In [1]:

```
import pandas as pd
import numpy as np

import seaborn as sns
from matplotlib import pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.naive_bayes import MultinomialNB as MB
from sklearn.naive_bayes import GaussianNB as GB
```

In [2]:

```
data_train = pd.read_csv("SalaryData_Train.csv")
data_test = pd.read_csv("SalaryData_Test.csv")
data_train.head()
```

Out[2]:

	age	workclass	education	educationno	maritalstatus	occupation	relationship	race
0	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White
1	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White
2	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White
3	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black
4	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black

In [3]:
```
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30161 entries, 0 to 30160
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   age            30161 non-null  int64 
 1   workclass      30161 non-null  object
 2   education      30161 non-null  object
 3   educationno    30161 non-null  int64 
 4   maritalstatus  30161 non-null  object
 5   occupation     30161 non-null  object
 6   relationship   30161 non-null  object
 7   race           30161 non-null  object
 8   sex            30161 non-null  object
 9   capitalgain    30161 non-null  int64 
 10  capitalloss    30161 non-null  int64 
 11  hoursperweek   30161 non-null  int64 
 12  native         30161 non-null  object
 13  Salary         30161 non-null  object
dtypes: int64(5), object(9)
memory usage: 3.2+ MB
```

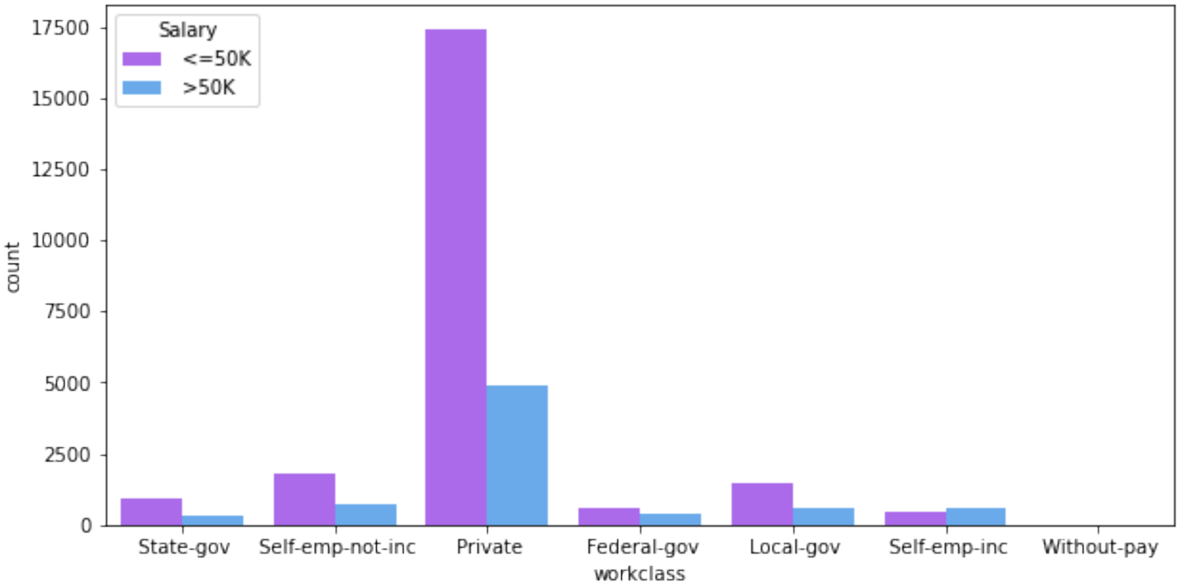
In [4]:
```
data_train.describe()
```

Out[4]:

	age	educationno	capitalgain	capitalloss	hoursperweek
count	30161.000000	30161.000000	30161.000000	30161.000000	30161.000000
mean	38.438115	10.121316	1092.044064	88.302311	40.931269
std	13.134830	2.550037	7406.466611	404.121321	11.980182
min	17.000000	1.000000	0.000000	0.000000	1.000000
25%	28.000000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	10.000000	0.000000	0.000000	40.000000
75%	47.000000	13.000000	0.000000	0.000000	45.000000
max	90.000000	16.000000	99999.000000	4356.000000	99.000000

