

**MIS 6324.003**

**Business Analytics with SAS**

**Project Report**



**GROUP 9**

Cheemalapati Jyothi, Rohini Reddy (rxc161830)

Danasekaran, Vignesh (vxd161130)

David Xavier Lourdu, Don Rudin (dxd161130)

Kumar, Akshara (axk160331)

Rajavel, Madhumitha (mpr160030)

# CONTENTS

CHAPTER	Page No.
1. INTRODUCTION.....	1
2. PROBLEM STATEMENT.....	1
3. DATA DESCRIPTION.....	1
4. HIGH LEVEL DATA SUMMARY.....	1
5. PROJECT DIAGRAM.....	3
6. DATA PRE- PROCESSING.....	4
7. FILE IMPORT.....	5
8. SAVE DATA.....	5
9. DATA SOURCE.....	5
10. STAT EXPLORER.....	6
11. DATA PARTITION.....	7
12. REPLACEMENT.....	8
13. IMPUTE.....	9
14. STAT EXPLORER_IMPUTED.....	10
15. DECISION TREE_ 2 BRANCH.....	11
16. DECISION TREE_ 3 BRANCH.....	14
17. DECISION TREE_ INTERACTIVE.....	16
18. GRADIENT BOOSTING.....	18
19. VARIABLE SELECTION.....	21
20. DECISION TREE_ VARIABLE SELECTION.....	22
21. TRANSFORM VARIABLES.....	24
22. STEPWISE_LOGISTIC REGRESSION.....	25
23. FORWARD_LOGISTIC REGRESSION.....	28
24. BACKWARD_LOGISTIC REGRESSION.....	31
25. NEURAL NETWORK.....	33
26. MODEL COMPARISON.....	36
27. CONCLUSION.....	38
28. REFERENCES.....	41

## Introduction

The most important source of survival for any living being on this planet is water. Though 71% of our planet is covered by water, safe water is still scarce on some parts of the planet. One among such hot and dry places is Tanzania, the largest country in East Africa with a population of over 52 million people. People here travel miles to get water from water pumps. Our project helps to draw a prediction about the functionality of these pumps and also helps the people to reach their nearest functional pump.

## Problem Statement

Our main aim is to check the functionality of water pumps in Tanzania for which the data sets have been collected. We would provide the following information after analyzing the data

1. The functional status of the pump
2. Factors affecting the functionality of water pumps

The analysis results can be used by various public and private sectors to yield fruitful results

1. It can help the Water Ministry of Tanzania to decide on which areas have the most functional failures and fix them to provide people with water.
2. Helps the locals to identify the nearest functional water pump
3. If proven that most of the pumps are not in a proper functional condition, social activists can come forward to help and assist maintain these pumps effectively.
4. Welfare societies and organizations who are willing to invest can use this data to determine where they can effectively use their funds to create the most impact

## Dataset Description

In our project we use second-hand data obtained from Drivendata. This multivariate dataset consists of a varied set of features about the water point like basin, extraction type, quantity of water, source type and so on. The data was collected by the Taarifa and the Tanzanian Ministry of Water. This dataset consists of 59400 observations across 40 attributes out of which 39 are predictive and 1 is non-predictive (ID).

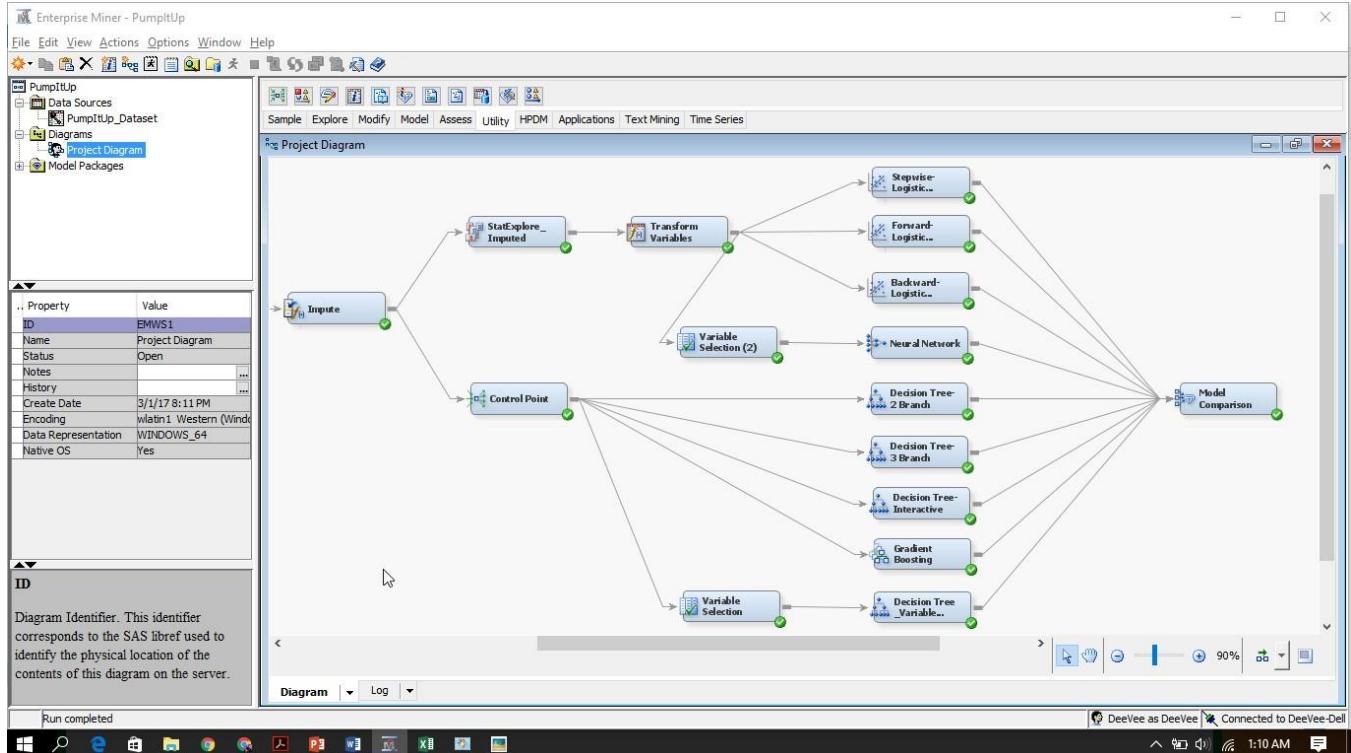
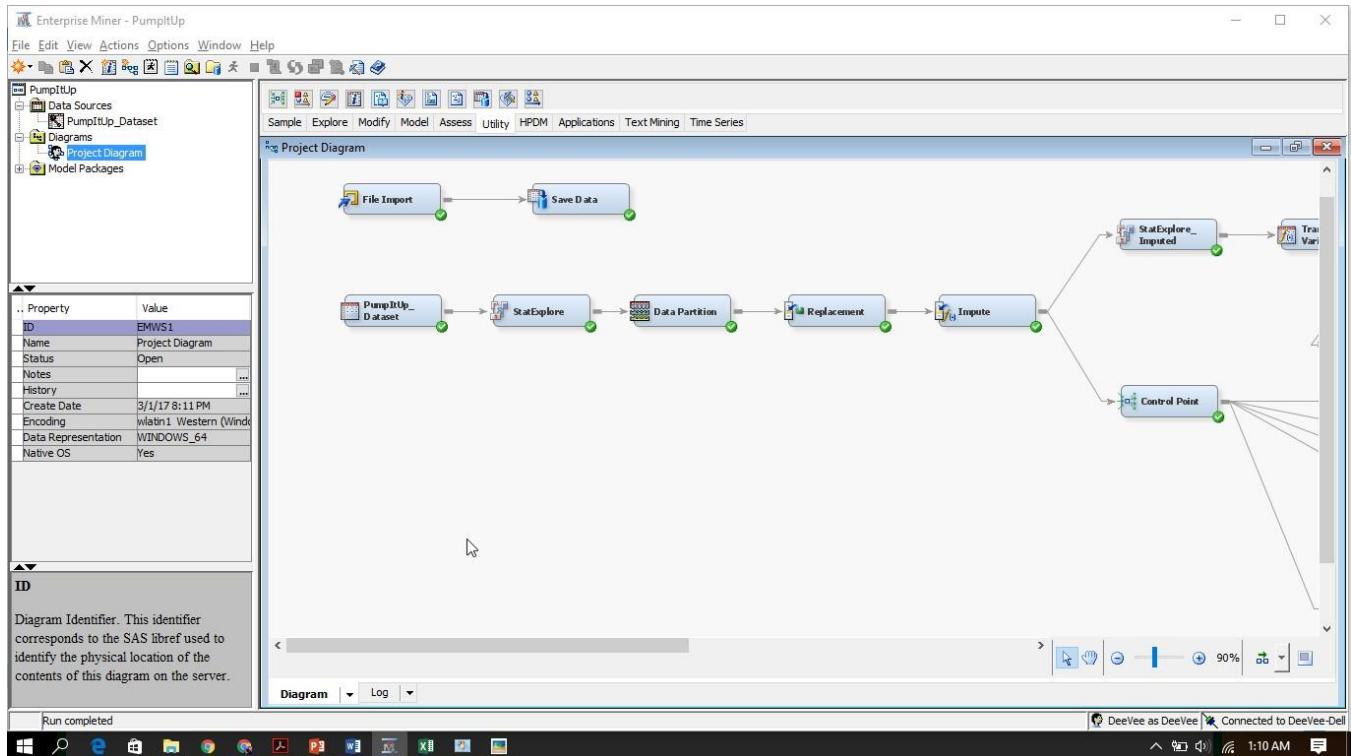
## High level Data summary:

Total observations in the dataset	59400
Total number of binary variables	0
Total number of nominal variables	29
Total number of interval variables	10
Outcome / target variable	status_group, quality_group
Level of the target variable (nominal, binary or interval)	Nominal
Percentage of the variables belong to each class.	Functional – 61.58% Non-functional – 38.42%

Given below is the description about each column used in our dataset:

1. amount\_tsh - Total static head (amount water available to waterpoint)
2. date\_recorded - The date the row was entered
3. funder - Who funded the well
4. gps\_height - Altitude of the well
5. installer - Organization that installed the well
6. longitude - GPS coordinate
7. latitude - GPS coordinate
8. wpt\_name - Name of the waterpoint if there is one
9. basin - Geographic water basin
10. subvillage - Geographic location
11. region - Geographic location
12. region\_code - Geographic location (coded)
13. district\_code - Geographic location (coded)
14. Iga - Geographic location
15. ward - Geographic location
16. population - Population around the well
17. public\_meeting - True/False
18. recorded\_by - Group entering this row of data
19. scheme\_management - Who operates the waterpoint
20. scheme\_name - Who operates the waterpoint
21. permit - If the waterpoint is permitted
22. construction\_year - Year the waterpoint was constructed
23. extraction\_type - The kind of extraction the waterpoint uses
24. extraction\_type\_group - The kind of extraction the waterpoint uses
25. extraction\_type\_class - The kind of extraction the waterpoint uses
26. management - How the waterpoint is managed
27. management\_group - How the waterpoint is managed
28. payment - What the water costs
29. payment\_type - What the water costs
30. water\_quality - The quality of the water
31. quality\_group - The quality of the water
32. quantity - The quantity of water
33. quantity\_group - The quantity of water
34. source - The source of the water
35. source\_type - The source of the water
36. source\_class - The source of the water
37. waterpoint\_type - The kind of waterpoint
38. waterpoint\_type\_group - The kind of waterpoint

## Project Diagram:



## Data Pre-Processing

Our dataset consists of 2 csv files (one for input variables and another for target variable). We combined these two into a single csv file.

### Feature selection:

- Initially **wpt\_name**, **scheme\_name**, **date\_recorded** and **recorded\_by** have been rejected since they cannot contribute to the actual prediction and do not have any predictive power
- Next **latitude** and **longitude** attributes have also been rejected since spatial data analysis is out of scope. We would also like to capture the location data using **basin**, **subvillage**, **region**, **Iga** and **ward**. We are working on clustering these five variables into a single variable with reduced class levels
- We are rejecting **funder** and **installer** as the data in these attributes contains more than 500 levels and Several inconsistencies have been found since the data entry was manual

E	F	G	H	I
gps_height	installer	longitude	latitude	wpt_name
	<div style="border: 1px solid #ccc; padding: 5px; display: inline-block;"> <span>A ↴ Sort A to Z</span>   <span>Z ↴ Sort Z to A</span>   <span>Sort by Color</span>   <span>Clear Filter From "installer"</span>   <span>Filter by Color</span>   <span>Text Filters</span>   <div style="border: 1px solid #ccc; padding: 2px; margin-top: 5px;">Search</div> <div style="border: 1px solid #ccc; padding: 2px; margin-top: 2px;"> <input checked="" type="checkbox"/> Got  <input checked="" type="checkbox"/> Gove  <input checked="" type="checkbox"/> Gover  <input checked="" type="checkbox"/> GOVERN  <input checked="" type="checkbox"/> GOVERN  <input checked="" type="checkbox"/> GOVERNME  <input checked="" type="checkbox"/> Governmen  <input checked="" type="checkbox"/> Government  <input checked="" type="checkbox"/> Government /Community           </div> </div>	39.1857105	-6.89259326	Chekanao
		39.52711433	-6.98874758	Msikitini
		39.15988725	-6.90254757	Kwa Chambuso
		39.17840407	-6.93801296	Ccm Kivule
		39.17884933	-6.97320593	Ofisi Ya Kata
		39.43371534	-7.09627911	Shuleni
		39.13540896	-6.91390781	none
		39.44163826	-6.90190842	Kwa Weso
		39.11655424	-6.89004147	Kwa Mzee Iddy
		39.0807633	-7.00018132	Office
		39.35517325	-6.86048297	Kwa Yohana
		39.11067764	-6.59136797	Sekondari Ya Mbweni Mpiji
		39.21996168	-6.91649866	Mzinga Ccm
		39.13603616	-6.6852803	Street Ofice Ground
		39.15886765	-6.89916957	Kijiweni
		39.12016025	-6.89327148	Kwa Mama Groly
		39.09728159	-6.96544389	Kwa Mzee Jongo
		39.434248	-7.09943384	Kwa Binti Hatibu
		39.12651326	-6.99545576	Shule Ya Msingi Yongwe
		39.04833589	-6.75092197	Zahanati Ya Kibwegere
55		0	39.09194412	Makao Makuu
48		0	39.1086116	Kwa Zainabu Sultan

All the values that are shown above point to Government. This leads to several unique levels for these attributes and will cause issues while predicting. And the name of the Funder and Installer will not contribute much for the prediction. This is the reason behind rejecting **funder** and **installer**.

- **amount\_tsh** has been rejected as its mode is 0 and it is difficult to determine whether it is missing data or not
- **payment\_type** and **payment** have exactly the same values and so we reject the **payment\_type**
- **num\_private** has too little information and so we rejected it

## 1. File Import:

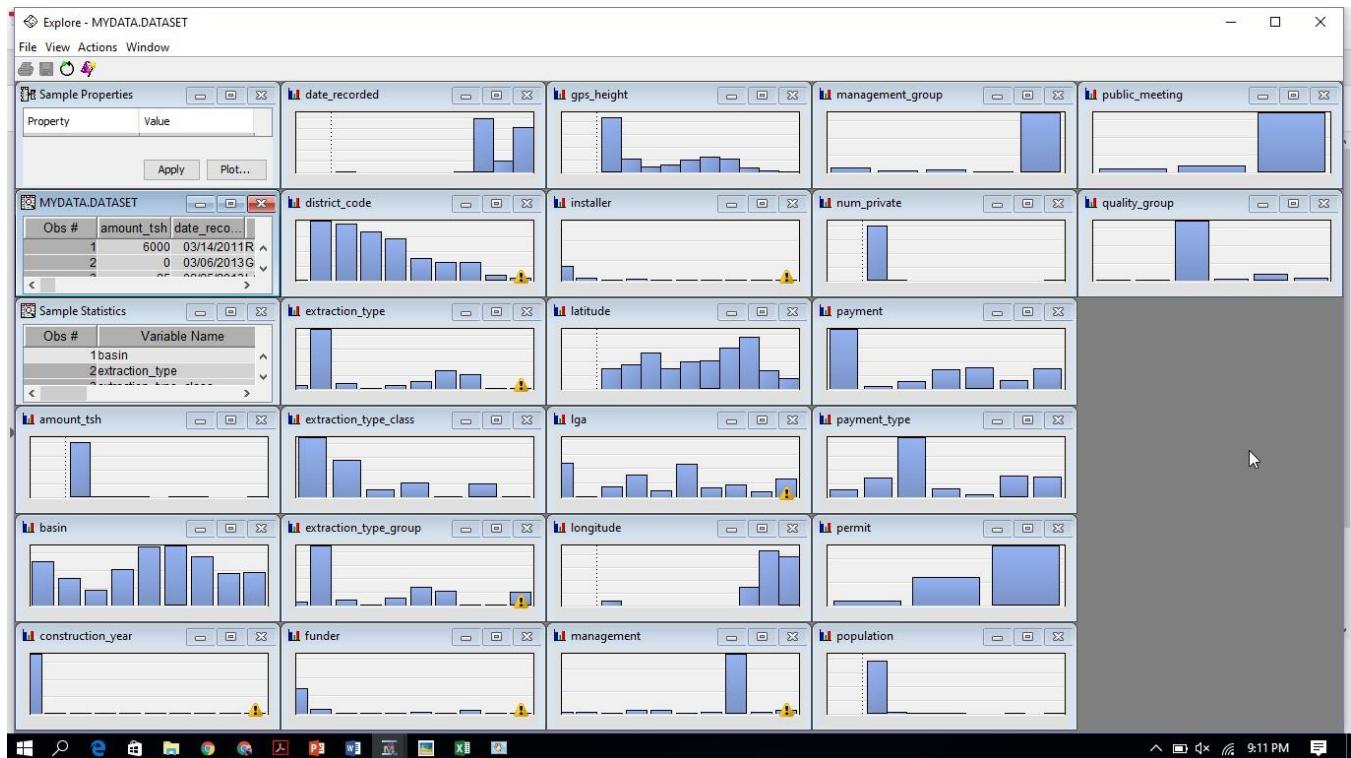
Our dataset consists of 2 csv files (one for input variables and another for target variable). We combined these two into a single csv file. This csv file was then imported using the File import node.

## 2. Save Data:

This node was used to convert the csv file to sas7bdat file.

