REPORT

Decision Tree Implementation



Major : Software Engineering

Student Number : 2013011167

Name : Jin Yong-Seok

1. Summary of Decision Tree Algorithm

This program creates ‘Decision Tree’ according to the datasets for training, and puts ‘for-test’ datasets into the created decision tree, and gets the result attribute value for each test dataset.

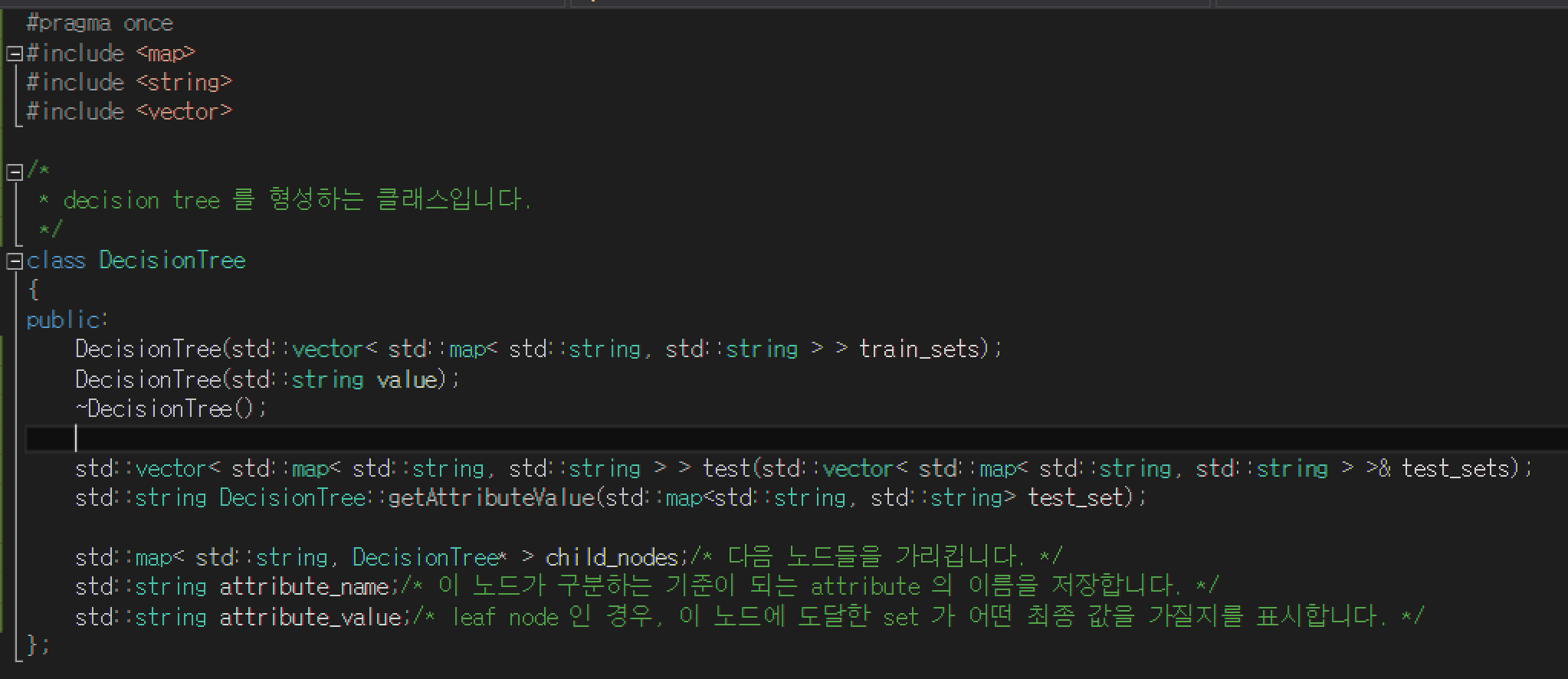
Every datasets including train and test datasets are stored in .txt file by such patteren :

First line of text file clarifies the ‘attribute name’, and the rest of file has the attribute value of each dataset. Each dataset is split by ‘\n’, and each value is split by ‘\t’, such as:

