Question 1

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

	f1	f2	f3	f4	class	color
0	3.62160	8.66610	-2.8073	-0.44699	0	green
1	4.54590	8.16740	-2.4586	-1.46210	0	green
2	3.86600	-2.63830	1.9242	0.10645	0	green
3	3.45660	9.52280	-4.0112	-3.59440	0	green
4	0.32924	-4.45520	4.5718	-0.98880	0	green
...	...	...	...	...	...	...
1367	0.40614	1.34920	-1.4501	-0.55949	1	red
1368	-1.38870	-4.87730	6.4774	0.34179	1	red
1369	-3.75030	-13.45860	17.5932	-2.77710	1	red
1370	-3.56370	-8.38270	12.3930	-1.28230	1	red
1371	-2.54190	-0.65804	2.6842	1.19520	1	red

1372 rows × 6 columns

2. for each class and for each feature f1, f2, f3, f4, compute its mean $\mu()$ 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)	σ(f3)	μ(f4)	σ(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

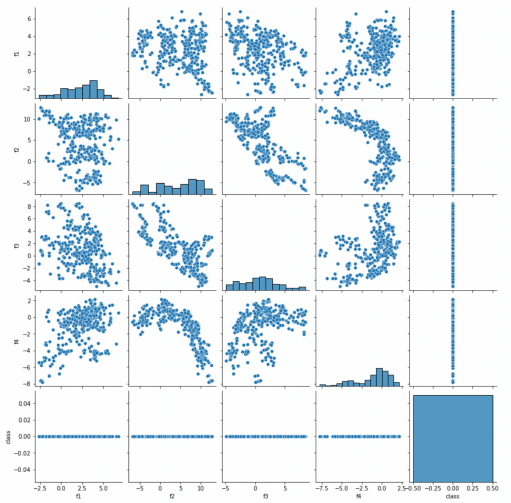
3. examine your table. Are there any obvious patterns in the distribution of banknotes in each class

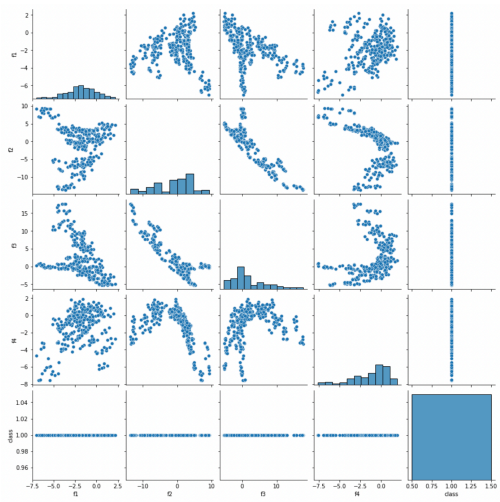
Yes, in class 1 and all, mean value of f4 is always negative, and in class 0 only mean of f4 is negative. Compare to class 0 and 1, the negative mean of f1 and f2 makes them different.

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”

   2 pdf has been saved in the homework3 folder.





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

   By visually examining my results, my 3 simple rules are:

```
if (row['f1'] < -3) and (row['f2'] < -8) and (row['f3'] > 10):
    predict = 'red'
else:
    predict = 'green'
```

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

	f1	f2	f3	f4	class	color	predict
202	-0.78689	9.5663	-3.78670	-7.50340	0	green	green
117	2.10800	6.7955	-0.17080	0.49050	0	green	green
733	3.82440	-3.1081	2.45370	0.52024	0	green	green
1325	-5.52500	6.3258	0.89768	-6.62410	1	red	green
1015	-0.90784	-7.9026	6.78070	0.34179	1	red	green
...	...	...	...	...	...	...	...
59	-0.78289	11.3603	-0.37644	-7.04950	0	green	green
1089	-2.98210	4.1986	-0.58980	-3.96420	1	red	green
1055	-0.60254	1.7237	-2.15010	-0.77027	1	red	green
527	2.53280	7.5280	-0.41929	-2.64780	0	green	green
1224	0.26877	4.9870	-5.15080	-6.39130	1	red	green

686 rows × 7 columns

4. comparing your predicted class labels with true labels, compute the following:

```
The value of TP is 390
The value of FP is 277
The value of TN is 19
The value of FN is 0
The value of TPR is 1.0
The value of TNR is 0.064189189189189
The value of accuracy is 0.5962099125364432
```

