WIE3007 DATA MINING & WAREHOUSING

Chiew Kian Khoon U2005256/1
Loi Yi Hang U2005366/1
Sim Jia Hong U2005316/1
Ooi Xie Gee U2005379/1

PRESENTATION LINK

INTRODUCTION

This study revolves around a comprehensive dataset obtained from a gas multisensor device deployed in an Italian city.

The dataset, spanning a year from March 2004 to February 2005, captures hourly responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multi Sensor Device. The device was strategically positioned at road level in an area marked by significant pollution. This research represents a valuable opportunity to delve into the intricate interplay of air quality factors within the urban landscape.

SAMPLING

	MeasurementID	DateID	TimeID	ReadingID	CO_Concentration	NMHC_Concentration	Benzene_Concentration	NOx_Concentration	NO2_Concentration
0	1	1	1	1	2.6	150.0	11.9	166.0	113.0
1	2	2	2	2	2.0	112.0	9.4	103.0	92.0
2	3	3	3	3	2.2	88.0	9.0	131.0	114.0
3	4	4	4	4	2.2	80.0	9.2	172.0	122.0
4	5	5	5	5	1.6	51.0	6.5	131.0	116.0
...	...	...	...	...	...	...	...	...	...
9466	9467	9467	9467	9467	NaN	NaN	NaN	NaN	NaN
9467	9468	9468	9468	9468	NaN	NaN	NaN	NaN	NaN
9468	9469	9469	9469	9469	NaN	NaN	NaN	NaN	NaN
9469	9470	9470	9470	9470	NaN	NaN	NaN	NaN	NaN
9470	9471	9471	9471	9471	NaN	NaN	NaN	NaN	NaN

9471 rows × 23 columns

Diagram 5.1: Raw Data

- Original datasets can be very large, leading to high computational costs.
- Sampling is employed to reduce dataset size.

FEATURE TOOLS

In [8]:

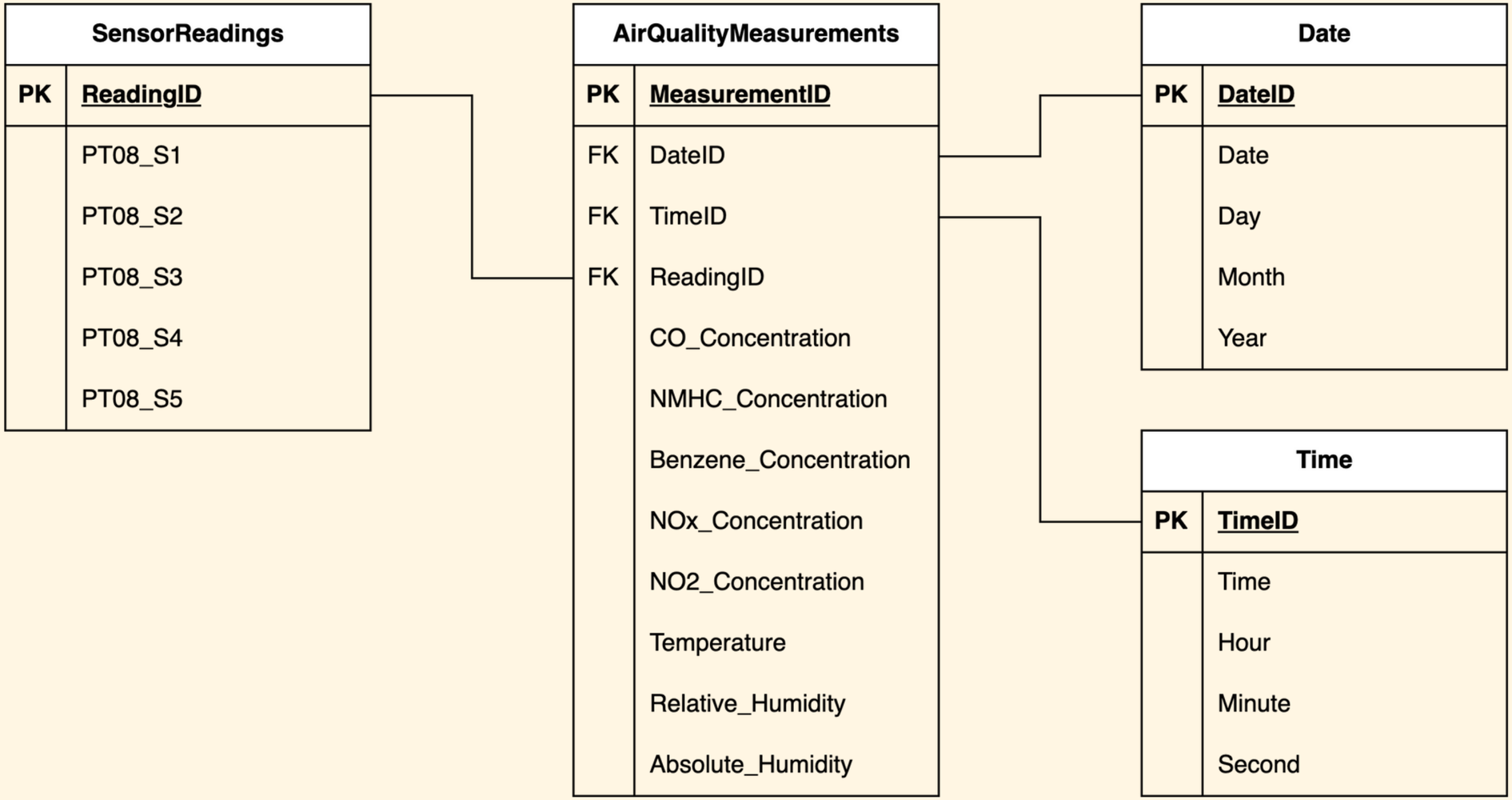
```
feature_matrix_SensorReadings, features_defs = ft.dfs(
    dataframes=dataframes,
    relationships=relationships,
    target_dataframe_name="SensorReadings",
)
feature_matrix_SensorReadings
```

Out[8]:

ReadingID	PT08_S1	PT08_S2	PT08_S3	PT08_S4	PT08_S5	COUNT(AirQuality)	MAX(AirQuality.TimeID)	MEAN(AirQuality.Time
1	1360	1046	1056	1692	1268	1	1.0	
2	1292	955	1174	1559	972	1	2.0	
3	1402	939	1140	1555	1074	1	3.0	
4	1376	948	1092	1584	1203	1	4.0	
5	1272	836	1205	1490	1110	1	5.0	
...	...	...	...	...	...	...	...	

- Creating entity sets
- Establishing relationships within the dataset
- Performing deep feature synthesis

STAR SCHEMA


SensorReadings
PK
ReadingID
PT08_S1
PT08_S2
PT08_S3
PT08_S4
PT08_S5
AirQualityMeasurements
PK
MeasurementID
FK
DateID
FK
TimeID
FK
ReadingID
CO_Concentration
NMHC_Concentration
Benzene_Concentration
NOx_Concentration
NO2_Concentration
Temperature
Relative_Humidity
Absolute_Humidity
Date
PK
DateID
Date
Day
Month
Year
Time
PK
TimeID
Time
Hour
Minute
Second

EXPLORE DATA ANALYSIS

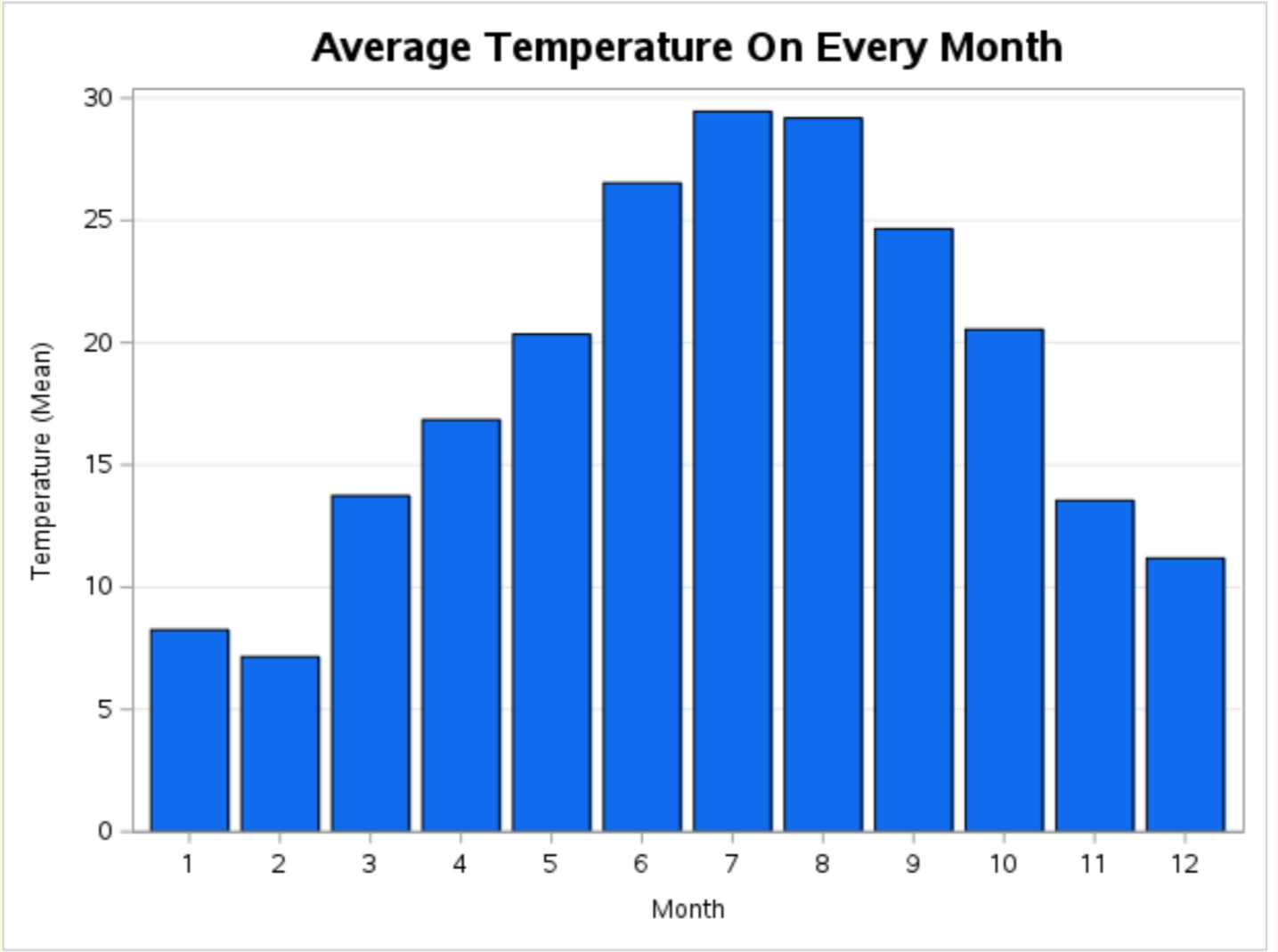


Diagram 7.1: Average Temperature Of Italy Base On Month

- July stands out with the highest temperature, peaking at approximately 29 degrees Celsius, while February records the lowest temperature at around 7 degrees Celsius.
- This clear distinction emphasizes the seasonal temperature fluctuations experienced in Italy, with summer months being notably warmer and winter months considerably cooler.

EXPLORE DATA ANALYSIS

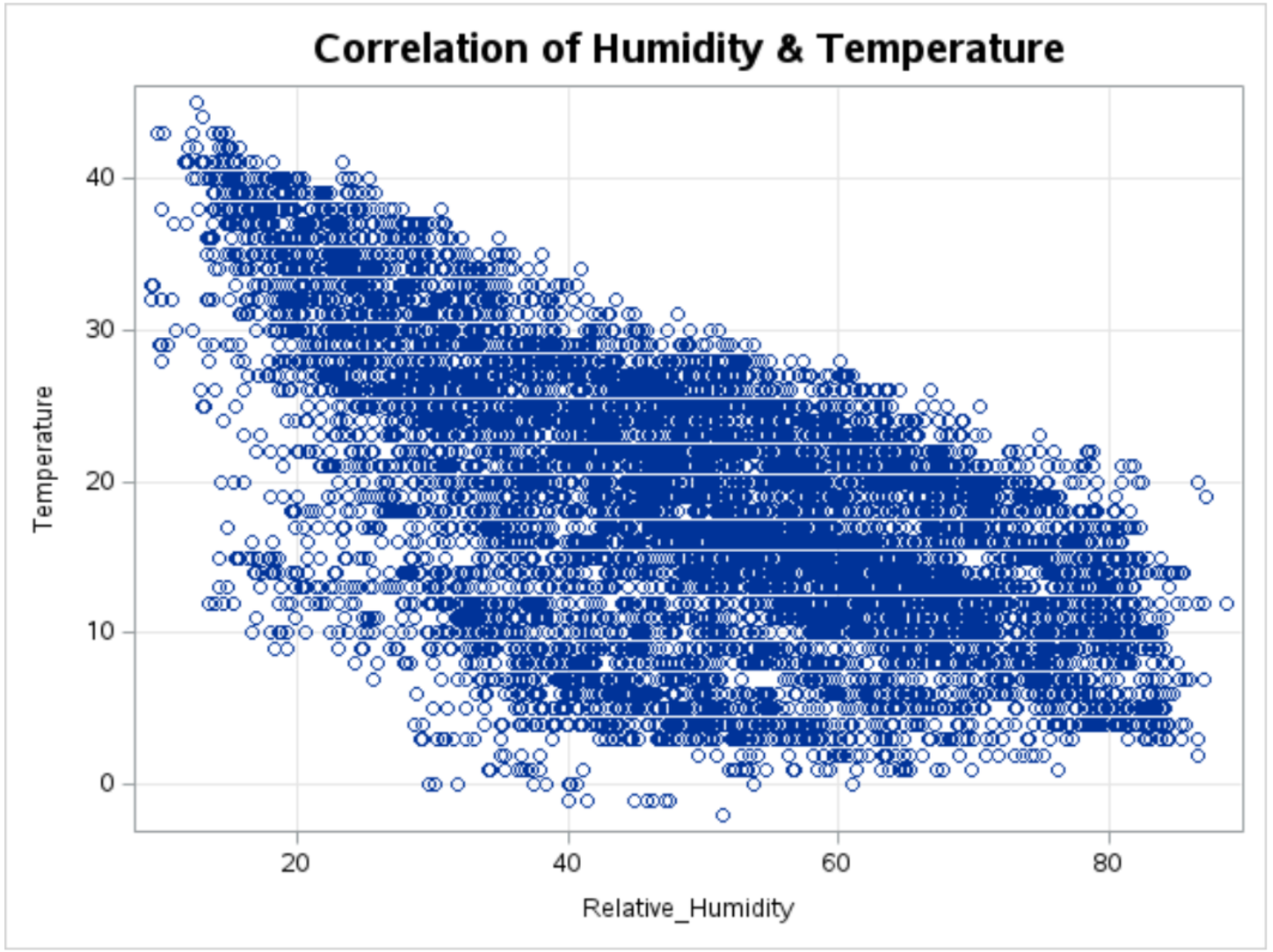


Diagram 7.2: Correlation of Humidity & Temperature

- This observation indicates an inverse relationship between humidity and temperature, suggesting that lower humidity levels are associated with cooler temperatures.

EXPLORE DATA ANALYSIS

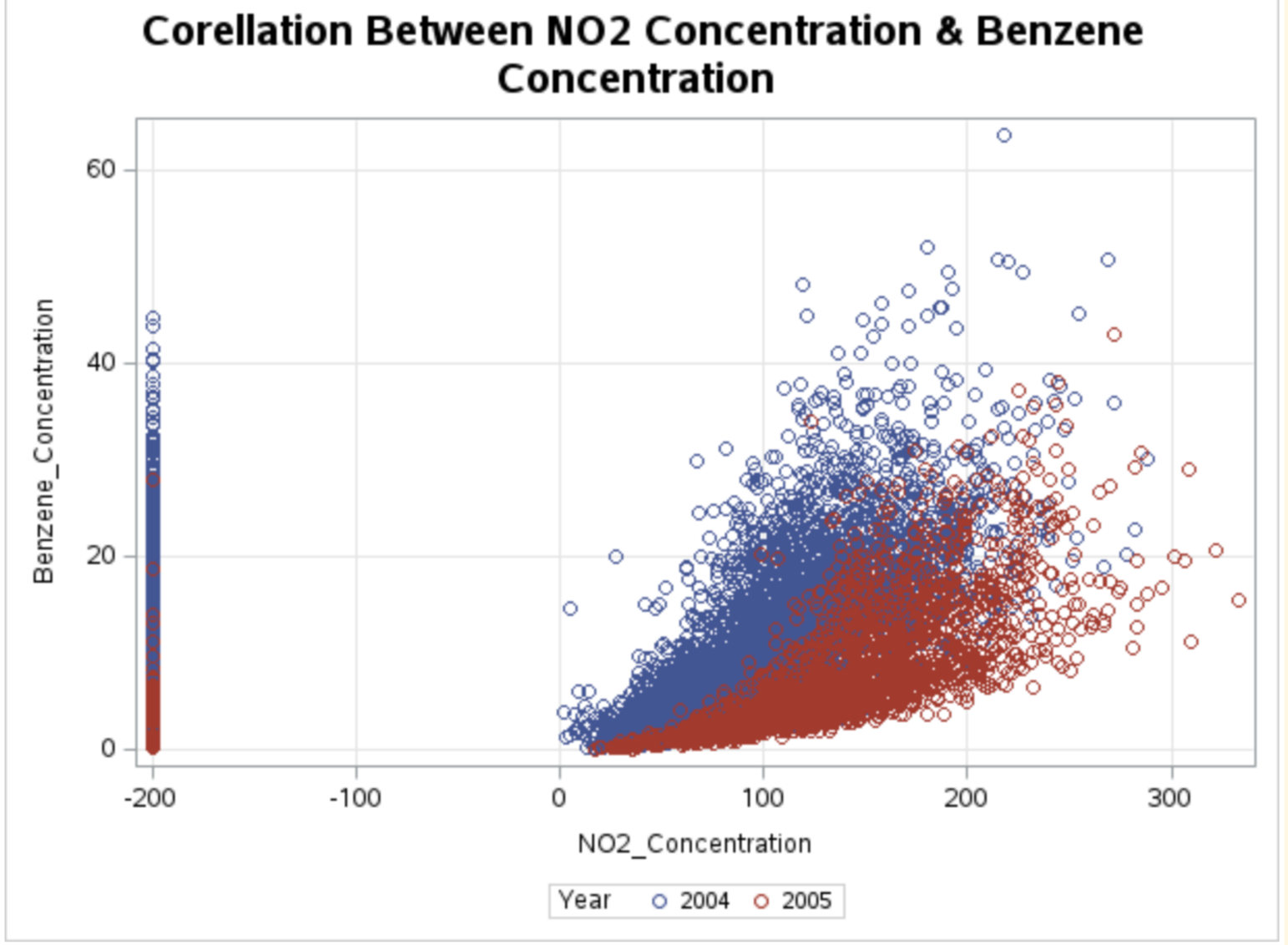


Diagram 7.3: Correlation Of NO2 & Benzene

- This observation suggests an overall reduction in the levels of these pollutants between the two years, indicating a potential improvement in air quality and environmental conditions during that period.

