Data Mining

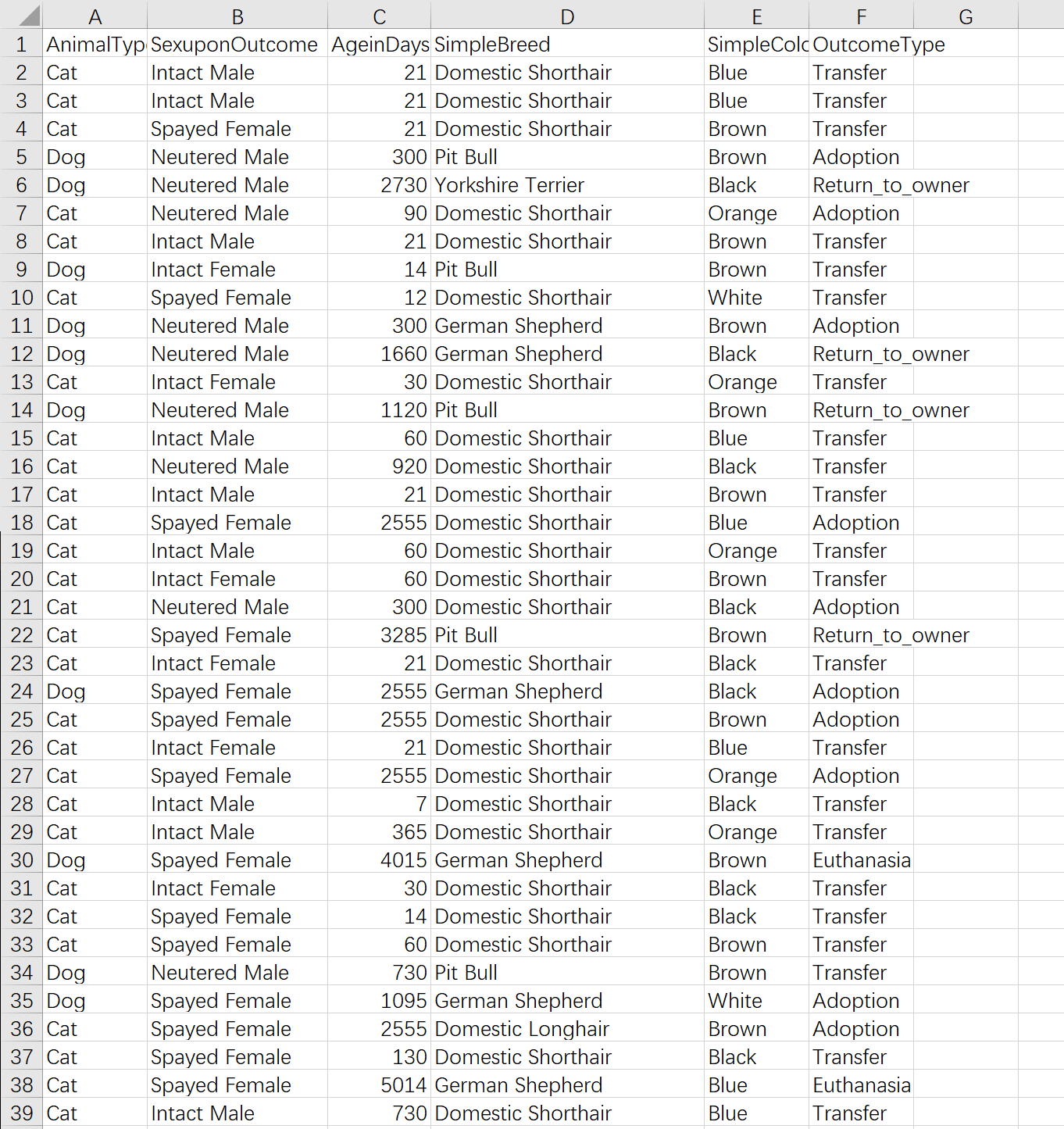
Project2 Classification

孫啟慧

P78063033

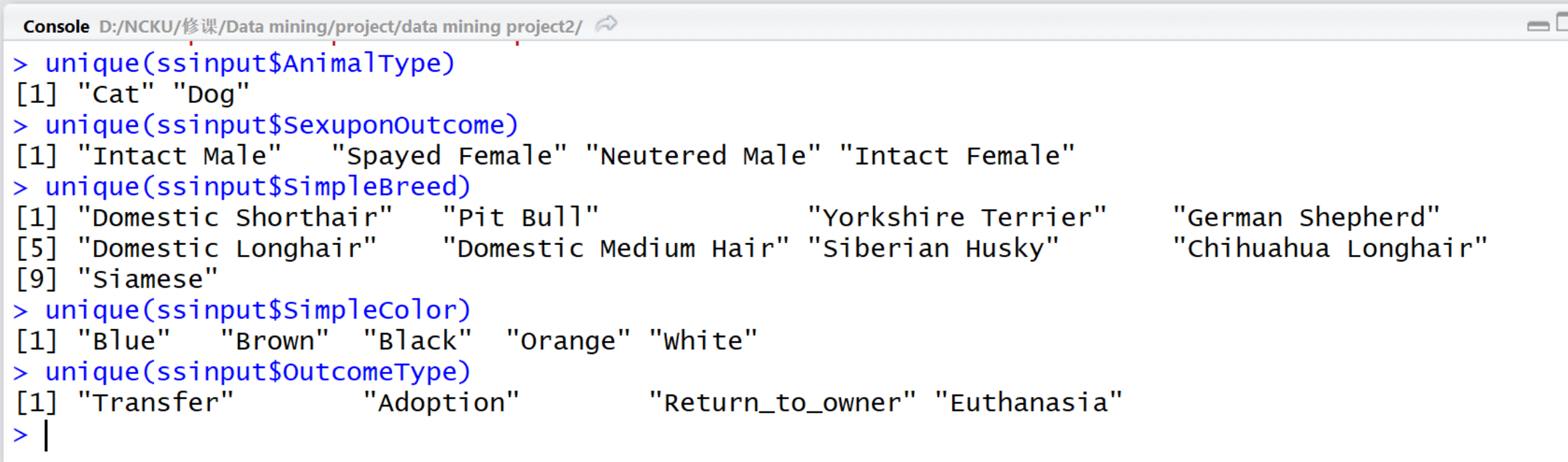
1. **實驗設計**

本次作業要求生成一組絕對規則（absolutely right rules）並使用其生成positive data和negative data。因為對於生成資料並沒有什麼頭緒，又想到kaggle上面有很多開源的資料集，所以我按照興趣選擇了一份關於動物收容所的記錄，他有記錄貓和狗兩種動物的詳細資訊以及最後的去向。通過使用我想出來的絕對規則對其進行資料進行篩選和改造，相應的部分實驗資料in.csv如下。



In.csv中的資料是按照絕對規則生成資料，主要包括五個features，也就是K為5，相應的欄位分別為：Animal Types（cat和dog兩種動物）、SexuponOutcome（Intact Male，Intact Female，Spayed Female和Neutered Male四種狀態，分別表示完整雄性，完整雌性，切除卵巢的雌性和被閹割的雄性）、AgeinDays（表示該動物的年齡，以天數來計算）、SimpleBreed（表示該動物的品種，值得注意的是，貓跟狗的品種不同，也就是說當已知品種的時候就可以表示被分析物件是貓還是狗）、SimpleColor（表示該動物的毛色，因為原有資料包含混色，所以我有簡單對其進行處理，選取主要毛色作為其毛色）、OutcomeType（則是表示該動物的結果，包含轉移到其他收容所，被人收養，年紀過大被安樂死以及送還給其原本的主人，這四種）。最後生成的data（不包括表頭）一共356項，也就是M為356。

