Assignment 3

**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”

**<AA>**

F1-Variance F2-Skewness F3-curtosis F4-Entropy F5-Class Color

0 4.54590 8.16740 -2.4586 -1.46210 0 Green

1 3.86600 -2.63830 1.9242 0.10645 0 Green

2 3.45660 9.52280 -4.0112 -3.59440 0 Green

3 0.32924 -4.45520 4.5718 -0.98880 0 Green

4 4.36840 9.67180 -3.9606 -3.16250 0 Green

... ... ... ... ... ... ...

1366 0.40614 1.34920 -1.4501 -0.55949 1 Red

1367 -1.38870 -4.87730 6.4774 0.34179 1 Red

1368 -3.75030 -13.45860 17.5932 -2.77710 1 Red

1369 -3.56370 -8.38270 12.3930 -1.28230 1 Red

1370 -2.54190 -0.65804 2.6842 1.19520 1 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:

<AA>

F1-Variance all Mean is 0.43

F1-Variance all STD is 2.84

F1-Variance for class 0 Mean is 2.27

F1-Variance for class 0 STD is 2.02

F1-Variance for class 1 Mean is -1.87

F1-Variance for class 1 STD is 1.88

F2-Skewness all Mean is 1.92

F2-Skewness all STD is 5.87

F2-Skewness for class 0 Mean is 4.25

F2-Skewness for class 0 STD is 5.14

F2-Skewness for class 1 Mean is -0.99

F2-Skewness for class 1 STD is 5.4

F3-curtosis all Mean is 1.4

F3-curtosis all STD is 4.31

F3-curtosis for class 0 Mean is 0.8

F3-curtosis for class 0 STD is 3.24

F3-curtosis for class 1 Mean is 2.15

F3-curtosis for class 1 STD is 5.26

F4-Entropy all Mean is -1.19

F4-Entropy all STD is 2.1

F4-Entropy for class 0 Mean is -1.15

F4-Entropy for class 0 STD is 2.13

F4-Entropy for class 1 Mean is -1.25

F4-Entropy for class 1 STD is 2.07

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

<AA>

The standard deviation of all 4 features are positive and on higher side indicate that the the data points are spread across and not closer to the mean value. The standard deviation of variance also indicate how individual data points differ largely from the mean.

The standard deviation of skewness is on the longer right tail and it is very skewed since its on the positive side away from 0 which indicates perfect symmetry.

Entropy is an indicate how the data is disordered or varied compared to other data. The standard deviation of entropy is almost same for both class 0 and 1.