[attribute name 1]\t[attribute name 2]\t…\t[attribute name n]\n\

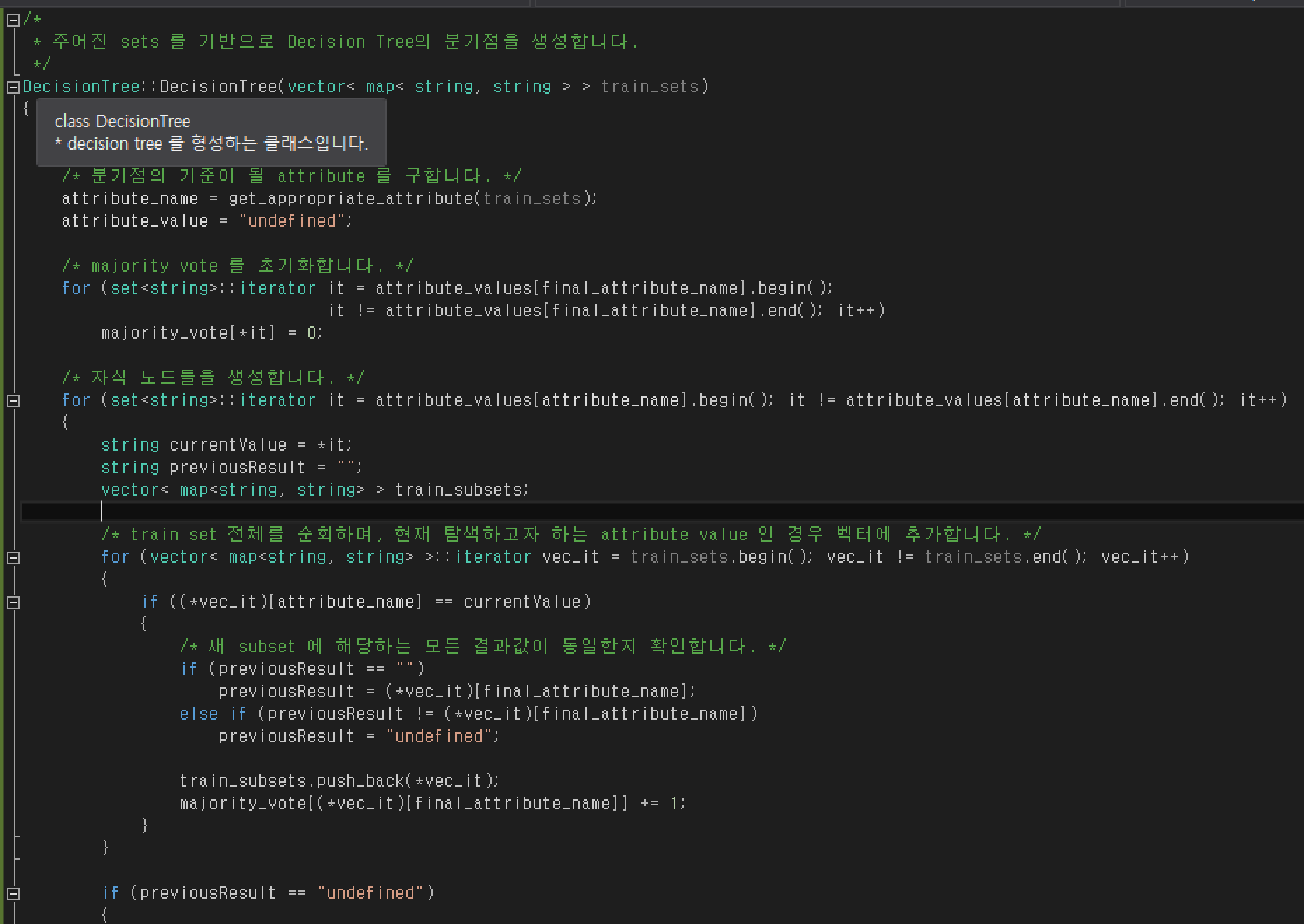
[attribute value 1]\t[attribute value 2]\t…\t[attribute value n]\n\

2. Description of Code

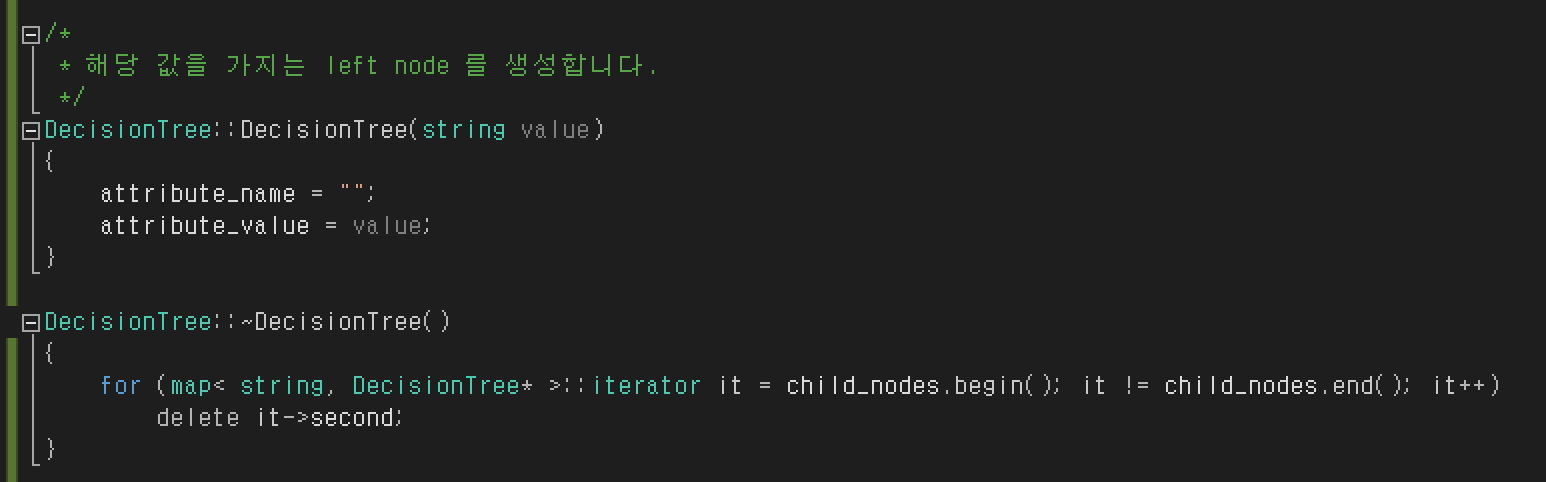


This class has 3 fields, and 2 methods.

‘child\_nodes’ has the pointer of child node pointer. ‘attribute\_name’ has the name of this node’s standard attribute. ‘attribute value’ is for leaf node of tree.