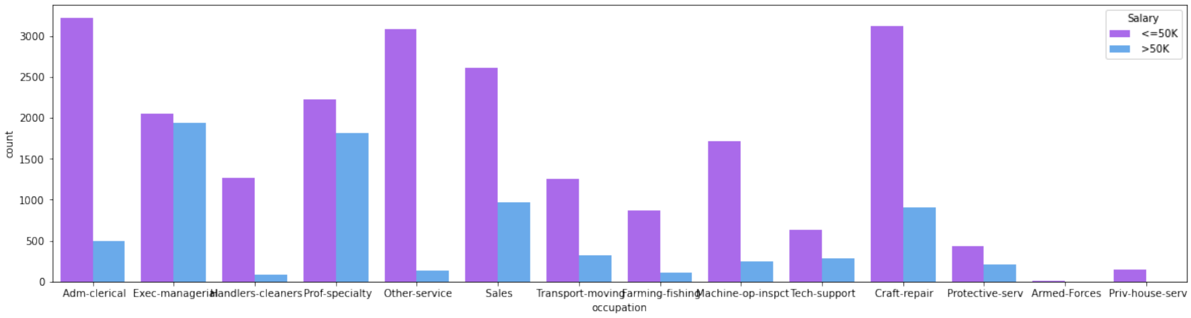
In [5]:
```
dims = (10,5)
fig, ax = plt.subplots(figsize=dims)
sns.countplot(ax = ax, data=data_train,x='workclass',hue='Salary',p
```

Out[5]:
```
<AxesSubplot:xlabel='workclass', ylabel='count'>
```



In [6]:
```
dims = (20,5)
fig, ax = plt.subplots(figsize=dims)
sns.countplot(data=data_train,x='occupation',hue='Salary',palette='
```

