Acme Money Analysis and Prediction Enterprises

Executive Summary

Prepared by Raakesh Sureshkumar (1218477559)

Problem Statement:

The following details are provided regarding the counterfeit notes in 1372 samples of data. This classification of data is to be used for further analysis and to perform the statistical analysis on the data. The degree of dependency of each variable on the others and the variable to be predicted (Counterfeit / Genuine) is to be determined using the statistical tools correlation and cross - covariance. Create a model with the best accuracy for determining the counterfeit notes.

The following data are given regarding the counterfeit bill:

1. Variance of wavelength transformed image (continuous)
2. Skewness of wavelength transformed image (continuous)
3. Curtosis of wavelength transformed image (continuous)
4. Entropy of image (continuous)
5. Class (integer)

Objective:

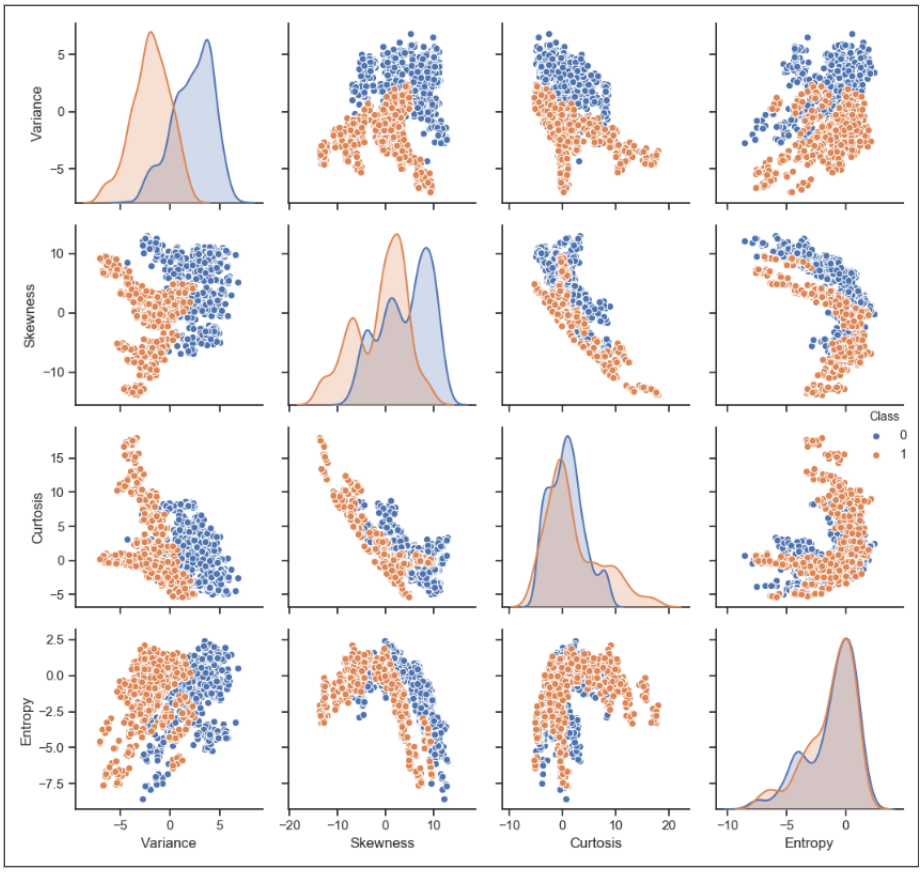
Create a machine learning model that learns based on the given information on the bills with their classification information. This model will be used for developing an app for the US Treasury to aid in the detection of counterfeit bills. Also, perform an analysis to choose the variables of significance from the 5 entities listed above that will have a profound impact in the design of the machine learning model.

