0	20.13	28.25	131.20	1261.0	
566	0	16.60	28.08	108.30	858.1	
567	0	20.60	29.33	140.10	1265.0	
568	1	7.76	24.54	47.92	181.0	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	0.11840	0.27760	0.30010	0.14710	
1	0.08474	0.07864	0.08690	0.07017	
2	0.10960	0.15990	0.19740	0.12790	
3	0.14250	0.28390	0.24140	0.10520	
4	0.10030	0.13280	0.19800	0.10430	
..	…	…	…	…	
564	0.11100	0.11590	0.24390	0.13890	
565	0.09780	0.10340	0.14400	0.09791	
566	0.08455	0.10230	0.09251	0.05302	
567	0.11780	0.27700	0.35140	0.15200	
568	0.05263	0.04362	0.00000	0.00000	

	symmetry_mean	…	radius_worst	texture_worst	perimeter_worst	\
0	0.2419	…	25.380	17.33	184.60	
1	0.1812	…	24.990	23.41	158.80	
2	0.2069	…	23.570	25.53	152.50	
3	0.2597	…	14.910	26.50	98.87	
4	0.1809	…	22.540	16.67	152.20	
..	…	…	…	…	…	
564	0.1726	…	25.450	26.40	166.10	
565	0.1752	…	23.690	38.25	155.00	
566	0.1590	…	18.980	34.12	126.70	
567	0.2397	…	25.740	39.42	184.60	

```
568         0.1587  …         9.456          30.37           59.16

     area_worst  smoothness_worst  compactness_worst  concavity_worst  \
0        2019.0           0.16220            0.66560           0.7119
1        1956.0           0.12380            0.18660           0.2416
2        1709.0           0.14440            0.42450           0.4504
3         567.7           0.20980            0.86630           0.6869
4        1575.0           0.13740            0.20500           0.4000
..          …                 …                  …                …
564      2027.0           0.14100            0.21130           0.4107
565      1731.0           0.11660            0.19220           0.3215
566      1124.0           0.11390            0.30940           0.3403
567      1821.0           0.16500            0.86810           0.9387
568       268.6           0.08996            0.06444           0.0000

     concave points_worst  symmetry_worst  fractal_dimension_worst
0                  0.2654          0.4601                  0.11890
1                  0.1860          0.2750                  0.08902
2                  0.2430          0.3613                  0.08758
3                  0.2575          0.6638                  0.17300
4                  0.1625          0.2364                  0.07678
..                    …               …                        …
564                0.2216          0.2060                  0.07115
565                0.1628          0.2572                  0.06637
566                0.1418          0.2218                  0.07820
567                0.2650          0.4087                  0.12400
568                0.0000          0.2871                  0.07039

[569 rows x 31 columns]
```

[12]:
```
df2.fillna(0,inplace=True)
```

[13]:
```
df2
```

[13]:
```
     diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0            0        17.99         10.38          122.80     1001.0
1            0        20.57         17.77          132.90     1326.0
2            0        19.69         21.25          130.00     1203.0
3            0        11.42         20.38           77.58      386.1
4            0        20.29         14.34          135.10     1297.0
..         …          …             …               …          …
564          0        21.56         22.39          142.00     1479.0
565          0        20.13         28.25          131.20     1261.0
566          0        16.60         28.08          108.30      858.1
567          0        20.60         29.33          140.10     1265.0
568          1         7.76         24.54           47.92      181.0
```

```
     smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0            0.11840           0.27760         0.30010              0.14710
1            0.08474           0.07864         0.08690              0.07017
2            0.10960           0.15990         0.19740              0.12790
3            0.14250           0.28390         0.24140              0.10520
4            0.10030           0.13280         0.19800              0.10430
..               …                 …               …                    …
564          0.11100           0.11590         0.24390              0.13890
565          0.09780           0.10340         0.14400              0.09791
566          0.08455           0.10230         0.09251              0.05302
567          0.11780           0.27700         0.35140              0.15200
568          0.05263           0.04362         0.00000              0.00000

     symmetry_mean  …  radius_worst  texture_worst  perimeter_worst  \
0           0.2419  …        25.380          17.33           184.60
1           0.1812  …        24.990          23.41           158.80
2           0.2069  …        23.570          25.53           152.50
3           0.2597  …        14.910          26.50            98.87
4           0.1809  …        22.540          16.67           152.20
..             …  …             …              …               …
564         0.1726  …        25.450          26.40           166.10
565         0.1752  …        23.690          38.25           155.00
566         0.1590  …        18.980          34.12           126.70
567         0.2397  …        25.740          39.42           184.60
568         0.1587  …         9.456          30.37            59.16

     area_worst  smoothness_worst  compactness_worst  concavity_worst  \
0        2019.0           0.16220            0.66560           0.7119
1        1956.0           0.12380            0.18660           0.2416
2        1709.0           0.14440            0.42450           0.4504
3         567.7           0.20980            0.86630           0.6869
4        1575.0           0.13740            0.20500           0.4000
..          …                 …                  …                …
564      2027.0           0.14100            0.21130           0.4107
565      1731.0           0.11660            0.19220           0.3215
566      1124.0           0.11390            0.30940           0.3403
567      1821.0           0.16500            0.86810           0.9387
568       268.6           0.08996            0.06444           0.0000

     concave points_worst  symmetry_worst  fractal_dimension_worst
0                  0.2654          0.4601                  0.11890
1                  0.1860          0.2750                  0.08902
2                  0.2430          0.3613                  0.08758
3                  0.2575          0.6638                  0.17300
4                  0.1625          0.2364                  0.07678
..                    …               …                        …
564                0.2216          0.2060                  0.07115
```

```
565                0.1628          0.2572                0.06637
566                0.1418          0.2218                0.07820
567                0.2650          0.4087                0.12400
568                0.0000          0.2871                0.07039

[569 rows x 31 columns]
```

