Sarah Nicholson

DATA 210 – Project 01

3/16/18

The first step is to calculate the entropy of the entire system/table. The entropy

for the entire system/table is **0.884**.

Now let's try to split by the attribute Gender:

The entropy for the sub-table Gender: Male (M) is **0.893**

The entropy for the Sub-table Gender: Female (F) is **0.8797**

The weighted/expected entropy is **0.884**, which means that the information gain is

**0.884 – 0.884 = 0**.

Now let's try to split by the attribute Alcoholism:

The entropy for the sub-table Alcoholism: Non-Alcoholic (0) is **0.882**

The entropy for the sub-table Alcoholism: Alcoholic (1) is **0.955**

The weighted/expected entropy is **0.884**, which means that the information gain is

**0.884 - 0.8838 = 0.0002**.

Now let's try to split by the attribute Diabetes:

The entropy for the sub-table Diabetes: Non-Diabetic (0) is **0.889**

The entropy for the sub-table Diabetes: Diabetic (1) is **0.817**

The weighted/expected entropy is **0.8834**, which means that the information gain is

**0.884 – 0.8834 = 0.0006**.

Since splitting by **Diabetes** yields the most information gain, then that attribute would

be our first split. Now, let's repeat the process for the left sub-table and the

right sub-table.

Splitting the left sub-table by Gender:

The entropy for the sub-table Gender (M) and Diabetes (0) is **0.89781**

The entropy for the sub-table Gender (F) and Diabetes (0) is **0.88472**

The weighted/expected entropy is **0.889097**, which means that the information gain is

**0.88919 – 0.889097 = 0.000097.**

Splitting the right sub-table by Gender:

The entropy for the sub-table Gender (M) and Diabetes (1) is **0.82022**

The entropy for the sub-table Gender (F) and Diabetes (1) is **0.81498**

The weighted/expected entropy is **0.81655**, which means that the information gain is

**0.817 – 0.81655 = 0.00045.**

Splitting the left sub-table by Alcoholism:

The entropy for the sub-table Alcoholism (0) and Diabetes (0) is **0.886909**

The entropy for the sub-table Alcoholism (1) and Diabetes (0) is **0.961385**

The weighted/expected entropy is **0.888702**, which means that the information gain is

**0.88919 – 0.888702 = 0.00049**

Splitting the right sub-table by Alcoholism:

The entropy for the sub-table Alcoholism (0) and Diabetes (1) is **0.813747**

The entropy for the sub-table Alcoholism (1) and Diabetes (1) is **0.88288**

The weighted/expected entropy is **0.8162386**, which means that the information gain is

**0.817 – 0.8162386 = 0.00076.**

Since splitting by **Alcoholism** yields the most information gain, then that attribute would

be our second split. Now, let's repeat the process for the final time since **Gender** is the only attribute remaining.

Splitting the far-left sub-table of Alcoholism by Gender:

The entropy for the sub-table Gender (M) Alcoholism (0) Diabetes (0) is **0.894806**

The entropy for the sub-table Gender (F) Alcoholism (0) Diabetes (0) is **0.882944**

The weighted/expected entropy is **0.886833**

Splitting the middle-left sub-table of Alcoholism by Gender:

The entropy for the sub-table Gender (M) Alcoholism (1) Diabetes (0) is **0.952712**

The entropy for the sub-table Gender (F) Alcoholism (1) Diabetes (0) is **0.972724**

The weighted/expected entropy is **0.960711**

Splitting the middle-right sub-table of Alcoholism by Gender:

The entropy for the sub-table Gender (M) Alcoholism (0) Diabetes (1) is **0.815451**

The entropy for the sub-table Gender (F) Alcoholism (0) Diabetes (1) is **0.813066**

