CS 180 Machine Problem 2:
Learning Decision Trees

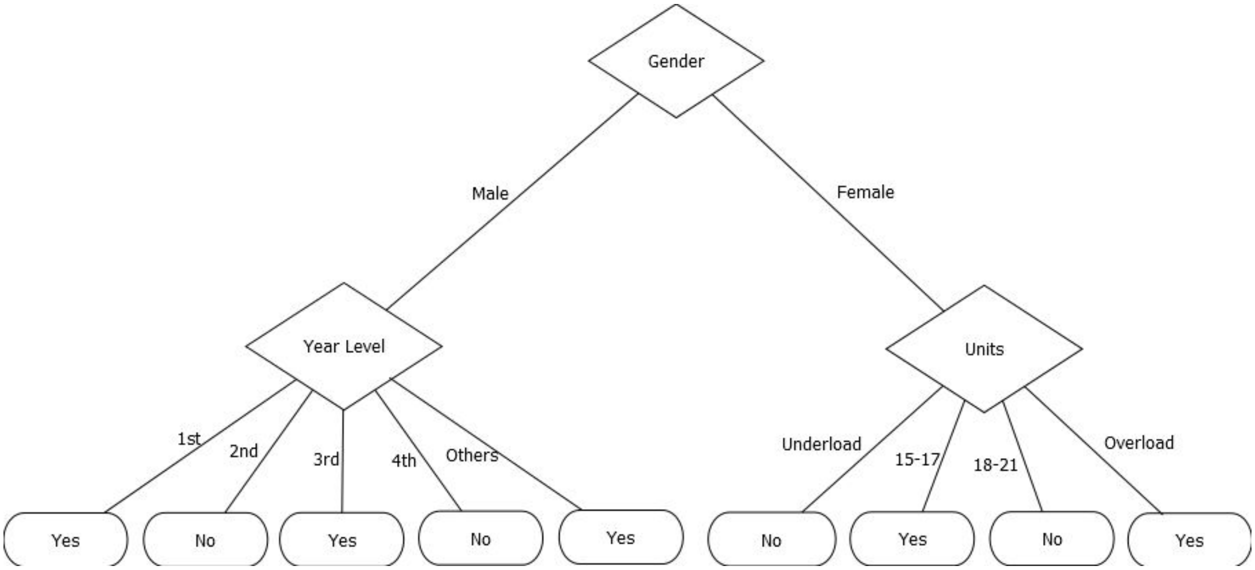
Written Report

Submitted by:
ESGUERRA, Regina Alyssa D.
KAPUNAN, Justine Mari R.

November 19, 2015

DECISION TREE LEARNING

Decision Tree Learning makes use of a decision tree structured based on a given set of training data. It is through these training data that it is able to construct the tree and “learn”. The internal nodes of the tree denote the attributes, the branches represent the attribute’s values, and the leaf nodes embody the results or classifications. This method of learning uses the decision tree as a predictive model to reach an inputted test data’s target classification. The decision tree determines the classification of a test instance through traversing the nodes of the tree that correspond or match its values until it reaches a leaf node, which provides the classification of the instance.

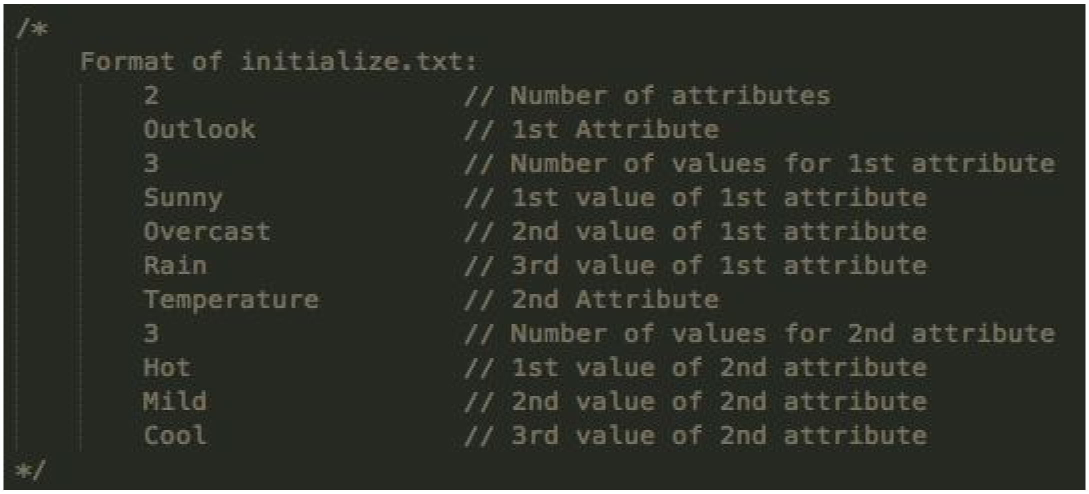


For example, in the decision tree given above, the internal nodes are the attributes, namely Units, Gender, and Year Level. The branches are the attribute’s values (Male and Female for Gender; 1st, 2nd, 3rd, 4th, and Others for Year Level; and Underload, 15-17, 18-21, and Overload for Units). The leaf nodes represent the classifications which are Yes and No. So if we want to classify a Male, 3rd, 15-17 we start at the root of the tree which is Gender. Since the instance is Male we go to the left node of Gender which is Year Level. And finally since the instance is in 3rd Year we follow that branch and end up with a leaf node with classification Yes. Hence, the classification for Male, 3rd, 15-17 is Yes.

IMPLEMENTATION

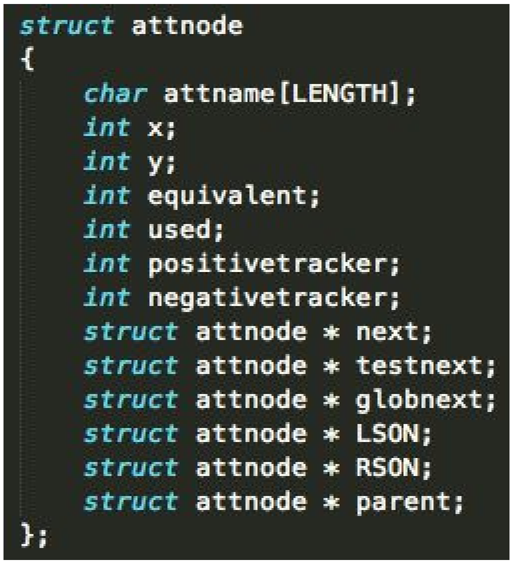
For the implementation of decision tree learning, we used the ID3 algorithm.

Storing data from the input files:



```
/*
    Format of initialize.txt:
        2                   // Number of attributes
        Outlook             // 1st Attribute
        3                   // Number of values for 1st attribute
        Sunny               // 1st value of 1st attribute
        Overcast            // 2nd value of 1st attribute
        Rain                // 3rd value of 1st attribute
        Temperature         // 2nd Attribute
        3                   // Number of values for 2nd attribute
        Hot                 // 1st value of 2nd attribute
        Mild                // 2nd value of 2nd attribute
        Cool                // 3rd value of 2nd attribute
*/
```

Shown above is the format for the initialize.txt file. The first line states the number of attributes that each training set contains. It is then followed by each attribute's name followed by each attribute's corresponding number of values. The succeeding lines contain the respective values of each attribute.



```
struct attnode
{
    char attname[LENGTH];
    int x;
    int y;
    int equivalent;
    int used;
    int positivetracker;
    int negativetracker;
    struct attnode * next;
    struct attnode * testnext;
    struct attnode * globnext;
    struct attnode * LSON;
    struct attnode * RSON;
    struct attnode * parent;
};
```

A linked list called **train** is used to store the data from initialize.txt. A node in the linked list corresponds to either an attribute (i.e. Outlook, Temperature), value of the attribute (i.e.

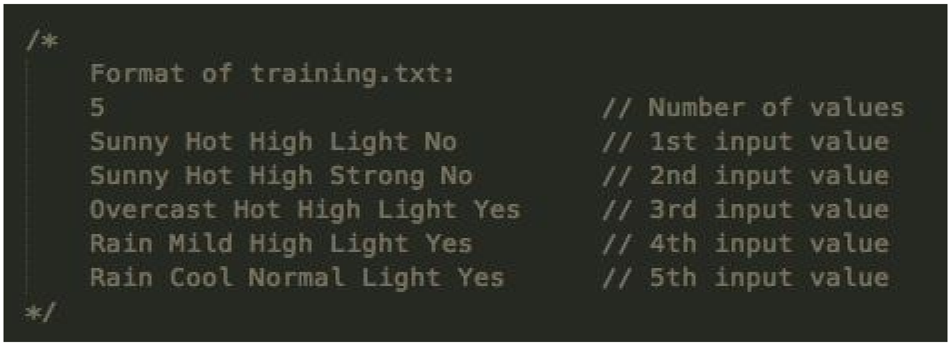
Sunny, Overcast, Rain for Outlook), or classification (i.e. Yes, No) from the text file. Each node has an attribute name, x-coordinate, y-coordinate, corresponding integer value, indicator if it's used, number of positive values, and number of negative values. Aside from these attributes, the node also has a next, testnext, globnext, LSON, RSON, and parent. *next* is used for the **train** linked list. *testnext* is used for the **test** linked list. *glob* is used for the global **glob** linked list. *LSON* and *RSON* is used for the binary decision tree. *parent* is used as a trace for the general decision tree.

Each node's distinct equivalent integer value is assigned as follows:

1. Attributes have negative integer values that decrement for each attribute.
2. Values have positive, non-zero integer values that increment for each attribute value.

For example, the attributes and values shown in the example for initialize.txt above will have the following equivalent integer values:

Outlook: -1
Sunny: 1
Overcast: 2
Rain: 3
Temperature: -2
Hot: 4
Mild: 5
Cool: 6



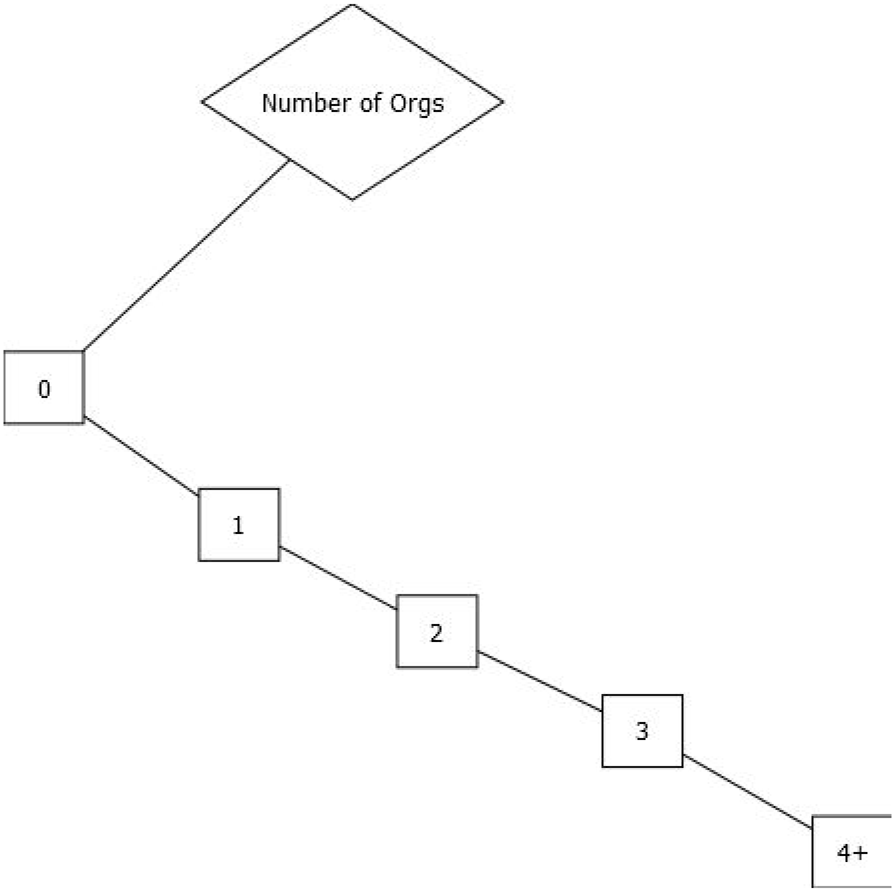
```
/*
    Format of training.txt:
    5                               // Number of values
    Sunny Hot High Light No         // 1st input value
    Sunny Hot High Strong No        // 2nd input value
    Overcast Hot High Light Yes     // 3rd input value
    Rain Mild High Light Yes        // 4th input value
    Rain Cool Normal Light Yes      // 5th input value
*/
```

Shown above is the format for the training.txt file. The first line states the total number of training instances to be inputted. The following lines contain the actual training set.

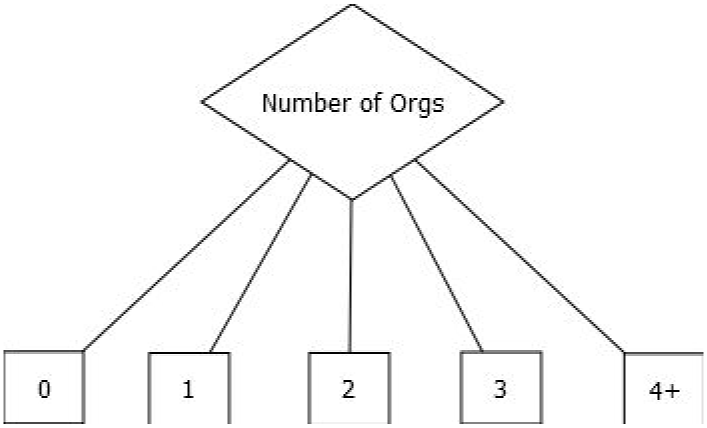
A 2D array is used to store the data from training.txt. This array contains almost the same contents as the training.txt file except that the values are translated into their corresponding integer equivalents. For example, if Sunny=1, Hot=4, High=7, Light=9, Strong=10, and No=11, then the first entry/row of the 2D array is {1, 4, 7, 9, 11} and the second row is {1, 4, 7, 10, 11}.

Construction of the decision tree:

We used a binary tree using a linked list for our decision tree. While the tree was still empty, the information gain of each attribute based on the given dataset was calculated*. The attribute with the highest information gain would be chosen as the decision tree's root. Connected as the root's left son is its first corresponding value, and the remaining values will be connected as right sons of the previous values. All of the values' parent pointers will point to the root. All of the root's values are added to a separate linked list **glob** to implement breadth first construction of the tree.