各項feature的值如下：



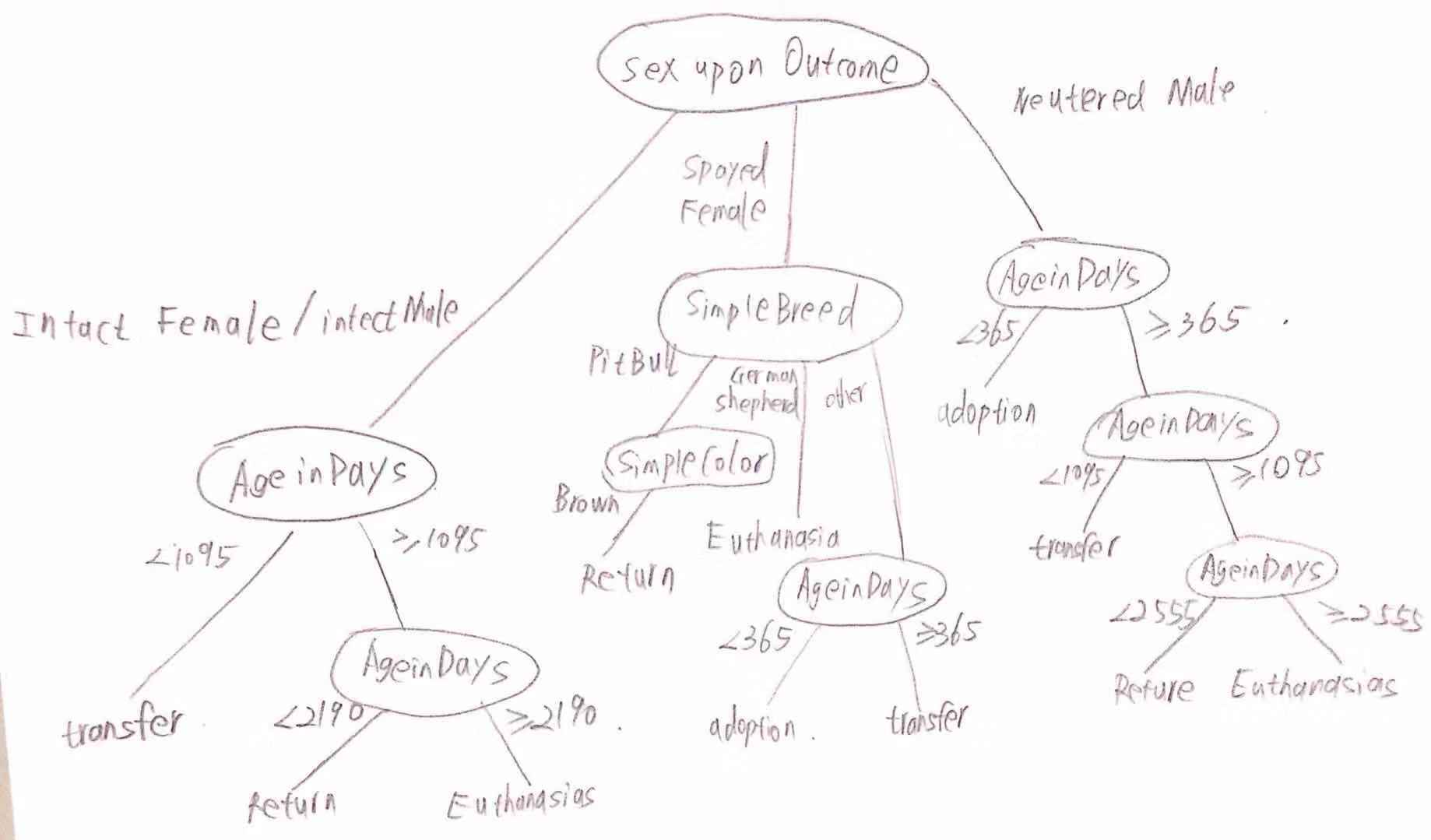
動物的種類一共有九種，而相應的color有五種，其他屬性的值上面已經有介紹。因為AgeinDays屬性有太多的可能，我就不一一列舉在這邊，具體可以參照in.csv。

絕對規則主要如下：

1.若該動物為Intact Male/Female，當其AgeinDays小於1095天的時候該動物會被transfer到其他動物收容所，若在1095和2190之間的時候該動物會被return給原主人，若大於2190則會因為年紀過大被安樂死。

2.若該動物為Spayed Female，則當動物的品種為pit bull，顏色為Brown的時候會被返還給原主人，其品種為German Shepherd的時候會被安樂死，對於其它品種的貓和狗，若其小於365天則會被adoption，而大於365天的則會被transfer去其他的動物收容所。