MODIFY

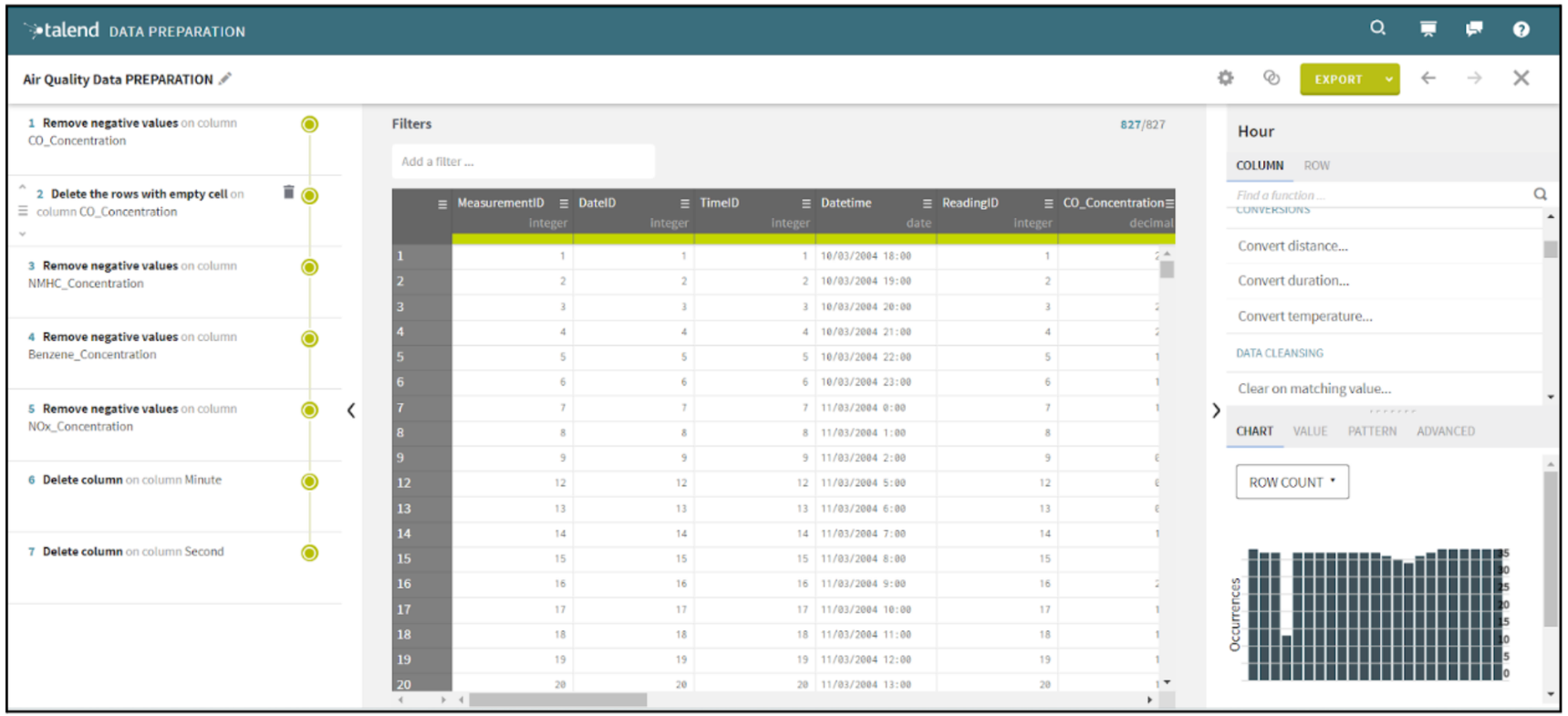
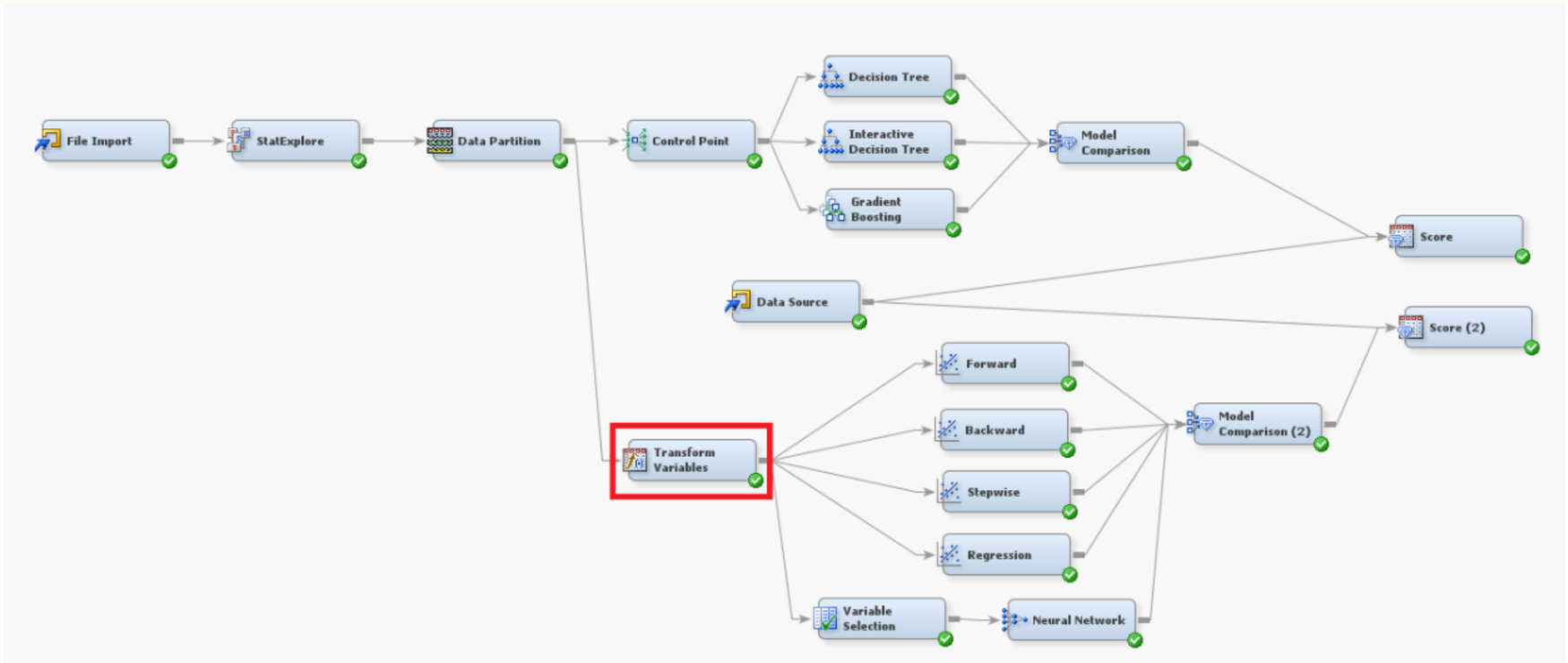


Diagram 8.1: Talend Data Preprocessing

1. Remove negative values on CO_Concentration column
2. Delete all the rows where consists of missing values
3. Remove negative values such as -200 on NMHC_Concentration column
4. Remove negative values on -200 Benzene_Concentration column
5. Remove negative values on -200 NOx_Concentration column
6. Delete unnecessary column Minute
7. Delete unnecessary column Second

MODIFY



Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	Absolute Humidity			579	0	0.4023	1.4852	0.832248	0.182761	0.760412	0.958562	
Input	Original	Benzene Concent...			579	0	0.5	38.4	10.6468	7.31004	0.955839	0.577603	
Input	Original	CO Concentration			579	0	0.3	8.1	2.339724	1.382382	0.979696	0.705202	
Input	Original	NMHC Concentra...			579	0	7	1084	228.6598	207.3925	1.445405	1.547807	
Input	Original	NO2 Concentration			579	0	20	196	100.0363	31.48148	0.112398	-0.12285	
Input	Original	NOx Concentration			579	0	12	478	143.342	81.31736	0.851289	0.433524	
Input	Original	PT08 S1			579	0	753	1975	1204.801	240.7075	0.515333	-0.25945	
Input	Original	PT08 S2			579	0	448	1737	962.0691	264.1464	0.345637	-0.49448	
Input	Original	PT08 S3			579	0	494	1935	965.9119	264.5125	0.81433	0.652152	
Input	Original	PT08 S4			579	0	955	2665	1595.301	300.9163	0.606134	0.032453	
Input	Original	PT08 S5			579	0	263	2359	1036.238	403.8245	0.316521	-0.40933	
Input	Original	Relative Humidity			579	0	14.9	83.2	49.13126	15.48082	0.014247	-0.9003	
Output	Computed	LG10 PT08 S1	log10(PT08 S1 +...		579	0	2.877371	3.295787	3.072823	0.085508	0.125483	-0.72195	Transformed PT0...
Output	Computed	LG10 PT08 S2	log10(PT08 S2 +...		579	0	2.652246	3.24005	2.966926	0.122097	-0.21199	-0.60217	Transformed PT0...
Output	Computed	LG10 PT08 S3	log10(PT08 S3 +...		579	0	2.694605	3.286905	2.96992	0.115352	0.160465	-0.45519	Transformed PT0...
Output	Computed	LG10 PT08 S4	log10(PT08 S4 +...		579	0	2.980458	3.42586	3.195644	0.080238	0.180058	-0.46249	Transformed PT0...
Output	Computed	LG10 PT08 S5	log10(PT08 S5 +...		579	0	2.421604	3.372912	2.979031	0.186829	-0.55778	-0.26754	Transformed PT0...
Output	Computed	OPT Absolute Hu...	Optimal Binning(4)	3	.	0	.	.	.	.	.	.	Transformed Abso...
Output	Computed	OPT Benzene C...	Optimal Binning(4)	3	.	0	.	.	.	.	.	.	Transformed Ben...
Output	Computed	OPT CO Concen...	Optimal Binning(4)	3	.	0	.	.	.	.	.	.	Transformed CO ...
Output	Computed	OPT NMHC Con...	Optimal Binning(4)	3	.	0	.	.	.	.	.	.	Transformed NMH...
Output	Computed	OPT NO2 Conce...	Optimal Binning(4)	3	.	0	.	.	.	.	.	.	Transformed NO2...
Output	Computed	OPT NOx Conce...	Optimal Binning(4)	2	.	0	.	.	.	.	.	.	Transformed NOx...
Output	Computed	OPT Relative Hu...	Optimal Binning(4)	4	.	0	.	.	.	.	.	.	Transformed Rela...

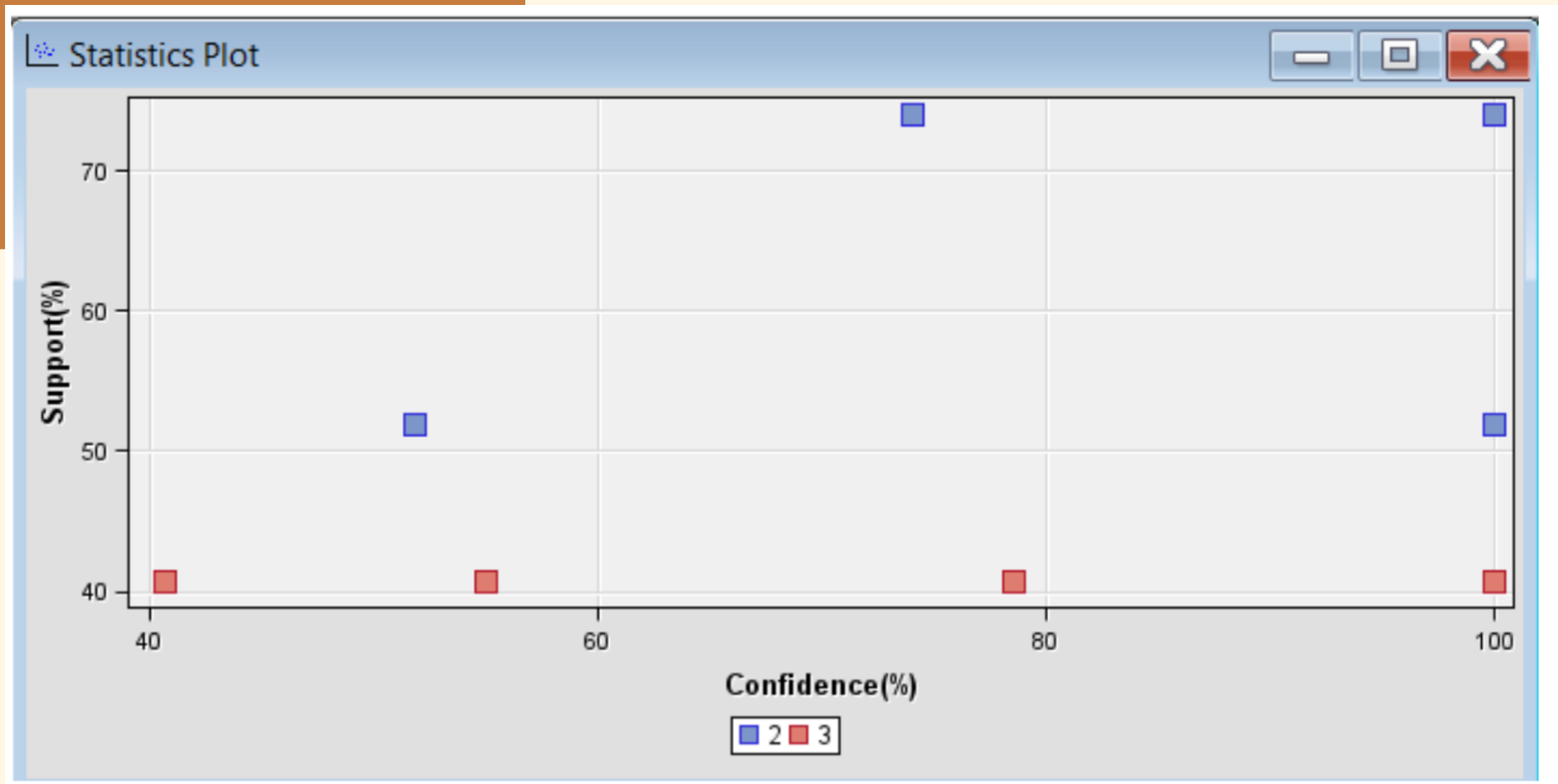
ASSOCIATION RULE

Diagram 9.1.1: Association Rule Diagram

Rule Description

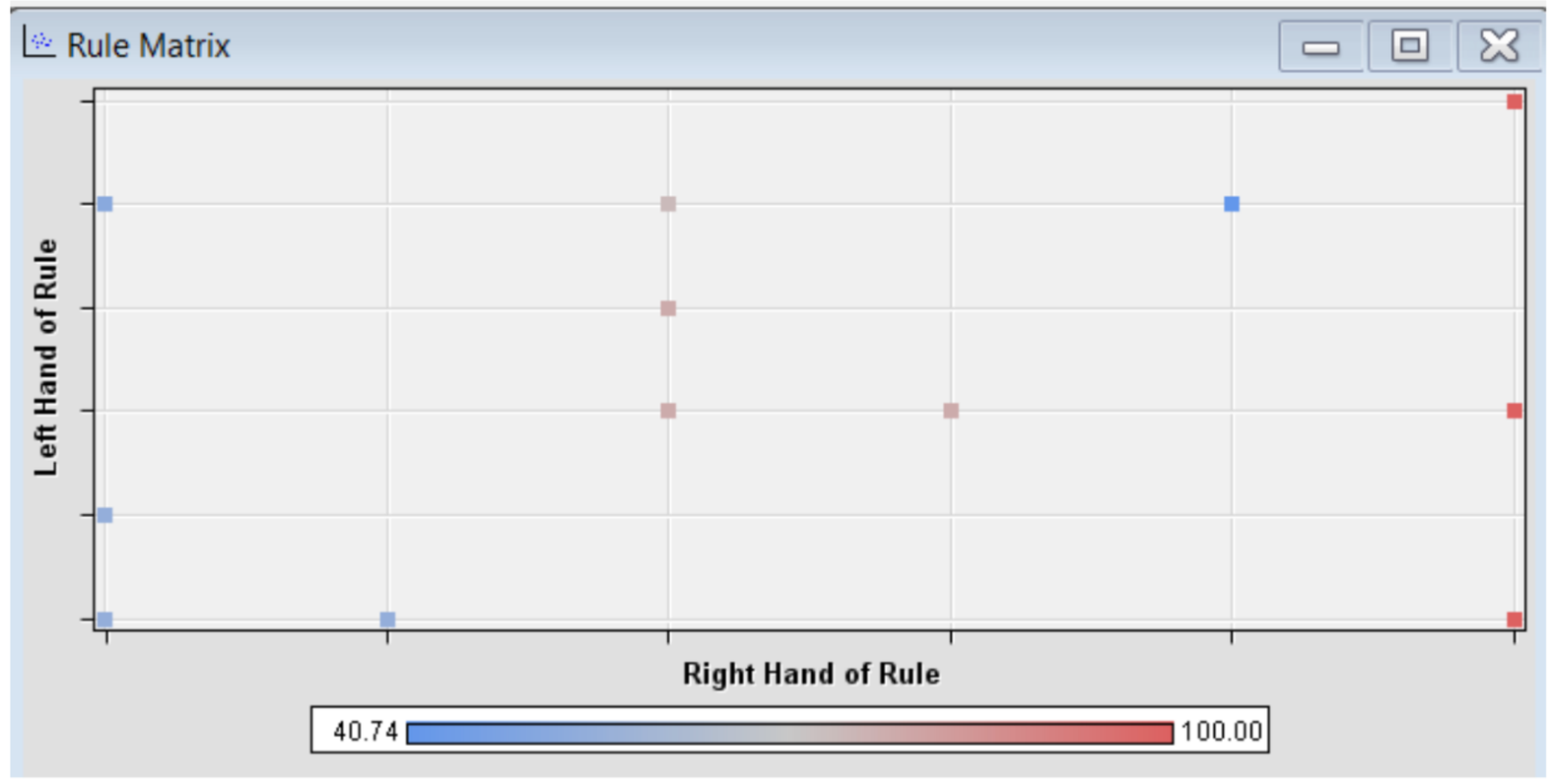
Map	Rule
RULE1	High ==> Low
RULE2	Mid & High ==> Low
RULE3	High ==> Mid & Low
RULE4	Low ==> High
RULE5	Mid & Low ==> High
RULE6	Low ==> Mid & High
RULE7	Mid ==> High
RULE8	Mid ==> Low
RULE9	Mid ==> Low & High
RULE10	High ==> Mid
RULE11	Low ==> Mid
RULE12	Low & High ==> Mid

ASSOCIATION RULE



- X-axis representing rule confidence and the Y-axis representing rule support.
- Each point on the graph corresponds to a specific association rule. Notably, a standout rule appears in the top right corner, symbolizing a robust rule with both high support and confidence.
- This rule is generally regarded as the most trustworthy. Conversely, the rule found in the bottom-left corner has the lowest support and confidence, suggesting a less reliable association rule.

