Take ALL_AML_original_data.zip file from Data and extract from it
Train file: data_set_ALL_AML_train.txt
Test file: data_set_ALL_AML_independent.txt
Sample and class data: table_ALL_AML_samples.txt
This data comes from pioneering work by Todd Golub et al at MIT Whitehead Institute (now MIT Broad Institute).

1. Rename the train file to ALL_AML_grow.train.orig.txt and test file to ALL_AML_grow.test.orig.txt .

Both train and test datasets are tab-delimited files with 7130 records.
The "train" file should have 78 fields and "test" 70 fields. The first two fields are
Gene Description (a long description like GB DEF = PDGFRalpha protein) and
Gene Accession Number (a short name like X95095_at)

The remaining fields are pairs of a sample number (e.g. 1,2,..38) and an Affymetix "call" (P is gene is present, A if absent, M if marginal).

Think of the training data as a very tall and narrow table with 7130 rows and 78 columns. Note that it is "sideways" from machine learning point of view. That is the attributes (genes) are in rows, and observations (samples) are in columns. This is the standard format for microarray data, but to use with machine learning algorithms like WEKA, we will need to do "matrix transpose" (flip) the matrix to make files with genes in columns and samples in rows.

Here is a small extract
Gene Description Gene Accession Number 1 call 2 call ...
GB DEF = GABAa receptor alpha-3 subunit A28102_at 151 A 263 P ...
... AB000114_at 72 A 21 A ...
... AB000115_at 281 A 250 P ...
... AB000220_at 36 A 43 A ...
Clean the data
Perform the following cleaning steps on both the train and test sets.
Document all the steps and create intermediate files for each step. After each step, report the number of fields and records in train and test files.
Microarray Data Cleaning Steps

1. Remove the initial records with Gene Description containing "control)".
(Those are Affymetrix controls, not human genes). Call the resulting files ALL_AML_grow.train.noaffy.tmp and ALL_AML_grow.test.noaffy.tmp. How many such control records are in each file?

2. Remove the first field (long description) and the "call" fields, i.e. keep fields numbered 2,3,5,7,9,...

3. Replace all tabs with commas

4. Change "Gene Accession Number" to "ID" in the first record.

5. Normalize the data: for each value, set the minimum field value to 20 and the maximum to 16,000. (Note: The expression values less than 20 or over 16,000 were considered by biologists unreliable for this experiment.)
6.Write a program to transpose the training data to get ALL_AML_gcol.test.tmp and ALL_AML_gcol.train.tmp ("gcol" stands for genes in columns). These files should each have 7071 fields, and 39 records in "train", 35 records in "test" datasets.

7. Extract from file table_ALL_AML_samples.txt tables ALL_AML_idclass.train.txt and ALL_AML_idclass.test.txt with sample id and sample labels, space separated.
Add a header row with "ID Class" to each of the files.
File ALL_AML_idclass.train.txt should have 39 records and two columns. First record (header) has "ID Class", next 27 records have class "ALL" and last 11 records have class "AML". Be sure to remove all spaces and tabs from this file.
ALL_AML_idclass.test.txt should have 20 "ALL" samples and 14 "AML" samples, intermixed.

8. Note that the sample numbers in ALL_AML_gcol*.csv files are in different order than in *idclass files. create combined files ALL_AML_gcol_class.train.csv and ALL_AML_gcol_class.test.csv which have ID as the first field, Class as the last field, and gene expression fields in between.

9.What accuracy and error rate do you get?
Now, excluding the field ID, build models using Decision tree, NaiveBayes Simple, and k-means, using training set only.