3.若該動物為Neutered Male則當其年紀小於365天的時候會被adoption，結餘365和1095之間的時候會被transfer去其他的收容所，介於1095和255之間的時候則會被return給原主人，當大於2555的時候會被安樂死。



1. **對比試驗**

為了比較絕對規則和現有的decision tree的演算法，我將絕對規則生成的in.csv作為input，使用R語言撰寫的決策樹演算法對資料進行建模，主要採用的是C5.0和rpart。相應的代碼如decision\_tree.r中所示。（因為R有撰寫過相應的演算法了，故直接install和library相應演算法所在package即可）

1. **C5.0 演算法**

C5.0相較於C4.5主要進行了如下的改進：增加了對Boosting的支援，它同時也用更少地記憶體。它與C4.5演算法相比，它構建了更小地規則集，因此它更加準確。

使用該演算法對絕對規則生成的data進行tree model建立的詳細結果如下

Call:

C5.0.formula(formula = OutcomeType ~ ., data = ssinput)

C5.0 [Release 2.07 GPL Edition] Mon Nov 16 17:05:53 2018

-------------------------------

Class specified by attribute `outcome'

Read 356 cases (6 attributes) from undefined.data

Decision tree:

AgeinDays > 1090:

:...AgeinDays > 3285: Euthanasia (17/1)

: AgeinDays <= 3285:

: :...SexuponOutcome in {Intact Male,Neutered Male,

: : Intact Female}: Return\_to\_owner (36/5)

: SexuponOutcome = Spayed Female: Adoption (28/1)

AgeinDays <= 1090:

:...SexuponOutcome in {Intact Male,Intact Female}: Transfer (96)

SexuponOutcome in {Spayed Female,Neutered Male}:

:...AgeinDays <= 33: Transfer (27)

AgeinDays > 33:

:...AgeinDays <= 330:

:...SexuponOutcome = Neutered Male: Adoption (77)

: SexuponOutcome = Spayed Female:

: :...AgeinDays <= 240: Transfer (14/2)

: AgeinDays > 240: Adoption (2)

AgeinDays > 330:

:...SexuponOutcome = Neutered Male: Transfer (30/1)

SexuponOutcome = Spayed Female:

:...SimpleBreed in {Domestic Shorthair,Yorkshire Terrier,

: German Shepherd,Domestic Longhair,

: Domestic Medium Hair,Siberian Husky,

: Chihuahua Longhair,Siamese}: Adoption (21)

SimpleBreed = Pit Bull:

:...SimpleColor in {Blue,Black,Orange,White}: Adoption (4)

SimpleColor = Brown: Return\_to\_owner (4)

Evaluation on training data (356 cases):

Decision Tree

----------------

Size Errors

12 10( 2.8%) <<

(a) (b) (c) (d) <-classified as

---- ---- ---- ----

131 3 (a): class Adoption

16 2 (b): class Euthanasia

1 35 (c): class Return\_to\_owner

1 3 164 (d): class Transfer

Attribute usage:

