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.

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).

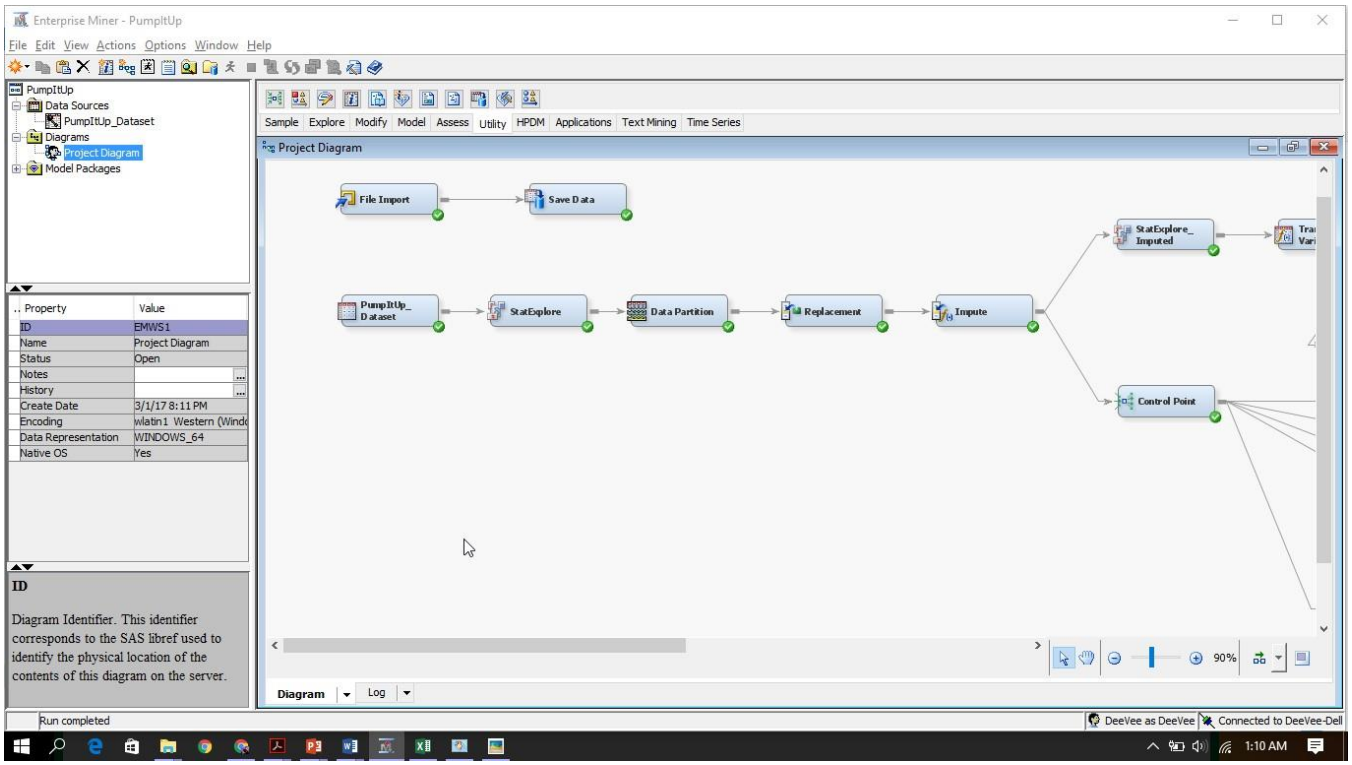
High level Data summary:

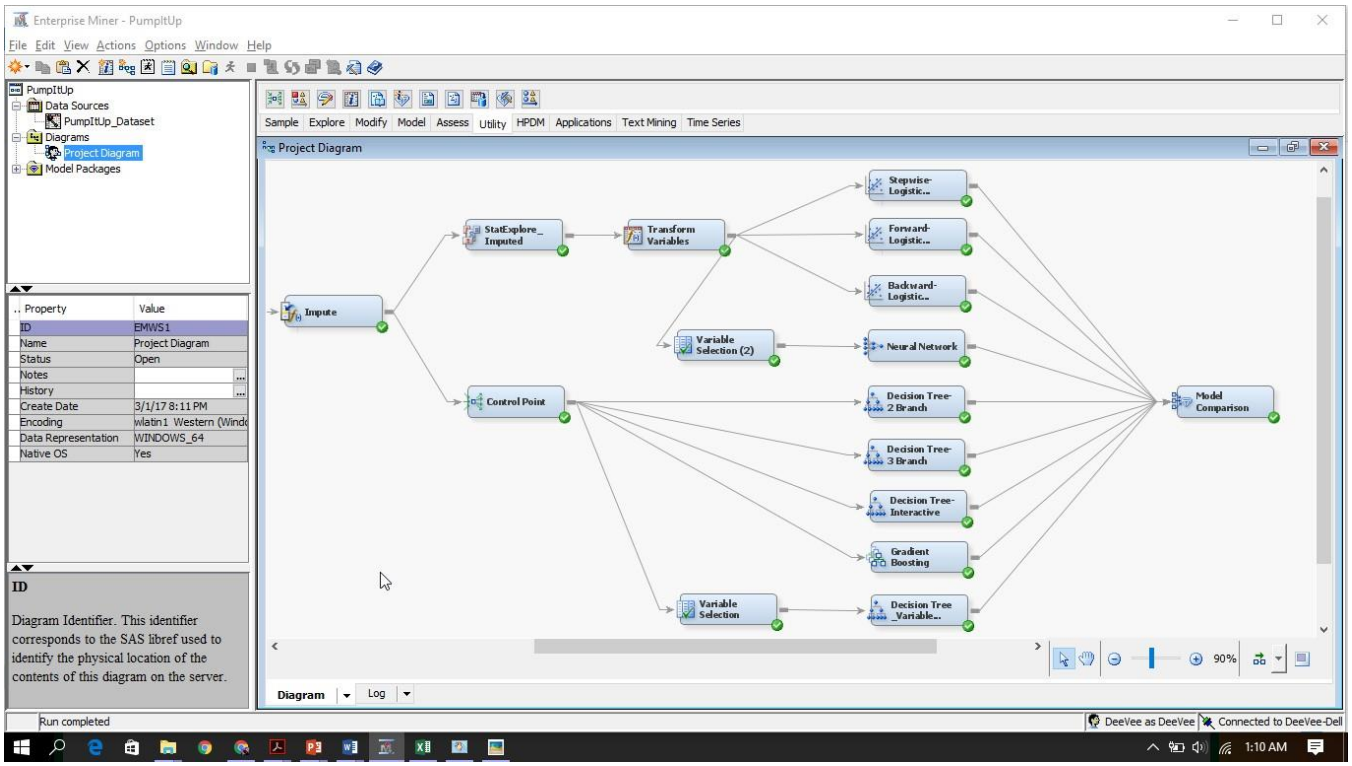
Total observations in the dataset	59400
Total number of binary variables	0
Total number of nominal variables	29
Total number of interval variables	10
Outcome / target variable	status_group, quality_group
Level of the target variable (nominal, binary or interval)	Nominal
Percentage of the variables belong 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. lga - 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:



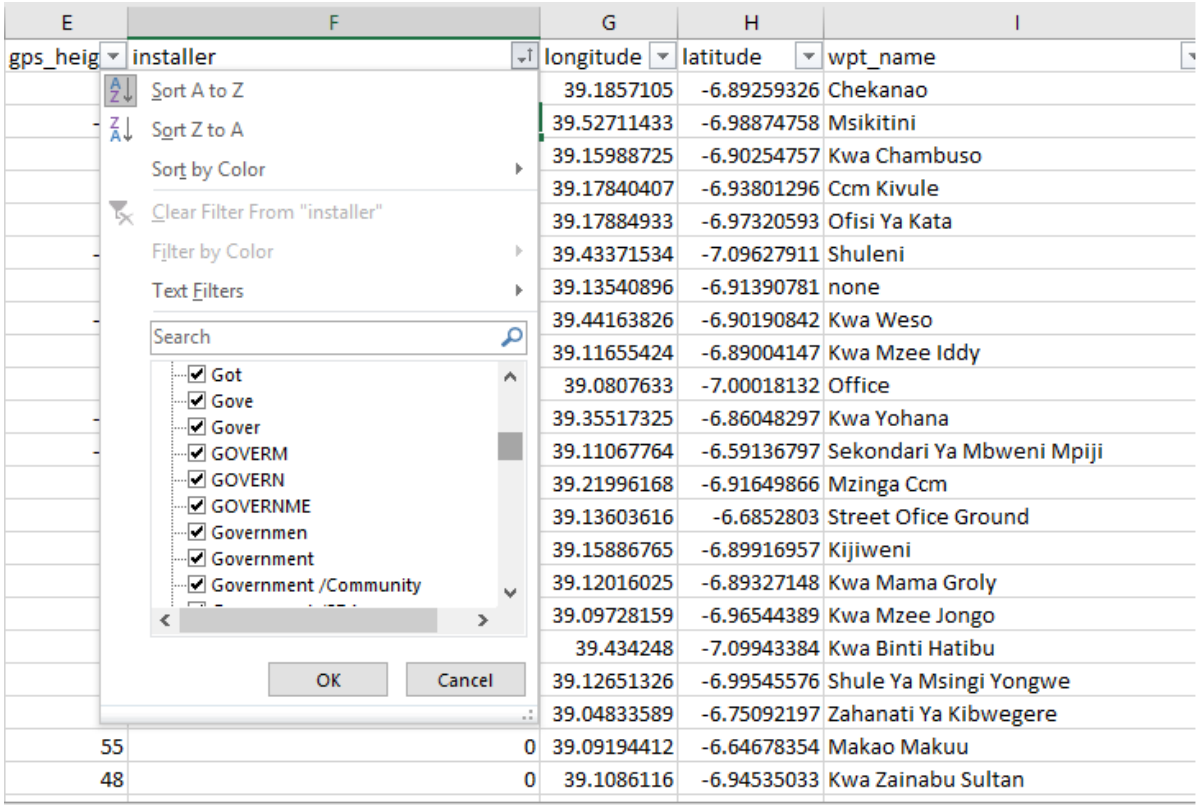


Data Pre-Processing

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, lga*** 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



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 it's 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:

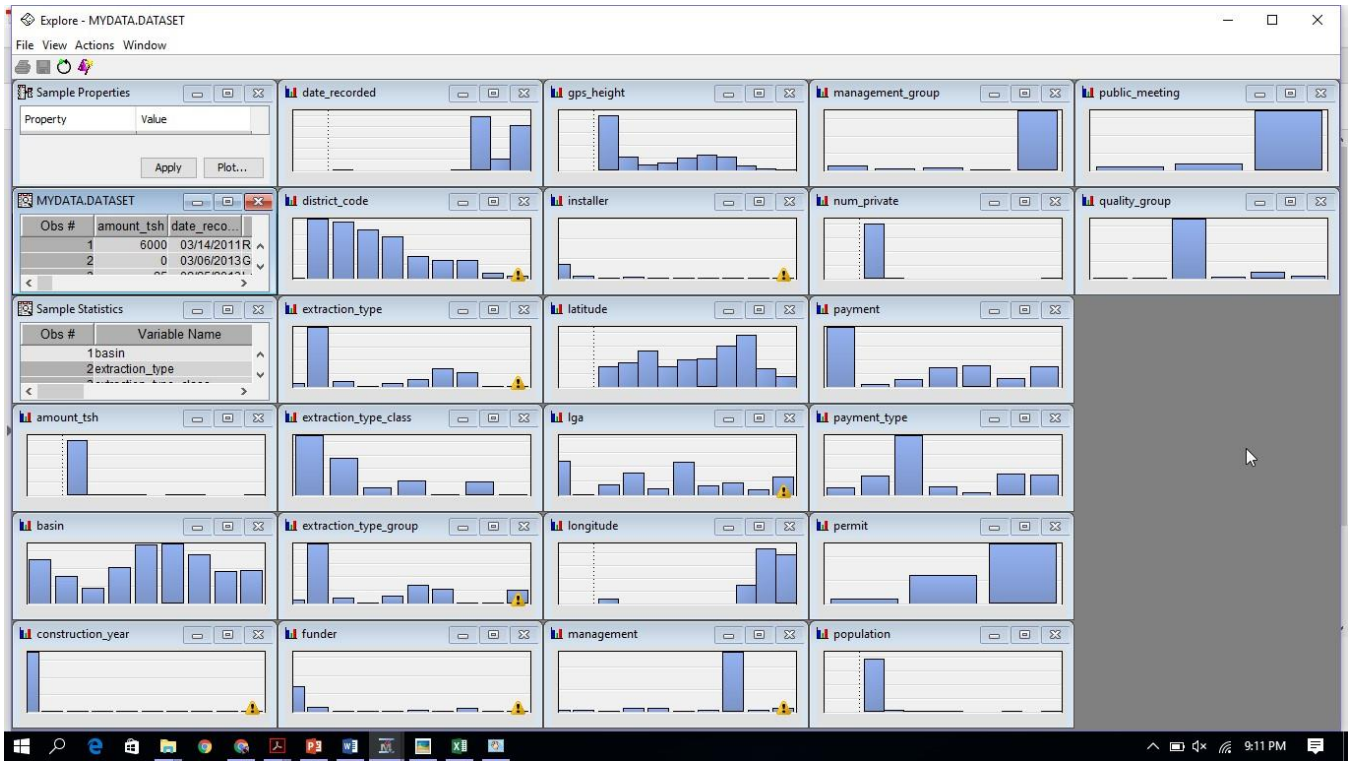
Our dataset consists of 2 csv files (one for input variables and another for target variable). We combined these two into a single csv file. This csv file was then imported using the File import node.

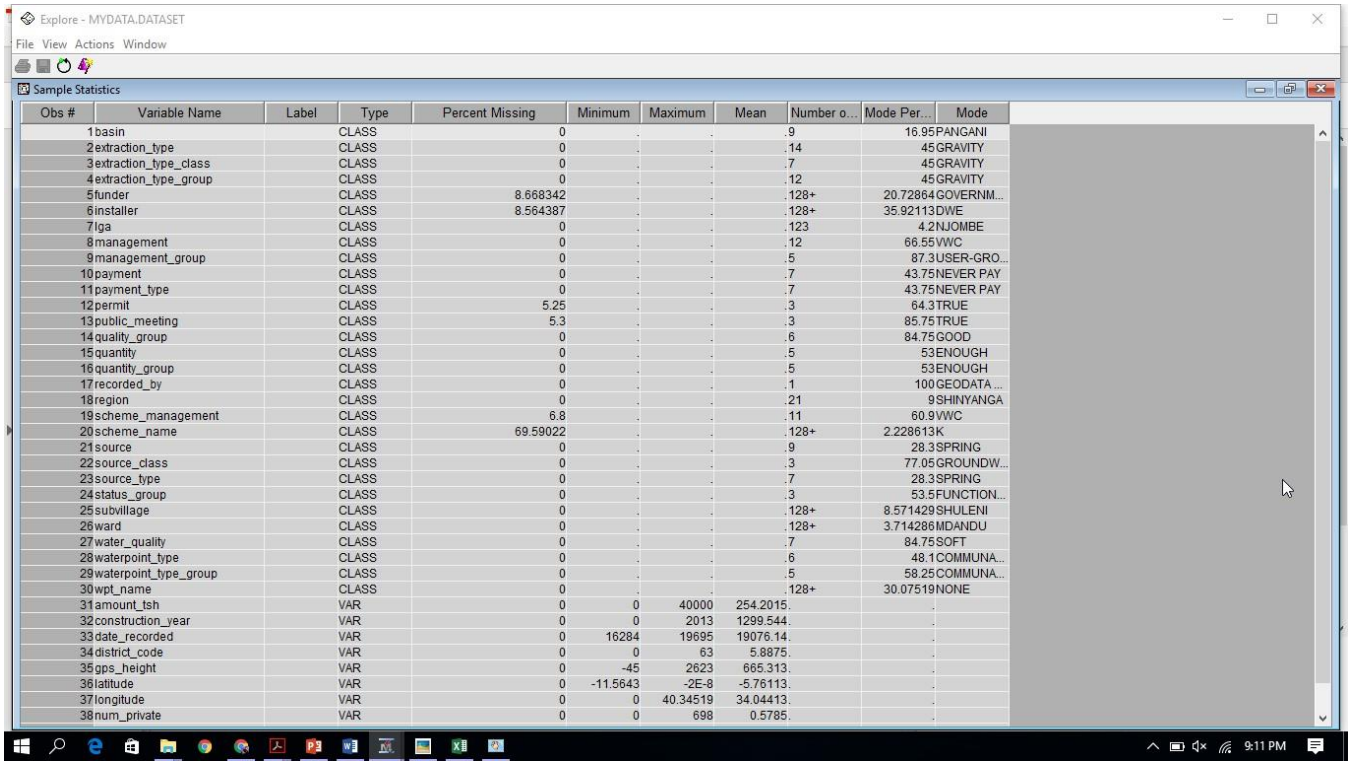
2. Save Data:

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.

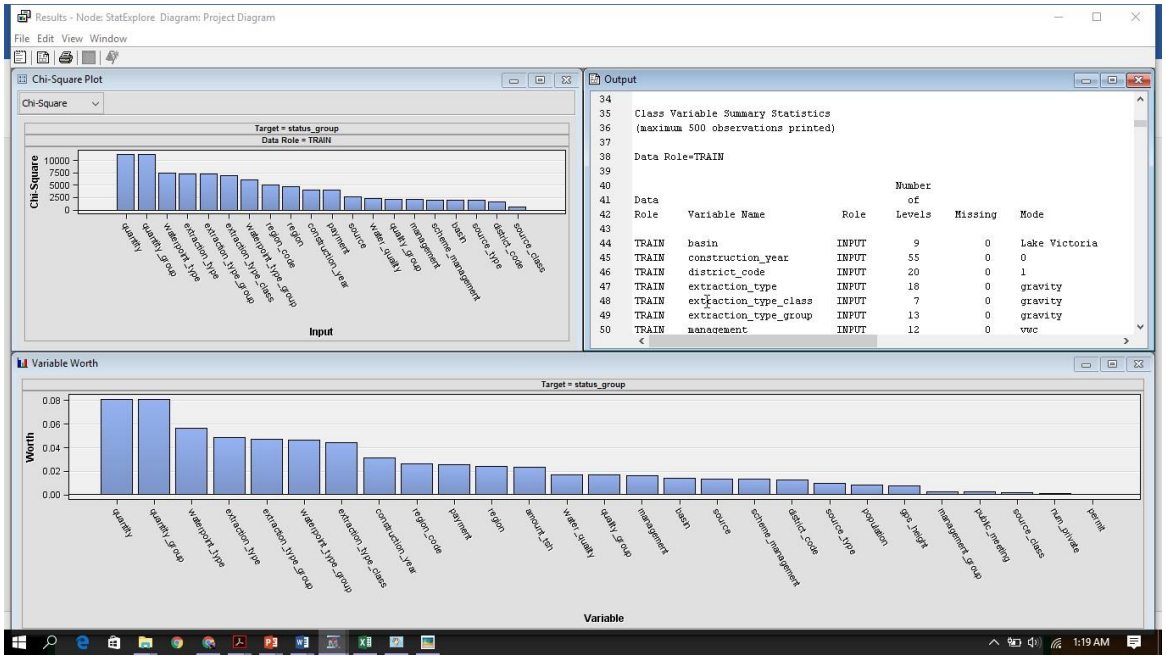




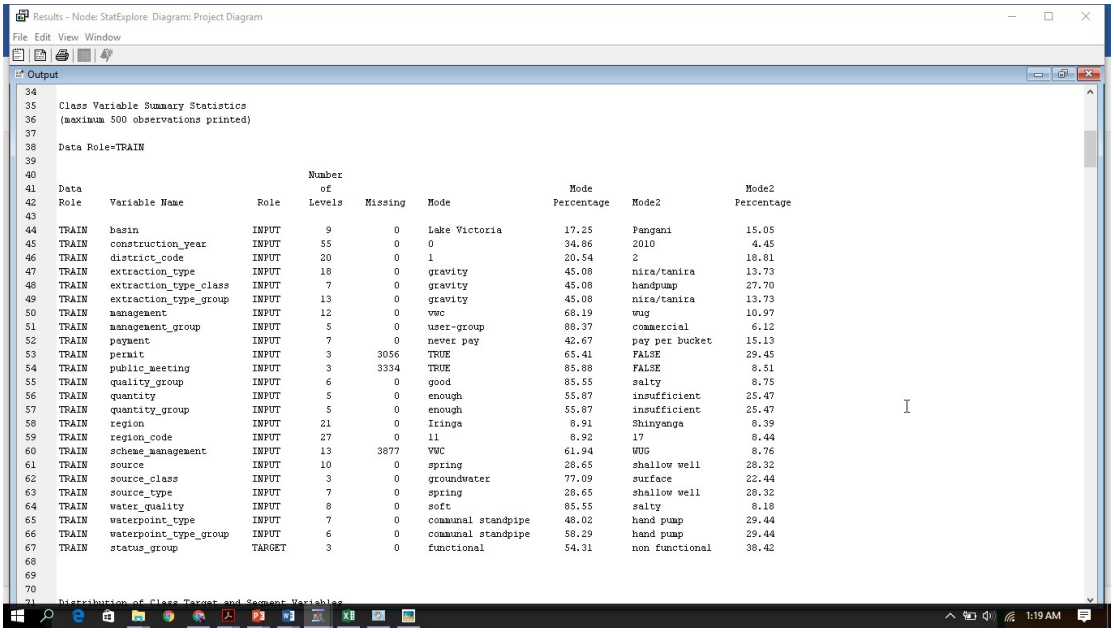
Obs #	Variable Name	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1	basin		CLASS	0	.	.	.	9	16.95	PANGANI
2	extraction_type		CLASS	0	.	.	.	14	45	GRAVITY
3	extraction_type_class		CLASS	0	.	.	.	7	45	GRAVITY
4	extraction_type_group		CLASS	0	.	.	.	12	45	GRAVITY
5	funder		CLASS	8.668342	.	.	.	128+	20.72864	GOVERNM...
6	installer		CLASS	8.564387	.	.	.	128+	35.92113	DWE
7	lga		CLASS	0	.	.	.	123	4.2	NJOMBE
8	management		CLASS	0	.	.	.	12	66.55	VWC
9	management_group		CLASS	0	.	.	.	5	87.3	USER-GRO...
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	GEODATA ...
18	region		CLASS	0	.	.	.	21	9	SHINYANGA
19	scheme_management		CLASS	6.8	.	.	.	11	60.9	VWC
20	scheme_name		CLASS	69.59022	.	.	.	128+	2.228613	K
21	source		CLASS	0	.	.	.	9	28.3	SPRING
22	source_class		CLASS	0	.	.	.	3	77.05	GROUNDW...
23	source_type		CLASS	0	.	.	.	7	28.3	SPRING
24	status_group		CLASS	0	.	.	.	3	53.5	FUNCTION...
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	COMMUNA...
29	waterpoint_type_group		CLASS	0	.	.	.	5	58.25	COMMUNA...
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:



Data Role	Variable Name	Role	Number of Levels	Missing	Mode
TRAIN	basin	INPUT	9	0	Lake Victoria
TRAIN	construction_year	INPUT	55	0	0
TRAIN	district_code	INPUT	20	0	1
TRAIN	extraction_type	INPUT	18	0	gravity
TRAIN	extraction_type_class	INPUT	7	0	gravity
TRAIN	extraction_type_group	INPUT	13	0	gravity
TRAIN	management	INPUT	12	0	vwc



Results - Node: StatExplore Diagram: Project Diagram

File Edit View Window

Output

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	basin	INPUT	9	0	Lake Victoria	17.25	Pangani	15.05
TRAIN	construction_year	INPUT	55	0	0	34.86	2010	4.45
TRAIN	district_code	INPUT	20	0	1	20.54	2	18.81
TRAIN	extraction_type	INPUT	18	0	gravity	45.08	nira/tanira	13.73
TRAIN	extraction_type_class	INPUT	7	0	gravity	45.08	handpump	27.70
TRAIN	extraction_type_group	INPUT	13	0	gravity	45.08	nira/tanira	13.73
TRAIN	management	INPUT	12	0	vwc	68.19	wug	10.97
TRAIN	management_group	INPUT	5	0	user-group	88.37	commercial	6.12
TRAIN	payment	INPUT	7	0	never pay	42.67	pay per bucket	15.13
TRAIN	permit	INPUT	3	3056	TRUE	65.41	FALSE	29.45
TRAIN	public_meeting	INPUT	3	3334	TRUE	85.88	FALSE	8.51
TRAIN	quality_group	INPUT	6	0	good	85.55	salty	8.75
TRAIN	quantity	INPUT	5	0	enough	55.87	insufficient	25.47
TRAIN	quantity_group	INPUT	5	0	enough	55.87	insufficient	25.47
TRAIN	region	INPUT	21	0	Iringa	8.91	Shinyanga	8.39
TRAIN	region_code	INPUT	27	0	11	8.92	17	8.44
TRAIN	scheme_management	INPUT	13	3877	VWC	61.94	WUG	8.76
TRAIN	source	INPUT	10	0	spring	28.65	shallow well	28.32
TRAIN	source_class	INPUT	3	0	groundwater	77.09	surface	22.44
TRAIN	source_type	INPUT	7	0	spring	28.65	shallow well	28.32
TRAIN	water_quality	INPUT	8	0	soft	85.55	salty	8.18
TRAIN	waterpoint_type	INPUT	7	0	communal standpipe	48.02	hand pump	29.44
TRAIN	waterpoint_type_group	INPUT	6	0	communal standpipe	58.29	hand pump	29.44
TRAIN	status_group	TARGET	3	0	functional	54.31	non functional	38.42

Distribution of Class Target and Segment 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.

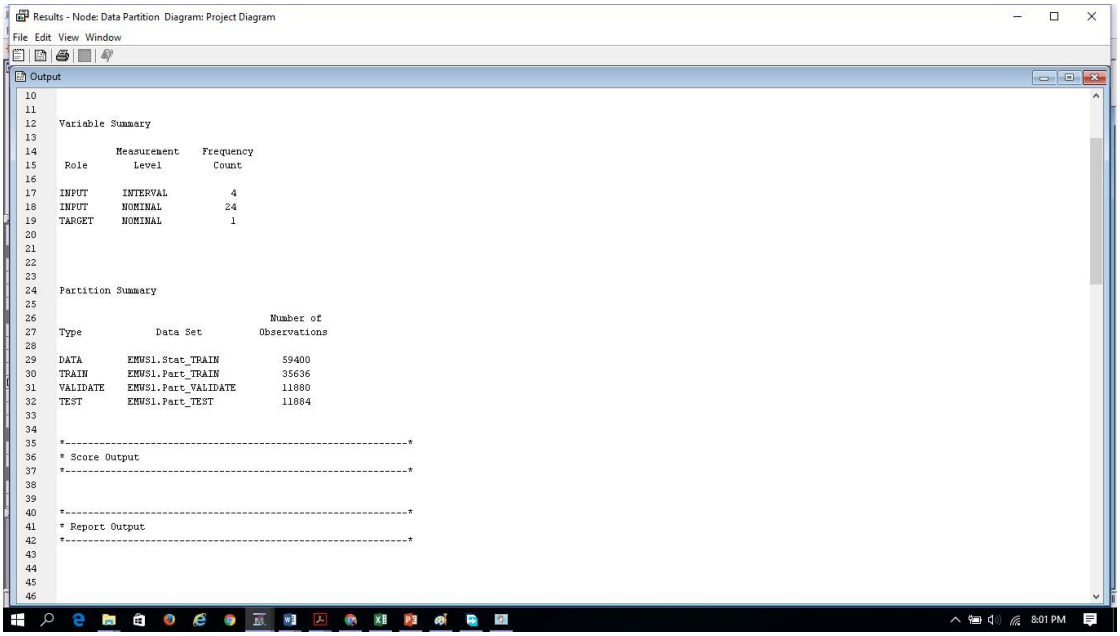
5. Data Partition:

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

Training set: 60%

Validation set: 20%

Test set: 20%



Results - Node: Data Partition Diagram: Project Diagram

File Edit View Window

Output

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	4
INPUT	NOMINAL	24
TARGET	NOMINAL	1

Partition Summary

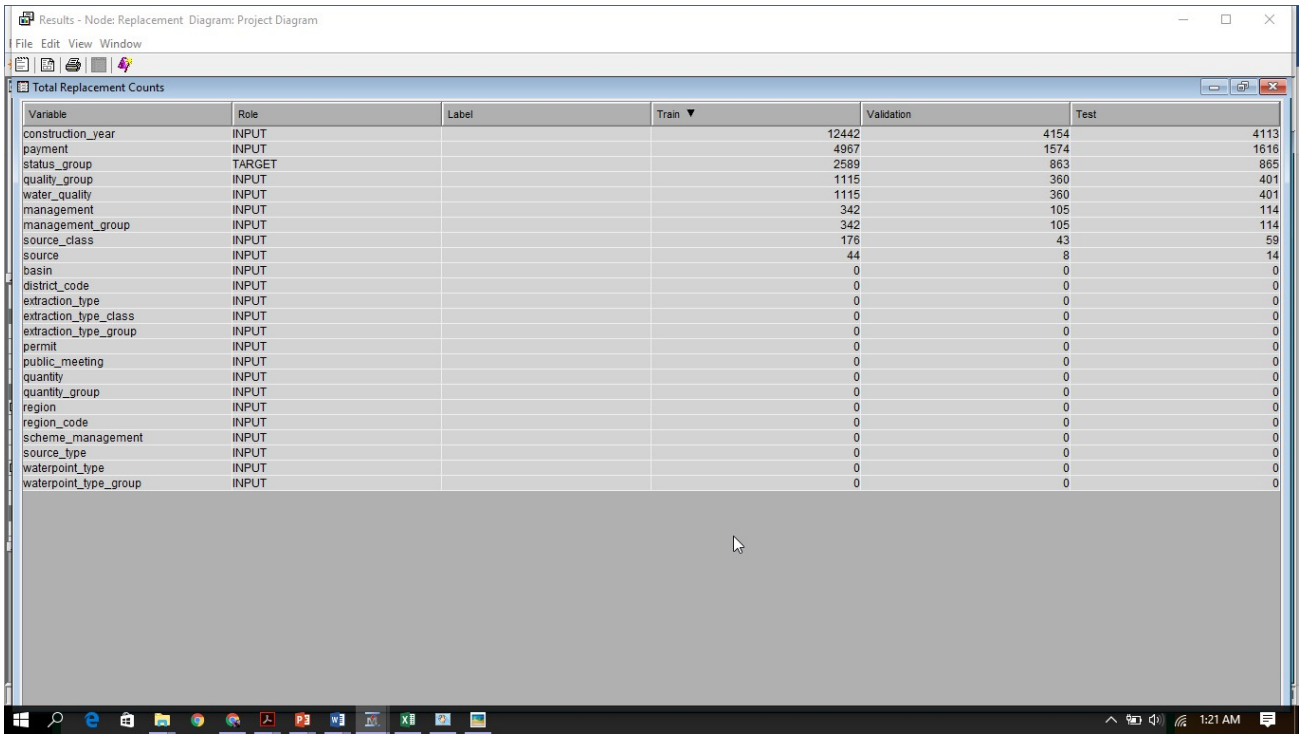
Type	Data Set	Number of Observations
DATA	EMWS1.Stat_TRAIN	59400
TRAIN	EMWS1.Part_TRAIN	35636
VALIDATE	EMWS1.Part_VALIDATE	11880
TEST	EMWS1.Part_TEST	11884

--
* Score Output
--

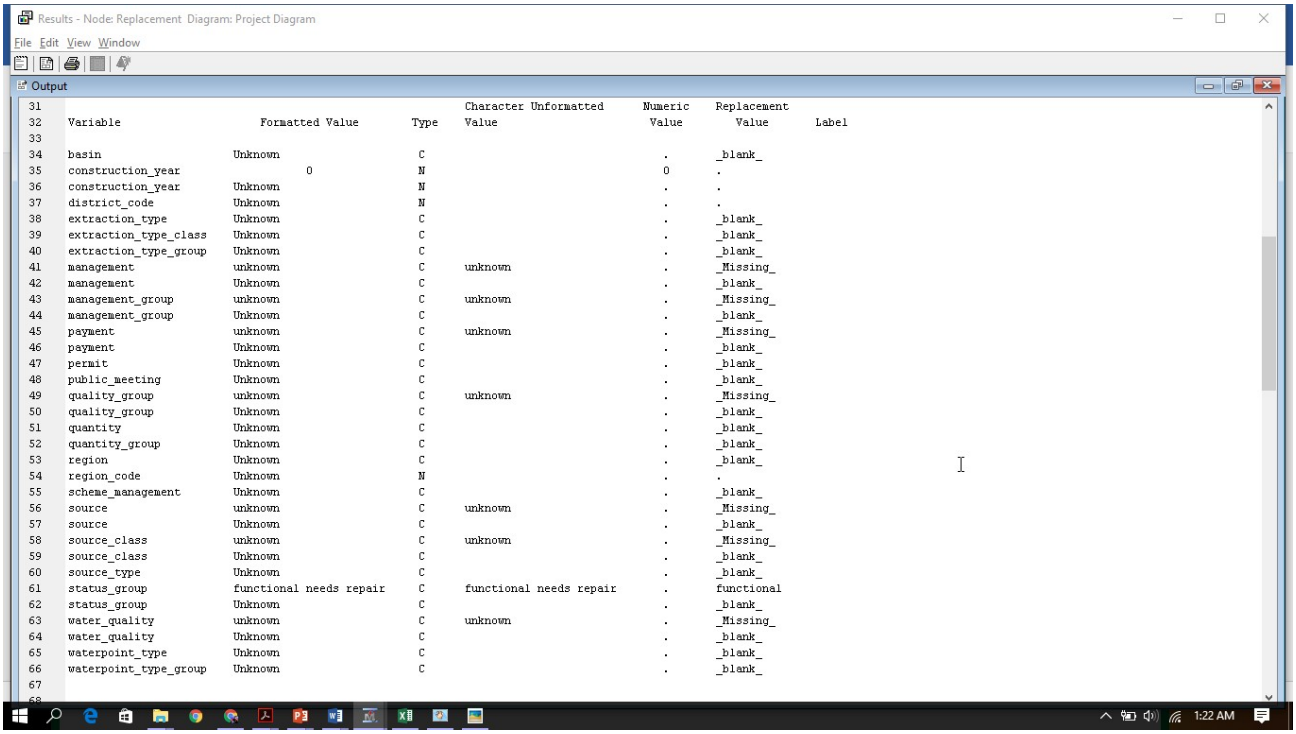
--
* Report Output
--

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”.



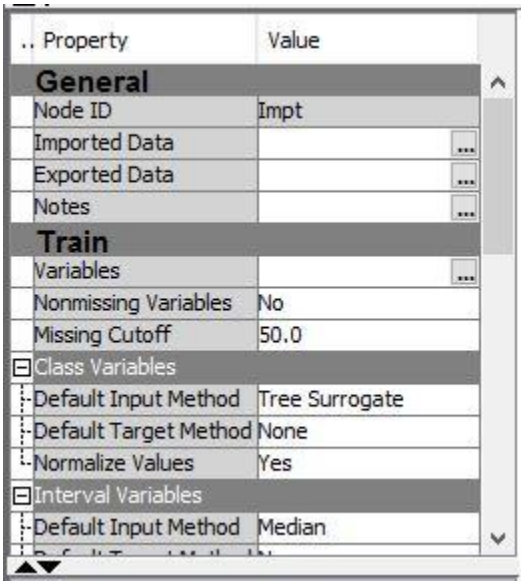
Variable	Role	Label	Train ▼	Validation	Test
construction_year	INPUT		12442	4154	4113
payment	INPUT		4967	1574	1616
status_group	TARGET		2589	863	865
quality_group	INPUT		1115	360	401
water_quality	INPUT		1115	360	401
management	INPUT		342	105	114
management_group	INPUT		342	105	114
source_class	INPUT		176	43	59
source	INPUT		44	8	14
basin	INPUT		0	0	0
district_code	INPUT		0	0	0
extraction_type	INPUT		0	0	0
extraction_type_class	INPUT		0	0	0
extraction_type_group	INPUT		0	0	0
permit	INPUT		0	0	0
public_meeting	INPUT		0	0	0
quantity	INPUT		0	0	0
quantity_group	INPUT		0	0	0
region	INPUT		0	0	0
region_code	INPUT		0	0	0
scheme_management	INPUT		0	0	0
source_type	INPUT		0	0	0
waterpoint_type	INPUT		0	0	0
waterpoint_type_group	INPUT		0	0	0

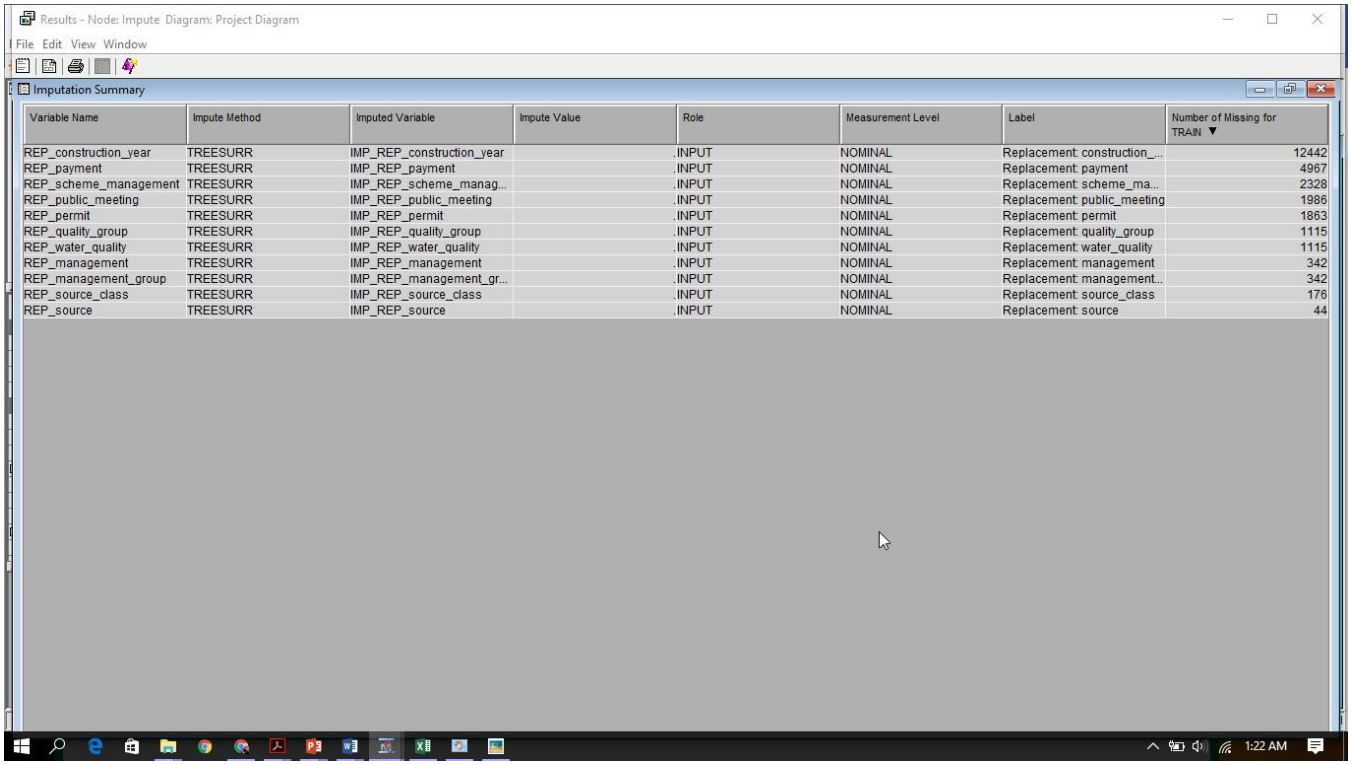


	Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value	Label
34	basin	Unknown	C		.	_blank_	
35	construction_year	0	N		0	.	
36	construction_year	Unknown	N		.	.	
37	district_code	Unknown	N		.	.	
38	extraction_type	Unknown	C		.	_blank_	
39	extraction_type_class	Unknown	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_	