ASSOCIATION RULE



- Utilizing this rule matrix facilitates a thorough comprehension of the patterns and connections between the components on the left-hand side and the results on the right-hand side within association rules.
- By referring to the results, we can identify that the right-hand side of the rule consists of 'Mid,' 'Low,' 'Mid & Low,' 'High,' 'Mid & High,' and 'Low & High,' while the left-hand side of the rule consists of 'Mid,' 'Mid & High,' 'High,' 'Mid & Low,' and 'Low & High.'

ASSOCIATION RULE

Diagram 9.1.5: Association Rule Line Plot

- The green line (confidence) exhibits a sudden drop from 80% to 40% but eventually achieves consistency at 100%. This indicates that, in the end, the presence of the antecedent guarantees the presence of the consequent

ASSOCIATION RULE

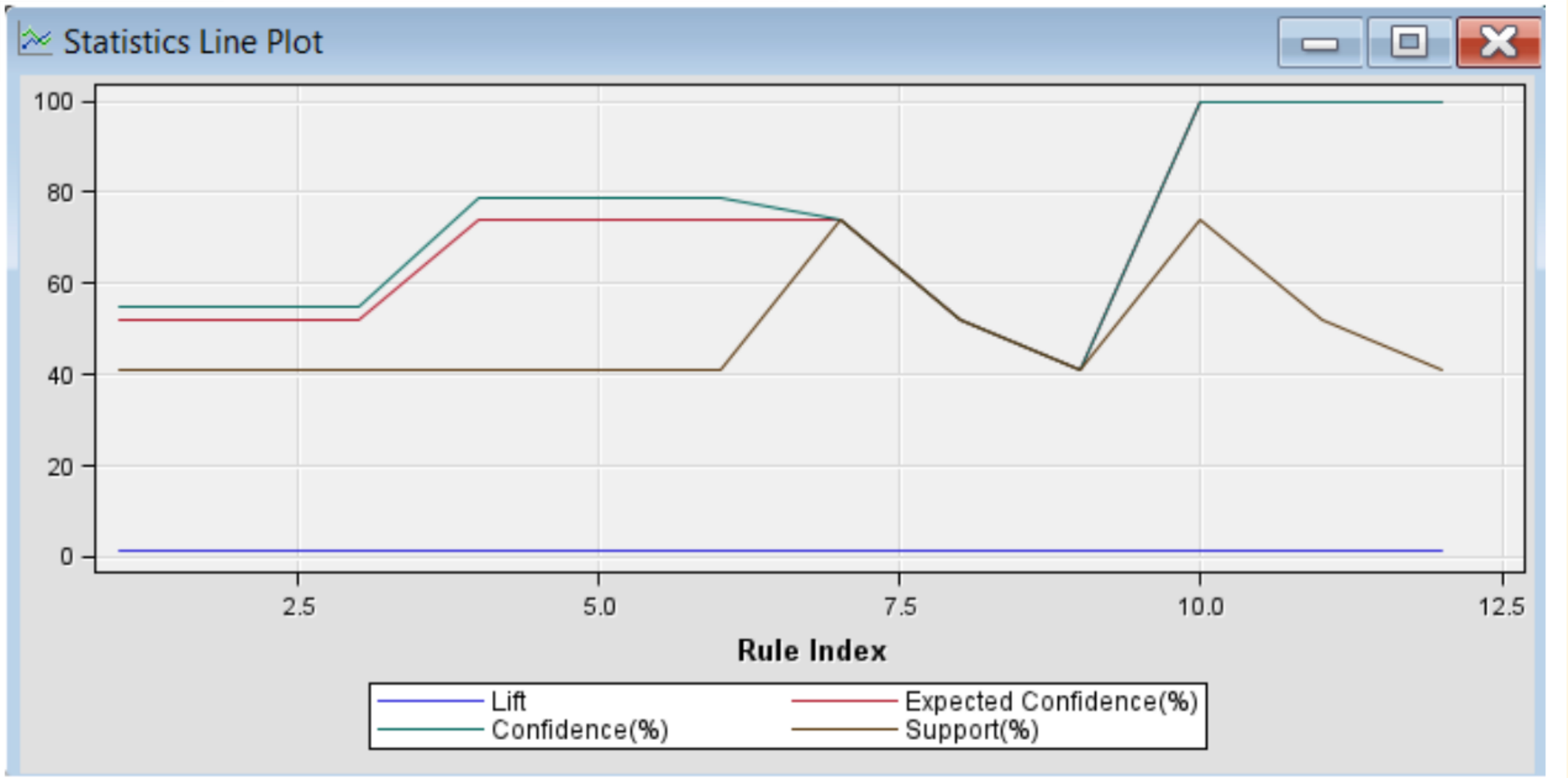


Diagram 9.1.5: Association Rule Line Plot

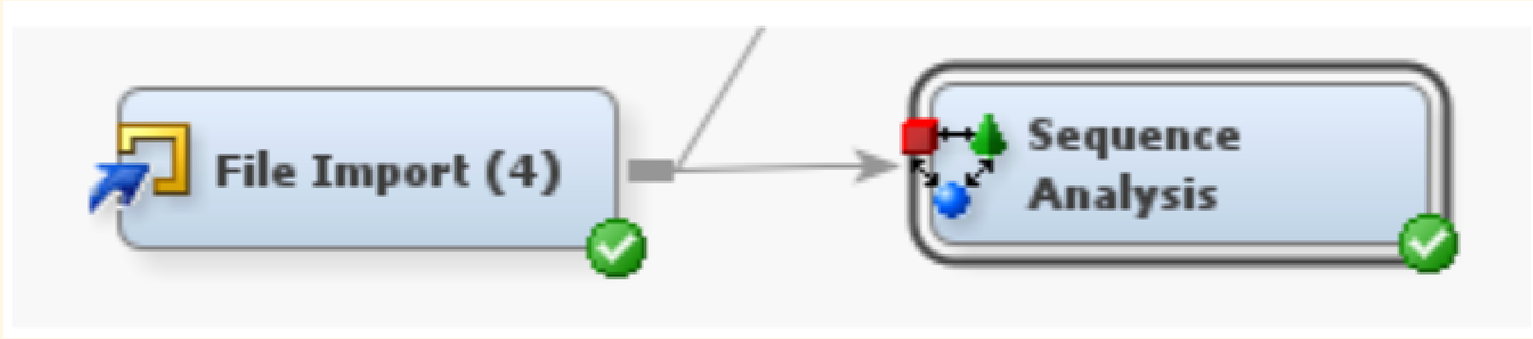
ASSOCIATION RULE

Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index	Transpose Rule
2	51.85	55.00	40.74	1.06	11.00	High ==>...	High	Low	High	===>	Low			1	1
3	51.85	55.00	40.74	1.06	11.00	Mid & Hig...	Mid & High	Low	Mid	High	===>	Low		2	1
3	51.85	55.00	40.74	1.06	11.00	High ==>...	High	Mid & Low	High	===>	Mid	Low		3	1
2	74.07	78.57	40.74	1.06	11.00	Low ==> ...	Low	High	Low	===>	High			4	1
3	74.07	78.57	40.74	1.06	11.00	Mid & Lo...	Mid & Low	High	Mid	Low	===>	High		5	1
3	74.07	78.57	40.74	1.06	11.00	Low ==> ...	Low	Mid & High	Low	===>	Mid	High		6	1
2	74.07	74.07	74.07	1.00	20.00	Mid ==> ...	Mid	High	Mid	===>	High			7	1
2	51.85	51.85	51.85	1.00	14.00	Mid ==> ...	Mid	Low	Mid	===>	Low			8	1
3	40.74	40.74	40.74	1.00	11.00	Mid ==> ...	Mid	Low & High	Mid	===>	Low	High		9	1
2	100.00	100.00	74.07	1.00	20.00	High ==>...	High	Mid	High	===>	Mid			10	1
2	100.00	100.00	51.85	1.00	14.00	Low ==> ...	Low	Mid	Low	===>	Mid			11	1
3	100.00	100.00	40.74	1.00	11.00	Low & Hi...	Low & High	Mid	Low	High	===>	Mid		12	1

Diagram 9.1.6: Association Rule Table

SEQUENCE ANALYSIS



Sequence Report

Chain Length	Transaction Count	Support (%)	Confidence (%)	Pseudo Lift	Rule	Chain Item 1	Chain Item 2	Chain Item 3	Rule Index	Left Hand of Rule	Right Hand of Rule
2	26	96.30	96.30	0.96	Mid ==> Mid	Mid	Mid		1	Mid	Mid
3	26	96.30	100.00	1.00	Mid ==> Mid ==> Mid	Mid	Mid	Mid	2	Mid ==> Mid	Mid
2	20	74.07	100.00	1.00	High ==> Mid	High	Mid		3	High	Mid
3	20	74.07	100.00	1.00	High ==> Mid ==> Mid	High	Mid	Mid	4	High ==> Mid	Mid
2	19	70.37	95.00	1.28	High ==> High	High	High		5	High	High
2	19	70.37	70.37	0.95	Mid ==> High	Mid	High		6	Mid	High
3	19	70.37	73.08	0.99	Mid ==> Mid ==> High	Mid	Mid	High	7	Mid ==> Mid	High
3	19	70.37	100.00	1.00	High ==> High ==> Mid	High	High	Mid	8	High ==> High	Mid
3	19	70.37	100.00	1.00	Mid ==> High ==> Mid	Mid	High	Mid	9	Mid ==> High	Mid
3	18	66.67	94.74	1.28	High ==> High ==> High	High	High	High	10	High ==> High	High
3	18	66.67	94.74	1.28	Mid ==> High ==> High	Mid	High	High	11	Mid ==> High	High
2	13	48.15	48.15	0.93	Mid ==> Low	Mid	Low		12	Mid	Low
2	13	48.15	92.86	0.93	Low ==> Mid	Low	Mid		13	Low	Mid
3	13	48.15	50.00	0.96	Mid ==> Mid ==> Low	Mid	Mid	Low	14	Mid ==> Mid	Low
3	13	48.15	100.00	1.00	Low ==> Mid ==> Mid	Low	Mid	Mid	15	Low ==> Mid	Mid
2	12	44.44	85.71	1.65	Low ==> Low	Low	Low		16	Low	Low
2	12	44.44	100.00	1.00	Mid & High ==> Mid	Mid & High	Mid		17	Mid & High	Mid
2	12	44.44	44.44	1.00	Mid ==> Mid & High	Mid	Mid & High		18	Mid	Mid & High
3	12	44.44	92.31	0.92	Mid ==> Low ==> Mid	Mid	Low	Mid	19	Mid ==> Low	Mid
3	12	44.44	100.00	1.00	Mid & High ==> Mid ==> Mid	Mid & High	Mid	Mid	20	Mid & High ==> Mid	Mid
3	12	44.44	100.00	1.00	Mid ==> Mid & High ==> Mid	Mid	Mid & High	Mid	21	Mid ==> Mid & High	Mid
3	12	44.44	46.15	1.04	Mid ==> Mid ==> Mid & High	Mid	Mid	Mid & High	22	Mid ==> Mid	Mid & High
2	11	40.74	78.57	1.06	Low ==> High	Low	High		23	Low	High
2	11	40.74	91.67	1.24	Mid & High ==> High	Mid & High	High		24	Mid & High	High
3	11	40.74	100.00	1.35	Mid & High ==> High ==> High	Mid & High	High	High	25	Mid & High ==> High	High

Diagram 9.2.2:Seaquence Analysis Report

- For conducting sequence analysis on our dataset, the initial step involves redefining the dataset role to "Transaction" and designating "Day" as the ID, "Category" as the Target, and "Hour" as the Sequence.
- Subsequently, an Association node is introduced to the diagram workspace, linked to the dataset node.

SEQUENCE ANALYSIS

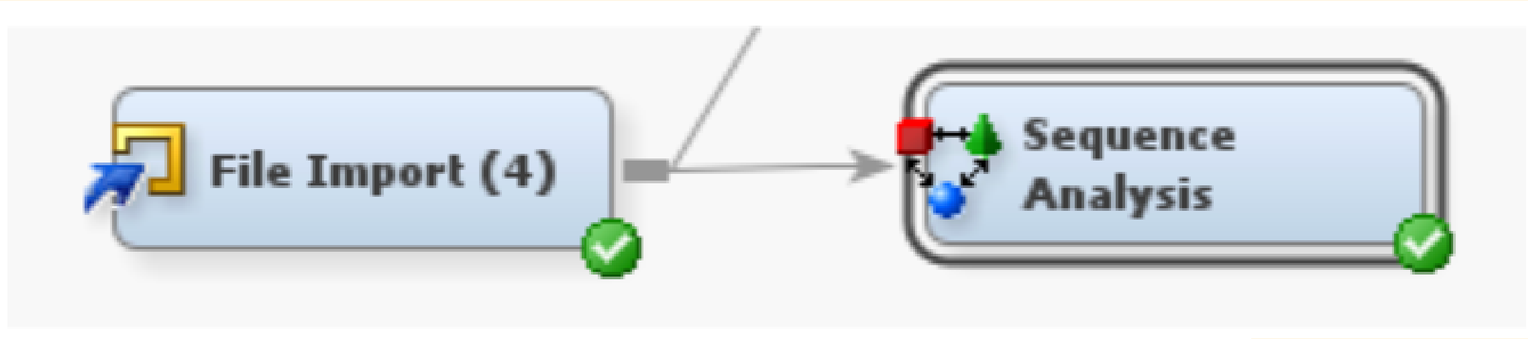


Diagram 9.2.3:Rule Matrix

SEQUENCE ANALYSIS

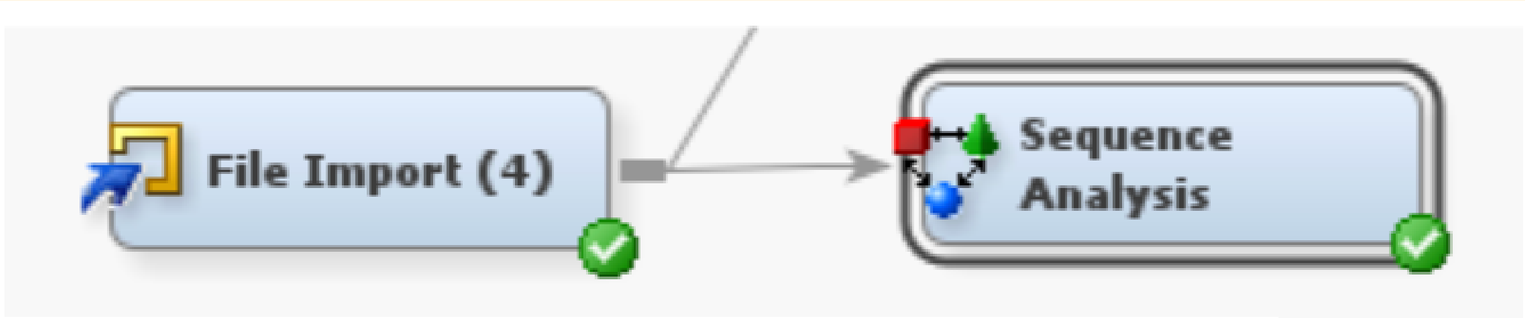
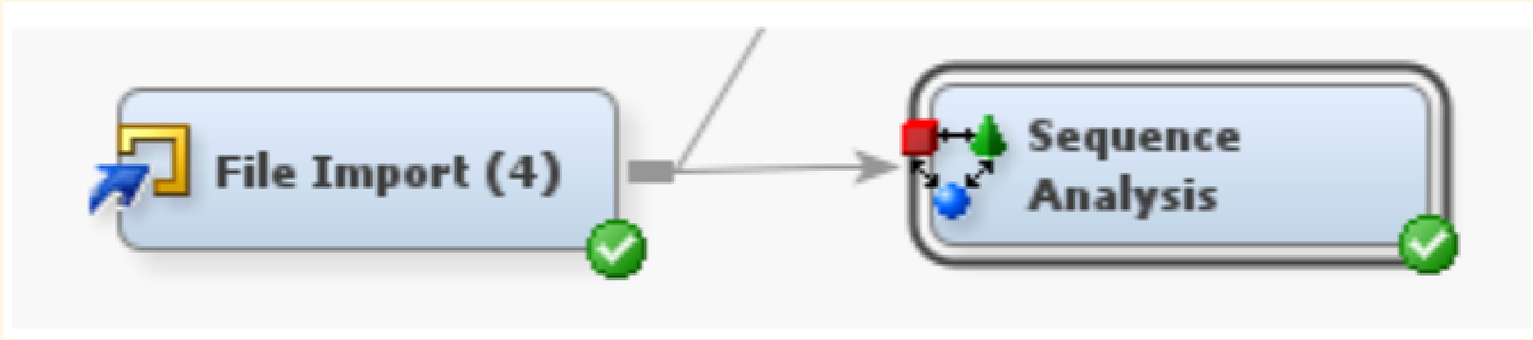


