Q4_WineQualityDataset_code

August 29, 2024

[93]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
df = pd.read_csv('C:/Users/qa284/Desktop/@IIT Bhilai cousre work/ML/
 ↪Assignment_01_theory/Dataset/WineQT.csv')
```

0.0.1 Original Dataset (without normalization)

[74]:
```
print("\nDataFrame before scaling:")
df.head()
```

```
DataFrame before scaling:
```

[74]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

0.0.2 Defining custom filter for "quality" feature

[77]:
```
def custom_filter(x):
    if x <= 6:
        return 0
    else:
        return 1
```

0.0.3 Normalizing the "quality" feature according to the condition given in the question

[80]:
```
df['quality'] = df['quality'].apply(custom_filter)
print("Scaled'quality' feature:")
df['quality']
```

```
Scaled'quality' feature:
```

[80]:
```
0       0
1       0
2       0
3       0
4       0
       ..
1138    0
1139    0
1140    0
1141    0
1142    0
Name: quality, Length: 1143, dtype: int64
```

0.0.4 Removing the "Id" and "quality" columns from the list of columns beacuse they dont require any min-max scaling

[83]:
```
print("Columns before removing 'Id' and 'quality':")
print("*"*150)
print(df.columns.tolist())

print("="*150)

columns_to_normalize = df.columns.tolist()
columns_to_normalize.remove('Id')            #Removing Id column
columns_to_normalize.remove('quality')       #Removing quality column

print("\nColumns to be Scaled:")
print("*"*150)
print(columns_to_normalize)
```

```
Columns before removing 'Id' and 'quality':
******************************************************************************************************************************************************
```

```
**********************************************************************
['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH',
'sulphates', 'alcohol', 'quality', 'Id']
==========================================================================
==========================================================================

Columns to be Scaled:
**************************************************************************
**************************************************************************
['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH',
'sulphates', 'alcohol']
```

0.0.5 Applying min-max scaling to the filtered columns

[86]:
```
df[columns_to_normalize] = (df[columns_to_normalize] - df[columns_to_normalize].
↪min()) / (df[columns_to_normalize].max() - df[columns_to_normalize].min())
```

[88]:
```
print("\nDataFrame after scaling:")
df.head()
```

```
DataFrame after scaling:
```

[88]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	0.247788	0.397260	0.00	0.068493	0.106845	
1	0.283186	0.520548	0.00	0.116438	0.143573	
2	0.283186	0.438356	0.04	0.095890	0.133556	
3	0.584071	0.109589	0.56	0.068493	0.105175	
4	0.247788	0.397260	0.00	0.068493	0.106845	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	0.149254	0.098940	0.567548	0.606299	0.137725	
1	0.358209	0.215548	0.494126	0.362205	0.209581	
2	0.208955	0.169611	0.508811	0.409449	0.191617	
3	0.238806	0.190813	0.582232	0.330709	0.149701	
4	0.149254	0.098940	0.567548	0.606299	0.137725	

	alcohol	quality	Id
0	0.153846	0	0
1	0.215385	0	1
2	0.215385	0	2
3	0.215385	0	3
4	0.153846	0	4