[14]:
```
scalar.fit(df2.drop(['diagnosis'],axis=1))
```

[14]:
```
StandardScaler()
```

[15]:
```
Scaled_features=scalar.transform(df2.drop(['diagnosis'],axis=1))
```

[16]:
```
Scaled_features
```

[16]:
```
array([[ 1.09706398, -2.07333501,  1.26993369, …,  2.29607613,
         2.75062224,  1.93701461],
       [ 1.82982061, -0.35363241,  1.68595471, …,  1.0870843 ,
        -0.24388967,  0.28118999],
       [ 1.57988811,  0.45618695,  1.56650313, …,  1.95500035,
         1.152255  ,  0.20139121],
       …,
       [ 0.70228425,  2.0455738 ,  0.67267578, …,  0.41406869,
        -1.10454895, -0.31840916],
       [ 1.83834103,  2.33645719,  1.98252415, …,  2.28998549,
         1.91908301,  2.21963528],
       [-1.80840125,  1.22179204, -1.81438851, …, -1.74506282,
        -0.04813821, -0.75120669]])
```

[17]:
```
df_feat=pd.DataFrame(Scaled_features)
```

[18]:
```
df_feat
```

[18]:
```
           0         1         2         3         4         5         6   \
0    1.097064 -2.073335  1.269934  0.984375  1.568466  3.283515  2.652874
1    1.829821 -0.353632  1.685955  1.908708 -0.826962 -0.487072 -0.023846
2    1.579888  0.456187  1.566503  1.558884  0.942210  1.052926  1.363478
3   -0.768909  0.253732 -0.592687 -0.764464  3.283553  3.402909  1.915897
4    1.750297 -1.151816  1.776573  1.826229  0.280372  0.539340  1.371011
..        …         …         …         …         …         …         …
564  2.110995  0.721473  2.060786  2.343856  1.041842  0.219060  1.947285
565  1.704854  2.085134  1.615931  1.723842  0.102458 -0.017833  0.693043
566  0.702284  2.045574  0.672676  0.577953 -0.840484 -0.038680  0.046588
567  1.838341  2.336457  1.982524  1.735218  1.525767  3.272144  3.296944
568 -1.808401  1.221792 -1.814389 -1.347789 -3.112085 -1.150752 -1.114873

           7         8         9   …        20        21        22  \
```

```
0    2.532475  2.217515  2.255747  …  1.886690 -1.359293  2.303601
1    0.548144  0.001392 -0.868652  …  1.805927 -0.369203  1.535126
2    2.037231  0.939685 -0.398008  …  1.511870 -0.023974  1.347475
3    1.451707  2.867383  4.910919  … -0.281464  0.133984 -0.249939
4    1.428493 -0.009560 -0.562450  …  1.298575 -1.466770  1.338539
..        …       …       …  …       …        …        …
564  2.320965 -0.312589 -0.931027  …  1.901185  0.117700  1.752563
565  1.263669 -0.217664 -1.058611  …  1.536720  2.047399  1.421940
566  0.105777 -0.809117 -0.895587  …  0.561361  1.374854  0.579001
567  2.658866  2.137194  1.043695  …  1.961239  2.237926  2.303601
568 -1.261820 -0.820070 -0.561032  … -1.410893  0.764190 -1.432735

           23        24        25        26        27        28        29
0    2.001237  1.307686  2.616665  2.109526  2.296076  2.750622  1.937015
1    1.890489 -0.375612 -0.430444 -0.146749  1.087084 -0.243890  0.281190
2    1.456285  0.527407  1.082932  0.854974  1.955000  1.152255  0.201391
3   -0.550021  3.394275  3.893397  1.989588  2.175786  6.046041  4.935010
4    1.220724  0.220556 -0.313395  0.613179  0.729259 -0.868353 -0.397100
..        …       …       …       …       …       …       …
564  2.015301  0.378365 -0.273318  0.664512  1.629151 -1.360158 -0.709091
565  1.494959 -0.691230 -0.394820  0.236573  0.733827 -0.531855 -0.973978
566  0.427906 -0.809587  0.350735  0.326767  0.414069 -1.104549 -0.318409
567  1.653171  1.430427  3.904848  3.197605  2.289985  1.919083  2.219635
568 -1.075813 -1.859019 -1.207552 -1.305831 -1.745063 -0.048138 -0.751207

[569 rows x 30 columns]
```

