**THE FOREST FIRE DATSET**

## Introduction

The main purpose of this project report is to analyse and understand the concepts, basic theory and different algorithms and put in SAS Enterprise Miner and Guide to solve given data set by Identify the business problems. This report will include analysis of various algorithms to given data. Here we be using dataset from the region of Portugal to analyse the predicted area that gets affected by the various precipitation factors defined in the report will use recent real-world data, collected from the northeast region of Portugal, with the aim of predicting the burned area (or size) of forest fires. Several experiments were carried out by considering five DM techniques (i.e. multiple regression, DT, RF, NN and SVM) and four feature selection setups (i.e. using spatial, temporal, the FWI system and meteorological data). The proposed solution includes only four weather variables (i.e. rain, wind, temperature and humidity) along with the predictive and descriptive analysis of dataset to predict the major factors and outliers in data. Such knowledge is particularly useful for fire management decision support (e.g. resource planning).

## Business Problems Addressed

The SAS approach to AI is to augment human efforts through algorithms and automation. The problem statement here is to formulate the factors affecting the burning area.

## Data

The forest Fire Weather Index (FWI) is the Canadian system for rating fire danger and it includes six components (Figure 1) : Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Although different scales are used for each of the FWI elements, high values suggest more severe burning conditions. Also, the fuel moisture codes require a memory (time lag) of past weather conditions: 16 hours for FFMC, 12 days for DMC and 52 days for DC.

The data set is about forest fires, the total number of samples are 512 and it has 13 attributes which Interval variables are:

1. X-axis spatial coordinate (range: 1 to 9)
2. Y-axis spatial coordinate (range: 2 to 9)
3. FFMC (range: 18.7 to 96.20)
4. DMC (range: 1.1 to 291.3)
5. DC (range: 7.9 to 860.6)
6. ISI (range: 0.0 to 56.10)
7. Temperature (range: 2.2 to 33.30)
8. Relative Humidity (range: 15.0 to 100)
9. Rain in mm/m2 (range: 0.0/6.4)
10. Wind km/h (range: 0.40 to 9.40)
11. Area in ha (range: 0.00 to 1090.84)

Nominal variables are:

1. Month (January to December)
2. Day (Monday to Sunday)

## Data Understanding

While I was going through the data set, I understand that among these attributes area is target or dependent variable according to this data. Based on remaining variable we should easily identify which factor is affecting mostly an “area” to burn.

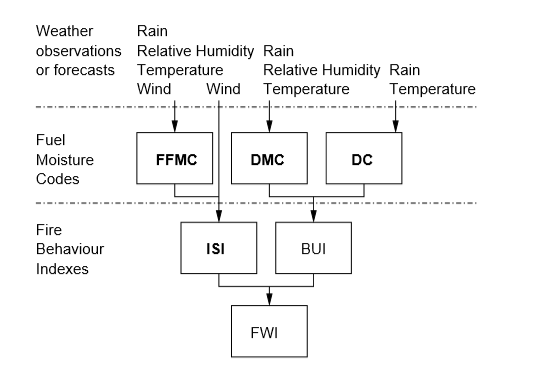


Figure 1

## Scatter plot to understand the data

Log area over temperature Log area over RH

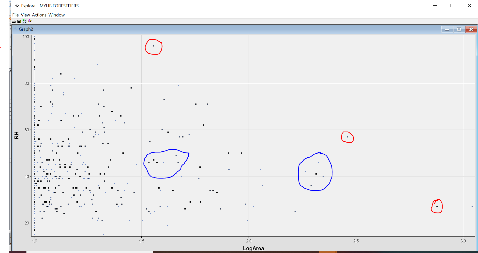
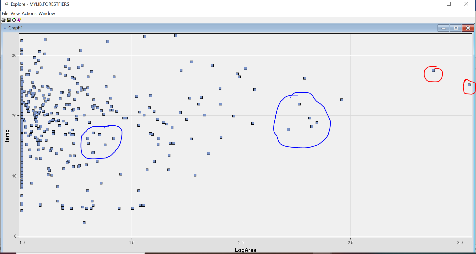
from the above figures we can see that data distributed outlier and cluster groups, the one circle with red color consider has outliers and with blue color are clusters.

