DSCI -5240 Course project

Analysis of Diabetes Re-Admission Rate

Prabhu Chaitanya 11379637

Table of Contents

List of figures

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.

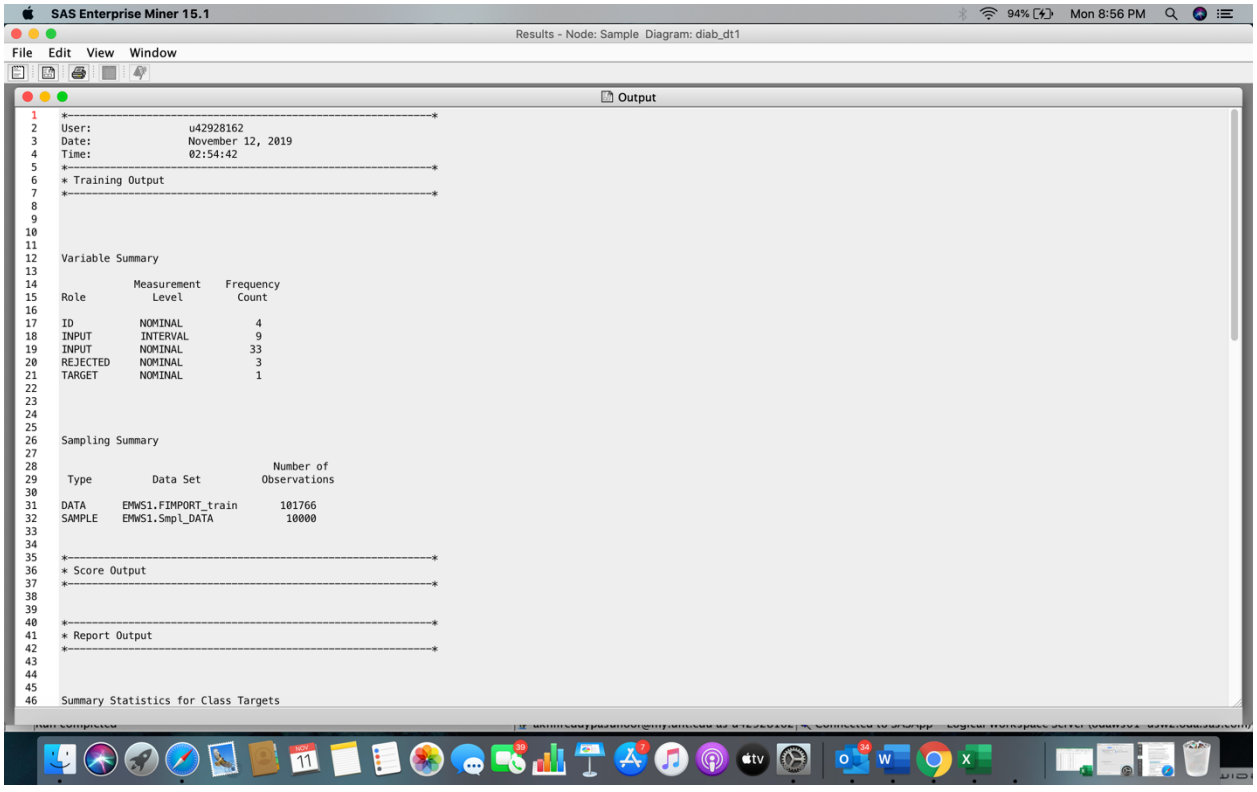


Figure 1: Sampling output

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)

