Discretization

Data discretization is a preprocessing technique used to transform continuous or numerical data into discrete categories or bins. This process can be particularly useful for machine learning algorithms that handle categorical data more effectively than continuous data. It can also help to improve model interpretability and handle outliers more robustly.

In [1]:
```
import pandas as pd
import seaborn as sns
from sklearn.preprocessing import KBinsDiscretizer


df = sns.load_dataset('titanic')

df.head()
```

Out[1]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	Fals
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	Fals
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	Tru
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	Fals
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	Tru

In [2]:
```
# impute missing values
df['age'] = df['age'].fillna(df['age'].median())
df['fare'] = df['fare'].fillna(df['fare'].median())


df.head()
```

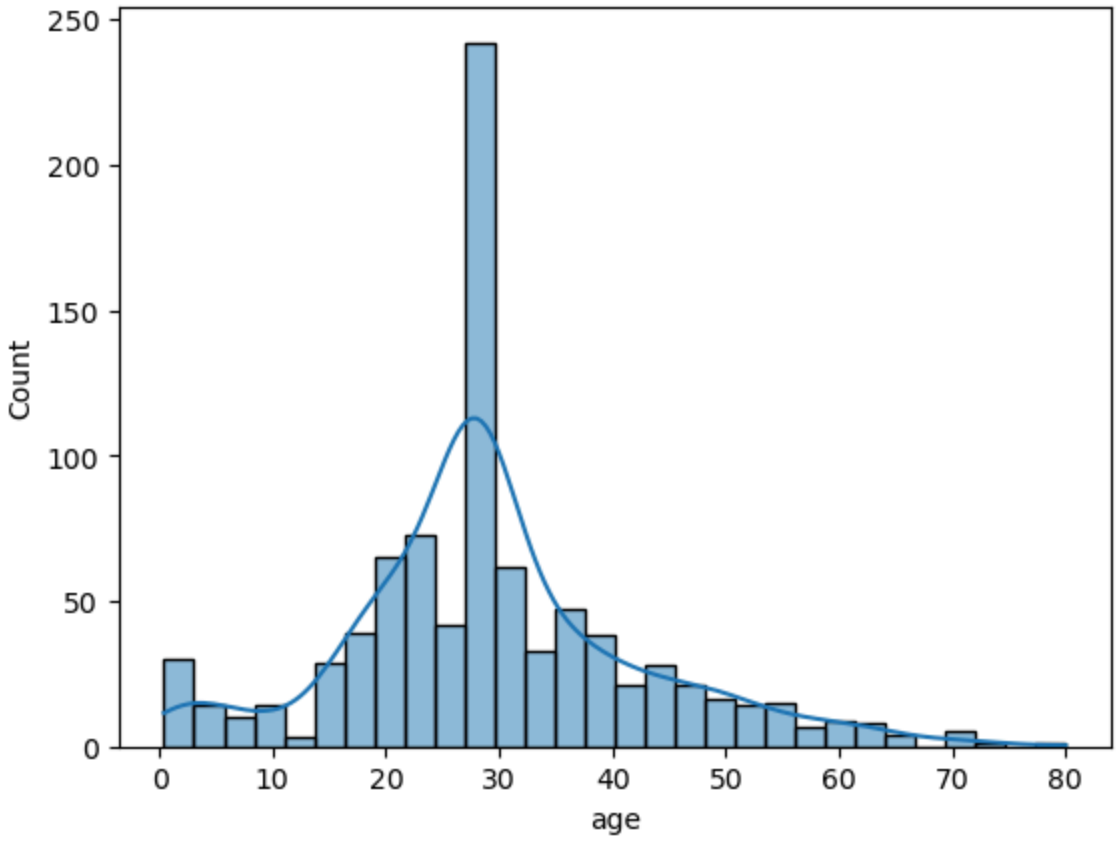
Out[2]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	Fals
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	Fals
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	Tru
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	Fals
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	Tru