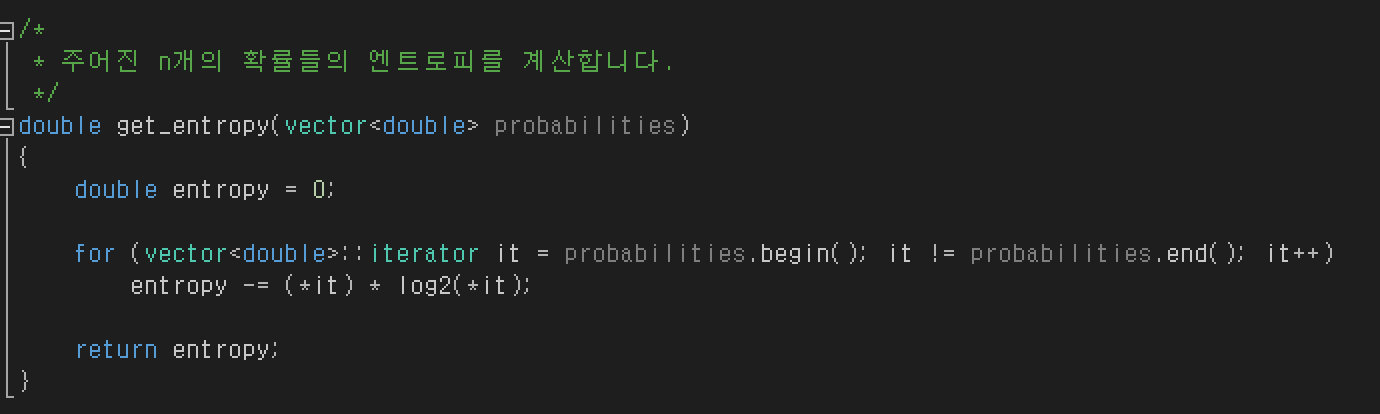
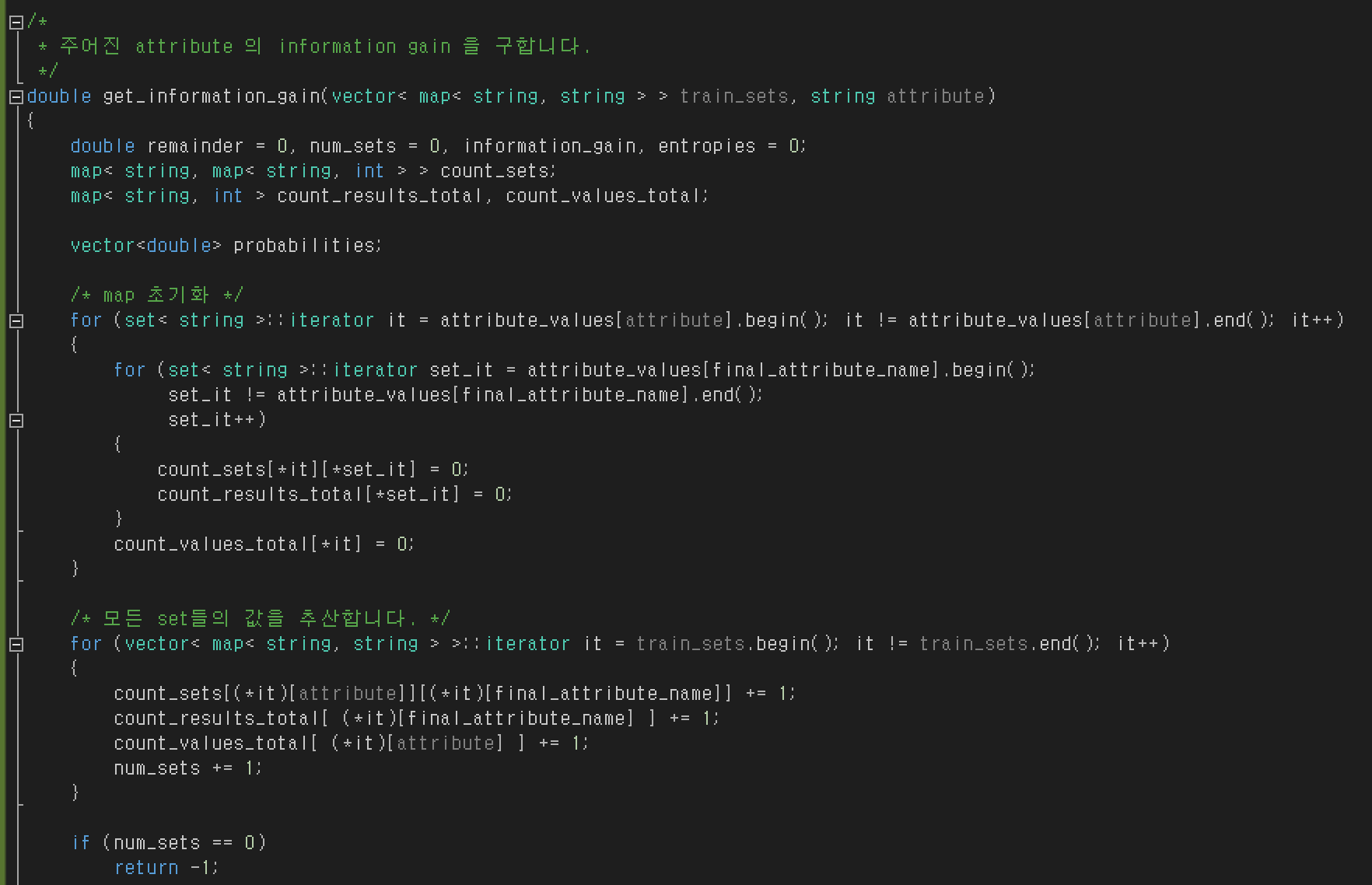
