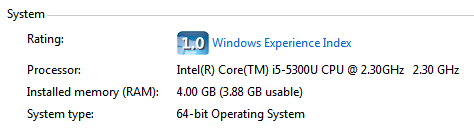
Assignment 2 Report

COMP 4331

PHAM Trung Kien

20553388

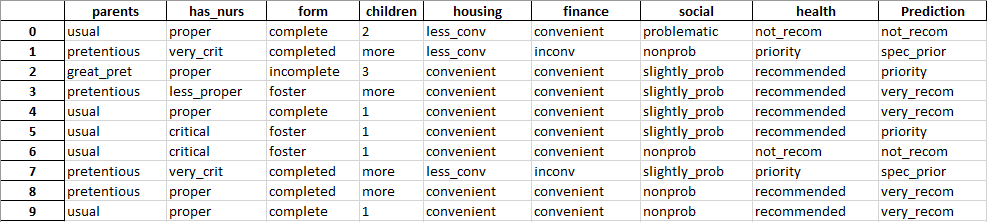
1. Environment



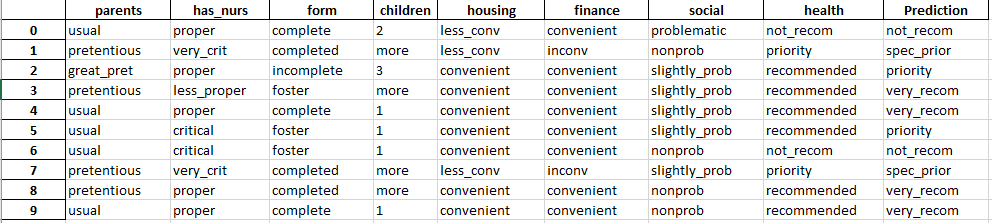
Besides, I use python 3. Other than these, there are some libraries I use, which are all imported in the source codes.

2. Result

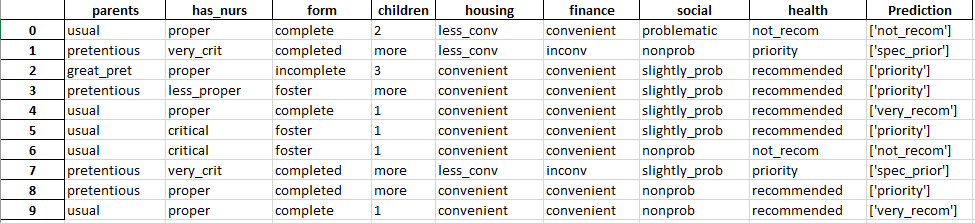
* ID3: Runtime is 16.3609356880188 s



* C4.5: Runtime is 16.70995569229126 s



* Naïve Bayes: Runtime is 1.3450770378112793 s



Note: For clearer result tables, please look at the 3 excel files submitted together with the report and source codes.

3. Comparison

* There is no differences between predictions of implementations with ID3 and C4.5
* However, differences occur between implementations with ID3 (as well as C4.5) and Naïve Bayes at the 3rd and 8th rows’ predictions. In details:
* 3rd row:

ID3 and C4.5 predict ‘very\_recom’

Naïve Bayes predicts ‘priority’

* 8th row:

ID3 and C4.5 predict ‘very\_recom’

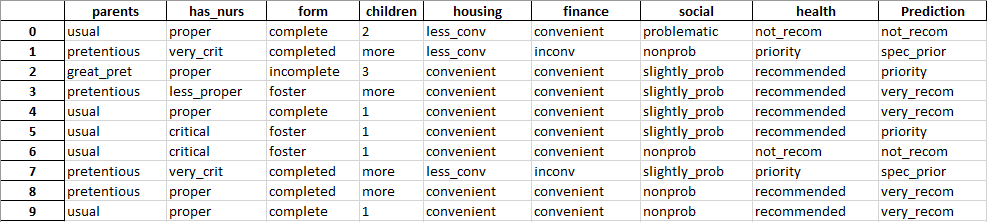
Naïve Bayes, again, predicts ‘priority’

Reasons that I believe to hold responsible for:

* No differences between predictions of implementations with ID3 and C4.5 : is based on the train dataset, after the training (calculating entropy and partitioning) processes, two identical Decision Trees for each implementation has been recursively constructed. In other words, comparing Info(T) – Info(Attribute, T) between different attributes will relatively resulted in the same as comparing between different attributes for this train dataset. As a result, they will produce the same predictions.
* Differences occur between implementations with ID3 (as well as C4.5) and Naïve Bayes at the 3rd and 8th rows’ predictions : is because unlike ID3 and C4.5 with main principle of partitioning train dataset by finding the most informative attribute with highest entropy and do not require class conditional independence, Naïve Bayes Classifier based on conditional probability for calculating prior probabilities in order to rank every class values when considering one valued combination of attributes and it does require class conditional independence. In other words, the class conditional independency between columns of the train dataset may hold responsible.

4. Implication

* Having observed the predicted values from all three classifiers, I believe the predicted output for each data point in the test file is as below:



, which are the predictions made by implementation with ID3 and C4.5.

The reasons for my conclusion are:

* By voting, 2/3 implementations (ID3 and C4.5) do generate those predictions above
* In addition, we know that although Naïve Bayes Classifier can produce good predictions in most of the cases, one of the huge disadvantages of Naïve Bayes Classifier is that when dependencies occur between class and attributes or among attributes, they can not be modeled by Naïve Bayes Classifier, subsequently resulting in loss of accuracy. From the train dataset, we can easily find some dependencies such as relation among attribute “health”, attribute “finance” and class “NURSERY”…
* As support to the first and second reasons, ID3 and C4.5 do not require class conditional independencies between attributes of their train datasets.