Shown above is a representation of the binary decision tree. Here, Number of Orgs has LSON of 0, and the rest of the attributes are set as RSONs of the previous nodes.



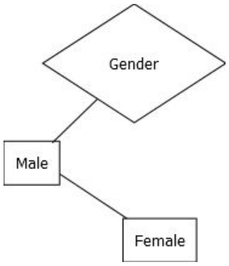
Shown above is a decision tree representation using the "parent" pointer. This shows that 0, 1, 2, 3, and 4+ have Number of Orgs as their parent, making it easy to trace the attribute where they came from.

We will traverse the linked list **glob**, then an attribute or a classification will be chosen to be appended to the current node to further construct the tree following these cases:

Case 1: If the current node does not have a positive example (i.e. [0+, 3-]) in the training data with respect to its parent node, a "no" classification node will be appended to it. This "no" node is a leaf node.

Case 2: If the current node does not have a negative example (i.e. [2+, 0-]) in the training data with respect to its parent node, a "yes" classification node will be appended to it. This "yes" node is a leaf node.

Case 3: If the current node has both positive and negative examples in the training data with respect to its parent node, every attribute not yet included in the tree will have its information gain* computed. The attribute with the highest information gain will be appended as the left son of the current node. Its first value will be appended as its left son, and the remaining values will be appended as following right sons of the previous values. For example, if "Gender" has the highest information gain, it will be the left son of the current node. "Male" will be appended as the right son of "Gender" while "Female" will be the right son of "Male".



Case 4: If the current node has both positive and negative examples in the training data but there are no remaining unused attributes to append, a yes or no classification node will be appended to it.

After appending the assigned attribute or classification node to the tree, its attribute values (if any) will be appended to the linked list **glob** and then the same procedure will be done to the next node in the **glob** linked list**.** This will be repeated until the whole linked list has been traversed and thus, the decision tree will be fully constructed.

* To calculate the information gain of an attribute A with respect to a set S:

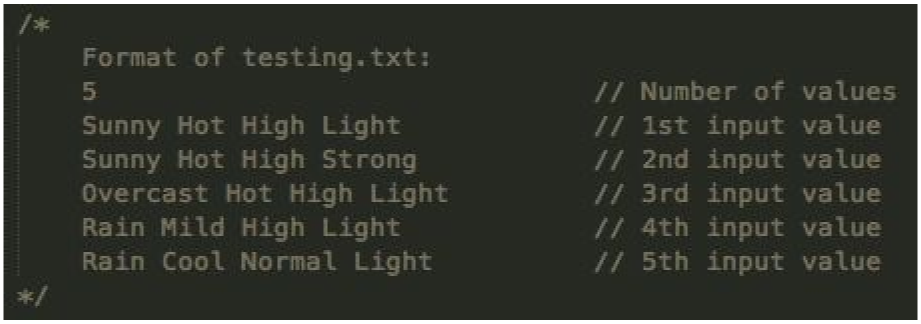
$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

* To calculate the entropy of an attribute A with respect to a set S:

$$Entropy(S, A) = -p_+log_2(p_+) - p_-log_2(p_-)$$

where p_+ is the proportion of positive examples and p_- is the proportion of negative examples with respect to the set S

Determining classification:



```
/*
    Format of testing.txt:
    5                               // Number of values
    Sunny Hot High Light            // 1st input value
    Sunny Hot High Strong           // 2nd input value
    Overcast Hot High Light         // 3rd input value
    Rain Mild High Light            // 4th input value
    Rain Cool Normal Light          // 5th input value
*/
```

Shown above is the format for the testing.txt file. The first line states the total number of test cases to be inputted. The following lines contain the actual test set.

An array is used to store the data from testing.txt. This array stores the equivalent integer data of one instance from the test data. The corresponding attributes, values, and classification of the instance are then stored as individual nodes in a linked list **test** with the same structure as **attnode**.

To classify a test case, we will start by getting the y-coordinate of the decision tree's root (decurr). A pointer decurr will be used to keep track of the current node in the decision tree. A pointer curr will be used to keep track of the current node in the linked list **test**. Traversing the linked list **test**, curr will be pointed to the node of the same y-coordinate as decurr. Curr will be compared to decurr's left son.

1. If the equivalent integer values of both decurr and curr are equal, decurr will be curr's left son.
2. If the equivalent integer values of decurr and curr are not equal, decurr will be curr's right son.

This will be repeated until a leaf node has been reached. The leaf node reached will contain the result classification for the inputted test instance.

TOPIC (JOINING AN ORGANIZATION)

The topic we have chosen for our MP2 is "Joining an Organization." The attributes involved are:

Attributes: Gender, Year Level, Free Time, Social Skills, Number of Orgs, Units

And each of these attributes have the following values:

Gender: Male, Female

Year Level: 1st, 2nd, 3rd, 4th, Others

Free Time in a Weekday (in hours): 0-1, 1-2, 2-4, 5+

Social Skills: Ambivert, Introvert, Extrovert

Number of Orgs: 0, 1, 2, 3, 4+

Number of Units: Underload, 15-17, 18-21, Overload

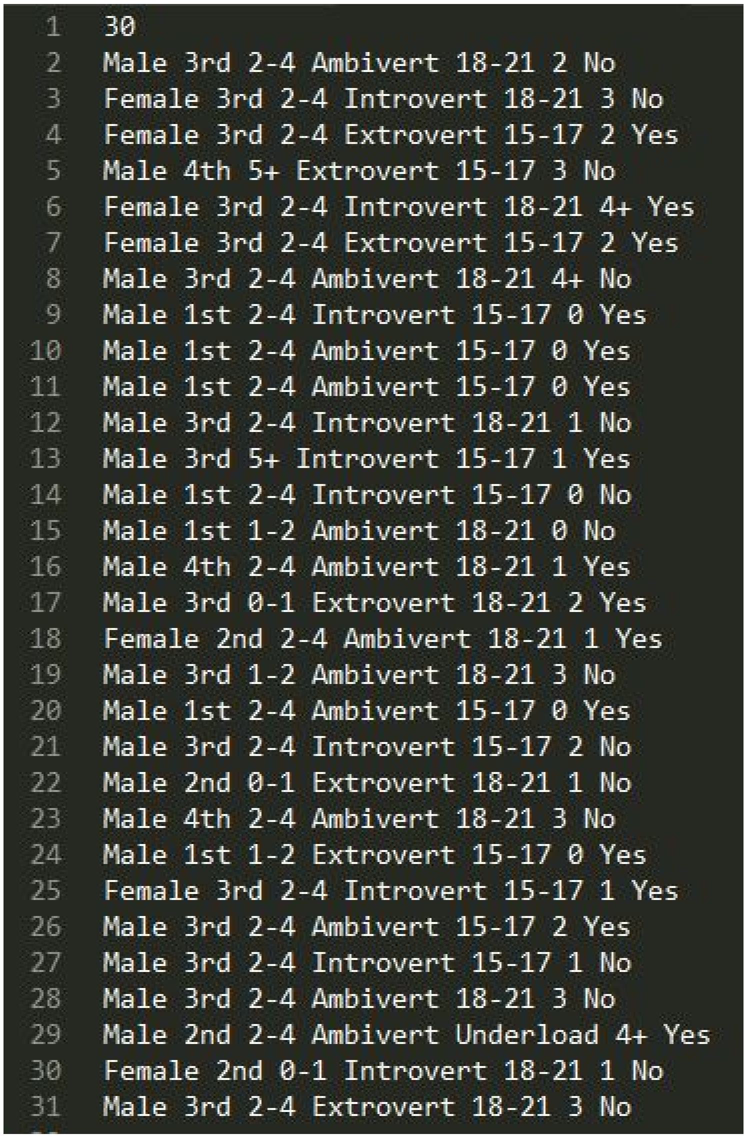
We chose this topic because we wanted to determine whether other people would (still) want to join an organization given certain factors such as their gender, year level, free time, social skills, number of organizations, and number of units for the academic semester. For example, we wanted to see if males were more inclined to join organizations than females, or if 1st Year students would be more likely to join an/another org than a 4th Year student because the 4th Year student would be graduating soon and will spend less time in the organization so he/she may not deem it worthy of his/her time anymore to consider joining an org because his/her college life is nearing to an end while a freshman has 4 more years to look forward to in college to spend being a part of an organization. We also wanted to see if free time was relevant in considering whether to join an org because if one has less free time, then he/she has less time to dedicate to an org. For social skills, we also wanted to see if introverts were less likely to join due to their shy nature or if they would be more likely to join in order to step out of their shell and learn to interact more with other people. Or maybe the current number of organizations a person has may also play a role in whether a person would still want to join an org because if he/she has 7 orgs, he/she already has a lot of responsibilities so he/she may not want to join another org, or if a person has no orgs, then we wanted to see whether he/she would prefer to remain "orgless" and focus on his/her academics than join an organization or if he/she would be more likely to join an organization because it's something new to them. We also chose units as a factor because we wanted to determine whether this affects their decision since if a person has a heavy academic load, then he/she may not be able to balance his/her time so he/she cannot handle the weight or another organization. It is for these reasons that we chose this topic, and the specific attributes enumerated above.

EXPERIMENT 1

The following involves n-fold cross-validation with n=5 wherein 30 samples were used for the training sets and 20 samples were used for the test sets.

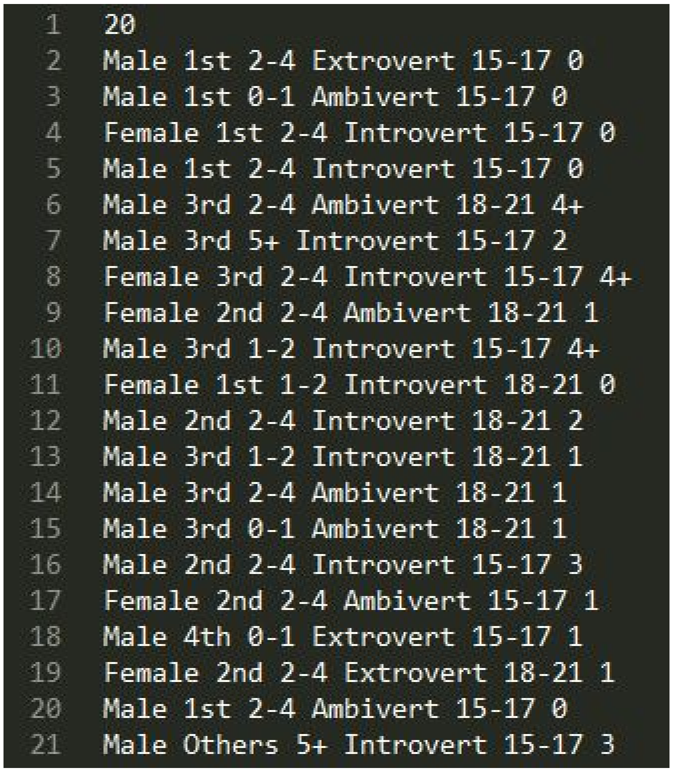
n=1:

Training Set



```
30
Male 3rd 2-4 Ambivert 18-21 2 No
Female 3rd 2-4 Introvert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 4th 5+ Extrovert 15-17 3 No
Female 3rd 2-4 Introvert 18-21 4+ Yes
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 1st 2-4 Introvert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 2-4 Introvert 18-21 1 No
Male 3rd 5+ Introvert 15-17 1 Yes
Male 1st 2-4 Introvert 15-17 0 No
Male 1st 1-2 Ambivert 18-21 0 No
Male 4th 2-4 Ambivert 18-21 1 Yes
Male 3rd 0-1 Extrovert 18-21 2 Yes
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 3rd 1-2 Ambivert 18-21 3 No
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 2-4 Introvert 15-17 2 No
Male 2nd 0-1 Extrovert 18-21 1 No
Male 4th 2-4 Ambivert 18-21 3 No
Male 1st 1-2 Extrovert 15-17 0 Yes
Female 3rd 2-4 Introvert 15-17 1 Yes
Male 3rd 2-4 Ambivert 15-17 2 Yes
Male 3rd 2-4 Introvert 15-17 1 No
Male 3rd 2-4 Ambivert 18-21 3 No
Male 2nd 2-4 Ambivert Underload 4+ Yes
Female 2nd 0-1 Introvert 18-21 1 No
Male 3rd 2-4 Extrovert 18-21 3 No
```