Diagram 9.2.4: Statistics Line Plot

SEQUENCE ANALYSIS



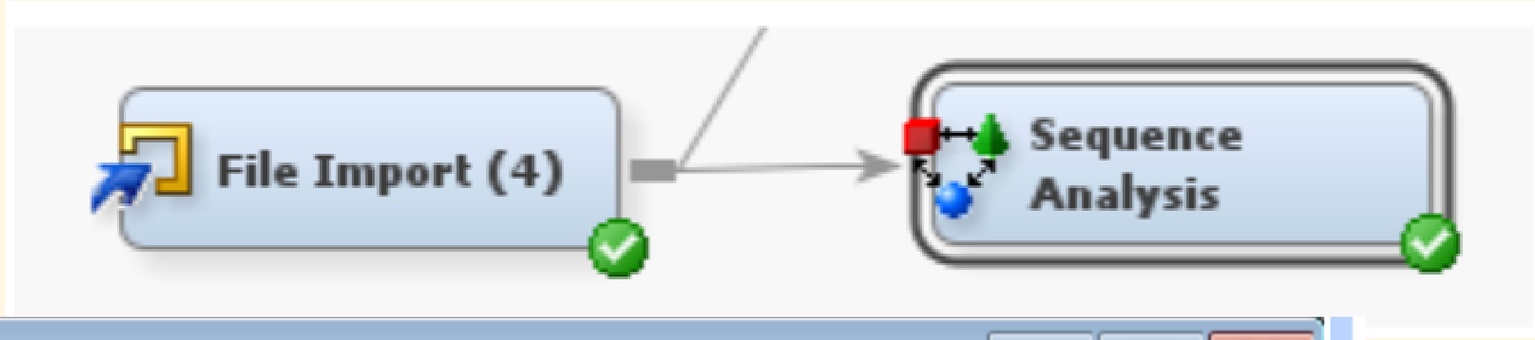

```
Rule Statistics

The MEANS Procedure

Variable    Label                          Minimum          Maximum             Mean
-------------------------------------------------------------------------------------
NITEMS      Chain Length                 2.0000000        3.0000000        2.8200000
COUNT       Transaction Count            1.0000000       26.0000000        6.1150000
SUPPORT     Support(%)                   3.7037037       96.2962963       22.6481481
CONF        Confidence(%)                5.0000000      100.0000000       63.7409009
-------------------------------------------------------------------------------------
```

Diagram 9.2.5: Rule Statistics

SEQUENCE ANALYSIS



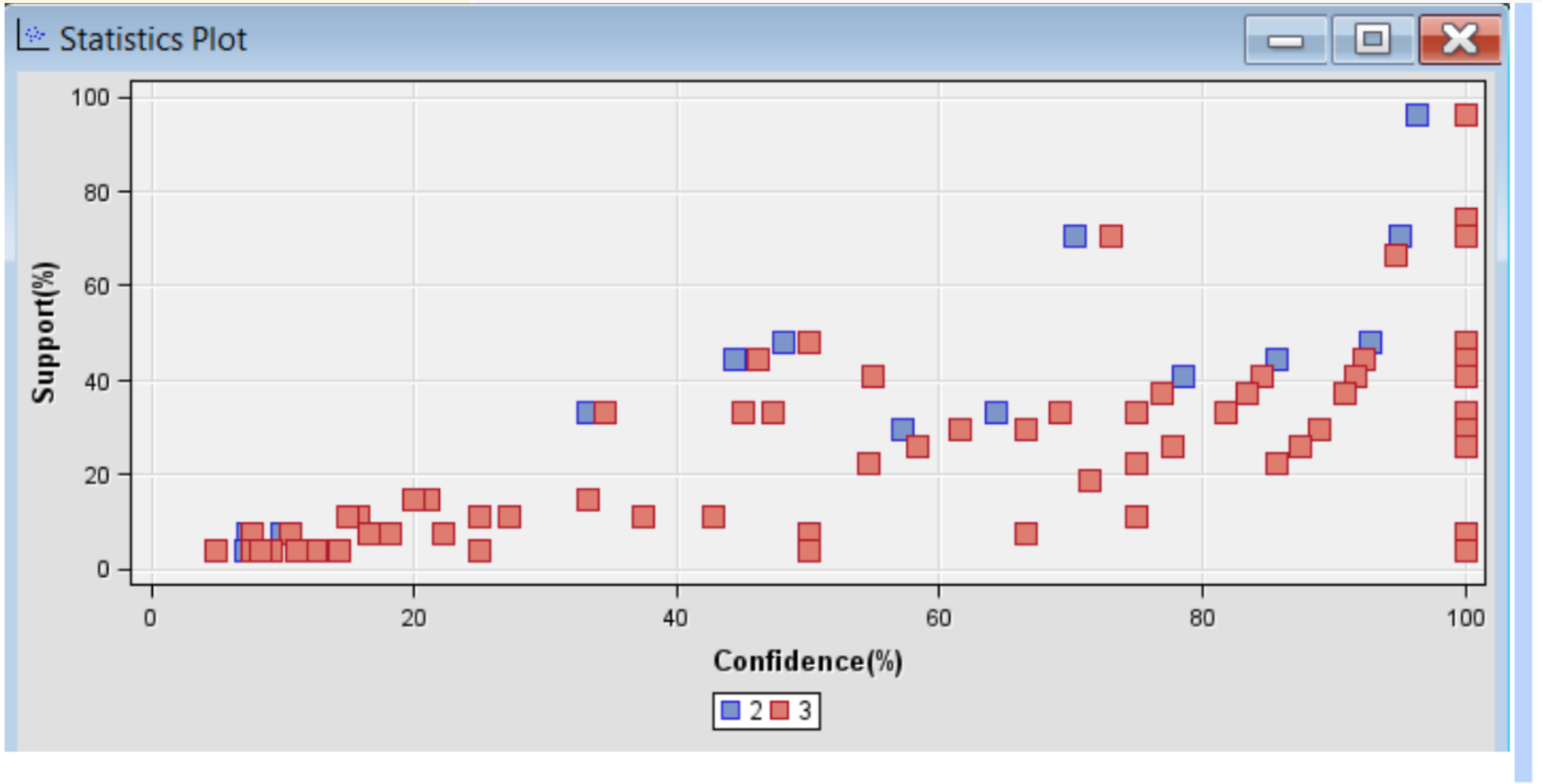
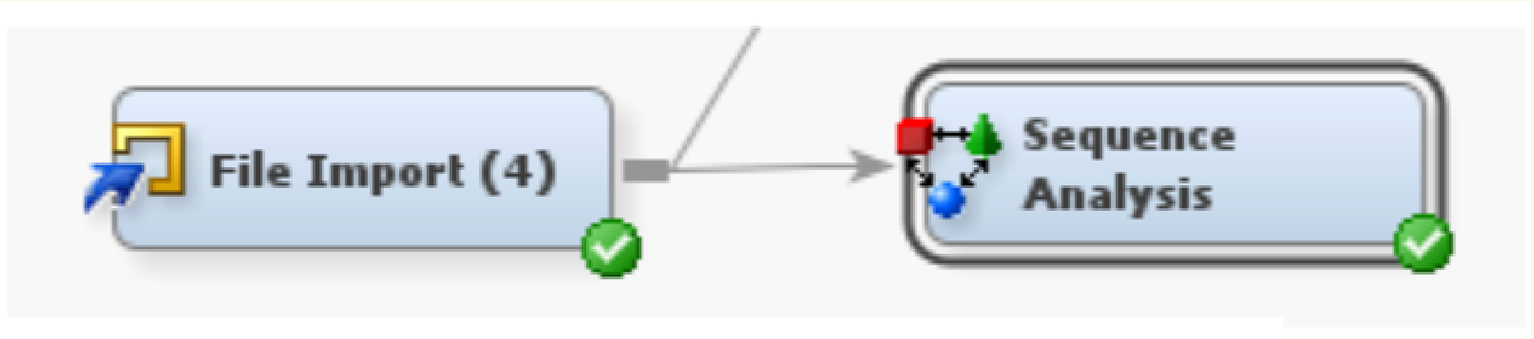


Diagram 9.2.6: Statistics Plot

SEQUENCE ANALYSIS



Sequence Report

The FREQ Procedure

Chain Length

NITEMS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	36	18.00	36	18.00
3	164	82.00	200	100.00

Diagram 9.2.7: Sequence Report

TIME SERIES CLUSTERING

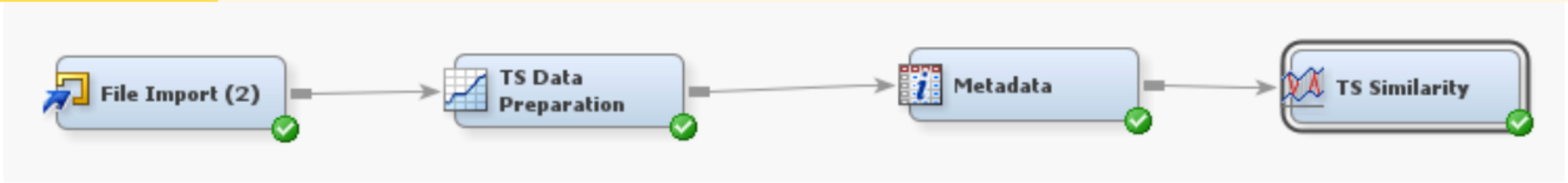


Diagram 9.3.1: Time Clustering Diagram

Report	
Similarity Plot Maximum	5
Preference of Similarity Plot	Most Similar
Output Data Set	Clustering Segment

Diagram 9.3.2: TS Similarity Node Properties

TIME SERIES CLUSTERING

Exported Attributes for TRAIN Port
(maximum 500 observations printed)

Variable Name	Role	Measurement Level	Creator	Label
Day	TIMEID	INTERVAL		
_TS_01	TARGET	INTERVAL	TSDP	Temperature 1
_TS_02	INPUT	INTERVAL	TSDP	Temperature 2
_TS_03	INPUT	INTERVAL	TSDP	Temperature 3
_TS_04	INPUT	INTERVAL	TSDP	Temperature 4
_TS_05	INPUT	INTERVAL	TSDP	Temperature 5
_TS_06	INPUT	INTERVAL	TSDP	Temperature 6
_TS_07	INPUT	INTERVAL	TSDP	Temperature 7
_TS_08	INPUT	INTERVAL	TSDP	Temperature 8
_TS_09	INPUT	INTERVAL	TSDP	Temperature 9
_TS_10	INPUT	INTERVAL	TSDP	Temperature 10
_TS_11	INPUT	INTERVAL	TSDP	Temperature 11
_TS_12	INPUT	INTERVAL	TSDP	Temperature 12
_TS_13	INPUT	INTERVAL	TSDP	Temperature 13
_TS_14	INPUT	INTERVAL	TSDP	Temperature 14
_TS_15	INPUT	INTERVAL	TSDP	Temperature 15
_TS_16	INPUT	INTERVAL	TSDP	Temperature 16
_TS_17	INPUT	INTERVAL	TSDP	Temperature 17
_TS_18	INPUT	INTERVAL	TSDP	Temperature 18
_TS_19	INPUT	INTERVAL	TSDP	Temperature 19
_TS_20	INPUT	INTERVAL	TSDP	Temperature 20
_TS_21	INPUT	INTERVAL	TSDP	Temperature 21
_TS_22	INPUT	INTERVAL	TSDP	Temperature 22
_TS_23	INPUT	INTERVAL	TSDP	Temperature 23

Diagram 9.3.3: Time Series Data Preparation

TIME SERIES CLUSTERING

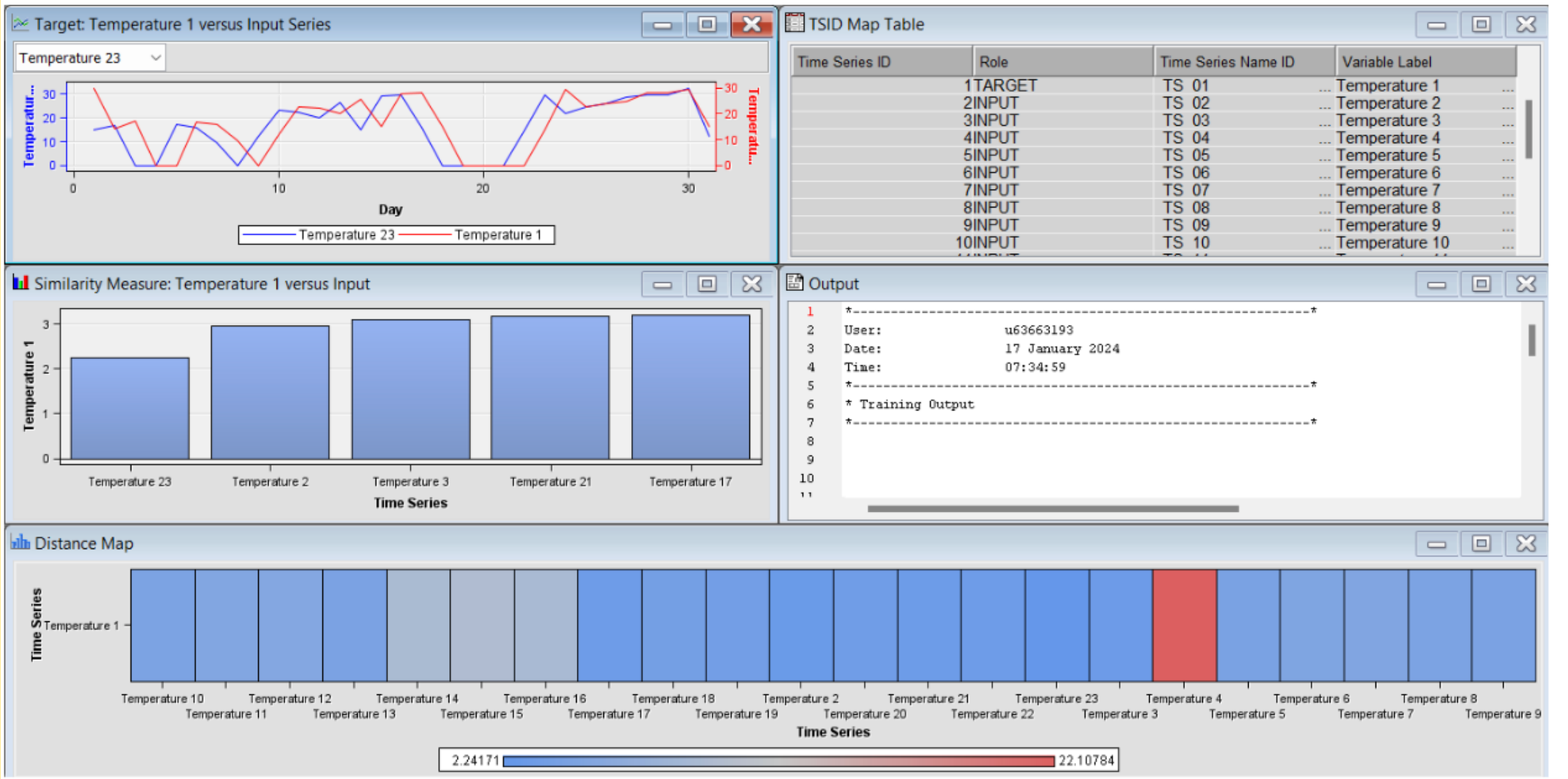


Diagram 9.3.4: Time Series Similarity Node

TIME SERIES CLUSTERING

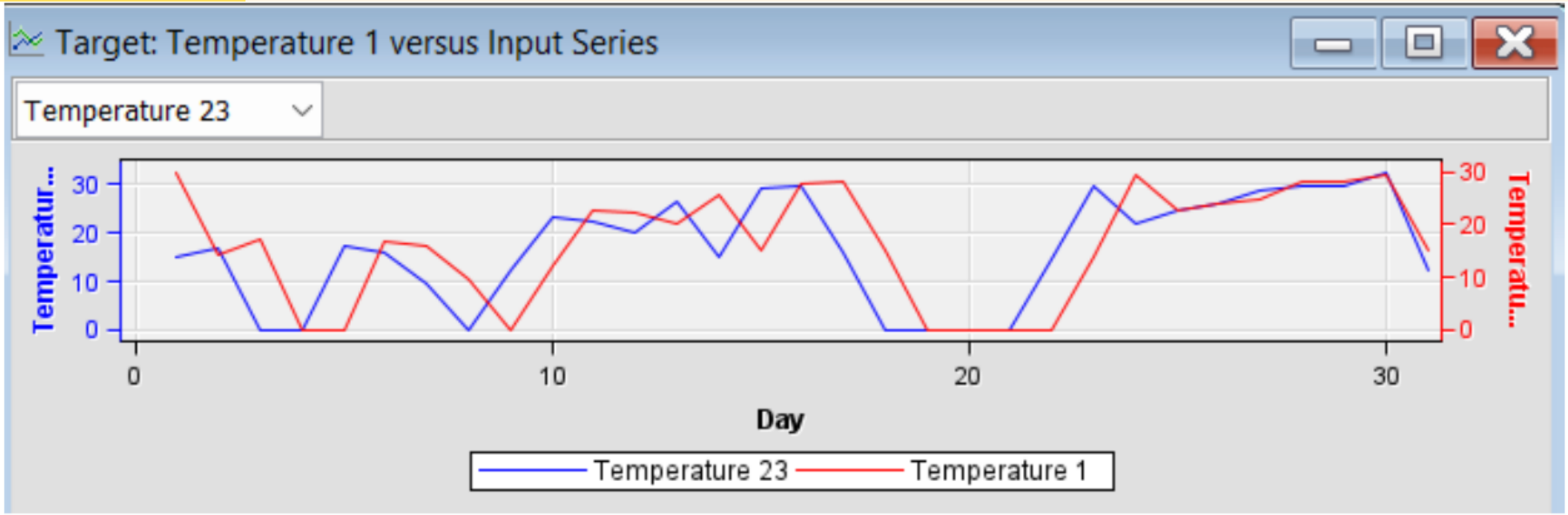


Diagram 9.3.5: Most Similar Time

MODEL (CLASSIFICATION)

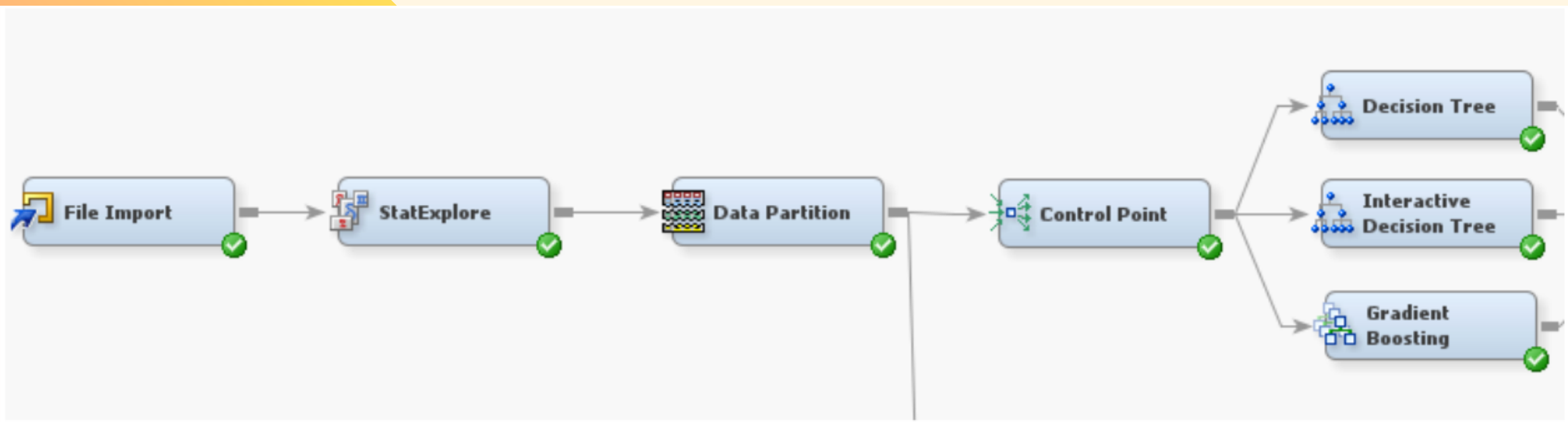


Diagram 10.1: Model Diagram

- The models are developed by analyzing the training data.
- The model is then used to predict the label or class of unlabeled objects.

