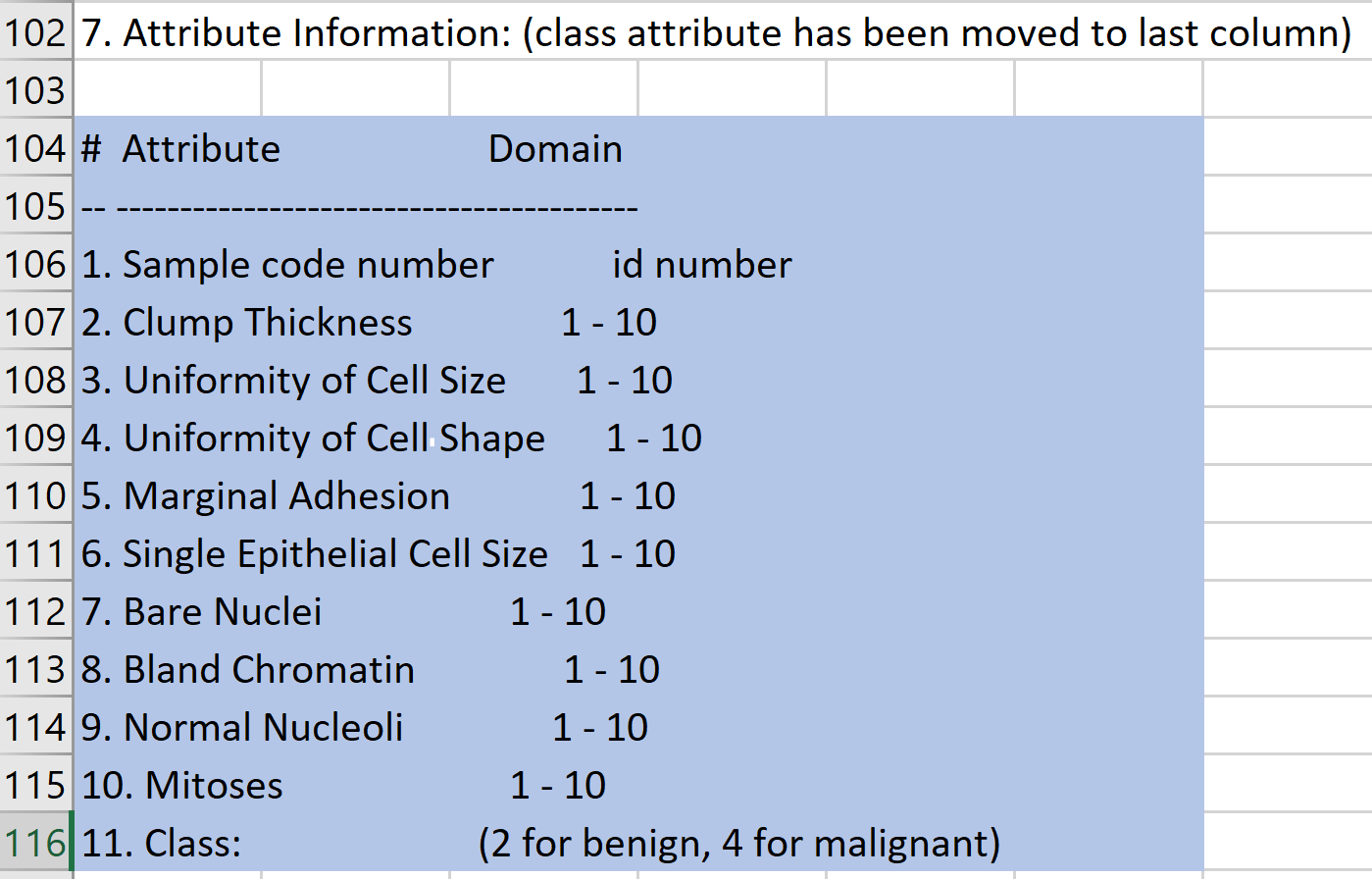
Classification Exercises

Datasets are attached as either XLS or CSV files.

1. Breast cancer data: Use the dataset “Breast Cancer Wisconsin.xls” to build a model to diagnose breast cancer.

The Description below provides a description of the data. Row 104+ contains description of the columns.



Build a classification model using various methods below and compare your models.

* + 1. KNN .. Try a few K values. Report your best K
    2. Logistic Regression
    3. Decision Tree.. Try a few “Max Number of Splits”. Report your best choice.
    4. SVM.
    5. Ensemble of your choice: Boosted Tree and Bagged Tree.

1. Handwritten digits recognition

MNIST in the attached csv’s is a version of the MNIST handwritten digits data. Each image is represented by 28x28 greyscale array that has been transformed to a 784 element vector. The first column represents the labeled classification of 0-9 for each image.

The data have already been split into Train and Test. Use the data to train a classification model to recognize the handwritten digits. Choose your best model, report performance metrics and confusion matrix.

Note: Later you can use the same data and build a Neural Net model to compare.

1. Telco Customer Churn: A telecom company wants to predict whether a customer would churn (switch to another telecom supplier). The customer data base (telco-customer-churn.xls) contains 19 potential features (first column is customer ID). The last column (U) is the binary target of Yes (Churn) or No (not churn). There are about 7000 instances.

* Split the data set and build binary classifiers using various methods of your choice.
* Use at least one of the Feature Selection methods we mentioned to select your features as part of the model building process.

Compare algorithms and choose your best model. Report the performance metrics and confusion matrix.