

DSCI -5240 Course project

Analysis of Diabetes Re-Admission Rate

Prabhu Chaitanya

11379637

Table of Contents

1.0 EXECUTIVE SUMMARY:	3
2.0 PROJECT MOTIVATION/BACKGROUND	4
3.0 DATA SET DESCRIPTION:	5
3.1 Predictors in Dataset:.....	5
4.0 DATA PREPROCESSING :	7
4.1 SAMPLING :	7
4.2 DATA EXPLORATION :	8
5.0 SEMMA DATA MINING MODELS :	10
5.1 DECISION TREES:	10
5.2 NEURAL NETWORKS :	14
5.3 LOGISTIC REGRESSION :	15
6.0 Findings :	16
7.0 CONCLUSION :	21
8.0 REFERENCES :	21

List of figures

1. SAMPLING OUTPUT.....	7
2. STAT EXPLORER.....	8
3. REPLACEMENT EDITOR.....	9
4. DATA PARTITION.....	10
5. DECISION TREE WITH INTERACTIVE MODEL.....	11
6. NUMBER OF SPLITS =2.....	12
7. NUMBER OF SPLITS =3.....	12
8. NUMBER OF SPLITS =4.....	13
9. Log10 OPERATION	14
10. LOGISTIC REGRESSION.....	15
11. SUBTREE ASSESSMENT	16
12. OUTPUT OF DECISION TREE MODEL1.....	17
13. COMPARISION OF OUTPUTS OF DECISON TRESS.....	17
14. OUTPUT OF NEURAL NETWORKS.....	18
15. OUTPUT OF LOGISTIC REGRESSION.....	20
16. OUTPUT OF MODEL COMPARISION.....	21

1.0 EXECUTIVE SUMMARY:

Diabetes is a common disease these days, there is a substantial increase in the number of cases detected in the hospitals every year. There are several factors that affect the diabetes like hyperglycaemia levels in the blood and the drugs prescribed by the hospitals to the patients diagnosed with different diseases. These factors either directly or indirectly affect the re-admission rate of the patients to the hospitals.

During patient hospitalisation with any disease or condition certain tests are conducted to check for any abnormal sugar levels in blood(Hyperglycaemia-Hb1Ac test), this data along with the drugs prescribed during hospitalisation will have an impact on readmission rate, which is used as a base for knowing the major causes related to diabetes and also to maintain the hospitals to follow certain guidelines to meet the need of patients.

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospital and integrate delivery networks. The dataset contains 50 variables. This dataset has been extracted from the health facts database(Cerner corporation, Kansas city, MO).

2.0 PROJECT MOTIVATION/BACKGROUND

According to WHO, for the past 25 years number of people with diabetes has risen from 108 million to 422 million and there is recorded 1.6 million deaths directly from diabetes in the year 2016 alone and this is increasing year on year.

There is a need to apply data driven approaches to diabetes data in order to estimate the diabetes patients and increase the patient care related based on it. Analyzing factors that affect diabetes will lead to improve patient care, and explores new ways to treat diabetes.

3.0 DATA SET DESCRIPTION:

This dataset has been acquired from center for clinical and translational research, virginia common health university(UCI machine learning repository).

Link for the data set:

<http://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospital and integrate delivery networks. The dataset contains 50 variables. This dataset has been extracted from the health facts database(cerner corporation,Kansas city,MO).

3.1 Predictors in Dataset:

1. HBA1C(Hemoglobin A1c test): glycated hemoglobin is a form of chemical formed that is linked to the sugar levels in blood, which indicates diabetes. With HB1Ac test it is easy to know the signs of diabetes in our body.
2. Diag 1, Diag 2, Diag 3 are the ICD9 coded values that represents the international classification of diseases which are codes to describe diagnosis. Ex. ICD9 code 250.xx represents diabetes, 390-459, 789 represents circulatory diseases etc..
3. Gluco serumtest: This test measures the amount of glucose dissolved in the blood. People with diabetes have high levels of glucose in their blood which can be known by this test.
4. Medications: 24 drugs were prescribed to the patients encountered with diabetes, The data we have shows the amount of medications prescribed to patients which may effects the readmission rate.
5. Readmission(Target Variable): Indicates the number of days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.
6. The other common variables are number of lab procedures ,Race, Gender, Age, Weight, Time in hospital, Discharge disposition, Number of lab procedures, Number of medications, Number of

outpatient visits, 24 Features of medications(Insulin, Metamorphin, glyburide, tolazamide, miglitol etc...)

After sampling the data to decrease the complexity and to speed up the process, we have reduced the rows from 101767 rows to 10176 rows.

The weight attribute has highest percentage of missing values(over 90), so we have rejected the variable 'weight' while applying regression, the 'payer code' and 'medical specialty' are also rejected since it has no effect on the target variable .

S.NO	Variable	Type	Description
1	Medications (Drugs)	Ordinal	Effect of Drugs given during the treatment of diabetes.
2	Gender	Nominal	Male or Female
3	Readmitted	Interval	Readmitted Patients
4	Medical Speciality	Nominal	Types of surgeon the patient consulted
5	No of lab tests	Interval	No of laboratory tests conducted during the treatment
6	HBA1c test result	Interval	Test which indicates how well your diabetes is controlled.
7	Admission type description	Nominal	Tells the condition of patient when admitted with diabetes (Emergency, New born, Urgent etc..)
8	Discharge description	Nominal	Tells us for what the patient has been discharged for (Discharged to home, expired, discharged to other hospital etc..)
9	Age, weight	Interval	Tells about age and weight.

4.0 DATA PREPROCESSING :

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

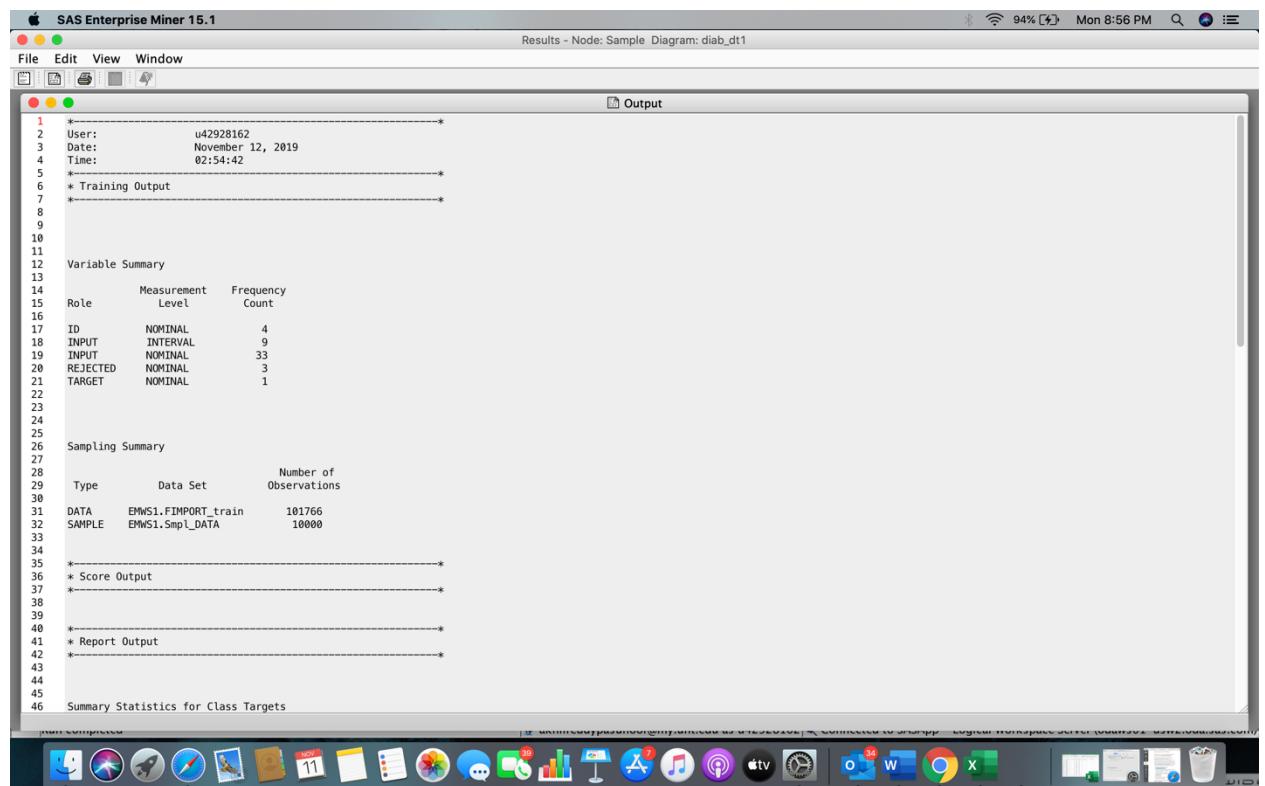
In Real world data are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.