MODEL (DECISION TREE)

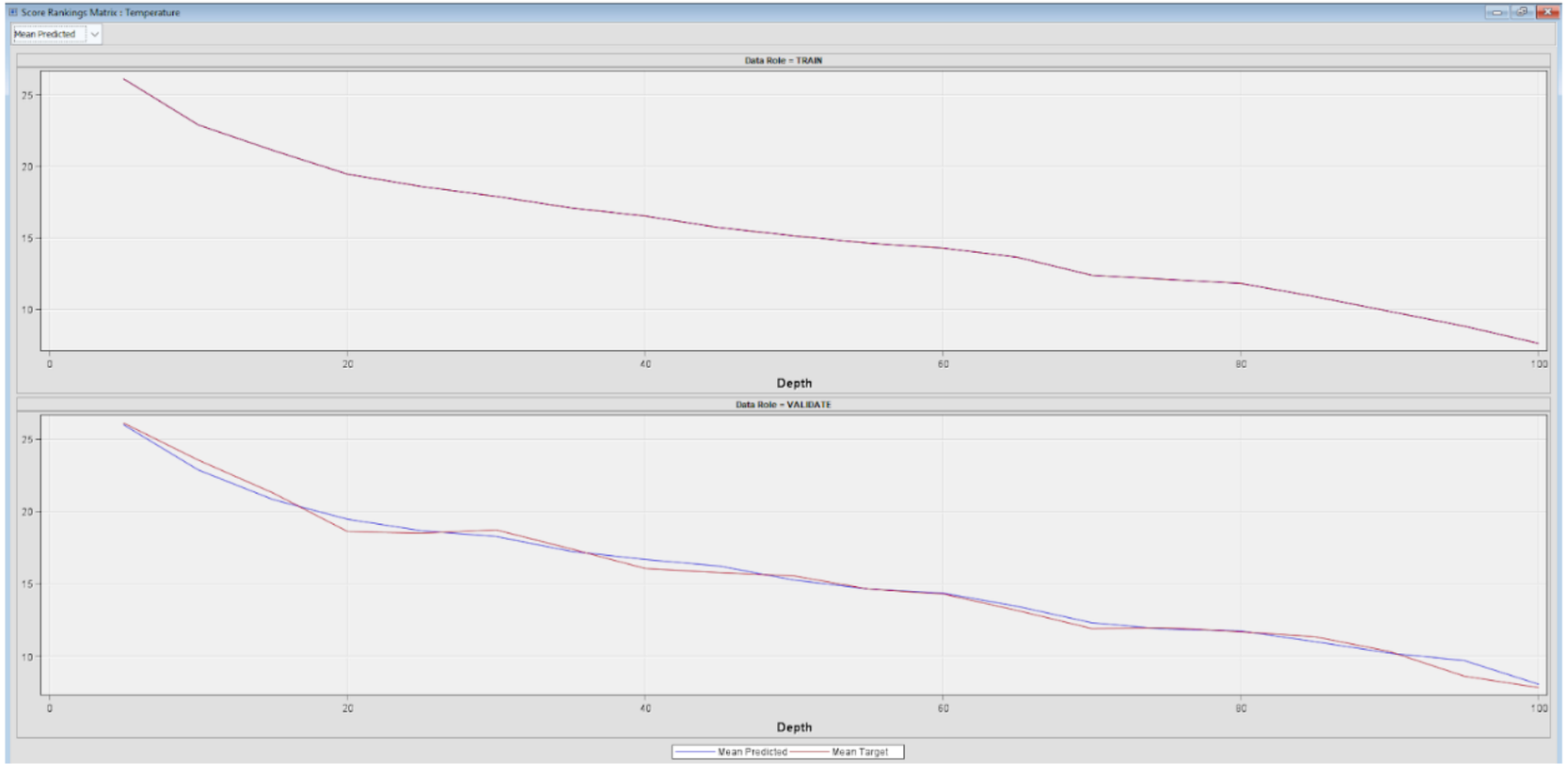


Diagram 10.2: Score Ranking Matrix (Decision Tree)

- In the TRAIN data model, the trends of the Score Ranking Matrix for Mean Predicted and Mean Target are overlapping
- In the VALIDATE train data model, there is a slight disparity between Mean Predicted and Mean Target, suggesting that the model may not be as accurate.

MODEL (DECISION TREE)

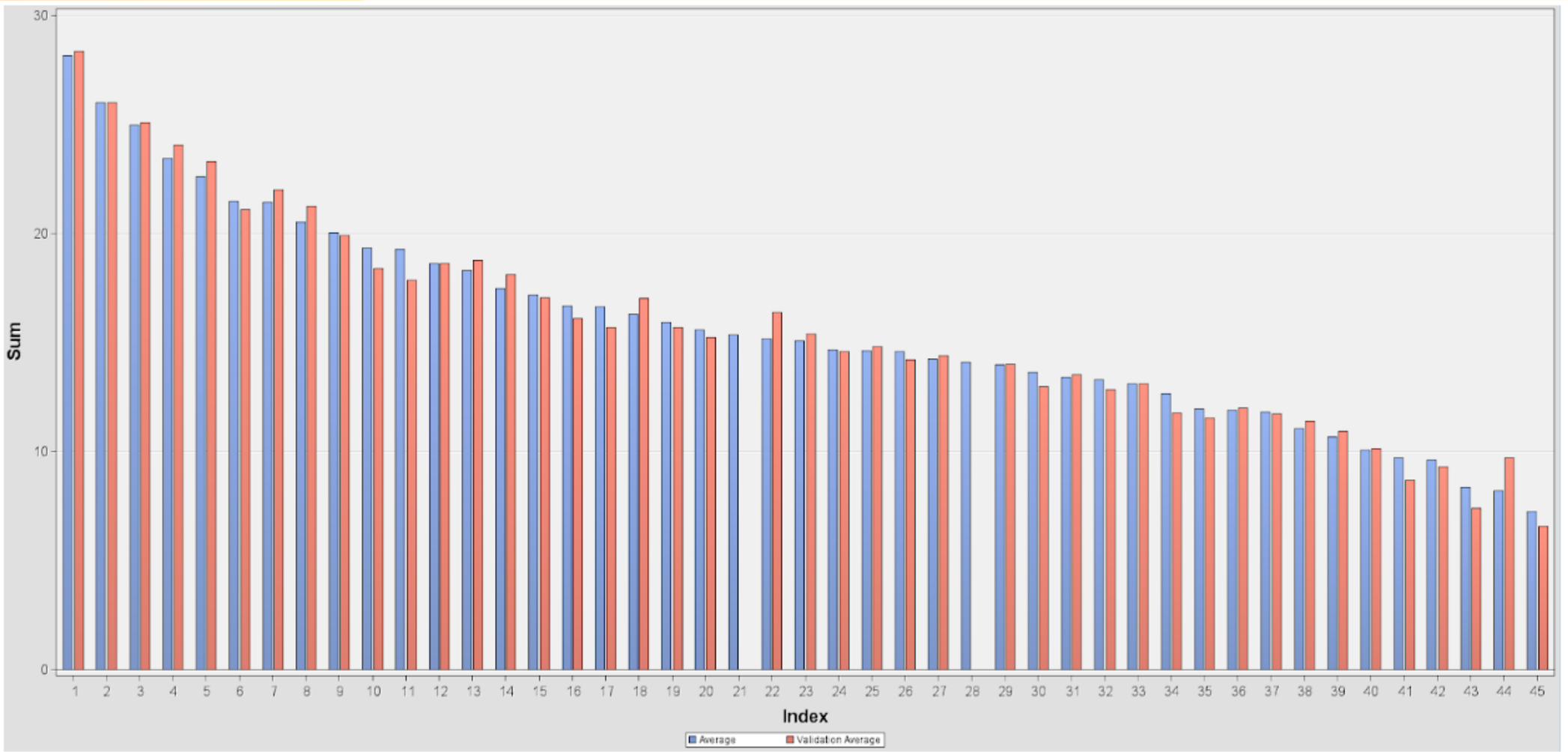


Diagram 10.3: Leaf Statistic (Decision Tree)

MODEL (DECISION TREE)

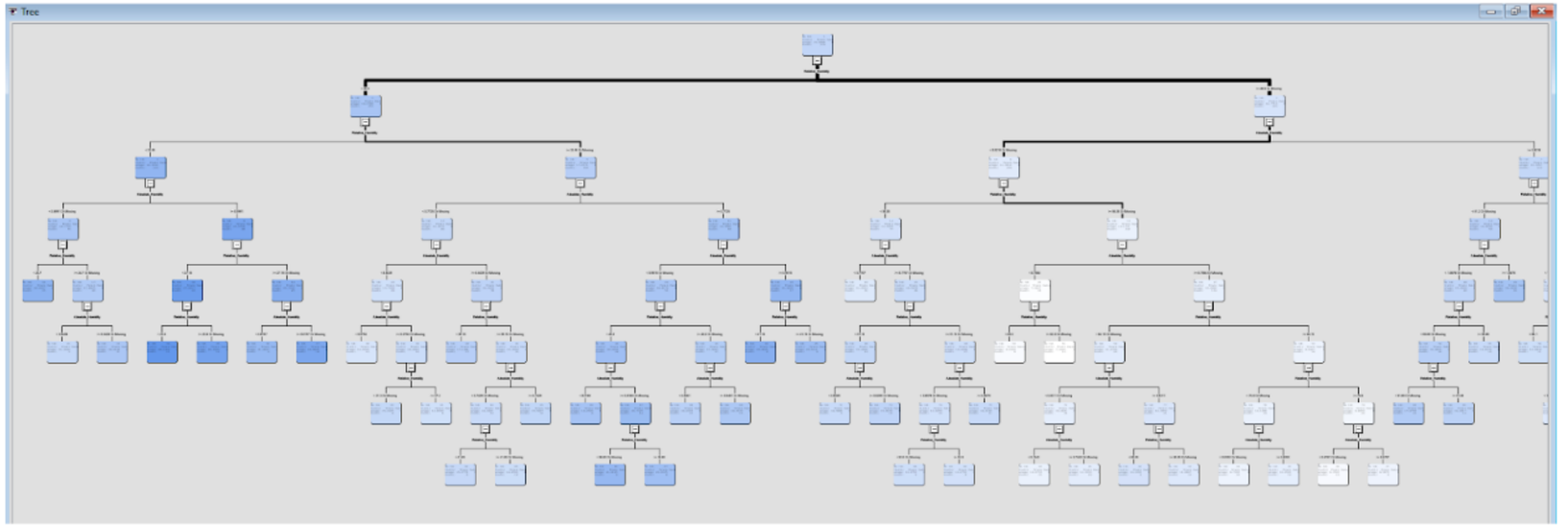


Diagram 10.4: Tree (Decision Tree)

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
Temperature	NOBS	Sum of Frequencies	579	165	83
Temperature	MAX	Maximum Absolute Error	5.893333	6.693333	2.4375
Temperature	SSE	Sum of Squared Errors	535.577	245.6968	71.99258
Temperature	ASE	Average Squared Error	0.925003	1.489071	0.86738
Temperature	RASE	Root Average Squared Error	0.961771	1.220275	0.931333
Temperature	DIV	Divisor for ASE	579	165	83
Temperature	DFT	Total Degrees of Freedom	579	.	.

Diagram 10.5: Fit Statistic (Decision Tree)

MODEL (DECISION TREE)

Data Role=TRAIN Target Variable=Temperature Target Label=' '

Depth	Number of Observations	Mean Target	Mean Predicted
5	39	26.1769	26.1769
10	23	22.9043	22.9043
15	26	21.1192	21.1192
20	32	19.4875	19.4875
25	31	18.6194	18.6194
30	27	17.9407	17.9407
35	27	17.1481	17.1481
40	38	16.5421	16.5421
45	22	15.7227	15.7227
50	26	15.1962	15.1962
55	30	14.6433	14.6433
60	30	14.2900	14.2900
65	26	13.6577	13.6577
70	58	12.4328	12.4328
80	33	11.8273	11.8273
85	38	10.8921	10.8921
90	35	9.8771	9.8771
95	20	8.8450	8.8450
100	18	7.6667	7.6667

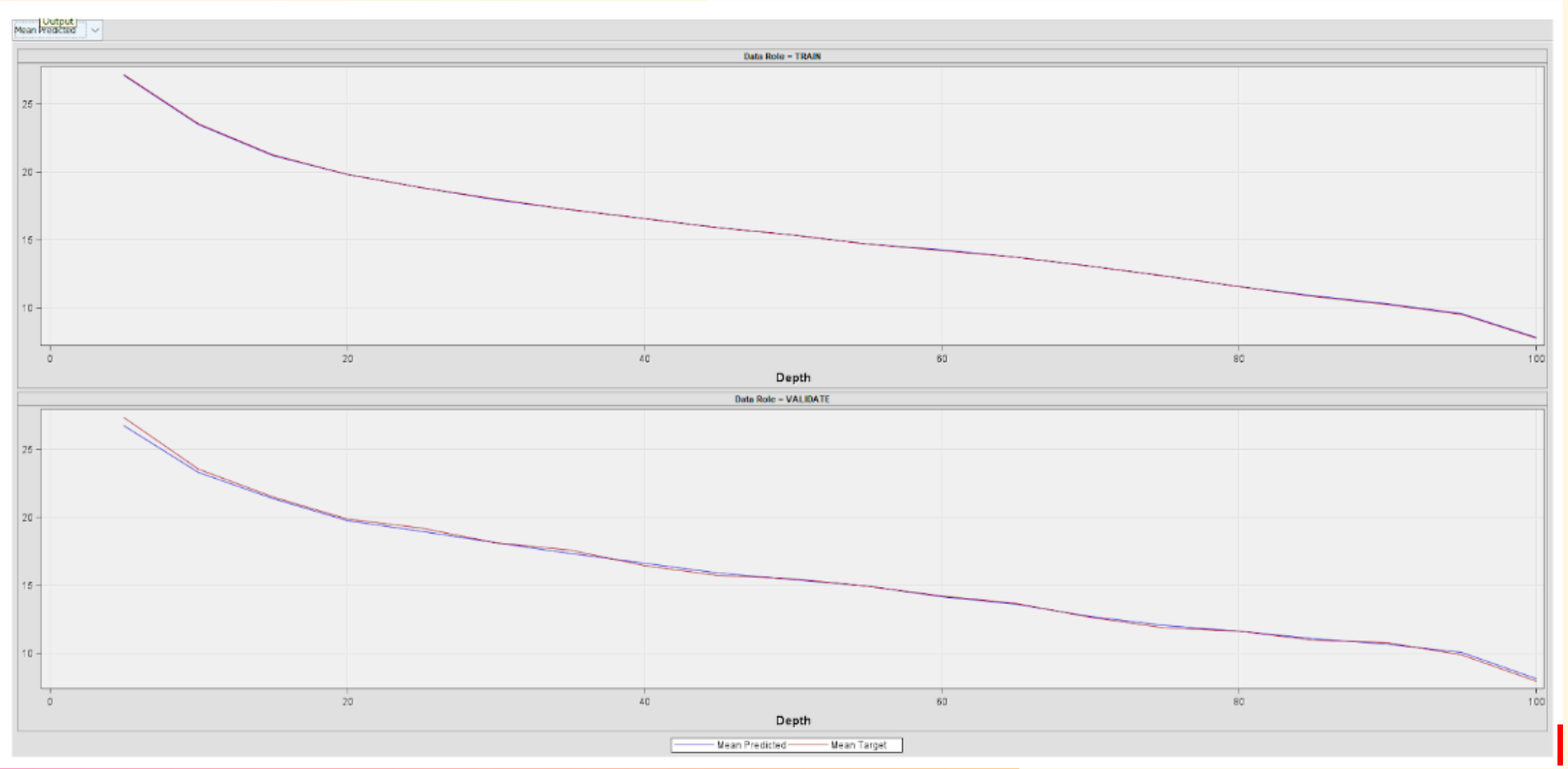
Diagram 10.6: Assessment Score Ranking - Train (Decision Tree)

Data Role=VALIDATE Target Variable=Temperature Target Label=' '

Depth	Number of Observations	Mean Target	Mean Predicted
5	11	26.1364	26.0167
10	6	23.5500	22.8919
15	12	21.2750	20.8251
20	6	18.6500	19.4552
25	14	18.4929	18.7094
30	2	18.7500	18.3133
35	13	17.4538	17.2739
40	3	16.1000	16.6786
45	9	15.8222	16.2739
50	8	15.5625	15.2823
55	9	14.6778	14.6402
60	6	14.3333	14.3630
65	10	13.2000	13.4980
70	12	11.9583	12.3551
75	4	12.0000	11.8667
80	7	11.7286	11.8048
85	13	11.3692	11.0333
90	11	10.3545	10.2306
95	4	8.6750	9.7300
100	5	7.9000	8.1235

Diagram 10.6: Assessment Score Ranking - Validate (Decision Tree)

MODEL (INTERACTIVE DECISION TREE)



- Both the TRAIN data model and the VALIDATE train data model, the trends of the Score Ranking Matrix for Mean Predicted and Mean Target are overlapping.
- This alignment suggests that the model is accurate and reliable.

