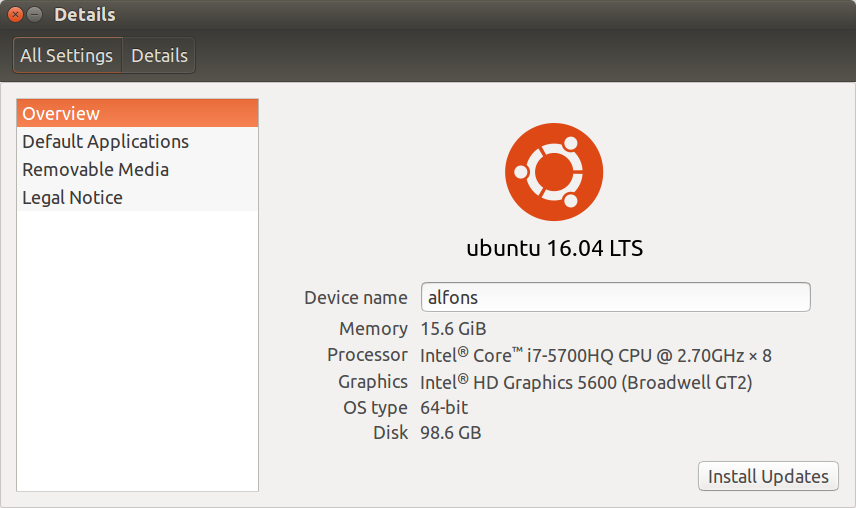
Introduction to Machine Learning Project 1

0416324 資工系 胡安鳳 An-Fong Hwu ,Dept of Computer Science

1.Programming Environment



Programming language used: C++ with -std=c++11 standard and library with #include <bits/stdc++.h>

2.What is Decision Tree and Random Forest?

A simple machine learning and training model for data prediction and analysis.

<https://en.wikipedia.org/wiki/Decision_tree>

<https://en.wikipedia.org/wiki/Random_forest>

3.How decision tree is built?

0.Store the data into the set of vector

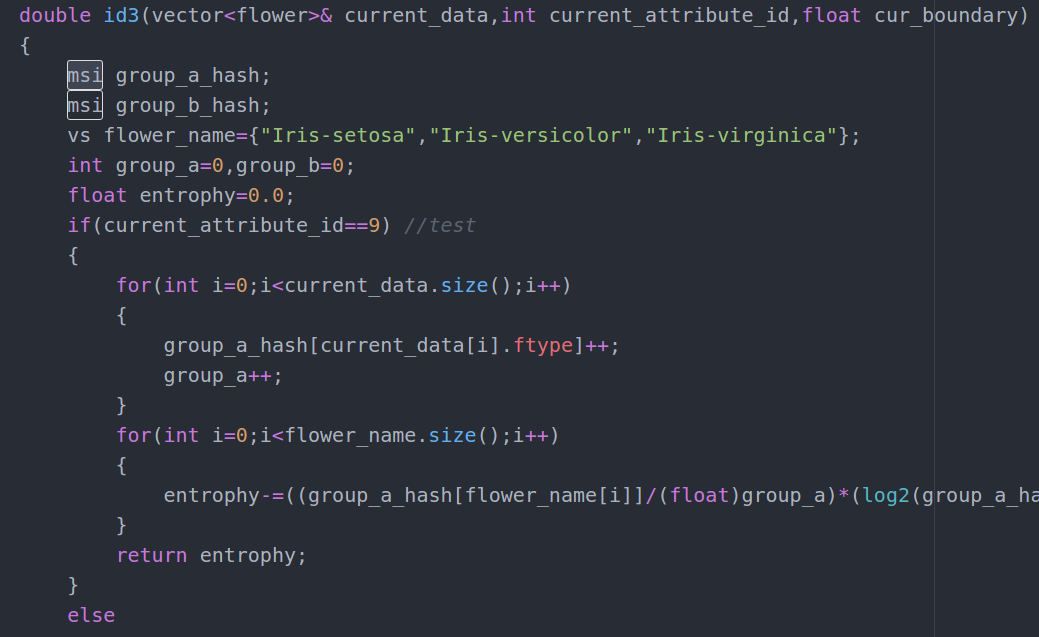
1.Sort according to the attribute (dose not matter which attribute will get the most information gain since all splitting and attribute will be calculated )

2.If different at index then we calculate at (or say split with (value[index]+value[index-1])/2)

3.Now the table has been split into 2 parts, then calculate “each splitting point”according to the ID3 algorithm

By using the std::map, we will increase the convenience to make a statistics of dataset.