Out[6]:
```
<AxesSubplot:xlabel='occupation', ylabel='count'>
```



In [7]:
```
data_train.Salary.value_counts()
```

Out[7]:
```
 <=50K    22653
 >50K      7508
Name: Salary, dtype: int64
```

In [8]:
```
data_test.Salary.value_counts()
```

Out[8]:
```
 <=50K    11360
 >50K      3700
Name: Salary, dtype: int64
```

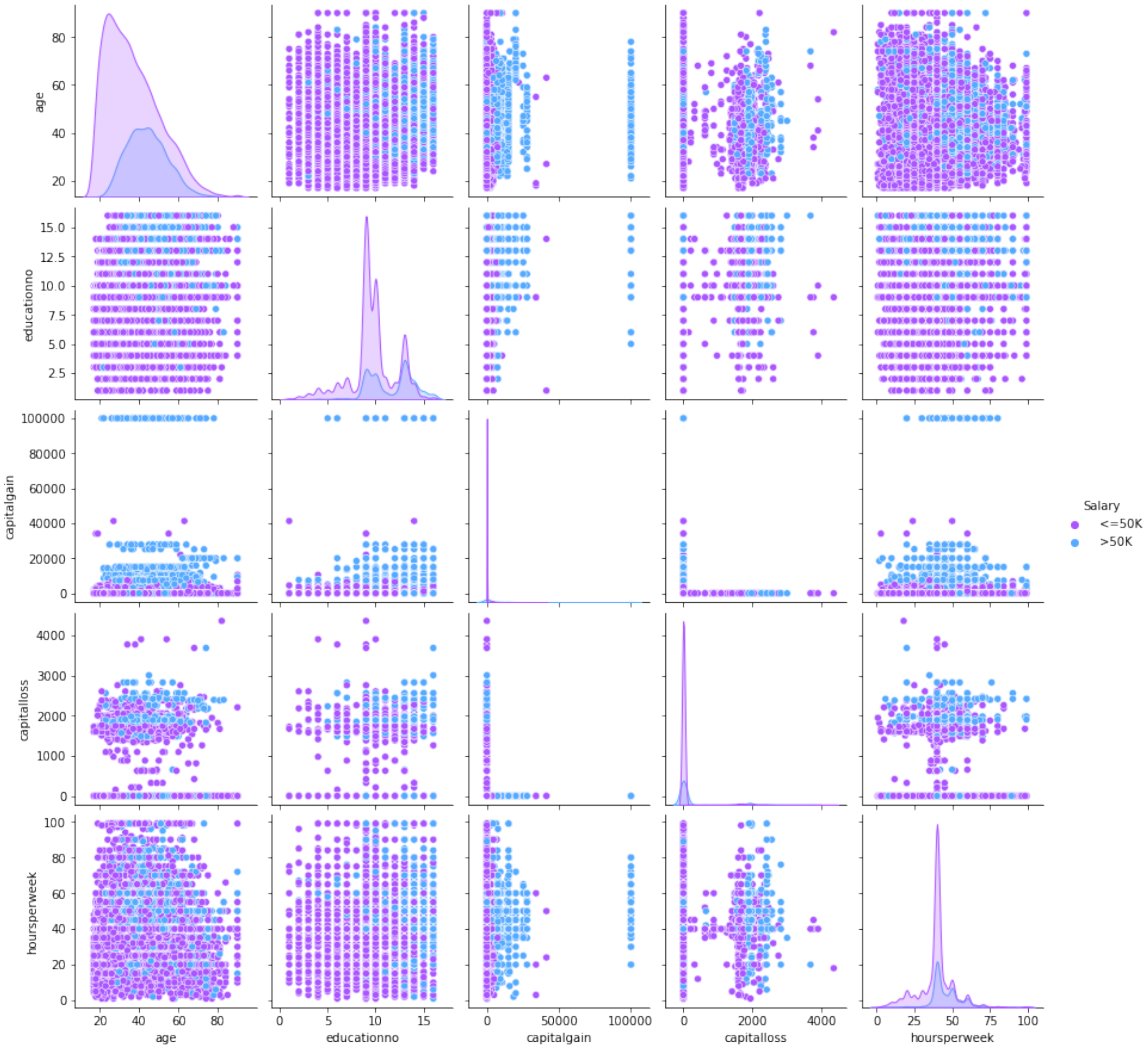
```
In [9]: data_train.occupation.value_counts()
```

```
Out[9]:  Prof-specialty       4038
 Craft-repair         4030
 Exec-managerial      3992
 Adm-clerical         3721
 Sales                3584
 Other-service        3212
 Machine-op-inspct    1965
 Transport-moving     1572
 Handlers-cleaners    1350
 Farming-fishing       989
 Tech-support          912
 Protective-serv       644
 Priv-house-serv       143
 Armed-Forces            9
Name: occupation, dtype: int64
```

Visualization EDA

```
In [10]: sns.pairplot(data_train,hue='Salary',palette='cool_r')
```

```
Out[10]: <seaborn.axisgrid.PairGrid at 0x7ff1cda83f70>
```



Feature Engineering

In [11]:

```
labels = ['workclass', 'education', 'maritalstatus', 'occupation',
dftrain = data_train.copy()
dftest = data_test.copy()
label_encoder = preprocessing.LabelEncoder()
for x in labels:
    dftrain[x] = label_encoder.fit_transform(dftrain[x])
    dftest[x] = label_encoder.fit_transform(dftest[x])
dftrain.head()
```

Out[11]:

	age	workclass	education	educationno	maritalstatus	occupation	relationship	race	s
0	39	5	9	13	4	0	1	4	
1	50	4	9	13	2	3	0	4	
2	38	2	11	9	0	5	1	4	
3	53	2	1	7	2	5	0	2	
4	28	2	9	13	2	9	5	2	

Train test split

In [12]:

```
X_train = dftrain.iloc[:,:-1]
y_train = dftrain['Salary']
X_test = dftest.iloc[:,:-1]
y_test = dftest['Salary']
```