7. Impute:

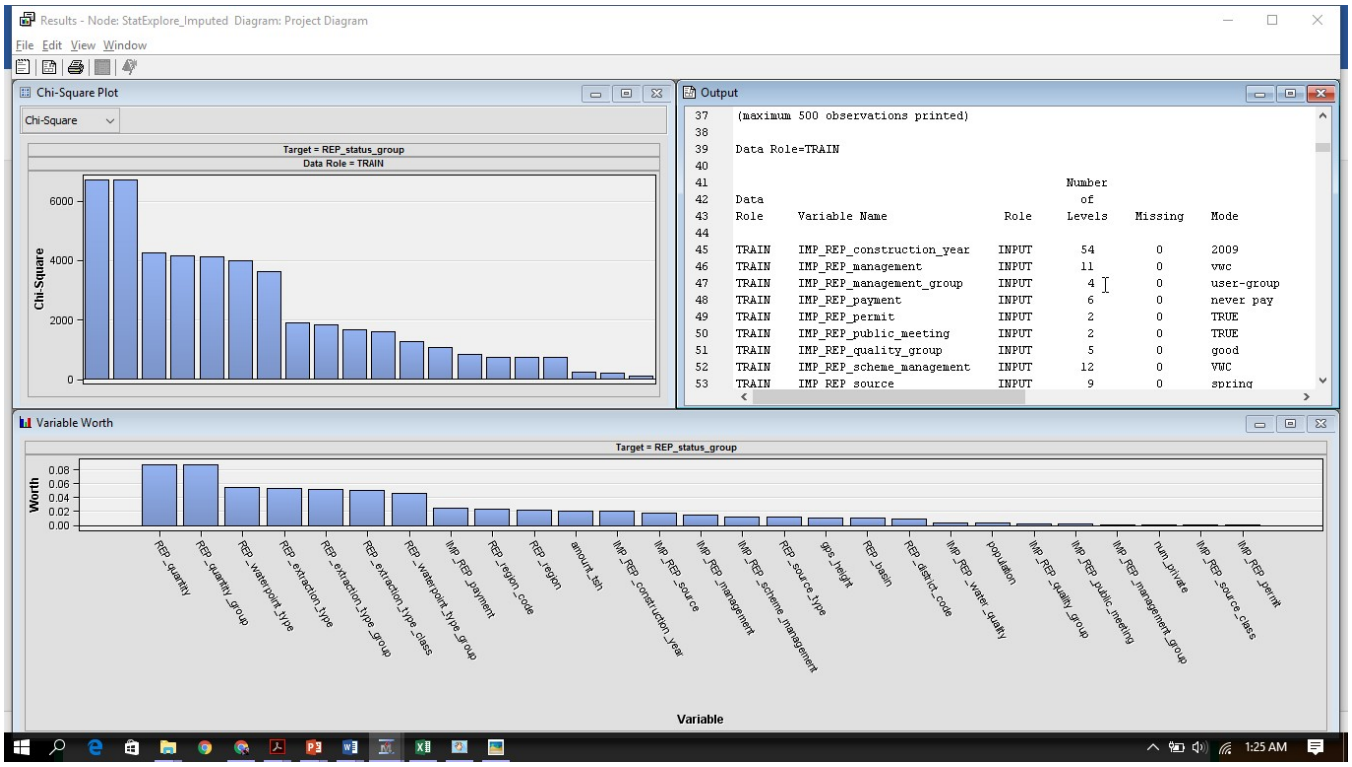
This node was used to impute the missing values for the following variables using **'Tree Surrogate'** method: ***Construction_year, payment, permit, public_meeting, scheme_management, quality_group, water_quality, management, management_group, source_class and source***.

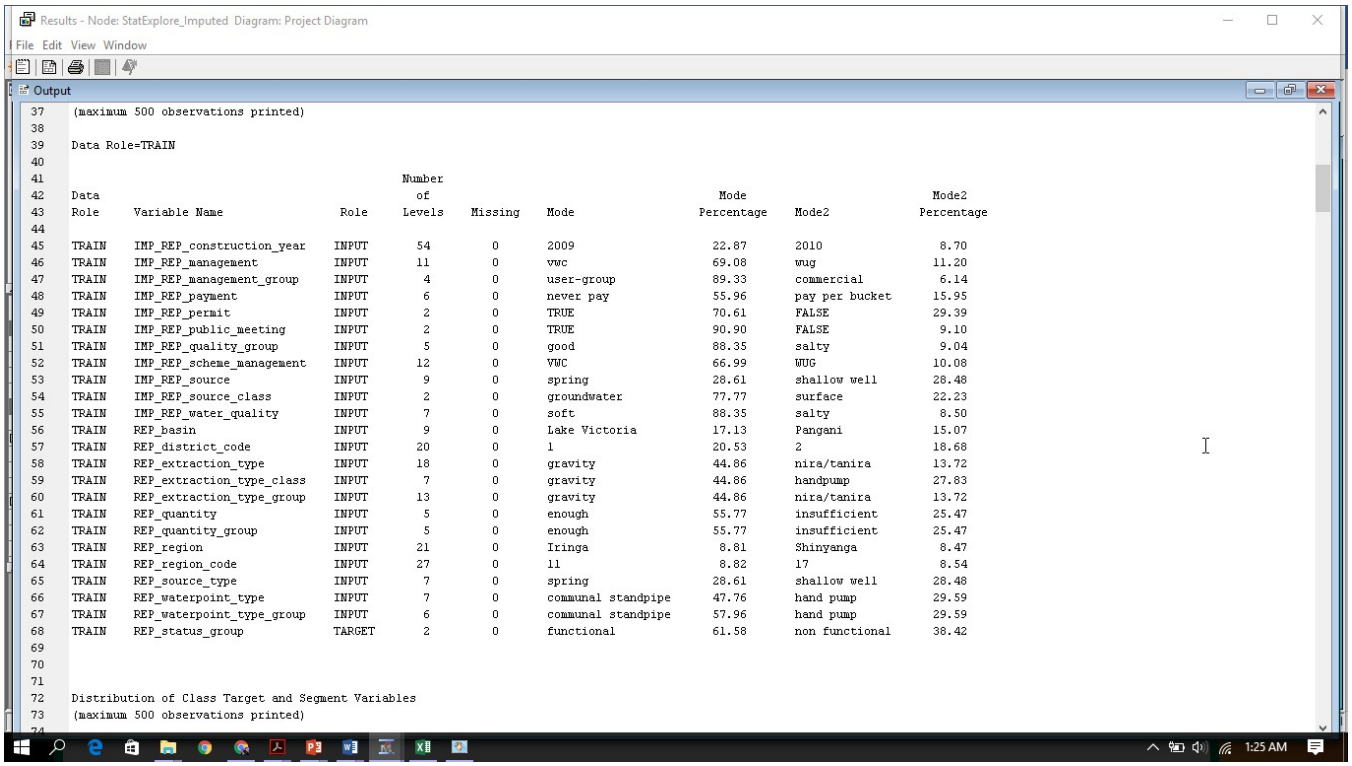




8. Stat Explore_Imputed:

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

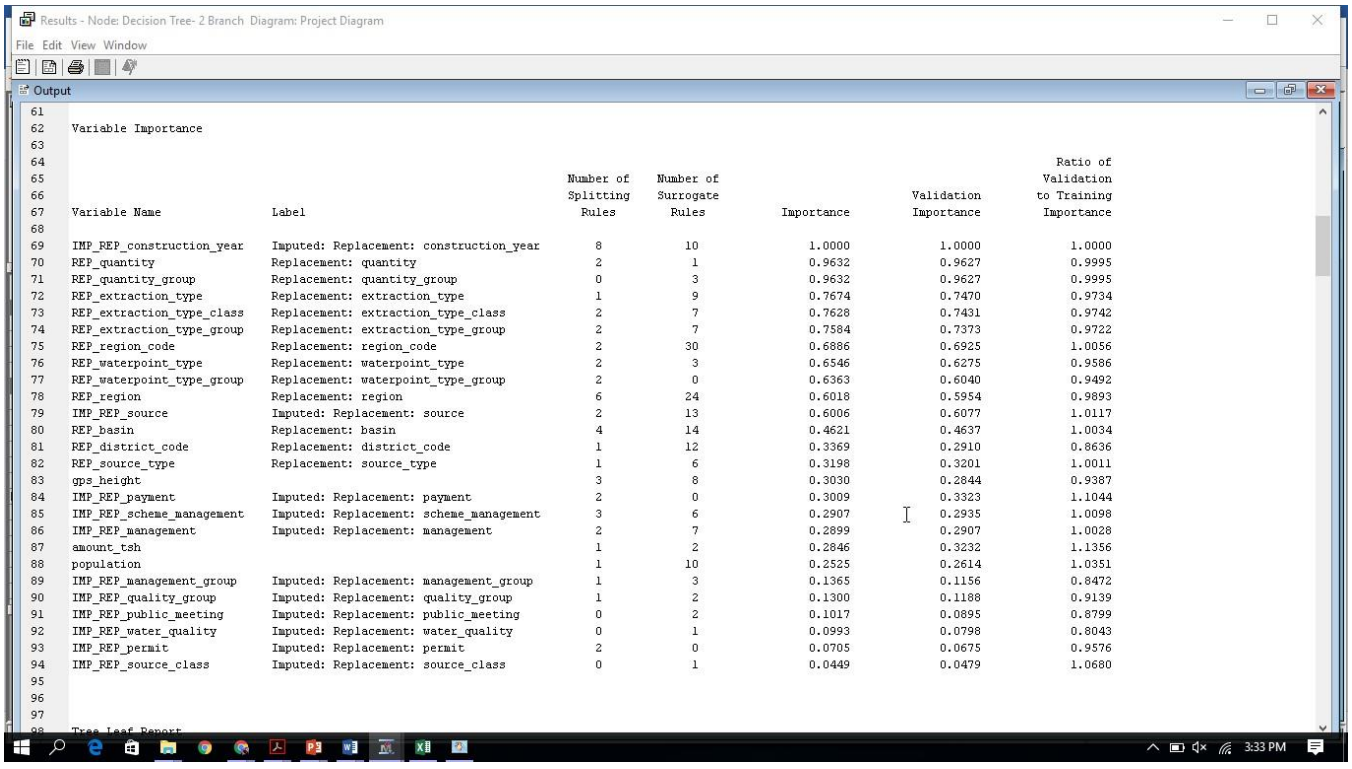




9. Decision Tree- 2 Branch:

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

Property	Value
Splitting Rule	
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
Node	
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

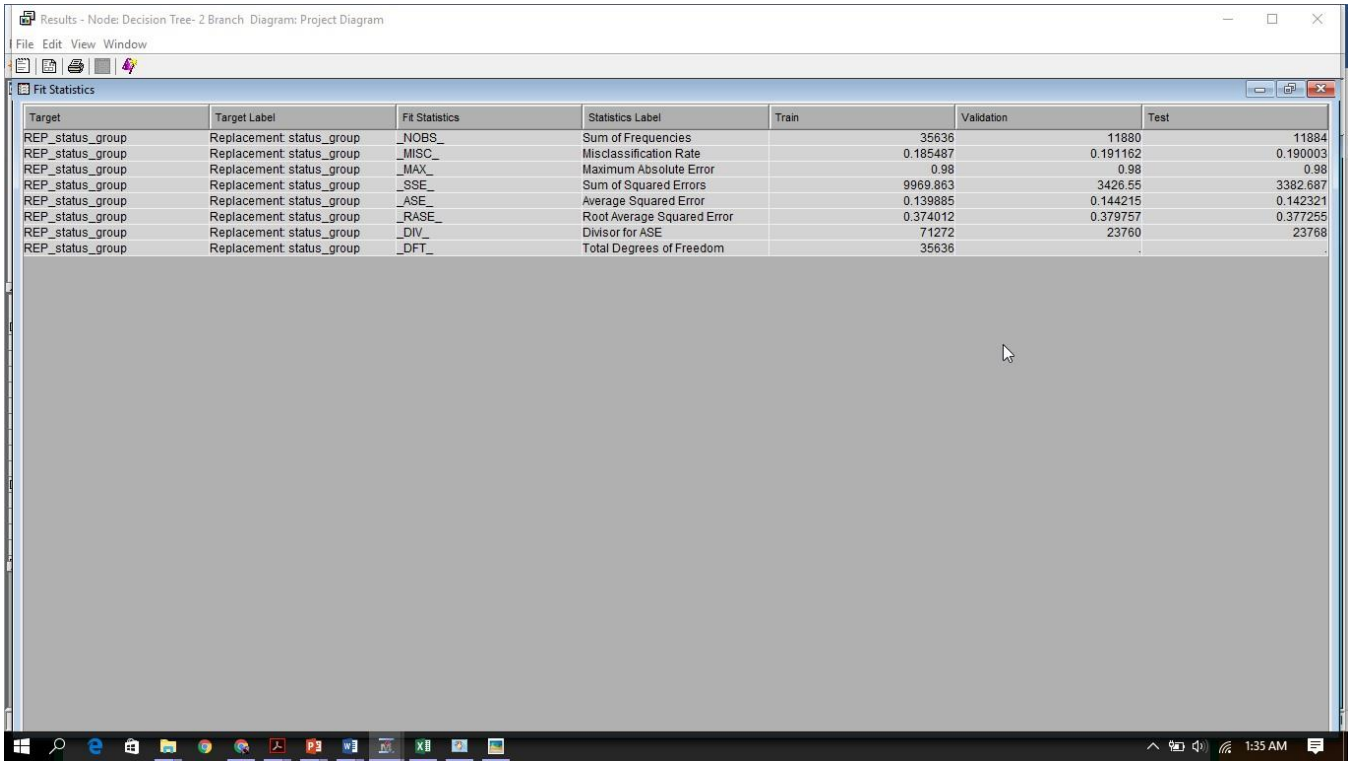
Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_REP_construction_year	Imputed: Replacement: construction_year	8	10	1.0000	1.0000	1.0000
REP_quantity	Replacement: quantity	2	1	0.9632	0.9627	0.9995
REP_quantity_group	Replacement: quantity_group	0	3	0.9632	0.9627	0.9995
REP_extraction_type	Replacement: extraction_type	1	9	0.7674	0.7470	0.9734
REP_extraction_type_class	Replacement: extraction_type_class	2	7	0.7628	0.7431	0.9742
REP_extraction_type_group	Replacement: extraction_type_group	2	7	0.7584	0.7373	0.9722
REP_region_code	Replacement: region_code	2	30	0.6886	0.6925	1.0056
REP_waterpoint_type	Replacement: waterpoint_type	2	3	0.6546	0.6275	0.9586
REP_waterpoint_type_group	Replacement: waterpoint_type_group	2	0	0.6363	0.6040	0.9492
REP_region	Replacement: region	6	24	0.6018	0.5954	0.9893
IMP_REP_source	Imputed: Replacement: source	2	13	0.6006	0.6077	1.0117
REP_basin	Replacement: basin	4	14	0.4621	0.4637	1.0034
REP_district_code	Replacement: district_code	1	12	0.3369	0.2910	0.8636
REP_source_type	Replacement: source_type	1	6	0.3198	0.3201	1.0011
gps_height		3	8	0.3030	0.2844	0.9387
IMP_REP_payment	Imputed: Replacement: payment	2	0	0.3009	0.3323	1.1044
IMP_REP_scheme_management	Imputed: Replacement: scheme_management	3	6	0.2907	0.2935	1.0098
IMP_REP_management	Imputed: Replacement: management	2	7	0.2899	0.2907	1.0028
amount_tsh		1	2	0.2846	0.3232	1.1356
population		1	10	0.2525	0.2614	1.0351
IMP_REP_management_group	Imputed: Replacement: management_group	1	3	0.1365	0.1156	0.8472
IMP_REP_quality_group	Imputed: Replacement: quality_group	1	2	0.1300	0.1188	0.9139
IMP_REP_public_meeting	Imputed: Replacement: public_meeting	0	2	0.1017	0.0895	0.8799
IMP_REP_water_quality	Imputed: Replacement: water_quality	0	1	0.0993	0.0798	0.8043
IMP_REP_permit	Imputed: Replacement: permit	2	0	0.0705	0.0675	0.9576
IMP_REP_source_class	Imputed: Replacement: source_class	0	1	0.0449	0.0479	1.0680

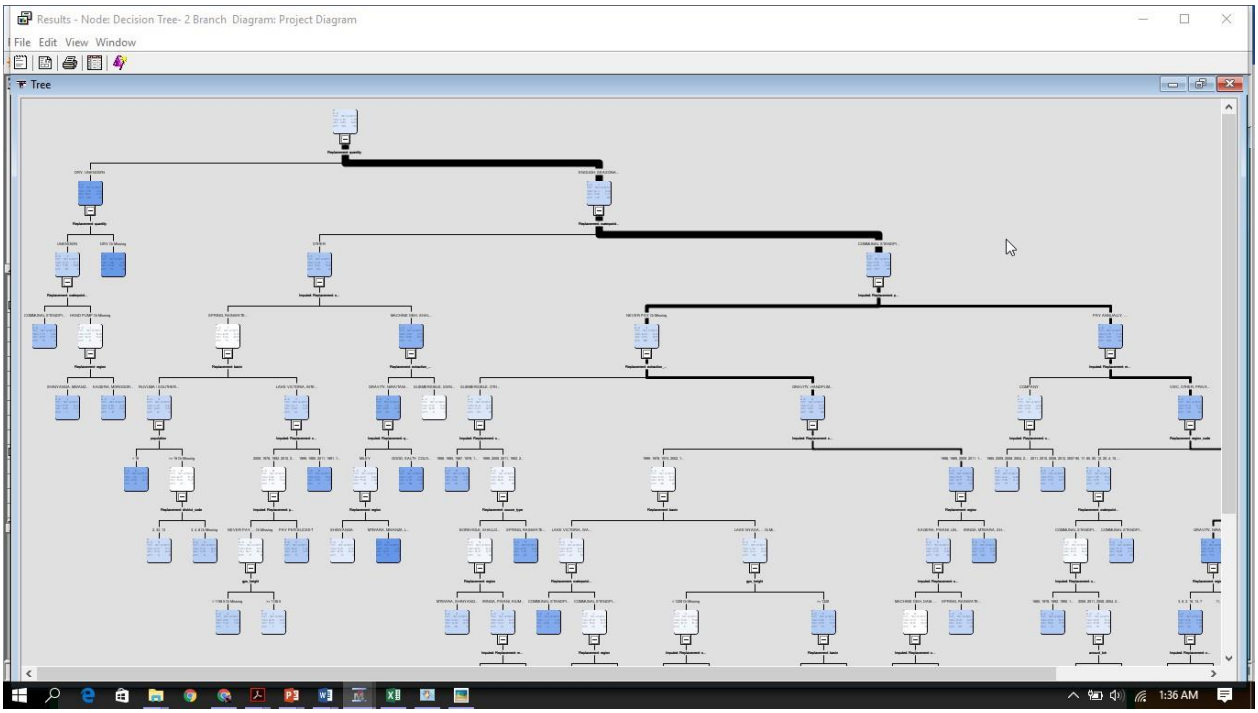
Tree Leaf Report

3:33 PM

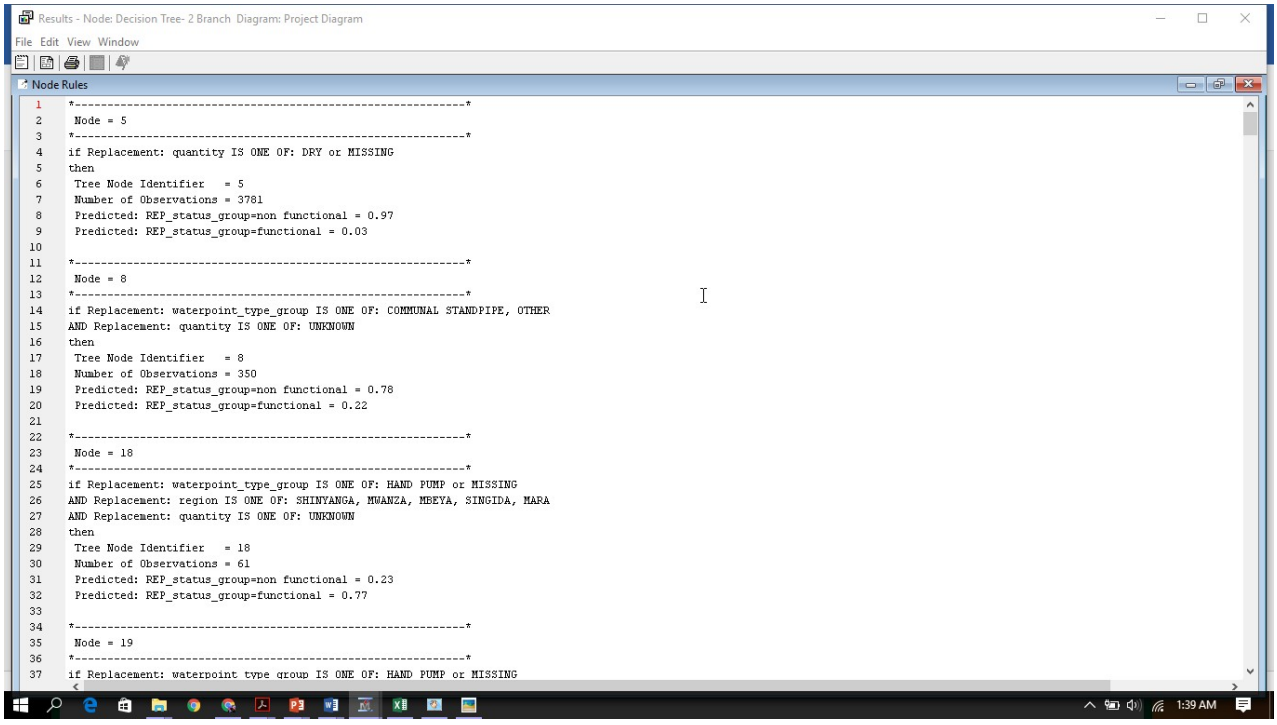
It can be seen that “Quantity” has high variable importance (splitting rules= 2) than the other variables to determine the “Status_Group”. It has a variable importance of 0.9627.