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

TP	FP	TN	FN	Accuracy	TPR	TNR
390	277	19	0	0.596	1.0	0.064

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

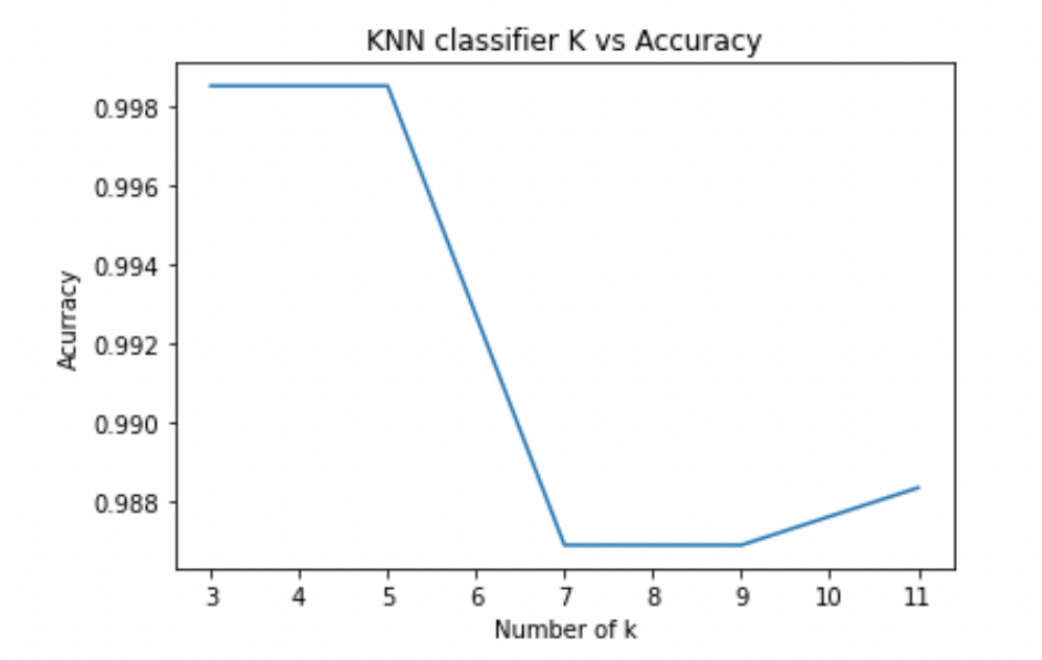
No, 59.6% is not a higher accuracy on identifying ”fake” bills or ”real” bills”, but it’s better than 50%.

Question 3 (use k-NN classifier using sklearn library)

1. take k = 3, 5, 7, 9, 11. Use the same Xtrain and Xtest as before. For each k, train your k-NN classifier on Xtrain and compute its accuracy for Xtest

```
Accuracy for k=3 is 0.9985422740524781
Accuracy for k=5 is 0.9985422740524781
Accuracy for k=7 is 0.9868804664723032
Accuracy for k=9 is 0.9868804664723032
Accuracy for k=11 is 0.9883381924198251
```

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 is 3 and 5.

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

TP	FP	TN	FN	Accuracy	TPR	TNR
388	0	296	2	0.997	0.995	1.0

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

Yes, my k-NN classifier is much better than my simple classifier for most of the measures from the previous table. Except for TP and TPR.

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 ∗ ?

```
The Prediction of my BUID as feature values by using my simple classifier is: green

The Prediction of my BUID as feature values by using best k* kNN classifier is: 0
```

Question 4:

1. take your best value k ∗ . For each of the four features f1, . . . , f4, drop that feature from both Xtrain and Xtest. Train your classifier on the ”truncated” Xtrain and predict labels on Xtest using just 3 remaining features. You will repeat this for 4 cases: (1) just f1 is missing,

(2) just f2 missing, (3) just f3 missing and (4) just f4 is missing. Compute the accuracy for each of these scenarious.

```
Accuracy for missing f1 feature is 0.9620991253644315
Accuracy for missing f2 feature is 0.9752186588921283
Accuracy for missing f3 feature is 0.9766763848396501
Accuracy for missing f4 feature is 0.9941690962099126
```

2. did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?

No, my best k=3 classifier gives accuracy 0.9985, after dropping any of the features, the highest accuracy is only 0.9942.

3. which feature, when removed, contributed the most to loss of accuracy?

When removing f1, it contributed the most to loss of accuracy.

4. which feature, when removed, contributed the least to loss of accuracy?

When removing f4, it contributed the least to loss of accuracy

Question 5 (use logistic (regression classifier using sklearn library)

1. Use the same Xtrain and Xtest as before. Train your logistic regression classifier on Xtrain and compute its accuracy for Xtest

```
The accuracy of Logistic Regression is 0.9868804664723032
```

2. summarize your performance measures in the table

TP	FP	TN	FN	Accuracy	TPR	TNR
383	2	294	7	0.987	0.982	0.993

3. is your logistic regression better than your simple classifier for any of the measures from the previous table?

My FP of logistic regression is less than my simple classifier, and my TN, TNR and Accuracy of logistic regression are more than my simple classifier, which are better.

4. is your logistic regression better than your k-NN classifier (using the best k $*$) for any of the measures from the previous table?

No, none of the measures from the logistic regression is better than my k-NN classifier with the best k*

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 x by logistic regression? Is it the same label as predicted by k-NN?

```
The Prediction of my BUID as feature values by using Logistic Regression classifier is: 0
```

Yes, it's the same label as predicted by k-NN.

Question 6:

1. For each of the four features f1, . . . , f4, drop that feature from both Xtrain and Xtest. Train your logistic regression classifier on the "truncated" Xtrain and predict labels on Xtest using just 3 remaining features. You will repeat this for 4 cases: (1) just f1 is missing, (2) just f2 missing, (3) just f3 missing and (4) just f4 is missing. Compute the accuracy for each of these scenarious.

```
Accuracy for missing f1 feature is 1.0
Accuracy for missing f2 feature is 0.9985422740524781
Accuracy for missing f3 feature is 1.0
Accuracy for missing f4 feature is 1.0
```

2. did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?

Yes, when f1, f3 or f4 missing respectively, the accuracy increases compared with accuracy when all features are used.

3. which feature, when removed, contributed the most to loss of accuracy?

When f2 is removed, it contributed the most to loss of accuracy

4. which feature, when removed, contributed the least to loss of accuracy?

When f1, f3 or f4 is removed, it contributed the most to loss of accuracy

5. is relative significance of features the same as you obtained using k-NN?

Both of them have a higher accuracy when f4 is removed.