[19]:
```
df_feat.isna().sum()
```

[19]:
```
0     0
1     0
2     0
3     0
4     0
5     0
6     0
7     0
8     0
9     0
10    0
11    0
12    0
13    0
14    0
15    0
16    0
17    0
```

```
18    0
19    0
20    0
21    0
22    0
23    0
24    0
25    0
26    0
27    0
28    0
29    0
dtype: int64
```

[20]:
```
from sklearn.model_selection import train_test_split
```

[21]:
```
x=df_feat
y=df2['diagnosis']
```

[22]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.
↪2,random_state=101)
```

[23]:
```
x_train.shape
```

[23]:
```
(455, 30)
```

[24]:
```
y_test.shape
```

[24]:
```
(114,)
```

[25]:
```
y_train.shape
```

[25]:
```
(455,)
```

[26]:
```
x_test.shape
```

[26]:
```
(114, 30)
```

[27]:
```
from sklearn.neighbors import KNeighborsClassifier
```

[28]:
```
KNN=KNeighborsClassifier(n_neighbors=3)
```

[29]:
```
KNN
```

[29]:
```
KNeighborsClassifier(n_neighbors=3)
```

[30]:
```
KNN.fit(x_train,y_train)
```

[30]:
```
KNeighborsClassifier(n_neighbors=3)
```

[31]:
```
y_pred=KNN.predict(x_test)
y_pred
```

[31]:
```
array([1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1,
       1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0], dtype=int64)
```

[32]:
```
from sklearn.metrics import accuracy_score
accuracy=accuracy_score(y_test,y_pred)
accuracy
```

[32]:
```
0.956140350877193
```

[33]:
```
import numpy as np
error_rate=[]
for val in range(1,30):
    knn=KNeighborsClassifier(n_neighbors=val)
    knn.fit(x_train,y_train)
    pred_i=knn.predict(x_test)
    error_rate.append(np.mean(pred_i!=y_test))
```

[34]:
```
error_rate
```

[34]:
```
[0.05263157894736842,
 0.07017543859649122,
 0.043859649122807015,
 0.03508771929824561,
 0.03508771929824561,
 0.03508771929824561,
 0.03508771929824561,
 0.02631578947368421,
 0.03508771929824561,
 0.017543859649122806,
 0.02631578947368421,
 0.017543859649122806,
 0.02631578947368421,
 0.02631578947368421,
 0.043859649122807015,
 0.043859649122807015,
 0.05263157894736842,
 0.043859649122807015,
 0.05263157894736842,
```