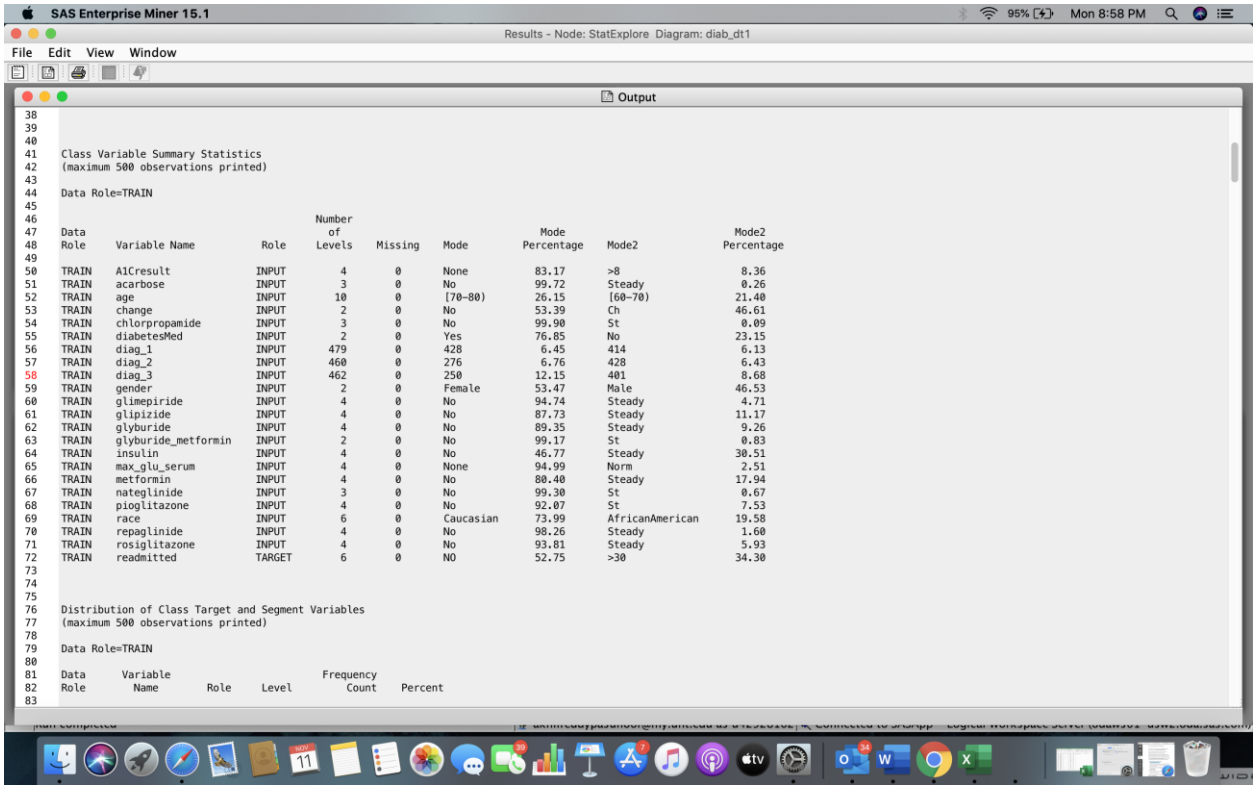


Figure 2:Stat Explorer

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

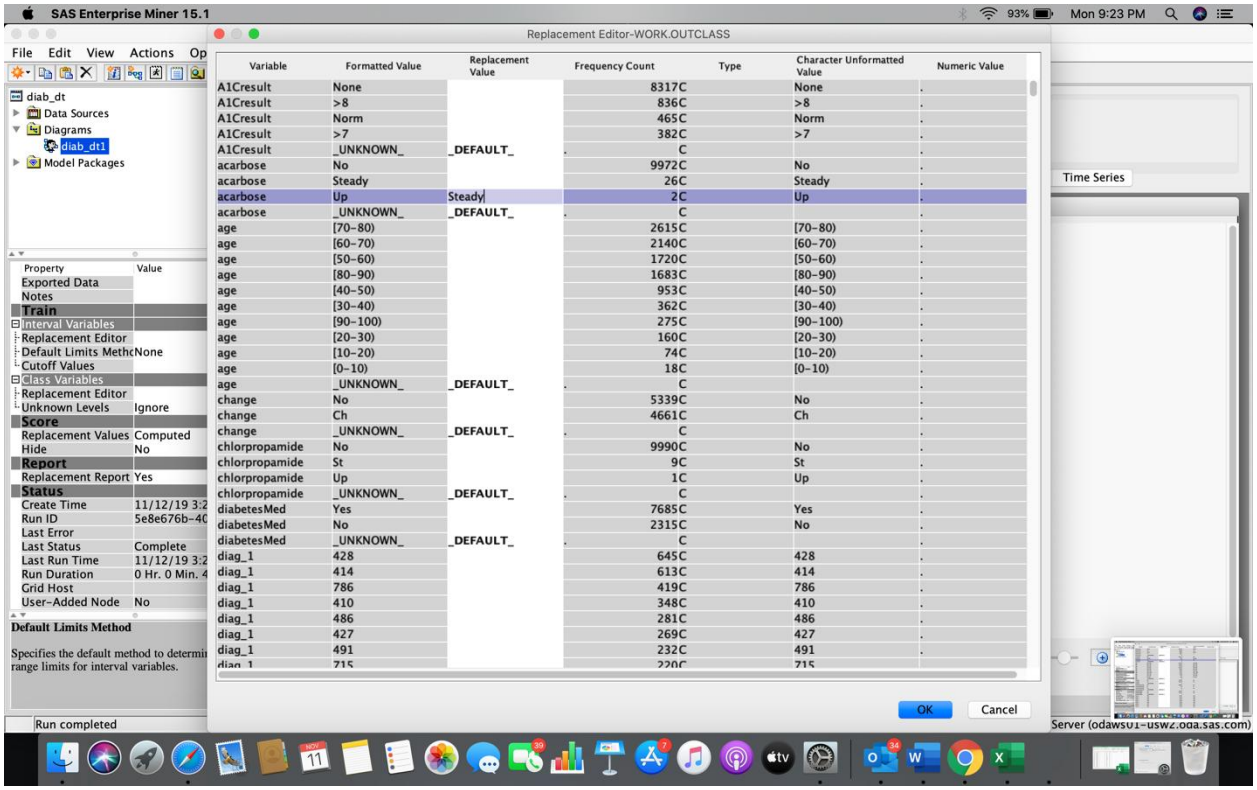


Figure 3: Replacement Editor

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 partioning the data. we have selected data partition node and alloted 70% of our data as training data and 15% of data each to validation and testing data.