Noisy: Containing errors or outliers

Inconsistent: containing discrepancies in codes or names.

4.1 SAMPLING :

As you can see in the below figure, we sampled the rows in the original dataset from 101,766 to 10,000.



```
1  *-----*
2  User:          u42928162
3  Date:          November 12, 2019
4  Time:          02:54:42
5  *-----*
6  * Training Output
7  *-----*
8
9
10
11
12 Variable Summary
13
14     Measurement   Frequency
15     Role       Level      Count
16
17  ID      NOMINAL      4
18  INPUT   INTERVAL      9
19  INPUT   NOMINAL     33
20  REJECTED NOMINAL      3
21  TARGET  NOMINAL      1
22
23
24
25
26 Sampling Summary
27
28
29     Type      Data Set      Number of
30
31  DATA      EMWS1.FIMPORT_train  101766
32  SAMPLE   EMWS1.Smpl_DATA      10000
33
34
35  *-----*
36  * Score Output
37  *-----*
38
39
40  *-----*
41  * Report Output
42  *-----*
43
44
45
46 Summary Statistics for Class Targets
```

Figure 1: Sampling output

4.2 DATA EXPLORATION :

We explored our variables using stat explorer and we found that there are no missing values and the most repeated value in our target is “NO” and the second most repeated value is “>30” (Patients who were admitted after 30 days)

Apple SAS Enterprise Miner 15.1

Results - Node: StatExplore Diagram: diab_dt1

File Edit View Window

Output

```

38
39
40
41  Class Variable Summary Statistics
42  (maximum 500 observations printed)
43
44  Data Role=TRAIN
45
46
47  Data
48  Role  Variable Name  Role  Number of Levels  Missing  Mode  Mode Percentage  Mode2 Percentage
49
50  TRAIN  A1Cresult  INPUT  4  0  None  83,17  >8  8,36
51  TRAIN  acarbose  INPUT  3  0  No  99,72  Steady  0,26
52  TRAIN  age  INPUT  10  0  [70-80]  26,15  [60-70]  21,40
53  TRAIN  change  INPUT  2  0  No  53,39  Ch  46,61
54  TRAIN  chlorpropamide  INPUT  3  0  No  99,98  St  0,09
55  TRAIN  diabetesMed  INPUT  2  0  Yes  76,85  No  23,15
56  TRAIN  diag_1  INPUT  479  0  428  6,45  414  6,13
57  TRAIN  diag_2  INPUT  469  0  276  6,76  420  6,43
58  TRAIN  diag_3  INPUT  462  0  359  12,15  401  8,69
59  TRAIN  gender  INPUT  2  0  Female  53,47  Male  46,53
60  TRAIN  glimepiride  INPUT  4  0  No  94,74  Steady  4,71
61  TRAIN  glipizide  INPUT  4  0  No  87,73  Steady  11,17
62  TRAIN  glyburide  INPUT  4  0  No  89,35  Steady  9,26
63  TRAIN  glyburide_metformin  INPUT  2  0  No  99,17  St  0,83
64  TRAIN  insulin  INPUT  4  0  No  46,77  Steady  30,51
65  TRAIN  max_glu_serum  INPUT  4  0  Non  94,99  Norm  2,51
66  TRAIN  metformin  INPUT  4  0  No  88,48  Steady  17,94
67  TRAIN  nateglinide  INPUT  3  0  No  99,30  St  0,67
68  TRAIN  pioglitazone  INPUT  4  0  No  92,07  St  7,53
69  TRAIN  race  INPUT  6  0  Caucasian  73,99  AfricanAmerican  19,58
70  TRAIN  repaglinide  INPUT  4  0  No  98,26  Steady  1,60
71  TRAIN  rosiglitazone  INPUT  4  0  No  93,81  Steady  5,93
72  TRAIN  readmittad  TARGET  6  0  No  52,75  >30  34,30
73
74
75  Distribution of Class Target and Segment Variables
76  (maximum 500 observations printed)
77
78  Data Role=TRAIN
79
80  Data
81  Role  Variable  Role  Frequency  Count  Percent
82  Role  Name  Level
83

```

Figure 2:Stat Explorer

The next step we did was Replacement. we have used replacement and replaced a few values like acarbose since it has very low frequency

Apple SAS Enterprise Miner 15.1

Replacement Editor-WORK.OUTCLASS

File Edit View Actions Op

diab_dt

File Sources Diagrams diab_dt1 Model Packages

Property Value

Exported Data Notes Train

Interval Variables

- Replacement Editor
- Default Limits Meth:None
- Cutoff Values

Class Variables

- Replacement Editor
- Unknown Levels: Ignore

Score

Replacement Values: Computed

Report

Replacement Report: Yes

Status

Create Time: 11/12/19 3:22

Run ID: 5e8e676b-40

Last Error

Last Status: Complete

Last Run Time: 11/12/19 3:22

Run Duration: 0 Hr. 0 Min. 4 Sec.

Grid Host

User-Added Node: No

Default Limits Method

Specifies the default method to determine range limits for interval variables.

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
A1Cresult	None		8317C	None	.	.
A1Cresult	>8		836C	>8	.	.
A1Cresult	Norm		465C	Norm	.	.
A1Cresult	>7		382C	>7	.	.
acarbose	No		9972C	No	.	.
acarbose	Steady		26C	Steady	.	.
acarbose	Up	Steady	2C	Up	.	.
acarbose	_UNKNOWN_	_DEFAULT_	C	.	.	.
age	[70-80]		2615C	[70-80]	.	.
age	[60-70]		2140C	[60-70]	.	.
age	[50-60]		1720C	[50-60]	.	.
age	[80-90]		1683C	[80-90]	.	.
age	[40-50]		953C	[40-50]	.	.
age	[30-40]		362C	[30-40]	.	.
age	[90-100]		275C	[90-100]	.	.
age	[20-30]		160C	[20-30]	.	.
age	[10-20]		74C	[10-20]	.	.
age	[0-10]		18C	[0-10]	.	.
age	_UNKNOWN_	_DEFAULT_	C	.	.	.
change	No		5339C	No	.	.
change	Ch		4661C	Ch	.	.
change	_UNKNOWN_	_DEFAULT_	C	.	.	.
chlorpropamide	No		9990C	No	.	.
chlorpropamide	St		9C	St	.	.
chlorpropamide	Up		1C	Up	.	.
chlorpropamide	_UNKNOWN_	_DEFAULT_	C	.	.	.
diabetesMed	Yes		7685C	Yes	.	.
diabetesMed	No		2315C	No	.	.
diabetesMed	_UNKNOWN_	_DEFAULT_	C	.	.	.
diag_1	428		645C	428	.	.
diag_1	414		613C	414	.	.
diag_1	786		419C	786	.	.
diag_1	410		348C	410	.	.
diag_1	486		281C	486	.	.
diag_1	427		269C	427	.	.
diag_1	491		232C	491	.	.
diag_1	715		220C	715	.	.

OK Cancel

Server (odawsu1-uswz.oda.sas.com)

Figure 3: Replacement Editor

5.0 SEMMA DATA MINING MODELS :

The following models were used in SAS Enterprise Miner to seek valuable insights from the dataset.

5.1 DECISION TREES:

The first thing we did in decision trees was partitioning the data. we have selected data partition node and allotted 70% of our data as training data and 15% of data each to validation and testing data.