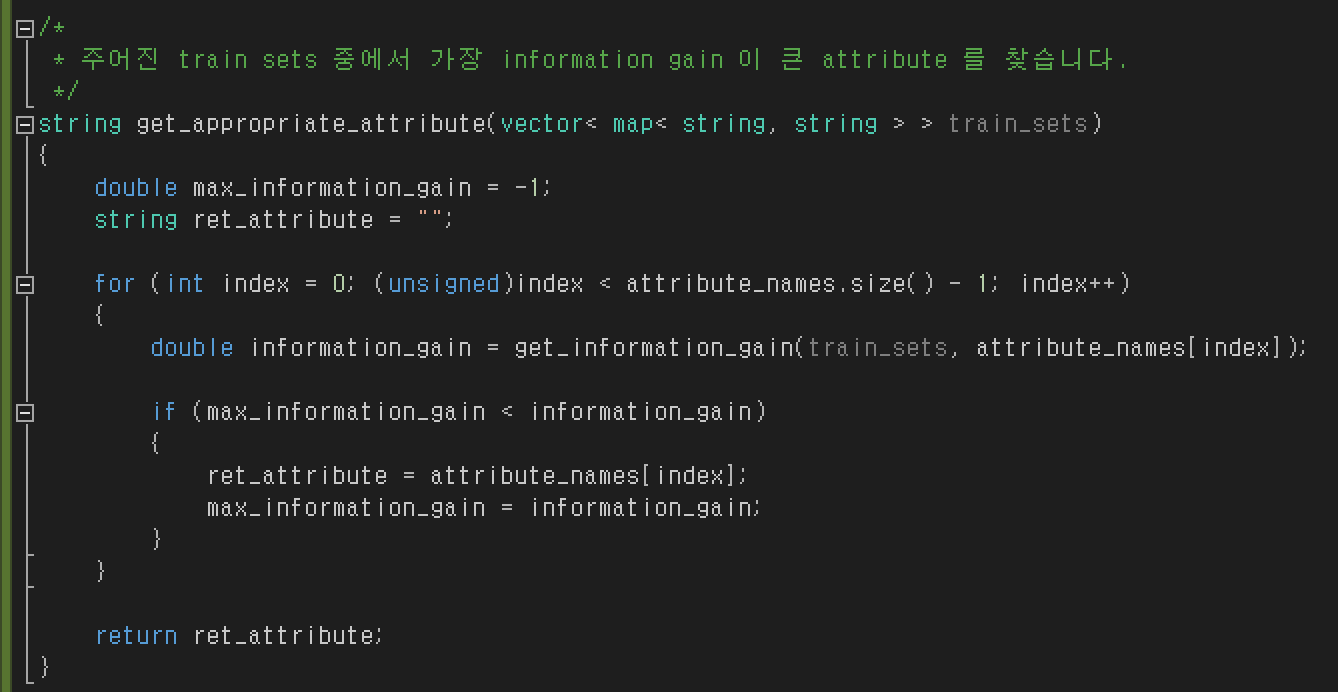


