**Molecular Biomarkers in Clinical Research**

**Assignment for lecture on “Bioinformatics and Biomarker Discovery”, 5/9/2012**

The objective of this assignment is to familiarize you with one of the commonly used tools in data analysis, viz. WEKA. The WEKA machine learning package is used in this assignment. Please download and install it from <http://www.cs.waikato.ac.nz/ml/weka/>.

Consider the sample data set in the table given below. The table is available as the file “confounding-factor.csv” in the datasets folder on the lecture webpage <http://www.comp.nus.edu.sg/~wongls/courses/mci5004/mci12/>.

|  |  |  |  |
| --- | --- | --- | --- |
| **sample\_id** | **sex** | **param** | **class** |
| **f1** | **f** | **60** | **dead** |
| **f2** | **f** | **62** | **dead** |
| **f3** | **f** | **63** | **dead** |
| **f4** | **f** | **64** | **dead** |
| **f5** | **f** | **65** | **dead** |
| **m1** | **m** | **71** | **dead** |
| **f6** | **f** | **71** | **alive** |
| **f7** | **f** | **72** | **alive** |
| **m2** | **m** | **73** | **dead** |
| **f8** | **f** | **73** | **alive** |
| **m3** | **m** | **75** | **dead** |
| **f8** | **f** | **75** | **alive** |
| **f9** | **f** | **76** | **alive** |
| **m4** | **m** | **80** | **dead** |
| **m5** | **m** | **85** | **dead** |
| **m6** | **m** | **90** | **alive** |
| **m7** | **m** | **95** | **alive** |
| **m8** | **m** | **98** | **alive** |
| **m9** | **m** | **99** | **alive** |
| **m10** | **m** | **100** | **alive** |

**Q1. Run the J48 decision tree classifier in WEKA on this data set, show the decision tree produced? [5 marks]**

**Q2. Suggest a more natural decision tree. [5 marks]**

**Q3. Explain why this is a more natural decision tree. [5 marks]**

**Q4. What feature of WEKA did you use to arrive at this more natural decision tree? If you did not use WEKA to obtain this alternative decision tree, how did you go about getting it? [5 marks]**