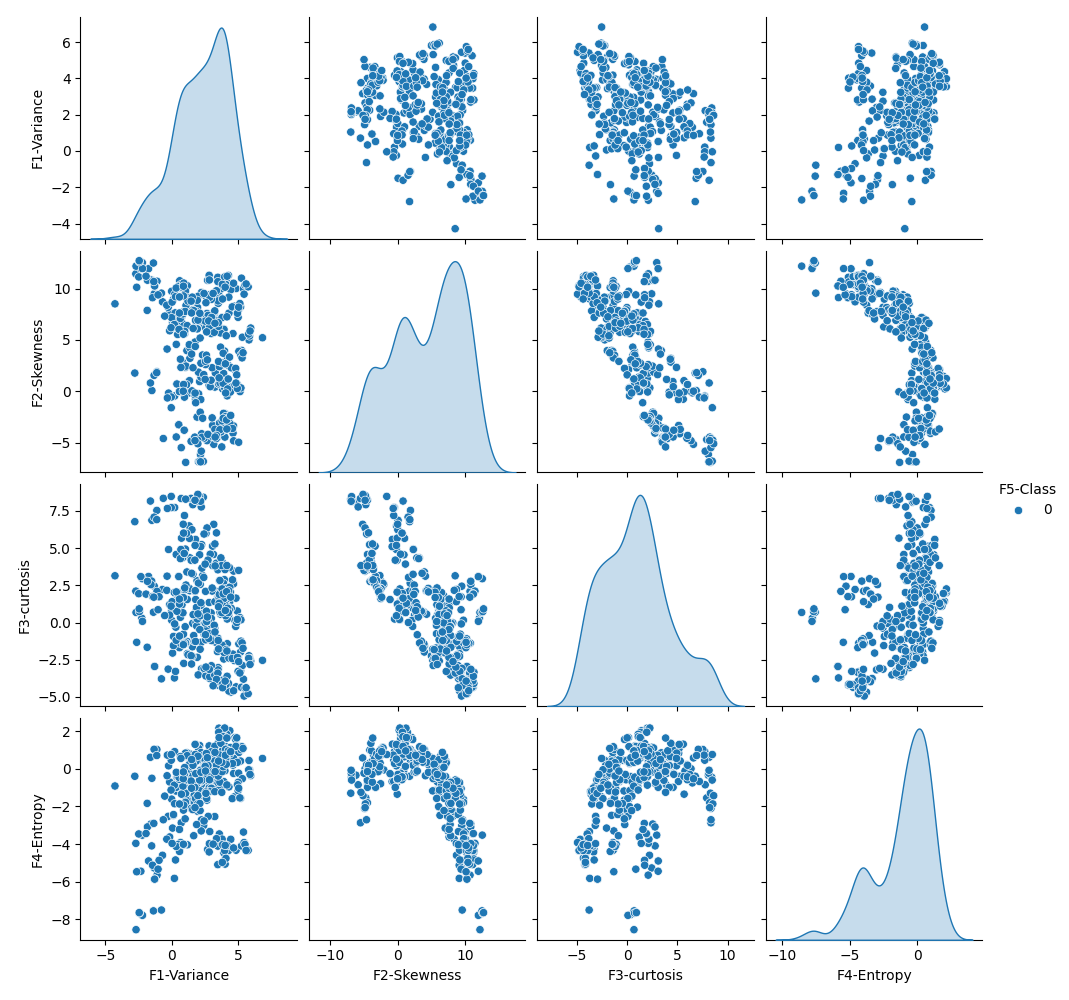
| **Class** | **µ(f1)**  **Variance** | **σ(f1)**  **Variance** | **µ(f2)**  **Skewness** | **σ(f2)**  **Skewness** | **µ(f3)**  **curtosis** | **σ(f3)**  **curtosis** | **µ(f4)**  **Entropy** | **σ(f4)**  **Entropy** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0(green)** | 2.27 | 2.02 | 4.25 | 5.14 | 0.8 | 3.24 | -1.15 | 2.13 |
| **1 (red)** | -1.87 | 1.88 | -0.99 | 5.4 | 2.15 | 5.26 | -1.25 | 2.07 |
| **all** | 0.43 | 2.84 | 1.92 | 5.87 | 1.4 | 4.31 | -1.19 | 2.1 |