If the constructor has parameter ‘vector’, this creates a new node that divides datasets by attribute values of chosen attribute. The attribute is chosen by ‘get\_appropriate\_attribute’ function. This initializes its fields, and creates the child node of each value.



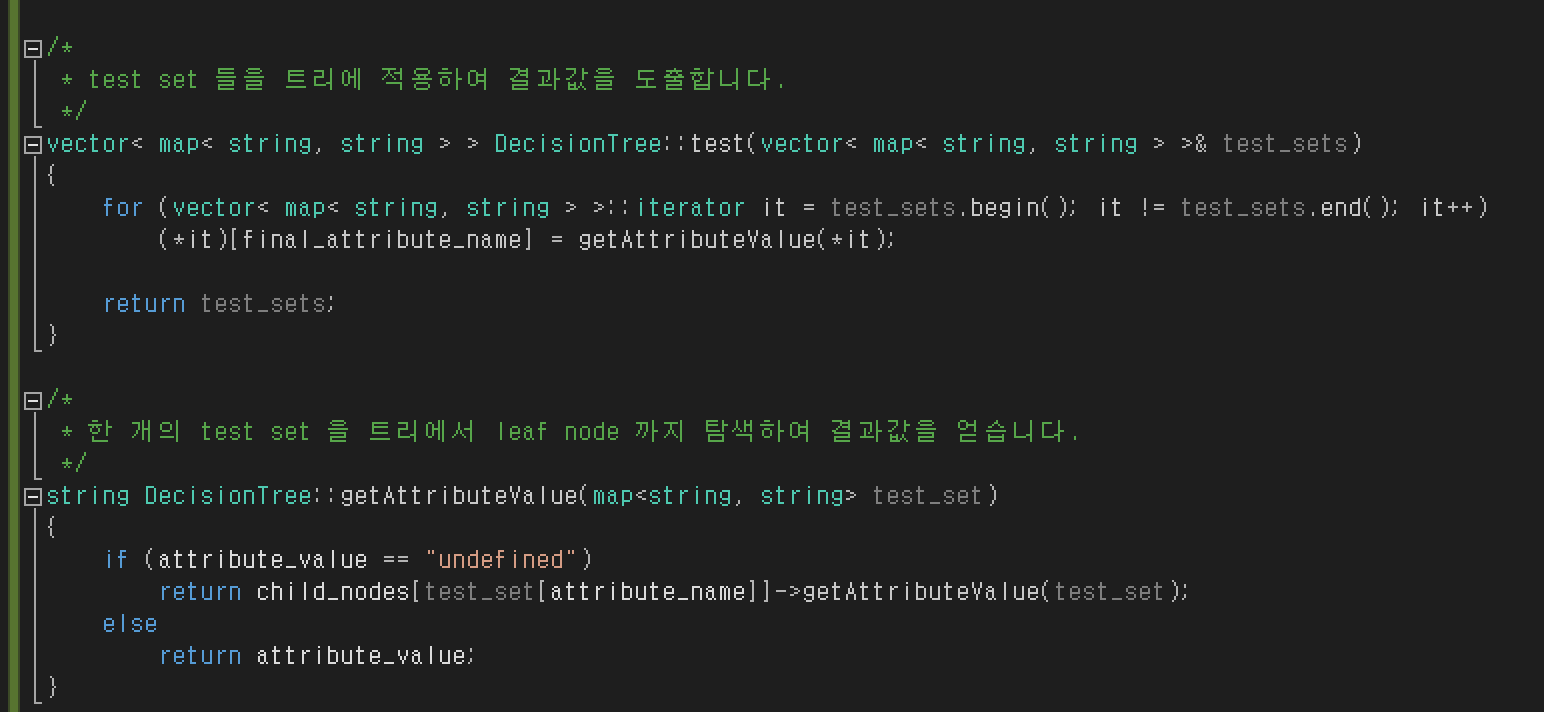
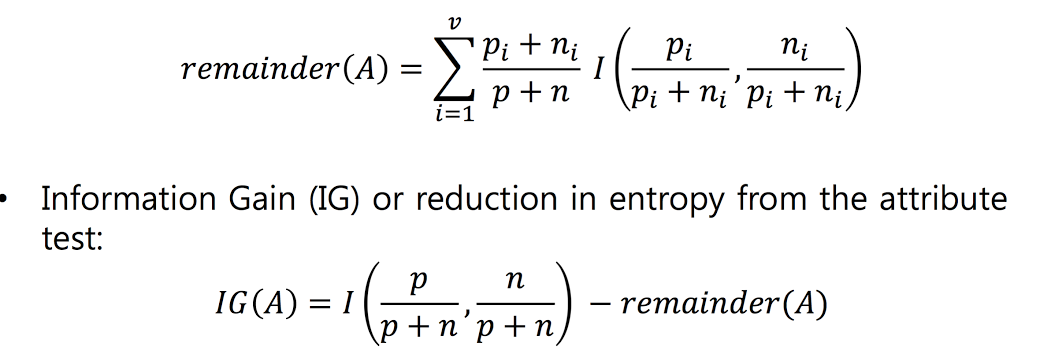
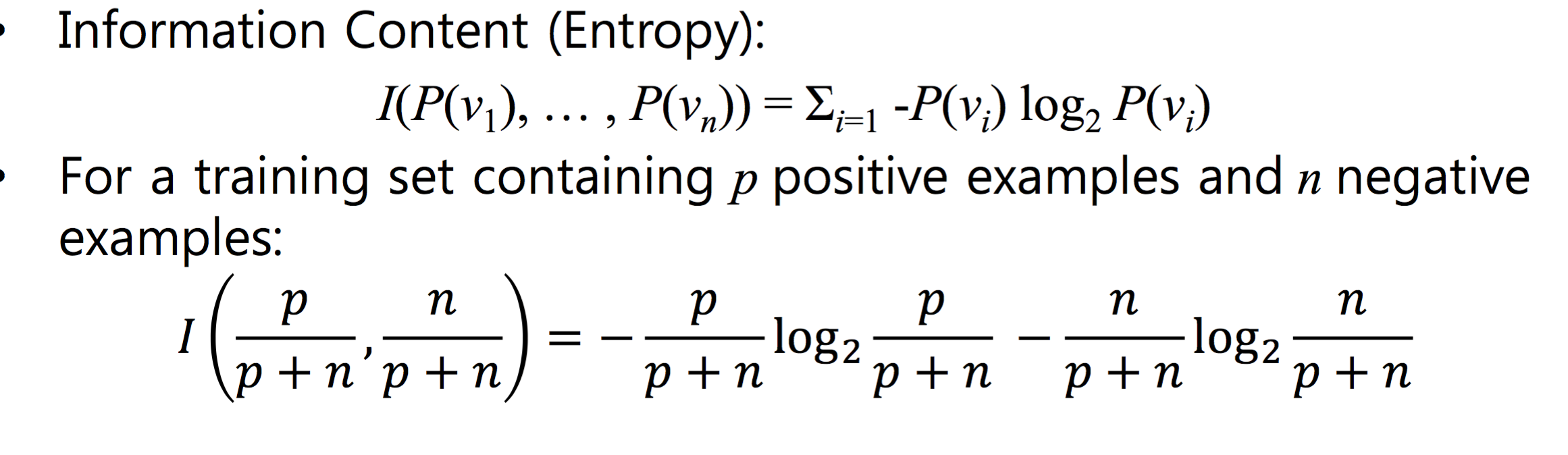
If the constructor has the parameter ‘string’, this node is the leaf node of tree. If the test dataset reaches this node, ‘attribute\_value’ would be the result of it.

The tree destructs every child nodes when destructs itself.