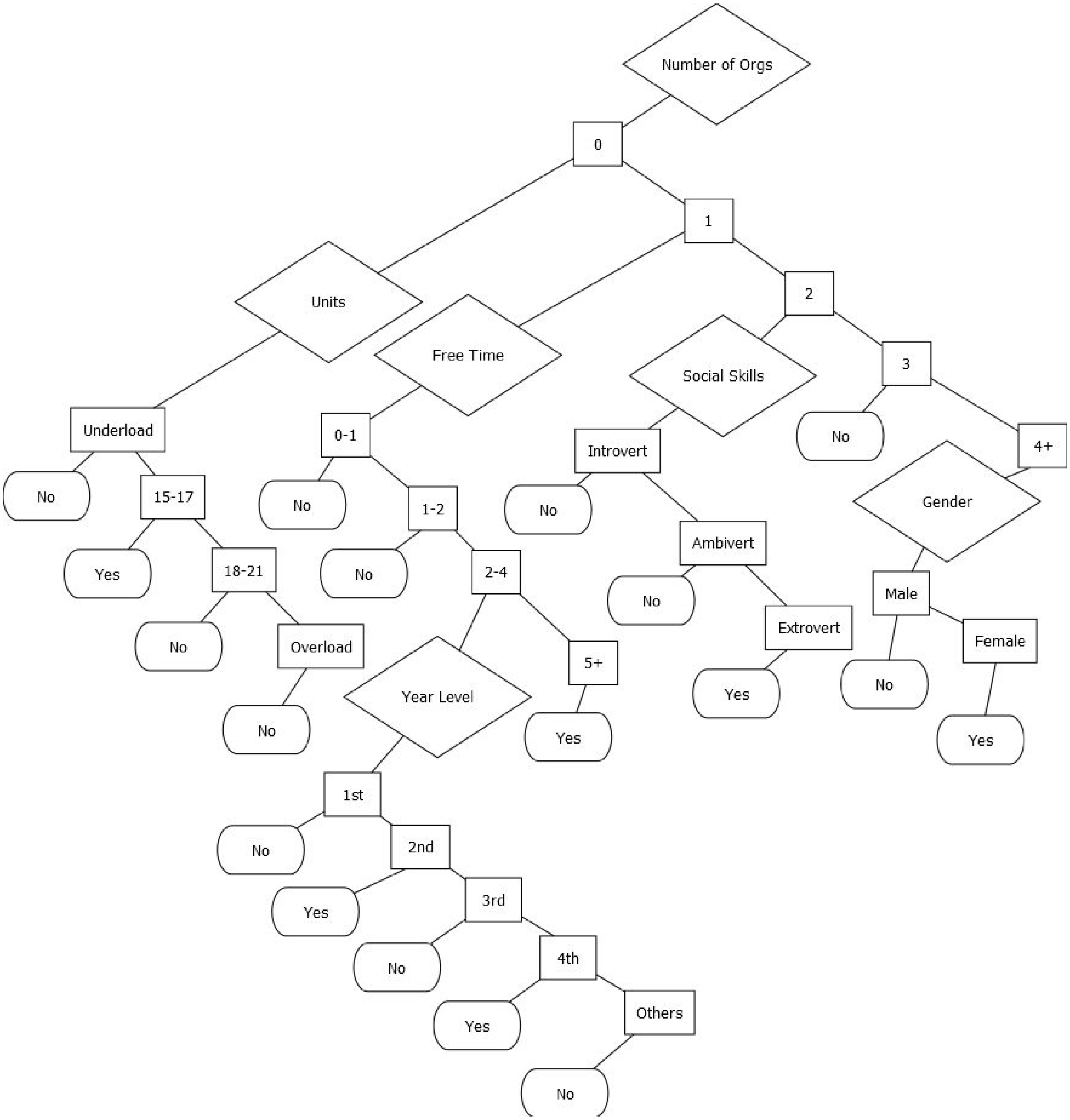
Test Set



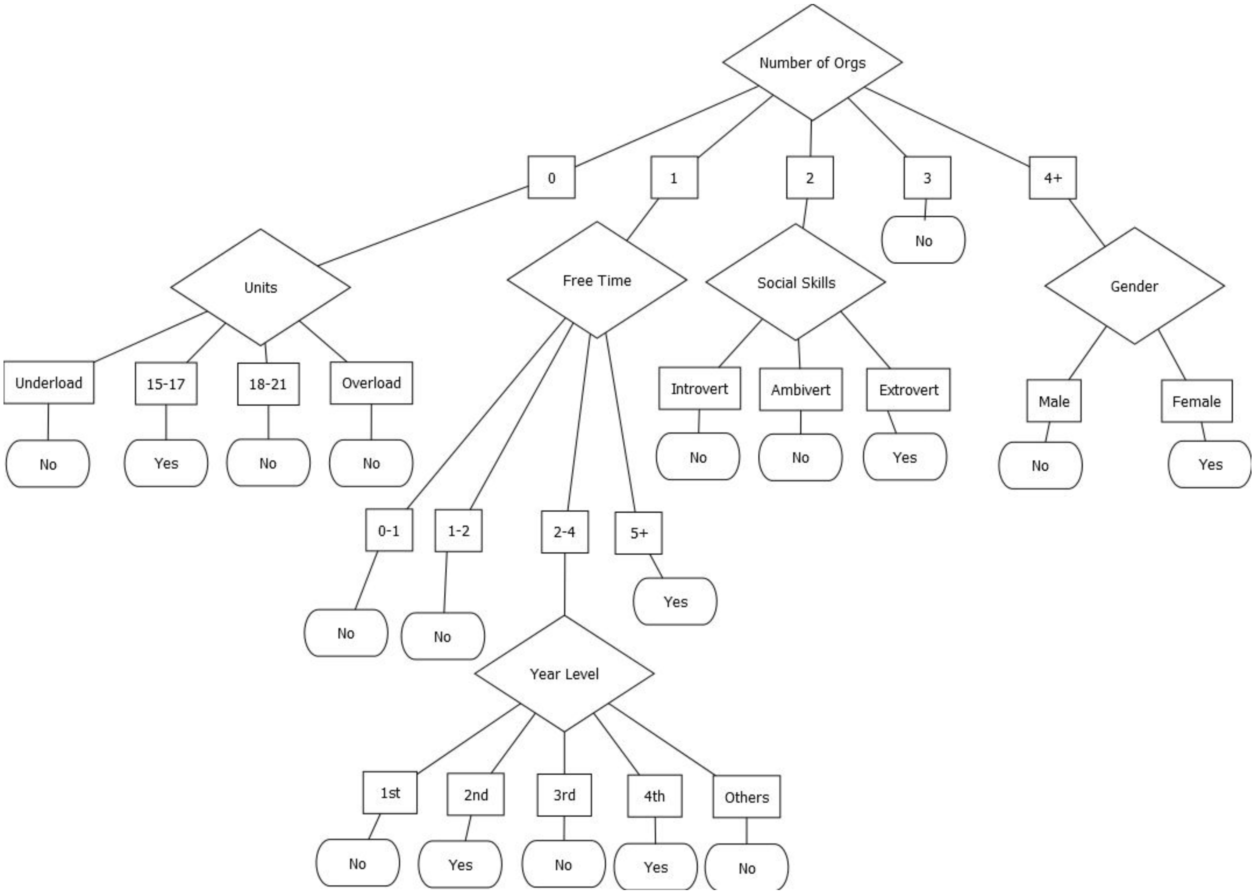
```
20
Male 1st 2-4 Extrovert 15-17 0
Male 1st 0-1 Ambivert 15-17 0
Female 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Male 3rd 2-4 Ambivert 18-21 4+
Male 3rd 5+ Introvert 15-17 2
Female 3rd 2-4 Introvert 15-17 4+
Female 2nd 2-4 Ambivert 18-21 1
Male 3rd 1-2 Introvert 15-17 4+
Female 1st 1-2 Introvert 18-21 0
Male 2nd 2-4 Introvert 18-21 2
Male 3rd 1-2 Introvert 18-21 1
Male 3rd 2-4 Ambivert 18-21 1
Male 3rd 0-1 Ambivert 18-21 1
Male 2nd 2-4 Introvert 15-17 3
Female 2nd 2-4 Ambivert 15-17 1
Male 4th 0-1 Extrovert 15-17 1
Female 2nd 2-4 Extrovert 18-21 1
Male 1st 2-4 Ambivert 15-17 0
Male Others 5+ Introvert 15-17 3
```

Decision Tree (Binary Implementation)

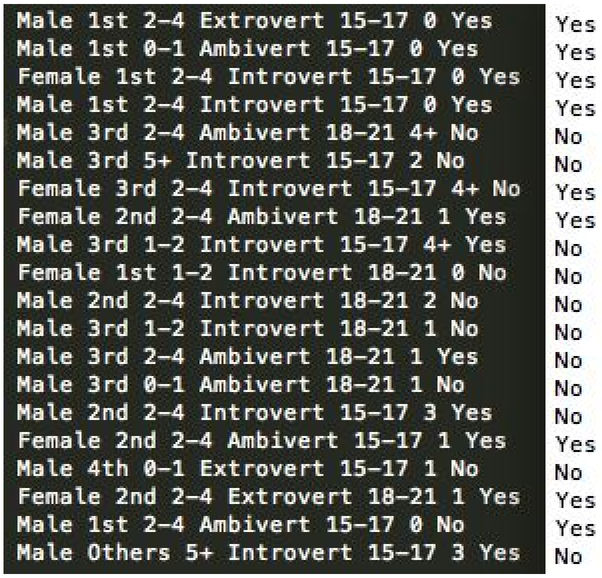
Just to show how the tree is actually implemented in the program, below is the diagram of the decision tree. On the next page is the human-readable decision tree. For the next experiments/cases, we would only display the human-readable/general decision tree instead of the binary tree.



Decision Tree



Results (Shows expected and actual outcome) - The one on the left shows the data from the survey while the Yes/No on the right (in white) shows the results from the program.



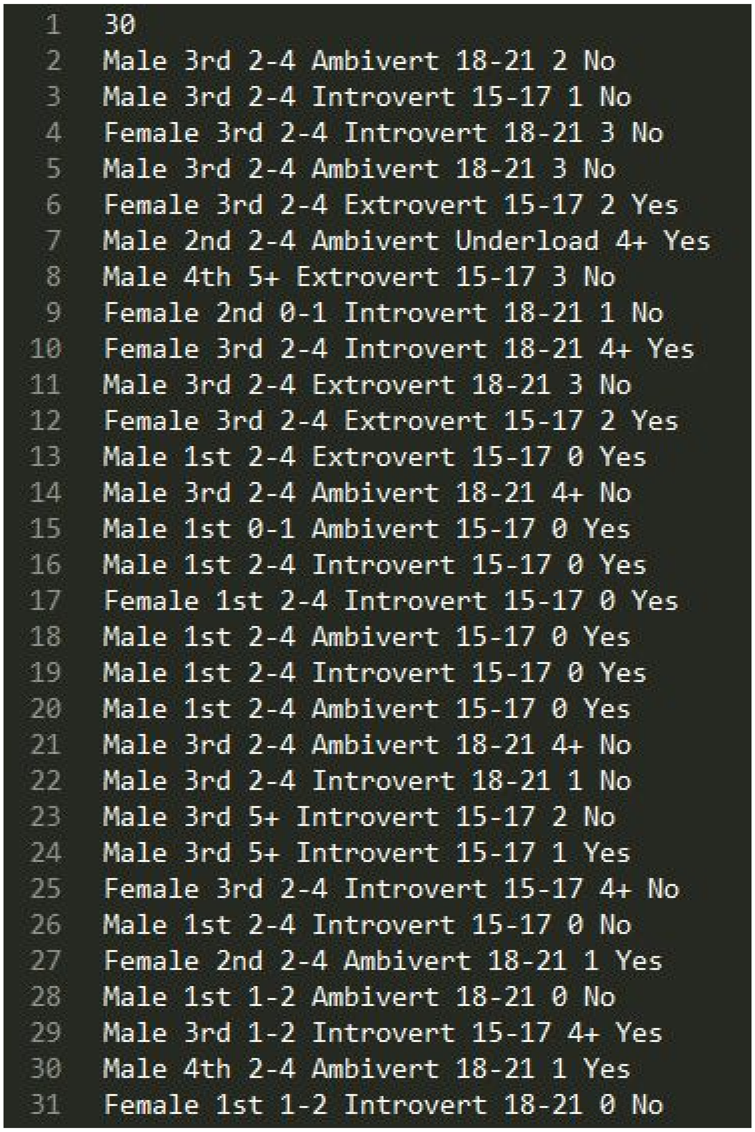
```
Male 1st 2-4 Extrovert 15-17 0 Yes       Yes
Male 1st 0-1 Ambivert 15-17 0 Yes        Yes
Female 1st 2-4 Introvert 15-17 0 Yes     Yes
Male 1st 2-4 Introvert 15-17 0 Yes       Yes
Male 3rd 2-4 Ambivert 18-21 4+ No        No
Male 3rd 5+ Introvert 15-17 2 No         No
Female 3rd 2-4 Introvert 15-17 4+ No     Yes
Female 2nd 2-4 Ambivert 18-21 1 Yes      Yes
Male 3rd 1-2 Introvert 15-17 4+ Yes      No
Female 1st 1-2 Introvert 18-21 0 No      No
Male 2nd 2-4 Introvert 18-21 2 No        No
Male 3rd 1-2 Introvert 18-21 1 No        No
Male 3rd 2-4 Ambivert 18-21 1 Yes        No
Male 3rd 0-1 Ambivert 18-21 1 No         No
Male 2nd 2-4 Introvert 15-17 3 Yes       No
Female 2nd 2-4 Ambivert 15-17 1 Yes      Yes
Male 4th 0-1 Extrovert 15-17 1 No        No
Female 2nd 2-4 Extrovert 18-21 1 Yes     Yes
Male 1st 2-4 Ambivert 15-17 0 No         Yes
Male Others 5+ Introvert 15-17 3 Yes     No
```

14 out of 20 test data were correctly classified.

Accuracy: 70%

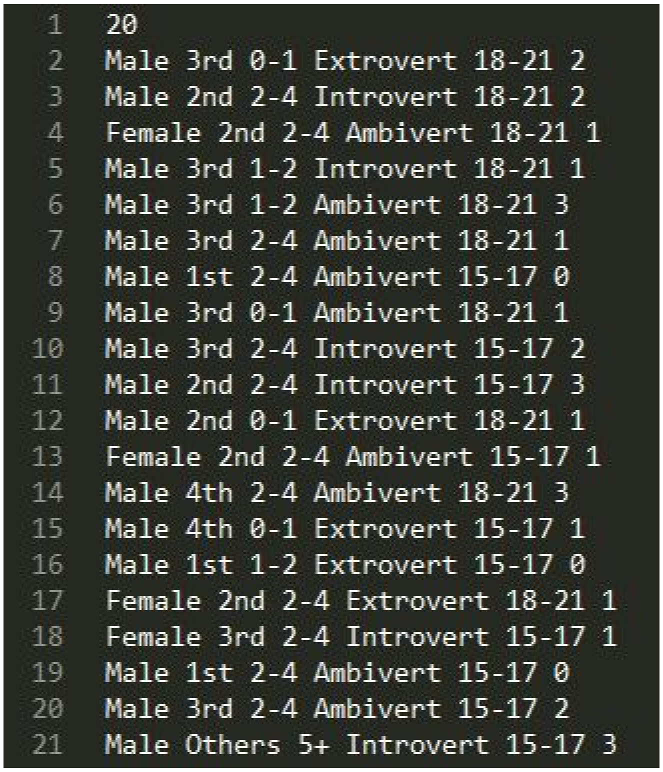
n=2:

Training Set



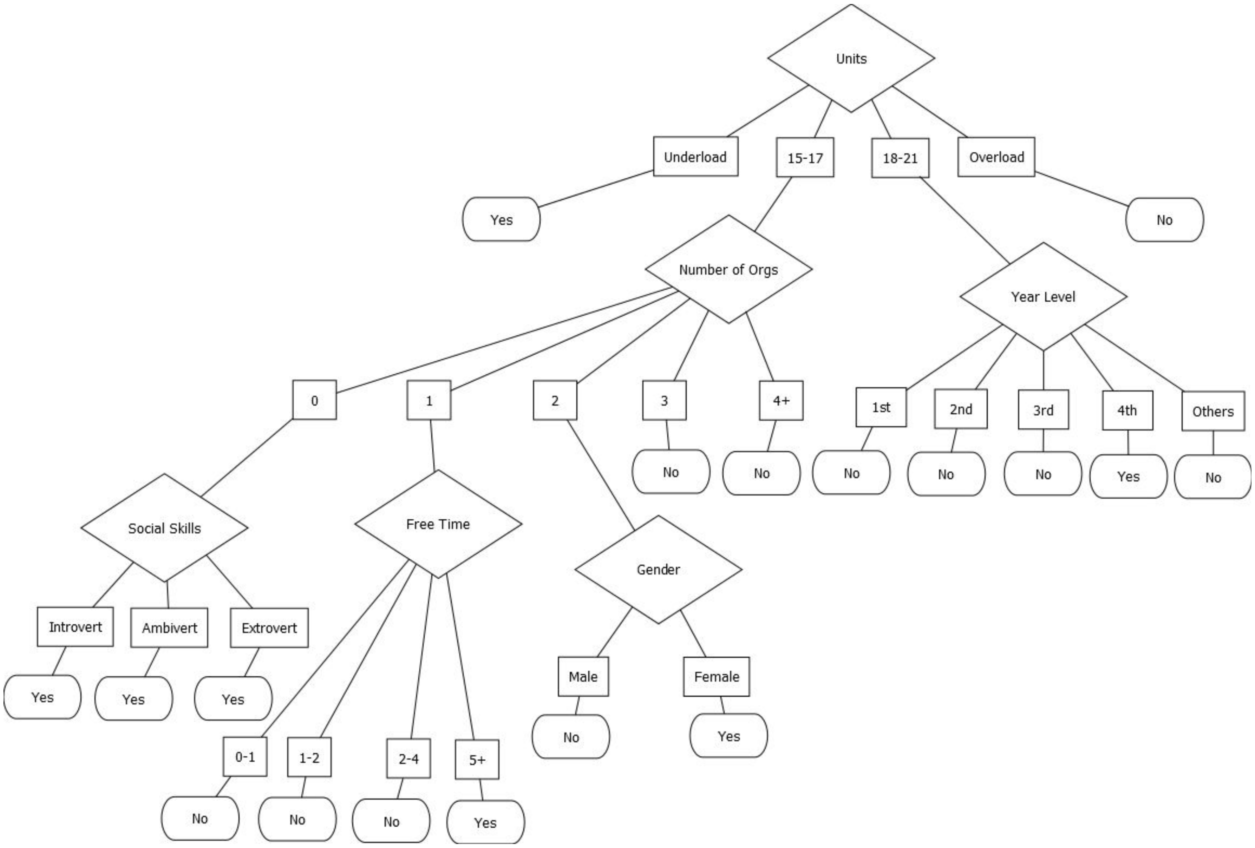
```
30
Male 3rd 2-4 Ambivert 18-21 2 No
Male 3rd 2-4 Introvert 15-17 1 No
Female 3rd 2-4 Introvert 18-21 3 No
Male 3rd 2-4 Ambivert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 2nd 2-4 Ambivert Underload 4+ Yes
Male 4th 5+ Extrovert 15-17 3 No
Female 2nd 0-1 Introvert 18-21 1 No
Female 3rd 2-4 Introvert 18-21 4+ Yes
Male 3rd 2-4 Extrovert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 1st 2-4 Extrovert 15-17 0 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 1st 0-1 Ambivert 15-17 0 Yes
Male 1st 2-4 Introvert 15-17 0 Yes
Female 1st 2-4 Introvert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 1st 2-4 Introvert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 3rd 2-4 Introvert 18-21 1 No
Male 3rd 5+ Introvert 15-17 2 No
Male 3rd 5+ Introvert 15-17 1 Yes
Female 3rd 2-4 Introvert 15-17 4+ No
Male 1st 2-4 Introvert 15-17 0 No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 1st 1-2 Ambivert 18-21 0 No
Male 3rd 1-2 Introvert 15-17 4+ Yes
Male 4th 2-4 Ambivert 18-21 1 Yes
Female 1st 1-2 Introvert 18-21 0 No
```

Test Set

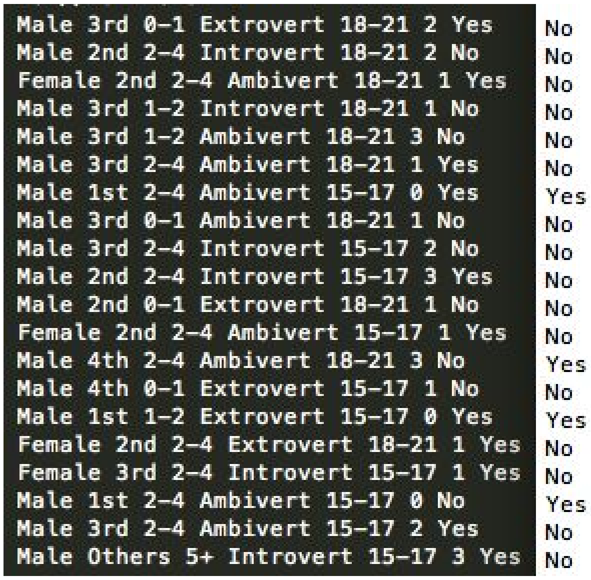


```
20
Male 3rd 0-1 Extrovert 18-21 2
Male 2nd 2-4 Introvert 18-21 2
Female 2nd 2-4 Ambivert 18-21 1
Male 3rd 1-2 Introvert 18-21 1
Male 3rd 1-2 Ambivert 18-21 3
Male 3rd 2-4 Ambivert 18-21 1
Male 1st 2-4 Ambivert 15-17 0
Male 3rd 0-1 Ambivert 18-21 1
Male 3rd 2-4 Introvert 15-17 2
Male 2nd 2-4 Introvert 15-17 3
Male 2nd 0-1 Extrovert 18-21 1
Female 2nd 2-4 Ambivert 15-17 1
Male 4th 2-4 Ambivert 18-21 3
Male 4th 0-1 Extrovert 15-17 1
Male 1st 1-2 Extrovert 15-17 0
Female 2nd 2-4 Extrovert 18-21 1
Female 3rd 2-4 Introvert 15-17 1
Male 1st 2-4 Ambivert 15-17 0
Male 3rd 2-4 Ambivert 15-17 2
Male Others 5+ Introvert 15-17 3
```

Decision Tree



Results



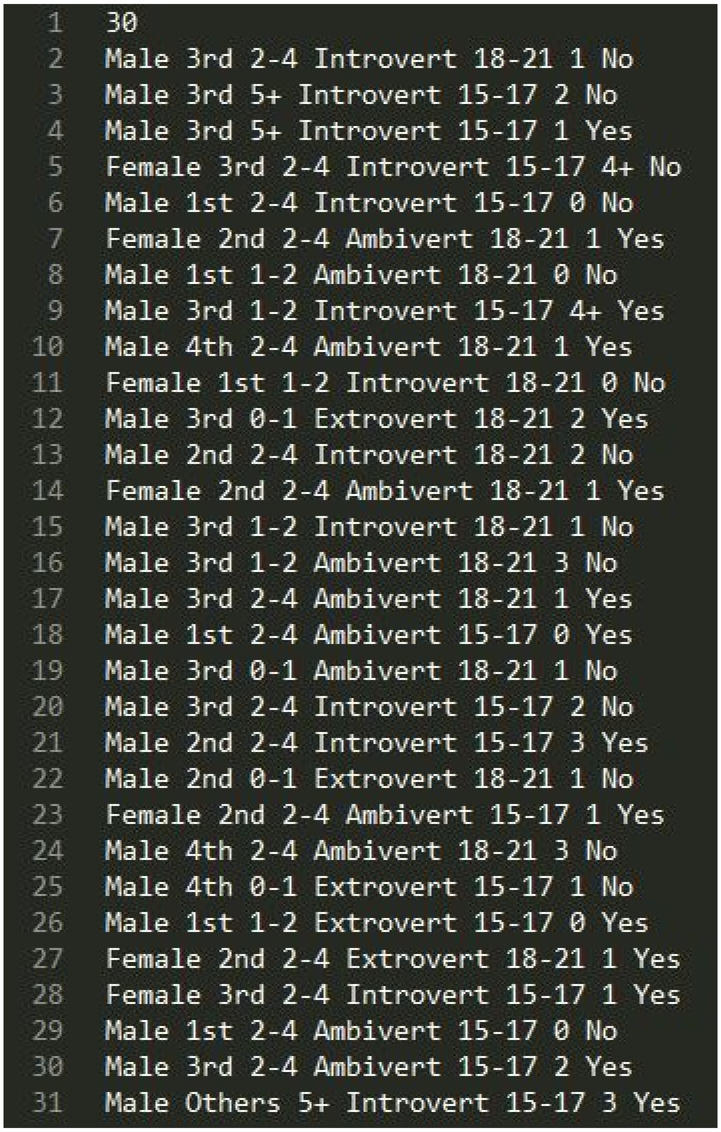
```
Male 3rd 0-1 Extrovert 18-21 2 Yes      No
Male 2nd 2-4 Introvert 18-21 2 No       No
Female 2nd 2-4 Ambivert 18-21 1 Yes     No
Male 3rd 1-2 Introvert 18-21 1 No       No
Male 3rd 1-2 Ambivert 18-21 3 No        No
Male 3rd 2-4 Ambivert 18-21 1 Yes       No
Male 1st 2-4 Ambivert 15-17 0 Yes       Yes
Male 3rd 0-1 Ambivert 18-21 1 No        No
Male 3rd 2-4 Introvert 15-17 2 No       No
Male 2nd 2-4 Introvert 15-17 3 Yes      No
Male 2nd 0-1 Extrovert 18-21 1 No       No
Female 2nd 2-4 Ambivert 15-17 1 Yes     No
Male 4th 2-4 Ambivert 18-21 3 No        Yes
Male 4th 0-1 Extrovert 15-17 1 No       No
Male 1st 1-2 Extrovert 15-17 0 Yes      Yes
Female 2nd 2-4 Extrovert 18-21 1 Yes    No
Female 3rd 2-4 Introvert 15-17 1 Yes    No
Male 1st 2-4 Ambivert 15-17 0 No        Yes
Male 3rd 2-4 Ambivert 15-17 2 Yes       No
Male Others 5+ Introvert 15-17 3 Yes    No
```

9 out of 20 test data were correctly classified.

Accuracy: 45%

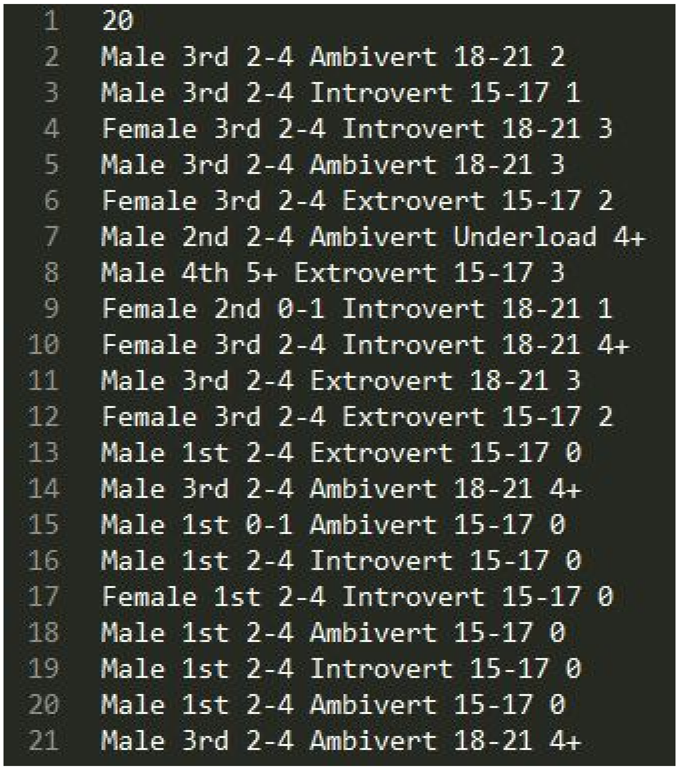
n=3:

