# Statistical analysis

* The 1st data set has 756 records, 754 attributes and 1 class.
* The attributes have very different ranges.

# Preprocessing

**Scaling**:

* We will suspect that scaling will improve classification results in the classifiers that may be affected by scaling, due to the different scales of the features in the dataset.
* NB: should not be very affected because the model’s priors determined by the count in each class and not by the actual value.
* **Distance-based** such as KNN methods are affected by scaling.
* **Tree-based** models are not distance-based and so can handle varying ranges of features. Thus, scaling should not affect much the quality of the trees model.
* Scaling should be important while performing PCA, because PCA tries to get the features with maximum variance and the variance is high for high magnitude features. This skews the PCA towards high magnitude features.
* In the Cover Type dataset, we reduced the records per category to 1000, by taking random samples from the original dataset

# Classification

1. **Naive Bayes**

* Multinomial and Bernoulli distributions are discrete probabilities distributions and are appropriate for counts. Since the 1st dataset has mostly real values, the MultinomialNB and BernoulliNB are not appropriate for this kind of classification, so we only use the GaussianNB.
* We did not change any parameters in GaussianNB.

1. **KNN**

* In KNN, we changed the 2 following parameters:
  + Number of neighbors
  + Distance calculation method
* The first dataset has approximately as many features as records. And so in KNN, it makes sense that the best n is 1, because it is very unlikely to have 2 records similar to a test instance, if the number of features is not reduced.
* In the second dataset, we have

1. **Decision Trees**

* In decision trees we changed the 3 following parameters:
  + Criteria: gini or entropy
  + Maximum tree depth
  + Minimum number of samples to be a leaf node
* In the 2nd dataset, as we have 7000 records, it makes sense that the classification with decision trees will improve if the minimum samples leaf parameter is low. The depth doesn’t influence much the accuracy of the model.

1. **Random Forests**

* In random forests we changed the 3 following parameters:
  + Number of trees in the forest
  + Maximum tree depth
  + Number of features to consider for the best split

1. **Gradient Boosting**

* In gradient boosting we changed the 4 following parameters:
  + Learning rate
  + Number of trees in the forest
  + Maximum tree depth
  + Number of features to consider for the best split

1. **XGBoost**

* In xgboost we changed the 2 following parameters:
  + Number of trees in the forest
  + Maximum tree depth

# Unsupervised

# Compare results