# 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.
* NB: should not 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 is not required while modelling trees.
* 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.

# Classification

1. **Naive Bayes**

Multinomial and Bernoulli distributions are discrete probabilities distributions and since the 1st dataset has mostly real values, the MultinomialNB and BernoulliNB are not appropriate for this kind of classification.

1. **KNN**

The first dataset has few examples 750 attributes +- for 750 examples. 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.

Best se

1. **Decision Trees**
2. **Random Forests**
3. **Gradient Boosting**
4. **XGBoost**

# Unsupervised

# Compare results