Step1 : Design Feature to classify data

我設計的資料說明如下：

首先我的分類問題是區分性別，目前是隨機產生100筆data（可調整），

每筆Data有七個Data Attributes，

Data的欄位有 ['gender' , 'age', 'blood type', 'height', 'weight', 'bmi', 'salary']

Gender 的值為0或1，0代表男生，1代表女生，隨機產生。

Age 的值會是1～50的整數。

Blood type的值0～3的整數，分別對應A、B、O、AB血型。

Height 的值會根據男女和年齡的部分產生：

* 男生平均身高為175,標準差5
* 女生平均身高為163,標準差4

根據gaussian distribution random產生。

另外，不足18歲的

* + 男生：會扣掉 (18-age)x2
  + 女生：會扣掉 (18-age)x1.2

Weight 的值會根據男女和年齡的部分產生：

* 男生平均體重為69, 標準差5
* 女生平均體重為57, 標準差4

根據gaussian distribution random產生。

另外，不足18歲的

* + 男生：會扣掉 (18-age)x2
  + 女生：會扣掉 (18-age)x1.2

BMI 的值是由前項的Height 及Weight算出來的，公式如下：



Salary 的值會根據男女和年齡的部分產生：

* 男生平均薪水為50, 標準差5
* 女生平均薪水為40, 標準差4

根據gaussian distribution random產生。

另外，不足18歲的都會設為0。

Step2 顯示某一次執行結果

下兩張圖分別是用GINI及 Entropy 做區分產生的分類樹。



Fig 1 . Decision Tree Result by GINI

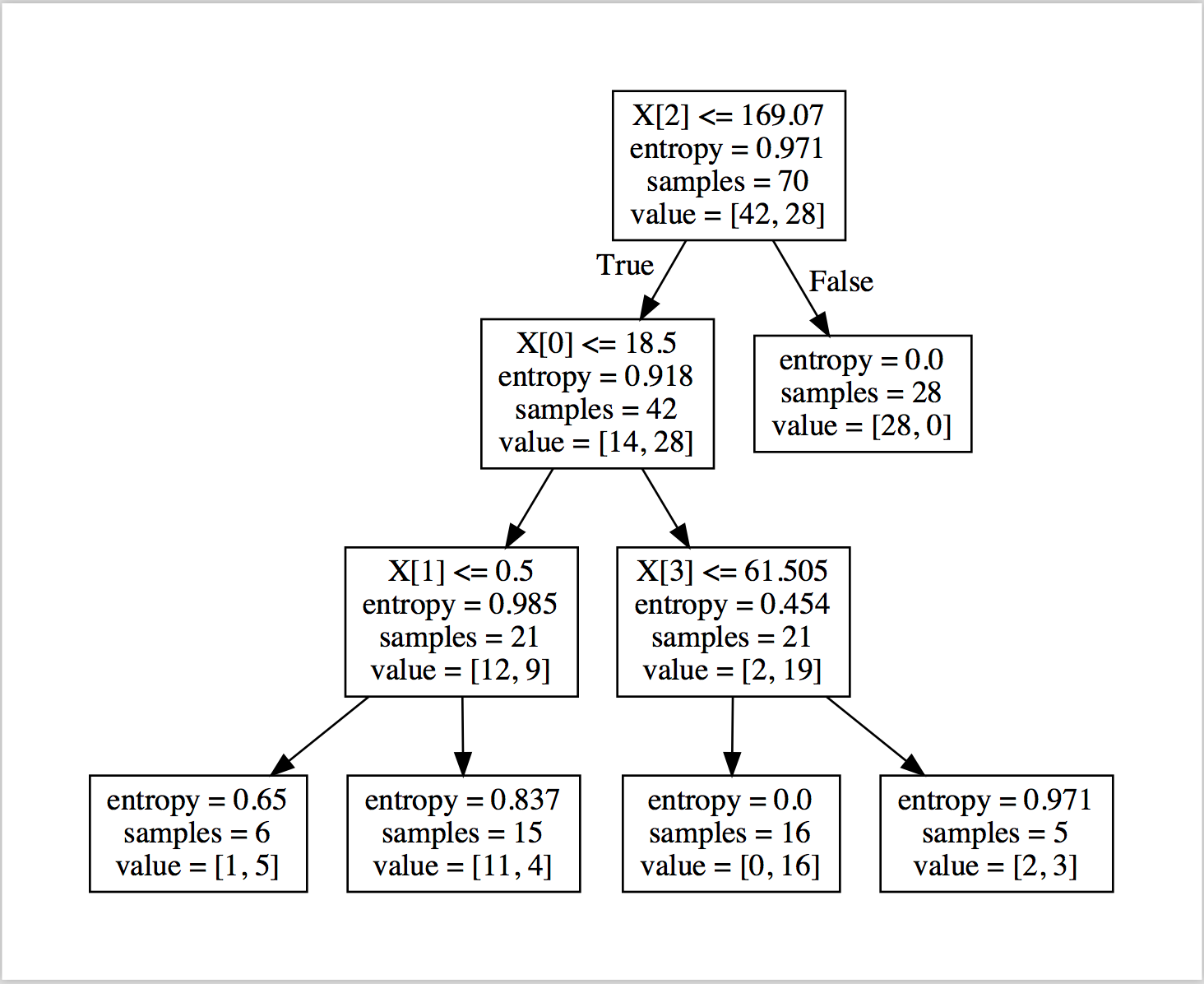


Fig 2 . Decision Tree Result by Entropy

Accuracy by Gini is 80.0

Accuracy by Entropy is 76.67

Step 3

在做了幾次的分類樹的過程中，大部分是會先用類身高、體重做為第一次分支的依據，這與我設計所用的規則，是較相符合的。但少部份會由BMI作為分類的依據，但通常出現這樣的情形，他算出來的Accuracy明顯是比其他次低的。

在分類的過程中，沒有出現過以血型作為分類標準的例子，顯然血型這個屬性的資料對於判斷性別這件事情是沒有幫助的。

薪水作為區分標準的次數相當的少，可能是在區分上無法區分的那麼好，因爲未成年者比沒有薪水的資料，且薪水的男女差距也沒有設定的如同身高及體重的差距那麼大，如果將差距調整的更大可能會被Decision Tree作為分類的標準。