# Appendix A

# Datasets

This appendix contains a summary of which datasets were used in various experiments for this thesis. The datasets described here can be divided into two categories: genomic datasets on one side and all other datasets on the other side. From this point, non-genomic datasets will be referred to as "standard" datasets.

## A.1 Standard datasets

| Name | Nb. Instances | Nb. Features | Categorical Feat. | Continuous Feat. | Class priors |
|---|---|---|---|---|---|
| Iris | 150 | 4 | 0 | 4 | 50/50/50 |
| Tic-tac-toe | 958 | 9 | 9 | 0 | 332/626 |
| Heart | 270 | 13 | 6 | 7 | 150/120 |
| Glass | 214 | 9 | 0 | 9 | 70/76/17/13/9/29 |
| Ionosphere | 351 | 34 | 2 | 32 | 126/225 |

## A.2 Genomic datasets

| Name | Nb. Instances | Nb. Features | Categorical Feat. | Continuous Feat. | Class priors |
|---|---|---|---|---|---|
| Alon | 62 | 2000 | 0 | 2000 | 22/40 |
| Lymphoma (lymph) | 45 | 4026 | 0 | 4026 | 23/22 |
| Golub | 72 | 7129 | 0 | 7129 | 47/25 |
| West | 49 | 7129 | 0 | 7129 | 25/24 |

**Note on the Lymphoma Dataset** The Lymphoma dataset used in the experiments is a modified version and not the original. Indeed, the original dataset contains missing values. Since the RIT algorithms proposed in this thesis do not deal with missing values, the dataset was modified by imputing missing values (cfr `rfImpute` function in the `randomForest` package).

## A.3 Sources

The Lymphoma dataset can be found at `http://eps.upo.es/aguilar/datasets.html`. The other genomic datasets can be found in the following Github repository: `https://github.com/ramhiser/datamicroarray`.

The standard datasets can be found on the UCI machine learning repository at `http://archive.ics.uci.edu/ml`.