In [3]:
```
sns.histplot(df['age'], kde=True)
```

```
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is dep
recated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Out[3]:
```
<Axes: xlabel='age', ylabel='Count'>
```



In [4]:
```
# age discretization
age_discretizer = KBinsDiscretizer(n_bins=3, encode='ordinal', strategy='uniform')
df['age_bin'] = age_discretizer.fit_transform(df[['age']])
df.head()
```

Out[4]:

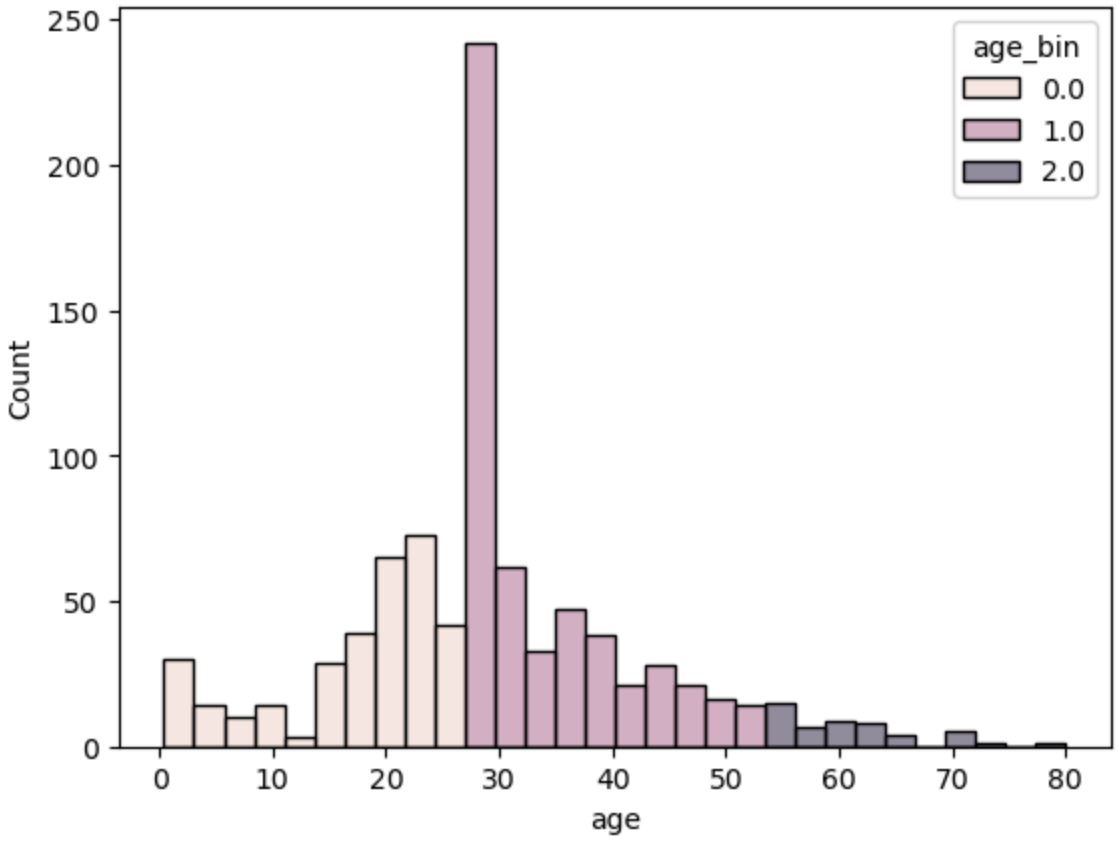
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alon
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	Fals
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	Fals
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	Tru
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	Fals
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	Tru

In [5]:

```
sns.histplot(df, x='age', hue = 'age_bin')
```

```
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is dep
recated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

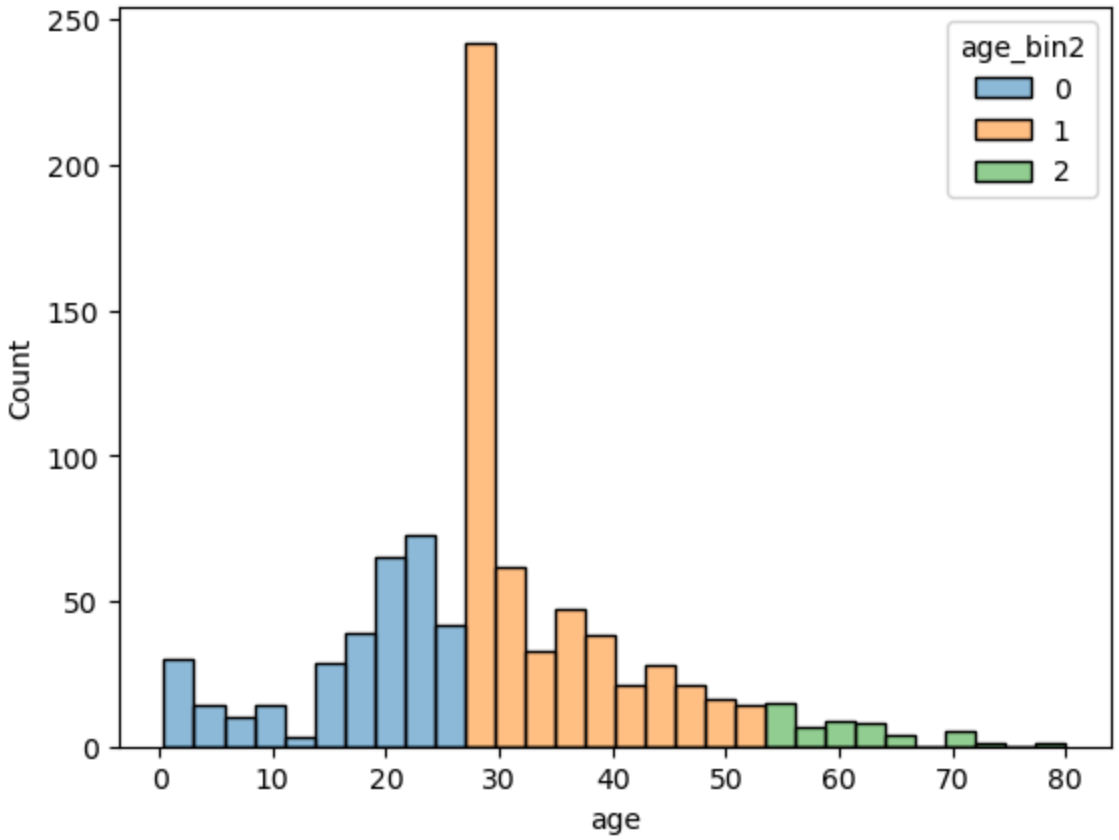
Out[5]: <Axes: xlabel='age', ylabel='Count'>