```
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842,
 0.05263157894736842]
```

[ ]:

[ ]:
```
PROJECT-2
```

[ ]:
```
Water Quality Testing using KNN
```

[35]:
```
import numpy as np
import pandas as pd
df=pd.read_csv(r"C:\Users\user\Downloads\water_potability.csv")
df
```

[35]:

	ph	Hardness	Solids	Chloramines	Sulfate	\
0	NaN	204.890455	20791.318981	7.300212	368.516441	
1	3.716080	129.422921	18630.057858	6.635246	NaN	
2	8.099124	224.236259	19909.541732	9.275884	NaN	
3	8.316766	214.373394	22018.417441	8.059332	356.886136	
4	9.092223	181.101509	17978.986339	6.546600	310.135738	
…	…	…	…	…	…	
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	
3272	7.808856	193.553212	17329.802160	8.061362	NaN	
3273	9.419510	175.762646	33155.578218	7.350233	NaN	
3274	5.126763	230.603758	11983.869376	6.303357	NaN	
3275	7.874671	195.102299	17404.177061	7.509306	NaN	

	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	564.308654	10.379783	86.990970	2.963135	0
1	592.885359	15.180013	56.329076	4.500656	0
2	418.606213	16.868637	66.420093	3.055934	0
3	363.266516	18.436524	100.341674	4.628771	0
4	398.410813	11.558279	31.997993	4.075075	0
…	…	…	…	…	…
3271	526.424171	13.894419	66.687695	4.435821	1
3272	392.449580	19.903225	NaN	2.798243	1
3273	432.044783	11.039070	69.845400	3.298875	1
3274	402.883113	11.168946	77.488213	4.708658	1
3275	327.459760	16.140368	78.698446	2.309149	1

```
[3276 rows x 10 columns]
```

[36]:
```
df.fillna(0,inplace=True)
```

[37]:
```
df
```

[37]:
```
             ph    Hardness        Solids  Chloramines     Sulfate  \
0      0.000000  204.890455  20791.318981     7.300212  368.516441
1      3.716080  129.422921  18630.057858     6.635246    0.000000
2      8.099124  224.236259  19909.541732     9.275884    0.000000
3      8.316766  214.373394  22018.417441     8.059332  356.886136
4      9.092223  181.101509  17978.986339     6.546600  310.135738
...         ...         ...           ...          ...         ...
3271   4.668102  193.681735  47580.991603     7.166639  359.948574
3272   7.808856  193.553212  17329.802160     8.061362    0.000000
3273   9.419510  175.762646  33155.578218     7.350233    0.000000
3274   5.126763  230.603758  11983.869376     6.303357    0.000000
3275   7.874671  195.102299  17404.177061     7.509306    0.000000

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436524       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
...            ...             ...              ...        ...         ...
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225         0.000000   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459760       16.140368        78.698446   2.309149           1

[3276 rows x 10 columns]
```

[38]:
```
df[df.Potability==0]
```

[38]:
```
             ph    Hardness        Solids  Chloramines     Sulfate  \
0      0.000000  204.890455  20791.318981     7.300212  368.516441
1      3.716080  129.422921  18630.057858     6.635246    0.000000
2      8.099124  224.236259  19909.541732     9.275884    0.000000
3      8.316766  214.373394  22018.417441     8.059332  356.886136
4      9.092223  181.101509  17978.986339     6.546600  310.135738
...         ...         ...           ...          ...         ...
3112   6.616731  195.096968  34277.760400     7.632639    0.000000
3113   7.734569  230.919506  21776.594455     6.908591    0.000000
3114   6.971577  185.906938  27959.987873     7.214510  349.743879
```

```
3115  4.709187  179.141018  22291.418577     6.774276  407.417977
3116  5.230003  176.714023  27971.891806     7.597981  413.914001

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436524       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
…              …              …                …          …           …
3112    417.465080       13.432557        47.945936   3.622379           0
3113    395.114961       15.033557        92.697369   3.821456           0
3114    414.067354       19.882917        36.179003   3.226349           0
3115    371.264843       18.186801        86.528627   3.860084           0
3116    440.355374       14.423614        72.837370   3.045612           0

[1998 rows x 10 columns]
```

[39]:
```
df[df.Potability==1]
```

[39]:
```
            ph    Hardness        Solids  Chloramines     Sulfate  \
250   9.445130  145.805402  13168.529156     9.444471  310.583374
251   9.024845  128.096691  19859.676476     8.016423  300.150377
252   0.000000  169.974849  23403.637304     8.519730    0.000000
253   6.800119  242.008082  39143.403329     9.501695  187.170714
254   7.174135  203.408935  20401.102461     7.681806  287.085679
…          …         …           …           …         …
3271  4.668102  193.681735  47580.991603     7.166639  359.948574
3272  7.808856  193.553212  17329.802160     8.061362    0.000000
3273  9.419510  175.762646  33155.578218     7.350233    0.000000
3274  5.126763  230.603758  11983.869376     6.303357    0.000000
3275  7.874671  195.102299  17404.177061     7.509306    0.000000