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.

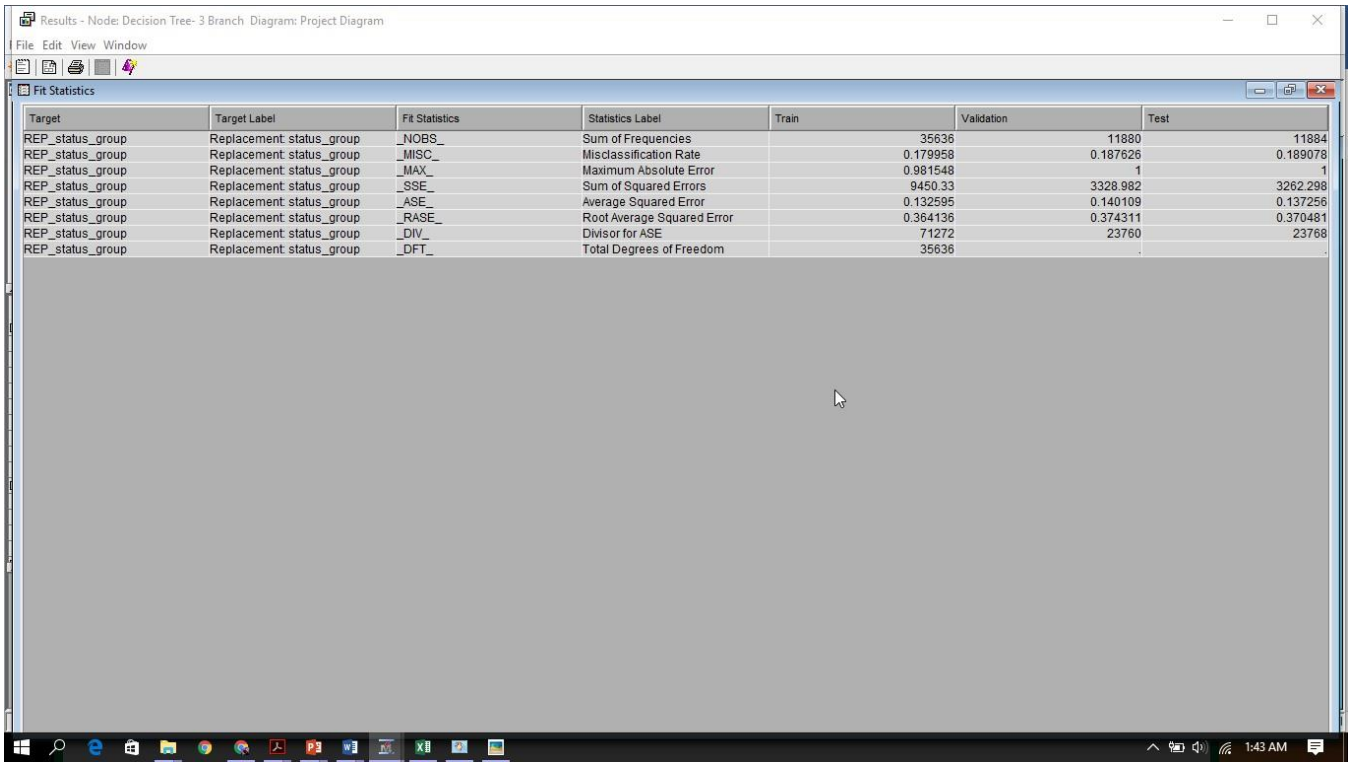


The misclassification rate for the above model is 0.1911.

10. Decision Tree- 3 Branch:

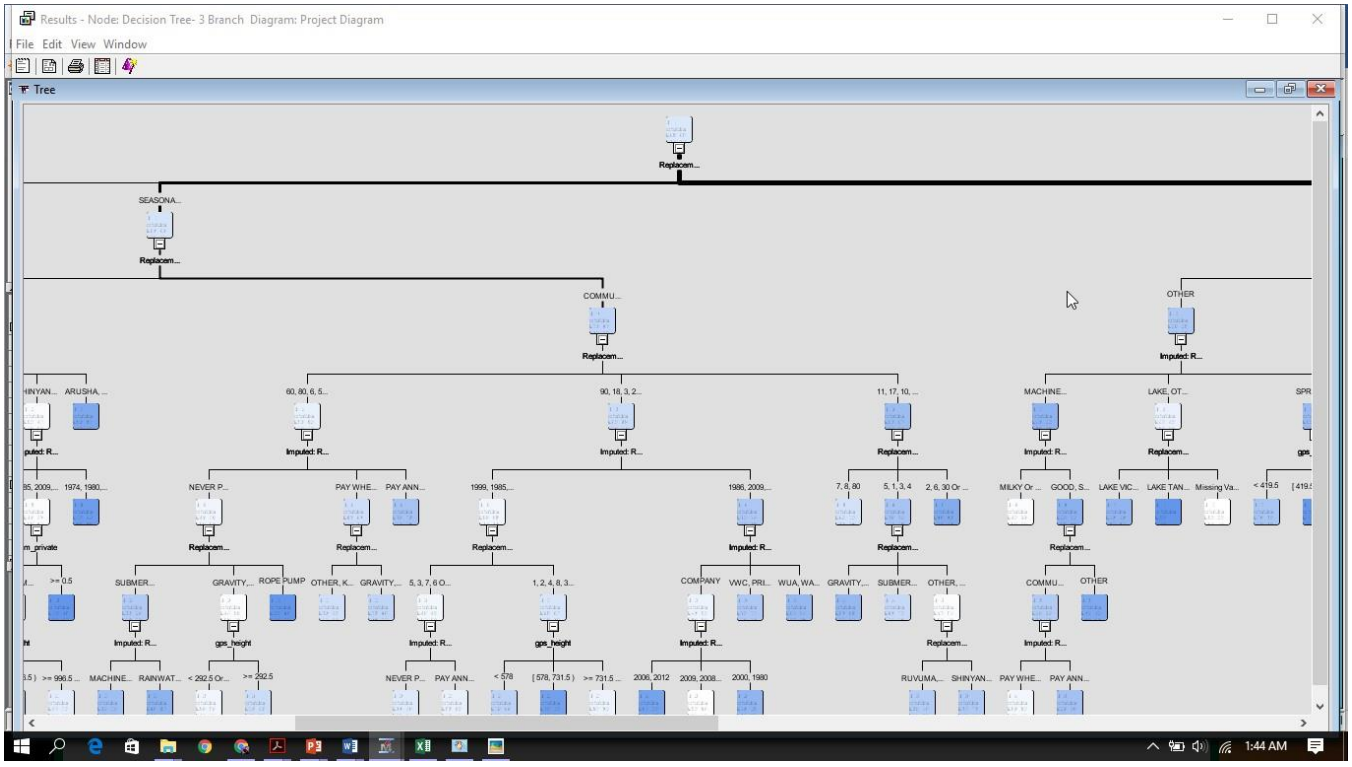
This tree was used to predict the ***status_group*** target variable from the imputed data with the following properties:

Property	Value
Splitting Rule	
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
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

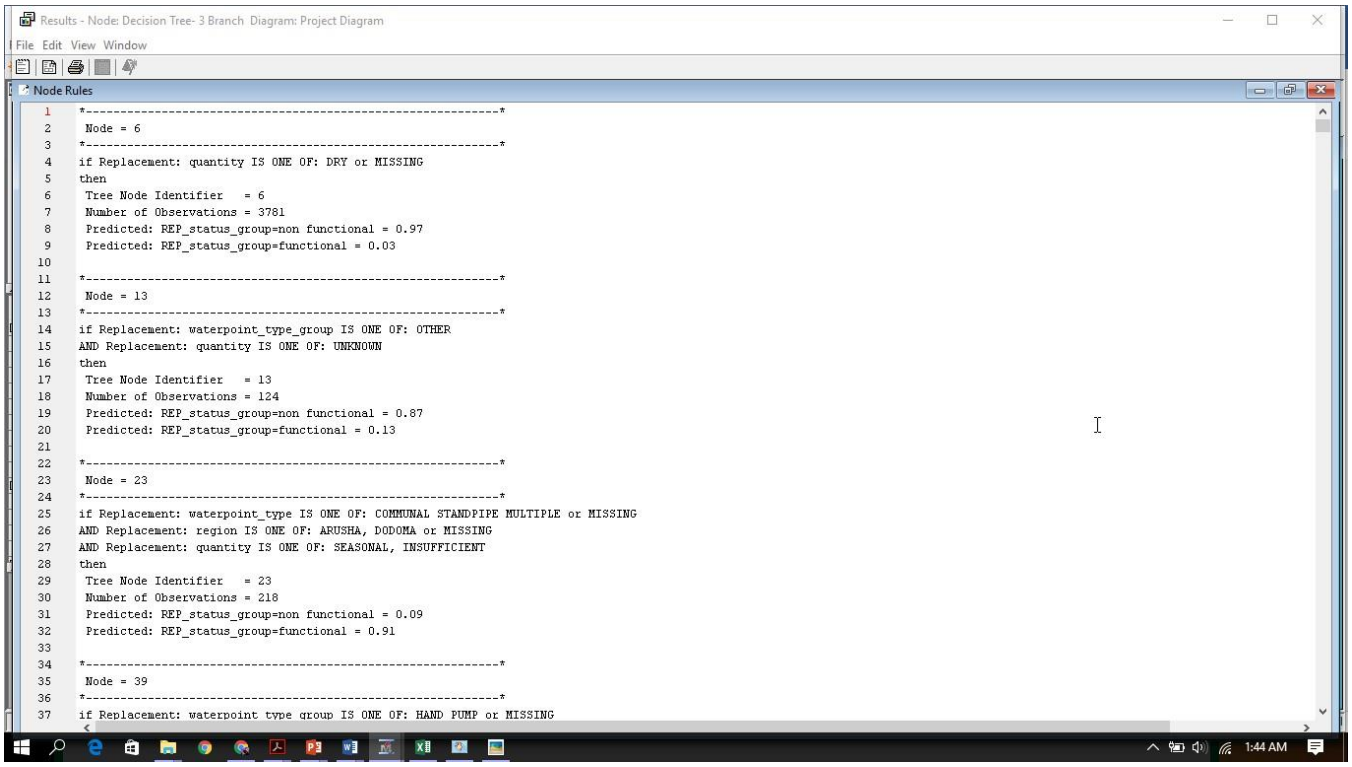


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_status_group	Replacement: status_group	_NOBS_	Sum of Frequencies	35636	11880	11884
REP_status_group	Replacement: status_group	_MISC_	Misclassification Rate	0.179958	0.187626	0.189078
REP_status_group	Replacement: status_group	_MAX_	Maximum Absolute Error	0.981548	1	1
REP_status_group	Replacement: status_group	_SSE_	Sum of Squared Errors	9450.33	3328.982	3262.298
REP_status_group	Replacement: status_group	_ASE_	Average Squared Error	0.132595	0.140109	0.137256
REP_status_group	Replacement: status_group	_RASE_	Root Average Squared Error	0.364136	0.374311	0.370481
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	.	.


Results - Node: Decision Tree- 3 Branch Diagram: Project Diagram
File Edit View Window
Score Rankings Overlay: Replacement: status_group
Cumulative Lift
Cumulative Lift
2.5
2.0
1.5
1.0
0
20
40
60
80
100
Depth
TRAIN
VALIDATE
1:43 AM


Results - Node: Decision Tree- 3 Branch Diagram: Project Diagram
File Edit View Window
Tree
1:44 AM

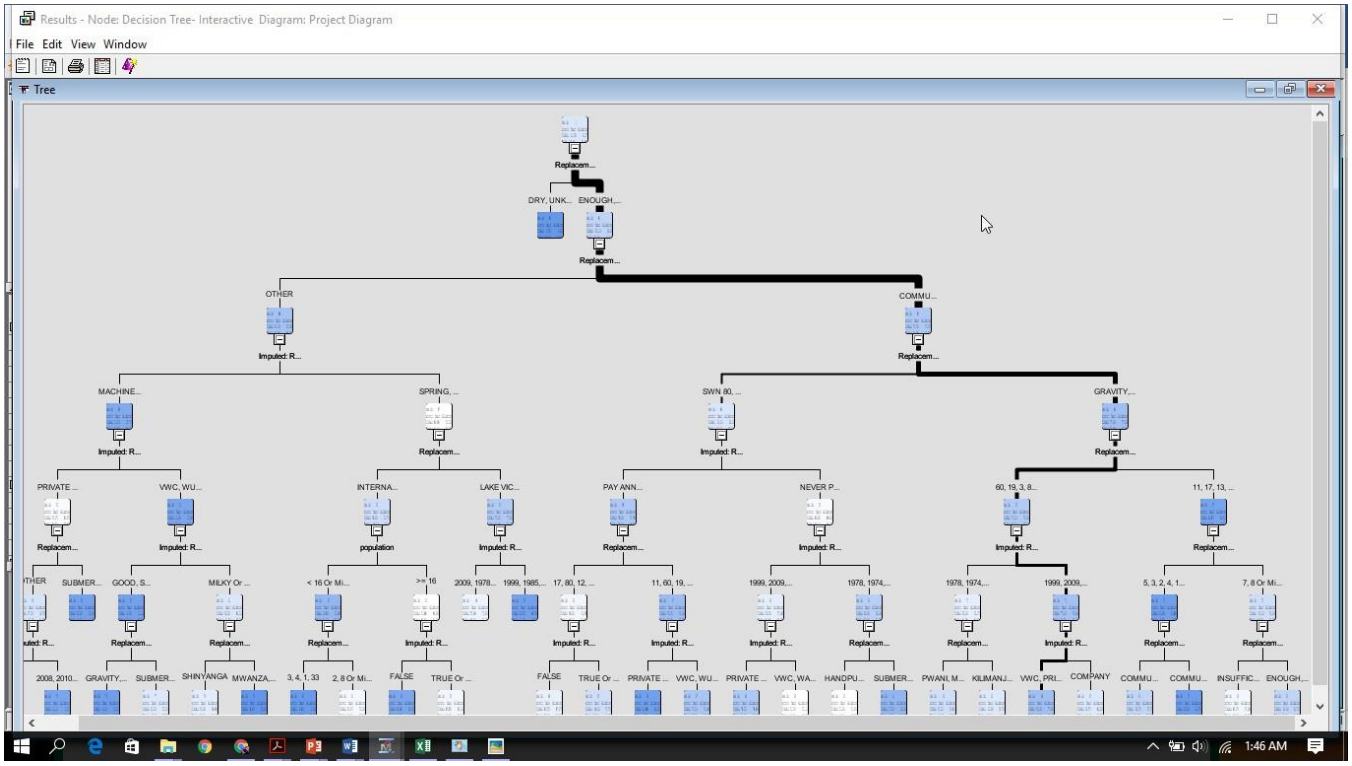
Below we can see the Node Rules window containing the IF-THEN logic that distributes observations into each leaf node of the decision tree.

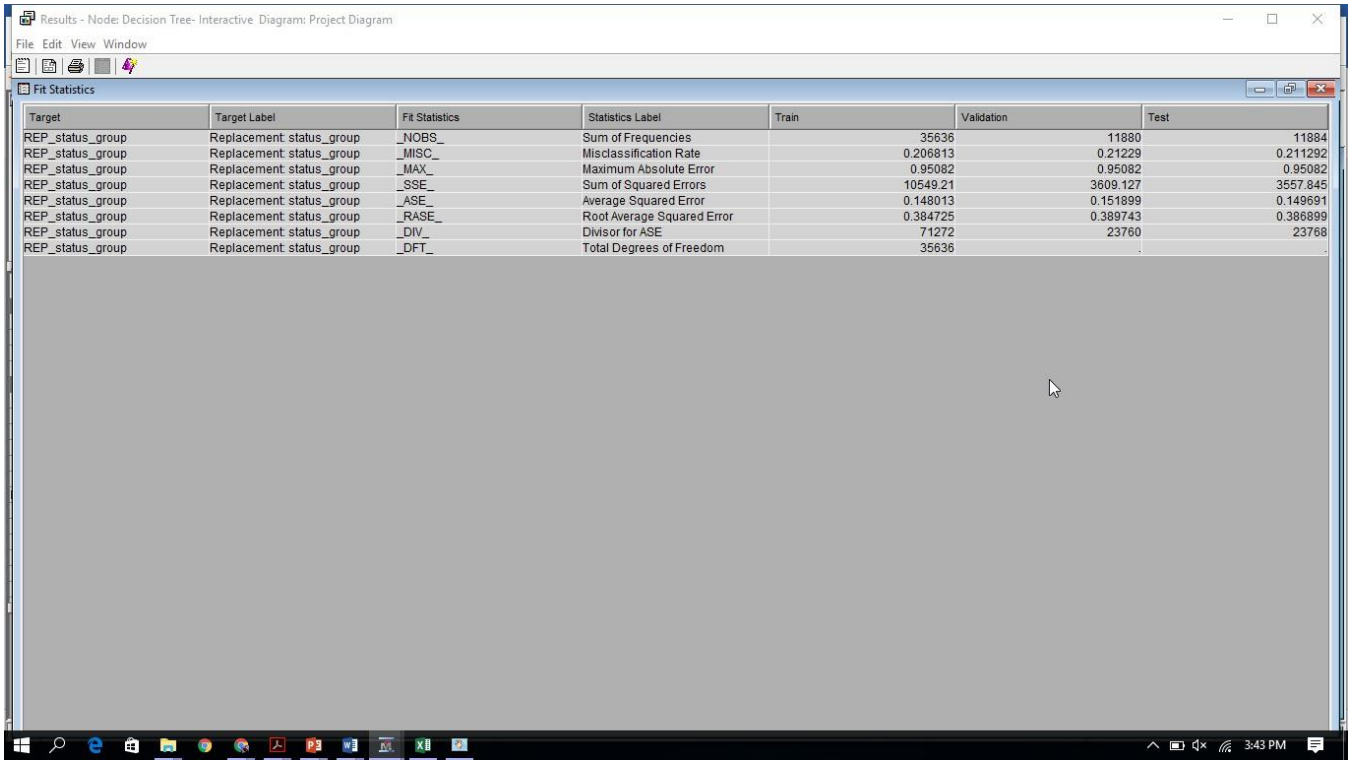


The misclassification rate for the above model is 0.1875.

