**CS539 Machine Learning. Fall 2014. Project 2: Decision Trees. Student’s Name: Daniel Fitzgerald**

**Dataset Description and initial data preprocessing if any (at most ½ page):**

As with project 1, there were several preprocessing steps taken to make the dataset more amenable to analysis. These were done in Weka and included:

* Removing the **weight** attribute which was too sparse for meaningful analysis.
* Removing the **payer\_code**  attribute which was considered irrelevant to the **readmitted** outcome.
* Removing the **encounter id** attribute, which is likewise irrelevant.
* Removing the **patient\_nbr** attribute because it is only useful for associating data from the same patient with each other, but the values have no meaningful reference, proportion, or distance from each other. Furthermore, no accurate medical prediction can be made or information gained from a new, unseen patient number, so this attribute should not be used for creating decision trees.
* The diagnosis attributes, **diag\_1**, **diag\_2**, and **diag\_3** were removed for a similar reason – they are completely categorical (no meaningful distance, reference, or proportion) and the codes observed in this data cannot be taken for the full range of codes that are possible, so any decision tree made from these codes could not handle new diagnosis codes that may be encountered in future data. They are not suitable for building a predictive model.

Most decision tree algorithms are applicable only to discrete attributes. Therefor, all remaining numeric attributes were discretized to ordinal, interval, or binary values. This was done using the supervised discretize in WEKA.

* **admisision\_type\_id** was converted directly to nominal (as it should have been)
* Likewise for **discharge\_disposition\_id** and **admission\_source\_id**
* **time\_in\_hospital** already have integer values in the range [1,14], following a rapid decay in frequency vs. time. This was discretized using WEKA’s supervised discretize filter, resulting in the ranges (-inf,1.5], (1.5,3.5], (3.5,inf).
* The attributes **num\_lab\_procedures**, **num\_ procedures**, **num\_medications**, **num\_outpatient**, **num\_emergency**, **num\_inpatient**, and **num\_diagnoses** were similarly discretized.

**Three Guiding Questions (at most ½ page):**

1. **…**
2. **…**

**Weka Code and Results (at most 1 page)**

**Description of Weka Code (at most 1/2 page):**

Notes from class

-can use continuous variables

-shorter trees from letting c4.5 discretize them – better than binning beforehand

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Weka Experiments.** *Weka code description and this table should take at most 1 page.* | | | | | | | | | | | |
| **Tech.** | **Guiding**  **questions** | **Train/**  **Test** | **Pre-process** | **Parameters** | **Post-process** | **Confusion matrix**  **& Accuracy** | **Time to build model** | **Size of model** | **Interesting patterns in the model** | **Analysis & observations about experiment** | **You can add other columns** |
| ZeroR?  OneR?  J4.8?  …? | 1? 2? 3? | 10-fold cross-val? |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |

**R Code and Results (at most 2 page)**

**Description of R Code, Functions, and Packages (at most 3/4 page):**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **R Experiments.** *R code description and this table should take at most 2 pages.* | | | | | | | | | | | |
| **Tech.** | **Guiding**  **questions** | **Train/**  **Test** | **Pre-process** | **Parameters** | **Post-process** | **Confusion matrix**  **& Accuracy** | **Time to build model** | **Size of model** | **Interesting patterns in the model** | **Analysis &**  **observations about experiment** | **You can add other columns** |
| ZeroR?  OneR?  J4.8?  …? | 1? 2? 3? | 10-fold cross-val? |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |

**Summary of Results and Visualizations. (at most 1 page)**

**Advanced Topic: (at most 1 page) <include name of the topic here>**

**List of sources/books/papers used for this topic (include URLs if available):**

* …
* …
* …

...

**Summary of what you learned:**

**How does this topic relate to trees and the material covered in this course?**