Master Data Science and Business Analytics - Exam: Machine Learning

Directions

Fill the empty code cells in order to implement the described tasks and reproduce the given output.

A slight difference of your output in values and number of rows does not affect a positive evaluation

The program must be *reproducible*: repeated executions must give the same results

Workflow

1. load the data in memory
2. drop the useless data
3. separe the predicting attributes X from the class attribute y
4. split X and y into training and test
5. train a classifier of your choice and find the best parameter setting using **cross validation**, optimize for best **accuracy**
6. show a classification report for the training set
7. test the optimized classifier with the *test set* and show a classification report

In [1]:

```
# insert your imports here
```

In [2]:

```
# insert here your initial variable settings and load the data
```

Have a quick look to the data.

- use the .shape attribute to see the size
- use the `.head()` function to see column names and some data
- use the `.hist()` method for an histogram of the numeric columns
- show an histogram of the target column
- use seaborn pairplot to show the numeric data, use the target values as color

In [3]:

```
Shape of the input data (1000, 6)
```

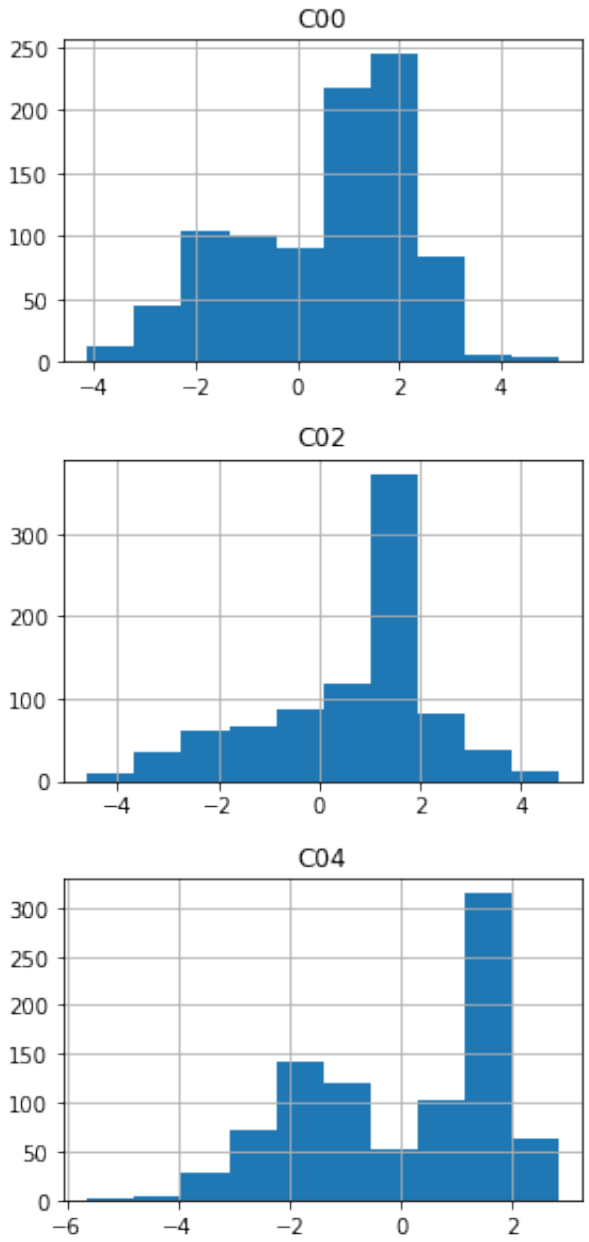
In [4]:

Out[4]:

	C00	C01	C02	C03	C04	Class
0	NaN	0.466367	-0.176765	1.546514	0.149219	c
1	NaN	-0.136792	1.551591	NaN	1.357674	a
2	2.712560	-0.495846	NaN	1.483562	1.656526	b
3	-2.166084	-0.582271	0.353011	-1.864210	-2.267033	b
4	2.848831	-0.507369	1.661752	1.466627	1.938519	b

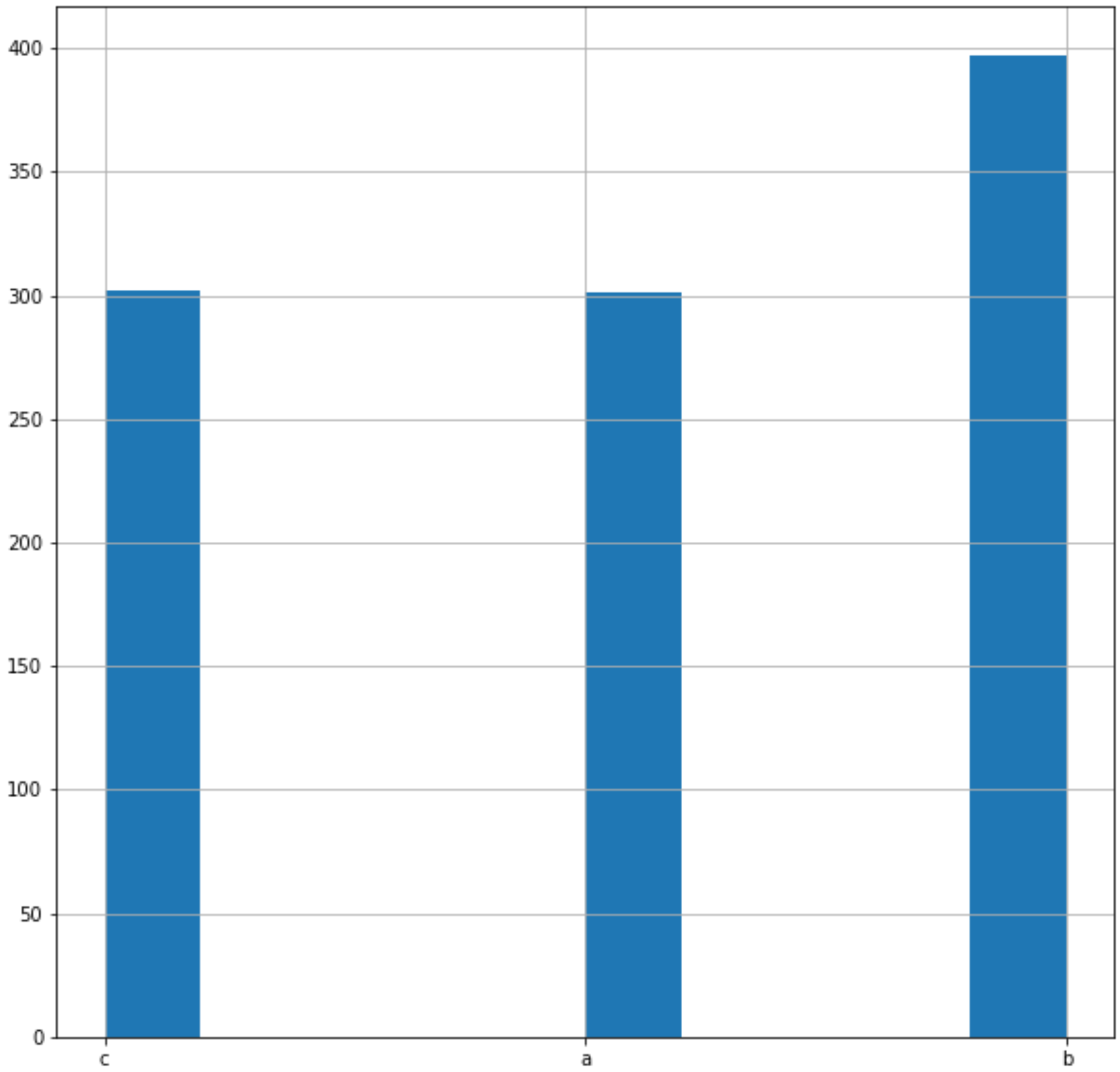
In [5]:

```
# generate histogram of numeric features
```

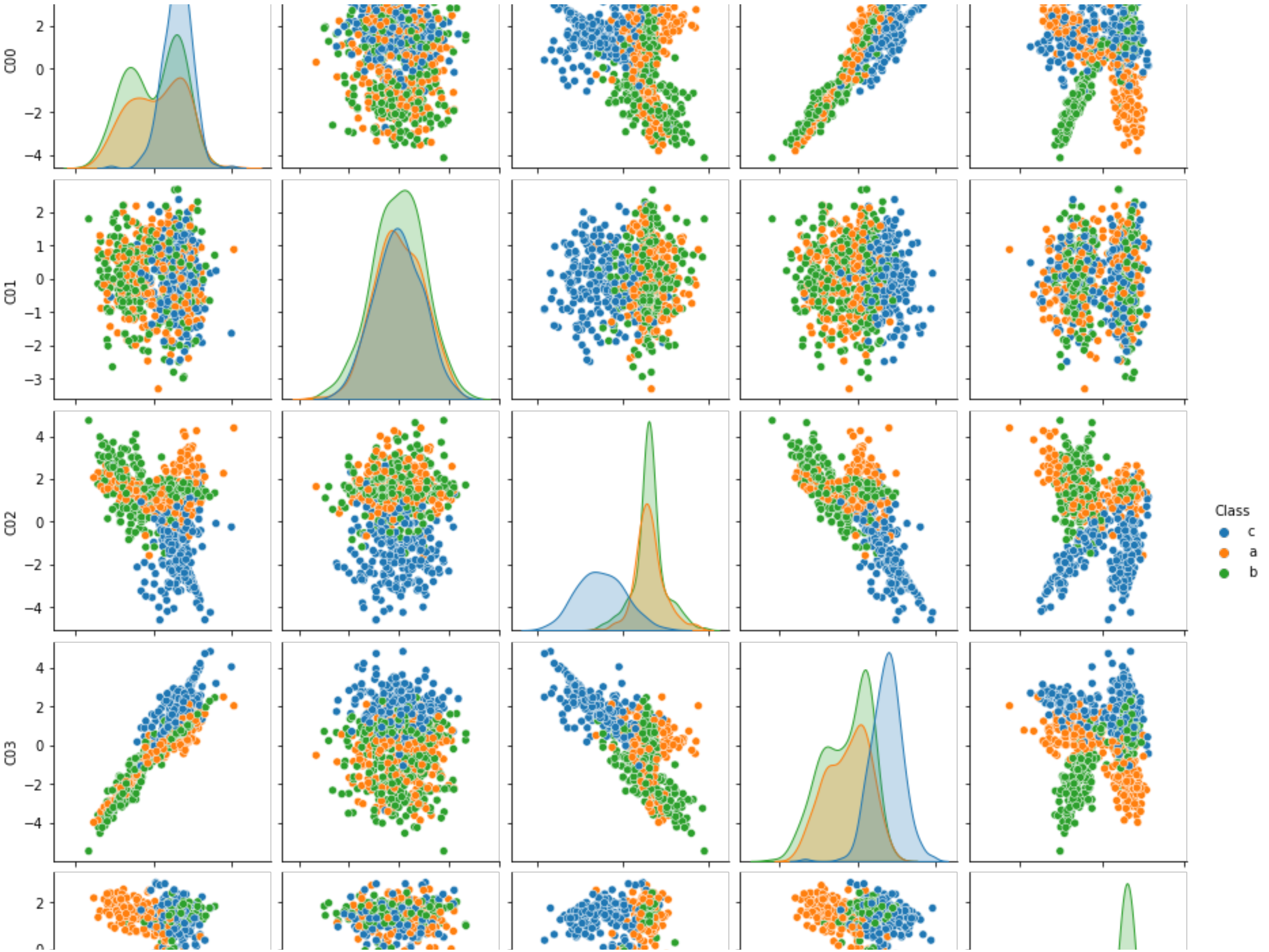


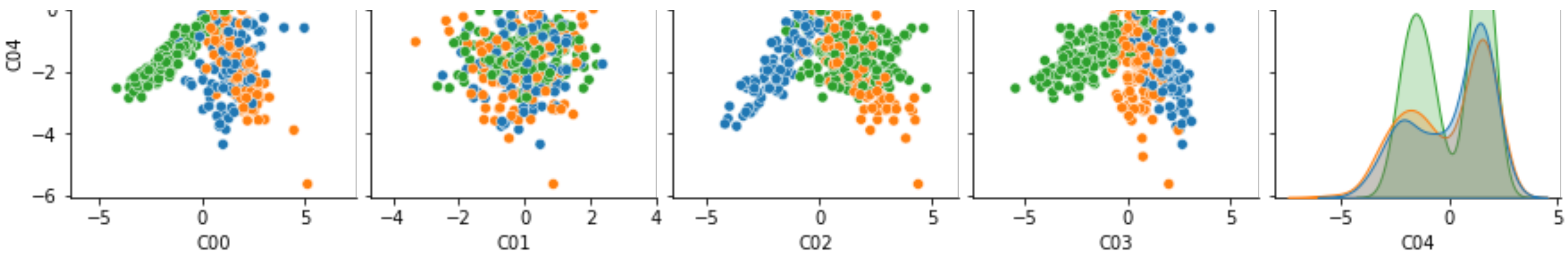
In [6]:

```
# generate histogram of target column
```



```
In [7]: # pairplot using target as color
```


C00
C01
C02
C03
2
0
−2
−4
1
−1
−3
4
Class
● c
● a
● b



Verify if there are nan values in the dataset, and, in case, drop rows with nan

In [8]:

```
There are 519 nan values
```

In [9]:

```
After drop there are 0 nan values
```

- Split predicting attributes and target into X and y
- Show the number of samples in train and test, show the number of features

In [10]:

```
There are 434 samples in the training dataset
There are 145 samples in the testing dataset
Each sample has 5 features
```

Optimising the estimator

- determine the range of the parameters for the estimator
- repeatedly fit the estimator with cross validation for each value of the parameter range and find the value of the parameter giving the best accuracy
- print the value of the best parameter

In [11]:

```
The best parameter value is 12
```

- fit the estimator using the `train` part
- use the fitted estimator to predict using the test features
- compute the accuracy on the test set and print it with the best parameter value
- print a classification report and the confusion matrix for the test set

In [12]:

```
```

```
The accuracy on test set tuned with cross_validation is 79.3% with parameter 12
```

In [13]:

```
# classification report on test set
```

```
              precision    recall  f1-score   support

           a       0.80      0.73      0.76        51
           b       0.80      0.89      0.84        54
           c       0.77      0.75      0.76        40

    accuracy                           0.79       145
   macro avg       0.79      0.79      0.79       145
weighted avg       0.79      0.79      0.79       145
```

In [14]:

```
# Confusion matrix for test set
```

```
[[37  8  6]
 [ 3 48  3]
 [ 6  4 30]]
```