11. Decision Tree- Interactive:

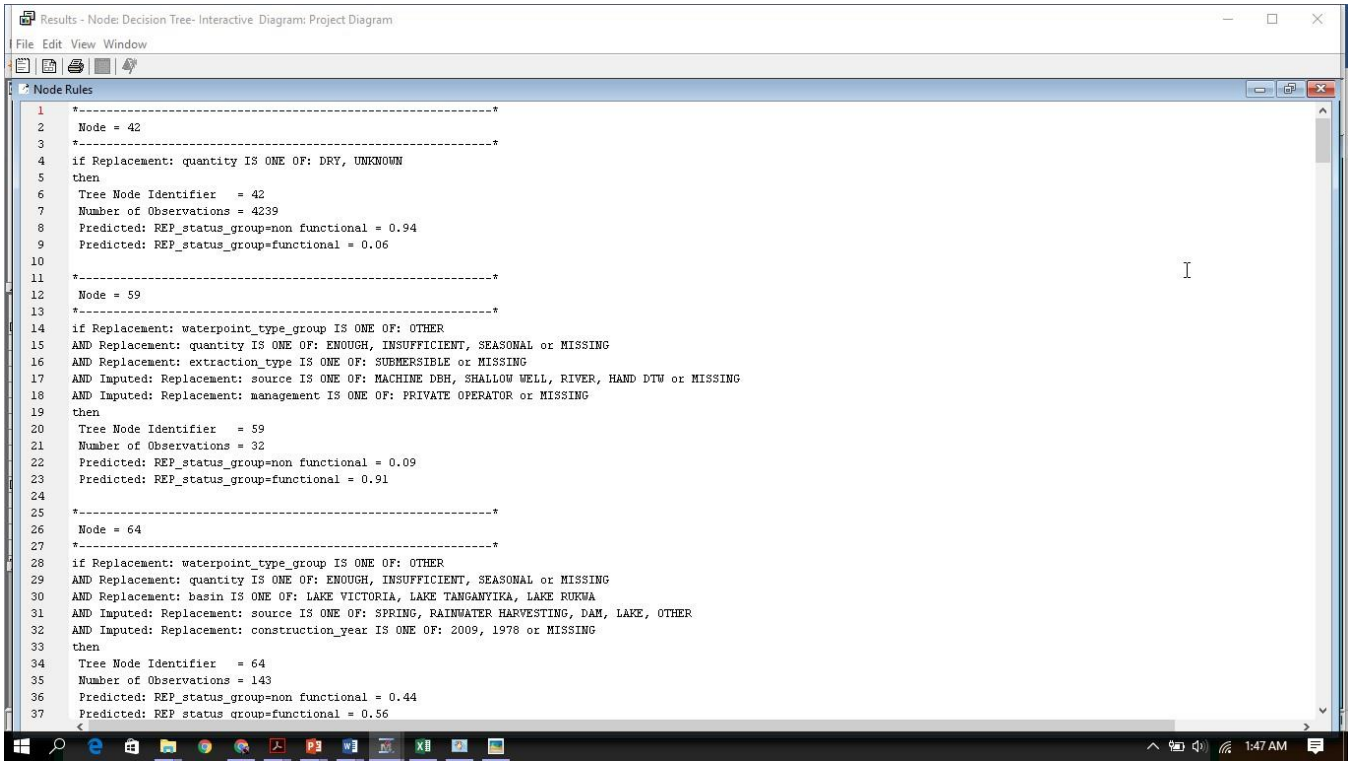
We create an interactive decision tree by deciding the split node based on the log worth of each variable and we choose the node with the highest log value. We then train a node to create the entire decision tree and we get the interactive decision tree. The interactive decision tree gives us an advantage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





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 misclassification rate for the above model is 0.2123.

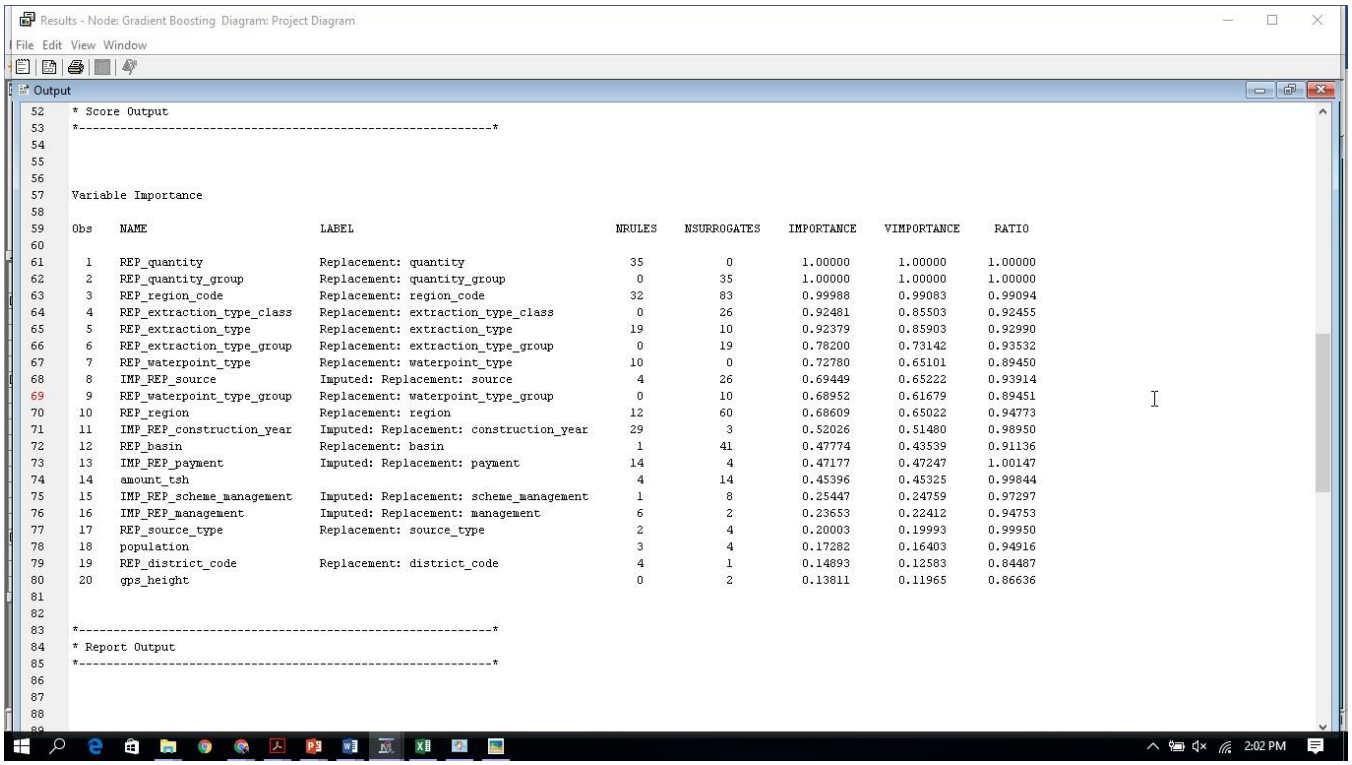
12. Gradient Boosting:

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
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
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
Node	
Leaf Fraction	0.1
Number of Surrogate Rules	2
Split Size	.

