**Machine Learning Assignment Group 71**
**(18CS10021) Hardik Aggarwal and (18CS30040) Sriyash Poddar**

## Assignment - 1 - Decision Tree

Procedure :

1. Preprocess Data - Convert Date to Integer, separate continuous and non continuous values, add in missing values, convert dataframe to array etc.
2. Separate out Data based on the country values.
3. Call out build decision tree method to recursively build the Decision tree
4. Find out the best attribute j and best split point s - Iterate over all j and sort our input vector according to attribute j. Now for every consecutive pair of samples find split on the basis of their midpoint and find out the variance gain from the formula :-

   *Variance Gain = Variance(parent) - Weighted_Average(Variance(Children))*

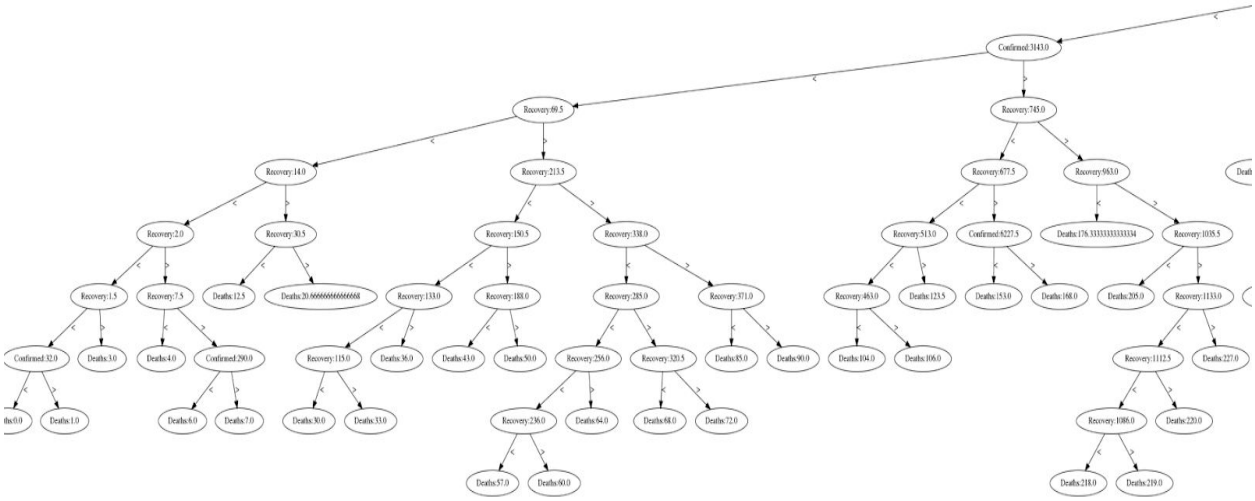
5. Find out the best (j, s) pair and return this to make our condition for splitting the decision tree.
6. Do the above procedure for 10 random splits of data and select the best tree based on the test error
7. The error used is Mean Square error.

   *Mean Squared Error = Sum (actual_value - predicted_value)**2/ (total_samples)*
8. Do the above procedure for different max depths of tree and plot a graph representing the variation of error with the height.
9. Prune the tree ->
   a. Method used for Pruning - >
      i. Post Pruning
      ii. Bottom Up
      iii. Reduced Error
   b. For the validation set -> at each node check if the error at current node is less than the weighted average of error of children. If it is less prune the tree from the current node and remove links to the children. **Using a 95% confidence interval for this as a statistical test .**

      *Error_Difference = mean(Error_Child - Error_Parent)*
      *Variance = Error_Parent*(1-Error_Parent)/n+Error_Child*(1-Error_Child)/n*
      *Limits = Error_Difference ±1.96*sqrt( Variance)*

   c. If the limits of 95% C.I. is positive, it implies that with 95% confidence the error difference is greater than 0 i.e its better to prune the node (use the parent hypothesis rather than the children.)
10. Print the Decision tree using the Graphviz package of python .

11. Decision tree :




- A small portion of pruned Decision tree
- Link for the full tree - > https://drive.google.com/file/d/17EfHJsI99QJNaymjoPAQmhlJInmkzNY3/view?usp=sharing

Observations :

| **Split Number** | **Mean Squared loss * (10^4)** |
|---|---|
| 0 | 1.754885 |
| 1 | 1.998219 |
| 2 | 1.780157 |
| 3 | 3.691383 |
| 4 | 3.187241 |
| 5 (minimum) | 1.302578 |
| 6 | 1.634223 |
| 7 | 1.738071 |
| 8 | 3.006433 |
| 9 | 1.985773 |

**Fig (a) : Variation of Mean squared error with different splits**

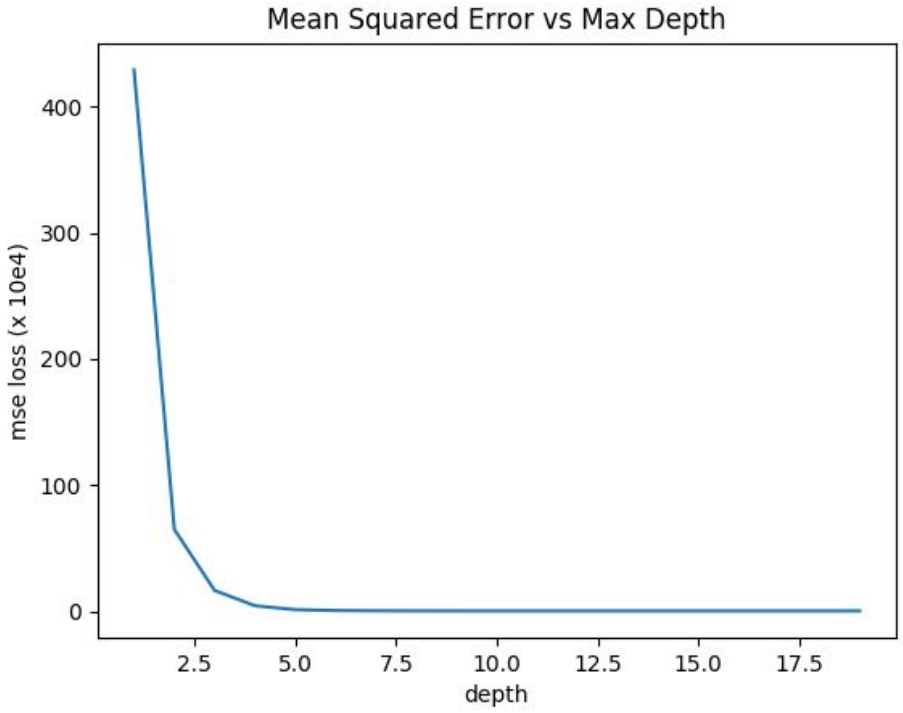



**Fig (b) Variation of Mean squared error with Max depth achieved**

**Best depth = 11 with MSE = 13025.769395905978**

Results :

| | |
|---|---|
| MSE Loss (Before Pruning) | **5059.583275413558** |
| MSE Loss (After Pruning) | **2803.63785048752** |
| Nodes in the tree (Before Pruning) | 18586 |
| Nodes pruned (removed) | 4440 |
| Percentage improvement in MSE | (5059.58 - 2803.63)/(5059.58) = **44%** |
| Final MSE of the tree | **2803.63** |