Training Set



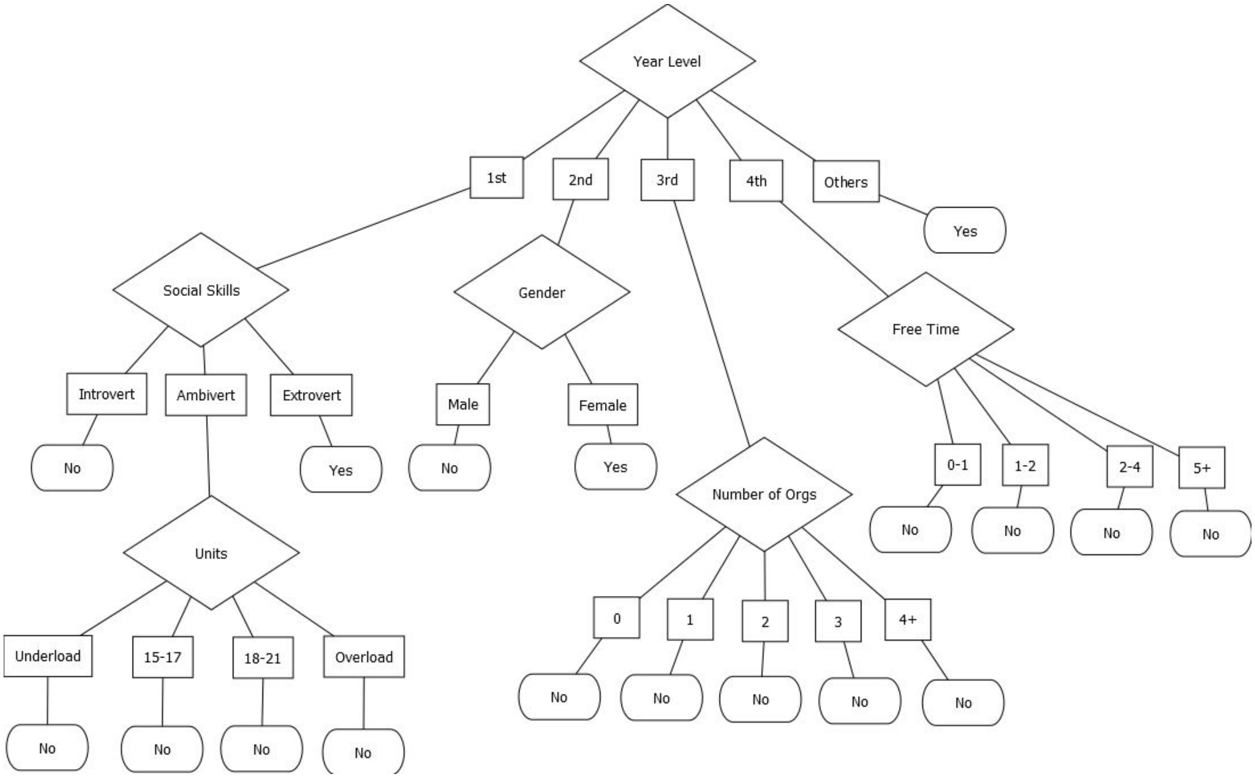
```
30
Male 3rd 2-4 Introvert 18-21 1 No
Male 3rd 5+ Introvert 15-17 2 No
Male 3rd 5+ Introvert 15-17 1 Yes
Female 3rd 2-4 Introvert 15-17 4+ No
Male 1st 2-4 Introvert 15-17 0 No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 1st 1-2 Ambivert 18-21 0 No
Male 3rd 1-2 Introvert 15-17 4+ Yes
Male 4th 2-4 Ambivert 18-21 1 Yes
Female 1st 1-2 Introvert 18-21 0 No
Male 3rd 0-1 Extrovert 18-21 2 Yes
Male 2nd 2-4 Introvert 18-21 2 No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 3rd 1-2 Introvert 18-21 1 No
Male 3rd 1-2 Ambivert 18-21 3 No
Male 3rd 2-4 Ambivert 18-21 1 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 0-1 Ambivert 18-21 1 No
Male 3rd 2-4 Introvert 15-17 2 No
Male 2nd 2-4 Introvert 15-17 3 Yes
Male 2nd 0-1 Extrovert 18-21 1 No
Female 2nd 2-4 Ambivert 15-17 1 Yes
Male 4th 2-4 Ambivert 18-21 3 No
Male 4th 0-1 Extrovert 15-17 1 No
Male 1st 1-2 Extrovert 15-17 0 Yes
Female 2nd 2-4 Extrovert 18-21 1 Yes
Female 3rd 2-4 Introvert 15-17 1 Yes
Male 1st 2-4 Ambivert 15-17 0 No
Male 3rd 2-4 Ambivert 15-17 2 Yes
Male Others 5+ Introvert 15-17 3 Yes
```

Test Set

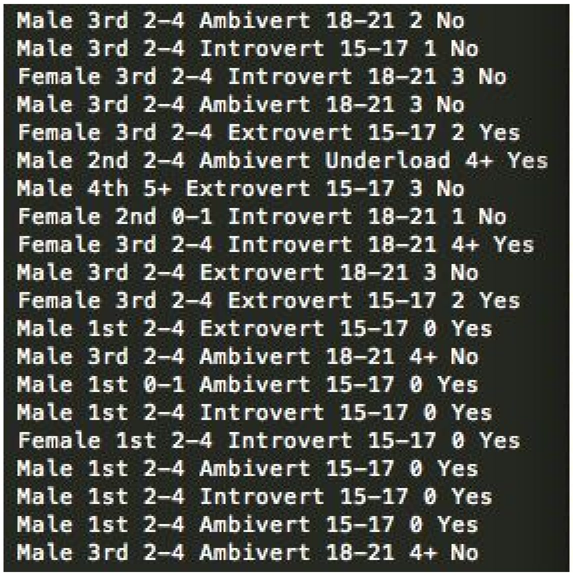


```
20
Male 3rd 2-4 Ambivert 18-21 2
Male 3rd 2-4 Introvert 15-17 1
Female 3rd 2-4 Introvert 18-21 3
Male 3rd 2-4 Ambivert 18-21 3
Female 3rd 2-4 Extrovert 15-17 2
Male 2nd 2-4 Ambivert Underload 4+
Male 4th 5+ Extrovert 15-17 3
Female 2nd 0-1 Introvert 18-21 1
Female 3rd 2-4 Introvert 18-21 4+
Male 3rd 2-4 Extrovert 18-21 3
Female 3rd 2-4 Extrovert 15-17 2
Male 1st 2-4 Extrovert 15-17 0
Male 3rd 2-4 Ambivert 18-21 4+
Male 1st 0-1 Ambivert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Female 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Ambivert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Ambivert 15-17 0
Male 3rd 2-4 Ambivert 18-21 4+
```

Decision Tree



Results

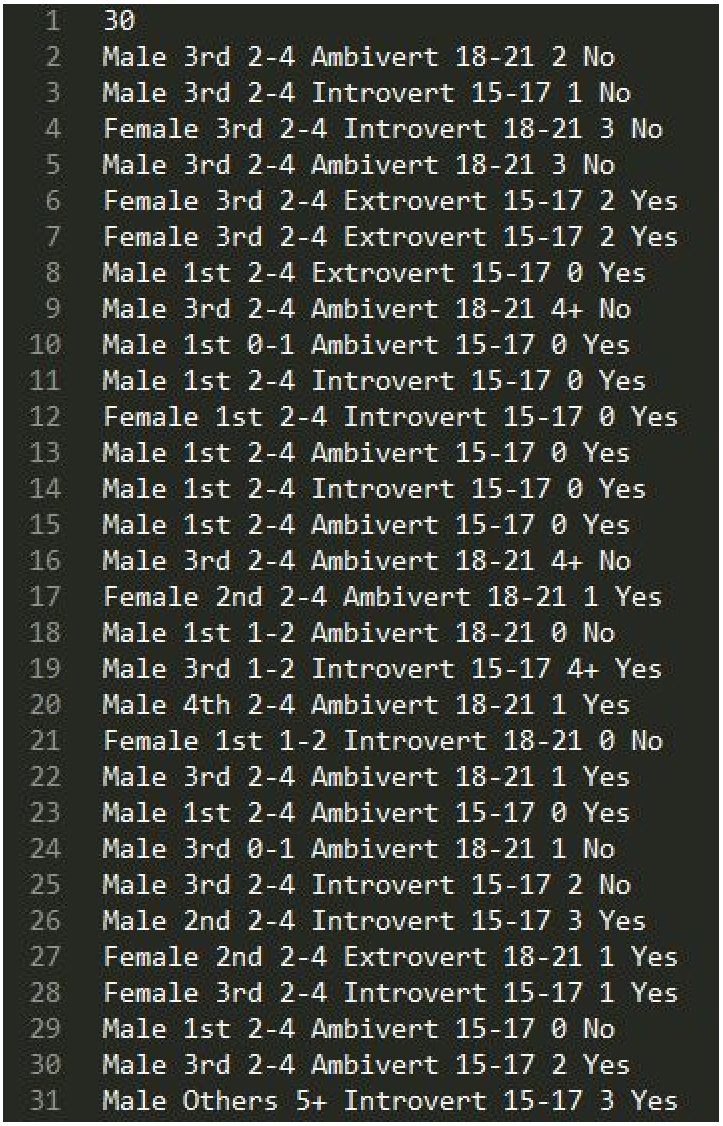


```
Male 3rd 2-4 Ambivert 18-21 2 No              No
Male 3rd 2-4 Introvert 15-17 1 No             No
Female 3rd 2-4 Introvert 18-21 3 No           No
Male 3rd 2-4 Ambivert 18-21 3 No              No
Female 3rd 2-4 Extrovert 15-17 2 Yes          No
Male 2nd 2-4 Ambivert Underload 4+ Yes        No
Male 4th 5+ Extrovert 15-17 3 No              No
Female 2nd 0-1 Introvert 18-21 1 No           Yes
Female 3rd 2-4 Introvert 18-21 4+ Yes         No
Male 3rd 2-4 Extrovert 18-21 3 No             No
Female 3rd 2-4 Extrovert 15-17 2 Yes          No
Male 1st 2-4 Extrovert 15-17 0 Yes            Yes
Male 3rd 2-4 Ambivert 18-21 4+ No             No
Male 1st 0-1 Ambivert 15-17 0 Yes             No
Male 1st 2-4 Introvert 15-17 0 Yes            No
Female 1st 2-4 Introvert 15-17 0 Yes          No
Male 1st 2-4 Ambivert 15-17 0 Yes             No
Male 1st 2-4 Introvert 15-17 0 Yes            No
Male 1st 2-4 Ambivert 15-17 0 Yes             No
Male 3rd 2-4 Ambivert 18-21 4+ No             No
```

9 out of 20 test data were correctly classified.
Accuracy: 45%

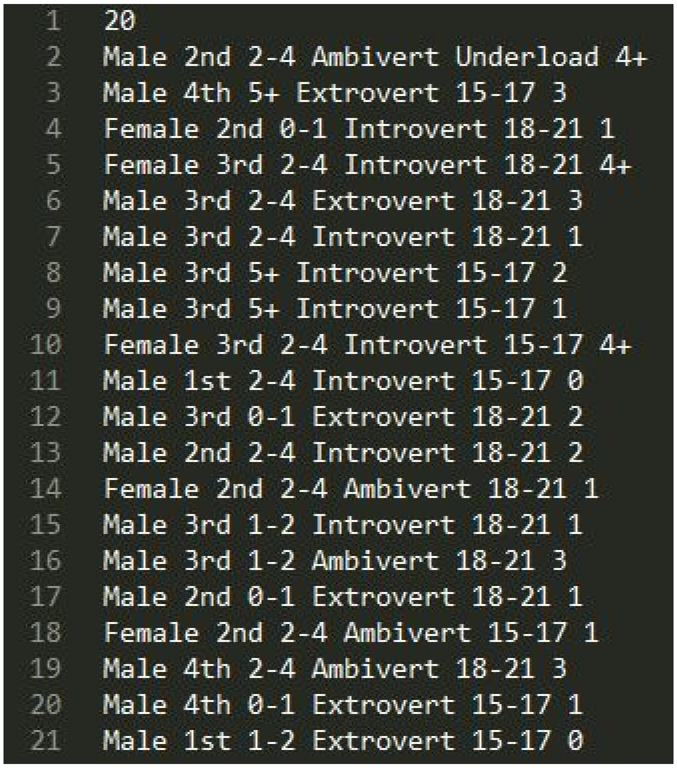
n=4:

Training Set



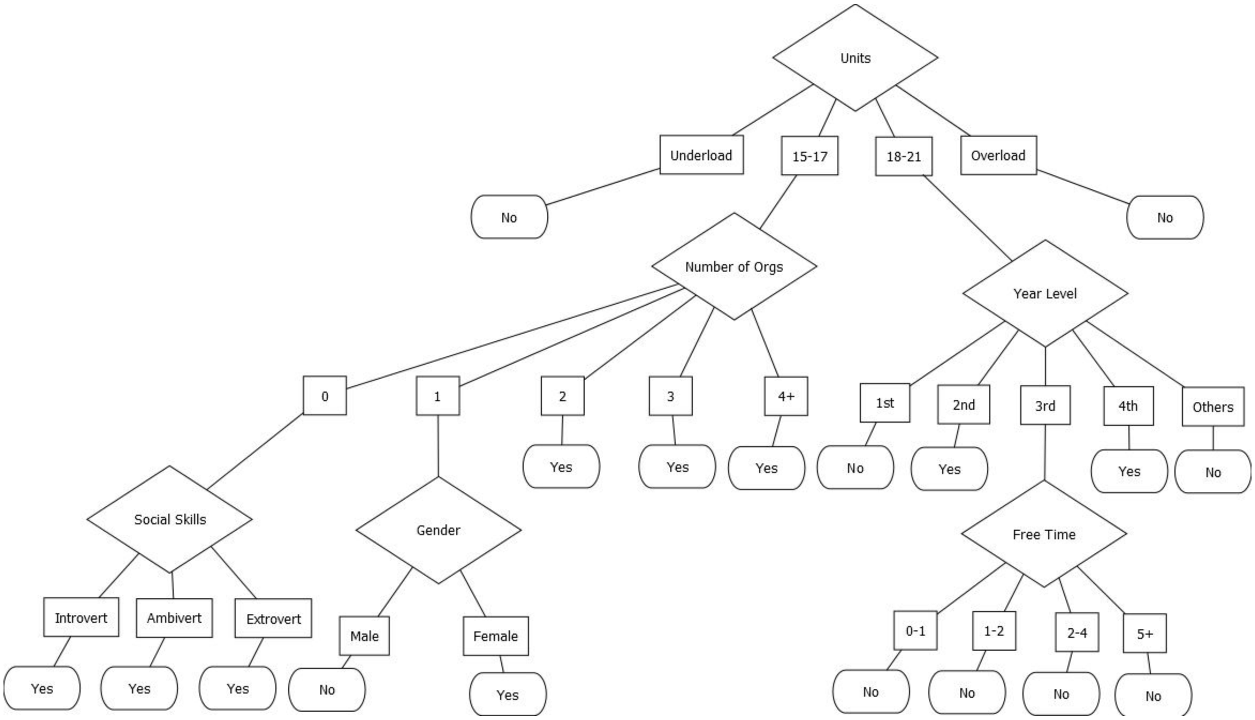
```
30
Male 3rd 2-4 Ambivert 18-21 2 No
Male 3rd 2-4 Introvert 15-17 1 No
Female 3rd 2-4 Introvert 18-21 3 No
Male 3rd 2-4 Ambivert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 1st 2-4 Extrovert 15-17 0 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 1st 0-1 Ambivert 15-17 0 Yes
Male 1st 2-4 Introvert 15-17 0 Yes
Female 1st 2-4 Introvert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 1st 2-4 Introvert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 1st 1-2 Ambivert 18-21 0 No
Male 3rd 1-2 Introvert 15-17 4+ Yes
Male 4th 2-4 Ambivert 18-21 1 Yes
Female 1st 1-2 Introvert 18-21 0 No
Male 3rd 2-4 Ambivert 18-21 1 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 0-1 Ambivert 18-21 1 No
Male 3rd 2-4 Introvert 15-17 2 No
Male 2nd 2-4 Introvert 15-17 3 Yes
Female 2nd 2-4 Extrovert 18-21 1 Yes
Female 3rd 2-4 Introvert 15-17 1 Yes
Male 1st 2-4 Ambivert 15-17 0 No
Male 3rd 2-4 Ambivert 15-17 2 Yes
Male Others 5+ Introvert 15-17 3 Yes
```