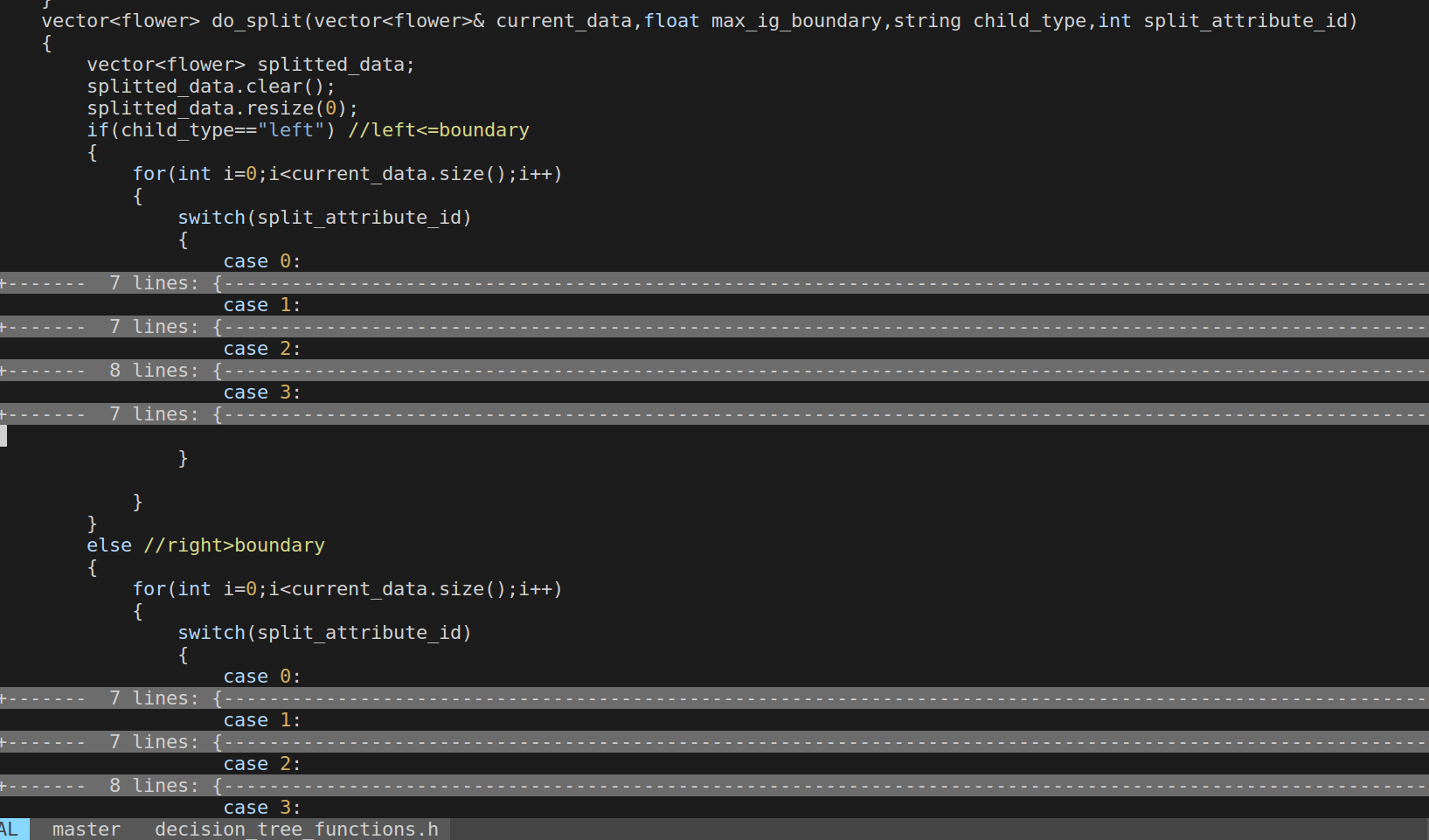
After calculating the position of splitting with LEAST ENTROPY , which means the “chaos” of data is the LEAST, then we split at such position to reduce the data inconsistency.