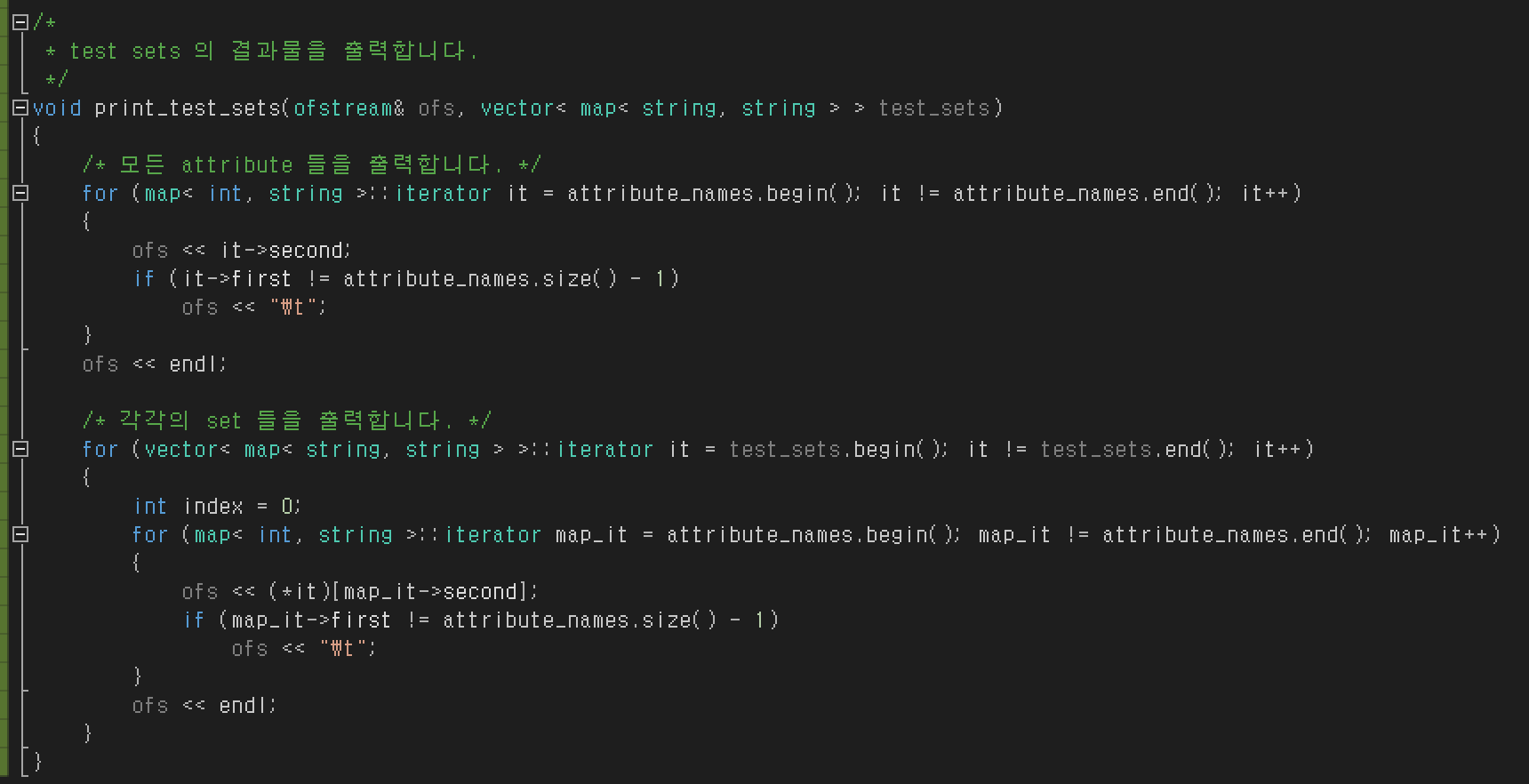


‘get\_appropriate\_attribute’ function returns the ‘appropriate’ attribute for dividing given datasets. It gets the ‘Information gain’ value for each attribute value, and gets the maximum value.

‘get\_information\_gain’ function returns the information gain value of given attribute by parameter. It first counts the number of sets of given datasets, and calculates entropy by ‘get\_entropy’ function, and calculate the information gain by these formula:



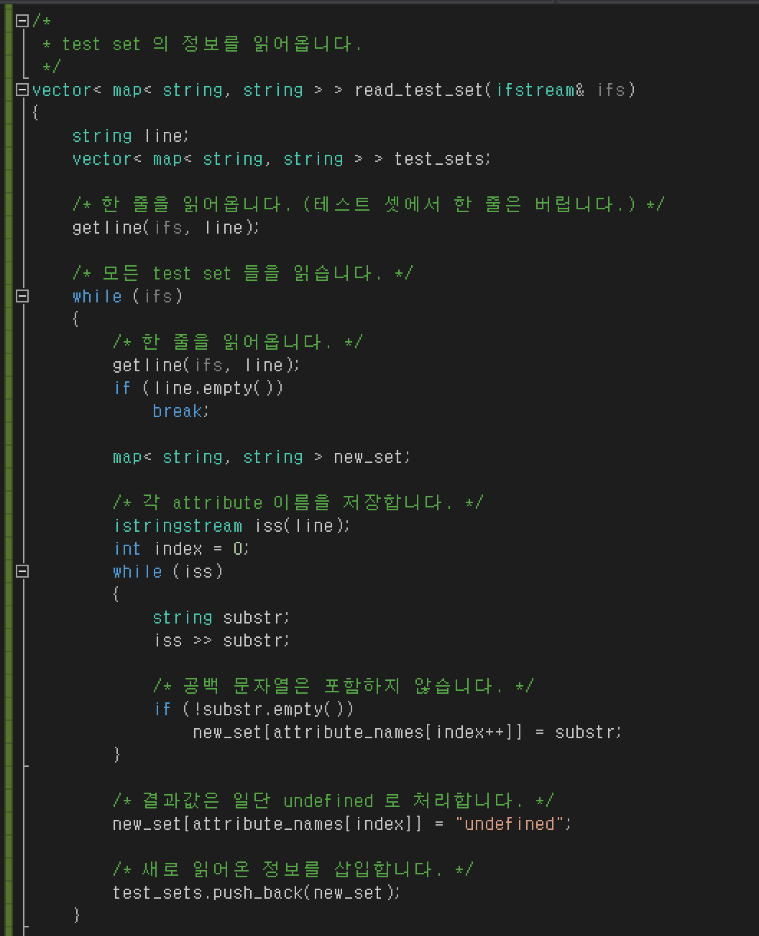
These functions are for evaluating test datasets. First, put test datasets into ‘test’ method of Decision Tree. Then, each dataset traverses into Decision Tree until it reaches leaf node. This works by ‘getAttributeValue’ method which is implemented recursively.