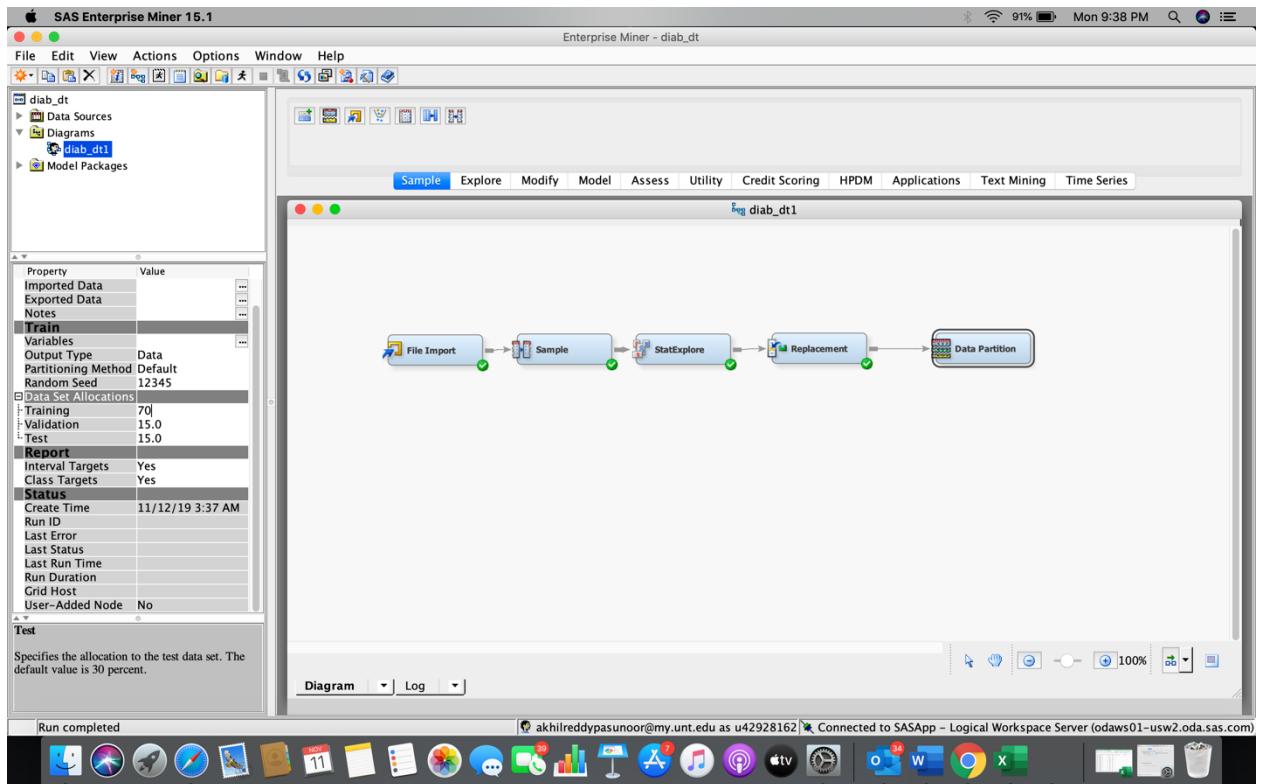


Figure 4: Data Partition

We modelled 4 different Trees to enable us identify the best fit for our data. The first Decision Tree was built with using interactive option and selecting frozen tree and we split the tree based on rep_race variable as mentioned by sas. The output tree was built with 9 leaves. We have built our second decision tree by using the default sas options and we got 5 leaves. We have now selected the maximum number of splits per each branch and we set it to 3 for the third decision tree and 4 for the fourth decision tree we have discussed the results in observations

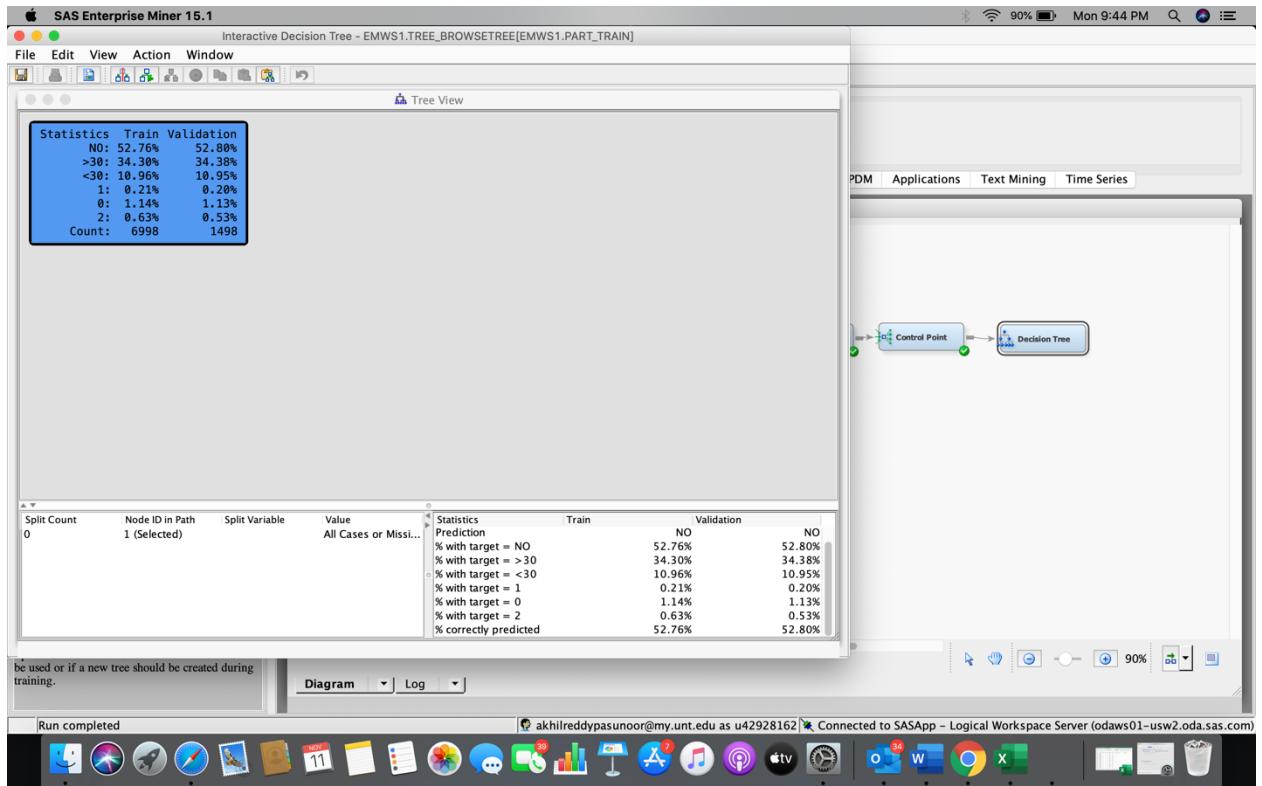
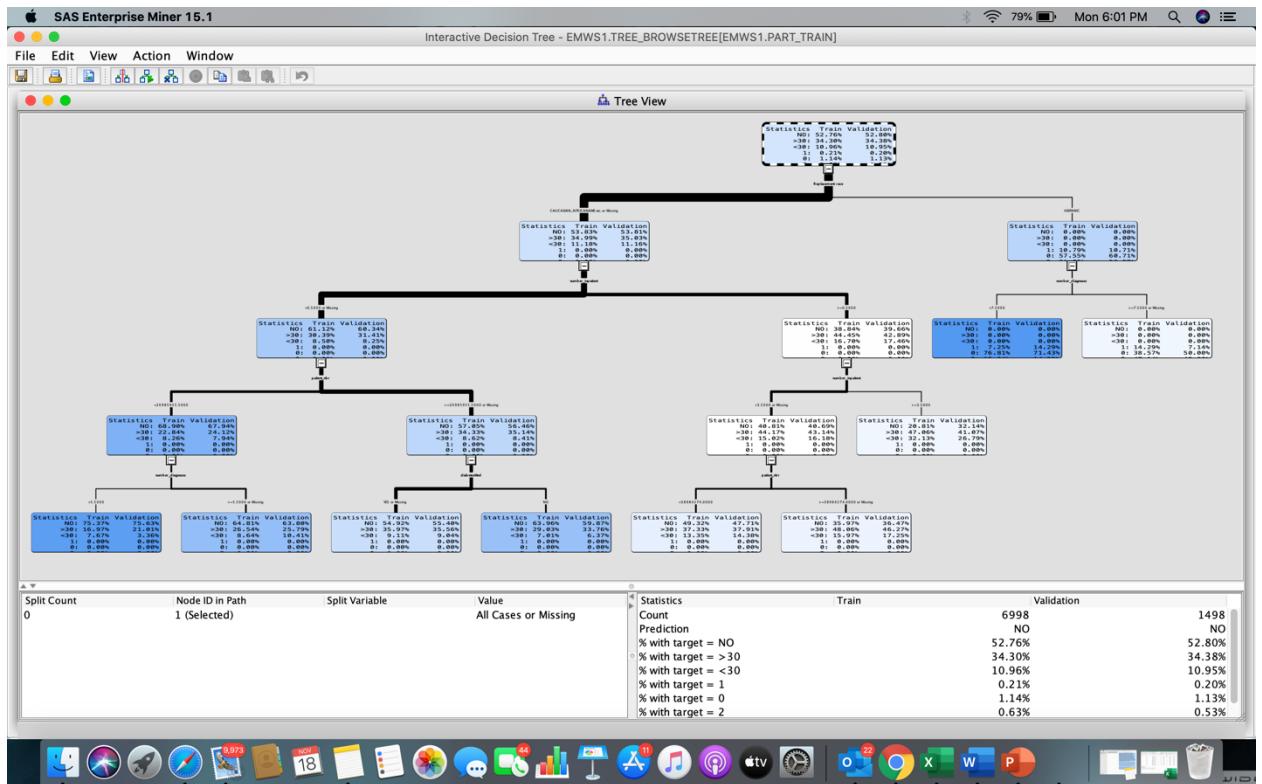