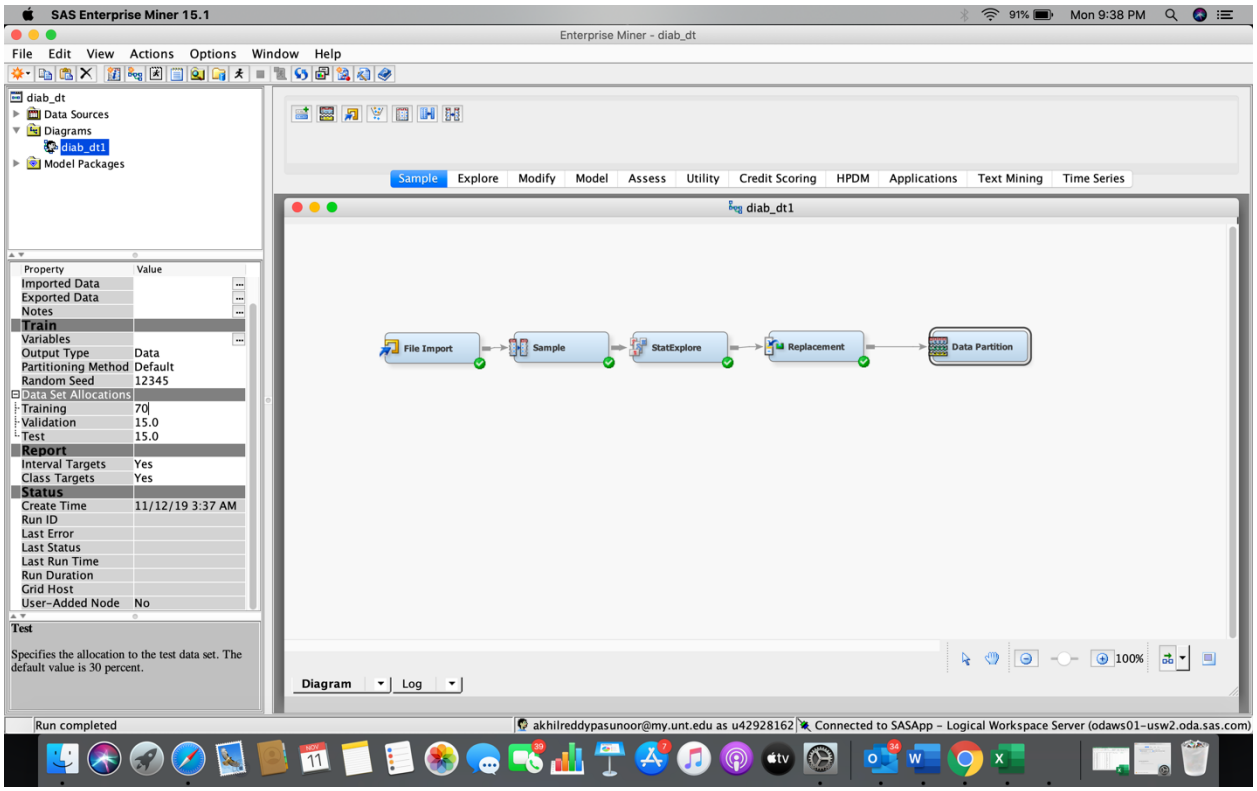


Figure 4: Data Partition

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

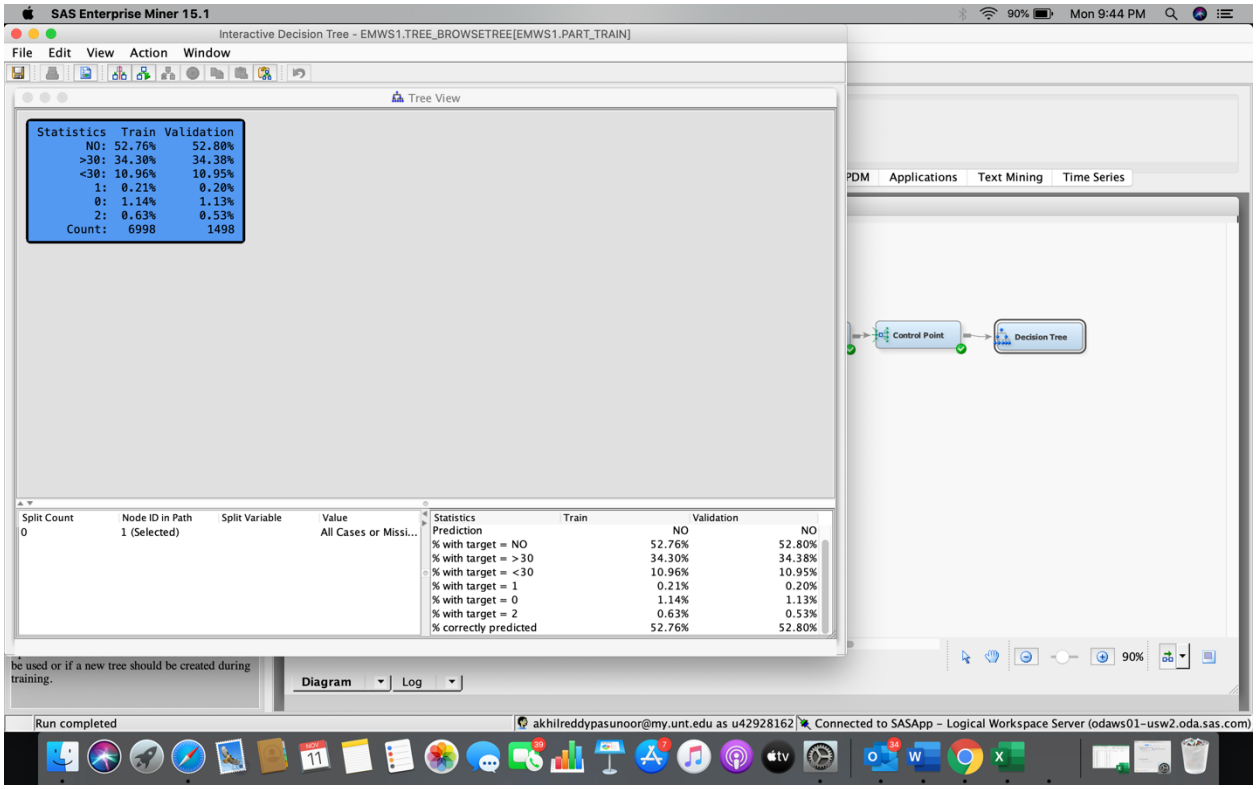
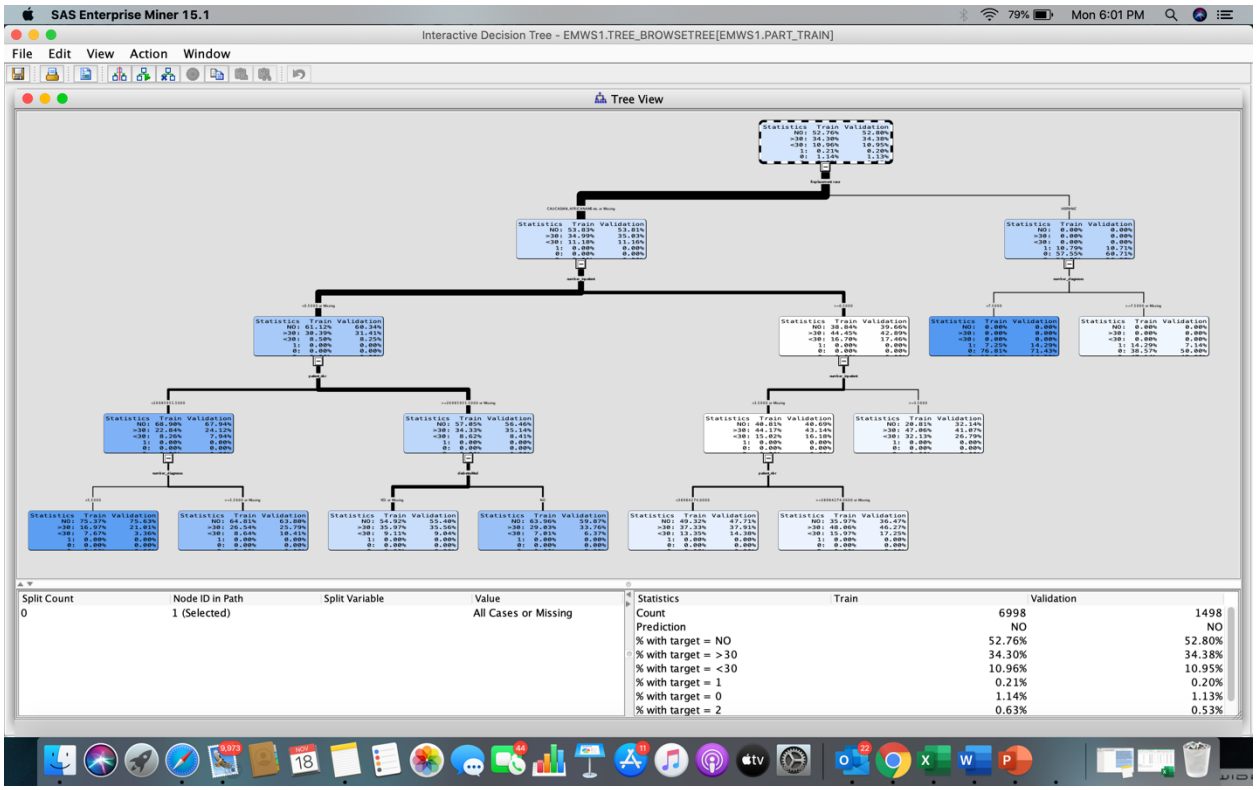


Figure 5: Decision tree with Interactive model

Output of Interactive Model:



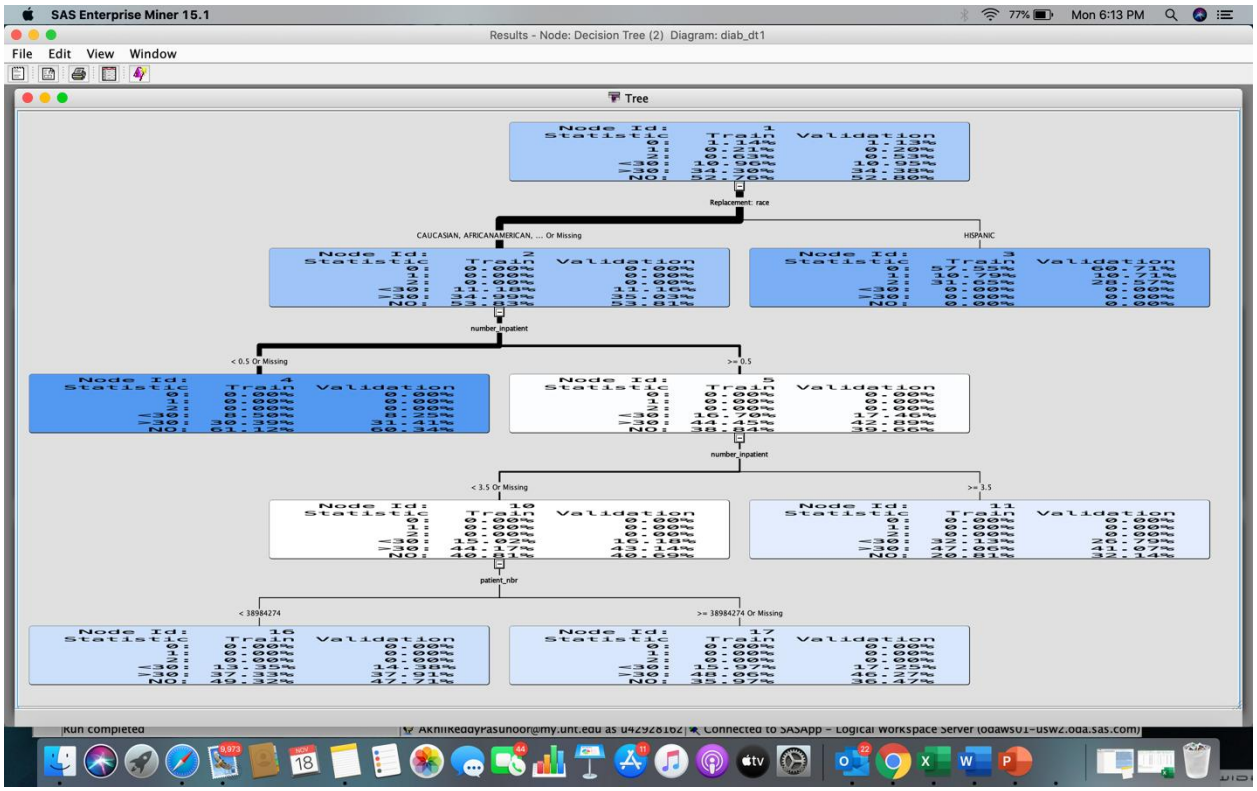


Figure 6: With no of splits=2

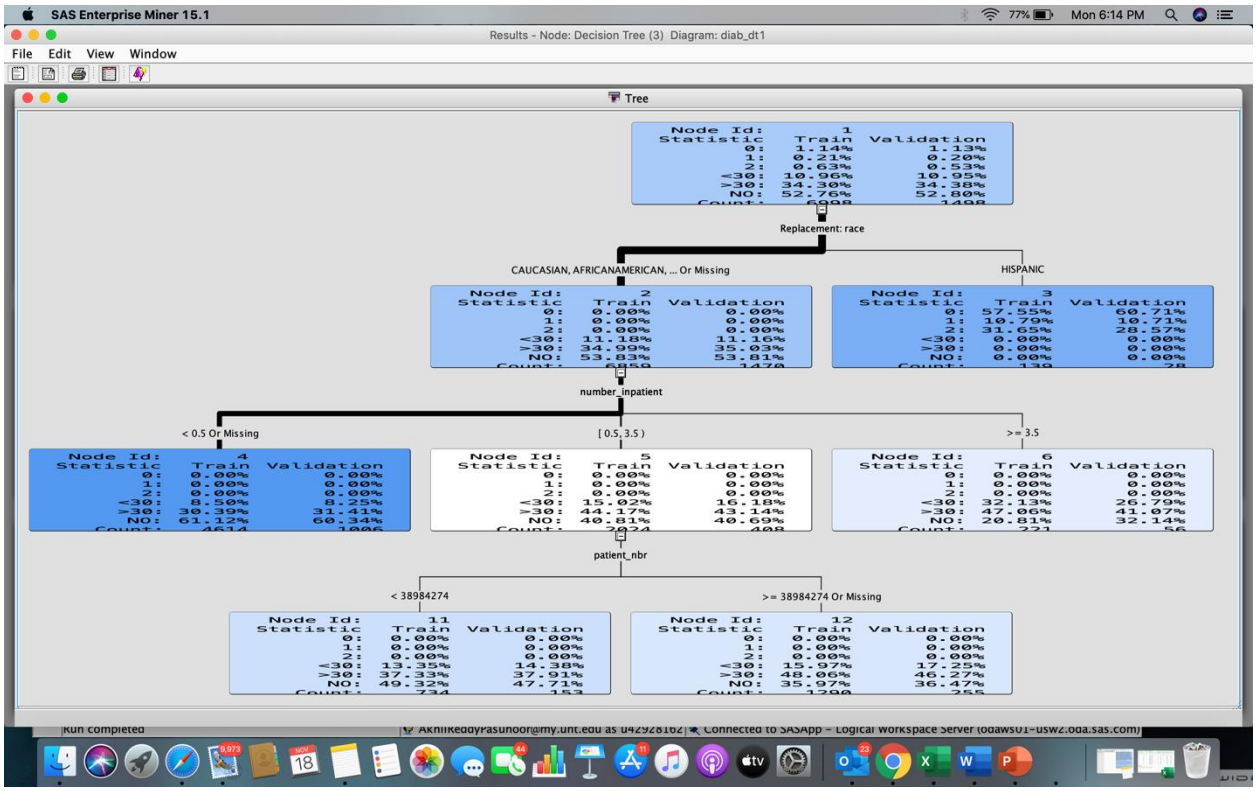


Figure 7: With no of splits =3

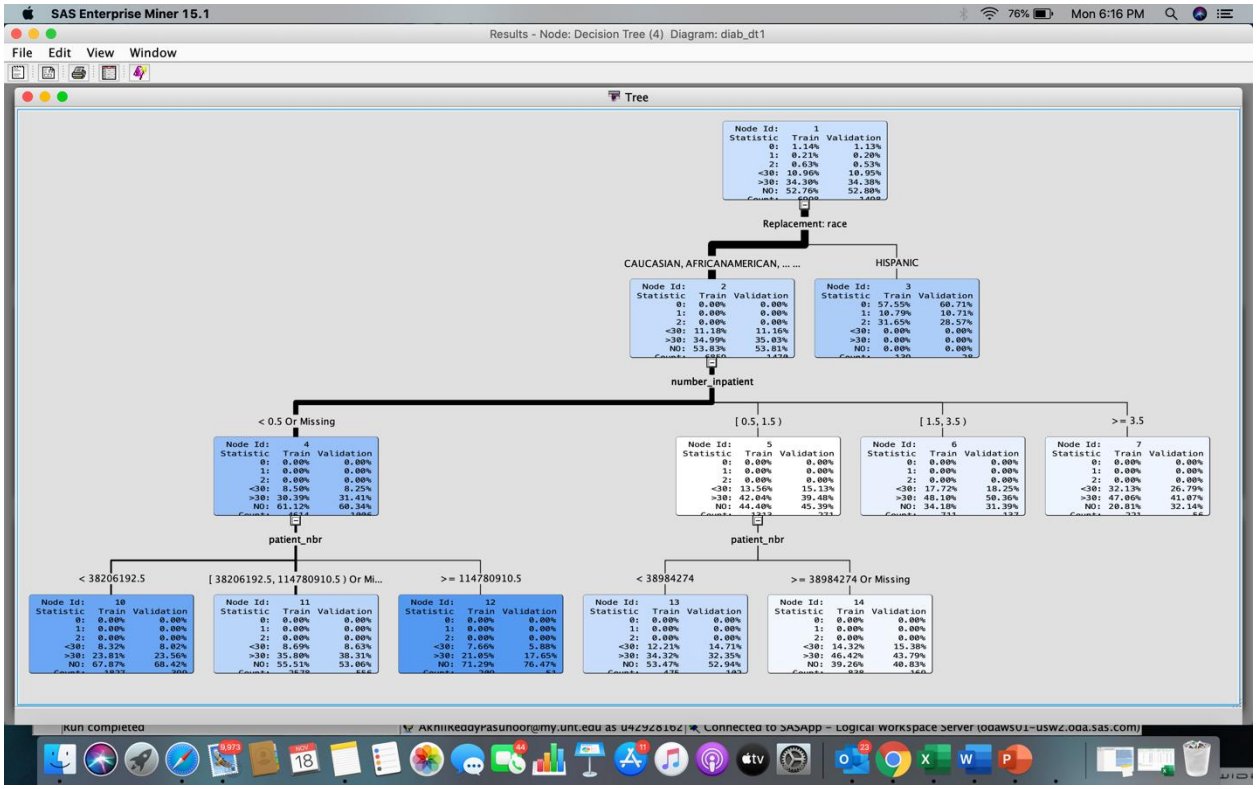


Figure 8: With no of splits=4

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.