MODEL (DECISION TREE)

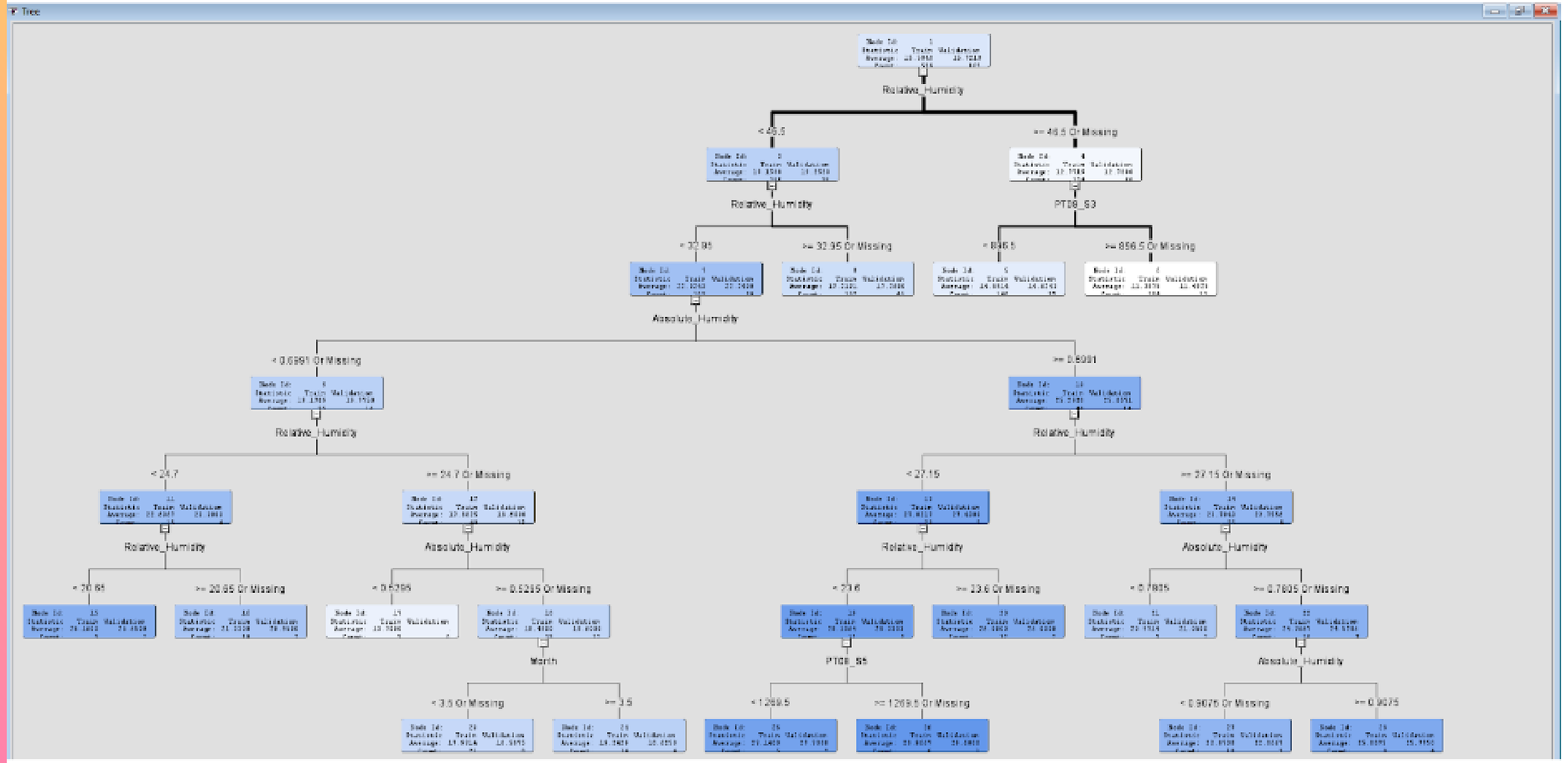
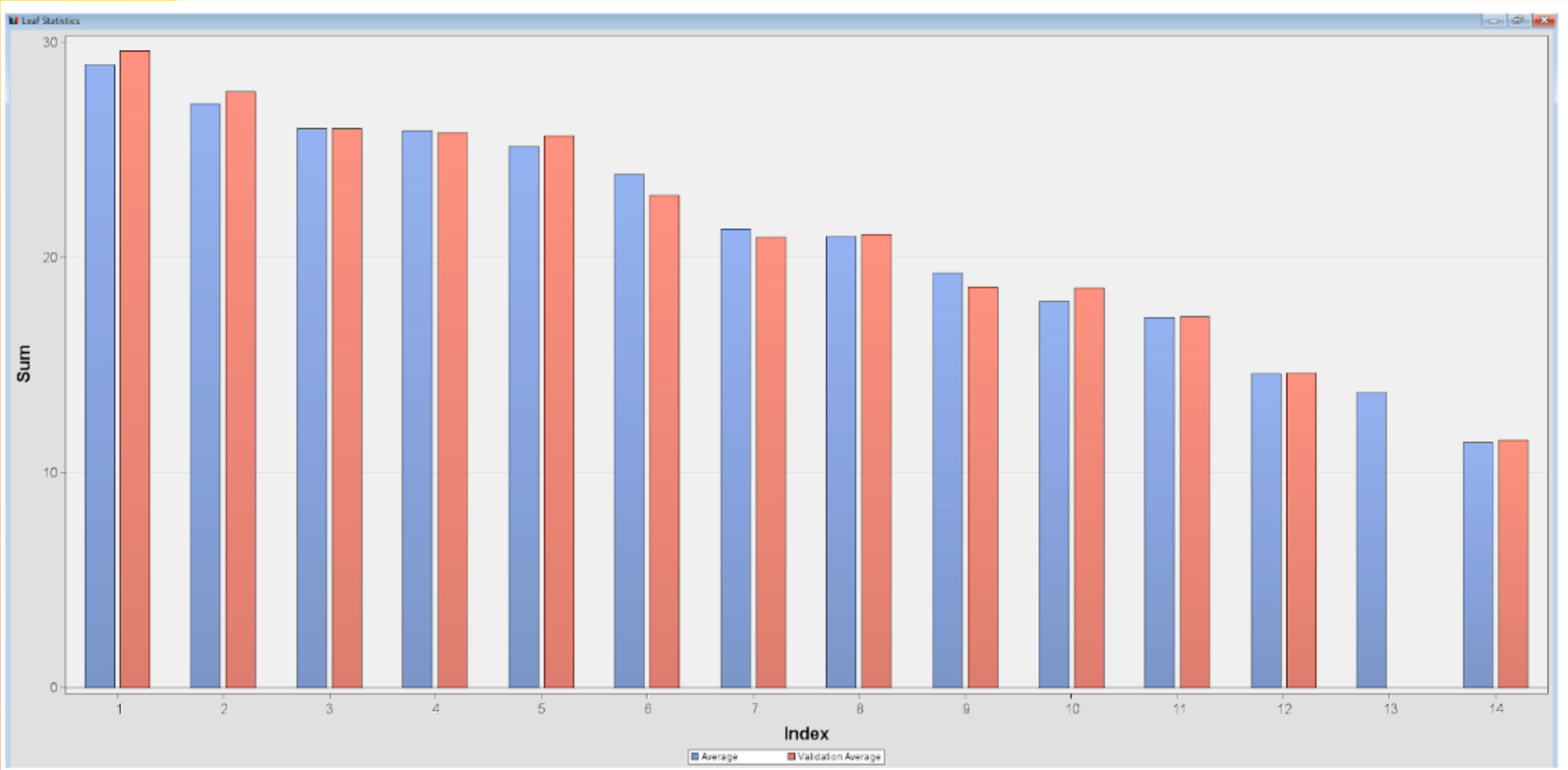


Diagram 10.9 Tree (Interactive Decision Tree)

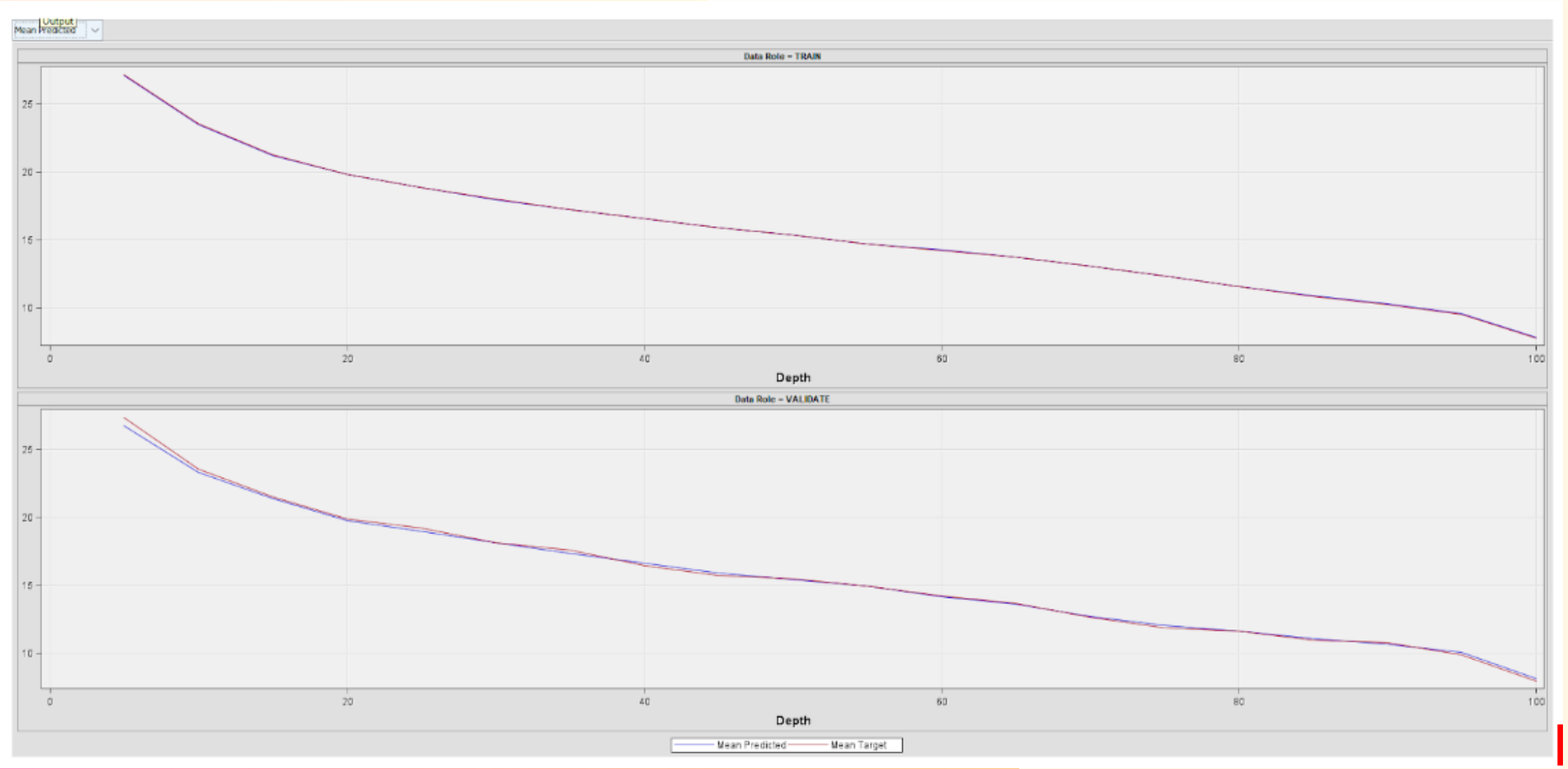
MODEL (DECISION TREE)



Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
Temperature	NOBS	Sum of Frequencies	579	165	83
Temperature	MAX	Maximum Absolute Error	8.382895	7.082895	6.491429
Temperature	SSE	Sum of Squared Errors	3635.586	1144.089	483.8932
Temperature	ASE	Average Squared Error	6.279078	6.933872	5.830039
Temperature	RASE	Root Average Squared Error	2.505809	2.633225	2.414547
Temperature	DIV	Divisor for ASE	579	165	83
Temperature	DFT	Total Degrees of Freedom	579	.	.

MODEL (GRADIENT BOOSTING)



- Trends of the Score Ranking Matrix for both the TRAIN data model (Mean Predicted and Mean Target) and the VALIDATE train data model (Mean Predicted and Mean Target) exhibit overlapping patterns.

MODEL (REGRESSION)

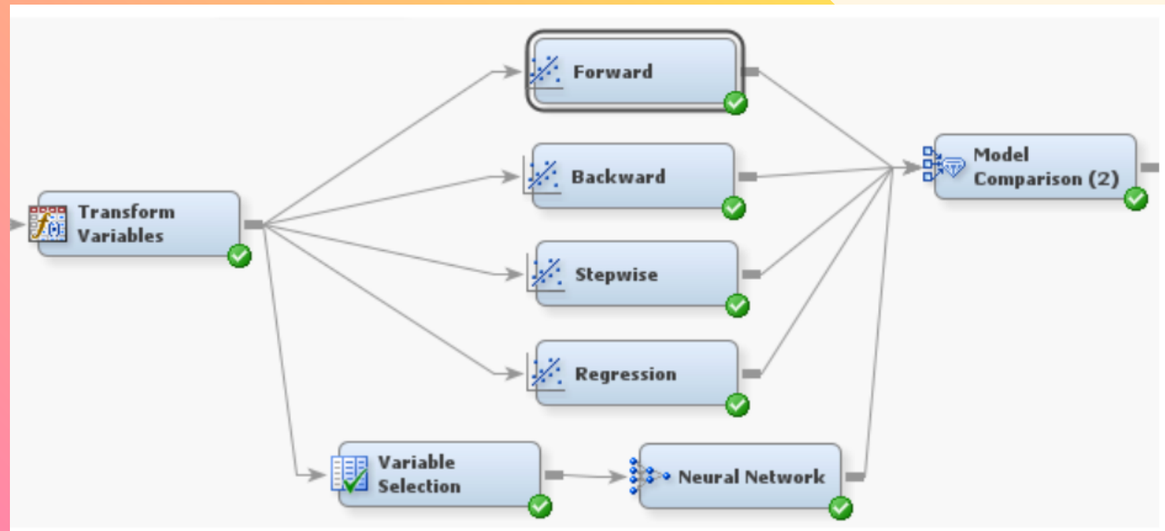


Diagram 10.5.1: Regression Node

- First, we will execute the Regression node with its default settings.
- SAS Enterprise Miner employs the Logistic Regression with a Logit link function as the default regression type. Logistic regression, often associated with the sigmoid function, is utilized for binary classification tasks, where the outcome variable has only two possible results (typically coded as 0 and 1).

- Additionally, we will employ a sequential selection approach within the Regression node by configuring the Selection Model to Forward, Backward, or Stepwise.
- This sequential selection method enhances the model's performance and identifies a subset of variables that offer the most effective explanation for the variations in the target variable.

MODEL (REGRESSION)

Regression

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	11403	518.318166	150.64	<.0001
Error	556	1913.026593	3.440695		
Corrected Total	578	13316			

Model Fit Statistics

R-Square	0.8563	Adj R-Sq	0.8507
AIC	737.9857	BIC	741.8852
SBC	838.2957	C(p)	23.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
DateID	1	324.2996	94.25	<.0001
Day	1	6.3414	1.84	0.1751
Hour	1	31.3069	9.10	0.0027
LG10_PT08_S1	1	52.5490	15.27	0.0001
LG10_PT08_S2	1	25.5728	7.43	0.0066
LG10_PT08_S3	1	12.1799	3.54	0.0604
LG10_PT08_S4	1	17.0758	4.96	0.0263
LG10_PT08_S5	1	48.8527	14.20	0.0002
MeasurementID	0	.	.	.
Month	0	.	.	.
OPT_Absolute_Humidity	2	302.9028	44.02	<.0001
OPT_Benzene_Concentration	2	27.0032	3.92	0.0203
OPT_CO_Concentration	2	0.7071	0.10	0.9024
OPT_NMHC_Concentration	2	11.4225	1.66	0.1911
OPT_NO2_Concentration	2	7.3641	1.07	0.3437
OPT_NOx_Concentration	1	2.1225	0.62	0.4325
OPT_Relative_Humidity	3	4364.9374	422.87	<.0001
ReadingID	0	.	.	.

Diagram 10.14: Output Result (Default Regression)

- In our regression model, we've obtained a Mean Square Error (MSE) value of 3.440695. During our Type 3 Analysis of Effects, we identified the input variables with the highest F values, revealing that Relative Humidity and Absolute Humidity exhibit the strongest correlations with temperature.
- These insights provide valuable information about the influential factors impacting our temperature predictions in the model.