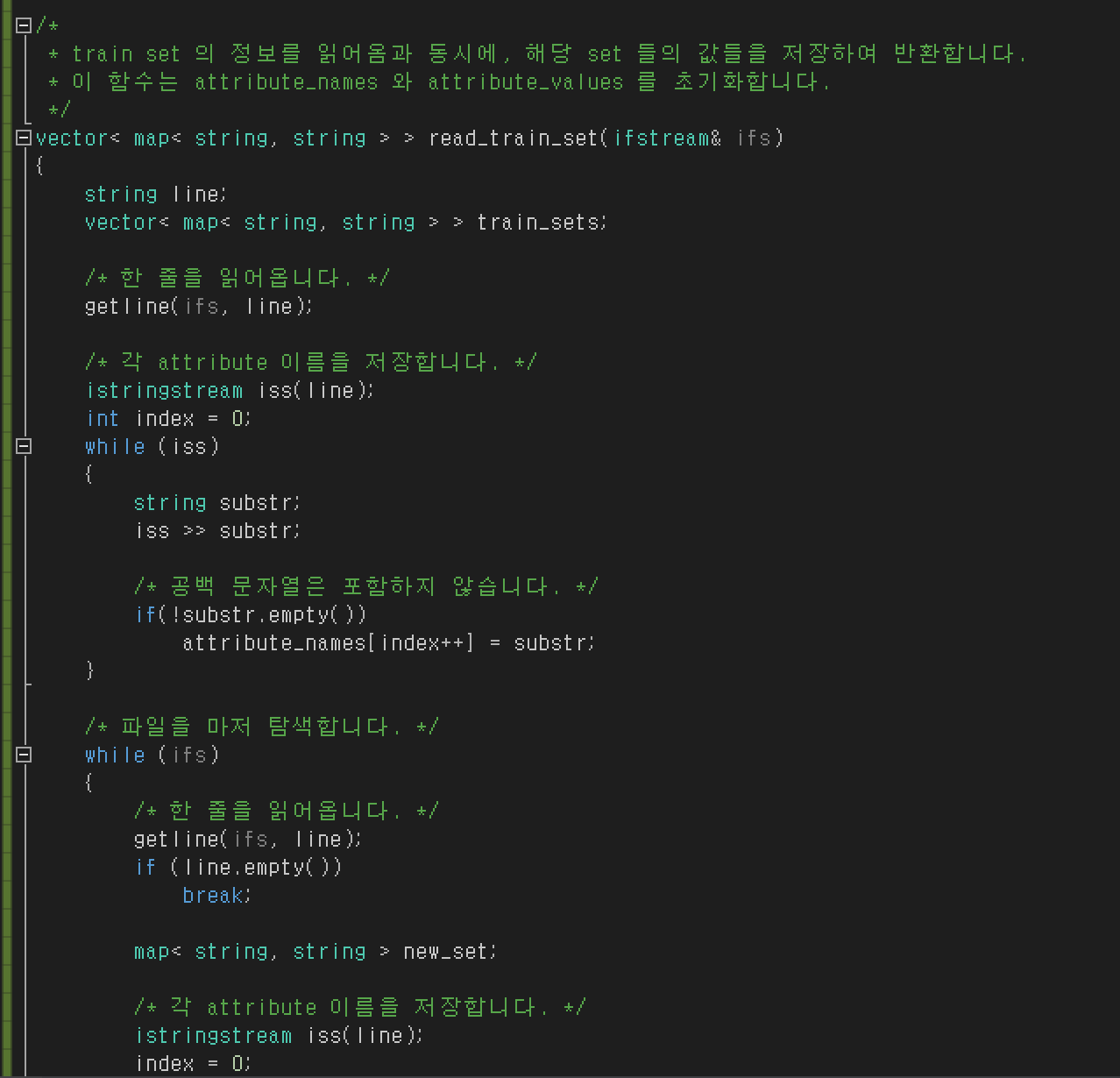


‘print\_test\_sets’ prints all datasets in vector in same format as input file :