Test Set

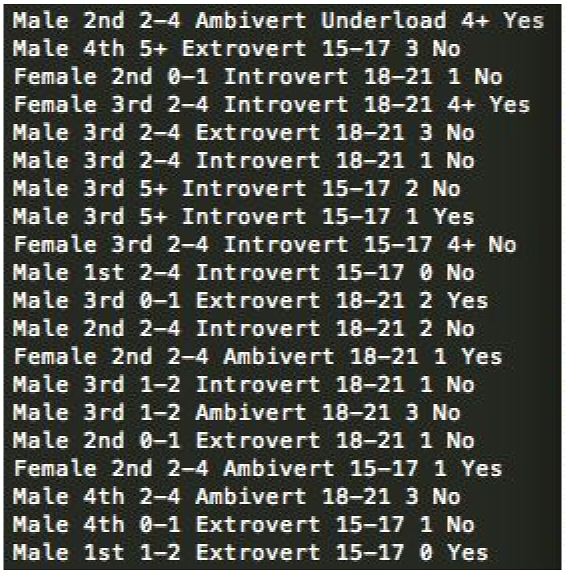


```
20
Male 2nd 2-4 Ambivert Underload 4+
Male 4th 5+ Extrovert 15-17 3
Female 2nd 0-1 Introvert 18-21 1
Female 3rd 2-4 Introvert 18-21 4+
Male 3rd 2-4 Extrovert 18-21 3
Male 3rd 2-4 Introvert 18-21 1
Male 3rd 5+ Introvert 15-17 2
Male 3rd 5+ Introvert 15-17 1
Female 3rd 2-4 Introvert 15-17 4+
Male 1st 2-4 Introvert 15-17 0
Male 3rd 0-1 Extrovert 18-21 2
Male 2nd 2-4 Introvert 18-21 2
Female 2nd 2-4 Ambivert 18-21 1
Male 3rd 1-2 Introvert 18-21 1
Male 3rd 1-2 Ambivert 18-21 3
Male 2nd 0-1 Extrovert 18-21 1
Female 2nd 2-4 Ambivert 15-17 1
Male 4th 2-4 Ambivert 18-21 3
Male 4th 0-1 Extrovert 15-17 1
Male 1st 1-2 Extrovert 15-17 0
```

Decision Tree



Results


```
Male 2nd 2-4 Ambivert Underload 4+ Yes
Male 4th 5+ Extrovert 15-17 3 No
Female 2nd 0-1 Introvert 18-21 1 No
Female 3rd 2-4 Introvert 18-21 4+ Yes
Male 3rd 2-4 Extrovert 18-21 3 No
Male 3rd 2-4 Introvert 18-21 1 No
Male 3rd 5+ Introvert 15-17 2 No
Male 3rd 5+ Introvert 15-17 1 Yes
Female 3rd 2-4 Introvert 15-17 4+ No
Male 1st 2-4 Introvert 15-17 0 No
Male 3rd 0-1 Extrovert 18-21 2 Yes
Male 2nd 2-4 Introvert 18-21 2 No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 3rd 1-2 Introvert 18-21 1 No
Male 3rd 1-2 Ambivert 18-21 3 No
Male 2nd 0-1 Extrovert 18-21 1 No
Female 2nd 2-4 Ambivert 15-17 1 Yes
Male 4th 2-4 Ambivert 18-21 3 No
Male 4th 0-1 Extrovert 15-17 1 No
Male 1st 1-2 Extrovert 15-17 0 Yes
```

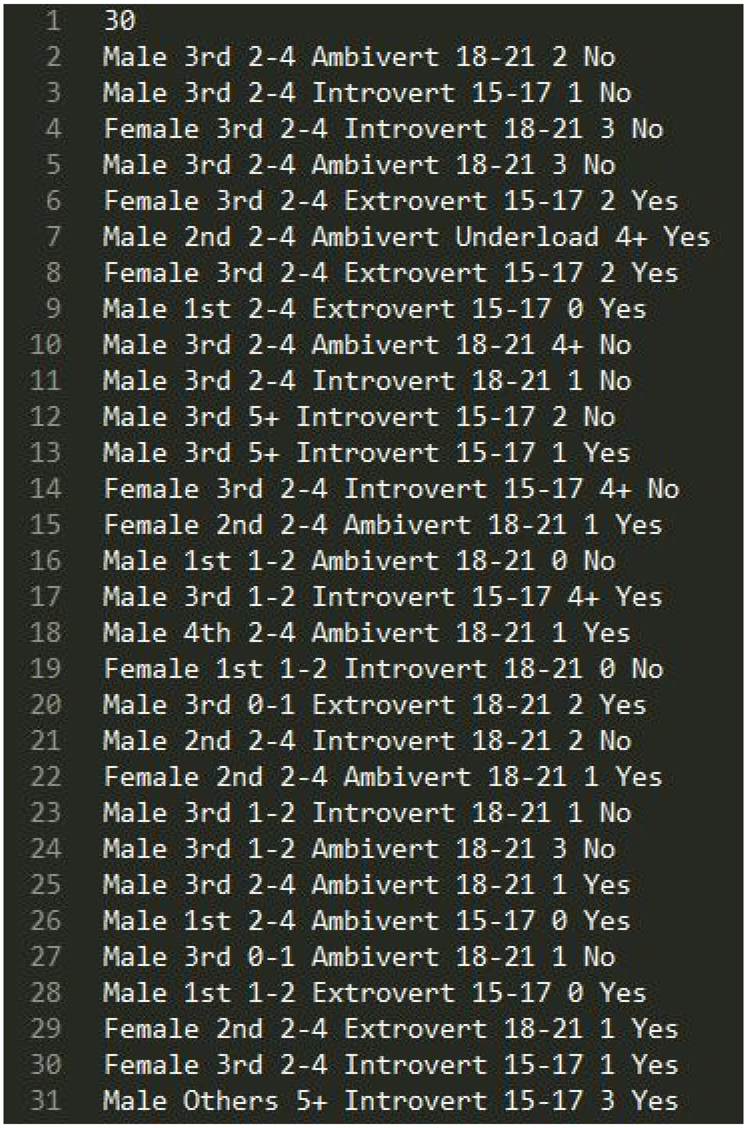
```
No
Yes
Yes
No
No
No
Yes
No
Yes
Yes
No
Yes
Yes
No
No
Yes
Yes
Yes
No
Yes
```

8 out of 20 test data were correctly classified.

Accuracy: 40%

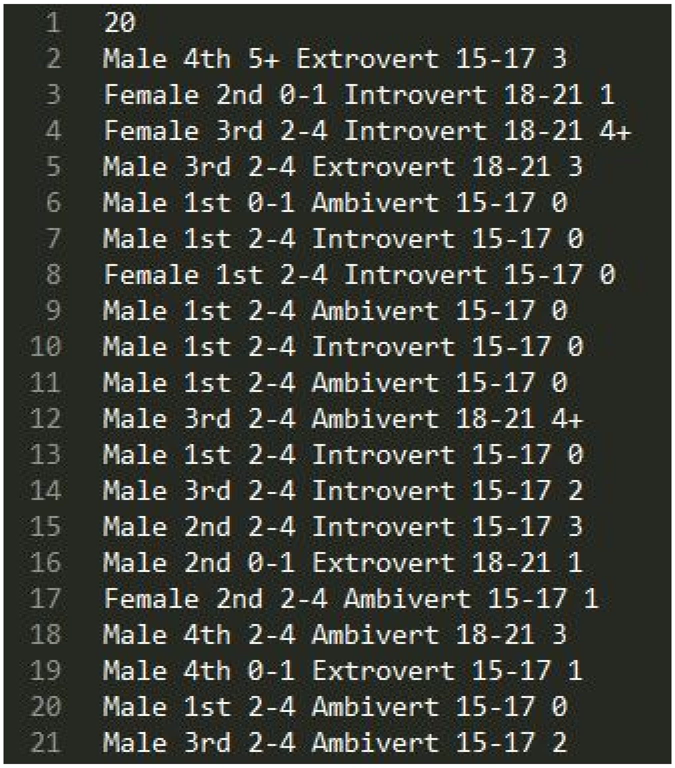
n=5:

Training Set



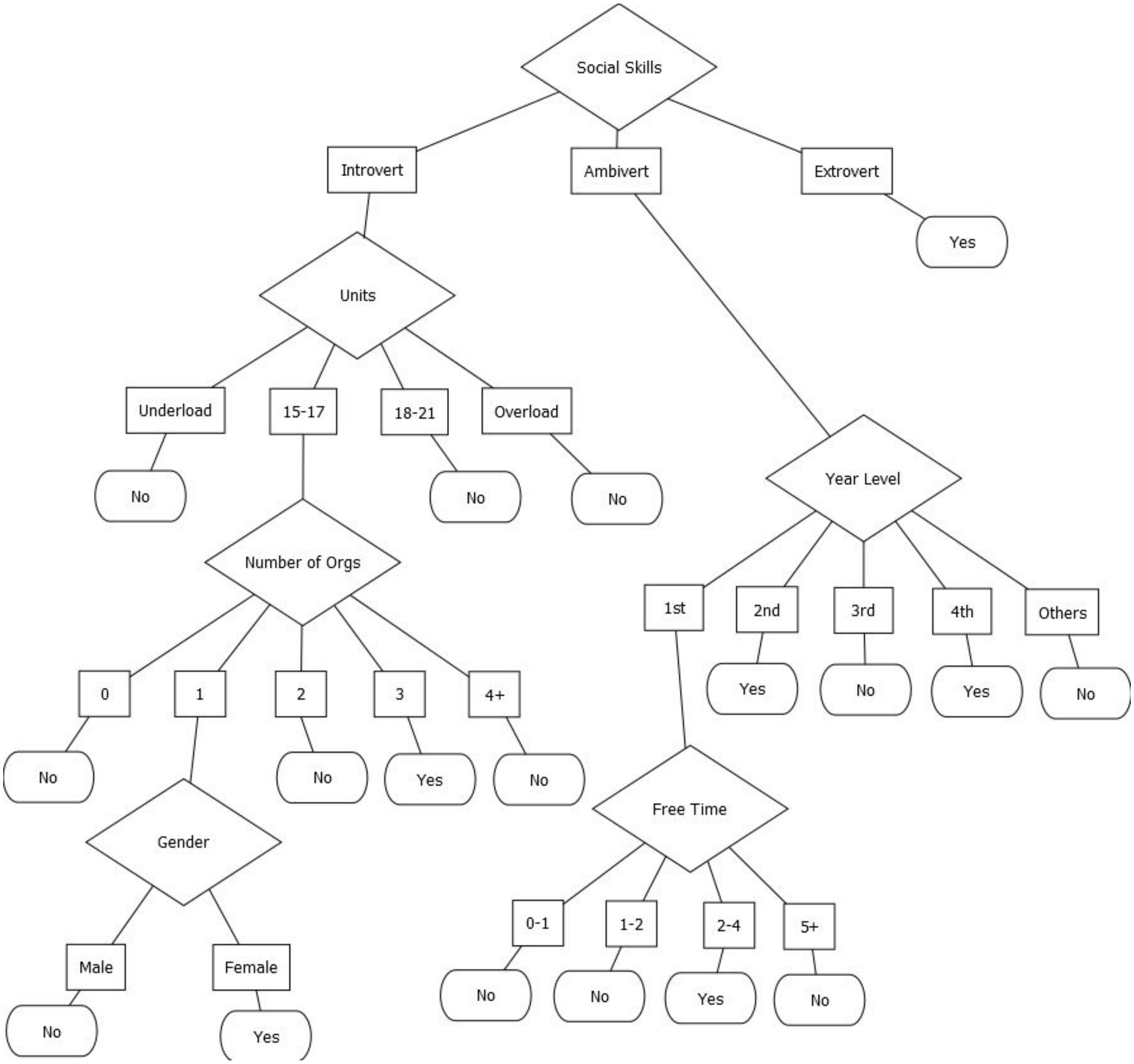
```
30
Male 3rd 2-4 Ambivert 18-21 2 No
Male 3rd 2-4 Introvert 15-17 1 No
Female 3rd 2-4 Introvert 18-21 3 No
Male 3rd 2-4 Ambivert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 2nd 2-4 Ambivert Underload 4+ Yes
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 1st 2-4 Extrovert 15-17 0 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 3rd 2-4 Introvert 18-21 1 No
Male 3rd 5+ Introvert 15-17 2 No
Male 3rd 5+ Introvert 15-17 1 Yes
Female 3rd 2-4 Introvert 15-17 4+ No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 1st 1-2 Ambivert 18-21 0 No
Male 3rd 1-2 Introvert 15-17 4+ Yes
Male 4th 2-4 Ambivert 18-21 1 Yes
Female 1st 1-2 Introvert 18-21 0 No
Male 3rd 0-1 Extrovert 18-21 2 Yes
Male 2nd 2-4 Introvert 18-21 2 No
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 3rd 1-2 Introvert 18-21 1 No
Male 3rd 1-2 Ambivert 18-21 3 No
Male 3rd 2-4 Ambivert 18-21 1 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 0-1 Ambivert 18-21 1 No
Male 1st 1-2 Extrovert 15-17 0 Yes
Female 2nd 2-4 Extrovert 18-21 1 Yes
Female 3rd 2-4 Introvert 15-17 1 Yes
Male Others 5+ Introvert 15-17 3 Yes
```