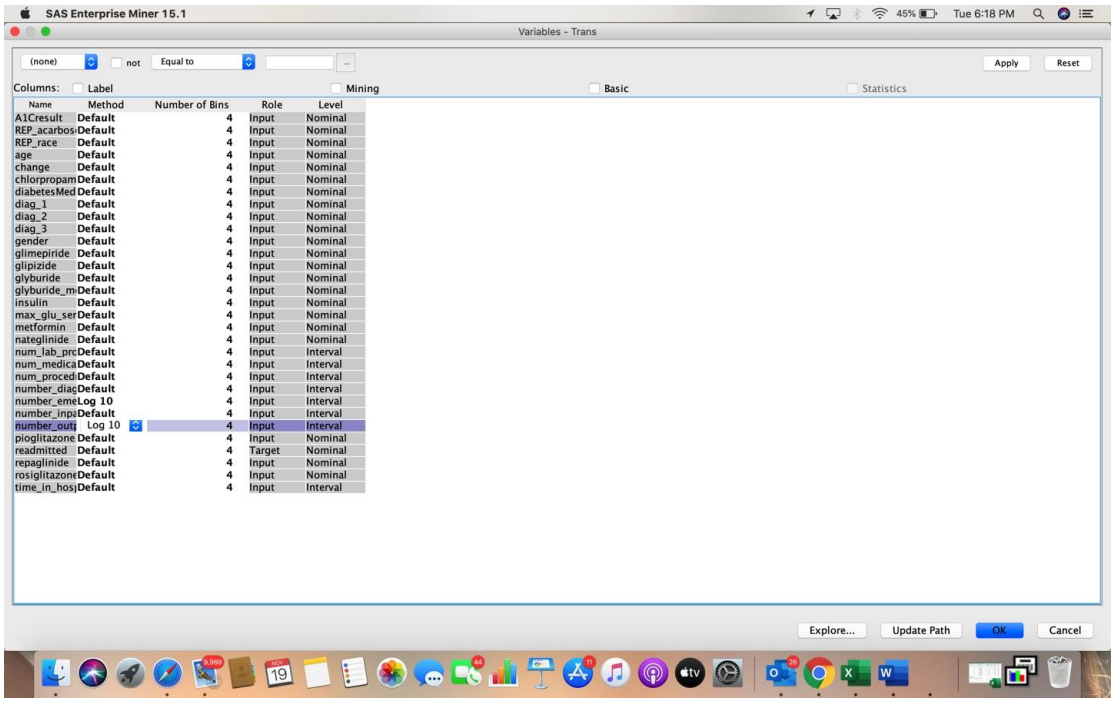
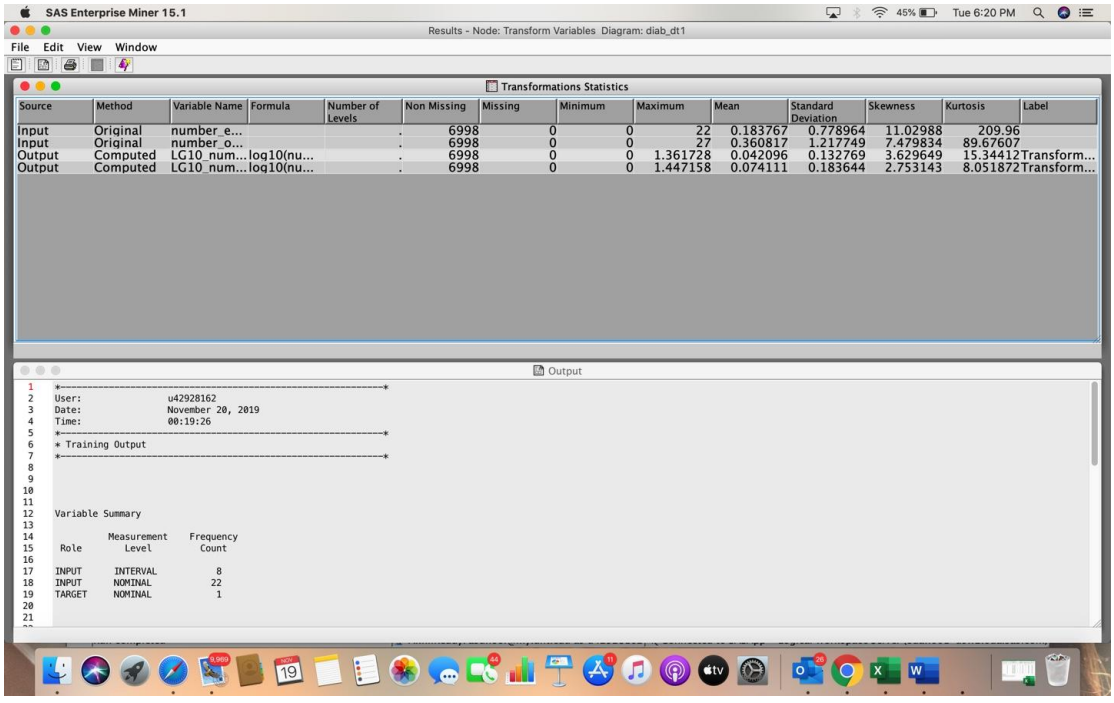


Figure 9: log10 operation

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 .