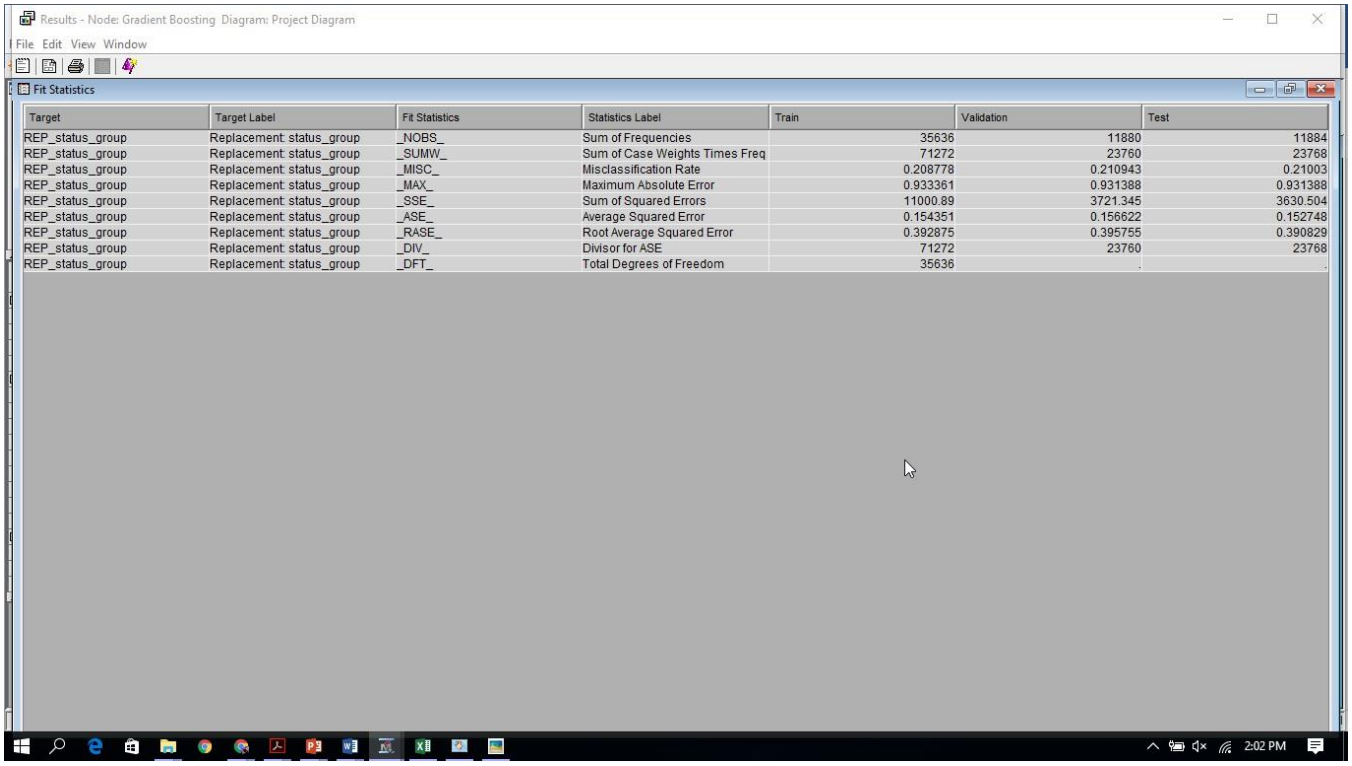


Results - Node: Gradient Boosting Diagram: Project Diagram

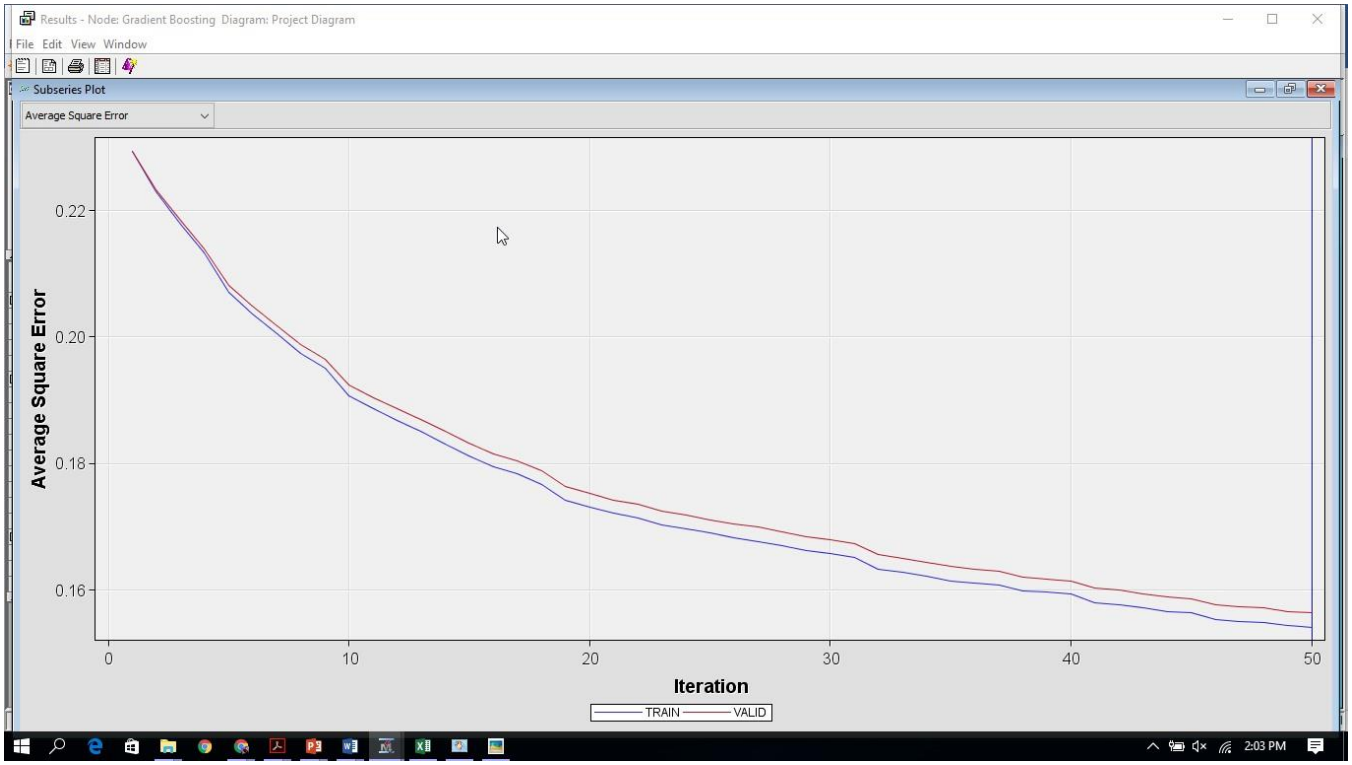
File Edit View Window

Output

```
* Score Output
*------------------------------------------------------------*


Variable Importance

Obs    NAME                        LABEL                                    NRULES    NSURROGATES    IMPORTANCE    VIMPORTANCE    RATIO

  1    REP_quantity                Replacement: quantity                        35           0         1.00000       1.00000      1.00000
  2    REP_quantity_group          Replacement: quantity_group                   0          35         1.00000       1.00000      1.00000
  3    REP_region_code             Replacement: region_code                     32          83         0.99988       0.99083      0.99094
  4    REP_extraction_type_class   Replacement: extraction_type_class            0          26         0.92481       0.85503      0.92455
  5    REP_extraction_type         Replacement: extraction_type                 19          10         0.92379       0.85903      0.92990
  6    REP_extraction_type_group   Replacement: extraction_type_group            0          19         0.78200       0.73142      0.93532
  7    REP_waterpoint_type         Replacement: waterpoint_type                 10           0         0.72780       0.65101      0.89450
  8    IMP_REP_source              Imputed: Replacement: source                  4          26         0.69449       0.65222      0.93914
  9    REP_waterpoint_type_group   Replacement: waterpoint_type_group            0          10         0.68952       0.61679      0.89451
 10    REP_region                  Replacement: region                          12          60         0.68609       0.65022      0.94773
 11    IMP_REP_construction_year   Imputed: Replacement: construction_year      29           3         0.52026       0.51480      0.98950
 12    REP_basin                   Replacement: basin                            1          41         0.47774       0.43539      0.91136
 13    IMP_REP_payment             Imputed: Replacement: payment                14           4         0.47177       0.47247      1.00147
 14    amount_tsh                                                               4          14         0.45396       0.45325      0.99844
 15    IMP_REP_scheme_management   Imputed: Replacement: scheme_management       1           8         0.25447       0.24759      0.97297
 16    IMP_REP_management          Imputed: Replacement: management              6           2         0.23653       0.22412      0.94753
 17    REP_source_type             Replacement: source_type                      2           4         0.20003       0.19993      0.99950
 18    population                                                               3           4         0.17282       0.16403      0.94916
 19    REP_district_code           Replacement: district_code                    4           1         0.14893       0.12583      0.84487
 20    gps_height                                                               0           2         0.13811       0.11965      0.86636


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```



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	_SUMW_	Sum of Case Weights Times Freq	71272	23760	23768
REP_status_group	Replacement: status_group	_MISC_	Misclassification Rate	0.208778	0.210943	0.21003
REP_status_group	Replacement: status_group	_MAX_	Maximum Absolute Error	0.933361	0.931388	0.931388
REP_status_group	Replacement: status_group	_SSE_	Sum of Squared Errors	11000.89	3721.345	3630.504
REP_status_group	Replacement: status_group	_ASE_	Average Squared Error	0.154351	0.156622	0.152748
REP_status_group	Replacement: status_group	_RASE_	Root Average Squared Error	0.392875	0.395755	0.390829
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	.	.



The misclassification rate for the above model is 0.2109.

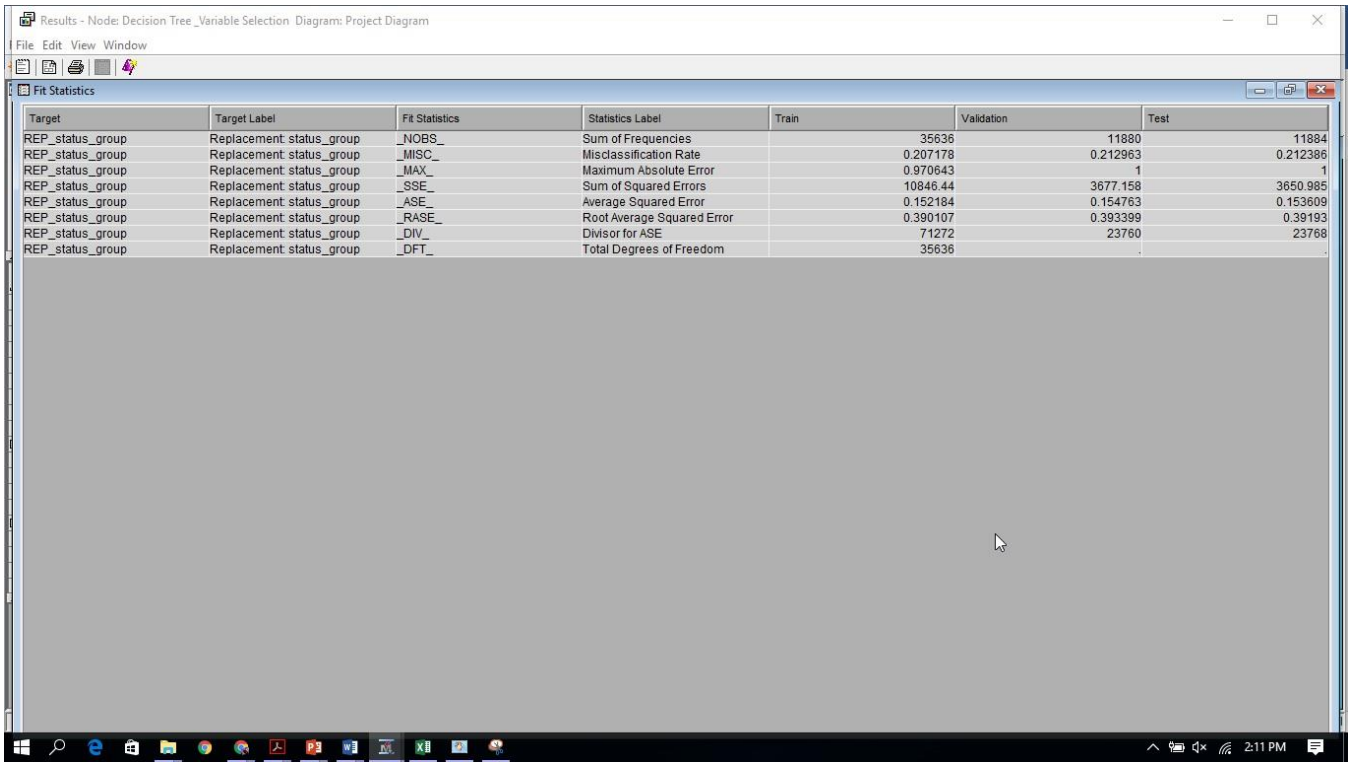
13. Variable Selection:

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

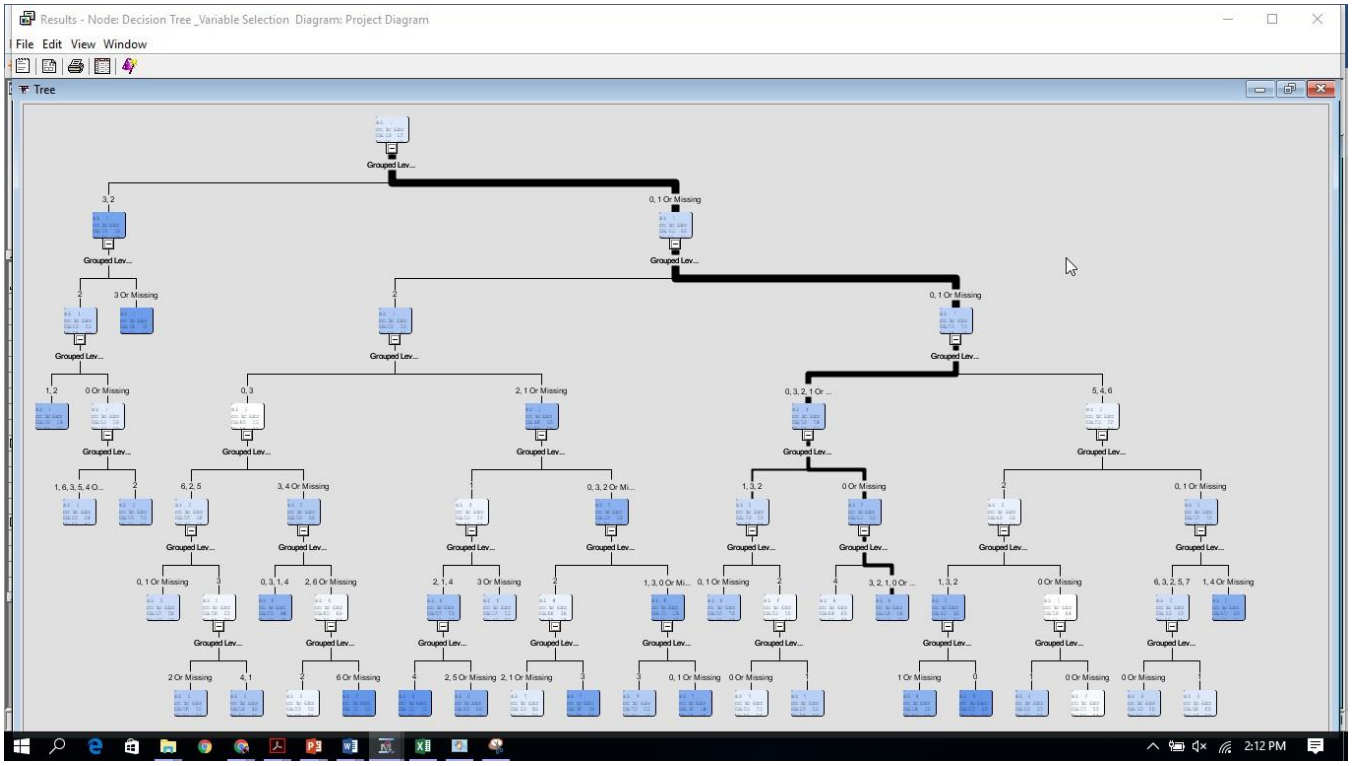
Property	Value
Train	
Variables	...
Max Class Level	200
Max Missing Percentage	50
Target Model	Default
Manual Selector	...
Rejects Unused Input	Yes
Bypass Options	
Variable	None
Role	Input
Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
R-Square Options	
Maximum Variable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

14. Decision Tree_ Variable Selection:

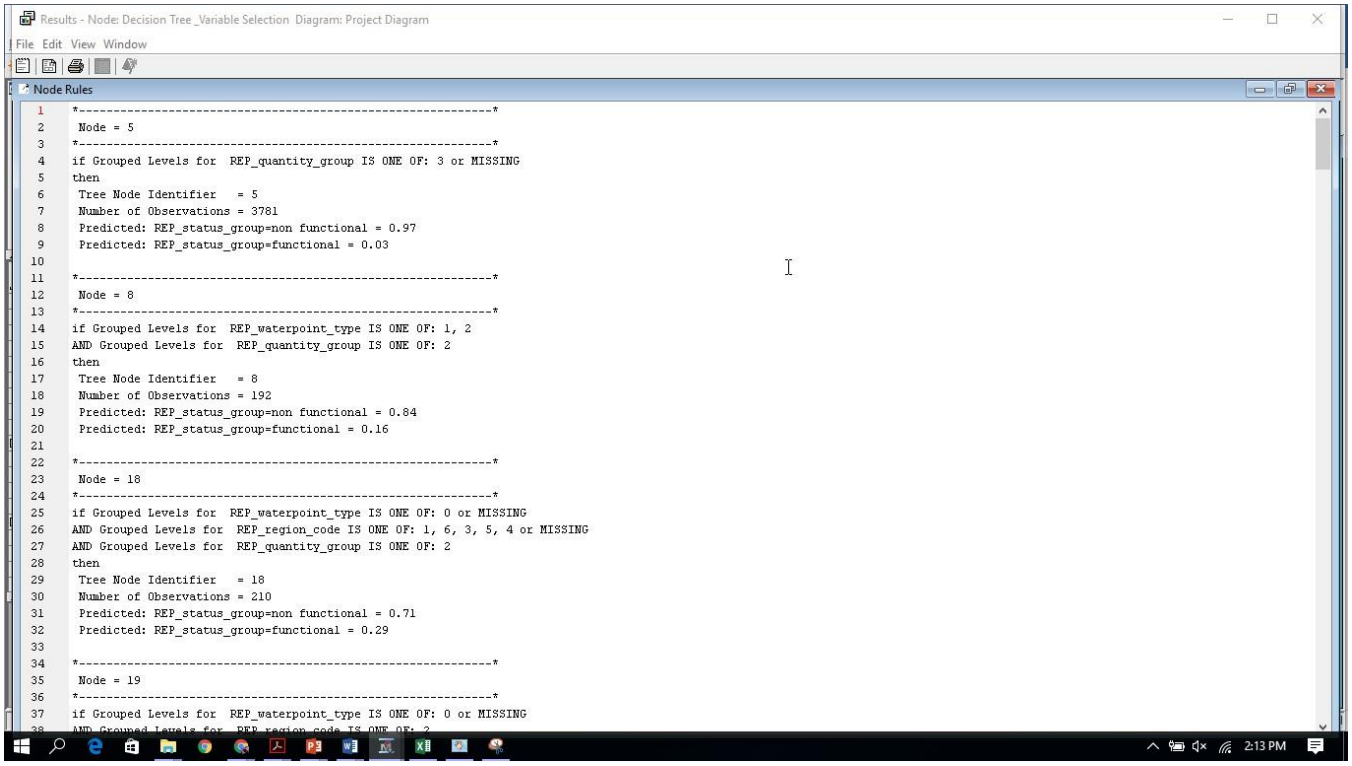
This node was used to predict the ***status_group*** target variable from the selected variables.



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	0.207178	0.212963	0.212386
REP_status_group	Replacement: status_group	_MAX_	Maximum Absolute Error	0.970643	1	1
REP_status_group	Replacement: status_group	_SSE_	Sum of Squared Errors	10846.44	3677.158	3650.985
REP_status_group	Replacement: status_group	_ASE_	Average Squared Error	0.152184	0.154763	0.153609
REP_status_group	Replacement: status_group	_RASE_	Root Average Squared Error	0.390107	0.393399	0.39193
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.

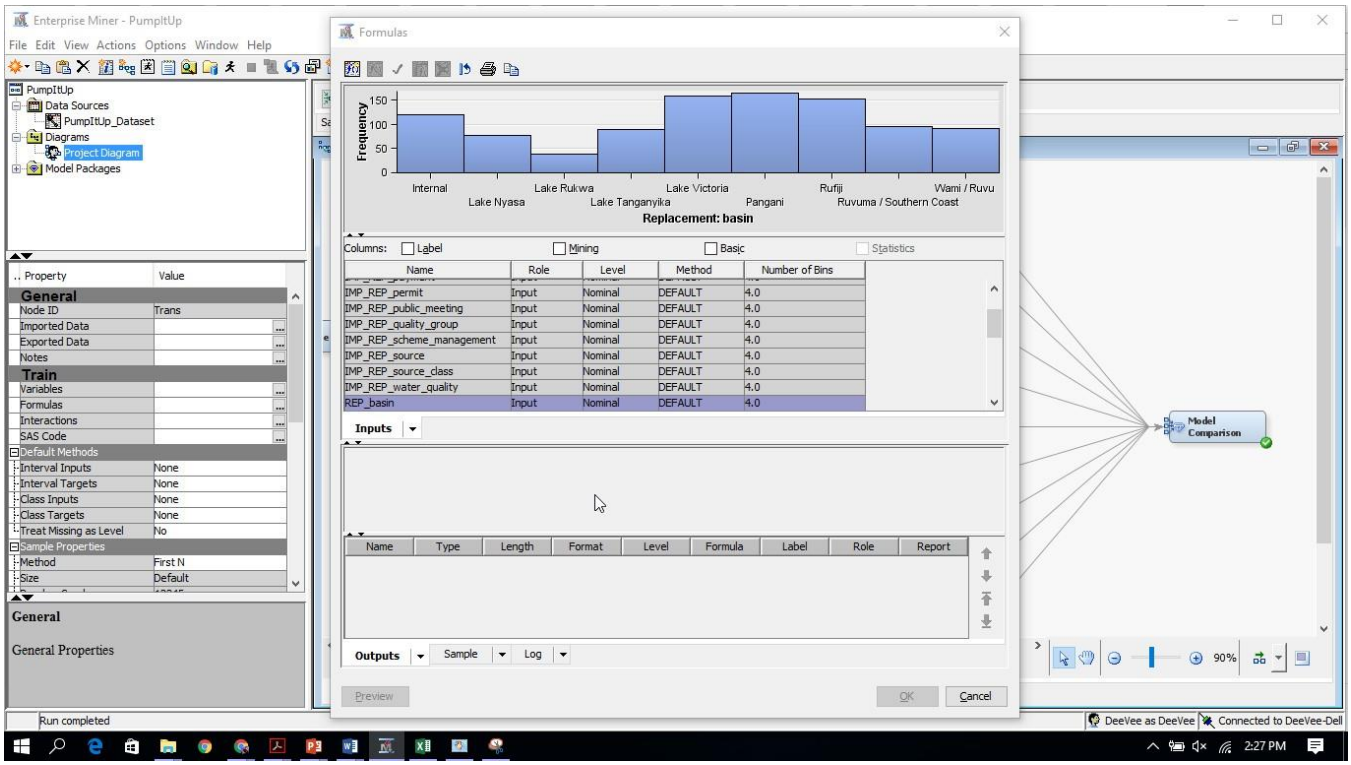


The misclassification rate for the above model is 0.2129.

15. Transform Variables:

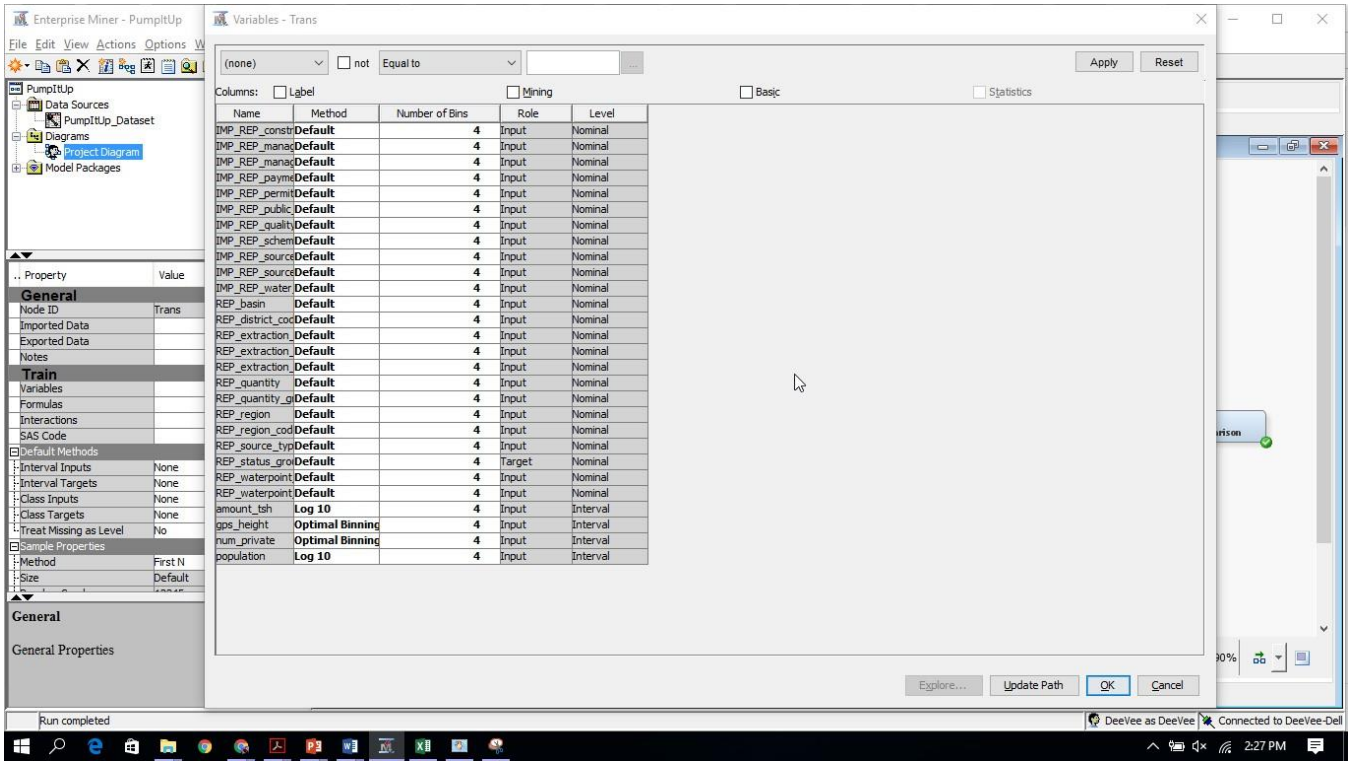
Transforming the data can improve model response. Transforming the data tends to stabilize variance, remove nonlinearity, improve additivity, and counter non- normality. For many models, transforming input data leads to better model fits. These transformations can be a function of one or more variables.

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.

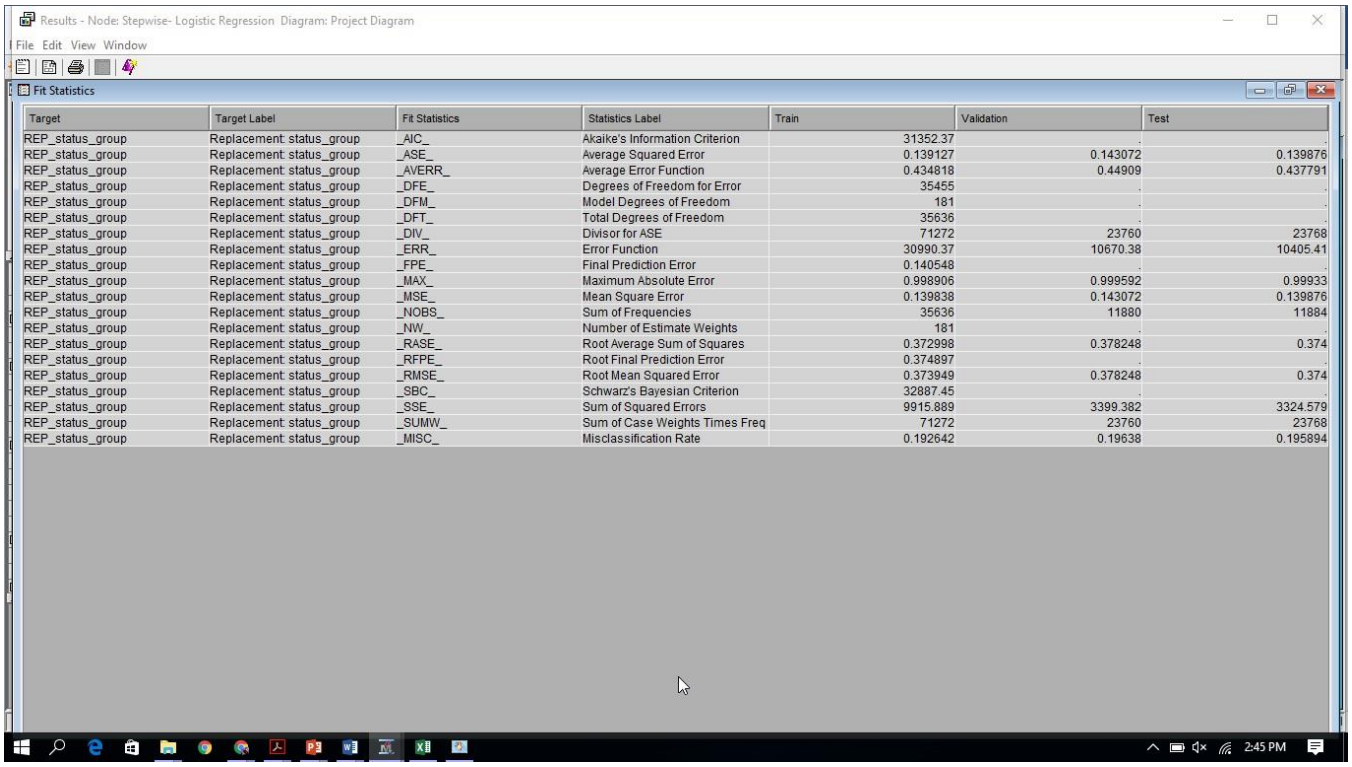
We have chosen log10 method for ***amount_tsh and population*** and Optimal Binning method for ***gps_height and num_private.***



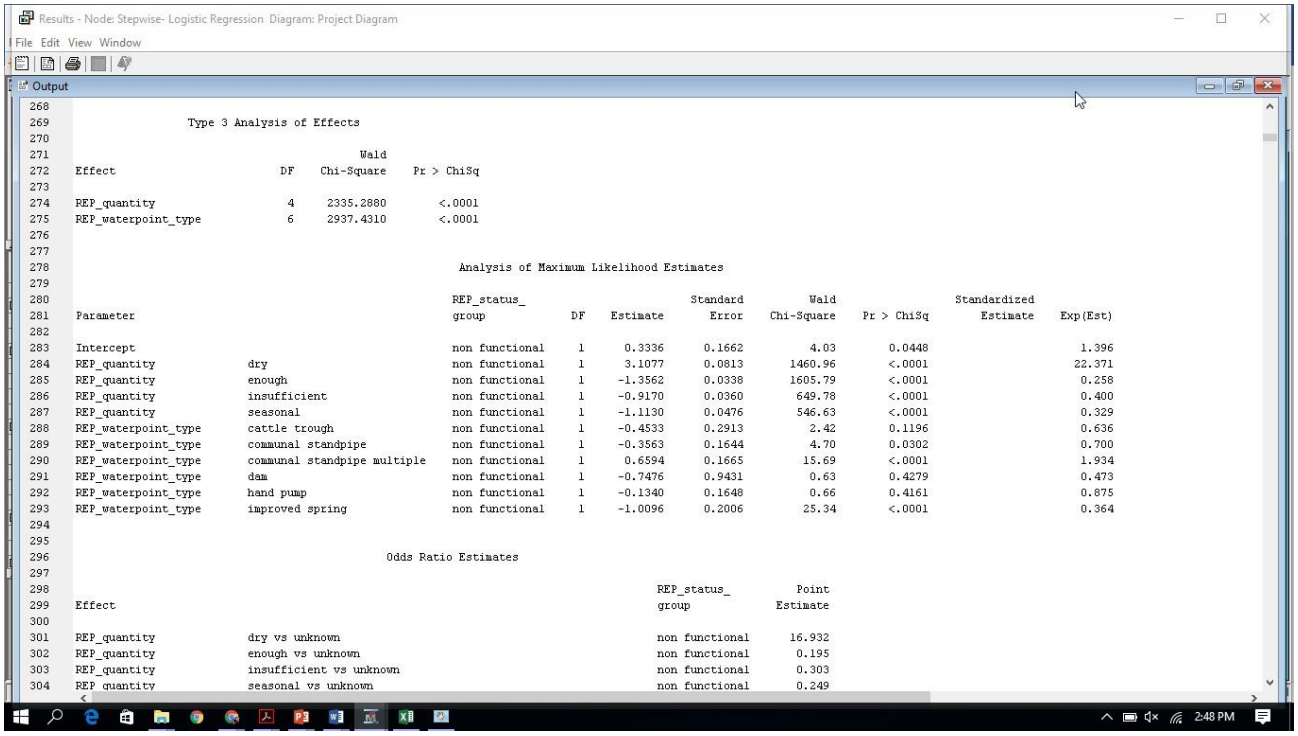
16. Stepwise – Logistic Regression:

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	



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	0.998906	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	32887.45	.	.
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

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
REP_quantity	4	2335.2880	<.0001
REP_waterpoint_type	6	2937.4310	<.0001

Analysis of Maximum Likelihood Estimates

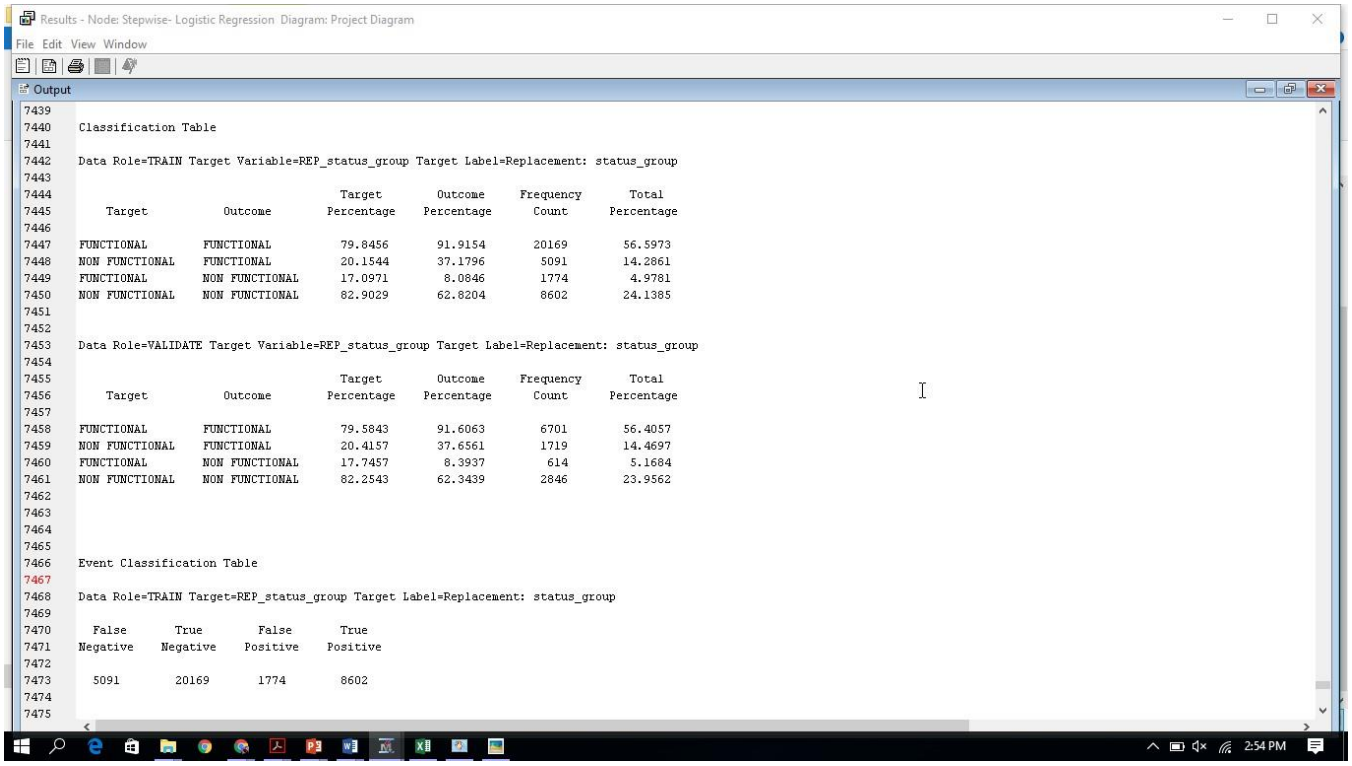
Parameter		REP_status_ group	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		non functional	1	0.3336	0.1662	4.03	0.0448		1.396
REP_quantity	dry	non functional	1	3.1077	0.0813	1460.96	<.0001		22.371
REP_quantity	enough	non functional	1	-1.3562	0.0338	1605.79	<.0001		0.258
REP_quantity	insufficient	non functional	1	-0.9170	0.0360	649.78	<.0001		0.400
REP_quantity	seasonal	non functional	1	-1.1130	0.0476	546.63	<.0001		0.329
REP_waterpoint_type	cattle trough	non functional	1	-0.4533	0.2913	2.42	0.1196		0.636
REP_waterpoint_type	communal standpipe	non functional	1	-0.3563	0.1644	4.70	0.0302		0.700
REP_waterpoint_type	communal standpipe multiple	non functional	1	0.6594	0.1665	15.69	<.0001		1.934
REP_waterpoint_type	dam	non functional	1	-0.7476	0.9431	0.63	0.4279		0.473
REP_waterpoint_type	hand pump	non functional	1	-0.1340	0.1648	0.66	0.4161		0.875
REP_waterpoint_type	improved spring	non functional	1	-1.0096	0.2006	25.34	<.0001		0.364

Odds Ratio Estimates

Effect		REP_status_ group	Point Estimate
REP_quantity	dry vs unknown	non functional	16.932
REP_quantity	enough vs unknown	non functional	0.195
REP_quantity	insufficient vs unknown	non functional	0.303
REP_quantity	seasonal vs unknown	non functional	0.249

2:48 PM

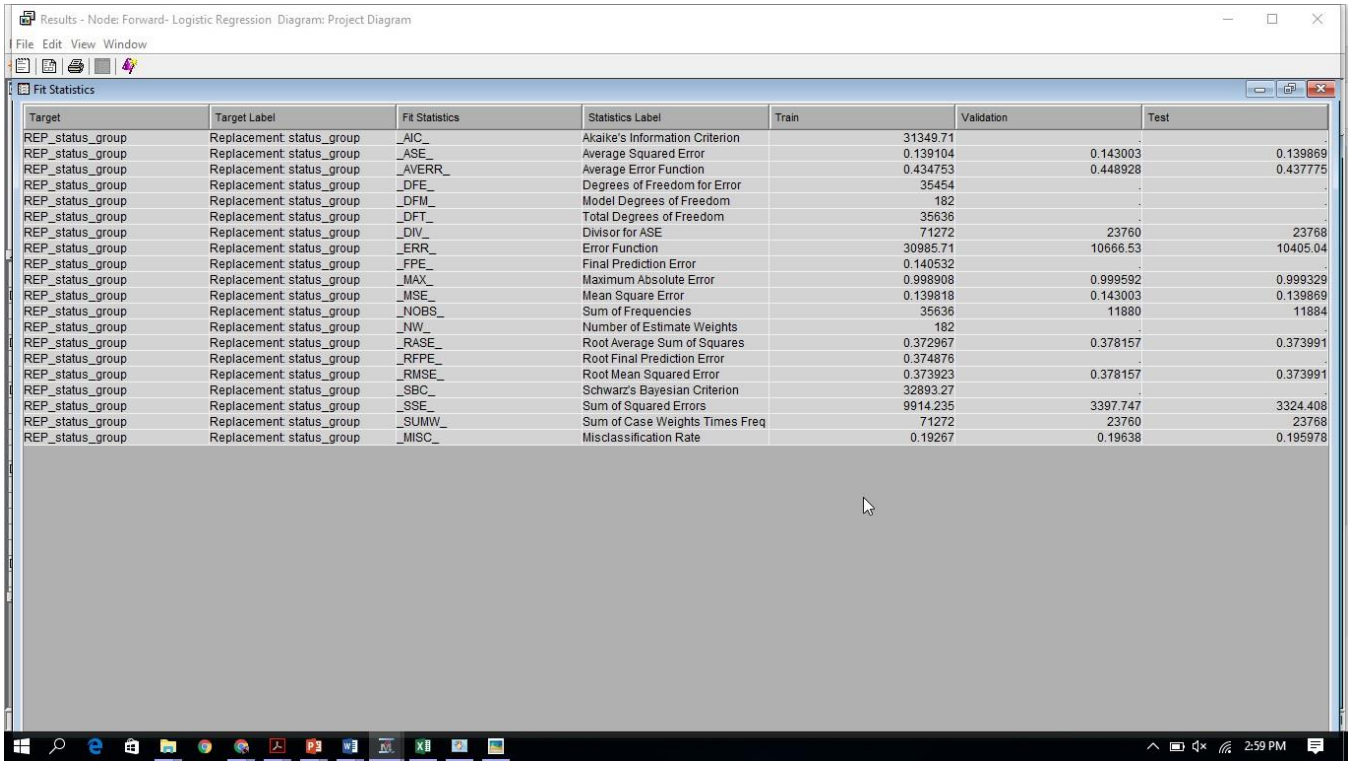
From the above result window, we can say that the parameters with chi- squared value is less than 0.05 are highly significant. For example, if the quantity of water is Dry, enough or Insufficient then the status group is likely to be non- functional and similarly so on.



The misclassification rate for the above model is 0.1963.

17. Forward- Logistic Regression:

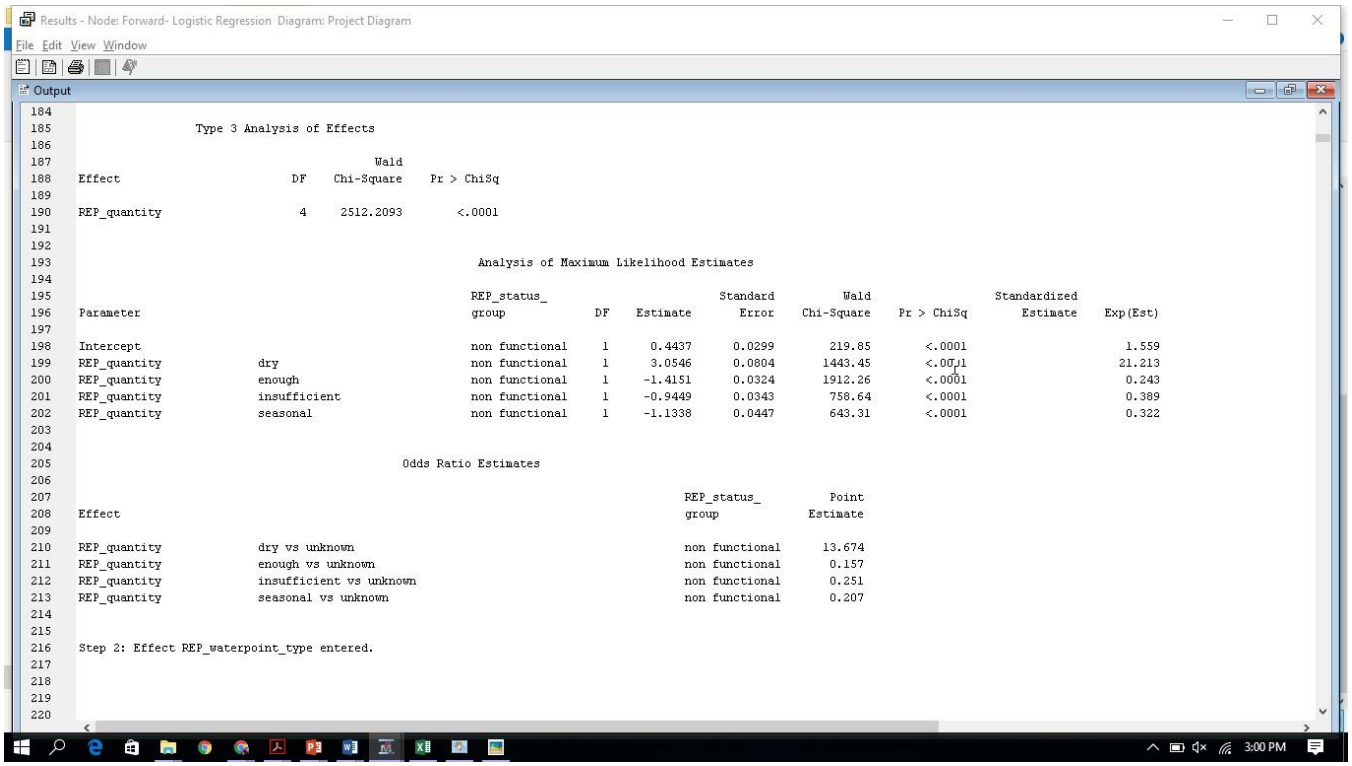
Property	Value
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default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


Results - Node: Forward- 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	31349.71	.	.
REP_status_group	Replacement: status_group	_ASE_	Average Squared Error	0.139104	0.143003	0.139869
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	0.139869
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

File Edit View Window

Output

Type 3 Analysis of Effects

Effect	DF	Wald 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	-1.1338	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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FUNCTIONAL	FUNCTIONAL	79.8448	91.9109	20168	56.5945
NON FUNCTIONAL	FUNCTIONAL	20.1552	37.1796	5091	14.2861
FUNCTIONAL	NON FUNCTIONAL	17.1051	8.0891	1775	4.9809
NON FUNCTIONAL	NON FUNCTIONAL	82.8949	62.8204	8602	24.1385

