**Notes on Ages and new Attributes added**

**As discussed last time we met, I played around with different ways to predict age for those who have missing age information. I created some variations on the approaches, so only one of the following attributes from each category should be used. I don’t have access to weka right now, so I can’t wait to see if any of these help! The corresponding files are Family with Age Training file.csv, and Family with Age Test file.csv**

**Category: AGE**

1. Original attribute “age” (265 missing values)
2. “age by group” (no missing values)

Uses these rules:

|  |  |  |
| --- | --- | --- |
| * If in a group of unrelated people, set missing age equal to the mean of the others | | |
| * If missing age has siblings, set equal to the mean of the sibling ages |  |  |
| * If person with missing age has spouse, set based on the spouse age (female -2, male +2) | | |
| * If person with missing age has child, set based on oldest child + 20 |  |  |
| * If person with missing age has parent, set based on youngest parent -20 | |  |

38 were filled in using these rules

65 could not be determined by their group (either all of the group were missing ages, or family relationship was indeterminate.

These 65 were filled in (as were the 161 without group) by using their demographics:



1. “age by demographics” (no missing values)

All people with missing ages were filled in based on the table above with no accounting for groups or relationships. (So, values are the same as in the “age by group” except for the 38 who were filled in based on the rules. Differences between the methods were almost centered on 0 for those 38, but there was a wide spread of about +/- 16 years.)

**Category: Age Bucket**

1. “child\_adult\_with\_est” (no missing values)

This is based on translating the ages into two buckets – child or adult. 14 was used as the cutoff (above 14 = adult.) The age is based on the “age by group” method

1. “child\_adult\_no\_est” (50 missing values)

This is based on translating ages into two buckets – child or adult, but with no assumptions about age. The reason it does not have as many missing values as the original age column is that moniker is used as a classifier for child and adult in some cases. The cutoff is still 14, but all “Master” monikers are marked child (except 2 that were 14.5), and all “Mr.” (except the one who is known to be 11) are marked adult. (It is possible that some boys with missing age information were inadvertently classified as Mr., especially those with non-anglo names.) All “Mrs.” are marked as adult (the youngest married woman on board with age information was 14.) Unfortunately, “Miss” cannot be subdivided as “Master” was, since it is used for unmarried adult women as well as girls. So, anyone with the moniker “Miss” and no age information is left as a missing value. The professions (Dr. Rev, Col.) and nobility were all marked as adult. So basically, the 50 missing values are all females with the moniker “Miss.”

1. child\_adult\_senior (256 missing values)

Here I added a 3rd category for senior, for anyone 60 years or older. This one has almost as many missing values as the original age attribute; I could no longer use moniker because it does not distinguish adults from seniors. Only the moniker “Master” could be used with any confidence.