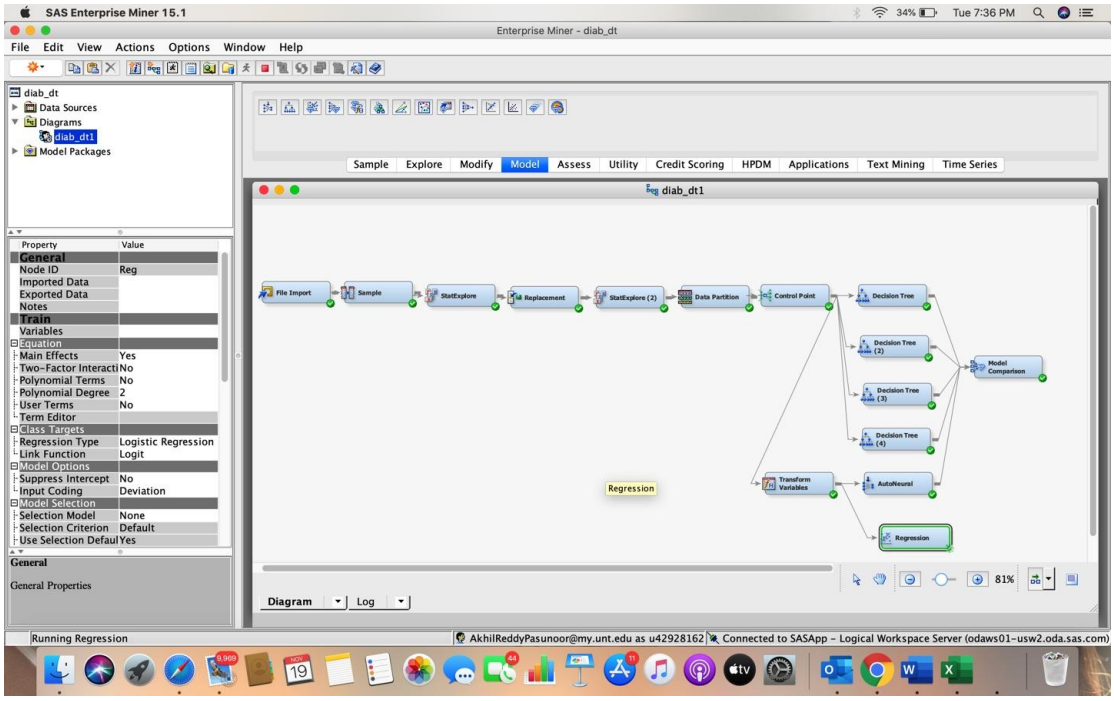


Figure 10: Logistic Regression

6.0 Findings :

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

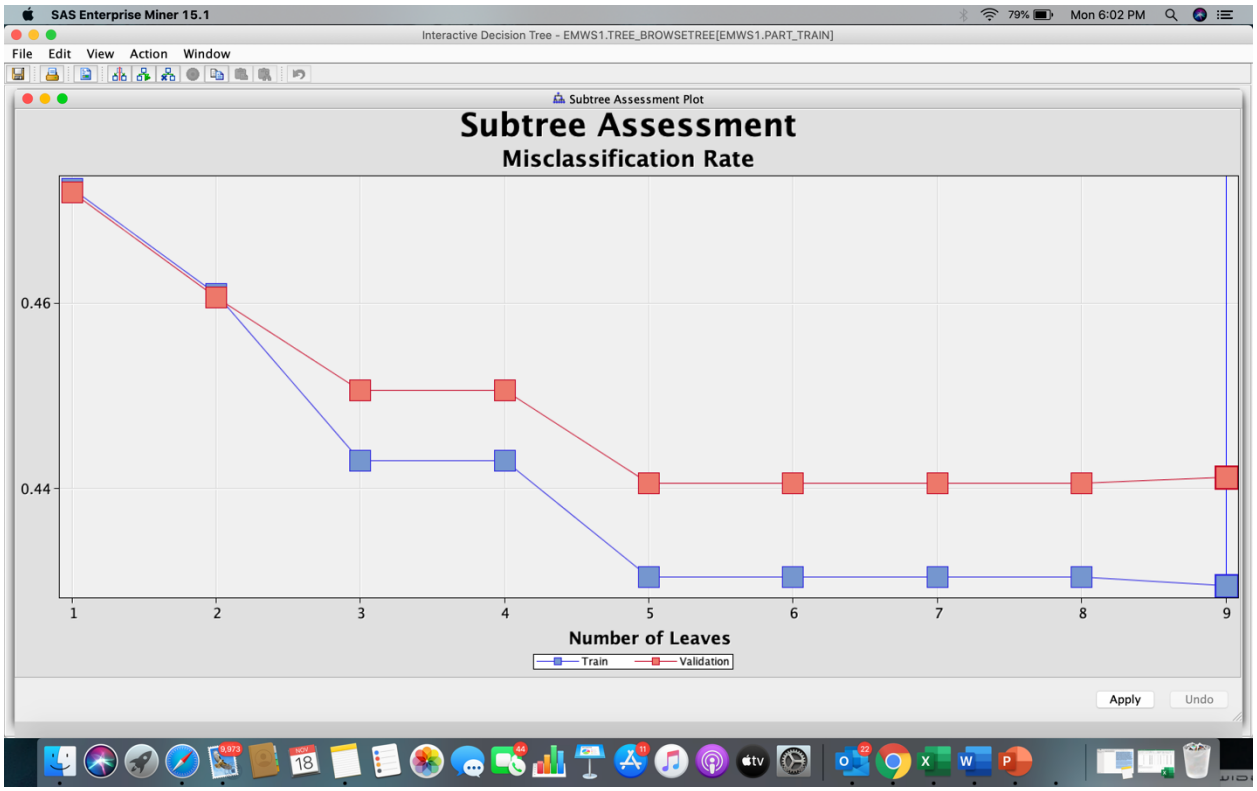


Figure 11: Subtree Assessment

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.

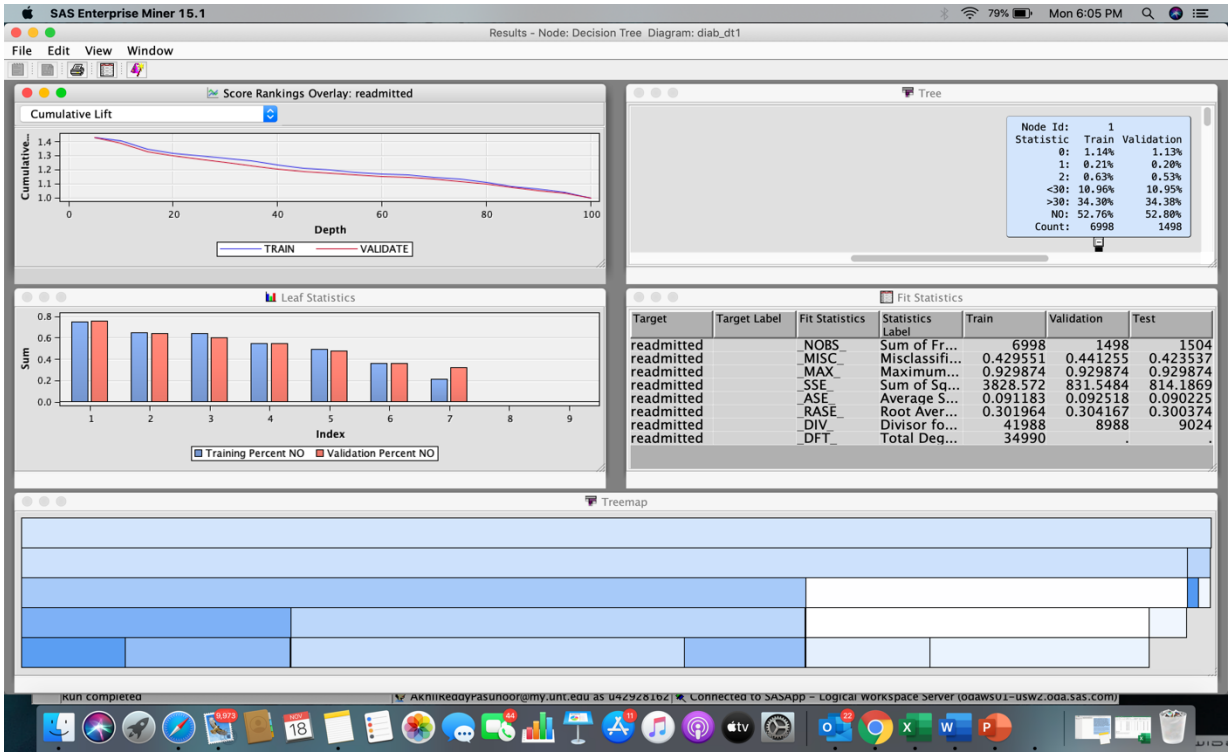


Figure 12: Output of Decision Tree model 1

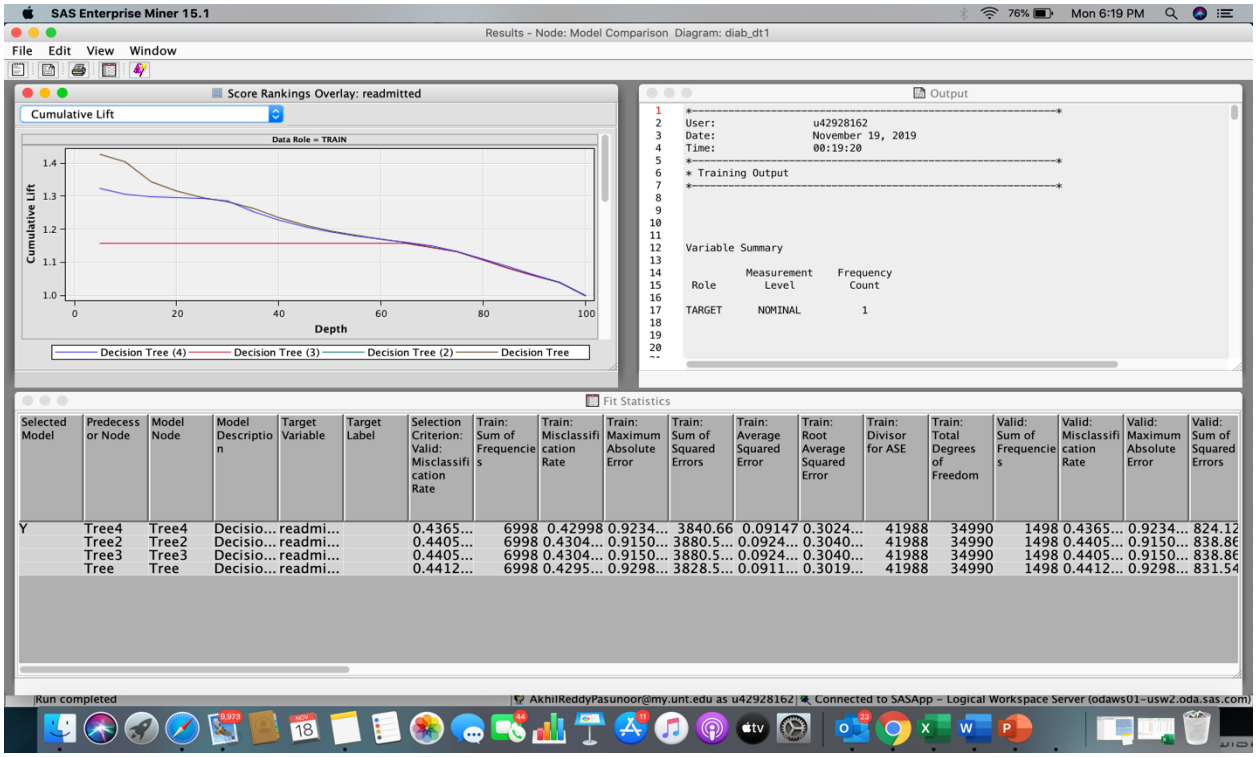


Figure 13: Comparision of 4 different models of Decision Trees

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 gradiant descent it estimates the weights which best fits the data with difference between the weighted sum and the actual output.

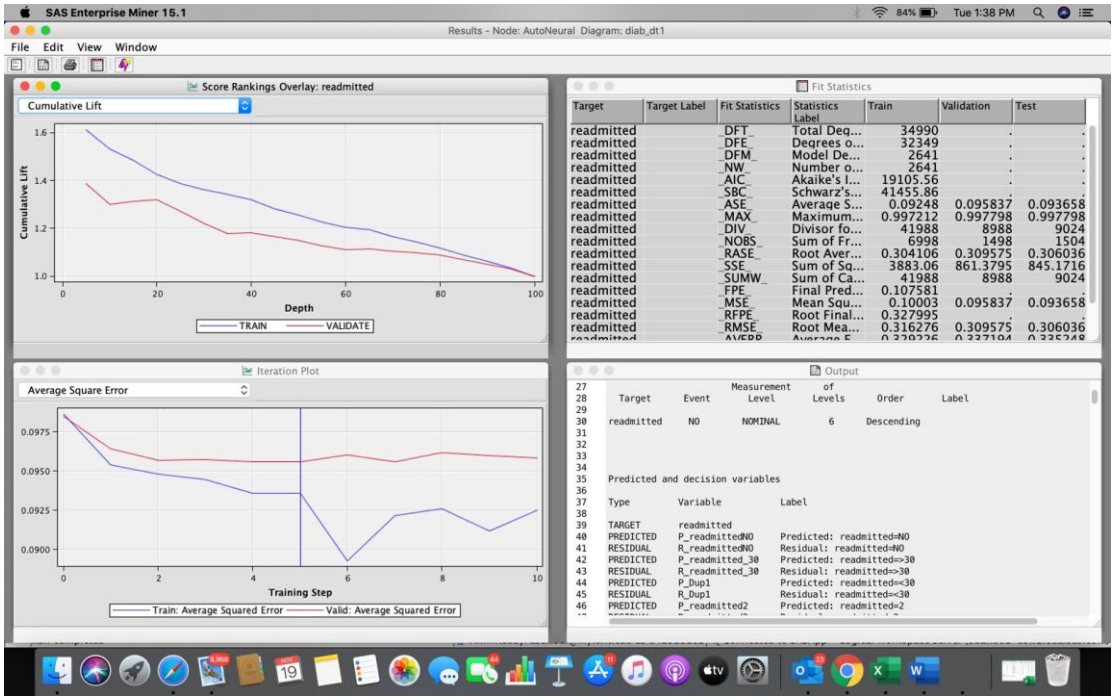
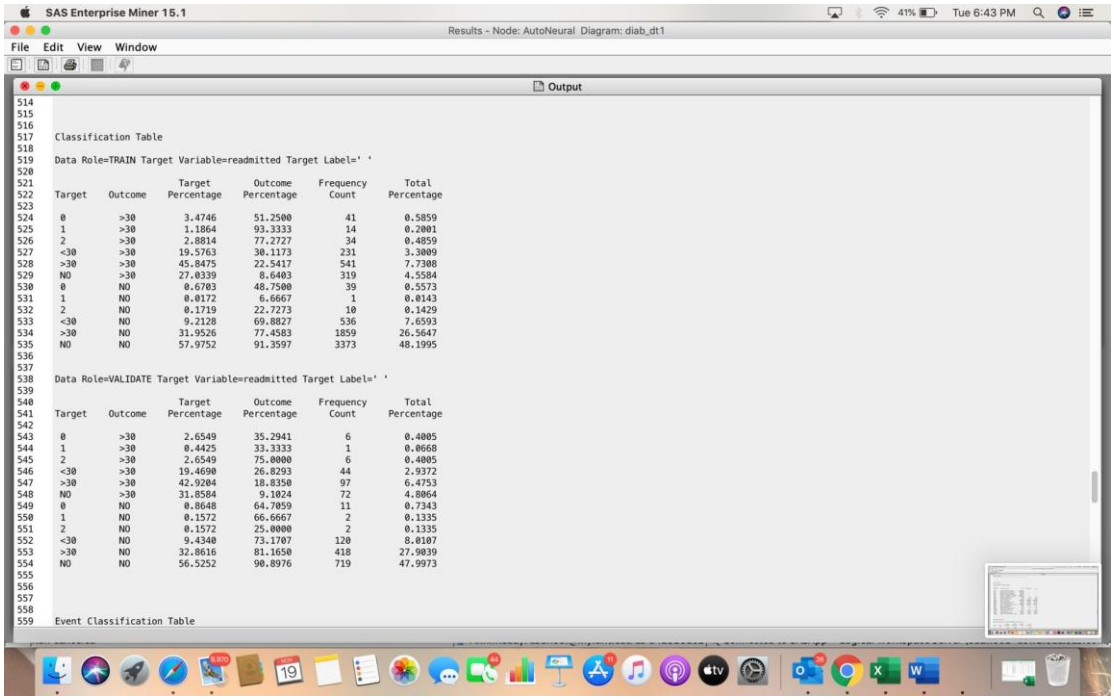
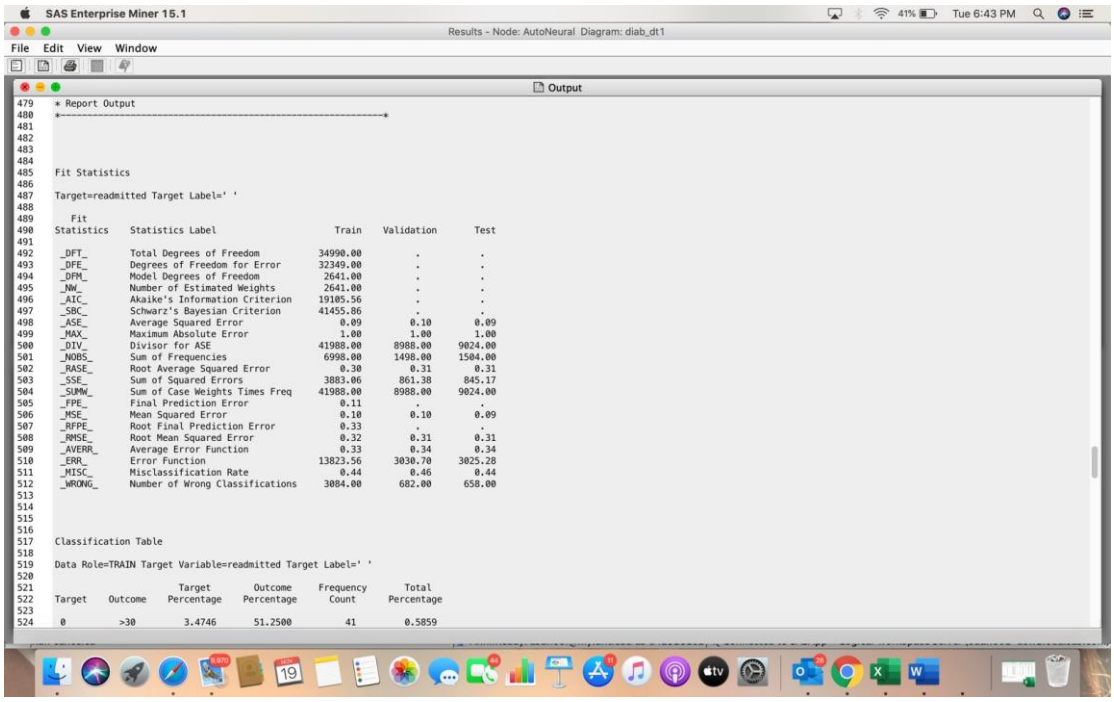


Figure 14: Output of neural networks




```
514
515
516
517  Classification Table
518
519  Data Role=TRAIN Target Variable=readmitted Target Label=' '
520
521                          Target       Outcome     Frequency       Total
522  Target    Outcome     Percentage    Percentage      Count      Percentage
523
524   0          >30         3.4746       51.2500         41         0.5859
525   1          >30         1.1864       93.3333         14         0.2001
526   2          >30         2.8814       77.2727         34         0.4859
527   <30        >30        19.5763       30.1173        231         3.3009
528   >30        >30        45.8475       22.5417        541         7.7308
529   NO         >30        27.0339        8.6403        319         4.5584
530   0          NO          0.6703       48.7500         39         0.5573
531   1          NO          0.0172        6.6667          1         0.0143
532   2          NO          0.1719       22.7273         10         0.1429
533   <30        NO          9.2128       69.8827        536         7.6593
534   >30        NO         31.9526       77.4583       1859        26.5647
535   NO         NO         57.9752       91.3597       3373        48.1995
536
537
538  Data Role=VALIDATE Target Variable=readmitted Target Label=' '
539
540                          Target       Outcome     Frequency       Total
541  Target    Outcome     Percentage    Percentage      Count      Percentage
542
543   0          >30         2.6549       35.2941          6         0.4005
544   1          >30         0.4425       33.3333          1         0.0668
545   2          >30         2.6549       75.0000          6         0.4005
546   <30        >30        19.4690       26.8293         44         2.9372
547   >30        >30        42.9204       18.8350         97         6.4753
548   NO         >30        31.8584        9.1024         72         4.8064
549   0          NO          0.8648       64.7059         11         0.7343
550   1          NO          0.1572       66.6667          2         0.1335
551   2          NO          0.1572       25.0000          2         0.1335
552   <30        NO          9.4340       73.1707        120         8.0107
553   >30        NO         32.8616       81.1650        418        27.9039
554   NO         NO         56.5252       90.8976        719        47.9973
555
556
557
558
559  Event Classification Table
```




```
479  * Report Output
480  *-------------------------------------------------------------*
481
482
483
484
485  Fit Statistics
486
487  Target=readmitted Target Label=' '
488
489     Fit
490  Statistics    Statistics Label                      Train    Validation       Test
491
492   _DFT_        Total Degrees of Freedom          34990.00          .          .
493   _DFE_        Degrees of Freedom for Error      32349.00          .          .
494   _DFM_        Model Degrees of Freedom           2641.00          .          .
495   _NW_         Number of Estimated Weights        2641.00          .          .
496   _AIC_        Akaike's Information Criterion    19105.56          .          .
497   _SBC_        Schwarz's Bayesian Criterion      41455.86          .          .
498   _ASE_        Average Squared Error                 0.09       0.10       0.09
499   _MAX_        Maximum Absolute Error                1.00       1.00       1.00
500   _DIV_        Divisor for ASE                   41988.00    8988.00    9024.00
501   _NOBS_       Sum of Frequencies                 6998.00    1498.00    1504.00
502   _RASE_       Root Average Squared Error            0.30       0.31       0.31
503   _SSE_        Sum of Squared Errors              3883.06     861.38     845.17
504   _SUMW_       Sum of Case Weights Times Freq    41988.00    8988.00    9024.00
505   _FPE_        Final Prediction Error                0.11          .          .
506   _MSE_        Mean Squared Error                    0.10       0.10       0.09
507   _RFPE_       Root Final Prediction Error           0.33          .          .
508   _RMSE_       Root Mean Squared Error               0.32       0.31       0.31
509   _AVERR_      Average Error Function                0.33       0.34       0.34
510   _ERR_        Error Function                    13823.56    3030.70    3025.28
511   _MISC_       Misclassification Rate                0.44       0.46       0.44
512   _WRONG_      Number of Wrong Classifications    3084.00     682.00     658.00
513
514
515
516
517  Classification Table
518
519  Data Role=TRAIN Target Variable=readmitted Target Label=' '
520
521                          Target       Outcome     Frequency       Total
522  Target    Outcome     Percentage    Percentage      Count      Percentage
523
524   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.

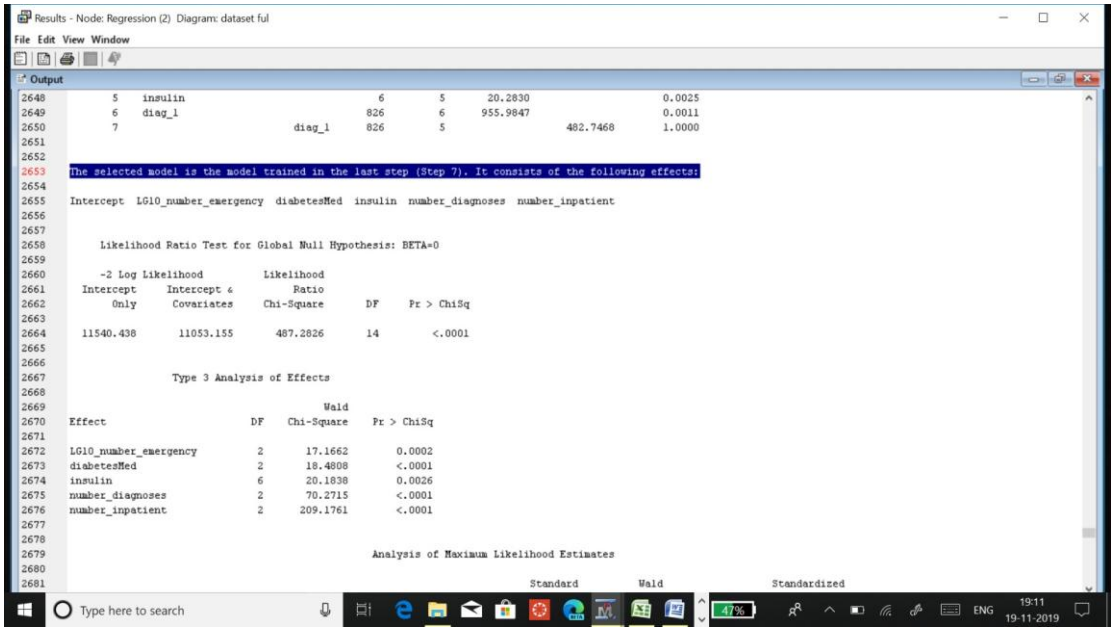


Figure 15: Output of Logistic Regression

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.

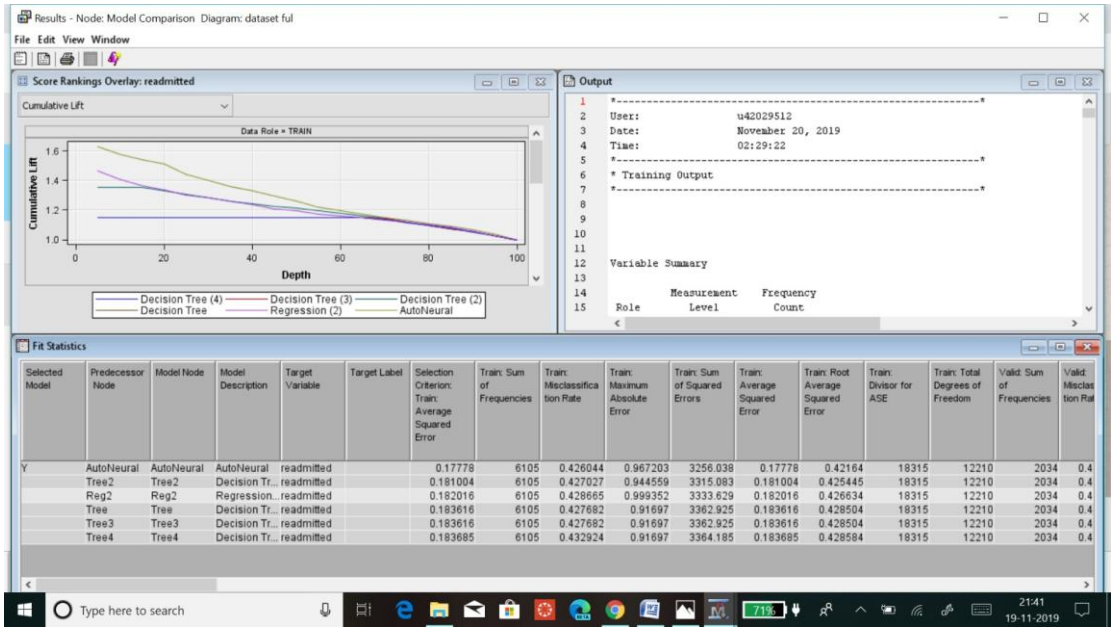


Figure 16: Output of model comparision

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