Test Set

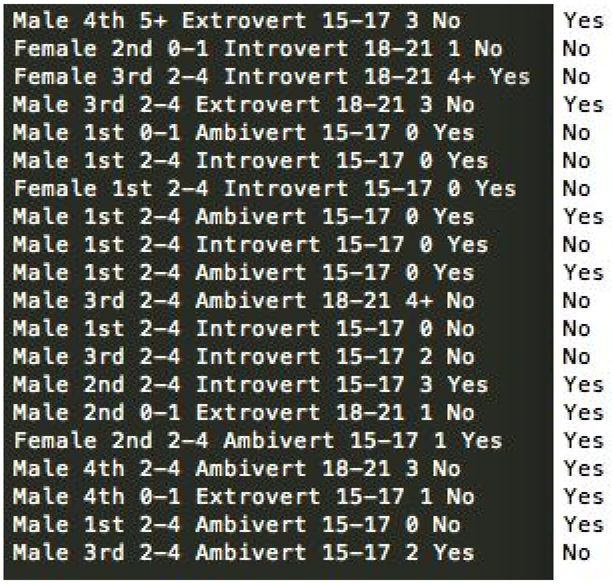


```
20
Male 4th 5+ Extrovert 15-17 3
Female 2nd 0-1 Introvert 18-21 1
Female 3rd 2-4 Introvert 18-21 4+
Male 3rd 2-4 Extrovert 18-21 3
Male 1st 0-1 Ambivert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Female 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Ambivert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Ambivert 15-17 0
Male 3rd 2-4 Ambivert 18-21 4+
Male 1st 2-4 Introvert 15-17 0
Male 3rd 2-4 Introvert 15-17 2
Male 2nd 2-4 Introvert 15-17 3
Male 2nd 0-1 Extrovert 18-21 1
Female 2nd 2-4 Ambivert 15-17 1
Male 4th 2-4 Ambivert 18-21 3
Male 4th 0-1 Extrovert 15-17 1
Male 1st 2-4 Ambivert 15-17 0
Male 3rd 2-4 Ambivert 15-17 2
```

Decision Tree



Results



```
Male 4th 5+ Extrovert 15-17 3 No            Yes
Female 2nd 0-1 Introvert 18-21 1 No         No
Female 3rd 2-4 Introvert 18-21 4+ Yes       No
Male 3rd 2-4 Extrovert 18-21 3 No           Yes
Male 1st 0-1 Ambivert 15-17 0 Yes           No
Male 1st 2-4 Introvert 15-17 0 Yes          No
Female 1st 2-4 Introvert 15-17 0 Yes        No
Male 1st 2-4 Ambivert 15-17 0 Yes           Yes
Male 1st 2-4 Introvert 15-17 0 Yes          No
Male 1st 2-4 Ambivert 15-17 0 Yes           Yes
Male 3rd 2-4 Ambivert 18-21 4+ No           No
Male 1st 2-4 Introvert 15-17 0 No           No
Male 3rd 2-4 Introvert 15-17 2 No           No
Male 2nd 2-4 Introvert 15-17 3 Yes          Yes
Male 2nd 0-1 Extrovert 18-21 1 No           Yes
Female 2nd 2-4 Ambivert 15-17 1 Yes         Yes
Male 4th 2-4 Ambivert 18-21 3 No            Yes
Male 4th 0-1 Extrovert 15-17 1 No           Yes
Male 1st 2-4 Ambivert 15-17 0 No            Yes
Male 3rd 2-4 Ambivert 15-17 2 Yes           No
```

8 out of 20 test data were correctly classified.
Accuracy: 40%

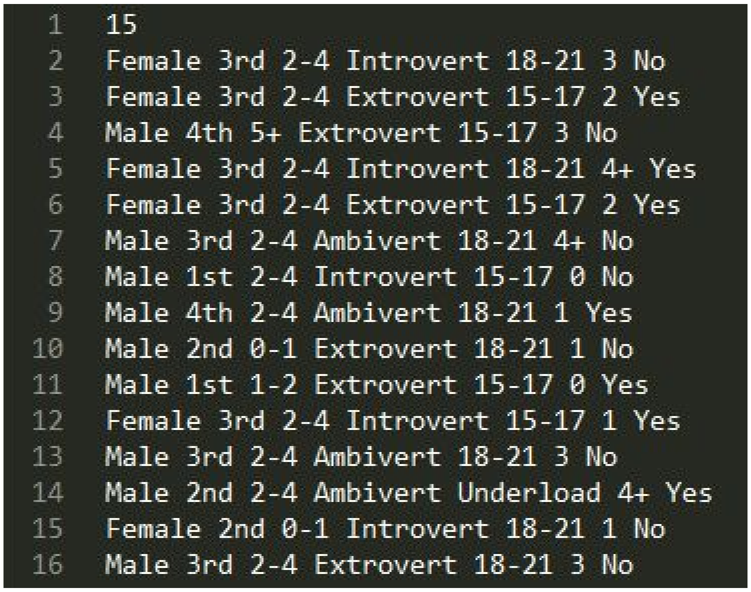
Experiment 1 General Results:
Shown below are the summarized results of the iterations.

n=1: Accuracy: 14/20 = 70%
n=2: Accuracy: 9/20 = 45%
n=3: Accuracy: 9/20 = 45%
n=4: Accuracy: 8/20 = 40%
n=5: Accuracy: 8/20 = 40%

Average Accuracy of the decision tree on the five datasets: (70+45+45+40+45)/5 = 48%

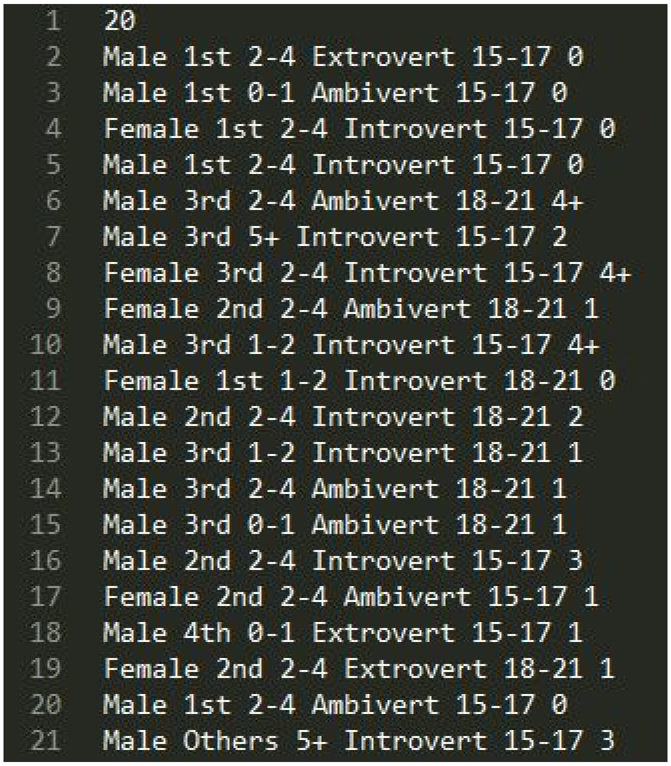
EXPERIMENT 2

Training Set



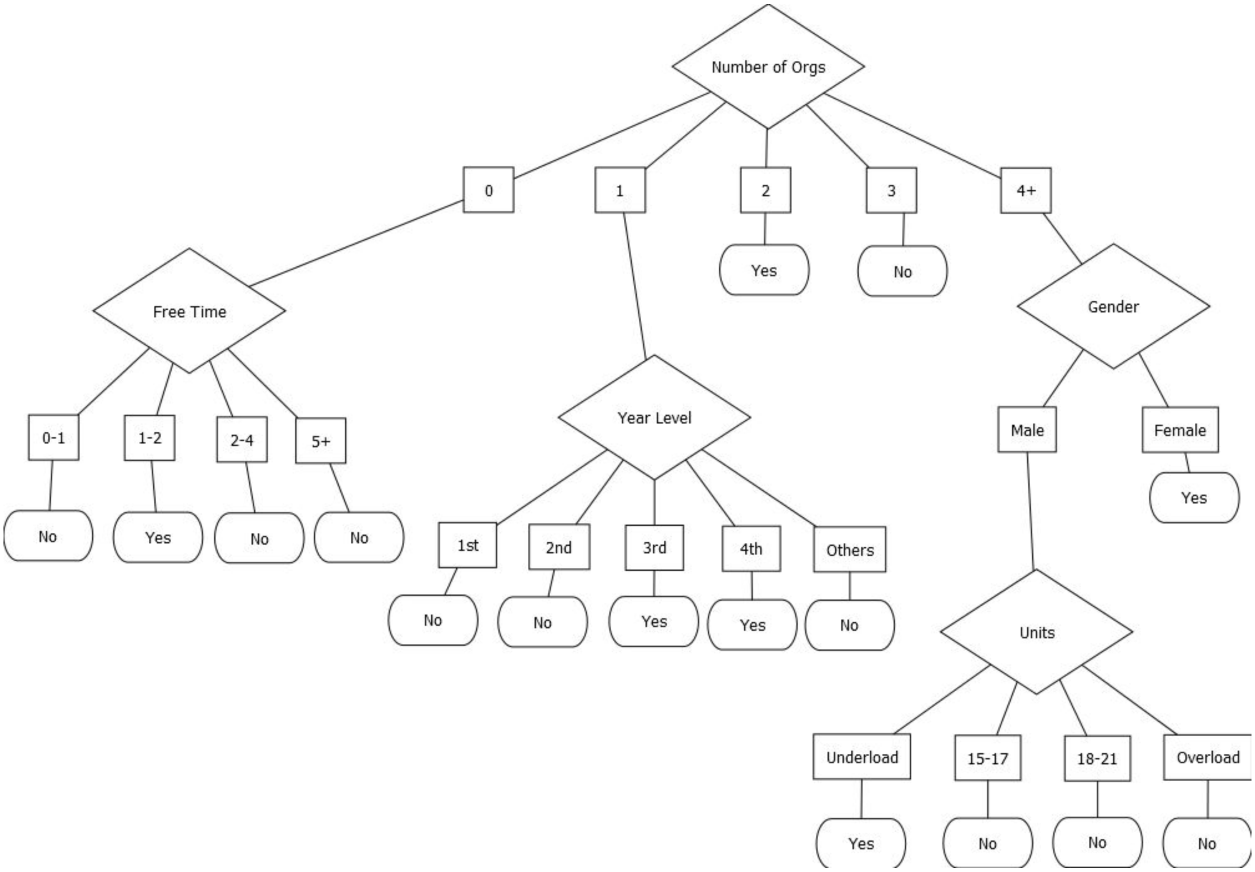
```
15
Female 3rd 2-4 Introvert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 4th 5+ Extrovert 15-17 3 No
Female 3rd 2-4 Introvert 18-21 4+ Yes
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 1st 2-4 Introvert 15-17 0 No
Male 4th 2-4 Ambivert 18-21 1 Yes
Male 2nd 0-1 Extrovert 18-21 1 No
Male 1st 1-2 Extrovert 15-17 0 Yes
Female 3rd 2-4 Introvert 15-17 1 Yes
Male 3rd 2-4 Ambivert 18-21 3 No
Male 2nd 2-4 Ambivert Underload 4+ Yes
Female 2nd 0-1 Introvert 18-21 1 No
Male 3rd 2-4 Extrovert 18-21 3 No
```

Test Set (same as in Experiment 1, n=1)

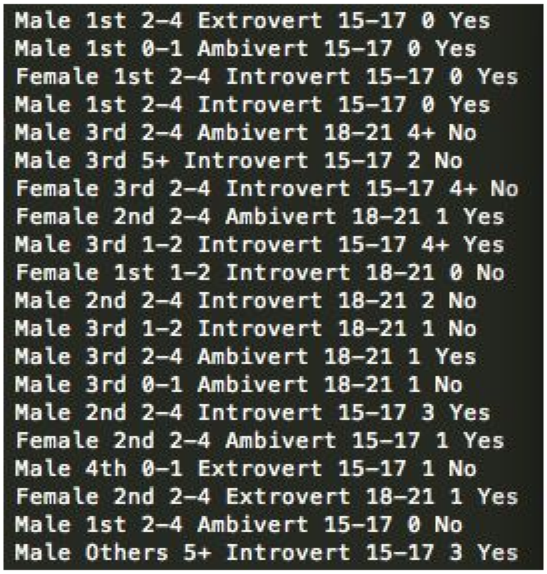


```
20
Male 1st 2-4 Extrovert 15-17 0
Male 1st 0-1 Ambivert 15-17 0
Female 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Male 3rd 2-4 Ambivert 18-21 4+
Male 3rd 5+ Introvert 15-17 2
Female 3rd 2-4 Introvert 15-17 4+
Female 2nd 2-4 Ambivert 18-21 1
Male 3rd 1-2 Introvert 15-17 4+
Female 1st 1-2 Introvert 18-21 0
Male 2nd 2-4 Introvert 18-21 2
Male 3rd 1-2 Introvert 18-21 1
Male 3rd 2-4 Ambivert 18-21 1
Male 3rd 0-1 Ambivert 18-21 1
Male 2nd 2-4 Introvert 15-17 3
Female 2nd 2-4 Ambivert 15-17 1
Male 4th 0-1 Extrovert 15-17 1
Female 2nd 2-4 Extrovert 18-21 1
Male 1st 2-4 Ambivert 15-17 0
Male Others 5+ Introvert 15-17 3
```

Decision Tree



Results



```
Male 1st 2-4 Extrovert 15-17 0 Yes       No
Male 1st 0-1 Ambivert 15-17 0 Yes        No
Female 1st 2-4 Introvert 15-17 0 Yes     No
Male 1st 2-4 Introvert 15-17 0 Yes       No
Male 3rd 2-4 Ambivert 18-21 4+ No        No
Male 3rd 5+ Introvert 15-17 2 No         Yes
Female 3rd 2-4 Introvert 15-17 4+ No     Yes
Female 2nd 2-4 Ambivert 18-21 1 Yes      No
Male 3rd 1-2 Introvert 15-17 4+ Yes      No
Female 1st 1-2 Introvert 18-21 0 No      Yes
Male 2nd 2-4 Introvert 18-21 2 No        Yes
Male 3rd 1-2 Introvert 18-21 1 No        Yes
Male 3rd 2-4 Ambivert 18-21 1 Yes        Yes
Male 3rd 0-1 Ambivert 18-21 1 No         Yes
Male 2nd 2-4 Introvert 15-17 3 Yes       No
Female 2nd 2-4 Ambivert 15-17 1 Yes      No
Male 4th 0-1 Extrovert 15-17 1 No        Yes
Female 2nd 2-4 Extrovert 18-21 1 Yes     No
Male 1st 2-4 Ambivert 15-17 0 No         No
Male Others 5+ Introvert 15-17 3 Yes     No
```