Figure 5: Decision tree with Interactive model

Output of Interactive Model:



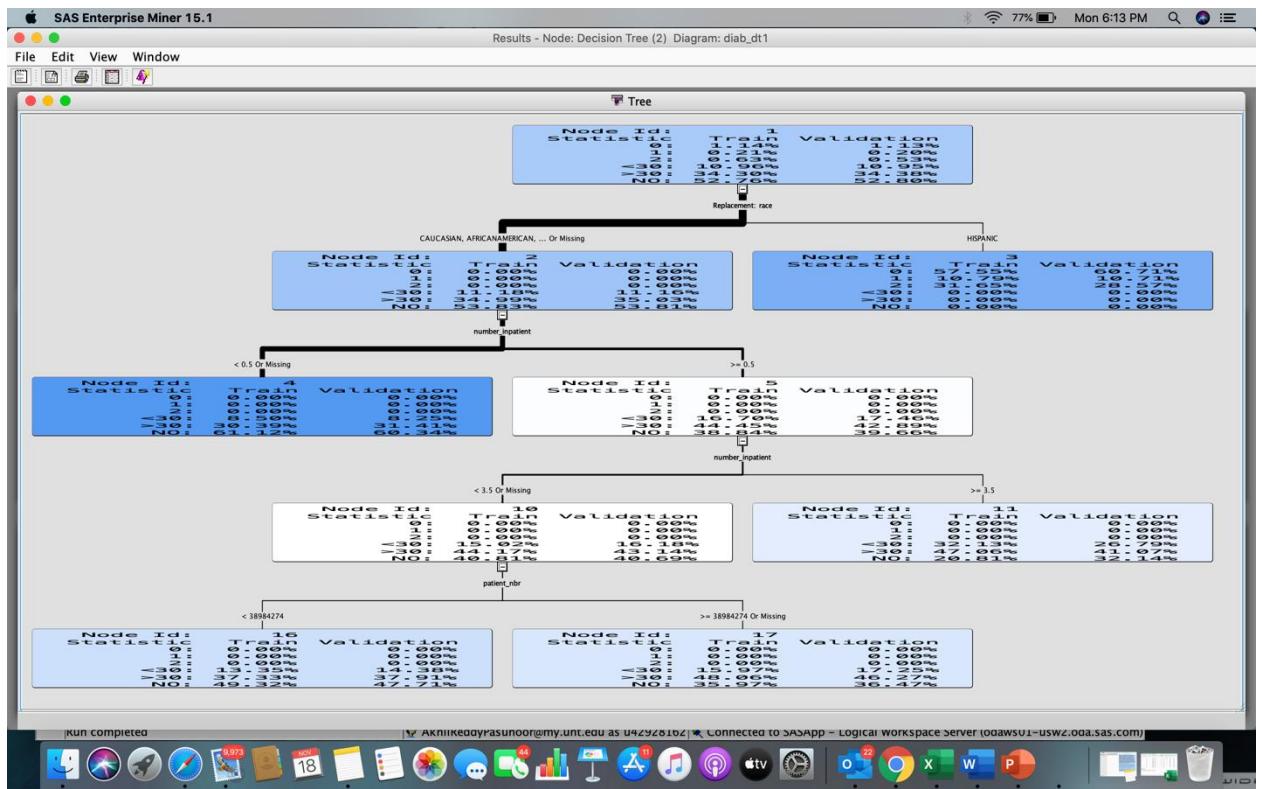


Figure 6: With no of splits=2

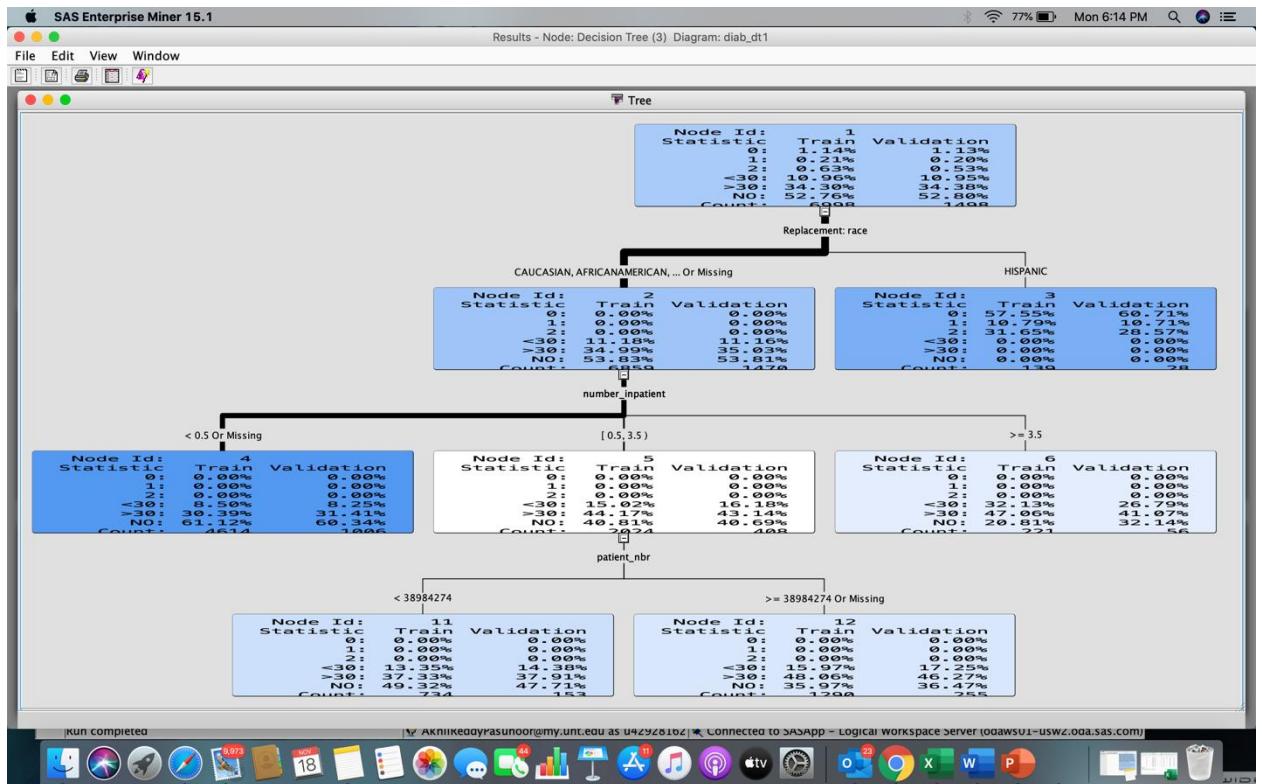
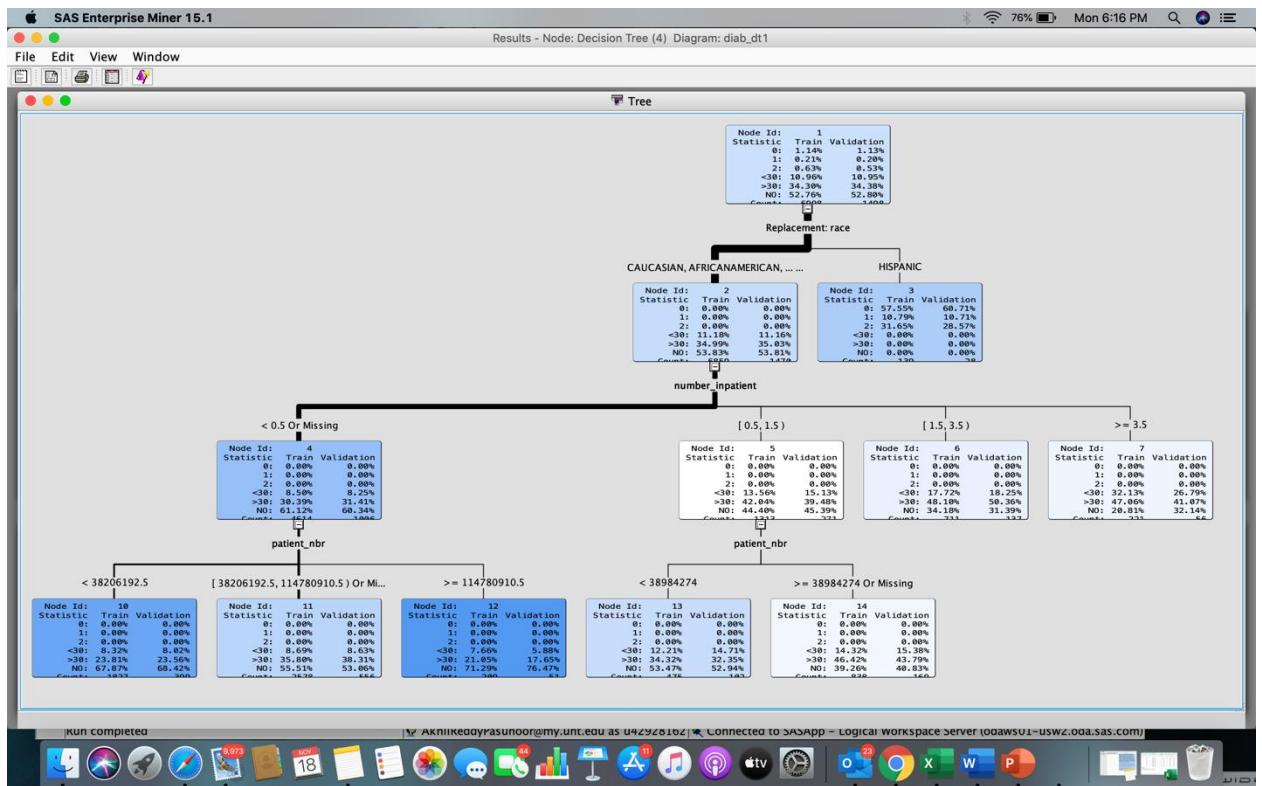


Figure 7: With no of splits =3



5.2 NEURAL NETWORKS :

Variables number out-patient and number emergency are positively skewed, which can be