

Thyroid classification using SAS and Tableau

Business Machine Learning -1

Shanmukha Venkata Sai Kiran Pathuri – U00877781



# Introduction

# Background

Thyroid disease is an illness that affects the thyroid gland. It is divided into two categories: hypothyroidism and hyperthyroidism. Hypothyroidism, also known as underactive thyroid, is a disorder in which the thyroid gland does not generate enough hormones to keep the body functioning normally. Hyperthyroidism, also known as hyperactive thyroid, is a disorder in which the thyroid gland generates an excessive amount of hormones required for regular body activities. Goiter, thyroid nodules, thyroid cancer, and autoimmune thyroid disease are all examples of thyroid disease. Thyroid disease can be caused by a number of reasons, including genetic predisposition, environmental exposure, and lifestyle choices (salman and Sonuç, 2021). Thyroid illness treatment varies depending on the kind of thyroid issue and may include drugs, dietary changes, hormone therapy, or surgery.

# Problem Statement

By analyzing patient data such as medical history, physical examination, and imaging results, machine learning can be used to diagnose thyroid diseases. The data can subsequently be used to train a machine learning model to diagnose thyroid problems. To accurately detect new patients with the same ailment, the algorithm can be trained on a dataset of individuals with known thyroid disorders. The model can also be used to forecast a patient's chance of developing particular illnesses based on their medical history, physical examination, and imaging results. This can lead to earlier and more accurate thyroid problem diagnosis and therapy.

# Project Aim

The goal of this project is to use machine learning techniques to accurately classify thyroid diseases in a given dataset. The dataset will include patient records such as age, gender, blood tests, and other relevant information. The aim is to accurately classify the thyroid diseases into one of the four categories: hyperthyroid, hypothyroidism as well as negative. The performance of the model will be evaluated based on its accuracy and precision.

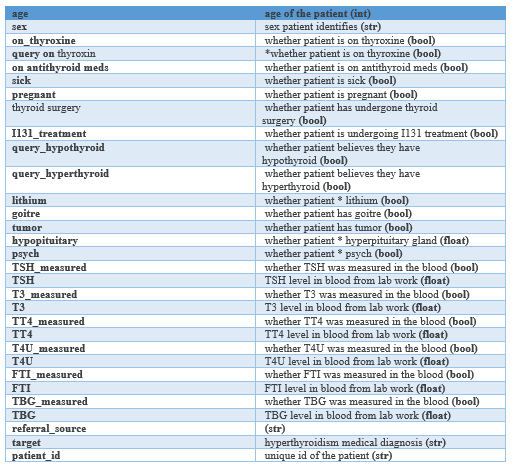
Different machine learning algorithms will be implemented with the help of SAS as well as visualization should be performed using Tableau. The visualizations will help in decision making for different organizations to detect thyroid based on different important features.

# Dataset used

The data have a total of 9172 observations with 3 attributes. The link of the data is given below

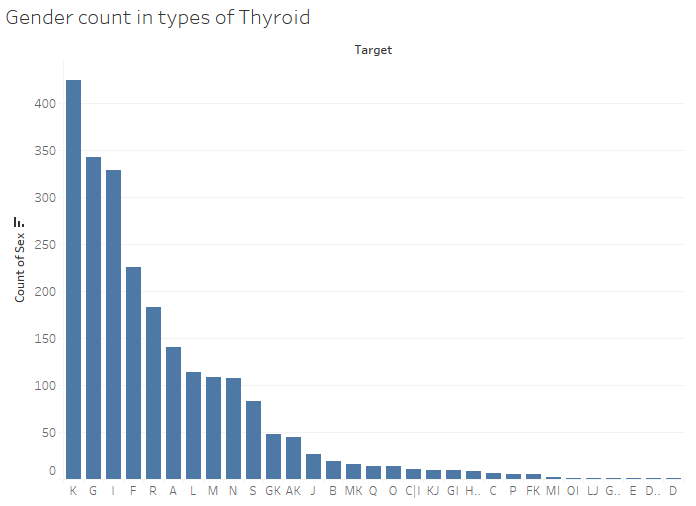
<https://www.kaggle.com/code/emmanuelfwerr/xgboost-multi-class-classification/data>

The details of the attributes present in the data are given below.

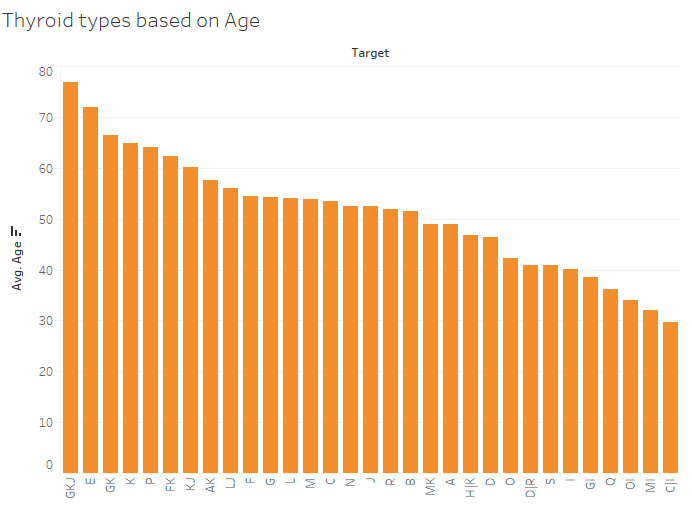


The outcome of the data is multiclass where thyroid is divided based on different diagnosed condition.

# Tableau Findings

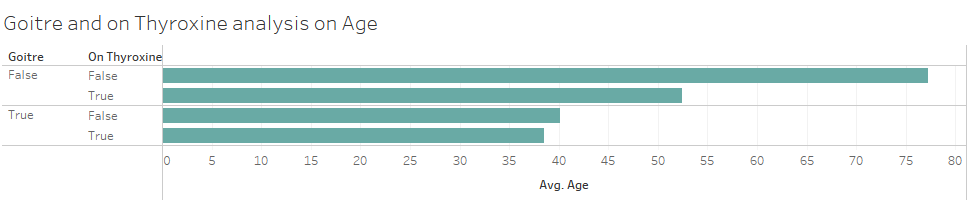


Most of the patients from this data set belong to level K of thyroid which stands for concurrent non thyroidal illness. There are different types of thyroid level present in the data in which most of the patients belong to this type.



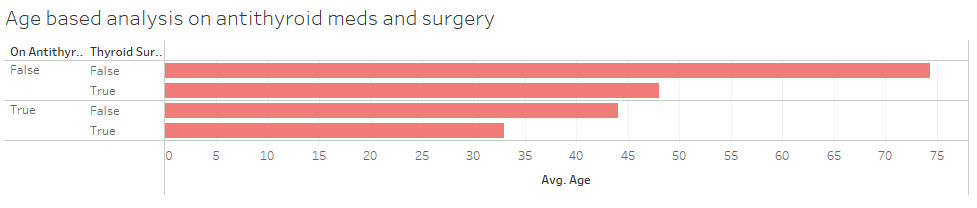
The average age of patients is analyzed in different types of thyroid where GKJ type of thyroid attacks old patients with average age higher than 75 years.

GKJ are 3 types where G belongs to compensated hypothyroid, K belongs to concurrent non thyroidal illness and J belongs to decreased binding protein. C and I types attacks patients in middle age of near to 40 years age.

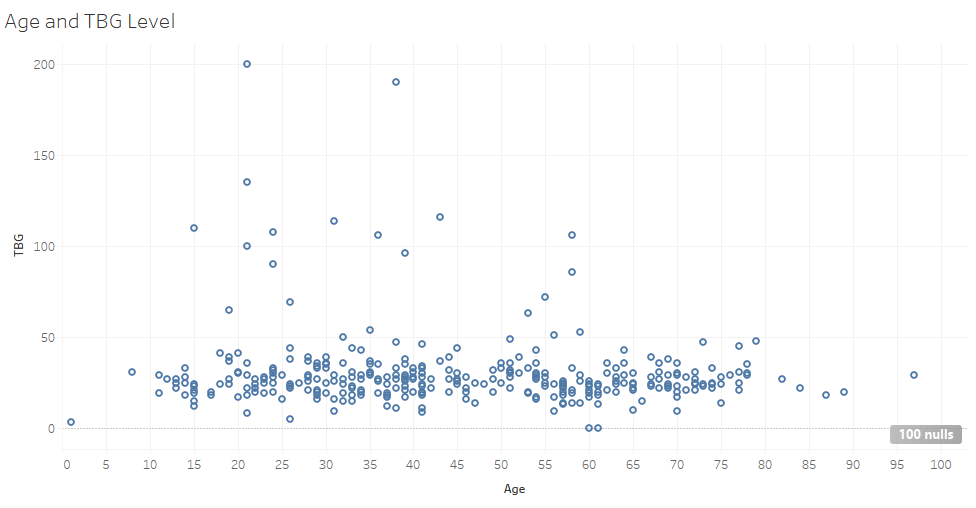


The average age of patients who have Goitre are near to 40 years and some patients below 40 years take thyroxine after having Goitre.

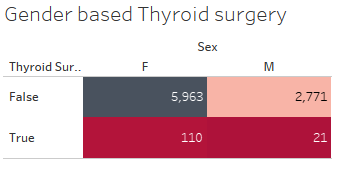
There are some patients who do not have Goitre but takes thyroxine whose average age is much higher than those patients who have Goitre.



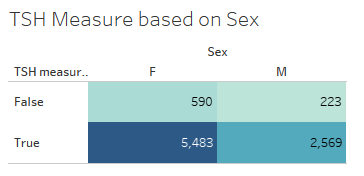
The average age who do not have anti thyroid medicines and surgery are higher than the patients who are on anti-thyroid medicine and surgery. There are older patients who are normal and do not take thyroid medicines in the data. However there are patients falling between 30 to 50 years who are on anti-thyroid medicine or had thyroid surgery.



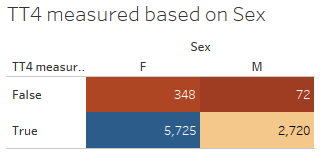
The TBG level of patients in all the age category seem to remain stable where increase of age of a patient do not increase the TBG level which seems to remain low below 50.



From the table, it seems there are more patients who did not have thyroid surgery in the data but based on gender, female patients had more thyroid surgery than male patients.



Based on the table, it shows that there are very few patients who did not measure the TSH in blood but female patients measured more TSH level compared to male patients.



The table shows female patients measured more TT4 compared to male and it is evident that there are more female patients in the data compared to male patients.

# Data Exploration and Preprocessing

The data is explored using Excel where the missing values and number of variables are found.

* **Missing Data**

There are some missing values which are found in variables such as Sex, TSH, T3, TT4, T4U, FTI and TBG. These missing values will be imputed with median of the features that has missing values.

* **Data Inconsistency**

The data has lot of labels which will affect the predictive performance. So all these labels should be aggregated under standard types of thyroid.

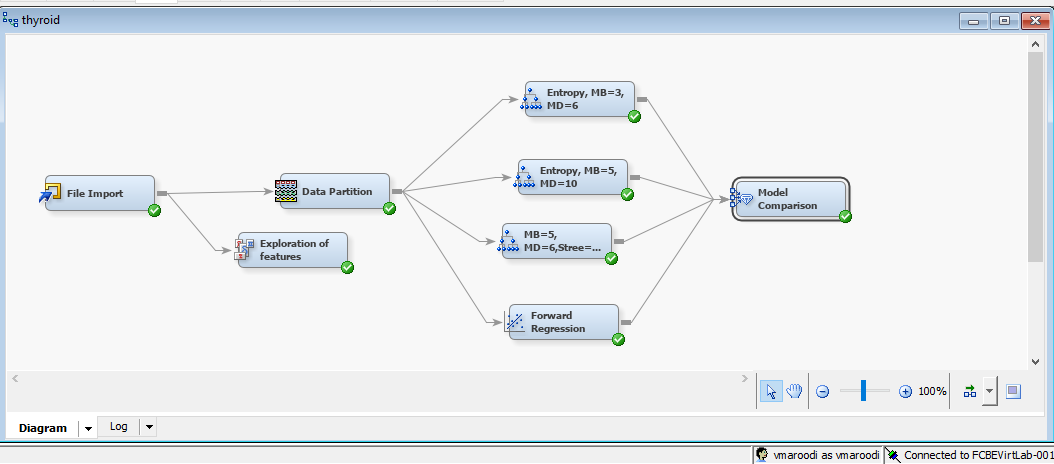
* **Data Reduction**

The data has some columns like **patient\_id** which do not bring any significance. So such columns will be reduced.

# Predictive modelling

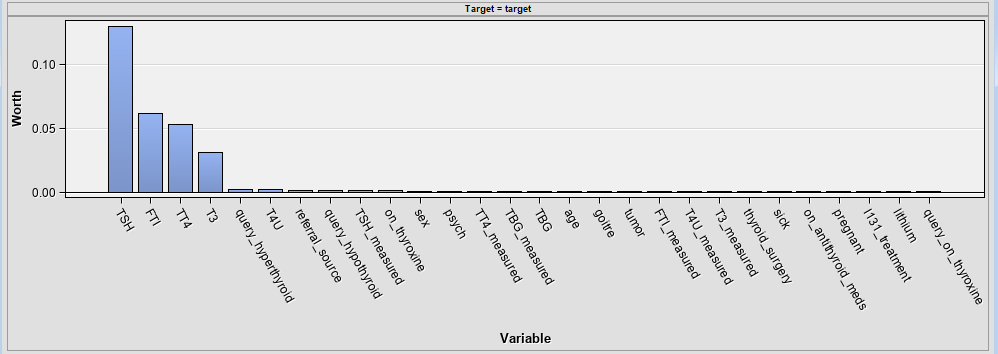
Different predictive techniques will be used such as classification algorithms where machine learning models will be applied in training data to classify thyroid. Also preprocessing of the levels will be done to segregate different levels of thyroid under standard levels. Decision tree classification will be used with different maximum depth and evaluated using different error metrics. Also the models will be compared and the best model will be chosen based on the least error in test data.

## Overview of the models built

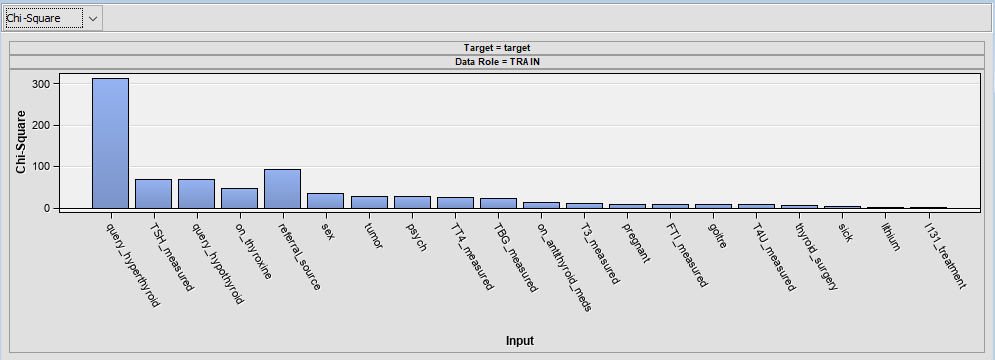


Different modelling techniques such as data partition, exploration of features and decision trees had been built and compared using model comparison. The decision trees are trained in the data using different maximum branches, maximum depth and impurity measures.

## Data Exploration



The effectiveness of the input variables are determined where TSH, FTI, TT4 and T3 are the most effective input variables in prediction of thyroid.

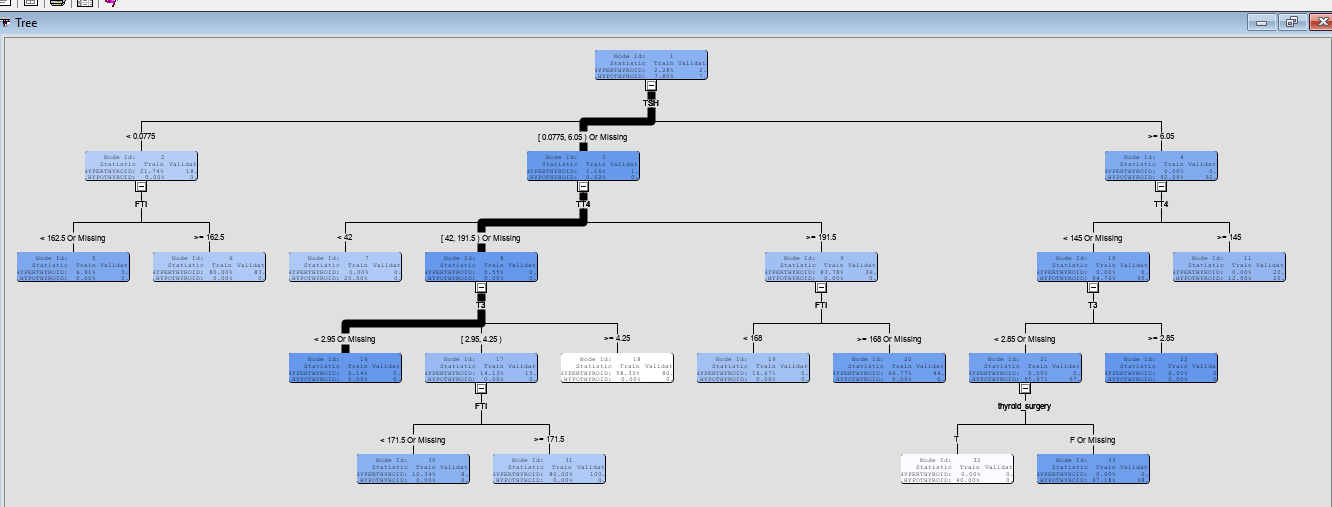


Chi Square test is performed to find out the effectiveness of categorical variables. The input variables such as query hyperthyroid, TSH measured, query Hypothyroid, on thyroxine are the important input categorical variables for prediction of thyroid.

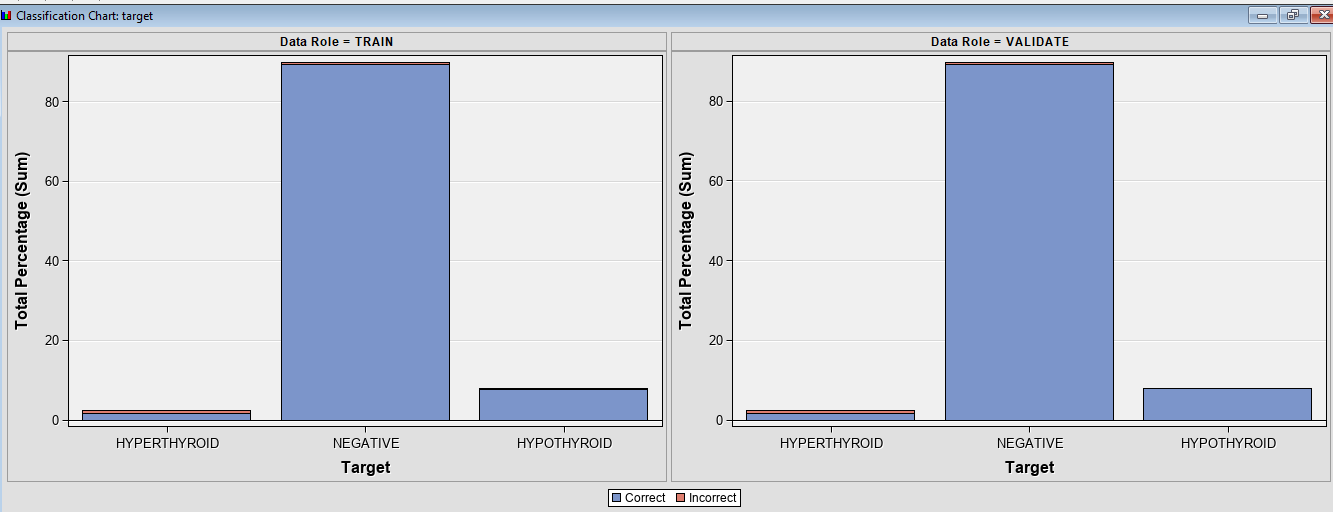
## Assessment of Models

### Maximum branch = 3, Max Depth = 6, Criterion-Entropy

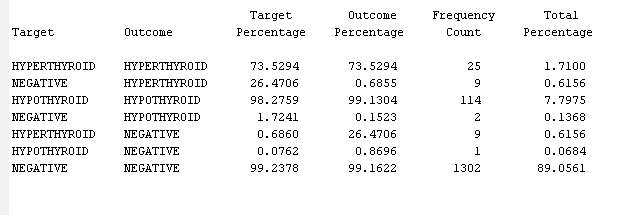
In the first tree, we have used 3 maximum branches and maximum depth of 6 with criterion as Gini.



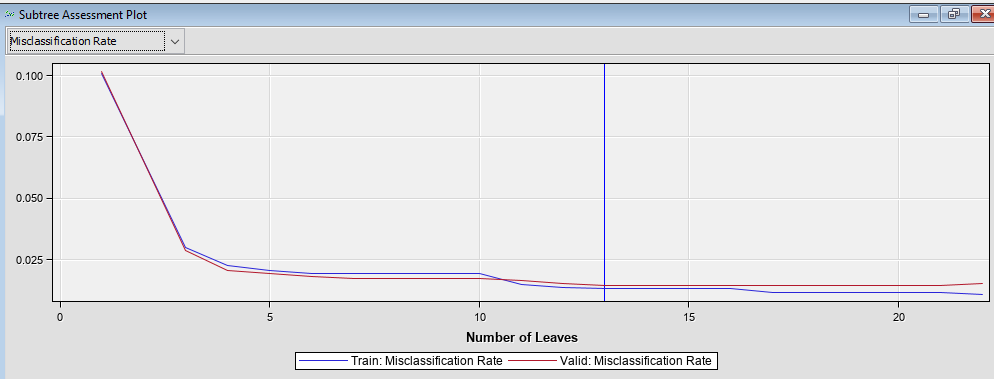
For splitting, TSH is used as the first variable. So TSH is considered as the most important input variables for classifying stages of thyroid.



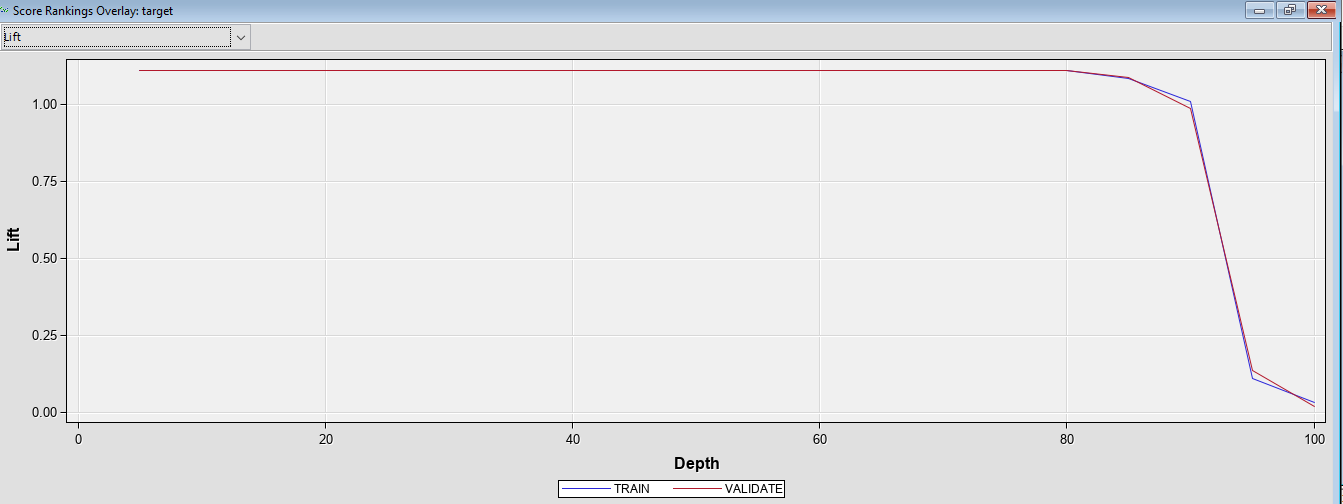
The chart shows the level of mis classification where we can see that negative samples are more present in both train and validation data. Decision tree gives few miss classification on all the classes which is seen from the plot.



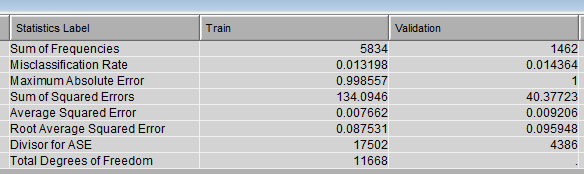
From the outcome, we can see there are some negative instances that are misclassified as hyperthyroid or hypothyroid and among the miss classification, major hyperthyroid instances are misclassified as negative. Only 1 Hypothyroid sample is misclassified as negative. So there is a good chance that negative samples can be often misclassified as hyperthyroid or hyperthyroid by the decision tree model.



At 13 leaves, the misclassification rate is lower in both train and valid samples. Also with the increase of leaves, we can see misclassification rate in valid sample is increasing higher at more number of leaves.



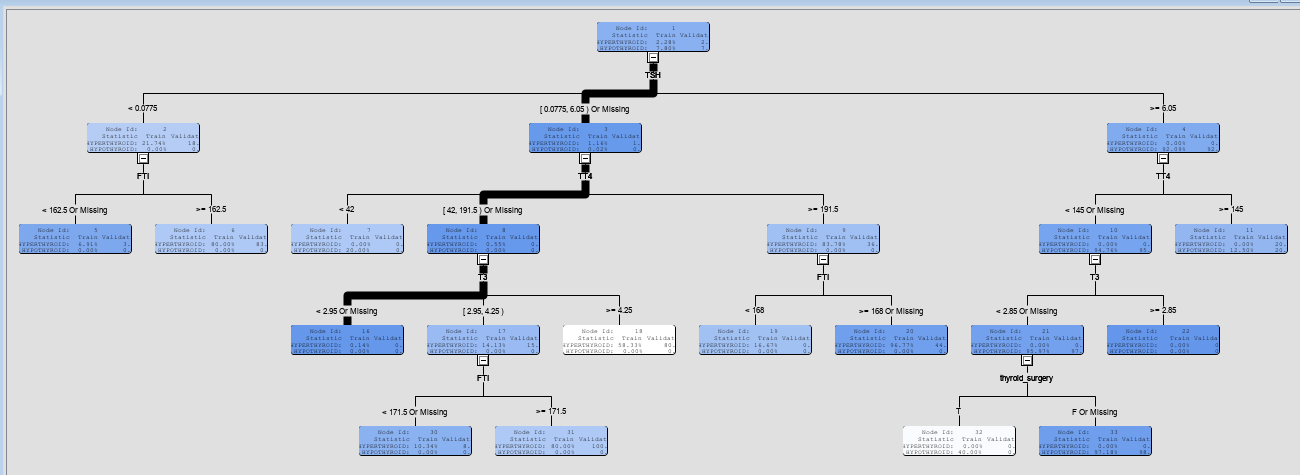
The lift shows that up to a depth of 90, the model shows a strong relationship and lower probability of predicting random samples.



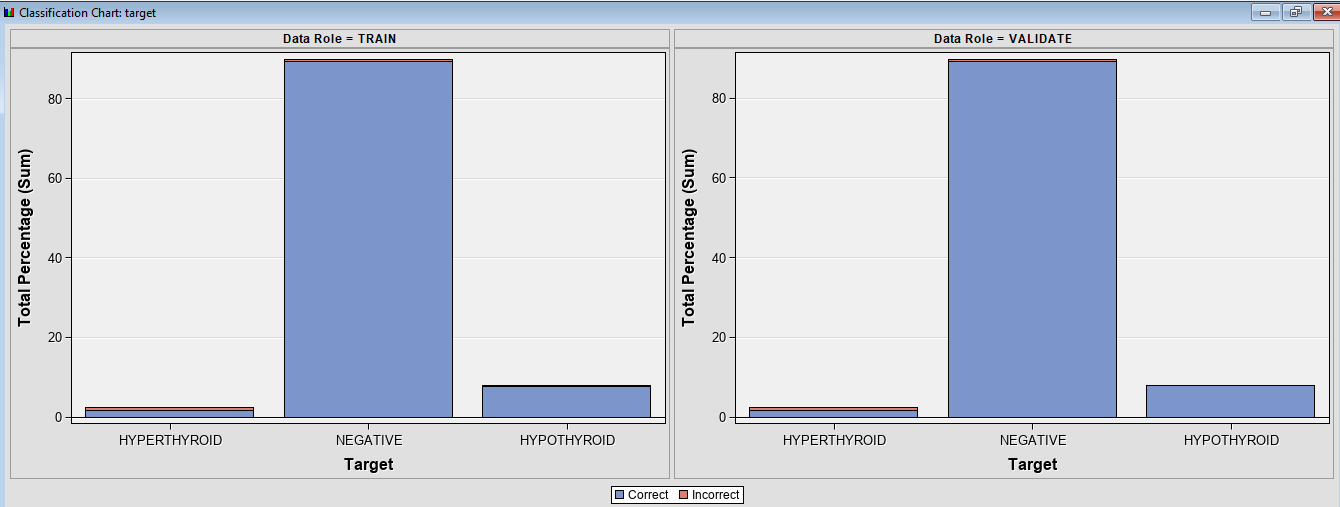
In the statistics, a higher misclassification rate in validation data shows that the model can give errors in classifying thyroid samples in real time data.

### Max Branch=5, Maximum Depth=10, Criterion - Entropy

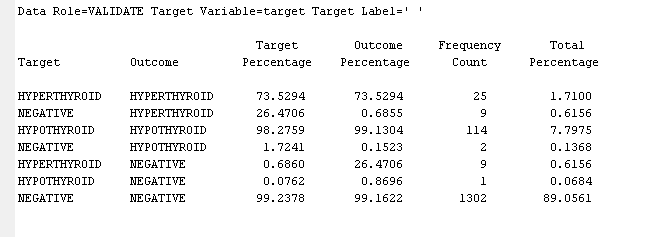
In this case, the maximum branch and maximum depth is increased and also the target criterion is changed to entropy.



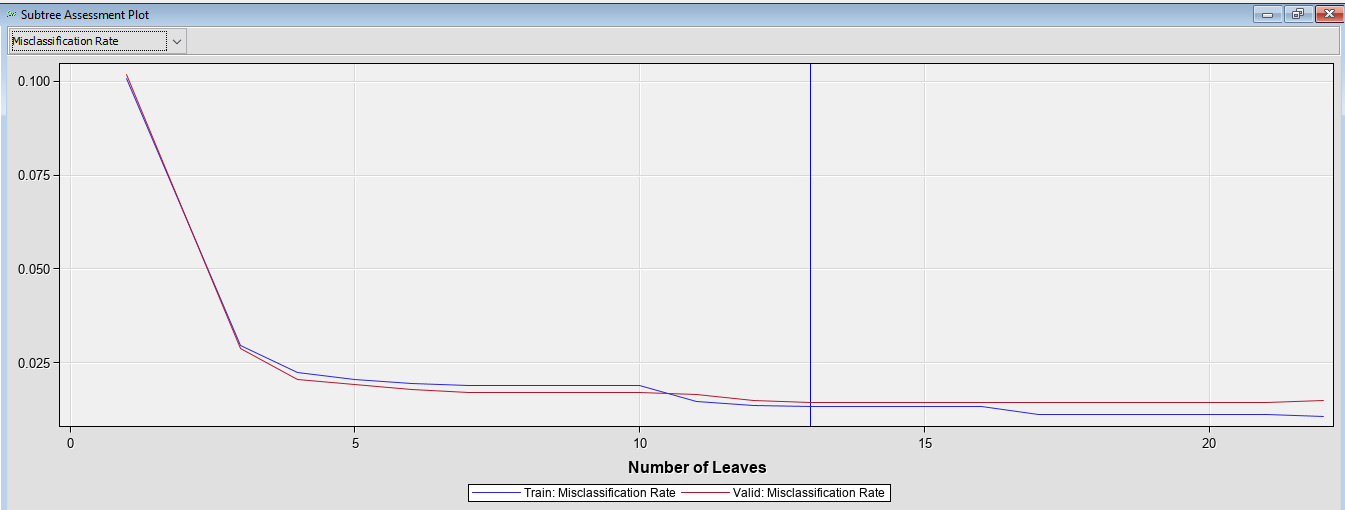
For the first split, TSH is used which is why it can be considered as the most important variable in classifying thyroid disease.



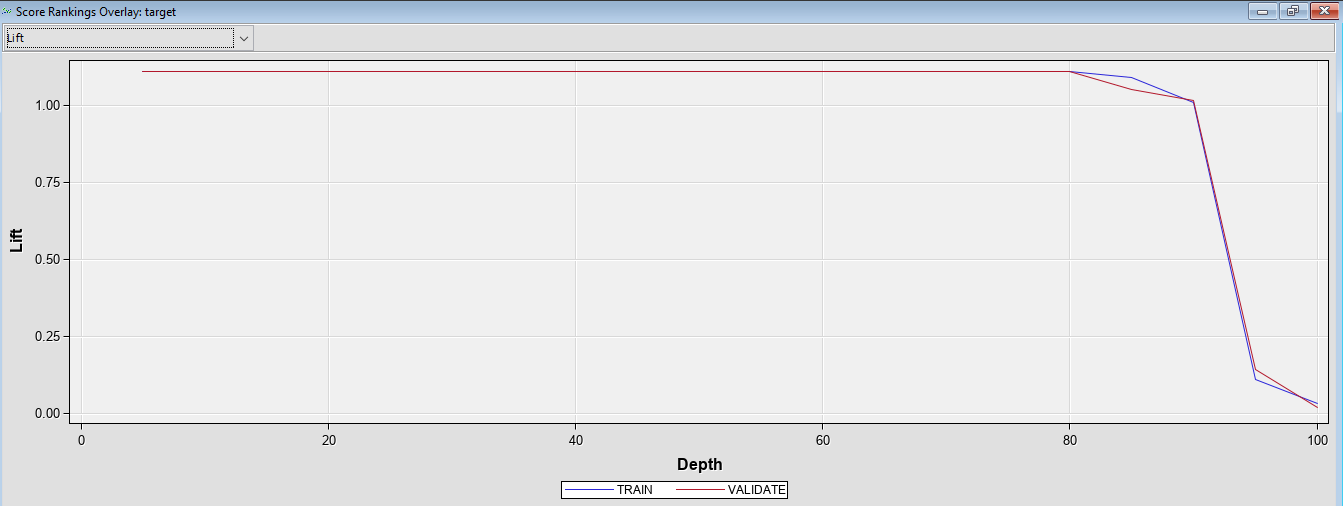
There are some instances that are not classified correctly in hyperthyroid and negative classes.



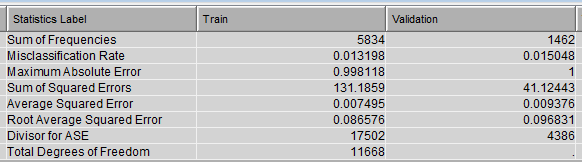
In hypothyroid, only 1 samples seem to be misclassified. This indicates this tree can classify hypothyroid with high accuracy.



The misclassification rate gives different values but at number of leaves which is 13, it gives the lowest misclassification rate in train and valid data. With increase of leaves, the misclassification rate is failing to converge which is a sign of overfitting.



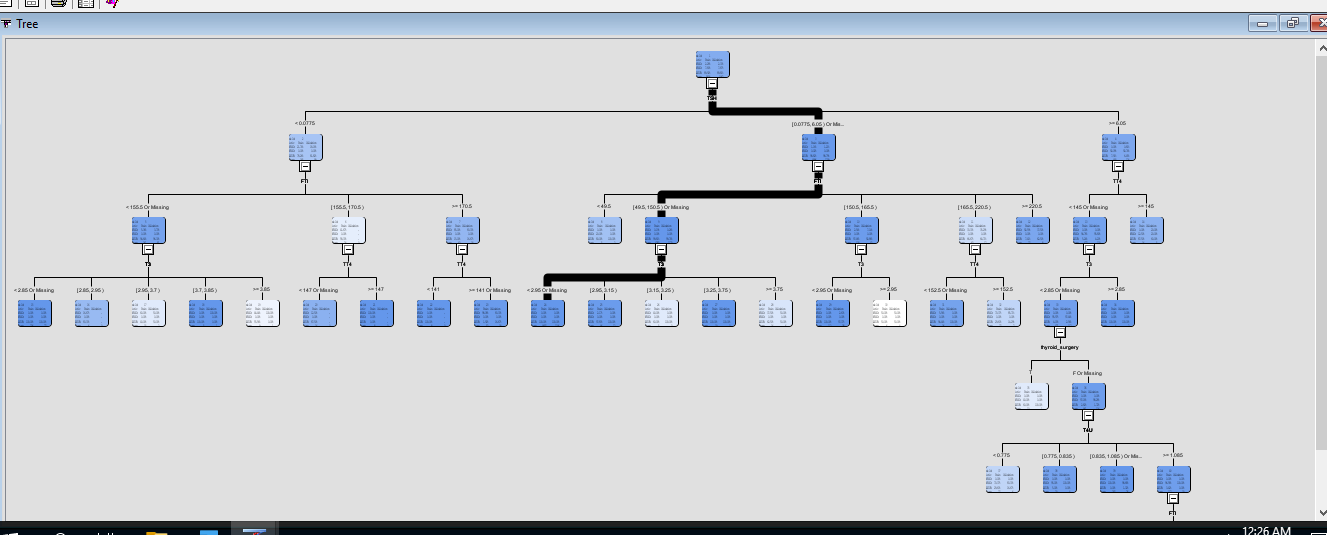
The lift shows same performance compared to the tree with lesser branches and depth. Hence we can expect similar performance from both.



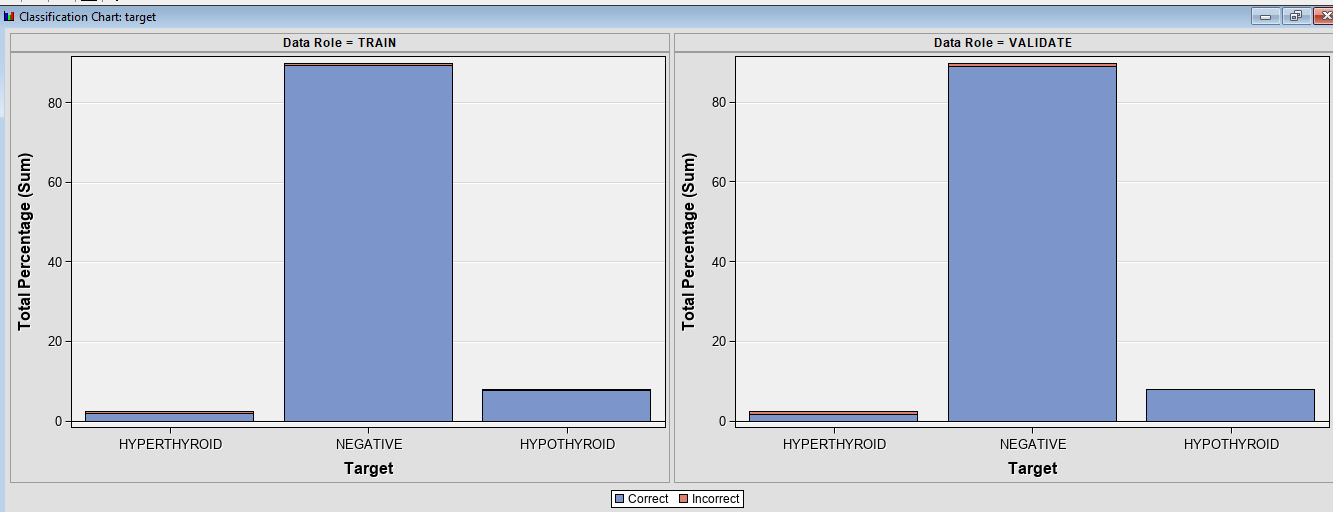
In valid data, the misclassification rate is higher than the last model. So with higher branches, the model can give more error in prediction.

### Maximum branch = 5, Max Depth = 6, Subtree Largest

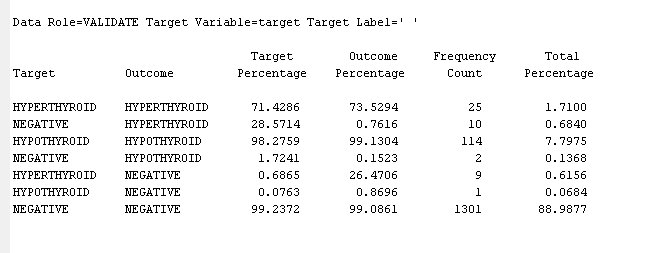
In this case, same branch is considered and we reduced the depth and set largest as the subtree assessment method.



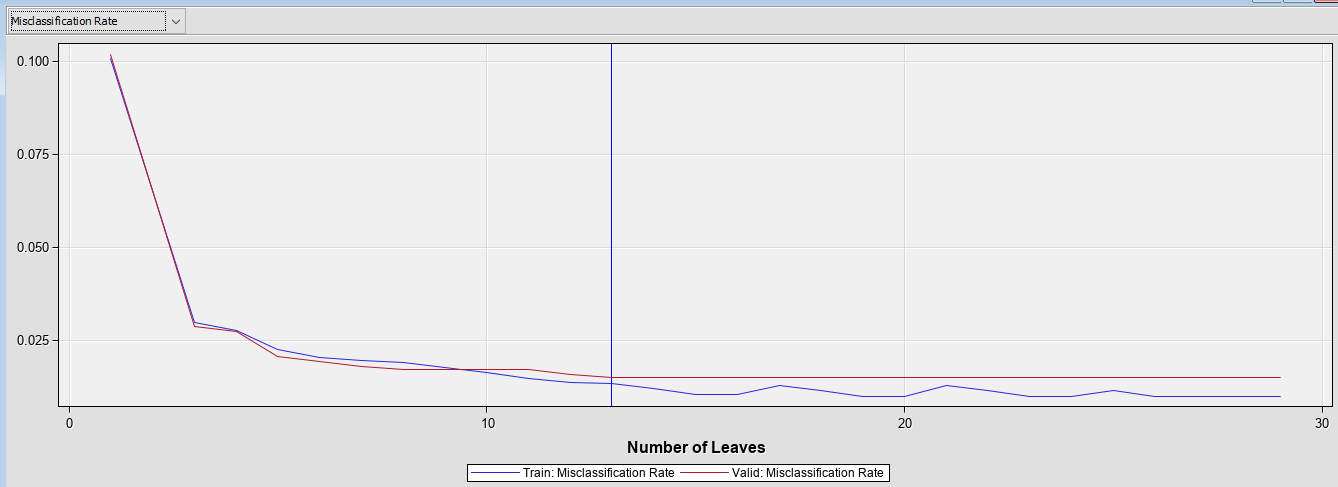
For the first split, TSH is used as the most important input variable for predicting thyroid disease. In all the case, TSH level in blood shows good importance.



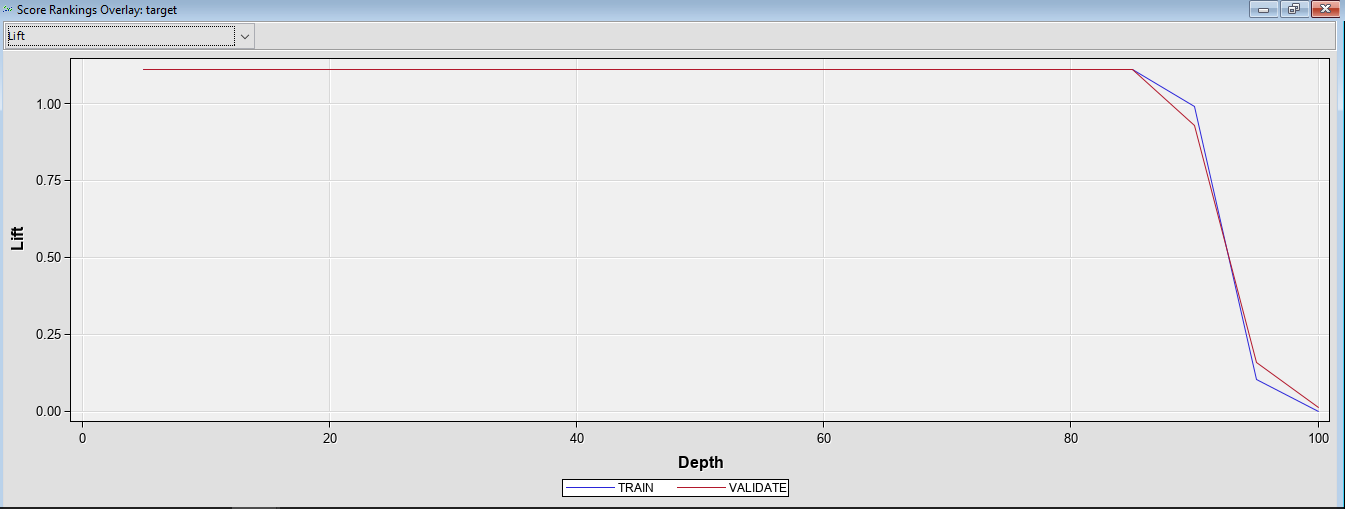
The plot shows in negative and hyperthyroid samples, we can see some samples are misclassified. In train samples of hypothyroid class, 1 sample got misclassified.



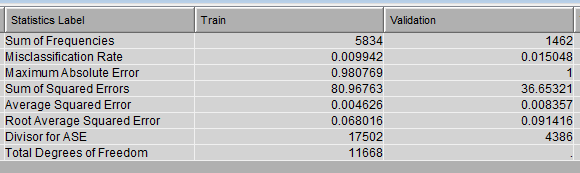
Most negative and hyperthyroid samples are misclassified but only 1 sample of hypothyroid are misclassified as negative. So this tree also gives good prediction rate in classifying hypothyroid samples.



The misclassification rate is also minimum in valid samples where at 13 leaves, it gives the best performance with the lowest error in train and valid data.



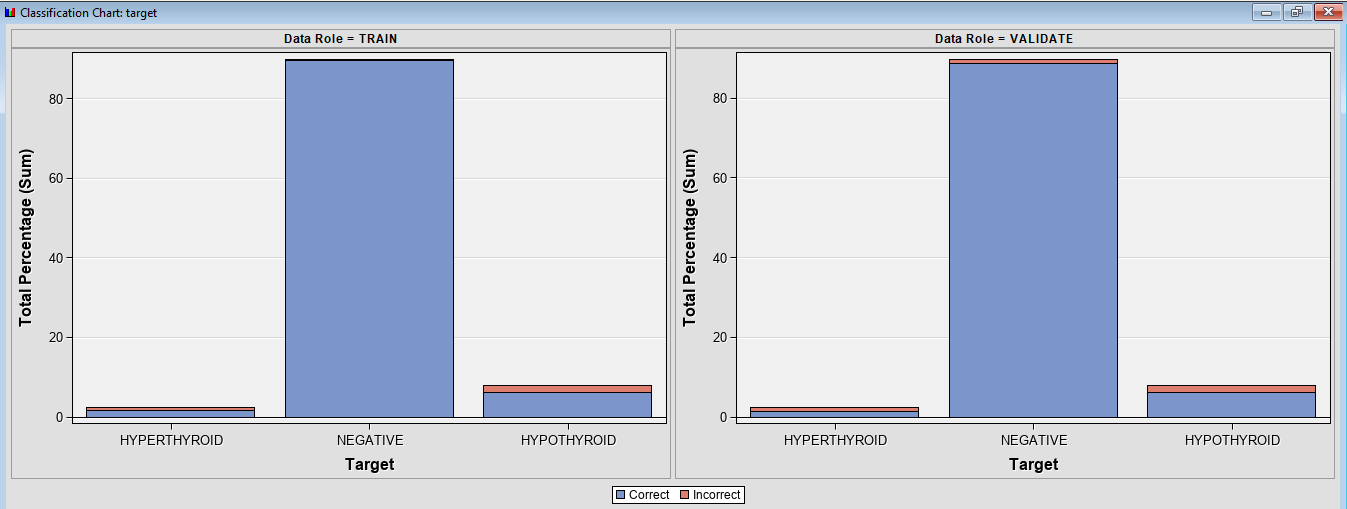
At a depth of 90, we can expect a good lift ratio higher than 1 and then the model becomes weaker.



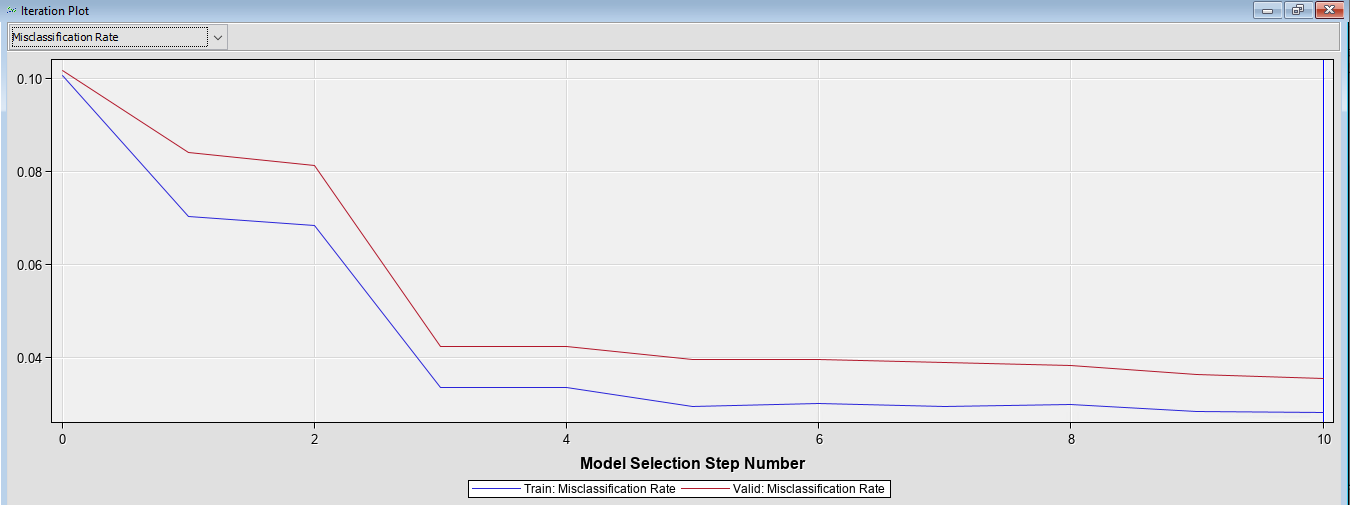
The validation error is similar compared to the last tree with 5 branches. So using subtree method as largest, we cannot improve the performance of classifying thyroid samples.

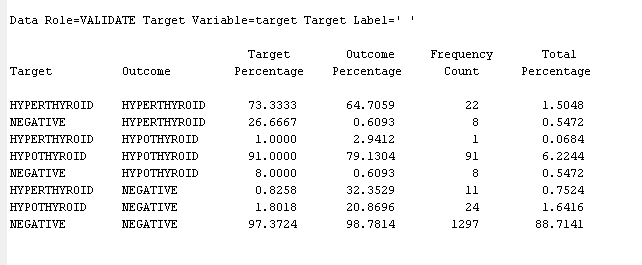
## Logistic Regression

In this case, forward regression is selected and trained in the data.

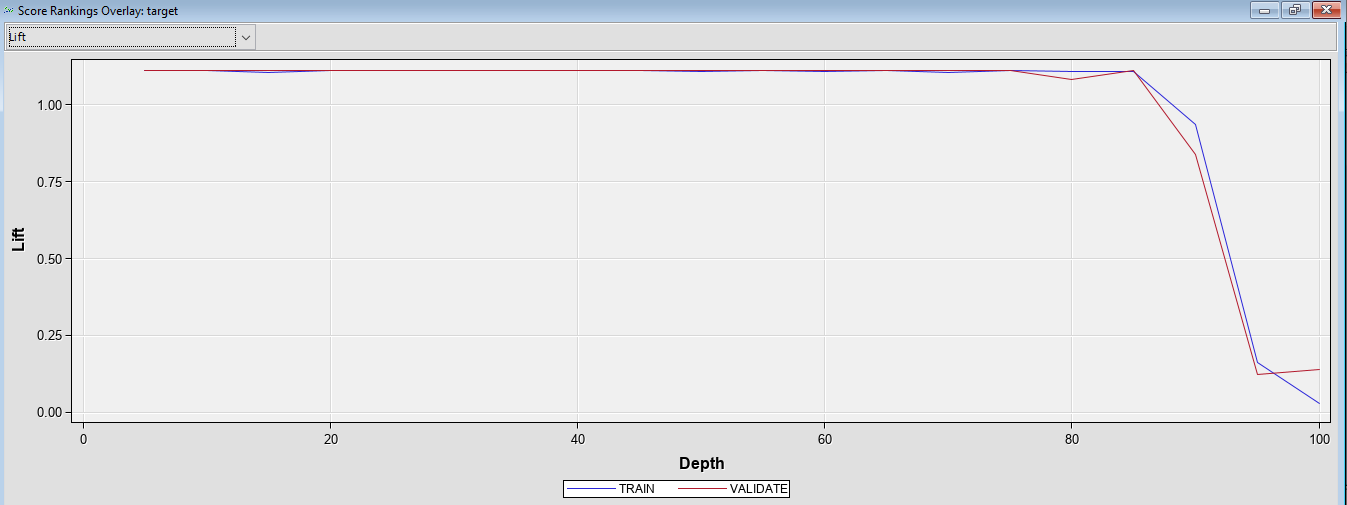


The chart shows that in hypothyroid samples, some observations are misclassified which implies logistic gives higher misclassification compared to decision tree models.

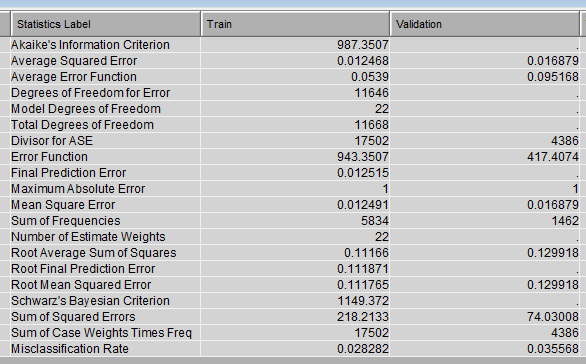


From the iteration plot, we can see at different step number logistic regression is giving an over fitting results with high misclassification rate in validation data. Logistic regression is not a good model for classifying thyroid as it will only perform poorly in test data while experimenting the model in real time application.

Nearly 24 hypothyroid samples are misclassified as negative. In decision tree, only 1 samples was misclassified. Considering such high misclassification rate, logistic regression model is not recommended.



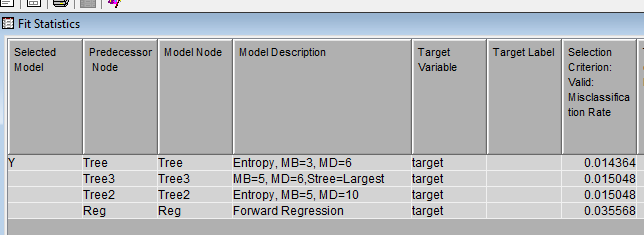
The lift ratio shows similar performance indicating logistic regression gives strong performance up to a depth of 90 and then the lift ratio goes below 1.

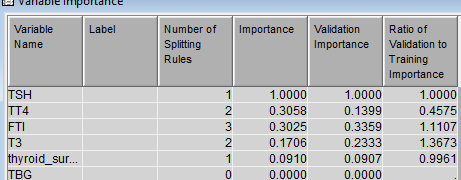


The misclassification rate in validation data is almost double than decision trees which make logistic regression the worst model for thyroid classification.

### Model Evaluation Criteria and Final Model

The mis classification rate in validation data is the evaluation criteria for all the models where the final model will be the one that will give the lowest misclassification rate from the model.

 Decision tree with maximum branch of 3, maximum depth of 6 and target criteria as entropy gave the best performance with the lowest mis classification rate in valid data. So this will be the final model for thyroid classification



From the best model, the variable importance is recorded where the variables such as TSH, TT4, FTI and T3 are the most important input considered for classifying thyroid disease.

## Interpretation of the findings and Recommendation

### Interpretation

* TSH is found to be significant for classifying thyroid disease
* Only 3 to 4 variables in the data are found give importance to the outcome
* Logistic regression gives worse results compared to decision trees
* When the branch of a trees is increased, the validation error increases
* With increase of leaves in the tree, the chance of overfitting increases

### Recommendation

* Pruning is necessary for the trees to combat overfitting problems in test data
* Neural Network techniques can be used for better accuracy of thyroid classification
* The model should be retrained on higher number of samples for indicating thyroid problems for bigger population
* Boosting techniques should be applied to enhance the performance compared to logistic regression
* Bias validation measures should be adopted to reduce bias in the models

# Summary and Lesson Learned

From the visualization, I have learnt that most of the samples are of female patients and also there are different types of thyroid present in the data which should be divided under standard thyroid levels.

I have learnt how to visualize the features and find the relation between two or more features using Tableau and how thyroid is dependent on different features such as age, gender, level of TSH, TT4, etc.

# Future Extension

1. In future work, more male patient should be updated to balance the data and more levels of thyroid patients should be added so that each level contain equal number of patients.
2. Neural network techniques can be used for better prediction of thyroid based on features and outliers handling can be done using different statistical analyses.
3. Standardization of features will help to scale the features in definite range and reduce the training time.