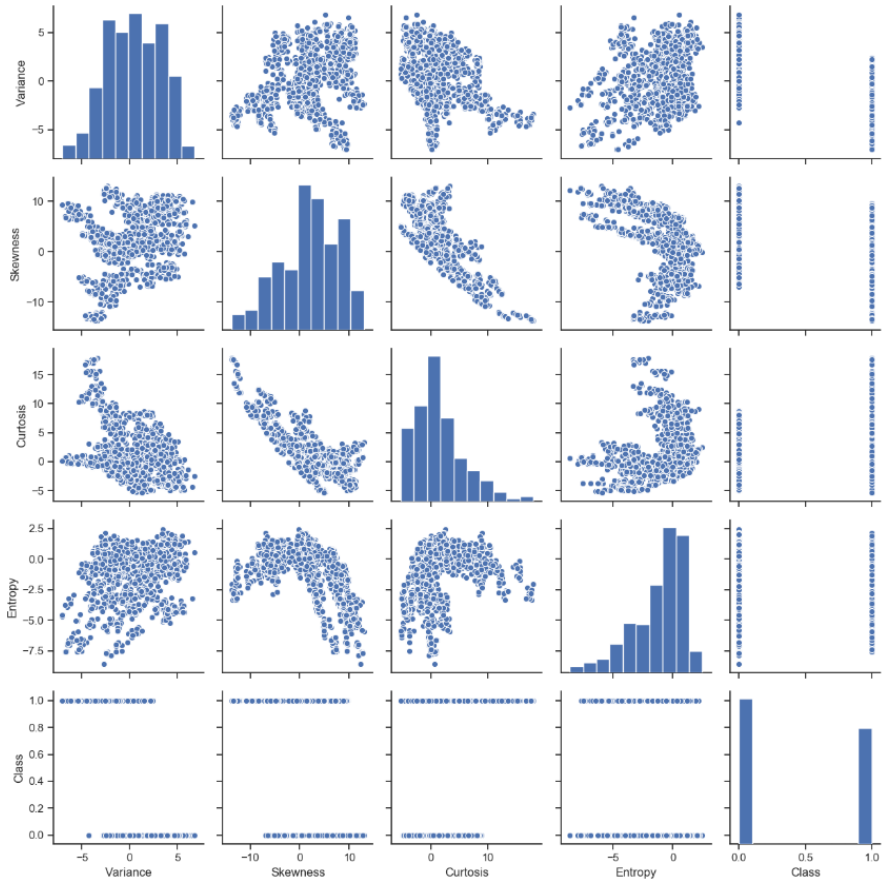
Problem 1:

Observations (Output of “data_analysis.py”)**:**

```
Database size:1372

Table 1: Initial Correlation Matrix - Table 3

          Variance  Skewness  Curtosis   Entropy     Class
Variance  1.000000  0.264026 -0.380850  0.276817 -0.724843
Skewness  0.264026  1.000000 -0.786895 -0.526321 -0.444688
Curtosis -0.380850 -0.786895  1.000000  0.318841  0.155883
Entropy   0.276817 -0.526321  0.318841  1.000000 -0.023424
Class    -0.724843 -0.444688  0.155883 -0.023424  1.000000

Table 2: Covariance Matrix

          Variance   Skewness   Curtosis   Entropy     Class
Variance  8.081299   4.405083  -4.666323  1.653338 -1.024310
Skewness  4.405083  34.445710 -19.905119 -6.490033 -1.297386
Curtosis -4.666323 -19.905119  18.576359  2.887241  0.333985
Entropy   1.653338  -6.490033   2.887241  4.414256 -0.024464
Class    -1.024310  -1.297386   0.333985 -0.024464  0.247112

Table 3: Most Highly Correlated

  FirstVariable SecondVariable  Correlation
0      Skewness       Curtosis    -0.786895
1      Variance          Class    -0.724843
2      Skewness        Entropy    -0.526321
3      Skewness          Class    -0.444688
4      Variance       Curtosis    -0.380850

Table 4: Correlation (with variable to be predicted - 'Class')

   Variable  Correlation with class
0  variance                   -0.72
1  skewness                   -0.44
2  curtosis                    0.16
3   entropy                   -0.02

Table 5: Highest Correlation Matrix (Element-wise distribution)

  Variable 1 Variable 2  Correlation
0   variance      class        -0.72
1   skewness   curtosis        -0.79
2   curtosis   skewness        -0.79
3    entropy   skewness        -0.53
4      class   variance        -0.72
```