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
250     592.659021        8.606397        77.577460   3.875165           1
251     451.143481       14.770863        73.778026   3.985251           1
252     475.573562       12.924107        50.861913   2.747313           1
253     376.456593       11.432466        73.777275   3.854940           1
254     315.549900       14.533510        74.405616   3.939896           1
…              …              …                …          …           …
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225         0.000000   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459760       16.140368        78.698446   2.309149           1

[1278 rows x 10 columns]
```

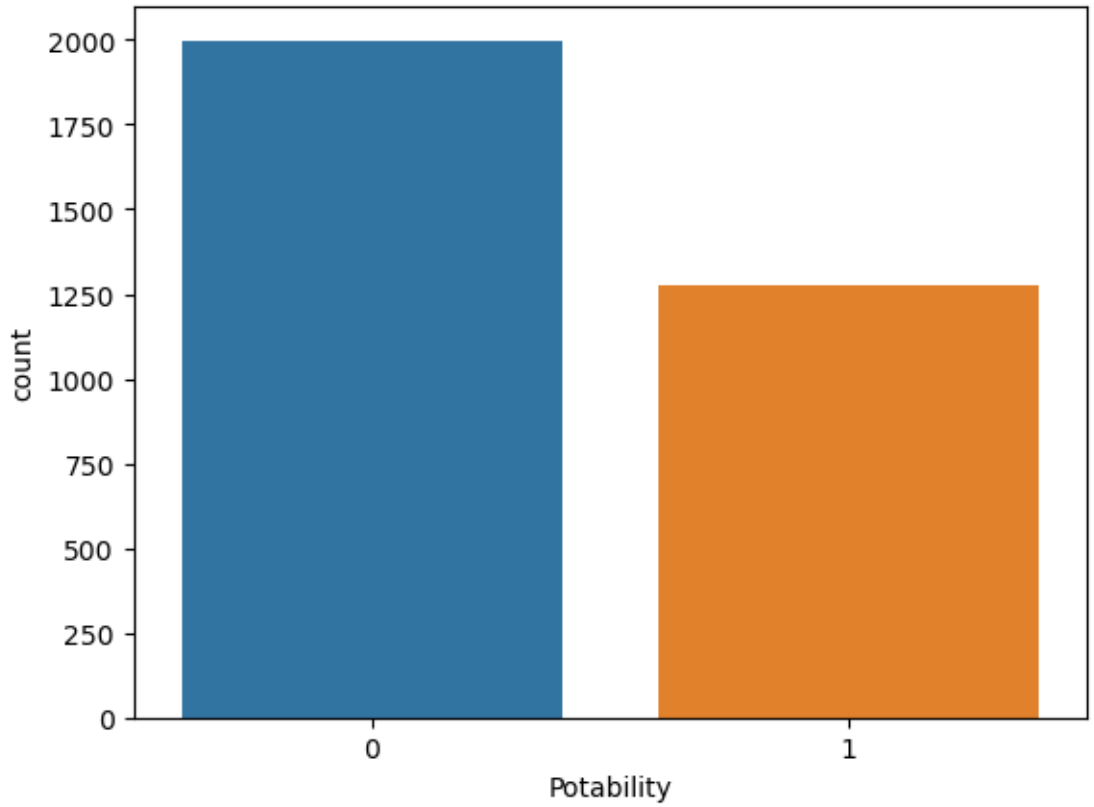
[40]:
```
df.isna().sum()
```

[40]:
```
ph                 0
Hardness           0
Solids             0
Chloramines        0
Sulfate            0
Conductivity       0
Organic_carbon     0
Trihalomethanes    0
Turbidity          0
Potability         0
dtype: int64
```

[75]:
```
print(df['Potability'].value_counts())