Naive Bayes Classifier

In [13]:

```
model_mb = MB()
model_mb.fit(X_train,y_train)
```

Out[13]:

```
▾ MultinomialNB
MultinomialNB()
```

In [14]:

```
model_gb = GB()
model_gb.fit(X_train,y_train)
```

Out[14]:

```
▾ GaussianNB
GaussianNB()
```

Evaluation

In [15]:
```
pip install scikit-plot
```

```
Requirement already satisfied: scikit-plot in /opt/anaconda3/lib/p
ython3.9/site-packages (0.3.7)
Requirement already satisfied: scipy>=0.9 in /opt/anaconda3/lib/py
thon3.9/site-packages (from scikit-plot) (1.7.1)
Requirement already satisfied: matplotlib>=1.4.0 in /opt/anaconda3
/lib/python3.9/site-packages (from scikit-plot) (3.4.3)
Requirement already satisfied: scikit-learn>=0.18 in /opt/anaconda
3/lib/python3.9/site-packages (from scikit-plot) (1.2.0)
Requirement already satisfied: joblib>=0.10 in /opt/anaconda3/lib/
python3.9/site-packages (from scikit-plot) (1.2.0)
Requirement already satisfied: pyparsing>=2.2.1 in /opt/anaconda3/
lib/python3.9/site-packages (from matplotlib>=1.4.0->scikit-plot)
(3.0.4)
Requirement already satisfied: cycler>=0.10 in /opt/anaconda3/lib/
python3.9/site-packages (from matplotlib>=1.4.0->scikit-plot) (0.1
0.0)
Requirement already satisfied: python-dateutil>=2.7 in /opt/anacon
da3/lib/python3.9/site-packages (from matplotlib>=1.4.0->scikit-pl
ot) (2.8.2)
Requirement already satisfied: numpy>=1.16 in /opt/anaconda3/lib/p
ython3.9/site-packages (from matplotlib>=1.4.0->scikit-plot) (1.20
.3)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/anaconda3
/lib/python3.9/site-packages (from matplotlib>=1.4.0->scikit-plot)
(1.3.1)
Requirement already satisfied: pillow>=6.2.0 in /opt/anaconda3/lib
/python3.9/site-packages (from matplotlib>=1.4.0->scikit-plot) (8.
4.0)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.9
/site-packages (from cycler>=0.10->matplotlib>=1.4.0->scikit-plot)
(1.16.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/anacon
da3/lib/python3.9/site-packages (from scikit-learn>=0.18->scikit-p
lot) (2.2.0)
Note: you may need to restart the kernel to use updated packages.
```

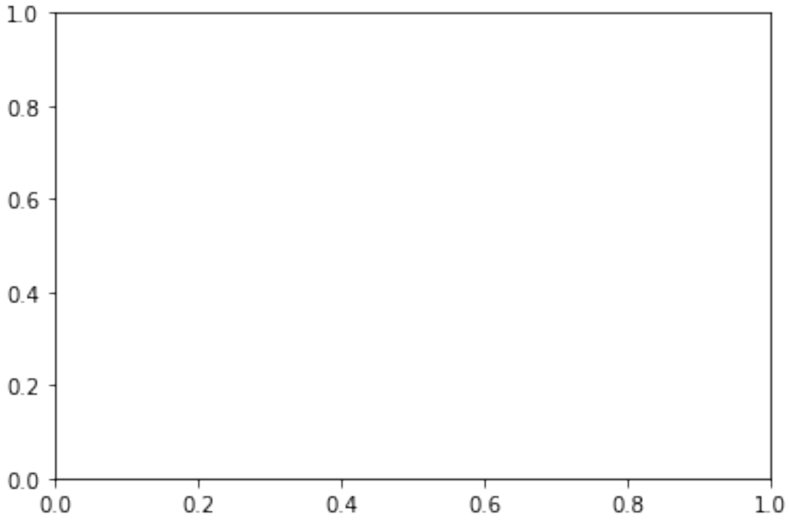
In [18]:
```
from scikitplot.estimators import plot_feature_importances
from scikitplot.metrics import plot_confusion_matrix, plot_roc
from sklearn.metrics import classification_report
```

