100/100

Management Science Project 3 - Group 5

I. Data:

This project involves the process of classification on multivariate data concerning authentication procedures for banknotes. The data was extracted from images taken, intended to assist in classifying determining if for digitization, an industrial camera can be used for print inspection of a banknote. To be clear, the final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tools were used to extract features from images. The four predictor/independent (x) variables include variance of Wavelet Transformed image (continuous), skewness of Wavelet Transformed image (continuous), curtosis of Wavelet Transformed image (continuous), and entropy of image (continuous), which are used for classification to determine the dependent (y) binary variable of class, so it may be an interesting data set for testing feature selection methods.

This project has a real world application, since millions of people, daily, use banknotes to make supposedly secure transactions. The security of these banknotes is in question, since the government and banks prioritize combating fraud. The issue now is that it is extremely tough to spot counterfeit banknotes and distinguish them from genuine notes. So, the overall goal is to develop a support system so organizations can classify fraudulent banknotes, accurately. There are 1372 instances (rows), and 5 variables (columns) in the data set. For the classification process, the instances are divided into training, selection and testing subsets. So, there are 824 instances for training (60%), 274 instances for selection (20%) and 274 instances for testing (20%).

We would want to indicate which factors better discriminate between authentic and false banknotes. So, after we do, we find that wavelet transformed variance is the most influential variable for the application, followed by wavelet transformed skewness, wavelet transformed curtosis, and image entropy. We would also want to reduce selection error. Some things to consider for this dataset include: class distribution, cost of misclassification, and size of training and test sets. Just to clarify, there were no noisy variables added. With our data set, given our sampling procedure, and the use of an induction algorithm, nmin is the size of the smallest viable training set. Additionally, models that have smaller training sets actually have lower accuracy than models that are formed from training sets of size nmin. Also, models that involve larger training sets, have no higher accuracy. In terms of our sampling method, we first normalized our features. Then, we separated out 20% of the cases of the test data set for our testing set, and the remaining 80% for our training set. Our coding analysis was done in Rstudio, using the R programming language.

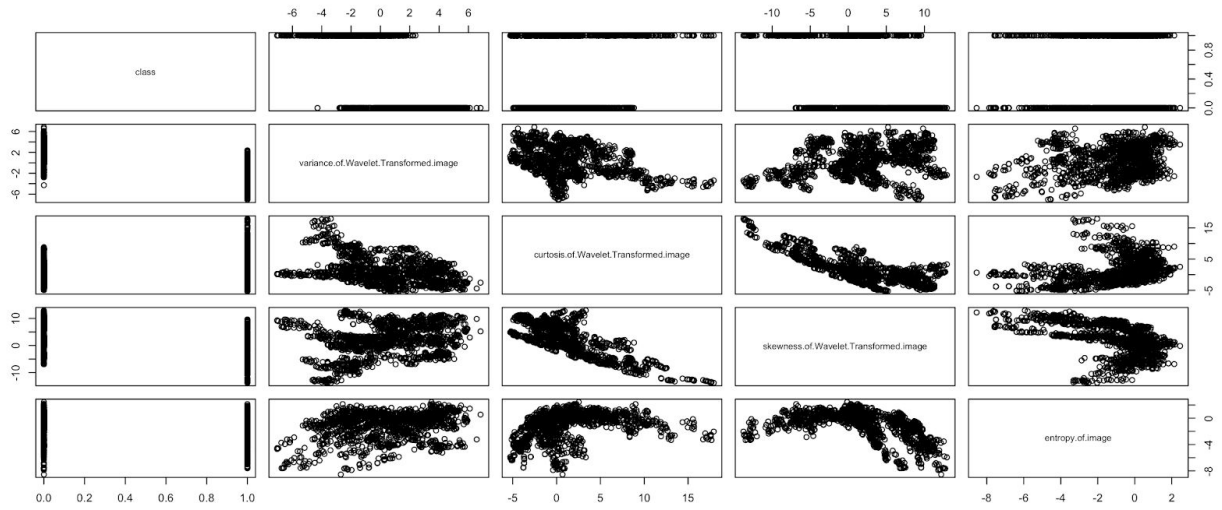
Our attributes are:

1. Variance of Wavelet Transformed image (continuous) -V1
2. Skewness of Wavelet Transformed image (continuous) -V2
3. Curtosis of Wavelet Transformed image (continuous) -V3

4. Entropy of image (continuous) -V4
5. Class (integer). It can only have two values: 0 (false) or 1 (true).

We used entropy of the categorical dataset, where entropy is the measure of disorder or of uncertainty within the machine learning classification model. Formula:
$I(E)=-\log_2(p(E))=\log_2(1/p(E))\}$

Here are the scatterplots of our data, Ys against X:

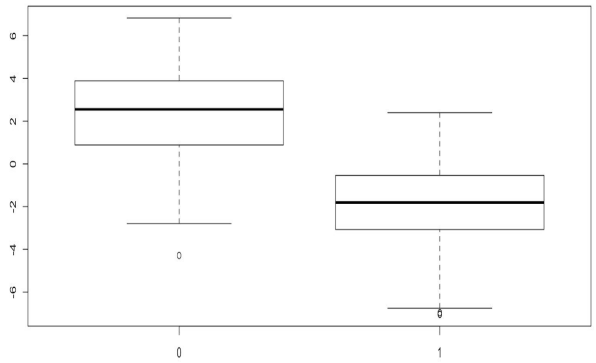


II. Box Plot:

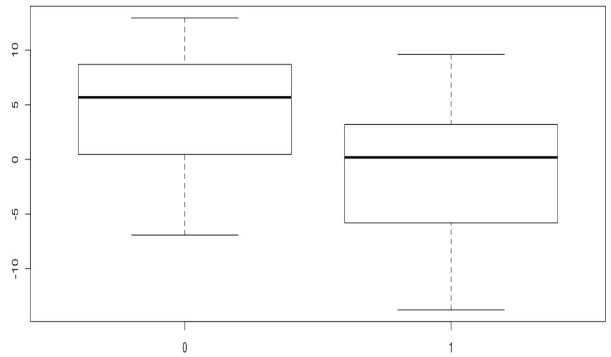
We use a box plot to show the approximate distribution of a feature. A boxplot quickly highlights several key quantiles -- min, 25th, median, 75th, and max -- of a dataset, which might indicate hints about skew or outliers in our data.

Since the class in our data set is a binary feature, a simple, overall box plot is not an effective representation of the distribution. By simply looking at the data, we know that 44.6% of the data belongs to the positive class. However, a box plot is more applicable to each of the independent factors, because they are continuous and not binary.

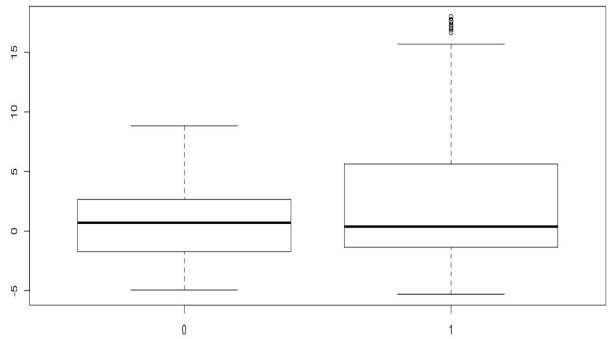
Below are the boxplots for each of the features V1 - V4:



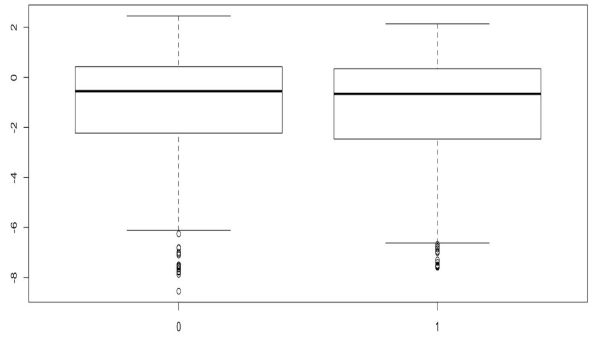
V1



V2



V3



V4

What we can see from these individual box plots is that for V1 (Variance of Wavelet Transformed Image) and V2 (Skewness of Wavelet Transformed Image), the data is relatively well behaved and all fall in a “normal” fashion, making it normally distributed. The V3 feature (Curtosis of Wavelet Transformed Image) shows that there are quite a few outlying data points at the upper end, demonstrated by the individual data points shown above. On the other hand, the boxplot for the V4 feature (Entropy of Image) shows that there are some outliers in the lower end, showing that there is a negative skew in the data of this feature.

III. Classification Analysis:

Next, we used K-Nearest Neighbor and Decision Tree algorithms to build a classification model using our training data set.

K-Nearest Neighbor Algorithm:

The K-nearest neighbor algorithm is a non-parametric supervised method of regression and classification to determine the k-closest training examples. Essentially, a point in a graph is assumed to be a part of a class if there is commonality with k nearest neighbors. It uses Minkowski distance, a variation of the Euclidean distance formula, which is a metric used to determine distance, and similarity of two graphed points.

Here's what we found from our KNN analysis:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	157
1	1.00	1.00	1.00	118
accuracy			1.00	275
macro avg	1.00	1.00	1.00	275
weighted avg	1.00	1.00	1.00	275

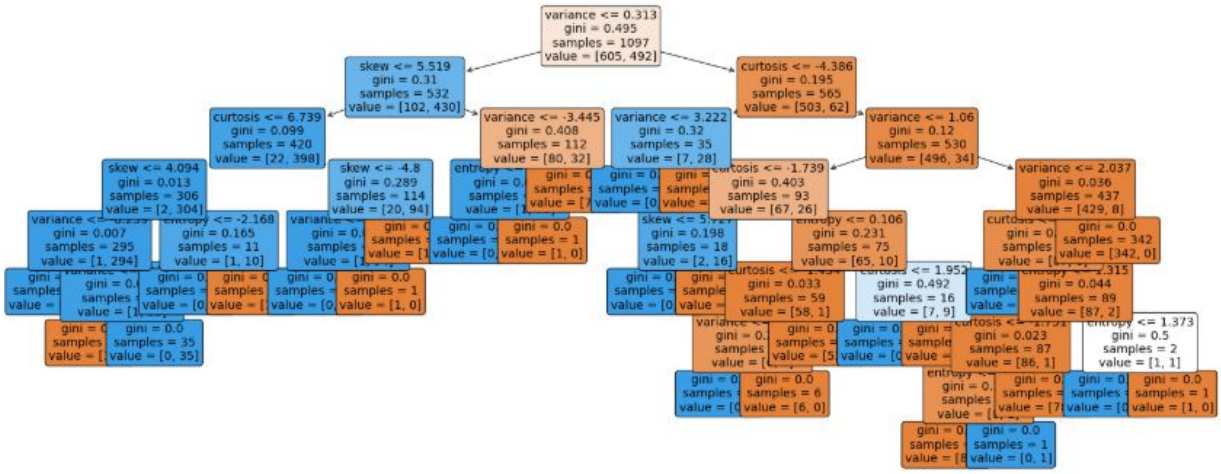
1.0

- Values of TP, TN, FP, and FN:
    - Accuracy: (TP+TN) / (TP +TN+FN+FP) = 1.0
    - Sensitivity/Recall: (TP)/ (TP+FN) = 1.0
    - Specificity (TN) /(TN+FP) = 1.0
    - Precision (TP)/(TP+FP) = 1.0
    - Error rate: (FN+FP) / (TP +TN+FN+FP) = 0.0

We didn't run into either a type I error (We claim it is a true positive, but it is actually a false positive) or a Type 2 error (We claim it is a true negative, but it is actually a false negative). This means there is no risk of experiencing high cost due to an inauthentic banknote, or the error of flagging an authentic banknote as inauthentic.

III. Decision Tree:

The decision tree algorithm is a supervised learning algorithm used for solving a classification problem, using a series of nodes, where the internal node corresponds to an attribute, and the leaf nodes represent class labels.



Here's what we found from our decision tree analysis:

	precision	recall	f1-score	support
0	0.99	0.98	0.99	157
1	0.97	0.99	0.98	118
accuracy			0.99	275
macro avg	0.98	0.99	0.99	275
weighted avg	0.99	0.99	0.99	275

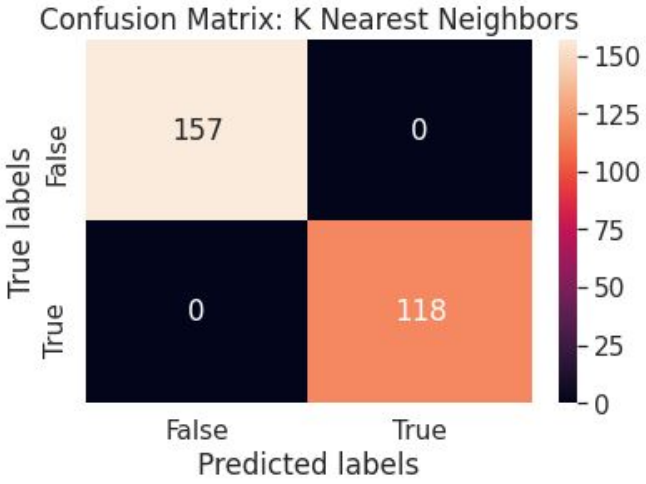
```
0.9854545454545455
```

Essentially, our results are a 99% precision rate, a 99% recall rate, and a 99% f1-score.

V. Confusion Matrix:

Cost Matrices are used for improving the classification knn and decision tree algorithms by getting a better model, especially when there is a skew, and to balance the class categorization distribution. In the confusion matrix, the rows represent target classes and the columns the output classes for the testing target data set. The diagonal cells in each table show the number of cases that were correctly classified, and the off-diagonal cells show the misclassified cases. The following table contains the elements of the confusion matrix:

KNN Confusion Matrix:




```
array([[157,   0],
       [  0, 118]])

Specificity :  1.0
Sensitivity :  1.0
```

Decision Tree Confusion Matrix:


```
array([[154,   3],
       [  1, 117]])

Sensitivity :  0.9745222929936306
Specificity :  0.9745762711864406
```

KNN vs Decision Tree:

Our KNN classifier confusion matrix shows that there are no false positive or false negative classifications as a result of using this algorithm. Comparatively, there are four false positive results and three false negative results when using the decision tree algorithm, where the frequency of correctly classified instances is 268 and the number of misclassified instances is seven. This gives an overall accuracy rate of 98.54%. What this shows is that due to there not being any misclassified patterns for the KNN classifier, the model evidently predicts the testing data better than the decision tree classifier. In conclusion, for digitization, using this model, an industrial camera can effectively complete print inspection of a banknote. For future tests, however, we should consider completing an ROC curve, which is a good measure for the precision of a binary classification model.

We *Mark Holtje, Ian George, Graciela Casanova, Apoorva Nori, Yash Tekwani* did not give or receive any assistance on this project."

Sources:

Owner of database: Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe, volker.lohweg '@' hs-owl.de)

Donor of database: Helene DÃ¶rksen (University of Applied Sciences, Ostwestfalen-Lippe, helene.doerksen '@' hs-owl.de)