Figure 2

## Exploratory data analysis

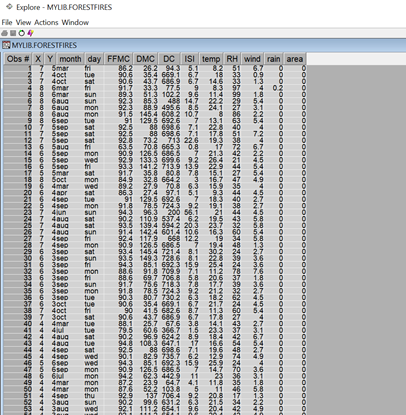


Figure 3

The data set here is divided into two groups the training data (70%) and validation data (30%) to predict the area accurately. There is no need to set probabilities in the dataset as already the target variable is same in both the datasets. The main objective is to predict the target variable and how independent variables effect the target variables using different classification models partitioned into train (70%) and validation (30%) subsets to provide more precise assessments. In this study, there is no need to set prior probabilities since the percentages of target events (fire =1 and fire =0) are pretty equal in both data sets. The high fluctuation between training and validation accuracy perhaps due to the small sample size of only about 512 observations, selected variables in classification models include: x and y coordinate, month of year, drought code index, relative humidity percentage, fine fuel moisture code index and outside rain. Of these variables, rain and x-y coordinates are also used to create treatment variables for incremental response analyses. While performing the initial exploration from the picture we can see that this data contains more than 20% “Zero Values” in Area and rain variable. The columns with zeros and missing values are ignored , but area cannot be ignored as it is the target variable.