Data Role=VALIDATE Target Variable=REP_status_group Target Label=Replacement: status_group

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FUNCTIONAL	FUNCTIONAL	79.5914	91.5926	6700	56.3973
NON FUNCTIONAL	FUNCTIONAL	20.4086	37.6342	1718	14.4613
FUNCTIONAL	NON FUNCTIONAL	17.7643	8.4074	615	5.1768
NON FUNCTIONAL	NON FUNCTIONAL	82.2357	62.3658	2847	23.9646

Event Classification Table

Data Role=TRAIN Target=REP_status_group Target Label=Replacement: status_group

False Negative	True Negative	False Positive	True Positive
5091	20168	1775	8602

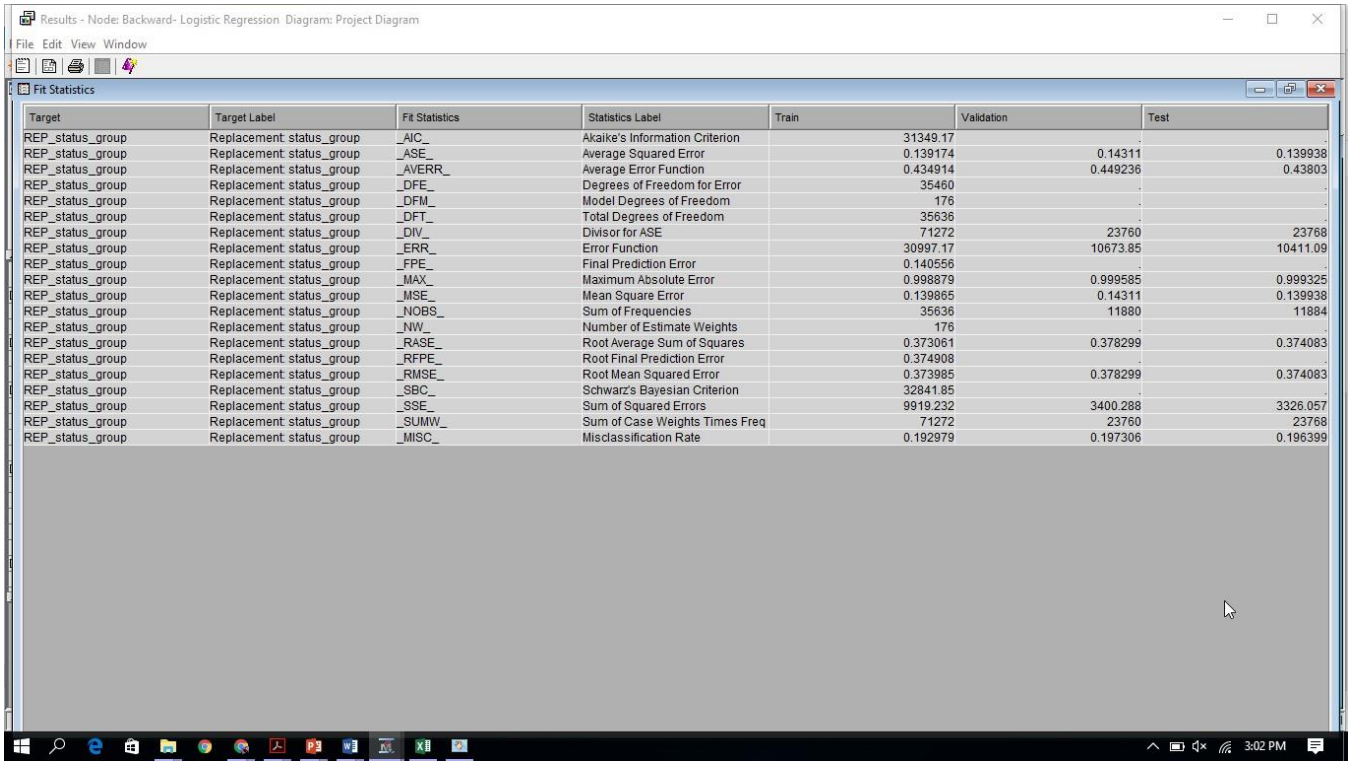
Data Role=VALIDATE Target=REP_status_group Target Label=Replacement: status_group

False Negative	True Negative	False Positive	True Positive

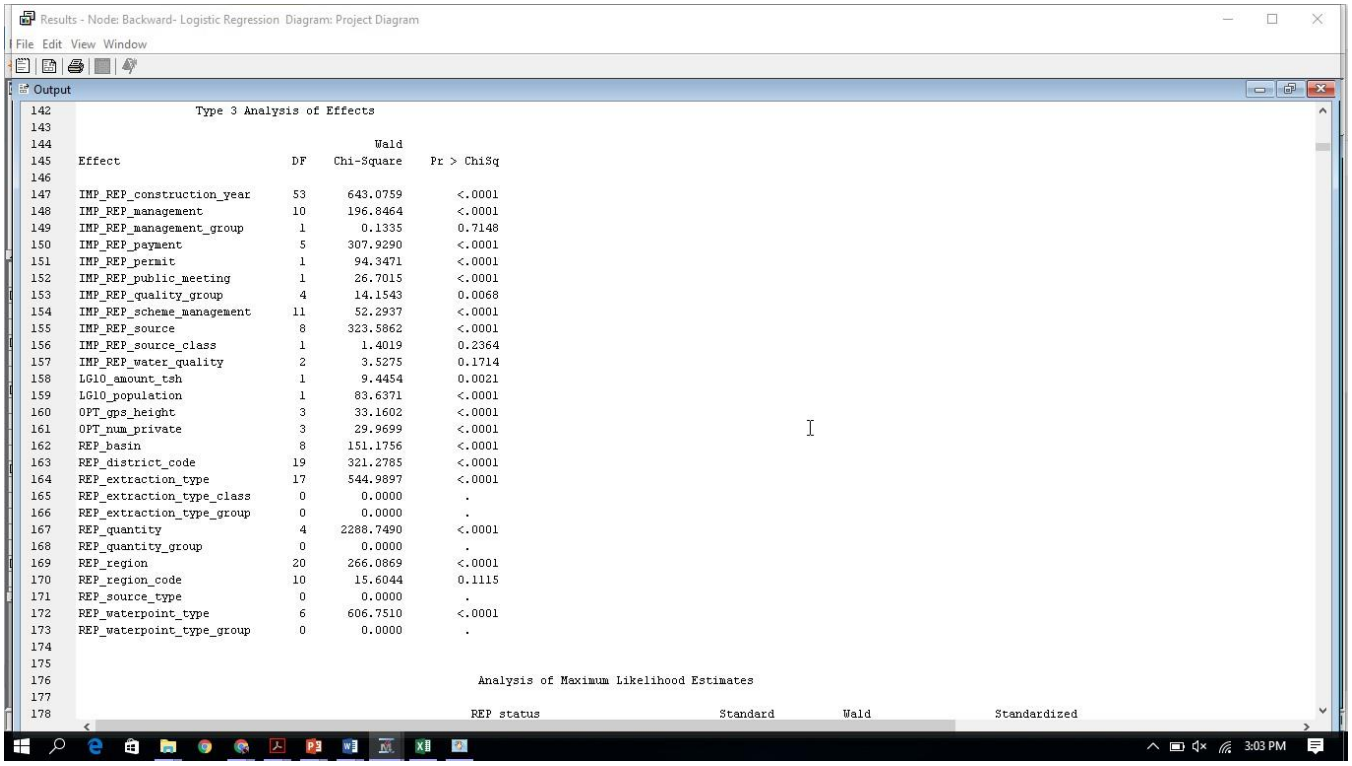
The misclassification rate for the above model is 0.1963.

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



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

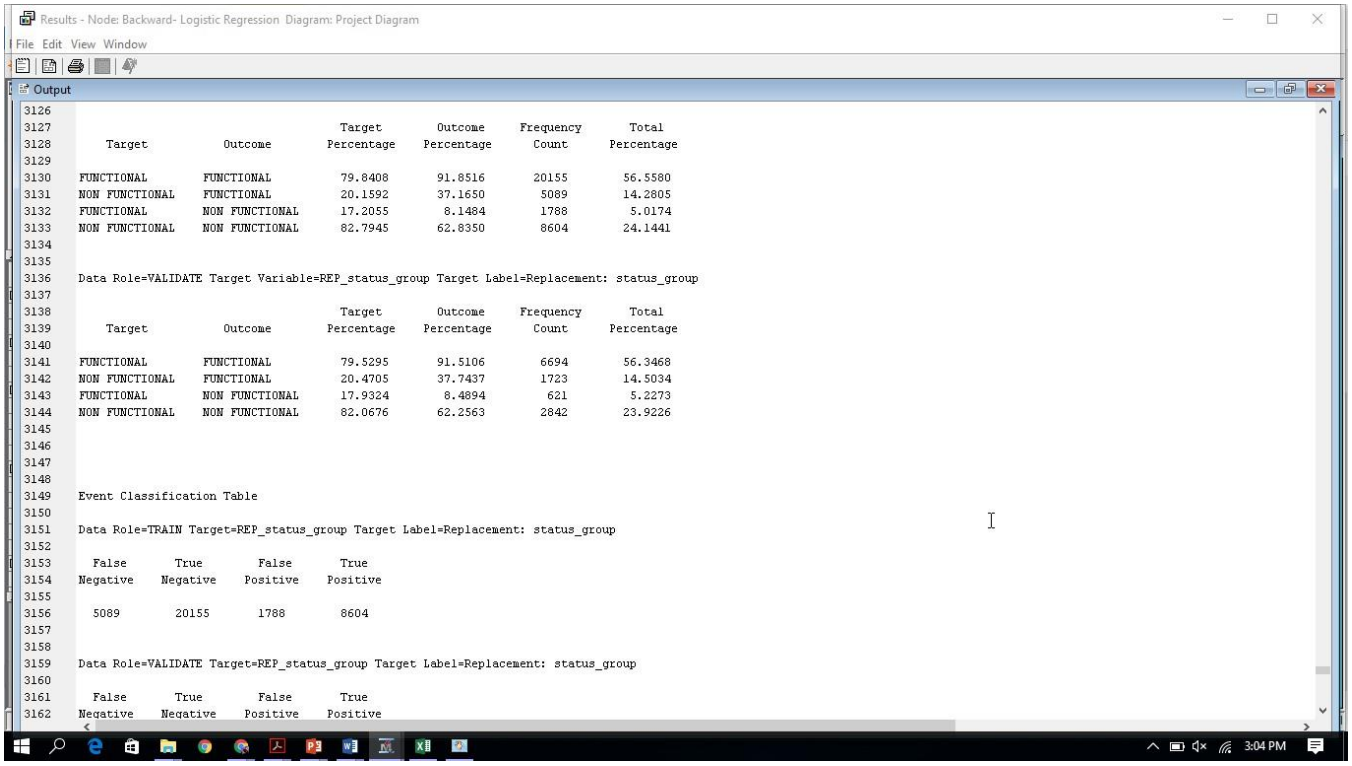


Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
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
LG10_amount_tsh	1	9.4454	0.0021
LG10_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

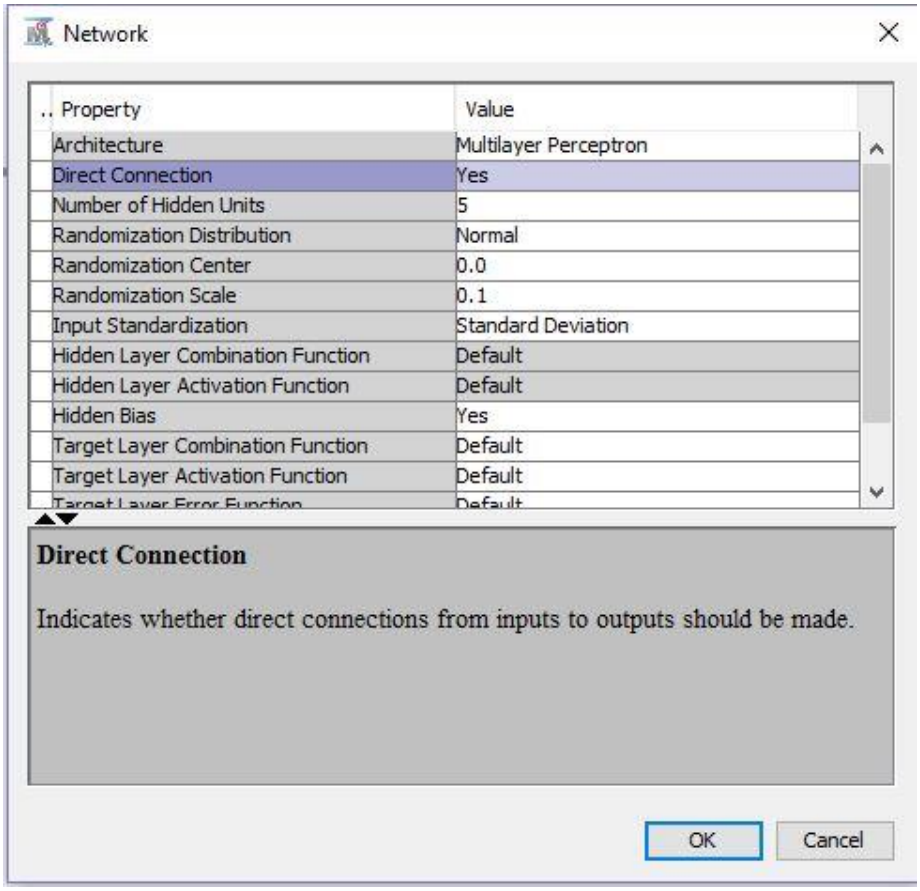


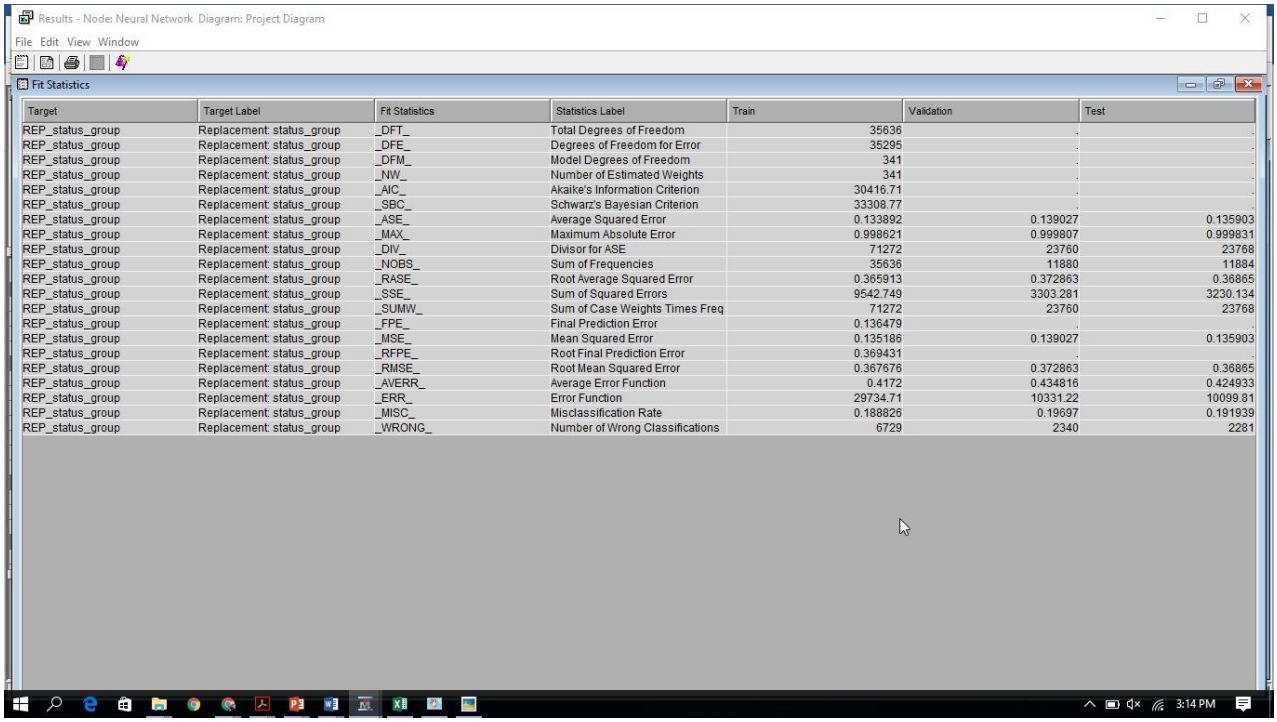
The misclassification rate for the above model is 0.1963.

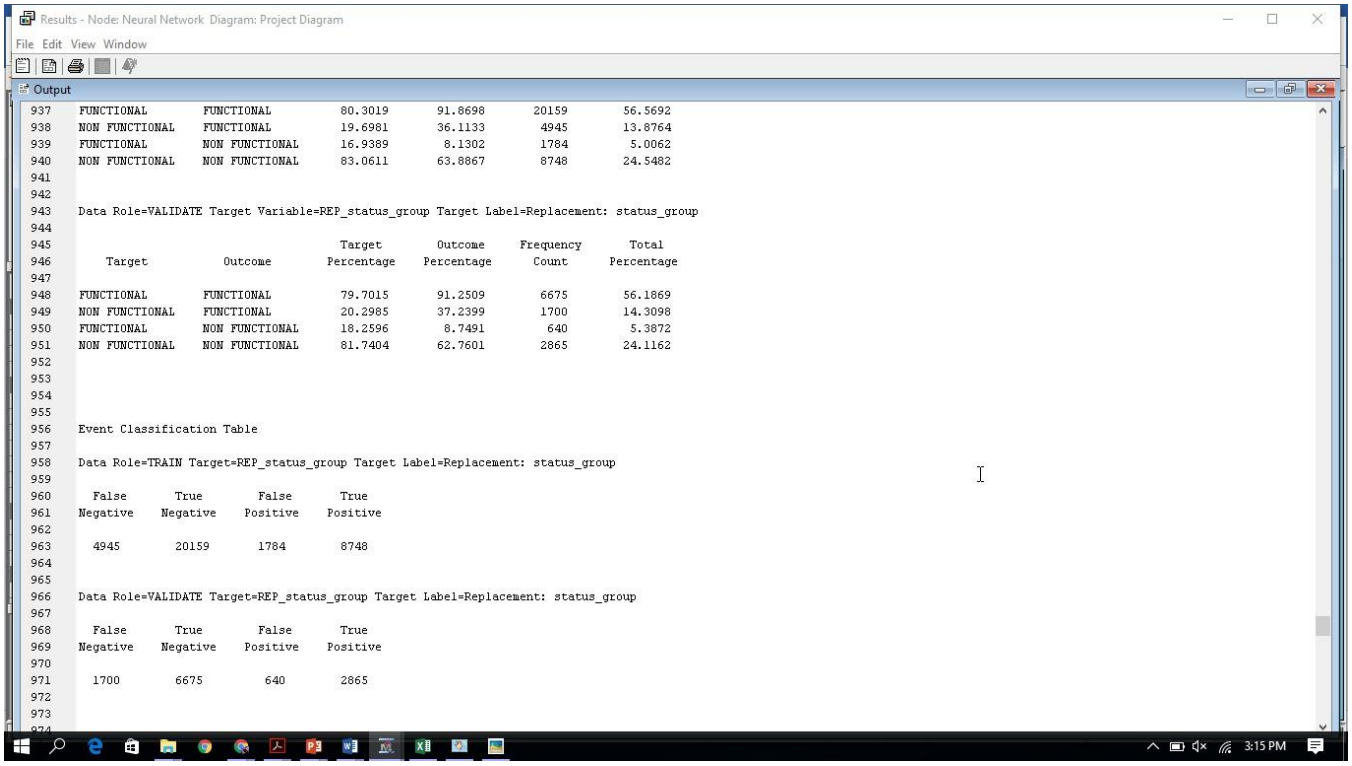
19. Neural Network:

Neural networks are a class of parametric models that can accommodate a wider variety of nonlinear relationships between a set of predictors and a target variable than can logistic regression. Building a neural network model involves two main phases. First, you must define the network configuration. You can think of this step as defining the structure of the model that you want to use. Then, you iteratively train the model. SAS Enterprise Miner has two nodes that fit neural network model: The Neural Network node and the AutoNeural node. We have used the Neural Network node for our modelling. To converge the model fitting process, we have set maximum iteration to 50.

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



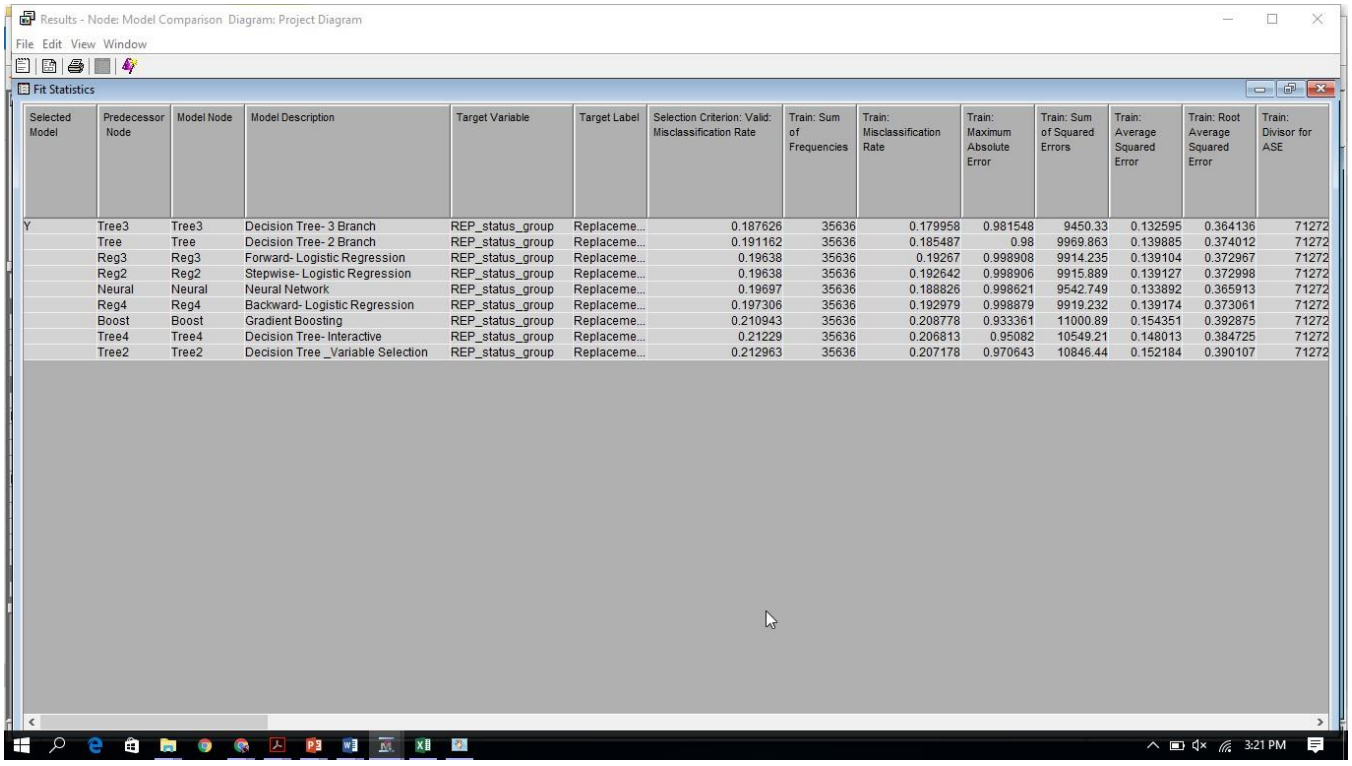




The misclassification rate for the above model is 0.1969.

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:



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 Average Squared Error	Train: 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	71272
	Tree	Tree	Decision Tree- 2 Branch	REP_status_group	Replaceme...	0.191162	35636	0.185487	0.98	9969.863	0.139885	0.374012	71272
	Reg3	Reg3	Forward- Logistic Regression	REP_status_group	Replaceme...	0.19638	35636	0.19267	0.998908	9914.235	0.139104	0.372967	71272
	Reg2	Reg2	Stepwise- Logistic Regression	REP_status_group	Replaceme...	0.19638	35636	0.192642	0.998906	9915.889	0.139127	0.372998	71272
	Neural	Neural	Neural Network	REP_status_group	Replaceme...	0.19697	35636	0.188826	0.998621	9542.749	0.133892	0.365913	71272
	Reg4	Reg4	Backward- Logistic Regression	REP_status_group	Replaceme...	0.197306	35636	0.192979	0.998879	9919.232	0.139174	0.373061	71272
	Boost	Boost	Gradient Boosting	REP_status_group	Replaceme...	0.210943	35636	0.208778	0.933361	11000.89	0.154351	0.392875	71272
	Tree4	Tree4	Decision Tree- Interactive	REP_status_group	Replaceme...	0.21229	35636	0.206813	0.95082	10549.21	0.148013	0.384725	71272
	Tree2	Tree2	Decision Tree _Variable Selection	REP_status_group	Replaceme...	0.212963	35636	0.207178	0.970643	10846.44	0.152184	0.390107	71272

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

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate
Y	Tree3	Decision Tree- 3 Branch	0.18763
	Tree	Decision Tree- 2 Branch	0.19116
	Reg3	Forward- Logistic Regression	0.19638
	Reg2	Stepwise- Logistic Regression	0.19638
	Neural	Neural Network	0.19697
	Reg4	Backward- Logistic Regression	0.19731
	Boost	Gradient Boosting	0.21094
	Tree4	Decision Tree- Interactive	0.21229
	Tree2	Decision Tree _Variable Selection	0.21296

Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
0.13260	0.17996	0.14011
0.13988	0.18549	0.14422
0.13910	0.19267	0.14300
0.13913	0.19264	0.14307
0.13389	0.18883	0.13903
0.13917	0.19298	0.14311
0.15435	0.20878	0.15662
0.14801	0.20681	0.15190
0.15218	0.20718	0.15476

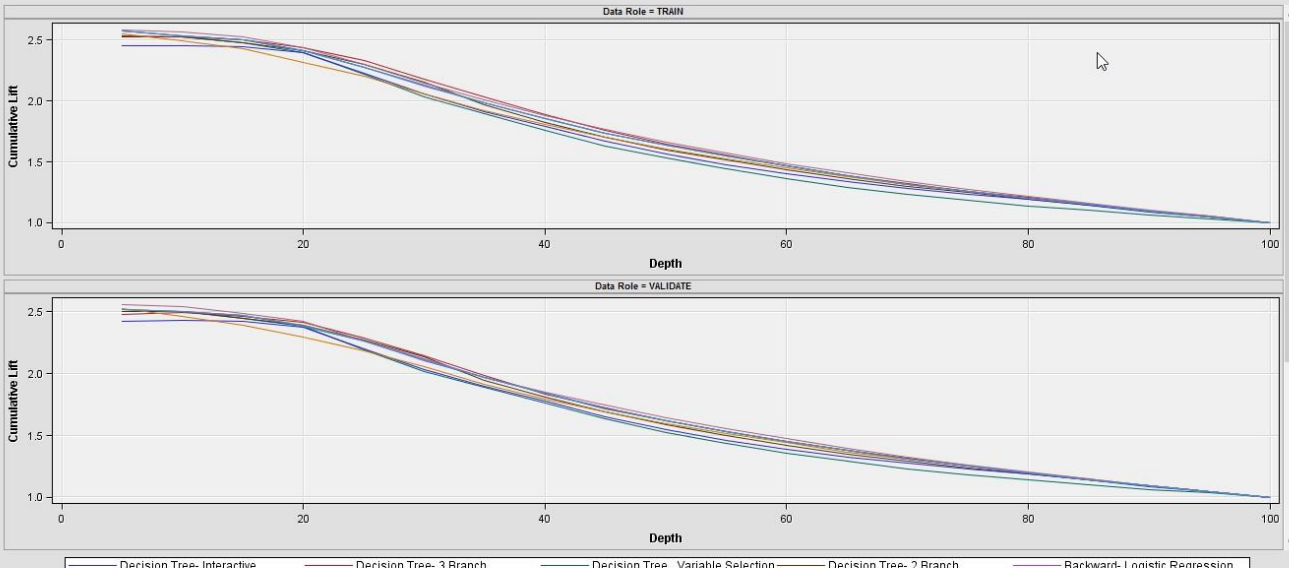
3:21 PM

Results - Node: Model Comparison Diagram: Project Diagram

File Edit View Window

Score Rankings Overlay: Replacement: status_group

Cumulative Lift



3:19 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
- Drivendata: https://www.drivendata.org/competitions/7/