1. Our dataset has four attributes to determine if a banknote is fake or real. These attributes are: the variance of Wavelet Transformed image (continuous), the skewness of Wavelet Transformed image (continuous), curtosis of Wavelet Transformed image (continuous) and the entropy of image (continuous).

These attributes together tell us if a banknote is real or fake which in this dataset is encoded as 1 and 0 respectively.

Our dataset has 4 attributes with 1372 instances and the attributes type are real numbers.

We are going to create several algorithms to learn from this dataset so that it can predict if a banknote is real or fake based on the 4 given attributes.

1. Our dataset is in a text file and we converted it to a csv file for our decision tree algorithm and used the text file directly for our logistic regression and Naïve Bayes algorithm. The value of the dataset didn’t have a range that was too big so there was no reason to normalize or standardize the data.

For the logistic regression algorithm we took the data in as a text file and made in into a 2d numpy array and the performed all our calculations and predations using numpy arrays.

For the decision tree algorithm we took the data in as a csv file and …

For the Naive Bayes algorithm we took the data in …

Like the notes said we split our data into 60%, 20%, and 20% as training, validation and testing data respectively.

1. We used 3 different classification algorithms: Decision trees, Naïve Bayes and Logistic Regression.

Logistic Regression: We chose this algorithm because our dataset has a binary outcome and the logistic regression formula gives a probabilistic output that is perfect for datasets that requires binary outputs. We did not regularize the algorithm because it was running fine without regularization and the weights are not getting too big so we felt regularization was not needed.

The basis function we used was: y = x0a0 + x1a1 + x2a2 + x3a3 + x4a4 where x0 is 1 and is our bias.

We trained the model using gradient decent method with a learning rate of 1e-6 and a tolerance of 0.001. After checking multiple values of the tolerance and learning rate we found those values gave us the highest accuracy.

Decision Trees:

Logistic Regression:

1. Logistic Regression: as said in the previous question a learning rate of 1e-6 and a tolerance of 0.001 worked best. Using this algorithm and this dataset the best possible performance we achieved was an accuracy in the mid 50%. We did that by training the model with over half of the dataset and further setting the parameters with the validation set.

I would recommend they try using a different algorithm like neural networks as this dataset is fun to work with and experimenting with an algorithm we didn’t use would be interesting