In [6]:
```
# pandas method for manual binning
df['age_bin2'] = pd.cut(df['age'], bins=3, labels=[0,1,2])
sns.histplot(df, x='age', hue = 'age_bin2')
```

```
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is dep
recated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1057: FutureWarning: The default of observed=Fal
se is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain current beh
avior or observed=True to adopt the future default and silence this warning.
  grouped_data = data.groupby(
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

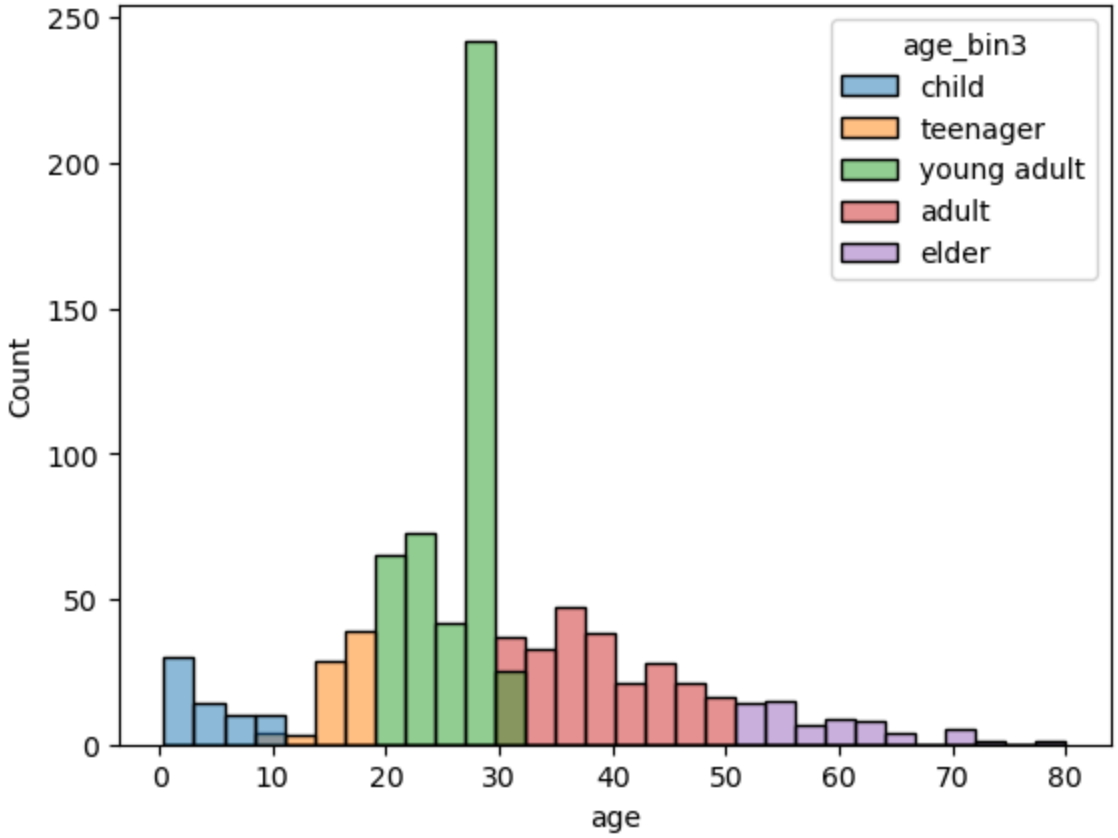
Out[6]:
```
<Axes: xlabel='age', ylabel='Count'>
```



In [7]:
```
# define the bins
bins = [0, 10, 18, 30, 50, 80]
labels = ['child', 'teenager', 'young adult', 'adult', 'elder']
df['age_bin3'] = pd.cut(df['age'], bins=bins, labels=labels)
sns.histplot(df, x='age', hue = 'age_bin3')
```

```
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is dep
recated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1057: FutureWarning: The default of observed=Fal
se is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain current beh
avior or observed=True to adopt the future default and silence this warning.
  grouped_data = data.groupby(
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
C:\Users\ustb\.anaconda\anwaar\Lib\site-packages\seaborn\_oldcore.py:1075: FutureWarning: When grouping with a length
-1 list-like, you will need to pass a length-1 tuple to get_group in a future version of pandas. Pass `(name,)` inste
ad of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

Out[7]:
```
<Axes: xlabel='age', ylabel='Count'>
```


age_bin3
child
teenager
young adult
adult
elder
Count
age
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80

In [ ]: