Data Mining and Decision Systems

08338

Assessed Coursework

Data Mining of Legacy Data

Stage 1. Data Description

Student Number: 201303310

PDF Due: 2pm 2 November 2015 via E-Bridge

Sunday, 29 November 2015

# Data Description

The given data (**BaseData-All**) is described in Table 1 below, based on the worksheet **GivenDataDescription**.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | **Classifier Type** | **Value Type** | **NumberOfValues** | **Values** | **Missing** | **Unknown** | **Null** |
| Id | Irrelevant | Numeric (unique) | 1020 Unique | [110,180270] | Zero | Zero | Zero |
| Indication | Input | Nominal | 8 Distinct | CVA|A-F|TIA|ASx|Asx | One | Two | One |
| Diabetes | Input | Nominal | 3 Distinct | Yes|No | Zero | Zero | One |
| IHD | Input | Nominal | 3 Distinct | Yes|No | One | One | Zero |
| Hypertension | Input | Nominal | 3 Distinct | Yes|No | Zero | Zero | One |
| Arrhythmia | Input | Nominal | 4 Distinct | Yes|No | Zero | One | One |
| History | Input | Nominal | 3 Distinct | Yes|No | Zero | One | Zero |
| IPSI | Input | Nominal (should be numeric) | 29 Distinct | [50,99] | One | One | One |
| Contra | Input | Nominal (should be numeric) | 25 Distinct | [40,60] | Zero | Two | Zero |
| Risk | Target | Nominal | 4 Distinct | Low|High | Two | Two | One |

*Table1. Summary Table of given ACW data*

Above is a table describing the raw data that has been supplied - before any cleaning has occurred. There are 1020 records with 1015 unique records, meaning there are 5 duplicated records in the data set. Of the total records there are also 25 records with issues of either duplicated entries, unknown, missing or null values. The value type for each attribute is mostly correct except for two, these are the ‘IPSI’ and ‘Contra’ attributes; the values are nominal where they should be numeric – this may prove to be problematic later when algorithms are used within Weka and will need to be rectified.