**Question 2:**

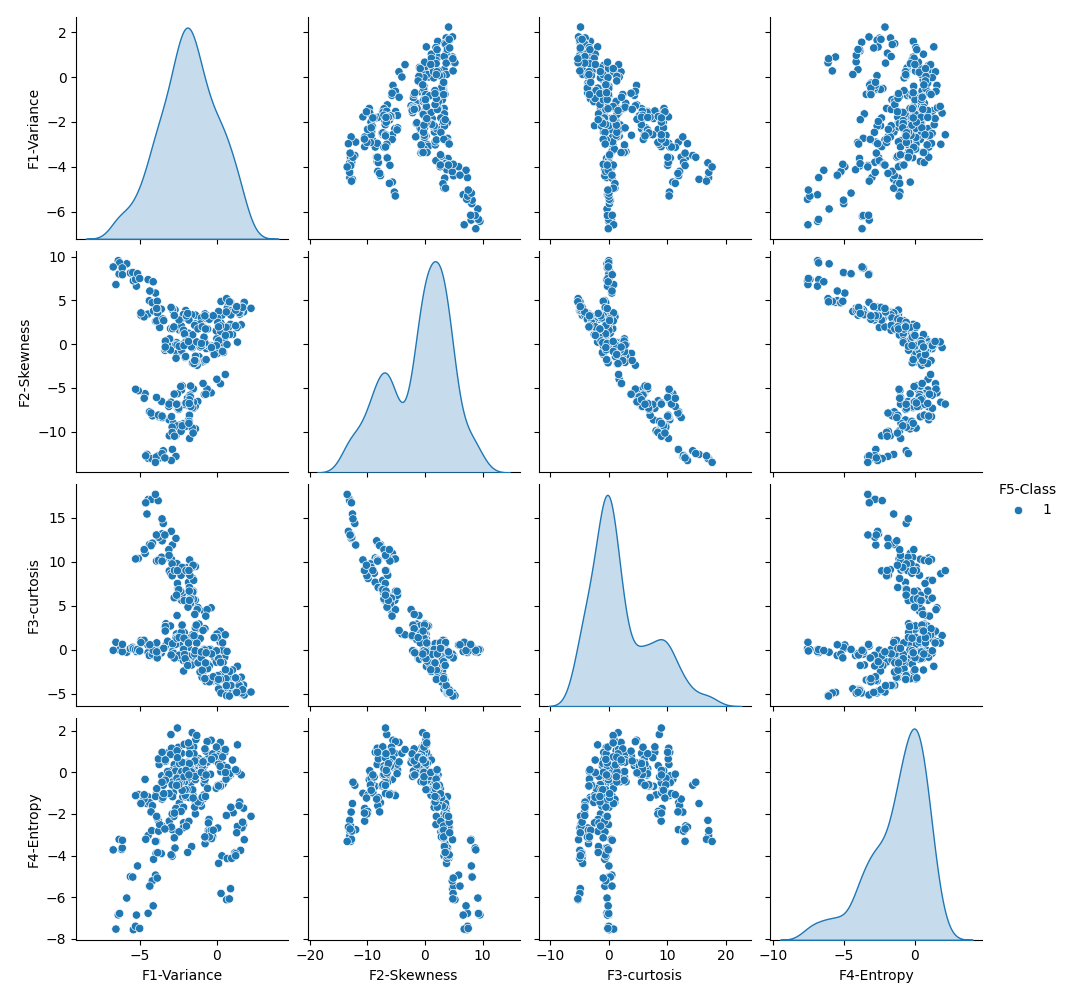
1. split your dataset X into training X train and X testing parts (50/50 split). Using ”pairplot” from seaborn package, plot pairwise relationships in Xtrain separately for class 0 and class 1. Save your results into 2 pdf files ”good bills.pdf” and ”fake bills.pdf”

<AA> Here is the scatter plot for class 0 and class 1 only for the training set.

**class0**



**class1**



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

<AA> Based on the analysis of the class labels, the following pattern is consistent with these two features. The remaining features such as entropy and skewness are evenly spread.

for i,r in X\_test.iterrows():

if r['F1-Variance'] and r['F2-Skewness'] >= 0 and (r['F3-curtosis'] <=0):

X\_test['Predicted Class'] = 0

else:

X\_test['Predicted Class'] = 1

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

<AA> Here is the results of the test data with predicted class based on the classifier from Q2.

### Test data with predictions

F1-Variance F2-Skewness F3-curtosis ... F5-Class Color Predicted Class

79 2.8033 9.0862 -3.36680 ... 0 Green 0

1212 -2.6200 -6.8555 6.21690 ... 1 Red 1

644 2.9233 6.0464 -0.11168 ... 0 Green 0

546 3.4647 -3.9172 3.97460 ... 0 Green 1

669 3.4820 -4.1634 3.50080 ... 0 Green 1

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

<AA>

Accuracy and confusion matrix for X\_test data

| TP | FP | TN | FN | accuracy | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- |
| 149 | 109 | 203 | 225 | 51.0 | 40.0 | 65.0 |

Accuracy and confusion matrix for full data set

| TP | FP | TN | FN | accuracy | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- |
| 307 | 250 | 360 | 454 | 49.0 | 40.0 | 59.0 |

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

<AA> With the manual classifier in Q2 using 3 features alone, the accuracy is 51% for X\_test data and 49% for full data set.. This indicates that with this classifier, 49% accuracy of true vs fake notes prediction.

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

<AA> No, this specific simple classifier does give a decent accuracy but only at 49% for identifying fake vs real bills.

**Question 3:**

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. Only 3 features are used to train using KNN model to keep it same with number of features used in simple classifier.

<AA> Here is the accuracy based on changes to the hyper parameter K.

