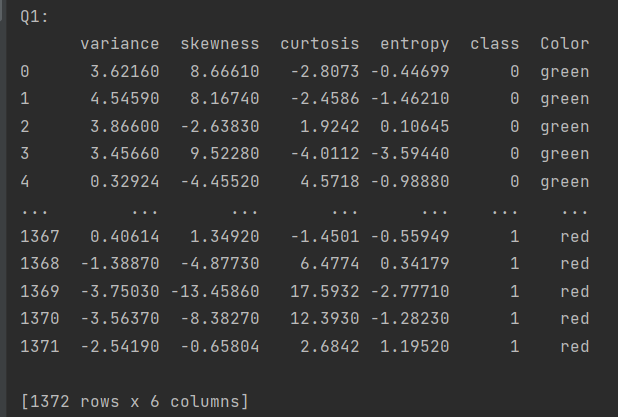
**Assignment 6**

# **Question 1:**

## **1. load the data into Pandas dataframe and add a column ”color”. For each class 0, this should contain ”green” and for each class 1 it should contain ”red”**



## **2. for each class and for each feature f1, f2, f3, f4, compute its mean µ() and standard deviation σ(). Round the results to 2 decimal places and summarize them in a table as shown below:**

| class | µ(f1) | σ(f1) | µ(f2) | σ(f2) | µ(f3) | µ(f3) | µ(f4) | µ(f4) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 2.28 | 2.02 | 4.26 | 5.14 | 0.80 | 3.24 | -1.15 | 2.13 |
| 1 | -1.87 | 1.88 | -0.99 | 5.40 | 2.15 | 5.26 | -1.25 | 2.07 |
| all | 0.43 | 2.84 | 1.92 | 5.87 | 1.40 | 4.31 | -1.19 | 2.10 |

## **3. examine your table. Are there any obvious patterns in the distribution of banknotes in each class**

**Answer:**

Yes, In my table, I find that the standard deviation is big, but the mean value is small.

# **Question 2:**

## **1. split your dataset X into training Xtrain and Xtesting parts (50/50 split). Using ”pairplot” from seaborn package, plot pairwise relationships in Xtrain separately for class 0 and class 1. Save your results into 2 pdf files ”good bills.pdf” and ”fake bills.pdf”**

**See the .py file and the .pdf file**

## **2. visually examine your results. Come up with three simple comparisons that you think may be sufficient to detect a fake bill.**

**See the Excel File attached (COL S - AA)**. We can know that if I choose >-1 >2.5 and >1 as the features to detect the fake bill. We can find that these features are more likely than the simple comparisons.

## **3. apply your simple classifier to Xtest and compute predicted class labels**

**See the Excel File attached (COL O - R).**

## **4. compare your predicted class labels with true labels in Xtest, compute the following: (a) TP (b) FP (c) TN (d) FN (e) TPR = TP/(TP + FN) (f) TNR = TN/(TN + FP)**

**The answer is together with q5**

## **5. summarize your findings in the table as shown below**:

| TP | FP | TN | FN | Accuracy | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- |
| 28 | 8 | 14 | 0 | 84% | 1 | 0.636364 |

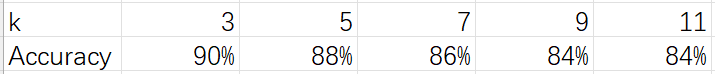
## **6. does you simple classifier gives you higher accuracy on identifying ”fake” bills or ”real” bills” Is your accuracy better than 50% (”coin” flipping)?**

YES, my accuracy is 84%

# **Question 3:**

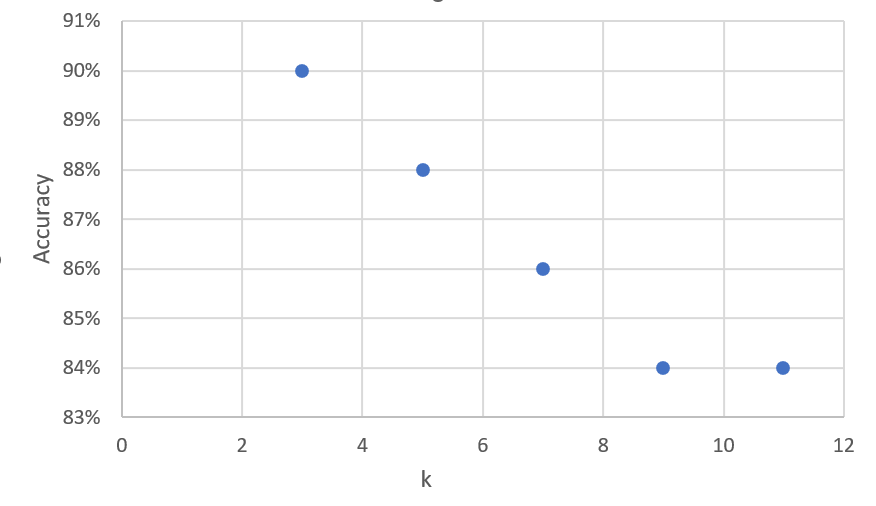
## **1. take k = 3, 5, 7, 9, 11. For each k, generate Xtrain and Xtest using 50/50 split as before. Train your k-NN classifier on Xtrain and compute its accuracy for Xtest**

According to my code and Excel File, the accuracy of each k is shown below



Note: All the data used to calculate are input to the code manually.

## **2. plot a graph showing the accuracy. On x-axis you plot k, and on y-axis, you plot accuracy. What is the optimal value k ∗ of k?**



The optimal value k\* of k is 3.

## **3. use the optimal value k ∗ to compute performance measures and summarize them in the table**

| TP | FP | TN | FN | Accuracy | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- |
| 24 | 5 | 21 | 0 | 90% | 1 | 0.807692 |

## **4. is your k-NN classifier better than your simple classifier for any of the measures from the previous table?**

Yes, because the accuracy of k-NN classifier is higher than my simple classifier, which is 90% > 80%.

## **5. consider a bill x that contains the last 4 digits of your BUID as feature values. What is the class label predicted for this bill by your simple classifier? What is the label for this bill predicted by k-NN using the best k ∗ ?**

For simple classifier, the labels for the last 4-digits of my BUID are 0,1,0,0

For k-NN , the label is 0 when I using best k\*.