The weighted/expected entropy is **0.813745**

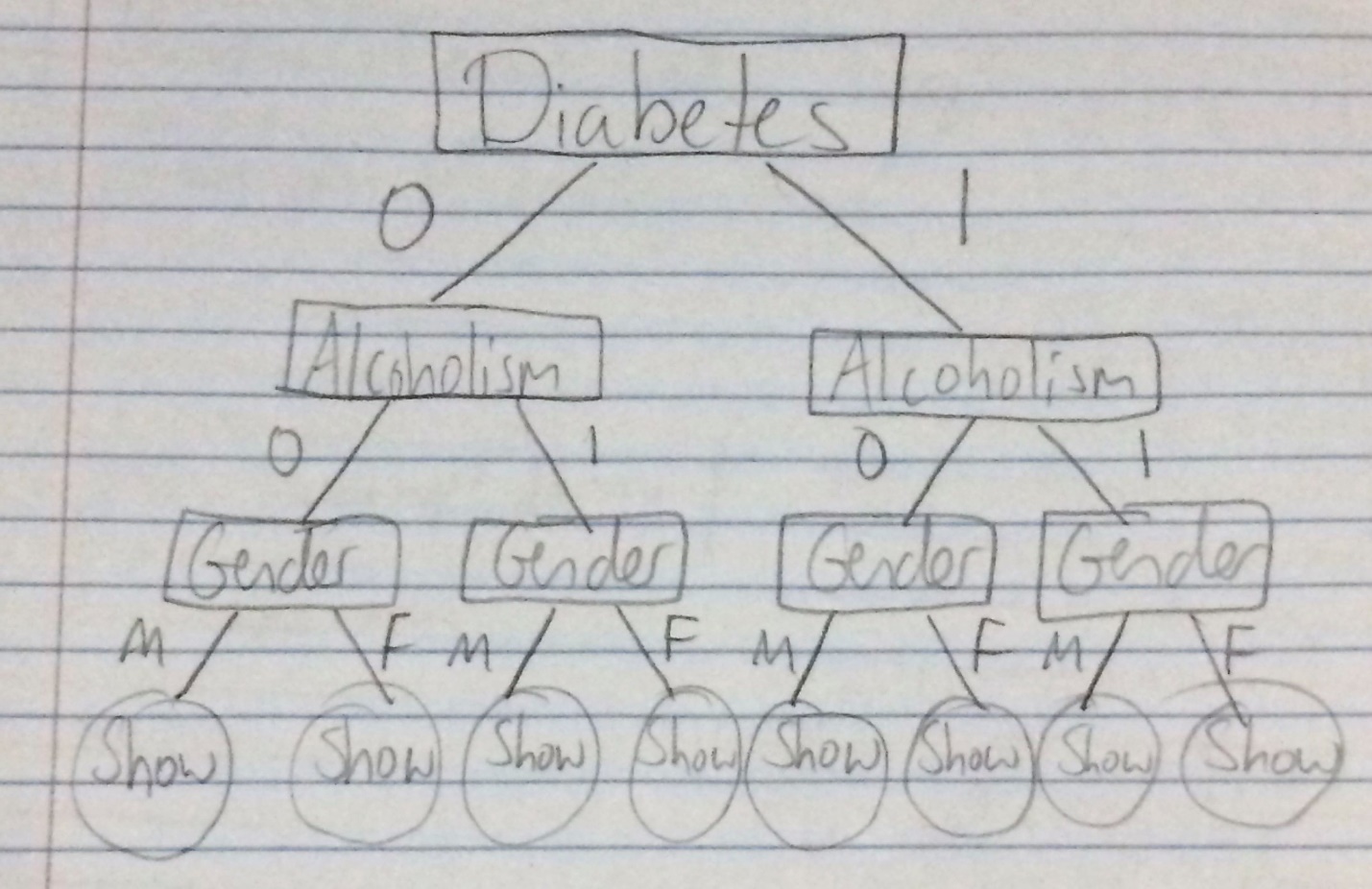
Splitting the far-right sub-table of Alcoholism by Gender:

The entropy for the sub-table Gender (M) Alcoholism (1) Diabetes (1) is **0.868275**

The entropy for the sub-table Gender (F) Alcoholism (1) Diabetes (1) is **0.911828**

The weighted/expected entropy is **0.881863**

**Final Tree:**

****