## 3. Data Source:

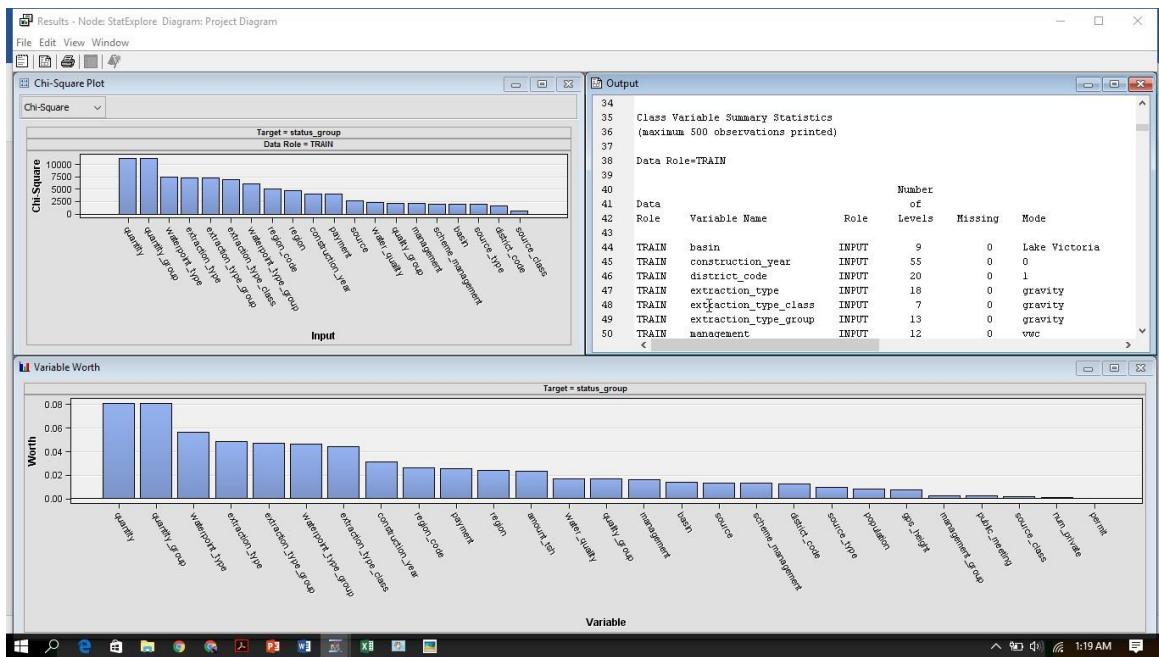
The above saved sas7bdat file was then created as a new data source.



Obs #	Variable Name	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1	basin		CLASS	0	.	.	.9		16.95	PANGANI
2	extraction_type		CLASS	0	.	.	.14		45	GRAVITY
3	extraction_type_class		CLASS	0	.	.	.7		45	GRAVITY
4	extraction_type_group		CLASS	0	.	.	.12		45	GRAVITY
5	funder		CLASS	8.668342	.	.	.128+		20.72864	GOVERNMENT
6	installer		CLASS	8.564387	.	.	.128+		35.92113	DWE
7	iga		CLASS	0	.	.	.123		4.2	JUMBE
8	management		CLASS	0	.	.	.12		66.55	WIC
9	management_group		CLASS	0	.	.	.5		87.3	USER-GROUP
10	payment		CLASS	0	.	.	.7		43.75	NEVER PAY
11	payment_type		CLASS	0	.	.	.7		43.75	NEVER PAY
12	permit		CLASS	5.25	.	.	.3		64.3	TRUE
13	public_meeting		CLASS	5.3	.	.	.3		85.75	TRUE
14	quality_group		CLASS	0	.	.	.6		84.75	GOOD
15	quantity		CLASS	0	.	.	.5		53	ENOUGH
16	quantity_group		CLASS	0	.	.	.5		53	ENOUGH
17	recorded_by		CLASS	0	.	.	.1		100	GEO DATA
18	region		CLASS	0	.	.	.21		9	SHINYANGA
19	scheme_management		CLASS	6.8	.	.	.11		60.9	WIC
20	scheme_name		CLASS	69.59022	.	.	.128+		2.228613K	
21	source		CLASS	0	.	.	.9		28.3	SPRING
22	source_class		CLASS	0	.	.	.3		77.05	GROUNDWATER
23	source_type		CLASS	0	.	.	.7		28.3	SPRING
24	status_group		CLASS	0	.	.	.3		53.5	FUNCTIONAL
25	subvillage		CLASS	0	.	.	.128+		8.571429	SHULENI
26	ward		CLASS	0	.	.	.128+		3.714286	MDANDU
27	water_quality		CLASS	0	.	.	.7		84.75	SOFT
28	waterpoint_type		CLASS	0	.	.	.6		48.1	COMMUNAL
29	waterpoint_type_group		CLASS	0	.	.	.5		58.25	COMMUNAL
30	wpt_name		CLASS	0	.	.	.128+		30.07519	NONE
31	amount_tsh		VAR	0	0	40000	254.2015			
32	construction_year		VAR	0	0	2013	1299.544			
33	date_recorded		VAR	0	16284	19695	19076.14			
34	district_code		VAR	0	0	63	5.8875			
35	gps_height		VAR	0	-45	2623	665.313			
36	latitude		VAR	0	-11.5643	-2E-8	-5.76113			
37	longitude		VAR	0	0	40.34519	34.04413			
38	num_private		VAR	0	0	698	0.5785			

#### 4. Stat Explore:

This node was used to generate a brief summary about the input variables. From the result, we found a few variables having missing or incorrect values. Shown below is the summary:



```

34
35 Class Variable Summary Statistics
36 (maximum 500 observations printed)
37
38 Data Role=TRAIN
39
40
41 Data Role Variable Name Role Number of Levels Missing Mode Percentage Node2 Percentage
42
43
44 TRAIN basin INPUT 9 0 Lake Victoria 17.25 Pangani 15.05
45 TRAIN construction_year INPUT 55 0 0 34.86 2010 4.45
46 TRAIN district_code INPUT 20 0 1 20.54 2 18.81
47 TRAIN extraction_type INPUT 18 0 gravity 45.08 nira/tanira 13.73
48 TRAIN extraction_type_class INPUT 7 0 gravity 45.08 handpump 27.70
49 TRAIN extraction_type_group INPUT 13 0 gravity 45.08 nira/tanira 13.73
50 TRAIN management INPUT 12 0 vvc 68.19 wug 10.97
51 TRAIN management_group INPUT 5 0 user-group 88.37 commercial 6.12
52 TRAIN payment INPUT 7 0 never pay 42.67 pay per bucket 15.13
53 TRAIN permit INPUT 3 3056 TRUE 65.41 FALSE 29.45
54 TRAIN public_meeting INPUT 3 3334 TRUE 85.88 FALSE 8.51
55 TRAIN quality_group INPUT 6 0 good 85.55 salty 8.75
56 TRAIN quantity INPUT 5 0 enough 55.87 insufficient 25.47
57 TRAIN quantity_group INPUT 5 0 enough 55.87 insufficient 25.47
58 TRAIN region INPUT 21 0 Iringa 8.91 Shinyanga 8.39
59 TRAIN region_code INPUT 27 0 ll 8.92 17 8.44
60 TRAIN scheme_management INPUT 13 3877 VVC 61.94 WUG 8.76
61 TRAIN source INPUT 10 0 spring 28.65 shallow well 28.32
62 TRAIN source_class INPUT 3 0 groundwater 77.09 surface 22.44
63 TRAIN source_type INPUT 7 0 spring 28.65 shallow well 28.32
64 TRAIN water_quality INPUT 8 0 soft 85.55 salty 8.18
65 TRAIN waterpoint_type INPUT 7 0 communal standpipe 48.02 hand pump 29.44
66 TRAIN waterpoint_type_group INPUT 6 0 communal standpipe 58.29 hand pump 29.44
67 TRAIN status_group TARGET 3 0 functional 94.31 non functional 38.42
68
69
70
71 Distribution Summary of Class, Target, and General Variables

```

From the above screenshot, we find that the variable ***construction\_year*** has a mode value of 0 which is incorrect. Replacement and Imputation of such values will take place in the later stages of this flow.

## 5. Data Partition:

This node has been used to partition our dataset into three parts namely,

Training set: 60%

Validation set: 20%

Test set: 20%

```

10
11
12 Variable Summary
13
14 Measurement Frequency
15 Role Level Count
16
17 INPUT INTERVAL 4
18 INPUT NOMINAL 24
19 TARGET NOMINAL 1
20
21
22
23
24 Partition Summary
25
26 Number of
27 Type Data Set Observations
28
29 DATA EMWS1.Stat_TRAIN 59400
30 TRAIN EMWS1.Part_TRAIN 35636
31 VALIDATE EMWS1.Part_VALIDATE 11800
32 TEST EMWS1.Part_TEST 11804
33
34
35 *-----*
36 * Score Output
37 *-----*
38
39
40 *-----*
41 * Report Output
42 *-----*

```

## 6. Replacement:

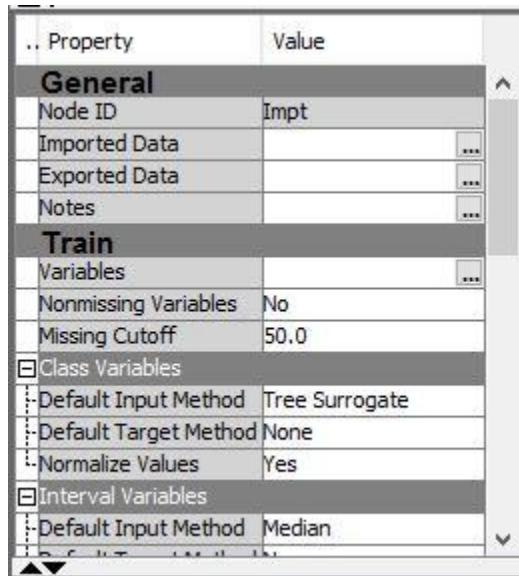
This node was used to replace the incorrect values like “unknown” with “\_Missing\_”. This node also helped us to replace the “functional needs repair” value of the **status\_group** variable with “functional”.

Variable	Role	Label	Train ▾	Validation	Test
construction_year	INPUT			12442	4113
payment	INPUT			4967	1616
status_group	TARGET			2589	865
quality_group	INPUT			1115	401
water_quality	INPUT			1115	401
management	INPUT			342	114
management_group	INPUT			342	114
source_class	INPUT			176	59
source	INPUT			44	14
basin	INPUT			0	0
district_code	INPUT			0	0
extraction_type	INPUT			0	0
extraction_type_class	INPUT			0	0
extraction_type_group	INPUT			0	0
permit	INPUT			0	0
public_meeting	INPUT			0	0
quantity	INPUT			0	0
quantity_group	INPUT			0	0
region	INPUT			0	0
region_code	INPUT			0	0
scheme_management	INPUT			0	0
source_type	INPUT			0	0
waterpoint_type	INPUT			0	0
waterpoint_type_group	INPUT			0	0

Line	Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value	Label
31							
32							
33							
34	basin	Unknown	C	.	.	_blank_	
35	construction_year	0	N	0	.	.	
36	construction_year	Unknown	N	.	.	.	
37	district_code	Unknown	N	.	.	.	
38	extraction_type	Unknown	C	.	.	_blank_	
39	extraction_type_class	Unknown	C	.	.	_blank_	
40	extraction_type_group	Unknown	C	.	.	_blank_	
41	management	unknown	C	unknown	.	_Missing_	
42	management	Unknown	C	.	.	_blank_	
43	management_group	unknown	C	unknown	.	_Missing_	
44	management_group	Unknown	C	.	.	_blank_	
45	payment	Unknown	C	unknown	.	_Missing_	
46	payment	Unknown	C	.	.	_blank_	
47	permit	Unknown	C	.	.	_blank_	
48	public_meeting	Unknown	C	.	.	_blank_	
49	quality_group	unknown	C	unknown	.	_Missing_	
50	quality_group	Unknown	C	.	.	_blank_	
51	quantity	Unknown	C	.	.	_blank_	
52	quantity_group	Unknown	C	.	.	_blank_	
53	region	Unknown	C	.	.	_blank_	
54	region_code	Unknown	N	.	.	.	
55	scheme_management	Unknown	C	.	.	_blank_	
56	source	unknown	C	unknown	.	_Missing_	
57	source	Unknown	C	.	.	_blank_	
58	source_class	unknown	C	unknown	.	_Missing_	
59	source_class	Unknown	C	.	.	_blank_	
60	source_type	Unknown	C	.	.	_blank_	
61	status_group	functional needs repair	C	functional needs repair	.	functional	
62	status_group	Unknown	C	.	.	_blank_	
63	water_quality	Unknown	C	unknown	.	_Missing_	
64	water_quality	Unknown	C	.	.	_blank_	
65	waterpoint_type	Unknown	C	.	.	_blank_	
66	waterpoint_type_group	Unknown	C	.	.	_blank_	
67							
68							

## 7. Impute:

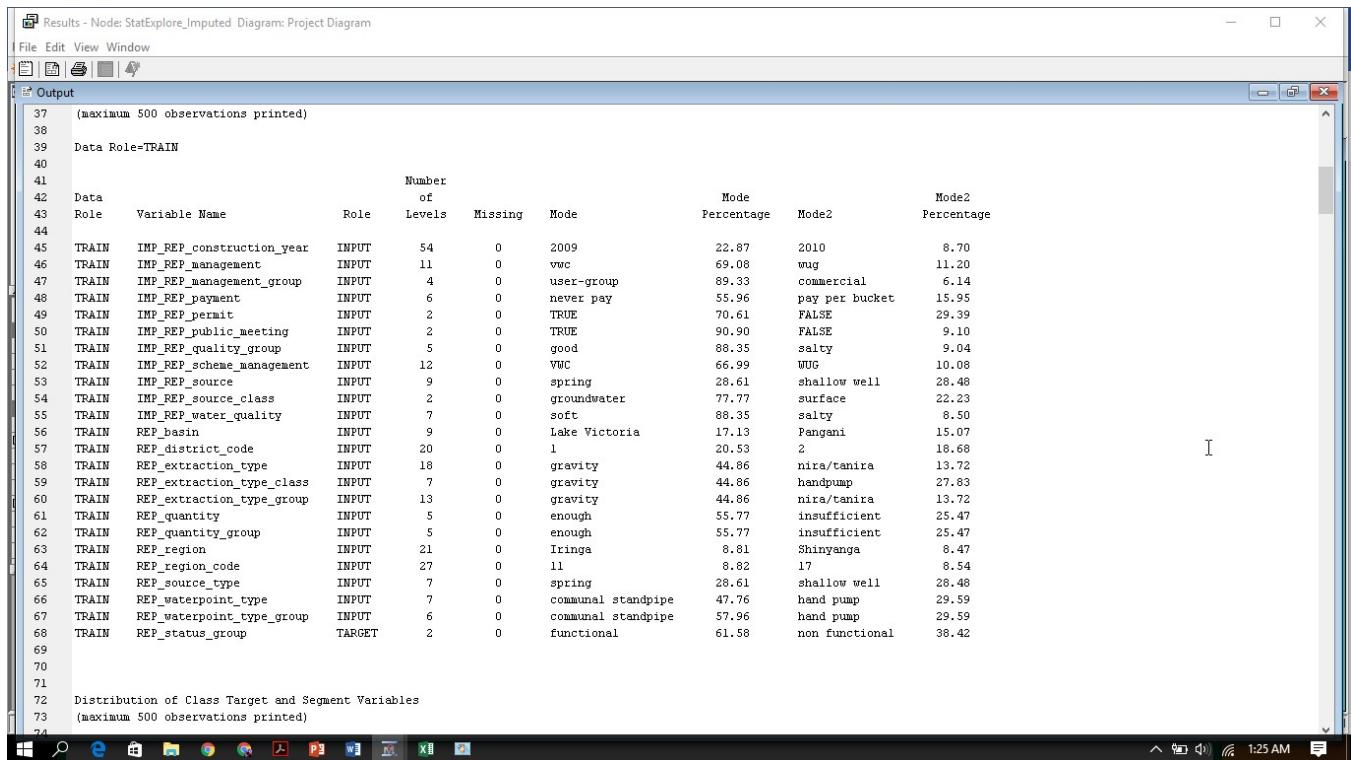
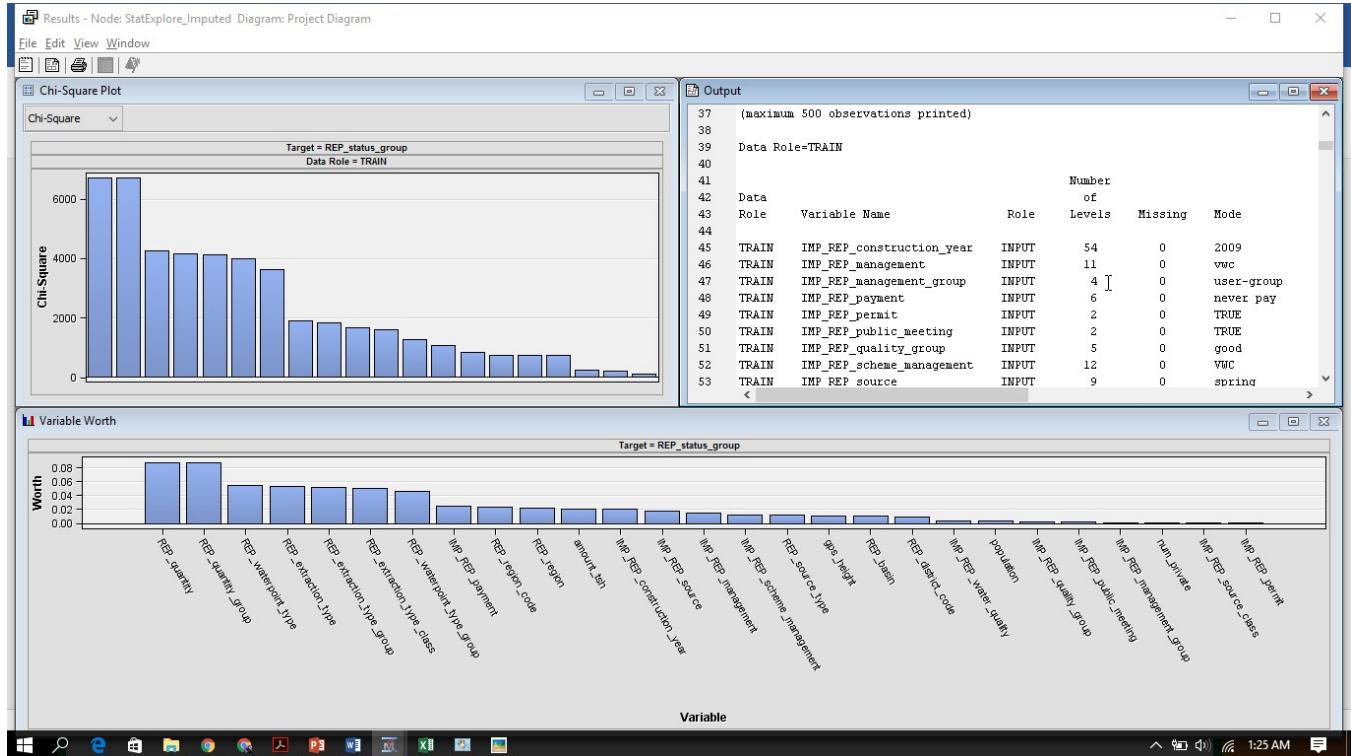
This node was used to impute the missing values for the following variables using 'Tree Surrogate' method: ***Construction\_year, payment, permit, public\_meeting, scheme\_management, quality\_group, water\_quality, management, management\_group, source\_class and source.***



Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP_construction_year	TREESURR	IMP REP_construction_year		.INPUT	NOMINAL	Replacement: construction_...	12442
REP_payment	TREESURR	IMP REP_payment		.INPUT	NOMINAL	Replacement: payment	4967
REP_scheme_management	TREESURR	IMP REP_scheme_manag...		.INPUT	NOMINAL	Replacement: scheme_ma...	2328
REP_public_meeting	TREESURR	IMP REP_public_meeting		.INPUT	NOMINAL	Replacement: public_meeting	1986
REP_permit	TREESURR	IMP REP_permit		.INPUT	NOMINAL	Replacement: permit	1863
REP_quality_group	TREESURR	IMP REP_quality_group		.INPUT	NOMINAL	Replacement: quality_group	1115
REP_water_quality	TREESURR	IMP REP_water_quality		.INPUT	NOMINAL	Replacement: water_quality	1115
REP_management	TREESURR	IMP REP_management		.INPUT	NOMINAL	Replacement: management	342
REP_management_group	TREESURR	IMP REP_management_gr...		.INPUT	NOMINAL	Replacement: management...	342
REP_source_class	TREESURR	IMP REP_source_class		.INPUT	NOMINAL	Replacement: source_class	176
REP_source	TREESURR	IMP REP_source		.INPUT	NOMINAL	Replacement: source	44

## 8. Stat Explore\_Imputed:

This node was used to generate a brief summary about the input variables after imputation.



