**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 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, scale, 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.

Decision trees 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

* **admisision\_type\_id** was converted directly to nominal (as it should have been)
* same for **discharge\_disposition\_id** and **admission\_source\_id**
* **time\_in\_hospital** already have integer values in the range [1,14] and was discretized using 7 bins of equal width 2
* **Patient\_nbr** was

**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):**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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?**