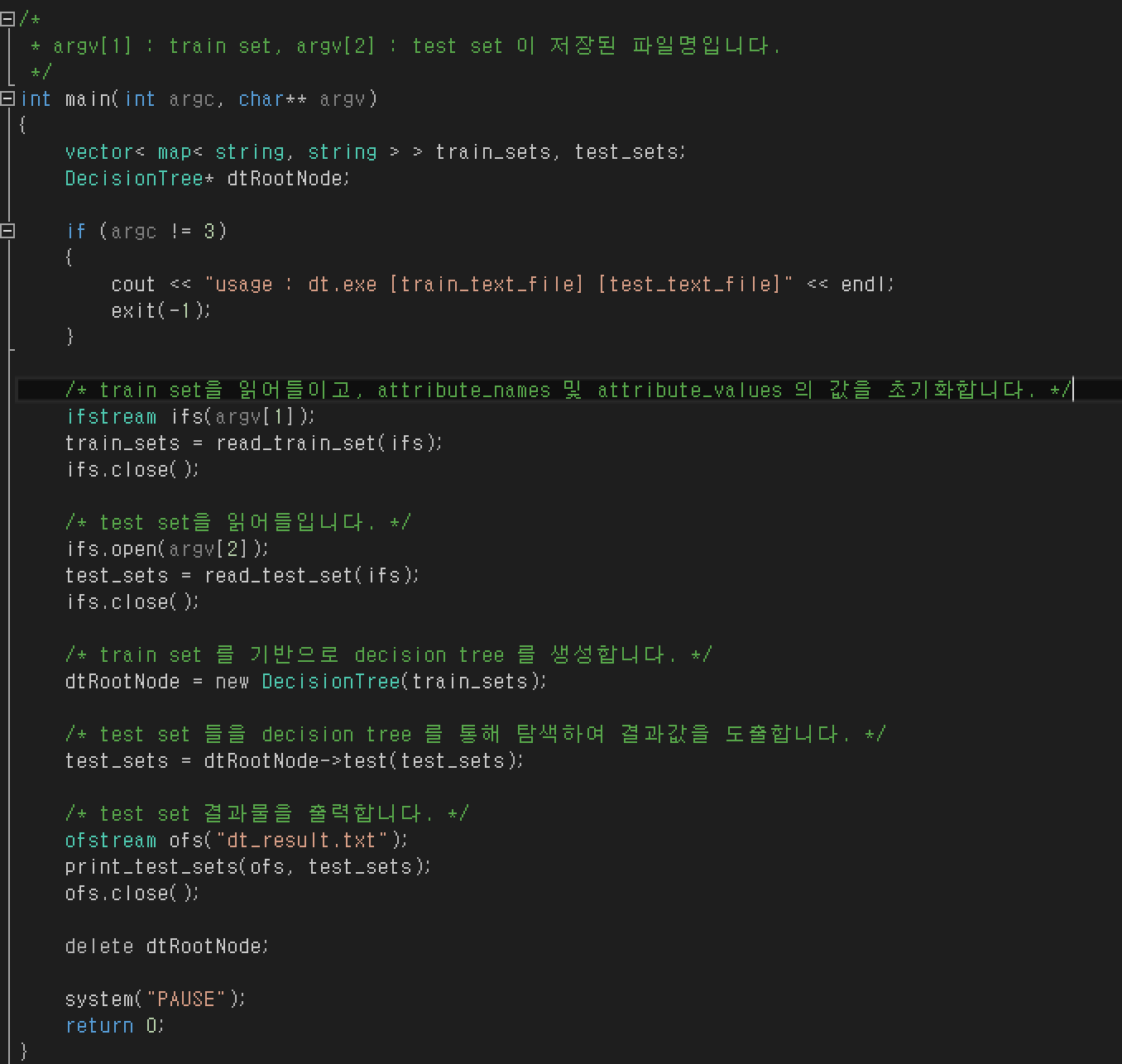
[attribute name 1]\t[attribute name 2]\t…\t[attribute name n]\n\

[attribute value 1]\t[attribute value 2]\t…\t[attribute value n]\n\



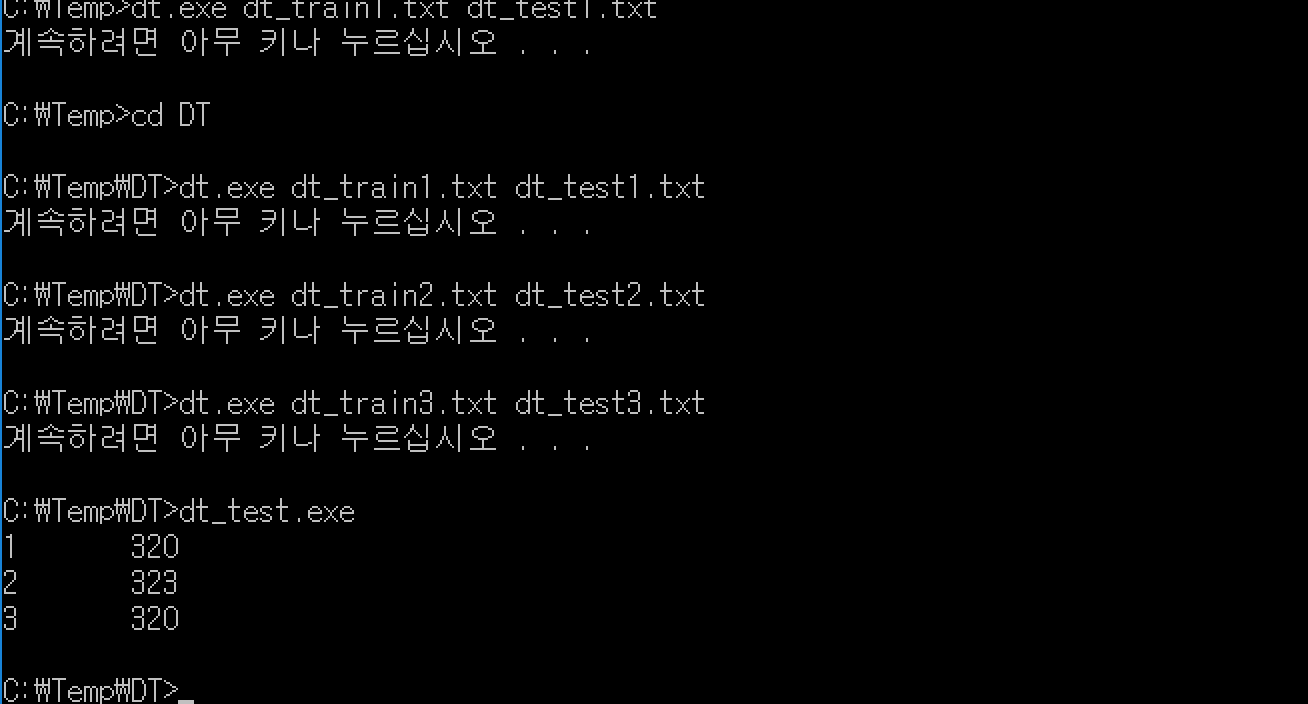


‘read\_test\_set’ and ‘read\_train\_set’ functions works similarly, since these functions read same format of text file. One thing different is that the ‘read\_train\_set’ function initializes the ‘attribute\_names’ vector and ‘attribute\_values’ vector for creating Decision Tree, etc.



This is the main function of the program. This reads each input files, and prints the output file to ‘dt\_result.txt’ file.

3. Result



This is the result of the test program uploaded on portal.

The result has about 93.2% ~ 94.1% accuracy, compared to the given answer text files.