4.We now have left\_child and right\_child. We split ,and new\* left\_child right child, connect them parent->newchild= something

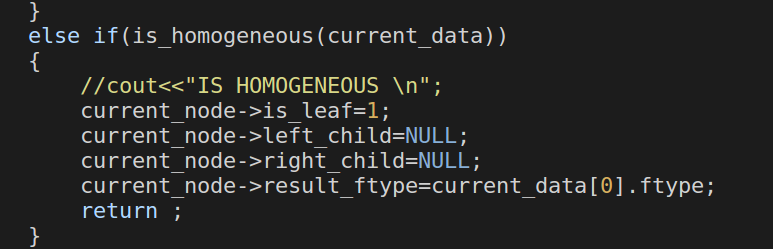
6.take the needed data into leftchild which for example <180cm , then take all the person whose height <180cm into left child and vice versa for splitting the data set according to the current criterion.

Here is the code for putting the data in left child and right child



5.If the node's data is homogeneous, stop (the node cannot be split even more).

6.the recursive algorithm is somehow like build\_decision\_tree(node\* left\_child) build\_decision\_tree(node\* right\_child) where the child is not null



Q:Which attribute to split first?

A:Does not matter, what matters is the boundary we split, the boundary has to bring us the most information gain

3.Extend to Random Forest

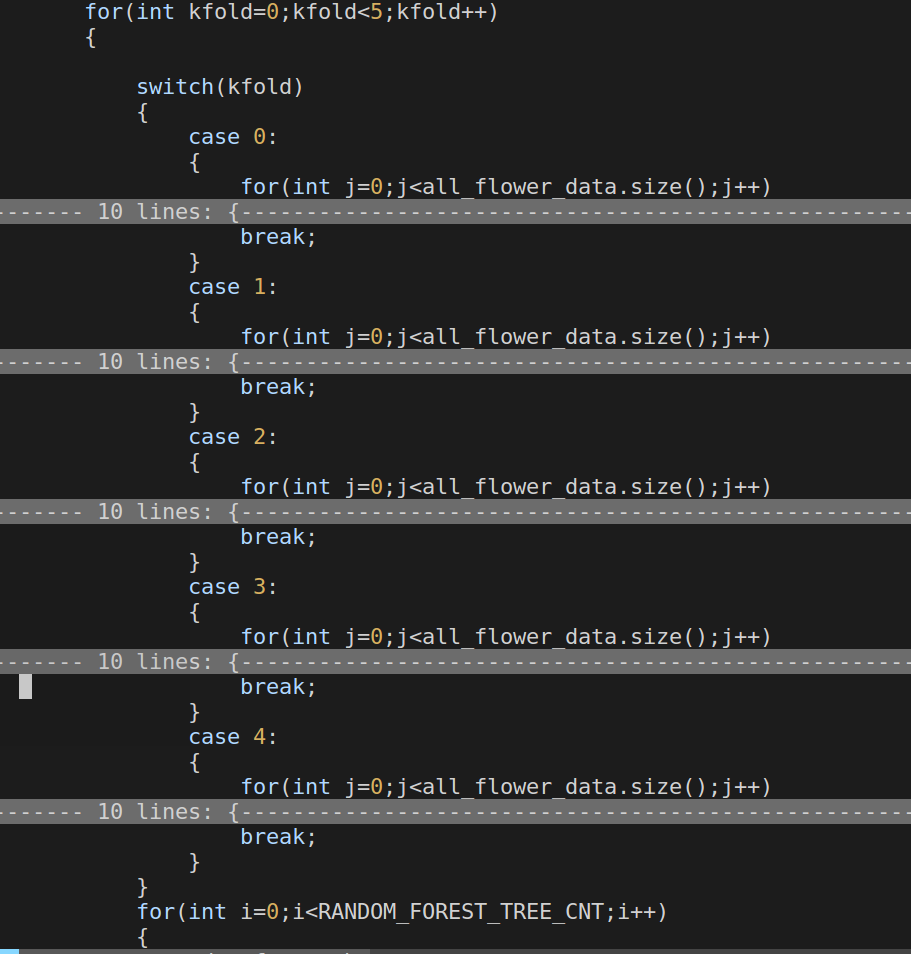
0.Build 5 decision tree where data picked from the data which require 120 training sets