In [19]:
```
def report(model):
    preds = model.predict(X_test)
    print(classification_report(y_test,preds))
    plot_confusion_matrix(model,X_test,y_test)
#MultinomialNB Evaluation
print('MultinomialNB')
report(model_mb) #model has high inbuilt bias
```

```
MultinomialNB
              precision    recall  f1-score   support
```

```
           0       0.79      0.96      0.87     11360
           1       0.62      0.21      0.32      3700

    accuracy                           0.77     15060
   macro avg       0.71      0.58      0.59     15060
weighted avg       0.75      0.77      0.73     15060


---------------------------------------------------------------------------
TypeError                                 Traceback (most recent c
all last)
/var/folders/9_/ckpgdd3s4qzg3w1zytsfvsmh0000gn/T/ipykernel_8571/28
20648027.py in <module>
      5 #MultinomialNB Evaluation
      6 print('MultinomialNB')
----> 7 report(model_mb) #model has high inbuilt bias


/var/folders/9_/ckpgdd3s4qzg3w1zytsfvsmh0000gn/T/ipykernel_8571/28
20648027.py in report(model)
      2     preds = model.predict(X_test)
      3     print(classification_report(y_test,preds))
----> 4     plot_confusion_matrix(model,X_test,y_test)
      5 #MultinomialNB Evaluation
      6 print('MultinomialNB')

/opt/anaconda3/lib/python3.9/site-packages/scikitplot/metrics.py
in plot_confusion_matrix(y_true, y_pred, labels, true_labels, pred
_labels, title, normalize, hide_zeros, hide_counts, x_tick_rotatio
n, ax, figsize, cmap, title_fontsize, text_fontsize)
    115         fig, ax = plt.subplots(1, 1, figsize=figsize)
    116
--> 117     cm = confusion_matrix(y_true, y_pred, labels=labels)
    118     if labels is None:
    119         classes = unique_labels(y_true, y_pred)

/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classi
fication.py in confusion_matrix(y_true, y_pred, labels, sample_wei
ght, normalize)
    315     (0, 2, 1, 1)
    316     """
--> 317     y_type, y_true, y_pred = _check_targets(y_true, y_pred
)
    318     if y_type not in ("binary", "multiclass"):
    319         raise ValueError("%s is not supported" % y_type)

/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classi
fication.py in _check_targets(y_true, y_pred)
     84     y_pred : array or indicator matrix
     85     """
---> 86     check_consistent_length(y_true, y_pred)
     87     type_true = type_of_target(y_true, input_name="y_true"
)
```

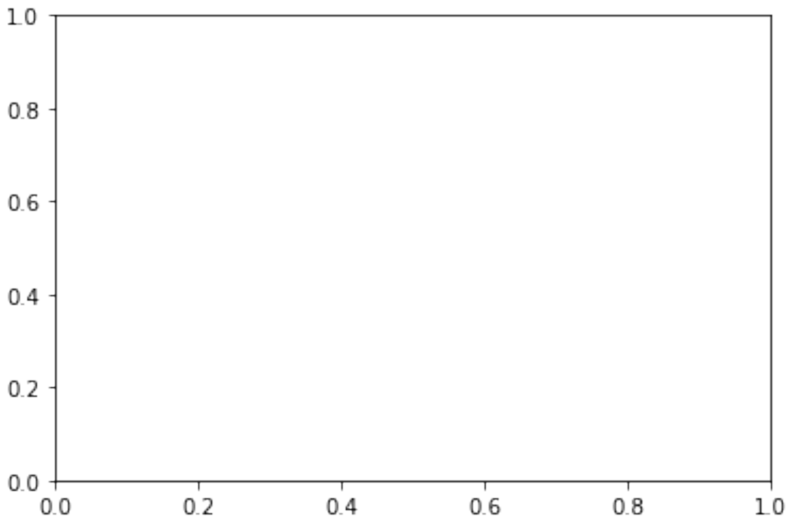
```
)
     88     type_pred = type_of_target(y_pred, input_name="y_pred"
)

/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/validatio
n.py in check_consistent_length(*arrays)
    392     """
    393 
--> 394     lengths = [_num_samples(X) for X in arrays if X is not
None]
    395     uniques = np.unique(lengths)
    396     if len(uniques) > 1:

/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/validatio
n.py in <listcomp>(.0)
    392     """
    393 
--> 394     lengths = [_num_samples(X) for X in arrays if X is not
None]

    395     uniques = np.unique(lengths)
    396     if len(uniques) > 1:

/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/validatio
n.py in _num_samples(x)
    333     if hasattr(x, "shape") and x.shape is not None:
    334         if len(x.shape) == 0:
--> 335             raise TypeError(
    336                 "Singleton array %r cannot be considered a
valid collection." % x
    337             )

TypeError: Singleton array array(MultinomialNB(), dtype=object) ca
nnot be considered a valid collection.
```