100.00% AgeinDays

95.22% SexuponOutcome

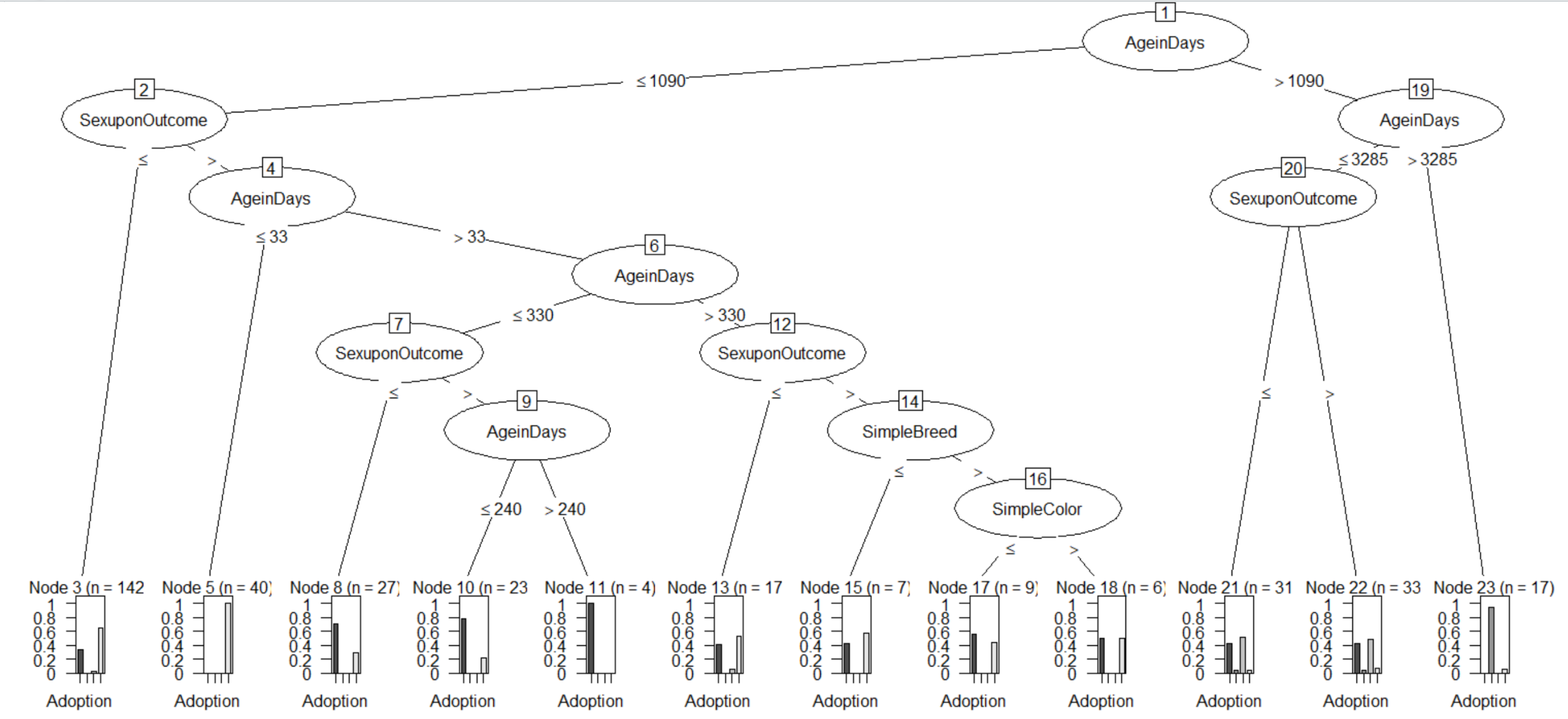
8.15% SimpleBreed

2.25% SimpleColor

Time: 0.0 secs

可以看出整個decision tree首先用AgeinDays來做屬性區分，之後針對於不同的區間再使用SexuponOutcome，SimpleBreed，SimpleColor三個屬性來生成最後decision tree，跟我用來生成data的絕對規則相比有所不同，這一點從AgeinDays的分割區間就可以看出來了。而且對於毛色還有品種這兩個屬性來看，C5.0生成的model和絕對規則也不一樣。相比較而言既包含了絕對規則，又因為某些屬性的不同分類準則，從而多了很多其他的分類準則。