MODEL (REGRESSION)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	11325	1132.454237	322.99	<.0001
Error	568	1991.483883	3.506134		
Corrected Total	578	13316			

Model Fit Statistics

R-Square	0.8504	Adj R-Sq	0.8478
AIC	737.2577	BIC	739.2652
SBC	785.2320	C(p)	21.8027

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
DateID	1	427.3215	121.88	<.0001
Hour	1	21.1499	6.03	0.0143
LG10_PT08_S1	1	112.0258	31.95	<.0001
LG10_PT08_S3	1	69.1998	19.74	<.0001
LG10_PT08_S5	1	44.2073	12.61	0.0004
OPT_Absolute_Humidity	2	621.4317	88.62	<.0001
OPT_Relative_Humidity	3	6893.3231	655.36	<.0001

Diagram 10.15: Output Result (Forward Regression)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	11379	812.789992	236.67	<.0001
Error	564	1936.966366	3.434338		
Corrected Total	578	13316			

Model Fit Statistics

R-Square	0.8545	Adj R-Sq	0.8509
AIC	729.1864	BIC	732.0397
SBC	794.6059	C(p)	13.9578

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
DateID	1	328.0336	95.52	<.0001
Hour	1	34.9978	10.19	0.0015
LG10_PT08_S1	1	56.6001	16.48	<.0001
LG10_PT08_S2	1	31.1009	9.06	0.0027
LG10_PT08_S3	1	13.5299	3.94	0.0476
LG10_PT08_S4	1	17.3899	5.06	0.0248
LG10_PT08_S5	1	52.5520	15.30	0.0001
OPT_Absolute_Humidity	2	361.7818	52.67	<.0001
OPT_Benzene_Concentration	2	31.1051	4.53	0.0112
OPT_Relative_Humidity	3	4835.8558	469.36	<.0001

Diagram 10.16: Output Result (Backward Regression)

MODEL (REGRESSION)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	11325	1132.454237	322.99	<.0001
Error	568	1991.483883	3.506134		
Corrected Total	578	13316			

Model Fit Statistics

R-Square	0.8504	Adj R-Sq	0.8478
AIC	737.2577	BIC	739.2652
SBC	785.2320	C(p)	21.8027

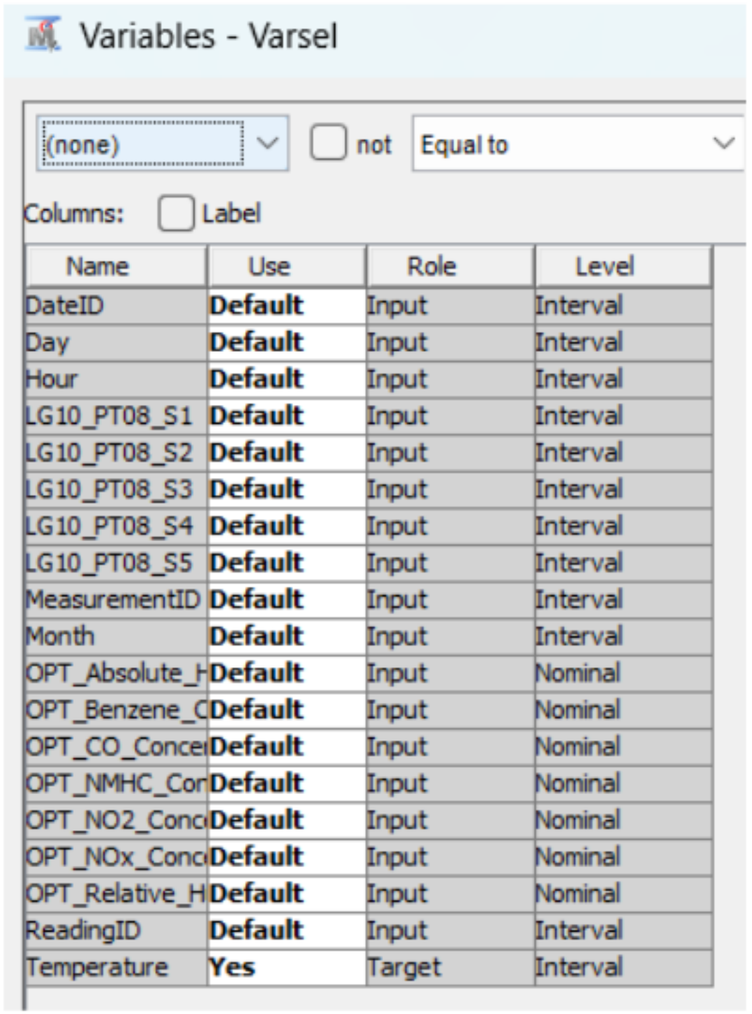
Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
DateID	1	427.3215	121.88	<.0001
Hour	1	21.1499	6.03	0.0143
LG10_PT08_S1	1	112.0258	31.95	<.0001
LG10_PT08_S3	1	69.1998	19.74	<.0001
LG10_PT08_S5	1	44.2073	12.61	0.0004
OPT_Absolute_Humidity	2	621.4317	88.62	<.0001
OPT_Relative_Humidity	3	6893.3231	655.36	<.0001

Diagram 10.17: Output Result (Stepwise Regression)

- In Stepwise Regression, we utilize the Regression node with a model selection set to Stepwise. Although the outcomes closely resemble those of other regression models,
- The relative slight variations occur due to distinct techniques in model training. The mean square error of the Backward Regression model is 3.506134.

MODEL (NEURAL NETWORK)



Name	Use	Role	Level
DateID	**Default**	Input	Interval
Day	**Default**	Input	Interval
Hour	**Default**	Input	Interval
LG10_PT08_S1	**Default**	Input	Interval
LG10_PT08_S2	**Default**	Input	Interval
LG10_PT08_S3	**Default**	Input	Interval
LG10_PT08_S4	**Default**	Input	Interval
LG10_PT08_S5	**Default**	Input	Interval
MeasurementID	**Default**	Input	Interval
Month	**Default**	Input	Interval
OPT_Absolute_H	**Default**	Input	Nominal
OPT_Benzene_C	**Default**	Input	Nominal
OPT_CO_Concei	**Default**	Input	Nominal
OPT_NMHC_Cor	**Default**	Input	Nominal
OPT_NO2_Conc	**Default**	Input	Nominal
OPT_NOx_Conc	**Default**	Input	Nominal
OPT_Relative_H	**Default**	Input	Nominal
ReadingID	**Default**	Input	Interval
Temperature	**Yes**	Target	Interval

Diagram 10.18: Variable Selection (Neutral Network)

Iter	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Step Size	Slope of Search Direction
1	0	6	0	3.34922	0.0309	0.2709	0.00543	-11.528
2	0	10	0	3.26807	0.0812	0.2528	1.528	-0.106
3	0	12	0	3.16432	0.1037	0.1679	1.000	-0.176
4	0	16	0	3.09773	0.0666	0.3045	2.174	-0.0613
5	0	18	0	3.03020	0.0675	0.1813	1.000	-0.119
6	0	20	0	2.95697	0.0732	0.1919	1.391	-0.121
7	0	23	0	2.91497	0.0420	0.1287	1.508	-0.0549
8	0	25	0	2.87198	0.0430	0.1799	2.847	-0.0369
9	0	27	0	2.84807	0.0239	0.2544	2.742	-0.0392
10	0	29	0	2.81013	0.0379	0.1052	0.975	-0.0614
11	0	31	0	2.76792	0.0422	0.2118	2.500	-0.0379
12	0	33	0	2.71086	0.0571	0.1616	2.289	-0.0461
13	0	36	0	2.68391	0.0269	0.1139	1.179	-0.0451
14	0	38	0	2.66655	0.0174	0.1709	3.225	-0.0209
15	0	42	0	2.61517	0.0514	0.0850	2.049	-0.0501
16	0	44	0	2.58166	0.0335	0.2016	3.756	-0.0342
17	0	46	0	2.54081	0.0408	0.0757	1.500	-0.0559
18	0	48	0	2.48690	0.0539	0.1517	3.458	-0.0295
19	0	51	0	2.46299	0.0239	0.1228	1.147	-0.0434
20	0	53	0	2.42921	0.0338	0.1146	1.507	-0.0397
21	0	55	0	2.38376	0.0455	0.1414	1.828	-0.0462
22	0	57	0	2.37174	0.0120	0.2618	2.029	-0.0560
23	0	61	0	2.31469	0.0570	0.1033	1.199	-0.0973
24	0	64	0	2.29211	0.0226	0.0907	1.237	-0.0372
25	0	66	0	2.26056	0.0316	0.0795	1.353	-0.0417
26	0	69	0	2.23947	0.0211	0.0911	1.342	-0.0313
27	0	71	0	2.21386	0.0256	0.1179	2.387	-0.0221
28	0	73	0	2.19967	0.0142	0.2386	2.550	-0.0251
29	0	75	0	2.17496	0.0247	0.0727	0.825	-0.0430
30	0	77	0	2.15706	0.0179	0.2163	2.241	-0.0276
31	0	79	0	2.13094	0.0261	0.0815	1.591	-0.0281
32	0	81	0	2.09830	0.0326	0.1617	2.309	-0.0283
33	0	83	0	2.08664	0.0117	0.2186	2.340	-0.0349
34	0	87	0	2.05825	0.0284	0.0645	1.195	-0.0488
35	0	89	0	2.03486	0.0234	0.2208	4.971	-0.0143
36	0	91	0	2.00266	0.0322	0.1518	0.972	-0.0601
37	0	94	0	1.98797	0.0147	0.0854	0.974	-0.0295
38	0	96	0	1.96846	0.0195	0.0811	2.030	-0.0181
39	0	98	0	1.94669	0.0218	0.1778	1.820	-0.0268
40	0	100	0	1.93085	0.0158	0.1038	1.964	-0.0277
41	0	102	0	1.91123	0.0196	0.0535	1.618	-0.0245
42	0	104	0	1.89588	0.0153	0.1603	3.463	-0.0142
43	0	106	0	1.87914	0.0167	0.0828	1.387	-0.0277
44	0	109	0	1.86765	0.0115	0.0771	1.976	-0.0113
45	0	111	0	1.85374	0.0139	0.1001	1.775	-0.0162
46	0	113	0	1.84900	0.00474	0.1036	2.476	-0.0140
47	0	117	0	1.83645	0.0126	0.0407	1.225	-0.0205
48	0	120	0	1.82791	0.00854	0.0711	2.873	-0.0058
49	0	122	0	1.81625	0.0117	0.0408	1.496	-0.0143
50	0	124	0	1.81327	0.00298	0.1053	3.110	-0.0094

Diagram 10.19: Optimization Results Neural Network

ACCESS

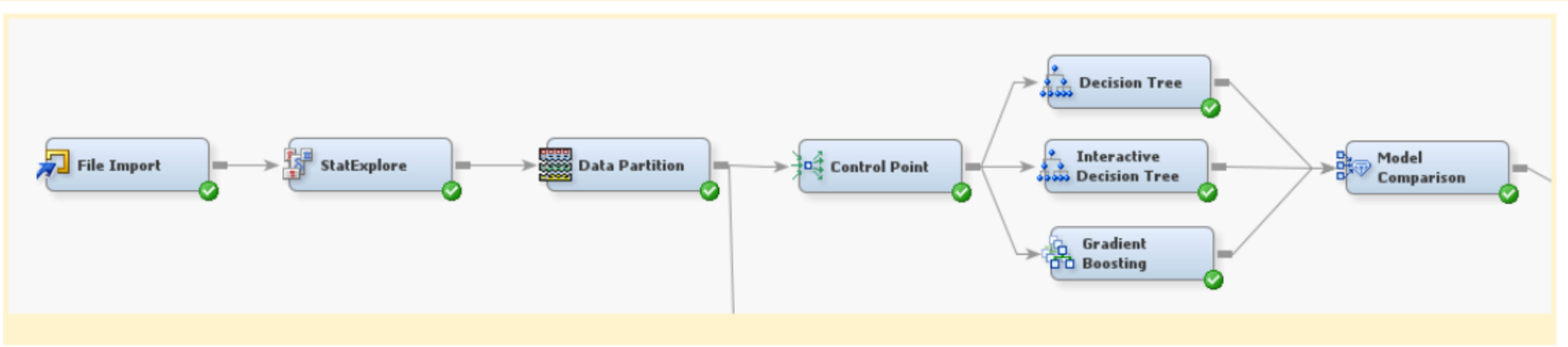


Diagram 11.1: Model Comparison Decision Tree & Gradient Boosting

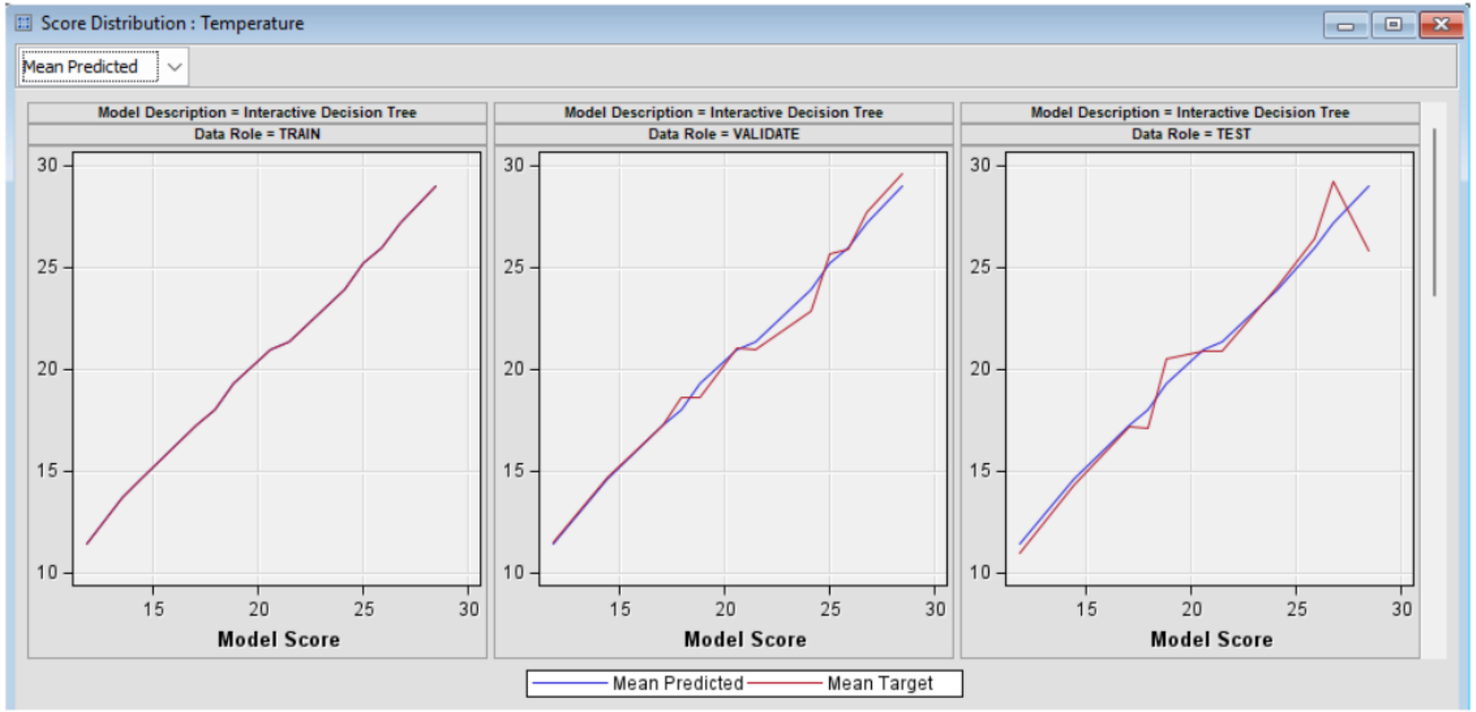


Diagram 11.2: Score Distribution Target Variable Interactive Decision tree

- Model comparison node is used for comparing the trainedd models

ACCESS

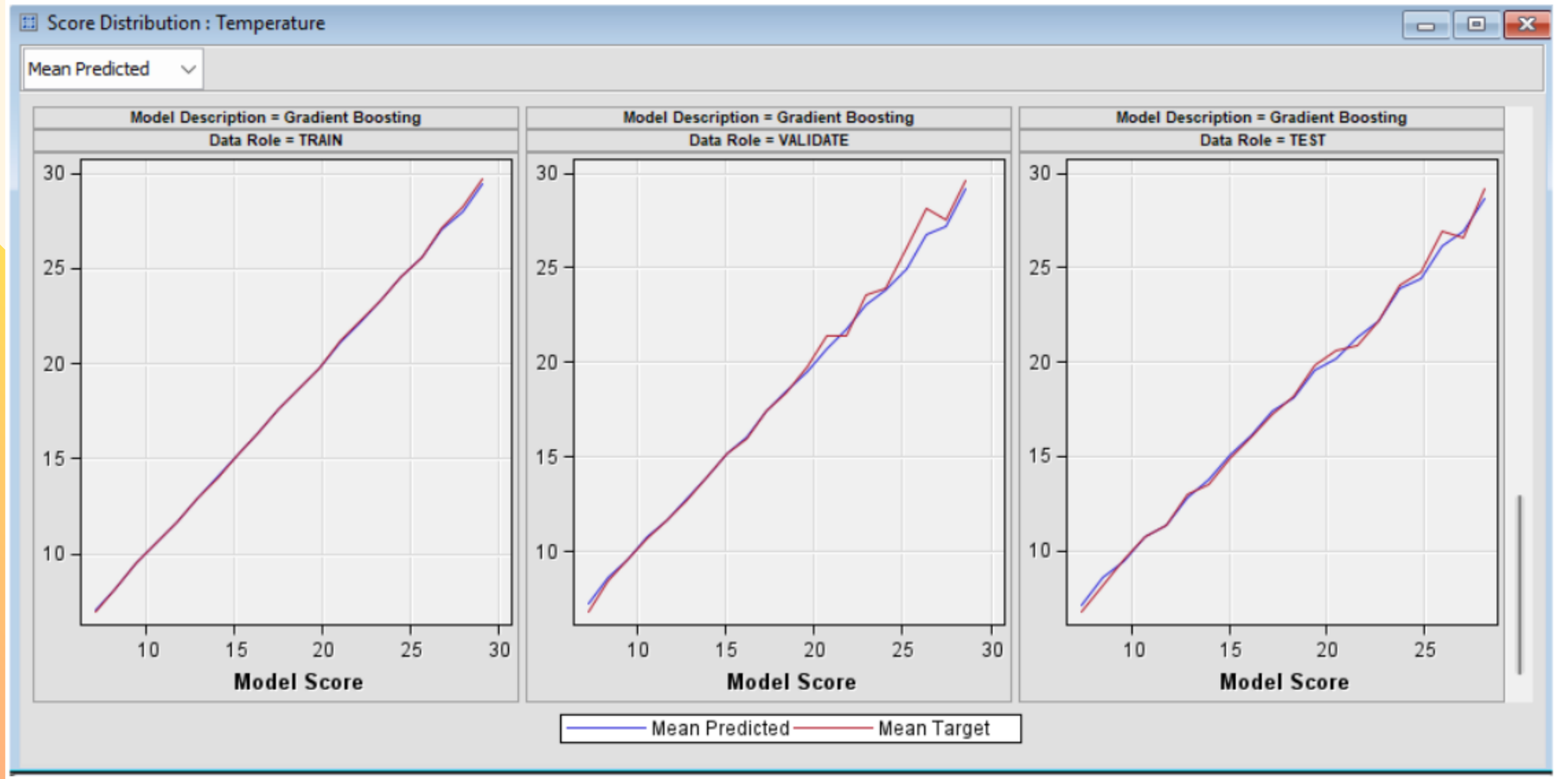
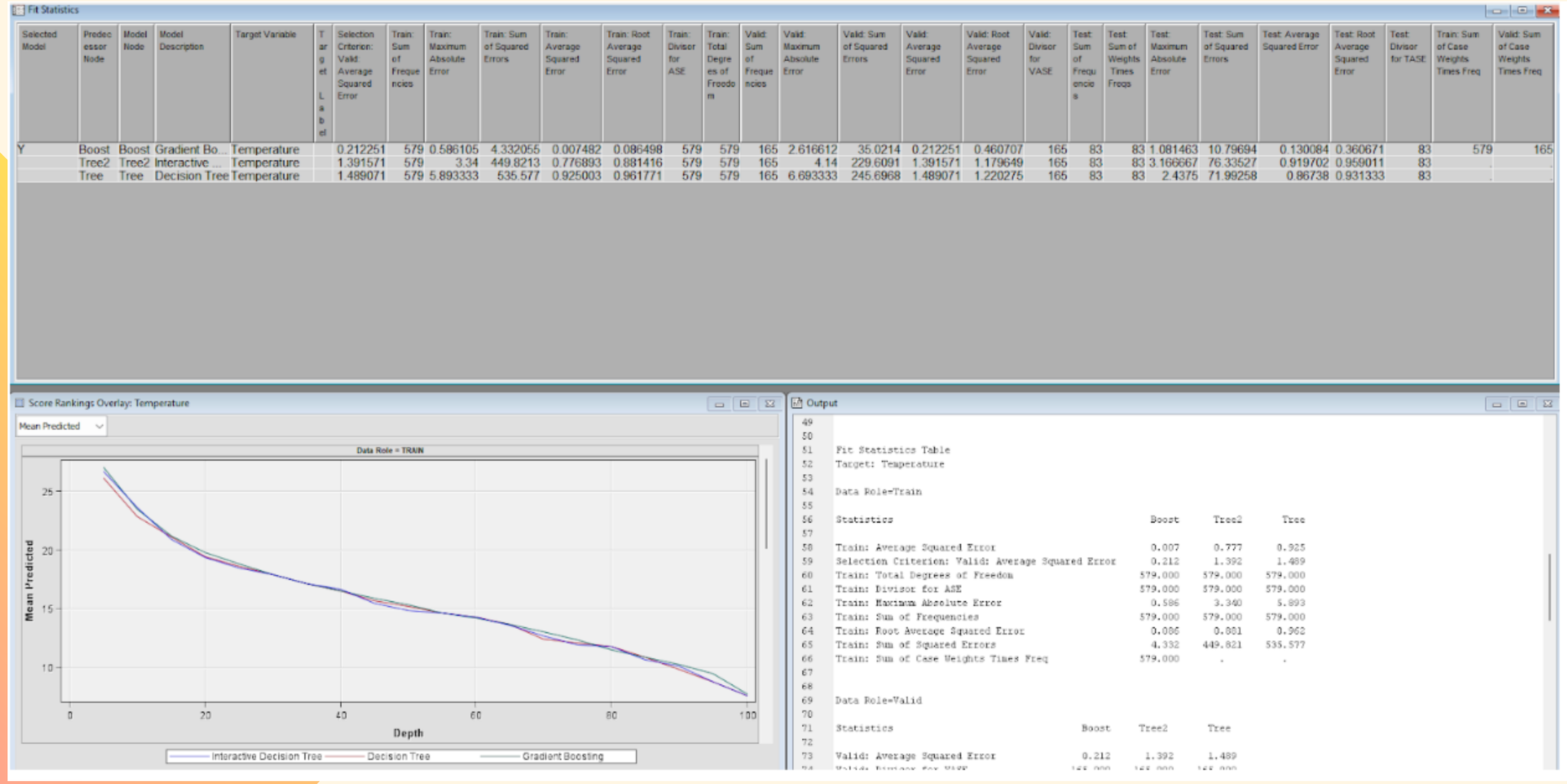