inferred from stat explorer. Inorder to remove this skewness, we performed log10 operation using transform variable block.

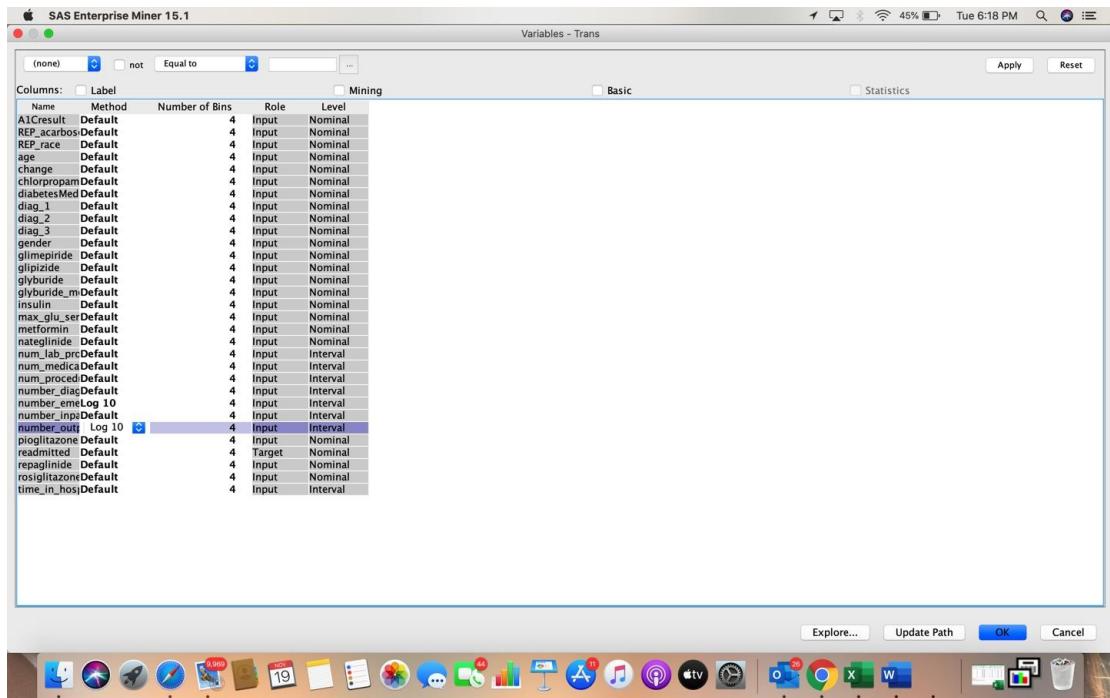
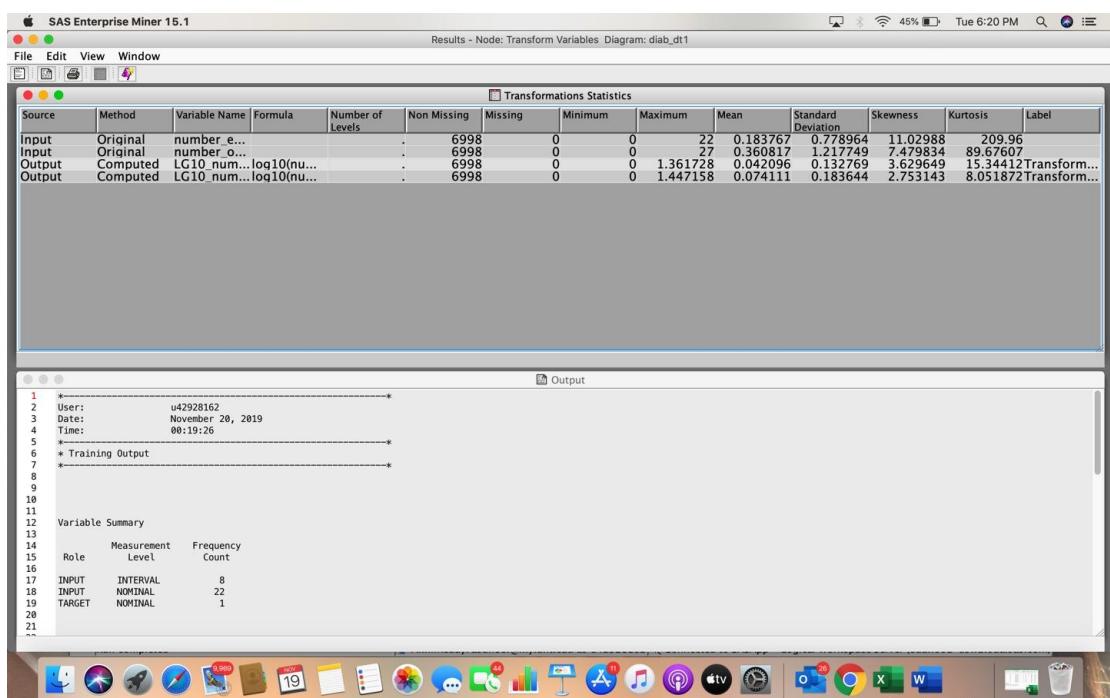


Figure 9: log10 operation

Output of log10 operation:



We performed the neural networks after removing the skewness with number of hidden variables 2 and maximum iterations of 8.

5.3 LOGISTIC REGRESSION :

We applied the stepwise variable selection to the logistic regression model. The weight attribute has highest percentage of missing values(over 90), so we have rejected the variable 'weight' while applying regression, the 'payer code' and 'medical specialty' are also rejected since it has no effect on the target variable .

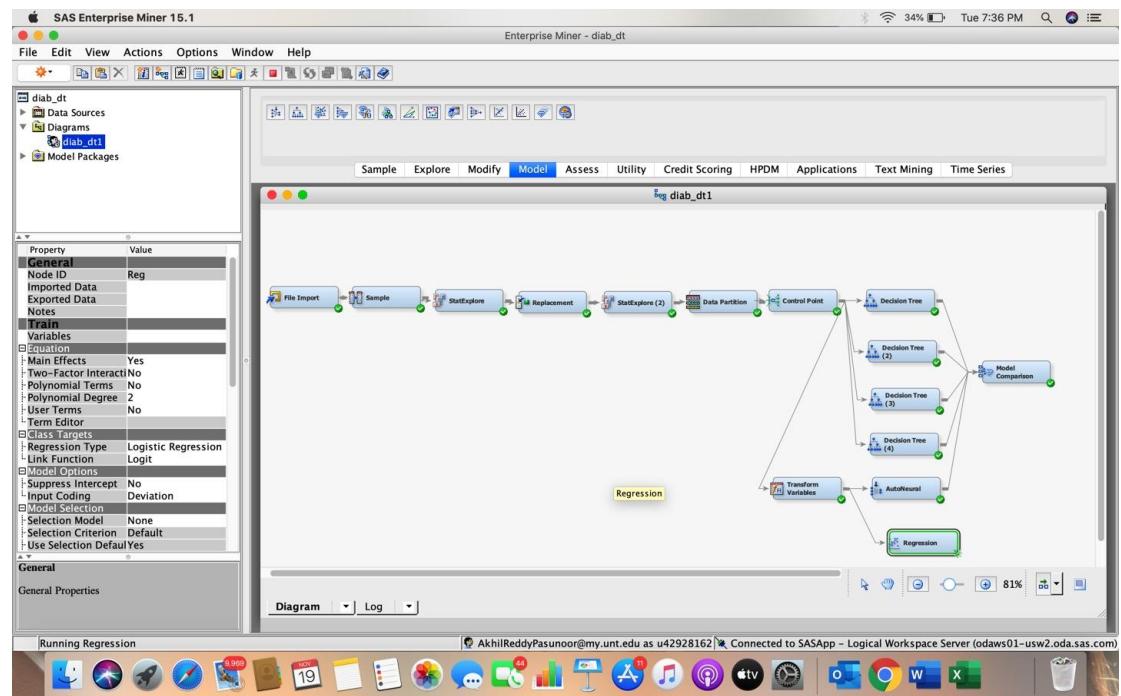


Figure 10: Logistic Regression

6.0 Findings :

In Decision tree model, we plotted the Subtree Assessment Plot for the first decision tree.

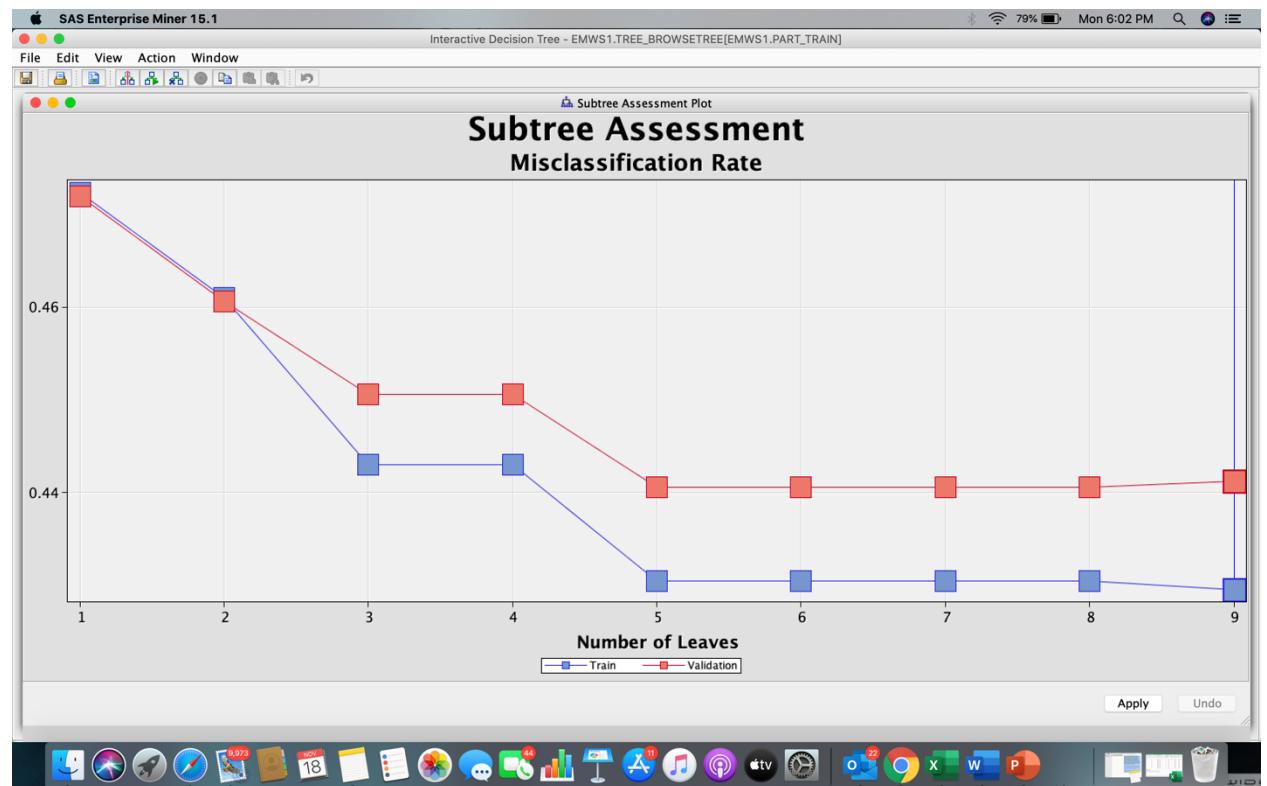


Figure 11: Subtree Assessment

The misclassification rate decreases for both training and validation data sets as the number of the leaves increases and it becomes constant.

The misclassification rate for training and validation sets are 0.429 and 0.441 respectively as you can see in the figure 11.

The misclassification rate for second decision tree validation data set is 0.4405 and for the third and fourth data sets are 0.4405 and 0.4365 respectively as you can see in the figure 12.

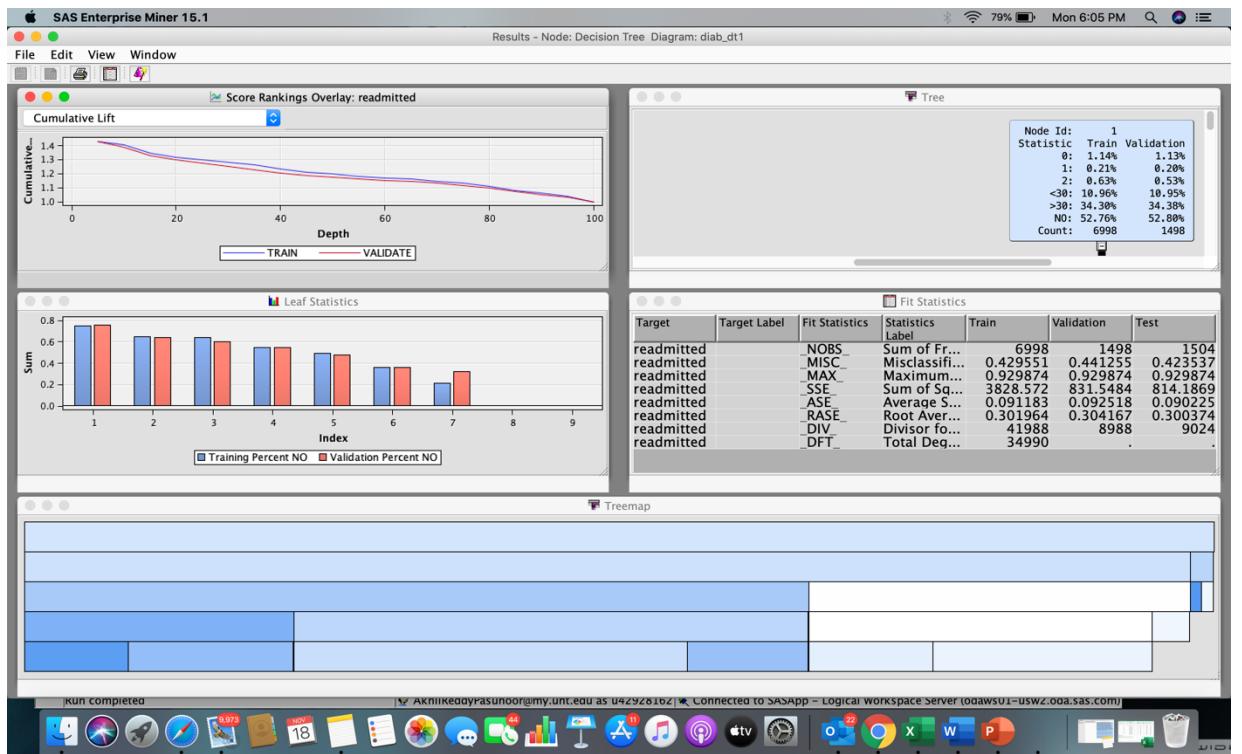


Figure 12: Output of Decision Tree model 1

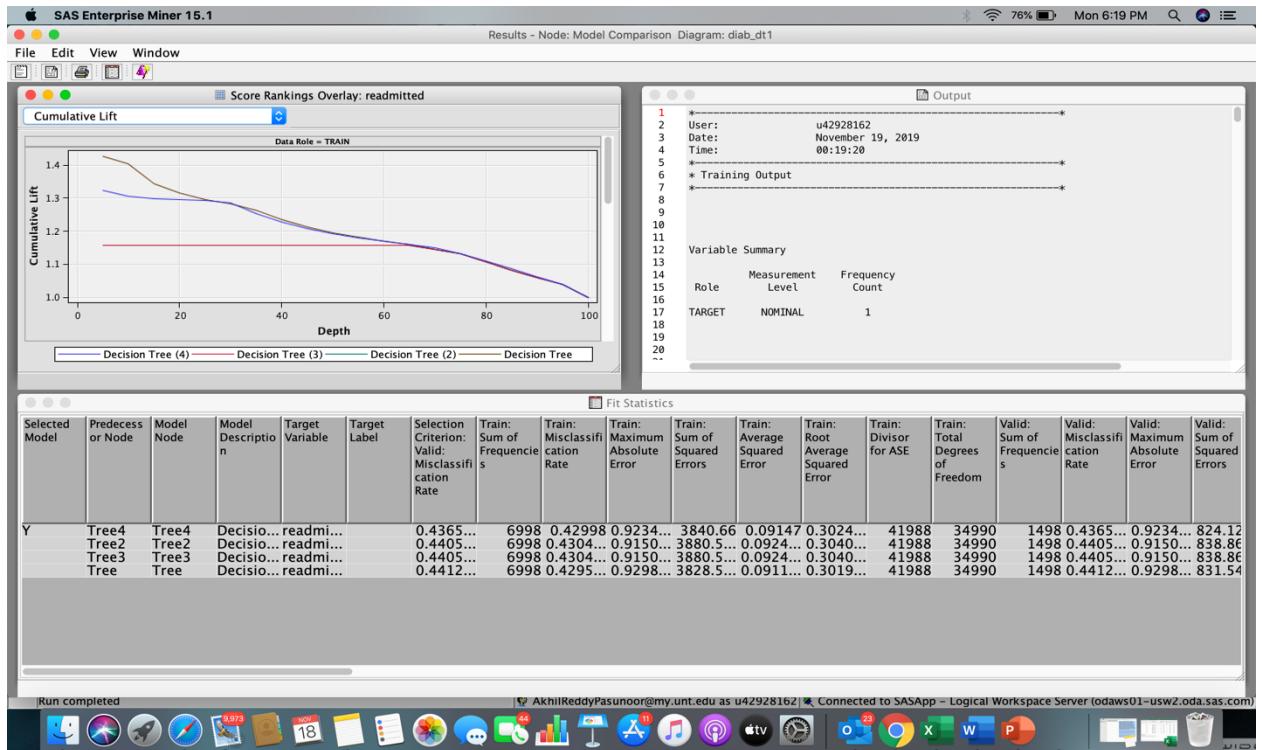


Figure 13: Comparison of 4 different models of Decision Trees

In Neural Networks we got the average square error of 0.092 as you can see in the below figure, for the 5 hidden units, in the classification table the levels of the target variable (Readmitted) is divided in to three levels and the outcome is calculated individually.

The weights generated during the process are important in predicting the target variables, thus the model initially assigns the weights randomly and by applying the tanh functions and gradient descent it estimates the weights which best fits the data with difference between the weighted sum and the actual output.



Figure 14: Output of neural networks

SAS Enterprise Miner 15.1

Results - Node: AutoNeural Diagram: diag_dt1

File Edit View Window

Classification Table

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency	Total Count	Total Percentage
0	>30	3.4746	51.2500	41	0.5859	
1	>30	1.1864	93.3333	14	0.2001	
2	>30	2.8814	77.2727	34	0.4859	
<30	>30	19.5763	30.1173	231	3.3009	
>30	>30	42.9775	22.1117	541	7.3538	
NO	>30	27.8329	8.6403	319	4.5584	
0	NO	8.6783	48.7500	39	0.5573	
1	NO	0.0172	6.6667	1	0.0143	
2	NO	0.1719	22.7273	10	0.1429	
<30	NO	9.2128	69.8827	536	7.6593	
>30	NO	31.9526	77.4583	1859	26.3647	
NO	NO	57.9752	91.3597	3373	48.1995	

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency	Total Count	Total Percentage
0	>30	2.6549	35.2941	6	0.4005	
1	>30	8.4425	33.3333	1	0.0668	
2	>30	2.5349	75.0000	6	0.4005	
<30	>30	19.4698	26.8293	44	2.8372	
>30	>30	42.9284	18.8356	97	6.4753	
NO	>30	31.8584	9.1024	72	4.8064	
0	NO	0.8648	64.7059	11	0.7343	
1	NO	0.1572	66.0007	2	0.1335	
2	NO	0.1572	25.8000	2	0.1335	
<30	NO	9.4348	73.1707	120	8.8107	
>30	NO	32.8616	81.1650	418	27.9839	
NO	NO	56.5252	90.8976	719	47.9973	