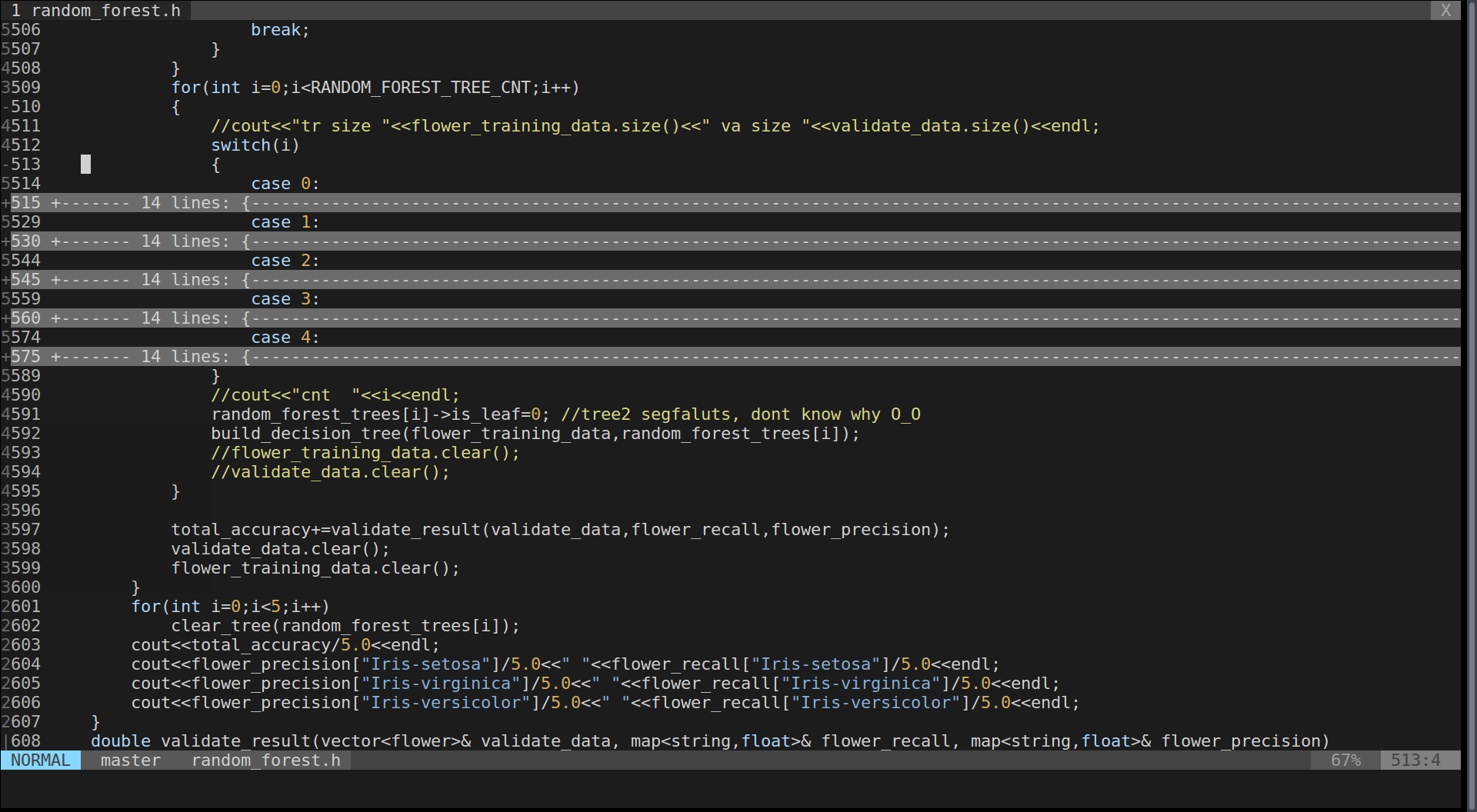
1.Each tree contains 96 datasets, reason is that 120/24=5 and 120-24=96, just like the K-Fold interval for one decision tree, but now we have the subinterval for 5 trees and 24 as a count number for interval.

2.Traverse the forest, the highest vote for the predicted class is the.

