Assignment 1

1.1 a) To get the next split attribute, we have to calculate squared residuals sum for every possible split point and we chose the point with the minimum sum to be the next split attribute. Then we repeat the process from above on reminder points to decide future split points.

b) One more accurate approach is to have a portion of the training data left out for squared error evaluation. This can minimize overfitting, but the drawback of this approach is that training a regression tree requires a sufficient amount of data, otherwise overfitting will still be possible.

2.1

```
TrainAndTestRandomForest (trainingdata, numberOfTrees, percentageOfAttributes, testdata)
        for i in range numberOfTrees do
                current_tree  = build_tree(trainingdata,percentageOfAttributes)  ## pseudo for
                                                                                ##  build_tree function below

                for row in testdata do
                        final_result(row) = make_predict(row, current_tree)



build_tree(trainingdata,percentageOfAttributes)

        information gain, column, value = Split_point(data,percentageOfAttributes) ## pseudo
                                                                                ## for  Split_point function below

        if information gain = 0
                return Leaf
        else
                Yes_data, No_data = separate(data, column, value)
                Yes_branch = build_tree(Yes_data,percentageOfAttributes)
                No_branch = build_tree(No_data,percentageOfAttributes)
```

```
Split_point(data,percentageOfAttributes)
        subset_id = generate_id(number of feature, percentageOfAttributes)
        for col in range(number of features ):
                for value in unique value of the column
                        if value is in subset_id
                                Yes_data, No_data = separate(data, column, value)
                                gain = info_gain(Yes_data, No_data, current_information_gain)
        return max(info_gain), column, value
```

2.2 See the code