Event Classification Table

SAS Enterprise Miner 15.1

Results - Node: AutoNeural Diagram: diag_dt1

File Edit View Window

* Report Output

Fit Statistics

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

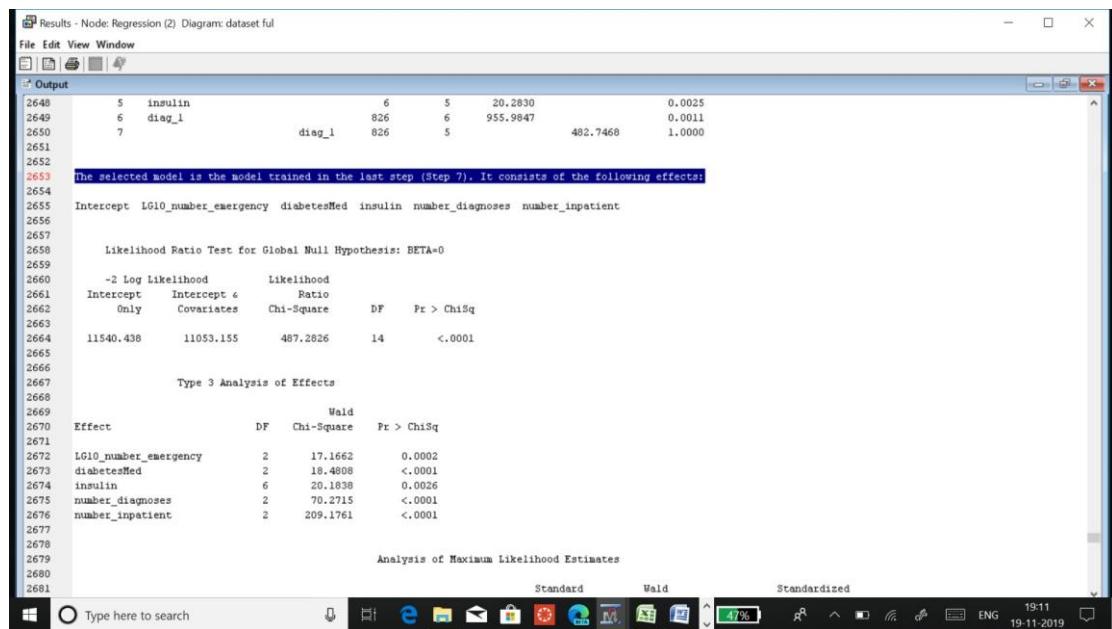
Statistics	Statistics Label	Train	Validation	Test
DFT	Total Degrees of Freedom	34998.00	.	.
DFE	Degrees of Freedom for Error	32349.00	.	.
DFM	Model Degrees of Freedom	2641.00	.	.
NW	Number of Estimated Weights	2641.00	.	.
AIC	Akaike's Information Criterion	19185.56	.	.
BIC	Schwarz's Bayesian Criterion	41455.56	.	.
ASE	Average Squared Error	0.09	0.10	0.09
MAX	Maximum Absolute Error	1.00	1.00	1.00
DIV	Divisor for ASE	41988.00	8988.00	9024.00
NORM	Sum of Frequencies	6998.00	1498.00	1584.00
PASE	Root Average Squared Error	0.30	0.31	0.31
SSE	Sum of Square Errors	3804.06	861.38	845.17
SUMW	Sum of Case Weights Times Freq	41988.00	8988.00	9024.00
FPE	Final Prediction Error	0.11	0.10	.
MSE	Mean Squared Error	0.10	0.10	0.09
RFE	Root Final Prediction Error	0.33	.	.
RMSE	Root Mean Squared Error	0.32	0.31	0.31
AVERR	Average Error Function	0.33	0.34	0.34
ERR	Error Function	13823.56	3030.70	3025.28
MISC	Misclassification Rate	0.44	0.46	0.44
WRONG	Number of Wrong Classifications	3804.00	682.00	658.00

Classification Table

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency	Total Count	Total Percentage
0	>30	3.4746	51.2500	41	0.5859	

In Logistic Regression the model is significant as chi-sq value is <0.0001, the ideal model is obtained in step 7 with all significant variables. The significant variables obtained are number_emergency, diabetMed, insulin, number_diagnosis and number_inpatient with having chi-squares <0.0001. This model is Stepwise logistic regression model.



```

Results - Node: Regression (2) Diagram: dataset ful
File Edit View Window
Output
2648      5  insulin          6      5    20.2830      0.0025
2649      6  diag_1           826     6    955.9847      0.0011
2650      7  diag_1           826     5    482.7468      1.0000
2651
2652
2653 The selected model is the model trained in the last step (Step 7). It consists of the following effects:
2654
2655 Intercept LG10_number_emergency diabetesMed insulin number_diagnoses number_inpatient
2656
2657
2658 Likelihood Ratio Test for Global Null Hypothesis: BETA=0
2659
2660      -2 Log Likelihood      Likelihood
2661 Intercept      Intercept &      Ratio
2662 Only Covariates      Chi-Square      DF      Pr > ChiSq
2663
2664 11540.438      11053.155      487.2826      14      <.0001
2665
2666
2667      Type 3 Analysis of Effects
2668
2669      Wald
2670 Effect      DF      Chi-Square      Pr > ChiSq
2671
2672 LG10_number_emergency      2      17.1662      0.0002
2673 diabetesMed      2      18.4808      <.0001
2674 insulin      6      20.1838      0.0026
2675 number_diagnoses      2      70.2715      <.0001
2676 number_inpatient      2      209.1761      <.0001
2677
2678
2679      Analysis of Maximum Likelihood Estimates
2680
2681

```

Figure 15: Output of Logistic Regression

7.0 CONCLUSION :

The three models are compared with respect to the average squared error. The average squared error of the neural networks model is less when compared to other models. So, it is the best model and also the re-admission rate depends on the variables diabetMed, insulin, number of diagnosis and number of inpatient.

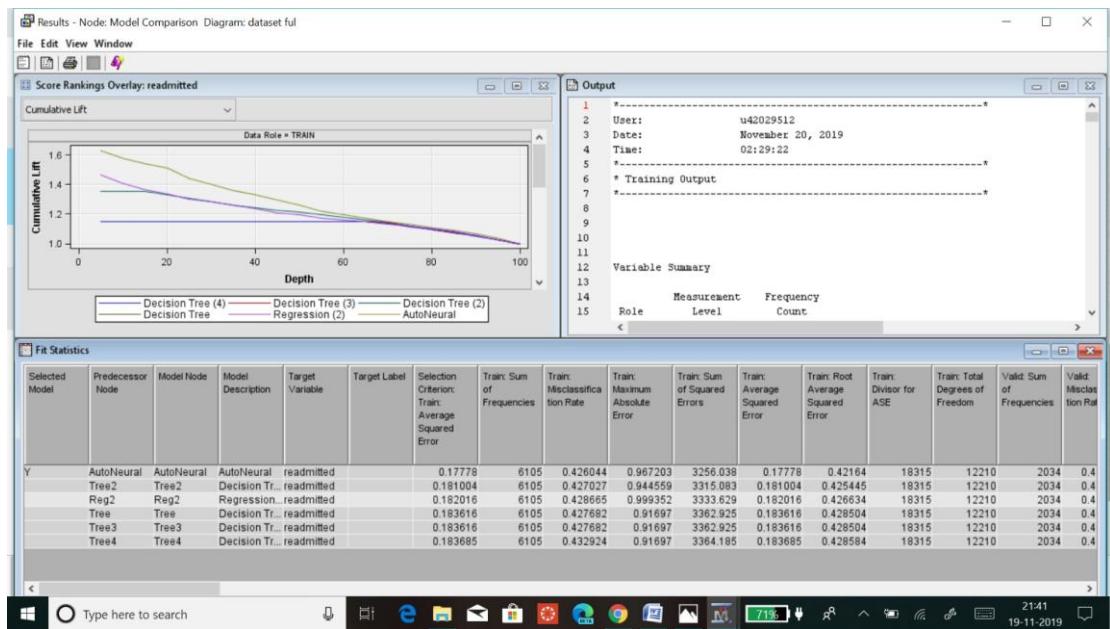


Figure 16: Output of model comparison

8.0 REFERENCES :

1. <https://www.hindawi.com/journals/bmri/2014/781670/>
2. <http://archive.ics.uci.edu/ml/datasets/diabetes+130+us+hospitals+for+years+1999-2008>
3. <https://www.who.int/news-room/fact-sheets/detail/diabetes>