## 9. Decision Tree- 2 Branch:

This tree was used to predict the ***status\_group*** target variable from the imputed data with the following properties:

.. Property	Value
<b>Splitting Rule</b>	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
<b>Node</b>	
Leaf Size	50
Number of Rules	5
Number of Surrogate Rules	4
Split Size	,

```

Results - Node: Decision Tree- 2 Branch Diagram: Project Diagram
File Edit View Window
Output
61
62 Variable Importance
63
64
65
66
67 Variable Name Label Number of Splitting Number of Surrogate Validation Ratio of
68 Rules Rules Importance Importance Validation to Training Importance
69 IMP REP_construction_year Imputed: Replacement: construction_year 8 10 1.0000 1.0000 1.0000
70 REP_quantity Replacement: quantity 2 1 0.9632 0.9627 0.9995
71 REP_quantity_group Replacement: quantity_group 0 3 0.9632 0.9627 0.9995
72 REP_extraction_type Replacement: extraction_type 1 9 0.7874 0.7470 0.9734
73 REP_extraction_type_class Replacement: extraction_type_class 2 7 0.7628 0.7431 0.9742
74 REP_extraction_type_group Replacement: extraction_type_group 2 7 0.7584 0.7373 0.9722
75 REP_region_code Replacement: region_code 2 30 0.6886 0.6925 1.0056
76 REP_waterpoint_type Replacement: waterpoint_type 2 3 0.6546 0.6275 0.9586
77 REP_waterpoint_type_group Replacement: waterpoint_type_group 2 0 0.6363 0.6040 0.9492
78 REP_region Replacement: region 6 24 0.6018 0.5954 0.9893
79 IMP REP_source Imputed: Replacement: source 2 13 0.6006 0.6077 1.0117
80 REP_basin Replacement: basin 4 14 0.4621 0.4637 1.0034
81 REP_district_code Replacement: district_code 1 12 0.3369 0.2910 0.8636
82 REP_source_type Replacement: source_type 1 6 0.3198 0.3201 1.0011
83 gps_height 3 8 0.3030 0.2844 0.9387
84 IMP REP_payment Imputed: Replacement: payment 2 0 0.3009 0.3323 1.1044
85 IMP REP_scheme_management Imputed: Replacement: scheme_management 3 6 0.2907 0.2935 1.0098
86 IMP REP_management Imputed: Replacement: management 2 7 0.2899 0.2907 1.0028
87 amount_tsh 1 2 0.2846 0.3232 1.1356
88 population 1 10 0.2525 0.2614 1.0351
89 IMP REP_management_group Imputed: Replacement: management_group 1 3 0.1365 0.1156 0.8472
90 IMP REP_quality_group Imputed: Replacement: quality_group 1 2 0.1300 0.1188 0.9139
91 IMP REP_public_meeting Imputed: Replacement: public_meeting 0 2 0.1017 0.0895 0.8799
92 IMP REP_water_quality Imputed: Replacement: water_quality 0 1 0.0993 0.0798 0.8043
93 IMP REP_permit Imputed: Replacement: permit 2 0 0.0705 0.0675 0.9576
94 IMP REP_source_class Imputed: Replacement: source_class 0 1 0.0449 0.0479 1.0680
95
96
97
98 Tree Leaf Report

```

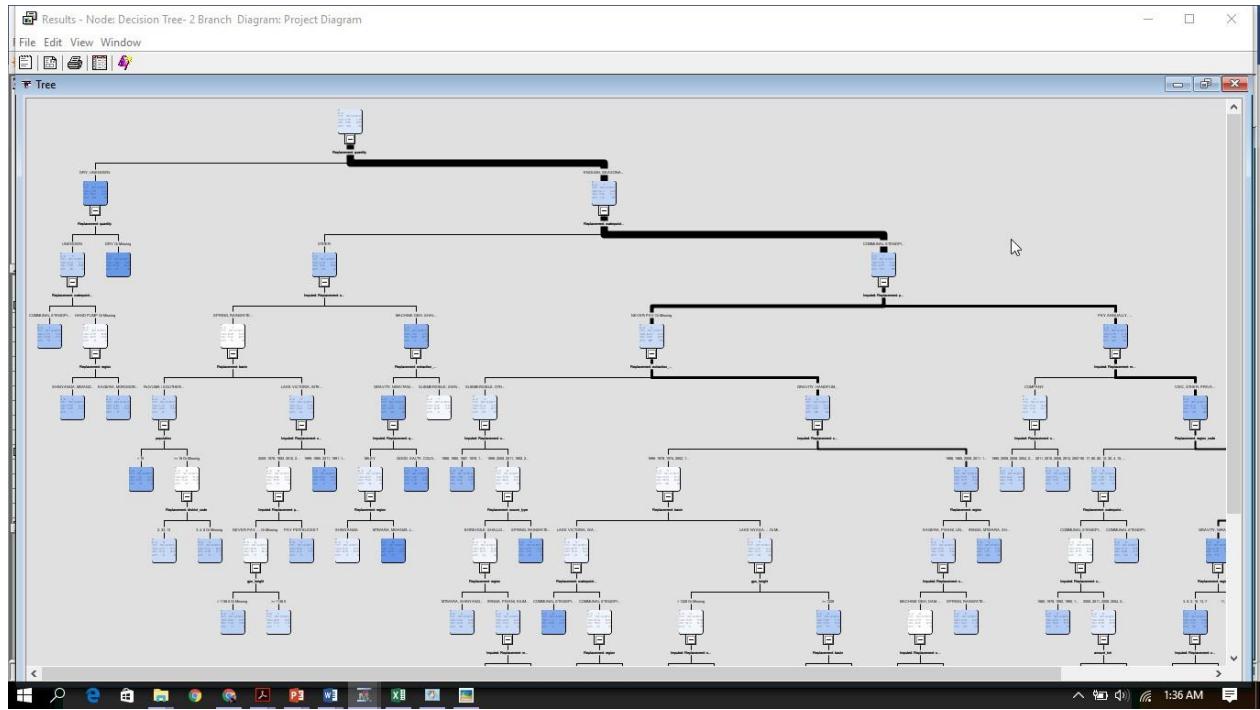
It can be seen that "Quantity" has high variable importance (splitting rules= 2) than the other variables to determine the "Status\_Group". It has a variable importance of 0.9627.

Results - Node: Decision Tree- 2 Branch Diagram: Project Diagram

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement: status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement: status_group	_MISC_	Misclassification Rate	0.185487	0.191162	0.190003
REP_status_group	Replacement: status_group	_MAX_	Maximum Absolute Error	0.98	0.98	0.98
REP_status_group	Replacement: status_group	_SSE_	Sum of Squared Errors	9969.863	3426.55	3382.687
REP_status_group	Replacement: status_group	_ASE_	Average Squared Error	0.139885	0.144215	0.142321
REP_status_group	Replacement: status_group	_RASE_	Root Average Squared Error	0.374012	0.379757	0.377255
REP_status_group	Replacement: status_group	_DIV_	Divisor for ASE	71272	23760	23768
REP_status_group	Replacement: status_group	_DFT_	Total Degrees of Freedom	35636	-	-





Below we can see the Node Rules window containing the IF-THEN logic that distributes observations into each leaf node of the decision tree.

```

Results - Node: Decision Tree- 2 Branch Diagram: Project Diagram
File Edit View Window
Tree Node Rules
1 -----
2 Node = 5
3 -----
4 if Replacement: quantity IS ONE OF: DRY or MISSING
5 then
6 Tree Node Identifier = 5
7 Number of Observations = 3781
8 Predicted: REP_status_group=non functional = 0.97
9 Predicted: REP_status_group=functional = 0.03
10 -----
11 -----
12 Node = 8
13 -----
14 if Replacement: waterpoint_type_group IS ONE OF: COMMUNAL STANDPIPE, OTHER
15 AND Replacement: quantity IS ONE OF: UNKNOWN
16 then
17 Tree Node Identifier = 8
18 Number of Observations = 350
19 Predicted: REP_status_group=non functional = 0.78
20 Predicted: REP_status_group=functional = 0.22
21 -----
22 -----
23 Node = 18
24 -----
25 if Replacement: waterpoint_type_group IS ONE OF: HAND PUMP or MISSING
26 AND Replacement: region IS ONE OF: SHINYANGA, MWANZA, MBEYA, SINGIDA, MARA
27 AND Replacement: quantity IS ONE OF: UNKNOWN
28 then
29 Tree Node Identifier = 18
30 Number of Observations = 61
31 Predicted: REP_status_group=non functional = 0.23
32 Predicted: REP_status_group=functional = 0.77
33 -----
34 -----
35 Node = 19
36 -----
37 if Replacement: waterpoint type group IS ONE OF: HAND PUMP or MISSING

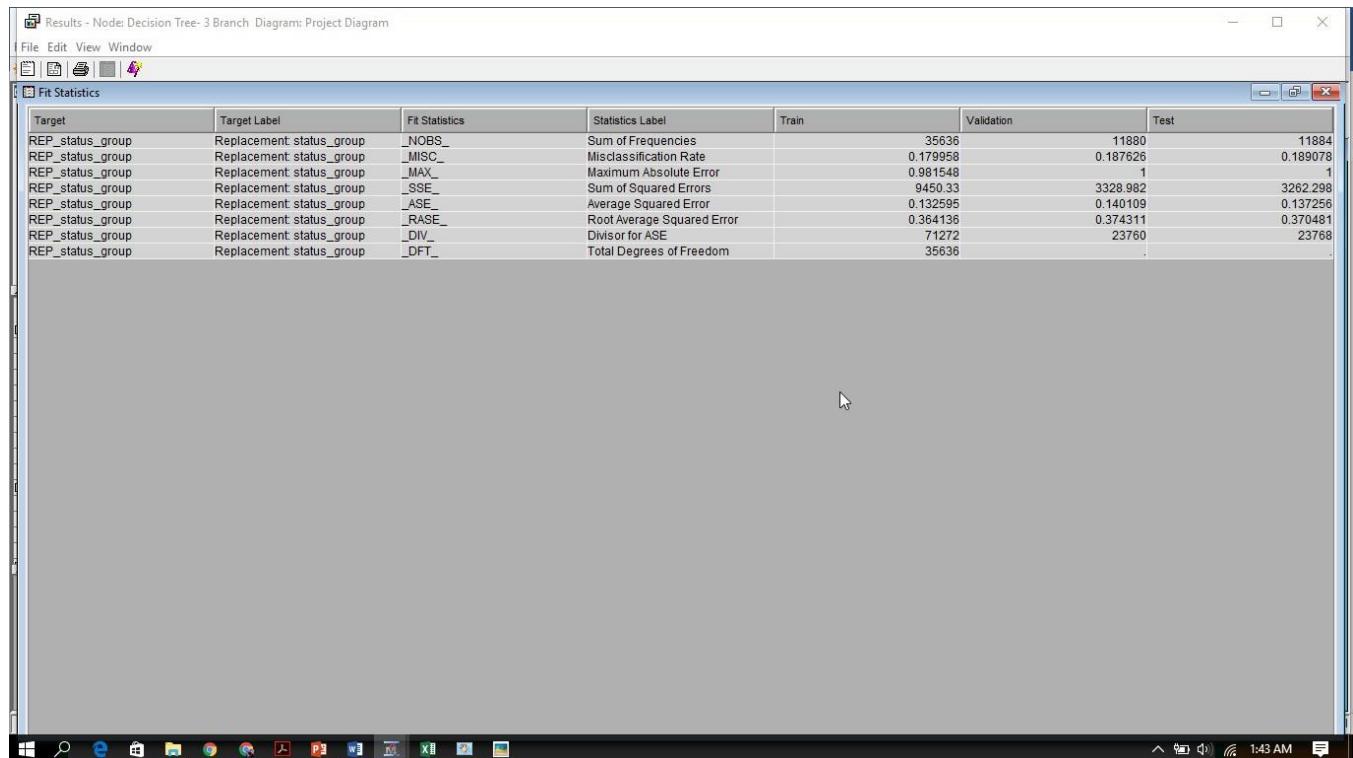
```

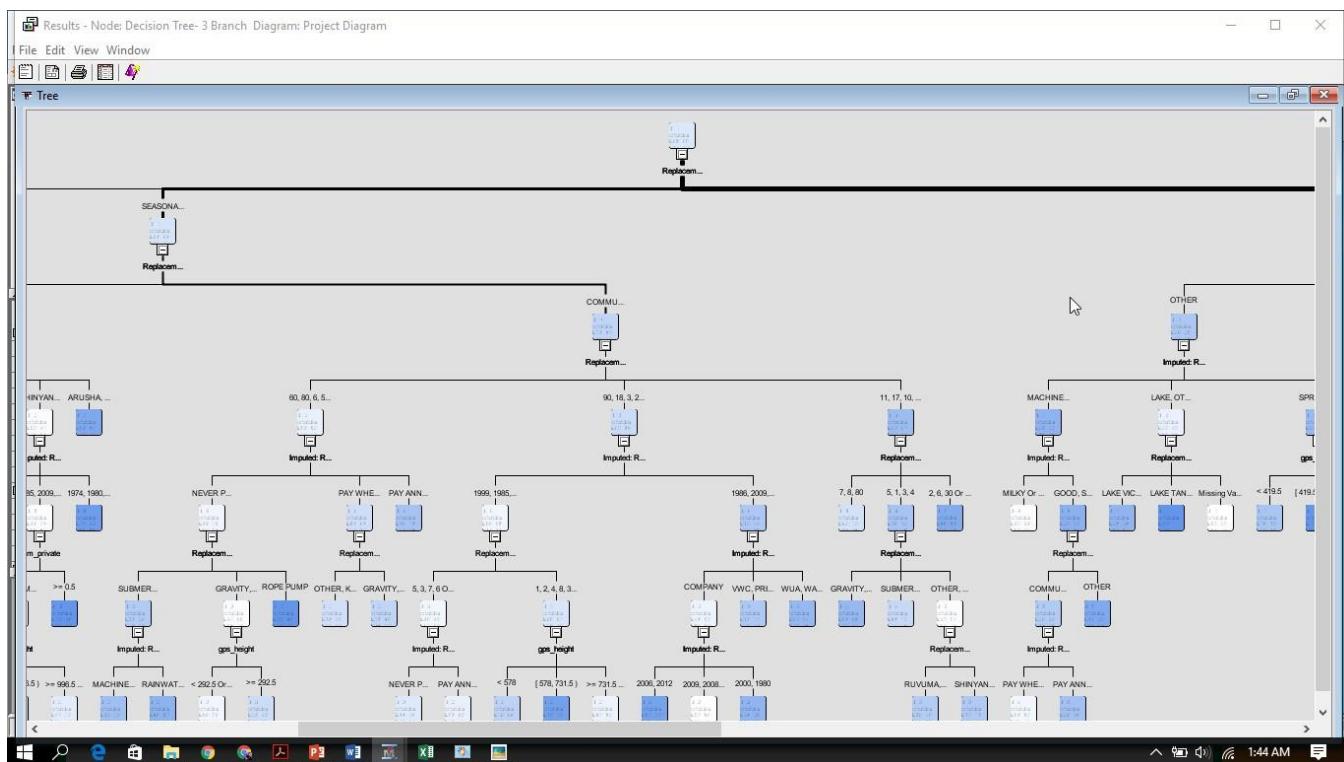
The misclassification rate for the above model is 0.1911.

## 10. Decision Tree- 3 Branch:

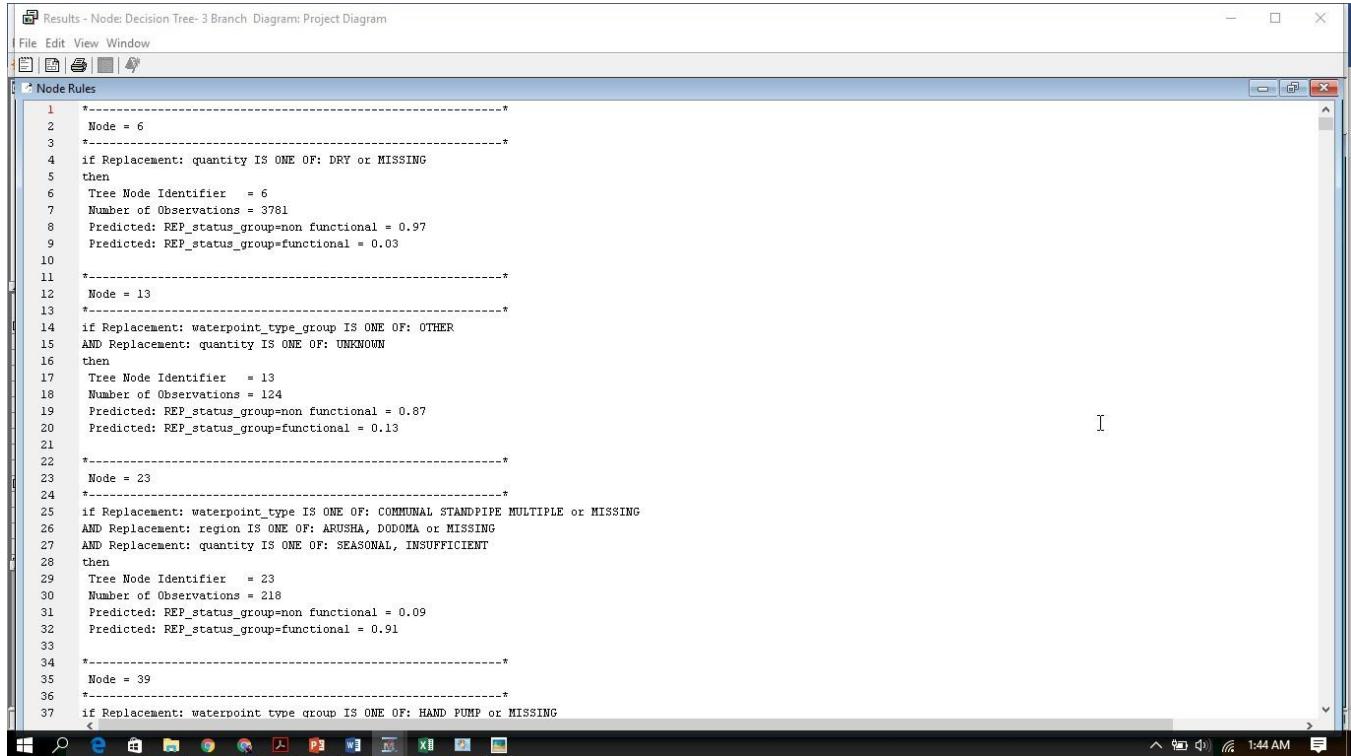
This tree was used to predict the ***status\_group*** target variable from the imputed data with the following properties:

.. Property	Value
<b>Splitting Rule</b>	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
<b>Node</b>	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.





Below we can see the Node Rules window containing the IF-THEN logic that distributes observations into each leaf node of the decision tree.



The screenshot shows a Windows application window titled "Results - Node: Decision Tree- 3 Branch Diagram: Project Diagram". The main area is titled "Node Rules" and displays the following code:

```

1  *-----*
2  Node = 6
3  *-----*
4  if Replacement: quantity IS ONE OF: DRY or MISSING
5  then
6  Tree Node Identifier = 6
7  Number of Observations = 3781
8  Predicted: REP_status_group=non functional = 0.97
9  Predicted: REP_status_group=functional = 0.03
10
11 *-----*
12 Node = 13
13 *-----*
14 if Replacement: waterpoint_type_group IS ONE OF: OTHER
15 AND Replacement: quantity IS ONE OF: UNKNOWN
16 then
17 Tree Node Identifier = 13
18 Number of Observations = 124
19 Predicted: REP_status_group=non functional = 0.87
20 Predicted: REP_status_group=functional = 0.13
21
22 *-----*
23 Node = 23
24 *-----*
25 if Replacement: waterpoint_type IS ONE OF: COMMUNAL STANDPIPE MULTIPLE or MISSING
26 AND Replacement: region IS ONE OF: ARUSHA, DODOMA or MISSING
27 AND Replacement: quantity IS ONE OF: SEASONAL, INSUFFICIENT
28 then
29 Tree Node Identifier = 23
30 Number of Observations = 218
31 Predicted: REP_status_group=non functional = 0.09
32 Predicted: REP_status_group=functional = 0.91
33
34 *-----*
35 Node = 39
36 *-----*
37 if Replacement: waterpoint type group IS ONE OF: HAND PUMP or MISSING
<

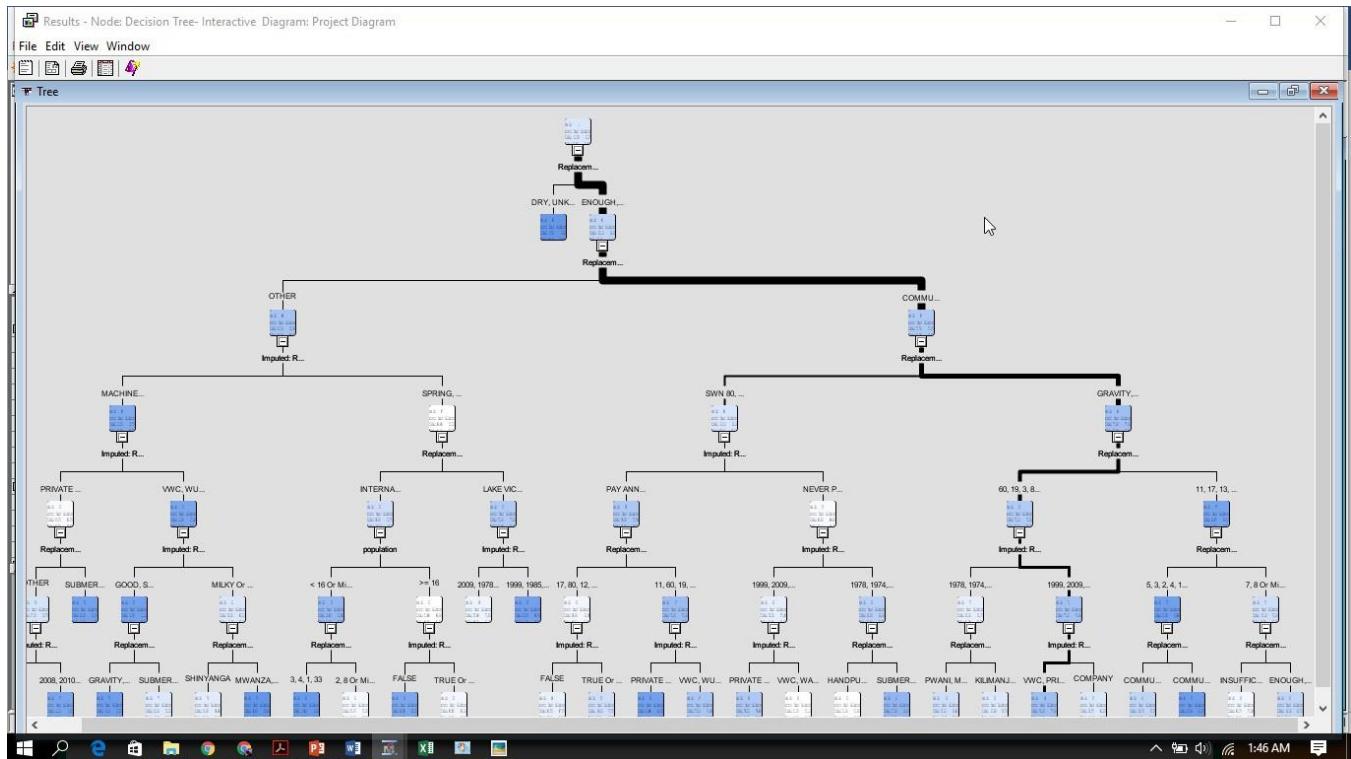
```

The window has a standard Windows title bar and menu bar. The status bar at the bottom shows the time as 1:44 AM.

The misclassification rate for the above model is 0.1875.

## 11. Decision Tree- Interactive:

We create an interactive decision tree by deciding the split node based on the log worth of each variable and we choose the node with the highest log value. We then train a node to create the entire decision tree and we get the interactive decision tree. The interactive decision tree gives us an advantage of choosing the split node and train the learning process so which would help us in producing better prediction models.



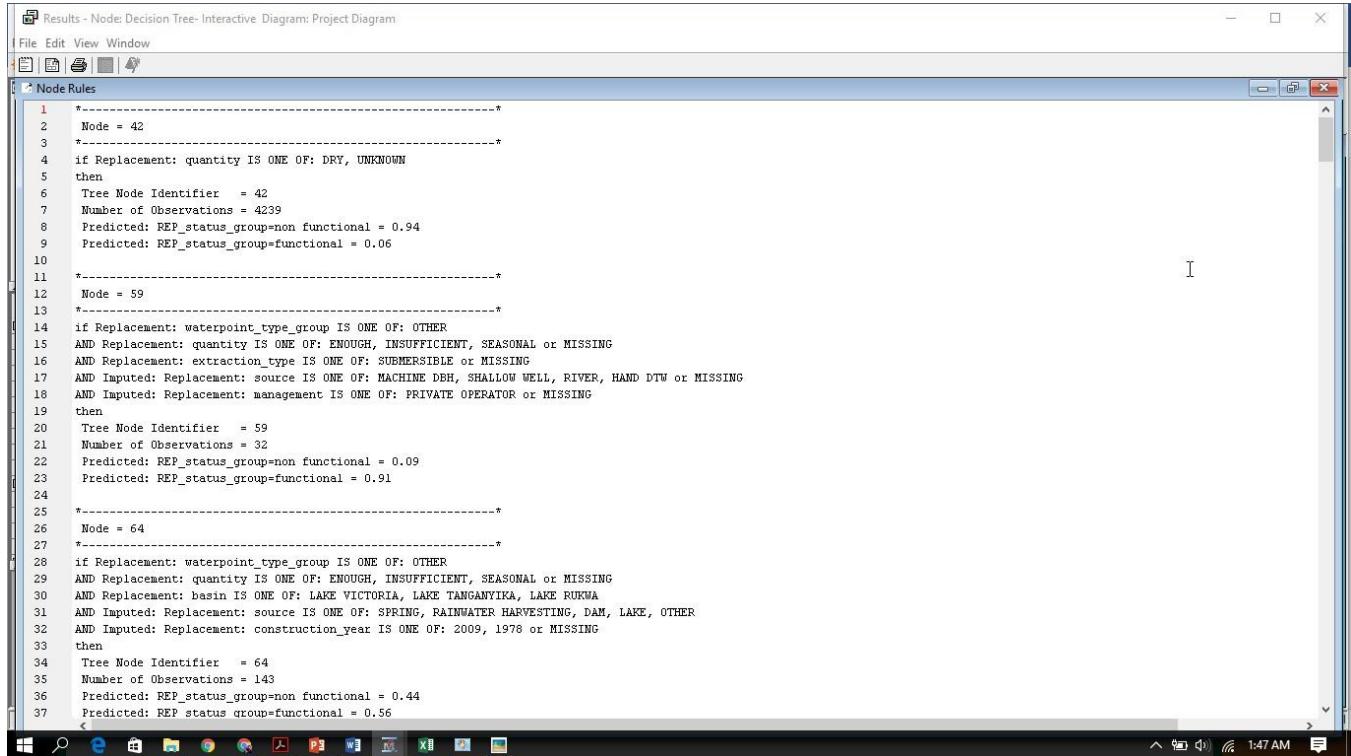
Results - Node: Decision Tree- Interactive Diagram: Project Diagram

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement status_group	_MISC_	Misclassification Rate	0.206813	0.21229	0.211292
REP_status_group	Replacement status_group	_MAX_	Maximum Absolute Error	0.95082	0.95082	0.95082
REP_status_group	Replacement status_group	_SSE_	Sum of Squared Errors	10549.21	3609.127	3557.845
REP_status_group	Replacement status_group	_ASE_	Average Squared Error	0.148013	0.151899	0.149691
REP_status_group	Replacement status_group	_RASE_	Root Average Squared Error	0.384725	0.389743	0.386899
REP_status_group	Replacement status_group	_DIV_	Divisor for ASE	71272	23760	23768
REP_status_group	Replacement status_group	_DFT_	Total Degrees of Freedom	35636	-	-

Below we can see the Node Rules window containing the IF-THEN logic that distributes observations into each leaf node of the decision tree.



The screenshot shows a Windows application window titled "Results - Node: Decision Tree- Interactive Diagram: Project Diagram". The main area is titled "Node Rules" and displays the following text:

```

1 -----#
2 Node = 42
3 -----
4 if Replacement: quantity IS ONE OF: DRY, UNKNOWN
5 then
6 Tree Node Identifier = 42
7 Number of Observations = 4239
8 Predicted: REP_status_group=non functional = 0.94
9 Predicted: REP_status_group=functional = 0.06
10
11 -----
12 Node = 59
13 -----
14 if Replacement: waterpoint_type_group IS ONE OF: OTHER
15 AND Replacement: quantity IS ONE OF: ENOUGH, INSUFFICIENT, SEASONAL or MISSING
16 AND Replacement: extraction_type IS ONE OF: SUBMERSIBLE or MISSING
17 AND Imputed: Replacement: source IS ONE OF: MACHINE DBH, SHALLOW WELL, RIVER, HAND DTW or MISSING
18 AND Imputed: Replacement: management IS ONE OF: PRIVATE OPERATOR or MISSING
19 then
20 Tree Node Identifier = 59
21 Number of Observations = 32
22 Predicted: REP_status_group=non functional = 0.09
23 Predicted: REP_status_group=functional = 0.91
24
25 -----
26 Node = 64
27 -----
28 if Replacement: waterpoint_type_group IS ONE OF: OTHER
29 AND Replacement: quantity IS ONE OF: ENOUGH, INSUFFICIENT, SEASONAL or MISSING
30 AND Replacement: basin IS ONE OF: LAKE VICTORIA, LAKE TANGANYIKA, LAKE RUKWA
31 AND Imputed: Replacement: source IS ONE OF: SPRING, RAINWATER HARVESTING, DAM, LAKE, OTHER
32 AND Imputed: Replacement: construction_year IS ONE OF: 2009, 1978 or MISSING
33 then
34 Tree Node Identifier = 64
35 Number of Observations = 143
36 Predicted: REP_status_group=non functional = 0.44
37 Predicted: REP_status_group=functional = 0.56
<

```

**The misclassification rate for the above model is 0.2123.**

## 12. Gradient Boosting:

The Gradient Boosting node uses partitioning algorithm to search for an optimal partition of the data for a single target variable. Gradient Boosting resamples the analysis data several times to generate results that form a weighted average of the resampled data set.

Tree boosting creates a series of decision trees that form a single predictive model. Boosting is less prone to overfitting the data than a single tree. If a decision tree fits the data fairly well, then boosting often improves the fit.

The following properties was used for this decision tree

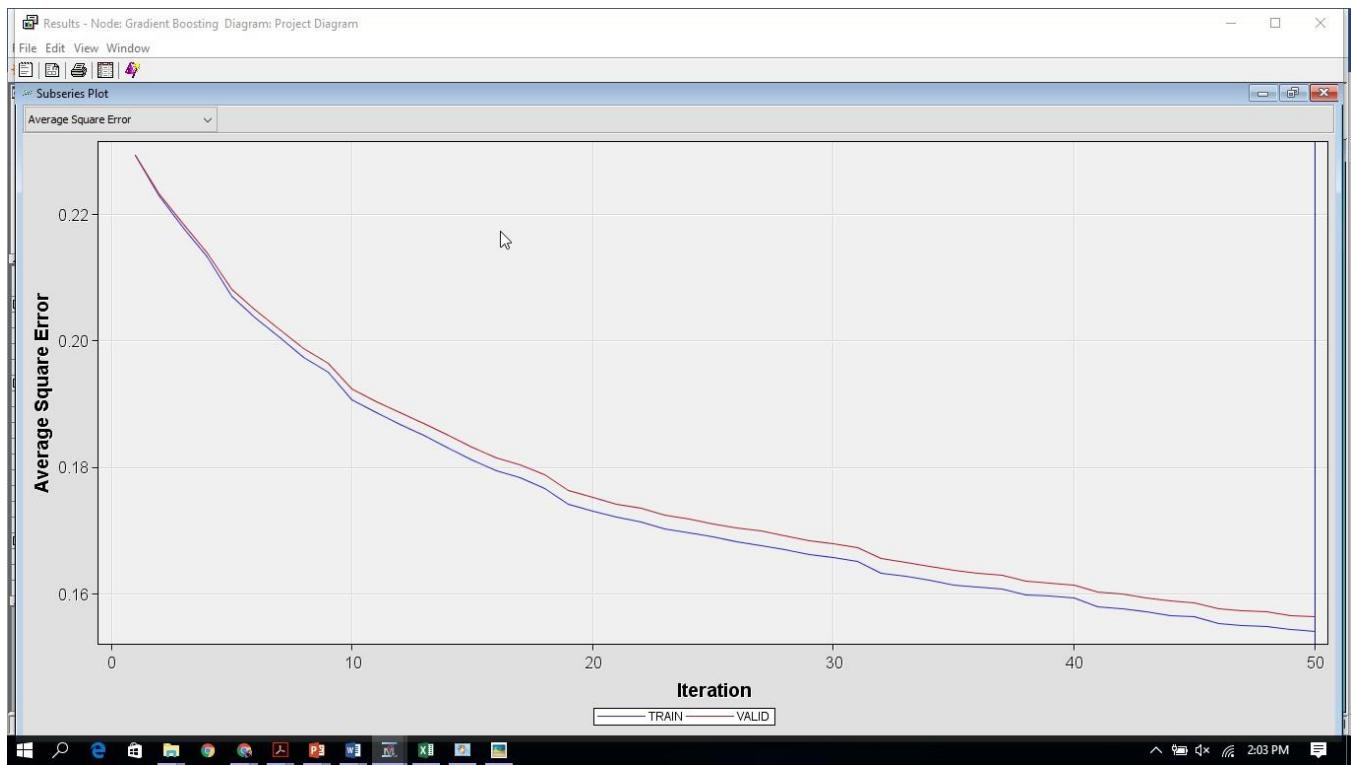
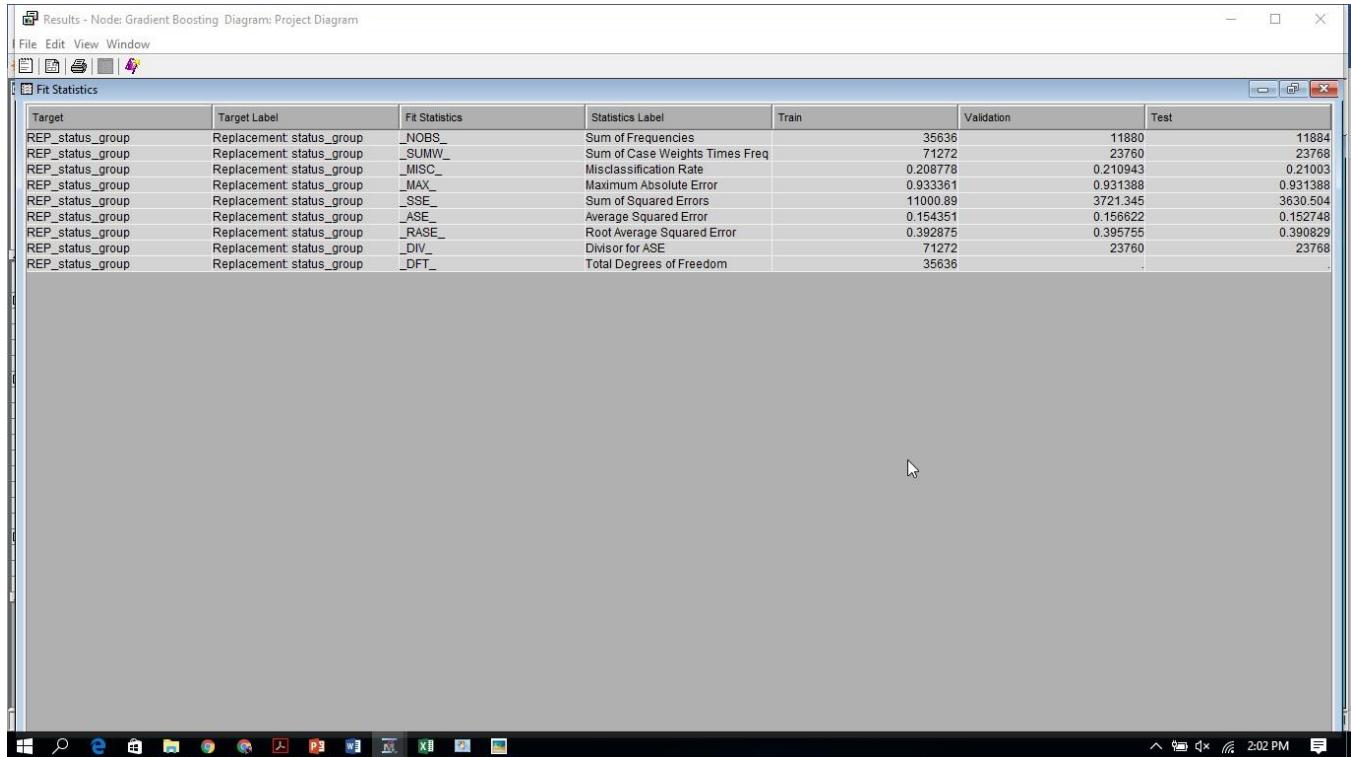
... Property	Value
<b>Series Options</b>	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
<b>Splitting Rule</b>	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
<b>Node</b>	
Leaf Fraction	0.1
Number of Surrogate Rules	2
Split Size	,

```

Results - Node: Gradient Boosting Diagram: Project Diagram
File Edit View Window
File|Edit|View|Window|Output|Print|Exit|X

* Score Output
*-----
52
53
54
55
56
57 Variable Importance
58
59 Obs NAME LABEL NRULES NSURROGATES IMPORTANCE VIMPORTANCE RATIO
60
61 1 REP_quantity Replacement: quantity 35 0 1.00000 1.00000 1.00000
62 2 REP_quantity_group Replacement: quantity_group 0 35 1.00000 1.00000 1.00000
63 3 REP_region_code Replacement: region_code 32 83 0.99988 0.99083 0.99094
64 4 REP_extraction_type_class Replacement: extraction_type_class 0 26 0.92481 0.85503 0.92455
65 5 REP_extraction_type Replacement: extraction_type 19 10 0.92379 0.85903 0.92990
66 6 REP_extraction_type_group Replacement: extraction_type_group 0 19 0.78200 0.73142 0.93532
67 7 REP_waterpoint_type Replacement: waterpoint_type 10 0 0.72780 0.65101 0.89450
68 8 IMP REP_source Imputed: Replacement: source 4 26 0.69449 0.65222 0.93914
69 9 REP_waterpoint_type_group Replacement: waterpoint_type_group 0 10 0.68952 0.61679 0.89451
70 10 REP_region Replacement: region 12 60 0.68609 0.65022 0.94773
71 11 IMP REP_construction_year Imputed: Replacement: construction_year 29 3 0.52026 0.51480 0.98950
72 12 REP_basin Replacement: basin 1 41 0.47774 0.43539 0.91136
73 13 IMP REP_payment Imputed: Replacement: payment 14 4 0.47177 0.47247 1.00147
74 14 amount_tsh 4 14 0.45396 0.45325 0.99844
75 15 IMP REP_scheme_management Imputed: Replacement: scheme_management 1 8 0.25447 0.24759 0.97297
76 16 IMP REP_management Imputed: Replacement: management 6 2 0.23653 0.22412 0.94753
77 17 REP_source_type Replacement: source_type 2 4 0.20003 0.19993 0.99950
78 18 population 3 4 0.17282 0.16403 0.94916
79 19 REP_district_code Replacement: district_code 4 1 0.14893 0.12583 0.84487
80 20 gps_height 0 2 0.13811 0.11965 0.86636
81
82
83 *-----
84 * Report Output
85 *-----
86
87
88
89

```

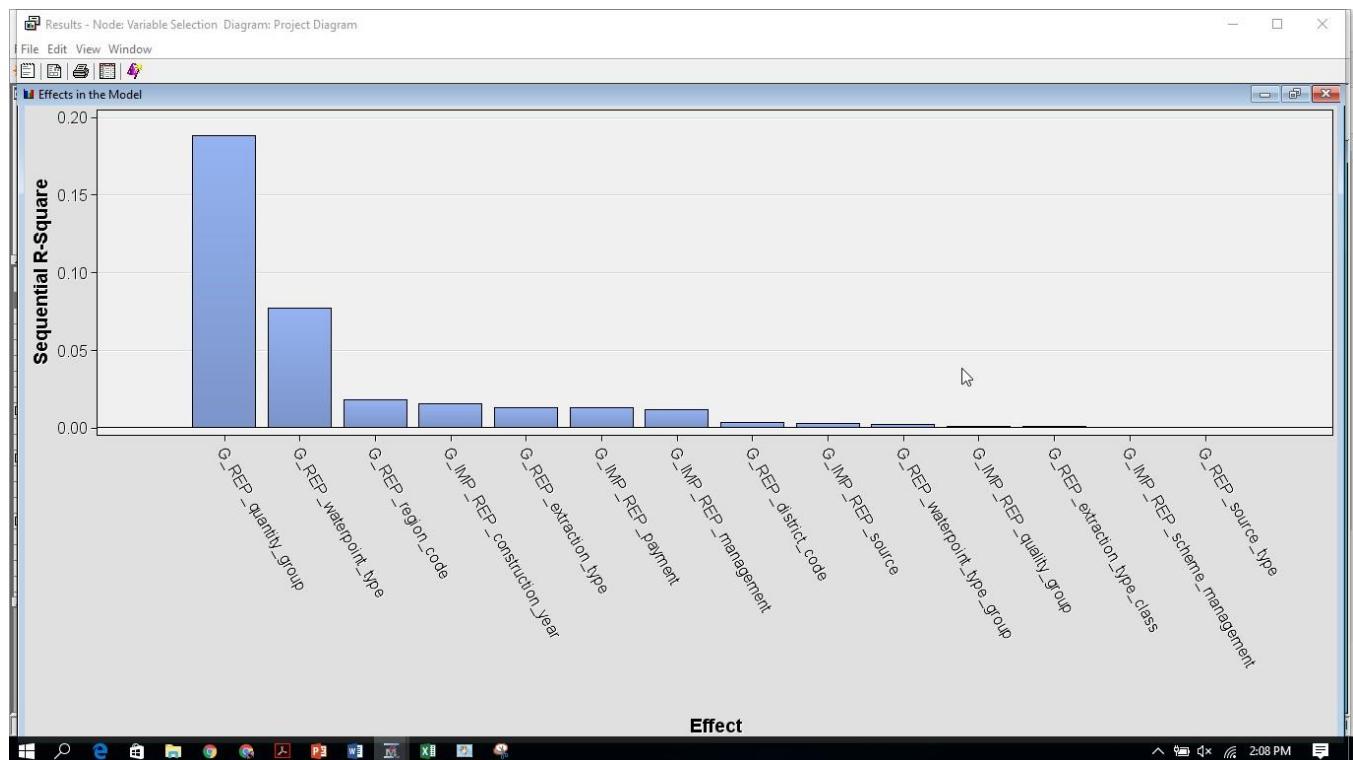


The misclassification rate for the above model is 0.2109.

### 13. Variable Selection:

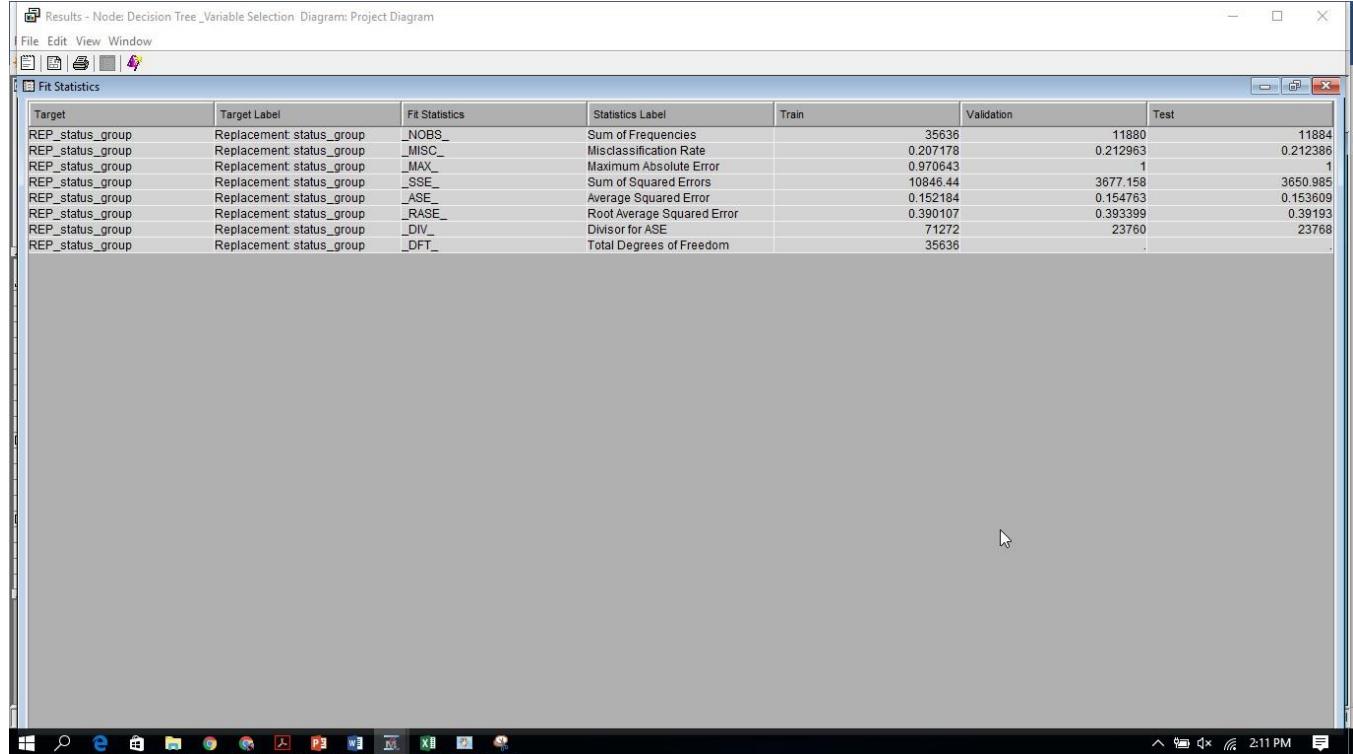
This node was used to filter variables based on their 'worth' and the maximum class levels was set to 200.

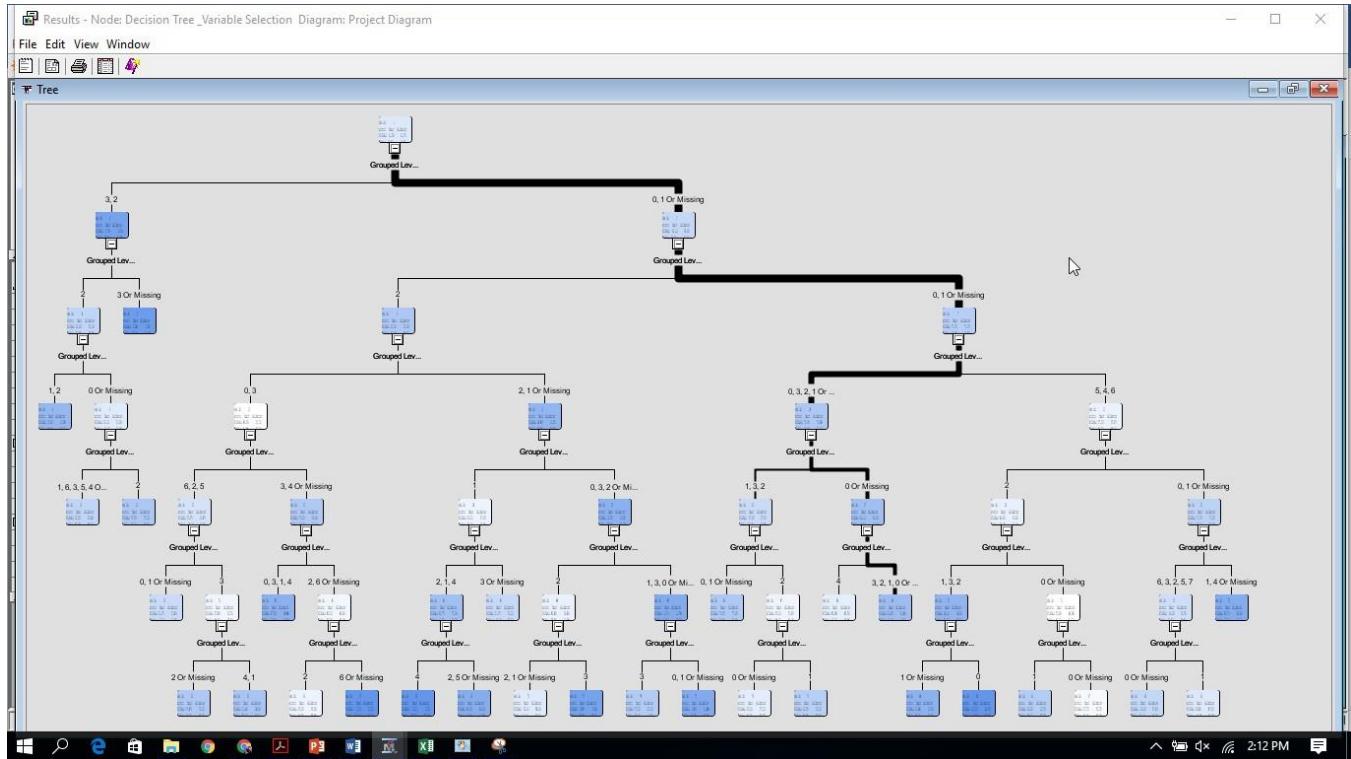
.. Property	Value
<b>Train</b>	
Variables	[...]
Max Class Level	200
Max Missing Percentage	50
Target Model	Default
Manual Selector	[...]
Rejects Unused Input	Yes
<input type="checkbox"/> Bypass Options	
Variable	None
Role	Input
<input type="checkbox"/> Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
<input type="checkbox"/> R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.005
Stop R-Square	5.0E-4
Use AOV16 Variables	No



## 14. Decision Tree \_ Variable Selection:

This node was used to predict the ***status\_group*** target variable from the selected variables.





Below we can see the Node Rules window containing the IF-THEN logic that distributes observations into each leaf node of the decision tree.

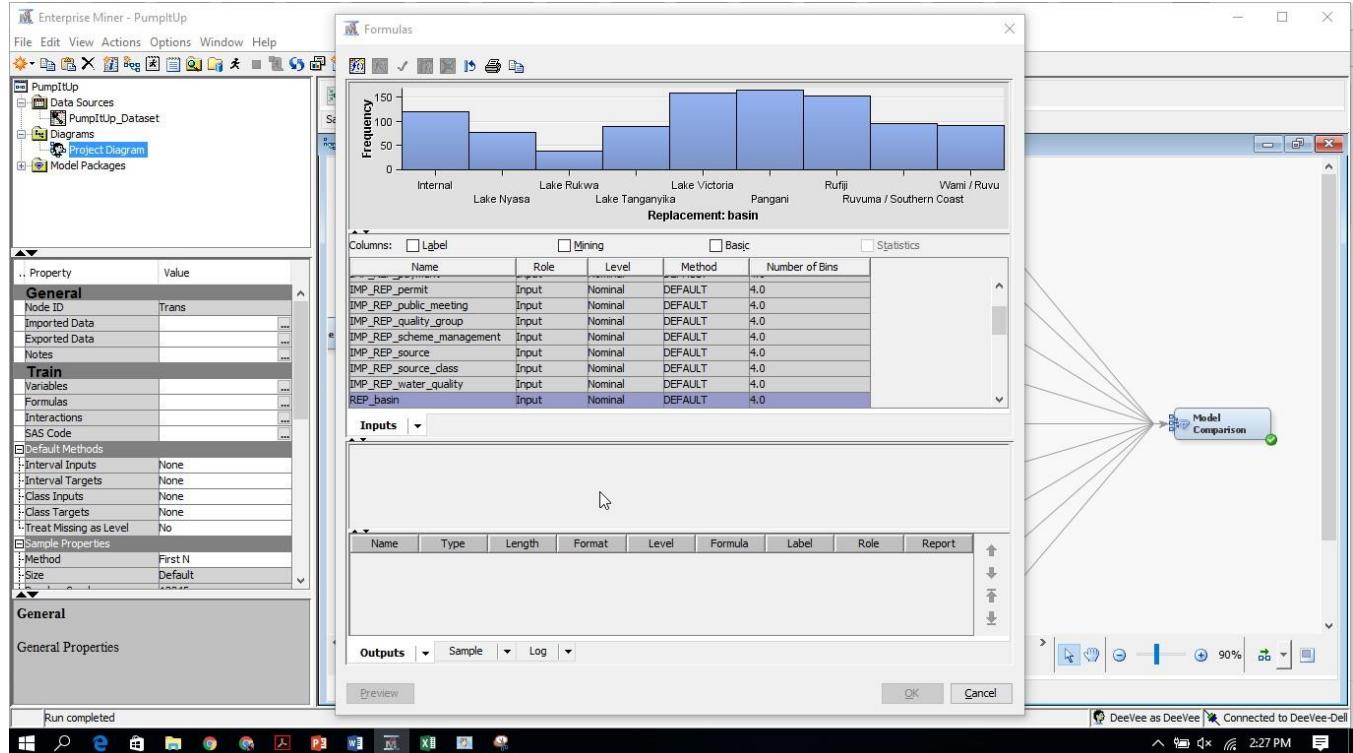
```
Results - Node: Decision Tree _Variable Selection Diagram: Project Diagram
File Edit View Window
Tree
Node Rules
1 -----#
2 Node = 5
3 -----
4 if Grouped Levels for REP_quantity_group IS ONE OF: 3 or MISSING
5 then
6 Tree Node Identifier = 5
7 Number of Observations = 3781
8 Predicted: REP_status_group=non functional = 0.97
9 Predicted: REP_status_group=functional = 0.03
10 -----
11 -----
12 Node = 8
13 -----
14 if Grouped Levels for REP_waterpoint_type IS ONE OF: 1, 2
15 AND Grouped Levels for REP_quantity_group IS ONE OF: 2
16 then
17 Tree Node Identifier = 8
18 Number of Observations = 192
19 Predicted: REP_status_group=non functional = 0.84
20 Predicted: REP_status_group=functional = 0.16
21 -----
22 -----
23 Node = 18
24 -----
25 if Grouped Levels for REP_waterpoint_type IS ONE OF: 0 or MISSING
26 AND Grouped Levels for REP_region_code IS ONE OF: 1, 6, 3, 5, 4 or MISSING
27 AND Grouped Levels for REP_quantity_group IS ONE OF: 2
28 then
29 Tree Node Identifier = 18
30 Number of Observations = 210
31 Predicted: REP_status_group=non functional = 0.71
32 Predicted: REP_status_group=functional = 0.29
33 -----
34 -----
35 Node = 19
36 -----
37 if Grouped Levels for REP_waterpoint_type IS ONE OF: 0 or MISSING
38 AND Grouped Levels for REP_region_code IS ONE OF: 2
```

The misclassification rate for the above model is 0.2129.

## 15. Transform Variables:

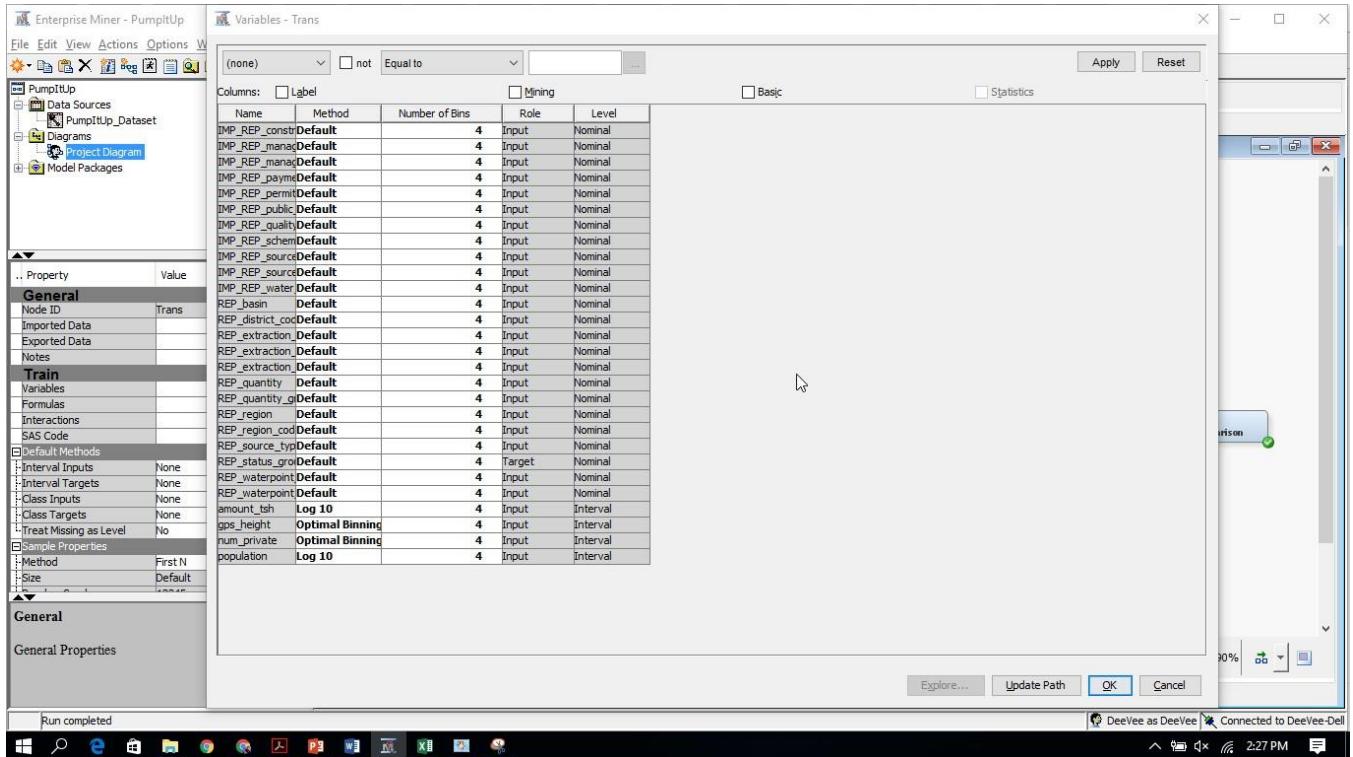
Transforming the data can improve model response. Transforming the data tends to stabilize variance, remove nonlinearity, improve additivity, and counter non-normality. For many models, transforming input data leads to better model fits. These transformations can be a function of one or more variables.

Below is the Formula window where we can browse the input variable distributions before specifying the type of transformation we want to perform.



The common log transformation is often used to control skewness.

We have chosen log10 method for ***amount\_tsh and population*** and Optimal Binning method for ***gps\_height and num\_private***.



## 16. Stepwise – Logistic Regression:

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion.

Property	Value
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	None
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	

Results - Node: Stepwise- Logistic Regression Diagram: Project Diagram

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement status_group	_AIC_	Akaike's Information Criterion	31352.37	-	-
REP_status_group	Replacement status_group	_ASE_	Average Squared Error	0.139127	0.143072	0.139876
REP_status_group	Replacement status_group	_AVERR_	Average Error Function	0.434818	0.44909	0.437791
REP_status_group	Replacement status_group	_DFE_	Degrees of Freedom for Error	35455	-	-
REP_status_group	Replacement status_group	_DFM_	Model Degrees of Freedom	181	-	-
REP_status_group	Replacement status_group	_DFT_	Total Degrees of Freedom	35636	-	-
REP_status_group	Replacement status_group	_DIV_	Divisor for ASE	71272	23760	23768
REP_status_group	Replacement status_group	_ERR_	Error Function	30990.37	10670.38	10405.41
REP_status_group	Replacement status_group	_FPE_	Final Prediction Error	0.140548	-	-
REP_status_group	Replacement status_group	_MAX_	Maximum Absolute Error	0.999806	0.999592	0.99933
REP_status_group	Replacement status_group	_MSE_	Mean Square Error	0.139838	0.143072	0.139876
REP_status_group	Replacement status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement status_group	_NW_	Number of Estimate Weights	181	-	-
REP_status_group	Replacement status_group	_RASE_	Root Average Sum of Squares	0.372998	0.378248	0.374
REP_status_group	Replacement status_group	_RFPE_	Root Final Prediction Error	0.374897	-	-
REP_status_group	Replacement status_group	_RMSE_	Root Mean Squared Error	0.373949	0.378248	0.374
REP_status_group	Replacement status_group	_SBC_	Schwarz's Bayesian Criterion	32867.45	-	-
REP_status_group	Replacement status_group	_SSE_	Sum of Squared Errors	9915.889	3399.382	3324.579
REP_status_group	Replacement status_group	_SUMW_	Sum of Case Weights Times Freq	71272	23760	23768
REP_status_group	Replacement status_group	_MISC_	Misclassification Rate	0.192642	0.19638	0.195894

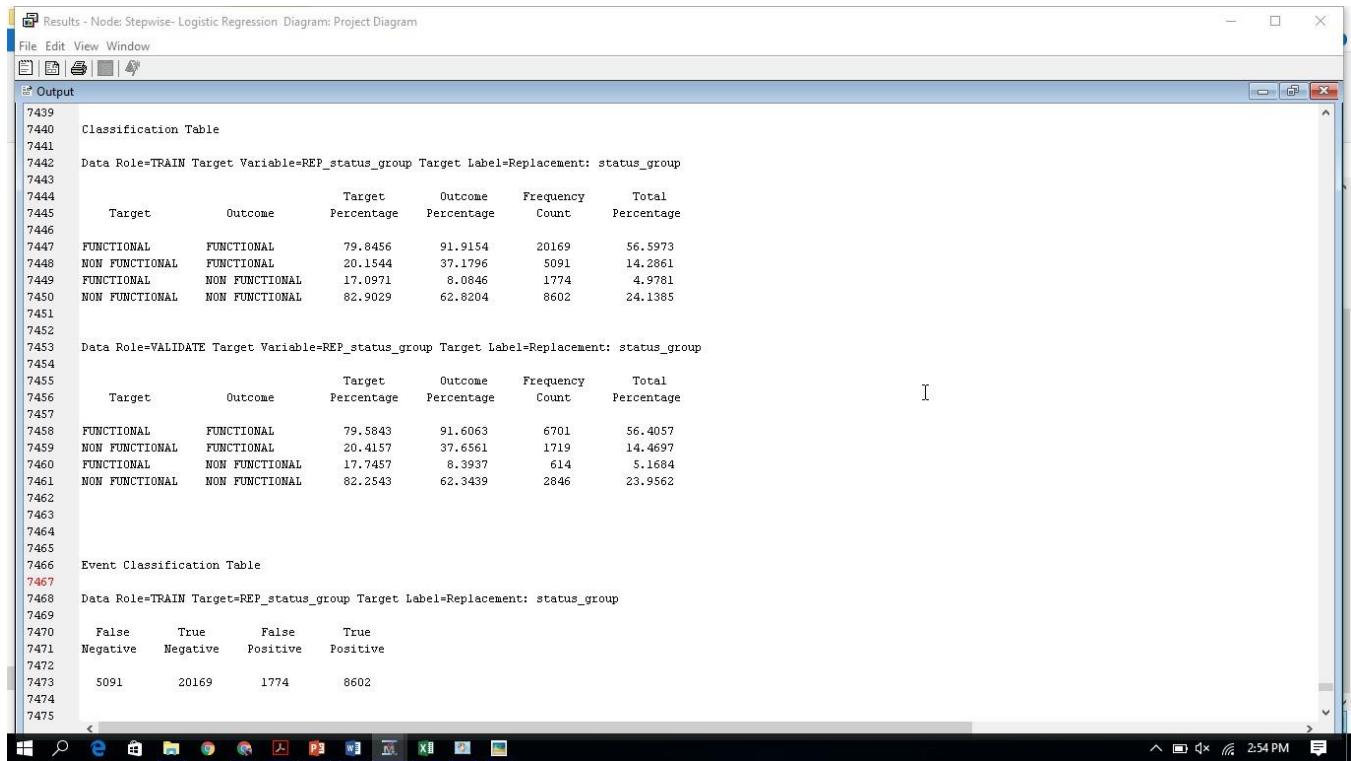
```

Results - Node: Stepwise- Logistic Regression Diagram: Project Diagram
File Edit View Window
Output
268
269      Type 3 Analysis of Effects
270
271      Effect          DF   Chi-Square   Pr > ChiSq
272      REP_quantity    4    2335.2880  <.0001
273      REP_waterpoint_type 6    2937.4310  <.0001
274
275
276
277      Analysis of Maximum Likelihood Estimates
278
279
280      Parameter
281      REP_status_group  DF   Estimate   Standard Error   Wald Chi-Square   Pr > ChiSq   Standardized Estimate   Exp(Est)
282
283      Intercept        non functional 1    0.3336    0.1662    4.03    0.0448    1.396
284      REP_quantity     dry           non functional 1    3.1077    0.0813   1460.96  <.0001   22.371
285      REP_quantity     enough        non functional 1    -1.3562   0.0338   1605.79  <.0001   0.258
286      REP_quantity     insufficient  non functional 1    -0.9170   0.0360   649.78   <.0001   0.400
287      REP_quantity     seasonal     non functional 1    -1.1130   0.0476   546.63   <.0001   0.329
288      REP_waterpoint_type  cattle trough  non functional 1    -0.4533   0.2913    2.42    0.1196   0.636
289      REP_waterpoint_type  communal standpipe  non functional 1    -0.3563   0.1644    4.70    0.0302   0.700
290      REP_waterpoint_type  communal standpipe multiple  non functional 1    0.6594   0.1665   15.69   <.0001   1.934
291      REP_waterpoint_type  dam          non functional 1    -0.7476   0.9431    0.63    0.4279   0.473
292      REP_waterpoint_type  hand pump    non functional 1    -0.1340   0.1648    0.66    0.4161   0.875
293      REP_waterpoint_type  improved spring  non functional 1    -1.0096   0.2006   25.34   <.0001   0.364
294
295
296      Odds Ratio Estimates
297
298
299      Effect
300
301      REP_quantity     dry vs unknown  non functional 16.932
302      REP_quantity     enough vs unknown  non functional 0.195
303      REP_quantity     insufficient vs unknown  non functional 0.303
304      REP quantity    seasonal vs unknown  non functional 0.249

```

From the above result window, we can say that the parameters with chi-squared value is less than 0.05 are highly significant. For example, if the quantity of water is Dry, enough or Insufficient then the status group is likely to be non-functional and similarly so on.





```

Results - Node: Stepwise- Logistic Regression Diagram: Project Diagram
File Edit View Window
Output
7439
7440 Classification Table
7441
7442 Data Role=TRAIN Target Variable=REP_status_group Target Label=Replacement: status_group
7443
7444 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
7445 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
7446
7447 FUNCTIONAL FUNCTIONAL 79.8456 91.9154 20169 56.5973
7448 NON FUNCTIONAL FUNCTIONAL 20.1544 37.1796 5091 14.2861
7449 FUNCTIONAL NON FUNCTIONAL 17.0971 8.0846 1774 4.9781
7450 NON FUNCTIONAL NON FUNCTIONAL 82.9029 62.8204 8602 24.1385
7451
7452
7453 Data Role=VALIDATE Target Variable=REP_status_group Target Label=Replacement: status_group
7454
7455 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
7456 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
7457
7458 FUNCTIONAL FUNCTIONAL 79.5843 91.6063 6701 56.4057
7459 NON FUNCTIONAL FUNCTIONAL 20.4157 37.6561 1719 14.4697
7460 FUNCTIONAL NON FUNCTIONAL 17.7457 8.3937 614 5.1684
7461 NON FUNCTIONAL NON FUNCTIONAL 82.2543 62.3439 2846 23.9562
7462
7463
7464
7465
7466 Event Classification Table
7467
7468 Data Role=TRAIN Target=REP_status_group Target Label=Replacement: status_group
7469
7470 False True False True
7471 Negative Negative Positive Positive
7472
7473 5091 20169 1774 8602
7474
7475

```

The misclassification rate for the above model is 0.1963.

## 17. Forward- Logistic Regression:

Property	Value
<b>Class Targets</b>	
Regression Type	Logistic Regression
Link Function	Logit
<b>Model Options</b>	
Suppress Intercept	No
Input Coding	Deviation
<b>Model Selection</b>	
Selection Model	Forward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
<b>Optimization Options</b>	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
<b>Convergence Criteria</b>	
Uses Defaults	Yes

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement status_group	_AIC_	Akaike's Information Criterion	31349.71		
REP_status_group	Replacement status_group	_ASE_	Average Squared Error	0.139104	0.143003	0.139669
REP_status_group	Replacement status_group	_AVERR_	Average Error Function	0.434753	0.448928	0.437775
REP_status_group	Replacement status_group	_DFE_	Degrees of Freedom for Error	35454		
REP_status_group	Replacement status_group	_DFM_	Model Degrees of Freedom	182		
REP_status_group	Replacement status_group	_DFT_	Total Degrees of Freedom	35636		
REP_status_group	Replacement status_group	_DIV_	Divisor for ASE	71272	23760	23768
REP_status_group	Replacement status_group	_ERR_	Error Function	30985.71	10666.53	10405.04
REP_status_group	Replacement status_group	_FPE_	Final Prediction Error	0.140532		
REP_status_group	Replacement status_group	_MAX_	Maximum Absolute Error	0.998908	0.999592	0.999329
REP_status_group	Replacement status_group	_MSE_	Mean Square Error	0.139818	0.143003	0.139669
REP_status_group	Replacement status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement status_group	_NW_	Number of Estimate Weights	182		
REP_status_group	Replacement status_group	_RASE_	Root Average Sum of Squares	0.372967	0.378157	0.373991
REP_status_group	Replacement status_group	_RFPE_	Root Final Prediction Error	0.374876		
REP_status_group	Replacement status_group	_RMSE_	Root Mean Squared Error	0.373923	0.378157	0.373991
REP_status_group	Replacement status_group	_SBC_	Schwarz's Bayesian Criterion	32893.27		
REP_status_group	Replacement status_group	_SSE_	Sum of Squared Errors	9914.235	3397.747	3324.408
REP_status_group	Replacement status_group	_SUMW_	Sum of Case Weights Times Freq	71272	23760	23768
REP_status_group	Replacement status_group	_MISC_	Misclassification Rate	0.19267	0.19638	0.195978

Results - Node: Forward- Logistic Regression Diagram: Project Diagram

Type 3 Analysis of Effects

Effect	DF	Chi-Square	Pr > ChiSq
REP_quantity	4	2512.2093	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	REP_status_group	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	non functional	1	0.4437	0.0299	219.85	<.0001	1.559	
REP_quantity	dry	non functional	1	3.0546	0.0804	1443.45	<.0001	21.213
REP_quantity	enough	non functional	1	-1.4151	0.0324	1912.26	<.0001	0.243
REP_quantity	insufficient	non functional	1	-0.9449	0.0343	758.64	<.0001	0.389
REP_quantity	seasonal	non functional	1	-1.1336	0.0447	643.31	<.0001	0.322

Odds Ratio Estimates

Effect	REP_status_group	Point Estimate	
REP_quantity	dry vs unknown	non functional	13.674
REP_quantity	enough vs unknown	non functional	0.157
REP_quantity	insufficient vs unknown	non functional	0.251
REP_quantity	seasonal vs unknown	non functional	0.207

Step 2: Effect REP\_waterpoint\_type entered.

```

Results - Node: Forward- Logistic Regression Diagram: Project Diagram
File Edit View Window
Output
6992
6993
6994 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
6995
6996 FUNCTIONAL FUNCTIONAL 79.8448 91.9109 20168 56.5945
6997 NON FUNCTIONAL FUNCTIONAL 20.1552 37.1796 5091 14.2861
6998 FUNCTIONAL NON FUNCTIONAL 17.1051 8.0891 1775 4.9809
6999 NON FUNCTIONAL NON FUNCTIONAL 82.8949 62.8204 8602 24.1385
7000
7001
7002 Data Role=VALIDATE Target Variable=REP_status_group Target Label=Replacement: status_group
7003
7004 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
7005
7006
7007 FUNCTIONAL FUNCTIONAL 79.5914 91.5926 6700 56.3973
7008 NON FUNCTIONAL FUNCTIONAL 20.4086 37.6342 1718 14.4613
7009 FUNCTIONAL NON FUNCTIONAL 17.7643 8.4074 615 5.1768
7010 NON FUNCTIONAL NON FUNCTIONAL 82.2357 62.3658 2847 23.9646
7011
7012
7013
7014
7015 Event Classification Table
7016
7017 Data Role=TRAIN Target=REP_status_group Target Label=Replacement: status_group
7018
7019 False True False True
7020 Negative Negative Positive Positive
7021
7022 5091 20168 1775 8602
7023
7024
7025 Data Role=VALIDATE Target=REP_status_group Target Label=Replacement: status_group
7026
7027 False True False True
7028 Negative Negative Positive Positive
    
```



The misclassification rate for the above model is 0.1963.

## 18. Backward- Logistic Regression:

.. Property	Value
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes

Results - Node: Backward- Logistic Regression Diagram: Project Diagram

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement status_group	_AIC_	Akaike's Information Criterion	31349.17		
REP_status_group	Replacement status_group	_ASE_	Average Squared Error	0.139174	0.14311	0.139938
REP_status_group	Replacement status_group	_AVERR_	Average Error Function	0.434914	0.449236	0.43803
REP_status_group	Replacement status_group	_DFE_	Degrees of Freedom for Error	35460		
REP_status_group	Replacement status_group	_DFM_	Model Degrees of Freedom	176		
REP_status_group	Replacement status_group	_DFT_	Total Degrees of Freedom	35636		
REP_status_group	Replacement status_group	_DIV_	Divisor for ASE	71272	23760	23768
REP_status_group	Replacement status_group	_ERR_	Error Function	30997.17	10673.85	10411.09
REP_status_group	Replacement status_group	_FPE_	Final Prediction Error	0.140556		
REP_status_group	Replacement status_group	_MAX_	Maximum Absolute Error	0.998879	0.999585	0.999325
REP_status_group	Replacement status_group	_MSE_	Mean Square Error	0.139865	0.14311	0.139938
REP_status_group	Replacement status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement status_group	_NW_	Number of Estimate Weights	176		
REP_status_group	Replacement status_group	_RASE_	Root Average Sum of Squares	0.373061	0.378299	0.374083
REP_status_group	Replacement status_group	_RFPE_	Root Final Prediction Error	0.374908		
REP_status_group	Replacement status_group	_RMSE_	Root Mean Squared Error	0.373985	0.378299	0.374083
REP_status_group	Replacement status_group	_SBC_	Schwarz's Bayesian Criterion	32841.85		
REP_status_group	Replacement status_group	_SSE_	Sum of Squared Errors	9919.232	3400.288	3326.057
REP_status_group	Replacement status_group	_SUMW_	Sum of Case Weights Times Freq	71272	23760	23768
REP_status_group	Replacement status_group	_MISC_	Misclassification Rate	0.192979	0.197306	0.196399

Type 3 Analysis of Effects				
Effect	DF	Chi-Square	Pr > ChiSq	Wald
IMP REP_construction_year	53	643.0759	<.0001	
IMP REP_management	10	196.8464	<.0001	
IMP REP_management_group	1	0.1335	0.7148	
IMP REP_payment	5	307.9290	<.0001	
IMP REP_permit	1	94.3471	<.0001	
IMP REP_public_meeting	1	26.7015	<.0001	
IMP REP_quality_group	4	14.1543	0.0068	
IMP REP_scheme_management	11	52.2937	<.0001	
IMP REP_source	8	323.5862	<.0001	
IMP REP_source_class	1	1.4019	0.2364	
IMP REP_water_quality	2	3.5275	0.1714	
LGI0_amount_teh	1	9.4454	0.0021	
LGI0_population	1	83.6371	<.0001	
OPT_gps_height	3	33.1602	<.0001	
OPT_num_private	3	29.9699	<.0001	
REP_basin	8	151.1756	<.0001	
REP_district_code	19	321.2785	<.0001	
REP_extraction_type	17	544.9897	<.0001	
REP_extraction_type_class	0	0.0000	.	
REP_extraction_type_group	0	0.0000	.	
REP_quantity	4	2288.7490	<.0001	
REP_quantity_group	0	0.0000	.	
REP_region	20	266.0869	<.0001	
REP_region_code	10	15.6044	0.1115	
REP_source_type	0	0.0000	.	
REP_waterpoint_type	6	606.7510	<.0001	
REP_waterpoint_type_group	0	0.0000	.	
Analysis of Maximum Likelihood Estimates				
	REP status	Standard	Wald	Standardized

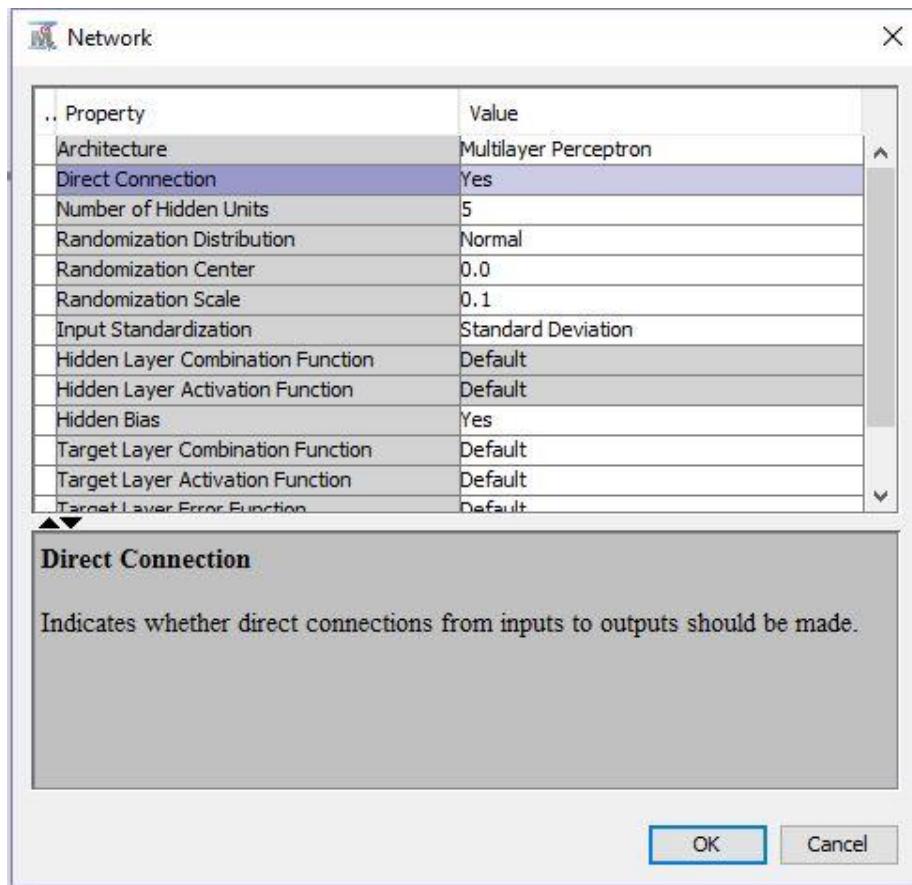


```

Results - Node: Backward- Logistic Regression Diagram: Project Diagram
File Edit View Window
Output
3126
3127
3128 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
3129
3130 FUNCTIONAL FUNCTIONAL 79.8408 91.8516 20155 56.5580
3131 NON FUNCTIONAL FUNCTIONAL 20.1592 37.1650 5089 14.2805
3132 FUNCTIONAL NON FUNCTIONAL 17.2055 8.1484 1788 5.0174
3133 NON FUNCTIONAL NON FUNCTIONAL 82.7945 62.8350 8604 24.1441
3134
3135
3136 Data Role=VALIDATE Target Variable=REP_Status_Group Target Label=Replacement: status_group
3137
3138 Target Outcome Target Percentage Outcome Percentage Frequency Count Total Percentage
3139
3140 FUNCTIONAL FUNCTIONAL 79.5295 91.5106 6694 56.3468
3141 NON FUNCTIONAL FUNCTIONAL 20.4705 37.7437 1723 14.5034
3142 FUNCTIONAL NON FUNCTIONAL 17.9324 8.4894 621 5.2273
3143 NON FUNCTIONAL NON FUNCTIONAL 82.0676 62.2563 2842 23.9226
3144
3145
3146
3147
3148
3149 Event Classification Table
3150
3151 Data Role=TRAIN Target=REP_Status_Group Target Label=Replacement: status_group
3152
3153 False True False True
3154 Negative Negative Positive Positive
3155
3156 5089 20155 1788 8604
3157
3158
3159 Data Role=VALIDATE Target=REP_Status_Group Target Label=Replacement: status_group
3160
3161 False True False True
3162 Negative Negative Positive Positive
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3450
3451
3452
3453
3454
3455
3456
3457
3458
3459
3460
3461
3462
3463
3464
3465
3466
3467
3468
3469
3470
3471
3472
3473
3474
3475
3476
3477
3478
3479
3480
3481
3482
3483
3484
3485
3486
3487
3488
3489
3490
3491
3492
3493
3494
3495
3496
3497
3498
3499
3500
3501
3502
3503
3504
3505
3506
3507
3508
3509
3510
3511
3512
3513
3514
3515
3516
3517
3518
3519
3520
3521
3522
3523
3524
3525
3526
3527
3528
3529
3530
3531
3532
3533
3534
3535
3536
3537
3538
3539
3540
3541
3542
3543
3544
3545
3546
3547
3548
3549
3550
3551
3552
3553
3554
3555
3556
3557
3558
3559
3560
3561
3562
3563
3564
3565
3566
3567
3568
3569
3570
3571
3572
3573
3574
3575
3576
3577
3578
3579
3580
3581
3582
3583
3584
3585
3586
3587
3588
3589
3590
3591
3592
3593
3594
3595
3596
3597
3598
3599
3600
3601
3602
3603
3604
3605
3606
3607
3608
3609
3610
3611
3612
3613
3614
3615
3616
3617
3618
3619
3620
3621
3622
3623
3624
3625
3626
3627
3628
3629
3630
3631
3632
3633
3634
3635
3636
3637
3638
3639
3640
3641
3642
3643
3644
3645
3646
3647
3648
3649
3650
3651
3652
3653
3654
3655
3656
3657
3658
3659
3660
3661
3662
3663
3664
3665
3666
3667
3668
3669
3670
3671
3672
3673
3674
3675
3676
3677
3678
3679
3680
3681
3682
3683
3684
3685
3686
3687
3688
3689
3690
3691
3692
3693
3694
3695
3696
3697
3698
3699
3700
3701
3702
3703
3704
3705
3706
3707
3708
3709
3710
3711
3712
3713
3714
3715
3716
3717
3718
3719
3720
3721
3722
3723
3724
3725
3726
3727
3728
3729
3730
3731
3732
3733
3734
3735
3736
3737
3738
3739
3740
3741
3742
3743
3744
3745
3746
3747
3748
3749
3750
3751
3752
3753
3754
3755
3756
3757
3758
3759
3760
3761
3762
3763
3764
3765
3766
3767
3768
3769
3770
3771
3772
3773
3774
3775
3776
3777
3778
3779
3780
3781
3782
3783
3784
3785
3786
3787
3788
3789
3790
3791
3792
3793
3794
3795
3796
3797
3798
3799
3800
3801
3802
3803
3804
3805
3806
3807
3808
3809
3810
3811
3812
3813
3814
3815
3816
3817
3818
3819
3820
3821
3822
3823
3824
3825
3826
3827
3828
3829
3830
3831
3832
3833
3834
3835
3836
3837
3838
3839
3840
3841
3842
3843
3844
3845
3846
3847
3848
3849
3850
3851
3852
3853
3854
3855
3856
3857
3858
3859
3860
3861
3862
3863
3864
3865
3866
3867
3868
3869
3870
3871
3872
3873
3874
3875
3876
3877
3878
3879
3880
3881
3882
3883
3884
3885
3886
3887
3888
3889
3890
3891
3892
3893
3894
3895
3896
3897
3898
3899
3900
3901
3902
3903
3904
3905
3906
3907
3908
3909
3910
3911
3912
3913
3914
3915
3916
3917
3918
3919
3920
3921
3922
3923
3924
3925
3926
3927
3928
3929
3930
3931
3932
3933
3934
3935
3936
3937
3938
3939
3940
3941
3942
3943
3944
3945
3946
3947
3948
3949
3950
3951
3952
3953
3954
3955
3956
3957
3958
3959
3960
3961
3962
3963
3964
3965
3966
3967
3968
3969
3970
3971
3972
3973
3974
3975
3976
3977
3978
3979
3980
3981
3982
3983
3984
3985
3986
3987
3988
3989
3990
3991
3992
3993
3994
3995
3996
3997
3998
3999
4000
4001
4002
4003
4004
4005
4006
4007
4008
4009
4010
4011
4012
4013
4014
4015
4016
4017
4018
4019
4020
4021
4022
4023
4024
4025
4026
4027
4028
4029
4030
4031
4032
4033
4034
4035
4036
4037
4038
4039
4040
4041
4042
4043
4044
4045
4046
4047
4048
4049
4050
4051
4052
4053
4054
4055
4056
4057
4058
4059
4060
4061
4062
4063
4064
4065
4066
4067
4068
4069
4070
4071
4072
4073
4074
4075
4076
4077
4078
4079
4080
4081
4082
4083
4084
4085
4086
4087
4088
4089
4090
4091
4092
4093
4094
4095
4096
4097
4098
4099
4100
4101
4102
4103
4104
4105
4106
4107
4108
4109
4110
4111
4112
4113
4114
4115
4116
4117
4118
4119
4120
4121
4122
4123
4124
4125
4126
4127
4128
4129
4130
4131
4132
4133
4134
4135
4136
4137
4138
4139
4140
4141
4142
4143
4144
4145
4146
4147
4148
4149
4150
4151
4152
4153
4154
4155
4156
4157
4158
4159
4160
4161
4162
4163
4164
4165
4166
4167
4168
4169
4170
4171
4172
4173
4174
4175
4176
4177
4178
4179
4180
4181
4182
4183
4184
4185
4186
4187
4188
4189
4190
4191
4192
4193
4194
4195
4196
4197
4198
4199
4200
4201
4202
4203
4204
4205
4206
4207
4208
4209
4210
4211
4212
4213
4214
4215
4216
4217
4218
4219
4220
4221
4222
4223
4224
4225
4226
4227
4228
4229
4230
4231
4232
4233
4234
4235
4236
4237
4238
4239
4240
4241
4242
4243
4244
4245
4246
4247
4248
4249
4250
4251
4252
4253
4254
4255
4256
4257
4258
4259
4260
4261
4262
4263
4264
4265
4266
4267
4268
4269
4270
4271
4272
4273
4274
4275
4276
4277
4278
4279
4280
4281
4282
4283
4284
4285
4286
4287
4288
4289
4290
4291
4292
4293
4294
4295
4296
4297
4298
4299
4300
4301
4302
4303
4304
4305
4306
4307
4308
4309
4310
4311
4312
4313
4314
4315
4316
4317
4318
4319
4320
4321
4322
4323
4324
4325
4326
4327
4328
4329
4330
4331
4332
4333
4334
4335
4336
4337
4338
4339
4340
4341
4342
4343
4344
4345
4346
4347
4348
4349
4350
4351
4352
4353
4354
4355
4356
4357
4358
4359
4360
4361
4362
4363
4364
4365
4366
4367
4368
4369
4370
4371
4372
4373
4374
4375
4376
4377
4378
4379
4380
4381
4382
4383
4384
4385
4386
4387
4388
4389
4390
4391
4392
4393
4394
4395
4396
4397
4398
4399
4400
4401
4402
4403
4404
4405
4406
4407
4408
4409
4410
4411
4412
4413
4414
4415
4416
4417
4418
4419
4420
4421
4422
4423
4424
4425
4426
4427
4428
4429
4430
4431
4432
4433
4434
4435
4436
4437
4438
4439
4440
4441
4442
4443
4444
4445
4446
4447
4448
4449
4450
4451
4452
4453
4454
4455
4456
4457
4458
4459
4460
4461
4462
4463
4464
4465
4466
4467
4468
4469
4470
4471
4472
4473
4474
4475
4476
4477
4478
4479
4480
4481
4482
4483
4484
4485
4486
4487
4488
4489
4490
4491
4492
4493
4494
4495
4496
4497
4498
4499
4500
4501
4502
4503
4504
4505
4506
4507
4508
4509
4510
4511
4512
4513
4514
4515
4516
4517
4518
4519
4520
4521
4522
4523
4524
4525
4526
4527
4528
4529
4530
4531
4532
4533
4534
4535
4536
4537
4538
4539
4530
4531
4532
4533
4534
4535
4536
4537
4538
4539
4540
4541
4542
4543
4544
4545
4546
4547
4548
4549
4540
4541
4542
4543
4544
4545
4546
4547
4548
4549
4550
4551
4552
4553
4554
4555
4556
4557
4558
4559
4550
4551
4552
4553
4554
4555
4556
4557
4558
4559
4560
4561
4562
4563
4564
4565
4566
4567
4568
4569
4560
4561
4562
4563
4564
4565
4566
4567
4568
4569
4570
4571
4572
4573
4574
4575
4576
4577
4578
4579
4570
4571
4572
4573
4574
4575
4576
4577
4578
4579
4580
4581
4582
4583
4584
4585
4586
4587
4588
4589
4580
4581
4582
4583
4584
4585
4586
4587
4588
4589
4590
4591
4592
4593
4594
4595
4596
4597
4598
4599
4590
4591
4592
4593
4594
4595
4596
4597
4598
4599
4600
4601
4602
4603
4604
4605
4606
4607
4608
4609
4600
4601
4602
4603
4604
4605
4606
4607
4608
4609
4610
4611
4612
4613
4614
4615
4616
4617
4618
4619
4610
4611
4612
4613
4614
4615
4616
4617
4618
4619
4620
4621
4622
4623
4624
4625
4626
4627
4628
4629
4620
4621
4622
4623
4624
4625
4626
4627
4628
4629
4630
4631
4632
4633
4634
4635
4636
4637
4638
4639
4630
4631
4632
4633
4634
4635
4636
4637
4638
4639
4640
4641
4642
4643
4644
4645
4646
4647
4648
4649
4640
4641
4642
4643
4644
4645
4646
4647
4648
4649
4650
4651
4652
4653
4654
4655
4656
4657
4658
4659
4650
4651
4652
4653
4654
4655
4656
4657
4658
4659
4660
4661
4662
4663
4664
4665
4666
4667
4668
4669
4660
4661
4662
4663
4664
4665
4666
4667
4668
4669
4670
4671
4672
4673
4674
4675
4676
4677
4678
4679
4670
4671
4672
4673
4674
4675
4676
4677
4678
4679
4680
4681
4682
4683
4684
4685
4686
4687
4688
4689
4680
4681
4682
4683
4684
4685
4686
4687
4688
4689
4690
4691
4692
4693
4694
4695
4696
4697
4698
4699
4690
4691
4692
4693
4694
4695
4696
4697
4698
4699
4700
4701
4702
4703
4704
4705
4706
4707
4708
4709
4700
4701
4702
4703
4704
4705
4706
4707
4708
4709
4710
4711
4712
4713
4714
4715
4716
4717
4718
4719
4710
4711
4712
4713
4714
4715
4716
4717
4718
4719
4720
4721
4722
4723
4724
4725
4726
4727
4728
4729
4720
4721
4722
4723
4724
4725
4726
4727
4728
4729
4730
4731
4732
4733
4734
4735
4736
4737
4738
4739
4730
4731
4732
4733
4734
4735
4736
4737
4738
4739
4740
4741
4742
4743
4744
4745
4746
4747
4748
4749
4740
4741
4742
4743
4744
4745
4746
4747
4748
4749
4750
4751
4752
4753
4754
4755
4756
4757
4758
4759
4750
4751
4752
4753
4754
4755
4756
4757
4758
4759
4760
4761
4762
4763
4764
4765
4766
4767
4768
4769
4760
4761
4762
4763
4764
4765
4766
4767
4768
4769
4770
4771
4772
4773
4774
4775
4776
4777
4778
4779
4770
4771
4772
4773
4774
4775
4776
4777
4778
4779
4780
4781
4782
4783
4784
4785
4786
4787
4788
4789
4780
4781
4782
4783
4784
4785
4786
4787
4788
4789
4790
4791
4792
4793
4794
4795
4796
4797
4798
4799
4790
4791
4792
4793
4794
4795
4796
4797
4798
4799
4800
4801
4802
4803
4804
4805
4806
4807
4808
4809
4800
4801
4802
4803
4804
4805
4806
4807
4808
4809
4810
4811
4812
4813
4814
4815
4816
4817
4818
4819
4810
4811
4812
4813
4814
4815
4816
4817
4818
4819
4820
4821
4822
4823
4824
4825
4826
4827
4828
4829
4820
4821
4822
4823
4824
4825
4826
4827
4828
4829
4830
4831
4832
4833
4834
4835
4836
4837
4838
4839
4830
4831
4832
4833
4834
4835
4836
4837
4838
4839
4840
4841
4842
4843
4844
4845
4846
4847
4848
4849
4840
4841
4842
4843
4844
4845
4846
4847
4848
4849
4850
4851
4852
4853
4854
4855
4856
4857
4858
4859
4850
4851
4852
4853
4854
4855
4856
4857
4858
4859
4860
4861
4862
4863
4864
4865
4866
4867
4868
4869
4860
4861
4862
4863
4864
4865
4866
4867
4868
4869
4870
4871
4872
4873
4874
4875
4876
4877
4878
4879
4870
4871
4872
4873
4874
4875
4876
4877
4878
4879
4880
4881
4882
4883
4884
4885
4886
4887
4888
4889
4880
4881
4882
4883
4884
4885
4886
4887
4888
4889
4890
4891
4892
4893
4894
4895
4896
4897
4898
4899
4890
4891
4892
4893
4894
4895
4896
4897
4898
4899
4900
4901
4902
4903
4904
4905
4906
4907
4908
4909
4900
4901
4902
4903
4904
4905
4906
4907
4908
4909
4910
4911
4912
4913
4914
4915
4916
4917
4918
4919
4910
4911
4912
4913