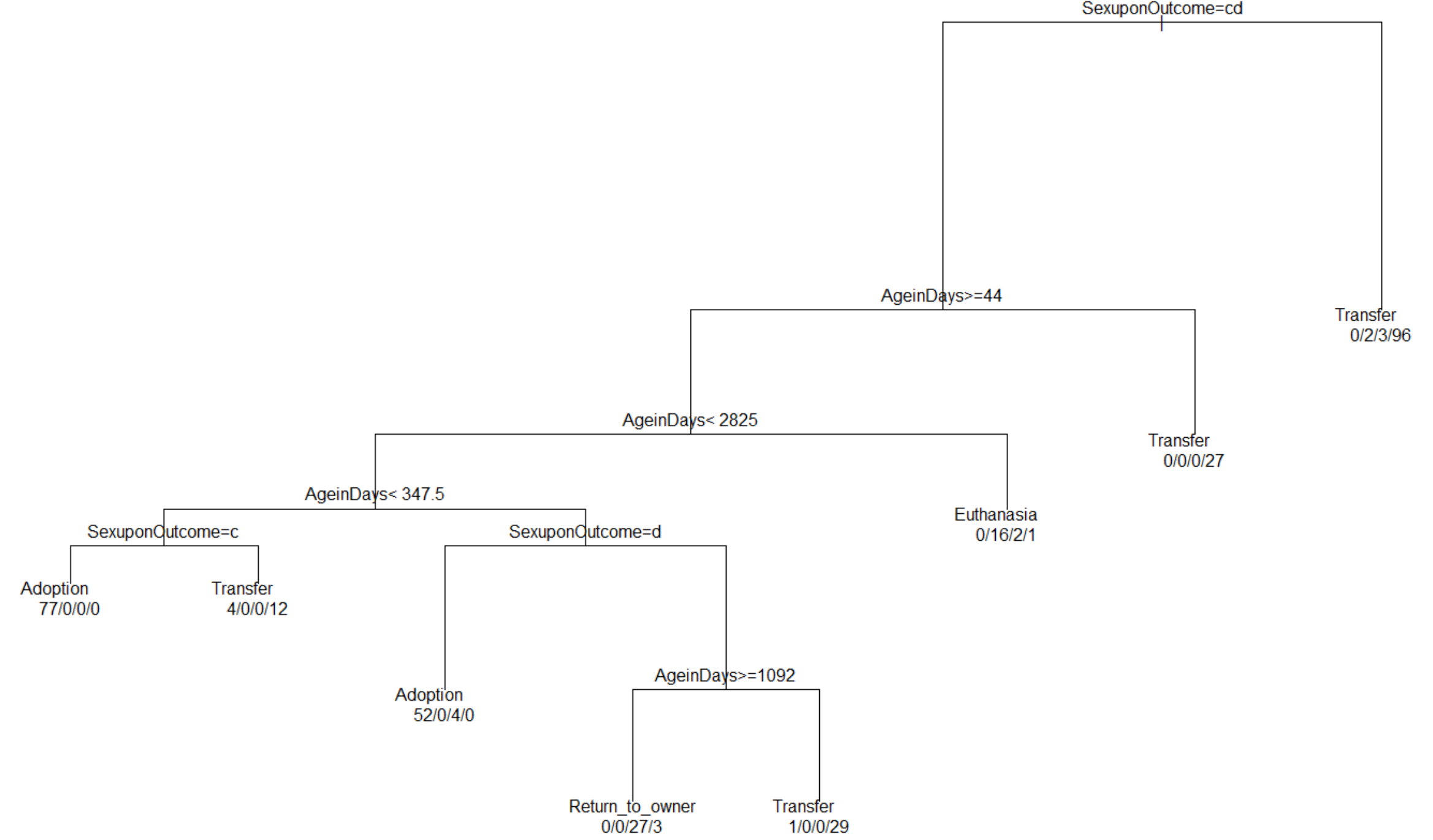
以圖的形式表示則如下，每各葉子節點的圖都表示預測為相應的outcome類別的可能性，值越高則可能性越大，因為結果總共有四種所以會有四條表示可能性的柱子，最高的一條表示最後分類的結果（由於圖的size的關係，所以部分文字有所缺失）。每一個內部結點都表示相應的分割依據。



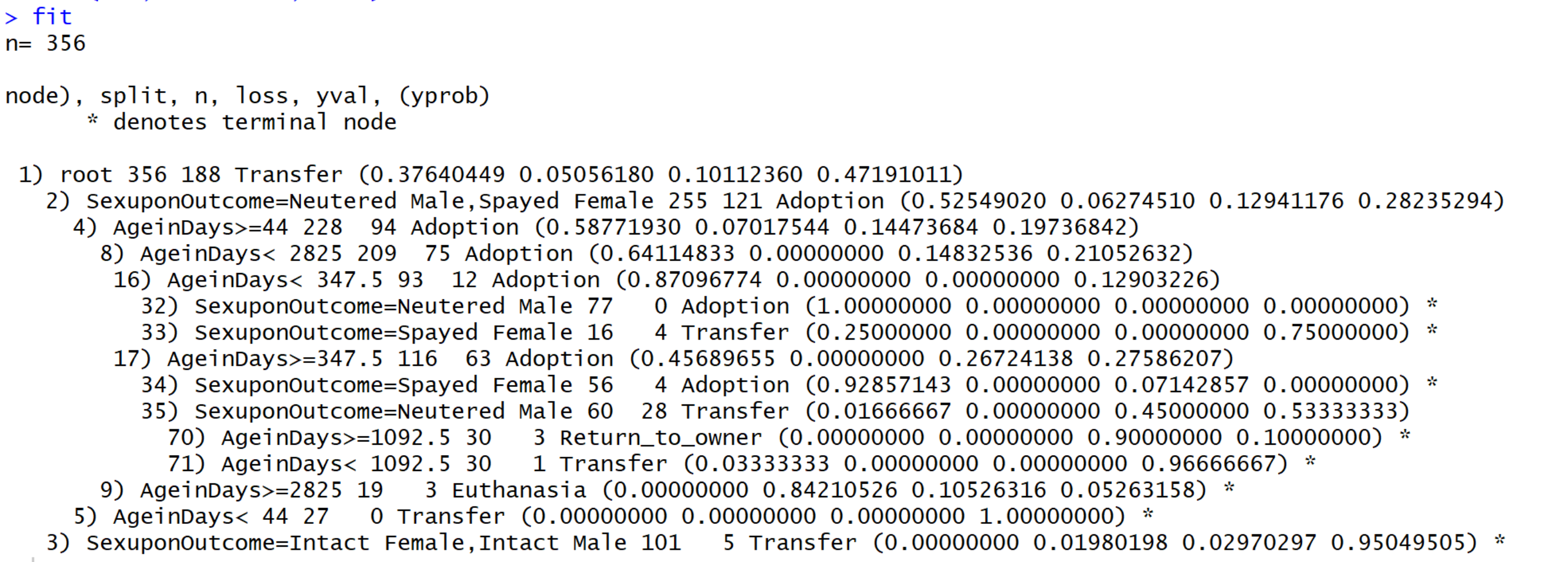
**CART分類演算法**

CART (Classification and Regression Trees)與C4.5演算法是非常相似的，但是CART支持預測連續的值（回歸）。CART用訓練集和交叉驗證集不斷地評估決策樹的性能來修剪決策樹，從而使訓練誤差和測試誤差達到一個很好地平衡點。

下圖是採用CART演算法對於絕對規則生成的資料進行訓練從而得到的二叉樹model，從結果可以看出主要採用了SexuponOutcome和AgeinDays兩個屬性來進行分類，關於毛色和種類的屬性已經被忽略了，結果的屬性主要是按照可能性來分配，最大可能性的一類就被當做最後的結果。在這裡跟我們的絕對規則相比較之後就會發現，分類的規則相差很多。



分類的詳細資訊如下圖所示，n表示有多少個結果，而node), split, n, loss, yval, (yprob)則是其最終的結果的形式，最大的可能性所在的結果類會被判斷成最後的結果。



1. **總結**

1.通過比較兩種decision tree（C5.0和CART）對於絕對數據生成的tree model可以發現兩種方式得到的分類準則與我生成資料時候所用的絕對準則不一致。

出現上述主要問題的原因在於，我在制定絕對規則的時候只是先按照SexuponOutcome的種類開始制定不同屬性情況下的收養結果的影響，之後再觀察其他feature的結果對於最後收養結果的影響來設計規則並篩選資料。而眾人熟知的ID3是使用Information Gain來計算，C4.5是使用gain ratio來計算，C5.0和C4.5類似，只是加入boosting，而CART則是使用gini index來劃分。不同的計算方法導致不同的特徵的權重不一樣，根據各個演算法的規定，相應的特徵排序也就不同，而後決策樹演算法得到的分類規則和我的絕對規則並不一樣讓我知道特徵排序的不同可能造就整個tree model完全不一致。

2.由於C5.0（C5.0是商業用的，所以其原始程式碼並沒有開源，只能直接進行使用，所以只能知道其官方文檔中公佈的針對C4.5的性能比較）和CART演算法均是R語言的package有寫好，所以根據其提供的結果可以看到他們均有被優化過，速度很快，跑我自己準備的in.csv只需要大約0秒。