After the analysis of the decision tree algorithm results, the obscured2 data set is believed to be the real file whose class labels was not modified.

First way to determine the signal is comparing two accuracy with the percentage of majority class in the data set. Even though one of the data set was randomly shuffled. The records with the class label C take up 33950/38160 = 88.97% in both data sets. This makes C the majority class in the data. However, when using J48 decision tree algorithm to classify the data, obscure1 gets an overall accuracy of 88.9675% and obscured gets a accuracy of 92.2327%. Comparing these two results with the original proportion of the majority class, the first result from obscured1 is identical while the obscured2’s result is higher than the majority proportion. This underlines that there is some deeper correlation between attributes than just looking at majority class in the obscured2 data set.

Another method used to determine the real data set is using the decision trees generated by the algorithm. First decision tree has only one node which means it will classify all records to the label C. However, the second decision generated from obscured2 data set has a height of 4 and overall 1623 leaves, which implies there are additional information provided by other attributes. This can also be checked with direct visualization of the attribute data.