3.K Fold Cross validation still implementable.

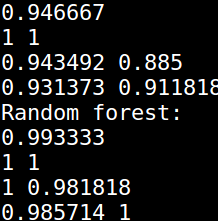
4.Implement a set of root and using the same method to build the random forest



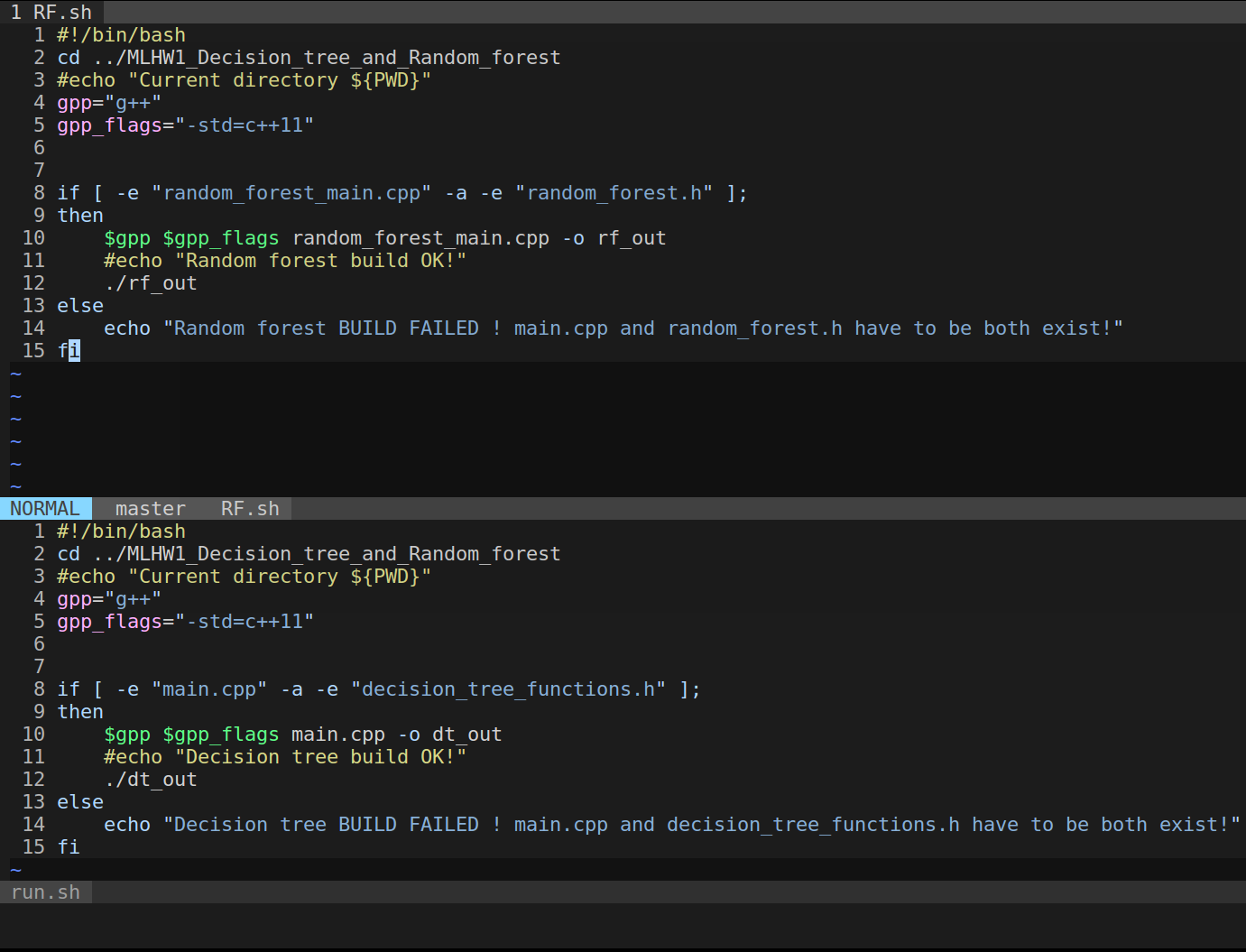


4.Validate the Result

There is a dramatically improved accuracy by using the Random Forest



5.Shell script



cd the relative path first ,setting the compiling flag, checking the availability of file then execute.