# plot class distribution
sns.countplot(x='Potability',data=df)
plt.show()
```

```
Potability
0    1998
1    1278
Name: count, dtype: int64
```



[ ]:

[41]:
```
from sklearn.preprocessing import StandardScaler
```

[42]:
```
scl=StandardScaler()
```

[43]:
```
scl.fit(df.drop('Potability',axis=1))
```

[43]:
```
StandardScaler()
```

[44]:
```
scaled_features=scl.transform(df.drop('Potability',axis=1))
```

[45]:
```
scaled_features
```

[45]:
```
array([[-2.05883481,  0.25919471, -0.13947087, …, -1.18065057,
         1.11839374, -1.28629758],
       [-0.78784149, -2.03641367, -0.38598665, …,  0.27059724,
        -0.31774225,  0.68421789],
       [ 0.7112704 ,  0.84766483, -0.24004734, …,  0.78111686,
         0.15489886, -1.16736546],
       …,
       [ 1.16287583, -0.62682923,  1.27080989, …, -0.98132923,
         0.31533277, -0.85600678],
       [-0.30535229,  1.0413545 , -1.14405809, …, -0.94206382,
         0.67330539,  0.95079738],
       [ 0.63450186, -0.03854623, -0.52581194, …,  0.56094007,
         0.72999007, -2.12445866]])
```

[46]:
```
df_feat=pd.DataFrame(scaled_features)
```

[47]:
```
df_feat
```

[47]:

	0	1	2	3	4	5	6
0	-2.058835	0.259195	-0.139471	0.112415	0.779002	1.708954	-1.180651
1	-0.787841	-2.036414	-0.385987	-0.307694	-1.732306	2.062575	0.270597
2	0.711270	0.847665	-0.240047	1.360594	-1.732306	-0.094032	0.781117
3	0.785709	0.547651	0.000493	0.592008	0.699746	-0.778830	1.255134
4	1.050935	-0.464429	-0.460249	-0.363698	0.381159	-0.343939	-0.824357
…	…	…	…	…	…	…	…
3271	-0.462226	-0.081758	2.916188	0.028027	0.720615	1.240155	-0.118075
3272	0.611991	-0.085667	-0.534295	0.593290	-1.732306	-0.417706	1.698560
3273	1.162876	-0.626829	1.270810	0.144017	-1.732306	0.072263	-0.981329
3274	-0.305352	1.041355	-1.144058	-0.517373	-1.732306	-0.288597	-0.942064
3275	0.634502	-0.038546	-0.525812	0.244515	-1.732306	-1.221919	0.560940

```
             7         8
0     1.118394 -1.286298
1    -0.317742  0.684218
2     0.154899 -1.167365
3     1.743711  0.848412
4    -1.457357  0.138786
…         …         …
3271  0.167433  0.601124
3272 -2.956073 -1.497627
3273  0.315333 -0.856007
3274  0.673305  0.950797
3275  0.729990 -2.124459

[3276 rows x 9 columns]
```

[48]:
```
from sklearn.model_selection import train_test_split
```

[49]:
```
x=df_feat
y=df['Potability']
```

[50]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.
↪2,random_state=42)
```

[51]:
```
x_train.shape
```

[51]:
```
(2620, 9)
```

[52]:
```
y_test.shape
```

[52]:
```
(656,)
```

[53]:
```
from sklearn.neighbors import KNeighborsClassifier
```

[54]:
```
KNN=KNeighborsClassifier(n_neighbors=3)
```

[55]:
```
KNN.fit(x_train,y_train)
```

[55]:
```
KNeighborsClassifier(n_neighbors=3)
```

[56]:
```
y_pred=KNN.predict(x_test)
y_pred
```

[56]:
```
array([1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1,
       1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0,
```

```
0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0,
1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1,
1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,
0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1,
1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0,
1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0,
0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0,
0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1,
1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0,
0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1,
0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0,
0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1,
1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0,
0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1,
1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1,
0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1,
1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0,
0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1], dtype=int64)
```

[58]:
```
from sklearn.metrics import accuracy_score
accuracy=accuracy_score(y_test,y_pred)
accuracy
```

[58]:
```
0.5487804878048781
```

[59]:
```
error_rate=[]
for val in range(1,30):
    knn=KNeighborsClassifier(n_neighbors=val)
    knn.fit(x_train,y_train)
    pred_i=knn.predict(x_test)
    error_rate.append(np.mean(pred_i!=y_test))
```

[60]:
```
error_rate
```

[60]:
```
[0.4573170731707317,
 0.39939024390243905,
 0.45121951219512196,
 0.4100609756097561,
 0.4329268292682927,
 0.3719512195121951,
```

```
0.4024390243902439,
0.3780487804878049,
0.3978658536585366,
0.3978658536585366,
0.3948170731707317,
0.3856707317073171,
0.38109756097560976,
0.3795731707317073,
0.35823170731707316,
0.3597560975609756,
0.36585365853658536,
0.37347560975609756,
0.37347560975609756,
0.36128048780487804,
0.36585365853658536,
0.36585365853658536,
0.36890243902439024,
0.3551829268292683,
0.36890243902439024,
0.36128048780487804,
0.3719512195121951,
0.36128048780487804,
0.36585365853658536]
```

[ ]: