Assignment 1 - Problem 1

CSC 732 Pattern Recognition and Neural Networks

Instructor: Prof. Natacha Gueorguieva

Contributors: Ali Necdet Guvercin, Benjamin Hermus, Kehang Wei

Date: February 26, 2024

Part 1.1

Import Libraries

```
# CSC 732 Hw1 - 1.1
# Ali Necdet Guvercin, Benjamin Hermus, Kehang Wei
# HW1 Part 1.1: Use matplotlib for plotting

# increase width of jupyter notebook cells
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))

# pandas is an open source, high-performance library with , easy-to-use data structures and data analysis tools for the Python programm
import pandas as pd
from pandas.plotting import scatter_matrix
# Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
import matplotlib.pyplot as plt
#sklearn Built on NumPy, SciPy, and matplotlib also used for data analysis
from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
#Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and info
import seaborn as sns
#numpy used to manipulate numerical data in python
import numpy as np
```

Import and explorer dataset

read_csv() pandas function

```
#Import the dataset
dataset = pd.read_csv ('/content/BankNote_Authentication.csv', header=None)
```

Setup column names

```
#Setup the column names
dataset.columns= ['variance','skewness','curtosis', 'entropy', 'class']
print (dataset)
print()
```

```
          variance  skewness  curtosis   entropy  class
     0     variance  skewness  curtosis   entropy  class
     1       3.6216    8.6661   -2.8073  -0.44699      0
     2       4.5459    8.1674   -2.4586   -1.4621      0
     3        3.866   -2.6383    1.9242   0.10645      0
     4       3.4566    9.5228   -4.0112   -3.5944      0
     ...        ...       ...       ...       ...    ...
     1368   0.40614    1.3492   -1.4501  -0.55949      1
     1369   -1.3887   -4.8773    6.4774   0.34179      1
     1370   -3.7503  -13.4586   17.5932   -2.7771      1
     1371   -3.5637   -8.3827    12.393   -1.2823      1
     1372   -2.5419  -0.65804    2.6842    1.1952      1

     [1373 rows x 5 columns]
```

Shape of dataset

```
# Print shape of dataset
print(dataset.shape)
print()
```

```
(1373, 5)
```

Peak at first 20 lines of dataset

```
# Peak at first 20 lines of dataset
print(dataset.head(20))
print()
```

```
    variance  skewness  curtosis   entropy  class
0   variance  skewness  curtosis   entropy  class
1     3.6216    8.6661   -2.8073  -0.44699      0
2     4.5459    8.1674   -2.4586   -1.4621      0
3      3.866   -2.6383    1.9242   0.10645      0
4     3.4566    9.5228   -4.0112   -3.5944      0
5    0.32924   -4.4552    4.5718   -0.9888      0
6     4.3684    9.6718   -3.9606   -3.1625      0
7     3.5912    3.0129   0.72888   0.56421      0
8     2.0922     -6.81    8.4636  -0.60216      0
9     3.2032    5.7588  -0.75345  -0.61251      0
10    1.5356    9.1772   -2.2718  -0.73535      0
11    1.2247    8.7779   -2.2135  -0.80647      0
12    3.9899   -2.7066    2.3946   0.86291      0
13    1.8993    7.6625   0.15394   -3.1108      0
14   -1.5768    10.843    2.5462   -2.9362      0
15     3.404    8.7261   -2.9915  -0.57242      0
16    4.6765   -3.3895    3.4896    1.4771      0
17    2.6719    3.0646   0.37158   0.58619      0
18   0.80355    2.8473    4.3439    0.6017      0
19    1.4479   -4.8794    8.3428   -2.1086      0
```

Generate descriptive statistics can be achieved with dataset.describe()

Descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.Analyzes both numeric and object series, as well as DataFrame column sets of mixed data types. The output will vary depending on what is provided.

```
# Print dataset describe
print(dataset.describe())
print()
```

```
        variance skewness curtosis  entropy class
count       1373     1373     1373     1373  1373
unique      1339     1257     1271     1157     3
top       0.5706  -4.4552   4.5718  -0.9888     0
freq           5        6        5        5   762
```

Class distribution of dataset

```
# Print class distribution of dataset
print(dataset.groupby('variance').size())
print()
```

```
variance
-0.0012852    1
-0.0068919    1
-0.014902     1
-0.016103     1
-0.023579     1
             ..
5.9374        1
```

```
6.0919        1
6.5633        1
6.8248        1
variance      1
Length: 1339, dtype: int64
```

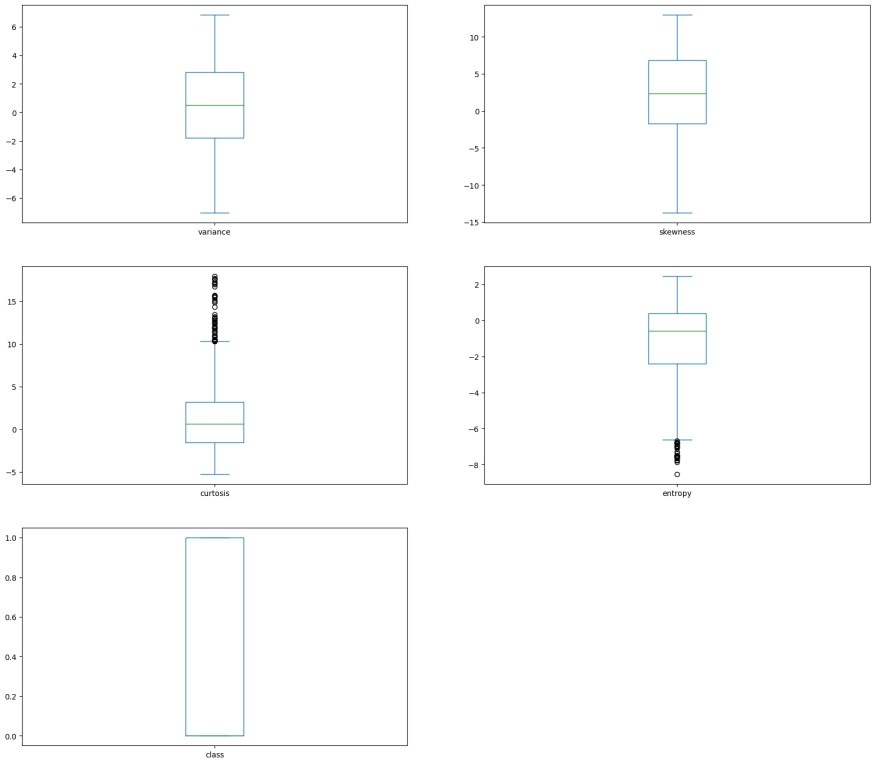
Visualize Dataset

Box or whisker Plots

A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median. Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.

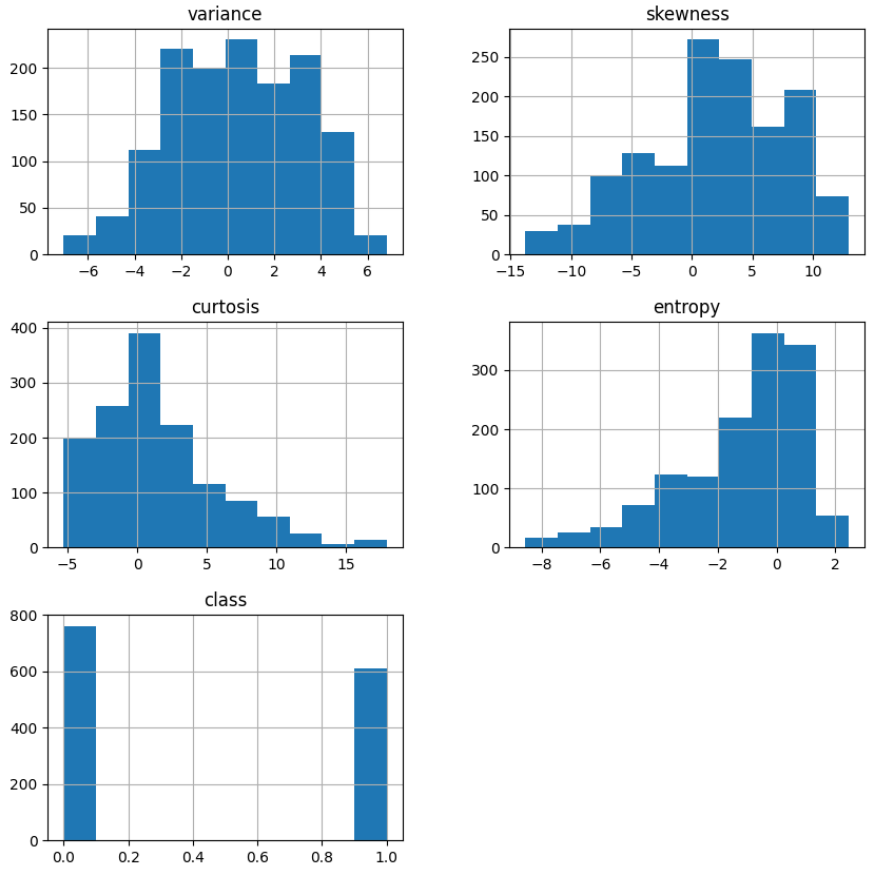
```
# convert all non numeric values to numeric
for col in dataset.columns:
    if dataset[col].dtype == 'object':
        dataset[col] = pd.to_numeric(dataset[col], errors='coerce')

#visualize dataset with box plot
dataset.plot(kind='box', subplots=True, layout=(5,2), sharex=False, sharey=False, figsize=(20,30))
plt.show()
print()
```



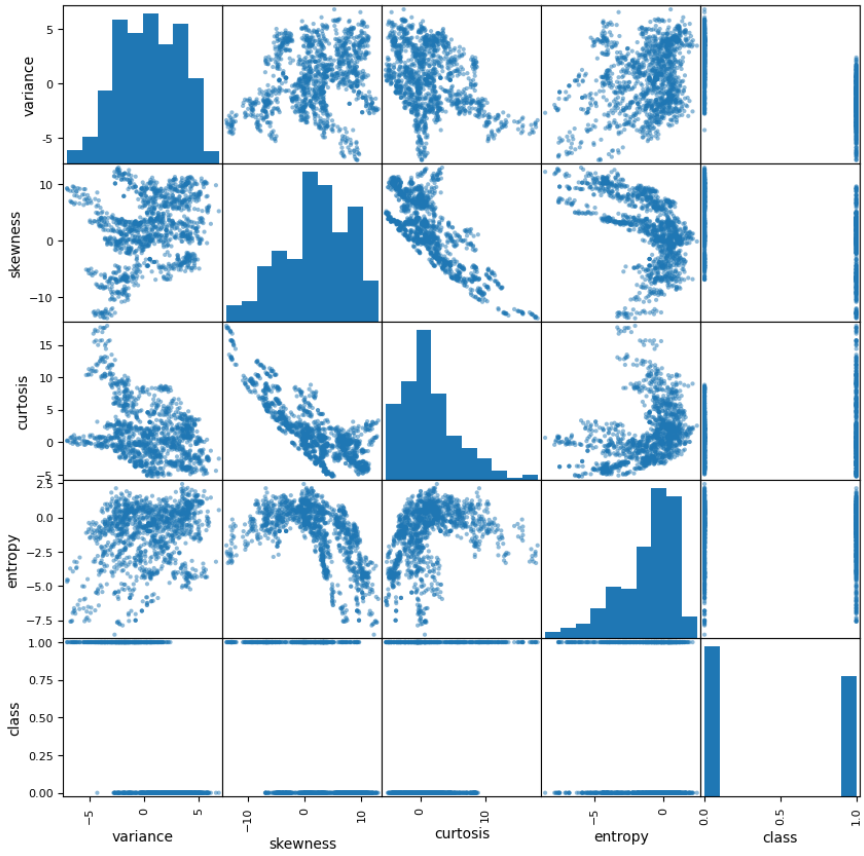
Histograms

```
# histograms
dataset.hist(figsize=(10,10))
plt.show()
print()
```



Scatter Plot Matrix

```
# scatter plot matrix
scatter_matrix(dataset, figsize=(10,10))
plt.show()
print()
```



There is a good concentration of the data on the regression line indicationg that most of the data is rightly placed. the outliers can also be seemainly in the assymetry coeficcient and the kernel groove length.