```

Shown below are the settings that was done in the Network window.



Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement: status_group	_DFT_	Total Degrees of Freedom	35636	.	.
REP_status_group	Replacement: status_group	_DFE_	Degrees of Freedom for Error	35295	.	.
REP_status_group	Replacement: status_group	_DFM_	Model Degrees of Freedom	341	.	.
REP_status_group	Replacement: status_group	_NW_	Number of Estimated Weights	341	.	.
REP_status_group	Replacement: status_group	_AIC_	Akaike's Information Criterion	30416.71	.	.
REP_status_group	Replacement: status_group	_SBC_	Schwarz's Bayesian Criterion	33308.77	.	.
REP_status_group	Replacement: status_group	_ASE_	Average Squared Error	0.133892	0.139027	0.135903
REP_status_group	Replacement: status_group	_MAX_	Maximum Absolute Error	0.998621	0.999807	0.999831
REP_status_group	Replacement: status_group	_DIV_	Divisor for ASE	71272	23760	23768
REP_status_group	Replacement: status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement: status_group	_RASE_	Root Average Squared Error	0.365913	0.372863	0.36865
REP_status_group	Replacement: status_group	_SSE_	Sum of Squared Errors	9542.749	3303.281	3230.134
REP_status_group	Replacement: status_group	_SUMW_	Sum of Case Weights Times Freq	71272	23760	23768
REP_status_group	Replacement: status_group	_FPE_	Final Prediction Error	0.136479	.	.
REP_status_group	Replacement: status_group	_MSE_	Mean Squared Error	0.135188	0.139027	0.135903
REP_status_group	Replacement: status_group	_RFPE_	Root Final Prediction Error	0.369431	.	.
REP_status_group	Replacement: status_group	_RMSE_	Root Mean Squared Error	0.367676	0.372863	0.36865
REP_status_group	Replacement: status_group	_AVERR_	Average Error Function	0.4172	0.434816	0.424933
REP_status_group	Replacement: status_group	_ERR_	Error Function	29734.71	10331.22	10099.81
REP_status_group	Replacement: status_group	_MISC_	Misclassification Rate	0.188826	0.19697	0.191939
REP_status_group	Replacement: status_group	_WRONG_	Number of Wrong Classifications	6729	2340	2281

Results - Node: Neural Network: Diagram: Project Diagram

File Edit View Window

**Output**

```

937 FUNCTIONAL      FUNCTIONAL      80.3019    91.8698    20159    56.5692
938 NON FUNCTIONAL  FUNCTIONAL      19.6981    36.1133    4945     13.8764
939 FUNCTIONAL      NON FUNCTIONAL 16.9389    8.1302     1784     5.0062
940 NON FUNCTIONAL  NON FUNCTIONAL 83.0611    63.8867    8748     24.5482
941
942
943 Data Role=VALIDATE Target Variable=REP_Status_Group Target Label=Replacement: status_group
944
945           Target      Outcome      Percentage   Frequency   Total
946           Target      Outcome      Percentage   Count       Percentage
947
948 FUNCTIONAL      FUNCTIONAL      79.7015    91.2509    6675     56.1869
949 NON FUNCTIONAL  FUNCTIONAL      20.2985    37.2399    1700     14.3098
950 FUNCTIONAL      NON FUNCTIONAL 18.2596    8.7491     640      5.3872
951 NON FUNCTIONAL  NON FUNCTIONAL 81.7404    62.7601    2865     24.1162
952
953
954
955
956 Event Classification Table
957
958 Data Role=TRAIN Target=REP_Status_Group Target Label=Replacement: status_group
959
960   False      True      False      True
961 Negative  Negative  Positive  Positive
962
963 4945      20159    1784      8748
964
965
966 Data Role=VALIDATE Target=REP_Status_Group Target Label=Replacement: status_group
967
968   False      True      False      True
969 Negative  Negative  Positive  Positive
970
971 1700      6675     640       2865
972
973
974

```



The misclassification rate for the above model is 0.1969.

## 20. Model Comparison:

This node was used to compare all the models. The main factor distinguishing these models was their misclassification rate. The following are the misclassification rates:

Fit Statistics												
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Divisor for ASE
Y	Tree3	Tree3	Decision Tree- 3 Branch	REP_status_group	Replaceme...	0.187626	35636	0.179958	0.981548	9450.33	0.132595	0.364136
	Tree	Tree	Decision Tree- 2 Branch	REP_status_group	Replaceme...	0.191162	35636	0.185487	0.98	9969.863	0.139885	0.374012
Reg3	Reg3		Forward- Logistic Regression	REP_status_group	Replaceme...	0.19638	35636	0.19267	0.998908	9914.235	0.139104	0.372967
Reg2	Reg2		Stepwise- Logistic Regression	REP_status_group	Replaceme...	0.19638	35636	0.192642	0.998806	9915.889	0.139127	0.372998
Neural	Neural		Neural Network	REP_status_group	Replaceme...	0.19697	35636	0.188826	0.998621	9542.749	0.133892	0.365913
Reg4	Reg4		Backward- Logistic Regression	REP_status_group	Replaceme...	0.197306	35636	0.192979	0.998879	9919.232	0.139174	0.373061
Boost	Boost		Gradient Boosting	REP_status_group	Replaceme...	0.210943	35636	0.208778	0.933361	11000.89	0.154351	0.392875
Tree4	Tree4		Decision Tree- Interactive	REP_status_group	Replaceme...	0.21229	35636	0.206813	0.95082	10549.21	0.148013	0.384725
Tree2	Tree2		Decision Tree _Variable Selection	REP_status_group	Replaceme...	0.212963	35636	0.207178	0.970643	10846.44	0.152184	0.390107

From the above model comparison result we find that **Decision tree with 3 branch** is the best model with a misclassification rate of **0.1876**

Results - Node: Model Comparison Diagram: Project Diagram

File Edit View Window

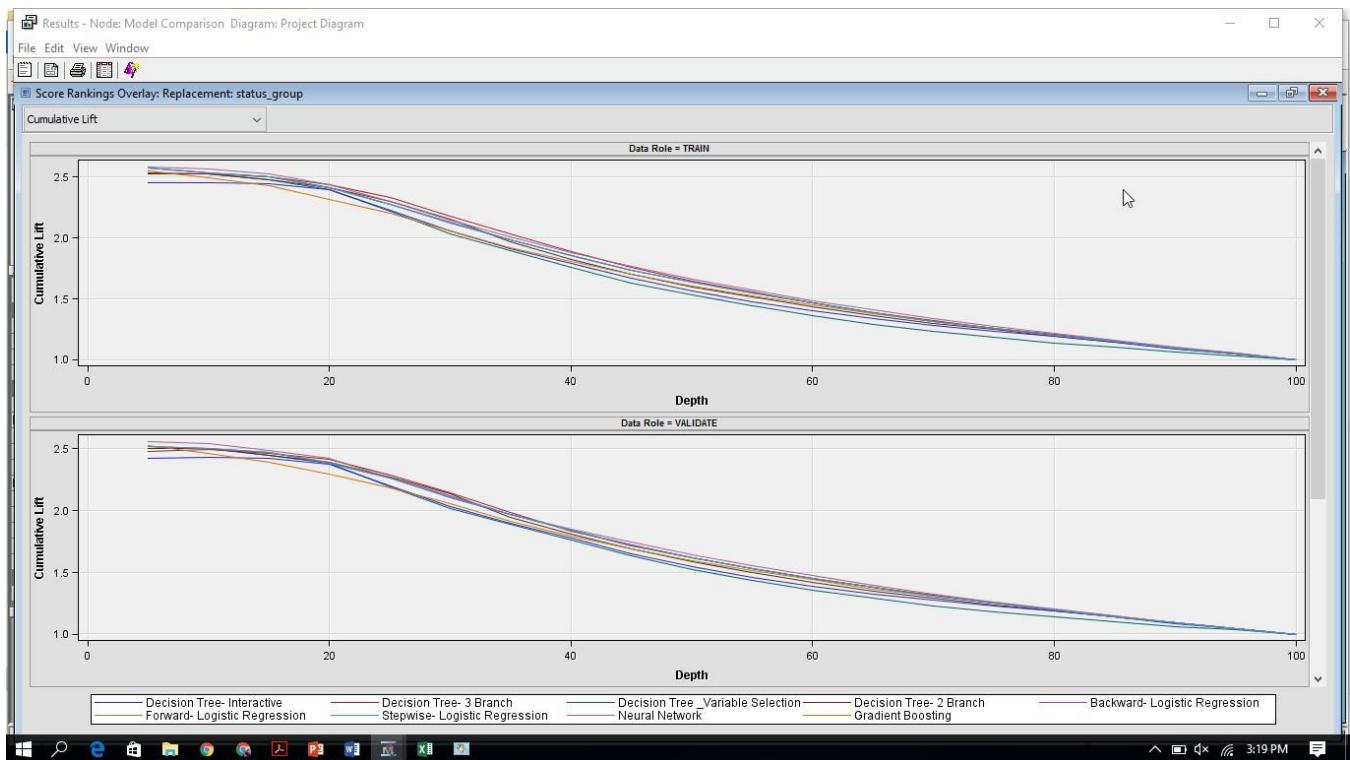
Output

```

28
29 Fit Statistics
30 Model Selection based on Valid: Misclassification Rate (_VMISC_)
31
32
33                               Valid:
34 Selected Model      Model Description      Misclassification
35 Model    Node          Rate
36
37     Y     Tree3      Decision Tree- 3 Branch      0.18763
38     Tree     Decision Tree- 2 Branch      0.19116
39     Reg3     Forward- Logistic Regression      0.19638
40     Reg2     Stepwise- Logistic Regression      0.19638
41     Neural    Neural Network      0.19697
42     Reg4     Backward- Logistic Regression      0.19731
43     Boost     Gradient Boosting      0.21094
44     Tree4     Decision Tree- Interactive      0.21229
45     Tree2     Decision Tree _Variable Selection      0.21296
46
47 Train:           Valid:
48 Average   Train:   Average
49 Squared   Misclassification   Squared
50 Error     Rate       Error
51
52 0.13260      0.17996      0.14011
53 0.13988      0.18549      0.14422
54 0.13910      0.19267      0.14300
55 0.13913      0.19264      0.14307
56 0.13889      0.18883      0.13903
57 0.13917      0.19298      0.14311
58 0.15435      0.20878      0.15662
59 0.14801      0.20681      0.15190
60 0.15218      0.20718      0.15476
61
62
63
64
65

```

Windows taskbar: Search, Start, File Explorer, Internet Explorer, Task View, Power, Task Manager, File Explorer, Task View, Power, Task Manager, 3:21 PM



## Conclusion:

- The Decision Tree with 3 branches model works best with a misclassification rate of approx. 18.76%
- From the IF- THEN logic of the Decision Tree with 3 branches we predict that the following are the main areas where the government or any other private organizations should concentrate:

\*-----\*

Node = 6

\*-----\*

if Replacement: quantity IS ONE OF: DRY or MISSING

then

Tree Node Identifier = 6

Number of Observations = 3781

Predicted: REP\_status\_group=non functional = 0.97

Predicted: REP\_status\_group=functional = 0.03

\*-----\*

Node = 47

\*-----\*

if Replacement: waterpoint\_type IS ONE OF: OTHER

AND Replacement: source\_type IS ONE OF: BOREHOLE, SHALLOW WELL, RIVER/LAKE  
or MISSING

AND Replacement: region\_code IS ONE OF: 90, 18, 10, 3, 19, 80, 6, 12, 16,  
13, 20, 5, 4, 7, 15, 14, 9, 8 or MISSING

AND Replacement: quantity IS ONE OF: SEASONAL, INSUFFICIENT

then

Tree Node Identifier = 47

Number of Observations = 715

Predicted: REP\_status\_group=non functional = 0.95

Predicted: REP\_status\_group=functional = 0.05

\*-----\*

Node = 53

\*-----\*

if Replacement: waterpoint\_type IS ONE OF: COMMUNAL STANDPIPE MULTIPLE or MISSING

AND Replacement: region IS ONE OF: MTWARA, KAGERA, MWANZA, LINDI

AND Replacement: quantity IS ONE OF: SEASONAL, INSUFFICIENT

AND Imputed: Replacement: payment IS ONE OF: NEVER PAY

then

Tree Node Identifier = 53

Number of Observations = 129

Predicted: REP\_status\_group=non functional = 0.98

Predicted: REP\_status\_group=functional = 0.02

\*-----\*

Node = 56

\*-----\*

if Replacement: waterpoint\_type IS ONE OF: COMMUNAL STANDPIPE MULTIPLE or MISSING

AND Replacement: region IS ONE OF: SHINYANGA, PWANI, RUVUMA, KILIMANJARO, KIGOMA, SINGIDA, MANYARA, MARA, TANGA, TABORA

AND Replacement: quantity IS ONE OF: SEASONAL, INSUFFICIENT

AND Imputed: Replacement: construction\_year IS ONE OF: 1974, 1980, 1970

then

Tree Node Identifier = 56

Number of Observations = 33

Predicted: REP\_status\_group=non functional = 0.97

Predicted: REP\_status\_group=functional = 0.03

```
*-----*
```

Node = 70

```
*-----*
```

if Replacement: quantity IS ONE OF: ENOUGH or MISSING  
AND Replacement: extraction\_type\_group IS ONE OF: OTHER  
AND Replacement: basin IS ONE OF: LAKE VICTORIA  
AND Imputed: Replacement: source IS ONE OF: LAKE, OTHER  
then

Tree Node Identifier = 70  
Number of Observations = 14  
Predicted: REP\_status\_group=non functional = 0.86  
Predicted: REP\_status\_group=functional = 0.14

```
*-----*
```

Node = 74

```
*-----*
```

if gps\_height < 715.5 AND gps\_height >= 419.5  
AND Replacement: quantity IS ONE OF: ENOUGH or MISSING  
AND Replacement: extraction\_type\_group IS ONE OF: OTHER  
AND Imputed: Replacement: source IS ONE OF: SPRING, RAINWATER HARVESTING,  
RIVER, DAM  
then

Tree Node Identifier = 74  
Number of Observations = 6  
Predicted: REP\_status\_group=non functional = 1.00  
Predicted: REP\_status\_group=functional = 0.00

- For the Entire Node Rules File, Click on the link below:
- [Node Rules](#)

**REFERENCES:**

- Jaro–Winkler distance – Wikipidea article:  
[https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler\\_distance](https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance)
- Drivendata:  
<https://www.drivendata.org/competitions/7/>