Pair plots:


Variance
Skewness
Curtosis
Entropy
Class
0
1


Variance
Skewness
Curtosis
Entropy
Class

Inferences:

1. Table 1 and Table 2 give the correlation and the cross co - variance between all the variables within the dataset.
2. From Table 3, the top 5 correlation values and their respective variables are taken into consideration. Here, the absolute values of the correlation represent the highest degree of correlation between the 2 variables within the data sample. So, skewness is considered to be one of the best parameters with good correlation with curtosis, entropy and class.
3. In Table 5, the highest correlation for each variable is taken into consideration (not the overall high among all the relation as in Table 3). Skewness and Curtosis seem to go hand in hand with one another with a correlation magnitude of 0.79. Variance and Class follow with a correlation of 0.72.
4. Table 4, represents the correlation of the individual variables with the variable to be predicted, that is, the Class variable. The class variable seems to be in good correlation with the variables Variance, Skewness, Curtosis and Entropy in the order of their ranking.
5. From Pairplot1, the correlation can be found by the scatter plots on against each and every variable and the distribution of the scatter. The more distributed it is or the more scattered it is, the lower is the correlation. From the histogram, we can infer the normal distribution of the variables across its mean. This can help us to decide if the data has to be normalized before applying the fit.
6. In Pairplot 2, the hue has been set to 'Class' variable. This provides much better distinction with the correlation of the variables.

Result:

The variables Variance seems to be in the best correlation with the variable to be predicted (Class). The variable pair Skewness and Curtosis exhibit the highest correlation. So, these variables (Variance, Skewness, Curtosis) are essentially the best variables to be considered for training the machine learning model.

Problem 2:

Test description:

The training and the test dataset are split as follows:
Training = 70% of samples
Test data = 30% of samples
Features considered: Variance, Skewness, Curtosis, Entropy
The train and test data are normalized before performing the transform, fit and prediction based upon the distribution of data as seem from problem 1. If the data has a standard normal distribution a standard scaler can be used to normalize the data to mean = 0 and standard deviation = 1
Classifier description:

1. Perceptron: iterations = 20, tolerance = 10^-3, Standard scaler = True
2. Logistic Regression: C = 10, solver = 'Library of Large Linear Classifiers', multi – class = 'One vs Rest Classifier', Standard scaler = True
3. Support Vector Classification: kernel = 'linear' , C = 1.0, Standard scaler = True
4. Decision Tree Classifier: criterion = 'entropy', max tree depth = 5, Standard scaler = True
5. Random Forest Classifier: criterion = 'entropy' , number of estimators = 10, number of jobs = 2, Standard scaler = True
6. K Nearest Neighbors: number of neighbors = 10, p = 2, metric = 'minkowski', Standard scaler = False

The above test cases are implemented in the python program "prediction.py"

Observations (Output of "prediction.py")**:**

```
Accuracy of Prediction

                    Combined Acc (in %)  Combined Acc(in #Samples)
Perceptron                        98.25                    1348.0
Logistic Regression               98.40                    1350.0
Support Vector                    98.47                    1351.0
Decision Tree                     99.27                    1362.0
Random Forest                     99.71                    1368.0
K Nearest Neighbor               100.00                    1372.0

                    Test Acc(in %)  Test Acc (in #Samples)
Perceptron                   98.06                   404.0
Logistic Regression          98.30                   405.0
Support Vector               98.54                   406.0
Decision Tree                98.54                   406.0
Random Forest                99.03                   408.0
K Nearest Neighbor          100.00                   412.0
```

Inferences:

The above table gives the accuracy on the combined dataset and the test dataset in terms of percentage and count of correct decisions made by the system based upon the model created using the training dataset.

This brings to the following inequality in terms of accuracy of the model (tested on both the test data and the combined data:

K Nearest Neighbor > Random Forest > Decision Tree > Support Vector Classifier > Logistic Regression > Perceptron

Result:

From the above inferences we can conclude that K Nearest Neighbor is the best method of predicting whether a bill is counterfeit with the current combination of input parameters to the classifier function.