Accuracy from KNN model for k=3 is: 99.56

Accuracy from KNN model for k=5 is: 99.71

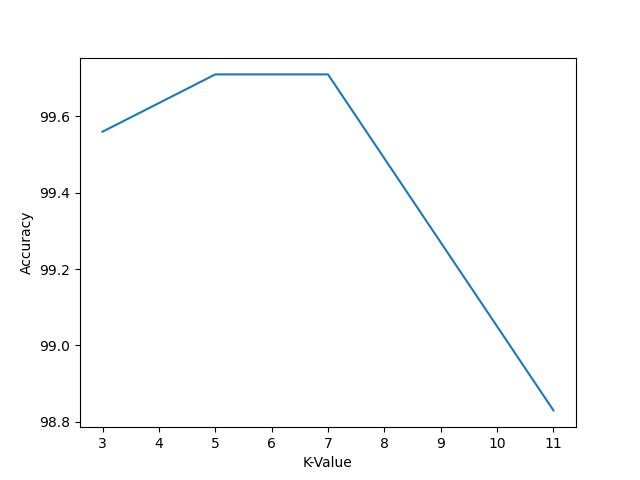
Accuracy from KNN model for k=7 is: 99.71

Accuracy from KNN model for k=9 is: 99.27

Accuracy from KNN model for k=11 is: 98.83

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?

<AA> Accuracy is best at K =5 and 7.



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

<AA> The optimal K value used here is 5.

| TP | FP | TN | FN | accuracy | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- |
| 372 | 0 | 312 | 2 | 99.71 | 99.0 | 100.0 |

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

<AA> Yes, KNN classifier is providing better result of 99% accuracy than manual classifier of 51%. Both the classifiers are using the same features.

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 ∗?

<AA> BU ID is U08370453

Prediction using simple classifier for last 4 of BUID is Red

Simple classifier:

if (row['F1-Variance'] and row['F2-Skewness'] >= 0) and (row['F3-curtosis'] < 0):

X\_test.at[index, 'Predicted Class'] = 0

else:

X\_test.at[index, 'Predicted Class'] = 1

F1-Variance = 0, F2-Skewness = 4, F3-curtosis = 5

With last 4 digits of my BU ID, the simple classifier would have marked it as fake bill (red label) i.e the predicted class is set to 1.

**KNN predictions:**

KNN prediction for K=5 for last 4 of BUID is ['Green']

**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.

<AA>

Accuracy from KNN model for k=5 after dropping F1-Variance: 94.75

Accuracy from KNN model for k=5 after dropping F2-Skewness: 97.38

Accuracy from KNN model for k=5 after dropping F3-curtosis: 97.52

Accuracy from KNN model for k=5 after dropping F4-Entropy: 99.71

Accuracy from KNN model for k=5 after dropping None: 99.85

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

<AA> No the accuracy is better with all 4 features enabled.

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

<AA> F-Variance caused the most loss of accuracy

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

<AA> F4-Entropy caused the least loss of accuracy.

**Question 5:**

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

<AA> Using same Xtrain and Xtest and with same set of features used as simple classifier and KNN classifier, here are the results from logistic regression classifier.

Accuracy from Logistic regression model is: 98.25

True Positive is 362

True Negative is 312

False Positive is 0

False Negative is 12

True Positive Rate is : 97.0

True Negative Rate is : 100.0

2. summarize your performance measures in the table

<AA>

| TP | FP | TN | FN | accuracy | TPR | TNR |
| --- | --- | --- | --- | --- | --- | --- |
| 362 | 0 | 312 | 12 | 98.25 | 97.0 | 100.0 |

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

<AA> Yes, logistic regression at 98.25% accuracy is better than simple classifier which was providing accuracy of 51%.

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

<AA> No, KNN with best K=5 value has accuracy of 99.71%, while logistic regression has little less accuracy of 98.25%.

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?

<AA> Yes. The same label is seen as KNN predictions.

Prediction using logistic regression for last 4 of BUID is ['Green']

**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.

<AA>

Accuracy from logistic regression model after dropping F1-Variance: 79.45

Accuracy from logistic regression model after dropping F2-Skewness: 89.94

Accuracy from logistic regression model after dropping F3-curtosis: 89.07

Accuracy from logistic regression model after dropping F4-Entropy: 98.25

Accuracy from logistic regression model after dropping None: 98.25

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

<AA> No, however the accuracy remained same with or without Entropy. In other words, entropy feature in both KNN and logistic regression added very little value in training the model.

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

<AA> In this case similar to KNN model, removing Variance feature caused big drop in accuracy.

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

<AA> In this case similar to KNN model, removing entropy feature caused no drop in accuracy.

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

<AA> Yes, the relative significance of features is same as KNN. In other words, in both training model, Variance feature and at most importance for accuracy and entropy had least importance to accuracy.