Data Mining and Decision Systems

08338

Assessed Coursework

Data Mining of Legacy Data

Stage 2. Data Cleaning

Student Number: 201303310

PDF Due: 2pm 16 November 2015 via E-Bridge

Sunday, 29 November 2015

# Data Cleaning

Data cleaning of the given data (**Data\_Base**) is described in Table 1 below, based on the worksheet **DataCleaning**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute Type** | **Value Type** | **Issue** | **Number of Records** | **Base Repair** |
|  |  |  |  |  |
| All | All | Duplicated Record | 5 | Delete duplicated records |
| All | Nominal | Homonym (case) | ASx 122 and Asx 22 (5 ASx and 1 Asx in records with other issues | Remove Asx records |
|  |  |  |  |  |
| Input | Nominal | Missing Value | 2 | Delete record |
| Input | Numeric | Missing Value | 1 | Delete record |
| Input | Nominal | Unknown Value | 5 (1 as numeric unknown) | Delete record |
| Input | Numeric | Unknown Value | 3 (1 has nominal unknown) | Delete record |
| Input | Nominal | Null Value | 4 | Delete record |
| Input | Numeric | Null Value | 1 | Delete record |
|  |  |  |  |  |
| Target (Risk) | Nominal | Missing Value | 2 | Move record to unknown risk |
| Target (Risk) | Nominal | Unknown Value | 2 | Move record to unknown risk |
| Target (Risk) | Nominal | Null Value | 1 | Move record to unknown risk |

*Table1. Summary Table of Data Cleaning of the given ACW data*

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Type** | **Value Type** | **Issue** | **Alternative Repairs** |
|  |  |  |  |
| All | All | Duplicated Record | none |
| All | Nominal | Homonym (case) | Change to lower case |
|  |  |  |  |
| Input | Nominal | Missing Value | Replace record with risk stratified mode |
| Input | Numeric | Missing Value | Replace record with risk stratified mean |
| Input | Nominal | Unknown Value | Replace record with risk stratified mode |
| Input | Numeric | Unknown Value | Replace record with risk stratified mean |
| Input | Nominal | Null Value | Replace record with risk stratified mode |
| Input | Numeric | Null Value | Replace record with risk stratified mean |
|  |  |  |  |
| Target (Risk) | Nominal | Missing Value | none |
| Target (Risk) | Nominal | Unknown Value | none |
| Target (Risk) | Nominal | Null Value | none |

Now add some words describing the issues with this data and table, for instance alternative means to obtain “clean” records.