Diagram 11.3: Score Distribution Target Variable Gradient Boosting

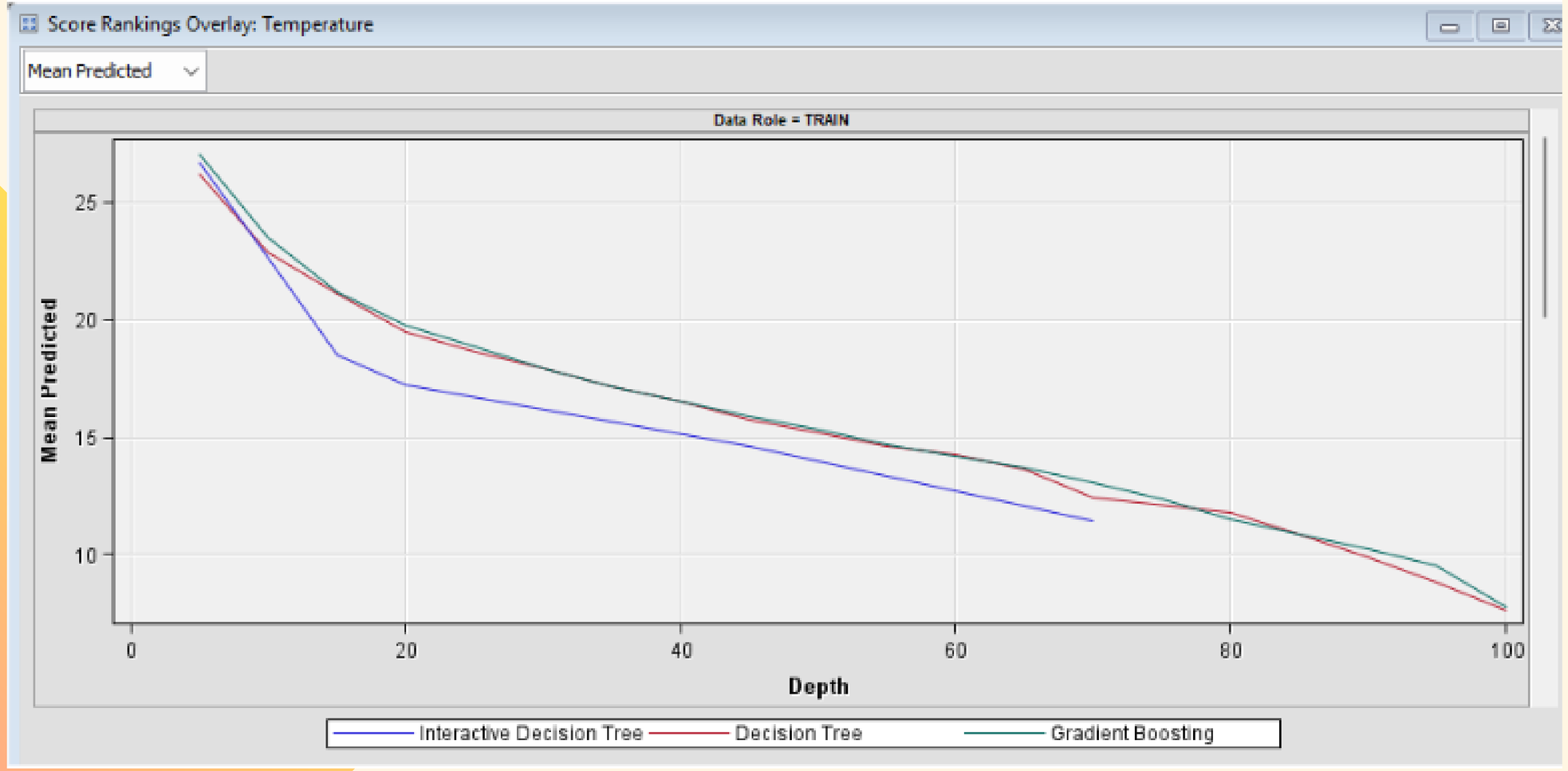
- The model's predictions are in agreement with the actual target values

ACCESS



- This diagram displays the results obtained from the Model Comparison Node

ACCESS



- The graph reveals that the decision tree and gradient boosting produced fairly similar results, but for gradient boosting showing a slightly higher curve

ACCESS

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train Divis ASE
Y	Boost	Boost	Gradient ...	Tempera...		0.212251	579	0.586105	4.332055	0.007482	0.086498	
	Tree	Tree	Decision ...	Tempera...		1.489071	579	5.893333	535.577	0.925003	0.961771	
	Tree2	Tree2	Interactiv...	Tempera...		6.933872	579	8.382895	3635.586	6.279078	2.505809	

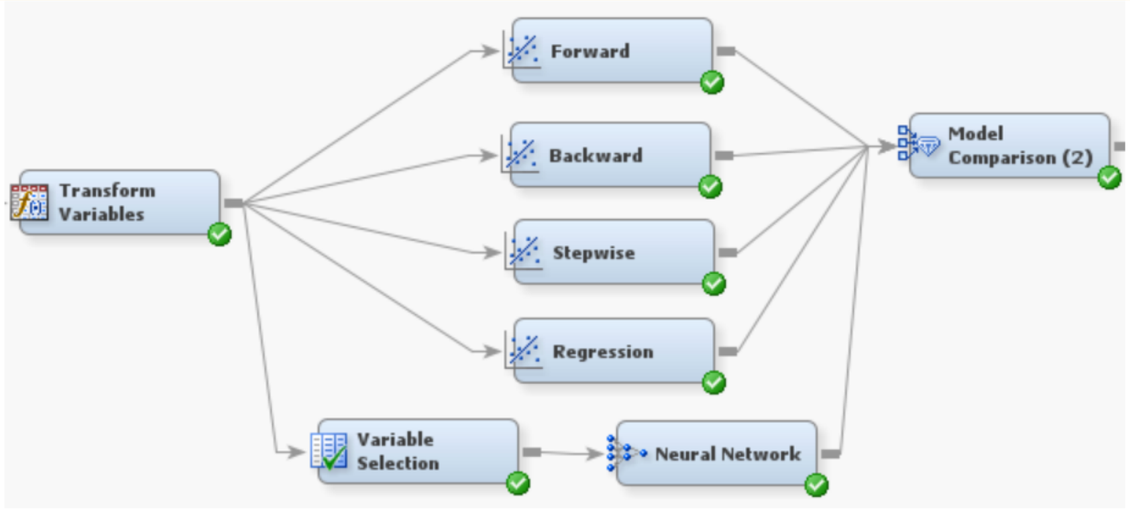
Fit Statistics

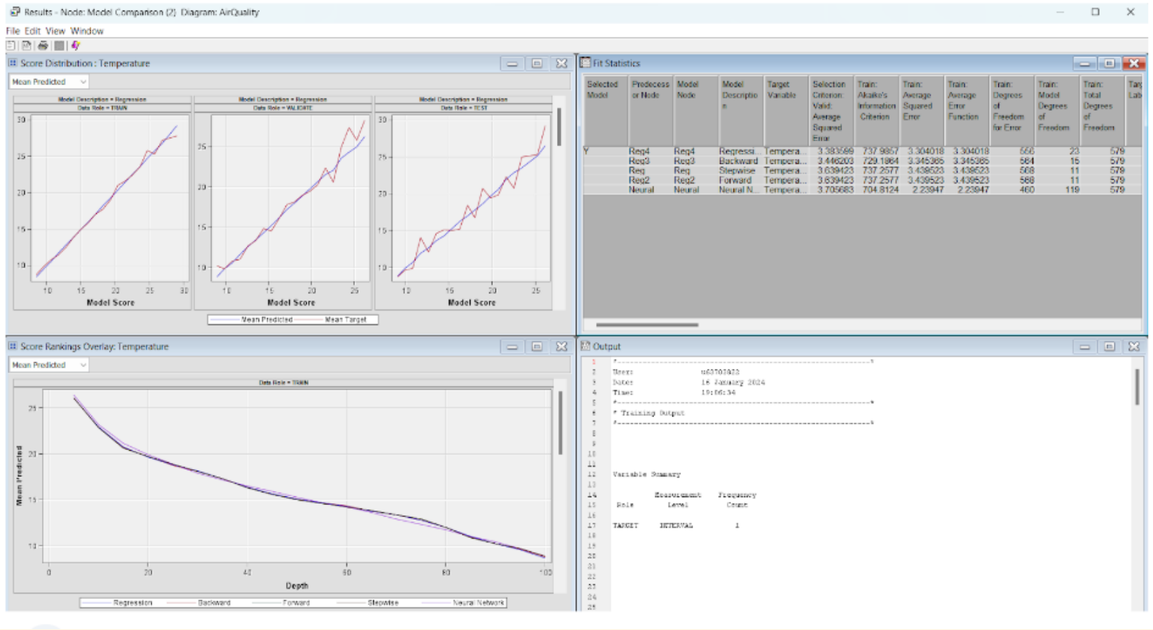
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Boost	Gradient Boosting	0.21225	0.00748
	Tree	Decision Tree	1.48907	0.92500
	Tree2	Interactive Decision Tree	6.93387	6.27908

- Gradient Boosting exhibits the lowest Average Squared Error when compared to the other models

ACCESS





- Comparison of comparison of Regression & Neural Network

ACCESS

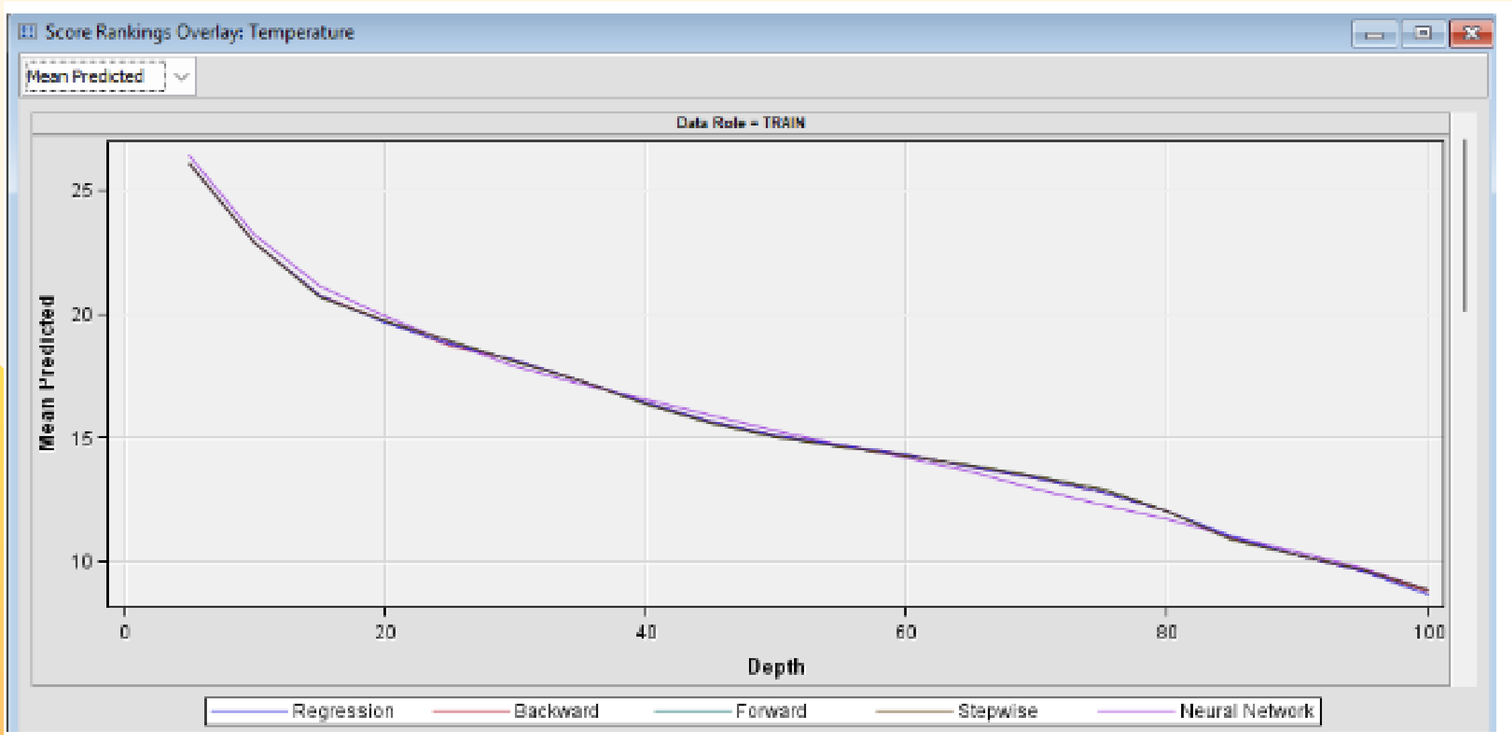


Diagram 11.11: Score Ranking Overlay Regression & Neural Network

- all models exhibit similar results on the line chart.
- However, a closer inspection reveals that our standard regression model outperforms the others.

ACCESS

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Trai Degi Free
Y	Reg4	Reg4	Regressi...	Tempera...		3.383599	737.9857	3.304018	3.304018	556	23	
	Reg3	Reg3	Backward	Tempera...		3.446203	729.1964	3.345365	3.345365	564	15	
	Reg	Reg	Stepwise	Tempera...		3.639423	737.2577	3.439523	3.439523	568	11	
	Reg2	Reg2	Forward	Tempera...		3.639423	737.2577	3.439523	3.439523	568	11	
	Neural	Neural	Neural N...	Tempera...		3.705683	704.8124	2.23947	2.23947	460	119	

Fit Statistics

Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate
Y	Reg4	Regression	3.38360	3.30402	.
	Reg3	Backward	3.44620	3.34537	.
	Reg	Stepwise	3.63942	3.43952	.
	Reg2	Forward	3.63942	3.43952	.
	Neural	Neural Network	3.70568	2.23947	.

- Regression emerges as the best model among Backward, Stepwise, Forward, and Neural Network this is because normal Regression has the lowest Average Squared Error

CONCLUSION

- SEMMA is playing an important role as it provides a good guide on handling the data and training correct and accurate models.
- For decision tree and gradient boosting, we determined that gradient boosting is the most suitable model for our dataset.
- For the regression and neural network aspects, we found that the original regression model is the most appropriate for our dataset.
- Each data mining model has its own strengths and limitations.

REFERENCES

Demir, F. (2022). Deep autoencoder-based automated brain tumor detection from MRI data. Artificial Intelligence-Based Bra
Computer Interface, 317-351. https://doi.org/10.1016/b978-0-323-91197-9.00013-8

Ghoson, A. M. A. (2011). Decision Tree Induction & Clustering Techniques In SAS Enterprise Miner, SPSS Clementine, And I
Intelligent Miner A Comparative Analysis. International Journal of Management & Information Systems, 14(3).
https://doi.org/10.19030/ijmis.v14i3.841

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Madhulatha, T. S. (2012). An overview on clustering methods. arXiv preprint arXiv:1205.1117,
https://doi.org/10.48550/arXiv.1205.1117.

Rithika, S. (2023, January 18). Sequence data in Data Mining Simplified 101 - learn. Hevo.
https://hevodata.com/learn/sequence-data-in-data-mining

THANKS FOR YOUR TIME