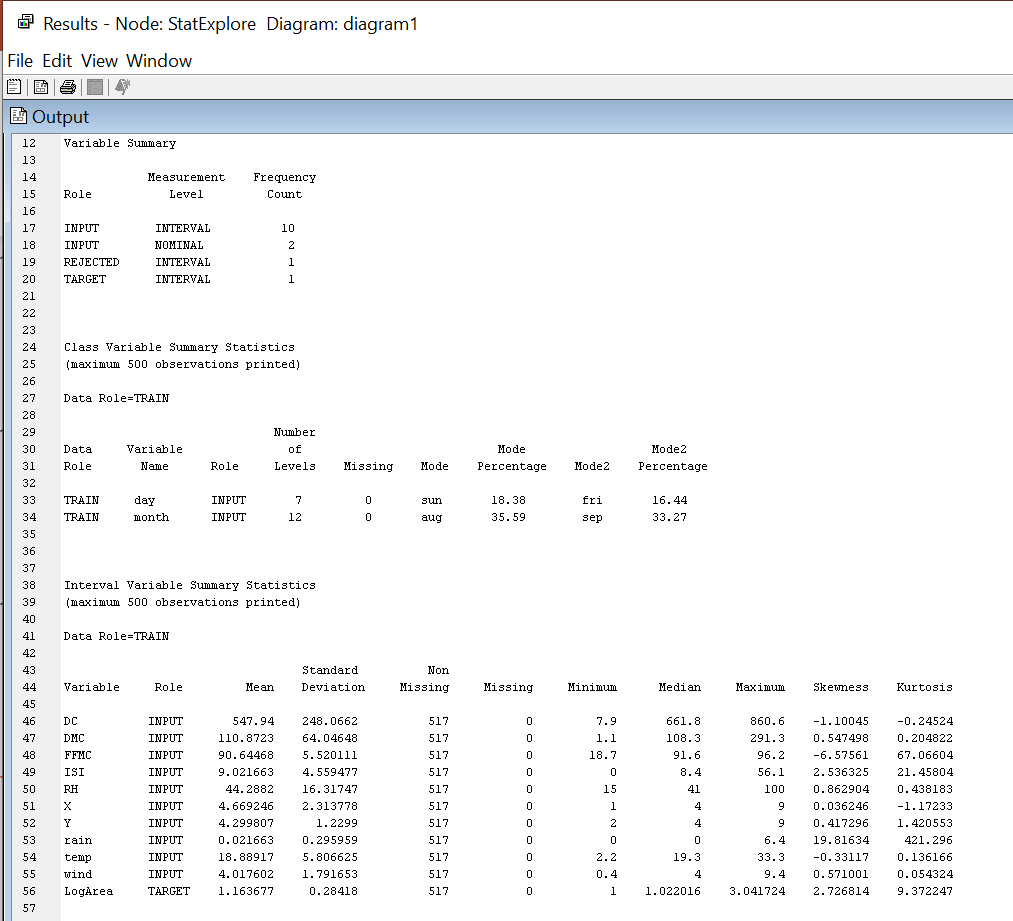


Figure 4

## Descriptive Modeling (Cluster analysis)

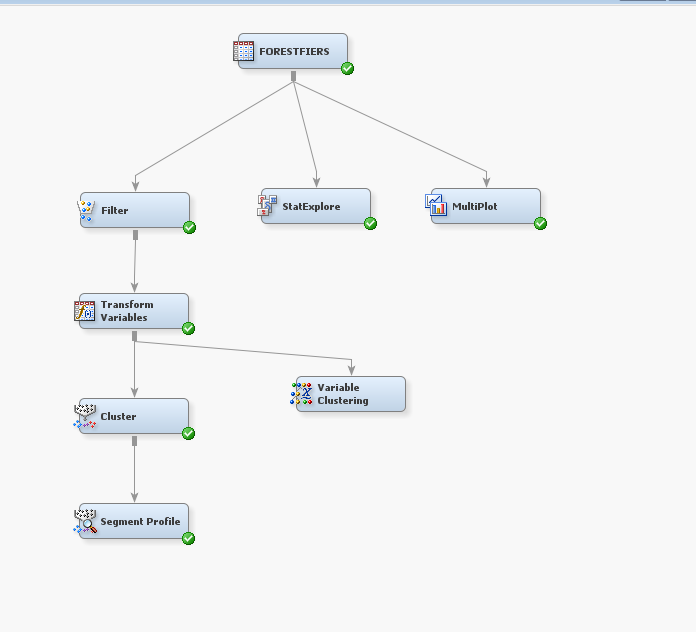


Figure 5

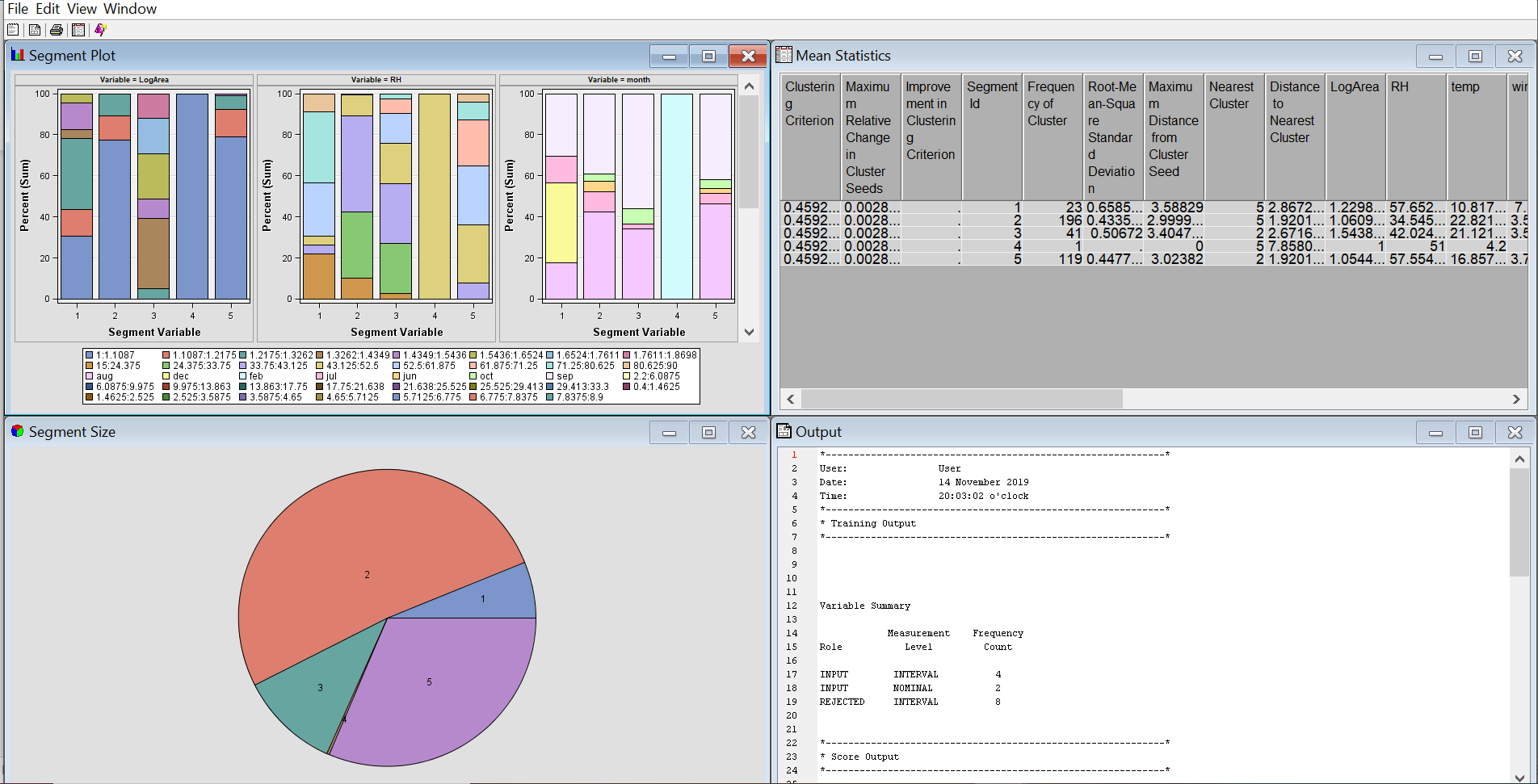


Figure 6

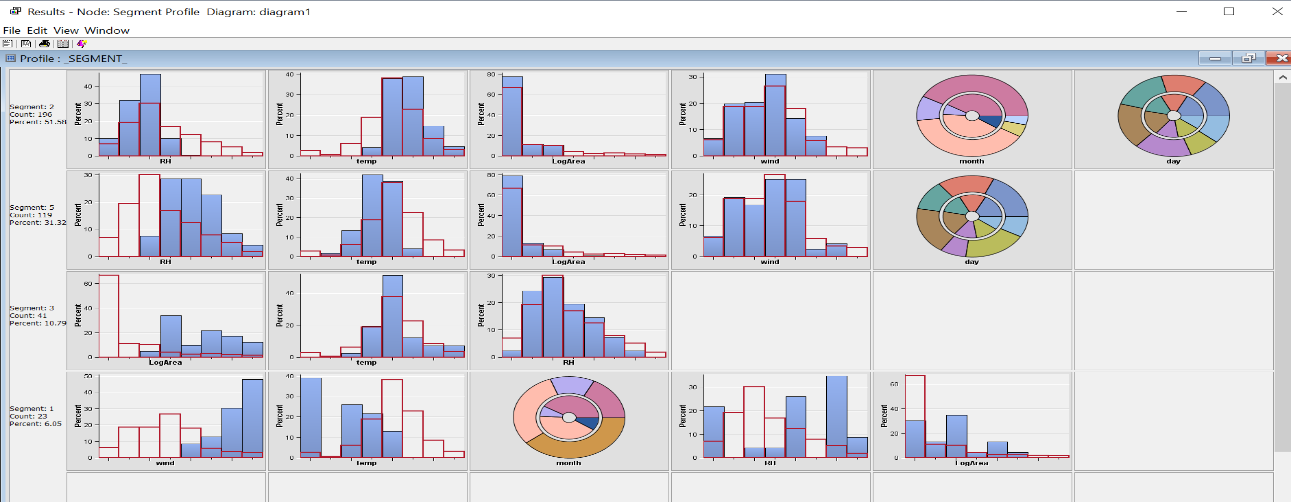
In the above figure we can see that our dataset forms 5 clusters.

Figure 7

## 

Cluster analysis creates a centroid around our dataset. The value of rh is highest in cluster one. And less is the similarity stronger is the link in clustering analysis. All the attributes are grouped according to similarity. In the above figure it shows how data is normalized to remove outliers in cluster analysis module. The R-squared value for the model is 0.89444 (>0.70). Hence, this a good fit model. The distance between the seed and observation of the first cluster distance is 18750, and the last cluster is the maximum value.

## Predictive Analysis

1. Decision Trees

## 

Figure 8

In the above analysis here decision tree reflects both categorical and continuous variables. At the top we have the root node. The root node predicts the leaf nodes using the defined rules. It finds relationship between our root and target variables. The DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form. This representation can interpret into a set of IF-THEN rules, which are easy to understand by humans. The RF is an ensemble of T unpruned DT, using random feature selection from bootstrap training samples. The RF predictor is built by averaging the outputs of the T trees. In general, RF exhibits a substantial improvement over a single DT.

## Linear Regression

## After finalizing the variables, we either perform linear or logistic regression depending upon our target variable. The fit statistics window tells us that which model would be selected ahead for prediction. We first want to run a cross-validation to assess the accuracy of the model with all the variables in using k-means techniques. This will split the dataset into train and test chunks in 89 different ways and evaluate the model error for each.

## 

Figure 9

Before fitting the models, some preprocessing was required by the MR, NN and SVM models. The nominal variables (i.e. discrete with more than two non-ordered values), such as the month and day, were transformed into a 1-of-C encoding. Also, for the NN and SVM methods, all attributes were standardized to a zero mean and one standard deviation. Next, the regression models were fitted. The MR parameters were optimized using a least squares algorithm, while the DT node split was adjusted for the reduction of the sum of squares. Regarding the remaining methods, the default parameters were adopted for the RF (e.g. T = 500).

## Methodology Applied

A regression dataset D is made up of k ∈ {1, ..., N} examples, each mapping an input vector (x k 1 , . . . , xk A) to a given target yk. The error is given by: ek = yk − ybk, where ybk represents the predicted value for the k input pattern. The overall performance is computed by a global metric, namely the Mean Absolute Deviation (MAD) and Root Mean Squared (RMSE), which can be computed as : MAD = 1/N × PN i=1 |yi − ybi | RMSE = qPN i=1 (yi − ybi) 2/N .

The lower values of these two values results in better prediction of the forest area. However, the RMSE is more sensitive to high errors. Another metric to compare our models is the regression error curve, which plots the error tolerance (x-axis), given in terms of the absolute deviation, versus the percentage of points predicted within the tolerance (y-axis). The ideal regressor should present a REC area close to 1.0. This report will consider five DM models. The Multiple Regression (MR) model is easy to interpret and this classical approach has been the widely used .The DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form. This representation can interpret into a set of IF-THEN rules, which are easy to understand by humans. The RF is an ensemble of T unpruned DT, using random feature selection from bootstrap training samples. The RF predictor is built by averaging the outputs of the T trees. In general, RF exhibits a substantial improvement over a single DT.

## Comparison and Evaluation Module

In this module the classification algorithms are compared to get the best goodness of fit model.

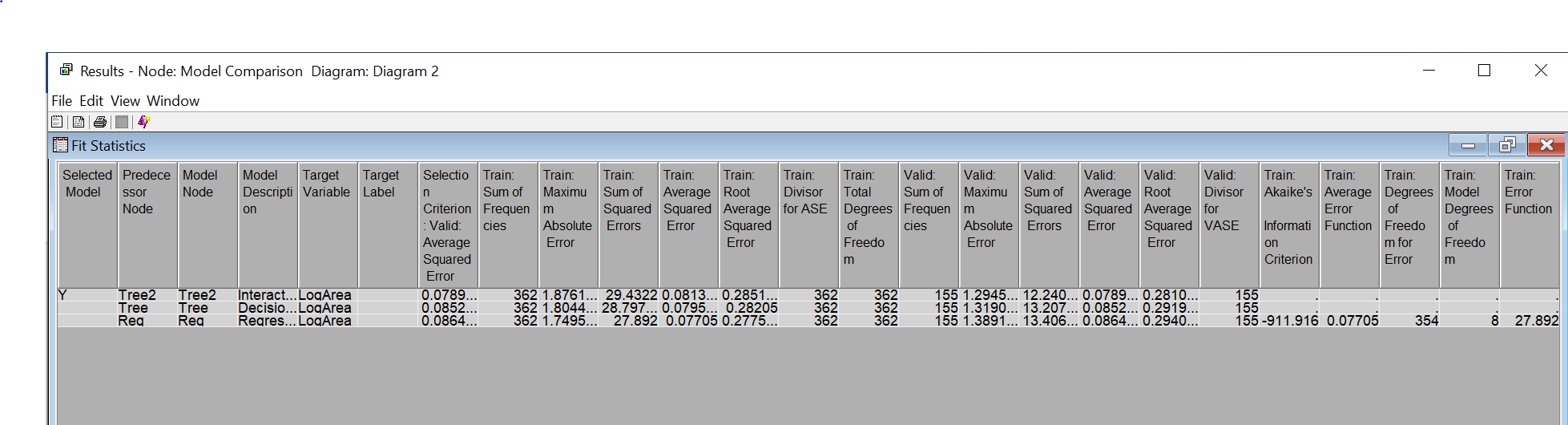


Figure 10

In this module we can see that root mean square value of the interactive decision tree algorithm is the least. The lower the value of root mean square better is the model. From the statistics window we can interpret that our interactive decision tree is the best model.