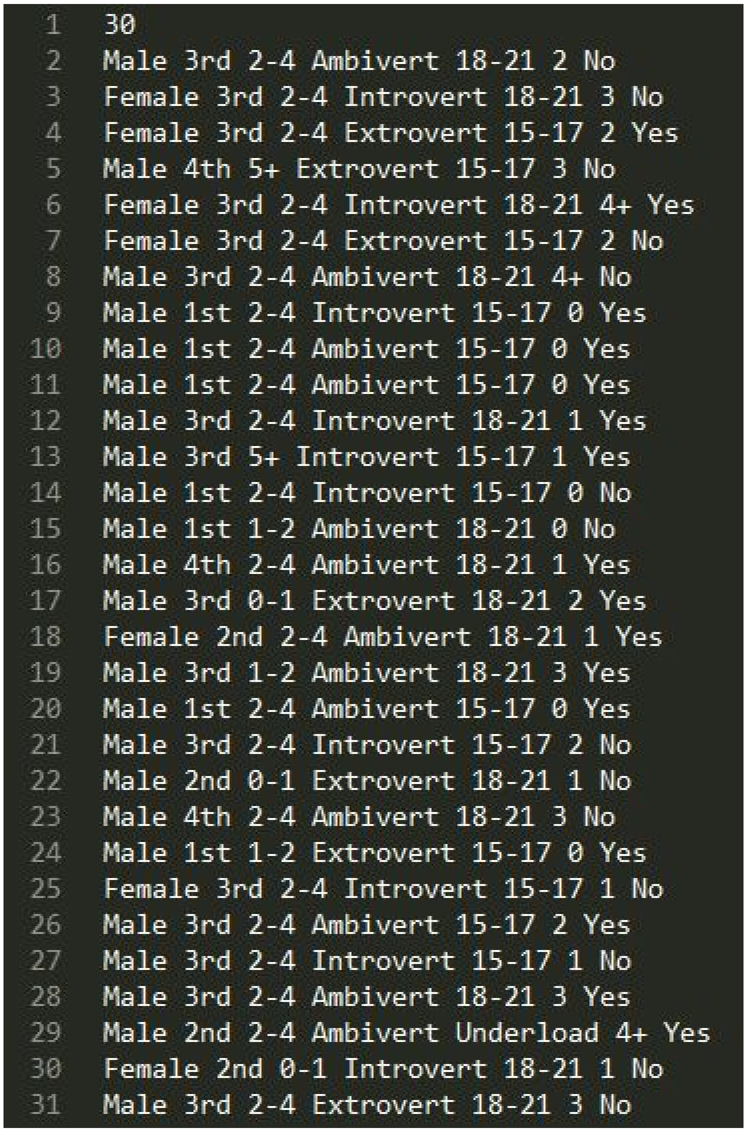
3 out of 20 test data were correctly classified.

Accuracy: 15%

The decision tree’s accuracy given a small training set (15%) is significantly lower than its accuracy given a larger training set (70%). A small training set gives the decision tree only a small number of instances to base its learning on, which is not reliable. More training instances help the tree to “learn” more which gives it a higher probability to arrive at an instance’s target value.

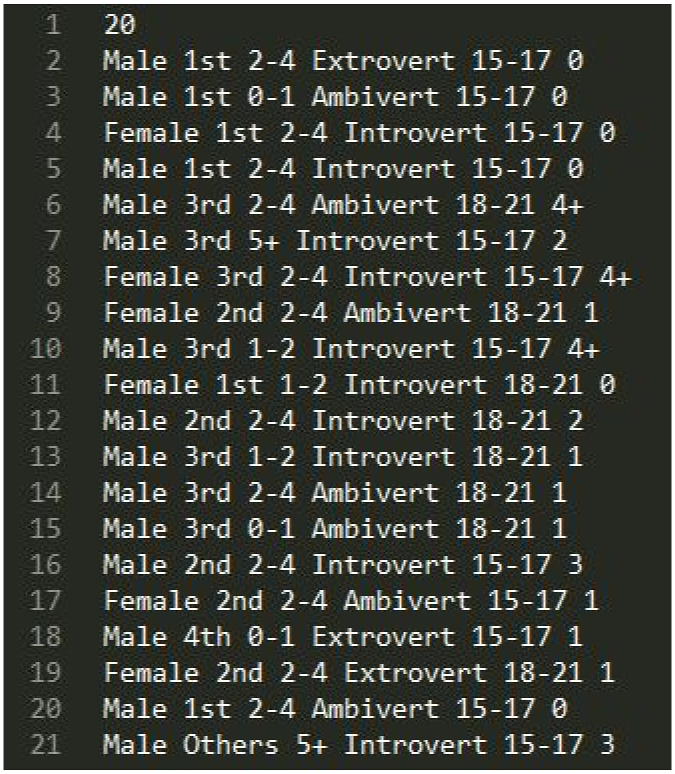
EXPERIMENT 3

Training Set



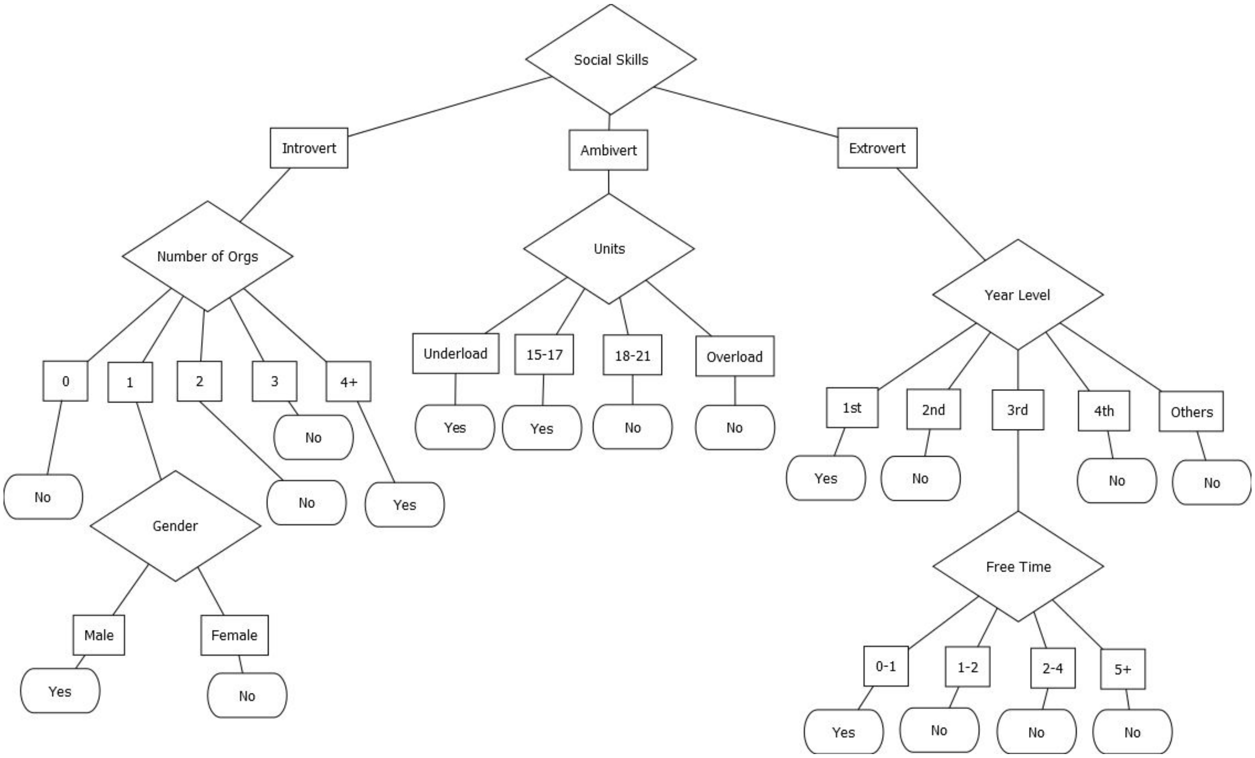
```
30
Male 3rd 2-4 Ambivert 18-21 2 No
Female 3rd 2-4 Introvert 18-21 3 No
Female 3rd 2-4 Extrovert 15-17 2 Yes
Male 4th 5+ Extrovert 15-17 3 No
Female 3rd 2-4 Introvert 18-21 4+ Yes
Female 3rd 2-4 Extrovert 15-17 2 No
Male 3rd 2-4 Ambivert 18-21 4+ No
Male 1st 2-4 Introvert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 2-4 Introvert 18-21 1 Yes
Male 3rd 5+ Introvert 15-17 1 Yes
Male 1st 2-4 Introvert 15-17 0 No
Male 1st 1-2 Ambivert 18-21 0 No
Male 4th 2-4 Ambivert 18-21 1 Yes
Male 3rd 0-1 Extrovert 18-21 2 Yes
Female 2nd 2-4 Ambivert 18-21 1 Yes
Male 3rd 1-2 Ambivert 18-21 3 Yes
Male 1st 2-4 Ambivert 15-17 0 Yes
Male 3rd 2-4 Introvert 15-17 2 No
Male 2nd 0-1 Extrovert 18-21 1 No
Male 4th 2-4 Ambivert 18-21 3 No
Male 1st 1-2 Extrovert 15-17 0 Yes
Female 3rd 2-4 Introvert 15-17 1 No
Male 3rd 2-4 Ambivert 15-17 2 Yes
Male 3rd 2-4 Introvert 15-17 1 No
Male 3rd 2-4 Ambivert 18-21 3 Yes
Male 2nd 2-4 Ambivert Underload 4+ Yes
Female 2nd 0-1 Introvert 18-21 1 No
Male 3rd 2-4 Extrovert 18-21 3 No
```

Test Set

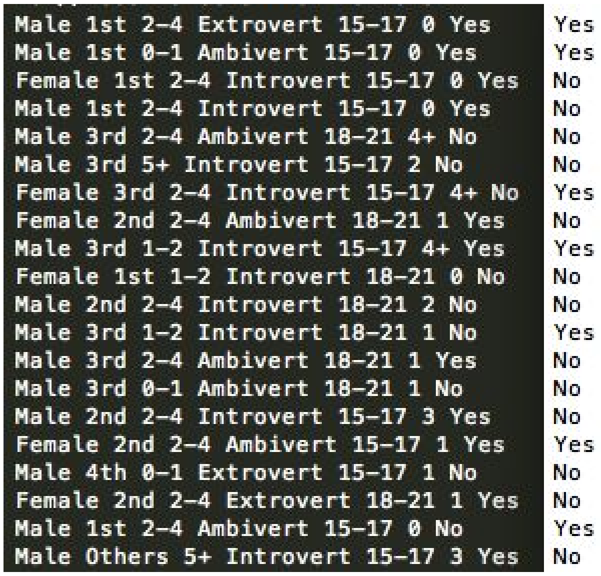


```
20
Male 1st 2-4 Extrovert 15-17 0
Male 1st 0-1 Ambivert 15-17 0
Female 1st 2-4 Introvert 15-17 0
Male 1st 2-4 Introvert 15-17 0
Male 3rd 2-4 Ambivert 18-21 4+
Male 3rd 5+ Introvert 15-17 2
Female 3rd 2-4 Introvert 15-17 4+
Female 2nd 2-4 Ambivert 18-21 1
Male 3rd 1-2 Introvert 15-17 4+
Female 1st 1-2 Introvert 18-21 0
Male 2nd 2-4 Introvert 18-21 2
Male 3rd 1-2 Introvert 18-21 1
Male 3rd 2-4 Ambivert 18-21 1
Male 3rd 0-1 Ambivert 18-21 1
Male 2nd 2-4 Introvert 15-17 3
Female 2nd 2-4 Ambivert 15-17 1
Male 4th 0-1 Extrovert 15-17 1
Female 2nd 2-4 Extrovert 18-21 1
Male 1st 2-4 Ambivert 15-17 0
Male Others 5+ Introvert 15-17 3
```

Decision Tree



Results



```
Male 1st 2-4 Extrovert 15-17 0 Yes      Yes
Male 1st 0-1 Ambivert 15-17 0 Yes       Yes
Female 1st 2-4 Introvert 15-17 0 Yes    No
Male 1st 2-4 Introvert 15-17 0 Yes      No
Male 3rd 2-4 Ambivert 18-21 4+ No       No
Male 3rd 5+ Introvert 15-17 2 No        No
Female 3rd 2-4 Introvert 15-17 4+ No    Yes
Female 2nd 2-4 Ambivert 18-21 1 Yes     No
Male 3rd 1-2 Introvert 15-17 4+ Yes     Yes
Female 1st 1-2 Introvert 18-21 0 No     No
Male 2nd 2-4 Introvert 18-21 2 No       No
Male 3rd 1-2 Introvert 18-21 1 No       Yes
Male 3rd 2-4 Ambivert 18-21 1 Yes       No
Male 3rd 0-1 Ambivert 18-21 1 No        No
Male 2nd 2-4 Introvert 15-17 3 Yes      No
Female 2nd 2-4 Ambivert 15-17 1 Yes     Yes
Male 4th 0-1 Extrovert 15-17 1 No       No
Female 2nd 2-4 Extrovert 18-21 1 Yes    No
Male 1st 2-4 Ambivert 15-17 0 No        Yes
Male Others 5+ Introvert 15-17 3 Yes    No
```

10 out of 20 test data were correctly classified.
Accuracy: 50%

The decision tree constructed based on a noisy training set has a lower accuracy (50%) than that of the tree constructed based on the original training set (70%). Noise in the training set increases the possibility of error in learning which leads to the construction of a less accurate decision tree.

SUMMARY

Experiment 1: When we performed a 5-fold cross-validation on 30 training instances and 20 test instances, the accuracy of the tree averaged to 48%.

Experiment 2: The decision tree constructed based on a smaller training set has a significantly lower accuracy (15%) than the tree constructed based on the original training set (70%).

Experiment 3: The decision tree constructed based on a noisy training set has a lower accuracy (50%) than that of the tree constructed based on the original training set (70%).

From the results of the different experiments, we have concluded that the decision tree can be constructed in multiple ways depending on the training set given. There are factors that can affect the reliability of the tree such as the number of training examples used. A smaller training set constructs a less reliable decision tree than that of a larger training set. This is because it is given less information to learn from. Noise also affects learning. The more noise a training set has, the less reliable the decision tree will turn out because the noise gives a higher possibility of error in the tree's learning.

We have also deduced from experiment 1 that cross-validation is a good technique for getting a decision tree's accuracy in learning. This is due to the observation that getting the average of several training sets' results is more reliable than only performing the experiment once (say, using the conventional 70% training 30% testing method of validation) because the measure of error tested only on one set is not representative of the entire model of the decision tree and its performance as a whole.