In [20]:
```
print('GaussianNB')
report(model_gb) #model has high inbuilt bias but better results as
```

```
GaussianNB
              precision    recall  f1-score   support

           0       0.81      0.95      0.87     11360
           1       0.67      0.33      0.44      3700

    accuracy                           0.79     15060
   macro avg       0.74      0.64      0.66     15060
weighted avg       0.78      0.79      0.77     15060


------------------------------------------------------------------
---------
TypeError                                 Traceback (most recent c
all last)
/var/folders/9_/ckpgdd3s4qzg3w1zytsfvsmh0000gn/T/ipykernel_8571/41
90215798.py in <module>
      1 print('GaussianNB')
----> 2 report(model_gb) #model has high inbuilt bias but better r
esults as compared to multinomial

/var/folders/9_/ckpgdd3s4qzg3w1zytsfvsmh0000gn/T/ipykernel_8571/28
20648027.py in report(model)
      2     preds = model.predict(X_test)
      3     print(classification_report(y_test,preds))
----> 4     plot_confusion_matrix(model,X_test,y_test)
      5 #MultinomialNB Evaluation
      6 print('MultinomialNB')

/opt/anaconda3/lib/python3.9/site-packages/scikitplot/metrics.py
in plot_confusion_matrix(y_true, y_pred, labels, true_labels, pred
_labels, title, normalize, hide_zeros, hide_counts, x_tick_rotatio
n, ax, figsize, cmap, title_fontsize, text_fontsize)
    115         fig, ax = plt.subplots(1, 1, figsize=figsize)
    116
--> 117     cm = confusion_matrix(y_true, y_pred, labels=labels)
    118     if labels is None:
    119         classes = unique_labels(y_true, y_pred)

/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classi
fication.py in confusion_matrix(y_true, y_pred, labels, sample_wei
ght, normalize)
    315     (0, 2, 1, 1)
    316     """
--> 317     y_type, y_true, y_pred = _check_targets(y_true, y_pred
)
    318     if y_type not in ("binary", "multiclass"):
    319         raise ValueError("%s is not supported" % y_type)

/opt/anaconda3/lib/python3.9/site-packages/sklearn/metrics/_classi
fication.py in _check_targets(y_true, y_pred)
     84     y_pred : array or indicator matrix
     85     """
---> 86     check_consistent_length(y_true, y_pred)
```

```
      87     type_true = type_of_target(y_true, input_name="y_true"
)
      88     type_pred = type_of_target(y_pred, input_name="y_pred"
)

/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/validatio
n.py in check_consistent_length(*arrays)
    392     """
    393
--> 394     lengths = [_num_samples(X) for X in arrays if X is not
None]
    395     uniques = np.unique(lengths)
    396     if len(uniques) > 1:


/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/validatio
n.py in <listcomp>(.0)
    392     """
    393
--> 394     lengths = [_num_samples(X) for X in arrays if X is not
None]
    395     uniques = np.unique(lengths)
    396     if len(uniques) > 1:

/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/validatio
n.py in _num_samples(x)
    333     if hasattr(x, "shape") and x.shape is not None:
    334         if len(x.shape) == 0:
--> 335             raise TypeError(
    336                 "Singleton array %r cannot be considered a
valid collection." % x
    337             )

TypeError: Singleton array array(GaussianNB(), dtype=object) canno
t be considered a valid collection.
```



In [ ]: